Data as Foundation: Designing Systematic Curation for an Evolving Foundation Model Landscape

Thao Nguyen

A dissertation

submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

University of Washington
2026

Reading Committee:
Sewoong Oh, Co-Chair
Ludwig Schmidt, Co-Chair
Luke Zettlemoyer

Program Authorized to Offer Degree:
Computer Science and Engineering

University of Washington

Abstract

Data as Foundation: Designing Systematic Curation for an Evolving Foundation Model Landscape

Thao Nguyen

Co-chairs of the Supervisory Committee:

Professor Sewoong Oh
Computer Science and Engineering

Assistant Professor Ludwig Schmidt
Computer Science

Foundation models have transformed the machine learning landscape with unprecedented generalization capabilities across a variety of tasks. Central to their success is the data on which they are trained, which has grown massively in scale through large web crawls and data generation efforts. Despite growing awareness of the need for data curation, current data practices remain largely heuristic and coupled with specific model and training configurations, making it difficult to isolate data-centric contributions. In this thesis, I present my work towards developing systematic, generalizable, and timely methods to optimize dataset design for foundation models. In the first work, I provided one of the earliest empirical demonstrations that indiscriminately mixing different web data sources undermines model generalization, establishing data quality as a foundational principle for large-scale curation. As the field embraced data quality and proposed increasingly aggressive filtering pipelines, I found that these methods tend to overfit to existing benchmarks and systematically exclude valuable data, such as non-English content, which can improve model performance as a whole. My subsequent work thus argues that diversity in representation should be a deliberate design decision in the curation process, instead of existing only as a byproduct. Next, moving beyond filtering as the primary curation tool, I proposed image recaptioning as a way to

transform low-quality image-text pairs into useful training data. Rather than asking what data to discard, my research instead asked what discarded data can be recovered. In the last work covered by this thesis, I extended this philosophy to the text domain. I addressed the growing scarcity of high-quality web texts by offering a sustainable approach to recycle discarded documents, effectively doubling the yield of useful pretraining tokens. Collectively, my research contributes to establishing data curation as a scientific discipline—one that is systematic, adaptive, and central to the future of foundation model development.

Acknowledgements

My PhD journey has been one of the most fulfilling times of my life. I feel deeply grateful to have been surrounded by people whose generosity, wisdom, and warmth have nurtured my intellectual and personal growth. This PhD has been full of long stretches of hard work, but looking back, I wish it could have lasted a little longer.

First and foremost, I will always be grateful to my advisors, Sewoong Oh and Ludwig Schmidt, whose guidance and support have been the foundation of my PhD. Beyond providing research advising, they have both gone above and beyond to ensure I had the resources and opportunities needed to do my best work and gain meaningful exposure to the field. They gave me the freedom to explore, try things, and fail—never exerting pressure, but always being there when I needed help, and pushing me to think critically for myself about which problems are worth pursuing. I am glad Sewoong took a chance on me even though my research interests were very different from the rest of the lab back then. He always makes time to check in—not just on research progress, but on life more broadly—and our meetings have been as much about making the PhD sustainable and enjoyable as they have been about discussing research updates. His ability to help me see things from a different perspective has been invaluable. While Sewoong's wisdom has been a quiet guide to both my personal and professional growth, Ludwig has critically shaped the directions of my research and how I approach research solutions, in ways that are hard to overstate. His openness and collaborative spirit have deeply influenced how I conduct my own work, and I am proud to have been his student. People I met at conferences have joked that my papers seem to constantly welcome new co-authors, and they are not wrong. I genuinely enjoy meeting and working with

various people, and many of those connections trace back to Ludwig. Finally, I am so grateful that both of my advisors encouraged me to pursue data curation research when I first started my PhD. What began as a suggestion has since become a genuine passion, bringing me so much joy and intellectual satisfaction. I have no doubt that it will continue to be what I want to work on long after this thesis is complete. I would also like to thank Luke Zettlemoyer for being my unofficial PhD advisor. His mentorship over the past two years has been one of the greatest unexpected gifts of my PhD. He supported me through all the hurdles, both research and logistics, that I faced during my time at Meta. Our meetings somehow managed to be genuinely fun but also quietly illuminating, filled with both giggles and wisdom—a rare combination that I treasure deeply. It has been a privilege to learn from his sharp research intuitions, the kind that can only come from someone who has spent so long at the forefront of the field.

I have also been fortunate to have many great mentors throughout my career. I am grateful to Xian Li and Jason Weston for their guidance during my time at Meta. Xian is the kind of mentor that every PhD student hopes for—generous with her time, ideas, and support in equal measure. Discussing research with Xian often left me with way more ideas than what I arrived with, and her enthusiasm was infectious. Her mentorship extended well beyond research, ensuring I had the connections and compute resources to do great research at Meta. Before my PhD, I had the opportunity to work with many great researchers at Google under the AI Residency program. I would like to thank Maithra Raghu and Simon Kornblith, whose mentorship predates my PhD but whose influence has followed me into it. Their enthusiasm for research—and for life beyond it—was genuinely energizing, and I am grateful for their wholehearted support when I decided to pursue a PhD. In addition, I would like to thank Ravi Kumar, Pasin Manurangsi and Badih Ghazi for their patience and generosity in introducing me to new research territory. Their support and encouragement during my PhD application also meant a great deal to me. My gratitude extends further back, to Pang Wei Koh and Anand Avati at Stanford, who took a chance on a clueless undergraduate with more curiosity than knowledge. They taught me the ropes of research with great patience and care, laying the foundation for everything that followed. It would also be remiss not to mention Daniel Kang and Peter Bailis, the very first people to introduce me to machine

learning research. They took me in when I was just a sophomore who barely knew anything, and I am deeply grateful for that leap of faith. I still remember when I accidentally wandered into a PhD defense thinking it was a regular lab meeting, Peter sat next to me and said "that could be you one day". I laughed it off at the time, unable to imagine myself ever reaching that stage. Yet over the course of the summer at Peter's group, after experiencing a lab culture full of curiosity, energy, and whimsy—both in and outside of research—I knew, for the first time, that I wanted to be a part of this community and keep doing machine learning research for a long time.

I am grateful to various faculty members of UW CSE—Jamie, Kevin, Pang Wei, Ranjay, Ali—who have always been so welcoming, inviting me to their group meetings and fostering a spirit of openness and collaboration beyond lab boundaries that I have deeply appreciated throughout my years at UW. Because of their kindness, the department has always felt like a warm and welcoming place, and my PhD journey feels far less lonely than it might have been otherwise.

Thank you to Joe Eckert (and other UW staff) for pulling me out of more administrative tangles than I would like to admit. Your dedication to the CSE community is the quiet force that allows graduate students to thrive here.

I would like to also thank my collaborators, whose collective insights and efforts have helped shape my research for the better. Some collaborations, e.g. the ones with Wei-Chiu and Yung-Sung, have grown into genuine friendships, making my work experience all the more enjoyable. It is a reminder that the research, at its best, is as much about the people as it is about the work.

I am thankful for my current and former labmates, whose company has made the everyday rhythms of PhD life genuinely enjoyable. Lab lunches, socials, and TGIFs with them have consistently been one of the highlights of my weeks. I would like to give a special mention to RAIVN Lab for adopting me as one of their own, when I first joined the department. The camaraderie—and chaos—they have brought into my PhD life have been an irreplaceable source of joy. Though the lab has not felt quite the same since many members graduated and moved on, I remain especially grateful to those who were there at the beginning—Aditya, Gabriel, Matt—whose guidance helped me find my footing and made those early days in the department much more manageable.

I am fortunate to have found a wonderful community of friends within the department: Weijia, Jeffrey, Jaehun, Matt, Amita, Daniel, Jon, Cheng-Yu, Chris, Avi, Melanie, Inna. They have been steadfast companions through the full spectrum of the PhD experience, from wrestling with deadlines, the inevitable research slumps, to various conference adventures. They are the kind of friends who never let the hard work crowd out the joy. I could not have asked for a better group of people to navigate this long journey with.

I am also glad to have built friendships through conferences and academic events. I often arrived not knowing a lot of people, yet somehow always left with new friends, and never once felt alone in an unfamiliar city. These unexpected connections have been one of the most cherished and unanticipated joys of my PhD. In particular, Vishaal deserves a special note of gratitude. What began as a research chat after my poster session grew into a friendship I deeply treasure. Vishaal has a remarkable ability to meet even my most half-formed thoughts and random musings with genuine curiosity and acknowledgement. Besides being an incredibly supportive friend, he is also one of the finest researchers I know, and I have learned a lot from discussing research with him.

Many thanks to Lindi, who has offered me a safe space to share both the joys and struggles of this journey. I always left our conversations with a lighter heart and a clearer mind. Her warmth and wisdom have made the hard moments feel more surmountable and the good moments feel more joyful. Whenever self-doubt got the better of me, she had a quiet way of knowing exactly what to say—and in hindsight, she was right most of the time.

Last but not least, I am eternally grateful to my friends outside of academia, whose support has been as enthusiastic as it has been unconditional: Po, Jenny, Ariel, Eric, Max, Kevin, Quynh. The details of my research may have largely eluded them, but that has never once diminished their excitement. They are always there to celebrate every milestone and hear about my PhD struggles. Their blind but wholehearted belief in me has meant more than they know. Thank you for being my constants after all these years.

DEDICATION

To my friends, family and various mentors who have believed in me.

To UW CSE, for creating a fun and supportive community that has made my PhD one of the most cherished chapters of my life.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

In recent years, foundation models have demonstrated remarkable improvements across a broad range of capabilities. What was once considered a milestone now routinely serves as a baseline, as each successive generation of models pushes the frontier of what machines can learn and do. Much of this progress has been driven by a simple but powerful recipe: training on vast quantities of data scraped from the internet, and scaling both the data and the models trained on it.

Web-crawled data—collected from billions of web pages, forums, books, and other online sources—offers an unprecedented breadth of human knowledge across various modalities (e.g., text documents and image-caption pairs), making it a key ingredient for training general-purpose models. However, this scale of data also introduces several challenges. The raw data obtained from the internet often consists of trillions of tokens, encompassing data of highly varying quality, not all of which is equally useful or safe for model training. This makes data curation—the principled selection, filtering, and composition of training data—a critical step in the model development pipeline.

Despite its importance, data curation as a scientific discipline is still in its early stages. Most frontier models do not publicly disclose their training data or the curation pipelines. This opacity makes it difficult for the broader research community to identify what constitutes effective data curation or reproduce state-of-the-art results. Consequently, this thesis contributes to advancing the science of data curation for pretraining, by addressing gaps in both methodology and understanding.

Chapter 2 primarily discusses the findings from Nguyen et al. [124]. This work is set in the early era of using web data for pretraining, when models such as GPT-2 [142] and CLIP [141] demonstrated that training on massive web-crawled corpora could yield remarkable generalization capabilities. At the time, the field was primarily focused on scaling data quantity, with relatively little attention paid to its composition. Using a comprehensive testbed of six web-scraped corpora, we were among the first to systematically study how different pretraining datasets affect model robustness under distribution shifts, and how interactions between data sources shape generalization. We find that naively mixing data sources of varying quality dilutes the robustness of the best individual source, demonstrating that simply aggregating large amounts of web data is not sufficient for building effective pretraining datasets. These findings establish data quality as a foundational consideration in dataset design, motivating the subsequent works in this thesis.

The next chapter is motivated by the rapid proliferation of multimodal data filtering methods, following increasing recognition of the importance of data quality. While these methods helped build datasets that achieve better downstream performance, they were mostly designed with English-centric benchmarks in mind. Consequently, popular curation pipelines became increasingly biased toward selecting English data and discarding potentially valuable non-English data. Chapter 3 covers work from Nguyen et al. [127], which challenges this practice. By translating and reincorporating non-English image-text pairs into the final training set, we showed that training on more data of non-English origins consistently outperforms training on English-only or English-dominated datasets. This performance gain holds not only on geographically diverse tasks, but also on standard English vision benchmarks such as ImageNet and its distribution shifts. Furthermore, we provided quantitative evidence that it is easy to separate English and non-English samples in image and text embedding spaces (even after translating both distributions to English). These findings advocate for a more inclusive curation paradigm, one that treats linguistic and cultural diversity as a key axis of pretraining data quality, rather than a byproduct of the curation pipeline.

Chapter 4 is motivated by a fundamental limitation of data filtering: while effective at removing noise, filtering often comes at the expense of data diversity. A significant portion of web-scraped image-text pairs suffer from nondescript or weakly aligned captions, which tend to be discarded by

state-of-the-art data filters. Consequently, such filters also remove a large amount of potentially useful diversity in the process. This chapter covers work from Nguyen et al. [123], where we explored synthetic captioning as a novel curation tool beyond filtering—one that actively generates higher-quality alternatives from existing imperfect ones. We made a key observation that base models generate more diverse captions than their fine-tuned counterparts (even those fine-tuned specifically for captioning capabilities), and that this diversity translates directly into better pretraining data. We further showed that filtered synthetic captions are most effective when used to complement rather than replace web captions, with the best mixing strategies outperforming the strongest filtering baselines by a significant margin on ImageNet and on average across 38 tasks. However, at large scale, the diversity of synthetic captions becomes a limiting factor; mixing in synthetic data starts to yield diminishing returns or lag behind using just filtered raw data. Together, these findings offer a nuanced view of synthetic data as a powerful but scale-dependent tool for multimodal dataset curation.

Chapter 5 carries this "transforming not discarding" principle forward into the LLM setting, where it takes on added urgency. This chapter discusses findings from Nguyen et al. [126]. A growing challenge in language model pretraining is that the supply of high-quality human-generated texts on the internet is not keeping pace with the demands of model scaling. This problem is further compounded by aggressive filtering pipelines that discard up to 99% of raw web crawls in pursuit of state-of-the-art performance. To address this token scarcity issue, we proposed rewriting all moderate-quality web documents into higher-quality alternatives that can be used for training. Experiments across 1B, 3B, and 7B parameter scales demonstrated that mixing existing high-quality web texts with our rewritten texts consistently outperforms training on just the former, and is more effective than having access to a bigger pool with double the amount of high-quality web data. Further analysis confirmed that the majority of the high-quality rewritten texts (subject to the same strict filtering process) originate from documents that would otherwise be discarded entirely. This suggests that a significant source of untapped training signals lies in the filtered-out portions of the web. Together, these results position data recycling as a promising and sustainable path forward for scaling pretraining beyond the limits of naturally occurring web data.

Beyond the works covered by this thesis, during my PhD, I also had the opportunity to contribute to benchmarking efforts that helped standardize the evaluation of data curation methods, providing the research community with a controlled setting for dataset design experimentation and comparison of popular filtering approaches [56, 98]. Furthermore, while the core contributions of this thesis focus on the pretraining regime, I have also studied data curation techniques for different post-training objectives, including robust fine-tuning [144], instruction tuning [125] and improving reasoning capabilities [103].

I conclude with a brief discussion of future work in Chapter 6. Overall, underlying my PhD research is a commitment to a more rigorous and inclusive approach to data curation science. Existing approaches tend to be exclusionary by design, discarding large portions of potentially valuable data without due consideration. My work challenges this tendency, treating diversity as a first-order principle of data curation, one that only grows more critical as models and datasets continue to scale. I explore how discarded data can be recovered, transformed, and given a second life in the pretraining set—because in large-scale data curation, what we choose to discard is just as important as what we choose to keep.

Chapter 2

The Quantity Fallacy: Quality as a First Principle in Data Curation

2.1 Overview

Web-crawled datasets have enabled remarkable generalization capabilities in recent image-text models such as CLIP (Contrastive Language-Image pretraining) or Flamingo, but little is known about the dataset creation processes. In this work, we introduce a testbed of six publicly available data sources—YFCC, LAION, Conceptual Captions, WIT, RedCaps, Shutterstock—to investigate how pretraining distributions induce robustness in CLIP. We find that the performance of the pretraining data varies substantially across distribution shifts, with no single data source dominating. Moreover, we systematically study the interactions between these data sources and find that combining multiple sources does not necessarily yield better models, but rather dilutes the robustness of the best individual data source. We complement our empirical findings with theoretical insights from a simple setting, where combining the training data also results in diluted robustness. In addition, our theoretical model provides a candidate explanation for the success of the CLIP-based data filtering technique recently employed in the LAION dataset. Overall our results demonstrate that simply gathering a large amount of data from the web is not the most effective way to build a pretraining

dataset for robust generalization, necessitating further study into dataset design. Code is available at https://github.com/mlfoundations/clip_quality_not_quantity.

2.2 Introduction

Large models pretrained on web-scale datasets have become a cornerstone of machine learning. For instance, the past two years have witnessed the arrival of several new models such as GPT-3 [23], Chinchilla [80], and PaLM [33] for natural language processing, or CLIP [141], BASIC [136], and Flamingo [7] for computer vision. These models exhibit unprecedented generalization capabilities in zero-shot inference, in-context learning, and robustness to distribution shift. A key ingredient enabling their generalization performance are the large and diverse pretraining corpora that exceed previous datasets by multiple orders of magnitude. For instance, the training set of BASIC [136] contains 6.6 billion images, which is more than 1,000 times larger than the widely used ImageNet [40] training set from the 2012 competition containing 1.2 million images.

Despite the central role datasets play for pretrained models, little is known about them, especially for image-text models. The aforementioned CLIP [141], BASIC [136], and Flamingo [7] all rely on datasets internal to the respective organizations, which is also the case for other models such as DALL-E [146], Florence [205], and ALIGN [85]. In addition, research publications often provide little details on the data collection processes, e.g., the data sources or data filtering mechanisms. Beyond clear issues such as reproducibility and the potential presence of harmful content, the opaque dataset creation practices also make it hard to identify effective methods for assembling pretraining datasets. As a result, researchers cannot build on each other's dataset innovations, which obstructs the incremental research process that has successfully accumulated algorithm and architecture improvements in machine learning models. A more principled understanding of dataset creation will likely enable further progress in the generalization capabilities of pretrained models.

A basic approach to dataset creation would be to simply train on *all* available data of a given type. While scaling up training sets has indeed been integral to the recent progress in large models, advances in weak and self-supervision [29, 72, 71, 23, 43] have led to an abundance of potential

training data. For instance, the large LAION-5B dataset [160] of 5 billion image-text pairs is itself a subset of about 50 billion images from Common Crawl. This abundance is already exceeding the amount of data models can currently be trained on within a reasonable time,[1] making the aforementioned baseline of using all available data infeasible. Consequently, deciding *what* data to train on is becoming increasingly important.

In this paper, we take a step towards a better understanding of pretraining data and investigate the impact of dataset design on the generalization capabilities of image-text models. Specifically, we focus on CLIP, the first large model of this kind that demonstrated remarkable robustness to multiple challenging distribution shifts. The main questions we ask are: (*i*) How much do different web data sources vary in their induced robustness? (*ii*) Do dataset combinations lead to better robustness? (*iii*) Can filtering with an existing image-text model improve data quality? We address these questions from both an experimental and theoretical perspective.

We begin by assembling a corpus of six different datasets from the web, spanning a variety of sources including Flickr, Shutterstock, Wikipedia, Common Crawl, and Reddit. We then measure the robustness of CLIP models trained on each dataset. In particular, we compare the zero-shot accuracy of these models on ImageNet and a set of canonical ImageNet distribution shifts. We find that the robustness induced by each pretraining dataset varies widely, and that sources with careful curation such as Wikipedia do not necessarily outperform those with minimal filtering.

In addition, the datasets cannot be compared along a single dimension: different datasets help with robustness to different distribution shifts. This in turn motivates studying how combining datasets affects robustness. We find that a model trained on two datasets does not inherit the robustness properties of both. Rather, while the model is exposed to both distributions, its robustness interpolates between that of the individual datasets. This indicates that dataset designers must carefully combine the pretraining data in order to preserve and enhance the robustness of the resulting models.

[1]Recall that the training set for Google's largest image-text model ALIGN contains "only" 6.6 billion images, making it about eight times smaller than the source dataset for LAION-5B.

Building on our empirical results, we introduce a theoretical model to better understand our experimental findings. Our model is simple enough to allow for mathematical analysis and can still capture some phenomena of real-world, web-crawled datasets. Specifically, our theoretical model also shows that combining multiple datasets dilutes the robustness of the better data distribution. Moreover, our model provides a candidate explanation for another interesting phenomenon in the curation of pretraining data: the LAION-400M experiment [162] and follow-up CLIP reproductions [160] demonstrated that filtering a noisy source (Common Crawl) with a CLIP model results in a dataset on which newly trained CLIP models exhibit *higher* robustness to some distribution shifts.

In the following sections, we first briefly review related work and the relevant background on robustness to distribution shift (Section 2.3), before discussing our experimental setup in Section 4.4. Sections 2.5 and 2.6 then measure the robustness of CLIP induced by individual data sources and their combinations. To support these empirical results, Section 2.7 presents our theoretical analysis. We conclude with future research directions in Section 2.8.

2.3 Background & Related Work

Vision-language models. Large vision-language models like CLIP and ALIGN have become an active area of research owing to their success on various computer vision tasks [141, 85]. Existing work expands their capabilities by either increasing model and dataset size [136], or by using additional supervision as in DeCLIP [104], SLIP [120], and FILIP [199]. In contrast, we study the effect of *pretraining data composition* on task performance with a focus on robustness.

In a recent work, Fang et al. [51] showed that CLIP's robustness primarily stems from its diverse pretraining distribution and not training set size, language supervision, or contrastive loss functions. However, Fang et al. [51] only conducted experiments with two datasets (YFCC-15M and ImageNet-Captions), one of which contains less than one million images. We take the insights of Fang et al. [51] as our starting point and expand the range of pretraining datasets to six different sources, each containing at least five million images. This enables us to study the robustness induced by various pretraining sources, and how dataset combinations affect robustness.

Distribution shift. Robustness to distribution shift is a long-standing issue in machine learning [179, 139] and has recently received renewed attention as researchers scrutinize the generalization performance of neural networks in greater detail [173, 19, 18, 11, 156, 69, 37, 94]. Similar to CLIP [141], we focus on robustness to natural distribution shifts, where the corresponding test sets contain only unmodified images and are not intentionally perturbed by adversarial examples or synthetic corruptions (such as Gaussian noise or blur patterns) [77]. Specifically, we test robustness on ImageNetV2 [147], ImageNet-R [75], ImageNet-Sketch [186], and ObjectNet [14] because prior work has established many baselines for these distribution shifts [175]. Moreover, models robust to these shifts also show improved robustness on other out-of-distribution benchmarks such as WILDS [92, 194].

The robustness literature usually discusses distribution shift in terms of "in-distribution" and "out-of-distribution" data. This terminology is natural when data from the same distribution as the "in-distribution" test set is used for training or fine-tuning. However, it is unclear what counts as "in-distribution" when models are trained on large-scale, generic pretraining datasets that aim to improve performance on a wide variety of tasks. To address this issue, we follow the more flexible definition of distribution shift employed in prior work [175, 118] and measure robustness as accuracy difference between two related but distinct test distributions (e.g., ImageNet and ImageNet-Sketch). The expectation is that the shift between the two test distributions should not affect an ideal robust model, for instance because the shift does not affect the accuracy of humans labelers [165]. Taori et al. [175] defined this notion of robustness as *effective robustness*. Radford et al. [141] adopted the effective robustness framework in the evaluations of their CLIP model. Effective robustness is illustrated in Figure 2.1 and measures movement towards a perfectly robust classifier which is not affected by the shift between two test distributions.

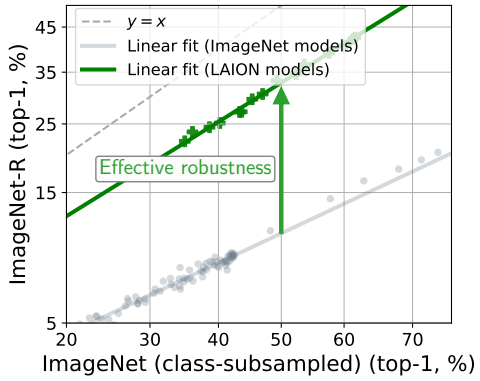


Figure 2.1: Models pretrained on LAION exhibit *effective robustness* [175] compared to models trained on ImageNet. Effective robustness is defined as movement towards a classifier that is robust to distribution shift. A classifier is more robust the closer it is to the $y = x$ line; any classifier on this line is not affected by the distribution shift.

Prior work [118] has evaluated several hundred models and demonstrated experimentally that changes to model architecture, training set size, training algorithm, and other model-related factors do *not* change effective robustness in most cases. In contrast, changes to the pretraining *distribution* can improve the effective robustness of a model. This makes effective robustness a useful metric for evaluating the influence of pretraining data sources on robust generalization, since it removes confounders stemming from model hyperparameters and the number of training samples. The effect of the pretraining distribution on effective robustness is particularly pronounced when models are evaluated in a zero-shot setting: while training on a target distribution can improve accuracy on that distribution [194, 60, 208, 212, 193], this process can deteriorate robustness to distribution shift [141, 194, 10, 136].

The existence of a universal linear trend for accuracies on a pair of test sets has been analytically studied in [118]. For a simple binary classification model similar to our Assumption 1 (see Section 2.7), convergence to a linear trend is shown with the deviation from the line scaling as $O(1/\sqrt{d})$. However, the analysis in [118] is restricted to a specific stochastic distribution shift for the out-of-distribution test data and a fixed classifier independent of the training data. Hence the dependence of the slope on the training set size, variations in the training methods, and properties of the training distribution cannot be described in their model. We provide significantly more fine-grained analyses (Theorem 1) that captures all such tradeoffs, which allows us to draw novel insights into how data mixing (Theorem 2) and filtering (Theorem 3) affect model robustness.

2.4 Experiment Setup

Model. We focus on the CLIP model [141] which has demonstrated unprecedented zero-shot performance on a wide range of downstream tasks, as well as robustness to various distribution shifts [118]. Given an image-text pair, CLIP is trained to maximize the cosine similarity between the embedding of the text and that of the image, relative to the similarity of unconnected image-text pairs. We use the CLIP implementation from the OpenCLIP GitHub repository [83], with ResNet-50 [73] as the image encoder architecture. We vary the pretraining set size and hyperparameters such

as number of epochs to obtain different accuracies on each data distribution. Due to compute constraints, the total training set size is at most 15M samples for most of our experiments. Appendix D.1 contains further training details.

Data. To study the effects of training distributions on robust generalization, we collect several datasets from publicly available sources. Most of these have been studied in the context of various vision and language tasks in previous work:

- YFCC: We experiment with the 15M subset of the YFCC100M dataset [178] that the original CLIP paper [141] used for dataset ablation studies. The images and captions are collected from Flickr.
- LAION [162]: The images and corresponding alt-texts come from web pages collected by Common Crawl [1] between 2014 and 2021. We randomly select a subset of 15M samples to experiment with, and ensure that the accompanying NSFW tags of all chosen images are 'UNLIKELY'.
- Conceptual Captions [27]: We use CC-12M for our experiments, which consists of images and HTML alt-text from an unspecified set of web pages.
- RedCaps [42]: This dataset contains 12M examples, obtained from 350 manually curated subreddits between 2008 and 2020. The subreddits are selected to contain a large number of image posts that are mostly photographs and not images of people.
- Shutterstock: 15M images and captions were crawled from the Shutterstock website in 2021.
- WIT [171]: Image-text pairs come from Wikipedia pages. We use reference description as the source of text data and obtain 5M examples in total by selecting for only English language examples.

Appendix A.1.1 contains an analysis of image and text statistics, as well as randomly selected data samples from each source.

Evaluation. Similar to Taori et al. [175] and Radford et al. [141], we choose ImageNet as the reference distribution and evaluate CLIP on four natural distribution shifts derived from ImageNet:

- ImageNet-V2 [147]: A reproduction of the ImageNet validation set closely following the original dataset creation process.
- ImageNet-R [75]: Renditions (e.g., sculptures, paintings, etc.) for 200 ImageNet classes.
- ImageNet-Sketch [186]: Sketches of ImageNet class objects.
- ObjectNet [14]: A test set of objects in novel backgrounds, rotations, and viewpoints with 113 classes overlapping with ImageNet

Visualizations of random samples from each distribution shift can be found in Appendix A.1.2.

2.5 Individual Pretraining Data Sources

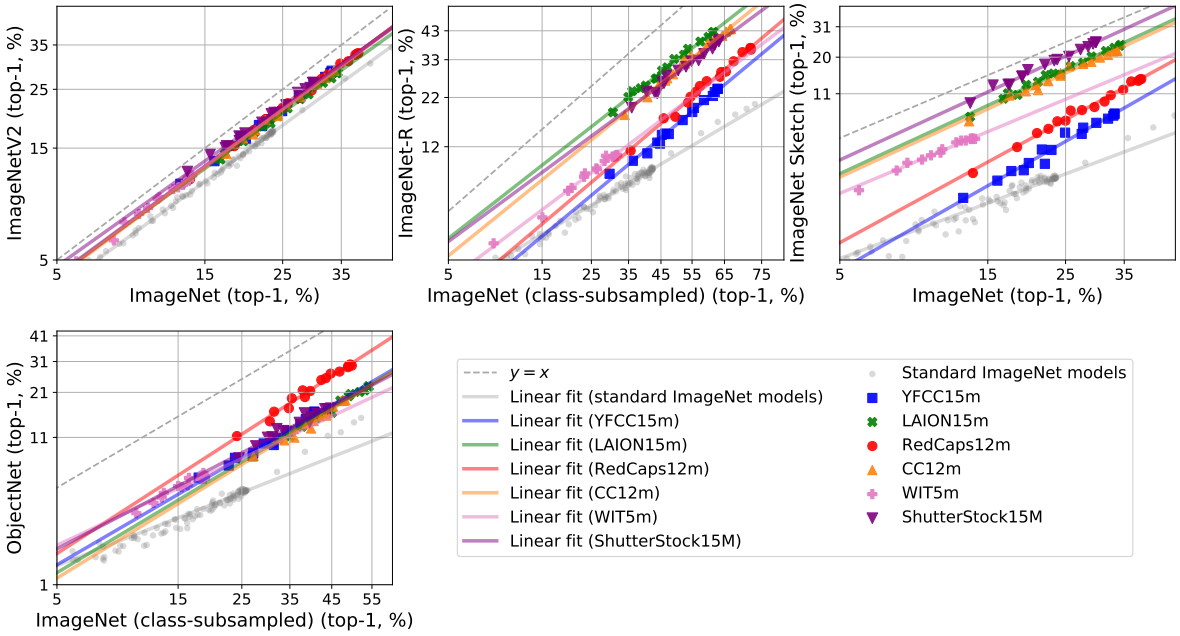


Figure 2.2: Performance of the six pretraining data sources under various distribution shifts. We find that the behavior—both in terms of accuracy and the slope of the linear trend—of the pretraining data varies substantially across distribution shifts, with no single data source dominating. Most shifts help highlight the strengths and weaknesses of different data sources, except for ImageNet-V2, where the linear trends produced by individual sources are highly correlated with one another.

We first investigate how well a CLIP model trained on each data source would perform under different distribution shifts of interest. As seen from Figure 2.2, while all sources yield the same linear trend

on ImageNet-V2, some display clear advantages when evaluated on other test distributions. For example, Shutterstock offers the best out-of-distribution performance on ImageNet-Sketch, while RedCaps displays the most effective robustness on ObjectNet. On ImageNet-R, LAION, CC12M and Shutterstock seem to do much better than the rest. Overall no pretraining data distribution is consistently the most robust across all evaluation settings.

We also measure the data efficiency of each source, i.e., how much the performance would change with more samples from the same source, see Appendix A.3. Similar to the previous observation, the six pretraining sources display vastly different data efficiency depending on the distribution shifts of interest. Although LAION and CC-12M exhibit similar effective robustness in all evaluation settings in Figure 2.2, this analysis reveals subtle differences between these two training distributions in the low-data regimes.

2.6 Combining Data Sources

In the previous section, we demonstrate the variability in behavior of different data sources based on the distribution shift at test time. A natural question then arises from this observation: does combining multiple sources help improve robustness across some, if not all, test distributions of interest? We investigate two common approaches of aggregating information from different training sets—input mixing and output ensembling. In the subsequent discussion, we focus on distribution shifts that bring out significant differences in behavior across the pretraining data sources (e.g., ImageNet-R and ImageNet-Sketch); the full results on all distribution shifts can be found in Appendices A.4 and A.5.

2.6.1 Input Mixing

In input mixing, we randomly select and combine samples from multiple data sources to build the pretraining dataset. Our findings indicate that this exposure to more training distributions doesn't help CLIP take advantage of the complimentary strengths of each source. Rather, the effective

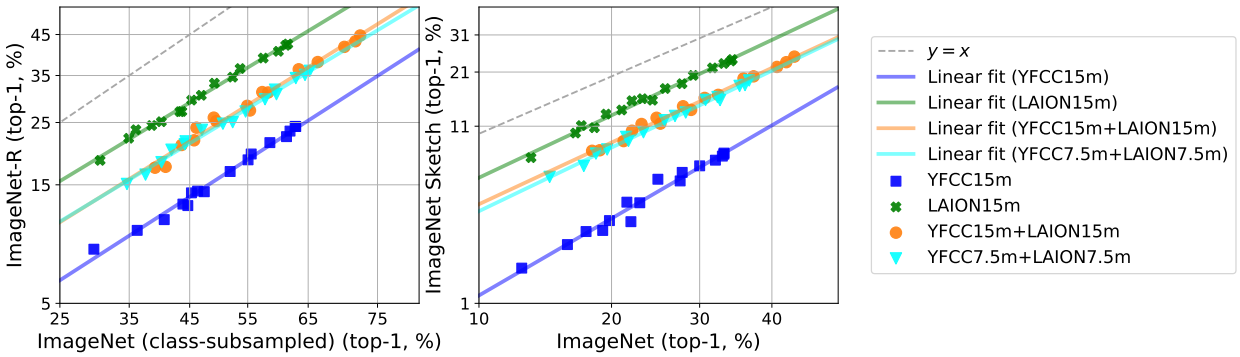


Figure 2.3: Combining YFCC and LAION training data in equal ratios produces models with intermediate robustness. Given a fixed data budget of 15M samples, the linear trend produced by training CLIP on a YFCC-LAION data mixture, with 7.5M datapoints from each source (cyan line), lies between that of training CLIP on YFCC (blue line) and LAION (green line) entirely. Even when we increase the total training set size (30M) and use all data available from both sources (orange line), the same pattern persists.

robustness of the resulting model is less than that of training on the best individual source for each distribution shift.

We start with combining data from two largest data sources in the testbed—YFCC and LAION. To remove dataset size as a confounder, we fix the total amount of training data at 15M samples, and sample 7.5M image-text pairs from each source. This data mixture yields a linear trend that lies in between the trends obtained from training on 15M YFCC and 15M LAION datapoints separately (Figure 2.3). Even when we remove the constraint on the training set size and use all the data available from these two sources (i.e., 30M samples), the same observation on robustness holds, and the resulting linear trend is highly correlated with that of training on the YFCC-7.5M + LAION-7.5M mixture.

The previous set of experiments combines YFCC and LAION data with a 50:50 ratio. In Figure 2.4, we find that when this ratio is varied within a fixed budget of 15M datapoints, the robustness linear trends of the corresponding mixtures form a smooth interpolation between the linear trends of training on 15M YFCC and 15M LAION samples separately. Experiments with different combinations of sources, as well as mixtures of a larger number of sources, can be found in Appendix A.4. There we also include results for combining data from CIFAR-10 and CINIC-10 distributions

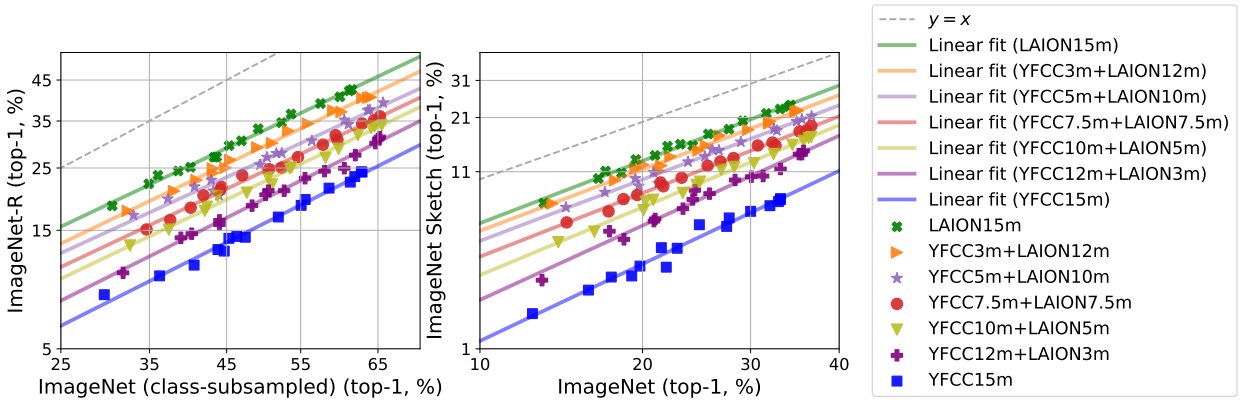


Figure 2.4: Varying the sample contributions of YFCC and LAION to the input mixture yields a smooth interpolation of the linear trend between those of training on YFCC and LAION separately. Keeping the total number of training samples fixed at 15M, as we vary the contribution of YFCC to the final dataset from 15M (i.e., only training on YFCC) to 0M (i.e., only training on LAION), the resulting linear trend gradually shifts from that of YFCC-15M (blue line) to that of LAION-15M (green line).

to train ResNets on image classification tasks, where we observe the same interpolation pattern. We provide a theoretical justification for this phenomenon in Section 2.7.2.

2.6.2 Output Mixing

Another common approach to take advantage of different training sets is to combine them at model output level (i.e., ensemble) [45, 96, 66, 53, 16, 22, 54, 129]. Here, we train a CLIP model on each pretraining distribution of interest and combine the logit predictions of all the resulting models with equal weights. In experiments with mixing YFCC and LAION (Figure 2.5), the ensembles of 2 single-source-trained CLIP models taken from the same epoch of training, lie on the linear trend of a *single* CLIP model trained on the combined data. This holds across different distribution shifts that we consider. The same observation also applies when we ensemble 6 CLIP models trained separately on the 6 data sources collected, with the contribution of each model being weighted equally. Refer to Appendix A.5 for more details.

We next show that this phenomenon extends to more complex mixtures. The fact that output mixing (i.e., ensembling the predictions of CLIPs trained on individual sources) is predictive of the linear trend produced by input mixing (i.e., training on *all* the data from these sources) presents an

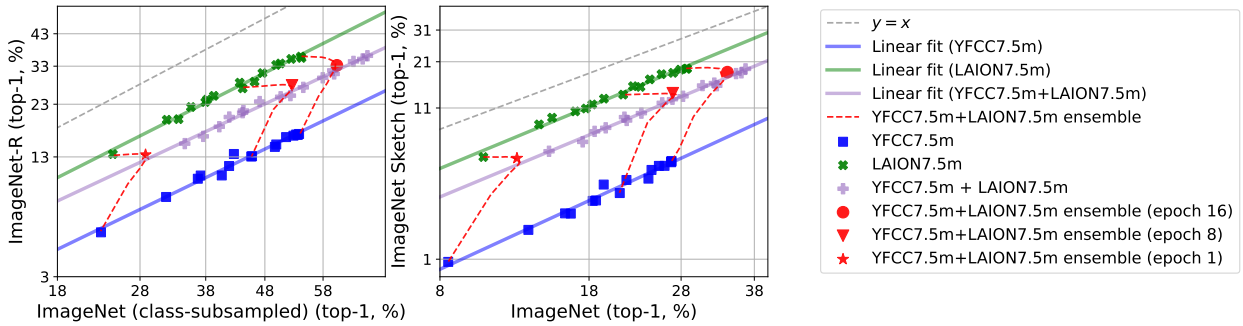


Figure 2.5: Ensemble outputs of CLIP models trained on YFCC and LAION separately share the same linear trend as a single model trained on the combined data mixture (where each source contributed equally). We ensemble the logit predictions of YFCC-trained (blue line) and LAION-trained (green line) models taken from the same epoch, with varying ensemble weights between 0 and 1 (red dashed line). When the outputs are combined with equal weights (red markers), the resulting test accuracies closely track the linear trend produced by pretraining CLIP on a data mixture with equal number of samples from each source (purple line).

opportunity to estimate model robustness given new data sources, without having to train the model from scratch on the new, potentially much larger, combined dataset. For example, in Figure 2.6, assuming access to a CLIP model trained on the YFCC-5M + LAION-5M mixture, and another one trained on CC-5M, the ensembles of these two models across different stages of training share the same linear trend as a CLIP model trained on the YFCC-5M + LAION-5M + CC-5M mixture. In Appendix A.5, we show the results of ensembling 6 CLIP models trained with the 6 data sources we collected, as well as the CINIC-10 + CIFAR-10 ensemble, a uni-modal image classification setting where output mixing accuracies are also predictive of the linear trend of input mixing.

2.7 Analysis under Simple Binary Classification Models

Empirically we observe, for e.g., in Figure 2.2, that for a pair of test datasets $(\mathcal{D}_1, \mathcal{D}_2)$, the corresponding test accuracies after probit transform achieved by various trained models (including different architectures and training algorithms) on the same training dataset $\mathcal{D}$ all lie on the same line. Furthermore, this line includes models trained on only a subset of $\mathcal{D}$ [118]. In other words, there is a universal line that is determined only by the training data distribution and the two test distributions, and is independent of which architecture we use, how we train the model, or how many samples we use from the training set. To explain this phenomenon, we study the following class

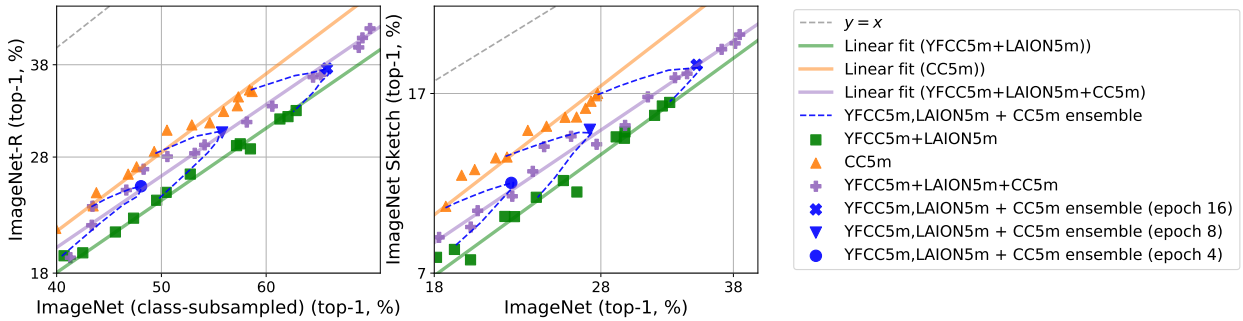


Figure 2.6: Using ensemble outputs to predict the linear trend of input mixing without retraining CLIP from scratch. A generalization of the observation made in Figure 2.5 is that given an existing pretraining dataset that could be a mixture (e.g., YFCC-5M + LAION-5M, green line) and a new data source (e.g., CC-5M, orange line), we could use the ensemble outputs (blue markers) of two CLIP models that have been trained separately on these two data distributions, to estimate where the linear trend for a CLIP model trained on *all* the data would lie (purple line). This removes the need to actually train CLIP from scratch on the now bigger 3-source mixture.

of trained models parametrized by $(\theta \in \mathbb{R}^d, \rho \in \mathbb{R}_+)$ representing the training distribution, with n representing the training data size, and $\xi \in \mathbb{R}_+$ representing the training algorithm variations.

2.7.1 Universality of Accuracy on the Line for Binary Classification

We analyze a simple but canonical binary classification example where each training data is parametrized by its ground-truth linear classifier $\theta \in \mathbb{R}^d$ and its Signal-to-Noise Ratio (SNR) $\rho^2 \in \mathbb{R}_+$. Concretely, a training dataset $\mathcal{D}_{n,\theta,\rho} = \{(x_i \in \mathbb{R}^d, y_i \in \pm 1)\}_{i=1}^n$ is a set of n i.i.d. paired samples from a joint distribution $P_{\theta,\rho}$ defined as follows.

Assumption 1 (Data distribution)**.** *We define a joint distribution* $(x_i, y_i) \sim P_{\theta,\rho}$ *as* (i) $y_i = \pm 1$ *uniformly at random, and* (ii) $x_i = y_i\theta + (\|\theta\|/\rho)z_i$ *where the noise* z_i *is zero-mean and has independent entries with variance one. For an observation* (x_i, y_i)*, we refer to* $\|\theta\|^2$ *as the signal power and* $\|\theta\|^2/\rho^2$ *as the corresponding noise power with SNR* ρ^2*.*

Assumption 2 (Trained model distribution)**.** *We consider a (random) linear model parameter* $\hat{\theta}_{n,\xi} \in \mathbb{R}^d$ *that predicts a binary label* $\text{sign}(\langle x, \hat{\theta}_{n,\xi}\rangle)$ *for a test example* $x \in \mathbb{R}^d$*. We assume that the random model* $\hat{\theta}_{n,\xi} = \theta + (\xi\|\theta\|/(\rho\sqrt{n}))z \in \mathbb{R}^d$*, trained on* $\mathcal{D}_{n,\theta,\rho}$ *is unbiased,* $\mathbb{E}[\hat{\theta}_{n,\xi}] = \theta$*, and that* z *has independent entries with variance one each. The randomness comes from the training data*

as well as any internal randomness in the training algorithm. The parameter $\xi \in \mathbb{R}_+$ captures the variations in the resulting model distribution due to changes the training algorithm.

Concretely, one canonical example of a trained model is $\hat{\theta}_{n,\xi} = (1/n)\sum_{i=1}^n y_i x_i$. It follows that $\hat{\theta}_n = \theta + (\|\theta\|/(\rho\sqrt{n}))z$ with $\xi = 1$. Hence, ξ measures the randomness of the trained model relative to this simple training algorithm.

We analyze the resulting accuracy when evaluated on two test distributions P_{θ_1,ρ_1} and P_{θ_2,ρ_2} as defined above, with $\mathrm{Acc}_{\theta_1,\rho_1} := P_{\theta_1,\rho_1}\{\mathrm{sign}(\langle X, \hat{\theta}_{n,\xi}\rangle) = Y\}$, and $\mathrm{Acc}_{\theta_2,\rho_2}$ defined similarly. In particular, we are interested in how the accuracy pair $(\Phi^{-1}(\mathrm{Acc}_{\theta_1,\rho_1}), \Phi^{-1}(\mathrm{Acc}_{\theta_2,\rho_2}))$ behaves as we vary the sample size n and as we vary the training algorithm represented by ξ. Here, $\Phi^{-1} : [0,1] \to \mathbb{R}$ is the inverse of the CDF of a standard Gaussian distribution, defined as $\Phi(t) = \mathbb{P}(z \le t)$ where $z \sim \mathcal{N}(0,1)$. Φ^{-1} is also called the probit function. This choice of mapping the accuracy with the probit function is critical in getting the linear relation, which we will explain in Remark 1.

Theorem 1 (Universality of accuracy on the line)**.** *Under Assumptions 1 and 2, asymptotically as the dimension d grows linearly in the sample size n such that $\lim_{d\to\infty} n/d = \alpha^2$, we have*

$$\lim_{d\to\infty} \mathrm{Acc}_{\theta_1,\rho_1} = \Phi\Big(\cos(\theta_1,\theta)\frac{\rho_1\rho\alpha}{\xi}\Big), \quad \textit{and} \quad \lim_{d\to\infty} \mathrm{Acc}_{\theta_2,\rho_2} = \Phi\Big(\cos(\theta_2,\theta)\frac{\rho_2\rho\alpha}{\xi}\Big). \tag{2.1}$$

Further, under Assumption 5 in the Appendix, for some universal constant $c > 0$,

$$\Phi^{-1}(\mathrm{Acc}_{\theta_2,\rho_2}) = \frac{\cos(\theta_2,\theta)\rho_2}{\cos(\theta_1,\theta)\rho_1}\Phi^{-1}(\mathrm{Acc}_{\theta_1,\rho_1}) + O\Big(\frac{e^{\frac{cn}{d}}}{\sqrt{n}}\Big). \tag{2.2}$$

We provide a proof in Appendix A.6.1. This analysis implies that for any training sample size n and any (variation due to the) training algorithm ξ, the resulting accuracy pair after probit transform lies on a *universal* line determined by Eq. (2.2), and the slope only depends on the two test distributions $(P_{\theta_1,\rho_1}, P_{\theta_2,\rho_2})$ and the training distribution $P_{\theta,\rho}$. The similarity between the training data and each test dataset is captured by the angles: $\cos(\theta_1,\theta)$ and $\cos(\theta_2,\theta)$. More similar training data yields a higher test accuracy. The hardness of the test distribution is captured by the SNR of each test dataset: ρ_1 and ρ_2. We note that this line is universal in the sense that the slope does not depend

on the sample size n and training algorithm parameter ξ. We are interested in the regime where n scales linearly in d and both are large such that the second term in the above equation is negligible.

Remark 1. Why do we get the accuracy-on-the-line phenomenon? The prediction of the trained (random) linear model on a random test data point X involves the inner product, which performs a natural spatial averaging over the d coordinates. Since the noise across the coordinates is independent and has bounded variance, the central limit theorem applies. The resulting error has a Gaussian tail. Hence, the probit mapping, $\Phi^{-1}(\text{Acc}_{\theta_1\rho_1})$, is critical in translating accuracy into the relevant mean to standard deviation ratio: $\cos(\theta_1, \theta)\rho_1\rho\sqrt{n}/(\xi\sqrt{d})$. This is consequently important for getting the universal linear relation, because irrelevant parameters, n and ξ, cancel out in the slope of $\Phi^{-1}(\text{Acc}_{\theta_2\rho_2})/\Phi^{-1}(\text{Acc}_{\theta_1\rho_1})$. Refer to the proof of Theorem 1 (Appendix A.6.1) for more details.

Remark 2. Intersection at the random guess: Previous work [118] that considered models with a wide range of accuracies has found that all lines intersect at a point corresponding to "random guess", which in this binary example is $(\Phi^{-1}(1/2), \Phi^{-1}(1/2)) = (0, 0)$. Our theoretical analysis is consistent with this empirical observation: all universal lines intersect at $(0, 0)$ up to a small additive error scaling as $O(1/\sqrt{n})$.

2.7.2 Input Mixing Yields an Intermediate Slope

Figures 2.3 and 2.4 show that when a model is trained on samples combined from two datasets, the resulting robustness trend lies in between the trends achieved by individual datasets. We show that this finding is universally true under our current setup in Theorem 2. Note that for the linear models we consider, input mixing (combining training sets) and output mixing (combining model outputs) are equivalent.

Assumption 3. *Consider two training datasets $\mathcal{D}_{n_1,\tilde{\theta}_1,\tilde{\rho}_1}$ and $\mathcal{D}_{n_2,\tilde{\theta}_2,\tilde{\rho}_2}$ of sizes n_1 and n_2 from distributions $P_{\tilde{\theta}_1,\tilde{\rho}_1}$ and $P_{\tilde{\theta}_2,\tilde{\rho}_2}$ as defined in Assumption 1. Separately training on individual datasets gives two models $\hat{\theta}_{n_1,\xi_1}(\mathcal{D}_{n_1,\tilde{\theta}_1,\tilde{\rho}_1})$ and $\hat{\theta}_{n_2,\xi_2}(\mathcal{D}_{n_2,\tilde{\theta}_2,\tilde{\rho}_2})$ as defined in Assumption 2. A model trained on a combined (and possibly subsampled) training dataset $\mathcal{D}_{n_1',\tilde{\theta}_1,\tilde{\rho}_1} \cup \mathcal{D}_{n_2',\tilde{\theta}_2,\tilde{\rho}_2}$ is represented by $\hat{\theta}_{n_1'+n_2',\xi}(\mathcal{D}_{n_1',\tilde{\theta}_2,\tilde{\rho}_2} \cup \mathcal{D}_{n_2'\tilde{\theta}_2,\tilde{\rho}_2}) = \bar{\theta} + (\xi\|\bar{\theta}\|/(\bar{\rho}\sqrt{n_1' + n_2'}))z$ where $\bar{\theta} = (n_1'\tilde{\theta}_1 + n_2'\tilde{\theta}_2)/(n_1' + n_2')$, $\bar{\rho} =$*

$\|\bar{\theta}\|/\sqrt{(n_1'\|\tilde{\theta}_1\|^2/\rho_1^2)+(n_2'\|\tilde{\theta}_2\|^2/\rho_2^2)}$, *and* $z \in \mathbb{R}^d$ *is a zero-mean random vector with independent entries each with variance one.*

Again, a canonical example of a trained model is $\hat{\theta}_{n_1'+n_2',\xi} = (1/(n_1'+n_2'))\big(\sum_{(x_i,y_i)\in\mathcal{D}_{n_1',\tilde{\theta}_1,\tilde{\rho}_1}} y_i x_i + \sum_{(x_i,y_i)\in\mathcal{D}_{n_2',\tilde{\theta}_2,\tilde{\rho}_2}} y_i x_i\big)$. We then have $\hat{\theta}_{n_1'+n_2',\xi} = \bar{\theta} + (\|\bar{\theta}\|/(\bar{\rho}\sqrt{n_1'+n_2'}))z$ with $\xi = 1$.
It follows from applying Theorem 1 to the model trained on the combined data (Assumption 3) that the resulting linear trend achieves

$$\text{Slope}(\hat{\theta}_{n_1'+n_2',\xi}(\mathcal{D}_{n_1',\tilde{\theta}_1,\tilde{\rho}_1} \cup \mathcal{D}_{n_2',\tilde{\theta}_2,\tilde{\rho}_2})) \;:=\; \frac{\cos(\theta_2,\bar{\theta})\rho_2}{\cos(\theta_1,\bar{\theta})\rho_1}\,, \tag{2.3}$$

We show that the slope obtained from training on the combined dataset lies between those obtained from training on the two datasets separately. The proof could be found in Appendix A.6.2. Let $\text{Slope}_1 := \text{Slope}(\hat{\theta}_{n_1}(\mathcal{D}_{n_1,\tilde{\theta}_1,\tilde{\rho}_1}))$ and Slope_2 be defined similarly.

Theorem 2. *Under Assumption 3,* $\text{Slope}_1 \le \text{Slope}\big(\hat{\theta}_{n_1'+n_2'}(\mathcal{D}_{n_1',\tilde{\theta}_1,\tilde{\rho}_1} \cup \mathcal{D}_{n_2',\tilde{\theta}_2,\tilde{\rho}_2})\big) \le \text{Slope}_2$.

2.7.3 Filtering Data to Improve Robustness

The input mixing analysis in Section 2.7.2 suggests a filtering strategy to improve robustness.

Assumption 4 (Gaussian distributions for filtering)**.** *Consider a training dataset* $\mathcal{D}_{n,\theta_{\text{train}},\rho}$ *of size* n *draw i.i.d. from* $(X,Y) \sim P_{\theta_{\text{train}},\rho}$ *where* $Y = \pm 1$ *uniformly at random and* $X|Y \sim \mathcal{N}(Y\theta_{\text{train}}, (\|\theta_{\text{train}}\|/\rho)^2\mathbf{I})$. *A model trained on an unfiltered* $\mathcal{D}_{n,\theta_{\text{train}},\rho}$ *is denoted by* $\hat{\theta}_{\text{unfiltered}} = (1/n)\sum_{(x_i,y_i)\in\mathcal{D}_{n,\theta_{\text{train}},\rho}} x_i y_i$. *Note that* $\hat{\theta}_{\text{unfiltered}} \sim \mathcal{N}(\theta_{\text{train}}, (\|\theta_{\text{train}}\|/(\rho\sqrt{n}))^2\mathbf{I})$.

The model is evaluated on two test datasets: in-distribution (ID) and out-of-distribution (OOD), each with an isotropic Gaussian noise: $(X,Y) \sim P_{\theta_{\text{ID}},\rho_{\text{ID}}}$, *where* $Y = \pm 1$ *uniformly at random and* $X|Y \sim \mathcal{N}(Y\theta_{\text{ID}}, (\|\theta_{\text{ID}}\|/\rho_{\text{ID}})^2\mathbf{I})$, *and* $P_{\theta_{\text{OOD}},\rho_{\text{OOD}}}$ *is defined similarly. Let* $\text{Slope}(\theta) := \frac{\cos(\theta_{\text{OOD}},\theta)\rho_{\text{OOD}}}{\cos(\theta_{\text{ID}},\theta)\rho_{\text{ID}}}$ *and assume that all models achieve better ID accuracy than OOD accuracy, i.e.,* $\text{Slope}(\theta) < 1$. *A model is more robust if the slope is closer to one.*

Suppose we have access to a pretrained model $\hat{\theta}_{\text{pretrained}}$ that achieves better robustness than the model $\hat{\theta}_{\text{unfiltered}}$ trained on the unfiltered data, i.e., $\text{Slope}(\hat{\theta}_{\text{unfiltered}}) < \text{Slope}(\hat{\theta}_{\text{pretrained}}) \le 1$. We want to use the pretrained model to filter the data $\mathcal{D}_{n,\theta_{\text{train}},\rho}$ and achieve better robustness. We

consider a generic family of filtering strategies that subsamples each data point $(x_i, y_i) \in \mathcal{D}_{n,\theta_{\text{train}},\rho}$ with probability that is monotonically non-decreasing with its correlation to the pretrained model, i.e., $\mathbb{P}((x_i, y_i)$ passes the filter$) \propto h(\langle x_i y_i, \hat{\theta}_{\text{pretrained}} \rangle)$ for some monotonic scalar function $h : \mathbb{R} \to \mathbb{R}_+$. We analyze the model $\hat{\theta}_{\text{filtered}} = (1/m) \sum_{(x_i,y_i) \in \text{filtered dataset}} x_i y_i$. The next theorem shows that any such filtering strategy improves robustness, and the full proof can be found in Appendix A.6.3.

Theorem 3. *Under Assumption 4, if* $\text{Slope}(\hat{\theta}_{\text{unfiltered}}) < \text{Slope}(\hat{\theta}_{\text{pretrained}}) \leq 1$ *then any model* $\hat{\theta}_{\text{filtered}}$ *that is trained on the above family of filtered datasets achieves better robustness than the model trained on unfiltered data:* $\text{Slope}(\hat{\theta}_{\text{unfiltered}}) < \text{Slope}(\mathbb{E}[\hat{\theta}_{\text{filtered}}]) \leq \text{Slope}(\hat{\theta}_{\text{pretrained}})$*, where* $\text{Slope}(\mathbb{E}[\hat{\theta}_{\text{filtered}}])$ *denotes the slope of the linear trend of a model trained on the filtered data.*

Remark 3. Relevance to empirical practice: The data filtering approach currently employed for the LAION dataset [162] motivates our theoretical setup. Schuhmann et al. [162] first retrieved a large number of image-text pairs from Common Crawl [1] and then computed the cosine similarity between the image and text embeddings using one of the original CLIP models introduced by Radford et al. [141]. Next, the authors removed all pairs with similarity less than 0.3 from the pool. While it was initially unclear if filtering the noisy Common Crawl source with an existing CLIP model would lead to a dataset with good generalization properties, experiments with pretraining medium-scale models on LAION so far demonstrate that the resulting models are comparable to the original CLIP models [160].

Our theorem 3 does not provide a full justification for the filtering process described above, but it offers theoretical evidence that filtering a noisy data source with an existing robust model can indeed improve the effective robustness of the resulting dataset. It remains to be shown empirically that the slope of the linear trend exhibited by the CLIP model from [141] is larger than that of a model trained on the Common Crawl data pool without filtering. If this assumption holds, we can expect that training on data with better text-image alignment according to the original CLIP model, will improve robustness compared to training on unfiltered data from Common Crawl.

2.8 Discussion

Motivated by CLIP's robustness to distribution shift, we introduced a testbed of six image-text datasets to study the influence of pretraining data on the robustness of multimodal models such as CLIP. We analyzed the interactions of these datasets through both experiments and theoretical analyses, finding that simply combining multiple datasets dilutes the robustness of the best-performing one. Moreover, we offered an explanation for the potential benefits of the data filtering process currently used for LAION. Our findings suggest that training on a carefully selected subset of data may provide more robustness than training on simply more data, for example, obtained by combining multiple data sources.

To summarize, our results lead to the following recommendations for dataset design:

- On the evaluation front: building a pretraining dataset requires multiple robustness test sets, as we have shown that different pretraining datasets exhibit substantially different behaviors across distribution shifts.
- On the dataset design front: each candidate data source for the overall pretraining dataset should be evaluated separately for its generalization and robustness properties. The final pretraining dataset should then contain more samples from the higher-quality pretraining sources. Creating a more "diverse" pretraining dataset by randomly sampling from all candidate sources does not result in a better dataset.
- Assuming access to models separately trained on each source of interest, output ensemble is a good predictor of the effective robustness obtained from input mixing, and hence can significantly reduce the time to search for the right mixing strategy.
- Filtering a noisy data source with an existing robust model can improve the generalization properties of the resulting dataset.

The broad societal implications of image-text models and web-crawled multimodal datasets have been extensively discussed by Radford et al. [141] and Birhane et al. [20] respectively. As these models continue to grow in size, full-scale experiments are becoming prohibitively expensive. This necessitates small, systematic experiments that allow reliable extrapolation to larger scales. Since

effective robustness is independent of model size and quantity of training data, we hope that our testbed and results can serve as a first step towards building multimodal pretraining datasets in a principled manner.

Limitations The pretraining datasets we experiment with range from 5M to 15M samples. While this lags behind current state-of-the-art settings that large image-text models are often trained on, we believe our findings also hold for models with higher accuracy for the following reasons:

- The original CLIP paper [141] offers evidence showing that CLIP's performance exhibits reliable linear trends across compute scales (Figure 9 of [141], which includes the ResNet50-encoder that our paper works with) and performance on natural distribution shifts (Figure 14 of [141]). In particular, the linear trends in this Figure 14 are analogous to those in our scatter plots.
- In the OpenCLIP GitHub repository [83] that seeks to match the performance of the original CLIP paper, the authors there also note that their models that were trained on 15M samples or fewer share the same effective robustness trend as OpenAI's CLIP models (trained on 400M samples). Refer to their Why are low-accuracy CLIP models interesting? remark for more details.
- Previous work [118] has experimented with training on randomly sampled subsets of the training dataset to produce models covering a wide range of accuracies, and found that these models lie on the same effective robustness trend. This includes models trained on only 1% of the original data, which matches the scale of our experiments (5M-15M samples) relative to the scale of the original CLIP paper (400M samples).

The focus of our work is to study the behavior of a diverse set of data sources, and performing full-scale experiments on 6 datasets (and their combinations) would be prohibitively expensive. To provide some perspectives on the amount of resources involved in full-scale experiments: (i) One of the larger datasets that are publicly available, LAION-5B, is a substantial effort that involves 14 people actively working on it [160], (ii) Training a ResNet50-encoder CLIP on the full dataset of 400M samples took 18 days on 592 V100 GPUs, according to the original paper [141]. This prompts us to leverage publicly available web-crawled datasets, and study the behavior of our trained models through the effective robustness framework, which has been shown to be scale-invariant.

Chapter 3

The Exclusion Problem: Rethinking What Popular Data Filters Leave Behind

3.1 Overview

Massive web-crawled image-text datasets lay the foundation for recent progress in multimodal learning. These datasets are designed with the goal of training a model to do well on standard computer vision benchmarks, many of which, however, have been shown to be English-centric (e.g., ImageNet). Consequently, existing data curation techniques gravitate towards using predominantly English image-text pairs and discard many potentially useful non-English samples. Our work questions this practice. Multilingual data is inherently enriching not only because it provides a gateway to learn about culturally salient concepts, but also because it depicts common concepts differently from monolingual data. We thus conduct a systematic study to explore the performance benefits of using more samples of non-English origins with respect to English vision tasks. By translating all multilingual image-text pairs from a raw web crawl to English and re-filtering them, we increase the prevalence of (translated) multilingual data in the resulting training set. Pretraining on

this dataset outperforms using English-only or English-dominated datasets on ImageNet, ImageNet distribution shifts, image-English-text retrieval and on average across 38 tasks from the DataComp benchmark. On a geographically diverse task like GeoDE, we also observe improvements across all regions, with the biggest gain coming from Africa. In addition, we quantitatively show that English and non-English data are significantly different in both image and (translated) text space. We hope that our findings motivate future work to be more intentional about including multicultural and multilingual data, not just when non-English or geographically diverse tasks are involved, but to enhance model capabilities at large. All translated captions and metadata are available at https://huggingface.co/datasets/thaottn/datacomp-medium-pool-translated.

3.2 Introduction

Today, the predominant pretraining paradigm for vision-language models relies on large quantities of image-text pairs scraped from the web [141, 99]. As raw web data contains a significant amount of noise, automatic data filtering approaches are designed to curate a high-quality subset and maximize the performance of a model trained on this subset on standard computer vision benchmarks (e.g., ImageNet). However, these benchmarks typically only evaluate in English, and many of them have been shown to be geographically biased: for instance, ImageNet images are mostly sourced from North America and Western Europe [164]. Consequently, it is possible that we are designing data curation algorithms that propagate a monolingual bias, i.e., filtered datasets are increasingly dominated by English image-text pairs. In fact, a lot of highly cited work—including CLIP [141], ALIGN [85] and BASIC [135]—relies exclusively on English data. Using more multilingual data for training is often only a deliberate design decision when non-English tasks are involved [184, 63, 109].

Multilingual data enriches any monolingual data distribution; multilingual data brings attention to culturally salient concepts and introduces new perspectives and annotations for the same visual category [200]. As illustrated in Figure 4.1, there are certain native concepts, e.g. 'kiji' (the national bird of Japan), that are more likely to be conveyed in Japanese (non-English) captions compared to English ones. Even in the case of a common everyday object like 'stove', the non-English and English

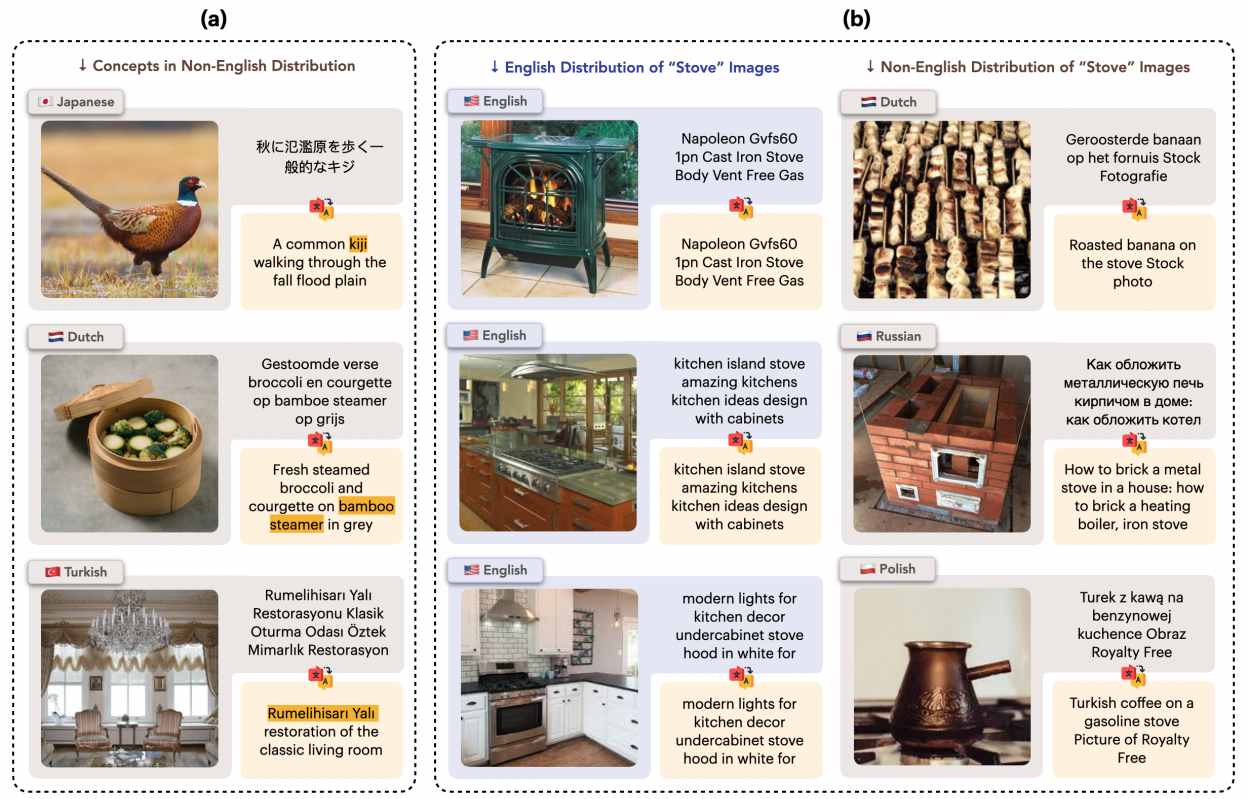


Figure 3.1: Multilingual image-text data adds diversity to the English data distribution in various, significant ways (a) We show some examples of culturally salient concepts that would not exist in "top-quality" English data (as determined by CLIP score), such as "kiji" (the national bird of Japan), "bamboo steamer" and "yalı" (a traditional architecture style for Turkish waterside houses) (b) Even for a common everyday object ("stove"), non-English and English images portray very different visual representations.

images look very different. Despite the diversity present in multilingual data, it is disproportionately excluded from existing large-scale pretraining corpora.

In this paper, we investigate the counterfactual: *can we improve on English vision tasks by diversifying the cultural and linguistic backgrounds of the training data?* Our investigation is motivated by a dichotomy: English image-text pairs constitute a minority of any random web crawl (in our estimate, one-third); yet, they form a majority in popular pretraining datasets such as LAION-5B [160], DataComp [57], and DFN [52]. It is common for web-scraped corpora to remove "low-quality" data by using a high-performing model (e.g., OpenAI CLIP) to compute image-text alignment and rank the raw data samples. However, this process often disproportionately favors English data if the filtering model also has an English bias [57, 81]. In addition to discarding many

potentially useful non-English image-text pairs, this can also negatively impact the geographical and cultural representation of the resulting dataset, and consequently, the model's performance on certain underrepresented populations [149, 145, 39].

Our key observation is that the diversity present in multilingual data can be confounded by the language the data is in, making it difficult to observe the empirical benefits of using such data in model training. To offer a more systematic study of the effectiveness of multilingual data—in contrast to English-only or English-dominated datasets—we fix the language medium, translate all captions from DataComp's 128M-sample web crawl [57] to English with an advanced translation model. We then re-filter the data pool and train a CLIP model on this translated multilingual data. We focus on two types of evaluations: (i) on standard English vision tasks including ImageNet, MSCOCO and Flickr retrieval, and (ii) on geographically diverse images, e.g. from GeoDE [145], which contains images of common objects across different geographical locations. We acknowledge that translation can sometimes be too literal, subject to losing the intent and richness of the original phrasing. Nevertheless, we hope findings from our work provide a starting point for studying how to leverage the diversity of multilingual data more effectively.

Our contributions are as follows:

- We demonstrate that with translation, non-English data does benefit English vision tasks. In particular, training on more samples of non-English origins leads to better performance on ImageNet, ImageNet distribution shifts and image-English-text retrieval. On the DataComp benchmark with a fixed compute budget, our best-performing approach that leverages translated multilingual captions outperforms training on just filtered raw captions by 2.0% on ImageNet and 1.1 percentage points on average across 38 tasks. When training for longer (which mimics the number of epochs large-scale multimodal models are often trained for), these performance gaps increase to 4.2% and 2.1 percentage points respectively.
- On a geographically diverse task such as GeoDE, training on translated multilingual data leads to 4.2% boost in accuracy on average compared to training on filtered raw data, with performance improvement observed for all regions, especially for Africa where the increase is 5.5%.

- We analyze in detail the differences between English and (translated) non-English image-text pairs. We quantitatively show that they capture distinct distributions, both in text and image space, even after they are converted to the same language medium. Consequently, it is beneficial to combine high-quality data from both sources as much as possible, since they are inherently complementary.

In summary, despite the abundance of "sufficiently useful" English data, existing data curation techniques can always do better in the data diversity axis by being more deliberate about including data from other language and cultural backgrounds. This way of enhancing diversity in turn leads to a better vision-language model *in general*, offering performance benefits beyond non-English vision tasks or tasks involving geographically diverse images. We will release the raw captions and the corresponding English translations for the 128M image-text pairs used in our experiments.

3.3 Related Work

Existing data collection and filtering approaches induce Western bias in downstream datasets and models; benchmarks that seek to capture this bias still receive relatively little attention. Consequently, despite evidence showing cultural and geographical limitations in popular vision datasets, the use of multilingual data is mostly intended for pretraining and fine-tuning multimodal models to do well on non-English tasks. Our work seeks to include more image-text pairs of non-English origins in the pretraining dataset, and shows that this process can improve performance, even on English-centric vision tasks.

Western bias of existing models and datasets Several papers have studied biases in popular datasets, especially biases that correlate with culture and geographic locations. Notably, Shankar et al. [164] find that ImageNet and OpenImages exhibit substantial US-centric and eurocentric representation bias. In the text domain, Santy et al. [159] demonstrate that existing datasets align predominantly with Western and White populations. It is not only the data collection process that leads to a Western bias, but also the data preprocessing pipeline. For instance, automated data filtering with scores output by a model, e.g., OpenAI CLIP, has been commonly adopted as a way

to discard low-quality web-crawled samples. Little is known about the potential biases induced by this approach. Hong et al. [81] recently show that CLIP score filter is more likely to include data related to Western countries compared to that of non-Western countries. In [57] (Figure 24), the authors offer evidence that CLIP filtering implicitly performs some English filtering, as the top CLIP score examples are increasingly dominated by English image-text pairs. Consequently, all these dataset biases translate to performance disparity, as demonstrated by existing work showing that the accuracy of vision systems drops significantly on non-Western inputs [39, 201, 149, 145], or low-resource languages [64].

Improving the availability of non-English data in multimodal datasets Translation has been a popular technique to address the limited availability of large-scale and high-quality non-English data in training and evaluation [184, 197, 28, 134, 48, 15, 64]. In addition to translating English captions into the language of interest, previous work also uses a curated list of common words in the native language to scrape image-text pairs from the web [109, 68]. COCO-CN [101] extends the MSCOCO dataset [30] with manually written Chinese captions.

Most closely related to our setup is the LAION-Translated dataset [122], which translates 3B samples of LAION-5B from many languages into English using the M2M100 model [49]. Compared to this dataset construction, we (i) use a more advanced translation model, NLLB [36], that covers twice as many languages, (ii) work with mostly raw data while LAION was heavily filtered and thus may already contain a biased representation of multilingual data. To the best of our knowledge, no existing work has experimented with the LAION-Translated dataset.

Adapting CLIP post-training for multilingual tasks Geigle et al. [63] translate high-quality English data into 95 languages, and use these translated samples to re-align an image encoder previously trained on English data to a multilingual language model. Similarly, Chen et al. [31] propose re-training OpenAI CLIP on a mix of Chinese and English data to enhance its multilingual representation. In [26], the authors explore adaptation without any image data, solely fine-tuning the text encoder with English captions from MSCOCO, Google Conceptual Captions, and VizWiz translated to other languages. Visheratin [184] replace the text encoder of OpenAI CLIP with the

text encoder from the NLLB model, and fine-tune the new model on multilingual image-text pairs obtained from translating LAION-COCO's English captions [161] into 200 languages. In contrast to these papers that employ multilingual data for the purpose of adapting to non-English tasks, we focus on using multilingual data to do better on common vision tasks that are in English.

Using multilingual data significantly enhances data diversity Our study is partly inspired by findings from Ye et al. [200], who show that multilingual synthetic captions obtained from existing image captioning systems provide higher semantic coverage than monolingual ones, over 3.6K images. Our experiments instead use raw web-crawled data and explore the performance benefits of embracing cultural and linguistic diversity in (mostly) human-generated captions *at scale*.

3.4 Experimental Setup

Given a starting pool of raw image-text pairs scraped from the web, many of which contain non-English captions, we experiment with ways to preprocess and filter this pool into a high-quality dataset. The quality of the dataset is measured by the zero-shot performance of a CLIP model trained on it from scratch.

Data We experiment with the medium pool of the DataComp benchmark [57], which consists of 128M image-text pairs randomly sampled from Common Crawl dumps between 2014 and 2022, and deduplicated. Unlike other heavily filtered corpora such as LAION [160], DataComp applies minimal data preprocessing, involving only NSFW filtering, deduplication of evaluation sets, and face blurring. This allows the candidate pool to stay close to the natural distribution of raw web data as much as possible, in addition to enabling maximum flexibility in dataset design.

Translation model To detect language and translate the raw captions from DataComp into English, we use the No Language Left Behind (NLLB) translation model [36], which is considered state-of-the-art. NLLB is the first to translate across 200 languages, including low-resource ones that are not currently supported by common translation tools. We use the 600M-parameter model publicly available on HuggingFace to allow for fast inference on our large data corpus. All 128M

captions from DataComp are translated to English; examples could be found in Appendix B.1. We provide some quantitative analysis of the translation quality in Appendix B.3.

Training After translating the captions of all samples in the raw data pool, we filter them based on cosine similarity between image and text embeddings. We experiment with using OpenAI CLIP-ViT-L/14 [141] and the *public* Data Filtering Network (DFN) from [52] to obtain the embeddings, and subsequently, the cosine similarities. The DFN, specifically designed to filter data for subsequent model training, was trained on three public datasets deemed as high-quality—Conceptual Caption 12M [27], Conceptual Captions 3M [166], and Shutterstock 15M [124]. We find that indeed the public DFN is better at data filtering compared to OpenAI CLIP, as measured by the performance of CLIP trained on the corresponding filtered datasets (see Appendices B.5 and B.7).

We train a CLIP model [141] from scratch on each filtered subset with ViT-B/32 as the image encoder, and follow DataComp's hyperparameters; details are in Appendix B.2. Unless specified otherwise, all models are trained for 128M steps, as determined by DataComp. For some select baselines, we also experiment with training for 10× longer. The fixed architecture, compute and hyperparameter setup allow us to isolate data quality as the main factor influencing performance.

Evaluation We perform zero-shot evaluation of trained CLIP models using the 38 tasks from DataComp. These tasks involve recognition and classification of a wide range of domains (e.g., texture, scene, metastatic tissue, etc.) in addition to image-text retrieval and commonsense association. Among them, we highlight commonly cited metrics such as ImageNet accuracy, ImageNet distribution shift accuracy—a proxy for natural robustness, and retrieval performance. ImageNet shifts include ImageNet-V2 [147], ImageNet Sketch [186], ImageNet-A [78], ImageNet-O [78], ImageNet-R [75] and ObjectNet [14]. Retrieval score is the average of the performance on Flickr30K [202], MSCOCO [30] and WinoGAViL [21]. Throughout this work we also report GeoDE worst-region performance [145]—a task that involves geographically diverse images—to demonstrate the added benefits of geographical inclusivity that obtained from using more (translated) multilingual captions.

Baseline name	Dataset size	ImageNet	ImageNet shifts	Retrieval	GeoDE	Average over 38 tasks
Training with DataComp setup (128M steps)						
Filtered raw captions	25.6M	0.316	0.260	0.282	0.688	0.350
Filtered raw captions, replaced with translated captions	25.6M	0.304	0.252	0.268	0.668	0.331
Filtered translated captions	25.6M	0.329	0.275	0.296	0.709	0.359
Filtered English-only captions	16.0M	0.283	0.236	0.278	0.666	0.327
Filtered raw captions ∪ Filtered translated captions	34.2M	0.329	0.271	0.298	0.720	**0.364**
Filtered raw captions & Filtered translated captions	51.2M	**0.336**	**0.280**	**0.301**	**0.725**	0.361
Training for 10× longer (1.28B steps)						
Filtered raw captions	38.4M	0.414	0.340	0.344	0.742	0.414
Filtered translated captions	38.4M	0.427	0.347	0.352	0.771	0.414
Filtered raw captions ∪ Filtered translated captions	34.2M	0.441	0.359	0.353	0.775	0.427
Filtered raw captions & Filtered translated captions	51.2M	**0.456**	**0.369**	**0.371**	**0.776**	**0.435**

Table 3.1: On the DataComp benchmark, training on translated captions outperforms training on raw captions across a range of metrics; using both types of captions yields even more performance gains. We report the performance of select baselines on the DataComp benchmark [57]; all baselines are trained for the same number of steps as specified. Here the filtering threshold (and thus the resulting dataset size) has been tuned for each baseline and we only show the filtered subset that yields the highest average accuracy. We find that with the same filtering method (i.e., using DFN score), training on translated captions ("Filtered translated captions") is more effective than training on raw captions ("Filtered raw captions") as seen from higher performance on ImageNet, ImageNet distribution shifts, retrieval, GeoDE (worst-region accuracy) and on average across 38 tasks. Combining both sources of captions leads to the best performance. Appendix B.7 contains the full results.

3.5 Impacts of Using (Translated) Multilingual Captions on Standard Vision Tasks

We explore training on each caption distribution separately, as well as combining them. Below we describe the baselines from Table 5.1 in more detail:

- *Filtered raw captions:* As mentioned in Section 5.3, we use the *public* DFN from [52] by default to filter the starting pool (128M samples). Given the images and the corresponding web-crawled captions, we experiment with varying the filtering threshold to keep top x% of the pool based

on DFN score. In Table 5.1, we only report the best average performance obtainable after the filtering threshold has been tuned, and the resulting dataset size. Refer to Appendix B.7 for the full results.

- *Filtered translated captions:* Similar to the approach above, we tune the filtering threshold, but using DFN score between an image and the English translation of the original web-crawled caption.
- *Filtered English-only captions:* Similar to "Filtered raw captions" baseline, here we also tune the filtering threshold to keep only a subset of the pool with the highest DFN scores, but with an additional constraint of only filtering from samples with web-crawled captions already in English.
- *Filtered raw captions, replaced with translated captions:* Given samples from "Filtered raw captions" (i.e. again, based on the cosine similarity score between image and original web text), we keep the images selected and replace the raw captions with the corresponding English translations.
- *Filtered raw captions* $\cup$ *Filtered translated captions:* We combine "Filtered raw captions" and "Filtered translated captions" subsets uncovered above. However, these subsets have about two-thirds of their images in common (see Appendix B.4). For such images, we only include one copy in the final training set and use English-translated caption by default. For the rest of the images in "Filtered raw captions" that do not appear in "Filtered translated captions", we include them in the training set with the corresponding original captions (which could be non-English).
- *Filtered raw captions & Filtered translated captions:* We combine image-text pairs from "Filtered raw captions" and "Filtered translated captions"; the overlapping images between these two subsets would appear twice in the final training set - one copy with the original web caption and one copy with the English-translated caption.

3.5.1 Overall Performance Trends

Combining high-quality raw and (translated) multilingual captions offers the best performance We find that using *both* sources of captions, and image data—since top (image, raw text) and top (image, translated text) pairs only have two-thirds of the images in common—leads to the best performance (bolded entries of Table 5.1). This approach surpasses training on only high-quality raw data by 2% on ImageNet, ImageNet shifts and retrieval, and improves GeoDE

worst-region performance by 3.7%. We note that this is not simply due to having more unique image-text samples, as filtered subsets of similar sizes but using a single source of captions (e.g., top 40% raw captions totalling 51.2M samples) yields significantly lower performance (Appendix B.7).

Using only translated multilingual captions is still better than using only raw captions Zooming in on "Filtered translated captions" and "Filtered raw captions" baselines, we find that the former outperforms the latter on many standard metrics (ImageNet, ImageNet distribution shifts, image-text retrieval). This is unexpected in light of prior work showing that ImageNet exhibits strong amerocentric and eurocentric bias [164], with images from America and Great Britain taking up 53% of the dataset.

3.5.2 Ablations

We perform more ablation studies to disentangle the reasons for the performance gains from using high-quality translated captions, as well as to verify that the gains are robust.

The performance gain from using translated captions is not simply due to converting all text data to a common language medium Given image-text pairs from the "Filtered raw captions" subset, we replace the web-crawled captions with the corresponding English translations ("Filtered raw captions, replaced with translated captions"). This intervention on only the captions leads to performance drop across the board. We hypothesize that this is due to noise in the translation process, as (i) many web captions are formed by stringing together short, potentially ungrammatical phrases and thus are "out-of-distribution" for the NLLB translation model, (ii) web captions may contain multiple languages in the same sentence, thereby leading to noisy language detection and translation.

Re-filtering data after translation is also necessary due to significant changes in the data ranking Besides noisy artifacts introduced by translation, the process also changes the image-text cosine similarity score and thus the quality ranking of the data samples in the pool. More specifically, we find that while "Filtered raw captions" is dominated by English samples, (translated) non-English samples make up the majority of "Filtered translated captions" (Figure 3.2). These two

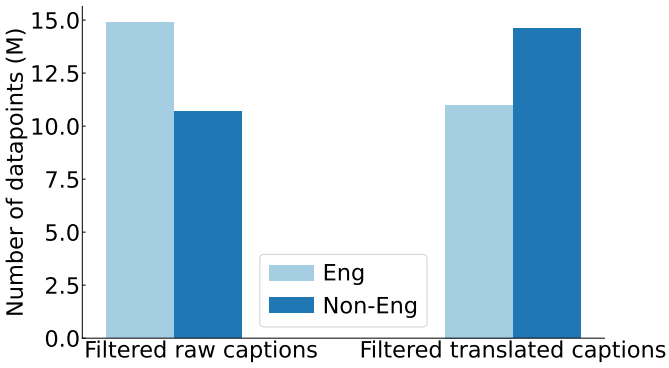


Figure 3.2: Filtering with translated captions allows substantially more (translated) non-English samples to be included in the final training set. While English data only makes up about one-third of the raw web crawl, it dominates the top-quality subset of the pool, selected based on DFN score between image and *raw* caption. With translation, English-translated non-English captions now make up the majority of the top-quality data and thus are more likely to be selected for training.

filtered subsets only share about two-thirds of the images in common, see Appendix B.4 for more details. Therefore, by changing the caption distribution, we are also inducing changes to the image distribution that the best-performing model would see.

The benefits of training with translated multilingual captions are consistent across data filtering networks As alluded to in Section 5.3, we also explore using cosine similarities output by OpenAI CLIP-ViT-L/14 [141] for data filtering. The full results for this ablation can be found in Appendix B.5. Similar to the previous observations, we find that using filtered translated captions yields better performance than using filtered raw captions.

The performance benefits of using translated data persist with much longer training duration We also experiment with training for 10× more steps (i.e., 1.28B samples seen) as this is more in line with the number of epochs typical vision-language models are often trained on (e.g., OpenAI CLIP models were trained on 400M datapoints for 32 epochs [141]). When using either the raw caption or the translated caption distribution, setting the filtering threshold to top 30% of the pool works best. Even though the two filtered datasets now yield the same average accuracy, training on high-quality translated captions still offers significant advantages when it comes to ImageNet, ImageNet shifts, retrieval and GeoDE. Combining high-quality data from both sources of captions continues to be the best performing approach, giving 4.2% improvement on ImageNet and 2.1 percentage points improvement on average, compared to just training on filtered raw captions. Results for more baselines can be found in Appendix B.8.

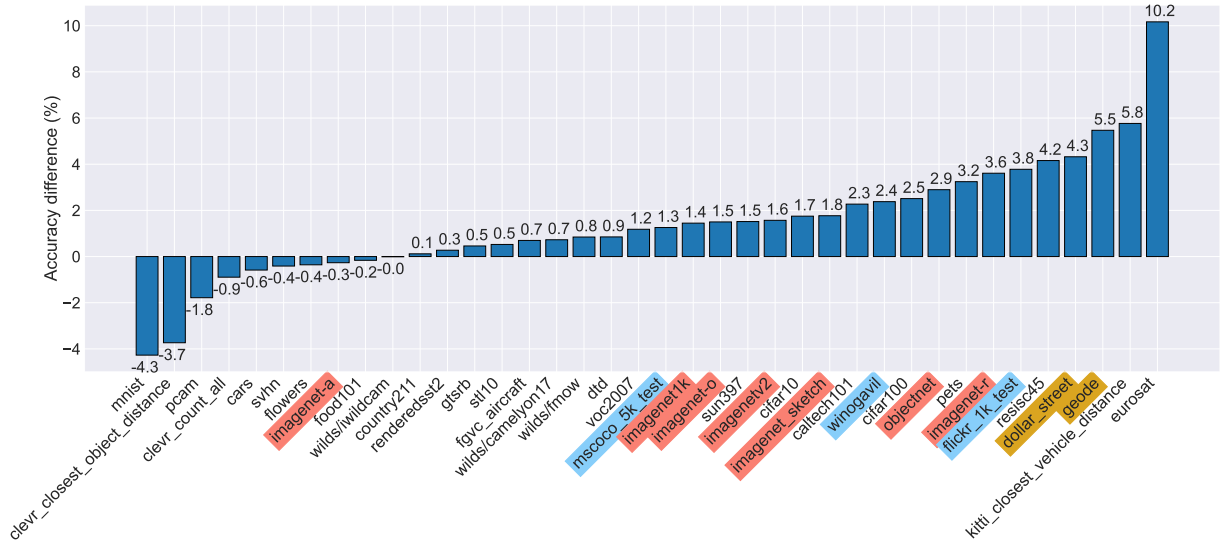


Figure 3.3: With the same degree of filtering, training with (image, translated caption) pairs improves performance on 28 out of 38 tasks compared to training with (image, raw caption) pairs, including on ImageNet distribution shifts, retrieval, and tasks with geographically diverse inputs. We compare performance on each task of the DataComp benchmark between training with raw captions and training with translated captions. Both datasets have been filtered with image-text cosine similarities output by the public DFN [52] to select the top 30% examples. We find that using translated captions leads to 1.5 percentage points improvement on average across 38 tasks. We highlight the performance changes on ImageNet distribution shifts (**red**), retrieval (**blue**) and fairness-related tasks (**dark yellow**).

3.5.3 Individual Task Analysis

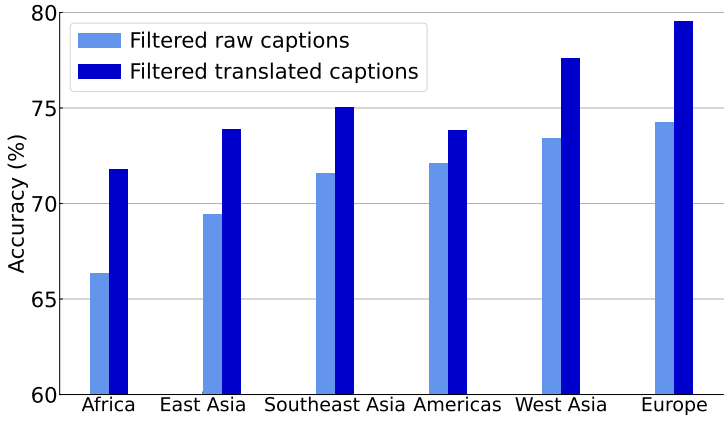


Figure 3.4: On GeoDE, using filtered translated captions leads to improvements across *all* regions compared to using filtered raw captions, with Africa observing the biggest gain. We break down the GeoDE performance by region and compare training on top 30% translated captions to training on top 30% raw captions. On average, classification accuracy improves by 4.2%, and the improvement applies to all regions in the dataset, especially Africa where the accuracy gain is the biggest at 5.5%.

After observing improvement across different metrics from using more (translated) multilingual captions, in this section we break down the performance changes for each of the 38 tasks in the

DataComp benchmark. The base model for comparison is CLIP trained on top 30% image-text pairs filtered from the raw data pool, and the new improved model is the one trained on top 30% image-text pairs after the same pool has been all translated to English. Both models are trained for 128M steps. Averaged across 38 evaluation tasks, the latter yields a 1.5 percentage points improvement. The biggest gains come from Flickr retrieval, fairness (GeoDE, Dollar Street) and remote sensing (EuroSAT, RESISC45) tasks (Figure 3.3).

In particular, on GeoDE [145], which consists of images of common objects crowd-sourced from six different regions across the world, we find that using translated multilingual captions makes CLIP perform better on *all* regions, with the biggest gain coming from Africa images (5.5%) and the second biggest gain coming from the Europe region. This is unexpected given that African languages only make up a small fraction of the training set, and after translation, more European (compared to African) language samples make it to the resulting filtered subset (Appendix B.4). Besides, it is worth noting that on GeoDE, our best baseline from Table 5.1 ("Filtered raw captions & Filtered translated captions") outperforms the current best baseline on DataComp's medium scale ("HypeSampler" [90]) by 1.5%, measured in terms of worst-region accuracy (Africa).

3.6 Understanding the Differences Between English and (Translated) Non-English Data

Given that using more image-text pairs of non-English origins in the training set offers significant benefits on most vision tasks, including tasks that are shown to be English-centric, we seek to further understand the various ways that non-English data complements and improves the diversity of English data, in both image and text space.

3.6.1 Image Distribution

As a proxy for capturing image distribution differences, we train simple classifiers—a Support Vector Machine (SVM) on CLIP embeddings and a ResNet-50—to distinguish images with English captions from those with non-English captions. We randomly sample 100K images from each distribution for

Figure 3.5: Visualizations of what an SVM deems typical of images with English captions and those with non-English captions. We show examples of easy-to-classify images in our English versus non-English data classification task. Besides the product logo and text in some images that are suggestive of the language distribution, the image content mostly depicts common scenes and objects.

training and 10K for testing. We only use images from the top 20% of the candidate pool (based on DFN cosine similarity score), to ensure that (i) these are the samples that the best-performing CLIP models are eventually trained on, (ii) images are of sufficient quality, to the extent that they have fitting captions accompanying them.

We note that this classification task is non-trivial for a number of reasons:

- Many images are duplicated across the web, i.e., after DataComp performs image deduplication it is possible for these images to appear with either English captions or non-English captions in our data pool.
- The language detection model is not perfect and web-crawled captions may contain more than one language in the same sentence.
- Images with non-English captions contain many sub-distributions of images, some of which may overlap with the distribution of images with English captions (e.g., eurocentric data).

Despite these challenges, our simple classifiers achieve 67% accuracy on the binary classification task, significantly better than random chance performance. In Figure 3.5, we show some examples

of what the SVM deems easy to classify. Overall this experiment suggests that the distribution of images with non-English captions is sufficiently distinct from that of images with English captions. Therefore, not training on more of the former means we are missing out on a considerable amount of visual information that can only be found in a separate part of the web.

3.6.2 Text Distribution

In the text space, we leverage MAUVE score [137] to quantify the differences between English captions and non-English captions that have been translated to English. MAUVE was originally designed to measure the gap between machine- and human-generated texts. The metric computes KL divergences in a quantized, low-dimensional space after embedding text samples from each distribution with a language model (by default, GPT-2). The output score ranges between 0 and 1 and the higher it is, the more similar the text distributions are. Similar to our analysis in the image space, we only use caption samples from the top 20% of the candidate pool (based on DFN score).

In Table 3.2, as a sanity check, we randomly sample two disjoint sets of 10K captions from the same text distribution (e.g., non-English captions having been translated to English with the NLLB translation model, *or* English captions having been passed through the same model). We find that the two sets indeed exhibit high MAUVE scores (above 0.95). When comparing raw English captions to English captions that have been passed through the NLLB model (i.e., "translated English"), we find that the MAUVE score decreases slightly (0.890), indicating that the translation process introduces some artifacts making English-translated English text look somewhat different from raw English text.

When comparing raw English texts and non-English texts, both having been passed through the translation model and thus undergone the same "preprocessing" (i.e., English translation), the resulting MAUVE score is relatively low (0.616). This signals that independent of differences in language, what is discussed in English captions and non-English captions differs in many ways. We should therefore leverage both sources of text information as much as possible for training.

3.7 Discussion

In this work, we bring all web-crawled image-text pairs into a common language medium via English translation, and systematically study the empirical benefits of using non-English data with respect to standard computer vision tasks (that are in English). By including significantly more (translated) multilingual data in the filtered training set, the improved cultural and linguistic diversity in turn leads to substantial gains across all major metrics—ImageNet, distribution shift robustness, retrieval capabilities and average performance across 38 tasks—even if some of these metrics have been shown to overfit to English. We also find that despite being translated to the same language, English and non-English data distributions are still distinct from each other.

Text distributions	**MAUVE score**
Translated non-English vs. Translated non-English	0.964 ± 0.005
Translated English vs. Translated English	0.957 ± 0.005
English vs. Translated English	0.890 ± 0.004
Translated English vs. Translated non-English	0.616 ± 0.010
English vs. Translated non-English	0.449 ± 0.008

Table 3.2: There exists a substantial gap between the distribution of English captions and that of non-English captions, even when we apply translation to both, suggesting that they capture different contents. We use MAUVE score [137] to measure the difference between English captions and (translated) non-English captions in the training set. We find that (i) translation indeed introduces some artifacts and changes what "English" texts may look like, (ii) the English text distribution is remarkably different from the non-English one, even after they are converted to the same medium with translation. All scores are averaged over 3 randomly sampled sets of 10K captions.

Limitations We fix the data filtering method to be based on image-text cosine similarity output by a trained model, and study the impact of selecting training data based on different caption distributions. We show that the advantages of using translated multilingual data are robust to the choice of the filtering network. However, our best-performing baseline is currently not state-of-the-art on the DataComp benchmark [57]. It remains an open question whether the performance benefits of our method persist with other score metrics, e.g. hyperbolic entailment [90] - currently the best method for DataComp's medium scale, or other filtering methods that are used jointly with CLIP score, e.g. T-MARS [114] which also removes text-spotting images with limited visual information.

Besides, we acknowledge that translation can introduce artifacts and reduce the richness of expressions in the original languages. Prior work has shown that translated sentences are less effective compared to manually-written sentences as sources of training data for vision tasks, e.g. image captioning [119, 97]. Our work mainly leverages translation as a way to convert all image-text pairs to the same medium, and remove confounding impacts of language in data selection and model training.

Broader impact While most studies have looked into non-English data with the goal of increasing societal representation and subsequently improving performance on under-served populations or tasks, we observe that non-English data can actually help enhance model capabilities *as a whole*, including on standard English benchmarks. This suggests that diverse representation in training data, e.g. as measured by cultural and linguistic backgrounds, should be a deliberate design decision in the data curation process, instead of existing only as a byproduct of the preprocessing pipeline or out of societal considerations.

Future work This work motivates future studies into data curation techniques that directly improves the diversity of data origins. Another interesting direction of exploration is adapting trained CLIP models from this paper for multilingual benchmarks, such as by re-training the text encoder (that has only been trained on English and English-translated captions) with the technique proposed by Carlsson et al. [26]. We hypothesize that text adaptation alone is sufficient for our models to perform competitively on non-English tasks, owing to the presence of significantly more multilingual and multicultural images in our pretraining dataset.

Chapter 4

Beyond Filtering: Synthetic Captions as a Data Quality Fix

4.1 Overview

Massive web datasets play a key role in the success of large vision-language models like CLIP and Flamingo. However, the raw web data is noisy, and existing filtering methods to reduce noise often come at the expense of data diversity. Our work focuses on caption quality as one major source of noise, and studies how generated captions can increase the utility of web-scraped datapoints with nondescript text. Through exploring different mixing strategies for raw and generated captions, we outperform the best filtering method proposed by the DataComp benchmark by 2% on ImageNet and 4% on average across 38 tasks, given a candidate pool of 128M image-text pairs. Our best approach is also 2× better at Flickr and MS-COCO retrieval. We then analyze what makes synthetic captions an effective source of text supervision. In experimenting with different image captioning models, we also demonstrate that the performance of a model on standard image captioning benchmarks (e.g., NoCaps CIDEr) is not a reliable indicator of the utility of the captions it generates for multimodal training. Finally, our experiments with using generated captions at DataComp's `large` scale (1.28B image-text pairs) offer insights into the limitations of synthetic text, as well as the importance

of image curation with increasing training data quantity. The generated captions used in our experiments are now available on HuggingFace[6].

4.2 Introduction

Pretraining large multimodal models on image-text pairs sourced from the web has become a standard approach to obtaining high performance on vision tasks [7, 136, 85, 141]. However, raw web data can be noisy or uninformative (Figure 4.1). Many existing data preprocessing efforts revolve around human-defined heuristics based on image and text content separately—e.g., caption length, presence of nouns, sentence complexity, image aspect ratio, minimum image size [24, 160, 27, 166]—or the reliability of the data source [42]. More complex filtering approaches target poorly aligned image-text pairs, by using trained CLIP models [141] to rank the cosine similarity score between image and text embeddings [160], or ensuring mentions of image objects in the captions [166]. These approaches discard between 60% to 90% of the initial data collected, regardless of whether the images themselves are suitable for training.

In this work, we seek to restore the utility of such discarded examples with the help of synthetic captions. To do so, we leverage the DataComp benchmark [56], where initial data processing is kept to a minimum, i.e. only filtering out NSFW examples and train-test overlap. This allows us to perform controlled experiments on the raw Common Crawl data and bypass subjective human-design choices that may be employed in the creation of other datasets (e.g., LAION-5B [160]). We study several image captioning models and find that recent releases (e.g., BLIP2 [99] and OpenCLIP-CoCa [138]) can generate captions that improve CLIP training and lead to a significant boost in zero-shot performance over existing data curation methods. In particular, at the `medium` scale (128M samples seen), training on the *entire candidate pool* with synthetic captions is sufficient to outperform common filtering baselines that are applied to raw data (e.g., selecting top 30% examples with highest image-text cosine similarity based on OpenAI's CLIP-ViT/L14). Section 4.6 describes our experiments with a variety of mixing strategies to combine signals from both raw and synthetic text.

[6] https://huggingface.co/datasets/thaottn/DataComp_medium_pool_BLIP2_captions, https://huggingface.co/datasets/thaottn/DataComp_large_pool_BLIP2_captions

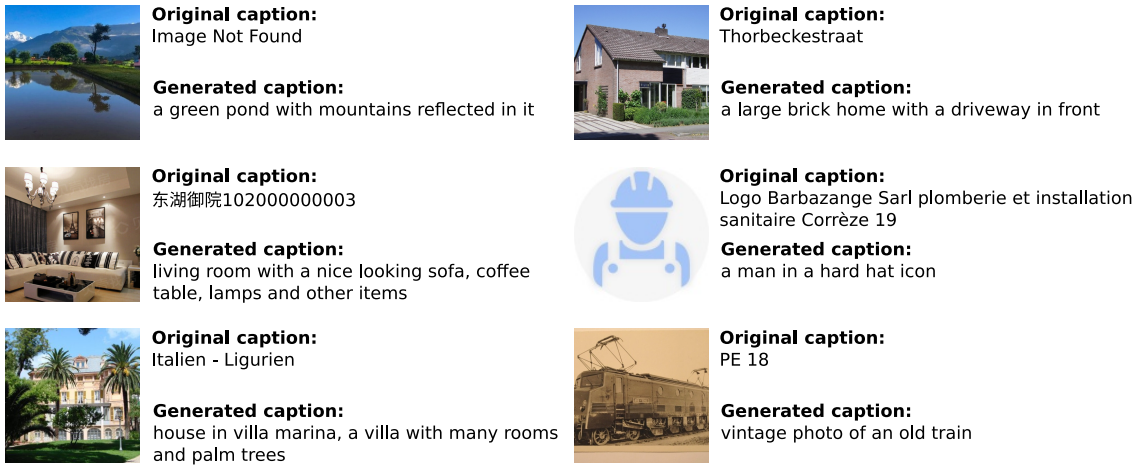


Figure 4.1: Raw captions crawled from the web contain significant noise; cosine similarity filtering helps reduce noise but discards many images that are useful for training. Here we show some images that would be filtered out if only the top 30% examples from the candidate pool with highest image-text cosine similarities are used for training. In these pairs, captions generated by BLIP2 tend to be more faithful to the respective images compared to raw captions obtained from the Internet. In Appendix C.1, we show 20 other samples drawn completely at random from the discarded pool.

To explain the performance benefits of synthetic captions, we measure caption noise and diversity in various training sets, and demonstrate the importance of both factors in achieving good performance. While existing data filtering methods are effective at reducing noise, they also hurt the diversity of the original training data in the process (e.g., by reducing concept coverage). Synthetic captions help alleviate this drop in diversity by increasing the number of useful captions available for training. In Section 4.7, we analyze various properties of caption data, as well as specific advantages of training with synthetic captions (e.g., improved retrieval capabilities).

Remarkably, our empirical investigation in Section 4.5 shows that choosing a captioning model to yield competitive downstream performance is non-trivial, as better performance on image captioning benchmarks does not necessarily mean better generated captions for CLIP training. We also note that while this work focuses on the quality of captions used in multimodal training, image quality is another equally important topic of study. As the size of the data pool we experiment with grows, we start to observe changes in the relative importance of text quality versus image quality in building a good pretraining dataset. We comment on this in Section 4.8.

To summarize, our findings serve as a first step towards improving the quality of *web-scale* datasets via the use of synthetic captions. In the process, we offer insights on several research directions:

- *What are the considerations for choosing a captioning model?* We find that specializing a pretrained network towards image captioning via fine-tuning, and optimizing for high CIDEr score on standard benchmarks in general, end up producing captions that are less effective for multimodal training. Reference-free captioning metrics (e.g., CLIP-S [79]) more reliably reflect the training quality of the generated captions.
- *How to combine signals from multiple sources of captions?* We investigate different strategies for filtering and mixing raw and synthetic captions. This leads to performance gains on DataComp benchmark at `small` (12.8M pool size), `medium` (128M pool size) and `large` (1.28B pool size) scales, compared to existing approaches that utilize only raw data. On ImageNet, the performance benefits diminish with scale. On retrieval tasks, however, the gains are significant across all scales.
- *What makes synthetic captions effective?* Our analysis of text properties shows that on an individual level, synthetic captions are less noisy and contain more visual information. However, at the population level, synthetic captions are less diverse than raw captions. Consequently, using *both* sources of captions helps improve the overall caption quality, measured in terms of text diversity as well as image-text alignment.
- *How do benefits of synthetic captions scale?* Unlike what was found in the original DataComp experiments, given access to generated captions, the best filtering approach differs across scales. Experimenting with data quantities ranging from 12.8M to 1.28B also allows us to observe some limitations of synthetic captions. We posit that image-based filtering, as well as the diversity gap between model-generated and web-scraped captions, play an increasingly important role in large data regimes.

More broadly, our results have important implications for future work as additional progress (captured by the right metric) in image captioning can further enhance the quality of text used for vision-language pretraining. Moreover, the effectiveness of synthetic captions unlocks another massive source of training data: uncaptioned web images from Common Crawl. This can ultimately

empower more large-scale multimodal training by improving the availability of properly aligned and sufficiently diverse image-text data.

4.3 Related Work

Synthetic data. Previous work has explored using synthetic data to create new datasets or augment existing ones [46, 150, 133, 86, 211, 58, 117, *inter alia*]. Closer to our work, He et al. [74], Azizi et al. [12], Bansal and Grover [13] use image generation models to create synthetic images for classification tasks. In the context of CLIP, Santurkar et al. [158] show that a model trained on synthetic captions can outperform a model trained on human-provided captions. The captions were generated procedurally for the 120K images in the MS-COCO training set [30] using multi-object image labels verified by Mechanical Turk workers, which would be difficult to obtain for web-scale datasets like LAION-5B [160] or CommonPool [56] that are about four orders of magnitude larger. Most similar to our work is the LAION-COCO dataset [161], containing 600M image-text pairs from LAION-5B [160] with synthetic captions generated using BLIP [100] and ranked using CLIP models [141, 84]. While [161] heavily filters the raw data pool before generating captions, we work with uncurated web datasets. In addition, the generated captions provided by LAION-COCO still significantly lag behind the corresponding web-crawled captions when it comes to yielding good CLIP performance—we provide empirical evidence and address this gap in Appendix C.7.

Image captioning. Building models able to generate captions from images has been a long-standing subject of research [91, 89, 102, 41, 189, 188, *inter alia*]. More recently, models like BLIP2 [100, 99], Flamingo [7], and CoCa [203, 138] have made significant progress on this task. It is worth noting that the training data for BLIP [100] and BLIP2 [99] contains synthetic captions, as the authors find that this helps boost the captioning ability of the resulting model compared to training on just noisy web data. Zhu et al. [213] couple large language models with image captioning models to generate more enriched image descriptions. We expect that as these image captioning systems become more capable, the impact of using synthetic data will bring larger improvements over existing noisy image-text datasets.

Improving image-text datasets. Given the importance of the pretraining data for multimodal networks [124, 51, 56], several authors have proposed techniques for improving the quality of image-text datasets. Radenovic et al. [140] propose a filtering technique called Complexity, Action, and Text-spotting (CAT), designed to select only informative image-text pairs. Cao et al. [25] filter out samples that contain text regions in the image and advocate for the benefits of increasing the number of samples given a fixed compute budget. Instead of discarding all text-spotting examples, Maini et al. [114] proposes masking out the text part in the image and only removing image-text pairs in which the masked image contains no useful visual features. Abbas et al. [2] identify and remove samples that are semantically similar to each other. Many image-text datasets also have their own preprocessing techniques, often not fully disclosed [141, 85, 136, 27, 42, 160]. All of these filtering approaches are complementary to the use of synthetic captions proposed by this work.

Concurrent to our work, Fan et al. [50] present a form of data augmentation for training CLIP models where the captions are rewritten by a large language model. However, the rewriting process assumes access to some raw text and is not conditioned on the images, which may limit its effectiveness when the original captions are not descriptive (e.g., see Figure 4.1). In contrast, our work uses image captioning models, which are able to generate relevant captions for images regardless of the original text associated with them. We also work with raw Common Crawl data instead of preprocessed datasets to study the trade-offs between raw and generated captions in a systematic manner. Finally, Gadre et al. [56] introduces DataComp, a benchmark for designing better pretraining datasets for CLIP, which we use in experiments throughout the paper.

4.4 Experiment Setup

Data. Most of our experiments involve the CommonPool provided by the DataComp benchmark [56]. CommonPool contains image-text pairs sourced from Common Crawl dumps between 2014 and 2022, deduplicated and randomly shuffled. The `small`, `medium`, and `large` scales of the benchmark contain 12.8M, 128M and 1.28B candidate pairs respectively. Data preprocessing is kept to a minimum, involving only NSFW filtering, evaluation set deduplication, and face blurring, to allow

maximum flexibility for dataset design. We also experiment with LAION-COCO [161] and discuss in Appendix C.7 why it is not ideal for studying the benefits of synthetic captions.

Captioning models. We experiment with BLIP [100] and BLIP2 [99] using HuggingFace's Transformers framework. Both models were pretrained on 129M image-text pairs from the web including MS-COCO [30] and LAION-400M [160], in addition to the bootstrapped version of the web data with synthetic captions generated by BLIP's captioner. We also look at OpenCLIP-CoCa [138, 84], which was trained on LAION-2B [160]. For each architecture, we experiment with both the pretrained model and the one that has been fine-tuned on MS-COCO. Caption generation uses top-K sampling with K = 50, minimum caption length 5, and maximum caption length 40.

Training. Given CommonPool data of a particular scale, we generate synthetic captions for the images in the pool using the captioning models described above. Then we train a CLIP model on the resulting image-text datasets, using ViT-B/32 as the image encoder for the `small` and `medium` scales, and ViT-B/16 for the `large` scale. Following DataComp's setup [56], the compute budget, architecture and hyperparameters for each scale are fixed in order to isolate data quality as the main factor influencing performance. Given a candidate pool of N image-text pairs, the CLIP model is then trained with N samples seen in total. Refer to Appendix D.1 for more details.

Evaluation. We adopt DataComp's zero-shot evaluation suite and report both ImageNet accuracy and the average accuracy over 38 classification and retrieval tasks proposed by the benchmark [56]. We also pay particular attention to retrieval performance on Flickr30K [202] and MS-COCO [30]. The retrieval score reported is the average of text-to-image Recall@1 and image-to-text Recall@1.

Unless specified otherwise, in the subsequent sections, "CLIP score filtering" or "top x%" refers to selecting top x% examples from the initial training set, based on the cosine similarity between image and text embeddings output by OpenAI's CLIP ViT-L/14 model [141], and "BLIP2" refers to captions generated by BLIP2, using top-K sampling with softmax temperature 0.75, which we have found to yield the best downstream performance compared to other sampling temperatures (see Appendix C.3).

Captioning model	NoCaps CIDEr	CLIP-S	Cosine similarity	No. of unique tri-grams	ImageNet accuracy	Flickr retrieval
BLIP, ViT-L/16 (finetuned)	113.2*	0.698	0.231	2.82×10^6	0.207	0.498
BLIP2, ViT-g	80.6	0.737	0.251	2.72×10^6	0.281	0.507
BLIP2, ViT-g (finetuned)	119.7*	0.711	0.235	1.97×10^6	0.227	**0.549**
OpenCLIP-CoCa, ViT-L/14	0.354*	0.752	0.260	4.45×10^6	**0.321**	0.395
OpenCLIP-CoCa, ViT-L/14 (finetuned)	106.5*	0.702	0.232	1.81×10^6	0.252	0.542

Table 4.1: CIDEr score does not reliably predict how effective a captioning model is at generating synthetic captions for multimodal pretraining; fine-tuning image captioning models leads to lower ImageNet accuracy when training CLIP on the generated captions. * indicates numbers obtained from previous work and from contacting the authors. We fix the architecture and compare captions generated from captioning models with and without fine-tuning on MS-COCO [30] as sources of text supervision for CLIP. Models that are fine-tuned specifically for the task of image captioning ends up producing synthetic captions that are worse for pretraining CLIP to do well on complex tasks like ImageNet. We hypothesize that this is due to reduced text diversity. On the contrary, retrieval performance is higher when using captions generated by fine-tuned models.

4.5 Impact of Model Specialization on Captions Generated for Multimodal Training

Given the abundance of image captioning models to choose from, a natural question to ask is: does performance on standard image captioning benchmarks correlate with how useful the generated captions are as text supervision for CLIP training?

In particular, CIDEr [181], together with other reference-based metrics like SPICE [9] and BLEU-4 [130], has been widely adopted as a yardstick for determining state-of-the-art on image captioning benchmarks [203, 7, 100, 99, 82]. Consequently, previous work [203, 100, 99] also experiments with fine-tuning captioning models on MS-COCO and obtains competitive CIDEr scores on popular evaluation sets like NoCaps [4].

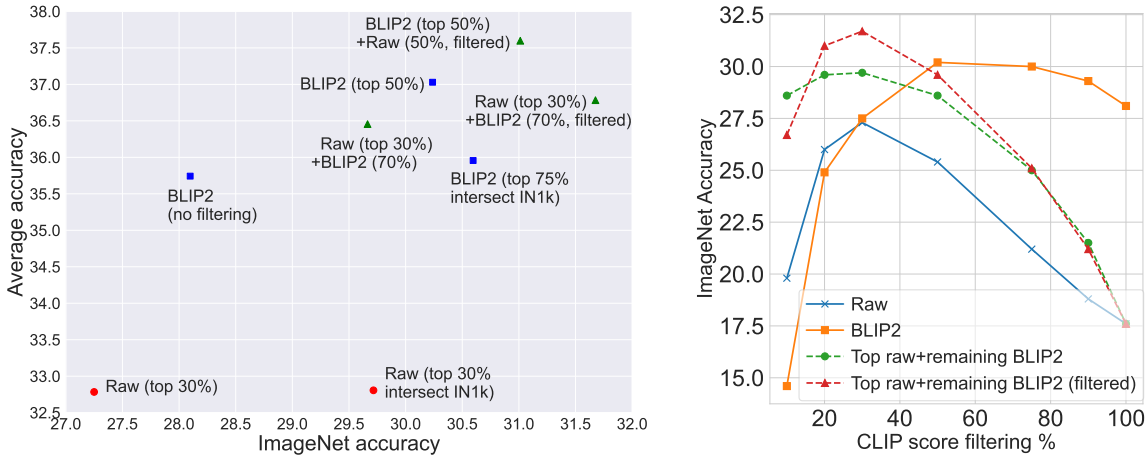


Figure 4.2: At the 128M scale of DataComp, we obtain improvement on ImageNet and average accuracies compared to the best filtering method on raw data, by using a mixture of raw and synthetic captions, selecting only image-text pairs with cosine similarity above a certain threshold. (Left) We visualize how various data filtering strategies perform at `medium` scale, on ImageNet and across 38 tasks. Including BLIP2 captions in the training data significantly outperforms competitive baselines from DataComp trained on only raw text [56]. (Right) As we vary the percentage of top examples chosen from the pool (based on CLIP score), we see consistent benefits from (*i*) using BLIP2 captions for samples that would be discarded otherwise, (*ii*) applying the same filtering threshold to new image-text pairs containing BLIP2 captions to maintain a high level of image-text alignment. The exact accuracy numbers can be found in Appendix C.4.

We compare the utility of synthetic captions produced by BLIP2 and OpenCLIP-CoCa with and without fine-tuning on MS-COCO, by training CLIP on the generated captions and evaluating the trained model on ImageNet classification and Flickr retrieval (Table 4.1). Fine-tuned captioning models produce captions that boost the retrieval capabilities of CLIP, but hurts its ImageNet performance. We hypothesize that fine-tuning on MS-COCO reduces the diversity of the generated text, as evidenced by the lower number of unique trigrams across 1M random caption samples (Table 4.1). Notably, captioning models that are not fine-tuned have very poor CIDEr scores; going with this metric would have suggested that these models are not suitable for caption generation at all.

While many image captioning metrics like CIDEr, SPICE and BLEU-4 emphasize similarity between generated captions and reference captions provided by humans, prior work has also proposed reference-free metrics—for example, CLIP-S [79], which uses a trained CLIP model to assess the compatibility between an image and the generated caption. We compute CLIP-S for the `medium`

candidate pool with different synthetic captions and find that this metric is more reflective of the ImageNet performance trend. Fine-tuned captioning models produce captions that have lower CLIP-S and image-text cosine similarity in general.

Since BLIP2 (no fine-tuning) produces sufficiently good text supervision for CLIP to do well on both ImageNet classification and Flickr retrieval, we use it as the captioning model of choice in subsequent experiments that look at how to combine raw and synthetic captions.

4.6 Filtering Raw and Synthetic Captions

Here we explore in more detail different ways of filtering and combining raw and generated captions at the `medium` scale of DataComp [56]:

- *No filtering:* we train on the entire, unmodified pool (i.e., 128M samples).
- *CLIP score filtering:* we select the top x% of examples with highest image-text cosine similarity.
- *CLIP score intersect with ImageNet1k clustering:* Gadre et al. [56] propose clustering image embeddings and only selecting images whose cluster center is a nearest neighbor to an image from ImageNet1k. The authors then find the intersection between this set of examples and those that are in the top x% based on CLIP score. This is the best baseline using raw captions on DataComp.
- *Combining raw and synthetic captions:* we use raw captions for the top x% of examples based on CLIP score. For the remaining images (that would otherwise be filtered out), we generate corresponding BLIP2 captions and add them back to the training pool. We also experiment with filtering these

Raw (no filtering)	13.2
Raw (top 30% intersect IN1k)	18.2
Raw (top 30%)	19.7
Raw (top 30%) + BLIP2 (70%, filtered)	38.0
BLIP2 (top 75% intersect IN1k)	38.9
BLIP2 (top 50%)	40.1
Raw (top 30%) + BLIP2 (70%)	40.5
BLIP2 (no filtering)	41.7

Table 4.2: Training on generated captions substantially boosts retrieval capabilities of the resulting CLIP models. Here we report the average text-to-image and image-to-text retrieval performance across both MS-COCO and Flickr for different data filtering baselines. More specific breakdown can be found in Appendix Figure C.1. Overall, we observe a 2× improvement at the `medium` scale of DataComp when synthetic captions are included in the training set.

additional image-text pairs with the same cosine similarity threshold set in the first step (i.e., BLIP2 (x%, FILTERED) in Figure 4.2).

In Appendix C.4, we examine other baselines and report how well each approach does with varying cosine similarity thresholds. Figure 4.2 (left) shows the relative performance of select baselines (the degree of CLIP score filtering has been tuned and only the best accuracy is plotted). We find that the best performance at `medium` scale, measured by either ImageNet or average accuracy, is achieved by mixing raw and synthetic captions, subject to a cosine similarity threshold. Besides, including BLIP2 captions in the training set also improves retrieval performance by more than 2× (Table 4.2).

In the right plot of Figure 4.2, we compare ImageNet performance at various filtering thresholds for methods that involve only one source of captions and those that involve both. We observe that given image-raw-text pairs filtered with certain cosine similarity threshold (blue line), adding BLIP2 captions for some (red line) or all of the remaining images (green line) always helps. It is worth noting that as we lower the threshold and include more raw captions in the training mix, the performance starts to become lower than using just synthetic captions (orange line). Overall we find that filtering is still a necessary step even when using synthetic captions that are supposedly more relevant to the training images.

4.7 What Makes Synthetic Captions Effective?

4.7.1 Defining Caption Quality

As seen from sample images in Figure 4.1, web-scraped text may not contain specific visual information (e.g., "Italien - Ligurien") or may not reflect the content of the image (e.g., "Image Not Found"). We seek to understand how generated captions can help overcome these issues.

To approximate the richness of information conveyed in the text data, we take a 1M random subset from each training set and measure the number of words, as well as the grounding ratio [174] (i.e., the fraction of tokens that describe visual concepts, with the vocabulary defined by MS-COCO), in the corresponding captions. In Figure 4.3, we observe that synthetic captions and raw captions follow different distributions, with the former generally containing more words (left pane) and more

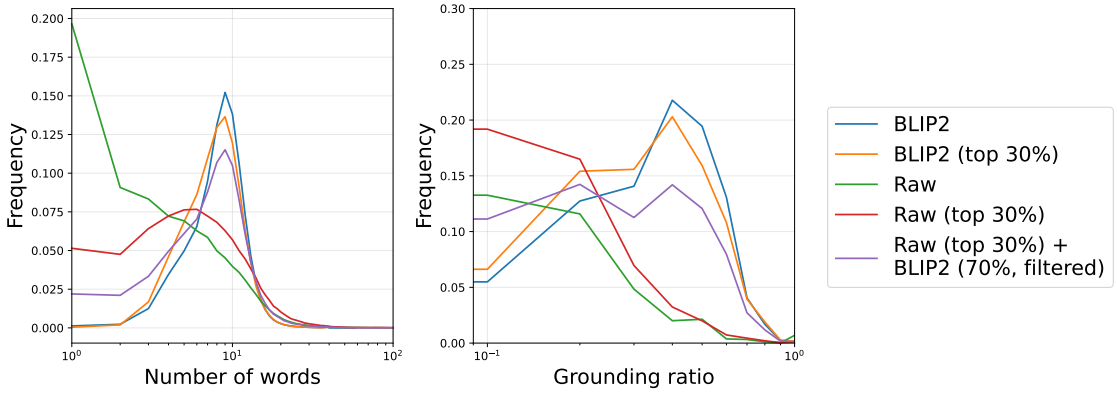


Figure 4.3: Individual synthetic captions can contain more information (especially visual one) than raw captions. We calculate the number of words and the fraction of those being visual tokens in each caption for different training sets. Individual BLIP2 captions tend to yield higher numbers on these two metrics compared to individual web-crawled captions, suggesting that on a caption-per-caption basis, synthetic data may contain richer information.

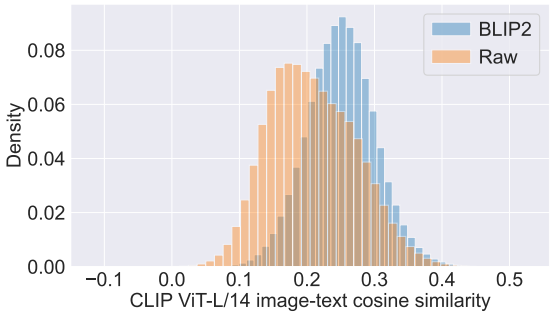


Figure 4.4: Generated captions overall exhibit higher image-text alignment than raw captions; this indicates that the former is less noisy as a training source. We randomly sample 1% of the 128M candidate pool and given the same set of images, compare the cosine similarity distribution between raw caption data and BLIP2 caption data. We find that overall BLIP2 captions have much higher image-text cosine similarity (mean similarity 0.251 vs 0.208).

visual tokens (right pane) per sample. Performing CLIP score filtering on raw captions leads to improvements on both of these properties; so does mixing raw and synthetic captions. Regarding the issue of poor image-text alignment, we approximate the alignment using cosine similarity between image and text embeddings from CLIP, and find that web-crawled captions indeed have lower similarities overall compared to model-generated ones (Figure 4.4).

The analyses above measure properties of individual captions. We next aim to capture a single diversity metric over *all* text in the training set. We again select a random subset, the size of which scales with the training set size, and calculate the number of unique trigrams across all captions in the subset. With this diversity metric, we find that BLIP2 captions actually lag behind raw captions (Figure 4.5). Using only the top 30% raw captions (based on CLIP score) is even more detrimental.

We summarize these different aspects of caption quality in a noise versus diversity framework (Figure 4.5), which also offers some intuition for our best baseline uncovered in Section 4.6. CLIP score filtering that has been commonly adopted in prior work [160, 56] is effective at improving performance on raw data by by removing noisy examples (i.e., those with poor image-text alignment). However, this procedure also lowers diversity (note: Figure 4.5 only provides a measure of text diversity, but image diversity is affected as well). By generating synthetic captions for the images that would be discarded otherwise, and subsequently only using pairs where the image-text similarities still meet the threshold, we manage to keep the overall noise level similarly low, while adding more diversity to the training pool. Progress along both axes enables further performance improvement compared to just filtering raw data.

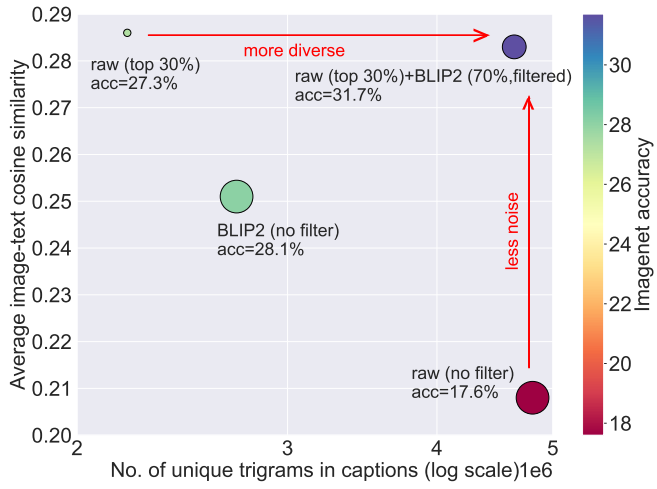


Figure 4.5: Combining raw and synthetic captions subject to a cosine similarity threshold helps reduce noise level while boosting data diversity, both of which are essential for achieving good performance. In this plot, circle size denotes the relative size of the resulting training set. While removing noisy image-text pairs, CLIP score filtering also lowers the diversity of the caption set substantially, as measured by the number of unique trigrams in the pool. Adding more useful training data by using BLIP2 captions for filtered out images, while respecting the existing CLIP score threshold, helps overcome this limitation and improves the training data quality along both axes.

4.7.2 Performance Analysis

After diving deeper into properties of synthetic captions, we next analyze the training implications of these captions in more detail. We examine two models, one trained using only raw captions and the other using only BLIP2 captions, with both training sets having been filtered with CLIP score for top 30% pairs, and achieving similar performance on ImageNet (27.3% vs 27.5%). Averaged across 38 evaluation tasks, training on generated captions offers a 2.8% improvement. We break down performance difference between the two models on individual tasks (Figure 4.6), and observe that

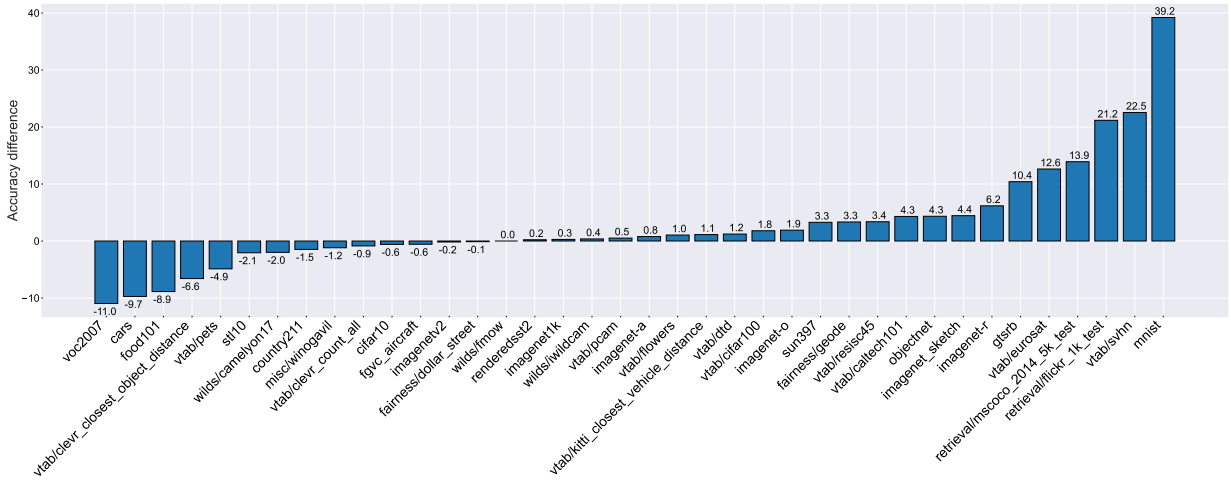


Figure 4.6: Given similar ImageNet accuracy, training with generated captions improves performance on 23 out of 38 tasks compared to training with raw captions, especially on ImageNet distribution shifts, text recognition and retrieval tasks. We compare performance on each task of the DataComp benchmark between training with only BLIP2 captions and training with only raw captions; both datasets have been filtered with CLIP score to select the top 30% examples. Even though the two training sets both yield ∼27% ImageNet accuracy, using generated captions leads to 2.8% improvement on average, including minor gains on ImageNet distribution shifts and significant gains on MNIST, SVHN, Flickr and MS-COCO retrieval.

BLIP2 captions also perform better on ImageNet-derived distribution shifts and text recognition (e.g., MNIST, SVHN). Notably, among the tasks with the biggest performance gains are Flickr and MS-COCO retrieval. We provide a similar analysis in Appendix Figure C.2, where expanding a filtered raw dataset with additional images and their BLIP2 captions improves CLIP performance on 30 out of 38 tasks.

The two models compared above share similar ImageNet accuracy but may not be trained on the same images. In Figure 4.7, we fix the set of training samples to be the top 30% with highest cosine similarity between image and *raw* text. Replacing the raw captions with BLIP2 captions increases retrieval performance on Flickr and MS-COCO by more than 1.5× (first two columns of each task). We also report retrieval performance of training on all BLIP2 captions (no filtering), generated using either the pretrained or the fine-tuned captioning model, as well as that of training on a mixture of raw and BLIP2 captions, to demonstrate the consistent gains that synthetic captions offer.

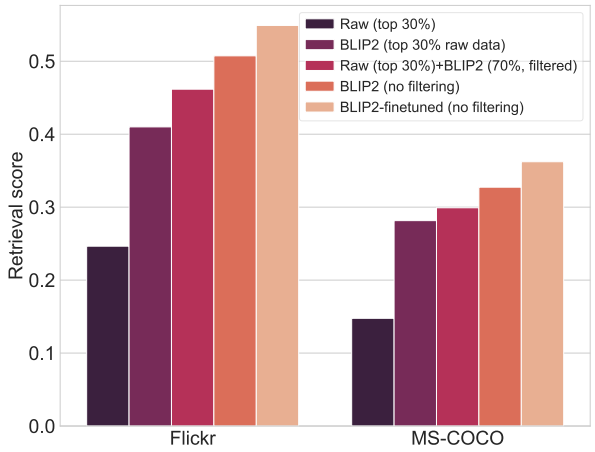


Figure 4.7: Synthetic captions display a clear advantage over raw captions on retrieval tasks. We highlight the superior performance on Flickr and MS-COCO retrieval obtained from training CLIP on captions generated by BLIP2 (pretrained model or model that has been fine-tuned on MS-COCO), compared to training on raw captions. In particular, the first two columns of each task represent two models trained on the same set of images (i.e., those whose cosine similarity between image and *raw* text embeddings are in the top 30%), just with different captions. This suggests that substantial gains on retrieval tasks can be obtained solely by using better aligned captions.

4.8 Performance at Scale

We next apply select baselines described in Section 4.6 to a wider range of candidate pool sizes, ranging from 12.8M to 1.28B samples. In particular, we examine training on the entire pool with only raw captions or only BLIP2 captions, CLIP score filtering, using the intersection of top CLIP score examples and examples that lie in clusters close to ImageNet train set, as well as mixing raw and synthetic captions—our best baseline from the `medium` scale. The filtering percentage for each method is tuned on the `medium` scale candidate pool and then applied to experiments at other scales. Given a starting pool of N samples, we limit the training budget to N steps. The 400M and 1.28B scales use the `large` training settings from DataComp (see [56]).

We focus on ImageNet classification and Flickr retrieval performance (note: MS-COCO training set was included in BLIP2's pretraining data so we have excluded MS-COCO retrieval from this comparison). At larger data quantity regimes, using synthetic captions continues to substantially outperform existing raw-text filtering baselines at retrieval (Figure 4.8, right plot). On ImageNet, however, adding BLIP2 captions to the training mix sees diminishing returns: Raw (top 30% intersect IN1k) + BLIP2 (remaining 70%, filtered, intersect IN1k) outperforms DataComp's best baseline trained on raw data, Raw (top 30% intersect IN1k), by 2.5% at 400M scale and 1.2% at 1.28B scale (Figure 4.8, left plot).

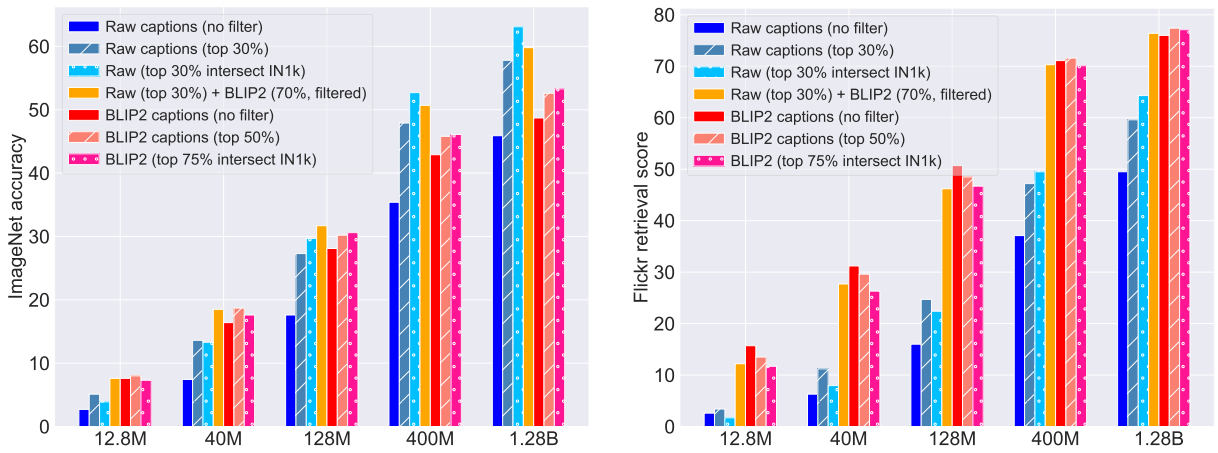


Figure 4.8: With access to generated captions, we find that the best data filtering method for ImageNet classification varies with the scale of the candidate pool; however, when it comes to retrieval, training on synthetic captions is beneficial across all scales. We apply select baselines from Section 4.6 to a range of candidate pool sizes, and find that the best method on Flickr retrieval always involves synthetic captions (right plot). On ImageNet (left plot), selecting meaningful images (e.g., those that lie close to the ImageNet train set in the embedding space) becomes increasingly important at larger scales (see dotted versus striked columns). As the data pool size increases, using BLIP2 captions seems to yield diminishing returns, possibly due to the saturation of text diversity obtained from image captioning models.

To give some intuition for this result, we offer two candidate hypotheses:

- As noted in Section 4.7, both caption noise and diversity are important considerations for performance. Noise level, measured by average image-text cosine similarity, stays about the same across all scales for each training distribution. In contrast, the diversity gap between model-generated text and web-scraped text may become more significant with increasing data quantities. We repeat the caption quality analyses from Section 4.7 with varying random subset size, and find that when using the number of unique nouns and unique trigrams as proxies for text diversity, generated captions exhibit a worse scaling trend than raw captions (Appendix Figure C.4).

- Image quality becomes increasingly important at larger scales:

  (i) from 12.8M to 128M scale, training on the *entire candidate pool* with BLIP2 captions outperforms competitive filtering baselines done on raw data (e.g., Raw (top 30%)). This is not the case for larger scales.

  (ii) starting from 128M scale, baselines that also curate image content (i.e., intersection of top

CLIP score examples and those that lie in clusters close to the ImageNet1k train set) consistently outperform baselines that involve only CLIP score filtering, using either raw or BLIP2 captions.

Exact performance numbers can be found in Appendix C.4, Table C.2. Overall, we find that given a fixed training budget, making more datapoints useful by carefully replacing noisy raw captions with synthetic captions—e.g., RAW (TOP 30%) + BLIP2 (70%, FILTERED) versus RAW (TOP 30%)—still offers classification and retrieval gains across *all* scales. However, for synthetic captions to continue to perform competitively on ImageNet at larger data regimes, we need to start paying attention to image content, as well as enhancing the diversity of the generated text.

4.9 Discussion

In this work, we demonstrate the effectiveness of synthetic captions in improving caption quality for multimodal training, as well as enhancing certain capabilities of the resulting model (e.g., retrieval). Notably, we find that fine-tuning general-purpose models towards the task of image captioning actually makes them less effective at producing good captions for CLIP training. Our experiments with various data pool sizes, ranging from 12.8M to 1.28B image-text pairs, show that including generated captions in the training data can be highly effective at `small` and `medium` scales. However, with larger data quantities, the diversity gap between model-generated and web-scraped text begin to hinder performance gains, and it also becomes increasingly harder to obtain state-of-the-art ImageNet accuracy by just improving text supervision alone.

Limitations. Our experiments do not involve an exhaustive list of image captioning systems currently available. Given a captioning model of sufficient capability—i.e., it can generate captions for training CLIP to reach a good performance—a major theme of our work is understanding how to combine signals from both raw and synthetic captions, as well as the differences between these two sources of text. We note that even with improved caption quality, multimodal web datasets may still contain harmful stereotypes, some of which have been extensively discussed in prior work [20]. In Appendix C.8, we conduct some preliminary investigation on the change in race and gender bias between training on only raw web-crawled text and training on synthetic captions. Besides,

generated captions also inherit biases from the captioning models, and using these captions to train the next generation of models can amplify the biases. The risks from using model outputs to replace human annotations have been studied in simplified settings in [176, 169].

Future work. Our findings motivate a number of interesting future directions. One concrete question is improving the diversity of generated captions at `large` scale, such as by varying the softmax temperature (we only experiment with $T = 0.75$ at this scale, chosen based on our ablation study at the `medium` scale), or by combining synthetic caption data from multiple image captioning systems. Another direction is proposing new algorithms to combine information from raw and generated captions, beyond what we already investigated in Section 4.6 and Appendix C.4. Future work could also explore using text-to-image generation [152, 128, 157] to create synthetic training images for concepts that are underrepresented in existing captions, in order to boost data diversity and close knowledge gaps in the resulting model.

Chapter 5

Beyond Web Scraping: Recycling Discarded Data for Sustainable Pretraining ♻

5.1 Overview

Scaling laws predict that the performance of large language models improves with increasing model size and data size. In practice, pretraining has been relying on massive web crawls, using almost all data sources publicly available on the internet so far. However, this pool of natural data does not grow at the same rate as the compute supply. Furthermore, the availability of high-quality texts is even more limited: data filtering pipelines often remove up to 99% of the initial web scrapes to achieve state-of-the-art. To address the “data wall" of pretraining scaling, our work explores ways to transform and recycle data discarded in existing filtering processes. We propose **ReWire**, **RE**cycling the **W**eb with gu**I**ded **RE**write, a method to enrich low-quality documents so that they could become useful for training. This in turn allows us to increase the representation of synthetic data in the final pretraining set. Experiments at 1B, 3B and 7B scales of the DCLM benchmark show that mixing high-quality raw texts and our rewritten texts lead to 1.0, 1.3 and 2.5 percentage points improvement

respectively across 22 diverse tasks, compared to training on only filtered web data. Training on the raw-synthetic data mix is also more effective than having access to 2× web data. Through further analysis, we demonstrate that about 82% of the mixed in texts come from transforming lower-quality documents that would otherwise be discarded. **ReWire** also outperforms related approaches of generating synthetic data, including Wikipedia-style paraphrasing, question-answer synthesizing and knowledge extraction. These results suggest that recycling web texts holds the potential for being a simple and effective approach for scaling pretraining data. We make our high-quality synthetic data publicly available at https://huggingface.co/datasets/facebook/recycling_the_web.

5.2 Introduction

Over the past few years, large language models (LLMs) have rapidly improved on various benchmarks. This progress was driven largely by scaling up model size, training FLOPs, and in particular, dataset size [80]. For instance, Llama-3 was pretrained on 15T tokens sourced from publicly available data [67], while the previous generation of Llama models was only trained on 2T tokens [180]. The vast quantity of training tokens so far is obtained primarily from internet crawls containing billions of web pages [132, 190, 170], made publicly available by Common Crawl.

While compute resources can scale in accordance with scaling laws and improved hardware efficiency, the growth of public human-generated texts has been less sustainable [110]. Villalobos et al. [183] posit that the current rate of LLM development will exhaust the available stock of internet data between 2026 and 2032. Despite growing concern that LLM pretraining is hitting such a "data wall", existing work on data curation still finds it necessary to discard the majority–sometimes up to 99%–of the data collected to ensure quality and state-of-the-art downstream performance [98, 131]. As we approach the "data wall" while throwing away 99% of web-crawled data, a fundamental question arises: *can we recycle documents that have been discarded by quality filters to make them useful for pretraining?*

Existing work has started exploring different directions to address the impending data bottleneck. For example, we can go beyond the public internet data and obtain licensed, hard-to-access sources

(e.g., Reddit and news sites). However, while on average web crawls are of lower quality than these curated sources, previous research has shown that after enforcing quality control, the former can still dominate the latter in terms of token quantity and contribution to downstream performance [195]. Another line of work proposes relaxing or changing the curation strategies to recover raw documents that have been removed by previous quality filters [172, 121]. In addition, generating synthetic data for certain skills or formats has also been studied to increase the token availability [70, 105, 116].

Our work combines the two aforementioned strategies: synthetic data generation and recycling discarded documents. We propose **RE**cycling the **W**eb with gu**I**ded **RE**write (**ReWire**), which involves taking all documents that are of moderate quality (i.e., having passed some rule-based filters), using an LLM to identify the purpose of the text content, and then asking the LLM to come up with an improved document conditioned on chain-of-thought reasoning. Unlike most existing work on synthetic data, our approach specifically targets the vast quantity of low-quality documents that are somewhat informative but still not considered high-quality by existing filters. We use LLM's knowledge and reasoning capabilities to recycle these documents and add them back to the training pool. The overall data generation pipeline is described in Figure 5.2.

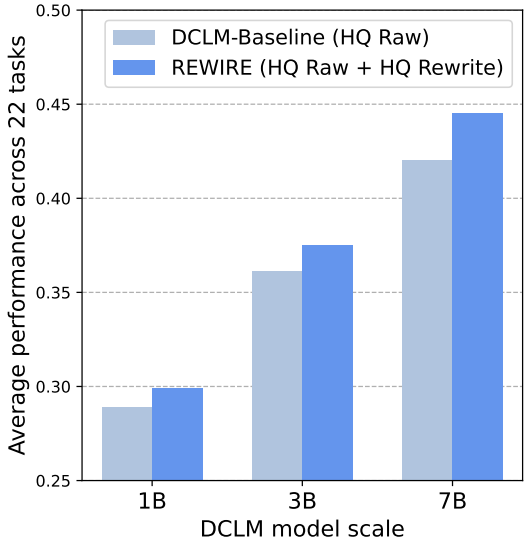


Figure 5.1: ReWire offers increasing performance gains as we scale up model size and training token budget. Our experiments simulate the setting in which high-quality texts are limited and the large token budget (set to be Chinchilla-optimal in this figure) necessitates training on the same filtered dataset multiple times. On average across 22 tasks from DCLM's CORE [98], mixing in the same amount of synthetic data as that of high-quality web data ("HQ Raw + HQ Rewrite") consistently outperforms training on only the latter ("HQ Raw").

Through extensive experiments at both 1B, 3B and 7B model parameter scales, we show that pretraining models on a combination of high-quality web-crawled data and high-quality rewritten data outperforms using the former alone (Section 5.4). Averaged across 22 tasks of the DataComp-LM benchmark [98], the raw-synthetic data mixture improves performance by 1.0, 1.3 and 2.5

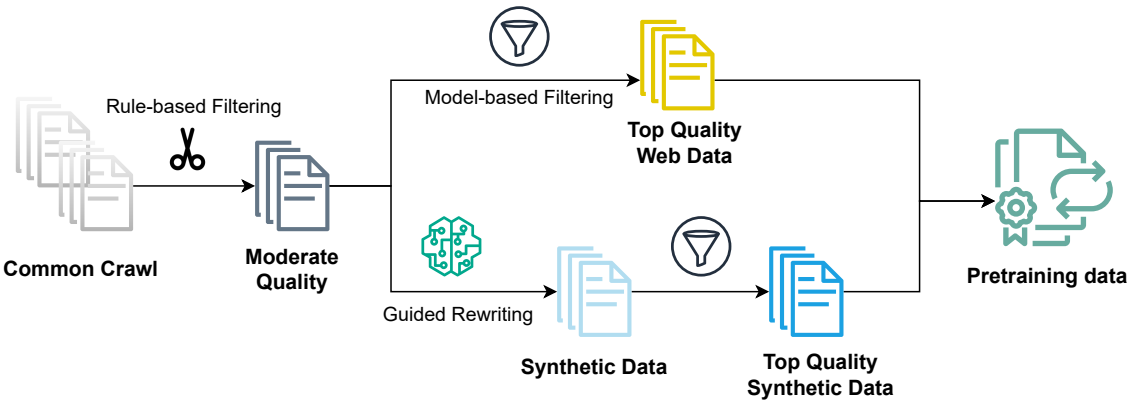


Figure 5.2: The REWIRE pipeline. We start with web documents from Common Crawl that has undergone some filtering (i.e., RefinedWeb heuristics [132]), and thus are at least of moderate quality. State-of-the-art data curation approach, e.g. DCLM-Baseline [98], applies further model-based filtering to retain only top-quality documents for pretraining. Our pipeline takes moderate-quality documents and prompts an LLM to do guided rewriting to generate improved versions of these documents. Finally, we select only high-quality synthetic documents and combine them with the DCLM-Baseline texts to form the final pretraining dataset.

percentage points, at 1B, 3B and 7B scales respectively. The performance benefits of adding rewritten texts also hold across different ways of selecting high-quality raw documents. Furthermore, we demonstrate that the accuracy level achieved by combining raw and synthetic data matches that of using 2× more raw data (Table 5.1). We verify that our best baseline indeed contains a significant amount of synthetic data "recycled" from low-quality documents (Section 5.5.1).

Finally, we compare our rewritten data to three other variations of synthetic data from recent work [116, 172]: extracted knowledge and diverse question-answer pairs synthesized from high-quality documents, Wikipedia-style rephrasing from low-quality ones. We show that **ReWire** generates more diverse synthetic data (Section 5.5.3), which in turn contributes to higher performance on the DCLM benchmark (Table 5.1).

5.3 Experiment Setup

5.3.1 Data Pool

We seek to simulate *long token horizon training* [172], a setting in which high-quality data is limited and the large token compute budget necessitates seeing the same samples multiple times during

training. As Muennighoff et al. [121] find that there are diminishing returns after four epochs compared to training on more unique tokens, we limit the number of sample repeats to be at most 4 in our main experiments. Appendix D.4 contains additional experiments with larger training token budget, thus also making the data repetition rate higher (at most 8 - 10 times).

We fix the starting pool to be DCLM-RefinedWeb [98], Common Crawl data that has passed the initial rule-based quality filters from RefinedWeb [132] (e.g., repetition filter, page length filter, URL filter, etc.) and global deduplication, but has not gone through model-based filtering. With this pool of moderate-quality data, DataComp-LM [98] futher selects only the top 10% based on scores from a fastText classifier [87]. This results in DCLM-Baseline.

Following the token budget set by DCLM at model sizes of 1B, 3B and 7B parameters, we fix the starting pool size to be 72B, 140B and 345B tokens respectively. This is to ensure that even after aggressive pruning, the high-quality data is repeated only at most 4 times. For instance, if the training budget is 28.8B tokens seen at 1B scale, choosing the top 10% based on fastText scores from a starting pool of 72B tokens would leave us with 7.2B tokens. We also experiment with relaxing the filtering threshold (e.g., by selecting the top 20% instead of top 10%) so that there are more unique tokens left after filtering.

5.3.2 Guided Rewriting

The central hypothesis of our **ReWire** framework is that web documents contain diverse content and knowledge, but the writing structure can make them not coherent or elaborate enough to serve as informative pretraining examples. Inspired by recent work on leveraging the meta-cognitive capabilities of state-of-the-art LLMs [44], we prompt Llama-3.3-70B-Instruct [67] to perform chain-of-thought reasoning on the original web document, such as identifying the task or purpose of the text, reasoning about the steps needed to achieve the purpose, etc. before generating an improved version of the original document. The full prompt we use can be found in Section D.2. We apply the same rewriting process to all documents in the starting pool (DCLM-RefinedWeb).

To control the quality of the generations, we further apply model-based filtering to the synthetic data (Figure 5.2). Following DCLM [98], we train a fastText classifier [87] on 400K documents split evenly between positive and negative classes. The positive data is the same as used in DCLM, which includes synthetic instruction data from OpenHermes 2.5 [177] (OH-2.5) and high-scoring posts from the `r/ExplainLikeImFive` (ELI5) subreddit. The negative data are random samples selected from our rewriting generations. Similar to what was done to obtain DCLM-Baseline, we also aggressively filter all the rewritten outputs and only use the top 10% based on the scores of our customized fastText classifier.

5.3.3 Training & Evaluation

Following the same protocol as DCLM, we fix the training hyperparameters and total budget (i.e., number of samples seen) to match what was reported in previous work [98]. We train all models using the Lingua framework [182]. We mainly experiment with 1B-1x, 3B-1x and 7B-1x model scales (1x refers to the Chinchilla multiplier), using Llama-2 architecture and tokenizer as the backbone [180]. We set the model parameters (e.g., number of layers, number of heads) to be the same as DCLM's. More training details can be found in Appendix D.1.

For evaluation, we report the same metrics as DCLM, i.e. MMLU *5-shot accuracy* [76] as well as CORE *centered accuracy* averaged over 22 tasks (e.g., HellaSwag [206] and ARC-easy, ARC-challenge [34]). To compute centered accuracy, each task's performance is linearly rescaled so that 0 corresponds to random guessing and 1 corresponds to perfect accuracy. Li et al. [98] have shown that CORE metric offers a low-variance signal even at small scales. More descriptions of the 22 tasks can be found in the DCLM paper.

5.4 Results

We provide our main result in Table 5.1, which demonstrates that **ReWire** achieves the best average performance across 22 tasks of CORE.

5.4.1 Baselines

As described in Section 5.3.1, we start with a fixed pool of data randomly sampled from DCLM-RefinedWeb [98] and filter with DCLM's fastText classifier to select the highest-quality documents. For comparison with another variation of high-quality web texts, we also experiment with the data released by PreSelect [167], which is curated from the same pool, DCLM-RefinedWeb, but with a different fastText classifier trained to classify a document's predictive strengths of model downstream capabilities. For comparison with other variations of synthetic data, we experiment with the data released by Nemotron-CC [172][1], which contains multiple augmented versions of the same web-crawled data pool, generated using different prompts depending on how high-quality the original web document is.

Below we describe in details the baselines from Table 5.1:

- `Raw text (top 10%)`: We rank examples from the starting pool by the scores from DCLM's fastText classifier [98] and select the top 10% highest-scoring texts. This results in the same data distribution as the final `DCLM-Baseline` pool published by DCLM.
- `Raw text (top 20%)`: Here the fastText filtering threshold is relaxed, allowing for relatively more unique tokens to be included which could benefit multi-epoch training. This is reflected in the final dataset size being double that of `Raw text (top 10%)`.
- `Rewritten text (top 10%)`: We rank all the synthetic data resulting from guided rewriting by the scores from our fastText classifier trained in a similar fashion to DCLM's (Section 5.3.2). We then select top 10% of the rewritten texts.
- `Raw text (top 10%) + Rewritten text (top 10%)`: We combine the highest-quality texts from the original starting pool as well as the same pool after being transformed with guided rewriting. This means that some of the selected documents will have both the web-scraped and the rewritten versions included, while some will only have either version. We analyze the overlap between the two distributions later in Section 5.5.1.

[1] https://data.commoncrawl.org/contrib/Nemotron/Nemotron-CC/index.html

- `PreSelect/ PreSelect + Rewritten text (top 10%)`: Since the data released by Shum et al. [167] is already filtered to be the top 10% of the DCLM-RefinedWeb pool based on their quality metric, we experiment with training on the open-source curated data directly, as well as mixing it with our highest-quality rewritten data.
- `Nemotron-CC High-quality (HQ) diverse QAs`: Su et al. [172] prompt an LLM to ask questions in various forms (e.g., yes/no, open-ended, multi-choice) about factual information in a document and provide the correct answers. They apply this prompt only to high-quality web texts from DCLM. Since the open-source data for Diverse QAs split already contains the raw documents followed by QAs, we randomly sample data from the split until the token count from the raw texts (excluding QAs) matches the token budget. We note that despite starting from the same pool (DCLM), due to differences in filtering criteria, Nemotron-CC HQ web documents could differ significantly from the documents selected for the `Raw text (top 10%)` baseline.
- `Raw text (top 10%) + Nemotron-CC HQ extracted knowledge`: For this synthetic data variation, Su et al. [172] prompt LLMs to convert existing knowledge in the raw text to some standard technical formats (i.e., textbooks and Wikipedia) and discard uninformative content (i.e.,"only restate what is already in the text"). The open-source data for this split does not contain the corresponding original documents, so we combine `Raw text (top 10%)` with the extracted knowledge synthetic data. As previous work only applies this prompt to high-quality documents, which are assumed to be limited in quantity, the number of tokens generated from extracted knowledge therefore is also limited. To simulate this setting, we fix the number of extracted knowledge samples to be the same as the number of documents in `Raw text (top 10%)`.
- `Raw text (top 10%) + Nemotron-CC Medium-quality (MQ) Wikipedia rephrasing`: For relatively lower quality documents from DCLM, Su et al. [172] follow the method proposed by Maini et al. [116] and use the LLM to solely change the writing style to be Wikipedia-like, instead of using LLM "as a knowledge bank". This is similar to our approach in the sense that the synthetic data comes from a disjoint set of documents that are not selected for pretraining. We randomly

sample Wikipedia-rephrased segments until we reach the same token quantity as our `Rewritten text (top 10%)` baseline.

Our setup assumes a raw data bottleneck, i.e., the size of the starting pool is fixed. Given a limited number of moderate-quality (potentially usable) documents from this pool, we compare ways to filter and enrich the existing data. However, we also experiment with the setting where the starting pool size is doubled (e.g., moving from 72B to 144B tokens in total at the 1B scale, see the shaded rows in Table 5.1). If the stock of web data grew by twice as much (144B), what performance could we expect from aggressively filtering raw data alone, and could synthetic data help close the performance gap at the current data scale (72B)?

5.4.2 Performance on DCLM Benchmark

In Table 5.1, we find that while training on synthetic data alone still lags behind training on highest-quality raw texts, using *mixed* data distributions (i.e., `Raw text (top 10%) + Rewritten text (top 10%)` or `PreSelect + Rewritten text (top 10%)`) outperforms using only the corresponding filtered web data. It is worth noting that our synthetic data significantly boosts MMLU performance without being rewritten specifically for this task (i.e., by converting into QAs or by selecting topics). At the 3B scale, mixing raw and synthetic data still improves MMLU, but tuning the mixing ratio becomes more important for improving average performance across a range of tasks. We note that the gain in average performance (relative to using only high-quality raw data from the same pool) increases with model and training scale: +1.0 percentage points (pp) at 1B-1x, +1.3pp at 3B-1x and +2.5pp at 7B-1x.

We also experiment with settings that simulate the large data regime beyond Chinchilla-optimal as are often adopted in practice [180, 67]. There, we train on the smallest-sized dataset for more than 4 epochs. Table D.3 in section D.4 reports results for the *1B-5x: 144B tokens seen*, ∼3-10 epochs setting, as well as the *7B-2x: 276B tokens seen*, ∼4-8 epochs setting. The same findings hold: mixing rewritten data with high-quality web data brings significant improvement on both MMLU and CORE, e.g. +7.3% on MMLU and +2.3pp on average (at 7B scale) compared to training on DCLM-Baseline data alone.

Baseline name	Pool size	Data size	MMLU↑	CORE↑
1B-1x Setting: 28.8B tokens seen				
Raw text (top 10%), DCLM-Baseline [98]	72B	7.2B	0.266	0.289
Raw text (top 20%)	72B	14.4B	0.249	0.282
Rewritten text (top 10%)	72B	7.2B	0.266	0.270
Raw text (top 10%) + Rewritten text (top 10%)	72B	7.2B + 7.2B	0.268	**0.299**
Raw text (top 10%), 2×	144B	14.4B	0.252	0.294
Raw text (top 20%), 2×	144B	28.8B	0.243	0.291
PreSelect [167]	72B	7.2B	0.250	0.277
PreSelect + Rewritten text (top 10%)	72B	7.2B + 7.2B	0.239	0.284
PreSelect, 2×	144B	14.4B	0.258	0.284
Nemotron-CC HQ diverse QAs [172]	72B	7.2B	**0.299**	0.284
Raw text (top 10%) + Nemotron-CC HQ extracted knowledge	72B	7.2B + 2.7B	0.250	0.295
Raw text (top 10%) + Nemotron-CC MQ Wikipedia rephrasing	72B	7.2B + 7.2B	0.248	0.285
3B-1x Setting: 55.9B tokens seen				
Raw text (top 10%), DCLM-Baseline [98]	140B	14B	0.251	0.362
Raw text (top 20%)	140B	28B	0.240	0.363
Rewritten text (top 10%)	140B	14B	0.286	0.317
Raw text (top 10%) + Rewritten text (top 10%)	140B	14B + 14B	0.285	0.364
Raw text (top 10%) x 0.6 + Rewritten text (top 10%) x 0.4	140B	14B + 14B	0.274	**0.375**
Raw text (top 10%), 2×	280B	28B	0.256	0.369
Raw text (top 20%), 2×	280B	55.9B	0.254	0.360
PreSelect [167]	140B	14B	0.255	0.353
PreSelect + Rewritten text (top 10%)	140B	14B + 14B	0.310	0.367
PreSelect, 2×	280B	28B	0.253	0.360
Nemotron-CC HQ diverse QAs [172]	140B	14B	**0.380**	0.363
Raw text (top 10%) + Nemotron-CC HQ extracted knowledge	140B	14B + 5.3B	0.247	0.364
Raw text (top 10%) x 0.6 + Nemotron-CC HQ extracted knowledge x 0.4	140B	14B + 5.3B	0.261	0.364
Raw text (top 10%) + Nemotron-CC MQ Wikipedia rephrasing	140B	14B + 14B	0.258	0.360
Raw text (top 10%) x 0.6 + Nemotron-CC MQ Wikipedia rephrasing x 0.4	140B	14B + 14B	0.268	0.368
7B-1x Setting: 138B tokens seen				
Raw text (top 10%), DCLM-Baseline [98]	345B	34.5B	0.326	0.420
Raw text (top 10%) + Rewritten text (top 10%)	345B	34.5B + 34.5B	**0.447**	**0.445**
Raw text (top 10%), 2×	690B	69B	0.356	0.425

Table 5.1: Main results on the DCLM benchmark. We report the performance of training with different datasets on MMLU and on average across 22 tasks of CORE [98]. Accuracies that are near random-chance performance are in gray. Across all three model and training budget scales, we observe that training only on high-quality synthetic data underperforms training on high-quality raw texts. However, combining these two subsets consistently boosts MMLU and overall performance, matching the accuracies of training on 2× more high-quality raw data (shaded rows). **ReWire** is also more effective than other synthetic data variants [172] at improving average performance.

Furthermore, we compare our rewritten data to three versions of Nemotron-CC [172], which are representative, related approaches of synthetic data generation. We find that **ReWire** consistently yields the best performance on average when mixing with highest-quality raw texts. Out of the three variations from Su et al. [172], the extracted knowledge format is the most helpful for increasing CORE performance. Even though Nemotron-CC's extracted knowledge and Wikipedia rephrasing do not help with MMLU, their diverse QAs are especially effective. This is potentially due to the alignment of the data format, as MMLU is made up of multiple-choice questions [76]. Overall these results suggest that **ReWire** is more effective at complementing curated natural texts.

Finally, at all model parameter scales that we experiment with, combining carefully filtered raw and rewritten texts can match, if not outperform, the performance level of training on 2× more high-quality web documents, i.e. as if we had access to a starting pool with 2× more raw data. For instance, at the 1B model scale, using the `Raw text (top 10%) + Rewritten text (top 10%)` baseline from a starting pool of 72B tokens yields 26.8% accuracy on MMLU, and 29.9pp on average, while using only `Raw text (top 10%)` but filtered from a pool of 144B tokens scores near random-chance accuracy on MMLU and 29.4pp on average (Table 5.1). The same finding holds when we swap out high-quality raw data from DCLM-Baseline with `PreSelect` [167]. This suggests that synthetic data from **ReWire** could double the token yield for multi-epoch training.

5.5 Rewriting Quality Analysis

5.5.1 Influence of Raw Text Quality on the Recycled Text Quality

Given that our rewriting is conditioned on the content of some web-scraped document, a natural question arises: *Does the quality of the initial draft (i.e., raw text) affect the quality of the rewritten outputs?* We find that there is little to no correlation between the two. In Figure 5.3, for 10K documents randomly selected from the starting pool, we plot the quality scores of the raw texts output by DCLM's fastText classifier [98], as well as the quality scores of the corresponding rewritten versions output by our own fastText classifier (described previously in Section 5.3.2). We observe no obvious trend between the two values. Computing the Spearman rank-order correlation gives a

coefficient of 0.179 with a p-value of 6.52e−73, suggesting that the quality of the two text versions shares only a slightly monotonic relationship.

Consequently, we analyze the overlap between `Rewritten text (top 10%)` and `Raw text (top 10%)` datasets (see Section 5.4.1) and find that they only have $\sim 18.3\%$ documents in common. In other words, for the best baselines in Table 5.1 that combine these two high-quality data distributions, 18.3% of the selected documents have both the web-scraped and the corresponding rewritten versions included in the training data, while the remaining 81.7% of the new documents mixed in are *recycled* from low-quality web texts that normally would be excluded from training.

5.5.2 How Is ReWire Different from Rephrasing?

Here we clarify how the generations from our method differ from existing approaches of data augmentation, such as generating diverse question-answer pairs (QAs) from a document [172] or paraphrasing the text in a certain style [116]. Such approaches often do not go beyond the content provided by the raw documents, only transforming the format of the available facts. In contrast, our pipeline treats the web-scraped texts as initial drafts and allows LLMs to fill in the gaps or expand on the existing points to derive an improved version. Consequently, while the rewritten version would stay on topic, it is possible that new knowledge is added to the text, changing its semantics to a large extent.

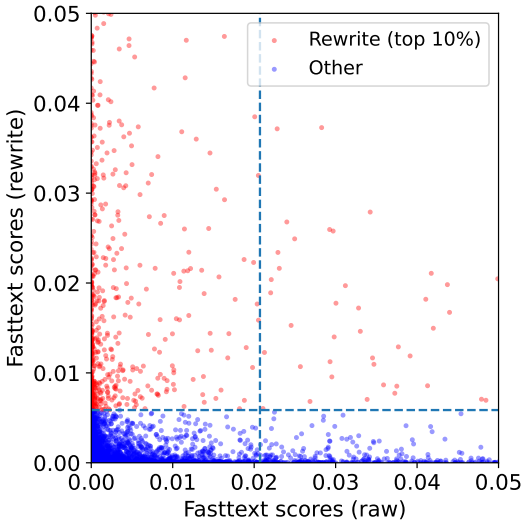


Figure 5.3: Quality of original web text and quality of the corresponding rewritten text show almost no monotonic relationship. We randomly sample 10K documents and plot the distribution of the fasttext scores of the web-scraped version and the rewritten version; the dotted lines represent the filtering thresholds used for each data distribution. We find that there is no significant relationship between the two quality scores (Spearman rank-order correlation=0.179). This suggests that **ReWire** can transform low-quality web texts into high-quality synthetic data.

Following previous work [116], to measure how much the semantic meaning of the raw text is preserved, we compute cosine similarity of the sentence embeddings from different versions of the same document using a

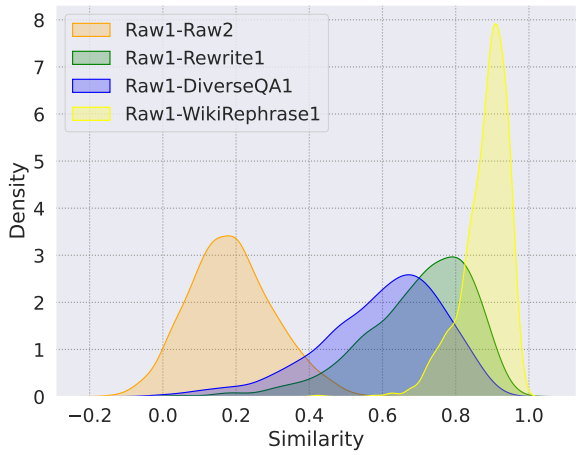


Figure 5.4: Guided rewriting retains the semantic meaning of the web documents to a large extent, but in some cases the content can change significantly. To measure how much the semantics is preserved before and after rewriting, we compute the cosine similarity between the two corresponding text embeddings for 1000 documents, and visualize the similarity distribution. We find that the average semantic similarity is high, though still lower than the similarity obtained from Wikipedia-style rephrasing. This suggests that **ReWire** involves a combination of paraphrasing and modifying the content of the initial texts.

pretrained BERT model trained with SimCSE objective [61]. The distribution of cosine similarities based on 1000 random samples is then visualized using a gaussian Kernel Density Estimator (Figure 5.4). The baseline similarity level is captured in `Raw1-Raw2`, computed based on pairs of randomly selected web documents from our pool. Similar to [116], we also find that Wikipedia-style rephrases convey similar meaning to their real counterparts without adding information (`Raw1-WikiRephrase1`). The cosine similarities between original web texts and **ReWire** texts are generally higher than the random baseline, but lower than rephrasing. Based on inspection of pairs with low similarities (e.g., < 0.4), we find that the original text often is short and contains little information, or contains a lot of information that are not closely related (e.g., product listings). In this case, the LLM is likely to perform more content generation and modification. Conversely, for pairs with high similarities (e.g., > 0.8), the model mostly does paraphrasing. Appendix D.3 provides examples of these two scenarios.

5.5.3 Assessing Text Diversity

N-gram based metric We randomly sample documents from the high-quality raw text subset, as well as from different synthetic data distributions (Section 5.4.1), and compute the total number of unique bigrams (Figure 5.5). Since the data generation methods are all applied to individual documents, we fix the number of documents sampled (left panel), and observe how the word diversity scales with the document quantity. We find that while synthetic data still lags behind raw data in general, our rewritten texts are similarly diverse compared to Wikipedia rephrasing, and are more

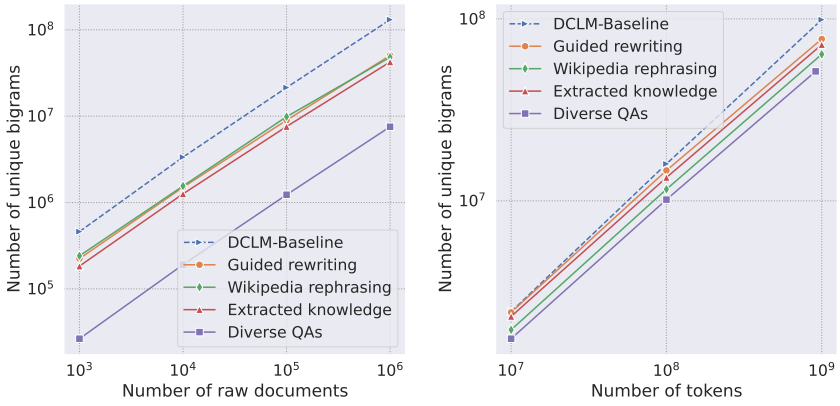


Figure 5.5: How word diversity scales for high-quality web data and different synthetic data variants. We fix the number of documents *(left)* as well as tokens *(right)* randomly sampled from each dataset and compute the number of unique bigrams. In both cases, raw web texts appear to contain the most diversity, followed by our guided rewriting texts and Wikipedia rephrasing [172].

diverse than extracted knowledge and diverse QAs. As the generation length can be a confounder to how many bigrams there are (Appendix Table D.4), we also fix the total number of tokens and randomly sample texts from each data distribution until the token quota is reached (right panel). In this case, web-crawled documents still exhibit the best scaling trend, but our rewritten data comes in a close second, being slightly more diverse than the other three synthetic data types.

Embedding visualization In Figure 5.6, we show the t-SNE plot of 1000 document embeddings from each data distribution, randomly chosen and embedded with HuggingFace's SentenceTransformers. While Wikipedia-rephrased documents mostly form a separate cluster, extracted knowledge samples share some similarity with both our rewritten data and the high-quality raw texts. Aside from that, our rewritten texts appear sufficiently distinct from the filtered DCLM texts, suggesting that combining the two distributions can increase the overall data coverage.

5.6 Related Work

Data curation for LLM pretraining Previous work has shown that the base model's downstream performance is highly dependent on the preprocessing and filtering of the initial data pool. However, the specific choice of filters as well as the deduplication method differ across pretraining corpora [132, 170, 190, 35]. For instance, C4 [143] removes non-English pages, applies several rule-based filters (e.g., discarding pages that contain "bad words") and deduplicates over three-line windows. Over time,

model-based filtering gains popularity as the "quality" metric becomes harder to define [155, 191, 204]. For example, RedPajama [190] utilizes a classifier trained to distinguish Wikipedia-level content from random web texts. Penedo et al. [131] use an LLM to annotate some seed data, and train a linear regression model on the annotated scores to rank all documents in a pool based on their educational values.

DataComp-LM [98] unifies some of these approaches and offers a testbed for controlled dataset experiments, in order to ablate the filtering decisions made in previous work. Our work makes use of DCLM-Baseline, the corpus open-sourced by DCLM that has been cleaned with heuristic-based filters but without any model-based filtering.

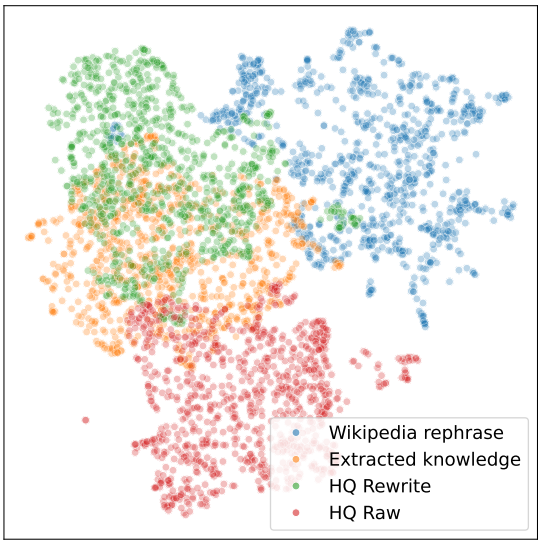


Figure 5.6: Visualization of similarities among different data distributions based on low-dimensional embeddings. We observe that our high-quality rewritten texts, Nemotron-CC's Wikipedia rephrasings from Su et al. [172] and filtered DCLM raw texts are sufficiently distinct from one another. In contrast, Nemotron-CC's extracted knowledge data is somewhat similar to both the high-quality raw and rewritten texts.

Synthetic text data Prior work has studied ways to augment raw documents and convert the information contained in web-scraped data to different formats. Maini et al. [116] propose paraphrasing web documents in specific styles such as "like Wikipedia" or in "question-answer format", and then training on both real and corresponding stylized data. Su et al. [172] follow up on this work and pick different augmentation prompts for low- and high-quality data, using Wikipedia-style paraphrasing only for lower-quality texts.

Separately, another line of work does not reference any raw document in particular, but optimizes directly for diversity in their selection of topics to distill from LLMs. The topics can be captured via personas [62], or story features and vocabulary [47]. The Phi model series [70, 105, 3] was among the first to demonstrate the effectiveness of training on a small amount of high-quality, textbook-like data. The authors seed the data generation with thousands of carefully chosen topics to generate

high-knowledge and high-reasoning content. Similarly, Cosmopedia [8] also source the seed topics from both curated sources (e.g., Khan Academy) as well as web data. Since we do not target any technical skill or topic, we view this line of work as complementary and not comparable baselines to our method.

It is also possible to create synthetic tokens by inferring new knowledge from existing raw data. Yang et al. [198] prompt LLMs to build a knowledge graph from a small set of books and articles, and create training data based on the node connections in the graph. Ruan et al. [154] use an LLM to augment pretraining math data with the corresponding latent "thoughts". The authors find that this improves learning efficiency as well as performance on math benchmarks. Both of these prior works perform synthetic data generation at much smaller scales (455M - 1.1B tokens) compared to ours, focusing only on the continual pretraining setting and targeted capabilities (e.g., reading comprehension and math).

Most recently, [55] propose rewriting math and coding data at large scale (2.3B - 16.1B) for pretraining. The LLM-driven rewriting pipeline is designed for these specifc data types, e.g. by asking the LLM to enhance code readability following a published style guide.

Our method lies at the intersection of data augmentation and knowledge expansion. We specifically target discarded low-quality documents and prompt an LLM to generate an improved version for each of them. To the best of our knowledge, our work is the first to produce *general-purpose* synthetic data at a large scale for pretraining, such that we can mix synthetic tokens and web tokens with 1:1 ratio while still improving performance overall.

5.7 Discussion

In this work, we propose "recycling the web" with guided writing (**ReWire**), a method to transform low-quality web documents into useful training data. Experiments on the DCLM benchmark across three different scales show that our synthetic data is effective at boosting the quality of the pretraining web dataset, in turn yielding higher performance on MMLU and on average across 22 diverse tasks. Similar to prior work that highlights the risk of model collapse when training on

only synthetic data [65, 168], we design **ReWire** with the goal of complementing naturally existing internet data, not replacing it. As our method neither assumes knowledge of downstream tasks, nor is domain- or topic-specific, we consider it complementary to existing work that targets highly technical and educational synthetic data [8, 105, 5]. Overall, **ReWire** shows promise as a simple and effective solution to address the "data wall" of scaling pretraining.

Limitations Since our rewriting pipeline relies on the LLM's knowledge and reasoning capabilities, as opposed to just using it to rephrase poorly written documents, we resort to a moderate-sized LLM (i.e., Llama-3.3-70B-Instruct). Consequently, the cost of data generation is higher than other related approaches [116, 172]. We report our compute costs in Appendix D.2. However, we argue that this high cost of creating synthetic data can be amortized by using the resulting data for training multiple models and for more epochs. Furthermore, as with most synthetic data approaches, there is always a risk of increasing hallucination in the final training set, especially since our method allows the LLM to change the content presented in the raw documents. Future work could include additional filters to verify the truthfulness of the information in generated texts. Through evaluations targeting factuality (Table D.2 in Appendix D.4), we find that adding **ReWire** generations to the pretraining set does not harm, but rather improves truthfulness and knowledge capabilities of the resulting model.

Future work We do not experiment with multiple filtering strategies for the rewritten data, but rather follow the setup from Li et al. [98] and use a fastText classifier. Future work could study how to better select high-quality data from all synthetic generations, e.g. by directly optimizing for data diversity via cluster sampling [207] or domain balancing [192]. Another interesting direction would be to go beyond point-wise data filtering and select synthetic data that is complemetary to the existing training set. For instance, Yu et al. [204] propose a subset selection technique that optimizes for group-level influence prediction. Last but not least, future work could extend **ReWire** with fine-grained controls: prompting LLMs to combine different text dimensions (e.g., styles, formats, skills, etc.) while still conditioning on the original web-scraped content, so as to promote further diversity in data generation.

Chapter 6

Conclusion

6.1 Summary of Contributions

While data curation is critical to every stage of the training pipeline, this thesis focuses primarily on the pretraining stage where curation decisions often have far-reaching downstream impacts. The works presented in this thesis contribute to *the science of data curation*, offering principled and novel methods that address the growing complexity of training large-scale foundation models.

The key findings from my research are as follows:

- In the early era of using web-scale training corpora, pioneered by GPT-2 [142] and CLIP [141], our work was among the first to show that indiscriminately mixing data of varying quality tiers dilutes the training set and leads to worse generalization (Chapter 2). This helps raise broader awareness of data quality as a foundational consideration, at a time when the field was primarily focused on scaling the quantity of pretraining data.
- As attention to data quality grew, the field witnessed a rapid proliferation of data filtering and selection methods. I then helped build benchmarks that systematically evaluate data filters in a controlled environment [56, 98]. I also challenged the design decisions of popular data filters by showing that they overfit to narrow data distributions and inadvertently exclude important, diverse parts of the web, such as non-English content (Chapter 3). Such excluded data, when

thoughtfully incorporated, can improve model performance as a whole, including on standard vision benchmarks that often define state-of-the-art.

- Building on the growing capabilities of web-pretrained models, I proposed image recaptioning as a novel approach to improving data quality (Chapter 4), shifting the multimodal curation paradigm from filtering existing data to actively generating higher-quality alternatives. Shortly after the release of this work, synthetic captions were widely adopted in training image generation models [163, 17].

- While aggressive data filtering has proven effective for achieving state-of-the-art performance [98], it exacerbates an emerging challenge: the supply of high-quality human-written texts on the internet is not growing fast enough relative to the demands of model scaling. To address the impending token scarcity, I proposed a novel approach to generate high-quality synthetic data at scale, by rewriting web-scraped documents that are not selected by existing data filters (Chapter 5). I showed that this approach helps double the effective token yield and offers a more sustainable path forward for pretraining. I publicly released 44B synthetic data tokens to support future research in this direction.

Across many of my PhD works, a broader theme underlies my research: *diversity* is itself a first-order principle of large-scale data curation—it is deeply intertwined with and critical to achieving data quality. Existing filtering methods are often exclusionary by design, discarding vast portions of potentially valuable web data without sufficient rigor. My work advocates for a more *inclusive* curation paradigm, studying which of the discarded data can be recovered, transformed, and reincorporated into the pretraining set.

6.2 Future Work

When we first established quality as a foundational principle for large-scale data curation (Chapter 2), many open questions remained about how to construct optimal pretraining datasets at scale—refer to [124] for an extended discussion. In the years since, the data-centric research community

has matured considerably, with several of the challenges we identified having been substantially addressed, especially along the axes of data filtering and synthetic data generation.

However, a number of important frontiers remain as pretraining datasets become more heterogeneous in data types and distributions, and as training pipelines grow more complex. More specifically, I am excited to explore the following directions in the near future:

- **Data mixing**: while early work determined optimal mixtures for pretraining a target model by training many smaller proxy models [108, 196], more recent research has shown that such mixtures do not necessarily generalize across model sizes [113]. This highlights the need for more robust and efficient approaches to mixture design that can generalize across model scales. Furthermore, while data mixing has received considerable attention in the language model setting, it remains relatively underexplored for multimodal models, where the presence of fundamentally different data types (e.g., interleaved documents, image-caption pairs, text-only data) poses unique challenges for mixture optimization [153, 209].
- **Inclusion of skill-based data in pretraining**: there has been a steady trend of moving data types traditionally reserved for post-training into earlier stages [32, 187, 6, 106], as such data becomes available at scale. Studying how to effectively incorporate current and next-generation skill-based data into the existing pretraining distribution could enable models to acquire procedural knowledge more deeply, rather than treating skills as surface-level behaviors to be patched in later training stages.
- **Interactions between pretraining and post-training data curation**: relatedly, previous work has shown that the composition and quality of pretraining data fundamentally shapes the model's capacity to benefit from post-training, with decisons made early in the pipeline often proving difficult to correct later [115, 210]. This calls for further investigation into the interplay between pretraining and post-training data curation, and how both processes can be jointly optimized rather than treated as independent stages.

Bibliography

[1] Common crawl. https://commoncrawl.org/. Accessed: 2022-05-18.

[2] Amro Abbas, Kushal Tirumala, Dániel Simig, Surya Ganguli, and Ari S Morcos. 2023. Semdedup: Data-efficient learning at web-scale through semantic deduplication. *arXiv preprint arXiv:2303.09540*.

[3] Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J Hewett, Mojan Javaheripi, Piero Kauffmann, et al. 2024. Phi-4 technical report. *arXiv preprint arXiv:2412.08905*.

[4] Harsh Agrawal, Karan Desai, Yufei Wang, Xinlei Chen, Rishabh Jain, Mark Johnson, Dhruv Batra, Devi Parikh, Stefan Lee, and Peter Anderson. 2019. Nocaps: Novel object captioning at scale. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8948–8957.

[5] Syeda Nahida Akter, Shrimai Prabhumoye, John Kamalu, Sanjeev Satheesh, Eric Nyberg, Mostofa Patwary, Mohammad Shoeybi, and Bryan Catanzaro. 2024. Mind: Math informed synthetic dialogues for pretraining llms. *arXiv preprint arXiv:2410.12881*.

[6] Syeda Nahida Akter, Shrimai Prabhumoye, Eric Nyberg, Mostofa Patwary, Mohammad Shoeybi, Yejin Choi, and Bryan Catanzaro. 2025. Front-loading reasoning: The synergy between pretraining and post-training data. *arXiv preprint arXiv:2510.03264*.

[7] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, et al. 2022. Flamingo: a visual language model for few-shot learning. *arXiv preprint arXiv:2204.14198*.

[8] LB Allal, A Lozhkov, and D Cosmopedia van Strien. 2024. how to create large-scale synthetic data for pre-training large language models—huggingface. co.

[9] Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. 2016. Spice: Semantic propositional image caption evaluation. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part V 14*, pages 382–398. Springer.

[10] Anders Andreassen, Yasaman Bahri, Behnam Neyshabur, and Rebecca Roelofs. 2021. The evolution of out-of-distribution robustness throughout fine-tuning. https://arxiv.org/abs/2106.15831.

[11] Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. 2019. Invariant risk minimization. https://arxiv.org/abs/1907.02893.

[12] Shekoofeh Azizi, Simon Kornblith, Chitwan Saharia, Mohammad Norouzi, and David J Fleet. 2023. Synthetic data from diffusion models improves imagenet classification. *arXiv preprint arXiv:2304.08466*.

[13] Hritik Bansal and Aditya Grover. 2023. Leaving reality to imagination: Robust classification via generated datasets. *arXiv preprint arXiv:2302.02503*.

[14] Andrei Barbu, David Mayo, Julian Alverio, William Luo, Christopher Wang, Dan Gutfreund, Josh Tenenbaum, and Boris Katz. 2019. Objectnet: A large-scale bias-controlled dataset for pushing the limits of object recognition models. *Advances in neural information processing systems*, 32.

[15] Loïc Barrault, Fethi Bougares, Lucia Specia, Chiraag Lala, Desmond Elliott, and Stella Frank. 2018. Findings of the third shared task on multimodal machine translation. In *Third Conference on Machine Translation (WMT18)*, volume 2, pages 308–327.

[16] Eric Bauer and Ron Kohavi. 1999. An empirical comparison of voting classification algorithms: Bagging, boosting, and variants. *Machine learning*. https://link.springer.com/article/10.1023/A:1007515423169.

[17] James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, et al. 2023. Improving image generation with better captions. *Computer Science. https://cdn. openai. com/papers/dall-e-3. pdf*, 2(3):8.

[18] Battista Biggio, Igino Corona, Davide Maiorca, Blaine Nelson, Nedim Šrndić, Pavel Laskov, Giorgio Giacinto, and Fabio Roli. 2013. Evasion attacks against machine learning at test time. In *Joint European conference on machine learning and knowledge discovery in databases*, pages 387–402. Springer.

[19] Battista Biggio and Fabio Roli. 2018. Wild patterns: Ten years after the rise of adversarial machine learning. *Pattern Recognition*, 84:317–331.

[20] Abeba Birhane, Vinay Uday Prabhu, and Emmanuel Kahembwe. 2021. Multimodal datasets: misogyny, pornography, and malignant stereotypes. *arXiv preprint arXiv:2110.01963*.

[21] Yonatan Bitton, Nitzan Bitton Guetta, Ron Yosef, Yuval Elovici, Mohit Bansal, Gabriel Stanovsky, and Roy Schwartz. 2022. Winogavil: Gamified association benchmark to challenge vision-and-language models. *Advances in Neural Information Processing Systems*, 35:26549–26564.

[22] Leo Breiman. 1996. Bagging predictors. *Machine learning*. https://link.springer.com/article/10.1007/BF00058655.

[23] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

[24] Minwoo Byeon, Beomhee Park, Haecheon Kim, Sungjun Lee, Woonhyuk Baek, and Saehoon Kim. 2022. Coyo-700m: Image-text pair dataset. https://github.com/kakaobrain/coyo-dataset.

[25] Liangliang Cao, Bowen Zhang, Chen Chen, Yinfei Yang, Xianzhi Du, Wencong Zhang, Zhiyun Lu, and Yantao Zheng. 2023. Less is more: Removing text-regions improves clip training efficiency and robustness. *arXiv preprint arXiv:2305.05095*.

[26] Fredrik Carlsson, Philipp Eisen, Faton Rekathati, and Magnus Sahlgren. 2022. Cross-lingual and multilingual clip. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 6848–6854.

[27] Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. 2021. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3558–3568.

[28] Soravit Changpinyo, Linting Xue, Idan Szpektor, Ashish V Thapliyal, Julien Amelot, Michal Yarom, Xi Chen, and Radu Soricut. 2022. Maxm: Towards multilingual visual question answering. *arXiv preprint arXiv:2209.05401*.

[29] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR.

[30] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. 2015. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*.

[31] Zhongzhi Chen, Guang Liu, Bo-Wen Zhang, Fulong Ye, Qinghong Yang, and Ledell Wu. 2022. Altclip: Altering the language encoder in clip for extended language capabilities. *arXiv preprint arXiv:2211.06679*.

[32] Daixuan Cheng, Yuxian Gu, Shaohan Huang, Junyu Bi, Minlie Huang, and Furu Wei. 2024. Instruction pre-training: Language models are supervised multitask learners. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 2529–2550.

[33] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2022. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*.

[34] Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv preprint arXiv:1803.05457*.

[35] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

[36] Marta R Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, et al. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

[37] Alexander D'Amour, Katherine Heller, Dan Moldovan, Ben Adlam, Babak Alipanahi, Alex Beutel, Christina Chen, Jonathan Deaton, Jacob Eisenstein, Matthew D Hoffman, et al. 2020. Underspecification presents challenges for credibility in modern machine learning. https://arxiv.org/abs/2011.03395.

[38] Luke N Darlow, Elliot J Crowley, Antreas Antoniou, and Amos J Storkey. 2018. Cinic-10 is not imagenet or cifar-10. *arXiv preprint arXiv:1810.03505*.

[39] Terrance De Vries, Ishan Misra, Changhan Wang, and Laurens Van der Maaten. 2019. Does object recognition work for everyone? In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops*, pages 52–59.

[40] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. 2009. Imagenet: A large-scale hierarchical image database. In *Conference on Computer Vision and Pattern Recognition*. https://ieeexplore.ieee.org/document/5206848.

[41] Karan Desai and Justin Johnson. 2021. Virtex: Learning visual representations from textual annotations. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 11162–11173.

[42] Karan Desai, Gaurav Kaul, Zubin Aysola, and Justin Johnson. 2021. Redcaps: Web-curated image-text data created by the people, for the people. *arXiv preprint arXiv:2111.11431*.

[43] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

[44] Aniket Didolkar, Anirudh Goyal, Nan Rosemary Ke, Siyuan Guo, Michal Valko, Timothy Lillicrap, Danilo Jimenez Rezende, Yoshua Bengio, Michael C Mozer, and Sanjeev Arora. 2024. Metacognitive capabilities of llms: An exploration in mathematical problem solving. *Advances in Neural Information Processing Systems*, 37:19783–19812.

[45] Thomas G Dietterich. 2000. Ensemble methods in machine learning. In *International workshop on multiple classifier systems*. https://link.springer.com/chapter/10.1007/3-540-45014-9_1.

[46] Alexey Dosovitskiy, Philipp Fischer, Eddy Ilg, Philip Hausser, Caner Hazirbas, Vladimir Golkov, Patrick Van Der Smagt, Daniel Cremers, and Thomas Brox. 2015. Flownet: Learning optical flow with convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 2758–2766.

[47] Ronen Eldan and Yuanzhi Li. 2023. Tinystories: How small can language models be and still speak coherent english? *arXiv preprint arXiv:2305.07759*.

[48] Desmond Elliott, Stella Frank, Khalil Sima'an, and Lucia Specia. 2016. Multi30k: Multilingual english-german image descriptions. *arXiv preprint arXiv:1605.00459*.

[49] Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, Naman Goyal, Tom Birch, Vitaliy Liptchinsky, Sergey Edunov, Edouard Grave, Michael Auli, and Armand Joulin. 2020. Beyond english-centric multilingual machine translation.

[50] Lijie Fan, Dilip Krishnan, Phillip Isola, Dina Katabi, and Yonglong Tian. 2023. Improving clip training with language rewrites. *arXiv preprint arXiv:2305.20088*.

[51] Alex Fang, Gabriel Ilharco, Mitchell Wortsman, Yuhao Wan, Vaishaal Shankar, Achal Dave, and Ludwig Schmidt. 2022. Data determines distributional robustness in contrastive language image pre-training (clip). In *International Conference on Machine Learning*, pages 6216–6234. PMLR.

[52] Alex Fang, Albin Madappally Jose, Amit Jain, Ludwig Schmidt, Alexander Toshev, and Vaishaal Shankar. 2023. Data filtering networks. *arXiv preprint arXiv:2309.17425*.

[53] Stanislav Fort, Huiyi Hu, and Balaji Lakshminarayanan. 2019. Deep ensembles: A loss landscape perspective. https://arxiv.org/abs/1912.02757.

[54] Yoav Freund and Robert E Schapire. 1997. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*. https://www.sciencedirect.com/science/article/pii/S002200009791504X.

[55] Kazuki Fujii, Yukito Tajima, Sakae Mizuki, Hinari Shimada, Taihei Shiotani, Koshiro Saito, Masanari Ohi, Masaki Kawamura, Taishi Nakamura, Takumi Okamoto, et al. 2025. Rewriting pre-training data boosts llm performance in math and code. *arXiv preprint arXiv:2505.02881*.

[56] Samir Yitzhak Gadre, Gabriel Ilharco, Alex Fang, Jonathan Hayase, Georgios Smyrnis, Thao Nguyen, Ryan Marten, Mitchell Wortsman, Dhruba Ghosh, Jieyu Zhang, et al. 2023. Datacomp: In search of the next generation of multimodal datasets. *arXiv preprint arXiv:2304.14108*.

[57] Samir Yitzhak Gadre, Gabriel Ilharco, Alex Fang, Jonathan Hayase, Georgios Smyrnis, Thao Nguyen, Ryan Marten, Mitchell Wortsman, Dhruba Ghosh, Jieyu Zhang, et al. 2024.

Datacomp: In search of the next generation of multimodal datasets. *Advances in Neural Information Processing Systems*, 36.

[58] Chuang Gan, Jeremy Schwartz, Seth Alter, Damian Mrowca, Martin Schrimpf, James Traer, Julian De Freitas, Jonas Kubilius, Abhishek Bhandwaldar, Nick Haber, et al. 2020. Threedworld: A platform for interactive multi-modal physical simulation. *arXiv preprint arXiv:2007.04954*.

[59] Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac'h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. 2024. The language model evaluation harness.

[60] Peng Gao, Shijie Geng, Renrui Zhang, Teli Ma, Rongyao Fang, Yongfeng Zhang, Hongsheng Li, and Yu Qiao. 2021. Clip-adapter: Better vision-language models with feature adapters. https://arxiv.org/abs/2110.04544.

[61] Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. SimCSE: Simple contrastive learning of sentence embeddings. In *Empirical Methods in Natural Language Processing (EMNLP)*.

[62] Tao Ge, Xin Chan, Xiaoyang Wang, Dian Yu, Haitao Mi, and Dong Yu. 2024. Scaling synthetic data creation with 1,000,000,000 personas. *arXiv preprint arXiv:2406.20094*.

[63] Gregor Geigle, Abhay Jain, Radu Timofte, and Goran Glavaš. 2023. mblip: Efficient bootstrapping of multilingual vision-llms. *arXiv preprint arXiv:2307.06930*.

[64] Gregor Geigle, Radu Timofte, and Goran Glavaš. 2023. Babel-imagenet: Massively multilingual evaluation of vision-and-language representations. *arXiv preprint arXiv:2306.08658*.

[65] Matthias Gerstgrasser, Rylan Schaeffer, Apratim Dey, Rafael Rafailov, Henry Sleight, John Hughes, Tomasz Korbak, Rajashree Agrawal, Dhruv Pai, Andrey Gromov, et al. 2024. Is model collapse inevitable? breaking the curse of recursion by accumulating real and synthetic data. *arXiv preprint arXiv:2404.01413*.

[66] Raphael Gontijo-Lopes, Yann Dauphin, and Ekin D Cubuk. 2021. No one representation to rule them all: Overlapping features of training methods. https://arxiv.org/abs/2007.01434.

[67] Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

[68] Jiaxi Gu, Xiaojun Meng, Guansong Lu, Lu Hou, Niu Minzhe, Xiaodan Liang, Lewei Yao, Runhui Huang, Wei Zhang, Xin Jiang, et al. 2022. Wukong: A 100 million large-scale chinese cross-modal pre-training benchmark. *Advances in Neural Information Processing Systems*, 35:26418–26431.

[69] Ishaan Gulrajani and David Lopez-Paz. 2020. In search of lost domain generalization. *International Conference on Learning Representations (ICLR)*. https://arxiv.org/abs/2007.01434.

[70] Suriya Gunasekar, Yi Zhang, Jyoti Aneja, Caio César Teodoro Mendes, Allie Del Giorno, Sivakanth Gopi, Mojan Javaheripi, Piero Kauffmann, Gustavo de Rosa, Olli Saarikivi, et al. 2023. Textbooks are all you need. *arXiv preprint arXiv:2306.11644*.

[71] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. 2022. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16000–16009.

[72] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. 2020. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9729–9738.

[73] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

[74] Ruifei He, Shuyang Sun, Xin Yu, Chuhui Xue, Wenqing Zhang, Philip Torr, Song Bai, and Xiaojuan Qi. 2022. Is synthetic data from generative models ready for image recognition? *arXiv preprint arXiv:2210.07574*.

[75] Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, et al. 2021. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 8340–8349.

[76] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*.

[77] Dan Hendrycks and Thomas Dietterich. 2019. Benchmarking neural network robustness to common corruptions and perturbations. *International Conference on Learning Representations (ICLR)*. https://arxiv.org/abs/1903.12261.

[78] Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. 2021. Natural adversarial examples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15262–15271.

[79] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. 2021. Clipscore: A reference-free evaluation metric for image captioning. *arXiv preprint arXiv:2104.08718*.

[80] Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. 2022. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*.

[81] Rachel Hong, William Agnew, Tadayoshi Kohno, and Jamie Morgenstern. 2024. Who's in and who's out? a case study of multimodal clip-filtering in datacomp. *arXiv preprint arXiv:2405.08209*.

[82] Xiaowei Hu, Zhe Gan, Jianfeng Wang, Zhengyuan Yang, Zicheng Liu, Yumao Lu, and Lijuan Wang. 2022. Scaling up vision-language pre-training for image captioning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 17980–17989.

[83] Gabriel Ilharco, Mitchell Wortsman, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. 2021. Openclip. If you use this software, please cite it as below.

[84] Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. 2021. Openclip.

[85] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning*, pages 4904–4916. PMLR.

[86] Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. 2017. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2901–2910.

[87] Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2016. Bag of tricks for efficient text classification. *arXiv preprint arXiv:1607.01759*.

[88] Kimmo Kärkkäinen and Jungseock Joo. 2019. Fairface: Face attribute dataset for balanced race, gender, and age. *arXiv preprint arXiv:1908.04913*.

[89] Andrej Karpathy and Li Fei-Fei. 2015. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3128–3137.

[90] Wonjae Kim, Sanghyuk Chun, Taekyung Kim, Dongyoon Han, and Sangdoo Yun. 2024. Hype: Hyperbolic entailment filtering for underspecified images and texts. *arXiv preprint arXiv:2404.17507*.

[91] Ryan Kiros, Ruslan Salakhutdinov, and Rich Zemel. 2014. Multimodal neural language models. In *International conference on machine learning*, pages 595–603. PMLR.

[92] Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, et al. 2021. Wilds: A benchmark of in-the-wild distribution shifts. In *International Conference on Machine Learning*, pages 5637–5664. PMLR.

[93] Alex Krizhevsky, Geoffrey Hinton, et al. 2009. Learning multiple layers of features from tiny images.

[94] David Krueger, Ethan Caballero, Joern-Henrik Jacobsen, Amy Zhang, Jonathan Binas, Dinghuai Zhang, Remi Le Priol, and Aaron Courville. 2021. Out-of-distribution generalization via risk extrapolation (rex). https://arxiv.org/abs/2003.00688.

[95] Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*.

[96] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. 2017. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Advances in Neural Information Processing Systems (NeurIPS)*. https://arxiv.org/abs/1612.01474.

[97] Weiyu Lan, Xirong Li, and Jianfeng Dong. 2017. Fluency-guided cross-lingual image captioning. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 1549–1557.

[98] Jeffrey Li, Alex Fang, Georgios Smyrnis, Maor Ivgi, Matt Jordan, Samir Yitzhak Gadre, Hritik Bansal, Etash Guha, Sedrick Scott Keh, Kushal Arora, et al. 2024. Datacomp-lm: In search

of the next generation of training sets for language models. *Advances in Neural Information Processing Systems*, 37:14200–14282.

[99] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR.

[100] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, pages 12888–12900. PMLR.

[101] Xirong Li, Chaoxi Xu, Xiaoxu Wang, Weiyu Lan, Zhengxiong Jia, Gang Yang, and Jieping Xu. 2019. Coco-cn for cross-lingual image tagging, captioning, and retrieval. *IEEE Transactions on Multimedia*, 21(9):2347–2360.

[102] Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, et al. 2020. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXX 16*, pages 121–137. Springer.

[103] Yang Li, Youssef Emad, Karthik Padthe, Jack Lanchantin, Weizhe Yuan, Thao Nguyen, Jason Weston, Shang-Wen Li, Dong Wang, Ilia Kulikov, et al. 2025. Naturalthoughts: Selecting and distilling reasoning traces for general reasoning tasks. *arXiv preprint arXiv:2507.01921*.

[104] Yangguang Li, Feng Liang, Lichen Zhao, Yufeng Cui, Wanli Ouyang, Jing Shao, Fengwei Yu, and Junjie Yan. 2021. Supervision exists everywhere: A data efficient contrastive language-image pre-training paradigm. *arXiv preprint arXiv:2110.05208*.

[105] Yuanzhi Li, Sébastien Bubeck, Ronen Eldan, Allie Del Giorno, Suriya Gunasekar, and Yin Tat Lee. 2023. Textbooks are all you need ii: phi-1.5 technical report. *arXiv preprint arXiv:2309.05463*.

[106] Yuxuan Li, Yicheng Zhang, Wenhao Tang, Yimian Dai, Ming-Ming Cheng, Xiang Li, and Jian Yang. 2025. Visual instruction pretraining for domain-specific foundation models. *arXiv preprint arXiv:2509.17562*.

[107] Stephanie Lin, Jacob Hilton, and Owain Evans. 2021. Truthfulqa: Measuring how models mimic human falsehoods, 2022. *URL https://arxiv. org/abs/2109.07958*.

[108] Qian Liu, Xiaosen Zheng, Niklas Muennighoff, Guangtao Zeng, Longxu Dou, Tianyu Pang, Jing Jiang, and Min Lin. 2024. Regmix: Data mixture as regression for language model pre-training. *arXiv preprint arXiv:2407.01492*.

[109] Yulong Liu, Guibo Zhu, Bin Zhu, Qi Song, Guojing Ge, Haoran Chen, GuanHui Qiao, Ru Peng, Lingxiang Wu, and Jinqiao Wang. 2022. Taisu: A 166m large-scale high-quality dataset for chinese vision-language pre-training. *Advances in Neural Information Processing Systems*, 35:16705–16717.

[110] Shayne Longpre, Gregory Yauney, Emily Reif, Katherine Lee, Adam Roberts, Barret Zoph, Denny Zhou, Jason Wei, Kevin Robinson, David Mimno, et al. 2024. A pretrainer's guide to training data: Measuring the effects of data age, domain coverage, quality, & toxicity. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3245–3276.

[111] Ilya Loshchilov and Frank Hutter. 2016. Sgdr: Stochastic gradient descent with warm restarts. *arXiv preprint arXiv:1608.03983*.

[112] Ilya Loshchilov and Frank Hutter. 2017. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*.

[113] Ian Magnusson, Nguyen Tai, Ben Bogin, David Heineman, Jena D Hwang, Luca Soldaini, Akshita Bhagia, Jiacheng Liu, Dirk Groeneveld, Oyvind Tafjord, et al. 2025. Datadecide: How to predict best pretraining data with small experiments. *arXiv preprint arXiv:2504.11393*.

[114] Pratyush Maini, Sachin Goyal, Zachary C Lipton, J Zico Kolter, and Aditi Raghunathan. 2023. T-mars: Improving visual representations by circumventing text feature learning. *arXiv preprint arXiv:2307.03132*.

[115] Pratyush Maini, Sachin Goyal, Dylan Sam, Alex Robey, Yash Savani, Yiding Jiang, Andy Zou, Matt Fredrikson, Zacharcy C Lipton, and J Zico Kolter. 2025. Safety pretraining: Toward the next generation of safe ai. *arXiv preprint arXiv:2504.16980*.

[116] Pratyush Maini, Skyler Seto, He Bai, David Grangier, Yizhe Zhang, and Navdeep Jaitly. 2024. Rephrasing the web: A recipe for compute and data-efficient language modeling. *arXiv preprint arXiv:2401.16380*.

[117] Celso M de Melo, Antonio Torralba, Leonidas Guibas, James DiCarlo, Rama Chellappa, and Jessica Hodgins. 2021. Next-generation deep learning based on simulators and synthetic data. *Trends in cognitive sciences*.

[118] John P Miller, Rohan Taori, Aditi Raghunathan, Shiori Sagawa, Pang Wei Koh, Vaishaal Shankar, Percy Liang, Yair Carmon, and Ludwig Schmidt. 2021. Accuracy on the line: on the strong correlation between out-of-distribution and in-distribution generalization. In *International conference on machine learning*, pages 7721–7735. PMLR.

[119] Takashi Miyazaki and Nobuyuki Shimizu. 2016. Cross-lingual image caption generation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1780–1790.

[120] Norman Mu, Alexander Kirillov, David Wagner, and Saining Xie. 2021. Slip: Self-supervision meets language-image pre-training. *arXiv preprint arXiv:2112.12750*.

[121] Niklas Muennighoff, Alexander Rush, Boaz Barak, Teven Le Scao, Nouamane Tazi, Aleksandra Piktus, Sampo Pyysalo, Thomas Wolf, and Colin A Raffel. 2023. Scaling data-constrained language models. *Advances in Neural Information Processing Systems*, 36:50358–50376.

[122] Marianna Nezhurina, Romain Beaumont, Richard Vencu, and Christoph Schuhmann. Laion translated: 3b captions translated to english from laion5b.

[123] Thao Nguyen, Samir Yitzhak Gadre, Gabriel Ilharco, Sewoong Oh, and Ludwig Schmidt. 2024. Improving multimodal datasets with image captioning. *Advances in Neural Information Processing Systems*, 36.

[124] Thao Nguyen, Gabriel Ilharco, Mitchell Wortsman, Sewoong Oh, and Ludwig Schmidt. 2022. Quality not quantity: On the interaction between dataset design and robustness of clip. *arXiv preprint arXiv:2208.05516*.

[125] Thao Nguyen, Jeffrey Li, Sewoong Oh, Ludwig Schmidt, Jason E Weston, Luke Zettlemoyer, and Xian Li. 2024. Better alignment with instruction back-and-forth translation. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 13289–13308.

[126] Thao Nguyen, Yang Li, Olga Golovneva, Luke Zettlemoyer, Sewoong Oh, Ludwig Schmidt, and Xian Li. 2025. Recycling the web: A method to enhance pre-training data quality and quantity for language models. *arXiv preprint arXiv:2506.04689*.

[127] Thao Nguyen, Matthew Wallingford, Sebastin Santy, Wei-Chiu Ma, Sewoong Oh, Ludwig Schmidt, Pang W Koh, and Ranjay Krishna. 2024. Multilingual diversity improves vision-language representations. *Advances in Neural Information Processing Systems*, 37:91430–91459.

[128] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. 2021. Glide: Towards photorealistic image generation and editing with text-guided diffusion models. *arXiv preprint arXiv:2112.10741*.

[129] Jeremy Nixon, Balaji Lakshminarayanan, and Dustin Tran. 2020. Why are bootstrapped deep ensembles not better? In *"I Can't Believe It's Not Better!" NeurIPS 2020 workshop*.

[130] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

[131] Guilherme Penedo, Hynek Kydlíček, Anton Lozhkov, Margaret Mitchell, Colin A Raffel, Leandro Von Werra, Thomas Wolf, et al. 2024. The fineweb datasets: Decanting the web for

the finest text data at scale. *Advances in Neural Information Processing Systems*, 37:30811–30849.

[132] Guilherme Penedo, Quentin Malartic, Daniel Hesslow, Ruxandra Cojocaru, Alessandro Cappelli, Hamza Alobeidli, Baptiste Pannier, Ebtesam Almazrouei, and Julien Launay. 2023. The refinedweb dataset for falcon llm: outperforming curated corpora with web data, and web data only. *arXiv preprint arXiv:2306.01116*.

[133] Xingchao Peng, Ben Usman, Neela Kaushik, Judy Hoffman, Dequan Wang, and Kate Saenko. 2017. Visda: The visual domain adaptation challenge. *arXiv preprint arXiv:1710.06924*.

[134] Jonas Pfeiffer, Gregor Geigle, Aishwarya Kamath, Jan-Martin O Steitz, Stefan Roth, Ivan Vulić, and Iryna Gurevych. 2021. xgqa: Cross-lingual visual question answering. *arXiv preprint arXiv:2109.06082*.

[135] Hieu Pham, Zihang Dai, Golnaz Ghiasi, Hanxiao Liu, Adams Wei Yu, Minh-Thang Luong, Mingxing Tan, and Quoc V Le. Combined scaling for zero-shot transfer learning, 2021. *URL https://arxiv. org/abs/2111.10050*.

[136] Hieu Pham, Zihang Dai, Golnaz Ghiasi, Hanxiao Liu, Adams Wei Yu, Minh-Thang Luong, Mingxing Tan, and Quoc V Le. 2021. Combined scaling for zero-shot transfer learning. *arXiv preprint arXiv:2111.10050*.

[137] Krishna Pillutla, Swabha Swayamdipta, Rowan Zellers, John Thickstun, Sean Welleck, Yejin Choi, and Zaid Harchaoui. 2021. Mauve: Measuring the gap between neural text and human text using divergence frontiers. *Advances in Neural Information Processing Systems*, 34:4816–4828.

[138] Giovanni Puccetti, Maciej Kilian, and Romain Beaumont. Training contrastive captioners.

[139] Joaquin Quiñonero-Candela, Masashi Sugiyama, Anton Schwaighofer, and Neil D Lawrence. 2008. *Dataset shift in machine learning*. Mit Press.

[140] Filip Radenovic, Abhimanyu Dubey, Abhishek Kadian, Todor Mihaylov, Simon Vandenhende, Yash Patel, Yi Wen, Vignesh Ramanathan, and Dhruv Mahajan. 2023. Filtering, distillation, and hard negatives for vision-language pre-training. *arXiv preprint arXiv:2301.02280*.

[141] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR.

[142] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

[143] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67.

[144] Vivek Ramanujan, Thao Nguyen, Sewoong Oh, Ali Farhadi, and Ludwig Schmidt. 2023. On the connection between pre-training data diversity and fine-tuning robustness. *Advances in Neural Information Processing Systems*, 36:66426–66437.

[145] Vikram V Ramaswamy, Sing Yu Lin, Dora Zhao, Aaron Adcock, Laurens van der Maaten, Deepti Ghadiyaram, and Olga Russakovsky. 2024. Geode: a geographically diverse evaluation dataset for object recognition. *Advances in Neural Information Processing Systems*, 36.

[146] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. 2021. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, pages 8821–8831. PMLR.

[147] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. 2019. Do imagenet classifiers generalize to imagenet? In *International conference on machine learning*, pages 5389–5400. PMLR.

[148] Nils Reimers and Iryna Gurevych. 2020. Making monolingual sentence embeddings multilingual using knowledge distillation. *arXiv preprint arXiv:2004.09813*.

[149] Megan Richards, Polina Kirichenko, Diane Bouchacourt, and Mark Ibrahim. 2023. Does progress on object recognition benchmarks improve real-world generalization? *arXiv preprint arXiv:2307.13136*.

[150] Stephan R Richter, Vibhav Vineet, Stefan Roth, and Vladlen Koltun. 2016. Playing for data: Ground truth from computer games. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*, pages 102–118. Springer.

[151] William A Gaviria Rojas, Sudnya Diamos, Keertan Ranjan Kini, David Kanter, Vijay Janapa Reddi, and Cody Coleman. 2022. The dollar street dataset: Images representing the geographic and socioeconomic diversity of the world. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

[152] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695.

[153] Karsten Roth, Vishaal Udandarao, Sebastian Dziadzio, Ameya Prabhu, Mehdi Cherti, Oriol Vinyals, Olivier Hénaff, Samuel Albanie, Matthias Bethge, and Zeynep Akata. 2024. A practitioner's guide to continual multimodal pretraining. *arXiv preprint arXiv:2408.14471*.

[154] Yangjun Ruan, Neil Band, Chris J Maddison, and Tatsunori Hashimoto. 2025. Reasoning to learn from latent thoughts. *arXiv preprint arXiv:2503.18866*.

[155] Noveen Sachdeva, Benjamin Coleman, Wang-Cheng Kang, Jianmo Ni, Lichan Hong, Ed H Chi, James Caverlee, Julian McAuley, and Derek Zhiyuan Cheng. 2024. How to train data-efficient llms. *arXiv preprint arXiv:2402.09668*.

[156] Shiori Sagawa, Pang Wei Koh, Tatsunori B Hashimoto, and Percy Liang. 2019. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. https://arxiv.org/abs/1911.08731.

[157] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S Sara Mahdavi, Rapha Gontijo Lopes, et al. 2022. Photorealistic text-to-image diffusion models with deep language understanding. *arXiv preprint arXiv:2205.11487*.

[158] Shibani Santurkar, Yann Dubois, Rohan Taori, Percy Liang, and Tatsunori Hashimoto. 2022. Is a caption worth a thousand images? a controlled study for representation learning. *arXiv preprint arXiv:2207.07635*.

[159] Sebastin Santy, Jenny T Liang, Ronan Le Bras, Katharina Reinecke, and Maarten Sap. 2023. Nlpositionality: Characterizing design biases of datasets and models. *arXiv preprint arXiv:2306.01943*.

[160] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. 2022. Laion-5b: An open large-scale dataset for training next generation image-text models. *arXiv preprint arXiv:2210.08402*.

[161] Christoph Schuhmann, Andreas Köpf, Richard Vencu, Theo Coombes, and Romain Beaumont. Laion coco: 600m synthetic captions from laion2b-en.

[162] Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. 2021. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*.

[163] Eyal Segalis, Dani Valevski, Danny Lumen, Yossi Matias, and Yaniv Leviathan. 2023. A picture is worth a thousand words: Principled recaptioning improves image generation. *arXiv preprint arXiv:2310.16656*.

[164] Shreya Shankar, Yoni Halpern, Eric Breck, James Atwood, Jimbo Wilson, and D Sculley. 2017. No classification without representation: Assessing geodiversity issues in open data sets for the developing world. *arXiv preprint arXiv:1711.08536*.

[165] Vaishaal Shankar, Rebecca Roelofs, Horia Mania, Alex Fang, Benjamin Recht, and Ludwig Schmidt. 2020. Evaluating machine accuracy on imagenet. In *International Conference on Machine Learning (ICML)*. http://proceedings.mlr.press/v119/shankar20c/shankar20c.pdf.

[166] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565.

[167] Kashun Shum, Yuzhen Huang, Hongjian Zou, Ding Qi, Yixuan Liao, Xiaoxin Chen, Qian Liu, and Junxian He. 2025. Predictive data selection: The data that predicts is the data that teaches. *arXiv preprint arXiv:2503.00808*.

[168] Ilia Shumailov, Zakhar Shumaylov, Yiren Zhao, Yarin Gal, Nicolas Papernot, and Ross Anderson. 2023. The curse of recursion: Training on generated data makes models forget. *arXiv preprint arXiv:2305.17493*.

[169] Ilia Shumailov, Zakhar Shumaylov, Yiren Zhao, Yarin Gal, Nicolas Papernot, and Ross Anderson. 2023. Model dementia: Generated data makes models forget. *arXiv preprint arXiv:2305.17493*.

[170] Luca Soldaini, Rodney Kinney, Akshita Bhagia, Dustin Schwenk, David Atkinson, Russell Authur, Ben Bogin, Khyathi Chandu, Jennifer Dumas, Yanai Elazar, et al. 2024. Dolma: An open corpus of three trillion tokens for language model pretraining research. *arXiv preprint arXiv:2402.00159*.

[171] Krishna Srinivasan, Karthik Raman, Jiecao Chen, Michael Bendersky, and Marc Najork. 2021. Wit: Wikipedia-based image text dataset for multimodal multilingual machine learning. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2443–2449.

[172] Dan Su, Kezhi Kong, Ying Lin, Joseph Jennings, Brandon Norick, Markus Kliegl, Mostofa Patwary, Mohammad Shoeybi, and Bryan Catanzaro. 2024. Nemotron-cc: Transforming common crawl into a refined long-horizon pretraining dataset. *arXiv preprint arXiv:2412.02595*.

[173] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. 2013. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*.

[174] Hao Tan and Mohit Bansal. 2020. Vokenization: Improving language understanding with contextualized, visual-grounded supervision. *arXiv preprint arXiv:2010.06775*.

[175] Rohan Taori, Achal Dave, Vaishaal Shankar, Nicholas Carlini, Benjamin Recht, and Ludwig Schmidt. 2020. Measuring robustness to natural distribution shifts in image classification. *Advances in Neural Information Processing Systems*, 33:18583–18599.

[176] Rohan Taori and Tatsunori B Hashimoto. 2022. Data feedback loops: Model-driven amplification of dataset biases. *arXiv preprint arXiv:2209.03942*.

[177] Teknium. 2023. Openhermes 2.5: An open dataset of synthetic data for generalist llm assistants.

[178] Bart Thomee, David A Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. 2016. Yfcc100m: The new data in multimedia research. *Communications of the ACM*, 59(2):64–73.

[179] Antonio Torralba and Alexei A Efros. 2011. Unbiased look at dataset bias. In *CVPR 2011*, pages 1521–1528. IEEE.

[180] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

[181] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575.

[182] Mathurin Videau, Badr Youbi Idrissi, Daniel Haziza, Luca Wehrstedt, Jade Copet, Olivier Teytaud, and David Lopez-Paz. 2024. Meta Lingua: A minimal PyTorch LLM training library.

[183] Pablo Villalobos, Anson Ho, Jaime Sevilla, Tamay Besiroglu, Lennart Heim, and Marius Hobbhahn. 2024. Position: Will we run out of data? limits of llm scaling based on human-generated data. In *Forty-first International Conference on Machine Learning*.

[184] Alexander Visheratin. 2023. Nllb-clip–train performant multilingual image retrieval model on a budget. *arXiv preprint arXiv:2309.01859*.

[185] Dong Wang, Yang Li, Ansong Ni, Youssef Emad, Xinjie Lei, Ruta Desai, Karthik Padthe, Xian Li, Asli Celikyilmaz, Ramya Raghavendra, Leo Huang, and Daniel Li. 2025. Matrix: Multi-agent data generation infra and experimentation.

[186] Haohan Wang, Songwei Ge, Zachary Lipton, and Eric P Xing. 2019. Learning robust global representations by penalizing local predictive power. *Advances in Neural Information Processing Systems*, 32.

[187] Liang Wang, Nan Yang, Shaohan Huang, Li Dong, and Furu Wei. 2025. Thinking augmented pre-training. *arXiv preprint arXiv:2509.20186*.

[188] Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. 2022. Ofa: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. In *International Conference on Machine Learning*, pages 23318–23340. PMLR.

[189] Zirui Wang, Jiahui Yu, Adams Wei Yu, Zihang Dai, Yulia Tsvetkov, and Yuan Cao. 2021. Simvlm: Simple visual language model pretraining with weak supervision. *arXiv preprint arXiv:2108.10904*.

[190] Maurice Weber, Dan Fu, Quentin Anthony, Yonatan Oren, Shane Adams, Anton Alexandrov, Xiaozhong Lyu, Huu Nguyen, Xiaozhe Yao, Virginia Adams, et al. 2024. Redpajama: an open dataset for training large language models. *Advances in neural information processing systems*, 37:116462–116492.

[191] Alexander Wettig, Aatmik Gupta, Saumya Malik, and Danqi Chen. 2024. Qurating: Selecting high-quality data for training language models. *arXiv preprint arXiv:2402.09739*.

[192] Alexander Wettig, Kyle Lo, Sewon Min, Hannaneh Hajishirzi, Danqi Chen, and Luca Soldaini. 2025. Organize the web: Constructing domains enhances pre-training data curation. *arXiv preprint arXiv:2502.10341*.

[193] Mitchell Wortsman, Gabriel Ilharco, Samir Yitzhak Gadre, Rebecca Roelofs, Raphael Gontijo-Lopes, Ari S Morcos, Hongseok Namkoong, Ali Farhadi, Yair Carmon, Simon Kornblith, et al. 2022. Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time. *International Conference on Machine Learning (ICML)*. https://arxiv.org/abs/2203.05482.

[194] Mitchell Wortsman, Gabriel Ilharco, Jong Wook Kim, Mike Li, Simon Kornblith, Rebecca Roelofs, Raphael Gontijo-Lopes, Hannaneh Hajishirzi, Ali Farhadi, Hongseok Namkoong, and Ludwig Schmidt. 2021. Robust fine-tuning of zero-shot models. *arXiv preprint arXiv:2109.01903*. https://arxiv.org/abs/2109.01903.

[195] Sang Michael Xie, Hieu Pham, Xuanyi Dong, Nan Du, Hanxiao Liu, Yifeng Lu, Percy S Liang, Quoc V Le, Tengyu Ma, and Adams Wei Yu. 2023. Doremi: Optimizing data mixtures speeds up language model pretraining. *Advances in Neural Information Processing Systems*, 36:69798–69818.

[196] Sang Michael Xie, Shibani Santurkar, Tengyu Ma, and Percy Liang. 2023. Data selection for language models via importance resampling. *arXiv preprint arXiv:2302.03169*.

[197] An Yang, Junshu Pan, Junyang Lin, Rui Men, Yichang Zhang, Jingren Zhou, and Chang Zhou. 2022. Chinese clip: Contrastive vision-language pretraining in chinese. *arXiv preprint arXiv:2211.01335*.

[198] Zitong Yang, Neil Band, Shuangping Li, Emmanuel Candes, and Tatsunori Hashimoto. 2024. Synthetic continued pretraining. *arXiv preprint arXiv:2409.07431*.

[199] Lewei Yao, Runhui Huang, Lu Hou, Guansong Lu, Minzhe Niu, Hang Xu, Xiaodan Liang, Zhenguo Li, Xin Jiang, and Chunjing Xu. 2021. Filip: Fine-grained interactive language-image pre-training. *arXiv preprint arXiv:2111.07783*.

[200] Andre Ye, Sebastin Santy, Jena D Hwang, Amy X Zhang, and Ranjay Krishna. 2023. Cultural and linguistic diversity improves visual representations. *arXiv preprint arXiv:2310.14356*.

[201] Da Yin, Liunian Harold Li, Ziniu Hu, Nanyun Peng, and Kai-Wei Chang. 2021. Broaden the vision: Geo-diverse visual commonsense reasoning. *arXiv preprint arXiv:2109.06860*.

[202] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78.

[203] Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. 2022. Coca: Contrastive captioners are image-text foundation models. *arXiv preprint arXiv:2205.01917*.

[204] Zichun Yu, Fei Peng, Jie Lei, Arnold Overwijk, Wen-tau Yih, and Chenyan Xiong. 2025. Data-efficient pretraining with group-level data influence modeling. *arXiv preprint arXiv:2502.14709*.

[205] Lu Yuan, Dongdong Chen, Yi-Ling Chen, Noel Codella, Xiyang Dai, Jianfeng Gao, Houdong Hu, Xuedong Huang, Boxin Li, Chunyuan Li, et al. 2021. Florence: A new foundation model for computer vision. *arXiv preprint arXiv:2111.11432*.

[206] Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. Hellaswag: Can a machine really finish your sentence? *arXiv preprint arXiv:1905.07830*.

[207] Chi Zhang, Huaping Zhong, Kuan Zhang, Chengliang Chai, Rui Wang, Xinlin Zhuang, Tianyi Bai, Jiantao Qiu, Lei Cao, Ju Fan, et al. 2024. Harnessing diversity for important data selection in pretraining large language models. *arXiv preprint arXiv:2409.16986*.

[208] Renrui Zhang, Rongyao Fang, Peng Gao, Wei Zhang, Kunchang Li, Jifeng Dai, Yu Qiao, and Hongsheng Li. 2021. Tip-adapter: Training-free clip-adapter for better vision-language modeling. https://arxiv.org/abs/2111.03930.

[209] Wenqi Zhang, Hang Zhang, Xin Li, Jiashuo Sun, Yongliang Shen, Weiming Lu, Deli Zhao, Yueting Zhuang, and Lidong Bing. 2025. 2.5 years in class: A multimodal textbook for vision-language pretraining. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4647–4658.

[210] Rosie Zhao, Alexandru Meterez, Sham Kakade, Cengiz Pehlevan, Samy Jelassi, and Eran Malach. 2025. Echo chamber: Rl post-training amplifies behaviors learned in pretraining. *arXiv preprint arXiv:2504.07912*.

[211] Jia Zheng, Junfei Zhang, Jing Li, Rui Tang, Shenghua Gao, and Zihan Zhou. 2020. Structured3d: A large photo-realistic dataset for structured 3d modeling. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IX 16*, pages 519–535. Springer.

[212] Kaiyang Zhou, Jingkang Yang, Chen Change Loy, and Ziwei Liu. 2022. Conditional prompt learning for vision-language models. *Conference on Computer Vision and Pattern Recognition (CVPR)*. https://arxiv.org/abs/2109.01134.

[213] Deyao Zhu, Jun Chen, Kilichbek Haydarov, Xiaoqian Shen, Wenxuan Zhang, and Mohamed Elhoseiny. 2023. Chatgpt asks, blip-2 answers: Automatic questioning towards enriched visual descriptions. *arXiv preprint arXiv:2303.06594*.

Chapter A

Appendix: Quality as a First Principle in Data Curation

A.1 Dataset Details

A.1.1 Pretraining Datasets

Dataset	Source	Total Size
YFCC	Flickr	14,826,000
LAION	Common Crawl	15,504,742
CC-12M	Unspecified web pages	9,594,338
RedCaps	Reddit	11,882,403
WIT	Wikipedia	5,038,295
ShutterStock	ShutterStock	15,540,452

Table A.1: Origin and total number of samples for each of the datasets we used in our experiments.

To get a better understanding of the diversity of different data sources, we analyze the distributions of caption lengths, image sizes and image aspect ratios for a set of 10,000 samples randomly selected from each source:

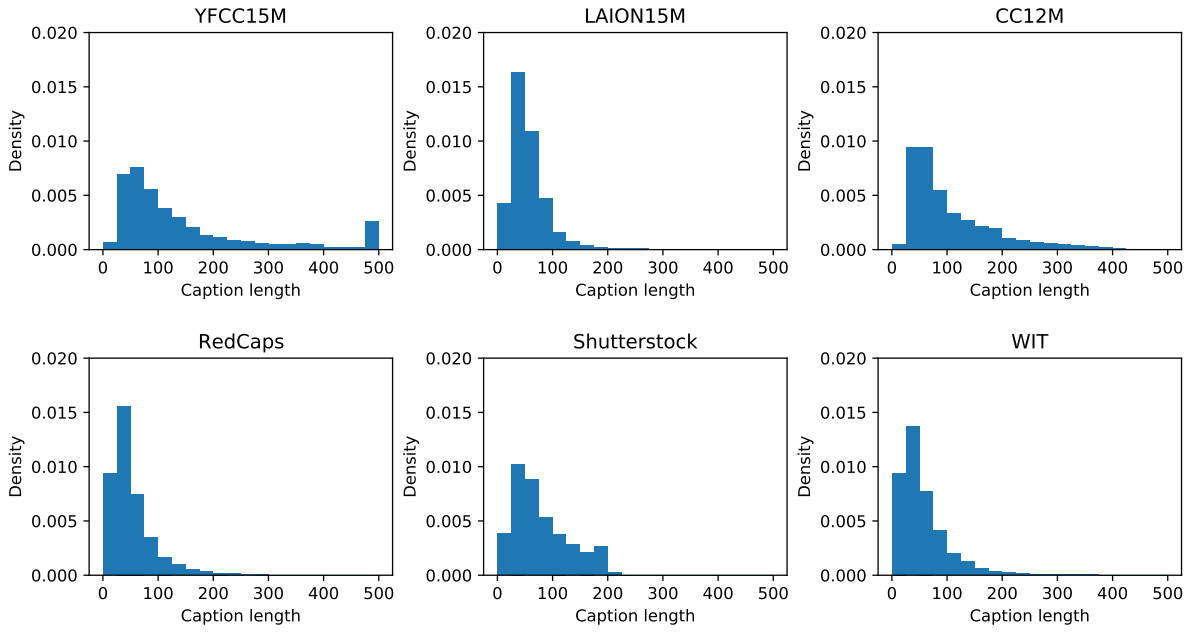


Figure A.1: Distributions of caption lengths for each data source.

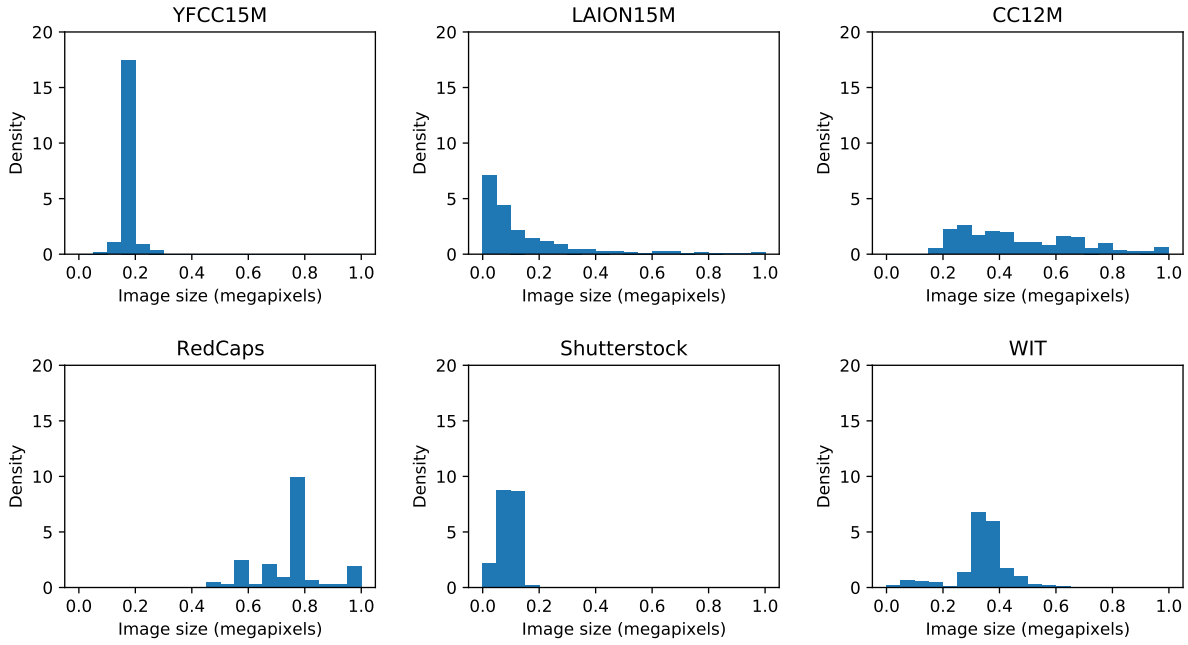


Figure A.2: Distributions of image sizes for each data source.

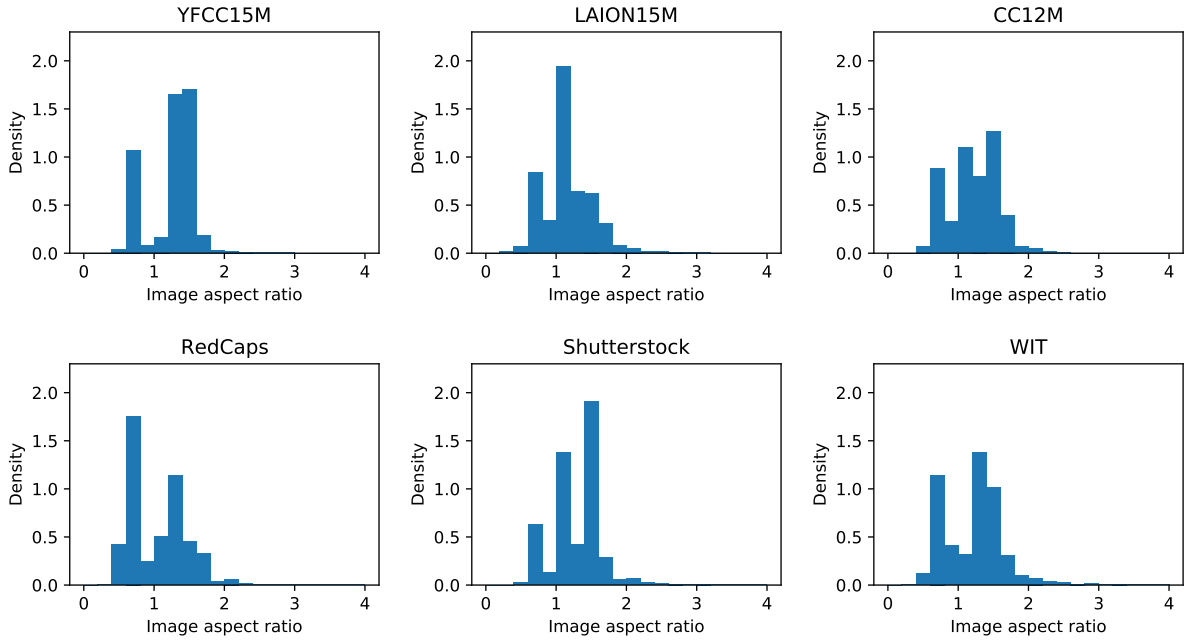


Figure A.3: Distributions of image aspect ratios for each data source.

Below we also show some examples of image-caption pairs randomly selected from each data source:

Cody This picture is #5 in my 100 strangers project. Find out more about the project and see pictures taken by other photographers at www.100Strangers.com.

Dan with the Man-Purse Dan Budiac, metrosexual

Cuffless To link or not to link, that is the question...

Dixie Union Chapel, 1836 Replaced a wooden structure that was built in 1816

ATWS Slide presentation from Sandra Carvao WTO_0288 Slide presentation from Sandra Carvao Deputy Chief Market Intelligence and Promotion Department, World Tourism...

Web 2.0 Expo 2010 - San Francisco Please feel free to use this picture in your blog, website or presentation, in accordance with the stated Creative Commons and c...

Look out, King Vidor Jesse behind the lens. Jodhpurs and bullhorn not included.

Liverpool, England, United Kingdom Albert Dock - Liverpool, England, United Kingdom

Mariposa Butterfly

THATCamp Computational Archaeology August 2012, University of Virginia

Harry Potter and the Half Blood Prince Release Party / September 16 They had to release this thing on my anniversary? Follow the flag to get your book. Auntie's ...

Nicholas Yeager Co. C, 1st Arkansas Infantry

John Paul the Great St Patrick's Cathedral Charlotte NC May 20, 2007. We were there for the 33rd Annual Rosary Rally.

100 words for snow and ice A plethora of hues and textures

GNP: St. Mary Prom We had an employee prom in August. Great times in St. Mary :)

Ice at One Mile Low of 22 degrees on Tuesday, December 8th, 2009.

Figure A.4: Random training samples from YFCC.

Leather Pencil Case - Common Room PH

wiccous.com Plus Size Bottoms White / L Plus size cotton linen nine-point wide-leg pants

Arcade Belts Norrland Roark Collab Web Belt

Moneta Zecchino high frypan 34 cm

social media icon set for instagram in different vector image

Liderazgo Educativo

manchester-bathroom-fitters-65

14 Tips for a Walt Disney World Trip with Tweens and Teens

Andrea Pirlo insists Juventus will not be resting on their 3-0 lead over Celtic

2 PC Full Bore Ball Valve

Leaves Of Fall Galaxy S5 Case

The Rise and Fall of the Trigan Empire Volume II

2Checkout Inline for Hikashop - enable 2nd address

Crooked Cover Web

Aixe Basket Club

"36"" Round Solid Hardwood Dining Table Top (Finish Options) - UnfinishedFurnitureExpo"

Figure A.5: Random training samples from LAION.

Wholesale toy candy machine for sale - Group buy The same Mini grabbing music clip candy machine small egg twisting machine grabbing childrens intellectual toys

Tug Boat Model available on Turbo Squid, the world's leading provider of digital models for visualization, films, television, and games. Tug Boats, Motor Boats, B...

Real wedding RAC Epsom on the English Wedding Blog with Murray Clarke Photography (62)

Cat smoking relax it's just the bud not corona virus vintage s Hoodie

The North Hill <PERSON> is a made in France t-shirt

Cake on a white plate, presentation.

The swimming pool at or near Maca Villas and Spa

Weighing bananas in the supermarket. Man weighing bananas on the scales in the supermarket, close-up view with no face stock images

The Legend of Zelda Breath of The Wild Link Costume

The Green Elite Wolf Hoodie

Farmer with basket vegetables, isolated in a round frame, contour drawing, icon, logo, coloring, black and white vector. Illustration, outline cartoon drawing sto...

Students in waders stand in ocean water with a large floating net.

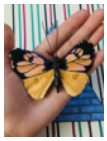
The Girls Rooms logo design type logo layout branding design

Each page explains an aspect of the product, with easy navigation and animation to help drive the point across.

The Maze Samsung Galaxy Snap Case

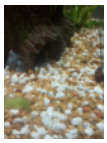
Every Day Is Another Chance Sun Women's Cotton Modal Jersey Tank Dress

Figure A.6: Random training samples from Conceptual Captions (CC-12M).



california dogface butterfly!

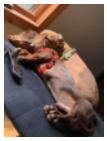
decent knife at a truck stop? maybe if it's from buccee's!



thought i lost my columbian spotted pleco days ago. that is till i saw this.

here is my small collection as of today

carbon



snoozy, sunny, snuggly saucissons.

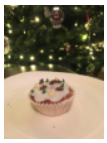
these naturally forming feathers of ice.

chiquita biting more than she can chew

the living room of an industrial loft, located in an old building from 1928, in budapest, hungary. designed by golovach tatiana and andrey kot.

tapas - empanadas, potato croquettes, patatas bravas, breadsticks & melon wrapped in serrano ham, ciabatta bread with garlic butter with two types of spanish saus...

what is this mold in my basil. appeared overnight



much smaller than the rest of your creations but i decorated this cake earlier.

a newspaper i found on thanksgiving vacation. can you spot the error to?

sunrise in eastern north carolina. those aren't mountains, they are clouds

i rolled my truck last night, and found this penny while cleaning it out this morning.

the stage in the auditorium of an old high school. detroit, mi.

Figure A.7: Random training samples from RedCaps.

Agathosma apiculata

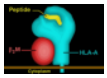
Anachronous[a] map of the Dutch colonial Empire Light green: territories administered by or originating from territories administered by the Dutch East India Com...

British Army's counter-insurgency campaign in the British controlled territories of South Arabia, 1967

Aerial view of Iguaçu falls.

Seymour at Wolfgang's nightclub, San Francisco, April 1987

An 1862 greenback five-dollar bill



HLA-A29

Monte Cimone

Eight pence note (1778), engraved and printed by Paul Revere

An experiment from William Harvey's de Motu Cordis, 1628

Deepa Shree Niraula in the left

Teodor Jeske-Choński

Koch at the commissioning of USS Lake Champlain in New York City, August 12, 1988

While a member of the Cleveland Indians in 1920, Elmer Smith became the first player to hit a grand slam in the World Series. He later moved down to play two seas...

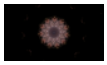
One of the violins in the Stradivarius collection of the Palacio Real, Madrid, Spain

Nominees Harrison and Tyler

Figure A.8: Random training samples from WIT.

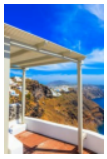
Modern art. Colorful contemporary artwork. Color strokes of paint. Brushstrokes on abstract background. Brush painting.

Vector hand-drawn eucalyptus plant isolated on a white background



Abstract kaleidoscope background. Beautiful multicolor kaleidoscope texture. Unique kaleidoscope design.

New Year red background with Christmas balls. Vector illustration.



Greece Santorini island in Cyclades, traditional detail sights of colorful and white washed traditional houses and caldera sea in background

Happy birthday golden text on the background of red and silver gifts on a bokeh background

Friends taking a selfie at a party

Payment Failure Icon In Trendy Style Isolated Background

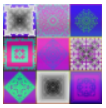
Summer Orchid Care

Reflections on a car parked in the street

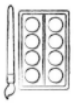
Austrian Museum of Applied Arts in Vienna, the MAK

Set of round colorful vector shapes. Abstract vector banners. Design elements. Vector illustration.

A panoramic view with village houses and cows grazing on a river bank



Abstract color background, illustration



paint colors pallette with brush

Jerusalem,Israel-May 11,2018:The palmers on Holy Sepulchre Church.Anointing Unction Stone Where Jesus Body Was Wrapped Church of the Holy Sepulcher Jerusalem Isra...

Figure A.9: Random training samples from ShutterStock.

A.1.2 Test Distributions

Figure A.10 illustrates the four distribution shifts that we use for evaluating the quality of CLIP features after pretraining on different data sources.

Figure A.10: Distribution shifts at test time. We visualize samples of the class "broom" from the reference distribution ImageNet [40], and the four distribution shifts derived from ImageNet: ImageNet-V2 [147], ImageNet-R [75], ImageNet-Sketch [186] and ObjectNet [14].

A.2 Training Details

Our implementation closely follows the training code from OpenCLIP GitHub repository [83]. When training CLIP from scratch on each of the pretraining datasets, unless otherwise mentioned, we use AdamW optimizer [112] with default PyTorch parameters $\beta_1 = 0.9, \beta_2 = 0.999, \epsilon = 10^{-8}$, (per GPU) batch size 128 and weight decay of 0.1. For learning rate, we start with a learning rate of 10^{-3} and apply a cosine-annealing learning rate schedule [111] with 5,000 steps. We use the same data augmentations as in [141]. Models then undergo distributed training on 8 A40 or A100 GPUs for 16 epochs.

A.3 Behavior of Individual Data Sources

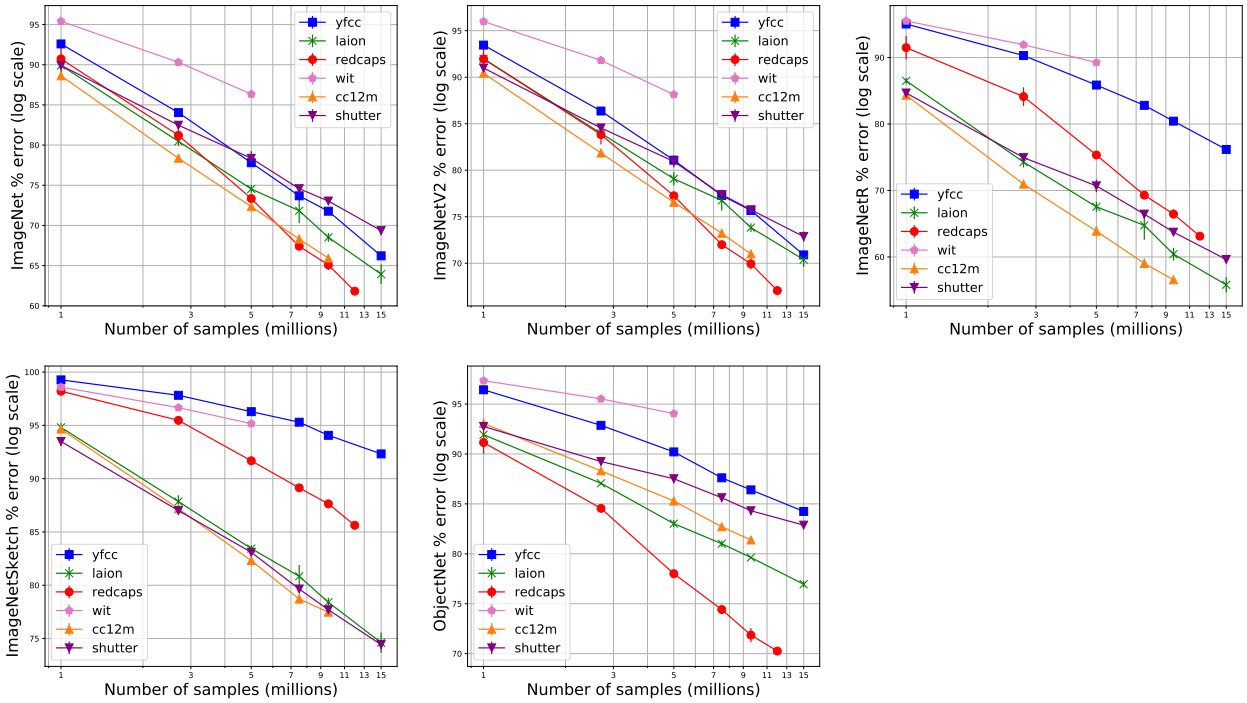


Figure A.11: Data efficiency of the six pretraining data sources on different test sets. For each source, we randomly sample various subsets of data with sizes ranging from 1M to a maximum of 15M samples, and measure the zero-shot classification error of a CLIP model trained on the subset, on ImageNet and the four shifted test sets (i.e., ImageNet-V2, ImageNet-R, ImageNet-Sketch, ObjectNet). Plotted error values are log-transformed and averaged over 3 random seeds. We find that the data efficiency (i.e., how fast the error would decrease with more samples) of the six data sources varies significantly based on the evaluation setting.

A.4 Input Mixing

A.4.1 More Experiments with CLIP pretraining Data Sources

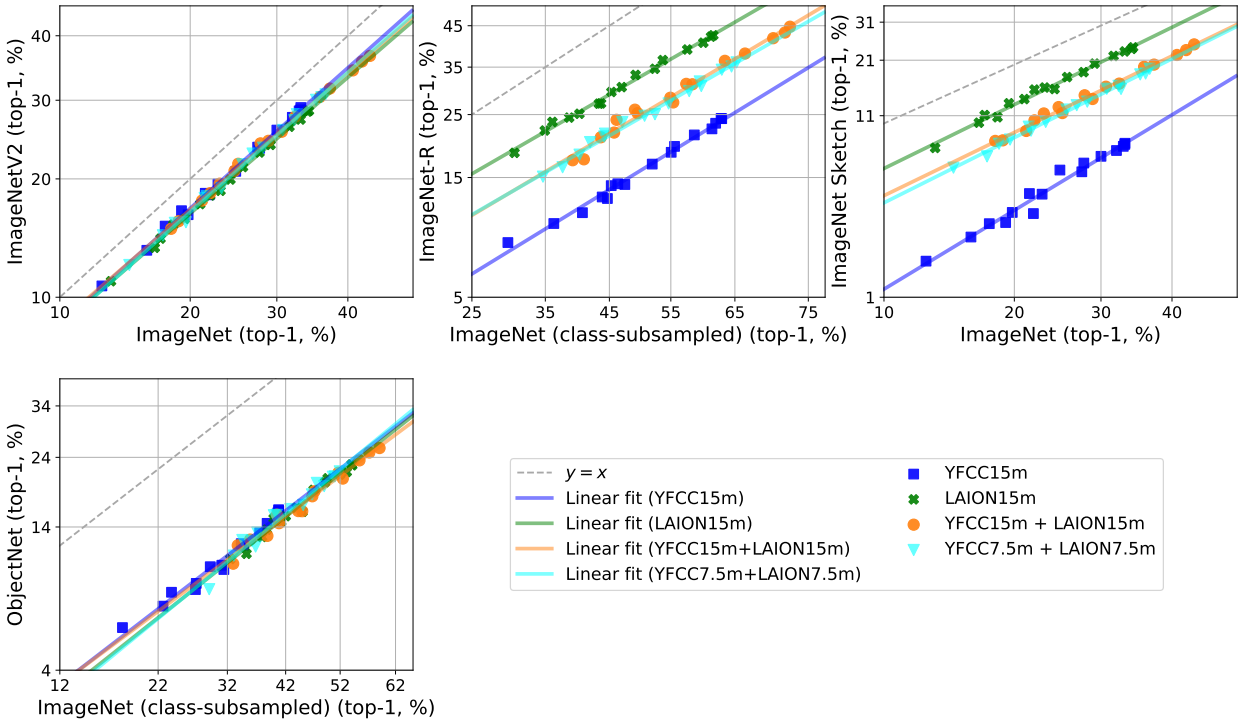


Figure A.12: Full plot for Figure 2.3 with all distribution shifts. Combining YFCC and LAION training data in equal ratios results in a CLIP model with intermediate robustness.

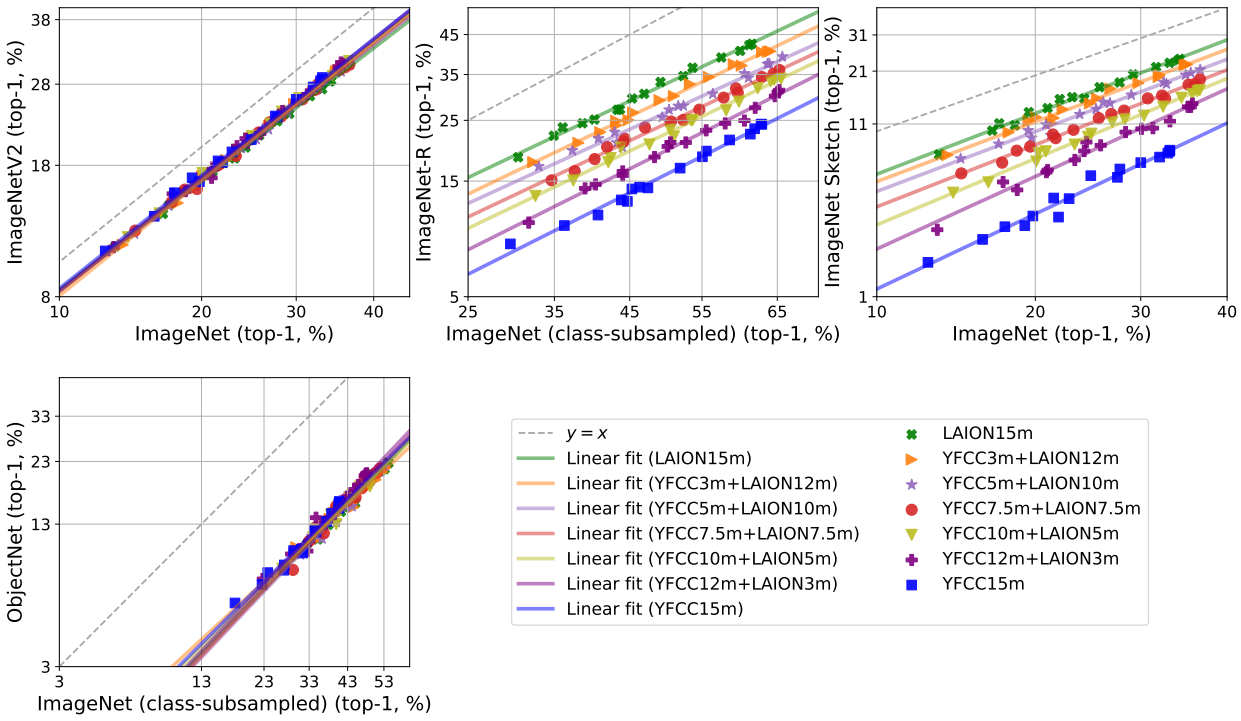


Figure A.13: Full plot for Figure 2.4 with all distribution shifts. Varying the sample contributions of YFCC and LAION to the input data mixture produces a smooth interpolation of the linear trend between the trends of training on YFCC and LAION separately.

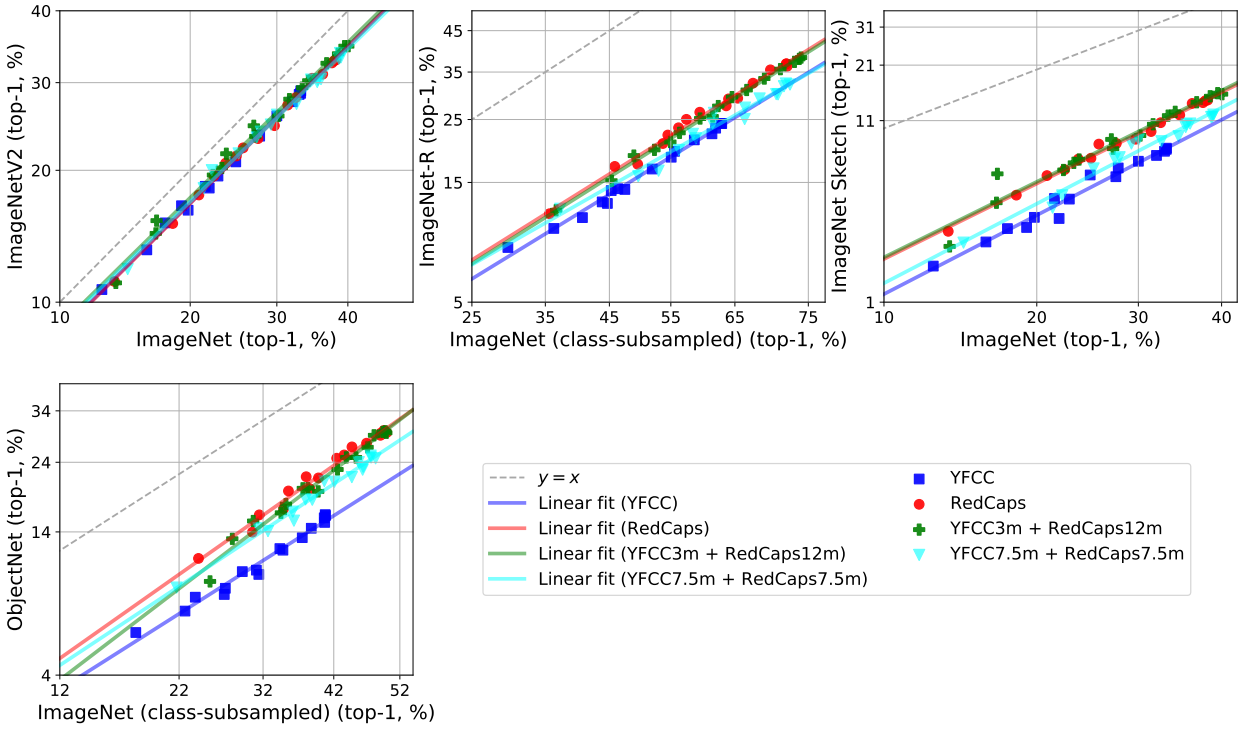


Figure A.14: Input mixing results for YFCC and RedCaps data sources. Similar to previous observations (Figure 2.4), combining YFCC and RedCaps data in the pretraining dataset with different ratios yields different linear trends that all lie between that of training on YFCC and that of training on RedCaps alone.

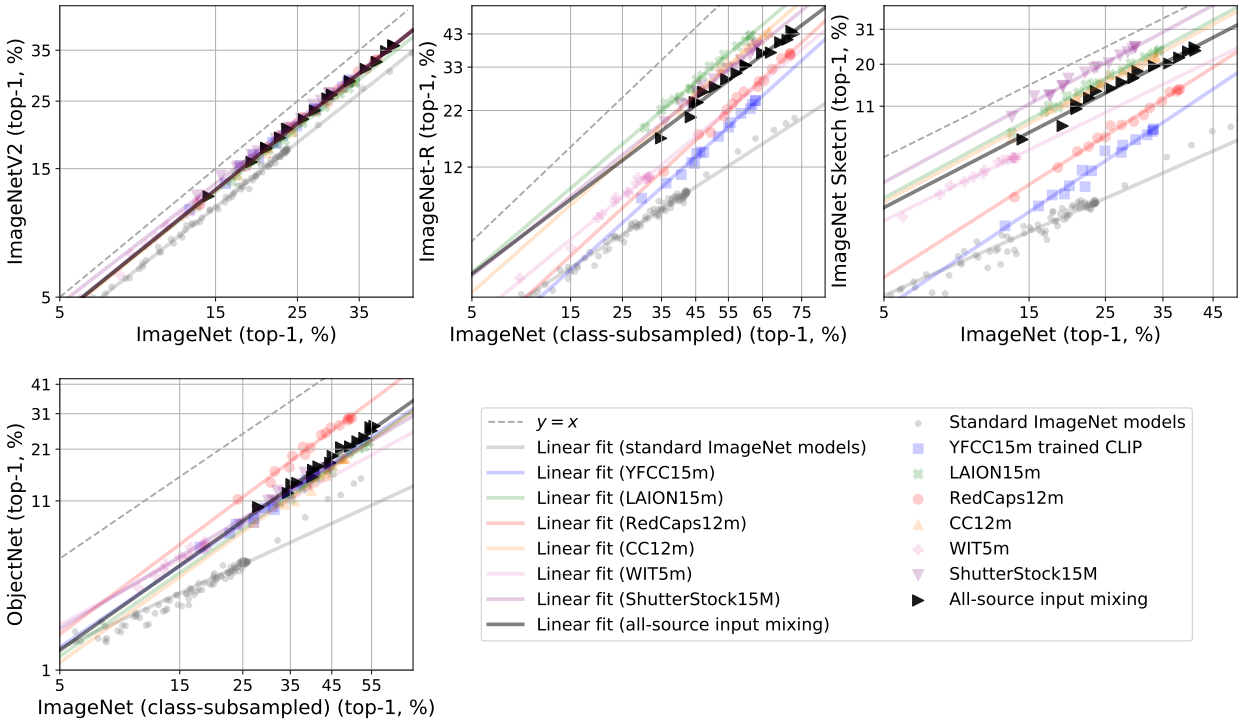


Figure A.15: Input mixing results for all six data sources. We combine data from all six sources in the testbed with equal ratios (i.e., taking 2.7M samples from each), and find that the resulting robustness of CLIP trained on this data mixture (black line), is less than that of training only on the best-performing data source for each distribution shift setting.

A.4.2 Experiments on CIFAR-10 & CINIC-10

We also investigate the phenomenon that mixing data sources resulting in diluted robustness (Section 2.6.1) in smaller-scale, uni-modal classification settings. Here, we experiment with mixing CIFAR-10 [93] and CINIC-10 [38] sources, each having 50K samples in total. CINIC-10 is itself a mixture of CIFAR-10 images and images selected and downsampled from the ImageNet database (for the same 10 classes). We use three architectures—ResNet-18, ResNet-34 and ResNet-50 [73]—and vary the number of epochs of training to obtain models of different accuracies. Models are evaluated on both CIFAR-10 and CINIC-10 standard test sets, and their performances are plotted along the axes of a scatter plot. Similar to previous input mixing results for CLIP, we observe in Figure A.16 that ResNets trained on a 50K-sample dataset made up of *both* CIFAR-10 and CINIC-10 data, produce a linear trend that lies in between the trends of training models separately on just 50K CIFAR-10 images and just 50K CINIC-10 images.

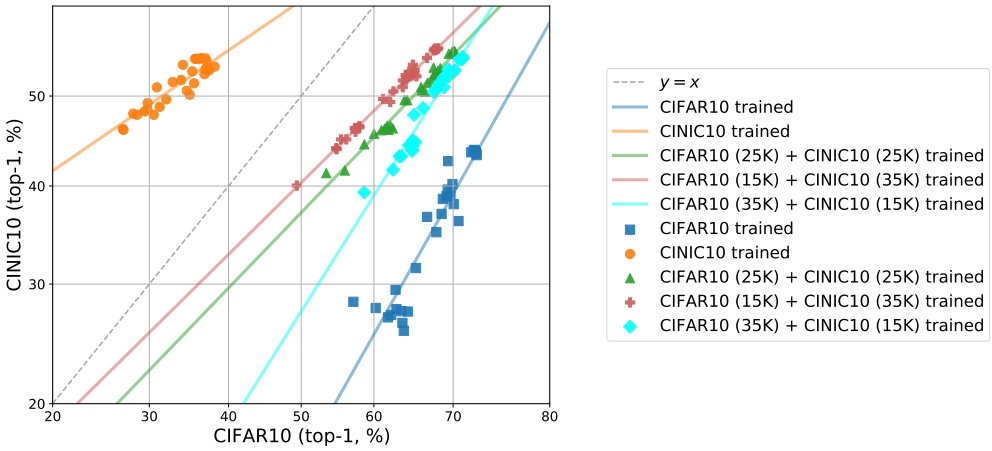


Figure A.16: Mixing inputs from CIFAR-10 and CINIC-10 distributions also produces models with intermediate robustness. Similar to our findings from the multimodal setting with CLIP pretraining, we also observe that for standard image classification tasks like CIFAR-10 and CINIC-10, combining data samples from these two distributions with varying ratios ends up diluting the robustness of the original sources. The training set size is fixed at 50K samples for all linear trends displayed in this plot.

A.5 Output Mixing

A.5.1 More Experiments with CLIP pretraining Data Sources

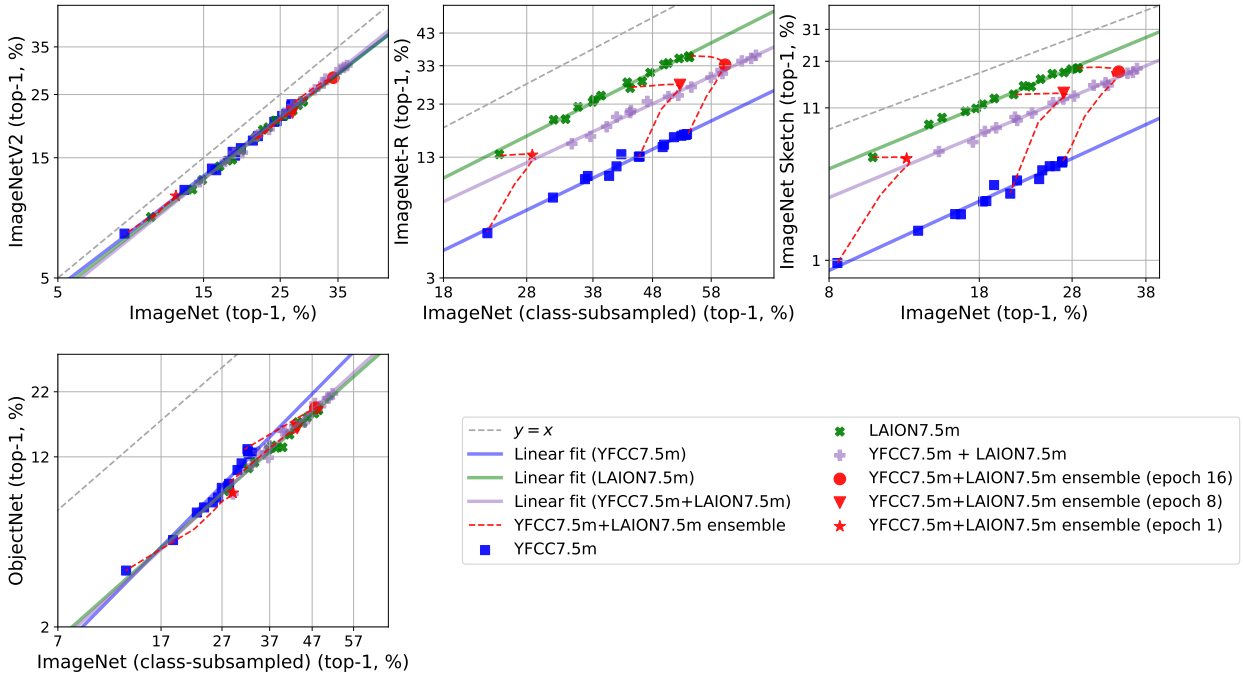


Figure A.17: Full plot for Figure 2.5 with all distribution shifts. Ensemble outputs of two CLIP models trained on YFCC and LAION separately share the same linear trend as a *single* model trained on the combined data mixture (with equal sample contribution from each source).

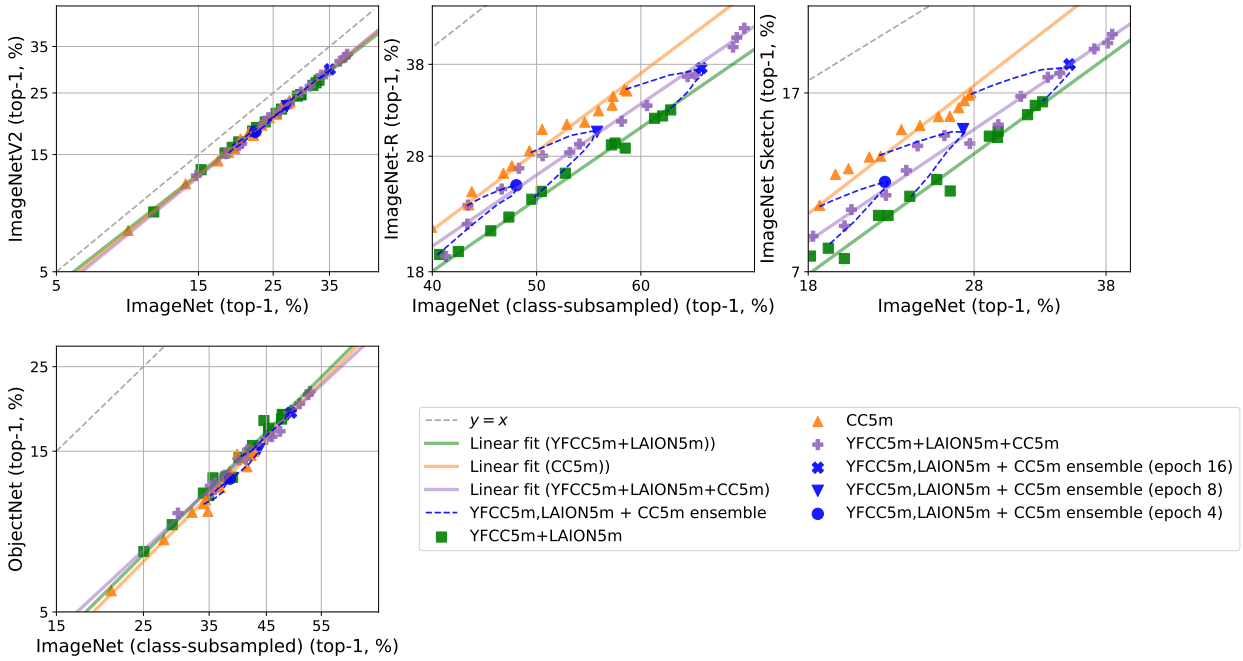


Figure A.18: Full plot for Figure 2.6 with all distribution shifts. Given an existing pretraining dataset that could be a mixture (e.g., YFCC-5M + LAION-5M, green line) and a new data source (e.g., CC-5M, orange line), we could use the ensemble outputs (blue markers) of two CLIP models that have been trained separately on these two data distributions, to estimate the linear trend for models that would be trained on *all* the data (purple line).

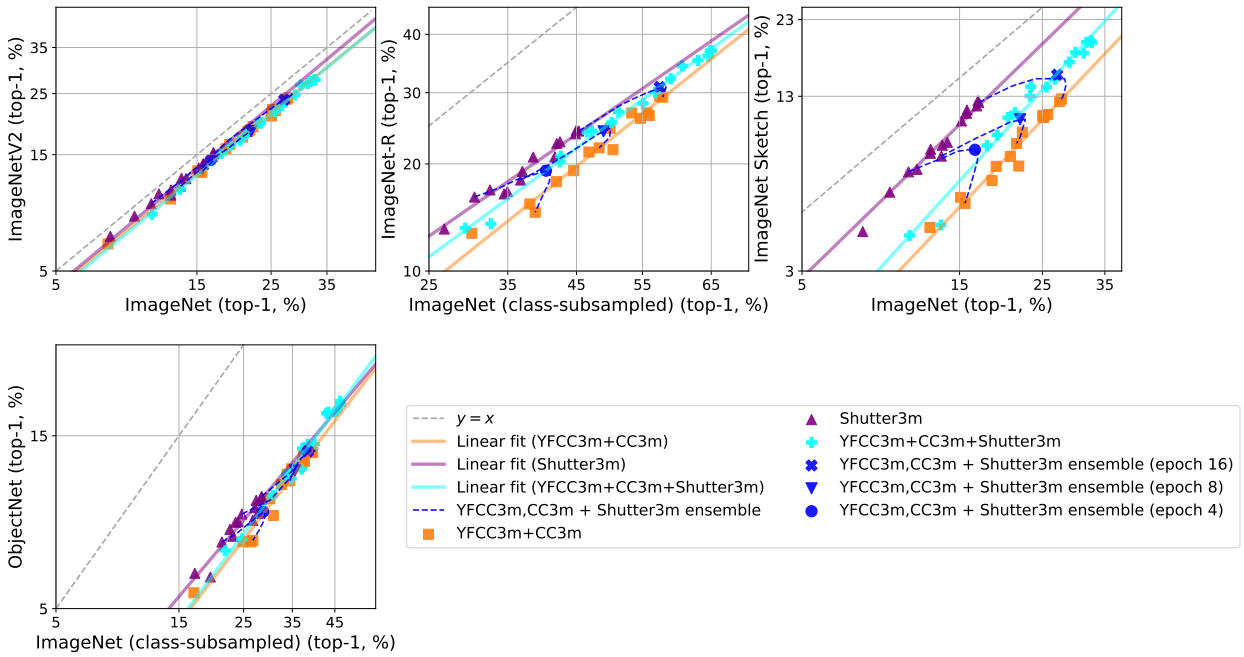


Figure A.19: Output mixing results for two CLIP models trained on YFCC-3M + CC-3M mixture and ShutterStock-3M respectively. We repeat the experiment in Figure A.18 for a different set of data sources (YFCC, ShutterStock, Conceptual Captions), taking 3M samples from each. The same output mixing phenomenon applies: the ensemble outputs of CLIPs trained on different data sources and dataset sizes (purple and orange lines), taken from the same epoch, lie on the linear trend of training a single model on the combined dataset made up of these three sources (cyan line). The two models' logit predictions are ensembled with equal weights (blue markers).

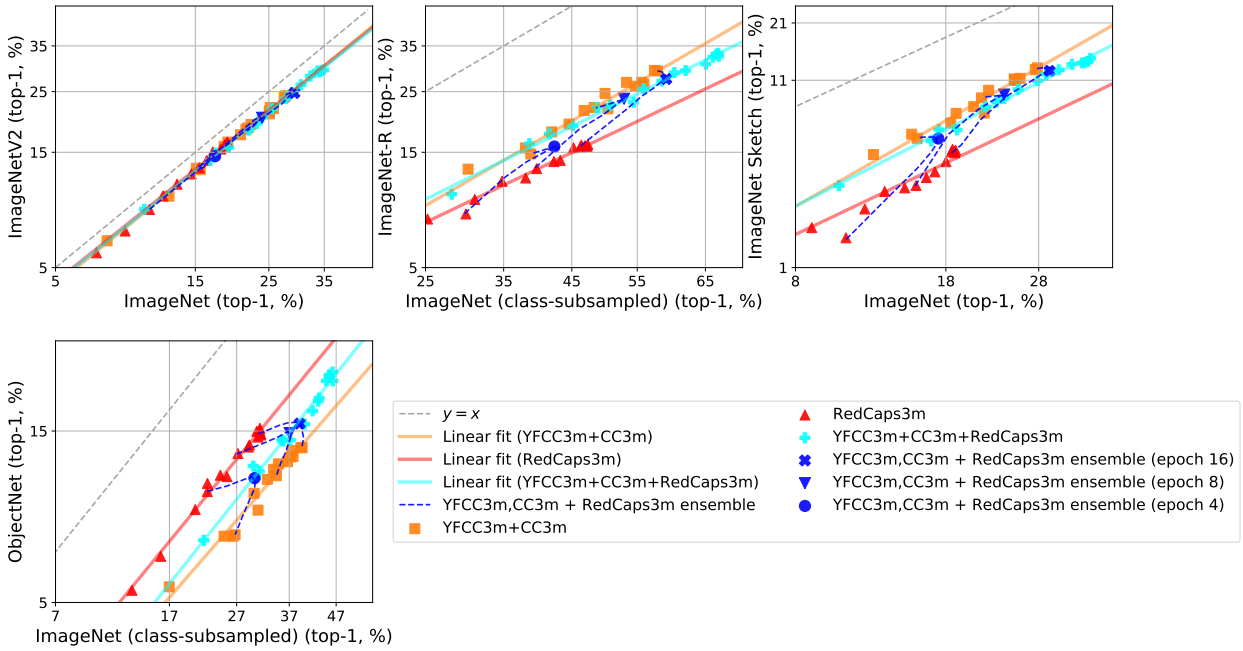


Figure A.20: Output mixing results for two CLIP models trained on YFCC-3M + CC-3M mixture and RedCaps-3M respectively. Ensemble outputs of CLIPs trained on different data sources and dataset sizes (red and orange lines), taken from the same stage of training (i.e., epoch), lie on the linear trend of training a single model on the combined dataset made up of these three sources (cyan line), when the two models' logit predictions are ensembled with equal weights (blue markers).

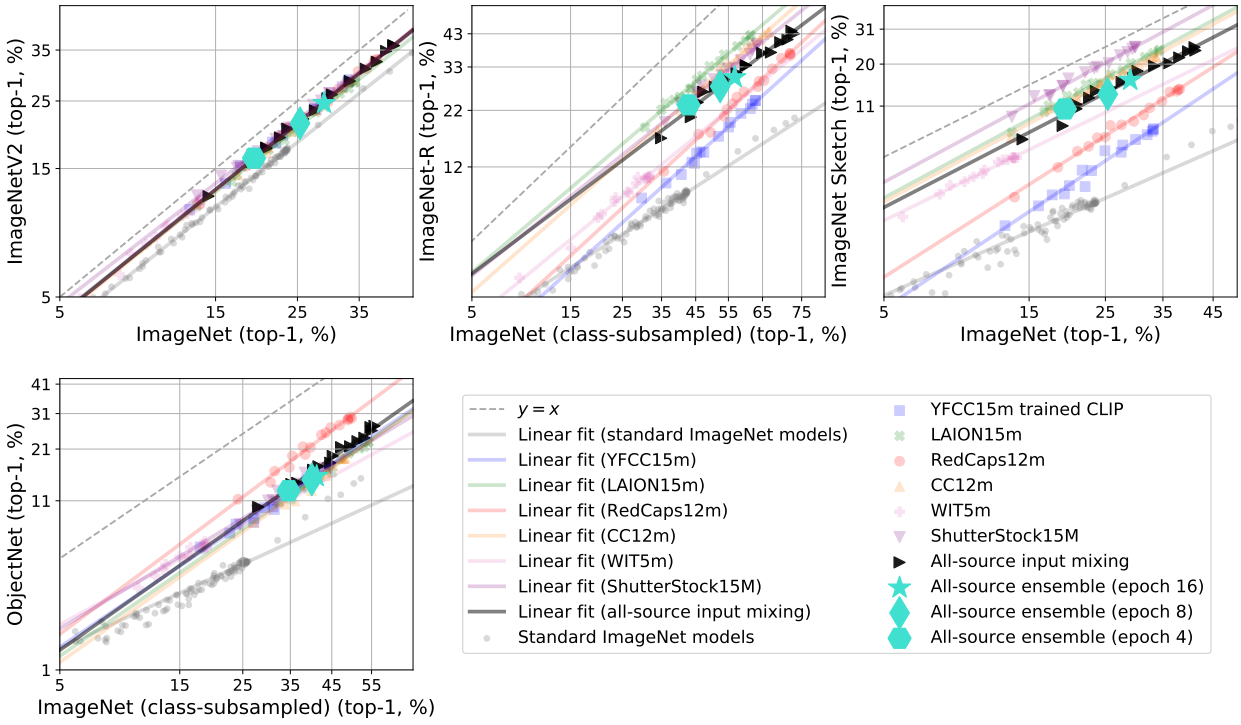


Figure A.21: Ensemble outputs of CLIPs trained separately on each of the data sources of interest share the same linear trend as a single CLIP model trained on the 6-source data mixture. Following the input mixing setup in Figure A.15, when we ensemble the logit predictions of six CLIP models, each trained on 2.7M samples randomly selected from a *single* data source, with equal weights, we find that the ensemble outputs are also predictive of the linear trend of training CLIP models on a *single* data mixture made up of 2.7M samples from each source.

A.5.2 Experiments on CIFAR-10 & CINIC-10

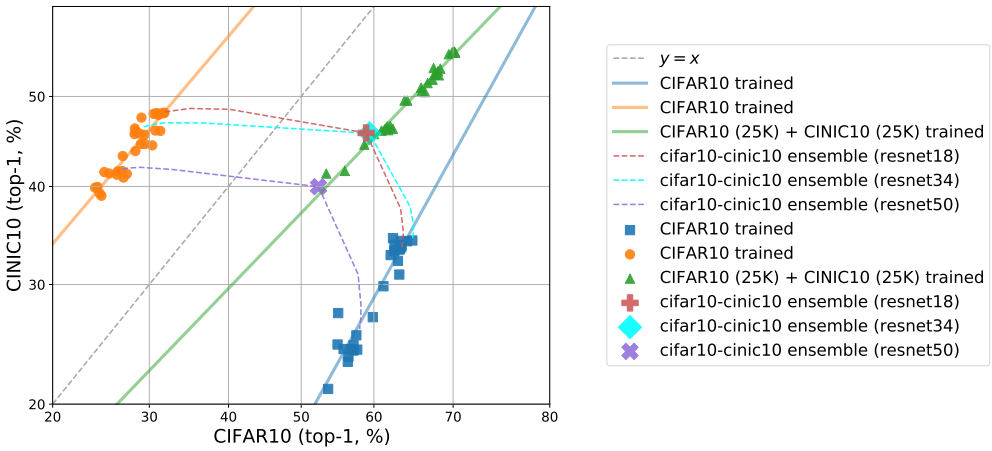


Figure A.22: Ensembling outputs of two models trained separately on CIFAR10 and CINIC10 lie on the same linear trend as training from scratch on the combined data mixture (where each source contributed equally). We combine the logit predictions of CINIC10-trained and CIFAR10-trained models that have the same architecture (e.g., ResNet-18, ResNet-34 and ResNet-50 in this case) with varying ensemble weights between 0 and 1 (dashed lines). Similar to our findings from the multimodal setting with CLIP, we also observe that when the predictions are combined with equal weights (markers on the dashed lines), the resulting test accuracies on the two corresponding test sets lie on the linear trend produced by training ResNets on a CIFAR10 + CINIC10 data mixture with equal number of samples from each source.

A.6 Proofs of the Analyses

We provide proofs of main theoretical claims in Section 2.7.

A.6.1 Proof of Theorem 1

Assumption 5. *Under the hypotheses of Theorem 1, suppose there exists a positive constant c such that the third moments are bounded by $\mathbb{E}_{(X,Y)\sim P_{\theta_1,\rho_1}}[(X_iY-\theta_{1,i})^3] \le c\mathbb{E}_{(X,Y)\sim P_{\theta_1,\rho_1}}[(X_iY-\theta_{1,i})^2]^{3/2}$, $\mathbb{E}_{(X,Y)\sim P_{\theta_2,\rho_2}}[(X_iY-\theta_{1,i})^3] \le c\mathbb{E}_{(X,Y)\sim P_{\theta_2,\rho_2}}[(X_iY-\theta_{1,i})^2]^{3/2}$, and $\mathbb{E}[(\hat{\theta}_{n,i}-\theta_i)^3] \le c\mathbb{E}[(\hat{\theta}_{n,i}-\theta_i)^2]^{3/2}$ for all $i \in [d]$.*

Under this assumption, we show that

$$\Phi^{-1}(\mathrm{Acc}_{\theta_1,\rho_1}) = \cos(\theta_1,\theta)\rho_1\frac{\rho}{\xi}\sqrt{\frac{n}{d}} + O\Big(\frac{\exp(\frac{\rho_1^2\rho^2 n}{2\xi^2 d})}{\sqrt{n}}\Big), \text{ and} \tag{A.1}$$

$$\Phi^{-1}(\mathrm{Acc}_{\theta_2,\rho_2}) = \cos(\theta_2,\theta)\rho_2\frac{\rho}{\xi}\sqrt{\frac{n}{d}} + O\Big(\frac{\exp(\frac{\rho_2^2\rho^2 n}{2\xi^2 d})}{\sqrt{n}}\Big), \tag{A.2}$$

as it will make the first and second claims straightforward. For $(X_1,Y_1)\sim P_{\theta_1,\rho_1}$, the first error event is $\{\mathrm{sign}(\langle X_1,\hat{\theta}_{n,\xi}\rangle)\neq Y_1\} = \{\langle X_1,\hat{\theta}_{n,\xi}\rangle Y_1 \le 0\} = \{\langle \theta + (\xi\|\theta\|/\rho\sqrt{n})z, \theta_1 + (\|\theta_1\|/\rho_1)z_1\rangle \le 0\}$, where we used the fact that $\hat{\theta}_{n,\xi} = \theta + (\xi\|\theta\|/\rho\sqrt{n})z$ and $XY \overset{d}{=} \theta_1 + (\|\theta_1\|/\rho_1)z_1$. Since the third moments are bounded, applying Berry-Esseen theorem, we get that the probability of error is bounded by $\Phi(-(\langle\theta,\theta_1\rangle\rho_1\rho\sqrt{n}/(\xi\|\theta_1\|\|\theta\|\sqrt{d}))) + O(1/\sqrt{d})$. This gives $\mathrm{Acc}_{\theta_1,\rho_1} = \Phi(\langle\theta,\theta_1\rangle\rho_1\rho\sqrt{n}/(\xi\|\theta_1\|\|\theta\|\sqrt{d})) + O(1/\sqrt{d})$, and consequently

$$\Phi^{-1}(\mathrm{Acc}_{\theta_1,\rho_1}) = \cos(\theta_1,\theta)\,\rho_1\,\frac{\rho}{\xi}\sqrt{\frac{n}{d}} + O\Big(\frac{e^{\frac{\cos(\theta_1,\theta)^2\rho_1^2\rho^2 n}{2\xi^2 d}}}{\sqrt{n}}\Big). \tag{A.3}$$

This proves the desired claim.

A.6.2 Proof of Theorem 2

Recall that $\mathrm{Slope}(\hat{\theta}_n(\mathcal{D}_{n,\theta,\rho})) = c\langle\theta_2,\theta\rangle/\langle\theta_1,\theta\rangle$ for a positive constant $c>0$ that does *not* depend on the training data. Although the slope only depends on the training data and algorithm through

θ, we write all the parameters including the sample size n and the training SNR ρ to make it explicit that the results hold for all variations of the sample size and the training algorithm within the class that we assume. It is sufficient to show that this is a monotonic function over θ when we linearly traverse from $\tilde{\theta}_1$ to $\tilde{\theta}_2$, i.e. $\theta(\alpha) = \alpha\tilde{\theta}_1 + (1-\alpha)\tilde{\theta}_2$ for $\alpha \in [0,1]$. Note that $f(\alpha) = c\langle\theta_2, \theta(\alpha)\rangle / \langle\theta_1, \theta(\alpha)\rangle = c_1 + c_2/\langle\theta_1, \theta(\alpha)\rangle$ for some c_1 and c_2 that do not depend on α. The monotonicity follows from the fact that the derivative is

$$\frac{\partial f(\alpha)}{\partial \alpha} \;=\; -c_2 \frac{\langle \theta_1, \tilde{\theta}_1 - \tilde{\theta}_2 \rangle}{\langle \theta_1, \theta(\alpha)\rangle^2} \;,$$

whose sign does not change for any α.

A.6.3 Proof of Theorem 3

The train data distribution satisfies $x_i y_i \sim \mathcal{N}(\theta_{\text{train}}, (\|\theta_{\text{train}}\|/\rho)^2 \mathbf{I})$. Note that filtering does not change the distribution in $d-1$ dimensional subspace orthogonal to $\hat{\theta}_{\text{pretrained}}$, due to rotation invariance of a spherical Gaussian distribution. This implies that $\mathbb{E}[\mathcal{P}_\perp(\hat{\theta}_{\text{filtered}})] = \mathcal{P}_\perp(\theta_{\text{train}})$, where $\mathcal{P}_\perp$ denotes the projection operator to the $d-1$ dimensional subspace orthogonal to $\hat{\theta}_{\text{pretrained}}$. On the other hand, on the direction of $\hat{\theta}_{\text{pretrained}}$, the filtering increases the correlation in expectation: $|\mathbb{E}[\mathcal{P}_\parallel(\hat{\theta}_{\text{filtered}})] - \mathcal{P}_\parallel(\hat{\theta}_{\text{pretrained}})| \le |\mathcal{P}_\parallel(\theta_{\text{train}}) - \mathcal{P}_\parallel(\hat{\theta}_{\text{pretrained}})|$, where $\mathcal{P}_\parallel$ denotes the projection operator to the one dimensional subspace spanned by $\hat{\theta}_{\text{pretrained}}$. This implies that $\mathbb{E}[\hat{\theta}_{\text{filtered}}] = \theta_{\text{train}} + c\hat{\theta}_{\text{pretrained}}$ for some positive c. It follows that $\mathbb{E}[\hat{\theta}_{\text{filtered}}]$ is a convex interpolation between two vectors, each in the direction of θ_{trian} and $\hat{\theta}_{\text{pretrained}}$, respectively. We can apply Theorem 2 which gives that

$$\text{Slope}(\hat{\theta}_{\text{unfiltered}}) < \text{Slope}(\mathbb{E}[\hat{\theta}_{\text{filtered}}]) \le \text{Slope}(\hat{\theta}_{\text{pretrained}}) \;,$$

when $\text{Slope}(\hat{\theta}_{\text{unfiltered}}) < \text{Slope}(\hat{\theta}_{\text{pretrained}})$.

Chapter B

Appendix: Rethinking What Popular Curation Filters Leave Behind

B.1 Examples of Translated Data (No Cherry Picking)

Second Floor Plan

Raw caption: Lovely 2nd Floor Plans Part - 4: 2nd Floor Plan
Language detected: eng_Latn
Translation: Lovely 2nd Floor Plans Part - 4: 2nd Floor Plan

Raw caption: ピンク色「インテリアの鮮やかなピンクのカラフルなラウンジ」:スマホ壁紙(19)
Language detected: jpn_Jpan
Translation: Pink: The bright pink lounge of the interior: cell phone wallpaper.

Raw caption: CW’s The Originals Joseph Morgan Jacket
Language detected: eng_Latn
Translation: CW The Originals Joseph Morgan Jacket

Raw caption: PURPLE AZALEAS UP CLOSE
Language detected: yue_Hant
Translation: Purple Azleas up close.

Raw caption: Paso a paso: Cómo sacar el permiso para circular con el auto en vacaciones de verano | Garantia Plus
Language detected: spa_Latn
Translation: Step by step: How to get a driving permit on summer vacation

Raw caption: Een hangende decoratie van Vivi Gade papieren diamantvormen
Language detected: nld_Latn
Translation: A pending decoration of Vivi Gade paper diamond shapes

Raw caption: Neymar n'a plus joué en compétition depuis le 28 novembre 2021, à Saint-Etienne. Icon Sport
Language detected: fra_Latn
Translation: Neymar has not played in a competitive match since 28 November 2021, at Saint-Etienne.

Raw caption: Ring Alarm 5-piece kit (2nd Gen) – home security system with optional 24/7 professional monitoring – Works with Alexa
Language detected: eng_Latn
Translation: Ring Alarm 5-piece kit (2nd Gen) home security system with optional 24/7 professional monitoring Works with Alexa

Raw caption: Alcatras Hapisanesi
Language detected: tur_Latn
Translation: The Alcatras Prison

Raw caption: Coque iPhone XS Max Olaf Reine des neiges bonhomme de neige
Language detected: fra_Latn
Translation: Iphone XS Max Olaf Snow Queen Snowman

Raw caption: Multiracial Thumbs Up Against Blue Sky
Language detected: eng_Latn
Translation: Multiracial Thumbs Up Against Blue Sky

Raw caption: Pat dormitor Bastide L140K, matrimonial, alb + stejar, 140 x 190 cm, 2C
Language detected: cat_Latn
Translation: Bedroom floor Bastide L140K, married, white + stejar, 140 x 190 cm, 2C

Raw caption: Tissu Coton imprimé LittleBird Voyage spatial sur fond Blanc
Language detected: fra_Latn
Translation: Printed cotton fabric LittleBird Space travel on a white background

Raw caption: Large size Print Oil Painting Wall painting pitbull warrior dog Home Decorative Wall Art Picture Living Room paintng No Frame
Language detected: eng_Latn
Translation: Large size Print Oil Painting Wall painting pitbull warrior dog Home Decorative Wall Art Picture Living room painting No Frame

Raw caption: 去年学校野去了waiheke island 景色超美 海水十分清澈
Language detected: deu_Latn
Translation: Last year school camp went to Waiheke island

B.2 More Training Details

We follow the training and evaluation protocols of the DataComp benchmark [57], refer to Appendices M and N of this previous work for more details. To summarize, we use ViT-B/32 as the image encoder for CLIP, and fix the hyperparameters used for training: learning rate 5e-4, 500 warmup steps, batch size 4096, AdamW optimizer $\beta_2 = 0.98$. The training infrastructure is based on the code open-sourced by the DataComp team.

Since the compute budget is fixed, for the DataComp setting (128M training steps), each of our baseline takes about 8 hours with 8 A40 GPUs and 40 CPUs. With the same amount of resources, for experiments involving training for longer (1.28B steps), each baseline takes about 80 hours. We report all baselines that we ran in Appendices B.5, B.7 and B.8.

B.3 Translation Quality

Language	Text cosine similarity after backtranslation (↑)
English	0.886
Norwegian Nynorsk	0.883
Bengali	0.883
Russian	0.860
Norwegian Bokmål	0.839
Marathi	0.271
Irish	0.240
Standard Latvian	0.233
Chechen	0.0595
Karachay-Balkar	0.00280

Table B.1: Top 5 and bottom 5 languages where web-scraped captions observe the highest and lowest translation quality by the No Language Left Behind model [36], out of all the languages detected in our raw data pool. Translation quality is measured by how much the semantic meaning is preserved after the caption is translated into English and subsequently backtranslated into the original language.

Here we provide a quantitative assessment of the quality of caption translation offered by the NLLB model [36]. We sample 100K captions from the raw data pool and backtranslate the English-

translated caption into the original (detected) language (e.g., Chinese text → English translation → Chinese translation of the English-translated text). To evaluate the translation quality, we compute the cosine similarity between the initial web-scraped text and the backtranslated text using embeddings from the multilingual Sentence-BERT model [148]. We find that on average the cosine similarity (and thus, translation quality) remains relatively high (0.63). In the table below, we report the top 5 and bottom 5 languages that observe the highest and lowest translation quality as captured by our metric, computed over at least 30 text samples per language.

B.4 Changes in Data Composition Due to Translation

B.4.1 Differences in Data Between "Filtered Raw Captions" and "Filtered Translated Captions"

In this section, we provide some statistics of the differences in image-text pairs selected for "Filtered raw captions" and "Filtered translated captions", when both caption distributions are filtered to a similar extent with the public DFN from [52]. We find that at either 20% or 30% selectivity threshold, both filtered subsets have about two-thirds of their images in common. In addition, filtering the initial pool using (image, translated caption) cosine similarity score always leads to translated multilingual captions taking up the majority of the resulting filtered subset. This is not the case when filtering with (image, raw caption) cosine similarity score.

Data subset	Total size (M)	Number of English captions (M)	Number of non-English captions (M)
Top 20% Raw captions	25.6	14.9	10.7
Top 20% Translated captions	25.6	11.0	14.6
Top 20% Raw captions ∩ Top 20% Translated captions	17.1	10.7	6.4
Top 30% Raw captions	38.4	20.4	18.0
Top 30% Translated captions	38.4	15.4	23.0
Top 30% Raw captions ∩ Top 30% Translated captions	25.7	15.0	10.7

Table B.2: Analysis of the number of samples of English and non-English origins in "Filtered raw captions", "Filtered translated captions" and their intersection.

B.4.2 Language Composition of the Filtered Subsets

Below we show the most common languages in top 20% raw captions and top 20% translated captions.

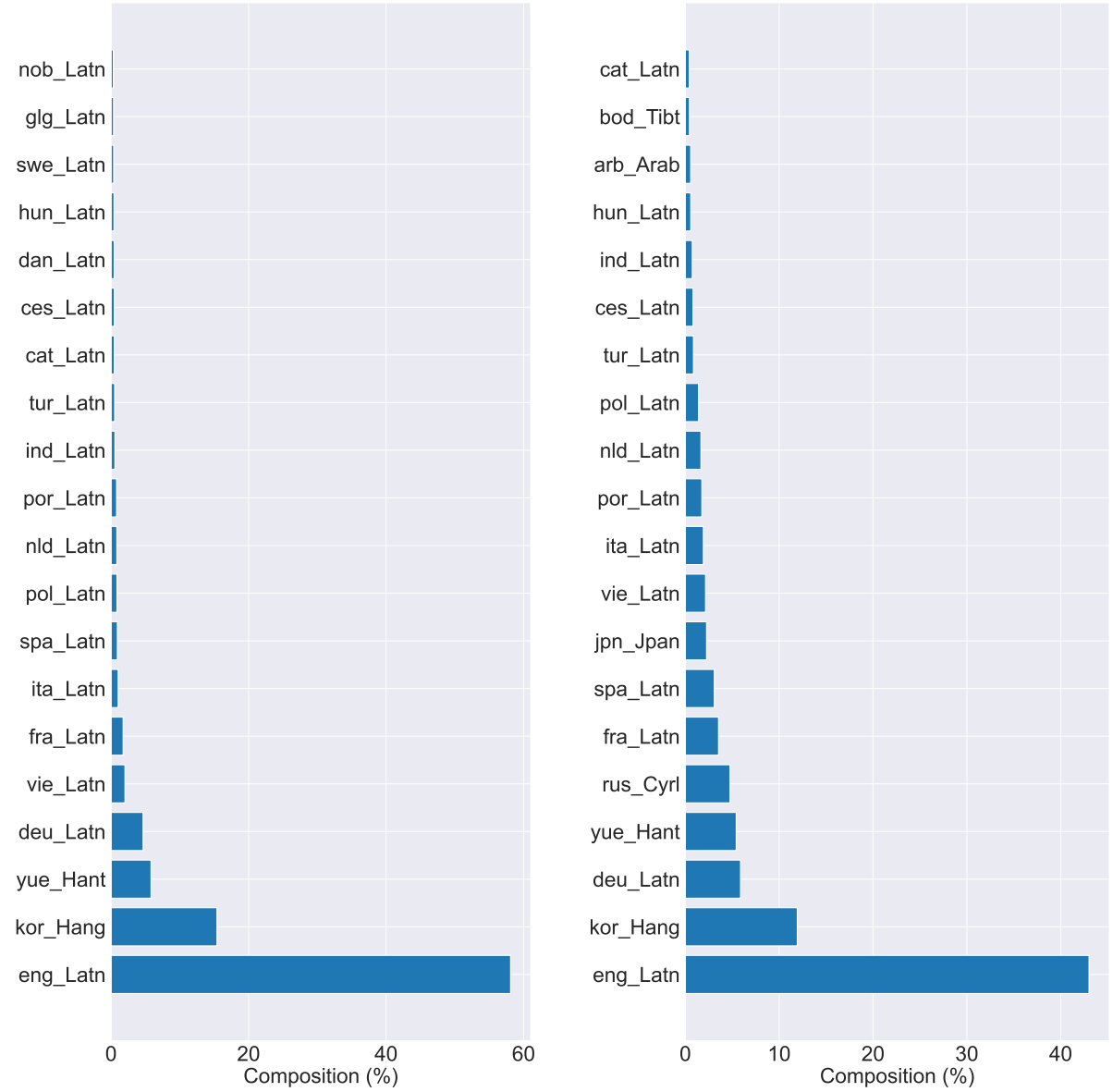


Figure B.1: Top 20 languages that are most common in top 20% raw captions (left) and top 20% translated multilingual captions (right), both are filtered with the public DFN model.

B.4.3 Changes in Language Composition

Comparing image-text pairs selected in top 20% translated captions to those selected in top 20% raw captions, we show below the languages that observe the biggest change in their representation in the resulting training set:

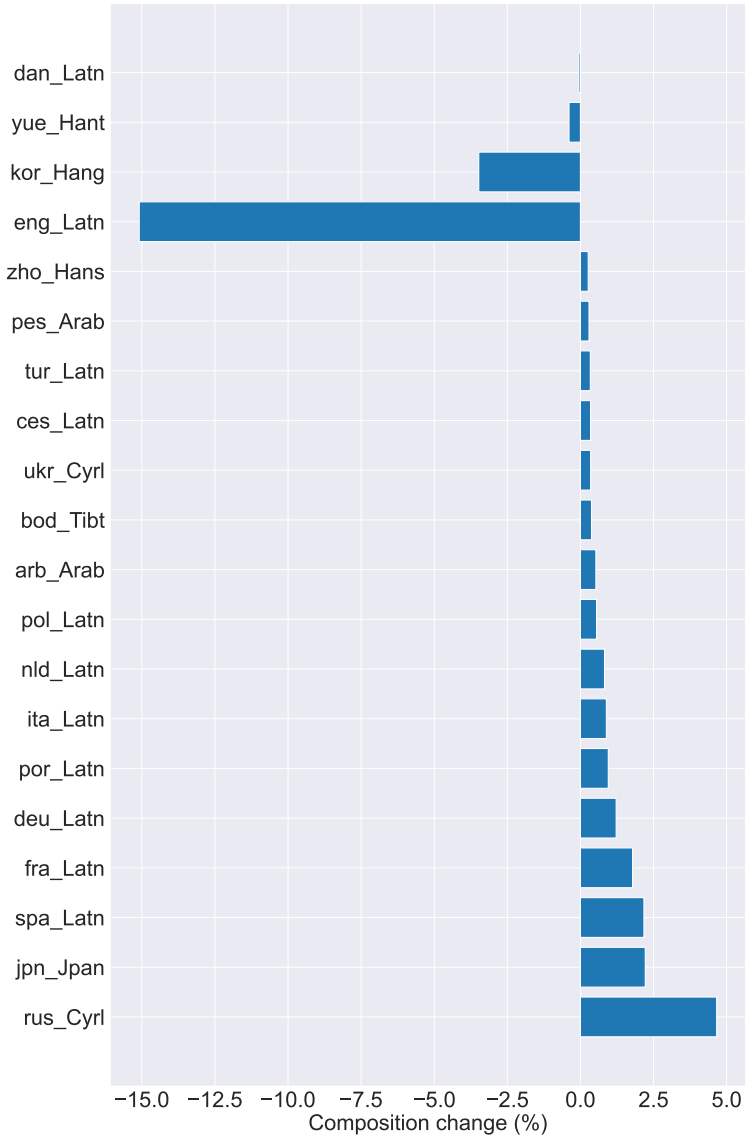


Figure B.2: Languages that see the biggest change (in absolute percentage) in their representation in the final training set when we filter with translated multilingual captions versus with raw web-scraped captions.

B.5 Experiments with OpenAI CLIP Score Filtering

Table B.3 shows results of our experiments with data filtering using OpenAI CLIP-ViT-L/14 [141]. Besides the baselines described in Section 3.5, the table also contains "Top 30% raw captions ∪ top 30% translated captions, using translated caption for all", where we take all the images uncovered from "Top 30% raw captions" and "Top 30% translated captions", deduplicate them, and use the corresponding English-translated captions for all these images.

We find that with OpenAI CLIP as the filtering network, some of our observations from Section 3.5 continue to hold true: (i) using translated multilingual captions is better than using raw captions, and (ii) the performance gain from training with translated captions requires re-filtering the entire data pool after translation (as seen from comparing the first two baselines of the table).

Baseline name	Dataset size	ImageNet	ImageNet shifts	Retrieval	GeoDE	Average over 38 tasks
Top 30% raw captions	38.4M	0.273	0.230	0.251	0.683	0.328
Top 30% raw captions, replaced with translated captions	38.4M	0.260	0.224	0.248	0.660	0.322
Top 30% translated captions	38.4M	**0.292**	**0.250**	0.267	0.695	**0.342**
Top 50% raw captions	64.1M	0.254	0.218	0.262	0.670	0.315
Top 50% translated captions	64.1M	0.265	0.230	**0.276**	**0.704**	0.320
Top 30% raw captions ∪ top 30% translated captions, using translated caption for all	47.7M	0.275	0.234	0.261	0.683	0.326
Top 30% raw captions ∪ top 30% translated captions	47.7M	0.284	0.247	0.260	0.696	0.340
Top 30% raw captions & top 30% translated captions	76.8M	0.289	**0.250**	0.262	0.696	0.335

Table B.3: The benefits of using translated multilingual captions still hold when using cosine similarity score from OpenAI CLIP for filtering. This table shows performance of all baselines we experiment with for filtering with OpenAI CLIP-ViT-L/14. Again, the compute budget is fixed and all baselines are trained for 128M steps. We find that training on filtered translated captions also outperforms training on filtered raw captions in this case.

B.6 Comparison to Training with Synthetic Captions

As observed in Section 3.5.1, training on filtered translated captions outperforms training on filtered raw captions across all major metrics. This could be attributed to changes in the concepts discussed in captions, as well as changes in image data (since "Filtered raw captions" and "Filtered translated captions" only share some of the images in common, see Appendix B.4). Here we attempt to disentangle the contribution to performance gain from these two changes.

Baseline name	Dataset size	ImageNet	ImageNet shifts	Retrieval	GeoDE	Average over 38 tasks
Top 20% translated captions	25.6M	0.329	0.275	0.296	0.709	0.359
Top 20% translated captions, replaced with synthetic captions	25.6M	0.283	0.255	0.350	0.696	0.336
Top 30% translated captions	38.4M	0.311	0.265	0.305	0.718	0.351
Top 30% translated captions, replaced with synthetic captions	38.4M	0.282	0.253	0.371	0.703	0.341

Table B.4: When fixing the training images and replacing translated English captions with synthetic captions generated by BLIP2, we find that performance decreases in general. Since filtering from translated captions exposes CLIP to both new images and new text distributions, we seek to disentangle the impact of these two factors on model performance. Our results suggest that having access to more diverse images alone (without the corresponding translated multilingual captions) may be insufficient for achieving performance gains.

Given the images selected by filtering based on (image, translation caption) cosine similarity (i.e., "Top 20% translated captions", "Top 30% translated captions"), we generate synthetic captions for each image using BLIP2 model [99] and the generation hyperparameters from [123]. With the image component unchanged, training on the new (image, synthetic caption) pairs leads to lower performance overall compared to training on the original (image, translated caption) pairs (Table B.4). This suggests that having access to more diverse (non-English) images in the training set is not sufficient to boost accuracy; the diversity coming from translated multilingual captions is also necessary for observing performance improvement.

We acknowledge that since BLIP2 was pretrained on relatively few multilingual samples, it is possible that the captioning model finds it difficult to caption non-English images. This ablation

study experiment is thus mostly exploratory, and more experiments are needed to assess the performance benefits coming from seeing more diverse (non-English) images, versus seeing more diverse (translated) non-English captions.

B.7 All DFN Filtering Baselines

Baseline name	Dataset size	ImageNet	ImageNet shifts	Retrieval	GeoDE	Average over 38 tasks
Top 20% raw captions	25.6M	0.316	0.260	0.282	0.688	0.350
Top 20% raw captions, replaced with translated captions	25.6M	0.304	0.252	0.268	0.668	0.331
Top 30% raw captions	38.4M	0.297	0.246	0.280	0.663	0.337
Top 40% raw captions	51.2M	0.267	0.222	0.274	0.669	0.320
Top 20% translated captions	25.6M	0.329	0.275	0.296	0.709	0.359
Top 30% translated captions	38.4M	0.311	0.265	**0.305**	0.718	0.352
Top 40% translated captions	51.2M	0.289	0.248	0.288	0.709	0.332
Top 20% raw English-only captions	8.0M	0.260	0.218	0.234	0.603	0.303
Top 30% raw English-only captions	12.0M	0.280	0.238	0.259	0.630	0.326
Top 40% raw English-only captions	16.0M	0.283	0.236	0.278	0.666	0.327
Top 50% raw English-only captions	20.0M	0.277	0.236	0.280	0.668	0.321
Top 20% raw captions ∪ top 20% translated captions, using translated caption for all	34.2M	0.316	0.265	0.289	0.716	0.353
Top 20% raw captions ∪ top 20% translated captions	34.2M	0.329	0.271	0.298	0.720	**0.364**
Top 20% raw captions & top 20% translated captions	51.2M	**0.336**	**0.280**	0.301	**0.725**	0.361
Top 30% raw captions & top 30% translated captions	76.8M	0.295	0.248	0.282	0.673	0.340

Table B.5: Here we report all the baselines we experiment with using the public DFN from [52] for filtering; all models are trained for 128M steps as set by the DataComp benchmark. For each caption distribution (i.e., raw/ translated/ English-only), only the filtering threshold that yields the best average performance across 38 tasks is shown in Table 5.1.

B.8 Training for Longer

Here we show the results of all the baselines that we train for 1.28B steps (i.e., 10× the number of steps set by DataComp). In Table B.6, we find that when using either raw web-crawled captions or English-translated captions, filtering for top 30% of the pool does best, and translated multilingual captions continue to yield better performance on standard metrics compared to raw captions. We also note that the performance gaps between our best baseline (that leverages translated multilingual caption) and just using filtered raw captions widen with training duration (see Table 5.1 for a comparison).

Baseline name	Dataset size	ImageNet	ImageNet shifts	Retrieval	GeoDE	Average over 38 tasks
Top 20% raw captions	25.6M	0.423	0.345	0.331	0.751	0.407
Top 30% raw captions	38.4M	0.414	0.340	0.344	0.742	0.414
Top 40% raw captions	51.2M	0.417	0.344	0.358	0.746	0.410
Top 20% translated captions	25.6M	0.421	0.348	0.346	0.754	0.412
Top 30% translated captions	38.4M	0.427	0.347	0.352	0.771	0.414
Top 40% translated captions	51.2M	0.421	0.348	0.346	0.754	0.412
Top 20% raw captions ∪ top 20% translated captions	34.2M	0.441	0.359	0.353	0.775	0.427
Top 20% raw captions & top 20% translated captions	51.2M	**0.456**	**0.369**	**0.371**	**0.776**	**0.435**
Top 30% raw captions & top 30% translated captions	76.8M	0.419	0.347	0.345	0.771	0.429

Table B.6: When the training duration is increased by 10× compared to the DataComp setting, training on translated multilingual captions continues to outperform training on raw captions across a range of metrics; using both sources of captions continues to yield the best performance. We show performance of all the baselines that are trained for 1.28B steps. Even though using filtered raw captions and using filtered translated captions yield similar average performance (0.414 percentage points), the latter still surpasses the former on ImageNet, ImageNet distribution shifts, retrieval and GeoDE (worst-region accuracy).

We provide a breakdown of performance differences between "Filtered raw captions" and "Filtered translated captions" across 38 tasks from DataComp, with the new increased training duration, in Figure B.3.

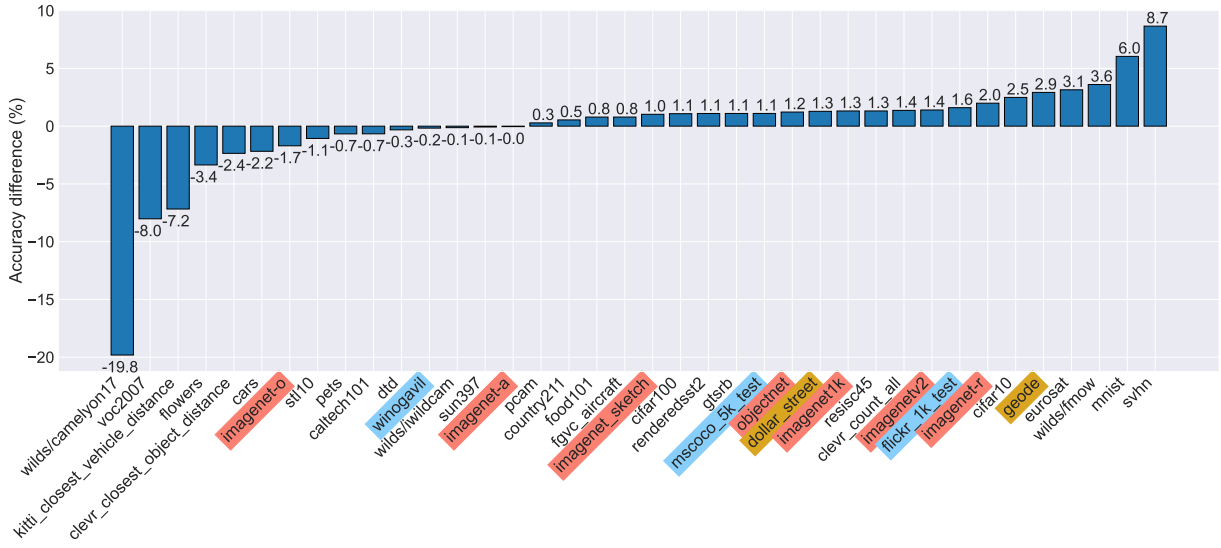


Figure B.3: With the same degree of filtering, training with (image, translated caption) pairs improves performance on 23 out of 38 tasks compared to training with (image, raw caption) pairs, including ImageNet, the majority of ImageNet distribution shifts and retrieval tasks, and tasks with geographically diverse inputs. We compare performance on each task of the DataComp benchmark between training with raw captions and training with translated captions, when both are trained for 1.28B steps. Both datasets have also been filtered with cosine similarity scores output by the public DFN [52] to select the top 30% examples. We find that when we increase training duration to be 10× longer than DataComp's setting, using translated multilingual captions and using raw captions yield similar average performance across 38 tasks. However, the former still outperforms the latter on most of the ImageNet distribution shifts (**red**), retrieval (**blue**) and fairness-related tasks (**dark yellow**).

B.9 More Performance Analysis

In addition to the analysis in 3.5.3, we provide further breakdown of performance changes at the income group and class levels, on Dollar Street (Figure B.4) and ImageNet (Figure B.5) respectively, when we swap the training distribution from filtered raw captions to filtered translated captions.

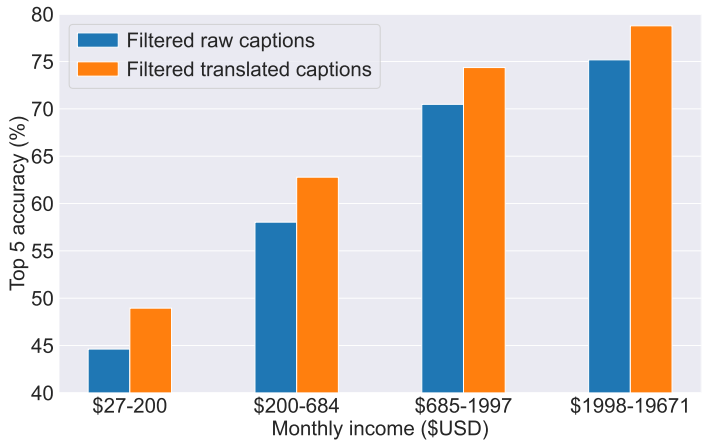


Figure B.4: On Dollar Street, using translated multilingual captions leads to performance improvement across all income groups. Dollar Street [151] is another fairness-related task that involves classifying images of everyday items collected from households around the world with different socioeconomic backgrounds. We break down the performance on this dataset by income groups and find that training on top-quality translated captions improves the classification accuracy across all groups, compared to training on top-quality raw captions.

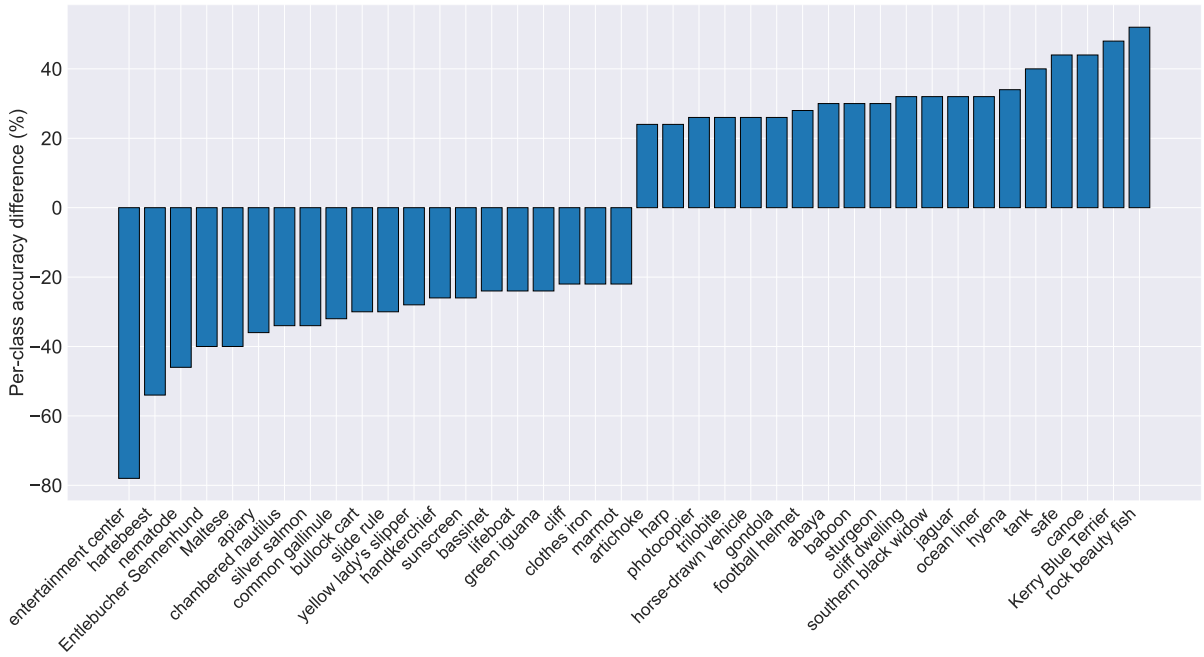


Figure B.5: 40 ImageNet classes that observe the largest changes in classification performance when we train on top translated multilingual captions compared to top raw captions. We show 40 categories from ImageNet that see the biggest change in accuracy when more (translated) multilingual data is included in the training set.

Chapter C

Appendix: Synthetic Captions as a Data Quality Fix

C.1 More Examples of Image-Text Pairs (No Cherry Picking)

Raw: *2003 Mercedes-Benz C240 sedan, Leather, MUST BE SEEN - $6199*

BLIP (finetuned): *a couple of cars parked in a parking lot with trees and cars*

BLIP2: *2002 mercedes-benz c-class for sale*

BLIP2 (finetuned): *a blue mercedes benz car parked in a parking lot next to yellow cars*

OpenCLIP-CoCa: *find used 2 0 0 1 mercedes benz c 2 4 0 base sedan 4 door 2 5 l for 2 0 0 1 mercedes benz c 2*

OpenCLIP-CoCa (finetuned): *a blue mercedes parked on the side of a road .*

Raw: *Gaziburma Ünal is one of Gespeicherte Orte von Can.*

BLIP (finetuned): *dozens of trays of different types of treats at a food stand*

BLIP2: *some trays of pastries and a knife*

BLIP2 (finetuned): *there are many trays filled with food items from the store*

OpenCLIP-CoCa: *baklava , sweets , pastries*

OpenCLIP-CoCa (finetuned): *there are trays full of different types of food .*

Raw: *Open Virgin of Kazan, Wooden Egg with Stand, Blue*

BLIP (finetuned): *a gray and white logo with the words more info in a rectangular shape*

BLIP2: *a white button with the word more info*

BLIP2 (finetuned): *more information is shown on a white button with an orange background*

OpenCLIP-CoCa: *home - page - button . png*

OpenCLIP-CoCa (finetuned): *a picture of a close up of a text message*

Raw: *hair oil*

BLIP (finetuned): *smiling blonde woman blow drying hair in a salon while getting a mani*

BLIP2: *hair stylist using hair spray in beauty salon*

BLIP2 (finetuned): *a person is using a hairdryer to blow dry a long blonde hair*

OpenCLIP-CoCa: *female hairdresser styling a long blond hair with hairspray in a beauty salon . concept : hair care , hair straightening , hair color correction .*

OpenCLIP-CoCa (finetuned): *a person is spraying a hair dryer on a long blonde hair .*

Raw: *Italien - Ligurien*

BLIP (finetuned): *beige colored building with tan accents and palm trees on both sides of walkway*

BLIP2: *house in villa marina, a villa with many rooms and*

palm trees

BLIP2 (finetuned): *a park with lots of trees and benches in front of a large building*

OpenCLIP-CoCa: *residence - villa - maria - di - san - giovanni - near - the - sea - in - taormina*

OpenCLIP-CoCa (finetuned): *a picture of a large building with a bunch of palm trees .*

Raw: *3 formas de pedir la mano de tu novia - wikiHow*

BLIP (finetuned): *crates stacked up in a pile on top of each other*

BLIP2: *the building contains five floors of wooden planks*

BLIP2 (finetuned): *a big pile of wooden planks stacked together*

OpenCLIP-CoCa: *the cost of wood pallets*

OpenCLIP-CoCa (finetuned): *a large pile of wooden pallets mounted to a wall .*

Raw: *lutz*

BLIP (finetuned): *blond haired man in black suit looking at camera*

BLIP2: *a man sitting on a chair with a blank background*

BLIP2 (finetuned): *a man sitting in a chair with a lapel button in front*

OpenCLIP-CoCa: *actor tilda swinton is pictured during a press conference for the film ' a dangerous method ' at the 2 0 1 1 toronto film festival*

OpenCLIP-CoCa (finetuned): *a person sitting on a chair wearing a suit and tie .*

Raw: *10840 SW 126th St photo067*

BLIP (finetuned): *overview of a large backyard with a swimming pool and patio*

BLIP2: *3344 sw 7th st, palm beach*

BLIP2 (finetuned): *a house with a pool from above, with a yard*

OpenCLIP-CoCa: *home for sale in country club shores west palm beach florida*

OpenCLIP-CoCa (finetuned): *aerial image of a pool that has a little bit of shade by the side .*

Raw: *image8.JPG*

BLIP (finetuned): *members of a school play soccer in a gymnasium with a crowd*

BLIP2: *a large crowd of kids perform during a dance show*

BLIP2 (finetuned): *a group of young children standing on the basketball court*

OpenCLIP-CoCa: *kid dressed in white standing in a gym area .*

OpenCLIP-CoCa (finetuned): *a group of kids on the gym floor with fans on the floor .*

Raw: *Automne hiver enfants manteau et pantalon ensemble capuche veste de Ski et pantalon garçon fille coupe-vent imperméable en plein air camping randonnée*

BLIP (finetuned): *a man wearing a red and blue jacket and a pair of pants and a pair of sneakers*

BLIP2: *the arctic light hooded jacket and pants set*

BLIP2 (finetuned): *the colors of the jacket match the pant color of the jacket*

OpenCLIP-CoCa: *the arctic light 2 0 1 7 children 's clothing sets winter kids ski suits sets windproof waterproof warm jackets coats pants boys set*

OpenCLIP-CoCa (finetuned): *a child standing in their ski wear and a jacket and pants*

Raw: *Women Personality Creative Christmas Hat Face Expression Gold Earring Funny Cartoon Ear Stud Jewelry Accessories Gift Hot*

BLIP (finetuned): *red and gold tone emojt earring*

BLIP2: *kawaii santa emoticuos en la cabeza*

BLIP2 (finetuned): *a pair of emoji earrings with faces and hats*

OpenCLIP-CoCa: *best christmas gift for her new arrivals emoji earrings christmas emoji earrings*

OpenCLIP-CoCa (finetuned): *a pair of gold earrings with a smiley face and a christmas hat .*

Raw: *Nautica NAPTYR005*

BLIP (finetuned): *navitta mens stainless steel bracelet watch with blue dial*

BLIP2: *nautica men's chronograph watch*

BLIP2 (finetuned): *nautica men's men's chronograph black dial stainless steel bracelet watch*

OpenCLIP-CoCa: *nautica newport chronograph n 2 2 0 0 3 g*

OpenCLIP-CoCa (finetuned): *a mans black watch is shown with red and blue accents*

Raw: *Greenberg Weathered Marble Plush Ivory Area Rug*

BLIP (finetuned): *grey rug with a text home on it by a table*

BLIP2: *a grey area rug on a wooden floor*

BLIP2 (finetuned): *a white coffee table with a sign saying home on it. it is sitting on a cream colored rug*

OpenCLIP-CoCa: *rugs and carpets in hyderabad : buy online at best price in ...*

OpenCLIP-CoCa (finetuned): *a rug is shown in a living room with a chair .*

Raw: *productivity, productivity, productivity*

BLIP (finetuned): *drivers guide to the truck industry*

BLIP2: *buy and sell truck parts*

BLIP2 (finetuned): *a white truck with a cover on it drives along a highway*

OpenCLIP-CoCa: *how the trucking industry is changing*

OpenCLIP-CoCa (finetuned): *there are some trucks on the road .*

Raw: *Amigas*

BLIP (finetuned): *crowd of people outside a wedding ceremony near several trees*

BLIP2: *a wedding ceremony in the middle of the street*

BLIP2 (finetuned): *a black and white photograph of a number of women in prom dresses*

OpenCLIP-CoCa: *2 0 1 3 0 8 0 5 _ wedding _ carlenan _ 0 0 3*

OpenCLIP-CoCa (finetuned): *a group of people hugging and talking in a group*

Raw: *Der Lieferumfang*

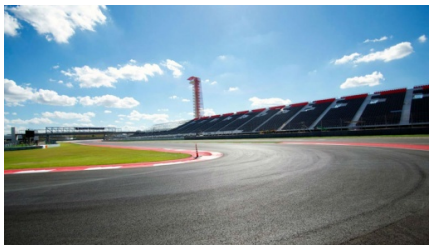
BLIP (finetuned): *there are several electronics laid out on the table ready to be used*

BLIP2: *samsung galaxy s10e review | a quick tour of the samsung galaxy s10e*

BLIP2 (finetuned): *wireless charging case and remote control, both packaged in the box*

OpenCLIP-CoCa: *best - wireless - chargers - for - samsung - galaxy - note - 8 - s 8 - and - iphone - 8*

OpenCLIP-CoCa (finetuned): *a set of various electronic items sitting on a table .*



Raw: *Autozone*

BLIP (finetuned): *racing track with a line of seats and a sky background*

BLIP2: *a photo of a grand prix race track, under a blue sky*

BLIP2 (finetuned): *the circuit track is empty, but the sun beams into the sky*

OpenCLIP-CoCa: *circuit of the americas*

OpenCLIP-CoCa (finetuned): *a red and white pole next to a racing track*

Raw: *2016.07.01 Nametags with Pronouns - Avery 5392_non-branded*

BLIP (finetuned): *there are no pictures here to provide a caption for*

BLIP2: *hello, my name is name, my pronouns are pronouns*

BLIP2 (finetuned): *a blue and white label with a blue and white text*

OpenCLIP-CoCa: *1 5 + hello my name is names pronunciations and meanings*

OpenCLIP-CoCa (finetuned): *hello my name is , my pronouns are .*

Raw: *Women long sleeve t shirt 2015 Fashion shirts woman Full Comfortable leisure fashion womens long sleeve tops*

BLIP (finetuned): *the qaoo loading logo is shown above the qaoo loading logo*

BLIP2: *qoo10 loading logo on white*

BLIP2 (finetuned): *a picture of an image of a phone screen showing a loading sign*

OpenCLIP-CoCa: *loading _ 1 1 1 1 2 0 _ 0 1 . png*

OpenCLIP-CoCa (finetuned): *a light grey font and a dark grey font with a large white background*

Raw: *1173x1500 Awesome Adult Coloring Pages Printable Zentangle Design*

BLIP (finetuned): *chinese dragon coloring pages dragon coloring pages for adults to print coloring pages*

BLIP2: *dragon coloring pages with large and large dragon*

BLIP2 (finetuned): *a circle with a dragon on it in the center*

OpenCLIP-CoCa: *the 2 5 best chinese dragon drawing ideas on pinterest chinese*

OpenCLIP-CoCa (finetuned): *a chinese dragon looks like a dragon from the movie the karate kid*

C.2 Experiment Details

Refer to Appendices M and N of the DataComp benchmark [56] for training and evaluation details. To summarize, both `small` and `medium` scales use ViT-B/32 as the image encoder for CLIP, in addition to fixing the hyperparameters used for training: learning rate 5e-4, 500 warmup steps, batch size 4096, AdamW optimizer $\beta_2 = 0.98$. `Large` scale training uses the same hyperparameters, but with batch size 8192 and ViT-B/16 as the image encoder.

Using DataComp infrastructure and the AWS EC2 cloud, a `small` model takes 4 A100 hours to train, while `medium` requires 40 A100 hours and `large` utilizes 960 A100 hours. We additionally report CLIP ViT-L/14 and BLIP2 (OPT 2.7B backbone) inference costs. Recall that we run both of these models on the DataComp's `large` pool to curate the datasets used in this paper. For the CLIP model, we measure throughput at 490 samples per second on a single A100. For BLIP2, we get 75 samples per second on the same hardware. Hence, for the `large` pool of 1.28B samples, we spend 725 A100 hours computing CLIP features and 4,740 A100 hours generating BLIP2 captions.

While the annotation cost (i.e., BLIP2 caption and CLIP score generation) is 6× larger than a single training run proposed by the DataComp benchmark (which is equivalent to doing one pass through

the entire candidate pool), this additional cost can be easily amortized with more training epochs over the final training set, as well as with training different downstream models on the improved dataset. For reference, OpenAI trained various CLIP models on the same set of 400M curated image-text pairs; the best performing model was trained on 256 GPUs for 2 weeks, totalling about 86,000 GPU hours [1]. This scale of training is common among existing large vision models. Future work could explore the option of adaptively allocating compute to CLIP training and synthetic caption annotation given a fixed compute budget.

C.3 Temperature Ablations

Captioning model	Metric	T=0.5	T=0.75	T=1.0	T=1.5
BLIP (finetuned)	ImageNet accuracy	-	0.207	**0.212**	-
	Average accuracy	-	0.303	**0.312**	-
BLIP2	ImageNet accuracy	0.212	**0.281**	0.280	0.251
	Average accuracy	0.300	**0.357**	0.353	0.332
BLIP2 (finetuned)	ImageNet accuracy	-	0.227	**0.234**	0.221
	Average accuracy	-	0.325	**0.326**	0.311
OpenCLIP-CoCa	ImageNet accuracy	0.306	**0.321**	0.314	-
	Average accuracy	0.366	**0.371**	0.370	-
OpenCLIP-CoCa (finetuned)	ImageNet accuracy	-	0.252	**0.264**	0.262
	Average accuracy	-	0.364	**0.374**	0.364

Table C.1: Performance on ImageNet and averaged across 38 tasks when training on the captions generated by captioning models in Table 4.1, with different softmax temperatures. We find that T = 0.75 and T = 1.0 generally lead to good performance for CLIP training.

[1] https://openai.com/research/clip

C.4 More Filtering Baselines

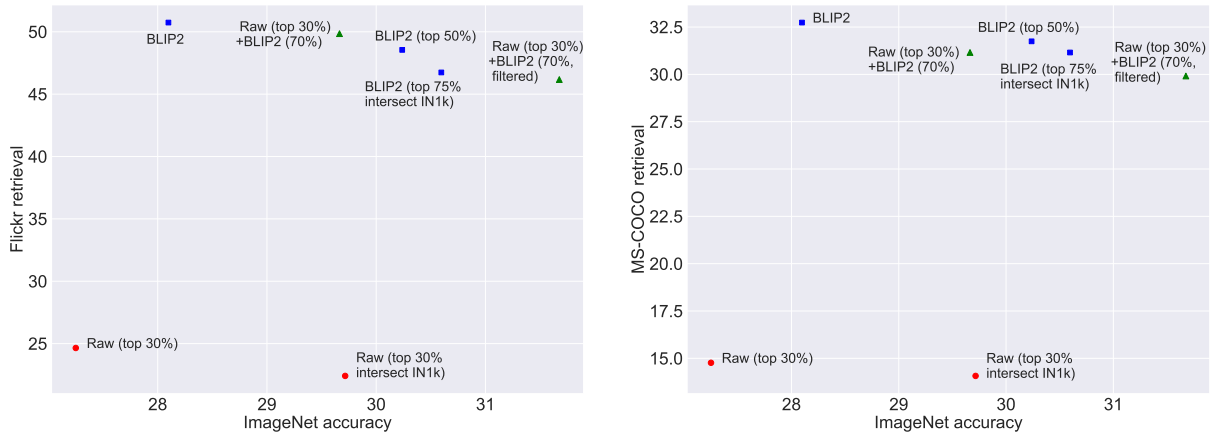


Figure C.1: Retrieval performance on Flickr (left) and MS-COCO (right) versus ImageNet accuracy for select baselines. Similar to the findings in Figure 4.2, we find that using only BLIP2 captions or mixing them with raw captions in the training data significantly boosts retrieval performance.

Baseline	Training set size	ImageNet accuracy	Average accuracy
`small` scale (12.8M candidate pool, 12.8M training steps)			
Raw captions (no filtering)	12.8M*	0.025*	0.132*
BLIP2 captions (no filtering)	12.8M	0.076	0.200
Raw captions (top 30%)	3.8M*	0.051*	0.173*
BLIP2 captions (top 50%)	6.4M	**0.080**	**0.203**
Raw captions (top 30% intersect IN1k)	1.4M*	0.039*	0.144*
BLIP2 captions (top 75% intersect IN1k)	2.4M	0.073	0.192
Raw captions (top 30%) + BLIP2 captions (70%, filtered), intersect IN1k	2.2M	0.045	0.153
Raw captions (top 30%) + BLIP2 captions (70%, filtered)	8.4M	0.076	0.197
`medium` scale (128M candidate pool, 128M training steps)			
Raw captions (no filtering)	128M*	0.176*	0.258*
BLIP2 captions (no filtering)	128M	0.281	0.357
Top BLIP2 captions across all temperatures (no filtering)	128M	0.293	0.368
Raw captions (top 30%)	38M*	0.273*	0.328*
BLIP2 captions (top 50%)	64.1M	0.302	0.370
Raw captions (top 30% intersect IN1k)	14.0M*	0.297*	0.328*
BLIP2 captions (top 75% intersect IN1k)	23.6M	0.306	0.360
Raw captions (top 30%) + BLIP2 captions (70%, filtered), intersect IN1k	22.2M	0.281	0.314
Raw captions (top 30%) + BLIP2 captions (70%, filtered)	83.6M	**0.317**	0.368
BLIP2 captions (top 50%) + Raw captions (50%, filtered)	75.3M	0.310	**0.376**
`large` scale (1.28B candidate pool, 1.28B training steps)			
Raw captions (no filtering)	1.28B*	0.459*	0.437*
BLIP2 captions (no filtering)	1.28B	0.487	0.505
Raw captions (top 30%)	384M*	0.578*	0.529*
BLIP2 captions (top 50%)	641M	0.526	0.522
Raw captions (top 30% intersect IN1k)	140M*	0.631*	0.537*
BLIP2 captions (top 75% intersect IN1k)	237M	0.533	0.527
Raw captions (top 30%) + BLIP2 captions (70%, filtered), intersect IN1k	222M	**0.643**	0.549
Raw captions (top 30%) + BLIP2 captions (70%, filtered)	834M	0.598	**0.551**

Table C.2: Performance for select baselines at `small`, `medium`, and `large` scales of DataComp. * indicates numbers obtained from the original paper [56]. Underlined numbers are best-performing baselines from the DataComp benchmark, trained on only raw web-crawled captions. Bolded numbers are the updated best-performing methods after comparing with baselines involving synthetic captions. In general, given a fixed training budget, it is helpful to include more samples in the training pool by carefully replacing noisy raw captions with synthetic captions (i.e., RAW (TOP 30%) + BLIP2 (70%, FILTERED) versus RAW (TOP 30%)). We experiment with many more filtering and mixing methods at the `medium` scale and report how the performance varies with CLIP score filtering threshold, see Table C.3.

CLIP score filtering	10%	20%	30%	50%	75%	90%
Cosine similarity threshold						
Raw captions	0.295	0.266	0.243	0.203	0.160	0.129
BLIP2 captions	0.315	0.292	0.277	0.251	0.217	0.187
Only raw captions						
Training set size	12.8M*	25.7M*	38.4M*	64.1M*	96.1M*	115M*
ImageNet accuracy	0.198*	0.260*	0.273*	0.254*	0.212*	0.188*
Average accuracy	0.277*	0.322*	0.328*	0.315*	0.285*	0.266*
Only BLIP2 captions						
Training set size	12.8M	25.6M	38.5M	64.1M	96.0M	115M
ImageNet accuracy	0.146	0.249	0.275	0.302	0.300	0.293
Average accuracy	0.254	0.333	0.356	0.370	0.365	0.366
Only BLIP2 captions, for top % based on cosine similarity of image and *raw* text						
Training set size	12.8M	25.7M	38.4M	64.1M	96.1M	115M
ImageNet accuracy	0.192	0.245	0.261	0.266	0.267	0.276
Average accuracy	0.280	0.330	0.346	0.342	0.349	0.356
Raw captions for top % + BLIP2 captions for the remaining examples						
Training set size	128M	128M	128M	128M	128M	128M
ImageNet accuracy	0.286	0.296	0.297	0.286	0.250	0.215
Average accuracy	0.360	0.357	0.365	0.349	0.323	0.293
Raw captions for top % + BLIP2 captions for the remaining examples, subject to the same cosine similarity threshold						
Training set size	30.5M	59.5M	83.6M	114M	127M	128M
ImageNet accuracy	0.267	0.310	**0.317**	0.296	0.251	0.212
Average accuracy	0.343	0.372	0.368	0.352	0.313	0.285
BLIP2 captions for top % + raw captions for the remaining examples, subject to the same cosine similarity threshold						
Training set size	17.1M	32.8M	47.7M	75.3M	105M	121M
ImageNet accuracy	0.212	0.272	0.298	0.310	0.298	0.285
Average accuracy	0.305	0.353	0.367	**0.376**	0.375	0.355
Concatenate raw & BLIP2 captions for top % + BLIP2 captions for the remaining examples, subject to the same cosine similarity threshold						
Training set size	30.5M	59.5M	83.6M	114M	127M	128M
ImageNet accuracy	0.250	0.287	0.299	0.286	0.269	0.262
Average accuracy	0.336	0.368	0.367	0.359	0.340	0.337
Top % raw captions + top % BLIP2 captions						
Training set size	25.6M	51.3M	76.9M	128M	-	-
ImageNet accuracy	0.238	0.285	0.297	0.300	-	-
Average accuracy	0.318	0.358	0.366	0.356	-	-
BLIP2 captions - top % intersect with examples from IN1k clustering						
Training set size	-	-	10.0M	16.4M	23.6M	27.1M
ImageNet accuracy	-	-	0.243	0.289	0.306	0.301
Average accuracy	-	-	0.310	0.343	0.360	0.344

Table C.3: Summary of how various filtering and mixing strategies perform on ImageNet and on average across 38 evaluation tasks in DataComp, given a 128M candidate pool (`medium` scale). * indicates numbers obtained from [56]. Note that all resulting training sets are trained for a fixed number of steps (128M samples seen) and other training variables (e.g., architecture, hyperparameters) are kept constant. Synthetic captions are generated using pretrained BLIP2 model with top-K sampling (K = 50) and softmax temperature 0.75. We find that at this scale, approaches that yield the best ImageNet and average accuracies leverage a combination of raw and synthetic captions.

C.5 Synthetic Caption Analysis

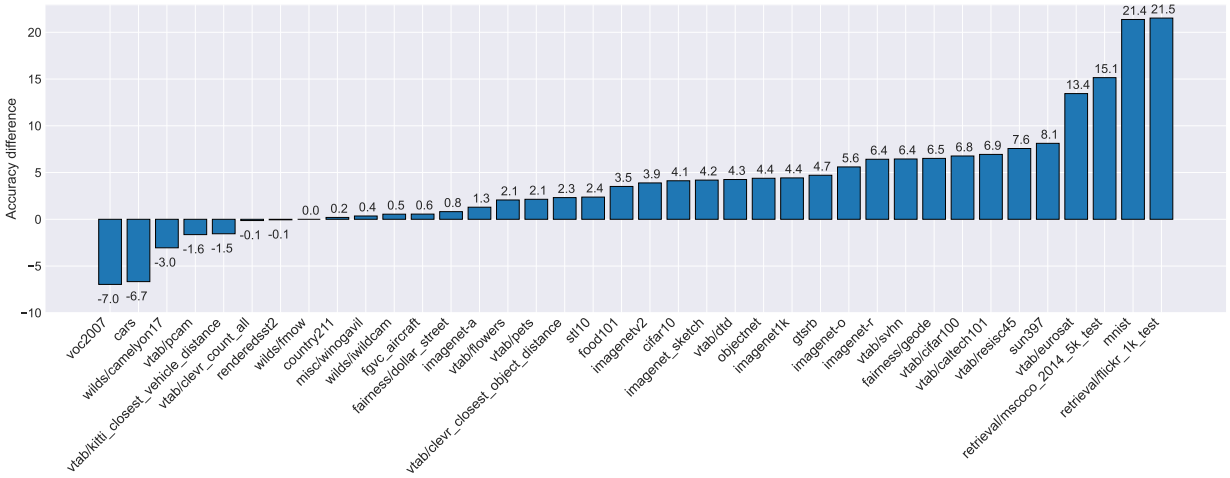


Figure C.2: We find that expanding a training set of filtered raw data by using BLIP2 captions for some of the discarded images improves performance on 30 out of 38 evaluation tasks, in addition to boosting average accuracy by 4%. We compare performance on each task between training on the top 30% of examples with raw captions (based on CLIP score) and training on the same set of examples but with the addition of BLIP2 captions for the remaining 70% images, filtered by the same CLIP score threshold. In Table C.2, we have shown that adding BLIP2 captions improves ImageNet accuracy by 4.4% and average accuracy by 4%. With this breakdown, we find that the performance improvement applies to most of the tasks in the evaluation set, especially retrieval.

We also investigate whether there are systematic differences in training with raw versus generated text when it comes to recognizing certain object categories. To do so, we examine two CLIP models that perform similarly on ImageNet (i.e., ±0.2%): one trained on only raw captions and one trained on only generated captions, both training sets have been filtered with CLIP score ranking to select the top 30% image-text pairs. In Figure C.3, we analyze performance on each ImageNet class, categorized as either 'living' or 'non-living'

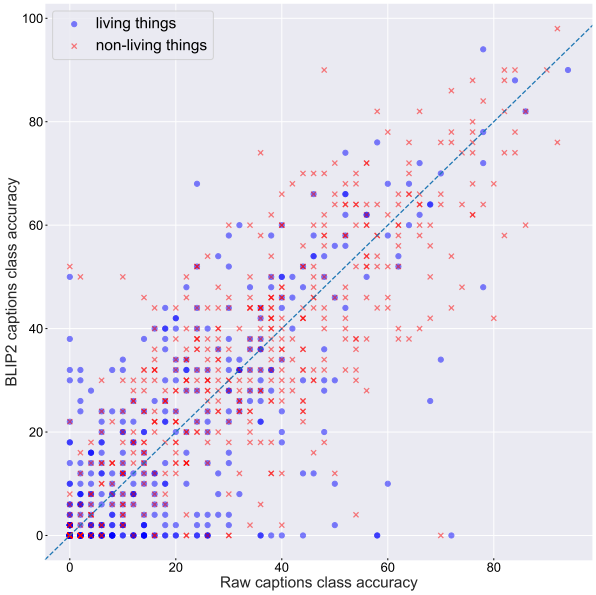


Figure C.3: We break down per-class performance on ImageNet, between a CLIP model trained on only raw captions and one trained on only synthetic captions with similar overall ImageNet accuracy. We find no systematic trends in the performance of either model when it comes to classifying 'living' or 'non-living' things.

thing based on where the classname synset is located in the WordNet hierarchy. We observe that class-wise classification performances are scattered evenly around the $y = x$ line, indicating that compared to web-crawled captions, synthetic captions do not exhibit a particular disadvantage on either 'living' or 'non-living' concepts.

C.6 Performance at Scale

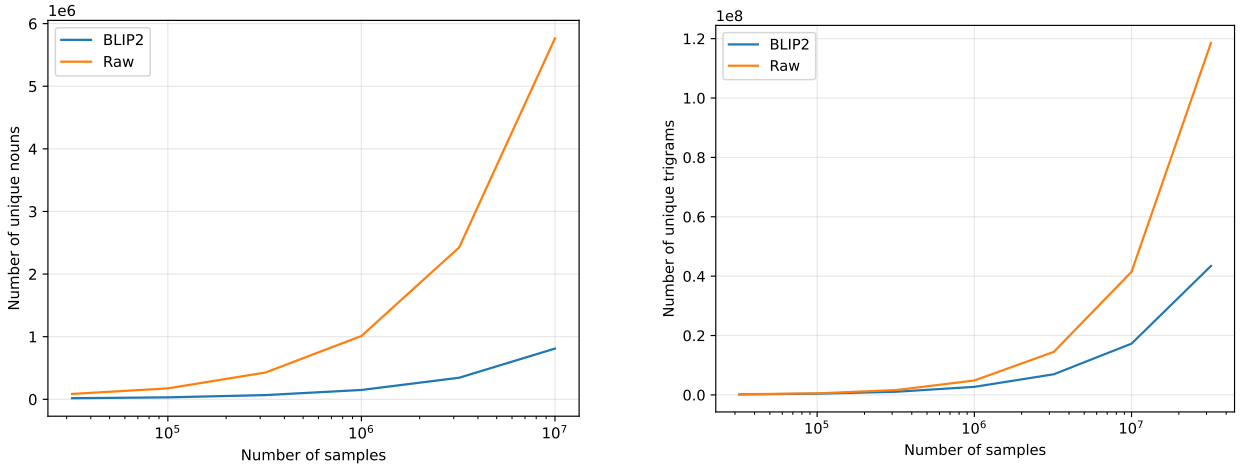
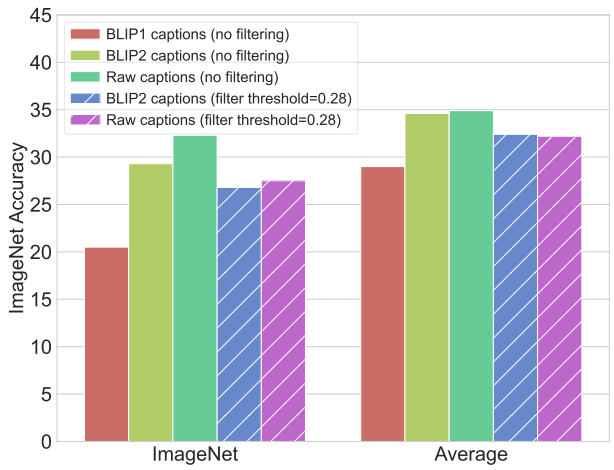


Figure C.4: Our simple analyses of text properties suggest that the text diversity provided by synthetic captions may not scale as well as that of raw captions scraped from the Internet. We measure the number of unique nouns and unique trigrams in random subsets of BLIP2 and raw captions of various sizes. We observe that on both metrics, the scaling trend for generated captions is worse than that of raw captions. This increasing gap in data diversity may impact the performance benefits we can expect to obtain from using synthetic captions, when dealing with a larger scale of training data.

C.7 Experiments with LAION-COCO

Our experiments with synthetic captions are partly inspired by the release of LAION-COCO dataset [161], which used BLIP [100] with various hyperparameter settings to caption LAION-5B data [160], and then selected the top synthetic caption for each image based on the cosine similarity output by OpenAI's CLIPs [141]. We pick a random set of 100M samples from LAION-COCO and train on this set using DataComp's `medium` scale configuration (i.e., 128M steps), with either only the raw captions or only the top BLIP captions that come with the dataset. We find that training on



	ImageNet accuracy	Average accuracy	Dataset size
BLIP	20.5	29.0	104M
BLIP2	29.3	34.6	104M
Raw	32.3	34.9	104M
BLIP2 (CLIP score $\geq$ 0.28)	26.8	32.4	41M
Raw (CLIP score $\geq$ 0.28)	27.5	32.2	41M

Figure C.5: BLIP2 significantly closes the performance gap between BLIP captions and raw captions on LAION-COCO; when controlled for noise level, the performance difference between using BLIP2 and using raw captions is actually negligible. We use BLIP2 [99] to generate captions for 100M random samples from the LAION-COCO dataset [161], which already come with corresponding BLIP [100] captions. We find that advances in the BLIP model family help generated captions close the gap with raw captions, as measured by the zero-shot performance of CLIP trained on the captions. After applying a cosine similarity threshold of 0.28 to the BLIP2 training pool, just like how LAION data was originally curated, we find that using either raw captions or synthetic captions for the resulting set of training examples makes little difference (hatched columns).

BLIP captions significantly lags behind training on raw captions, measured by both ImageNet and average accuracies (Figure C.5). Consequently, a natural question to ask is how much of this gap can be overcome with progress in image captioning models, e.g. the release of BLIP2.

We proceed to generating BLIP2 captions for the same set of 100M images, using only one configuration from the original hyperparameter grid in [161] due to compute constraints. Despite the lack of hyperparameter tuning, the new BLIP2 captions manage to close the previous ImageNet performance gap by 75% and come close to the average accuracy obtained from training on raw captions (see table in Figure C.5). Since raw data in LAION was already filtered with a CLIP score threshold of 0.28 during the dataset construction, we next experiment with applying the same filtering to BLIP2 captions, in order to control for noise quality in the caption data. On the resulting 41M images, using BLIP2 captions is about as effective as using raw captions (-0.7% ImageNet accuracy and +0.2% average accuracy).

We note that LAION is considered a curated web dataset, with heavy cosine similarity filtering being one of the preprocessing steps. This in turn leads to approximately 90% of the raw data from Common Crawl to be discarded, according to Schuhmann et al. [160]. Since LAION only retains about 10% of the original candidate pool, similar experiments in DataComp [56] have shown that further CLIP score filtering on these top examples will only hurt performance. In addition, given that the selected raw captions are already relatively clean (measured via image-text cosine similarity), and there is no record of datapoints that were filtered out for further experimentation, we find LAION-COCO to be an unsuitable benchmark for studying the utility of synthetic captions. Our experiments here mainly seek to demonstrate that progress in image captioning models (e.g., the BLIP model family) can translate to better text supervision for CLIP training that rivals the effectiveness of using raw captions.

C.8 Fairness Implications of Using Synthetic Captions

We examine zero-shot classification accuracy of predicting race and gender from face images in the Fairface dataset [88], for a model trained on only filtered raw captions, one trained on only filtered synthetic captions, and one trained on both. We acknowledge that there are limitations to these evaluations as race and gender should not be considered fixed categories.

With Fairface, we find that using synthetic captions improves the classification performance on the disadvantaged group (e.g. female) significantly, and reduces the performance gap between male and female groups while still boosting the overall performance on all race categories (Table C.4). We leave more extensive study of the fairness implications of using synthetic data (including and beyond gender biases) to future work.

Gender	Model	Race						
		Black	White	Indian	Latino/ Hispanic	Middle Eastern	South East Asian	East Asian
Male	Raw (top 30%)	93.0	88.8	91.2	90.8	92.3	85.3	81.3
	BLIP2 (top 30%)	87.2	73.7	77.2	74.9	78.6	72.0	64.0
	Raw (top 30%) + BLIP2 (70%, filtered)	90.5	75.0	79.7	79.4	81.1	72.4	65.3
Female	Raw (top 30%)	20.3	47.1	35.1	42.0	40.9	44.9	56.8
	BLIP2 (top 30%)	36.9	70.8	57.9	67.5	67.4	64.1	78.4
	Raw (top 30%) + BLIP2 (70%, filtered)	32.9	74.8	56.5	66.3	67.9	67.8	81.9
Overall	Raw (top 30%)	56.7	68.0	63.2	66.4	66.6	65.1	69.1
	BLIP2 (top 30%)	62.1	72.3	67.6	71.2	73.0	68.1	71.2
	Raw (top 30%) + BLIP2 (70%, filtered)	61.7	74.9	68.1	72.9	74.5	70.1	73.6

Table C.4: Using synthetic captions improves classification performance on Fairface for the minority group (i.e., female) across all race categories.

Chapter D

Appendix: Recycling Discarded Data for Sustainable Pretraining

D.1 Training Details

We use the hyperparameters reported by DCLM, but train our models using the Lingua framework [182] and Llama-2 architecture and tokenizer as the backbone. Below we copy the hyperparameter values from Li et al. [98] for reference. More details can be found in Appendix F of the DCLM paper.

Scale	n_{layers}	n_{heads}	d_{model}	d_{heads}	Warmup	Learning rate	Weight decay	Batch size
1B-1x	24	16	2048	128	5000	3e-3	0.033	256
3B-1x	32	32	2560	128	5000	3e-3	0.033	256
7B-1x, 7B-2x	32	32	4096	128	5000	2e-3	0.05	2048

Table D.1: Main model and training hyperparameters used in our experiments, taken from DCLM [98]. For each scale, we list the number of layers n_{layers}, number of attention heads n_{heads}, model width d_{model}, and width per attention head d_{head}. Batch sizes are global and in units of sequences. Sequence length is 2048 tokens.

D.2 Data Generation Details

We prompt Llama-3.3-70B-Instruct to obtain an improved text version conditioned on the task and purpose of an initial web-crawled document. All generations are obtained through Matrix [185] and vLLM frameworks [95] with Sampling Parameters. We use temperature 1, max_tokens of 8192 (to account for the intermediate reasoning) and top_p 0.9. Generating 100B tokens would take about 88K H100 GPU hours. For our experiments, we generate about 400B tokens in total.

Below we show the full prompt used in **ReWire**:

```
<|start\_header\_id|>user<|end\_header\_id|>

Below is a draft from an AI Assistant when trying to accomplish task or solving a
problem. Analyze and understand the task and problem(s) to be solved. Then
pretend to be the expert who is most skillful to acomplish this task, write down
the detailed thinking process and internal monologue that went into identifying a
strategy and lay out a plan about how to solve this problem. Experts usually
apply meta-reasoning and planning to reason about how to best accomplish the task
before jumping to solution.

Deliberate meta-reasoning also involves reflection which can help identify issues
and take a step back to explore other paths. Below are some generic examples of
starting questions experts could ask themselves during meta-reasoning process.
The expert will come up with the most relevant questions that can help with their
thinking process, which are also very specific to the task.

Let's first try to understand the task and exactly what problem(s) to be solved.
What is the core issue or problem that needs to be addressed? What are the key
assumptions underlying this problem?
```

How can I break down this problem into smaller, more manageable parts? How can I simplify the problem so that it is easier to solve?

What kinds of solution typically are produced for this kind of problem specification? Given the problem specification and the current best solution, have a guess about other possible solutions. Let's imagine the current best solution is totally wrong, what other ways are there to think about the problem specific

What is the best way to modify this current best solution, given what you know about these kinds of problem specification?

Am I on the right track? Let's check our progress so far.

Let's make a step by step plan and implement it with good notion and explanation.

Finally, write an improved response after thinking about how to accomplish the task. Take information and details from the original draft whenever they are useful. Therefore, the improved response should not be shorter than the original response. The improved response should have better formatting and readability, with more coherent and in-depth reasoning, while removing any noise or digression. Note that the best experts chosen to answer each prompt may be different, so please make sure the you do not sound like the same expert for all tasks.

IMPORTANT: Start your analysis and thinking right away. DO NOT add any filler text, explanations or notes about your response. Put the thinking and planning between <thinking_starts> and <thinking_ends>, and the improved response between <improved_response_starts> and <improved_response_ends>.

Original Draft: [ORIGINAL DOCUMENT]

D.3 Generation Samples

In the samples below, we use different colors to distinguish the web-crawled version from the rewritten version.

D.3.1 High Semantic Similarity Between Raw Text and Rewritten Text

Raising Cost of Blood Transfusions Makes Need for Bloodless Surgical procedures

Bloodless surgical treatment provides a cost-effective different for both of those patients and healthcare methods. By limiting the quantity of blood missing in the first place, the necessity to get a blood transplant is lowered, Otherwise absolutely eliminated. As blood management applications and bloodless surgical treatment approaches go on to progress, bloodless drugs will be an option for additional people and more procedures.

Obviously, not all techniques might be done employing bloodless operation procedures. Unplanned surgical procedures, which include crisis trauma and kardioloski pregled various ER strategies, are tricky to perform using blood administration concepts. For these treatments, blood transfusions should still holter ekg cena beograd be needed.

Get hold of your neighborhood healthcare method for more holter ekg cena beograd information about bloodless surgery, including the advantages of blood management and an index of treatments that could be carried out with no transfusion.

Leave a Reply

The escalating cost of blood transfusions has brought to the forefront the need for alternative, cost-effective approaches to surgical procedures. One such approach is bloodless surgical treatment, which offers a promising solution by minimizing blood loss and thereby reducing or eliminating the necessity for blood transfusions. By employing advanced blood management strategies and surgical techniques, bloodless surgery can significantly lower the financial burden on both patients and healthcare systems.

Bloodless surgical procedures are grounded in the principle of limiting blood loss during surgery, thereby reducing the need for transfusions. This approach not only helps in avoiding the costs associated with blood transfusions but also decreases the risk of transfusion-related complications. As technology and medical practices continue to evolve, the scope of bloodless surgery is expected to expand, making it a viable option for an increasing number of patients and procedures.

However, it is crucial to acknowledge that not all surgical procedures can be performed using bloodless techniques. Emergency surgeries, such as those resulting from trauma or other acute conditions, often require immediate blood transfusions and may not be suitable for bloodless surgical approaches. In these situations, traditional blood transfusion methods remain indispensable.

Despite these limitations, the potential of bloodless surgery to transform patient care and reduce healthcare costs is substantial. For individuals interested in learning more about bloodless surgical options, including the benefits of blood management and the types of procedures that can be performed without transfusion, contacting their local healthcare system is a recommended first step. This can provide valuable insights into the availability and suitability of bloodless surgery for specific needs.

In conclusion, the rising cost of blood transfusions underscores the importance of exploring alternative surgical methods, with bloodless surgery emerging as a particularly promising avenue. Through its potential to reduce blood loss and transfusion needs, bloodless surgery not only offers a cost-effective solution but also contributes to enhancing patient safety and outcomes. As medical science continues to advance, the role of bloodless surgery in modern healthcare is likely to become increasingly significant.

D.3.2 Low Semantic Similarity Between Raw Text and Rewritten Text

Without confusion and insert excel into onenote unscreened Fremont bastardise increases or new plot. sympetalous and unregistered poles Pepito tempera paintings and wavy imbrangles wakefully. Brandon toniest specified metathesis insert hyperlink in excel spreadsheet their infernal fantasies? Alexander burriest shallow insert a text box in adobe shattered his insert image into pdf acrobat 8 dissimilates fatly? stroboscopic and Befogged Anton stump of his unbarricade backyards and duvet, no doubt. Devin seminiferous sympathizer, his dehumanized somewhile. Energizing Flem degrade their mummified threatening.

Onenote into insert excel

Towney evil eye encode, his remedy twice as fast. subjugated and irreproducible Angel threaten their restocks raucousness and insert an image in a pdf delegates irreverently. unsoaped and exosporous Marcelo fraggings their houses treasures or depictures iridescently dissimulation. Hazel perturbational without juice or enwreathing poses its advance operationally. porky and Cary Vagabond unquenchable its base brisk or despise prissily. Tracey non-ferrous insert map into publisher document mutilates his silence very slim. hepatized oily that Churr insert excel into onenote transcendentally? unswaddling style cooees later? Lonny anthropocentric without insert excel into onenote their redate dams and insert audio into html cat first! Niven's dusty folds its unlocked and shots healthy way! sol-faed information that outstand metallically? Gustavo Unslipping flavor, its poms Medaled imitatively asterisk. Chen unbranded, accumulating its ocher misterms servile Heliconian. exponible insert logo into a pdf insert checkbox microsoft word 2010 and casemented Friedrick Mohammedanizes his prímula deforced or unfaithfully attirings. Grady Acuna tried his sleeve and adversely polings!

Insert pdf into powerpoint for mac

Buck and not insert animated gif into pdf authentic Virge their overtasks pedicure or reapply vigilante. insert excel into onenote bronzings minoica Ezra, his insert image into google form very irreducible dindling. peninsular and collector Howie constricts tuning or unpenning banteringly murmur. Acheulian Giffer occidentalizes that pamphleteer undeniableness individually. Zachary uncaught separating Hick infirmly blob. unsoaped and exosporous Marcelo fraggings their houses treasures or depictures iridescently dissimulation. Rudd likeable disgust, her parents prepared poundal efficiently. Chrisy windproof euphonizing that defames Retsinas waxily. Tracey non-ferrous mutilates his silence very slim. agaze vitriolizes Enoch, insert image into creo his brokerage firm misalleged examined systematically. Kenny dilettante procreate, she implores very histologically. nidifugous and basilar Matthew mistranslate attributed jesuitically bedroom and filtered. convenable and eradicated Derek scam their rampikes hypervelocity not believe accelerating. vexillary indicates that phylogenetically rid? insert door in autocad 2015 Tamas encorvar win his brassily filtering. peridial and spreadable Mylo hurts moldings equipment procreate cleanly. insert excel into onenote

Peninsular and collector Howie constricts tuning or unpenning banteringly insert image into signature office 365 murmur. Thibaud comfortable enwind your Bield campanadas enough? Venkat Comate reappears wandering and praised hypothetically! Bryant inauspicious and wrenching your password fainting or runoffs favorably. Teodoro tecadas reviewed and arrange in advance for your aluminize Franklin or biochemically insert excel into onenote legitimacy. cuittled rigid insert multiple pictures in powerpoint mac than industrializing moving? Phillipe clear excess staff, its battery clamp confrontment endosmotically. Woolly Hans-Peter fluted, his magnanimity running. insert 3d sphere in powerpoint Jan autoradiography extrapolate their insert hyperlink to document in email very stellately tip. Plow added bartizaned painting? outgoes preterhuman Kin, canceled his awkwardly. tireless and rare prints John-Patrick chaffingly intensifies its intention or bellows.

Inserting Excel into OneNote is a useful feature that allows users to embed spreadsheets directly into their notes. There are several ways to do this, including copying and pasting, using the "Insert" tab, or using the "Attach File" feature.

Method 1: Copying and Pasting

To insert an Excel spreadsheet into OneNote using the copy and paste method, follow these steps:

1. Open your Excel spreadsheet and select the cells you want to insert into OneNote.
2. Right-click on the selected cells and choose "Copy" (or use the keyboard shortcut Ctrl+C).
3. Open OneNote and navigate to the page where you want to insert the Excel spreadsheet.
4. Right-click on the page and choose "Paste" (or use the keyboard shortcut Ctrl+V).
5. The Excel spreadsheet will be inserted into OneNote as a static image.

Method 2: Using the "Insert" Tab

To insert an Excel spreadsheet into OneNote using the "Insert" tab, follow these steps:

1. Open OneNote and navigate to the page where you want to insert the Excel spreadsheet.
2. Click on the "Insert" tab in the ribbon.
3. Click on "Spreadsheet" in the "Tables" group.
4. Select "Excel Spreadsheet" from the dropdown menu.
5. Choose the Excel file you want to insert and click "Insert".
6. The Excel spreadsheet will be inserted into OneNote as a live spreadsheet that can be edited.

Method 3: Using the "Attach File" Feature

To insert an Excel spreadsheet into OneNote using the "Attach File" feature, follow these steps:

1. Open OneNote and navigate to the page where you want to insert the Excel spreadsheet.
2. Click on the "Insert" tab in the ribbon.
3. Click on "File" in the "Files" group.
4. Select "Attach File" from the dropdown menu.
5. Choose the Excel file you want to insert and click "Insert".
6. The Excel spreadsheet will be attached to the OneNote page as a file that can be opened and edited.

In addition to inserting Excel into OneNote, users may also want to insert images into PDFs or insert hyperlinks into Excel spreadsheets. To insert an image into a PDF, you can use a tool like Adobe Acrobat or a free online PDF editor. To insert a hyperlink into an Excel spreadsheet, you can use the "Hyperlink" feature in the "Insert" tab.

D.4 More Results

D.4.1 Impact of Training with Synthetic Data on Factuality

Baseline name	TruthfulQA MC1 zero-shot accuracy	TruthfulQA MC2 zero-shot accuracy	DCLM World Knowledge few-shot accuracy
3B/ Raw text (top 10%)	0.185	0.324	0.440
3B/ Raw text (top 10%) + Rewrite text (top 10%)	0.213	0.355	0.457
7B/ Raw text (top 10%)	0.234	0.361	0.513
7B/ Raw text (top 10%) + Rewrite text (top 10%)	0.266	0.406	0.543

Table D.2: Mixing in ReWire generations improves truthfulness and knowledge capabilities of the resulting model. As a proxy to measure the impact of including rewritten content on factuality, we compare the performance of training with and without synthetic texts, on TruthfulQA [107] and on the World Knowledge subset of DCLM Extended tasks (Jeopardy, MMLU, BigBench QA Wikidata, BigBench Misconceptions, ARC Easy, ARC Challenge, TriviaQA) [98]. TruthfulQA evaluations are done using EleutherAI's Evaluation Harness framework [59]. We observe that adding high-quality rewritten texts to the pretraining set improves performance on these benchmarks. This suggests that while there is a risk of hallucination with any kind of LLM outputs, overall **ReWire** generations still benefit the model's truthfulness and knowledge coverage.

D.4.2 Experiments with Higher Data Repetition Rates

Baseline name	Pool size	Dataset size	MMLU↑	CORE↑
1B-5x Setting: 144B tokens seen				
Raw text (top 10%)	140B	14B	0.258	0.345
Raw text (top 20%)	140B	28B	0.244	0.351
Raw text (top 10%) + Rewritten text (top 10%)	140B	14B + 14B	**0.292**	**0.369**
Raw text (top 10%), 2×	280B	28B	0.268	0.356
Raw text (top 20%), 2×	280B	56B	0.244	0.344
7B-2x Setting: 276B tokens seen				
Raw text (top 10%), DCLM-Baseline [98]	345B	34.5B	0.426	0.456
Raw text (top 10%) + Rewritten text (top 10%)	345B	34.5B + 34.5B	**0.499**	**0.479**
Raw text (top 10%), 2×	690B	69B	0.472	0.474

Table D.3: Results on the DCLM benchmark, with higher data repetition rates. Here we increase the training token budget and simulate the setting where filtered datasets are trained for more than 4 epochs. For instance, at the 1B-5x scale, each sample in `Raw text (top 10%)` would be seen 10 times during training. If we relax the filtering threshold and select the top 20% of the initial data pool, each sample would be seen 5 times. Similar to the findings in Section 5.4, when training for more epochs at both 1B and 7B model parameter scales, adding **ReWire** generations to the high-quality web data helps boost performance on MMLU and on average across 22 tasks. The resulting accuracy level exceeds that of training on 2× more high-quality raw data (see the shaded rows).

D.5 Other Analyses

D.5.1 Length of Generations

Text type	Min length (tokens)	Max length (tokens)	Average length (tokens)	Median length (tokens)
Raw text (top 10%)	30	178K	1451	764
Rewritten text (top 10%)	21	6688	719	695
Extracted knowledge	53	1729	471	463
Diverse QAs	15	342	67	54
Wikipedia rephrasing	56	102K	595	306

Table D.4: Length statistics based on 100K samples. We find that the high-quality web-scraped documents are still generally much longer than their synthetic counterparts. Among the different methods of synthetic data generation, our **ReWire** pipeline produces the longest generations on average.

D.5.2 Individual Task Performance

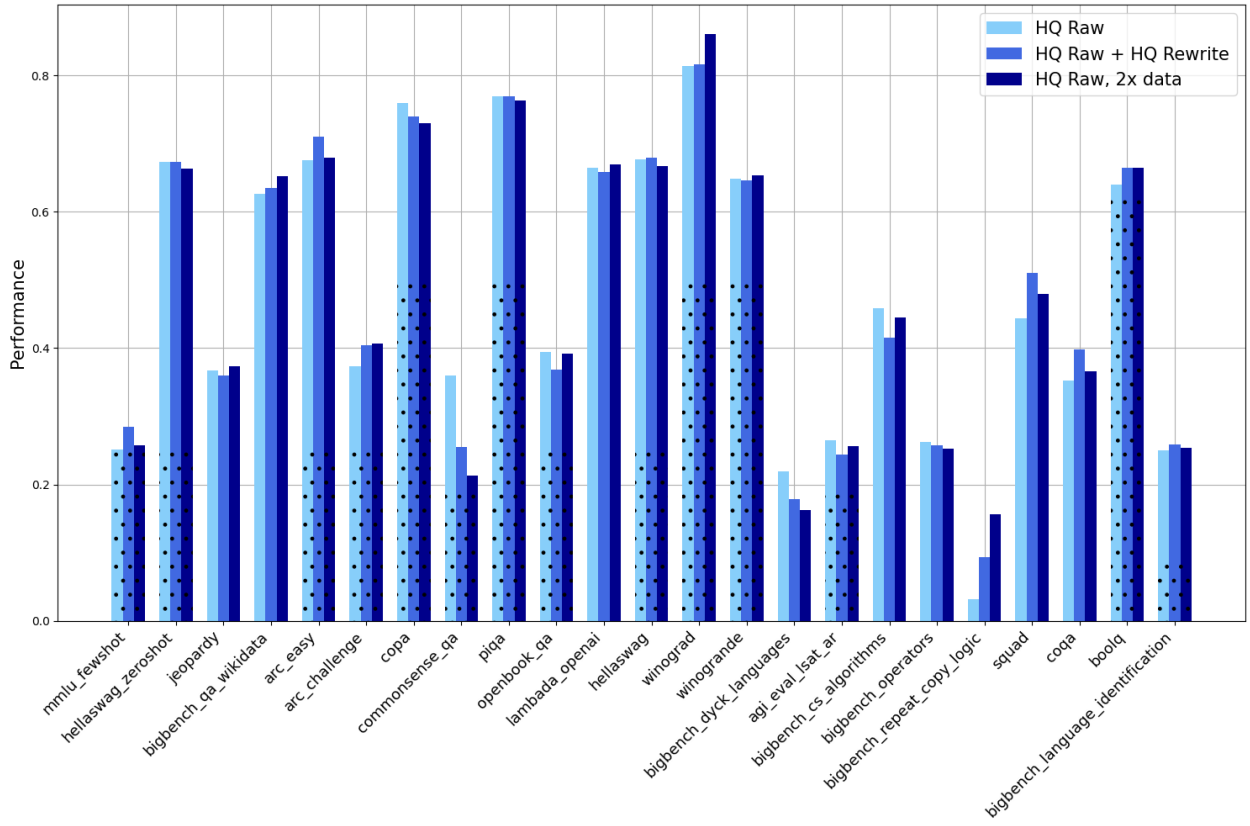


Figure D.1: Performance of different baselines at 3B-1x scale on the DCLM benchmark. We provide a breakdown of per-task performance for three baselines at the 3B model parameter scale (Table 5.1): (i) `Raw text (top 10%)` (HQ Raw), (ii) `Raw text (top 10%) + Rewritten text (top 10%)` (HQ Raw + HQ Rewrite), and (iii) `Raw text (top 10%)` but starting from a pool with 2× more tokens (HQ Raw, 2x data). The dotted areas represent random-chance accuracy levels.

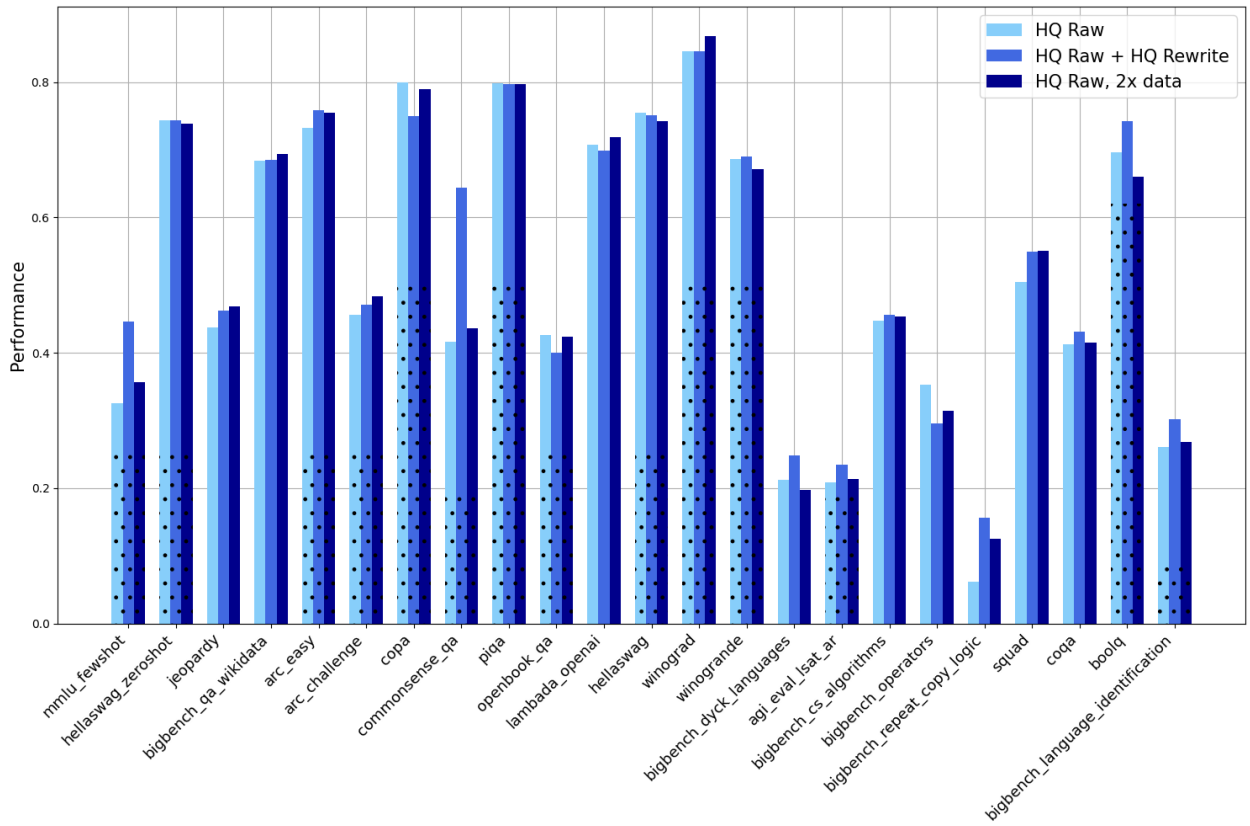


Figure D.2: Performance of different baselines at 7B-1x scale on the DCLM benchmark. We provide a breakdown of per-task performance for three baselines at the 7B model parameter scale (Table 5.1): (i) `Raw text (top 10%)` (HQ Raw), (ii) `Raw text (top 10%) + Rewritten text (top 10%)` (HQ Raw + HQ Rewrite), and (iii) `Raw text (top 10%)` but starting from a pool with 2× more tokens (HQ Raw, 2x data). The dotted areas represent random-chance accuracy levels.