Hidden Capabilities and Counterintuitive Limits in Large Language Models

Peter West

A dissertation

submitted in partial fulfillment of the

requirements for the degree of

Doctor of Philosophy

University of Washington

2024

Reading Committee:

Yejin Choi, Chair

Luke Zettlemoyer

Tim Althoff

Program Authorized to Offer Degree:

Paul G. Allen School of Computer Science & Engineering

University of Washington

Abstract

Hidden Capabilities and Counterintuitive Limits
in Large Language Models

Peter West

Chair of the Supervisory Committee:
Yejin Choi
Paul G. Allen School of Computer Science & Engineering

As massive language models like GPT-4 dominate NLP and AI, extreme-scale has become a clear and frequent theme for success. My research envisions a world where alternative approaches, efficient methods working on small to medium-scale models, work alongside extreme-scale models at the forefront of AI. In pursuit of this goal, the work described in this dissertation develops learning and inference algorithms that unlock hidden capabilities in compact language models. In parallel, I describe the underlying nature of model capabilities, and the limits that even scale-driven frontier models continue to suffer from. Concretely, this dissertation will explore three interconnected threads. First, Decoding-time Algorithms for Unlocking Out-of-the-box Capabilities. I have worked to develop a suite of inference-time algorithms that unlock capabilities in off-the-shelf, compact language models. Next, Symbolic Knowledge Distillation for Compact Expert Models. I study the way that useful knowledge can be extracted from general LMs, and incorporated into efficient expert models. Towards this goal, I introduce Symbolic Knowledge Distillation, a framework for distilling domain/task-specific knowledge from frontier LMs. Finally, Limits of LMs. I investigate the limits of LMs that even extreme scale has yet to overcome. Here, I pose the Generative AI Paradox: despite impressive generation capabilities, strong LMs and other generative models can exhibit much weaker understanding performance than we would expect from a human with the same ability to generate.

For Holly, who will never read this.

ACKNOWLEDGMENTS

There are many people without whom I would not be where I am–without whom I would be nowhere right now. The greatest privilege in my already very privileged life has been an endless network of people who I can call friends and family, the thought of whom brings a smile to my face. Below are the names of a few of those people most connected to this particular achievement, but it belongs to everyone on this list and beyond that held me on their shoulders. I give thanks to you.

First, to my family. My immediate family: Mom, Dad, Chris, Julian, Shannon (+ Martina and Leo), I have never seen another family like ours. To be born into a group of people that I trust and truly connect with, to have that from the moment I was born to the moment I'm writing this as an unquestioned home to go back to, gave me the foundation to do everything I've ever done. To my extended family, who I always felt the support of and with whom I have had ceaseless fun, including those in Edmonton, Squamish, Ireland, and scattered everywhere. To the Christiani/Holowaychuks, for precious times and the finest meat and produce which fueled my brain and soul the whole way through. To the Glovers, whose humor and warmth have been a piece of my life for as long as I can remember.

To Yejin, who accepted me, believed in me, and was open to me being myself as a researcher and an academic. There are so many countless things I could thank you for, but I will just say this: you taught me so much, showed me what I can be, and you made me believe in how exciting, fun, creative, and interactive scientific research is. Thank you for being patient with me, letting me grow, and giving me a chance 5 years ago.

To my letter writers: Luke, Allyson, Greg, and Yejin. Your kind words and guidance brought me to the dream that I achieved. Thank you for going up to bat for me.

To my committee: Luke, Aylin, Tim, Yejin. You made this final step of my degree comfortable, fun, and engaging. Thank you for giving me your time, thoughts, and many interesting questions.

To so many friends that I made during grad school. To those that I started together with: Nicasia, Willie, Brian, Ian, Naveena, Chris, Philip, Nick, Kay, Tal, Sam, and many others. Our cohort was something special and I will always treasure being a part of that. You made me feel like a person and a friend first, and a researcher after that. To my office mates from 260: Liang, Manaswi, Annie, Keunhong, Chungyi. You gave me a hilarious and supportive home to start everything from, and I cherish those countless hours I spent getting to know you. Many friends who invited me into their lives, including Charlie Blue, Mara, Mike, Margaret, Niloofar, Gus, Tony, Kentrell, Alane, Sarah, Galen, and so many more. Your friendship is a gift.

To those who gave me mentorship and nurtured me. Max (+ Julie), Julian, Antoine, Ari. You came with such humor and wisdom–you felt like older siblings who truly believed in me. To mentors in xlab and Mosaic: Ronan, Chandra, Jena, Alane, Sean, Faeze, Niloofar, Allyson, Vered, Jan. Even when you were shepherding me through hard questions and challenging times, you made me feel like an independent researcher and a colleague. To Chris and Michel at Microsoft, it was a true pleasure to know and work with you both. The environment you created made me question my decision not to go to industry more than anything else.

To all of the outstanding colleagues in the xlab that I saw and worked with (in no particular order): Max, Ximing, Niloofar, Alisa, Liwei, Faeze, Vered, Jack, Melanie, Jaehun, Sean, Raj, Alane, Abhilasha, Valentina, Hyunwoo, James, Jiacheng, Jillian, Taylor, Ben, Lindsey, Yuntian, Nouha, Yuchen, Lorraine, Youngjae, Daniel, Swabha, Rachel, Yonatan, Jan, Xiujun, Lianhui, Saadia, Maarten, Rowan, Antoine, Hannah, Eunsol, and everyone. You created an environment with none of the callousness, jadedness, competition, or mean spirit that I feared I would live through in grad school. Perhaps this was sheltering me, but all I saw was kindness, excitement, and genuine interest. This is something I once again thank Yejin for, in bringing all of us together and creating the space for us.

To many small and unique groups that I had the pleasure of being a part of. To my housemates at the Fjord, it seems to be a rare case to be excited to run into your roommates whenever you get home, but one that I lived. To my poetry group, we bared our souls and I will never forget what I'm doing here because of you all. To the biscotti club, I wish we had more time before everything stopped but my memories together are some of the most precious. To the TV club, which has been and remains what I look forward to every week. To the turbo team, in which I died laughing. To the SAC, who stared into the abyss with me. There is nothing like someone who understands those struggles you can't express.

To those friends who hear the small and sacred things on a long meandering walk: Nick, An, Marcus, Charlie Blue, Mara, Kenard, Ari. Know that I feel seen by you, and look forward to every chance I have to share a thought and a moment. My mind feels most at home in these talks.

To so many wonderful friends in Vancouver: Paul, Kaitlyn, Patrick, Kirsty, Annika, Raf, Peter, Jack, Kenard, Tony, Mohammad, Martin and many others. I not only survived but felt true joy through a complicated time because of you. You have made a new home for me in Vancouver long before I officially move back there. To Conrad who makes Vancouver my home even when he's in Montreal.

To Ari. You know what you did for me, I do not need to say it here in front of everyone. I'll just say, I already had a family and didn't need a new one, but you became my brother just the same.

Finally, to Holly. You know what you did even more than Ari. I love you.

CONTENTS

1 INTRODUCTION

1.1 LANGUAGE, MODELS, AND SCALE

Contemporary Large Language Models (LLMs) are driving an undeniable boom in AI popularity, both in terms of public knowledge and use (Chiang, 2023; Roose, 2024), as well as industrial development and application (AI@Meta, 2024; Jiang et al., 2023; OpenAI, 2023; Team et al., 2024). Yet these models are drawing on a long and rich history of language modeling, that was not always focused on general purpose AI assistants.

Formally, language models are defined as probabilistic estimators of textual probability; in other words, given some span of text t, a language model is something that attempts to estimate $P(t)$ or the true probability of t occurring in natural text. They are trained with a cross entropy objective (Jurafsky and Martin, 2009) which encourages these probabilities to be accurate, and can also be interpreted in terms of concepts such as optimal compression (Del'etang et al., 2023). Earlier development of language models was motivated from this probabilistic angle–creating an accurate estimator of textual probability, as a useful component in statistical systems for applications such as automatic translation (Brown et al., 1988) and speech recognition (Jelinek, 1997). A more powerful language model results in a more effective system overall, naturally driving improvement.

What began as reasonable probabilistic estimators became extremely powerful models of language, capable of handling complex text and prompts, through a series of innovations. Language models evolved from being statistical, count-based models (Kneser and Ney, 1995) to basic feed-forward neural networks; then to more complicated recurrent neural networks (Hochreiter and Schmidhuber, 1997) and finally to the transformer architecture (Vaswani et al., 2017) which is ubiquitous in modern LLMs. Methods for training models improved as well, with learning objectives becoming more stable (Raffel et al., 2019) and incorporating human feedback (Ouyang et al., 2022). Countless computational and algorithmic improvement were necessary to drive all of these changes.

Yet, the most important factor that brought language models from statistical components to the general purpose AI models of today is widely accepted to be *scale*. Models became both better probability estimators, and general AI systems as they became larger (Kaplan et al., 2020). Following the intuition of Rich Sutton's *The Bitter Lesson* (Sutton, 2019), it seemed that making language models larger and larger was the best and most consistent path towards both better probability estimation, and human-level capabilities.

The goal of this thesis will be to investigate this approach, particularly the recent and frequent refrain that "scale is all you need" for effective AI. While the extreme-scale

models of today have inarguably demonstrated completely new capabilities for AI, I will explore whether the benefits of scale are so clear cut. Scale has many downsides: high cost, excessive energy use, and lack of accessibility due to the complexity of running extreme-scale models. Besides this, many mysteries remain both within the largest scale models, and much smaller ones. What capabilities are yet undiscovered in even very compact language models? As well, what is the true nature of the capabilities of the largest models? Are these truly human-like, or do they break down in interesting and unexpected ways? These questions will define the lines of research discussed here.

1.2 UNEXPECTED CAPABILITIES OF COMPACT MODELS

As stated above, compact models have the advantages of being efficient and much more accessible to the general public than extreme-scale models. Many can even work on a personal computer, and so the prospect of strongcore abilities in these models would have a multiplicative benefit in their broad usability. As part of this dissertation, I will discuss methods for unlocking such abilities.

One major theme in this study will be latent capabilities–those that exist within such models but are not immediately accessible, or may require some extra ingredient. A motivation in pushing the scale of models has been to make capabilities *trivially accessible*. Extreme-scale models can often simply be prompted with textual instructions for specific behaviors, or perhaps finetuned a small amount. Yet, this ignore what much greater abilities may be lurking within these models, but are not accessible with shallow approaches.

I first will explore the combination of language models with principles from *information theory*. As previously stated, language models can be thought of as estimators of probability, the core primitive of information theory. Under this interpretation, we can embed language models into information theoretic approaches. I will particularly describe **BottleSum** (West et al., 2019a), a method I introduced to operationalize the information bottleneck principle (Tishby, Pereira, and Bialek, 2001) for textual summarization. This approach will allow models that are orders of magnitude smaller than the state of the art to do complex summarization tasks, even without access to human-written data.

I will also explore the potential of *knowledge* to unlock new capabilities in compact models. Knowledge, and particularly *commonsense knowledge*, became much more accessible in extreme scale models. I will discuss **Symbolic Knowledge Distillation** (West et al., 2021a), a method I introduced to transfer this knowledge from extreme-scale models and into much more compact ones. This yields more efficient and deploy-able domain expert models, which can even surpass the quality of their extreme-scale teachers.

Besides information theory and knowledge, there are many possible ingredients that can unlock latent capabilities of compact models. While this is beyond the scope of this dissertation, it is a core area for my future research.

1.3 COUNTERINTUITIVE LIMITS AT EXTREME SCALE

The capabilities of extreme-scale models have proven to be very useful, and inspire or work within approaches I have developed. Yet, are these capabilities really consistent and human-like? The ease of use, and tendency for LLMs to handle common cases well can be very convincing of robust competencies, yet this is often not the case.

In this line of work, I study the ways that even the most extreme-scale models can break down, and diverge from human intuitions. It is most important to study these limits for the extreme-scale, at which most users are interacting with models and many of the most impressive capabilities are evident, yet these limits are broadly relevant across model scales and methods, and can necessarily inspire new methods and solutions.

I will go into depth into one such limit that I propose, **Generative AI Paradox**. This questions whether the impressive abilities of models to generate content belies a genuine understanding. In humans, we often assume understanding is a prerequisite of generation (to write an essay on World War II, one must have some understanding of the topic; to draw a horse, one must understand what a horse is). Yet, I will discuss how this is often not the case in generative AI models. Oftentimes, they can excel at generation even when they struggle to demonstrate basic understanding.

Proposing notions such as the *Generative AI Paradox* can push users to have more realistic expectations of AI systems. We should not expect that one ability in an LLM implies another–generation does not imply understanding, although it often does in humans. This can also help inspire future work to understand these models more deeply. Particularly, this may imply that models generate content in a *fundamentally different* way from how humans do this. Rather than building up from base understanding, perhaps models largely work down from memorized examples. This does not necessarily imply that this capability is *weak*–models can handle many new inputs with surprising ease and creativity. Rather, this implies that there is a *divergence* from humans that demands further study, and new proposals of underlying mechanism.

1.4 SCOPE OF THIS DISSERTATION

The previous sections can be distilled into asking "is scale all you need?" in the pursuit of effective language models and AI, from two angles. First, *necessity*–is extreme-scale a necessary component in achieving the competencies we are interested in for models, or can we give more compact models these abilities through other ingredients such as algorithms and knowledge? Second, *sufficiency*–is extreme-scale alone creating models capable of human-level tasks in the ways that we would hope or expect? If not, how do models differ from humans and what are the most salient and functional differences to explore? This dissertation will explore these two angles as (respectively) **Unexpected Capabilities** and **Counterintuitive Limit** of large language models.

The chapters are as follows:

UNEXPECTED CAPABILITIES

- Chapter 2 presents **BottleSum**, an inference-time algorithm demonstrating how even weak LMs can have impressive capabilities out-of-the-box with information theory.

  > This chapter was previously published as: Peter West, Ari Holtzman, Jan Buys, and Yejin Choi (2019b). "BottleSum: Unsupervised and Self-supervised Sentence Summarization using the Information Bottleneck Principle." In: *ArXiv* abs/1909.07405. URL: https://api.semanticscholar.org/CorpusID:202583464.

- Chapter 3 presents **Symbolic Knowledge Distillation**, a method for distilling the benefits of scale–specifically domain-specific knowledge–into smaller and seemingly weaker models, to make compact expert models that may surpass their teachers.

  > This chapter was previously published as: Peter West, Chandrasekhar Bhagavatula, Jack Hessel, Jena D. Hwang, Liwei Jiang, Ronan Le Bras, Ximing Lu, Sean Welleck, and Yejin Choi (2021b). "Symbolic Knowledge Distillation: from General Language Models to Commonsense Models." In: *North American Chapter of the Association for Computational Linguistics*. URL: https://api.semanticscholar.org/CorpusID:238857304.

COUNTERINTUITIVE LIMITS

- Chapter 4 presents **The Generative AI Paradox**, a hypothesis about the ways that AI models seem to violate human intuitions about capabilities and intelligence, wherein models can often generate well without demonstrating understanding.

  > This chapter was previously published as: Peter West, Ximing Lu, Nouha Dziri, Faeze Brahman, Linjie Li, Jena D. Hwang, Liwei Jiang, Jillian R. Fisher, Abhilasha Ravichander, Khyathi Raghavi Chandu, Benjamin Newman, Pang Wei Koh, Allyson Ettinger, and Yejin Choi (2023). "The Generative AI Paradox: "What It Can Create, It May Not Understand"." In: *ArXiv* abs/2311.00059. URL: https://api.semanticscholar.org/CorpusID:264832736.

Part I

UNEXPECTED CAPABILITIES

2

BOTTLESUM

2.1 INTRODUCTION

Recent approaches based on neural networks have brought significant advancements for both extractive and abstractive summarization (Nallapati et al., 2016; Rush, Chopra, and Weston, 2015). However, their success relies on large-scale parallel corpora of input text and output summaries for direct supervision. For example, there are ~280,000 training instances in the CNN/Daily Mail dataset (Hermann et al., 2015; Nallapati et al., 2016), and ~4,000,000 instances in the sentence summarization dataset of Rush, Chopra, and Weston (2015). Because it is too costly to have humans write gold summaries at this scale, existing large-scale datasets are based on naturally occurring pairs of summary-like text paired with source text, for instance using news titles or highlights as summaries for news-text. A major drawback to this approach is that these pairs must already exist in-domain, which is often not true.

The sample inefficiency of current neural approaches limits their impact across different tasks and domains, motivating the need for unsupervised or self-supervised alternatives (Artetxe et al., 2017; LeCun, 2018; Schmidhuber, 1990). Further, for summarization in particular, the current paradigm requiring millions of supervision examples is almost counter-intuitive; after all, humans don't need to see a million summaries to know how to summarize, or what information to include.

In this paper, we present *BottleSum*, consisting of a pair of novel approaches, $\textit{BottleSum}^{Ex}$ and $\textit{BottleSum}^{Self}$ for *unsupervised extractive* and *self-supervised abstractive* summarization, respectively. Core to our approach is the principle of the Information Bottleneck (Tishby, Pereira, and Bialek, 1999), producing a summary for information X optimized to predict some other relevant information Y. In particular, we map (conditional) language modeling objectives to the Information Bottleneck principle to guide the unsupervised model on what to keep and what to discard.

The key intuition of our bottleneck-based summarization is that a good sentence summary contains information related to the broader context while discarding less significant details. Figure 2.1 demonstrates this intuition. Given input sentence "Hong Kong, a bustling metropolis with a population over 7 million, ...", which is followed by the next sentence "The city returned to Chinese control in 1997", the information bottleneck would suggest that minute details such as the city's population being over 7 million are relatively less important to keep. In contrast, the continued discussion of the city's governance in the next sentence suggests its former British rule is important here.

This intuition contrasts with that of autoencoder-based approaches where the goal is to minimize the reconstruction loss of the input sentence when constructing the summary (Baziotis et al., 2019; Fevry and Phang, 2018; Miao and Blunsom, 2016;

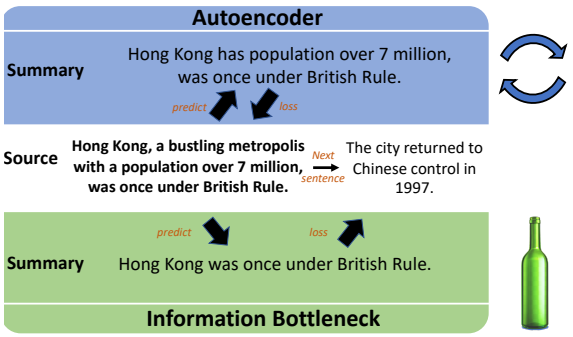


Figure 2.1: Example contrasting the Autoencoder (AE) and Information Bottleneck (IB) approaches to summarization. While AE (top) preserves any detail that helps to reconstruct the original, such as population size in this example, IB (bottom) uses context to determine which information is relevant, which results in a more appropriate summary.

Wang and Lee, 2018). Under the reconstruction loss, minute but specific details such as the city's population being over 7 million will be difficult to discard from the summary, because they are useful for reconstruction.

Concretely, $\mathcal{BottleSum}^{Ex}$ is an extractive and unsupervised sentence summarization method using the next sentence, a sample of nearby context, as guidance to *relevance*, or what information to keep. We capture this with a conditional language modelling objective, allowing us to benefit from powerful deep neural language models that are pre-trained over an extremely large-scale corpus. Under the Information Bottleneck objective, we present an iterative algorithm that searches gradually shorter subsequences of the source sentence while maximizing the probability of the next sentence conditioned on the summary. The benefit of this approach is that it requires no domain-specific supervision or fine-turning.

Building on our unsupervised extractive summarization, we then present $\mathcal{BottleSum}^{Self}$, a new approach to self-supervised abstractive summarization. This method also uses a pretrained language model, but turns it into an abstractive summarizer by fine-tuning on the output summaries generated by $\mathcal{BottleSum}^{Ex}$ paired with their original input sentences. The goal is to generalize the summaries generated by an extractive method by training a language model on them, which can then produce abstractive summaries as its generation is not constrained to be extractive.

Together, $\mathcal{BottleSum}^{Ex}$ and $\mathcal{BottleSum}^{Self}$ are $\mathcal{BottleSum}$ methods for unsupervised sentence summarization. Empirical results demonstrate that $\mathcal{BottleSum}^{Ex}$ outperforms other unsupervised methods on multiple automatic metrics, closely followed by $\mathcal{BottleSum}^{Self}$. Furthermore, testing on a large unsupervised corpus, we find $\mathcal{BottleSum}^{Self}$ outperforms unsupervised baselines (including our own $\mathcal{BottleSum}^{Ex}$) on human evaluation along multiple attributes.

2.2 THE INFORMATION BOTTLENECK PRINCIPLE

Unsupervised summarization requires formulating an appropriate learning objective that can be optimized without supervision (example summaries). Recent work has

treated unsupervised summarization as an autoencoding problem with a reconstruction loss (Baziotis et al., 2019; Miao and Blunsom, 2016). The goal is then to produce a compressed summary from which the source sentence can be accurately predicted, i.e. to maximize:

$$\mathbb{E}_{p(\tilde{s}|s)} \log p(s|\tilde{s}), \tag{2.1}$$

where s is the source sentence, $\tilde{s}$ is the generated summary and $p(\tilde{s}|s)$ the learned summarization model. The exact form of this loss may be more elaborate depending on the system, for example including an auxiliary language modeling loss, but the main aim is to produce a summary from which the source can be reconstructed.

The intuitive limitation of this approach is that it will always prefer to retain all informative content from the source. This goes against the fundamental goal of summarization, which crucially needs to forget all but the "relevant" information. It should be detrimental to keep tangential information, as illustrated by the example in Figure 2.1. As a result, autoencoding systems need to introduce additional loss terms to augment the reconstruction loss (e.g. length penalty, or the topic loss of Baziotis et al. (2019)).

The premise of our work is that the Information Bottleneck (IB) principle (Tishby, Pereira, and Bialek, 1999) is a more natural fit for summarization. Unlike reconstruction loss, which requires augmentative terms to summarize, IB naturally incorporates a tradeoff between information selection and pruning. These approaches are compared directly in section 2.4.3.

At its core, IB is concerned with the problem of maximal compression while defining a formal notion of information relevance. This is introduced with an external variable Y. The key is that $\tilde{S}$, the summary of source S, contains only information useful for predicting Y. This can be posed formally as learning a conditional distribution $p(\tilde{S}|S)$ minimizing:

$$I(\tilde{S}; S) - \beta I(\tilde{S}; Y), \tag{2.2}$$

where I denotes mutual information between these variables.

A notion of information relevance comes from the second term, the *relevance term*: with a positive coefficient β, this is encouraging summaries $\tilde{S}$ to contain information shared with Y. The first term, or *pruning term*, ensures that irrelevant information is discarded. By minimizing the mutual information between summary $\tilde{S}$ and source S, any information about the source that is not credited by the *relevance term* is thrown away. The statistical structure of IB makes this compressive by forcing the summary to only contain information shared with the source.[1]

In sum, IB relies on 3 principles:

1. Encouraging relevant information with a *relevance term.*
2. Discouraging extra information with a *pruning term.*
3. Strictly summarizing the source.

To clarify the difference from a reconstructive loss, suppose there is irrelevant information in S (i.e. unrelated to relevance variable Y), call this Z. With the IB objective

1 In IB, this is a strict statistical relationship.

(eq 2.3), there is no benefit to keeping any information from Z, which strictly makes the first term worse (more mutual information between source and summary) and does not affect the second (Z is unrelated to Y). In contrast, because Z contains information about S, including it in $\tilde{S}$ could easily benefit the reconstructive loss (eq. 2.1) despite being irrelevant.

As a relevance variable we will use the sentence following the source in the document in which it occurs. This choice is motivated by linguistic cohesion, in which we expect more broadly relevant information to be common between consecutive sentences, while less relevant information and details are often not carried forward.

We use these principles to derive two methods for sentence summarization. Our first method (§2.3) enforces strict summarization through being extractive. Additionally, it does not require any training, so can be applied directly without the availability of domain-specific data. The second method (§2.4) generalizes IB-based summarization to *abstractive* summarization that can be trained on large unsupervised datasets, learning an explicit summarization function $p(\tilde{s}|s)$ over a distribution of inputs.

2.3 UNSUPERVISED EXTRACTIVE SUMMARIZATION

We now use the Information Bottleneck principle to propose *BottleSum*^{Ex}, an unsupervised extractive approach to sentence summarization. Our approach does not require any training; only a pretrained language model is required to satisfy the IB principles of (2.2), and the stronger the language model, the stronger our approach will be. In section 2.4.3 we demonstrate the effectiveness of this method using GPT-2, the pretrained language model of Radford et al. (2019). [2]

2.3.1 *IB for Extractive Summarization*

Here, we take advantage of the natural parallel between the Information Bottleneck and summarization developed in section 2.2. Working from the 3 IB principles stated there, we derive a set of actionable principles for a concrete sentence summarization method.

We approach the task of summarizing a single sentence s using the following sentence s_{next} as the relevance variable. The method will be a deterministic function mapping s to the summary $\tilde{s}$, so instead of learning a distribution over summaries, we take $p(\tilde{s}|s) = 1$ for the summary we arrive at. Our goal is then to optimize the IB equation (Eq 2.2) for a single example rather a distribution of inputs (as in the original IB method).

In this setting, to minimize equation 2.2 we can equivalently minimize:

$$-\log p(\tilde{s}) - \beta_1 p(s_{next}|\tilde{s}) p(\tilde{s}) \log p(s_{next}|\tilde{s}), \tag{2.3}$$

where coefficient $\beta_1 > 0$ controls the trade-off between keeping relevant information and pruning. Similar to eq 2.2, the first term encourages pruning, while the second

2 We use the originally released "small" 117M parameter version.

encourages information about the relevance variable, s_{next}. Both unique values in eq 2.3 ($p(\tilde{s})$ and $p(s_{next}|\tilde{s})$) can be estimated directly by a pretrained language model, a result of the summary being natural language as well as our choice of relevance variable. This will give us a direct path to enforcing IB principles 1 and 2 from section 2.2.

To interpret principle 3 for text, we consider what attributes are important to strict textual summarization. Simply, a strict textual summary should be shorter than the source, while agreeing semantically. The first condition is straightforward but the second is currently infeasible to ensure with automatic systems, and so we instead enforce extractive summarization to ensure the first and encourage the second.

Without a supervised validation set, there is no clear way to select a value for β_1 in Eq 2.3 and so no way to optimize this directly. Instead, we opt to ensure both terms improve as our method proceeds. Thus, we are not comparing the pruning and relevance terms directly (only ensuring mutual progress), and so we optimize simpler quantities monotonic in the two terms instead: $p(\tilde{s})$ for pruning and $p(y|\tilde{s})$ for relevance.

We perform extractive summarization by iteratively deleting words or phrases, starting with the original sentence. At each elimination step, we only consider candidate deletions which decrease the value of the pruning term, i.e., increase the language model score of the candidate summary. This ensures progress on the pruning term, and also enforces the notion that word deletion should reduce the information content of the summary. The relevance term is optimized through only expanding candidates that have the highest relevance scores at each iteration, and picking the candidate with the highest relevance score as final summary.

Altogether, this gives 3 principles for extractive summarization with IB.

1. Maximize *relevance term* by maximizing $p(s_{next}|\tilde{s})$.
2. Prune information and enforce compression by bounding: $p(\tilde{s}_{i+1}) > p(\tilde{s}_i)$.
3. Enforce strict summarization by extractive word elimination.

2.3.2 *Method*

We turn these principles into a concrete method which iteratively produces summaries of decreasing length by deleting consecutive words in candidate summaries (Algorithm 1). The relevance term is optimized in two ways: first, only the top-scoring summaries of each length are used to generate new, shorter summaries. Second, the final summary is chosen explicitly by this measure.

In order to satisfy the second condition, each candidate must contain less self-information (i.e., have higher probability) than the candidate that derives it. This ensures that each deletion (line 9) strictly removes information. The third condition, strict extractiveness, is satisfied per definition.

The algorithm has two parameters: m is the max number of consecutive words to delete when producing new summary candidates (line 9), and k is the number of candidates at each length used to generate shorter candidates by deletion (line 5).

Algorithm 1 *BottleSum*^{Ex} method

Require: sentence s and context s_{next}

$C \leftarrow \{s\}$ ▷ set of summary candidates
for l in $length(s)...1$ **do**
  $C_l \leftarrow \{s' \in C | len(s') = l\}$
  sort C_l descending by $p(s_{next}|s')$
  for s' in $C_l[1:k]$ **do**
    $l' \leftarrow length(s')$
    for j in $1...m$ **do**
      for i in $1...(l'-j)$ **do**
        $s'' \leftarrow s'[1:i-1] \circ s'[i+j:l']$
        if $p(s'') > p(s')$ **then**
          $C \leftarrow C + \{s''\}$
        end if
      end for
    end for
  end for
end for

2.4 ABSTRACTIVE SUMMARIZATION WITH EXTRACTIVE SELF-SUPERVISION

Next, we extend the unsupervised summarization of *BottleSum*^{Ex} to abstractive summarization with *BottleSum*^{Self}, based on a straightforward technique for self-supervision. Simply, a large corpus of unsupervised summaries is generated with *BottleSum*^{Ex} using a strong language model, then the same language model is tuned to produce summaries from source sentences on that dataset.

The conceptual goal of *BottleSum*^{Self} is to use *BottleSum*^{Ex} as a guide to learn the notion of information relevance as expressed through IB, but in a way that (a) removes the restriction of extractiveness, to produce more natural outputs and (b) learns an explicit compression function not requiring a next sentence for decoding.

2.4.1 *Extractive Dataset*

The first step of *BottleSum*^{Self} is to produce a large-scale dataset for self-supervision using the *BottleSum*^{Ex} method set out in §2.3.2. The only requirement for the input corpus is that next sentences need to be available.

In our experiments, we generate a corpus of 100,000 sentence-summary pairs with *BottleSum*^{Ex}, using the same parameter settings as in section 2.3. The resulting summaries have an average compression ratio (by character length) of approximately 0.55.

2.4.2 *Abstractive Fine-tuning*

The second step of *BottleSum*^{Self} is fine-tuning the language model on its extractive summary dataset. The tuning data is formed by concatenating source sentences with generated summaries, separated by a delimiter and followed by an end token. The

model (GPT-2) is fine-tuned with a simple language modeling objective over the full sequence.

As a delimiter, we use `TL;DR:`, following Radford et al. (2019) who found that this induces summarization behavior in GPT-2 even without tuning. We use a tuning procedure closely related to Radford et al. (2018), training for 10 epochs. We take the trained model weights that minimize loss on a held-out set of 7000 extractive summaries.

To generate from this model, we use a standard beam search decoder, keeping the top candidates at each iteration. Unless otherwise specified, assume we use a beam size of 5. We restrict produced summaries to be at least 5 tokens long, and no longer than the source sentence.

2.4.3 *Experiments*

We evaluate our *BottleSum* methods using both automatic metrics and human evaluation. We find our methods dominant over a range of baselines in both categories.

2.4.3.1 *Setup*

We evaluate our methods and baselines using automatic ROUGE metrics (1,2,L) on the DUC-2003 and DUC-2004 datasets (Over, Dang, and Harman, 2007), similar to the evaluation used by Baziotis et al. (2019). DUC-2003 and DUC-2004 consist of 624 and 500 sentence-summary pairs respectively. Sentences are taken from newstext, and each summary consists of 4 human-written reference summaries capped at 75 bytes. We recover next-sentences from DUC articles for $\textit{BottleSum}^{Ex}$.

We also employ human evaluation as a point of comparison between models. This is both to combat known issues with ROUGE metrics (Schluter, 2017) and to experiment beyond limited supervised domains. Studying unsupervised methods allows for comparison over a much wider range of data where training summary pairs are not available, which we take advantage of here by summarizing sentences from the non-anonymized CNN corpus (Hermann et al., 2015; Nallapati et al., 2016; See, Liu, and Manning, 2017).

We use Amazon Mechanical Turk (AMT) for human evaluation, summarizing on 100 sentences sampled from a held out set. Evaluation between systems is primarily done as a pairwise comparison between *BottleSum* models and baselines, over 3 attributes: coherence, conciseness, and agreement with the input. AMT workers are then asked to make a final judgement of which summary has higher overall quality. Each comparison is done by 3 different workers. Results are aggregated across workers and examples.

2.4.3.2 *Models*

In both experiments, $\textit{BottleSum}^{Ex}$ is executed as described in section 2.3.2. In experiments on DUC datasets, next-sentences are recovered from original news sources, while we limit test sentences in the CNN dataset to those with an available next-sentence (this includes over 95% of sentences). We set parameter $k = 1$ (i.e. expand a single candidate

at each step) with up to $m = 3$ consecutive words deleted per expansion. GPT-2 (small) is used as the method's pretrained language model, with no task-specific tuning. To clarify, the only difference between how $\mathcal{BottleSum}^{Ex}$ runs on the datasets tested here is the input sentences; no data-specific learning is required.

As with $\mathcal{BottleSum}^{Ex}$, we use GPT-2 (small) as the base for $\mathcal{BottleSum}^{Self}$. To produce source-summary pairs for self supervision, we generate over 100,000 summaries using $\mathcal{BottleSum}^{Ex}$ with the parameters above, on both the Gigaword sentence dataset (for automatic evaluation) and CNN training set (for human evaluation). $\mathcal{BottleSum}^{Self}$ is tuned on the respective set for 10 epoch with a procedure similar to Radford et al. (2019). When generating summaries, $\mathcal{BottleSum}^{Self}$ uses beam-search with beam size of 5, and outputs constrained to be at least 5 tokens long.

We include a related model, Recon^{Ex} as a simple autoencoding baseline comparable in setup to $\mathcal{BottleSum}^{Ex}$. Recon^{Ex} follows the procedure of $\mathcal{BottleSum}^{Ex}$, but replaces the next-sentence with the source sentence. This aims to take advantage of the tendency of language models to semantically repeat in to substitute the Information Bottleneck objective in $\mathcal{BottleSum}^{Ex}$ with a reconstruction-inspired loss. While this is not a perfect autoencoder by any means, we include it to probe the role of the next-sentence in the success of $\mathcal{BottleSum}^{Ex}$, particularly compared to a reconstructive method. As Recon^{Ex} tends to have a best reconstructive loss by retaining the entire source as its summary, we constrain its length to be as close as possible to the $\mathcal{BottleSum}^{Ex}$ summary for the same sentence.

As an unsupervised neural baseline, we include SEQ^3 (Baziotis et al., 2019), which is trained with an autoencoding objective paired with a topic loss and language model prior loss. SEQ^3 had the highest comparable unsupervised results on the DUC datasets that we are aware of, which we cite directly. For human evaluation, we retrained the model with released code on the training portion of the CNN corpus.

We use the ABS model of Rush, Chopra, and Weston (2015) as a baseline for automatic and human evaluation. For automatic evaluation, this model is the best published supervised result we are aware of on the DUC-2003 dataset, and we include it as a point of reference for the gap between supervised and unsupervised performance. We cite their results directly. For human evaluation, this model demonstrates the performance gap for out-of-domain summarization. Specifically, it requires supervision (unavailable for the CNN dataset), and so we use the model as originally trained on the Gigaword sentence dataset. This constitutes a significant domain-shift from the first-sentences of articles with limited vocabulary to arbitrary article sentences with diverse vocabulary.

We include the result of Li et al. (2017) on DUC-2004, who achieved the best supervised performance we are aware of. This is intended as a point of reference for supervised performance.

Finally, for automatic metrics we include common baseline PREFIX, the first 75 bytes of the source sentence. To take into account lack of strict length constraints and possible bias of ROUGE towards longer sequences, we include INPUT, the full input sentence. Because our model is extractive, we know its outputs will be no longer than the input, but may exceed the length of other methods/baselines.

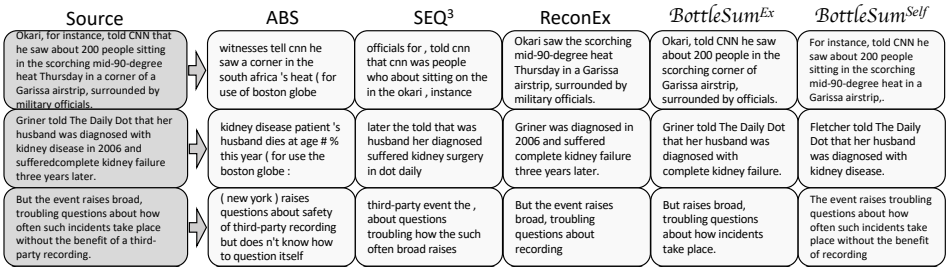


Figure 2.2: Representative example generations from the summarization systems compared

	DUC-2004			DUC-2003		
Method	R-1	R-2	R-L	R-1	R-2	R-L
Supervised						
ABS	28.18	8.49	23.81	28.48	8.91	23.97
Li et al. (2017)	31.79	10.75	27.48	-	-	-
Unsupervised						
PREFIX	20.91	5.52	18.20	21.14	**6.35**	18.74
INPUT	22.18	**6.30**	19.33	20.83	6.15	18.44
SEQ^3	22.13	6.18	19.3	20.90	6.08	18.55
Recon^{Ex}	21.97	5.70	18.81	21.11	5.77	18.33
BottleSum^{Ex}	**22.85**	5.71	**19.87**	**21.80**	5.63	**19.19**
BottleSum^{Self}	22.30	5.84	19.60	21.54	5.93	18.96

Table 2.1: Averaged ROUGE on the DUC-2004 and DUC-2003 dataset

2.4.3.3 *Results*

In automatic evaluation, we find *BottleSum*^{Ex} achieves the highest R-1 and R-L scores for unsupervised summarization on both datasets. This is promising in terms of the effectiveness of the Information Bottleneck (IB) as a framework. *BottleSum*^{Self} achieves the second highest scores in both of these categories, further suggesting that the tuning process used here is able to capture some of this benefit. The superiority of *BottleSum*^{Ex} suggests possible benefit to having access to a relevance variable (next-sentence) to the effectiveness of IB on these datasets.

The R-2 scores for *BottleSum*^{Ex} on both benchmark sets were lower than baselines, possibly due to a lack of fluency in the outputs of the extractive approach used. PREFIX and INPUT both copy human text directly and so should be highly fluent, while Rush, Chopra, and Weston (2015) and Baziotis et al. (2019) have the benefit of abstractive summarization, which is less restrictive in word order. Further, the fact that *BottleSum*^{Self} is abstractive and surpasses R-2 scores of both new extractive methods tested here (*BottleSum*^{Ex}, Recon^{Ex}) supports this idea. Recon^{Ex}, also extractive, has similar R-2 scores to *BottleSum*^{Ex}.

Models		**Attributes**			**Overall**		
Model	Comparison	cohere	concise	agreement	better	equal	worse
BottleSum^{Ex} vs.	ABS	+0.45	+0.48	+0.52	60%	31%	9%
	SEQ^3	+0.61	+0.57	+0.56	61%	34%	5%
	Recon^{Ex}	-0.05	+0.01	-0.05	37%	22%	41%
BottleSum^{Self} vs.	ABS	+0.47	+0.39	+0.48	62%	26%	12%
	SEQ^3	+0.56	+0.45	+0.53	65%	26%	9%
	Recon^{Ex}	+0.11	-0.05	+0.09	47%	14%	39%
	BottleSum^{Ex}	+0.14	+0.06	+0.11	43 %	27%	30%

Table 2.2: Human evaluation on 100 CNN test sentences (pairwise comparison of model outputs). Attribute scores are averaged over a scale of 1 (better), 0 (equal) and -1 (worse). We also report the overall preferences as percentages.

The performance of Recon^{Ex}, our simple reconstructive baseline, is mixed. It does succeed to some extent (e.g. surpassing R-1 for all other baselines but PREFIX on DUC-2003) but not as consistently as either *BottleSum* method. This suggests that while some benefit may come from the extractive process of *BottleSum*^{Ex} alone (which Recon^{Ex} shares), there is significant benefit to using a strong relevance variable (specifically in contrast to a reconstructive loss).

Next, we consider model results on human evaluation. *BottleSum*^{Self} and *BottleSum*^{Ex} both show reliably stronger performance compared to models from related work (ABS and SEQ^3 in Table 2.2). While *BottleSum*^{Self} seems superior to Recon^{Ex} other than in conciseness (in accordance with their compression ratios in Table 2.3), *BottleSum*^{Ex} appears roughly comparable to Recon^{Ex} and slightly inferior to *BottleSum*^{Self}.

The inversion of dominance between *BottleSum*^{Ex} and *BottleSum*^{Self} on automatic and human evaluation may cast light on competing advantages. *BottleSum*^{Ex} captures reference summaries more effectively, while *BottleSum*^{Self}, through a combination of abstractivness and learning a cohesive underlying mechanism of summarization, writes more favorable summaries for a human audience. Further analysis and accounting for known limitations of ROUGE metrics may clarify these competing advantages.

In comparing these models, there are also practical considerations ABS can be quite effective, but requires learning on a large supervised training set (as demonstrated by its poor out-of-domain performance in Table 2.2). While SEQ^3 is unsupervised, it still needs extensive training on a large corpus of in-domain text. *BottleSum*^{Ex}, whose outputs were preferred over both by humans, requires neither of these. Given a strong pretrained language model (GPT-2 small is used here) it only requires a source and next-sentence to summarize. *BottleSum*^{Self} requires in-domain text for self-supervision, but its superior performance by human evaluation and summarization without next-sentence are clear advantages. Further, its beam-search decoding is more computationally efficient than *BottleSum*^{Ex}, which requires evaluating conditional next-sentence perplexity over a large grid of extractive summary candidates.

Model	Abstractive Tokens %	Compression Ratio %
$\mathcal{BottleSum}^{Ex}$	-	51
Recon^{Ex}	-	52
$\mathcal{BottleSum}^{Self}$	5.8	56
SEQ^3	12.6	58
ABS	60.4	64

Table 2.3: Abstractiveness and compression of CNN summaries. Abstractiveness is omitted for strictly extractive approaches

Another difference from $\mathcal{BottleSum}^{Ex}$ is the ability of $\mathcal{BottleSum}^{Self}$ to be abstractivene (Table 2.3). Other baselines have a higher degree of abstractiveness than $\mathcal{BottleSum}^{Self}$, but this can be misleading. Consider the examples in figure 2.2. While many of the phrases introduced by other models are technically abstractive, they are often off-topic and confusing.

This hints at an advantage of $\mathcal{BottleSum}$ methods. In only requiring the base model to be a (tunable) language model, they are architecture-agnostic and can incorporate as powerful a language model as is available. Here, incorporating GPT-2 (small) carries benefits like strong pretrained weights and robust vocabulary handling by byte pair encoding, allowing them to process the diverse language of the non-anonymized CNN corpus with ease. The specific benefits of GPT-2 are less central, however; any such language model could be used for $\mathcal{BottleSum}^{Ex}$ immediately, and $\mathcal{BottleSum}^{Self}$ with some tuning. This is in contrast architecture-specific models like ABS and SEQ^3, which would require significant restructuring to fully incorporate a new model.

As a first work to study the Information Bottleneck principle for unsupervised summarization, our results suggest this is a promising direction for the field. It yielded two methods with unique performance benefits (Table 2.1, 2.2) and practical advantages. We believe this concept warrants further exploration in future work.

2.5 RELATED WORK

2.5.1 *Sentence Compression and Summarization*

Rush, Chopra, and Weston (2015) first proposed abstractive sentence compression with neural sequence to sequence models, trained on a large corpus of headlines with the first sentences of articles as supervision. This followed early work on approaching headline generation as statistical machine translation (Banko, Mittal, and Witbrock, 2000). Subsequently, recurrent neural networks with pointer-generator decoders became standard for this task, and focus shifted to the document-level (Nallapati et al., 2016; See, Liu, and Manning, 2017).

Pointer-based neural models have also been proposed for extractive summarization (Cheng and Lapata, 2016). The main limitations of this approach are that the training data is constructed heuristically, covering a specific type of sentence summarization

(headline generation). Thus, these supervised models do not generalize well to other kinds of sentence summarization or domains. In contrast, our method is applicable to any domain for which example inputs are available in context.

2.5.2 *Unsupervised Summarization*

Miao and Blunsom (2016) framed sentence compression as an autoencoder problem, where the compressed sentence is a latent variable from which the input sentence is reconstructed. They proposed extractive and pointer-generator models, regularizing the autoencoder with a language model to encourage compression and optimizing with the REINFORCE algorithm. While their extractive model does not require supervision, results are only reported for semi-supervised training, using less supervised data than purely supervised training. Fevry and Phang (2018) applied denoising autoencoders to fully unsupervised summarization, while Wang and Lee (2018) proposed an autoencoder with a discriminator for distinguising well-formed and ill-formed compressions in a Generative Adversarial Network (GAN) setting, instead of using a language model. However, their discriminator was trained using unpaired summaries, so while they beat purely unsupervised approaches like ours their results are not directly comparable. Recently Baziotis et al. (2019) proposed a differentiable autoencoder using a gumbel-softmax to represent the distribution over summaries. The model is trained with a straight-through estimator as an alternative to reinforcement learning, obtaining better results on unsupervised summarization. All of these approaches have in common autoencoder-based training, which we argue does not naturally capture information relevance for summarization.

Recently, Zhou and Rush (2019) introduced a promising method for summarization using contextual matching with pretrained language models. While contextual matching requires pretrained language models to generate contextual vectors, *BottleSum* methods do not have specific architectural constraints. Also, like Wang and Lee (2018) it trains with unpaired summaries and so is not directly comparable to us.

2.5.3 *Mutual Information for Unsupervised Learning*

We take inspiration from an exciting direction leveraging mutual information for unsupervised learning. Recent work in this area has seen success in natural language tasks (McAllester, 2018; Oord, Li, and Vinyals, 2018), as well as computer vision (Bachman, Hjelm, and Buchwalter, 2019; Hjelm et al., 2019) by finding novel ways to measure and optimize mutual information. Within this context, our work is a further example suggesting mutual information is an important element stimulating progress in unsupervised learning and modelling.

2.6 CONCLUSION

We have presented *BottleSum*^{Ex}, an unsupervised extractived approach to sentence summarization, and extended this to *BottleSum*^{Self}, a self-supervised abstractive approach. *BottleSum*^{Ex}, which can be applied without any training, achieves competitive performance on automatic and human evaluations, compared to unsupervised baselines. *BottleSum*^{Self}, trained on a new domain, obtains stronger performance by human evaluation than unsupervised baselines as well as *BottleSum*^{Ex}. Our results show that the Information Bottleneck principle, by encoding a more appropriate notion of relevance than autoencoders, offers a promising direction for progress on unsupervised summarization.

3

SYMBOLIC KNOWLEDGE DISTILLATION

3.1 INTRODUCTION

Prior works have suggested that pre-trained language models possess limited understanding of commonsense knowledge (Davis and Marcus, 2017; Merrill et al., 2021; Talmor et al., 2021) despite otherwise stellar performance on leaderboards. As a result, symbolic commonsense knowledge graphs (Hwang et al., 2021; Sap et al., 2019a; Speer, Chin, and Havasi, 2017) and corresponding neural representations (Bosselut et al., 2019; Hwang et al., 2021; Zhang et al., 2020b) have supplemented past models with commonsense capabilities. This has enabled diverse downstream applications, including interactive learning through a conversational interface (Arabshahi et al., 2021), persona- and affect-aware conversation models (Kearns et al., 2020), figurative language understanding (Chakrabarty et al., 2020; 2021), story telling (Ammanabrolu et al., 2021a) and fantasy games (b).

The common practice for commonsense knowledge graph construction sees humans spell out as many pieces of knowledge as possible. This pipeline goes *from–human–to–corpus–to–machine*, with commonsense models trained from human-authored knowledge graphs. Yet, high-quality, human-authored knowledge is expensive to scale, limiting coverage; this motivates an alternative: *from–machine–to–corpus–to–machine*. Prior efforts toward automatic commonsense knowledge graphs have resulted in considerably lower quality than human-written data (Hwang et al., 2021; Zhang et al., 2020b), which in turn leads to less reliable neural models (Hwang et al., 2021). Broad literature consistently shows machine-authored knowledge graphs underperform human-authored graphs (Bollacker et al., 2008; Etzioni et al., 2011; Mitchell et al., 2015).

In this work, we propose **Symbolic knowledge distillation**, a new conceptual framework towards high-quality automatic knowledge graphs for commonsense, leveraging state-of-the-art models and novel methodology. Most prior art for automatic knowledge graph construction extracts knowledge from raw text (Bhakthavatsalam, Anastasiades, and Clark, 2020; Li et al., 2020; Zhang et al., 2020a; b; Zhou et al., 2020). In contrast, our approach is motivated by knowledge distillation (Hinton, Vinyals, and Dean, 2015) wherein a larger teacher model transfers knowledge to a compact student model (§3.2.1). Our method differs from prior knowledge distillation in key ways: we distill a symbolic knowledge graph (i.e., generated text) in addition to a neural model, and we distill only a selective aspect of the teacher model. This selectively allows the student model to be of a different type (commonsense model), compared to the teacher (general language model), enriching the scope of distillation. An added benefit is that knowledge distilled as text is human readable: it can be understood and evaluated.

A general language model–GPT-3 in our case–is an imperfect commonsense teacher on its own, and the ability to evaluate distilled knowledge is useful in improving it. We

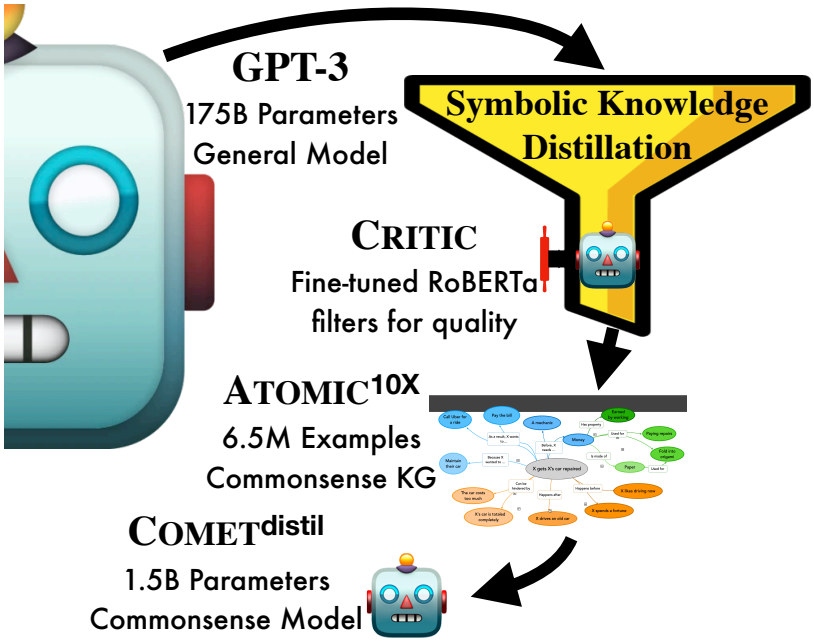


Figure 3.1: **Symbolic knowledge distillation** extracts the commonsense from the large, general language model GPT-3, into 2 forms: a large commonsense knowledge graph **ATOMIC^{10x}**, and a compact commonsense model **COMET$^{DIS}_{TIL}$**. The quality of this knowledge can be controlled and improved by adding a **critic** model, making GPT-3 a stronger teacher.

empirically demonstrate that, by training a separate critic model to judge symbolic generation quality, a more precise teacher can be defined. Knowledge from this critical teacher is higher quality–even exceeding human-authored knowledge. Yet even before training a critic, our study makes the unexpected finding that the student model surpasses the commonsense of GPT-3, our knowledge source.

To test symbolic knowledge distillation against the *human–to–corpus–to–machine* paradigm, we compare with ATOMIC$^{20}_{20}$ (Hwang et al., 2021), which is a human-authored commonsense knowledge graph. We find that **ATOMIC^{10x}**, our machine-generated corpus, exceeds the human generated corpus in *scale, accuracy,* and *diversity* with respect to 7 commonsense inference types that we focus on in this study. The resulting commonsense model, **COMET$^{DIS}_{TIL}$**, not only surpasses the human-trained equivalent COMET$^{20}_{20}$, but is also smaller, more efficient, and produces commonsense at a higher accuracy than its own teacher–GPT-3.

Symbolic knowledge distillation offers a promising new role for general language models, as commonsense knowledge sources, and humans, as small-scale evaluators to train critic models rather than authors of commonsense knowledge. Our work demonstrates that humans and LMs can be effective collaborators for curating commonsense knowledge graphs and training efficient and performant commonsense models.

X starts running	**xEffect** *so, X*	gets in shape	X sings a song	**HinderedBy** *but not if*	X can't remember the lyrics
X and Y engage in an argument	**xWant** *so, X wants*	to avoid Y	X is not well liked	**xReact** *so, X feels*	lonely
X learns to type fast	**xNeed** *X needed*	to have taken typing lessons	X takes care of a monkey	**xAttr** *X is seen as*	kind
X steals his grandfather's sword	**xEffect** *so, X*	is punished by his grandfather	X butts in	**HinderedBy** *but not if*	X is too shy to speak up
X takes up new employment	**xIntent** *because X wants*	to be self sufficient	X waits for the storm to break	**xEffect** *so, X*	is safe from the storm

Figure 3.2: Example **automatically generated** ATOMIC triples from our ATOMIC$^{\mathbf{10x}}$ commonsense knowledge graph. Each example includes a generated **event**, **relation** (with natural language interpretation), and generated **inference**.

3.2 OVERVIEW AND KEY FINDINGS

Throughout our work, we describe the *machine–to–corpus–to–machine* methodology of symbolic knowledge distillation. We first go *machine–to–corpus* (§3.3), by decoding from GPT-3, then improve our knowledge with a specialized critic model (§3.4), and finally distill this knowledge into an efficient commonsense model (§3.5), going *corpus–to–machine*. Throughout this process, we evaluate against a human knowledge source, comparing our automatic knowledge graph ATOMIC$^{\mathbf{10x}}$ and commonsense model COMET$^{\text{DIS}}_{\text{TIL}}$ to the human-authored ATOMIC$^{20}_{20}$ and resulting model COMET$^{20}_{20}$ ().

3.2.1 *Symbolic Knowledge Distillation*

Our proposed methodology parallels knowledge distillation (Hinton, Vinyals, and Dean, 2015), a method for compressing a large or complicated teacher distribution P_t into a smaller/simpler student distribution P_s. Key to knowledge distillation[1] is the notion of minimizing the cross-entropy between P_t and P_s:

$$H(P_t, P_s) = -\sum_{y \in Y} P_t(y) \log P_s(y) \tag{3.1}$$

Knowledge is transferred to the student by encouraging it to match teacher predictions. Hinton, Vinyals, and Dean (2015) apply this to conditional classification: for each training input, P_t and P_s are model predictions over label set Y. Typically Y is a tractable set, over which this sum can reasonably be calculated.

For distilling the knowledge of generative models, we can think of an unconditional language model (LM e.g. GPT-3) as P_t. This makes Y the set of all strings, over which LMs define probability. Unfortunately Y is an exponential set, intractable to sum over in Eq 3.1. Kim and Rush (2016) address this problem by simply taking the mode of P_t over Y, truncating most of the teacher distribution to the most likely sequence and discarding information.

Instead, we consider a sampling-based interpretation of the same objective:

$$H(P_t, P_s) = \mathop{\mathbb{E}}_{y \sim P_t(y)} [-\log P_s(y)] \tag{3.2}$$

1 In its simplest case, with temperature set to 1.0

which exactly equals the cross-entropy of Eq 3.1, at the limit under pure sampling from P_t.[2]

Yet distilling *all knowledge* from the teacher may not be desirable–our work is specifically focused on distlling commonsense knowledge from GPT-3. The ideal teacher P_t is a commonsense expert, but GPT-3 can approximate such a teacher, off-the-shelf, via prompting. This ability to select information is one explicit benefit of the sampling-based interpretation of Eq 3.2: while Eq 3.1 uses continuous logits over existing data, sampling gives discrete control over transferred information, by selecting which samples are elicited and used. For the general language model GPT-3, We encourage domain/quality with prompting, and sample truncation (Holtzman et al., 2020). We call this the *loose teacher* P_t^L–knowledge is generated and transferred from GPT-3, but without critical assessment of correctness (§3.3).

In fact, sampling knowledge in Eq 3.2 offers even more control, as generations can be individually interpreted and judged. Given an indicator function $A(x)$ for which knowledge x is *correct*, we can define a stronger teacher model. Using a Product of Experts (Hinton, 2002) between the loose teacher P_t^L and and the critic $A(x)$, we define a *critical teacher*:

$$P_t(x) \propto P_t^L(x|p) \cdot A(x) \tag{3.3}$$

In practice, $A(x)$ is a textual classifier learned on human judgements, 1 for knowledge predicted to be correct and 0 otherwise. Thus, the critic gives control over the correctness and confidence of the knowledge that is transferred (§3.4).

3.2.2 *Key Findings*

Applying symbolic knowledge distillation in practice results in promising and surprising findings:

1. LEARNING SYMBOLIC KNOWLEDGE FROM LANGUAGE MODELS CAN BE FRAMED AS A SYMBOLIC EXTENSION TO KNOWLEDGE DISTILLATION. In §3.2.1, we describe learning commonsense as a symbolic extension to knowledge distillation, with GPT-3 a knowledge source. We elaborate on this process with positive results in §3.3,3.4, and 3.5.

2. SYMBOLIC KNOWLEDGE DISTILLATION CONSTRUCTS A HIGH QUALITY KNOWLEDGE GRAPH AT SCALE. Our method naturally yields a machine-generated commonsense knowledge graph, which can achieve impressive quality (§3.4), beyond that of human-authored data. An effective critic which filters incorrect generated knowledge is key.

2 A useful consequence of this framing is that access to the full model distribution is not required. Our experiments (§3.3) use GPT-3, for which the distribution is **not available**, thus our method is applicable while knowledge distillation is not.

3. A CRITICAL TEACHER RESULTS IN A HIGHER QUALITY STUDENT. In §3.4, we show that making the teacher more critical results in higher quality knowledge, even as it reduces the scale of knowledge transferred. This demonstrates that *quality* matters, not just *quantity*, as higher quality knowledge results in a higher quality commonsense model in §3.5 despite smaller scale data.

4. CRITICAL TEACHER OR NOT, A STUDENT CAN OUTPERFORM THE KNOWLEDGE SOURCE. In §3.5, we show the unexpected result that all student models exceed the quality of GPT-3, the knowledge source.

5. MACHINES CAN WIN OVER HUMANS FOR AUTOMATIC KNOWLEDGE GRAPH CONSTRUCTION. In §3.4 and §3.5, we show that machine generated knowledge and the resulting commonsense model can outperform their equivalents that use a human knowledge source. Our symbolic knowledge exceeds humans at scale, quality, and diversity. The resulting commonsense model achieves the most accurate commonsense KG completions.

3.3 MACHINE-TO-CORPUS VERBALIZATION

Symbolic knowledge distillation begins by going *machine–to–corpus*, i.e. generating many commonsense facts, which results in a commonsense knowledge graph. §3.2.1 frames this as sampling to estimate the knowledge distillation objective–a student commonsense model learns from the generations of a teacher (GPT-3).

We start with a *loose teacher*, transferring knowledge by prompted generation with truncated sampling alone–this is in contrast to the *critical teacher* (§3.4) which explicitly judges and filters the generated samples. The loose teacher uses few-shot prompting as in Brown et al. (2020). We use a few-shot template:

<TASK-PROMPT>
<EX$_1$-INP><EX$_1$-OUT>
...
<EX$_{N-1}$-INP><EX$_{N-1}$-OUT>
<EX$_N$-INP>

where <EX$_i$-INP>/<EX$_i$-OUT> are human-authored, natural language ATOMIC entries, and <TASK-PROMPT> is a description of the problem. Given such a prompt, GPT-3 generates the *missing piece*, output <EX$_N$-OUT> for input <EX$_N$-INP>, following the pattern of earlier examples (1 to N-1). We find important aspects for producing high-quality commonsense knowledge:

- Examples should be numbered. e.g. <EX$_5$-INP> might begin with "5)" to indicate it is the 5th example.
- The format of <EX$_i$-INP> and <EX$_i$-OUT> should linguistically imply the relationship between them. See below for examples.
- <TASK-PROMPT> can be used to give extra specification to complicated problems.

3.3.1 *Data: ATOMIC*

We demonstrate symbolic knowledge distillation on the ATOMIC *if-then* resource (Sap et al., 2019a). This follows an event-relation-inference (triple) format. The corpus links *events* (e.g. *X attacks Y*) to relations, e.g. **HinderedBy** which describes what might hinder an event. For a relation/event, the goal is to generate a resulting inference, e.g. *X attacks Y* **HinderedBy** *X is restrained.*

Of the 23 relations from the most recent version–ATOMIC$^{20}_{20}$–we limit our investigation to 7 relations that correspond to *causal* commonsense knowledge: **xAttr** (how X is perceived after *event*), **xReact** (how X reacts in response to *event*), **xEffect** (what X does after *event*), **xIntent** (X's intent in *event*), **xWant** (what X wants after *event*), **xNeed** (what X needed for *event* to take place) and **HinderedBy**. We describe how **verbalization** is applied to ATOMIC data in 2 steps: generating underlying events (heads), then full examples (inference given event).

3.3.2 *Event Generation*

Events are context-free premises in ATOMIC involving `PersonX` (and sometimes a second `PersonY`) in various scenarios. These events form heads in knowledge graph triples. We generate events by filling in the elements of our template:

```
1. Event: X overcomes evil with good
2. Event: X does not learn from Y
...
10. Event: X looks at flowers
11.
```

The format is simple, as events are generated *unconditionally*. We use 100 high-quality events from the ATOMIC$^{20}_{20}$ corpus for our prompt, selected to avoid grammatical or logical errors, and minimize semantic overlap. We randomly sample 10 of these seed events for each generation batch, resulting in randomized prompts. We use nucleus sampling ($p = 0.9$) (Holtzman et al., 2020), and presence/frequency penalties of 0.5 from the GPT-3 interface. We generate 165K unique events using the 175B-parameter Davinci model[3] from Brown et al. (2020) (human-authored ATOMIC$^{20}_{20}$ contains only 6.2K events).

3.3.3 *Inference Generation*

Generating ATOMIC inferences requires reasoning about events and relations together. We design verbalization templates fo reach relation, with iterative design and small-scale verification by the authors e.g. we prompt the **xNeed** relation as follows:

```
What needs to be true for this event to take place?
...
```

3 the largest available version of GPT-3

```
Event <i>: X goes jogging Prerequisites: For this to happen, X
needed to wear running shoes

...

Event <N>: X looks at flowers Prerequisites: For this to happen,
```

The language of this template implies the relation-specific task, both "Prerequisites:" and beginning with "for this to happen" suggest the **xNeed** relation. As well, we include an xNeed-specific <TASK-PROMPT>. We use 10 few-shot examples for each prompt.[4]

For each event/relation (165K X 7) we generate 10 inferences with the Curie GPT-3 model[5] and earlier hyperparameters. Removing duplicate and degenerate (e.g. fewer than 3 characters) generations yields 6.46M ATOMIC-style data triples (examples in Figure 3.2). We call this ATOMIC$^{\mathbf{10x}}$, as it contains an order of magnitude more triples than ATOMIC$^{20}_{20}$ for the 7 relations we study.

3.3.4 *Evaluating a Generated Commonsense Knowledge Graph*

Machine generation enables a large scale of unique generations at a much lower cost than human-authored knowledge (Table 3.1), but what kind of examples are produced by GPT-3, and how does it differ from knowledge produced by humans? In this section, we conduct an in-depth analysis to answer these questions.

LEXICAL DIFFERENCES: DIVERSITY AND UNIQUENESS Recent work finds that machine generations can be repetitive and lack diversity (; Welleck et al., 2020); one way generated knowledge may differ from human-authored is less creative word choice, diversity, or more repetition.

To test this, we begin with lexical diversity (i.e. unique words used, Table 3.2). While there is variation by relation, the diveristy of ATOMIC$^{\mathbf{10x}}$ actually exceeds ATOMIC$^{20}_{20}$ here, 5.2M unique words to 1.5M. In addition, it contains significantly more strictly unique generated inferences (Table 3.2, unique tails).

BLEU SOFT UNIQUENESS. Exact match (above) fails to capture the notion of *similar* text. Following the intuition of self-BLEU (Zhu et al., 2018), we define *soft uniqueness* to describe diversity of generations in a corpus. An inference x is softly-unique if:

$$BLEU_2(C, x) < 0.5$$

where C is the set of inferences for a given input (in our case, event + relation), and 0.5 is an empirical threshold. To find soft-uniqueness of a corpus, we iteratively remove examples until all are softly unique, i.e. low mutual lexical overlap; higher diversity means more such examples (thus a larger softly unique corpus is preferable). Softly-unique corpus sizes are given in Table 3.4 ("Size (div)"). ATOMIC$^{\mathbf{10x}}$ has a smaller *fraction* of softly-unique examples than ATOMIC$^{20}_{20}$, yet it contains many more such examples.

4 We also replace anonymous names ("X") with sampled generic names as this improved quality.
5 for the largest, Davinci, 12M generations is computationally/monetarily intractable.

Relation	ATOMIC$^{20}_{20}$	ATOMIC$^{\mathbf{10x}}$
HinderedBy	77,616	**1,028,092**
xNeed	100,995	**760,232**
xWant	109,098	**730,223**
xIntent	54,839	**965,921**
xReact	62,424	**1,033,123**
xAttr	113,096	**884,318**
xEffect	90,868	**1,054,391**
Total Count	608,936	**6,456,300**
Est Total Cost	~$40,000	**~$6,000**
Est Cost Per Triple	~$0.06	**~$0.001**

Table 3.1: Number of unique triples with the given relation, $|(\cdot, \texttt{relation}, \cdot)|$. The estimated cost for ATOMIC$^{\mathbf{10x}}$ comes at a fraction of a conservative estimation for ATOMIC$^{20}_{20}$ crowdsourcing costs.

	Length		**Unique Tokens (K)**		**Unique Tails (K)**	
	$\mathbf{A}^{20}_{20}$	$\mathbf{A}^{\mathbf{10x}}$	$\mathbf{A}^{20}_{20}$	$\mathbf{A}^{\mathbf{10x}}$	$\mathbf{A}^{20}_{20}$	$\mathbf{A}^{\mathbf{10x}}$
xWant	4.69	5.16	322	784	69	152
xAttr	1.42	2.73	15	21	11	8
xEffect	3.92	4.66	216	864	55	185
xIntent	4.59	5.92	136	800	30	135
xNeed	4.51	5.97	289	1378	64	231
xReact	4.03	1.77	48	5	12	2
HinderedBy	7.93	7.49	522	1775	290	874
Events	5.20	**5.32**	109	**881**	6.2	**165**

Table 3.2: Average length, total unique tokens and total unique examples (in K, i.e. 1000s) by relation type and in events (bottom row) from ATOMIC$^{20}_{20}$ (A^{20}_{20}) and ATOMIC$^{\mathbf{10x}}$ ($\mathrm{A}^{\mathbf{10x}}$).

ATOMIC$^{\mathbf{10x}}$ contains 4.38M such examples (full size 6.5M) vs. ATOMIC$^{20}_{20}$, which has 560K (full size 600K).

MODEL-BASED DIVERSITY MEASUREMENT. Lexical notions of diversity reward differences in surface form, which may not always reflect diversity of *information*, only format. Thus, we next study information-theoretic measures for diversity. Intuitively, diverse information should be less predictable, or higher entropy. With GPT-2 XL models finetuned on ATOMIC$^{20}_{20}$ and ATOMIC$^{\mathbf{10x}}$ (§3.5) we estimate **entropy**–roughly, how difficult it is for a model to capture the corpus information (Table 3.3). This is 4 times higher for ATOMIC$^{\mathbf{10x}}$, suggesting more content from a modeling perspective. We

Entropy	Cross Entropy	KL Divergence
$H(D_1) = 1.27$	$H(D_1, D_2) = 9.31$	$D_{KL}(D1\|\|D2) = 8.04$
$H(D_2) = 7.80$	$H(D_2, D_1) = 41.48$	$D_{KL}(D_2\|\|D_1) = 33.68$

Table 3.3: Entropy, cross-entropy, and divergence of ATOMIC$^{20}_{20}$ (D_1) and ATOMIC$^{\mathbf{10x}}$ (D_2).

also estimate **cross-entropy**–how well a model trained on one corpus describes the other. From ATOMIC$^{\mathbf{10x}}$ to ATOMIC$^{20}_{20}$, this is 9.31, only 2 points higher than its entropy suggesting ATOMIC$^{20}_{20}$ is describable with information from ATOMIC$^{\mathbf{10x}}$. In reverse, this is 41.48 suggesting much of ATOMIC$^{\mathbf{10x}}$ is not captured by ATOMIC$^{20}_{20}$–ATOMIC$^{\mathbf{10x}}$ is surprising given only information from ATOMIC$^{20}_{20}$.

HUMAN EVALUATION OF QUALITY. Perhaps most importantly, we study the *quality* of knowledge in each corpus. We conduct human evaluation with Amazon Mechanical Turk. 3 annotators rate each triple resulting in "accepted", "rejected" or "no judgement". We evaluate 3000 examples[6] from ATOMIC$^{\mathbf{10x}}$, and 1000 from ATOMIC$^{20}_{20}$ (Table 3.4). We find Fleiss' kappa (Fleiss, 1971) of 40.8 indicating moderate agreement (Landis and Koch, 1977), and 90.5% accuracy agreement. We require workers meet an Amazon Mechanical Turk qualification for annotation quality based on past commonsense evaluations. We compensate workers $0.17 per task, which we estimate require 30 seconds.

For the *loose teacher*, consider the top row of ATOMIC$^{\mathbf{10x}}$ in Table 3.4 (other rows add the critic §3.4). ATOMIC$^{\mathbf{10x}}$ exceeds ATOMIC$^{20}_{20}$ in scale, but is somewhat less acceptable by human raters–by roughly 8 percentage points. Yet, the larger scale of ATOMIC$^{\mathbf{10x}}$ implies a significantly higher *number* of accurate examples. Increasing the proportion of these is the main objective of the critic (§3.4).

HOW DO KNOWLEDGE SOURCES COMPARE? To understand the robustness of our approach, we assess other language models as the knowledge source (i.e. loose teacher): GPT-J (Wang and Komatsuzaki, 2021) and T5-11B adapted for language modelling (Lester, Al-Rfou, and Constant, 2021). We substitute both for GPT-3 as in §3.3.2,3.3.3, generating a small-scale corpus to evaluate. We conduct human evaluation on 1000 examples as above (Table 3.4). Both models attain roughly 72% accuracy, 6 points below GPT-3 (78.5). This suggests strong potential, but higher quality from GPT-3.

3.4 MAKING THE TEACHER MORE CRITICAL

Symbolic knowledge distillation requires a strong teacher model to maximize the quality of the generated knowledge graph and resulting student model (§3.5). While the *loose teacher* (GPT-3 alone) results in a viable commonsense knowledge graph, evaluation shows this isn't a perfect commonsense teacher. Thus, we multiply in a

6 this ensures at least 1000 after filtering by the critic §3.4)

7 Size of ATOMIC$^{20}_{20}$ is given as the number of comparable datapoints, i.e. those with the same relations as ATOMIC$^{\mathbf{10x}}$.

Corpus	Accept	Reject	N/A	Size	Size (div)
ATOMIC$^{20}_{20}$	86.8	11.3	1.9	0.6M	0.56M
ATOMIC$^{\mathbf{10x}}$	78.5	18.7	2.8	**6.5M**	**4.38M**
	88.4	9.5	2.1	5.1M	3.68M
(critic$_{\text{low}}$)	91.5	6.8	1.7	4.4M	3.25M
	95.3	3.8	1.0	3.0M	2.33M
(critic$_{\text{high}}$)	**96.4**	2.7	0.8	2.5M	2.00M
+ GPT-J	72.0	27.6	0.4	-	-
+ T5-11B LM	71.7	26.9	1.4	-	-

Table 3.4: Attributes of ATOMIC$^{\mathbf{10x}}$ and ATOMIC$^{\mathbf{10x}}$ (row 2) including the critic model (§3.4, rows 3 - 6) with various filtering cutoffs. Accept and Reject are by majority human vote unless any mark N/A. Size is in unique examples[7]. The highest precision corpus is ATOMIC$^{\mathbf{10x}}$ with (critic$_{\text{high}}$), but multiple versions surpass ATOMIC$^{20}_{20}$. We also include alternate models (GPT-J and T5-11B) as the loose teacher.

critic model, to filter lower-quality knowledge, *correcting the teacher* (§3.2.1). With modest supervision (a small-scale human evaluation) we train a classifier to predict and discriminate unacceptable examples. We multiply this with the *loose teacher* §3.3, creating a *critical teacher* product of experts. In practice this means filtering ATOMIC$^{\mathbf{10x}}$ to create new corpora that are higher quality, yet still larger scale than human-authored ATOMIC$^{20}_{20}$.

TRAINING A KNOWLEDGE CRITIC We gather a training set of *correct vs. incorrect* human judgments on a randomly-sampled set of 10K entries of ATOMIC$^{\mathbf{10x}}$, as in §3.3.4 but with one annotation per example. We take a (random) train/dev/test split of 8k/1k/1k. While this step requires human annotation, humans take on the role of high-level supervisors here–critiquing a small number of generations rather than authoring the entire knowledge graph as in previous work. Indeed, the cost/complexity of this step is similar to a typical human evaluation, making it far cheaper/easier than eliciting human-authored knowledge in past work.

We train binary classifiers (critics) for human acceptability using RoBERTa-Large (Liu et al., 2019). We find pretraining on MNLI results in the best model in terms of precision and recall, and we suggest this technique for future studies. Our best model vastly improves the accuracy of ATOMIC$^{\mathbf{10x}}$ (Table 3.4), demonstrating that a small amount of human supervision can consistently help to correct GPT-3's mistakes.

SIZE-ACCURACY TRADE-OFF Using our critic to filter knowledge results in a natural trade-off between size and accuracy. We test several cutoffs for ATOMIC$^{\mathbf{10x}}$, i.e. confidence at which the critic rejects examples. We report human-measured accuracy (Accept/Reject column Table 3.4) following §3.3.4. We compare the loose teacher (unfiltered) to critical teachers. Discarding 20% of instances that the critic judges as least acceptable (reducing corpus size from 6.5M to 5.1M), ATOMIC$^{\mathbf{10x}}$'s accuracy rises 78.5

	Random	Inf	Event	EMAP	Full
AP	79.3	81.9	86.2	87.1	**94.0**

Table 3.5: Average Precision for ablated critic models. The critic not only filters *awkward phrasings* which can be identified by either the event (**Event**) or inference (**Inf**) in isolation (EMAP only identifies these), but also *logical misalignments*, which require modeling interactions between event/inference, i.e. the full critic (**Full**).

→ 88.4; human-authored ATOMIC$^{20}_{20}$ contains 600K entries at 86.8% accuracy. Reducing to total size to 2.5M examples (38% of full size), we attain 96.4% accuracy, nearly 10 points above ATOMIC$^{20}_{20}$ while still 4X larger.

WHAT GETS FILTERED OUT? We qualitatively identify two types of filtered triples: 1) *logical misalignments*, events/inferences joined in an inconsistent manner. Recognizing these requires understanding events-inference interactions, e.g., *X cannot find his shirt* ***as a result X*** *is wearing a shirt*; 2) *awkward phrasings*, in which events/inferences are individually incoherent e.g. *PersonX has a fire in the bath*–resulting triples are invalid as the event is implausible.

To understand what is filtered, we ablate the critic (Table 3.5): our full model is compared to a random predictor, event-only model, and inference-only model. We also compare to an EMAP (Hessel and Lee, 2020) version, i.e. an ensemble of event and inference-only, without interactions between event/inference (needed for *logical misalignments*).

We find GPT-3 produces both independent awkwardly-phrased events/inferences (filtered by X-only models) and logical misalignments. The classifier, trained on validated knowledge triples, helps in both cases. The EMAP of our full model (identifies only awkward phrasings) achieves 87% AP, and our full model (which additionally identifies logical misalignments) improves to 94% AP.

DOES FILTERING HURT DIVERSITY? One concern is that the critic may keep only similar "safe" examples, lacking novelty. We repeat our diversity analysis (§3.3.4) for critical corpora (Table 3.4, "Size (div)", higher=better). As we filter, we surprisingly observe proportionally *more* diverse examples: full ATOMIC$^{\mathbf{10x}}$ has a diverse subset 68% of its size; rising to 80% with the most extreme filtering. One possibility is that GPT-3 gravitates towards common sentence structures for inconsistent knowledge. These would be recognizable to the critic, and removing them would increase both quality and diversity. This surprising result warrants further study.

3.5 CORPUS-TO-MACHINE: DISTILLATION

The final step of symbolic knowledge distillation trains a compact model on the generated natural language knowledge graph. Our base model is GPT2-XL trained on all of ATOMIC$^{\mathbf{10x}}$: we denote this model by COMET$^{\text{DIS}}_{\text{TIL}}$. We additionally train the model on

CKG Completion Model	Train Corpus Acc	Accept	Reject	N/A
GPT2-XL zero-shot	-	45.1	50.3	4.6
GPT-3	-	73.3	24.1	2.6
COMET$^{20}_{20}$	86.8	81.5	16.3	2.2
COMET$^{DIS}_{TIL}$	78.5	78.4	19.2	2.4
+critic$_{low}$	91.5	82.9	14.9	2.2
+critic$_{high}$	96.4	**87.5**	10.2	2.3

Table 3.6: Model performance on knowledge base completion, measured by human judgement. Inferences are generated on held-out events from ATOMIC$^{20}_{20}$. Models besides GPT-3 use GPT-2 XL architecture. COMET$^{DIS}_{TIL}$ with a strong critic (+critic$_{high}$) achieves the highest acceptance rate overall–87.5.

critical versions of ATOMIC$^{\mathbf{10x}}$–crit$_{low}$ denotes training on the corpus achieving 91.5% accuracy, and crit$_{high}$ on the 96.4% accuracy corpus. Models are trained for 1 epoch, with default parameters using the Huggingface Transformers library (Wolf et al., 2019).

3.5.1 *Evaluating a Symbolically Distilled Model*

Evaluation follows past work (Bosselut et al., 2019; Hwang et al., 2021; Sap et al., 2019a) testing the ability of models to do knowledge base completion, i.e. generating inferences for test events, specifically from the ATOMIC$^{20}_{20}$ test set. We use human evaluation[8] following Section 3.3.4, on 1000 inputs (event + relation), with results in Table 3.6. We compare to the GPT2-XL-based COMET$^{20}_{20}$ model trained on human-generated ATOMIC$^{20}_{20}$, and GPT-3 using the same generation method as §3.3–in effect, comparing the student COMET$^{DIS}_{TIL}$ to the *loose teacher* GPT-3. We omit the *critical teacher* (GPT-3 + critic), which is not assured to produce an inference for each input, as the critic may reject all tails for some inputs. We also compare to zero-shot GPT2-XL (Radford et al., 2019a) using the same methodology (Table 3.6).

HOW DOES COMET$^{DIS}_{TIL}$ COMPARE TO GPT-3? In knowledge distillation, the student model often deteriorates in performance (Hinton, Vinyals, and Dean, 2015; Kim and Rush, 2016) compared to its teacher. Comparing our base teacher–GPT-3–to the simplest version of COMET$^{DIS}_{TIL}$ (top-row COMET$^{DIS}_{TIL}$ of Table 3.6) surprisingly shows the student *surpasses* GPT-3, the model that generates its training data[9]. We posit that the superior performance of COMET$^{DIS}_{TIL}$ may have to do with mistakes of GPT-3 being filtered by verbalization and training of GPT-2, and possibly the focus of COMET$^{DIS}_{TIL}$

8 We find Fleiss' kappa (Fleiss, 1971) of 47.1 for acceptance, indicating moderate agreement. (Landis and Koch, 1977), and accuracy agreement of 88.7%.

9 The slight difference in acceptability for GPT-3 from Table 3.4 is likely due to variance in raters between rounds of evaluation, and a different distribution of events–Table 3.4 uses generated events while Table 3.6 uses events from ATOMIC$^{20}_{20}$.

on one commonsense domain while GPT-3 covers a more general domain. We leave further study of this effect for future work.

HOW DOES COMET$^{\text{DIS}}_{\text{TIL}}$ COMPARE TO HUMAN KNOWLEDGE? While COMET$^{\text{DIS}}_{\text{TIL}}$ *without the critic* is slightly outperformed by COMET$^{20}_{20}$ in terms of accuracy, this reverses with the critic. For both cutoffs tested, COMET$^{\text{DIS}}_{\text{TIL}}$ surpasses COMET$^{20}_{20}$, with more filtering resulting in a wider gap.

USEFULNESS OF COMET$^{\text{DIS}}_{\text{TIL}}$ For on-demand inference, where a single high quality inference for some input event/relation is required, COMET$^{\text{DIS}}_{\text{TIL}}$ is the **best available model**: the most performant version surpasses COMET$^{20}_{20}$ by 5 points and GPT-3 by over 10. The critical teacher (GPT-3 + critic) yields a more accurate *corpus*, but may filter all inferences for an input, giving no output.

3.6 RELATED WORK

COMMONSENSE KNOWLEDGE GRAPHS (CKG) CKGs provide knowledge for commonsense reasoning. Some are manually constructed, e.g. ATOMIC (Hwang et al., 2021; Sap et al., 2019a). ConceptNet (Speer, Chin, and Havasi, 2017) contains taxonomy and physical commonsense, authored by humans or compiled from such sources. Some CKGs are automatically constructed: TransOMCS (Zhang et al., 2020a) extracts 18.48M tuples from syntactic parses and CausalBank (Li et al., 2020) extracts 314M cause-effect pairs by pattern-matching. In contrast, we *generate* commonsense.

EXTRACTING KNOWLEDGE FROM LMS Past work uses models for automatic knowledge graph completion (Bosselut et al., 2019; Hwang et al., 2021; Li et al., 2020). Yet, models are trained on *existing* resources; ATOMIC$^{\mathbf{10x}}$ is generated without these. Other works mine factual/commonsense knowledge directly from off-the-shelf LMs (Davison, Feldman, and Rush, 2019; Petroni et al., 2019; Xiong et al., 2020), but not resulting in the quality at scale of ATOMIC$^{\mathbf{10x}}$.

KNOWLEDGE DISTILLATION Other works use knowledge distillation (Hinton, Vinyals, and Dean, 2015) for generation. (Sanh et al., 2019) follow a label smoothing formulation, while Kim and Rush (2016) follow a similar formulation to us (§3.2.1), but use the mode of the teacher distribution rather than sampling. Our work is unique in distilling *specific* information (commonsense) from a general language model.

DATA GENERATION While manual dataset creation is expensive and complex (Agrawal et al., 2018; Bras et al., 2020; Schwartz et al., 2017; Tsuchiya, 2018), crowdsourcing is the most popular method for goal-oriented, high quality datasets.

Past automatic data mainly use extractive approaches, e.g. syntactic parsing (Zhang et al., 2020a) or pattern matching (Li et al., 2020) from unstructured text (Buck, Heafield, and Van Ooyen, 2014; Lehmann et al., 2015). These scale, but are noisy and limited

in format–ATOMIC knowledge will not appear simply in natural text. Some works explore automatic data synthesis/expansion by finetuning LMs on existing labeled data (Anaby-Tavor et al., 2020; Kumar, Choudhary, and Cho, 2020; Papanikolaou and Pierleoni, 2020; Yang et al., 2020), but are limited by data quality.

3.7 CONCLUSIONS

We introduce symbolic knowledge distillation, a *machine–to–corpus–to–machine* pipeline for commonsense that does not require human-authored knowledge–instead, using machine generation. Knowledge is transferred from a large, general model to a compact commonsense model, through a commonsense corpus–yielding a commonsense knowledge graph and model. Our resulting symbolic knowledge graph has greater scale, diversity, and quality than human authoring. symbolic knowledge distillation offers an alternative to human-authored knowledge in commonsense research.

Part II

COUNTERINTUITIVE LIMITS

4

PARADOX OF GENERATIVE AI

4.1 INTRODUCTION

"What I cannot create, I do not understand." – Richard Feynman

The recent wave of generative AI, from ChatGPT to GPT4 to DALL-E 2/3 to Midjourney, has sparked unprecedented global attention—with equal parts excitement about the expansive potential applications, and deep concern about the dangers of "intelligence[1]" that seems even to exceed that of humans. Indeed, in both language and visual domains, current generative models take only seconds to produce outputs that could challenge experts with years of skill and knowledge, providing compelling motivation for claims that models have surpassed human intelligence (Bubeck et al., 2023; Surameery and Shakor, 2023). At the same time, probing of models' outputs continues to uncover basic errors in understanding that would be unexpected even for non-expert humans (Arkoudas, 2023; Dziri et al., 2023; Qin et al., 2023). This presents us with an apparent paradox: how do we reconcile the seemingly superhuman capabilities of these models with the persistent presence of fundamental errors that most humans could correct?

We posit that this tension arises because the configuration of capabilities in today's generative models diverges from the configuration of intelligence in humans. Specifically, in this work we propose and test the **Generative AI Paradox** hypothesis: generative models, having been trained directly to reproduce expert-like outputs, acquire generative capabilities that are not contingent upon—and can therefore exceed—their ability to understand those same types of outputs. This contrasts with humans,

1 "Intelligence" and "understanding" here refer particularly to demonstrable aspects of models and technology (as in "Artificial Intelligence" or "Natural Language Understanding").

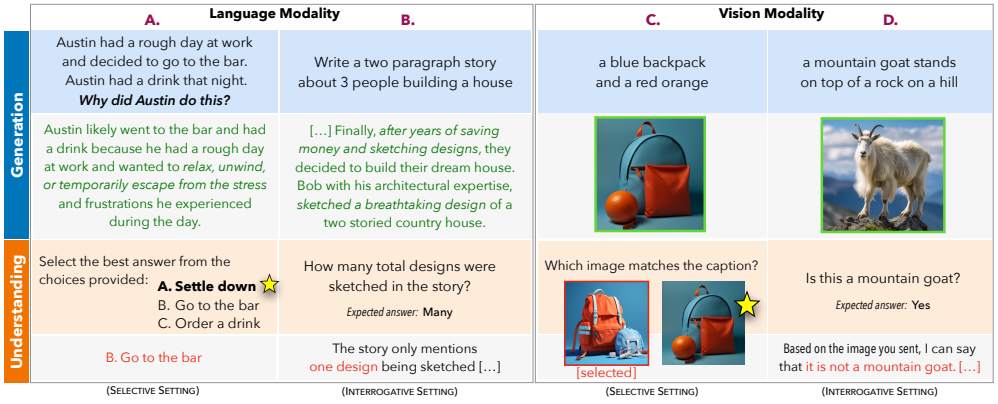


Figure 4.1: Generative AI in language and vision can produce high-quality generations. Paradoxically, however, models have trouble demonstrating selective (A,C) or interrogative (B,D) understanding of these modalities.

for whom basic understanding nearly always serves as a prerequisite to the ability to generate expert-level outputs (Alexander, 2003; Berliner, 1994; Gobet, 2017).

We test this hypothesis through controlled experiments analyzing generation and understanding capabilities in generative models, across language and visual modalities. We conceptualize "understanding" relative to generation via two angles: 1) given a generative task, to what extent can models select correct responses in a discriminative version of that same task? and 2) given a correct generated response, to what extent can models answer questions about the content and appropriateness of that response? This results in two experimental settings, *selective* and *interrogative*, respectively.

Though our results show variation across tasks and modalities, a number of clear trends emerge. In selective evaluation, models often match or even outperform humans on generative task settings, but they fall short of human performance in discriminative (understanding) settings. Further analysis shows that discrimination performance is more tightly linked to generation performance in humans than in GPT4, and human discrimination performance is also more robust to adversarial inputs, with the model-human discrimination gap increasing with task difficulty. Similarly, in interrogative evaluation, though models can generate high-quality outputs across tasks, we observe frequent errors in models' ability to answer questions about those same generations, with model understanding performance again underperforming human understanding. We discuss a number of potential reasons for this divergence in capability configurations for generative models versus humans, including model training objectives, and size and nature of input.

Our findings have a number of broader implications. First, the implication that existing conceptualizations of intelligence, as derived from experience with humans, may not be able to be extrapolated to artificial intelligence—although AI capabilities in many ways appear to mimic or exceed human intelligence, the contours of the capability landscape may diverge fundamentally from expected patterns in human cognition. On the flip side, our findings advise caution when studying generative models for insights into human intelligence and cognition, as seemingly expert human-like outputs may belie non-human-like mechanisms. Overall, the generative AI paradox encourages studying models as an intriguing counterpoint to human intelligence, rather than as a parallel.

4.2 THE GENERATIVE AI PARADOX

We begin by outlining the Generative AI Paradox and an experimental design to test it.

4.2.1 *Operational Definitions*

Figure 4.1 offers examples of the seemingly paradoxical behavior of generative models. In language (column B), GPT4 is able to generate a compelling story about 3 friends building a house, but when pressed on details of its *own generated story*, fails to correctly answer a simple question: GPT4 asserts that only one design was sketched in the story

despite writing about years of "sketching designs". In vision (column C), a generator produces a correct image beyond average human capabilities, yet the understanding model is unable to single out that correct generation against plausible alternatives, despite selection being the seemingly "easier" task. In both cases, models meet or exceed human generation abilities but lag in understanding.

Observations such as these motivate the Generative AI Paradox:

> *Generative models seem to acquire generation abilities more effectively than understanding, in contrast to human intelligence where generation is usually harder.*

Testing this hypothesis requires an operational definition of each aspect of the paradox. First, what it means for generation to be "more effective" than understanding for a given generative model m_g, understanding model m_u and task t, with human intelligence as a baseline. Taking $\mathbf{g}$ and $\mathbf{u}$ to be some *performance measures* of generation and understanding, we formally state the Generative AI Paradox hypothesis as:

$$\mathbf{g}(\text{human}, t) = \mathbf{g}(\mathrm{m}_g, t) \implies \mathbf{u}(\text{human}, t) - \mathbf{u}(\mathrm{m}_u, t) > \epsilon \tag{4.1}$$

Put simply, the hypothesis holds for a task t if a human who achieves the same generation performance $\mathbf{g}$ as a model m_g would be expected to achieve significantly ($> \epsilon$ for a reasonably large ϵ) higher understanding performance $\mathbf{u}$ than a model m_u does[2]. In simpler terms, models perform worse on understanding than we would expect of humans with similarly strong generative capabilities. In the language domain,

Generation is straightforward to operationally define: given a task input (question/prompt), generation is the production of observable content to satisfy that input. Thus, performance $\mathbf{g}$ can be evaluated automatically or by humans (e.g. style, correctness, preference). While understanding is not defined by some observable output, it can be tested by explicitly defining its effects.Thus, we measure performance $\mathbf{u}$ by asking the following questions:

1. **Selective evaluation.** For a given task, which can be responded to generatively, to what extent can models also select accurate answers among a provided candidate set in a discriminative version of that same task? A common example of this is multiple choice question answering, which is one of the most common ways to examine both human understanding and natural language understanding in language models (Wang et al., 2019) (Figure 4.1, columns A, C). This tests the performance aspect of understanding, i.e. the ability to identify the answer to a human input.

2. **Interrogative evaluation.** For a given generated model output, to what extent can models accurately respond to questions about the content and appropriateness of that output? This is akin to an oral examination in education (Sabin, Jin, and Smith, 2021). (Figure 4.1, columns B, D) This tests the explainability aspect of understanding, i.e. the ability to comprehend one's own answer.

2 To clarify, the paradox hypothesis is not restricted to the use of a single model to assess both generative and understanding capabilitites; different models can be employed to test these two aspects independently.

These definitions of understanding provide us with a blueprint for evaluating the Generative AI Paradox, allowing us to test whether Hypothesis 4.1 holds across modalities, tasks, and models.

4.2.2 *Experimental Overview*

Here, we provide a high-level road map for experiments informed by the definitions above. We propose 2 sub-hypotheses to test across experimental settings, and provide cross-experiment details.

4.2.2.1 *Hypotheses*

Evaluating whether Hypothesis 4.1 holds for a given task requires establishing a human baseline, specifically, the understanding performance we expect from a human with the same generation capabilities as the model. We define how such a baseline is established for both kinds of understanding above, resulting in 2 sub-hypotheses.

SELECTIVE EVALUATION. Here, we explicitly measure human generation and understanding performance to establish a baseline. We say Hypothesis 4.1 holds if models underperform in understanding compared to humans with equivalent generation performance (or lower generation performance, assuming that if humans *matched* model generation they would do even better at understanding. The sub-hypothesis is simply:

sub-hypothesis 1: *models meet or exceed humans at generation while lagging at discrimination.*

INTERROGATIVE EVALUATION. For the human baseline here, we assume that humans *can answer simple questions of understanding about their own generations.* For a given task input, we test how accurate models are at answering questions on AI generated outputs and as the human baseline, assume near-perfect accuracy on such questions for their own generations. The sub-hypothesis in this case is:

sub-hypothesis 2: *models struggle to answer simple questions about generated content, which humans could answer for their own generations.*

4.2.2.2 *Models and Experiments*

We focus our study on the strongest current generative models, i.e., those driving interest and concern among experts and the public. We investigate language and vision, modalities where recent impressive progress has been made. We test language models for both generative and understanding capabilities given strong performance in both areas, i.e. taking $m_u = m_g$. We test GPT4 (`gpt-4`) and GPT3.5 (`GPT3.5-turbo`) in a zero-shot setting where we instruct models to output a response given some background information. In contrast, for vision, image generators show weaker understanding (Li et al., 2023a) than dedicated understanding models, and so we assume $m_u \neq m_g$ for vision. We use Midjourney (Inc., 2023) to generate, CLIP (Radford et al., 2021) and OpenCLIP (Il-

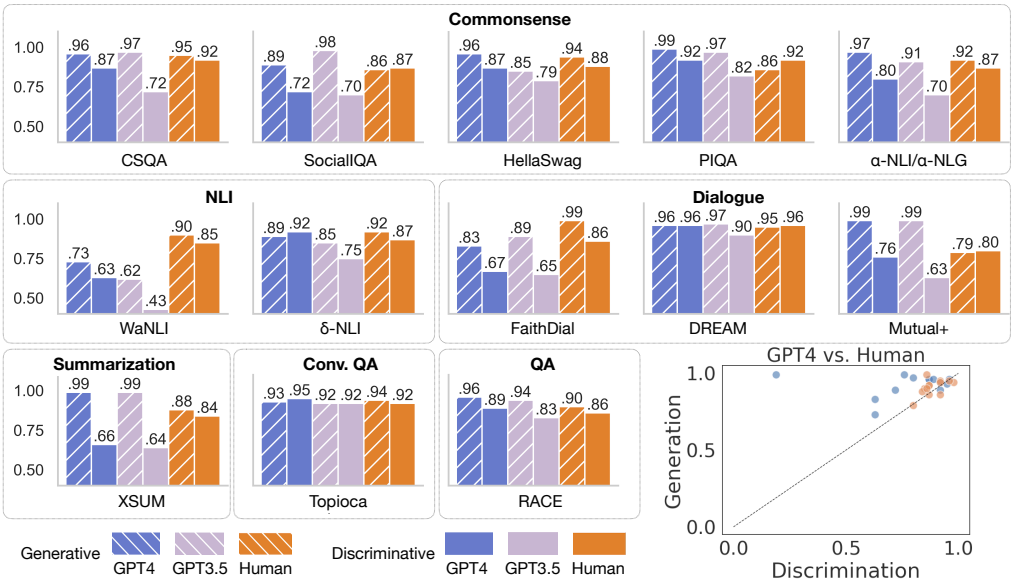


Figure 4.2: Discriminative and generative performance of GPT3.5 and GPT4 vs Humans. Models outperform humans in generation but underperform them in discrimination for most of the cases. The scatter plot in the bottom right summarizes GPT4's performance vs. human performance (using the hard negatives from Section 4.3.2 to measure discriminative accuracy for XSUM and FaithDial); each point represents a different task. Humans have a larger positive slope between their discrimination and generation abilities compared to GPT4.

harco et al., 2021) as understanding models for selective evaluation, and BLIP-2 (Li et al., 2023b), BingChat (Microsoft, 2023), and Bard (Google, 2023) for interrogative evaluation. All results on vision models are obtained in zero-shot fashion.

We conduct experiments across both sub-hypotheses, investigating tasks with selective evaluation of understanding (sub-hypothesis 1) in §4.3 and investigating tasks with interrogative evaluation of understanding (sub-hypothesis 2) in §4.4. Both sections include both language and vision tasks.

4.3 CAN MODELS DISCRIMINATE WHEN THEY CAN GENERATE?

First, in our *selective* evaluation, we conduct a side-by-side performance analysis on generative and discriminative variants of tasks to assess models' generation and understanding capabilities in language and vision modalities. We compare this generative and discriminative performance to that of humans. For our tasks we draw on diverse source benchmarks, detailed below:

Language benchmarks. For *dialogue*, we explore two open-ended datasets—**Mutual**$^+$ (Cui et al., 2020) and **DREAM** (Sun et al., 2019), and a document-grounded benchmark, **Faithdial** (Dziri et al., 2022). These tasks require generating coherent continuations based on conversation history (faithful to the document in grounded dialogue). For *reading comprehension*, we include **Topioca** ((Adlakha et al. 2022); conversational QA) and **RACE** ((Lai et al. 2017); factual QA). For *summarization*, we consider **XSUM** (Narayan, Cohen, and Lapata, 2018). We also include the *commonsense*

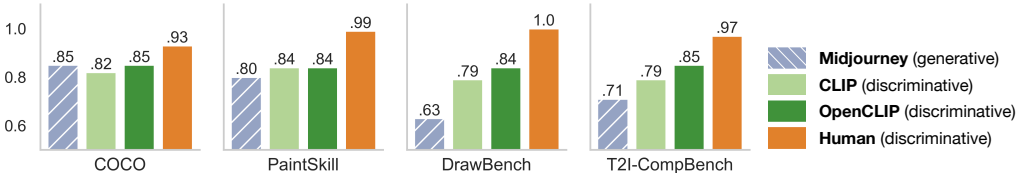


Figure 4.3: Model and human performance under the generative and discriminative settings on the **vision** modality. We observe models fall short of human accuracy in discriminative performance, and their generative accuracy also lags behind their discriminative accuracy.

benchmarks **CommonSenseQA** (Talmor et al., 2019), **SocialIQA** (Sap et al., 2019b), **HellaSwag** (Zellers et al., 2019), **PIQA** (Seo et al., 2018), and α**NLG**/α**NLI** (Bhagavatula et al., 2020). Lastly, we consider the *natural language inference* tasks **WaNLI** (Liu et al., 2022) and δ-**NLI** (Rudinger et al., 2020).

Vision benchmarks. For image generation, we source text prompts from four benchmarks: these range from descriptions of natural scenes, (likely in-domain for the model) to out-of-distribution scenes with specific attributes and relationships that rarely exist in real images. Prompts are sourced from: **COCO** (Lin et al., 2014), **PaintSkill** (Cho, Zala, and Bansal, 2022), **DrawBench** (Saharia et al., 2022) and **T2ICompBench** (Huang et al., 2023).

Experimental setup. For each task and modality, we consider two settings: **i) generative:** we prompt models to generate a response given task-specific inputs (e.g., dialogue history, document, image caption), and **ii) discriminative:** we require task-specific models to select the correct answer from a set of candidates, using existing candidates where available and otherwise generating options.

For the generative setting, we conduct human evaluations using Amazon Mechanical Turk (AMT) to judge the correctness of the generated responses (i.e, text or image) and report percentage of successful responses satisfying task requirements. For example, for the language domain, we present humans with examples from the language benchmarks.

For the discriminative setting, we report the accuracy of choosing the ground-truth response among the candidate options. To establish a human performance baseline, we ask workers to perform all discriminative tasks and evaluate the correctness of the ground-truth responses for each task.[3]

4.3.1 *Generative and Discriminative Capabilities in Models vs. Humans*

Language. Figure 4.2 presents a comparison of GPT3.5, GPT4, and human generative and discriminative performances. We see that for 10 of the 13 datasets, Sub-hypothesis 1 is supported in at least one model, with models outperforming humans in generation but underperforming humans in discrimination. For 7 of the 13 datasets, this sub-hypothesis is supported in both models.

Vision. It is not practical to ask humans to produce detailed images as we do with vision models, but we assume that an average human could not achieve the stylistic

3 Ground-truth responses were initially written by humans for the language tasks, while ground-truth images are generated by Midjourney.

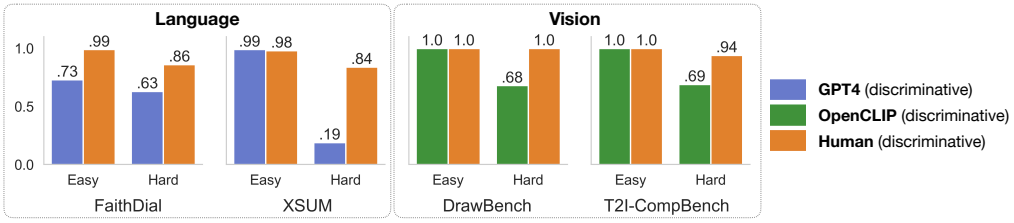


Figure 4.4: Model vs. human performance across varying levels of answer difficulty on discriminative tasks.

quality of models like Midjourney and thus assume human generation performance is lower. Therefore, we only compare models' generative and discriminative accuracy to humans' discriminative accuracy. Similar to the language domain, Figure 4.3 shows that CLIP and OpenCLIP[4] fall short of human accuracy in discriminative performance. Assuming human generation is worse, this agrees with sub-hypothesis 1: Vision AI exceeds average humans at generation but lags at understanding.

4.3.2 *Models fall further short of human performance with harder discrimination tasks*

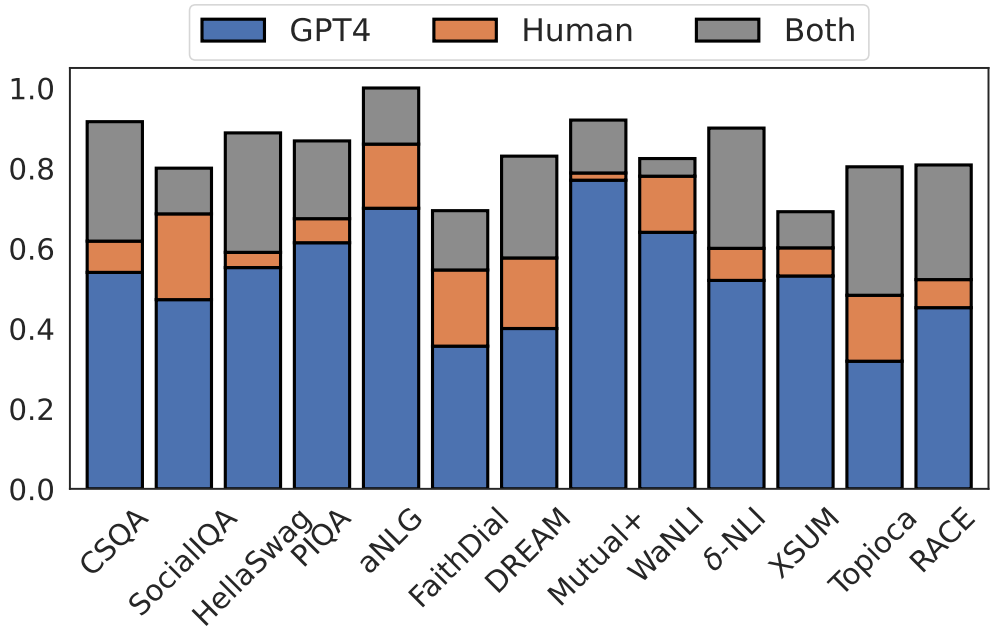
We take a closer look at the gap in discriminative performance between humans and models by manipulating the difficulty of the negative candidates. Two types of negatives are considered: **i) Hard negatives**: challenging examples that deter models from relying on data biases and artifacts to produce an answer. These negatives are wrong in subtle and challenging ways; recognizing them may require profound understanding of the task. **ii) Easy negatives**: these candidates are semantically distant from the topic of the question, providing a clear contrast to the correct answer.

Figure 4.4 (left) shows the comparison between GPT4 and humans[5]. Notably, as the complexity of the candidate answers increases, model performance gradually declines. For instance, in the XSUM task, GPT4 achieves 100% accuracy when selecting the correct answer from easy negatives, but this drops to 19% when confronted with hard negatives. XSUM exhibits a substantial difference in performance compared to FaithDial. Upon inspection, we observe that models tend to make the most mistakes in discrimination tasks when the responses are lengthy and challenging, such as summarizing lengthy documents. In contrast, humans can maintain a consistently high level of accuracy across different levels of difficulty.

Figure 4.4 (right) shows the discriminative performance of OpenCLIP, in comparison to humans, across difficulty levels. Consistent with the language results, and even more robustly across tasks, we see that while humans show versatile performance across hard and easy negative settings, model performance drops substantially when confronted with hard negatives (from 100% to ~69%). Overall, these results highlight that humans have the ability to discern correct answers even when faced with challenging or adversarial examples, but we see that this capability is not as robust in LMs. This discrepancy raises questions about the true extent of these models' understanding.

4 We report the best results on CLIP (`clip-vit-large-patch14`) and OpenCLIP (`CLIP-ViT-bigG-14-laion2B-39B-b160k`)

5 The same trend also applies for GPT3.5.



r0.45

Figure 4.5: Human's preference scores between human-generated vs. GPT4-generated responses

4.3.3 *Model generations are preferred over human generations*

To better understand the gap between humans and language models, we asked AMT workers to provide their preferences between machine and human-generated answers in the language-related tasks, along with a rationale for their choices.While both sets of responses score high in correctness (Figure 4.2), Figure 4.5 shows a notable trend: workers often favor responses from GPT4 over those generated by humans. The same applies for GPT3.5. The rationales provided by humans often indicate a preference for GPT4 due to longer response length, more elegant writing style, and being more informative, while human choice is preferred for brevity and conciseness. This makes the divergence in capabilities–with models excelling in relative terms at generation and humans at understanding-based tasks–even more apparent.

4.4 CAN MODELS UNDERSTAND WHAT MODELS GENERATE?

In the previous section, we showed that models often excel at generating accurate answers while lagging behind humans in the discriminative task. Now, in our *interrogative* evaluation, we investigate to what extent models can demonstrate meaningful understanding of generations—something humans are highly capable of—by directly asking models questions about generated content.

Language experimental setup. In language, we first prompt models to generate a paragraph using task-specific background information. Then using its generation as context, we ask the model multiple-choice questions about its own generated informa-

tion.[6] For example, for **XSUM** (Narayan, Cohen, and Lapata, 2018) (summarization) we prompt the model to generate an article based on a ground-truth summary, and then ask the model to select the best summary (same choices as §4.3) for the generated article. For **Mutual**$^+$ (Cui et al., 2020) (dialogue), the model generates the conversation history that leads to a given dialogue, and then is asked to choose the best dialogue continuing that history. In **HellaSwag** (Zellers et al., 2019) (commonsense), the model generates the context preceding a given sentence and then selects the most fitting continuation for that generated context. We only perform selective evaluation on the *correct generations* verified by humans.

We use zero-shot GPT3.5 and GPT4 for all of the evaluations, both generating and question answering. We report the model generation performance, the selection performance based on content generated by the model, and human selection performance using the model's generated content. As an implicit baseline, we assume that humans can answer such questions about their own generations with high accuracy, and so refrain from the complex process of eliciting these human generations.

Vision experimental setup. We conduct interrogative evaluation on image understanding models via visual question answering in an open-ended setting. We consider **TIFAv1.0** (Hu et al., 2023) as the evaluation benchmark, with text prompts from **COCO**, **PaintSkill**, **DrawBench** and **Parti** (Yu et al., 2022). TIFAv1.0 includes questions automatically generated by a language model, only concerning the content specified in the text prompt (*e.g.*, about existence/attributes of an object and relative position between objects). We first ask Midjourney to generate images, based on the text prompts. Then, we interrogate the understanding models (*e.g.*, BLIP-2) with answerable questions (verified by AMT workers) about the generated images. AMT is used to collect human responses, and judge the correctness of human/model outputs.

Results. Results for the language modality are shown in Figure 4.6 (left). We observe that while the models excel at generation, they make frequent errors in answering questions about their own generations, indicating failures in understanding. Humans, who we assume could not generate such text at the same speed or scale, consistently achieve higher accuracy in QA compared to the model, despite the fact that questions are about the model's own output. As stated in sub-hypothesis 2, we expect humans would achieve even higher accuracy for their own generations. We note that the humans in this study are not experts; producing text as sophisticated as the model's output could be a significant challenge. We anticipate that the performance gap in understanding one's own generation would widen even more when comparing the model to human experts, who are likely to answer such questions with near-perfect accuracy.

Figure 4.6 (right) shows the interrogative results in the visual modality.[7] We see that image understanding models still fall short of human accuracy in answering simple questions about elements in the generated images. At the same time, state-of-the-art image generation models can generate images at a quality and speed beyond most

6 Unlike §4.3, questions here are about the generation, rather than taking the generation as a potential answer.

7 We report performance of BingChat, Bard and the best BLIP-2 model (`BLIP2-flan-t5-xxl`) on two subsets.

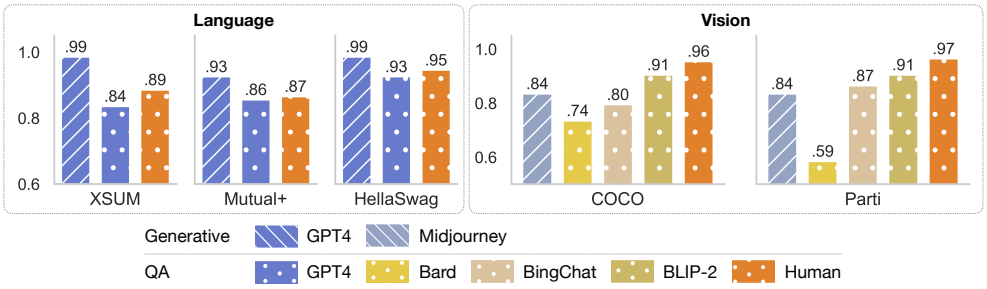


Figure 4.6: Models vs. human performance on language/visual QA based on model generated texts/images.

average humans (who we expect will have trouble generating comparable realistic images), indicating a relative gap between generation (stronger) and understanding (weaker) in vision AI compared to humans. Surprisingly, the performance gap between models and humans is smaller for simpler models than advanced multimodal LLMs (*i.e.*, Bard and BingChat), which have some intriguing visual understanding abilities, but still struggle to answer simple questions about generated images.

4.5 DISCUSSION

Assessing the generative AI paradox. Broadly, we find significant experimental evidence of the Generative AI Paradox: though models can regularly outperform humans in text and image generation, they fall short of human performance in discriminative versions of generative tasks, and when answering questions about generated content. Furthermore, our analyses show that discrimination performance is more tightly linked to generation performance in humans than in GPT4, and that human discrimination performance is also more robust to challenging inputs. These trends vary across tasks and modalities, but in general our results robustly support the hypothesis that generative capability can outstrip understanding capability in models, especially compared with humans.

Proposed explanations and points of future study. Given the above evidence in support of the Generative AI Paradox, the next question is: *what factors could lead to models that excel at generation even when they cannot demonstrate strong understanding?* We propose some hypotheses below, and encourage future work to explore this question.

Generative AI is defined by the generative learning objective, explicitly encouraging reconstruction/generation of the training distribution, while only implicitly encouraging understanding if it furthers this goal. Human learning, while not completely understood, likely diverges from this by encouraging behavior beyond pure reconstruction of stimuli.

Although we often query generative models as if they were individuals, they typically model a *medium* (e.g. text over many authors in language models). Providing context may push models closer to emulating a specific individual (Andreas, 2022), but they tend towards behavior that looks *distributionally correct* rather than *individually correct*, prioritizing stylistic and document-wide features over details necessary for understanding tasks. Training on many documents (e.g. huge swaths of internet text) also contrasts

with humans: it would take an average human reader e.g. over 32 years just to read all the pages of Wikipedia (Brysbaert, 2019; contributors, n.d.). This obvious discrepancy in not only quantity, but also diversity of knowledge could encourage models to use existing solutions to problems, which they have seen already, whereas humans have not and therefore need to exercise understanding and reasoning to answer the same questions correctly.

Evolutionary and economic pressures can affect the way that AI develops. For instance, popular language model architectures have shown a preference for languages like English (Ravfogel, Goldberg, and Linzen, 2019) which has seen the most attention in NLP (Bender, 2019) and thus the most reward for improvement. Similar pressures could encourage architectures, training paradigms, and other decisions that favor generation over understanding, as generation is harder for humans and thus more useful/valuable. Designing systems that are not affected by the Generative AI Paradox will require understanding its cause. Given the potential explanations above, promising paths forward may involve alternative optimization objectives, limiting the memorization in models to force reasoning, and even incentivizing stronger understanding at a field level.

Limitations. Dataset/benchmark contamination is a potential limitation with proprietary models, but this should have similar effects on generation *and* discriminative evaluation in §4.3, and our evaluation in §4.4 uses novel generations which would not be seen at training time. Also, we focus on a small set of the most popular/widely used models. Future work should investigate a wider range of models, including smaller or weaker models, for which we hypothesize the paradox may be even more pronounced as we often saw with GPT3.5 vs GPT4 (§4.3).

While our evaluation of human performance is focused, future work can explore more extensive comparisons between model and human performance. We also advocate for adopting comparison to humans as a widespread practice, to carefully judge when model capabilities extrapolate with human capabilities, and when they do not. Finally, we only investigate *one* divergence between humans and models. Proposing and testing other points of divergence between artificial and natural intelligence exceeds our scope but will be imperative to calm concerns and calibrate excitement.

4.6 RELATED WORK

Generative paradoxes in large language model behavior. Prior work paradoxically employs large language models to *improve their own generations*, finding that models successfully identify mistakes (despite these mistakes being generated by the models themselves). (Madaan et al., 2023) prompt models to critique and improve their own generations. (Agrawal, Mackey, and Kalai, 2023) find that models can identify hallucinated content in their own generations, and (Gero et al., 2023) show that models can identify erroneously omitted elements in generated in clinical extraction data.

Inconsistencies in large language models. Past work suggests that large language models (LMs) lack a robust concept representation. (Dziri et al., 2023) show that strong models often struggle at solving basic tasks like multiplication. (Elazar

et al., 2021) and (Ravichander et al., 2020) show that LMs make inconsistent predictions when prompted with similar statements. (Ribeiro, Guestrin, and Singh, 2019) find that QA systems often generate contradictory answers. (Kassner and Schütze, 2020) and (Ettinger, 2020) find that models can generate correct facts but also their negations. (Jang, Kwon, and Lukasiewicz, 2022) construct a benchmark showing large LMs often make inconsistent predictions. (Berglund et al., 2023) demonstrate that while models can correctly recognize factual knowledge present in their training data, they fail to make inferences related to those facts.

Generative models and human cognitive mechanisms. While the reasoning mechanism of models is unknown, prior work has investigated if models possess similar competencies with humans. (Stojnić et al., 2023) evaluate commonsense psychology, finding that while infants can reason about the causes of actions by an agent, models are not capable cannot emulating this. (Sap et al., 2022) find that language models fail to demonstrate Theory-of-Mind. (Storks et al., 2021) and (Bisk et al., 2020) show discrepancies between human and model capacities in physical commonsense reasoning.

4.7 CONCLUSIONS

In this work, we propose the Generative AI Paradox hypothesis, which posits that impressive generation abilities in generative models, by contrast to humans, may not be contingent upon commensurate understanding capabilities. We test this through controlled experiments in language and vision modalities, and though our results show variation depending on task and modality, we find robust support for this hypothesis. Our findings have a number of broader implications. In particular, they imply that existing conceptualizations of intelligence, as derived from experience with humans, may not be applicable to artificial intelligence—although AI capabilities may resemble human intelligence, the capability landscape may diverge in fundamental ways from expected patterns based on humans. Overall, the generative AI paradox suggests that the study of models may serve as an intriguing counterpoint to human intelligence, rather than a parallel.

ACKNOWLEDGEMENTS

This work was funded in part by NSF (DMS-2134012), DARPA MCS program through NIWC Pacific (N66001-19-2-4031), Darpa SemaFor, and the Allen Institute for AI. We thank OpenAI for offering access to various models through the API.

5

CONCLUSION

This dissertation set out to probe the question "is scale all your need?" to create performant AI models with human-level competencies. Although this question will likely remain open indefinitely–the future of scale remains unclear–the findings of this dissertation paint a more complex picture of the relationship between scale and AI models.

Part i explores the question of *necessity*–is scale a required ingredient in useful and human-like capabilities we would like to see in real AI systems? This part offers two alternative ingredients to scale: information theory and commonsense knowledge. Using structure from information theory, I demonstrate that even very small language models, orders of magnitude smaller than the current state of the art, can carry out complex tasks such as summarization without explicit human training. I also find that these compact models can learn very useful information, such as commonsense knowledge, using extreme-scale language models as teachers. In fact, compact student models are able to exceed extreme-scale teacher models in this domain, despite being much more efficient and in fact learning from those teacher models. Part i makes a strong case that the current prevailing preference for scale as the key ingredient is misdirected, and in fact compact language models–those lacking this ingredient of extreme-scale–can carry out many complex core skills by incorporating other concepts and ideas.

Part ii explores the question of *sufficiency*–is the extreme-scale of contemporary models enough to produce consistent and human like capabilities in models? Particularly, I explore the relationship between associated human competencies: generation and understanding for a given task. While understanding is typically thought of as a prerequisite for generation in humans–you must understand a topic before you write an essay on it–the same does not seem to hold for models. This line of work highlights the fact that models often violate human intuitions, and likely complete complex tasks in ways that diverge from human abilities. In the most extreme scale models, which seem to possess many human-like capabilities, this is particularly important in combating hype and unreasonable expectations for current and future systems.

Overall, this dissertation makes the case that research on language models and AI should be more scale-skeptical: questioning scale both as a sufficient and a necessary ingredient. Compact models are likely far more useful than expected, and there is a wide open field for method development in this space. At the same time, extreme-scale models break down in extremely complex ways, often diverging from human intuitions about which tasks are easy and which are difficult. This by no means implies that scale is not useful–indeed, extreme scale models factor into this work frequently as useful, enriching, and informative tools. Rather, I propose a future of AI research that works *up* and *down*, simultaneously learning and expanding the frontier of extreme-scale models

while pushing those benefits to more efficient and accessible small models, and most importantly digging into the limits and conceptual space of these models at every level. Below, I propose pressing next steps in this pursuit.

5.0.1 *Future Work*

SCALING LAWS OF SYMBOLIC KNOWLEDGE DISTILLATION Methods which use artificial, model-generated knowledge such as Symbolic Knowledge Distillation (West et al., 2022) offer a way to both leverage and reduce scale. Extreme scale models can produce very high quality knowledge and data to inform and enrich much more compact models, but what are the underlying dynamics of this transfer, and what are its limits? This can best be understood through the lens of neural scaling laws (Kaplan et al., 2020) to describe how this process can be improved or diminished by core factors, such as the scale of teacher and student model, quantity and filtration of generated knowledge, and more subtle factors such as underlying model generation technique. Deepening our understanding of these dynamics will both facilitate the broad impact of this paradigm, as well as deepen our fundamental understanding of model knowledge and learning.

SCIENCE OF SCALE AND SAFETY Methods to empower compact models and discover hidden capabilities will open a variety of pressing research questions intersecting AI safety. On one side, along with potential benefits, the risks of AI are multiplied as capabilities are made more broadly available in smaller, more accessible models. How can these risks be quantified, and importantly mitigated, when designing and deploying methods for compact models? For example, can inference-time algorithms discussed in Chapter 2 be designed to limit unintended or dangerous use? On the other side, methods for unlocking capabilities may also be useful tools for exploring and discovering unknown risks, particularly those that may be overlooked. The same family of methods that access complex knowledge and the ability to summarize in compact, seemingly weak models could also help to expose pressing dangers such as private or dangerous memorized data, and harmful failure points towards toxic or offensive text that are otherwise hidden.

LIMITS AND CONCEPTS IN MODEL LEARNING Another key point of interest is understanding how extreme-scale interacts with other aspects of the most effective models currently available. As described in Chapter 4, the generative AI paradox suggests that the base generative learning objective central to most extreme-scale models may result in fundamentally different capabilities than what we see in humans. Nascent work on this family of limits must be expanded, both to more precisely understand where model capabilities diverge from humans, and importantly to understand what underlying mechanisms would produce this unique set of capabilities and limits. Besides the base learning objective, another key area of research will be the effect of *alignment techniques*, which aim to make models better at following human instructions and are core to many state-of-the-art models. What kinds of limits do these result in (Li et al.,

2024), and to what extent do they access the full potential of extreme-scale models or simply superficial abilities?

CAPABILITIES AND LIMITS FOR MODEL UNDERSTANDING An enduring, underlying goal of my work outlined in these proposed future areas of study is to push forward fundamental understanding of AI models and particularly language models. This too interacts with scale: as models become larger and can be more easily queried for useful behavior, the risk naturally grows to oversimplify and over-hype these capabilities. My work will continue presenting a two-pronged approach to produce a broad, intuitive understanding of models in all of their uniqueness. A robust study of limits will work to make expectations of models more realistic, while drawing out more deliberate boundaries on what models are capable of and where they break down. While limits seek to draw better understanding by producing an *upper bound* on certain skills, the study of capabilities will continue to push the *lower bound* by discovering unexpected spikes in the landscape of useful model behavior. Together, these bounds will form a scaffolding, and a set of useful empirical and conceptual borders, closing in on more accurate theories that can form the basis of a new and sophisticated understanding of AI models.

BIBLIOGRAPHY

AI@Meta (2024). "Llama 3 Model Card." In: URL: https://github.com/meta-llama/llama3/blob/main/MODEL_CARD.md.

Adlakha, Vaibhav, Shehzaad Dhuliawala, Kaheer Suleman, Harm de Vries, and Siva Reddy (2022). "Topiocqa: Open-domain conversational question answering with topic switching." In: *Transactions of the Association for Computational Linguistics* 10, pp. 468–483.

Agrawal, Aishwarya, Dhruv Batra, Devi Parikh, and Aniruddha Kembhavi (2018). "Don't Just Assume; Look and Answer: Overcoming Priors for Visual Question Answering." In: *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4971–4980.

Agrawal, Ayush, Lester Mackey, and Adam Tauman Kalai (2023). "Do Language Models Know When They're Hallucinating References?" In: *arXiv preprint arXiv:2305.18248*.

Alexander, Patricia A (2003). "The development of expertise: The journey from acclimation to proficiency." In: *Educational researcher* 32.8, pp. 10–14.

Ammanabrolu, Prithviraj, Wesley Cheung, William Broniec, and Mark O. Riedl (2021a). "Automated Storytelling via Causal, Commonsense Plot Ordering." In: *AAAI*.

Ammanabrolu, Prithviraj, Jack Urbanek, Margaret Li, Arthur D. Szlam, Tim Rocktaschel, and Jason Weston (2021b). "How to Motivate Your Dragon: Teaching Goal-Driven Agents to Speak and Act in Fantasy Worlds." In: *NAACL*.

Anaby-Tavor, Ateret, Boaz Carmeli, Esther Goldbraich, Amir Kantor, George Kour, Segev Shlomov, Naama Tepper, and Naama Zwerdling (Apr. 2020). "Do Not Have Enough Data? Deep Learning to the Rescue!" In: *Proceedings of the AAAI Conference on Artificial Intelligence* 34, pp. 7383–7390. DOI: 10.1609/aaai.v34i05.6233.

Andreas, Jacob (2022). "Models of meaning?" The 11th Joint Conference on Lexical and Computational Semantics at NAACL.

Arabshahi, Forough, Jennifer Lee, Antoine Bosselut, Yejin Choi, and Tom. Mitchell (2021). "Conversational Multi-Hop Reasoning with Neural Commonsense Knowledge and Symbolic Logic Rules." In: *EMNLP*.

Arkoudas, Konstantine (2023). "GPT-4 Can't Reason." In: *arXiv preprint arXiv:2308.03762*.

Artetxe, Mikel, Gorka Labaka, Eneko Agirre, and Kyunghyun Cho (2017). "Unsupervised Neural Machine Translation." In: *CoRR* abs/1710.11041. arXiv: 1710.11041. URL: http://arxiv.org/abs/1710.11041.

Bachman, Philip, R. Devon Hjelm, and William Buchwalter (2019). "Learning Representations by Maximizing Mutual Information Across Views." In: *ArXiv* abs/1906.00910.

Banko, Michele, Vibhu O. Mittal, and Michael J. Witbrock (Oct. 2000). "Headline Generation Based on Statistical Translation." In: *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics*. Hong Kong: Association for Computational Linguistics, pp. 318–325. DOI: 10.3115/1075218.1075259. URL: https://www.aclweb.org/anthology/P00-1041.

Baziotis, Christos, Ion Androutsopoulos, Ioannis Konstas, and Alexandros Potamianos (2019). "SEQ^3: Differentiable Sequence-to-Sequence-to-Sequence Autoencoder for Unsupervised Abstractive Sentence Compression." In: *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL:HLT)*. To appear. Minneapolis, USA. URL: https://arxiv.org/abs/1904.03651.

Bender, Emily (2019). "High Resource Languages vs Low Resource Languages." In: *The Gradient*. URL: https://thegradient.pub/the-benderrule-on-naming-the-languages-we-study-and-why-it-matters/#fn4.

Berglund, Lukas, Meg Tong, Max Kaufmann, Mikita Balesni, Asa Cooper Stickland, Tomasz Korbak, and Owain Evans (2023). "The Reversal Curse: LLMs trained on"A is B"fail to learn"B is A"." In: URL: https://api.semanticscholar.org/CorpusID:262083829.

Berliner, David C (1994). "Expertise: The wonder of exemplary performances." In: *Creating powerful thinking in teachers and students*, pp. 161–186.

Bhagavatula, Chandra, Ronan Le Bras, Chaitanya Malaviya, Keisuke Sakaguchi, Ari Holtzman, Hannah Rashkin, Doug Downey, Wen-tau Yih, and Yejin Choi (2020). "Abductive Commonsense Reasoning." In: *ICLR*.

Bhakthavatsalam, Sumithra, Chloe Anastasiades, and Peter E. Clark (2020). "GenericsKB: A Knowledge Base of Generic Statements." In: *ArXiv* abs/2005.00660.

Bisk, Yonatan, Rowan Zellers, Jianfeng Gao, Yejin Choi, et al. (2020). "Piqa: Reasoning about physical commonsense in natural language." In: *Proceedings of the AAAI conference on artificial intelligence*. Vol. 34. 05, pp. 7432–7439.

Bollacker, Kurt D., Colin Evans, Praveen K. Paritosh, Tim Sturge, and Jamie Taylor (2008). "Freebase: a collaboratively created graph database for structuring human knowledge." In: *SIGMOD Conference*.

Bosselut, Antoine, Hannah Rashkin, Maarten Sap, Chaitanya Malaviya, A. Çelikyilmaz, and Yejin Choi (2019). "COMET: Commonsense Transformers for Automatic Knowledge Graph Construction." In: *ACL*.

Bras, Ronan Le, Swabha Swayamdipta, Chandra Bhagavatula, Rowan Zellers, Matthew E. Peters, Ashish Sabharwal, and Yejin Choi (2020). "Adversarial Filters of Dataset Biases." In: *ICML*.

Brown, P., J. Cocke, S. Della Pietra, V. Della Pietra, F. Jelinek, R. Mercer, and P. Roossin (1988). "A statistical approach to language translation." In: *Proceedings of the 12th Conference on Computational Linguistics - Volume 1*. COLING '88. Budapest, Hungry: Association for Computational Linguistics, 71–76. ISBN: 963 8431 56 3. DOI: 10.3115/991635.991651. URL: https://doi.org/10.3115/991635.991651.

Brown, Tom B., Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei (2020). *Language Models are Few-Shot Learners*. arXiv: 2005.14165 [cs.CL].

Brysbaert, Marc (2019). "How many words do we read per minute? A review and meta-analysis of reading rate." In: *Journal of Memory and Language* 109, p. 104047. ISSN: 0749-596X. DOI: https://doi.org/10.1016/j.jml.2019.104047. URL: https://www.sciencedirect.com/science/article/pii/S0749596X19300786.

Bubeck, Sébastien, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al. (2023). "Sparks of artificial general intelligence: Early experiments with gpt-4." In: *arXiv preprint arXiv:2303.12712*.

Buck, Christian, Kenneth Heafield, and Bas Van Ooyen (2014). "N-gram Counts and Language Models from the Common Crawl." In: *LREC*. Vol. 2. Citeseer, p. 4.

Chakrabarty, Tuhin, Debanjan Ghosh, Smaranda Muresan, and Nanyun Peng (July 2020). "Rˆ3: Reverse, Retrieve, and Rank for Sarcasm Generation with Commonsense Knowledge." In: *ACL*.

Chakrabarty, Tuhin, Xurui Zhang, Smaranda Muresan, and Nanyun Peng (June 2021). "MERMAID: Metaphor Generation with Symbolism and Discriminative Decoding." In: *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Online: Association for Computational Linguistics, pp. 4250–4261. DOI: 10.18653/v1/2021.naacl-main.336. URL: https://aclanthology.org/2021.naacl-main.336.

Cheng, Jianpeng and Mirella Lapata (Aug. 2016). "Neural Summarization by Extracting Sentences and Words." In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Berlin, Germany: Association for Computational Linguistics, pp. 484–494. DOI: 10.18653/v1/P16-1046. URL: https://www.aclweb.org/anthology/P16-1046.

Chiang, Ted (2023). *CHATGPT is a blurry JPEG of the web*. URL: https://www.newyorker.com/tech/annals-of-technology/chatgpt-is-a-blurry-jpeg-of-the-web.

Cho, Jaemin, Abhay Zala, and Mohit Bansal (2022). "DALL-Eval: Probing the Reasoning Skills and Social Biases of Text-to-Image Generative Transformers." In: arXiv: 2202.04053 [cs.CV].

Cui, Leyang, Yu Wu, Shujie Liu, Yue Zhang, and Ming Zhou (July 2020). "MuTual: A Dataset for Multi-Turn Dialogue Reasoning." In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, pp. 1406–1416. DOI: 10.18653/v1/2020.acl-main.130. URL: https://aclanthology.org/2020.acl-main.130.

Davis, Ernest and Gary Marcus (2017). "Causal generative models are just a start." In: *Behavioral and Brain Sciences* 40.

Davison, Joe, Joshua Feldman, and Alexander M Rush (2019). "Commonsense knowledge mining from pretrained models." In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 1173–1178.

Del'etang, Gr'egoire, Anian Ruoss, Paul-Ambroise Duquenne, Elliot Catt, Tim Genewein, Christopher Mattern, Jordi Grau-Moya, Wenliang Kevin Li, Matthew Aitchison, Laurent Orseau, Marcus Hutter, and Joel Veness (2023). "Language Modeling Is

Compression." In: *ArXiv* abs/2309.10668. URL: https://api.semanticscholar.org/CorpusID:262054258.

Dziri, Nouha, Ehsan Kamalloo, Sivan Milton, Osmar Zaiane, Mo Yu, Edoardo M Ponti, and Siva Reddy (Dec. 2022). "FaithDial: A Faithful Benchmark for Information-Seeking Dialogue." In: *Transactions of the Association for Computational Linguistics* 10, pp. 1473–1490. DOI: 10.1162/tacl_a_00529.

Dziri, Nouha, Ximing Lu, Melanie Sclar, Xiang Lorraine Li, Liwei Jian, Bill Yuchen Lin, Peter West, Chandra Bhagavatula, Ronan Le Bras, Jena D Hwang, et al. (2023). "Faith and Fate: Limits of Transformers on Compositionality." In: *arXiv preprint arXiv:2305.18654*.

Elazar, Yanai, Nora Kassner, Shauli Ravfogel, Abhilasha Ravichander, Eduard Hovy, Hinrich Schütze, and Yoav Goldberg (2021). "Measuring and Improving Consistency in Pretrained Language Models." In: *Transactions of the Association for Computational Linguistics* 9, pp. 1012–1031. DOI: 10.1162/tacl_a_00410. URL: https://aclanthology.org/2021.tacl-1.60.

Ettinger, Allyson (2020). "What BERT Is Not: Lessons from a New Suite of Psycholinguistic Diagnostics for Language Models." In: *Transactions of the Association for Computational Linguistics* 8, 34–48. ISSN: 2307-387X. DOI: 10.1162/tacl_a_00298. URL: http://dx.doi.org/10.1162/tacl_a_00298.

Etzioni, Oren, Anthony Fader, Janara Christensen, Stephen Soderland, et al. (2011). "Open information extraction: The second generation." In: *Twenty-Second International Joint Conference on Artificial Intelligence*.

Fevry, Thibault and Jason Phang (Oct. 2018). "Unsupervised Sentence Compression using Denoising Auto-Encoders." In: *Proceedings of the 22nd Conference on Computational Natural Language Learning*. Brussels, Belgium: Association for Computational Linguistics, pp. 413–422. URL: https://www.aclweb.org/anthology/K18-1040.

Fleiss, Joseph L (1971). "Measuring nominal scale agreement among many raters." In: *Psychological bulletin* 76.5, p. 378.

Gero, Zelalem, Chandan Singh, Hao Cheng, Tristan Naumann, Michel Galley, Jianfeng Gao, and Hoifung Poon (2023). "Self-Verification Improves Few-Shot Clinical Information Extraction." In: *arXiv preprint arXiv:2306.00024*.

Gobet, Fernand (2017). *Understanding expertise: A multi-disciplinary approach*. Bloomsbury Publishing.

Google (2023). *Bard*. https://bard.google.com. Accessed before: 2023-09-28.

Hermann, Karl Moritz, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom (2015). "Teaching machines to read and comprehend." In: *Advances in neural information processing systems*, pp. 1693–1701. URL: https://papers.nips.cc/paper/5945-teaching-machines-to-read-and-comprehend.pdf.

Hessel, Jack and Lillian Lee (2020). "Does my multimodal model learn cross-modal interactions? It's harder to tell than you might think!" In: *EMNLP*.

Hinton, Geoffrey E (2002). "Training products of experts by minimizing contrastive divergence." In: *Neural computation* 14.8, pp. 1771–1800.

Hinton, Geoffrey, Oriol Vinyals, and Jeffrey Dean (2015). "Distilling the Knowledge in a Neural Network." In: *NIPS Deep Learning and Representation Learning Workshop*. URL: http://arxiv.org/abs/1503.02531.

Hjelm, Devon, Alex Fedorov, Samuel Lavoie-Marchildon, Karan Grewal, Phil Bachman, Adam Trischler, and Yoshua Bengio (2019). "Learning deep representations by mutual information estimation and maximization." In: *ICLR 2019*. ICLR. URL: https://www.microsoft.com/en-us/research/publication/learning-deep-representations-by-mutual-information-estimation-and-maximization/.

Hochreiter, Sepp and Jürgen Schmidhuber (1997). "Long Short-Term Memory." In: *Neural Computation* 9.8, pp. 1735–1780. DOI: 10.1162/neco.1997.9.8.1735.

Holtzman, Ari, Jan Buys, Maxwell Forbes, and Yejin Choi (2020). "The Curious Case of Neural Text Degeneration." In: *International Conference on Learning Representations*.

Hu, Yushi, Benlin Liu, Jungo Kasai, Yizhong Wang, Mari Ostendorf, Ranjay Krishna, and Noah A Smith (2023). "Tifa: Accurate and interpretable text-to-image faithfulness evaluation with question answering." In: *arXiv preprint arXiv:2303.11897*.

Huang, Kaiyi, Kaiyue Sun, Enze Xie, Zhenguo Li, and Xihui Liu (2023). "T2I-CompBench: A Comprehensive Benchmark for Open-world Compositional Text-to-image Generation." In: *arXiv preprint arXiv:2307.06350*.

Hwang, Jena D., Chandra Bhagavatula, Ronan Le Bras, Jeff Da, Keisuke Sakaguchi, Antoine Bosselut, and Yejin Choi (2021). "COMET-ATOMIC 2020: On Symbolic and Neural Commonsense Knowledge Graphs." In: *AAAI*.

Ilharco, Gabriel, 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 (July 2021). *OpenCLIP*. Version 0.1. If you use this software, please cite it as below. DOI: 10.5281/zenodo.5143773. URL: https://doi.org/10.5281/zenodo.5143773.

Inc., Midjourney (2023). *Midjourney*. https://midjourney.com. Accessed before: 2023-09-28.

Jang, Myeongjun, Deuk Sin Kwon, and Thomas Lukasiewicz (Oct. 2022). "BECEL: Benchmark for Consistency Evaluation of Language Models." In: *Proceedings of the 29th International Conference on Computational Linguistics*. Gyeongju, Republic of Korea: International Committee on Computational Linguistics, pp. 3680–3696. URL: https://aclanthology.org/2022.coling-1.324.

Jelinek, Frederick (1997). "Statistical methods for speech recognition." In: URL: https://api.semanticscholar.org/CorpusID:12495425.

Jiang, Albert Q., Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed (2023). *Mistral 7B*. arXiv: 2310.06825 [cs.CL]. URL: https://arxiv.org/abs/2310.06825.

Jurafsky, Daniel and James H. Martin (2009). *Speech and Language Processing (2Nd Edition)*. Upper Saddle River, NJ, USA: Prentice-Hall, Inc. ISBN: 0131873210.

Kaplan, Jared, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei (2020). *Scaling Laws for Neural Language Models*. arXiv: 2001.08361 [cs.LG]. URL: https://arxiv.org/abs/2001.08361.

Kassner, Nora and Hinrich Schütze (2020). "Negated and Misprimed Probes for Pretrained Language Models: Birds Can Talk, But Cannot Fly." In: Association for Computational Linguistics.

Kearns, William R., Neha Kaura, Myra Divina, Cuong Viet Vo, Dong Si, Teresa M. Ward, and Weichao Yuwen (2020). "A Wizard-of-Oz Interface and Persona-based Methodology for Collecting Health Counseling Dialog." In: *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*.

Kim, Yoon and Alexander M. Rush (2016). "Sequence-Level Knowledge Distillation." In: *EMNLP*.

Kneser, R. and H. Ney (1995). "Improved backing-off for M-gram language modeling." In: *1995 International Conference on Acoustics, Speech, and Signal Processing*. Vol. 1, 181–184 vol.1. DOI: 10.1109/ICASSP.1995.479394.

Kumar, Varun, Ashutosh Choudhary, and Eunah Cho (Dec. 2020). "Data Augmentation using Pre-trained Transformer Models." In: *Proceedings of the 2nd Workshop on Life-long Learning for Spoken Language Systems*. Suzhou, China: Association for Computational Linguistics, pp. 18–26. URL: https://www.aclweb.org/anthology/2020.lifelongnlp-1.3.

Lai, Guokun, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy (Sept. 2017). "RACE: Large-scale ReAding Comprehension Dataset From Examinations." In: *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Copenhagen, Denmark: Association for Computational Linguistics, pp. 785–794. DOI: 10.18653/v1/D17-1082. URL: https://aclanthology.org/D17-1082.

Landis, J Richard and Gary G Koch (1977). "The measurement of observer agreement for categorical data." In: *biometrics*, pp. 159–174.

LeCun, Yann (2018). *Self-supervised learning: could machines learn like humans?* https://www.youtube.com/watch?v=7I0Qt7GALVk.

Lehmann, Jens, Robert Isele, Max Jakob, Anja Jentzsch, D. Kontokostas, Pablo N. Mendes, Sebastian Hellmann, M. Morsey, Patrick van Kleef, S. Auer, and C. Bizer (2015). "DBpedia - A large-scale, multilingual knowledge base extracted from Wikipedia." In: *Semantic Web* 6, pp. 167–195.

Lester, Brian, Rami Al-Rfou, and Noah Constant (2021). "The Power of Scale for Parameter-Efficient Prompt Tuning." In: *EMNLP*.

Li, Alexander C, Mihir Prabhudesai, Shivam Duggal, Ellis Brown, and Deepak Pathak (2023a). "Your diffusion model is secretly a zero-shot classifier." In: *arXiv preprint arXiv:2303.16203*.

Li, Junnan, Dongxu Li, Silvio Savarese, and Steven Hoi (2023b). "Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models." In: *arXiv preprint arXiv:2301.12597*.

Li, Margaret, Weijia Shi, Artidoro Pagnoni, Peter West, and Ari Holtzman (2024). *Predicting vs. Acting: A Trade-off Between World Modeling Agent Modeling*. arXiv: 2407.02446 [cs.CL]. URL: https://arxiv.org/abs/2407.02446.

Li, Piji, Wai Lam, Lidong Bing, and Zihao Wang (Sept. 2017). "Deep Recurrent Generative Decoder for Abstractive Text Summarization." In: *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Copenhagen, Denmark: Association for Computational Linguistics, pp. 2091–2100. DOI: 10.18653/v1/D17-1222. URL: https://www.aclweb.org/anthology/D17-1222.

Li, Zhongyang, Xiao Ding, Ting Liu, J. Edward Hu, and Benjamin Van Durme (2020). "Guided Generation of Cause and Effect." In: *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*.

Lin, Tsung-Yi, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick (2014). "Microsoft coco: Common objects in context." In: *European conference on computer vision*. Springer, pp. 740–755.

Liu, Alisa, Swabha Swayamdipta, Noah A. Smith, and Yejin Choi (Dec. 2022). "WANLI: Worker and AI Collaboration for Natural Language Inference Dataset Creation." In: *Findings of the Association for Computational Linguistics: EMNLP 2022*. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, pp. 6826–6847. DOI: 10.18653/v1/2022.findings-emnlp.508. URL: https://aclanthology.org/2022.findings-emnlp.508.

Liu, Yinhan, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov (2019). *RoBERTa: A Robustly Optimized BERT Pretraining Approach*. arXiv: 1907.11692 [cs.CL].

Madaan, Aman, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, et al. (2023). "Self-refine: Iterative refinement with self-feedback." In: *arXiv preprint arXiv:2303.17651*.

McAllester, David (2018). *Information Theoretic Co-Training*. arXiv: 1802.07572 [cs.LG].

Merrill, William, Yoav Goldberg, Roy Schwartz, and Noah A. Smith (2021). "Provable Limitations of Acquiring Meaning from Ungrounded Form: What Will Future Language Models Understand?" In: *Transactions of the Association for Computational Linguistics* 9, pp. 1047–1060.

Miao, Yishu and Phil Blunsom (Nov. 2016). "Language as a Latent Variable: Discrete Generative Models for Sentence Compression." In: *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Austin, Texas: Association for Computational Linguistics, pp. 319–328. DOI: 10.18653/v1/D16-1031. URL: https://www.aclweb.org/anthology/D16-1031.

Microsoft (2023). *BingChat*. https://bing.com/chat. Accessed before: 2023-09-28.

Mitchell, Tom Michael, William W. Cohen, Estevam R. Hruschka, Partha P. Talukdar, Bo Yang, Justin Betteridge, Andrew Carlson, Bhavana Dalvi, Matt Gardner, Bryan Kisiel, Jayant Krishnamurthy, N. Lao, Kathryn Mazaitis, Thahir Mohamed, Ndapandula Nakashole, Emmanouil Antonios Platanios, Alan Ritter, Mehdi Samadi, Burr Settles, Richard C. Wang, D. Wijaya, Abhinav Gupta, Xinlei Chen, Abulhair Saparov, Malcolm Greaves, and Joel Welling (2015). "Never-Ending Learning." In: *Communications of the ACM* 61, pp. 103 –115.

Nallapati, Ramesh, Bowen Zhou, Cicero dos Santos, Caglar Gulcehre, and Bing Xiang (2016). "Abstractive Text Summarization using Sequence-to-sequence RNNs and Beyond." In: *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*, pp. 280–290. URL: https://www.aclweb.org/anthology/K16-1028.

Narayan, Shashi, Shay B. Cohen, and Mirella Lapata (2018). "Don't Give Me the Details, Just the Summary! Topic-Aware Convolutional Neural Networks for Extreme Summarization." In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Brussels, Belgium: Association for Computational Linguistics, pp. 1797–1807. DOI: 10.18653/v1/D18-1206. URL: https://aclanthology.org/D18-1206.

Oord, Aäron van den, Yazhe Li, and Oriol Vinyals (2018). "Representation Learning with Contrastive Predictive Coding." In: *ArXiv* abs/1807.03748.

OpenAI (2023). *GPT-4 Technical Report*. arXiv: 2303.08774 [cs.CL]. URL: https://arxiv.org/pdf/2303.08774.pdf.

Ouyang, Long, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe (2022). *Training language models to follow instructions with human feedback*. arXiv: 2203.02155 [cs.CL]. URL: https://arxiv.org/abs/2203.02155.

Over, Paul, Hoa Dang, and Donna Harman (2007). "DUC in context." In: *Information Processing & Management* 43.6, pp. 1506–1520.

Papanikolaou, Yannis and A. Pierleoni (2020). "DARE: Data Augmented Relation Extraction with GPT-2." In: *ArXiv* abs/2004.13845.

Petroni, Fabio, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller (2019). "Language Models as Knowledge Bases?" In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 2463–2473.

Qin, Chengwei, Aston Zhang, Zhuosheng Zhang, Jiaao Chen, Michihiro Yasunaga, and Diyi Yang (2023). "Is ChatGPT a general-purpose natural language processing task solver?" In: *arXiv preprint arXiv:2302.06476*.

Radford, Alec, 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*. PMLR, pp. 8748–8763.

Radford, Alec, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever (2018). "Improving Language Understanding by Generative Pre-Training." In.

Radford, Alec, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever (2019a). "Language Models are Unsupervised Multitask Learners." In.

Radford, Alec, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever (2019b). *Language models are unsupervised multitask learners*. Unpublished manuscript.

URL: https://d4mucfpksywv.cloudfront.net/better-language-models/language_models_are_unsupervised_multitask_learners.pdf.

Raffel, Colin, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu (2019). "Exploring the limits of transfer learning with a unified text-to-text transformer." In: *arXiv preprint arXiv:1910.10683*.

Ravfogel, Shauli, Yoav Goldberg, and Tal Linzen (2019). "Studying the Inductive Biases of RNNs with Synthetic Variations of Natural Languages." In: *North American Chapter of the Association for Computational Linguistics*. URL: https://api.semanticscholar.org/CorpusID:80628431.

Ravichander, Abhilasha, Eduard Hovy, Kaheer Suleman, Adam Trischler, and Jackie Chi Kit Cheung (Dec. 2020). "On the Systematicity of Probing Contextualized Word Representations: The Case of Hypernymy in BERT." In: *Proceedings of the Ninth Joint Conference on Lexical and Computational Semantics*. Barcelona, Spain (Online): Association for Computational Linguistics, pp. 88–102. URL: https://aclanthology.org/2020.starsem-1.10.

Ribeiro, Marco Tulio, Carlos Guestrin, and Sameer Singh (July 2019). "Are Red Roses Red? Evaluating Consistency of Question-Answering Models." In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, pp. 6174–6184. DOI: 10.18653/v1/P19-1621. URL: https://aclanthology.org/P19-1621.

Roose, Kevin (2024). *A.I.'s black boxes just got a little less mysterious*. URL: https://www.nytimes.com/2024/05/21/technology/ai-language-models-anthropic.html.

Rudinger, Rachel, Vered Shwartz, Jena D. Hwang, Chandra Bhagavatula, Maxwell Forbes, Ronan Le Bras, Noah A. Smith, and Yejin Choi (Nov. 2020). "Thinking Like a Skeptic: Defeasible Inference in Natural Language." In: *Findings of the Association for Computational Linguistics: EMNLP 2020*. Online: Association for Computational Linguistics, pp. 4661–4675. DOI: 10.18653/v1/2020.findings-emnlp.418. URL: https://aclanthology.org/2020.findings-emnlp.418.

Rush, Alexander M, Sumit Chopra, and Jason Weston (2015). "A Neural Attention Model for Abstractive Sentence Summarization." In: *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pp. 379–389. URL: https://www.aclweb.org/anthology/D15-1044.

Sabin, Mihaela, Karen H Jin, and Adrienne Smith (2021). "Oral exams in shift to remote learning." In: *Proceedings of the 52nd ACM Technical Symposium on Computer Science Education*, pp. 666–672.

Saharia, Chitwan, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. (2022). "Photorealistic text-to-image diffusion models with deep language understanding." In: *Advances in Neural Information Processing Systems* 35, pp. 36479–36494.

Sanh, Victor, Lysandre Debut, Julien Chaumond, and Thomas Wolf (2019). "DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter." In: *ArXiv* abs/1910.01108.

Sap, Maarten, Ronan Le Bras, Emily Allaway, Chandra Bhagavatula, Nicholas Lourie, Hannah Rashkin, Brendan Roof, Noah A Smith, and Yejin Choi (2019a). "Atomic: An atlas of machine commonsense for if-then reasoning." In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 33. 01, pp. 3027–3035.

Sap, Maarten, Ronan LeBras, Daniel Fried, and Yejin Choi (2022). "Neural theory-of-mind? on the limits of social intelligence in large lms." In: *arXiv preprint arXiv:2210.13312*.

Sap, Maarten, Hannah Rashkin, Derek Chen, Ronan Le Bras, and Yejin Choi (Nov. 2019b). "Social IQa: Commonsense Reasoning about Social Interactions." In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Hong Kong, China: Association for Computational Linguistics, pp. 4463–4473. DOI: 10.18653/v1/D19-1454. URL: https://aclanthology.org/D19-1454.

Schluter, Natalie (2017). "The limits of automatic summarisation according to rouge." In: *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pp. 41–45.

Schmidhuber, Jurgen (1990). "Making the World Differentiable: On Using Self-Supervised Fully Recurrent Neural Networks for Dynamic Reinforcement Learning and Planning in Non-Stationary Environments (TR FKI-126-90)." In.

Schwartz, Roy, Maarten Sap, Ioannis Konstas, Leila Zilles, Yejin Choi, and Noah A. Smith (Aug. 2017). "The Effect of Different Writing Tasks on Linguistic Style: A Case Study of the ROC Story Cloze Task." In: *Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017)*. Vancouver, Canada: Association for Computational Linguistics, pp. 15–25. DOI: 10.18653/v1/K17-1004. URL: https://www.aclweb.org/anthology/K17-1004.

See, Abigail, Peter J. Liu, and Christopher D. Manning (2017). "Get To The Point: Summarization with Pointer-Generator Networks." In: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*, pp. 1073–1083. DOI: 10.18653/v1/P17-1099. URL: https://doi.org/10.18653/v1/P17-1099.

Seo, Minjoon, Tom Kwiatkowski, Ankur P Parikh, Ali Farhadi, and Hannaneh Hajishirzi (2018). "Phrase-Indexed Question Answering: A New Challenge for Scalable Document Comprehension." In: *EMNLP*.

Speer, Robyn, Joshua Chin, and Catherine Havasi (2017). "Conceptnet 5.5: An open multilingual graph of general knowledge." In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 31. 1.

Stojnić, Gala, Kanishk Gandhi, Shannon Yasuda, Brenden M Lake, and Moira R Dillon (2023). "Commonsense psychology in human infants and machines." In: *Cognition* 235, p. 105406.

Storks, Shane, Qiaozi Gao, Yichi Zhang, and Joyce Chai (Nov. 2021). "Tiered Reasoning for Intuitive Physics: Toward Verifiable Commonsense Language Understanding." In: *Findings of the Association for Computational Linguistics: EMNLP 2021*. Punta Cana, Dominican Republic: Association for Computational Linguistics, pp. 4902–4918. DOI: 10.18653/v1/2021.findings-emnlp.422. URL: https://aclanthology.org/2021.findings-emnlp.422.

Sun, Kai, Dian Yu, Jianshu Chen, Dong Yu, Yejin Choi, and Claire Cardie (2019). "DREAM: A Challenge Data Set and Models for Dialogue-Based Reading Comprehension." In: *Transactions of the Association for Computational Linguistics* 7, pp. 217–231. DOI: 10.1162/tacl_a_00264. URL: https://aclanthology.org/Q19-1014.

Surameery, Nigar M Shafiq and Mohammed Y Shakor (2023). "Use chat gpt to solve programming bugs." In: *International Journal of Information Technology & Computer Engineering (IJITC) ISSN: 2455-5290* 3.01, pp. 17–22.

Sutton, Richard S. (2019). *The Bitter Lesson*. URL: http://www.incompleteideas.net/IncIdeas/BitterLesson.html (visited on 11/17/2021).

Talmor, Alon, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant (June 2019). "CommonsenseQA: A Question Answering Challenge Targeting Commonsense Knowledge." In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, pp. 4149–4158. DOI: 10.18653/v1/N19-1421. URL: https://aclanthology.org/N19-1421.

Talmor, Alon, Ori Yoran, Ronan Le Bras, Chandra Bhagavatula, Yoav Goldberg, Yejin Choi, and Jonathan Berant (2021). "CommonsenseQA 2.0: Exposing the Limits of AI through Gamification." In.

Team, Gemini et al. (2024). *Gemini: A Family of Highly Capable Multimodal Models*. arXiv: 2312.11805 [cs.CL]. URL: https://arxiv.org/abs/2312.11805.

Tishby, Naftali, Cicero Pereira, and William Bialek (July 2001). "The Information Bottleneck Method." In: *Proceedings of the 37th Allerton Conference on Communication, Control and Computation* 49.

Tishby, Naftali, Fernando C. Pereira, and William Bialek (1999). "The information bottleneck method." In: *Proc. of the 37-th Annual Allerton Conference on Communication, Control and Computing*, pp. 368–377. URL: https://www.cs.huji.ac.il/labs/learning/Papers/allerton.pdf.

Tsuchiya, Masatoshi (May 2018). "Performance Impact Caused by Hidden Bias of Training Data for Recognizing Textual Entailment." In: *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. Miyazaki, Japan: European Language Resources Association (ELRA). URL: https://www.aclweb.org/anthology/L18-1239.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin (2017). "Attention is all you need." In: *Advances in Neural Information Processing Systems*, pp. 5998–6008.

Wang, Alex, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman (2019). "GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding." In: *International Conference on Learning Representations*. URL: https://openreview.net/forum?id=rJ4km2R5t7.

Wang, Ben and Aran Komatsuzaki (May 2021). *GPT-J-6B: A 6 Billion Parameter Autoregressive Language Model*. https://github.com/kingoflolz/mesh-transformer-jax.

Wang, Yaushian and Hung-yi Lee (2018). "Learning to Encode Text as Human-Readable Summaries using Generative Adversarial Networks." In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Brussels, Belgium: Association for Computational Linguistics, pp. 4187–4195. URL: https://www.aclweb.org/anthology/D18-1451.

Welleck, Sean, Ilia Kulikov, Stephen Roller, Emily Dinan, Kyunghyun Cho, and Jason Weston (2020). "Neural Text Generation With Unlikelihood Training." In: *International Conference on Learning Representations*. URL: https://openreview.net/forum?id=SJeYe0NtvH.

West, Peter, Chandra Bhagavatula, Jack Hessel, Jena D. Hwang, Liwei Jiang, Ronan Le Bras, Ximing Lu, Sean Welleck, and Yejin Choi (2021a). *Symbolic Knowledge Distillation: from General Language Models to Commonsense Models*. arXiv: 2110.07178 [cs.CL].

West, Peter, Chandra Bhagavatula, Jack Hessel, Jena Hwang, Liwei Jiang, Ronan Le Bras, Ximing Lu, Sean Welleck, and Yejin Choi (July 2022). "Symbolic Knowledge Distillation: from General Language Models to Commonsense Models." In: *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*. Seattle, United States: Association for Computational Linguistics, pp. 4602–4625. DOI: 10.18653/v1/2022.naacl-main.341. URL: https://aclanthology.org/2022.naacl-main.341.

West, Peter, Chandrasekhar Bhagavatula, Jack Hessel, Jena D. Hwang, Liwei Jiang, Ronan Le Bras, Ximing Lu, Sean Welleck, and Yejin Choi (2021b). "Symbolic Knowledge Distillation: from General Language Models to Commonsense Models." In: *North American Chapter of the Association for Computational Linguistics*. URL: https://api.semanticscholar.org/CorpusID:238857304.

West, Peter, Ari Holtzman, Jan Buys, and Yejin Choi (Nov. 2019a). "BottleSum: Unsupervised and Self-supervised Sentence Summarization using the Information Bottleneck Principle." In: *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Hong Kong, China: Association for Computational Linguistics, pp. 3752–3761. DOI: 10.18653/v1/D19-1389. URL: https://aclanthology.org/D19-1389.

– (2019b). "BottleSum: Unsupervised and Self-supervised Sentence Summarization using the Information Bottleneck Principle." In: *ArXiv* abs/1909.07405. URL: https://api.semanticscholar.org/CorpusID:202583464.

West, Peter, Ximing Lu, Nouha Dziri, Faeze Brahman, Linjie Li, Jena D. Hwang, Liwei Jiang, Jillian R. Fisher, Abhilasha Ravichander, Khyathi Raghavi Chandu, Benjamin Newman, Pang Wei Koh, Allyson Ettinger, and Yejin Choi (2023). "The Generative AI Paradox: "What It Can Create, It May Not Understand"." In: *ArXiv* abs/2311.00059. URL: https://api.semanticscholar.org/CorpusID:264832736.

Wolf, Thomas, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, R'emi Louf, Morgan Funtowicz, and Jamie Brew (2019). "HuggingFace's Transformers: State-of-the-art Natural Language Processing." In: *ArXiv* abs/1910.03771.

Xiong, Wenhan, Jingfei Du, William Yang Wang, and Veselin Stoyanov (2020). "Pretrained Encyclopedia: Weakly Supervised Knowledge-Pretrained Language Model." In: *International Conference on Learning Representations*. URL: https://openreview.net/forum?id=BJlzm64tDH.

Yang, Yiben, Chaitanya Malaviya, Jared Fernandez, Swabha Swayamdipta, Ronan Le Bras, Ji-Ping Wang, Chandra Bhagavatula, Yejin Choi, and Doug Downey (Nov. 2020). "Generative Data Augmentation for Commonsense Reasoning." In: *Findings of the Association for Computational Linguistics: EMNLP 2020*. Online: Association for Computational Linguistics, pp. 1008–1025. DOI: 10.18653/v1/2020.findings-emnlp.90. URL: https://www.aclweb.org/anthology/2020.findings-emnlp.90.

Yu, Jiahui, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, et al. (2022). "Scaling autoregressive models for content-rich text-to-image generation." In: *arXiv preprint arXiv:2206.10789* 2.3, p. 5.

Zellers, Rowan, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi (July 2019). "HellaSwag: Can a Machine Really Finish Your Sentence?" In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, pp. 4791–4800. DOI: 10.18653/v1/P19-1472. URL: https://www.aclweb.org/anthology/P19-1472.

Zhang, Hongming, Daniel Khashabi, Y. Song, and D. Roth (2020a). "TransOMCS: From Linguistic Graphs to Commonsense Knowledge." In: *IJCAI*.

Zhang, Hongming, Xin Liu, Haojie Pan, Y. Song, and C. Leung (2020b). "ASER: A Large-scale Eventuality Knowledge Graph." In: *Proceedings of The Web Conference 2020*.

Zhou, Ben, Qiang Ning, Daniel Khashabi, and Dan Roth (July 2020). "Temporal Common Sense Acquisition with Minimal Supervision." In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, pp. 7579–7589. DOI: 10.18653/v1/2020.acl-main.678. URL: https://aclanthology.org/2020.acl-main.678.

Zhou, Jiawei and Alexander M Rush (2019). "Simple Unsupervised Summarization by Contextual Matching." In: *Proceedings of the 57th Conference of the Association for Computational Linguistics*, pp. 5101–5106.

Zhu, Yaoming, Sidi Lu, Lei Zheng, Jiaxian Guo, Weinan Zhang, Jun Wang, and Yong Yu (2018). "Texygen: A benchmarking platform for text generation models." In: *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, pp. 1097–1100.

contributors, Wikipedia (n.d.). *Wikipedia:Size of Wikipedia - Wikipedia*. en. URL: https://en.wikipedia.org/wiki/Wikipedia:Size_of_Wikipedia#:~:text=As%20of%2022%20September%202023,of%20all%20pages%20on%20Wikipedia..

COLOPHON

This dissertation was typeset using the typographical look-and-feel `classicthesis`, developed by André Miede and Ivo Pletikosić, and further refined by Amrita Mazumdar.

Final Version as of August 20, 2024 (classicthesis v4.6).