Investigating Generalization of Unlike Coordination in Language Models via Filtered-Corpus Training

Jiamu Luo

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Master of Science

University of Washington

2026

Reading Committee:
Shane Steinert-Threlkeld
Barbara Citko

Program Authorized to Offer Degree:
Computational Linguistics

University of Washington

Abstract

Investigating Generalization of Unlike Coordination in Language Models via Filtered-Corpus Training

Jiamu Luo

Chair of the Supervisory Committee:
Shane Steinert-Threlkeld
Linguistics

A long-standing debate in theoretical linguistics concerns the nature of coordination. A common view holds that only same-category constituents can be conjoined, which has been challenged by the many grammatical cases of unlike coordination found in natural language. This thesis investigates how language models (LMs) learn and generalize the linguistic phenomenon of coordination by Filtered-Corpus Training, which creates an environment where models have only been exposed to alike coordination during training. We evaluate and compare these models to counterparts trained on an unfiltered corpus. Our results suggest unlike coordination is not a general exception to LMs and can be learned only with indirect information, although direct exposure may still be needed for the more challenging cases. They further indicate that LMs process unlike coordination by treating the conjoined elements as belonging to similar structural categories or through a mechanism akin to deletion, both of which appear learnable from exposure to alike coordination alone. This work contributes to the growing understanding of how language models internally represent linguistic structure, while also adding to the broader debate on coordination by showing that how models generalize and process unlike coordination without direct exposure.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

The idea for this thesis came from two seminars. The first was a seminar on coordination, in which Barbara, my thesis reader, introduced us to a range of fascinating coordination phenomena and theories. The second was a seminar on language models and cognitive science, where Shane, my advisor, introduced me to the core methodology and philosophical thinking that ultimately shaped this thesis.

I would first like to thank my advisor, Shane Steinert-Threlkeld, for his support, encouragement, and mentorship, throughout this thesis, the application cycle, and even more, my time in this program. I am especially grateful to him for agreeing to supervise this project and for providing constant feedback and guidance even while on sabbatical. I would also like to thank my thesis reader, Barbara Citko, whose seminar sparked both intellectual curiosity and enjoyment.

Thank you, the friends I met at the CLMBR club. It's been fun working with you. And to my cohort friends: thank you for making this experience so memorable. I would like to extend my heartfelt thanks to my family and my friends besides this program, whose unwavering support, patience, and love have been the foundation of my academic journey. Without them, this work would not have been possible.

Chapter 1

INTRODUCTION

Coordination allows speakers to combine multiple constituents into one larger unit, with additional rough semantic implications that the conjuncts should be similar and equal in various perspectives. As a reflection of and perhaps also a motivation for this equal status of conjuncts, in common coordination instances, the conjuncts are of the same syntactic categories. Theoretical treatments of coordination thus often assume that two elements may be coordinated only if they share the same category, a principle known as the Law of Coordination of Likes (LCL) (Chomsky, 1957; Williams, 1978, 1981). However, there are apparent counterexamples. Conjuncts can differ syntactically, as in *you can depend on* [NP *my assistant*] *and* [CP *that he will be on time*](Sag et al. 1985:165). Such cases appear to have "coordination of unlikes." Importantly, these cannot be straightforwardly reduced to an ellipsis of coordination of two underlying CPs. This puzzling phenomenon, among many others, makes coordination a theoretically challenging domain whose precise nature remains actively debated.

Recent advances in language models (LMs) offer a novel avenue for investigating such questions. Despite their impressive performance, we still have limited understanding of how they acquire and encode complex linguistic structures. This thesis is situated at the intersection of computational language modeling and theoretical linguistics, aiming to explore how contemporary LMs learn and represent the principles of coordination. Specifically, it seeks to answer the following questions: Can LMs generalize to handle both alike and unlike coordination when they are trained only on coordination of likes, and what are their generalization patterns and internal representations? By observing and evaluating the generalization behaviors of LMs and analyzing the representations of coordination structures

in LMs in details, this thesis tries to shed light on the interaction of LM with linguistic theories of coordination, and the issue of how the behavior of LMs may inform the theories of coordination.

By examining how LMs process coordination and make generalization, this thesis connects the computational behaviors of language models directly to the theoretical debates in linguistics, and fosters a bidirectional exchange between them, allowing each to inform the other. This study uses established concepts from linguistics as a lens to interpret LMs behaviors. The generalization processes and outcomes of LMs can be framed under candidate theories of coordination, so this study can provide an account of how LMs learn coordination internally. Then, observed behaviors of the language model are intended to provide empirical evidence that may suggest theories of coordination.

Overall, this investigation contributes to understanding both the core linguistic properties of coordination and the learning biases and generalization capabilities of modern computational language models.

Chapter 2
BACKGROUND AND RELATED WORK

This research connects several lines of inquiry on various topics. It engages with a long-standing debate in theoretical linguistics concerning the nature of coordination, specifically the tension between the Law of Coordination of Likes (LCL) (Chomsky, 1957; Williams, 1978, 1981) and the vast existence of unlike coordination where conjoined elements belong to different syntactic categories. It leverages methodologies from computational linguistics, where researchers have developed sophisticated techniques to probe the syntactic abilities of large language models and assess their capacity to learn complex linguistic phenomena. It also adopts concepts and controlled experimental paradigms from the growing field at the intersection of computational linguistics and cognitive science. The following review will explore these areas to situate the present study's contribution to theoretical linguistics, computational linguistics, and cognitive science.

2.1 Treatments of Coordination

Multiple constituents can be combined into one coordinating structure by coordination as conjuncts. Additionally, there are rough semantic implications that the conjuncts should be similar and equal in various perspectives. The LCL (Chomsky, 1957; Williams, 1978, 1981) reflects this equal status of conjuncts, assuming that two elements may be coordinated only if they are of the same syntactic category or properties.

The LCL offers one of the first and the most influential description of coordination. It summarizes an intuitive and basic idea that only constituents of the same syntactic category or properties can be coordinated. Indeed, we may consider this ungrammatical sentence **I run and apple* as grammatical only if we interpret both conjuncts as verbs. Schachter

(1977) added likeness of semantic functions to the LCL. Pullum and Zwicky (1986) provided a more detailed reformulation of the LCL as "Wasow's Generalization", which states that in any syntactic position where a coordinate structure appears, each conjunct must be able to occur there on its own, and those feature values must be the same for all conjuncts.

The spirit of the LCL is true for common coordination instances. However, there are counterexamples of unlike coordination. Conjuncts may be of different syntactic categories, e.g., *you can depend on* [NP *my assistant*] *and* [CP *that he will be on time*] (Sag et al. 1985:165). This example also rejects the possibility of simple ellipsis of two CPs, because the second conjunct on its own violates the selectional restriction of the verb phrase *depend on*. There are many more apparent counterexamples that seem to involve coordination of unlikes, though they do not exclude a potential ellipsis analysis as cleanly as the preceding example. NP and AP can be conjoined, as in *Danny became a political radical and very antisocial* (Bruening and Al Khalaf, 2020), and so does AdvP and AP, *the once and future king* (Bruening and Al Khalaf, 2020). NP and AdvP, and also argument and modifier, can be conjoined, as in *Robin knows Kim and intimately* (Zoerner, 1995). The list of such cases is extensive and continues to grow.

The distinction between the common cases, i.e., alike coordination, and the less frequent but also natural cases, i.e., unlike coordination, has been extensively discussed in theoretical linguistics (Haspelmath, 2004; Sag et al., 1985) and leads to different theoretical treatments of coordination. In general, some theories posit a unified syntactic mechanism integrating the unlike cases and either partly require, or entirely build upon the LCL, while others do not presuppose the LCL so the distinction between alike/unlike coordination does not have theoretical impact.

2.1.1 Coordination Selects Equal Conjuncts

This general characterization of theoretical treatments to coordination represents a group of approaches that essentially consider all coordinations as coordination of likes in certain ways or need some part of the LCL (Chomsky, 1957; Williams, 1978, 1981) to hold. Key studies

that treat coordination in this way are summarized below.

Categories as Feature Bundles

Sag et al. (1985) provided a full analysis and grammar for coordination under the framework of Generalized Phrase Structure Grammar (GPSG). They proposed a perspective of viewing grammatical categories as complex objects, which are sets of feature-value pairs, instead of atomic symbols, with a new coordination rule similar to the LCL but tailored to this setup of grammatical categories, to maintain the likeness in unlike coordination.

This treatment of coordination requires conjuncts of a category introduced by a phrase structure rule to be its superset. Sag et al. (1985) listed multiple grammatical unlike coordination sentences involving *be* and an ungrammatical unlike coordination sentence. For example, *I am hoping to get an invitation and optimistic about my chances* (Sag et al. 1985:118) is grammatical, but *John sang beautifully and a carol* (Peterson 1981:449) is not. This can be explained by acknowledging that the phrase structure rule introducing *be* requires only an underspecified XP as its complement, but for *sing* the rule requires an optional NP as the complement. Phrases like VP and ADJP can be a superset of XP and be conjoined in the first sentence, but ADVP is not a superset of NP, so the second sentence is ungrammatical. In this view, unlike coordination, where conjuncts have different grammatical categories, is not an exception to the likeness requirement, because the apparent different categories are the same category in the sense that they are all supersets of a certain category, and categories themselves are nothing but feature bundles.

This method explains many of the unlike coordination to be in fact alike coordination. But sentences like *You can depend on my assistant and that he will be on time* are still problematic because somehow one conjunct, i.e., *that he will be on time*, is not acceptable in that complement position when extracted in a standalone sentence: * *You can depend on that he will be on time*. Sag et al. (1985) posited a solution, but they admitted that this may be highly speculative. Indeed, their solution is adding some features and rules to let the that-clause, as a sentential NP, to be an object of a preposition only when conjoined, which

seems to be quite ad hoc.

Multi-dimensional Analyses

Some scholars consider the syntactic structure behind coordination as multi-dimensional, which involves multiple planes of syntactic trees merged into one. Goodall (1987) put coordination in a parallel level, as a union of phrase markers. In this view, coordination operates over multiple trees, combining them by stacking one on top of another and merging identical nodes, then the conjunction is inserted between conjuncts during linearization. A syntactic toolkit for symmetric merge, called the multi-dominant structure, aligns with this kind of analyses. Citko (2011) motivated this symmetric structure to account for various kinds of symmetric behaviors of coordination, but noticed that the notion of multi-dominance and this structure created by symmetric merge are independent of coordination. The multi-dominant structure has a shared node, or a pivot in Muadz's (1991) term, and some parallel planes each having a tree containing the pivot.

These multi-dimensional analyses can explain many unique phenomena of coordination. For example, Right Node Raising can be easily captured within this framework. Sentences like *Lin-Manuel wrote, and I watched, Hamilton* would be derived from two planes merging the pivot: 1. *Lin-Manuel wrote Hamilton*; 2. *I watched Hamilton.* This treatment does not need the full version of LCL, because it does not require conjuncts to be of the category and only requires that they are in the right position in each plane. However, it still cares if the conjuncts can occur in the tree for each dimension. So, sentences like *You can depend on my assistant and that he will be on time* are again difficult to explain under these analyses. They also struggle with asymmetric behaviors of coordination, irreversible order of conjuncts in some cases, for example.

Supercategories

Supercategories are higher-level categories that can capture various grammatical categories. Proponents of strategies positing supercategories argue that constituents which appear to

be of different categories (e.g., an Adjective Phrase and a Prepositional Phrase) can be coordinated only if, at a more abstract level of representation, they are instances of the same supercategory.

Bruening and Al Khalaf (2020) classified unlike coordination into two types. The first involves a category mismatch with predicates and non-nominal modifiers, which is further divided into two subtypes: (a) arguments conjoined with modifiers, and (b) conjoined predicates and conjoined modifiers. The second type involves arguments of different syntactic categories to be coordinated. They proposed different analyses for the different types of unlike coordination with a same conclusion that they are all coordination of likes and coordination requires conjoined elements to match in category; the counterexamples are highly restrained to an extent that they motivate theories in which they are not counterexamples.

Sentences like *John eats only pork and only at home* (Grosu 1985:232) are examples of the first type and first subtype. Bruening and Al Khalaf (2020) treated this type as coordination of larger categories plus ellipsis, and for this example, what are conjoined are two VPs with the ellipsis of *eat them* in the second conjunct: *John* [VP *eats only pork*] *and* [VP ~~*eats them*~~ *only at home*]. Sentences like *Pat is a Republican and proud of it* (Sag et al. 1985:117) and *Danny became a political radical and very antisocial* (Bruening and Al Khalaf 2020:1) are examples of the first type and second subtype. They treat this type as coordination of PredP (Bowers, 1993) and ModP (Rubin, 1994), with their arrangement of supercategory, following Sag et al. (1985), to be in parallel with usual grammatical categories and coordination checks for supercategory only if the conjuncts have this feature. These sentences then have structure of coordination of likes: *Danny became* [PredP:NP *a political radical*] *and* [PredP:AP *very antisocial*].

You can depend on my assistant and that he will be on time, which is the sentence showing up many times in this chapter, is an example of the second type. They find that this type only contains two scenarios that also match those observed in ellipsis and displacement: a CP acts like an NP and conjoined with an NP, or a non-*ly* adverb acts like an AP and conjoined with an AP (as *once* in *the once and future king*). Their account of this unified

(a) CP → NP with null head

(b) AdvP → AP with null head

Figure 2.1: Reanalysis via null heads

with the cases in ellipsis and displacement: the apparent CP is an NP with a null head, and similarly, the adverb is an AP with a null head, as shown in Figure 2.1

Across different theories, however, the status of supercategories varies. For scholars like Bowers (1993) and Rubin (1994), supercategories arise from a null functional head. So PredP and ModP are headed by a null Pred head or Mod head. This makes supercategories entirely the same as usual grammatical categories like NP or VP. However, this may not fully explain selectional constraints that some verbs, while needing a predicate, also select certain grammatical categories. For example, *Danny became* [NP *a political radical*] *and* [AP *very antisocial*] is grammatical but **Danny became* [NP *a political radical*] *and* [PP *under suspicion*] (Bruening and Al Khalaf 2020:10) is infelicitous. Sag et al. (1985) posited category features [Pred] and [Mod] to share between predicates or modifiers that can be selected. In this view, the supercategories are not reducible to normal grammatical categories, but have their own feature bundles that enter selectional relations.

2.1.2 Theories without Matching of Conjuncts

Unlike theories that mandate category matching between conjuncts, some approaches consider this constraint irrelevant. These frameworks employ a range of mechanisms to account for coordination and may or may not explicitly address cases of unlike coordination. A

significant body of work (Thiersch 1994; Munn 1992; Collins 1988; Woolford 1987; Kayne 1994; Zoerner 1995; Camacho 1997, i.a.) arises from the view that conjunctions are heads of a phrase, a configuration readily accommodated in the X-bar schema. Advocates of this approach often reject the existence of the LCL, maintaining that if conjunctions are treated as heads, the LCL can be disregarded without any theoretical consequences. But a common failure mode for these theories is that they tend to overgenerate and license ungrammatical unlike coordinations as possible.

Recursive Complements in Conjunction Phrases

Some analyses posit that conjuncts occupy the specifier and complement positions of a conjunction phrase (&P), with the complement itself having a potentially recursive structure. One such analysis proposed by Johannessen (1998) was that the non-final conjuncts function as specifiers within a conjunction phrase, while the final conjunct serves as its complement. In cases with more than two conjuncts, silent conjunctions are assumed to occupy the specifier positions of successive &Ps within the complement. A similar structure was given by Zoerner (1995). However, in Zoerner's analysis of multiple coordination, only one overt conjunction appears in its canonical position, while the remaining conjunction heads are supplied through head movement.

An advantage of this analysis is that handles unlike coordination effortlessly. The structure itself eliminates the problem in the first place, requiring no additional mechanisms. Moreover, Johannessen (1998) proposes that the features of &P percolate from its specifier via Spec/Head agreement, enabling the account to accommodate challenging cases such as conjoined predicates or conjoined modifiers. In this example, *you can depend on* [*my assistant*] *and* [*that he will be on time*], the first conjunct is shared by the &P, allowing subsequent conjuncts to differ in form, which explains both the grammaticality of this order by virtue of the features of the phrase *my assistant* as the first conjunct, and the ungrammaticality of the reverse by the violation of selectional restrictions of the features of the phrase *that he will be on time*. However, a crucial challenge to this type of analysis is that it seems to be

too lenient, allowing for more unlike coordinations than we want.

Adjunction-Based Analyses

These analyses also assume that conjuncts head &Ps. The difference is that some conjuncts are viewed as adjuncts in the coordination. Munn (1993) proposes that the second conjunct, along with its &P, right-adjoins to the first conjunct. Kayne (1994) suggests the reverse: the first conjunct left-adjoins to the &P containing the second conjunct. Additionally, Munn (1993) and Progovac (1997) propose that each conjunct, within its own &P, is adjoined to an abstract phrase.

Munn's (1993) approach neatly accounts for the observation that an &P formed from NPs behaves like an NP—something that must be derived in frameworks where both conjuncts are contained within a single &P. However, in his structure, the first conjunct c-commands the subsequent ones, a prediction at odds with much empirical evidence. Kayne's (1994) revision, which makes the first conjunct the adjunct, avoids this problem and produces a configuration similar to Johannessen's (1998), handling the data in a comparable way. Analyses in which each conjunct is an adjunct have the clear advantage of capturing the absence of c-command between conjuncts. Nonetheless, they also predict the presence of multiple &Ps where only one is expected.

In Defense of Category Mismatches (Patejuk and Przepiórkowski, 2023)

Patejuk and Przepiórkowski (2023) (P&P) defended the reality of category mismatches in coordination, responding to Bruening and Al Khalaf (2020) (B&AK), who reject mismatches and treat all coordination as involving identical categories. B&AK enforce the Law of the Coordination of Likes (LCL) by offering analyses that reinterpret apparent mismatches so that conjuncts share the same category.

P&P directly challenge this position, arguing that category mismatches in coordination are a genuine grammatical phenomenon and that the analyses used to eliminate them are often unparsimonious and empirically flawed. They argue that solutions like ellipsis and

the postulation of null heads are overly abstract and lack independent motivation. Positing "invisible" structures (e.g., an elided verb or a null noun) just to save the LCL makes the theory less constrained and harder to falsify. For instance, while an ellipsis analysis might work for *John eats only pork and only at home*, it becomes far less plausible for other examples where reconstructing the elided material is awkward or impossible.

They also critique the supercategory approach. Although B&AK offer an apparently elegant account of type-1 unlike coordination using supercategories like PredP and ModP, P&P note that the theoretical mechanism for grouping base categories into sets is unclear, as are the properties of complex categories such as Pred: NP, AP. This vagueness becomes more problematic when PredP and ModP fail to account for certain cases, at which point additional supercategories (e.g., LocP, DurationP, MannerP) can be freely introduced to fit the data. Such flexibility renders the claim that only identical categories can be coordinated effectively unfalsifiable. In practice, conjuncts are deemed "the same" simply by virtue of sharing a syntactic position—having the same grammatical function, semantic role, or, in some cases, the same information-structural status. Given the proliferation of functional projections proposed since the 1980s, it is often possible to posit a projection for any shared property between unlike conjuncts. Yet, a principled definition of supercategory remains unarticulated. P&P also present a case in which the shared dependent of two conjoined verbs, *in Rome*, functions both as a Mod:PP and as a simple PP, which is an argument and cannot be a modifier: *St. Peter did reside and die in Rome* (Patejuk and Przepiórkowski 2023:332). In this case, it is unclear what category the phrase *in Rome* is, because positing supercategories makes it both a Mod:PP for *die* and a PP for *reside*, whereas if not, this phrase is just a PP and nothing more.

However, readers may worry that without clear principles governing when unlike coordination is permitted, like those that license alike coordination, theories that do not require conjuncts to share the same category would struggle to exclude many ungrammatical cases of unlike coordination. This concern partly motivates the analyses discussed in §2.1.1, which attempt to specify conditions requiring category matching. P&P propose that a grammar

should directly license the coordination of non-identical categories. In this view, any constituents can be coordinated, provided each is independently licensed in the syntactic position of the coordination. The LCL, they argue, is illusory: coordination does not require category identity, only that all conjuncts satisfy the external constraints of the position they occupy. Sometimes these constraints enforce categorial sameness; other times they are underspecified or disjunctive, yielding legitimate category mismatches.

2.2 Assessing Linguistic Knowledge in Language Models

The impressive performance of large language models (LMs) has spurred critical investigation into the scope of their linguistic knowledge. Their ability to generate fluent text does not necessarily imply that they possess the underlying knowledge that supports human language competence. In response, computational linguistics has developed a range of evaluation techniques and uncovered important insights. This section reviews the main methodologies, from early probing approaches to the now-standard practice of targeted behavioral evaluation with minimal pairs, and highlights several key findings.

2.2.1 Probing Internal Representations

Early efforts to understand the linguistic knowledge captured by large language models focused on their internal representations. The core question was whether the dense, contextualized vectors (or embeddings) generated by models like BERT implicitly encoded the abstract syntactic properties of language, even without explicit supervision on syntactic tasks. This led to the development of probing.

This group of approaches starts with training a simple "diagnostic classifier" on top of a model's frozen representations to predict a linguistic feature (Hupkes et al., 2018). The logic holds that if the simple probe succeeds, the information must be explicitly and accessibly encoded in the vectors rather than learned by the probe itself (Belinkov and Glass, 2019). While the methodology has caused debates about its interpretation, for instance, whether a successful probe truly indicates explicitly encoded information or simply exploits

a sufficiently expressive representation, it remains a foundational technique for analyzing model representations.

Probing studies have grown from simple diagnostic classifiers to all kinds of probes and successfully uncovered a hierarchy of linguistic knowledge. They have shown that contextualized representations of language models encode increasingly abstract layers of linguistic information (Conneau et al., 2018): from surface-level phenomena like word length or morphological features, through syntactic categories such as part-of-speech and dependency relations (Giulianelli et al., 2018; Liu et al., 2019; Şahin et al., 2020), to deeper semantic and relational structure, including long-range dependencies and world-knowledge facts (Klafka and Ettinger, 2020; Youssef et al., 2023). Across languages, probing tasks reveal that representations capture typology-sensitive properties and that simple classifiers can often recover these signals (Şahin et al., 2020). A landmark finding by Hewitt and Manning (2019) demonstrated that the dependency tree of a sentence could be recovered from the geometry of embedding spaces using simple linear transformations, which are what they called a structural probe, suggesting that models trained only on text prediction spontaneously learn an internal representation that systematically mirrors hierarchical syntax.

2.2.2 Targeted Syntactic Evaluations

Moving from the analysis of internal representations to that of external behavior, a complementary and highly influential methodology involves targeted linguistic evaluations. This approach addresses a key limitation of probing: while a successful probe demonstrates that syntactic information is present in a model's representations, it does not confirm that the model uses this information for its primary task of predicting text. To bridge this gap, targeted evaluations assess the model's functional grammatical knowledge by examining its output probabilities in highly controlled settings.

The cornerstone of this methodology is the minimal-pair paradigm, borrowed from experimental linguistics. Researchers create pairs of sentences that are identical except for a single, syntactically critical element, rendering one sentence grammatical and the other

ungrammatical. A consistent preference for the well-formed option across many such pairs is taken as strong evidence that the model has acquired the underlying abstract rule.

This paradigm, first widely popularized for LSTMs by Linzen et al. (2016), has become a standard for evaluating modern LMs. It has been scaled up from single phenomena to comprehensive benchmarks that test a wide array of complex syntactic structures, including argument structure, island constraints, and negative polarity item licensing (Marvin and Linzen, 2018; Warstadt et al., 2019; Futrell et al., 2019, i.a.). By directly measuring the model's performance on these targeted challenges, this approach provides a clear behavioral test of a model's linguistic capability.

2.3 Cognitive Science Paradigms for Model Investigation

A growing body of research seeks to evaluate language models through the lens of human cognition, employing paradigms and principles from cognitive science and psycholinguistics. This approach shifts the goal from simply determining if a model knows a rule to investigating how its learning processes and final knowledge state compare to those of humans. Treating LLMs as computational models of language acquisition and processing, this work employs methodologies of controlled experimental designs, often inspired by cognitive science, aiming to develop a richer and more nuanced account of the cognitive plausibility of LMs.

2.3.1 Controlled Experimental Design

Some researchers adopt the principles of controlled experimental design from psycholinguistics and experimental syntax to study LMs' behaviors. The fundamental goal of this approach is to isolate the specific linguistic variable under investigation, ensuring that a model's behavior can be attributed to its knowledge of a grammatical rule rather than to confounding factors such as lexical frequencies, semantic plausibility, or statistical artifacts in the training data. The minimal-pair paradigm, discussed previously, represents the simplest form of this experimental control.

Some research goes beyond curating test sets by manipulating the model's entire learning

environment (Warstadt and Bowman, 2022; Patil et al., 2024; Misra and Mahowald, 2024; Leong and Linzen, 2026, i.a.). One approach, termed Filtered Corpus Training (FiCT) by Patil et al. (2024), involves training a model from scratch on a dataset that has been intentionally and systematically altered. If a model trained on such an impoverished dataset can nevertheless succeed in testing items requiring structures absent from the training data, it provides powerful evidence that the model has acquired an abstract representation of the rule rather than simply memorizing surface patterns. This technique is particularly crucial for diagnosing whether models are learning robust grammatical knowledge or relying on superficial heuristics, a common failure mode identified by McCoy et al. (2019). At the same time, FiCT can be used to probe the minimal and sufficient conditions for acquiring a linguistic phenomenon. As Warstadt and Bowman (2022) emphasize, this approach enables precise and hypothesis-driven investigations into the mechanisms of language learning in neural networks.

2.3.2 Artificial Grammar Learning

Some researchers have adopted the Artificial Grammar Learning (AGL) paradigm to isolate the learning capacity of LMs from the statistical and semantic complexities of natural language. AGL provides a clean and controlled test bed for evaluating a model's ability to acquire novel structural patterns. In the classic AGL experiment pioneered by Reber (1967), human participants were first exposed to strings of symbols generated by a hidden formal grammar. After this training phase, they were tested on their ability to distinguish new, grammatically correct strings from ungrammatical ones. This paradigm translates directly to the evaluation of LMs. A model can be trained or fine-tuned on the grammatical strings of an artificial language and then tested on its ability to assign higher probabilities to novel grammatical strings than to ungrammatical ones. This methodology strips away all semantic and real-world knowledge confounds, allowing for a pure test of sequence learning and generalization.

Papadimitriou and Jurafsky (2023) applied a version of this paradigm to investigate inductive biases in model learning. They structurally biased models during pretraining, introducing specific formal patterns such as recursion or context-sensitive dependencies, and then assessed how these biases influence the model's ability to learn natural languages. They found that non-context-free structural biases provide stronger inductive advantages than purely recursive ones.

Similarly, Kallini and Fellbaum (2021) evaluated how transformer-based models learn artificial languages that are intentionally constructed to be "impossible" for humans, ranging from extreme cases, e.g., random word shuffles, to more plausible systems, e.g., count-based strings and attested languages. Their results indicate that models struggle with such "impossible" languages compared to natural or possible ones, suggesting that LMs' inductive biases may align in part with human linguistic plausibility. Later cross-linguistic work by Yang et al. (2025) shows that while GPT-2 can often distinguish natural languages from impossible ones, its ability to do so is imperfect, which indicates that some human-like inductive constraints are present but weaker.

Chapter 3
METHODOLOGY

This chapter discusses the methodology of this study. It consists of the following parts: an overview of the methodology underlying this research, the filters and data, the training setup and implementations, and the detailed evaluation design. All code and data can be found at https://github.com/Ljia1009/unlike-coordination-in-lm-fict.

3.1 Overview: Filtered Corpus Training

This research employs Filtered Corpus Training (FiCT) as its core methodology. As outlined in the literature review (§2.3.1), FiCT is a powerful paradigm that allows for precise, hypothesis-driven testing by systematically controlling the data a model learns from. Instead of merely evaluating pre-trained models, which have been exposed to a massive, uncontrolled deluge of text, FiCT enables us to create computational "poverty of the stimulus" experiments.

The central logic of this study involves training multiple models from scratch on two distinct types of corpora and evaluating them on curated test sets:

1. A baseline corpus, which is an original and unfiltered collection of text.

2. One or more filtered corpora, from which some text is removed according to some criteria. In this study, we remove occurrences of unlike coordination in two variants: all unlike coordination, and unlike coordination connected specifically by *and* (unlike conjunction).

This experimental design allows us to isolate the effect of exposure to unlike coordination on a model's ability to learn and generalize coordination structures. By comparing

the performance of models trained on these different datasets, we can investigate how language models acquire and process coordination and interpret these results through linguistic theories. If the models trained on the filtered corpus still succeed in generalizing to accept grammatical unlike coordination, it would indicate that unlike coordination is learnable from alike coordination for language models and support the hypothesis that they have acquired some structural rules akin to a "unified mechanism", which will be under further investigation. Conversely, if they fail on these test cases while the baseline model succeeds, then the immediate implication is unlike coordination is distinct from alike coordination and must be learned from such pattern itself, which aligns more with the theories without matching of conjuncts (§2.1.2) where such structures are not governed by any specific rules and should be learned as distinct and valid constructions.

3.2 Data

The training data are derived from the training corpus of the English corpus used and released by Gulordava et al. (2018), which was constructed from English Wikipedia dumps. We also use their vocabulary for the tokenizers for our models. To obtain filtered variants, we apply two filtering procedures to the base corpus. Filter I produces the corpus *filtered-all*, which excludes all sentences containing unlike coordination. Filter II produces the corpus *filtered-and*, which excludes all sentences containing unlike conjunctions.

To ensure comparability across datasets, we further downsample the base corpus and *filtered-and* to match the number of sentences in *filtered-all*, yielding three training corpora: *original*, *filtered-and*, and *filtered-all*. Downsampling was performed by randomly removing sentences in *original* and *filtered-and*. Each corpus consists of 1,948,104 sentences, though the total number of tokens varies. Using token counts measured with a GPT-2 tokenizer, *filtered-and* retains 91.02% of the tokens in *original*, while *filtered-all* retains 88.49%.

3.3 Filters

As mentioned briefly in §3.1, this study creates two filters targeting different types of unlike coordination. Filter I filters all sentences that have unlike coordination in the original corpus and filter II only filters out those with like conjunction. For example, consider these two sentences: a. *Is he available or at lunch*, b. *Sandy is a Republican and proud of it*. Filter I will remove both sentences while filter II will remove only sentence b. The goal of designing two filters is to also examine whether different levels of exposure to unlike coordination and whether the form of the coordinating structure plays a role in learning unlike coordination for the models.

The filters are built on top of the Berkeley Neural Parser (Kitaev and Klein, 2018), which is an off-the-shelf constituency parser integrated into spaCy using the same annotation standard as the English Web Treebank and OntoNotes. Specifically, the filters utilize the model benepar_en3 of Berkeley Neural Parser, which has 95.40 F1 score on revised WSJ test set. This neural parser is responsible for parsing each sentence in the training corpus into a constituency parse tree, from which coordination structures are automatically identified based on the presence of the CC (coordinating conjunction) label. The filters then inspect the syntactic categories of the conjuncts, by checking the parsing labels at the same level as the CC node, to determine whether they are alike or unlike coordination. Sentences containing unlike coordination are removed according to the criteria defined by each filter type: filter I simply removes them, and filter II removes them if the node labeled CC is the word *and*. If they are alike coordination, the filters then recursively do the previous process until reaching the end node. The filters also handle a special case involving phrases headed by *be*-verbs. When the parser identifies a phrase as being headed by a form of *be*, the filter determines its grammatical category based on the head of its complement. This excludes the leakage of a group of unlike coordination sentences. One example is that *the task is difficult and requires patience* could be otherwise unfiltered, because benepar_en3 labels *is difficult* as a verb phrase (VP). While this labeling is consistent with certain syntactic conventions,

it does not align with the objective of this study, which treats such structures as instances of unlike coordination and thus excludes them from the filtered corpus.

The filtering component is fundamental to this study, as all subsequent analyses rely on the filtered corpora. Ideally, the filter would perfectly identify and retain all sentences containing unlike coordination while leaving all other sentences untouched. However, since no established benchmark currently exists for evaluating such filters on this topic, we assess their performance using our own evaluation sets to determine how effectively they perform this task. The evaluation sets for the filter performance contain a set of 100 sentences of unlike coordination (*unlike*) and a set of the corresponding 100 sentences of alike coordination (*alike-from-unlike*). Details of the two evaluation sets are shown in Section 3.5. The filters should filter out as many targeted to-be filtered sentences as possible for *unlike*, and keep as many as possible for *alike*. We report the results in Table 3.1.

Filters	*U-All*	*U-Wrong*	*A-All*	*A-Kept*
Filter I	100	1	100	83
Filter II	100	1	100	83

Table 3.1: Filtering results on our evaluation sets. U = Unlike set, A = Alike set; *All* = total sentences, *Wrong* = wrongly retained, *Kept* = correctly kept.

As suggested by the evaluation results, these two filters are strong filters. Each filter wrongly retains only a single sentence in *unlike* due to parsing errors, yet also filters away some sentences in *alike*, which ideally should remain. Closer investigation reveals that these wrongly filtered sentences are because of parsing error, the complex case of *be*, and the strict implementation of the filters, which discard conjoined phrases that are not labeled identically, even if they might actually belong to the same category by human judgment.

This kind of strong filters may be, in fact, favorable because their implementation is designed to remove as many instances of unlike coordination as possible. In principle, such

aggressive filtering could be problematic: if too many false positives, i.e., sentences that are actually alike coordination, are filtered, the representativeness of the data could be reduced and potentially degrade the performance of models, making the evaluation results less convincing. However, our results suggest that this is not the case. On the other hand, the filters demonstrate high precision, effectively eliminating undesired unlike coordination. Consequently, these filters produce high-quality filtered corpora, providing a solid foundation for subsequent model training, evaluation, and further analysis.

During filtering, we also record the distribution of conjoined phrase labels in unlike coordination to provide statistics analogous to those reported by Kallini and Fellbaum (2021), serving as a side result of this thesis. These statistics also confirm that our filters are strong filters. For example, Filter I identifies 15,380 occurrences of the conjoined phrase combination NML & NP as the ninth most frequent, where NML stands for nominal modifier phrase, This combination likely includes many instances of alike coordination, since the phrases tagged as NML can actually be just noun phrases. Table 3.2 and Table 3.3 report the ten most frequent combinations identified by each filter, excluding those with a high probability of representing actual alike coordination (e.g., NML & NP). This allows readers to compare our findings with the COCA and PTB statistics reported by Kallini and Fellbaum (2021). Readers may refer to Appendix C for their results.

3.4 Model Training

3.4.1 Architecture

This study uses the Transformer (Vaswani et al., 2017) as the model architecture. Specifically, we employ a causal decoder-only Transformer following the General Purpose Transformer (GPT) framework introduced by Radford et al. (2019). The model configuration is comparable to gpt-2-large, with a maximum sequence length of 1024, an embedding dimension of 1280, 36 layers, and 20 attention heads. This model configuration contains approximately 774 million trainable parameters. Combining this architecture with the three distinct train-

Conjoined phrases	# of occurrences	% of identified unlike coordination
ADVP & VP	67,716	19.41%
NP & VP	48,766	13.98%
ADVP & NP	25,162	7.21%
ADJP & VP	23,274	6.67%
ADJP & NP	22,844	6.55%
NP & PP	21,121	6.05%
NP & S	20,577	5.90%
S & VP	18,722	5.37%
PP & VP	15,728	4.51%
ADVP & PP	12,908	3.70%

Table 3.2: Ten most frequent combinations of conjoined phrases for unlike coordination in training corpus by filter I, after adjustment.

ing corpora yields three separate models, each trained on a different corpus. We implemented these models using the Huggingface `transformers` package (Wolf et al., 2020).

3.4.2 Training

We trained three models variants: *original*, *filtered-and*, and *filtered-all*, each on its corresponding training corpus of the same name. To account for variability introduced by initialization, we trained multiple models with different random seeds for each model variant. We used a random number generator to generate three random numbers: 291, 1543, and 9071, which serve as the random seeds. So, in total we trained 9 models, 3 models of different random seeds for each variant.

All models were trained using the same strategy. We employed the `Trainer` class from

Conjoined phrases	# of occurrences	% of identified unlike conjunction
ADVP & VP	54,841	20.69%
NP & VP	42,987	16.21%
ADJP & VP	18,915	7.13%
NP & S	17,717	6.68%
ADVP & NP	16,766	6.32%
NP & PP	16,393	6.18%
S & VP	14,502	5.47%
PP & VP	12,823	4.84%
ADVP & PP	10,277	3.88%
ADJP & NP	7,903	2.98%

Table 3.3: Ten most frequent combinations of conjoined phrases for unlike conjunction in training corpus by filter II, after adjustment.

the `transformers` package. Training of each model was conducted on one NVIDIA L40 GPU for up to 10 epochs, using the default AdamW optimizer (Loshchilov and Hutter, 2017) with mixed-precision. The effective batch size was set to 16, the initial learning rate to 5e-5, and the weight decay to 0.01. A complete list of training parameters is provided in Appendix A.

Model evaluation was performed every 1,000 training steps. We applied early stopping based on evaluation loss, terminating training when performance failed to improve (threshold for improvement was set to 0, so any decrease in evaluation loss counts as improvement) for three consecutive evaluations and selecting the best-performing checkpoint. This resulted in nine individual models: *original-291*, *original-1543*, *original-9071*, *filtered-and-291*, *filtered-and-1543*, *filtered-and-9071*, *filtered-all-291*, *filtered-all-1543*, and *filtered-all-9071*, which were saved at 332,000, 218,000, 265,000, 252,000, 248,000, 256,000, 248,000,

239,000, and 239,000 steps, respectively.

3.5 Evaluation

3.5.1 Evaluation Corpora

We manually constructed five evaluation corpora for a general investigation of coordination: *unlike*, *alike-from-unlike*, *supercat-deletion*, *deletion*, and *unlike-gj*. Each corpus targets a specific aspect of model performance and analysis related to unlike coordination.

- *Unlike*: This evaluation corpus contains 100 sentences that have unlike coordination. Following Kallini and Fellbaum (2021) findings on the ten most frequent unlike coordination patterns and their relative frequencies, the corpus includes instances of the following patterns: NP + SBAR, NP + VP, AP + VP, AdvP + PP, NP + AP, PP + VP, PP + AdvP, NP + PP, PP + NP, and VP + NP. The number of sentences for each pattern was adjusted to approximate the relative frequencies reported by them, resulting in the following distribution: 18, 15, 11, 10, 10, 9, 9, 7, 6, and 5 sentences, respectively. Table 3.4 presents selected examples from this evaluation corpus.

- *Alike-from-unlike*: This evaluation corpus consists of 100 sentences featuring alike coordination, created on the basis of the sentences in *unlike*. Each sentence was adapted from a corresponding sentence in *unlike* with only minor modifications to convert an unlike coordination into an alike one, while maintaining most of its meaning, ensuring comparability between the two sets. This evaluation corpus can thus be viewed as a loose minimal-pair counterpart to *unlike*. For example, *You can depend on my assistant and that he will be on time* → *You can depend on my assistant and his ability to be on time*. Additional examples are shown in Table 3.5.

- *Supercat-deletion*: This evaluation corpus consists of sentences from *unlike* that show clear evidence of presupposing a supercategory. These sentences can also be interpreted as involving the deletion of shared elements across conjuncts. For example, in

Pattern	Sentence
NP+SBAR	You can depend on my assistant and that he will be on time .
NP+VP	Voids are a nightmare and initialed by the employees .
AP+VP	That was manipulative and designed to keep her in the dark .
AdvP+PP	The phenomenon fall in to place organically and with ease .
NP+AP	My boss is a perfectionist and demanding .
PP+VP	Those leaders insist that they are still in charge and leading the people .
PP+AdvP	I don't know if I should take this exam in class or online .
NP+PP	I punched him a lot of times and with all my might .
PP+NP	More Americans work out of the house and longer hours .
VP+NP	Erosion has left the brick structures crumbling and a clear safety hazard .

Table 3.4: Examples of sentences in *unlike*.

The phenomenon fall into place organically and with ease, we can posit a shared supercategory of modifier phrases spanning both conjuncts. Alternatively, this sentence may be viewed as derived from the deletion of the repeated phrase *the phenomenon fall into place* at the beginning of the second conjunct. In total, the corpus contains 62 sentences.

- *Deletion*: This evaluation corpus consists of sentences from *unlike* that can be interpreted as involving deletion. These sentences are incompatible with the same-supercategory hypothesis. For example, *The movie was funny and made everyone laugh* can be interpreted as involving deletion of the shared subject *the movie* from the beginning of the second conjunct; however, it is difficult to identify a convincing supercategory that accounts for this coordination. In total, this evaluation corpus contains 24 sentences.

  It is worth noting that the combined number of sentences in this corpus and *supercat-*

Sentence in *Alike-from-unlike*
You can depend on my assistant and his ability to be on time .
Voids are a nightmare and a misery, which is initialed by the employees .
That was manipulative and was designed to keep her in the dark .
The phenomenon fall in to place organically and easily .
My boss is a perfectionist and a demanding person .
Those leaders insist that they are still in charge and are leading the people .
I don't know if I should take this exam in class or on Canvas .
I punched him repeatedly and forcefully .
More Americans work out of the house and in longer hours .
Erosion has left the brick structures crumbling and dangerous .

Table 3.5: Examples of sentences in *alike-from-unlike*. Sentence in each row is a counterpart to sentence in each row in Table 3.4

deletion does not exactly match the number of sentences in *unlike*. This is because some sentences in *unlike* do not clearly belong to either category. For instance, the sentence *You can depend on my assistant and that he will be on time* does not fit into *deletion*, since *you can depend on that ...* is ungrammatical. One might posit a supercategory unifying an NP and a CP that can function as arguments, or assume that the CP functions as an NP; however, the second conjunct still violates the selectional requirements of *depend on*, and further explanation is needed. Because of such complications, this sentence also does not qualify for inclusion in *supercat-deletion*, which only contains sentences that provide clear evidence for a convincing supercategory.

- *Unlike-judgment*: This evaluation corpus consists of 22 grammaticality judgment tests related to unlike coordination. The sentences were collected from relevant literature on coordination (Bruening and Al Khalaf, 2020; Patejuk and Przepiórkowski, 2023; Sag

et al., 1985; Zhang, 2009; Peterson, 1981; Grosu, 1985) and adapted for this study and context. In each test, when multiple sentence variants are included, they are designed to be as similar to one another as possible. A model with adequate knowledge of unlike coordination should be able to correctly classify the grammaticality of these sentences. These tests also allow us to inspect the model's knowledge of unlike coordination. Note that 2 tests contain only grammatical sentences; in such cases, we focus on which sentence the model prefers rather than whether they can reject ungrammatical ones. Table 3.6 shows some selected tests in this corpus.

3.5.2 Methods and Metrics

Different methods and metrics were employed to evaluate the performance of the models and to analyze their understanding of unlike coordination. These include *perplexity*, *accuracy*, *average attention score*, *similarity*, and *clustering analysis*.

- *Perplexity*: This is used to evaluate the performance of models in the evaluation corpora excluding *unlike-judgment*. Given a sequence of tokens, perplexity is defined as

$$\mathrm{PPL}(X) = \exp\left\{-\frac{1}{t}\sum_{i=1}^{t}\log p_\theta(x_i \mid x_{<i})\right\}$$

  where t is the number of tokens in the sequence, and $p_\theta(x_i \mid x_{<i})$ is the model probability of token x_i given its context. Then, the mean perplexity across M samples is defined as

$$\overline{\mathrm{PPL}} = \frac{1}{M}\sum_{j=1}^{M}\mathrm{PPL}_j$$

  In this case, M will be the total number of sentences in each evaluation corpus.

- *Average Surprisal*: Average surprisal is used to determine if the models prefer the grammatical sentences over the ungrammatical sentences in tests in *unlike-judgment*. Surprisal is the log-space version of perplexity, and average surprisal is defined as

$$\overline{S}(w_{1:N}) = -\frac{1}{N}\sum_{i=1}^{N}\log p(w_i \mid w_{<i})$$

where $\overline{S}$ is average surprisal of a sentence, N is the number of tokens in the sentence, and $p(w_i \mid w_{<i})$ is the conditional probability of token w_i assigned by the model.

- *Accuracy*: Accuracy of passing the grammaticality judgment tests. Passing a test means the model assigns a higher average surprisal to the ungrammatical sentence than to the grammatical one. The sample size of computing accuracy is 20, which is the number of relevant grammaticality judgment tests.

- *Average Attention Score*: Average attention score is defined as

$$\bar{A}(q, K) = \frac{1}{L \cdot H \cdot |K|} \sum_{\ell=1}^{L} \sum_{h=1}^{H} \sum_{k \in K} A_{q,k}^{(\ell,h)}$$

where $\bar{A}(q, K)$ denotes the average attention from a query token q to a key phrase K, L is the number of Transformer layers considered, H is the number of attention heads, $|K|$ is the number of tokens in the key phrase, and $A_{q,k}^{(\ell,h)}$ represents the attention weight from query token q to key token k in layer ℓ and head h. In this study, the score is computed by averaging attention values from the middle layer to the final layer, so as to capture all relevant contextual information involved in processing and forming the representation of the query token.

Averaging attention score provides a summary signal to measure the overall degree to which the model allocates attention from the query token to the key phrase across layers, heads, and phrase tokens. Another common method in studying attention is to use the maximum, which captures the strongest individual attention link, but also can be sensitive to spikes that might arise from irrelevant reasons. Since this study is concerned with the general attention patterns rather than identifying specialized heads, the average is the more appropriate summary statistic characterizing the overall attention allocated to the phrase of interest.

For the sentences in *deletion* and an equal number of sentences in *supercat-deletion*, the average attention score is used to investigate how the models process these sentences.

Given a query word, specifically, the first word in the second conjunct of the sentences in *deletion* and *supercat-deletion*, this score can be interpreted as

1. The model has a more supercategory-based interpretation of generating the coordination if the average attention score from the query to the first conjunct is greater than the score to the phrase that occupies a potential deletion position.
2. The model has a more deletion-based interpretation of generating the coordination if the opposite is true (excluding cases of equality).
3. Undefined if an equality appears.

For example, in the sentence *She mentioned the project and that it was already approved*, the query word is *that*, and the two key phrases of interest are the first conjunct *the project* and the phrase occupying the potential deletion position *She mentioned*. We compare the average attention scores from the query word to the two key phrases to determine whether the coordination is better explained by a supercategory-based analysis or by a deletion-based analysis.

- *Cosine Similarity*: In this study the cosine similarity is computed as

$$\mathrm{Sim}(C_1, C_2) = \frac{h_{C_1}^{(L)} \cdot h_{C_2}^{(L)}}{\|h_{C_1}^{(L)}\| \, \|h_{C_2}^{(L)}\|}$$

where $\mathrm{Sim}(C_1, C_2)$ denotes the cosine similarity between two conjuncts C_1 and C_2, and $h^{(L)}C_1$ and $h^{(L)}C_2$ are the averaged token embeddings of the two conjuncts, obtained from the last Transformer layer indexed by L. For sentences in *deletion* and an equal number of sentences in *supercat-deletion* and *alike-from-unlike*, the cosine similarity score is computed between the conjuncts of each sentence. They contrast with each other such as the deletion-based analysis should indicate the lowest similarity between conjuncts among the three for sentences in *deletion*, and the supercategory-based account predicts increased similarity between conjuncts for sentences in *supercat-deletion*, and *alike-from-unlike* condition serves as a control in which conjunct similarity should

be the highest. For each corpus, we then calculate the mean cosine similarity across all sentences. This measure reflects whether the model tends to represent phrases similarly when they appear in coordination.

- *Clustering*: The last-layer representations of the conjuncts are also used for a clustering analysis employing the k-means algorithm with $k = 3$. For each model variant, we extract its last-layer representations of conjunct phrases in a sentence in a corpus, and average the representations if needed, then use these vectors as input to train the classifier. For each of the three model variants, and each corpus in *deletion*, an equal number of sentences in *supercat-deletion*, and an equal number of sentences in *alike-from-unlike*, we train a clustering system in this way. This results in a total of $3 \times 3 \times 3$ clustering systems (for each model variant there are three models, each trained on the three corpora described above, and three model variants in total). The clustering results may reflect the representations of conjuncts.

Category	Test
Selection Violation	You can depend on my assistant and that he will be on time .
	* You can depend on that he will be on time .
Position Effect	You can depend on my assistant and that he will be on time .
	* You can depend on that he will be on time and my assistant .
CP as NP requirement	That he will be on time, you can depend on .
	* You can depend on that he will be on time .
Mismatch Restriction	Hobbs ended up liking Rhodes and hating Barnes .
	* Hobbs ended up liking Rhodes and to hate Barnes .
Mismatch Restriction	Danny became a political radical and very antisocial .
	* Danny became a political radical and under suspicion .
Dual Roles	St. Peter did reside in Rome .
	St. Peter did die in Rome .
	St. Peter did reside and die in Rome .
Unlike Supercategories	Reluctantly and in embarrassment, the white officer released the Black man .
	Reluctantly and embarrassed, the white officer released the Black man .
Deletion	I eat meat and at home .
	* Meat and at home do I eat .

Table 3.6: Examples of tests in *unlike-judgment*.

Chapter 4

RESULTS

4.1 General Performance

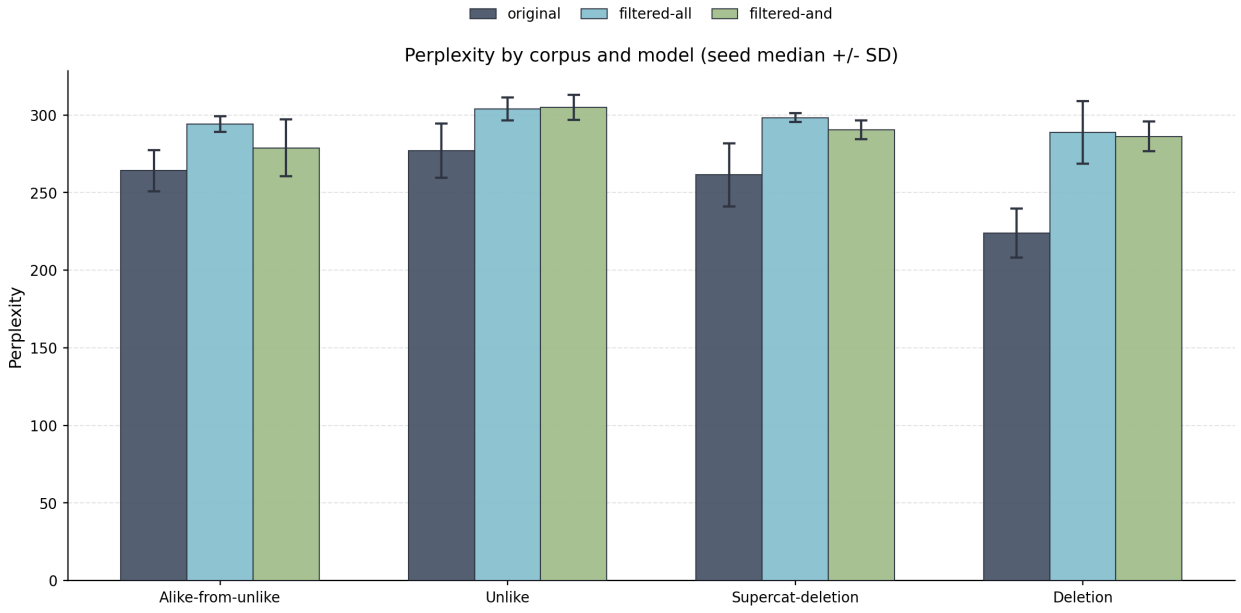

Figure 4.1: Mean perplexity of models on the evaluation corpora, excluding *unlike-judgment*.

Figure 4.1 summarizes the mean perplexity of each model across the evaluation corpora. Two observations emerge. First, the three model variants do not show a clear preference for either alike or unlike coordination. This comparison is valid because the *alike-from-unlike* corpus was constructed to be directly comparable to the *unlike* corpus. Second, the *original* model performs slightly better than the filtered models on unlike coordination, but shows less advantage on alike coordination. The perplexity results further suggest a performance ordering of *original* > *filtered-and* >= *filtered-all*, while also indicating that the numerical

advantage of the *original* model over the filtered models is relatively modest.

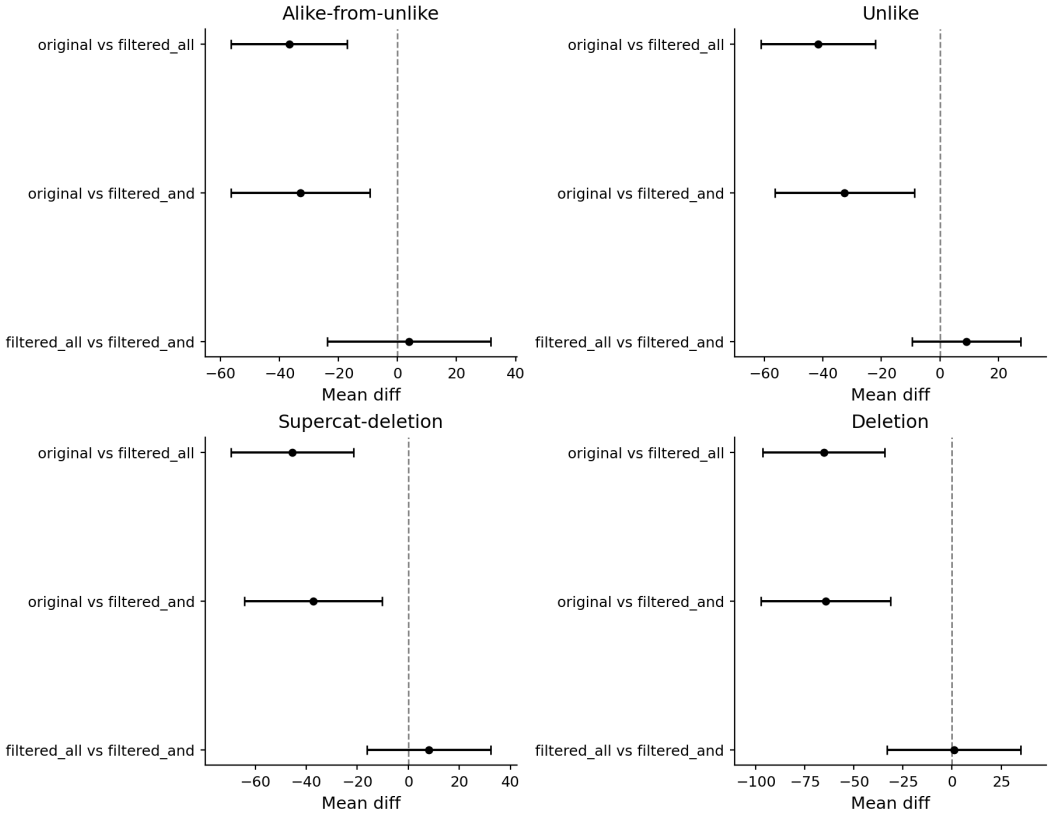

Figure 4.2: Forest plots from pairwise t-tests comparing three model variants on each evaluation corpus. The x-axis shows the mean difference, computed as the average raw perplexity difference across the same sentences. Each dot with its horizontal line represents the 95% confidence interval of the difference. The dotted vertical line marks the line of no effect.

In Figure 4.2, we present the forest plot of pairwise t-test results between all model variants on each evaluation corpus. The *original* model has significantly better performance than both the *filtered-all* and *filtered-and* models. In contrast, the line of no effect intersects the comparison between the two filtered models across all corpora, indicating no statistically significant difference in their performance, although the *filtered-and* model shows a slight numerical advantage.

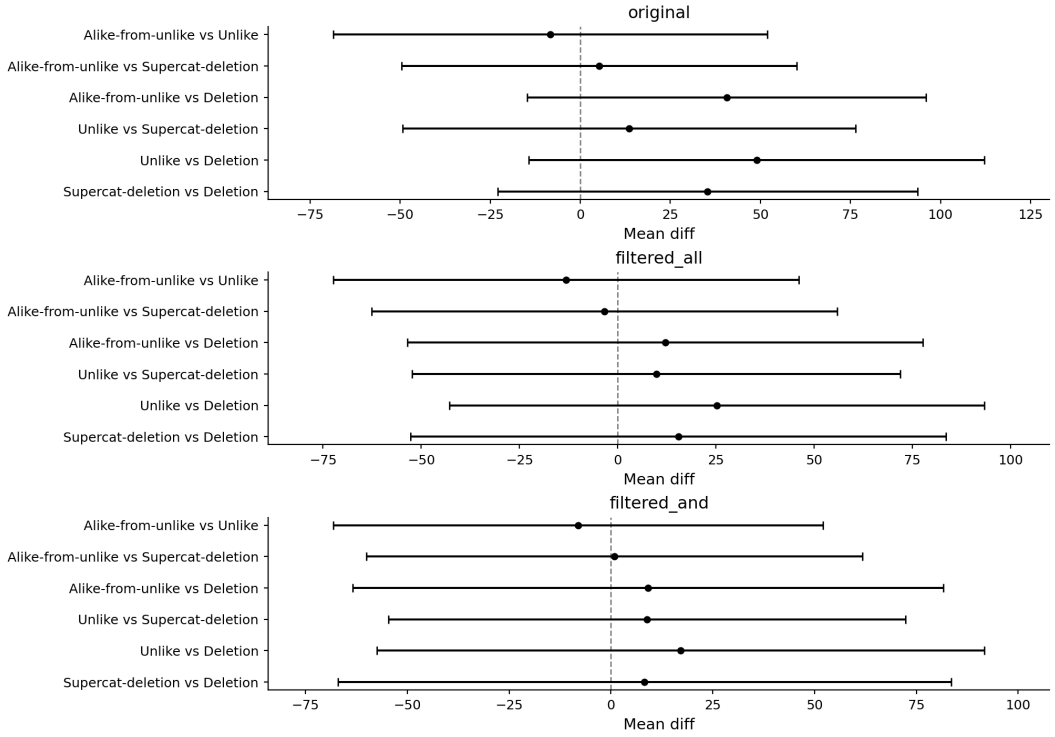

Figure 4.3: Forest plots from Welch's t-tests comparing each model on each pair of evaluation corpora. The x-axis shows the mean difference, computed as the average raw perplexity difference across the same sentences. Each dot with its horizontal line represents the 95% confidence interval of the difference. The dotted vertical line marks the line of no effect.

We further conduct a within-model analysis and present the forest plot of Welch's t-test results for each model across all pairs of evaluation corpora in Figure 4.3. For all three model variants, the line of no effect intersects every pairwise comparison, which indicates that each model responds similarly to alike and unlike coordination, since they show no significant perplexity difference attributable to coordination type. In other words, once a model is fixed, its perplexity remains stable across coordination types, suggesting that any observed differences in the between-model analysis stem from model variation rather than inherent sensitivity to coordination type within a model.

4.2 Grammaticality Judgment

In this section, we turn to how the models perform on the grammaticality judgment tests. As explained in Section 3.5, this is determined by whether the models prefer the grammatical sentence over the ungrammatical one, which is evaluated by comparing the average surprisal the models assign to the sentences. Table 4.1 reports the summary statistics for each model variant, and Appendix B presents the per-test results. These grammaticality tests provide a more diagnostic perspective on model behavior, complementary to the previous perplexity-based evaluation.

	Original	Filtered-all	Filtered-and
Median Row Acc.	0.75	0.70	0.70
Mean Acc. (± SD)	0.70 ± 0.04	0.67 ± 0.02	0.65 ± 0.08
Seed Corr.	0.71	0.79	0.78

Table 4.1: Summary statistics for the grammaticality judgment tests. The second row reports the median row accuracy, computed by comparing the median average surprisal assigned by a model to grammatical and ungrammatical sentences. The third row shows the mean accuracy across different random seeds, together with the corresponding standard deviation. The final row reports the mean Pearson correlation across random seeds. All values are rounded to two decimal places. The total number of tests involved is 20.

As shown in Table 4.1, the models demonstrate a moderate ability to capture the grammatical constraints on unlike coordination, although their performance remains far from perfect. Both the median row accuracy and the mean accuracy indicate that the *original* model achieves the strongest grammaticality judgments, while the two filtered variants show broadly comparable performance. The standard deviation further reveals a systematic effect of filtering on the stability of learned representations. The *filtered-all* variant, which receives

no exposure to unlike coordination patterns, shows the highest stability, suggesting that the absence of potentially conflicting evidence leads the model to adopt more uniform, but not necessarily accurate, generalizations. In contrast, the *original* models, trained on fully natural data, exhibit moderate variability, consistent with the fact that they learn multiple competing structural patterns present in the input. The *filtered-and* variant, which receives partial exposure, is the least stable, indicating that selective removal of coordination structures may introduce inconsistencies in the statistical learning signal, resulting in less coherent internal representations. Overall, the high mean seed correlation indicates substantial consistency between random seeds, supporting the use of the median row accuracy as a reliable summary measure.

Zooming in on the individual tests, where we use the median row accuracy to decide if a variant in general passes a test, clear success and error patterns emerge. All three models perform reliably on verb dependency (Test 5) and on most deletion possibility contrasts (Tests 13, 19, 20, 21). They correctly identify that the unlike coordination is sensitive to the verb: *...depend on my assistant and that he will be on time* is grammatical, but *...strengthens the claim and that he will be on time* is infelicitous. They also recognize that certain unlike coordination constructions involve deletion, so *She eats beans and with her hands* is acceptable, and they correctly dislike the counterpart that cancels out the deletion, as in *Beans and with her hands, she eats.* Although this example alone may be caused by the models' general aversion to topicalization, this pattern is consistent in many other cases. For example, the models recognize *He eats only fish raw and only at home*, while rejecting the counterpart *He eats only fish and only at home raw.* The models also do well in detecting that the conjunct position can affect grammaticality. In this category, the *original* variant succeeds in every test, and the two filtered variants succeed in 3 out of 4 tests.

The shared error pattern is that the models consistently fail on two categories: CP-as-NP requirement (Test 3) and several category mismatch restricted cases (Tests 6, 10, 11). These are not borderline reversals. In many cases, margins are large and directionally wrong. For test 3, the error could be introduced by topicalization, and the model prefers the wrong

sentence *You can depend on that he will be on time* instead of the correct sentence *That he will be on time, you can depend on*. Because in test 4, all the models correctly identify *That he will be on time, you can depend on* as the correct one instead of *He will be on time, you can depend on*, which suggests the models have at least partial grasp of CP functioning as an NP. In the category mismatch restricted cases, the models' failure appears more substantive. For example, they do not prefer *Hobbs ended up liking Rhodes and hating Barnes* over *Hobbs ended up liking Rhodes and to hate Barnes*, indicating genuine difficulty with constraints on category mismatch. Linguistically, these items require sensitivity to fine-grained category licensing and compatibility, and this is where the models fall short.

Model-specific weaknesses vary with filtering conditions. *Filtered-all* models uniquely fail on Test 7 (conjunct position) and Test 12 (mismatch topicalization): they do not prefer the grammatical sentence *the once and future king* over the position-swapped infelicitous sentence *the future and once king*, and they do not identify that category mismatched constituent cannot undergo topicalization, as in *Bill and with Sarah, she finally met*. *Filtered-and* models uniquely fail on Test 1 (selection violation) and Test 8 (conjunct position). They fail to detect the selectional violation in *you can depend on that...*, incorrectly treating it as preferable to the grammatical one *...depend on my assistant and that...*, and also fail to recognize the ungrammaticality of *Sandy is proud of it and a Republican* as opposed to the position-swapped counterpart *Sandy is a Republican and proud of it*. *Original* models uniquely fail on Test 22 (deletion possibility), while both filtered variants succeed: they fail to recognize that *John sang beautifully and a carol* is ungrammatical as opposed to *John sang a carol and beautifully*, which can be interpreted as that they allow a more permissive deletion process in their internal representations.

Two tests in *unlike-judgment* do not include an ungrammatical sentence. Their purpose is to assess how the models respond to challenging constructions associated with unlike coordination. These two tests can be found in Table 3.6. Specifically, one test introduces a dual-role instance of the PP *in Rome*, which can function either as a modifier of *die*, and thus as a Modifier Phrase, or as an argument of *reside*, in which case it is an ordinary

PP and does not fall under the supercategory of modifiers. The other test presents a case in which unlike supercategories can nonetheless be conjoined: in the first sentence, the two conjuncts share the same supercategory (both are Manner Phrases), whereas in the second sentence the conjuncts differ, with *reluctantly* functioning as a Manner Phrase and *embarrassed* functioning as a predicate.

For the dual role test, both the filtered variants show a similar pattern, consistently assigning the worst score by a large margin to the third sentence having a dual-role PP *Peter did reside and die in Rome*, effectively treating it as ungrammatical. In contrast, the *original* variant assigns comparable scores to the three sentences: *Peter did reside in Rome*, *Peter did die in Rome*, and *Peter did reside and die in Rome*, suggesting that the *original* variant tolerates a broader range of unlike coordination patterns and, in this case, it is correct.

In the unlike supercategory test, all model variants exhibit a similar pattern, strongly disfavoring the second sentence, *Reluctantly and embarrassed...*, relative to the first one, *Reluctantly and in embarrassment....* In other words, the models uniformly prefer cases in which the conjoined phrases can plausibly be analyzed as sharing a higher-level supercategory, while rejecting constructions that involve coordination across genuinely mismatched supercategories.

Stepping back from the individual tests, these grammaticality judgment tests provide a unique perspective on each model variant's behavior and internal representations of unlike coordination, as well as their similarities and biases. All three variants have moderate success in these tests, which agrees with the understanding of the previous perplexity results, showing that they encode several grammatical contrasts robustly for unlike coordination. Across conditions, the models perform well on constraints that operate primarily at the surface-syntactic level, including verb dependency, deletion, and conjunct position effects. In contrast, the models show difficulty with cases in which category mismatch is governed by more fine-grained constraints. The models exhibit a bias of processing unlike coordination the same as alike coordination. The *original* variant, trained on unfiltered natural data, can

be even more generous.

Comparison of the error and success patterns across variants further suggests that all models develop a shared tendency toward organizing coordinated elements according to broad supercategorical similarities, which enables them to process many instances of unlike coordination. This tendency is supported positively by the models' success on cases in which unlike conjuncts share a plausible supercategory, and negatively by their failure on more nuanced constraints where such constituents nevertheless cannot be conjoined. Moreover, in a test for unlike supercategory, all model variants exhibit a similar pattern, strongly disfavoring *Reluctantly and embarrassed...*, relative to *Reluctantly and in embarrassment...*. In other words, the models uniformly prefer cases in which the conjoined phrases can plausibly be analyzed as sharing a higher-level supercategory, while rejecting constructions that involve coordination across genuinely mismatched supercategories. As filtering decreases, the models appear to have an increasing capability to license deletion-based analyses. While this flexibility improves performance in some cases, it can also lead to overly permissive generalizations.

4.3 Internal Representations

In this section, we present the results of metrics for investigating the internal representations and processing of unlike coordination.

4.3.1 Attention Score

Average attention scores are used to examine how the models process sentences of unlike coordination in *deletion* and an equal number of sentences in *supercat-deletion*. Readers may refer to Section 3.5.2 for a detailed description of this metric and its application. Table 4.2 provides a concise summary of the results. Overall, the models exhibit different attention patterns for *deletion* versus *supercat-deletion* sentences. Models exposed to unlike coordination show a larger distinction in how they process deletion-type versus supercategory-type unlike coordination, whereas the filtered-all model exhibits a smaller difference.

Model	*Deletion*		*Supercat-deletion*	
	deletion-based	supercat-based	deletion-based	supercat-based
original	15.00 ± 1.41	9.00 ± 1.41	10.00 ± 0.47	14.00 ± 0.47
filtered-all	12.00 ± 1.25	12.00 ± 1.25	10.00 ± 2.45	14.00 ± 2.45
filtered-and	14.00 ± 0.82	10.00 ± 0.82	10.00 ± 1.25	14.00 ± 1.25

Table 4.2: Overall results from the attention scores of a query word to the two key phrases on the two evaluation corpora. Numbers are median of seeds with the standard deviation. The numbers indicate how many sentences are classified under each processing type based on attention-score interpretation. Deletion-based means the attention score to the phrase occupying a potential deletion position exceeds that of the first conjunct; supercat-based means the opposite.

We conducted within-model Welch's t-tests for each of the three models and between-model paired t-tests for every pair of variants on each evaluation corpus. For each sentence, we computed the difference between the attention scores assigned to the two key phrases. Figure 4.4 presents the within-model comparisons, while Figure 4.5 reports the between-model results.

The within-model results shown in Figure 4.4 clearly support the observed tendency: all three model variants exhibit differences in their attention patterns between *deletion* and *supercat-deletion*. For the *original* and *filtered-and* variants, this within-model difference is statistically significant, whereas no significant difference is observed for the *filtered-all* variant.

The between-model comparisons in Figure 4.5 further indicate that the models do not differ significantly in their processing of *supercat-deletion* sentences. In contrast, significant differences emerge in the *deletion* condition. The contrast between *filtered-all* and *filtered-and*, which represents no exposure versus partial exposure, is nearly significant, while the *original* model differs significantly from both filtered variants.

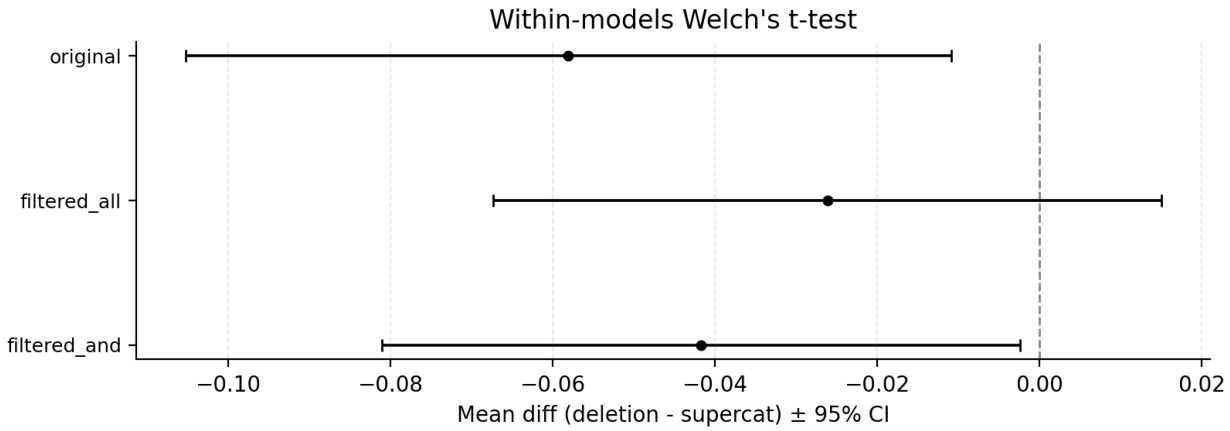

Figure 4.4: Within-model Welch's t-test results. The x-axis reports the mean difference between attention scores for the same sentence, and the y-axis lists the three models. Horizontal lines denote 95% confidence intervals, with dots marking the mean estimates. The vertical dotted line indicates the line of no effect.

4.3.2 Similarity

Using the same two corpora described in Section 4.3.1, along with an equal number of sentences from *alike-from-unlike*, we compute cosine similarity to measure the similarity between the conjoined phrases in each sentence for each corpus. For details of the computation, readers may refer to Section 3.5.2.

Table 4.3 reports the average similarity between conjuncts, along with the standard deviation, for each model across the evaluation corpora. The similarity scores clearly indicate that the models treat the three corpora differently: conjuncts in *alike* show the highest similarity, those in *supercat-deletion* exhibit moderate similarity, and those in *deletion* show the lowest similarity.

Table 4.4 presents the results of a linear mixed-effects model that predicts similarity scores from two categorical fixed-effect predictors, model and corpus, as well as their interaction. The model predictor had three levels (*original*,*filtered-all*, and *filtered-and*), and the corpus predictor had three levels (*supercat-deletion*, *alike-from-supercat*, and *deletion*). In

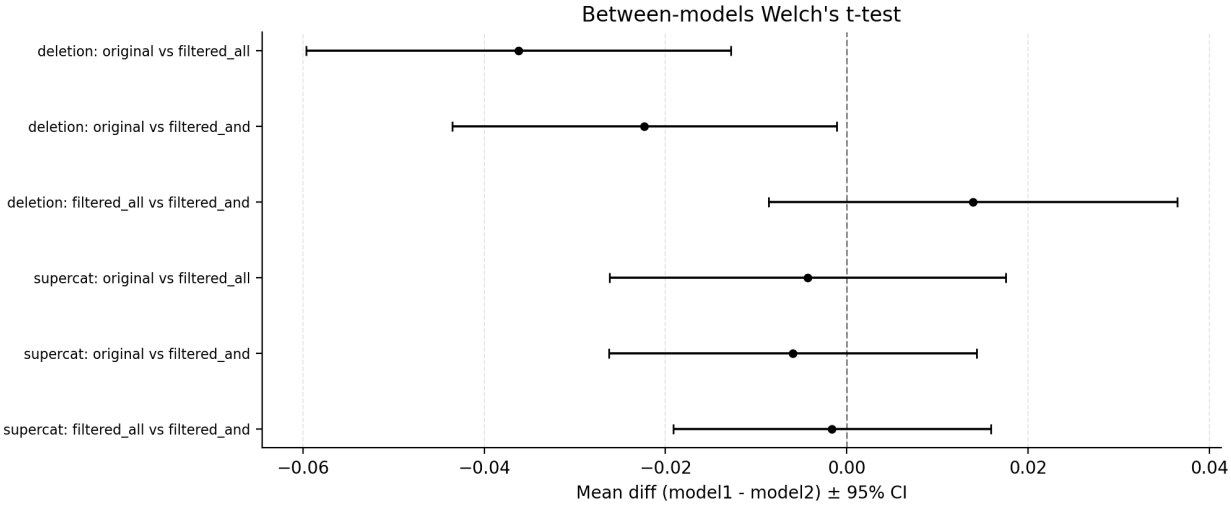

Figure 4.5: Between-model paired t-test results. The x-axis reports the mean difference between attention scores for the same sentence, and the y-axis lists the three model comparisons along with the evaluation corpus. Horizontal lines denote 95% confidence intervals, with dots marking the mean estimates. The vertical dotted line indicates the line of no effect.

this parameterization, *original* and *supercat-deletion* were used as the reference levels, so the intercept represents the predicted similarity score for *original* on the *supercat-deletion* corpus. The corpus-effect results show more evidence of the observed patterns: relative to *supercat*, the alike corpus produced significantly higher similarity scores ($b = 0.173$, $p < .001$), whereas the *deletion* corpus produced significantly lower scores ($b = -0.084$, $p < .001$). The interaction results further suggest a pattern that aligns with what is observed by the attention patterns: neither model shows a significant interaction with the *alike* corpus, suggesting that the increase in similarity for *alike* was stable across conditions. However, *filtered-all* shows a significant positive interaction with deletion ($b = 0.039$, $p = .002$), indicating that the reduction in similarity under the *deletion* condition was smaller for *filtered-all* variant than for the *original* variant. No such effect was observed for *filtered-and* ($p = .987$).

Overall, these results suggest that *alike* produces the highest similarity scores, *deletion* reduces similarity, and *filtered-all* is specifically less affected by the deletion manipulation.

Model/Test Corpus	*Alike-from-unlike*	*Supercat-deletion*	*Deletion*
original	0.6573 ± 0.0027	0.4880 ± 0.0086	0.4002 ± 0.0165
filtered-all	0.6512 ± 0.0200	0.4716 ± 0.0159	0.4264 ± 0.0123
filtered-and	0.6925 ± 0.0198	0.5113 ± 0.0117	0.4402 ± 0.0325

Table 4.3: Median value of average similarity for each seed of each model variant on each corpus with standard deviation across seeds. Average similarity is computed by taking average of all cosine similarity scores between conjuncts of all sentences in a corpus.

4.3.3 Clustering Analysis

Based on the representations obtained before computing similarity, we applied the k-means algorithm ($k = 3$) to produce 3×3 clustering systems. Details of the procedure are provided in Section 3.5.2.

Figure 4.6 provides an illustration of the clustering patterns. The clustering reveals that the models treat both alike coordination in *alike-from-unlike* and unlike coordination that can be considered alike within the same supercategory in *supercat-deletion* as alike coordination, since phrases grouped together are mostly conjuncts from the same sentence. In contrast, for sentences in *deletion*, where it is difficult to find an overarching alike category, the models recognize unlike coordination. Here, conjuncts from different sentences but sharing the same grammatical category are grouped together, showing a clear structure: nouns cluster with nouns, verbs with verbs, adjectives with adjectives, and so on.

Term	Coef.	SE	z	p
Fixed Effects				
Intercept	0.485	0.016	29.91	$< .001$
Model: filtered-all	-0.016	0.009	-1.75	.080
Model: filtered-and	0.019	0.009	2.19	.290
Corpus: alike	0.173	0.023	7.52	$< .001$
Corpus: deletion	-0.084	0.023	-3.66	$< .001$
Model: filtered-all × Corpus: alike	0.015	0.013	1.24	.217
Model: filtered-and × Corpus: alike	0.004	0.013	0.28	.778
Model: filtered-all × Corpus: deletion	0.039	0.013	3.08	.002
Model: filtered-and × Corpus: deletion	-0.000	0.013	-0.02	.987
Random Effects				
Group (Intercept Var)	0.016			

Table 4.4: Linear mixed-effects regression predicting similarity scores. Baseline levels are *original* for Model and *supercat-deletion* for Corpus. Positive coefficients indicate higher similarity relative to the baseline.

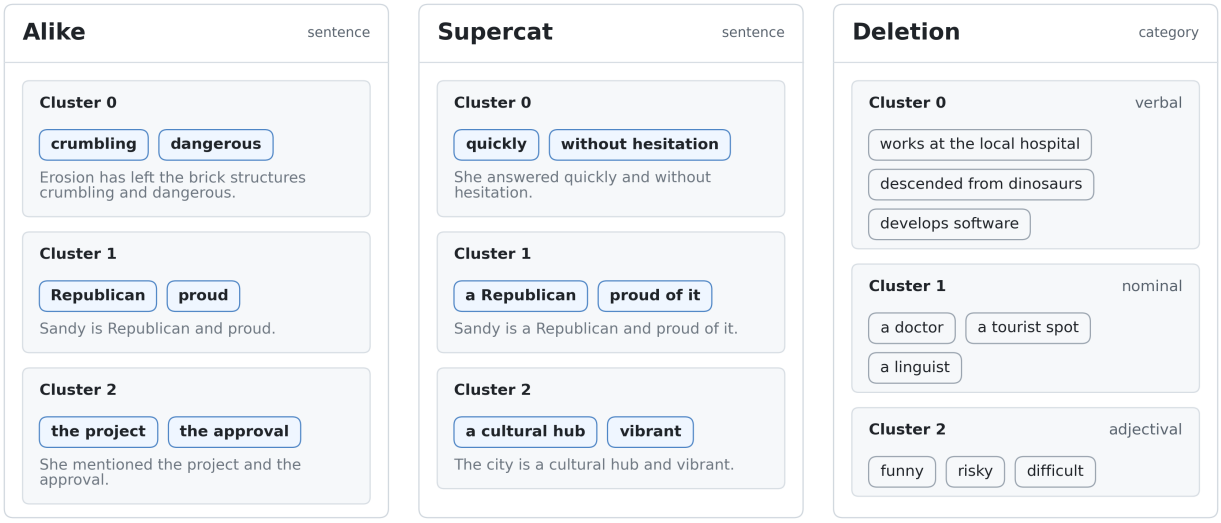

Figure 4.6: Illustration of clustering results. For each corpus, three clustering systems are trained with different random seeds. As they produce largely similar clusterings, we collapse them into a single representative system for visualization purposes.

Chapter 5

DISCUSSION

At a high level, this study investigates unlike coordination in language models along two complementary dimensions: first, how the models acquire knowledge of this linguistic construction, and second, how they process and represent it internally. Beyond language models, this investigation is partly motivated by ongoing debates in linguistic theory regarding the nature of coordination and the status of the LCL. In this chapter, we discuss these issues in light of our results, and also present our understanding of the limitations and promising directions of future research on this line of investigation.

5.1 Learning Unlike Coordination

Our investigation yields a clear finding that unlike coordination does not constitute an exception or a special case for LMs. The results from the filtered models demonstrate that direct exposure to unlike coordination is not strictly necessary for a model to accept and process it. *Filtered-all* model, despite being trained on a corpus where every instance of unlike coordination was removed (approximately 36% of the original tokens filtered out in the process to ensure purity), achieved perplexity scores on unlike coordination tasks that were not statistically significantly different from the *original* model in many pairwise comparisons.

Our results also suggest that LMs do not possess an inherent inductive bias towards a strict interpretation of the Law of Coordination of Likes (LCL). The models do not appear to have constraints that prevent them from conjoining constituents of different categories. If models strictly adhered to the LCL as a default assumption, presupposing that only constituents of identical categories can be conjoined, the *filtered-all* model should have catastrophically failed on the *unlike* evaluation corpus. Instead, it generalized effectively, in terms

of both the perplexity measurement and the grammaticality judgment results, implying that the constraints governing coordination in LMs are permissive rather than restrictive. We may further interpret this as indicating that the data of English also do not contain restrictive constraints that force the LCL to strictly hold.

The performance of the *filtered-all* model provides compelling evidence that direct exposure to unlike coordination is not strictly necessary for a model to handle it. While the *original* model performed slightly better than the filtered models, both *filtered-all* and *filtered-and* maintained a robust performance level. This suggests that the ability to process unlike coordination emerges from the model's general compositional capabilities, likely learned through alike coordination and other syntactic structures, rather than requiring explicit positive evidence of category mismatches.

Furthermore, the data suggest that alike coordination provides sufficient latent signal for models to license unlike coordination. When we isolate *supercat-deletion* and *deletion* from the broader *unlike* evaluation set, we find that all models perform robustly on these subsets. Detailed analyses suggest that the models all rely on the similarity between conjuncts at the level of supercategories, which offers a broadly available structural cue for unlike coordination. With greater exposure to unlike coordination in the training data, however, models become increasingly capable of exploiting the deletion-based mechanism for interpreting these structures. This is reflected in a consistent performance ranking of *original* > *filtered-and* > *filtered-all*, and a consistent pattern in which *filtered-all* shows less sensitivity to *deletion*. One plausible explanation is that if a model learns that a particular functional projection (e.g., a modifier) can occupy a given syntactic position, and separately learns the mechanics of conjunction through alike coordination, it can recombine these pieces to generate well-formed unlike coordination. Because functional projections are distributed everywhere, they can be learned independently of explicit exposure to unlike coordination and later integrated into the model's generalization process. Because deletion-based constructions are likely less frequent, the absence of exposure to unlike coordination provides a less ideal environment for models to generalize these patterns. The models can still learn deletion, but less effectively.

The filtered models generally tend to reject the dual-role case: sentences containing a dual-role phrase elicit significantly higher perplexity, indicating that these configurations are not readily licensed. However, with sufficient exposure to unlike coordination, as in the *original* model, the models begin to accommodate these structures. In contrast, combinations of unlike supercategories are strongly rejected by all model variants. This pattern suggests that the training data, even when they contain instances of unlike coordination, include very few cases that bridge different supercategories. As a result, the dominant strategy of relying on similarity between conjuncts makes it difficult for the models to accommodate such opposing possibilities. Taken together, these results suggest that while the general mechanics of unlike coordination are learnable without direct exposure, the grammaticality judgment tests indicate that direct exposure is beneficial for learning specific constraints. Moreover, greater exposure appears necessary for the models to learn more nuanced cases of unlike coordination, which are licensed under conditions that are difficult for a statistical learner to generalize from limited evidence.

Based on the results, we are certain that unlike coordination is not an exception to LMs. They are also not in general a special case, although direct exposure does increase their performance on unlike coordination. The models do not have a bias towards the strict LCL, which means that they do not presuppose a kind of structure where only phrases of the same syntactic category can be conjoined into a larger unit.

5.2 Processing Unlike Coordination

Beyond the question of *if* models learn unlike coordination, our internal representation analysis illuminates *how* they process it. The analysis of internal representations reveals that models may employ distinct mechanisms, specifically mechanisms akin to supercategory derivation and deletion, to process these structures.

The differences in similarity scores across the three corpora reveal distinct representational patterns in how the models encode conjuncts. Conjuncts in *alike-from-unlike*, which are genuinely alike to begin with, receive the highest similarity scores, indicating that the

models consistently encode them as closely related. Conjuncts in *supercat-deletion* receive intermediate similarity scores: although these pairs are not instances of true alike coordination, they can be construed as alike at the level of functional projection or supercategory, and the models appear to capture this partial structural overlap. In contrast, conjuncts in *deletion* exhibit the lowest similarity scores. These items share very little beyond their status as conjuncts, making them difficult to interpret as alike by any syntactic or functional criterion, and the models' representations reflect this lack of common structure.

The clustering analysis continues to reveal a sharp distinction between sentences in *supercat-deletion* and *deletion*. For *supercat-deletion* (and *alike-from-unlike*), the models cluster conjuncts based on the sentence they belong to, rather than their grammatical category. This indicates that in these cases, the model integrates the conjuncts into a unified semantic or syntactic representation, effectively a "supercategory", where the likeness of the conjuncts is derived from their shared function within the specific context. In contrast, for sentences in the *deletion* corpus, the models cluster conjuncts by their grammatical category (e.g., noun with noun, verb with verb) across different sentences. This suggests that when a unifying supercategory is unavailable, the model treats the conjuncts as distinct syntactic entities. This aligns with theories that posit deletion or ellipsis, where the conjuncts remain structurally independent and are joined only after parts of the structure are elided.

The attention scores further corroborate this distinction. Models with exposure to unlike coordination (*original* and *filtered-and*) exhibit clear and significant differences in their attention patterns when processing *deletion* versus *supercat-deletion* sentences. These patterns suggest that the models are indeed deploying mechanisms that resemble deletion-like operations: in the *deletion* condition, the second conjunct draws more information from the subject rather than from the first conjunct. Although the *filtered-all* model does not show statistically significant differences between these two sentence types, it nonetheless displays a noticeable shift in its attention behavior. This indicates that direct exposure to unlike coordination is not strictly required for processing these structures, since *filtered-all* still performs well and can employ different strategies. But such exposure does refine the model's internal

processing strategies. In particular, it enables the model to more sharply distinguish between structures that call for a unified representation and those that rely on implicit deletion.

5.3 Connection to Linguistic Theory

By observing how LMs acquire and represent unlike coordination under controlled training conditions, we can infer which theoretical analyses best align with the models' inductive biases and generalization patterns, and interpret the models accordingly. On the other hand, the models provide a usage-based and data-driven perspective on the plausibility of these theories—though, of course, such interpretations should be approached with a grain of salt.

The most immediate implication of our findings is a challenge to the necessity of the Law of Coordination of Likes (LCL) as a fundamental constraint on grammar. The models' successful performance suggests that it does not presuppose a constraint where conjuncts must be of identical categories. This supports part of the theoretical perspective of Patejuk and Przepiórkowski (2023), who argue that the LCL is illusory and that coordination is licensed not by the identity of the conjuncts, but by whether each conjunct independently satisfies the external constraints of the syntactic position. The LMs appear to operate on this principle: having learned from alike coordination that a specific position accepts specific categories, the model permits any combination of those categories in that position, effectively deriving unlike coordination from the independent validity of the conjuncts

While the LMs reject a strict LCL, our investigation into internal representations suggests a hybrid view that incorporates elements of general LCL-like explanation. Clustering analysis revealed that the models grouped conjuncts from the same sentence together, rather than grouping them by grammatical category across sentences. This implies that the model constructs a unified representation for these conjuncts, treating them as functionally equivalent in that specific context. This behavior aligns with the "Feature Bundles" proposal by Sag et al. (1985) and the Supercategory view, where disparate categories are viewed as instances of a shared, underspecified higher-level category. The models seem to operate on

the principle that conjuncts must match in semantic similarity or functional meaning, a matching requirement the models prioritize over strict syntactic labels.

Conversely, for sentences in the *deletion* corpus, the models clustered conjuncts according to their specific grammatical categories (e.g., verbs with verbs) rather than as a unified sentence-specific group. This distinct representation supports theories that view such coordination as the result of ellipsis or deletion, as discussed by Bruening and Al Khalaf (2020) regarding coordination of larger categories. Here, the model appears to treat the conjuncts as structurally independent units that are joined via a mechanism akin to deletion, rather than forcing them into a shared supercategory.

Most importantly, the fact that language models are able to generalize from alike coordination to unlike coordination suggests that unlike coordination does not require a separately motivated grammatical mechanism. In our experiments, the models successfully extended patterns learned from alike coordination to unlike coordination, employing multiple representational strategies, including the two we identified here, namely a supercategory-based strategy and a deletion-based strategy. Consequently, theoretical accounts that posit individually-motivated constraints or mechanisms for unlike coordination may be unnecessarily strong. Unlike coordination appears to emerge naturally from the same representational resources that underlie coordination more broadly, together with general strategies, such as category underspecification, or deletion-like operations.

5.4 Limitations

This study operates under constraints regarding model size and computational resources. We utilized a GPT-2 architecture with approximately 774 million parameters. While sufficient for analyzing fundamental syntactic generalizations, larger state-of-the-art models might exhibit different behaviors or more robust abstractions. It is also possible that with more training data and epochs, these models we evaluate can have a better performance or different behaviors.

Additionally, our reliance on automatic parsing for the filters introduces potential noise.

Although our evaluation shows the filters are "strong", favoring false negatives over false positives. A more fine-grained filtering pipeline could reduce the inadvertent removal of complex alike coordination structures. At the same time, the conservativeness of our current filter arguably strengthens our findings: the filtered models succeed despite being trained on an aggressively impoverished dataset. If a model can generalize unlike coordination from input that systematically removes anything resembling it, this provides even stronger evidence that direct exposure is not necessary for acquiring the construction.

5.5 Future Work

There are several promising directions for future work that go beyond extending the present study. One broader avenue involves improving the theoretical understanding of how language models encode structural and semantic constraints. Although current research focuses on specific coordination phenomena, future work could investigate whether similar patterns emerge across a wider range of syntactic constructions and languages. Such cross-linguistic and cross-phenomenon comparisons would help reveal whether the observed patterns reflect general representational tendencies or are specific to particular constructions.

The second direction concerns the role of training data in shaping these generalizations. The filtered corpora used in this study provided strong indirect evidence that conjuncts typically match in syntactic category, but the results show that the models do not fully recognize this pattern as a categorical constraint. This suggests that language models may rely on alternative cues, such as contextual compatibility or semantic similarity between conjuncts. Future work could directly test this hypothesis by systematically manipulating the training data, for example, by constructing corpora that control for semantic similarity while varying syntactic category distributions, or by introducing targeted contrasts between syntactically matched but semantically incompatible conjuncts. Such experiments would help clarify which cues the models prioritize when learning coordination patterns.

Finally, an important long-term direction is to connect these findings to human language processing. At present, there appears to be relatively little psycholinguistic data specifically

targeting unlike coordination, which makes direct comparisons between model behavior and human processing difficult. Collecting targeted experimental data, such as acceptability judgments, reading-time studies, or eye-tracking experiments involving systematically varied coordination structures, would make it possible to evaluate whether the strategies observed in language models correspond to patterns in human sentence processing. Such work could help clarify whether the mechanisms identified here reflect general properties of language processing.

Chapter 6

CONCLUSION

This thesis set out to investigate the nature of unlike coordination at the intersection of theoretical linguistics and natural language processing, and examine how language models acquire and represent unlike coordination. By systematically manipulating exposure to different coordination structures and analyzing the resulting representations and processing patterns, we uncovered several key findings.

Our findings provide empirical evidence that unlike coordination is not a special exception for language models, nor does it require direct exposure to be learned. The model trained on a corpus entirely void of unlike coordination (*filtered-all*) demonstrated a remarkable ability to generalize, performing comparably to, and in specific grammaticality judgments, better than, the baseline model trained on unfiltered text. This result challenges the strict interpretation of the LCL in the context of computational learning.

Furthermore, our analysis of internal representations suggests that LMs do not treat all unlike coordination uniformly. Instead, they appear to recruit different processing strategies for different structures. Cases amenable to a "supercategory" analysis are represented similarly to alike coordination, implying a unified structural interpretation. Conversely, cases that theoretically involve deletion are processed with distinct representations for each conjunct.

Also, attention patterns, perplexity measures, and grammaticality tests indicate that while models can recombine known structural elements to generate novel coordination, their internal mechanisms remain sensitive to training distribution, highlighting both strengths and limitations in their structural generalization capabilities.

The study also underscores the importance of targeted, linguistically informed prob-

ing as a method for interpreting model behavior, offering insights not readily captured by standard benchmarks. This work highlights broader directions for research. Future studies should explore scaling these methodologies to larger and more sophisticated language models, developing linguistically nuanced evaluation frameworks, and comparing model behavior to human judgments to better understand the cognitive plausibility of neural language representations. By bridging controlled empirical investigation, theoretical linguistics, and computational modeling, this line of work contributes to a more nuanced understanding of both the capabilities and limitations of contemporary language models.

BIBLIOGRAPHY

Yonatan Belinkov and James Glass. 2019. Analysis Methods in Neural Language Processing: A Survey. Transactions of the Association for Computational Linguistics, 7:49–72.

John Bowers. 1993. The Syntax of Predication. Linguistic Inquiry, 24(4):591–656.

Benjamin Bruening and Eman Al Khalaf. 2020. Category Mismatches in Coordination Revisited. Linguistic Inquiry, 51(1):1–36.

Jose Antonio Camacho. 1997. The Syntax of NP Coordination. Ph.D. thesis, University of Southern California, United States – California.

Noam Chomsky. 1957. Syntactic Structures. Syntactic Structures. Mouton, Oxford, England.

Barbara Citko. 2011. Symmetry in Syntax: Merge, Move and Labels.

Chris Collins. 1988. Conjunction adverbs. Unpublished MS.

Alexis Conneau, German Kruszewski, Guillaume Lample, Loïc Barrault, and Marco Baroni. 2018. What you can cram into a single $&!#* vector: Probing sentence embeddings for linguistic properties. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 2126–2136, Melbourne, Australia. Association for Computational Linguistics.

Richard Futrell, Ethan Wilcox, Takashi Morita, Peng Qian, Miguel Ballesteros, and Roger Levy. 2019. Neural language models as psycholinguistic subjects: Representations of syntactic state. 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), pages 32–42. Association for Computational Linguistics.

Mario Giulianelli, Jack Harding, Florian Mohnert, Dieuwke Hupkes, and Willem Zuidema. 2018. Under the Hood: Using Diagnostic Classifiers to Investigate and Improve how Language Models Track Agreement Information. In Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP, pages 240–248, Brussels, Belgium. Association for Computational Linguistics.

Grant Goodall. 1987. Parallel Structures in Syntax: Coordination, Causatives, and Restructuring. Cambridge University Press.

Alexander Grosu. 1985. SUBCATEGORIZATION AND PARALLELISM. 12(2-3):231–240.

Kristina Gulordava, Piotr Bojanowski, Edouard Grave, Tal Linzen, and Marco Baroni. 2018. Colorless Green Recurrent Networks Dream Hierarchically. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), pages 1195–1205, New Orleans, Louisiana. Association for Computational Linguistics.

Martin Haspelmath. 2004. Coordinating constructions. Typological Studies in Language, v.58 (2004).

John Hewitt and Christopher D. Manning. 2019. A Structural Probe for Finding Syntax in Word Representations. 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), pages 4129–4138, Minneapolis, Minnesota. Association for Computational Linguistics.

Dieuwke Hupkes, Sara Veldhoen, and Willem Zuidema. 2018. Visualisation and 'Diagnostic Classifiers' Reveal How Recurrent and Recursive Neural Networks Process Hierarchical Structure. Journal of Artificial Intelligence Research, 61:907–926.

Janne Bondi Johannessen. 1998. Coordination. Oxford University Press.

Julie Kallini and Christiane Fellbaum. 2021. A Corpus-based Syntactic Analysis of Two-termed Unlike Coordination. In Findings of the Association for Computational Linguistics: EMNLP 2021, pages 3998–4008, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Julie Kallini, Isabel Papadimitriou, Richard Futrell, Kyle Mahowald, and Christopher Potts. 2024. Mission: Impossible Language Models. In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 14691–14714, Bangkok, Thailand. Association for Computational Linguistics.

R.S. Kayne. 1994. The Antisymmetry of Syntax. Linguistic Inquiry Monographs. MIT Press.

Nikita Kitaev and Dan Klein. 2018. Constituency Parsing with a Self-Attentive Encoder. In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 2676–2686, Melbourne, Australia. Association for Computational Linguistics.

Josef Klafka and Allyson Ettinger. 2020. Spying on Your Neighbors: Fine-grained Probing of Contextual Embeddings for Information about Surrounding Words. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 4801–4811, Online. Association for Computational Linguistics.

Cara Su-Yi Leong and Tal Linzen. 2026. Manipulating language models' training data to study syntactic constraint learning: The case of english passivization. Journal of Memory and Language, 149:104751.

Tal Linzen, Emmanuel Dupoux, and Yoav Goldberg. 2016. Assessing the Ability of LSTMs to Learn Syntax-Sensitive Dependencies. Transactions of the Association for Computational Linguistics, 4:521–535.

Nelson F. Liu, Matt Gardner, Yonatan Belinkov, Matthew E. Peters, and Noah A. Smith. 2019. Linguistic Knowledge and Transferability of Contextual Representations. 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), pages 1073–1094, Minneapolis, Minnesota. Association for Computational Linguistics.

I. Loshchilov and F. Hutter. 2017. Decoupled Weight Decay Regularization. In International Conference on Learning Representations.

Rebecca Marvin and Tal Linzen. 2018. Targeted Syntactic Evaluation of Language Models. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 1192–1202, Brussels, Belgium. Association for Computational Linguistics.

R. Thomas McCoy, Ellie Pavlick, and Tal Linzen. 2019. Right for the Wrong Reasons: Diagnosing Syntactic Heuristics in Natural Language Inference. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pages 3428–3448, Florence, Italy. Association for Computational Linguistics.

Kanishka Misra and Kyle Mahowald. 2024. Language models learn rare phenomena from less rare phenomena: The case of the missing AANNs. In Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 913–929, Miami, Florida, USA. Association for Computational Linguistics.

Husni Muadz. 1991. Coordinate Structures: A Planar Representation. Ph.D. thesis, The University of Arizona, United States – Arizona.

Alan Boag Munn. 1992. A null operator analysis of ATB gaps. The Linguistic Review, 9(1):1–26.

Alan Boag Munn. 1993. Topics in the syntax and semantics of coordinate structures. Ph.D. thesis, University of Maryland.

Isabel Papadimitriou and Dan Jurafsky. 2023. Injecting structural hints: Using language models to study inductive biases in language learning. In Findings of the Association for Computational Linguistics: EMNLP 2023, pages 8402–8413, Singapore. Association for Computational Linguistics.

Agnieszka Patejuk and Adam Przepiórkowski. 2023. Category Mismatches in Coordination Vindicated. Linguistic Inquiry, 54(2):326–349.

Abhinav Patil, Jaap Jumelet, Yu Ying Chiu, Andy Lapastora, Peter Shen, Lexie Wang, Clevis Willrich, and Shane Steinert-Threlkeld. 2024. Filtered Corpus Training (FiCT) Shows that Language Models Can Generalize from Indirect Evidence. Transactions of the Association for Computational Linguistics, 12:1597–1615.

P. G. Peterson. 1981. Problems with Constraints on Coordination. Linguistic Analysis, 8:449–460.

Ljiljana Progovac. 1997. Slavic and the Structure for Coordination. In Formal Approaches to Slavic Linguistics, volume 5, pages 207–223.

Geoffrey K. Pullum and Arnold M. Zwicky. 1986. Phonological Resolution of Syntactic Feature Conflict. Language, 62(4):751–773.

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

Arthur S. Reber. 1967. Implicit learning of artificial grammars. Journal of Verbal Learning & Verbal Behavior, 6(6):855–863.

Edward Jon Rubin. 1994. Modification: A Syntactic Analysis and Its Consequences. Ph.D. thesis, Cornell University, United States – New York.

Ivan A. Sag, Gerald Gazdar, Thomas Wasow, and Steven Weisler. 1985. Coordination and how to distinguish categories. Natural Language & Linguistic Theory, 3(2):117–171.

Gözde Gül Şahin, Clara Vania, Ilia Kuznetsov, and Iryna Gurevych. 2020. LINSPECTOR: Multilingual Probing Tasks for Word Representations. Computational Linguistics, 46(2):335–385.

Paul Schachter. 1977. Constraints on Coördination. Language, 53(1):86–103.

Craig Thiersch. 1994. On some formal properties of coordination. In Carlos Martín-Vide, editor, North-Holland Linguistic Series: Linguistic Variations, volume 56, pages 171–180. Elsevier.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In Advances in Neural Information Processing Systems, volume 30. Curran Associates, Inc.

Alex Warstadt and Samuel R. Bowman. 2022. What Artificial Neural Networks Can Tell Us about Human Language Acquisition. In Algebraic Structures in Natural Language. CRC Press.

Alex Warstadt, Amanpreet Singh, and Samuel R. Bowman. 2019. Neural Network Acceptability Judgments. Transactions of the Association for Computational Linguistics, 7:625–641.

Edwin Williams. 1978. Across-the-Board Rule Application. Linguistic Inquiry, 9(1):31–43.

Edwin S. Williams. 1981. Transformationless Grammar. Linguistic Inquiry, 12(4):645–653.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, and 3 others. 2020. Transformers: State-of-the-art natural language processing. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, pages 38–45, Online. Association for Computational Linguistics.

Ellen Woolford. 1987. An ECP Account of Constraints on Across-the-Board Extraction. Linguistic Inquiry, 18(1):166–171.

Xiulin Yang, Tatsuya Aoyama, Yuekun Yao, and Ethan Wilcox. 2025. Anything Goes? A Crosslinguistic Study of (Im)possible Language Learning in LMs. In Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 26058–26077, Vienna, Austria. Association for Computational Linguistics.

Paul Youssef, Osman Koraş, Meijie Li, Jörg Schlötterer, and Christin Seifert. 2023. Give Me the Facts! A Survey on Factual Knowledge Probing in Pre-trained Language Models. In Findings of the Association for Computational Linguistics: EMNLP 2023, pages 15588–15605. Association for Computational Linguistics.

Niina Ning Zhang. 2009. Coordination in Syntax, volume 123. Cambridge University Press.

Cyril Edward Zoerner. 1995. Coordination: The syntax of &P. Ph.D. thesis, University of California, Irvine.

Appendix A

TRAINING PARAMETERS

num_train_epochs	10
per_device_train_batch_size	8
gradient_accumulation_steps	2
per_device_eval_batch_size	16
learning_rate	5e-5
warmup_steps	1000
weight_decay	0.01
max_grad_norm	1.0
fp16	True
dataloader_drop_last	True
prediction_loss_only	True
metric_for_best_model	'eval_loss'
save_strategy	'steps'
save_steps	1000
save_total_limit	3

Table A.1: A list of the selected training parameters. Any values not listed, except for logging, are assumed to be the default settings in the `transformers` package, version 4.56.2.

Appendix B

RESULTS OF GRAMMATICALITY JUDGMENT TESTS

Number & Category	Original	Filtered-all	Filtered-and
1. selection violation	correct	correct	incorrect
2. position effect	correct	correct	correct
3. CP as NP requirement	incorrect	incorrect	incorrect
4. CP as NP requirement	correct	correct	correct
5. verb dependency	correct	correct	correct
6. mismatch restricted	incorrect	incorrect	incorrect
7. position effect	correct	incorrect	correct
8. position effect	correct	correct	incorrect
9. mismatch restricted	correct	correct	correct
10. mismatch restricted	incorrect	incorrect	incorrect
11. mismatch restricted	incorrect	incorrect	incorrect
12. wrong topicalization	correct	incorrect	correct
13. deletion	correct	correct	correct
14. position effect	correct	correct	correct
15. mismatch restricted	correct	correct	correct
16. mismatch restricted	correct	correct	correct
17. dual roles	n/a	n/a	n/a
18. unlike supercategory	n/a	n/a	n/a
19. deletion	correct	correct	correct
20. deletion	correct	correct	incorrect
21. deletion	correct	correct	correct
22. deletion	incorrect	correct	correct

Table B.1: Results of the three models on each test in *unlike-judgment*.

Appendix C

STATISTICS OF THE COMBINATION OF UNLIKE CATEGORIES

This appendix presents two tables of the statistics of the combination of unlike categories as in Kallini and Fellbaum (2021).

Conjoined phrases	Relative frequency
NP & SBAR	0.121
NP & VP	0.099
ADJP & VP	0.073
ADVP & PP	0.065
NP & ADJP	0.065
PP & VP	0.057
PP & ADVP	0.057
NP & PP	0.047
PP & NP	0.045
VP & NP	0.039

Table C.1: Ten most frequent combinations of conjoined phrases for unlike coordination in COCA.

Conjoined phrases	Relative frequency
ADVP & PP	0.113
PP & ADVP	0.110
NP & ADJP	0.096
ADJP & NP	0.066
NP & ADVP	0.060
ADJP & VP	0.060
ADJP & PP	0.056
PP & SBAR	0.056
NP & SBAR	0.050
VP & ADJP	0.050

Table C.2: Ten most frequent combinations of conjoined phrases for unlike coordination in PTB.