Scope Ambiguities in Large Language Models

Gaurav Kamath[α,β] **Sebastian Schuster**[γ] **Sowmya Vajjala**[δ] **Siva Reddy**[α,β,ϵ]
[α]McGill University, Canada [β]Mila - Quebec AI Institute, Canada [γ]University College London, UK
[δ]National Research Council Canada [ϵ]Facebook CIFAR AI Chair, Canada
gaurav.kamath@mail.mcgill.ca s.schuster@ucl.ac.uk
sowmya.vajjala@nrc-cnrc.gc.ca siva.reddy@mila.quebec

Abstract

Sentences containing multiple semantic operators with overlapping scope often create ambiguities in interpretation, known as *scope ambiguities*. These ambiguities offer rich insights into the interaction between semantic structure and world knowledge in language processing. Despite this, there has been little research into how modern large language models treat them. In this paper, we investigate how different versions of certain autoregressive language models—GPT-2, GPT-3/3.5, Llama 2, and GPT-4—treat scope ambiguous sentences, and compare this with human judgments. We introduce novel datasets that contain a joint total of almost 1,000 unique scope-ambiguous sentences, containing interactions between a range of semantic operators, and annotated for human judgments. Using these datasets, we find evidence that several models (i) are sensitive to the meaning ambiguity in these sentences, in a way that patterns well with human judgments, and (ii) can successfully identify human-preferred readings at a high level of accuracy (over 90% in some cases).[1]

1 Introduction

Sentences like 'every farmer owns a donkey' are systematically ambiguous between two readings: one in which the embedded noun phrase (NP) (e.g., 'a donkey') is interpreted within the scope of the quantifier that precedes it ('every'), and another in which the embedded NP is interpreted outside its scope. As shown in Figure 1, 'every farmer owns a donkey', for example, could either mean (i) that each farmer simply owns their own (possibly unique) donkey, or (ii) that there is a specific donkey in question that all farmers jointly own.

[1]Data and code are available at: https://github.com/McGill-NLP/scope-ambiguity.

Such constructions are examples of what are known as *scope ambiguities*. They are called so because the standard account of these ambiguities is that they arise when the respective scope of multiple semantic operators in the expression is ambiguous, yielding more than one possible semantic structure. Consider the following example:

(1) a. Every farmer owns a donkey.

b. Surface Scope: $\forall y[farmer(y) \rightarrow \exists x[donkey(x) \wedge owns(y,x)]]$

c. Inverse Scope: $\exists x[donkey(x) \wedge \forall y[farmer(y) \rightarrow owns(y,x)]]$

(1a), in logical form, involves a universal quantifier (introduced by 'every'), and an existential quantifier (introduced by 'a'). The ambiguity lies in the order of application (and thereby scopes) of these two operators. The surface scope reading of the sentence, (1b), involves the universal quantifier outscoping the existential quantifier. The inverse scope reading, (1c), involves the reverse.

Importantly for this present work, English speakers (i) have access to both kinds of readings, and (ii) generally disambiguate between them to arrive at a preferred reading (see Kurtzman and MacDonald, 1993). For example, although (1a) has two possible interpretations, without further context, most people would prefer the surface reading, due to at least the surface positions of 'a' and 'every' in the sentence, as well as background world knowledge about farmers and donkeys (see Kurtzman and MacDonald, 1993; Saba and Corriveau, 2001; Anderson, 2004, for insights into how such factors affect reading preferences).

The focus of this paper is how large language models (LLMs) treat such ambiguities. Assessing how they do so offers important insights into

Transactions of the Association for Computational Linguistics, vol. 12, pp. 738–754, 2024. https://doi.org/10.1162/tacl_a_00670
Action Editor: Kristina Toutanova. Submission batch: 11/2023; Revision batch: 3/2024; Published 6/2024.
© 2024 Association for Computational Linguistics. Distributed under a CC-BY 4.0 license.

Downloaded from http://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00670/2377773/tacl_a_00670.pdf by guest on 23 February 2026

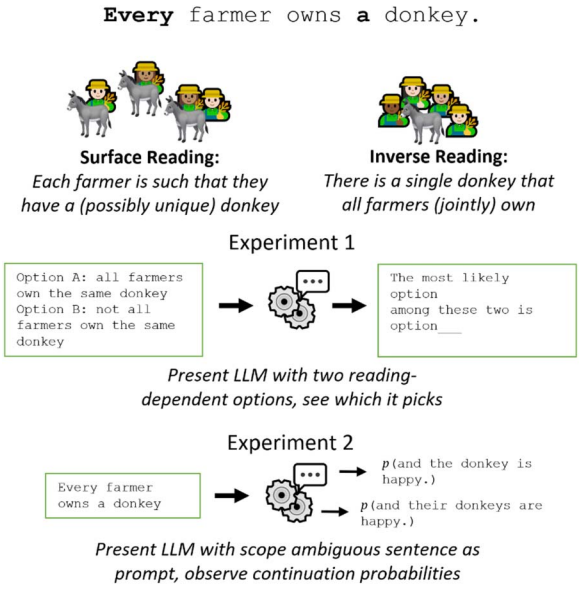

Figure 1: A high-level overview of our study, showing our approaches to our first (see Section 4) and second (see Section 5) experiments.

interactions between semantic structure and world knowledge, as well as the representation of scope in LLMs.

Semantic Structure and World Knowledge Scope disambiguation lies at the interface between natural language semantics and background world knowledge. Scope ambiguous sentences like (1) are ambiguous between two semantic structures; disambiguating between these two possible structures (and the different readings they yield), however, often requires background world knowledge (Saba and Corriveau, 2001). Take the following two sentences:

(2) a. Every conference attendee ate a Big Mac.

b. Every conference attendee attended a networking event.

Both examples in (2) are scope-ambiguous in a similar way to (1)—each offers two possible semantic structures yielding different readings. However, choosing the preferred reading is easy in both cases: in (2a), the surface scope reading (every attendee ate a potentially different Big Mac) is preferred, while in (2b), the inverse scope reading (there was a single networking event that all attendees attended) is preferred. These preferences are a result of the general knowledge we have about conference attendees, networking events, and Big Macs.

LLMs have been shown to be able to capture aspects of world knowledge (e.g., Roberts et al., 2020; Heinzerling and Inui, 2021; AlKhamissi et al., 2022), and, separately, to capture some properties of natural language semantics (e.g., Jawahar et al., 2019; Ettinger, 2020; Pavlick, 2022). Scope ambiguities present an opportunity to assess how they might integrate the two.

Scope Representation in LLMs Model weights are largely uninterpretable, so despite generally high performance on language-based tasks, many questions remain about the abstract linguistic structures they capture (Belinkov and Glass, 2019; Hewitt and Manning, 2019; Baroni, 2022). The ambiguities discussed here arise out of a crucial component of linguistic structure: scope. Analyzing how LLMs treat them helps us gain insight into how well they capture this component of structure. This is particularly interesting because while formal logic, as in (1), allows for a straightforward, symbolic representation of scope ambiguities, it remains an open question whether vector-based LLM representations can adequately capture the multiple readings of such constructions.

This paper therefore attempts to answer two questions:

Q1: *Do LLMs exhibit similar preferences to humans in the interpretation of scope ambiguous sentences?*

Q2: *Are LLMs sensitive to the presence of more than one reading of scope ambiguous sentences?*

We conduct two experiments to investigate these questions. From these experiments, we present evidence that the answer to these questions—at least for the more powerful models—is ‘yes’.

2 Related Work

Scope ambiguities have been the focus in research within computational linguistics and natural language processing (NLP) primarily through the task of quantifier scope disambiguation, which involves the proper selection of a preferred scope reading given a scope-ambiguous sentence.

Downloaded from http://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00670/2377773/tacl_a_00670.pdf by guest on 23 February 2026

Early examples of NLP research on this task, such as Higgins and Sadock (2003) and Andrew and MacCartney (2004), frame it as a classification task, and find models that outperform naive heuristics. Such work predates modern neural language models; Rasmussen (2022), however, builds on this approach, framing quantifier scope disambiguation as a span-pair classification task. They test RoBERTa (Liu et al., 2019) on this task, and find that the model achieves higher accuracy on it than a majority-prediction baseline. This work, however, does not directly test the model's underlying linguistic capabilities; it tests the model on the classification task only after it is trained on examples from the dataset used. As a result, it is unclear to what degree the model's performance on the test set is due to linguistic capabilities that emerged from its pretraining.

Manshadi and Allen (2011) and Tsiolis (2020) approach the problem differently, as neither frames it as a classification task. Manshadi and Allen (2011) represent scope relations as graphs, and frame the task as one of graph construction; they present a support vector machine that beats a naive heuristic baseline. Tsiolis (2020), on the other hand, attempts to reframe the task as a natural language inference task, Q&A task, or one in which probabilities of continuations are compared. They use a large language model—GPT-2 (Radford et al., 2019)—but present mixed results.

Other research focuses on the linguistic factors that determine scope reading preferences in a corpus. AnderBois et al. (2012) find that linear order, grammatical roles, and lexical effects determine these preferences; Leczkowski et al. (2022) build on this work and find that prepositions and preposition senses also affect scope reading preference.

The only two instances of work assessing how LLMs treat scope ambiguities in zero-shot contexts are, to our knowledge, recent work by Liu et al. (2023), and Stengel-Eskin et al. (2023). The latter assesses how LLMs treat ambiguous inputs in terms of semantic parsing. The authors use templates to generate ambiguous sentences—including scope-ambiguous sentences—along with logical parses of them, and assess the abilities of LLMs to properly produce the two logical parses of each ambiguous sentence, in both few-shot and zero-shot contexts. They find that models are poor at generating both parses of ambiguous sentences in zero-shot contexts, but can more accurately generate both parses in few-shot contexts. Liu et al. (2023), on the other hand, assess how LLMs treat linguistic ambiguity in terms of entailment relations. Using prompting approaches to the task, as well as observing probabilities assigned to continuations of ambiguous sentences, they present evidence suggesting LLMs struggle to model ambiguity.

Both studies, though they do not primarily focus on it, do include scope ambiguity data, and are thus relevant to our work. Where we diverge from these works, however, is in our data and experimental methods. While the templates Stengel-Eskin et al. (2023) use allow for the generation of hundreds of sentences, they do limit the diversity of these stimuli; moreover, the scope ambiguities in their datasets are limited to instances of quantifier-quantifier interactions. Similarly, Liu et al. (2023) estimate from a random sample that roughly 7.6% of their data involves scope ambiguity; manually inspecting all 579 ambiguous sentences in their dataset, however, we find that the dataset contains a total of around 20 instances of scope ambiguity. We also employ different experimental set-ups (see Sections 4.1 and 5.1) than those used in the aforementioned works. Crucially, these experimental methods may be what provide us opposite findings from both of them; we discuss this difference in Section 8.

More broadly, our work belongs to a growing body of literature evaluating how well neural language models capture a range of semantic phenomena (see Pavlick, 2022, for an overview). This includes work assessing the capacity of such models to capture compositionality (see, e.g., Ettinger et al., 2018; Shwartz and Dagan, 2019; Jawahar et al., 2019; Yu and Ettinger, 2020, 2021), as well as specific features such as negation (see, e.g., Ettinger et al., 2018; Kim et al., 2019; Ettinger, 2020; Jang and Lukasiewicz, 2023), quantification (e.g., Kim et al., 2019; Richardson et al., 2020; Cui et al., 2022), and monotonicity (e.g., Yanaka et al., 2019, 2020; Wijnholds, 2023).

3 Background

We focus on scope ambiguities involving quantifiers such as '*some*', '*every*', and '*most*', as well as quantifier-like determiners, like indefinites and numbers. (1) is an example of scope ambiguity arising out of quantifier-quantifier interactions. But scope ambiguities can also arise

Downloaded from http://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00670/2377773/tacl_a_00670.pdf by guest on 23 February 2026

out of quantifier-negation and quantifier-adverb interactions, as shown below:

Quantifier + Negation:

(3) Sita **doesn't** like **a** classmate of hers.

(4) a. Surface Reading: There is no classmate that Sita likes.

b. Inverse Reading: There is a specific classmate that Sita does not like.

Quantifier + Adverb:

(5) Bachi **usually** meets **two** professors.

(6) a. Surface Reading: Usually, Bachi meets any two professors, who are possibly different each time.

b. Inverse Reading: There are two professors who Bachi meets regularly.

In all three cases, the different readings have different truth conditions, and each is therefore logically compatible with a different set of propositions. As an illustration, consider our original example, reproduced here as (7):

(7) Every farmer owns a donkey.

(8) a. Each farmer has a different donkey.

b. All farmers have the same donkey.

(8a) is logically compatible with (7) only given the surface scope reading of the sentence, which states that each farmer has a potentially unique donkey. It is not logically compatible with the inverse scope reading of the sentence, which states that there is an individual donkey that all farmers (jointly) have. (8b), however, is also logically compatible with the inverse scope reading of the sentence. In Experiments 1A and 1B, we use these differences in logical compatibility to assess whether LLMs exhibit similar preferences to humans in the interpretation of scope ambiguous sentences.

Similarly, different scope readings often yield different effects in a discourse setting. Consider (5): Under the inverse scope reading, two professors are introduced as constant across the instances of Bachi's meetings. Consequently, they can therefore be further referred to in the discourse, as in (9a).

(9) a. He likes those two professors.

b. It's a different pair each time.

But under the surface scope reading of (5), there aren't necessarily two professors that are constant across instances, and who can therefore be further referred to in the discourse. As a result, (9a) is not an acceptable follow-up. The possible variability of the professors across multiple instances, however, does mean that (9b) is an acceptable follow-up (which it is not given the inverse scope reading). In Experiments 2A and 2B, we use such patterns of acceptable and unacceptable follow-ups to assess whether LLMs are sensitive to the presence of multiple readings of scope ambiguous sentences.

4 Experiment 1A

4.1 Method

In our first experiment, we assess whether LLMs show similar preferences to humans in how scope ambiguous sentences are interpreted. We frame this as a Q&A task. We present the LLM with sentences that are technically scope ambiguous, but have one strongly preferred scope reading (in some cases, this is a surface scope reading, and in others, it is an inverse scope reading). We then present the model with two possible statements based on this ambiguous sentence. One statement is compatible with only the surface scope reading of the ambiguous sentence, while the other statement is compatible with the inverse scope reading. In the case of chat-optimized models, we then ask the model which option is more likely; in the case of models not optimized for chat, we then obtain this answer through next token prediction. Finally, we observe whether these responses align with the reading preferred by most humans.

Figure 2 shows an example of how we conduct this method, using the dataset we develop for this experiment (details in Section 4.2). We concatenate, with newlines as shown in Figure 2, (i) a test sentence that is technically scope-ambiguous; (ii) an explanation that there are two options; (iii) two statements, labeled Option A and Option B, where one is compatible only with the surface scope reading, and the other is compatible with the inverse scope reading; and (iv) a

Downloaded from http://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00670/2377773/tacl_a_00670.pdf by guest on 23 February 2026

There are exactly six chairs evenly spaced around a circular table.

On the basis of this phrase/statement alone, and with no further context, there are two options:

Option A: the six chairs are all around the same table

Option B: the six chairs aren't all around the same table

Specifically in relation to this context, [the most likely option among these two is option] [which of these two options is most likely?]

Figure 2: An example of stimuli provided to models in Experiments 1A and 1B. The sections highlighted in bold are taken from our Experiment 1A dataset, and vary between individual stimuli presented to the models. The non-highlighted sections, which act as a prompt frame, remain fixed. For chat-optimized models, we solicit the model's response using the question highlighted in blue; for plain autoregressive models, we solicit the model's response by seeing what it predicts after the sequence highlighted in orange. In the control setting, the ambiguous sentence is dropped.

prompt that elicits the model's preferred choice. In the case of chat-optimized models, we observe the model's response to a question asking it to choose between the options. For other models, we observe the next token predicted by the model after the text '`the most likely option among these two is option`': this is either 'A' or 'B', corresponding to Option A and Option B, respectively. We treat these as the model's 'answer', and evaluate the model based on whether it aligns with preferred human readings.

Control: No Sentence in Prompt

One possible issue with this approach is that answers may depend more on the likeliness of the two options as general statements than on their likeliness *given* the ambiguous sentence. We therefore add a control: We remove the ambiguous sentence altogether from the stimulus, and present the rest of it to the model, conducting the same task as before. In this setting, the model should do significantly worse, as it is not exposed to the sentence that is being evaluated with respect to the two options. For instance, in the example in Figure 2, both options appear plausible in the absence of any context; following the scope-ambiguous sentence, however, only Option A is plausible. If model performance does not drop significantly when the ambiguous sentence is dropped, the model's performance in the original setting is likely unrelated to its processing of the ambiguous sentence, and instead reflects the background likeliness of each option.

4.2 Dataset

We build upon the quantifier scope dataset presented by AnderBois et al. (2012). We chose this dataset as a starting point because among the few existing scope ambiguity datasets (see Section 2), it was the dataset that had datapoints most appropriate to the focus of this study. This dataset contains around 1,700 sentences and phrases scraped from LSAT (Law School Admission Test) logic puzzles, and marked for quantifier scope where present. We filtered the dataset for instances of two interacting 'quantifiers'.[2] This narrowed it down to around 400 datapoints. Next, we manually constructed contrasting 'options' based on surface and inverse scope readings for whichever of these roughly 400 datapoints allowed this approach, giving us 186 sentences with accompanying contrasting statements.

To further ensure that these datapoints had strong scope reading preferences, we then conducted two rounds of human validation. In both rounds, we recruited participants via Prolific. Participants (38 in each round) were presented the filtered scope-ambiguous sentences along with two accompanying options, and were asked to pick the most likely option. In the first round, we reworded datapoints with low subject agreement; in the second round, we dropped any datapoints

[2]These included constructions such as '*per*' in '*one person per appointment*'—constructions that have some quantificational force, even if they aren't quantifiers in the strict sense—as well as instances of negation such as '*none*' in '*none of the cities*'.

Downloaded from http://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00670/2377773/tacl_a_00670.pdf by guest on 23 February 2026

Interaction Type	Example	Count	
		Experiment 1	Experiment 2
Quantifier-Quantifier	`Every laptop is facing a glitch`	136/153 (88.9%) **653/837 (78%)***	7/29 (24.1%) **38/110 (34.5%)**
Quantifier-Negation	`I didn't pass all of my exams`	11/153 (7.2%) **184/837 (22%)**	11/29 (37.9%) **35/110 (31.8%)**
Quantifier-Adverb	`I generally spar with two boxers`	– –	11/29 (37.9%) **37/110 (33.6%)**
Quantifier-Misc.	`Each truck is either green or red (but not both)`	6/153 (3.9%) –	– –

Table 1: Original Experiment 1A and 2A datasets (regular), as well as expanded versions used in Experiments 1B and 2B (bold), broken down by interaction type. Our original Experiment 1A dataset consisted of a few examples involving disjunction, as shown in the example above. We label these as a miscellaneous type of interaction, as they are not our primary focus. *These include a balanced set of 'quantifiers', including numbers, indefinites, and quantificational determiners.

with less than 75% agreement (all datapoints received at least 4 evaluations). For the datapoints that remained, gold labels (i.e., the correct option for each datapoint, and consequently, the preferred scope reading) were taken as the majority vote from study participants. This process ultimately yielded 153 scope ambiguous sentences, each with a pair of options.

Of these, 41 had an inverse scope reading preferred, while the remaining 112 had a surface scope reading preferred. Almost all were examples of scope ambiguities arising from quantifier-quantifier interactions, with a handful involving quantifier-negation interactions, and even fewer involving other types of interactions (see Table 1 for a breakdown by interaction type). As a final step, we duplicated each datapoint, but with a flipped order of options (i.e., 'Option A' was labeled 'Option B', and vice versa). This meant that while the distribution of preferred scope-readings remained skewed, the final dataset—which contains 306 datapoints covering 153 unique sentences—had an even distribution of correct answers (50% 'A' and 50% 'B').

4.3 Models

For all experiments, we choose to use autoregressive language models, due to their growing prevalence in practical applications using prompting.

Specifically, for this experiment, we use chat and vanilla versions of Llama 2 (Touvron et al., 2023) at 7B, 13B, and 70B sizes;[3] three variants of GPT-3/3.5 (Brown et al., 2020; Ouyang et al., 2022): `davinci`, `text-davinci-002`, and `text-davinci-003`; and GPT-4 (OpenAI, 2023). See Table 2 for a summary of key differences between these models.

Model	Size(s)	RLHF?
GPT-3-davinci	175B	×
GPT-3.5-text-davinci-002	Unclear	*
GPT-3.5-text-davinci-003	Unclear	✓
GPT-3.5-turbo	Unclear	✓
GPT-4	Unclear; possibly ensemble model	✓
Llama 2	7B, 13B, 70B	×
Llama 2 Chat	7B, 13B, 70B	✓
GPT-2	117M, 345M, 774M, 1.5B	×

Table 2: Summaries of models used for Experiments 1 and 2, including size and whether they were fine-tuned with reinforcement learning from human feedback (RLHF). *`text-davinci-002` is not fine-tuned with RLHF, but is fine-tuned on human demonstrations and highly rated model outputs.

4.4 Human Baselines

After the human feedback-based filtering mentioned above, we conducted another round of the same experiment with humans to get human baselines on this dataset. We are testing models on their ability to choose the scope readings preferred by most people—but how good are human

[3]For Llama 2 at 70B, we use a version of the model loaded in 8-bit (see Dettmers et al., 2022).

Downloaded from http://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00670/2377773/tacl_a_00670.pdf by guest on 23 February 2026

Source	Accuracy		Surface Acc.		Inverse Acc.	
	test	*control*	*test*	*control*	*test*	*control*
Human*	0.90	–	0.89	–	0.91	–
Llama2-7b	0.58	0.67	0.57	0.69	0.61	0.63
Llama2-7b-chat	0.54	0.55	0.54	0.58	0.54	0.49
Llama2-13b	0.71	0.63	0.73	0.65	0.65	0.59
Llama2-13b-chat	0.72	0.67	0.74	0.69	0.67	0.61
Llama2-70b	0.88	0.70	0.88	0.71	0.87	0.68
Llama2-70b-chat	0.85	0.71	0.88	0.72	0.78	0.67
GPT-3-davinci	0.58	0.58	0.58	0.58	0.57	0.57
GPT-3.5-td002	0.80	0.72	0.83	0.72	0.73	0.71
GPT-3.5-td003	0.91	0.75	0.94	0.79	0.84	0.67
GPT-3.5-turbo	0.80	0.67	0.81	0.66	0.77	0.66
GPT-4	**0.98**	**0.75**	**0.98**	**0.79**	**0.99**	**0.65**

Table 3: Results from Experiment 1A: model accuracy, as well as accuracy on sentences that had a preferred surface or inverse reading. In the test setting, the ambiguous sentence is present in the prompt; in the control setting it is dropped. *Values for humans are averaged across all participants' responses.

themselves at choosing the scope readings preferred by most other people? Our human baselines should provide a sense of the answer to this question. A total of 68 native speakers of English were recruited via Prolific for a repeat of the experimental set-up described in Section 4.2, but this time with the final dataset. Each participant was presented with 18 datapoints and evaluated on their answers. We then calculated overall accuracy as the total proportion of correct responses out of all responses.

4.5 Results

The results of Experiment 1A are shown in Table 3. Human responses yield an average accuracy of around 90%, suggesting that English speakers can, with a high degree of accuracy, arrive at scope reading preferences shared by most other people.

When it comes to model responses, although models like `davinci` and `Llama2-7b` do not perform far above chance (50%, since correct answers in the dataset were balanced through duplication), several other models do achieve high performance—both versions of `Llama2-70b`, as well as all the GPT-3.5 models achieve an accuracy of 80% or more in the test setting, while GPT-4 achieves 98%, close to the ceiling.

The control setting adds further insights to these results. Neither `davinci` nor the versions of Llama 2 at 7B see their accuracy scores drop when prompts are provided without the actual scope ambiguous sentence (`Llama2-7b` actually produces a higher accuracy in this setting), suggesting that performance in the test setting is not a result of the models' processing of the ambiguous input, but primarily driven by the background likeliness of the two options. The models with higher accuracy scores, however, see more severe drop-offs in the control setting, most notably with GPT-4, which sees its accuracy drop to 75% in the control setting.

4.6 Discussion

These results suggest that the more advanced LLMs evaluated—GPT-3.5, Llama 2 at 70B, and most notably GPT-4—are able to exhibit similar scope reading preferences as humans, with a high level of accuracy. Smaller or less advanced models, however, such as Llama 2 at 7B, appear to fail.

Also worth noting is the fact that, for almost all models, performance on sentences that had a preferred inverse scope reading was lower than on those that had a preferred surface scope reading. This aligns with literature suggesting that inverse scope readings are generally harder to access than surface readings (see, e.g., Kurtzman and MacDonald, 1993; AnderBois et al., 2012), but curiously, does not align with the behavior of humans in this experiment, who showed no such dispreference.

The deeper implication of some of the models' high performance, however, is that LLMs can not only capture different types of readings—surface and inverse, which correspond to different semantic structures—but also integrate background world knowledge in their behavioral preferences when confronted with scope ambiguous constructions.

5 Experiment 2A

5.1 Method

Our first experiment showed us that, where there is a clear preferred reading, LLMs can mimic human preferences in the interpretation of scope ambiguous sentences. It did not, however, indicate whether or not LLMs were sensitive to the fact that each such scope ambiguous sentence had more than one reading. This question is the focus of our second experiment.

Downloaded from http://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00670/2377773/tacl_a_00670.pdf by guest on 23 February 2026

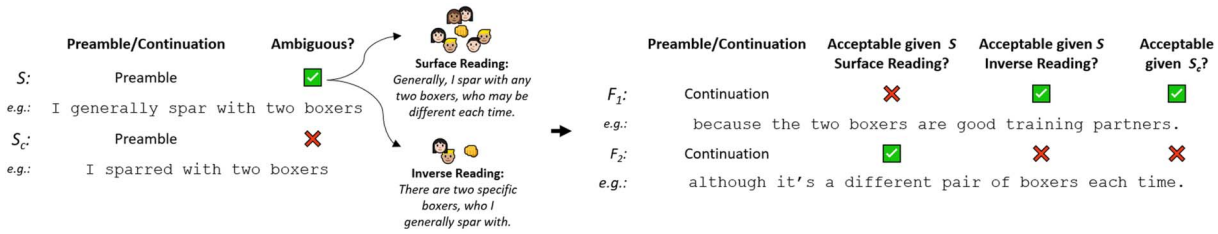

Figure 3: Experiment 2A and 2B set-up, comprising of an ambiguous sentence *S*, unambiguous control S_c, and two follow-ups, F_1 and F_2, demonstrated using an example from our manually constructed dataset. We compare the probabilities a model assigns to F_1 and F_2 as continuations to *S*, versus as continuations to S_c.

Here, we assess whether models exhibit different behavior for scope ambiguous sentences than they do for similar, *non* scope ambiguous sentences, in a manner that indicates a sensitivity to the meaning ambiguity in the former but not the latter. We do not frame this as a conventional goal-oriented task such as Q&A.[4] Instead, following work that brings psycholinguistic methods to language model analysis (see Linzen et al., 2016; Futrell et al., 2019; Ettinger, 2020; Baroni, 2022; Schuster and Linzen, 2022), we investigate the question by observing the probabilities a model assigns to different types of continuations given a scope-ambiguous sentence.

Figures 1 and 3 illustrate the general set-up we employ: We begin by presenting the LLM with a scope ambiguous sentence S. We then observe the probabilities the model assigns to two follow-ups to S, labeled F_1 and F_2. F_1 is an acceptable continuation to S only given the inverse scope reading of S, while F_2 is an acceptable continuation to S only given the surface scope reading of S. We then compare these probabilities with those the model assigns to F_1 and F_2 given a control sentence, S_c. S_c is highly similar in both syntax and semantics to S, but differs in that it is not scope ambiguous; F_1 remains an acceptable continuation to S_c, though F_2 does not.

If an LLM successfully captures the meaning ambiguity of a sentence like S, we would expect the ratio of probabilities it assigns to F_1 and F_2 as continuations to S to be smaller than the ratio of probabilities it assigns to the same follow-ups as continuations to S_c. In other words, we expect the following inequality to hold:

$$(10) \quad \frac{P(F_1|S)}{P(F_2|S)} < \frac{P(F_1|S_c)}{P(F_2|S_c)}$$

This is because while S is ambiguous between two readings, and therefore allows for both F_1 and F_2 as continuations, S_c has only one reading, and allows only for F_1 as a continuation (see Figure 3). If an LLM upholds (10), it thus provides evidence of capturing the fact that S is ambiguous between two readings, in a way that non scope-ambiguous sentences, even if syntactically and semantically similar, are not.

For a more thorough analysis, we also observe the degree to which $P(F_1|S_c) : P(F_2|S_c)$ is greater than $P(F_1|S) : P(F_2|S)$. We measure the difference in the log ratios of F_1 and F_2 given S, and F_1 and F_2 given S_c, and use it to calculate what we call the model's ambiguity recognition score, or 'α-score':

$$(11) \quad \alpha = -[[logP(F_1|S) - logP(F_2|S)] \\ -[logP(F_1|S_c) - logP(F_2|S_c)]]$$

If the inequality in (10) holds, the α-score will be positive; the larger its value, the greater the difference in ratios.

5.2 Dataset

Existing scope ambiguity datasets (see Section 2) are (i) few in number, and (ii) generally involve examples where one scope reading is strongly preferred over the other. While we made use of this second observation in our first experiment, where we assessed whether LLMs exhibited similar scope reading preferences as humans, it is

[4]This was mainly due to our focus on examples without one strongly preferred reading (see Section 5.2). Without one strongly preferred reading, tasks like Q&A, which require a 'right' and 'wrong' answer, are difficult to implement.

Downloaded from http://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00670/2377773/tacl_a_00670.pdf by guest on 23 February 2026

a significant problem for the current experiment, due to its aims.

In this experiment we aim to test LLMs for their sensitivity to the presence of multiple readings of a scope ambiguous sentence. But in such cases, if humans themselves find one of these readings very hard to access without further context, it would be unfair to expect models to do so. For a fair evaluation, therefore, it is crucial that we use sentences which do not have one reading strongly preferred over another.

We therefore construct a small-scale dataset, consisting of 38 manually handcrafted datapoints, where each datapoint includes a scope ambiguous sentence (S), a matching non scope-ambiguous control sentence (S_c), and two follow-up phrases (F_1 and F_2), yielding a total of 152 sentence-continuation pairs. For further validation, these datapoints were then filtered through our human baselines: Any datapoints that yielded negative human-derived scores (details in Section 5.3) were dropped, as such scores indicated that these were datapoints for which the inequality in (10) did not align with human judgments. This left us with 29 unique datapoints, yielding 116 unique sentence-continuation pairs (see Table 1 for a breakdown by interaction type).

5.3 Human Baselines

Since the current experiment involves the analysis of probabilities assigned to text sequences—something not directly replicable with humans—we use proxy scores derived from a crowdsourced judgment task as our human baselines. We conducted a crowdsourced study via Prolific, involving 140 native speakers of English; each was presented random sentence-continuation pairs from the dataset, and asked to provide ratings from 1 to 7 on how 'natural-sounding' the continuation was to the sentence. From these ratings, we computed the mean score for each sentence-continuation pair, and normalized them to be in an interval between 0 and 1. We then treat these normalized scores as we treat model-assigned probabilities when calculating α-scores and label the negative difference of log ratios our proxy α-score.

This proxy score gives us an indirect means by which to compare human judgments of scope-ambiguous sentences and continuations with LLM-assigned probabilities of the latter given the former. Just as in the case of α-scores for models, we would expect human proxy scores to be positive.

5.4 Models

As in Experiment 1A, we work with autoregressive LLMs. Unlike in Experiment 1A, however, the current experimental set-up allows us to also work with models ill-suited to zero-shot contexts. We therefore ran this experiment not only on the models tested before, but also on several smaller variants of GPT-2 (Radford et al., 2019): small (117M params), medium (345M params), large (774M params), and XL (1.5B params).[5] The reliance on probabilities, however, forces us to omit `GPT-3.5-turbo` and GPT-4, for which sequence log-probabilities are not accessible.

5.5 Results

5.5.1 Mean Scores

We first compute mean α and proxy scores, along with p-values derived from paired t-tests.[6] Positive, statistically significant mean scores point to an overall sensitivity to the meaning ambiguity of the sentences in the dataset; comparing between models, higher mean scores also suggest stronger overall sensitivity.[7] Table 4 shows our results. All models yield positive mean α-scores; and barring the case of `GPT-2-small`, all are statistically significant at a threshold of $p < 0.05$.

5.5.2 Correlations Between Model and Human Scores

We also compute the correlation between model-derived α-scores and human-derived proxy scores, to see how model behavior aligns with human judgments; if the two do align well, we expect to see a strong, positive correlation between them.

Table 4 shows model-wise Pearson correlation coefficients between α-scores and human proxy scores, along with corresponding p-values. Many

[5]Probabilities from GPT-2 and Llama 2 models were extracted using the `minicons` library (Misra, 2022).

[6]We choose this statistical measure as both α and proxy scores are calculated as paired differences of differences.

[7]While the latter is true between models, this relationship is less clear when comparing model scores and human scores. The main reason is that while human scores are derived from a bounded set of ratings between 1 and 7, model log-probabilities practically have no negative bound, allowing for more extreme differences between them. Since α-scores are computed as differences of log differences (see (11)), it is thus possible for them to be much higher than derived human scores, purely on account of being unbounded below zero.

Downloaded from http://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00670/2377773/tacl_a_00670.pdf by guest on 23 February 2026

Source	Mean Scores		Correlations		$\alpha > 0$
	α/proxy score	p-value	R-value	p-value	
Human	1.22	4.43e-06	1.0	–	1.0
GPT-2-small	0.29	0.43	0.32	0.09	0.52
GPT-2-med	0.97	0.03	0.38	0.04	0.62
GPT-2-large	1.51	1.97e-03	0.33	0.08	0.69
GPT-2-xl	1.79	9.78e-04	0.29	0.12	0.76
Llama2-7b	3.89	2.9e-07	0.17	0.39	0.93
Llama2-7b-chat	5.03	9.45e-07	0.27	0.16	0.86
Llama2-13b	3.62	6.67e-07	0.49	7.38e-03	0.90
Llama2-13b-chat	4.54	1.28e-05	0.53	3.31e-03	0.83
Llama2-70b	3.97	5.74e-08	0.38	0.04	0.93
Llama2-70b-chat	4.72	3.52e-06	0.43	0.02	0.90
GPT-3-davinci	3.77	4.95e-08	0.20	0.31	0.93
GPT-3.5-td002	4.06	1.5e-04	0.41	0.03	0.83
GPT-3.5-td003	**8.36**	**1.44e-06**	**0.62**	**2.93e-04**	**1.0**

Table 4: Results from Experiment 2A. Mean Scores: Mean α (for models) and proxy (for humans) scores with p-values from paired t-tests ($df = 28$). Correlations: Pearson correlation coefficients (R-values) between each model's α scores and human proxy scores, with derived p-values ($n = 29$). $\alpha > 0$: Proportion of datapoints where α/proxy score was positive.

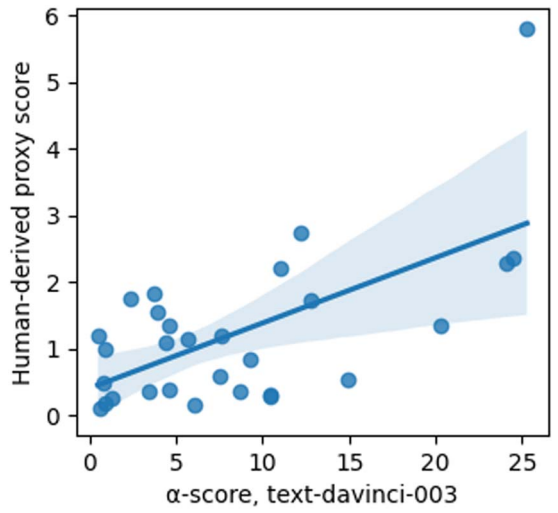

Figure 4: From Experiment 2A—scatterplot of α-scores produced by `text-davinci-003`, against human proxy scores for the same datapoints.

models fail to produce correlations that are significant at $p < 0.05$; `text-davinci-003` and Llama 2 at 13B, however, both produce highly significant correlation scores. The former also produces the highest correlation score, at around 0.62. See Figure 4 for a scatterplot of its α-scores against corresponding human proxy scores; the plot lends further evidence to this correlation.

5.5.3 Proportion of Positive α-Scores

Lastly, we compute the proportion of datapoints for which models produced positive α-scores, which allows us to assess whether the models behave consistently across datapoints. Like with correlations, we observe an effect of model size: Larger models perform well, with several yielding positive α-scores for over 90% of the data, and once again, `text-davinci-003` performs the best.

5.6 Discussion

These results suggest that a wide range of LLMs may be sensitive to the meaning ambiguity in scope ambiguous sentences. The positive mean α-scores provide evidence that larger or more powerful models (i.e., those besides GPT-2 small) distinguish between scope ambiguous and non-scope ambiguous sentences in a manner consistent with their meanings. Similarly, the statistically significant correlations we see between some models' α-scores and human proxy scores suggest that, at least for certain models, this behavior correlates well with human judgments. These models also produce positive α-scores for a high proportion of the data, indicating a high level consistency in this behavior.

Comparing chat and vanilla versions of Llama 2 also reveals an interesting pattern. As Table 4 shows, chat versions of Llama 2 produce slightly higher mean α-scores and correlations than their non-chat equivalents, but also lower proportions of positive α-scores—indicating increased alignment with human judgments on several sentences, but lower overall consistency.

The broader takeaway, however, is that several LLMs appear sensitive to a meaning ambiguity that arises from the presence of different possible semantic structures, which vary vis-à-vis scope relations. Consequently, although the current work does not investigate how or where models represent scope, these results suggest that LLMs capture scope-related phenomena.

6 Re-evaluation on Expanded Datasets

The results from Experiments 1A and 2A are promising, but rely on relatively small datasets that contain 153 and 29 unique datapoints, respectively—raising questions of how generalizable our results are. As a follow-up, we therefore considerably expand these two datasets, and rerun the same experiments described in Sections 4 and 5 on the expanded datasets.

Downloaded from http://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00670/2377773/tacl_a_00670.pdf by guest on 23 February 2026

6.1 Dataset Expansion Process

Experiment 1 Dataset Expansion To expand our Experiment 1A dataset, we begin by annotating it for both semantic operator and quantifier types: whether an ambiguity arose out of a negation-quantifier or quantifier-quantifier interaction, as well as whether the quantifier was an existential quantifier, universal quantifier, number or indefinite. Combined with scope reading preference labels (see Section 4.2), this gave us 13 categories of scope reading and operator combinations (e.g., `negation-indefinite_surface`, or `number-universal_inverse`). Following this categorization, we add manually handcrafted examples to any sparse categories, so each contains at least 10 unique datapoints.

We then use GPT-4 to expand the dataset. For each category in our annotated dataset, we randomly sample 5 datapoints, and instruct GPT-4 to produce 10 novel datapoints based on them. We repeat this process ten times, such that we have 100 datapoints generated from each of our annotated categories. We then manually inspect the combined 1,300 generated datapoints, removing duplicates, and dropping or editing low-quality datapoints; this left us with 1,062 datapoints. Finally, we run a crowd-sourced study via Prolific (278 participants, ∼5 ratings per datapoint), similar to those described in Sections 4.2 and 4.4, to obtain our gold labels for preferred scope readings, and filter out any datapoints that received low inter-subject agreement. This process eventually yields 837 unique scope-ambiguous sentences (with accompanying 'options', and human preference labels); 534 receive a preferred surface scope reading, and 303 an inverse scope reading.

Experiment 2 Dataset Expansion We use a process similar to the one for Experiment 1A.

We first split our Experiment 2A dataset into categories based on whether the datapoints involve negation-quantifier, adverb-quantifier or quantifier-quantifier interactions. We then run the same sampling, generation and manual filtering process as with the Experiment 1A dataset, giving us 126 datapoints from 300 GPT-4 generated datapoints. Finally, we run another study via Prolific (223 participants, ∼8 ratings per sentence-follow-up pair), and use this human judgement data to further filter the dataset. Our final dataset consists of 110 unique datapoints, where each datapoint consists of a scope-ambiguous sentence, control sentence, follow-up supporting an inverse scope reading, and follow-up supporting a surface scope reading.

Source	Accuracy		Surface Acc.		Inverse Acc.	
	test	*control*	*test*	*control*	*test*	*control*
Llama2-7b	0.64	0.64	0.62	0.63	0.66	0.65
Llama2-7b-chat	0.57	0.58	0.58	0.58	0.56	0.58
Llama2-13b	0.71	0.66	0.75	0.66	0.65	0.67
Llama2-13b-chat	0.75	0.67	0.77	0.64	0.73	0.73
Llama2-70b	0.89	0.72	0.91	0.74	0.84	0.69
Llama2-70b-chat	0.83	0.65	0.83	0.64	0.82	0.67
GPT-3-davinci	0.64	0.60	0.68	0.62	0.58	0.58
GPT-3.5-td002	0.84	0.70	0.89	0.69	0.74	0.72
GPT-3.5-td003	0.87	0.70	0.90	0.72	0.80	0.68
GPT-3.5-turbo	0.79	0.65	0.86	0.65	0.68	0.65
GPT-4*	**0.96**	**0.72**	**0.97**	**0.73**	**0.93**	**0.70**

Table 5: Results from Experiment 1B: model accuracy, as well as accuracy on sentences that had a preferred surface or inverse reading. In the test setting, the ambiguous sentence is present in the prompt; in the control setting it is dropped. *Expanded dataset was also generated by `GPT-4`, albeit in a different setting and with a different system prompt.

6.2 Experiment 1B

We re-run Experiment 1A on the expanded dataset; Table 5 shows our results. As can be seen, the general patterns observed in Experiment 1A (see Section 4.5) continue to hold true even when the models are evaluated on a much larger dataset. While some models either perform around chance or do not show a major accuracy drop in the control setting, models like GPT-4, `text-davinci-003` and `Llama2-70b` show both high performance (all above 85% accuracy in the test setting, with GPT-4 achieving ∼96% accuracy, albeit on data it produced in a separate context), as well as a drop-off in performance in the control setting.

6.3 Experiment 2B

Similarly, we re-run Experiment 2A on the expanded dataset; Table 6 shows our results. As with Experiments 1A and 1B, the general patterns observed continue to hold for the expanded dataset. All models still produce positive mean α-scores. Though not as high as on the original dataset, `text-davinci-003` once again produces the highest correlation with human data, with a R-value of roughly 0.48. Similarly,

Downloaded from http://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00670/2377773/tacl_a_00670.pdf by guest on 23 February 2026

Source	Mean Scores		Correlations		$\alpha > 0$
	α/proxy score	p-value	R-value	p-value	
Human	1.34	5.36e-23	1.0	–	1.0
GPT-2-small	1.38	3.78e-09	0.25	7.73e-03	0.80
GPT-2-med	1.88	1.58e-11	0.29	1.88e-03	0.79
GPT-2-large	1.98	1.38e-11	0.37	5.87e-05	0.76
GPT-2-xl	2.87	7.08e-17	0.32	5.59e-04	0.86
Llama2-7b	3.94	1.67e-19	0.37	7.95e-05	0.88
Llama2-7b-chat	5.21	3.05e-20	0.42	4.38e-06	0.89
Llama2-13b	4.31	5.02e-23	0.44	1.31e-06	**0.92**
Llama2-13b-chat	5.12	4.45e-21	0.46	3.37e-07	0.89
Llama2-70b	4.64	2.26e-22	0.36	1.32e-04	0.88
Llama2-70b-chat	5.56	1.01e-20	0.46	3.65e-07	0.85
GPT-3-davinci	4.16	2.84e-18	0.37	6.01e-05	0.85
GPT-3.5-td002	4.69	1.99e-20	0.48	1.51e-07	0.87
GPT-3.5-td003	**7.05**	**7.17e-22**	**0.48**	**1.09e-07**	0.90

Table 6: Results from Experiment 2B. Mean Scores: Mean α (for models) and proxy (for humans) scores with p-values from paired t-tests ($df = 109$). Correlations: Pearson correlation coefficients (R-values) between each model's α scores and human proxy scores, with derived p-values ($n = 110$). $\alpha > 0$: Proportion of datapoints where α/proxy score was positive.

most models show a high level of consistency in their behavior, producing positive α-scores from, in the case of `text-davinci-003` and `Llama2-13b`, over 90% of the data.[8] Once again, Llama 2 chat models produce higher correlations and mean scores than their vanilla counterparts, but in two out of three cases, lower proportions of positive α-scores.

7 Discussion

Our results, which indicate that LLMs are both proficient at choosing the scope readings preferred by most humans, and sensitive to the meaning ambiguity in scope ambiguous constructions, offer further evidence of the capacity of large language models to induce semantic structure (see Pavlick, 2022), and linguistic structure more generally (see Linzen and Baroni, 2021; Baroni, 2022).

On the other hand, these results contrast with closely related work by Liu et al. (2023) and Stengel-Eskin et al. (2023), who both find that LLMs struggle to model ambiguity in zero-shot

[8]GPT-2 models also do much better on the expanding dataset. This may be because in the expansion process, we ensured that the contrast between acceptable and unacceptable sentence-follow-up pairs (see Figure 3) was more clear cut than in the original dataset, often coming from grammatical cues, rather than world knowledge cues. GPT-2 may recognize the former more than the latter, and thus perform better here.

Model	Mean T/F Accuracy	Mean T/F Prob. Density
Llama2-7b	0.54	0.43
Llama2-7b-chat	0.44	0.99
Llama2-13b	0.51	0.24
Llama2-13b-chat	0.58	0.99
Llama2-70b	0.59	0.31
Llama2-70b-chat	0.57	0.98
GPT-3.5-turbo	0.64	NA
GPT-4	0.64	NA

Table 7: Results from running Liu et al.'s (2023) T/F evaluation. Mean T/F Accuracy: Average accuracy of model's responses. Mean T/F Prob. Density: Average probability density of the union of 'True' and 'False' tokens as responses given the prompt input.

contexts. What explains this contrast? One possible explanation is the difference in methodologies used. We assess models using Q&A- and probability-based approaches (see Sections 4.1 and 5.1) that implicitly test models' access to different scope readings. Liu et al. (2023), on the other hand, mostly use prompting-based approaches that elicit model responses on what an ambiguous sentence may mean or entail, and Stengel-Eskin et al. (2023) assess models in terms of their abilities to logically parse ambiguous inputs. It is possible that LLMs *implicitly* capture meaning ambiguities and human-preferred interpretations, but cannot reliably produce meta-linguistic judgments or logical translations consistent with this information. This is would be in line with findings from Hu and Levy (2023), which suggest meta-linguistic prompting-based approaches may underestimate LLMs' linguistic abilities.

To test this theory, we adapt a random sample of our Experiment 2B dataset to the format Liu et al. (2023) use in their True/False evaluation of models (see Section 4.2 of Liu et al., 2023). In this format, models are prompted to answer whether it is true or false that, given one of its disambiguations, an ambiguous input *may*, *may not*, *cannot*, or *can only* mean the disambiguation.

We rerun their experiment on this subset of our data; our results, shown in Table 7, are similar to the authors' findings on their own data. Most models do poorly on this task, performing around chance (50%), with GPT-4 and `GPT-3.5-turbo` only achieving 64% accuracy. As shown in Section 6.3, however, using the dataset in our experimental format yields positive

Downloaded from http://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00670/2377773/tacl_a_00670.pdf by guest on 23 February 2026

results that contrast this poor performance. This divergence highlights the importance of diverse approaches to investigating the linguistic capacities of language models; our results suggest that probability- and prompting-based methods may yield differing conclusions.

8 Conclusion

In this paper, we investigated how different autoregressive language models treat scope ambiguities. In doing so, we introduced novel datasets that contain a joint total of roughly 1,000 unique and diverse scope-ambiguous sentences, annotated for human judgments—the largest of this kind. Our results indicate that LLMs are able to exhibit behavior in line with human preferences of interpretation—informed at least in part by background knowledge—as well as compatible with different types of semantic structures. Finally, the contrast between our findings and those of other recent work emphasizes the need for diverse approaches in assessing the linguistic capacities of large language models.

9 Limitations

Aside from its focus only on English, one constraint of this work is that it does not assess how context affects scope reading preferences.

(12) Ada often studies with a few of her friends.

a. Context: Ada finds it hard to study alone, so she generally invites others for joint study sessions.

b. Context: Ada, Rohan, and Jo are good friends in the same program, and prepare for exams together.

(12) is ambiguous between a surface scope reading ((12) refers to no friends in particular) and an inverse scope reading ((12) refers to some specific friends). Different background contexts can prompt different readings: (12a) prompts the surface scope reading, while (12b) prompts the inverse scope reading. Our work does not address such effects.

At a higher level, while this work shows how LLMs treat scope-ambiguous inputs, it also does not reveal how or where models represent scope. Parallel work on model interpretability (such as causal mediation analysis, e.g., Vig et al., 2020; Finlayson et al., 2021; Geiger et al., 2022) could provide exciting insights to this question.

Ethics Statement

There are no obvious risks or harms associated with this experimental study. Experiments that involved human participants were approved by our university's ethics board. Human participants in our studies were recruited online via Prolific, and were paid on average a minimum of US$15.00 per hour. Wherever anonymous participant responses were to be made publicly accessible, participants were informed of how their responses would be used for the purposes of the study, and their rights with regard to their submitted data.

Acknowledgments

This work was partly funded by a Doctoral Training Award from the *Fonds de Recherche du Québec—Société et Culture*. We also thank Scott AnderBois, Justyna Grudzińska, and the Law School Admission Council for access to the materials used in Experiment 1A, as well as Kristina Toutanova and the anonymous reviewers for their valuable feedback on prior versions of this work.

References

Badr AlKhamissi, Millicent Li, Asli Celikyilmaz, Mona Diab, and Marjan Ghazvininejad. 2022. A review on language models as knowledge bases. *arXiv preprint arXiv:2204.06031*.

Scott AnderBois, Adrian Brasoveanu, and Robert Henderson. 2012. The pragmatics of quantifier scope: A corpus study. In *Proceedings of Sinn und Bedeutung*, volume 16, pages 15–28.

Catherine Anderson. 2004. *The Structure and Real-time Comprehension of Quantifier Scope Ambiguity*. Northwestern University.

Galen Andrew and Bill MacCartney. 2004. Statistical resolution of scope ambiguity in natural language. *Unpublished Manuscript*.

Marco Baroni. 2022. On the proper role of linguistically-oriented deep net analysis in linguistic theorizing. *Algebraic Structures in*

Downloaded from http://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00670/2377773/tacl_a_00670.pdf by guest on 23 February 2026

Natural Language, pages 1–16. https://doi.org/10.1201/9781003205388-1

Yonatan Belinkov and James Glass. 2019. Analysis methods in neural language processing: A survey. *Transactions of the Association for Computational Linguistics*, 7:49–72. https://doi.org/10.1162/tacl_a_00254

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D. Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33:1877–1901.

Ruixiang Cui, Daniel Hershcovich, and Anders Søgaard. 2022. Generalized quantifiers as a source of error in multilingual NLU benchmarks. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4875–4893, Seattle, United States. Association for Computational Linguistics.

Tim Dettmers, Mike Lewis, Younes Belkada, and Luke Zettlemoyer. 2022. GPT3.int8(): 8-bit matrix multiplication for transformers at scale. In *Advances in Neural Information Processing Systems*, volume 35, pages 30318–30332. Curran Associates, Inc.

Allyson Ettinger. 2020. What BERT is not: Lessons from a new suite of psycholinguistic diagnostics for language models. *Transactions of the Association for Computational Linguistics*, 8:34–48. https://doi.org/10.1162/tacl_a_00298

Allyson Ettinger, Ahmed Elgohary, Colin Phillips, and Philip Resnik. 2018. Assessing composition in sentence vector representations. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1790–1801, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Matthew Finlayson, Aaron Mueller, Sebastian Gehrmann, Stuart Shieber, Tal Linzen, and Yonatan Belinkov. 2021. Causal analysis of syntactic agreement mechanisms in neural language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1828–1843, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.acl-long.144

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, Minneapolis, Minnesota. Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1004

Atticus Geiger, Zhengxuan Wu, Hanson Lu, Josh Rozner, Elisa Kreiss, Thomas Icard, Noah Goodman, and Christopher Potts. 2022. Inducing causal structure for interpretable neural networks. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 7324–7338. PMLR.

Benjamin Heinzerling and Kentaro Inui. 2021. Language models as knowledge bases: On entity representations, storage capacity, and paraphrased queries. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1772–1791, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.eacl-main.153

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.

Downloaded from http://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00670/2377773/tacl_a_00670.pdf by guest on 23 February 2026

Derrick Higgins and Jerrold M. Sadock. 2003. A machine learning approach to modeling scope preferences. *Computational Linguistics*, 29(1):73–96. https://doi.org/10.1162/089120103321337449

Jennifer Hu and Roger Levy. 2023. Prompt-based methods may underestimate large language models' linguistic generalizations. *arXiv preprint arXiv:2305.13264*.

Myeongjun Jang and Thomas Lukasiewicz. 2023. Consistency analysis of ChatGPT. *arXiv preprint arXiv:2303.06273*.

Ganesh Jawahar, Benoît Sagot, and Djamé Seddah. 2019. What does BERT learn about the structure of language? In *ACL 2019-57th Annual Meeting of the Association for Computational Linguistics*. https://doi.org/10.18653/v1/P19-1356

Najoung Kim, Roma Patel, Adam Poliak, Patrick Xia, Alex Wang, Tom McCoy, Ian Tenney, Alexis Ross, Tal Linzen, Benjamin Van Durme, Samuel R. Bowman, and Ellie Pavlick. 2019. Probing what different NLP tasks teach machines about function word comprehension. In *Proceedings of the Eighth Joint Conference on Lexical and Computational Semantics (*SEM 2019)*, pages 235–249, Minneapolis, Minnesota. Association for Computational Linguistics. https://doi.org/10.18653/v1/S19-1026

Howard S. Kurtzman and Maryellen C. MacDonald. 1993. Resolution of quantifier scope ambiguities. *Cognition*, 48(3):243–279. https://doi.org/10.1016/0010-0277(93)90042-T, PubMed: 8269698

Aleksander Leczkowski, Justyna Grudzińska, Manuel Vargas Guzmán, Aleksander Wawer, and Aleksandra Siemieniuk. 2022. Prepositions matter in quantifier scope disambiguation. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3960–3970.

Tal Linzen and Marco Baroni. 2021. Syntactic structure from deep learning. *Annual Review of Linguistics*, 7:195–212. https://doi.org/10.1146/annurev-linguistics-032020-051035

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. https://doi.org/10.1162/tacl_a_00115

Alisa Liu, Zhaofeng Wu, Julian Michael, Alane Suhr, Peter West, Alexander Koller, Swabha Swayamdipta, Noah A. Smith, and Yejin Choi. 2023. We're afraid language models aren't modeling ambiguity. *arXiv preprint arXiv:2304.14399*. https://doi.org/10.18653/v1/2023.emnlp-main.51

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*.

Mehdi Manshadi and James Allen. 2011. Unrestricted quantifier scope disambiguation. In *Proceedings of TextGraphs-6: Graph-based Methods for Natural Language Processing*, pages 51–59.

Kanishka Misra. 2022. minicons: Enabling flexible behavioral and representational analyses of transformer language models. *arXiv preprint arXiv:2203.13112*.

OpenAI. 2023. GPT-4 technical report.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

Ellie Pavlick. 2022. Semantic structure in deep learning. *Annual Review of Linguistics*, 8(1):447–471. https://doi.org/10.1146/annurev-linguistics-031120-122924

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019.

Downloaded from http://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00670/2377773/tacl_a_00670.pdf by guest on 23 February 2026

Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9.

Nathan Ellis Rasmussen. 2022. *Broad-domain Quantifier Scoping with RoBERTa*. Ph.D. thesis, The Ohio State University.

Kyle Richardson, Hai Hu, Lawrence Moss, and Ashish Sabharwal. 2020. Probing natural language inference models through semantic fragments. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8713–8721. https://doi.org/10.1609/aaai.v34i05.6397

Adam Roberts, Colin Raffel, and Noam Shazeer. 2020. How much knowledge can you pack into the parameters of a language model? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5418–5426, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.emnlp-main.437

Walid S. Saba and Jean-Pierre Corriveau. 2001. Plausible reasoning and the resolution of quantifier scope ambiguities. *Studia Logica*, 67:271–289. https://doi.org/10.1023/A:1010503321412

Sebastian Schuster and Tal Linzen. 2022. When a sentence does not introduce a discourse entity, transformer-based models still sometimes refer to it. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 969–982, Seattle, United States. Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.naacl-main.71

Vered Shwartz and Ido Dagan. 2019. Still a pain in the neck: Evaluating text representations on lexical composition. *Transactions of the Association for Computational Linguistics*, 7:403–419. https://doi.org/10.1162/tacl_a_00277

Elias Stengel-Eskin, Kyle Rawlins, and Benjamin Van Durme. 2023. Zero and few-shot semantic parsing with ambiguous inputs. *arXiv preprint arXiv:2306.00824*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

K. C. Tsiolis. 2020. Quantifier scope disambiguation. *Unpublished manuscript*.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Yaron Singer, and Stuart Shieber. 2020. Investigating gender bias in language models using causal mediation analysis. *Advances in Neural Information Processing Systems*, 33:12388–12401.

Gijs Wijnholds. 2023. Assessing monotonicity reasoning in Dutch through natural language inference. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 1464–1470. https://doi.org/10.18653/v1/2023.findings-eacl.110

Hitomi Yanaka, Koji Mineshima, Daisuke Bekki, and Kentaro Inui. 2020. Do neural models learn systematicity of monotonicity inference in natural language? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6105–6117, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.543

Hitomi Yanaka, Koji Mineshima, Daisuke Bekki, Kentaro Inui, Satoshi Sekine, Lasha

Downloaded from http://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00670/2377773/tacl_a_00670.pdf by guest on 23 February 2026

Abzianidze, and Johan Bos. 2019. Can neural networks understand monotonicity reasoning? In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 31–40, Florence, Italy. Association for Computational Linguistics. https://doi.org/10.18653/v1/W19-4804

Lang Yu and Allyson Ettinger. 2020. Assessing phrasal representation and composition in transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4896–4907, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.emnlp-main.397

Lang Yu and Allyson Ettinger. 2021. On the interplay between fine-tuning and composition in transformers. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2279–2293, Online. Association for Computational Linguistics.

Downloaded from http://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00670/2377773/tacl_a_00670.pdf by guest on 23 February 2026