Disentangling Form and World Knowledge in LLM Interpretation: Evidence from Quantifier Scope Disambiguation

Anonymous ACL submission

Abstract

We investigate how large language models (LLMs) construct meaning using Quantifier Scope Disambiguation (QSD) as a controlled probing task. We introduce (i) a balanced English corpus designed to neutralize classical surface heuristics and systematically manipulate the relationship between formal predictors and preferred readings, and (ii) a pseudosentence dataset that removes real-world referents to isolate formal cues from world-knowledge support. In Experiment 1, we evaluate a range of LLMs in a zero-shot question–answering setup and compare them to human baselines. While models achieve high accuracy on the balanced corpus, performance drops substantially on pseudosentences. Crucially, robust performance is retained only when classical formal predictors converge on the preferred reading; once these predictors partially diverge or fully conflict, accuracy approaches chance levels and becomes highly unstable across models. These findings suggest that the strong QSD performance reported for current LLMs may reflect a masking effect caused by the systematic alignment of surface-level heuristics and world knowledge in naturalistic datasets. In Experiment 2, we manipulate access to external world knowledge via retrieval-augmented generation (RAG) while keeping the task and prompt fixed. RAG yields only limited overall gains, but selectively improves the hardest configuration, in which both classical surface predictors conflict with the preferred reading. Taken together, our results suggest that high QSD performance in LLMs remains strongly dependent on convergence between formal cues and world knowledge, despite the apparent semantic robustness suggested by naturalistic QSD results.

1 Introduction

Large language models (LLMs) have achieved remarkable success across a wide range of NLP tasks (e.g., Devlin et al. 2019; Liu et al. 2019). In particular, prior work has shown that LLMs capture substantial amounts of world knowledge (e.g., Roberts et al. 2020; Heinzerling and Inui 2021; AlKhamissi et al. 2022) and exhibit sensitivity to various aspects of natural language semantics (e.g., Jawahar et al. 2019; Ettinger et al. 2018; Ettinger 2020; Richardson et al. 2020; Sevastjanova et al. 2021; Baroni 2022; Kalouli et al. 2022). At the same time, several recent studies—including Liu et al. (2023) and Stengel-Eskin et al. (2023)—provide evidence that LLMs struggle with linguistic ambiguity in zero-shot settings. In contrast, Kamath et al. (2024) show that multiple versions of GPT models can resolve quantifier scope ambiguity with high accuracy, exceeding 90% in some cases. The authors argue that this strong performance reflects the integration of background world knowledge into model preferences when interpreting scope-ambiguous constructions. However, existing work leaves open whether high QSD performance genuinely reflects robust semantic competence or instead emerges from the interaction of shallow formal heuristics and world-knowledge cues that systematically converge in naturalistic data. The present paper addresses precisely this question by using QSD as a controlled probe of how LLMs balance formal interpretive pressures with background expectations during scope interpretation.

QSD arises in sentences containing multiple quantified noun phrases (NPs):

(1) <u>Every kid</u> lives in <u>a house</u>.

Sentence (1) contains two quantified NPs—the universal *every kid* and the existential *a house*—and admits two possible interpretations depending on their relative scope. Under the surface-scope reading, the first quantifier takes wide scope, yielding the interpretation that every kid lives in a possibly different house (*every kid* > *a house*). Under the inverse-scope reading, the second quantifier takes

wide scope, resulting in the less plausible interpretation that there is a single house in which all kids live (*a house* > *every kid*).

Prior work on statistical and automatic QSD (e.g., Higgins and Sadock 2003; AnderBois et al. 2012; Evang and Bos 2013; Manshadi and Allen 2011; Manshadi et al. 2013; Sayeed 2016) has identified several influential surface predictors. Linear order plays a role, with earlier quantifiers tending to take wider scope, and grammatical function and lexical realization also matter: subjects often outscope objects[1], and certain quantifiers (e.g., the universals *each* and *every*) exhibit a general preference for wide scope. However, surface-level cues alone are insufficient to account for scope preferences. A substantial body of work argues that pragmatic and world knowledge about how situations typically unfold strongly influences scope interpretation (e.g., Saba and Corriveau 2001; Srinivasan and Yates 2009; Manshadi and Allen 2011; AnderBois et al. 2012; Schuler and Wheeler 2014; Tsiolis 2020; Rasmussen 2022; Kamath et al. 2024). This makes QSD particularly suitable as a controlled probe of LLM interpretation. QSD allows formal predictors and world knowledge to be systematically aligned or placed in conflict, creating a setting in which the relative contribution of different interpretive factors can be experimentally manipulated. More broadly, we argue that QSD provides a uniquely controlled framework for investigating how formal cues and world knowledge interact during semantic interpretation in LLMs.

Using QSD as a controlled probing paradigm, we introduce two complementary datasets. First, we construct a balanced English corpus designed to neutralize standard surface heuristics and systematically manipulate the relationship between formal predictors and preferred readings. Second, we introduce a pseudosentence dataset that removes real-world referents to isolate formal cues from world-knowledge support. These resources let us compare interpretation under conditions where world knowledge is available (balanced data) versus largely unavailable (pseudosentences), while keeping the scope decision problem fixed. We then evaluate a range of LLMs in a zero-shot question–answering (Q&A) setup and collect human baselines under the same protocol. Finally, we test the causal contribution of external world knowledge by equipping the same models with retrieval-augmented generation (RAG) over ConceptNet and Simple Wikipedia, using exactly the same prompt template.

Our results reveal a consistent pattern in LLM interpretation: strong performance on the balanced corpus systematically co-occurs with degraded accuracy on pseudosentences. Crucially, models retain robust performance only in configurations where classical formal predictors converge on the preferred interpretation, whereas performance on the remaining configurations drops sharply and becomes highly unstable across models. Introducing retrieval leads to limited overall gains; however, these improvements are highly selective, concentrating on configurations in which classical surface predictors conflict with the preferred reading. Taken together, these findings suggest that strong QSD performance in current LLMs emerges most robustly when surface-level heuristics and world knowledge converge, while configurations lacking such convergence give rise to substantially more fragile and unstable behavior.

The paper is structured as follows. Section 2 introduces the datasets and describes their construction. Section 3 presents Experiment 1 on implicit world knowledge, including the evaluation protocol, model and human baselines, and results. Section 4 presents Experiment 2, which investigates the contribution of retrieval-augmented world knowledge, together with the corresponding results. Sections 5 and 6 conclude the paper with a summary of findings, a discussion in relation to prior work, limitations, and directions for future research.

2 Datasets

For our experiments, we constructed two complementary datasets: (i) a balanced English corpus designed to suppress surface-level heuristics in QSD and systematically manipulate the relationship between formal predictors and preferred readings, and (ii) a novel dataset of pseudosentences that eliminate real-world referents and encourage reliance on formal cues.

The English dataset was constructed to be strictly balanced across four canonical quantifier-combination patterns. These patterns are defined by the interaction of quantifier type, linear order, and the human-preferred interpretation. Table 1 illustrates the four combination classes with both real and pseudosentence examples.

[1] In English, the effects of grammatical function and linear order are strongly correlated due to relatively rigid word order.

Type	Pattern	Examples (real / pseudo)
I	U–E, surface	Each child gets one balloon. Each nimen brims one bifisior.
II	E–U, surface	A teacher grades every assignment. A tevor fuperos every zikekoor.
III	E–U, inverse	Exactly one flag flies on every building. Exactly one gefumi lidaharas te every fuve.
IV	U–E, inverse	Every member belongs to one team. Every luhahado nudotecas li one zorp.

Table 1: Quantifier-combination classes in the QSD dataset, illustrated with real and pseudosentence examples.

We began with the 837 scope-ambiguous English sentences from Kamath et al. (2024), each containing exactly two quantified noun phrases and annotated with a human-preferred interpretation (surface or inverse). We restricted this corpus to sentences involving two interacting quantifiers—one universal and one existential. Universal quantifiers (U) include *each*, *every*, *all*, and *any*[2]. Existential quantifiers (E) include bare numerals (e.g., *two*, *five*), modified numerals (e.g., *at least one*, *exactly three*), as well as *a(n)*, *some*, *several*, and *a few*. After filtering, 390 sentences remained.

These sentences were assigned to the four quantifier-combination classes. Their distribution was highly imbalanced: Combination I contained 145 items, Combination II 68 items, Combination III 171 items, and Combination IV only six items. To obtain a fully balanced dataset, we applied controlled augmentation. For the sparsely represented Combination IV, candidate sentences were evaluated in a crowdsourced survey with 50 native English speakers recruited via Prolific, and only items with at least 75% agreement on the intended reading were retained. The dataset was further expanded using two rounds of GPT-4–based sentence generation seeded with previously curated and validated datapoints. To mitigate paraphrastic redundancy, we applied SBERT-based semantic clustering to identify and remove near-duplicate items. All sentences subsequently underwent manual quality control to ensure grammaticality, interpretability, and consistency with the intended scope reading.

The resulting balanced dataset contains 440 sentences, with 110 items per combination type.[3] This balancing procedure is critical because naturally occurring QSD corpora are heavily skewed toward configurations in which formal predictors align with preferred interpretations.

To construct the pseudosentence condition, we randomly sampled 160 items from the balanced dataset (40 per combination type). Each sentence was transformed via systematic one-to-one lexical substitution: all open-class lexical items and all prepositions[4] were replaced with nonce forms, while preserving quantifier identity, linear order, and argument structure relevant for scope interpretation.[5] Thus, the pseudo condition removes familiar lexical-semantic and world-knowledge cues while keeping intact the formal cues underlying quantifier-scope resolution.

Together, the balanced dataset and the pseudosentence dataset allow us to treat QSD as a controlled probe of LLM interpretation.

3 Experiment 1: Implicit World Knowledge

In Experiment 1, we test whether LLMs can achieve high QSD accuracy using only *implicit* world knowledge, without external retrieval. We evaluate models in a zero-shot setup on (i) the balanced corpus, where real-world referents are available, and (ii) pseudosentences, which remove such referents and thus isolate formal cues.

3.1 Method

Our experimental design closely follows the methodology introduced by Kamath et al. (2024). We formulate QSD as a Q&A task. Each model is

[2]We included only uses of *any* that receive a universal interpretation.

[3]The balanced dataset is derived from the corpus presented in Kamath et al. (2024) and may be freely used in accordance with the MIT license governing this repository.

[4]Prepositions were replaced intentionally rather than preserved. This decision was theory-driven: previous work suggests that prepositions can strongly influence scope preferences (Leczkowski et al., 2022). Retaining real prepositions would therefore reintroduce semantic predictors that our manipulation aimed to remove.

[5]In the final set of pseudosentences, items containing passive constructions were replaced with active variants to avoid dissociations between linear order, subject position, and semantic role assignment. This modification did not materially affect the overall pattern of results.

```
A book is given to each student in the class.

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

Option A: In total, more than one book is given.

Option B: In total, one book is given.

Which of the two options is most likely? Answer with either 'Option A'
(for Option A) or 'Option B' (for option B).
```

Figure 1: An example of a stimulus provided to the model.

Model	Accuracy	Surface	Inverse
GPT-5.1	0.809 ± 0.007	0.873 ± 0.009	0.746 ± 0.009
GPT-4o	0.836 ± 0.005	0.914 ± 0.004	0.759 ± 0.011
GPT-4o-mini	0.830 ± 0.005	0.918 ± 0.005	0.741 ± 0.007
Qwen-Max	0.830 ± 0.003	0.932 ± 0.002	0.727 ± 0.005
Qwen3-8B	0.793 ± 0.003	0.918 ± 0.003	0.668 ± 0.005
Llama 3.1 70B	0.850 ± 0.003	0.936 ± 0.003	0.764 ± 0.005
Llama 3.1 8B	0.689 ± 0.009	0.777 ± 0.013	0.600 ± 0.011
Average	$\mathbf{0.805 \pm 0.003}$	$\mathbf{0.895 \pm 0.003}$	$\mathbf{0.715 \pm 0.004}$

Table 2: Accuracy on the balanced QSD dataset (440 items). Values report accuracy with SD across five runs.

Model	Accuracy	Surface	Inverse
GPT-5.1	0.606 ± 0.022	0.594 ± 0.025	0.619 ± 0.027
GPT-4o	0.650 ± 0.007	0.619 ± 0.007	0.681 ± 0.008
GPT-4o-mini	0.644 ± 0.004	0.713 ± 0.003	0.575 ± 0.009
Qwen-Max	0.713 ± 0.006	0.831 ± 0.004	0.594 ± 0.012
Qwen3-8B	0.500 ± 0.006	0.513 ± 0.010	0.488 ± 0.003
Llama 3.1 70B	0.675 ± 0.009	0.825 ± 0.014	0.525 ± 0.011
Llama 3.1 8B	0.497 ± 0.006	0.506 ± 0.010	0.488 ± 0.012
Average	$\mathbf{0.612 \pm 0.002}$	$\mathbf{0.657 \pm 0.004}$	$\mathbf{0.567 \pm 0.004}$

Table 3: Accuracy on pseudosentences (160 items). Values report accuracy with SD across five runs.

presented with a sentence that is technically scope-ambiguous but has a single human-preferred interpretation, which may correspond either to a surface-scope or an inverse-scope reading. For each such sentence, the model is then given two candidate statements representing the two possible scope interpretations: one consistent with the surface reading and the other consistent with the inverse reading.

The prompt is structured in three parts: (i) the input sentence, (ii) two explicitly labeled interpretation options (Option A and Option B), and (iii) instructions specifying the task and the required output format, as illustrated in Figure 1. Models are instructed to indicate which interpretation is more likely by responding with `Option A` or `Option B`, corresponding to the preferred interpretation. To ensure a balanced distribution of correct answers across options, each datapoint in the dataset is duplicated with the order of the two interpretations reversed (i.e., Option A is relabeled as Option B, and vice versa). This design prevents models from exploiting positional or response-format biases.

Crucially, the task formulation and prompt template are identical across the balanced and pseudosentence conditions; only the lexical content differs in whether it supports access to real-world referents. This makes Experiment 1 a controlled test of how far models can rely on *implicit* world knowledge during interpretation.

3.2 Models

We evaluate seven autoregressive LLMs spanning three model families and a range of capacities. From the OpenAI family, we include GPT-5.1 and GPT-4o as high-capability reference points, and GPT-4o-mini as a smaller, cost-efficient variant. To cover strong non-OpenAI systems, we include Qwen-Max (high tier) and Qwen3-8B (compact baseline) from the Qwen family. Finally, we include two open-weight models from the Llama family, Llama 3.1 70B and Llama 3.1 8B, to provide a within-family contrast between a large and a smaller model.

All models were queried via hosted APIs through a unified evaluation wrapper (OpenAI API for GPT models, DashScope for Qwen models, and OpenRouter-hosted endpoints for Llama models), using the same prompt template and answer format across conditions. Results were averaged over five independent runs.

3.3 Model Results[6]

Tables 2 and 3 report zero-shot accuracy on the balanced QSD dataset and on pseudosentences, respectively. On the balanced dataset, all evaluated models achieve relatively high overall accuracy, indicating that scope ambiguity can often be resolved successfully under naturalistic conditions.

In contrast, performance decreases consistently on pseudosentences, which remove world-knowledge cues. All model families show declines in overall accuracy relative to the balanced condition, as well as reductions for both surface- and inverse-scope interpretations.

3.4 Discussion

Experiment 1 provides a controlled test of how far quantifier scope interpretation in LLMs can be supported by implicit world knowledge in interaction with formal cues. Across models, performance decreases consistently when real-world referents are removed (Table 4). Averaged across models, overall accuracy drops by -0.193, surface-scope

[6]The code for all experiments and the datasets used in this paper will be made publicly available in our GitHub repository.

Model	Δ Overall	Δ Surface	Δ Inverse
GPT-5.1	−0.203	−0.279	−0.127
GPT-4o	−0.186	−0.295	−0.078
GPT-4o-mini	−0.186	−0.205	−0.166
Qwen-Max	−0.117	−0.101	−0.133
Qwen3-8B	−0.293	−0.405	−0.180
Llama 3.1 70B	−0.175	−0.111	−0.239
Llama 3.1 8B	−0.192	−0.271	−0.112
Average	**−0.193**	**−0.238**	**−0.148**

Table 4: Change in accuracy (Δ) from the balanced QSD dataset to pseudosentences, computed separately for overall, surface-scope, and inverse-scope accuracy.

Combination	Balanced	Pseudosentences
I (U–E, surface)	0.96	0.79
II (E–U, surface)	0.85	0.52
III (E–U, inverse)	0.86	0.59
IV (U–E, inverse)	0.58	0.55

Table 5: Average accuracy by quantifier-combination type on the balanced QSD dataset and on pseudosentences.

accuracy by -0.238, and inverse-scope accuracy by -0.148. We resampled base items with replacement and recomputed the balanced–pseudo accuracy difference across models for each bootstrap sample. Item-level cluster bootstrapping confirms that all three decreases are statistically reliable: overall accuracy ($\Delta = -0.193$, 95% CI $[-0.233, -0.153]$, $p < 10^{-4}$), surface-scope accuracy ($\Delta = -0.238$, 95% CI $[-0.292, -0.185]$, $p < 10^{-4}$), and inverse-scope accuracy ($\Delta = -0.148$, 95% CI $[-0.205, -0.091]$, $p < 10^{-4}$). Thus, the transition from naturalistic sentences to pseudosentences affects all evaluated dimensions of performance. This pattern indicates that high QSD performance in naturalistic settings depends not only on formal cues, but also on broad background expectations associated with familiar real-world referents.

A more fine-grained picture emerges when performance is examined across the four quantifier-combination types introduced in Section 2. As shown in Table 5, on the balanced QSD dataset the highest accuracy is observed for Combination I (U–E, surface), where both classical predictors of scope preference converge: the first quantifier is universal and the preferred interpretation assigns it wide scope. In contrast, the lowest accuracy is observed for Combination IV (U–E, inverse), where the preferred reading conflicts with both predictors, requiring the existential quantifier to outscope the initial universal quantifier. Combinations II (E–U, surface) and III (E–U, inverse) fall between these extremes, since only one classical predictor supports the preferred interpretation.

This pattern becomes substantially more pronounced in the pseudosentence condition (Table 5). Combination I remains relatively robust, reaching 0.79 accuracy on average, and this robustness is highly consistent across stronger models (Table 6). Importantly, however, robustness in the pseudo condition is not uniform across scope configurations. Models retain relatively high performance only when both classical formal predictors continue to support the preferred reading. Once these predictors partially diverge (II–III) or fully conflict (IV), performance drops sharply toward chance levels and becomes highly unstable across models.

Crucially, no single consistent strategy appears to dominate model behavior in these more difficult configurations. Instead, different models appear to rely on different partial heuristics, yielding highly heterogeneous performance patterns across combinations (Table 6). For example, GPT-5.1 remains highly accurate on Combination III but performs poorly on Combinations II and IV, whereas Qwen-Max exhibits the opposite pattern, including near-ceiling performance on Combination IV. Such divergences suggest reliance on partially overlapping heuristic strategies that differ across model families and capacities.

Taken together, these findings suggest a formal-cue convergence mechanism in LLM scope interpretation: models remain robust when world knowledge and converging formal cues stabilize the same preferred reading, but exhibit unstable and model-specific heuristic fallback when this convergence breaks down.

3.5 Human baselines

To establish human performance baselines, we conducted a set of psycholinguistic experiments designed to closely mirror the model evaluation protocol.

For the balanced QSD dataset, we ran three online survey studies. In total, 150 native English speakers were recruited via Prolific (50 per study). Written informed consent was obtained from all participants prior to data collection. Participants were paid approximately £3 for completing the survey (average completion time: 15 minutes). Participants completed a brief demographic questionnaire, and the samples were balanced with respect to gender across studies. Each study included 40 randomly sampled, non-overlapping experimental items from the balanced dataset (120 items total,

Model	I (U–E, surf)	II (E–U, surf)	III (E–U, inv)	IV (U–E, inv)
GPT-5.1	0.9250	0.2625	0.8625	0.3750
GPT-4o	0.8250	0.4125	0.7875	0.5750
GPT-4o-mini	0.7750	0.6500	0.5000	0.6500
Qwen-Max	0.8750	0.7875	0.3000	0.8875
Qwen3-8B	0.6500	0.3750	0.5125	0.4625
Llama 3.1 70B	1.0000	0.6500	0.6750	0.3750
Llama 3.1 8B	0.5125	0.5000	0.4750	0.5000
Average	**0.7946**	**0.5196**	**0.5875**	**0.5464**

Table 6: Accuracy on pseudosentences by quantifier-combination type and model.

out of 440), supplemented with 10 filler items. Fillers were drawn from the dataset of Kamath et al. (2024) and consisted of sentences containing two numerical quantifiers. Each study therefore comprised 50 items in total.

For the pseudosentence dataset, we recruited a separate group of 50 native English speakers. As in the balanced-dataset studies, participants completed a demographic questionnaire, and the sample was balanced with respect to gender. The experiment followed the same overall procedure as the balanced-dataset studies, and participants received the same compensation. The study included 40 randomly sampled, non-overlapping experimental items drawn from the 160-item pseudosentence dataset. Unlike the balanced-dataset studies, the pseudosentence experiment did not include filler items.

3.6 Human Results and Comparison to Models

Table 7 reports human accuracy on the balanced QSD dataset and on pseudosentences, alongside averaged results for LLMs. On the balanced dataset, human performance is high across conditions, indicating that quantifier scope can be reliably resolved when both linguistic cues and real-world knowledge are available. As in the case of language models, accuracy decreases substantially on pseudosentences, where real-world referents are removed.

This overall pattern parallels the behavior observed in LLMs and supports the view that both humans and models rely, at least in part, on interactions between formal constraints and world knowledge during scope interpretation, in line with prominent psycholinguistic accounts of ambiguity resolution (e.g., Kurtzman and MacDonald 1993; Trueswell et al. 1994; Dwivedi 2013; Feiman and Snedeker 2016; Dwivedi et al. 2018). At the same time, important differences emerge between human and model behavior.

Most notably, humans do not exhibit the same strong preference for configurations in which classical surface predictors converge. As shown in Table 8, human performance on the balanced dataset remains relatively stable across quantifier-combination types, including the most difficult Combination IV, in which the preferred interpretation conflicts with both classical predictors. In contrast, averaged LLM performance shows a much sharper advantage for Combination I, where both classical predictors support the preferred interpretation, and substantially lower accuracy for Combination IV.

This contrast becomes particularly informative in the pseudosentence condition (Table 9). While both humans and models show degraded performance once world-knowledge cues are removed, LLMs continue to display a comparatively strong advantage for Combination I relative to the remaining configurations. Human performance, by contrast, remains more evenly distributed across combinations, despite the overall decrease in accuracy. These findings suggest that, although world knowledge contributes importantly to scope interpretation in both humans and models, LLMs rely substantially more heavily than humans on the convergence of formal predictors and lexical-semantic support during scope interpretation.

4 Experiment 2: World Knowledge via Retrieval

In Experiment 2, we manipulate access to *external* world knowledge via retrieval, while keeping the task, prompt, and scoring procedure exactly the same as in Experiment 1.

4.1 Method

Experiment 2 extends the zero-shot setup from Experiment 1 by equipping models with retrieval-augmented generation (RAG). We evaluate the same models on the balanced QSD dataset, but augment the prompt with externally retrieved world

	Balanced			Pseudosentences		
Group	**Overall**	**Surface**	**Inverse**	**Overall**	**Surface**	**Inverse**
LLMs (avg.)	0.805	0.895	0.715	0.612	0.657	0.567
Humans	0.870	0.850	0.890	0.576	0.617	0.535

Table 7: LLM vs. human accuracy on balanced QSD and pseudosentences.

Group	**I**	**II**	**III**	**IV**
LLMs (avg.)	0.96	0.85	0.86	0.58
Humans	0.91	0.77	0.87	0.89

Table 8: LLM vs. human accuracy by quantifier-combination type on the balanced QSD dataset.

Group	**I**	**II**	**III**	**IV**
LLMs (avg.)	0.79	0.52	0.59	0.55
Humans	0.66	0.57	0.48	0.59

Table 9: LLM vs. human accuracy by quantifier-combination type on pseudosentences.

knowledge relevant to the input sentence. We implement dense retrieval over a combined ConceptNet and Simple Wikipedia corpus using Sentence-Transformer embeddings and a FAISS index. For each input sentence, the stimulus sentence itself is used as the retrieval query. The retrieved passages are appended to the prompt before the model produces its answer.

For example, consider the sentence *Every chapter was written by one author*. Here, the preferred reading (*one author* > *every chapter*) conflicts with both classical surface predictors: linear order favors the first quantifier, and universal quantifiers generally prefer wide scope. Retrieval returns information such as *chapter* → *partOf* → *book* and *book* → *createdBy* → *author*. This introduces a broader structural constraint—chapters typically belong to the same book, and books are commonly associated with a single author—thereby such retrieved information supports the interpretation, in which one author wrote all chapters, despite the fact that this reading conflicts with classical surface predictors.

Crucially, the task formulation, prompt template, and scoring procedure are identical to those used in Experiment 1, allowing performance differences to be attributed directly to the availability of externally retrieved world knowledge.

4.2 Results

Table 10 reports a per-model comparison between simple prompting and RAG-augmented variants, broken down by overall accuracy, surface-scope interpretations, and inverse-scope interpretations on the balanced QSD dataset.

Across models, the effects of RAG are mixed and relatively small, with modest increases and decreases observed across overall, surface-scope, and inverse-scope accuracy.

4.3 Discussion

Table 11 reports averaged results across models and confirms that RAG yields only limited changes in overall, surface-scope, and inverse-scope accuracy. A more informative pattern emerges when performance is examined by quantifier-combination type (Table 12). For Combination I (U–E, surface), where both classical predictors—quantifier type and linear order—support the preferred interpretation, performance is already at ceiling and remains unchanged under RAG. This indicates that when formal cues are fully sufficient, additional world knowledge does not alter model behavior.

For configurations in which exactly one classical predictor conflicts with the preferred reading, RAG has little effect. In Combination III (E–U, inverse), where linear order disfavors the preferred inverse interpretation but quantifier type supports it, performance remains stable under RAG. In Combination II (E–U, surface), where quantifier type conflicts with the preferred surface reading while linear order supports it, accuracy shows a slight decrease. Together, these patterns suggest that when partial structural support is already available, external knowledge does not systematically improve scope resolution and may introduce mild interference.

In contrast, the largest gains under RAG are observed for Combination IV (U–E, inverse), which represents the most challenging configuration. In this case, both classical predictors actively disfavor the preferred interpretation: the universal quantifier appears first, yet the correct reading requires the existential quantifier to take wide scope. Here, RAG yields a statistically significant improvement, increasing average accuracy from 0.58 to 0.65 across models ($\Delta = +0.091$, 95% CI $[0.047, 0.133]$, $p = 5 \times 10^{-5}$). These gains emerge precisely in the condition where models must override surface-level expectations and rely on broader knowledge about plausible event structures.

Model	Overall$_{Simple}$	Overall$_{RAG}$	Surface$_{Simple}$	Surface$_{RAG}$	Inverse$_{Simple}$	Inverse$_{RAG}$
GPT-5.1	0.809 ± 0.007	0.825 ± 0.008	0.873 ± 0.009	0.882 ± 0.007	0.746 ± 0.009	0.768 ± 0.010
GPT-4o	0.836 ± 0.005	0.811 ± 0.005	0.914 ± 0.004	0.859 ± 0.008	0.759 ± 0.011	0.764 ± 0.005
GPT-4o-mini	0.830 ± 0.005	0.811 ± 0.004	0.918 ± 0.005	0.859 ± 0.005	0.741 ± 0.007	0.764 ± 0.009
Qwen-Max	0.830 ± 0.003	0.821 ± 0.005	0.932 ± 0.002	0.914 ± 0.005	0.727 ± 0.005	0.727 ± 0.008
Qwen3-8B	0.793 ± 0.003	0.811 ± 0.002	0.918 ± 0.003	0.896 ± 0.003	0.668 ± 0.005	0.727 ± 0.005
Llama 3.1 70B	0.850 ± 0.003	0.823 ± 0.003	0.936 ± 0.003	0.891 ± 0.005	0.764 ± 0.005	0.755 ± 0.003
Llama 3.1 8B	0.689 ± 0.009	0.761 ± 0.009	0.777 ± 0.013	0.832 ± 0.012	0.600 ± 0.011	0.691 ± 0.009

Table 10: Comparison of simple prompting and RAG-augmented models on the balanced QSD dataset. Values report accuracy with SD across five runs.

Metric	Simple	RAG	Δ
Overall	0.805	0.809	$+0.004$
Surface	0.895	0.876	-0.019
Inverse	0.715	0.742	$+0.027$

Table 11: Average effect of RAG augmentation on balanced QSD performance across models.

Combination	LLM	LLM+RAG	Δ
I (U–E, surface)	0.96	0.96	$+0.00$
II (E–U, surface)	0.85	0.83	-0.02
III (E–U, inverse)	0.86	0.86	$+0.00$
IV (U–E, inverse)	0.58	0.65	$+0.07$

Table 12: Average accuracy across quantifier-combination types on the balanced QSD dataset, averaged over all models.

5 Conclusions and related work

We used QSD as a controlled probe of how LLMs integrate linguistic form and world knowledge during interpretation. Across models, zero-shot performance is high on our balanced corpus, where real-world referents are available, but drops substantially on pseudosentences that remove such referents. Crucially, robust performance in the pseudo condition is retained only in configurations where classical formal predictors converge on the preferred reading. Once these predictors partially diverge or fully conflict, performance deteriorates sharply and becomes highly unstable across models. This pattern suggests that strong QSD performance depends on interactions between formal heuristics and world-knowledge support.

Experiment 2 strengthens this conclusion by manipulating access to external knowledge while keeping the task and prompt fixed. Retrieval-augmented generation yields only limited gains in overall accuracy, but these gains are highly selective: they most clearly affect the hardest configuration (U–E, inverse), where both classical surface predictors conflict with the preferred reading. Importantly, retrieval does not act as a uniform performance enhancer. Instead, it selectively improves interpretations that are otherwise disfavored by formal surface cues, while leaving structurally well-supported configurations largely unchanged and occasionally introducing interference elsewhere.

Taken together, our results suggest that the strong QSD performance reported for current LLMs may reflect a masking effect arising from the systematic convergence of surface-level heuristics and world knowledge in naturalistic datasets. What initially appears to be robust semantic competence instead reveals considerable fragility once this convergence between formal cues and background knowledge breaks down.

Our findings are consistent with results reported by Kamath et al. (2024), who show that advanced LLMs can exhibit scope-reading preferences similar to those of humans and achieve high accuracy on naturalistic data, whereas smaller models tend to fail. At the same time, our results substantially refine this picture by demonstrating that high naturalistic performance can mask strong dependence on converging formal heuristics and lexical-semantic regularities. Overall accuracy on our balanced dataset is lower than the near-ceiling performance reported for GPT-4 in Kamath et al., which is expected given important differences in dataset design. In particular, the Kamath et al. dataset is heavily skewed toward surface-scope readings, whereas our dataset is explicitly balanced and augmented to include naturally underrepresented configurations in which formal predictors conflict.

Our results also resonate with recent work on retrieval-augmented generation and semantic interpretation (e.g., Cuconasu et al. 2024; Chatzikyriakidis 2025). The limited overall gains observed in Experiment 2 are therefore theoretically informative rather than unexpected. Retrieval improves accuracy primarily in precisely those configurations where implicit world knowledge is insufficient and formal predictors are misleading. This suggests that retrieved knowledge modulates interpretation through its interaction with linguistic form, selectively enabling readings that are otherwise disfavored by surface-level heuristics.

6 Limitations

All experiments in this paper are conducted on English data. Although quantifier scope ambiguity is widely attested cross-linguistically, languages differ in how scope is constrained and resolved. Consequently, the present results do not directly generalize beyond English. Extending the proposed experimental paradigm to typologically diverse languages remains an important direction for future work.

In addition to this language-specific limitation, our retrieval-augmented experiments employ a single, fixed RAG configuration, designed to enable a controlled comparison between retrieval and no-retrieval conditions. While this setup isolates the effect of adding external knowledge, it does not explore the broader space of retrieval strategies or hyperparameter choices. Model behavior may vary under different retrieval configurations, and a more systematic investigation of how RAG design choices affect semantic interpretation is left for future 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*, pages 15–28.

Marco Baroni. 2022. On the proper role of linguistically-oriented deep net analysis in linguistic theorizing. *Algebraic structures in natural language*, pages 1–16.

Stergios Chatzikyriakidis. 2025. Reasoning with ragged events: Rag-enhanced event knowledge base construction and reasoning with proof-assistants. *arXiv preprint arXiv:2506.07042*.

Florin Cuconasu, Giovanni Trappolini, Federico Siciliano, Simone Filice, Cesare Campagnano, Yoelle Maarek, Nicola Tonellotto, and Fabrizio Silvestri. 2024. The power of noise: Redefining retrieval for rag systems. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 719–729.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. 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 4171–4186.

Veena D. Dwivedi. 2013. Interpreting quantifier scope ambiguity: Evidence of heuristic first, algorithmic second processing. *PloS ONE*, 8(11):e81461.

Veena D. Dwivedi, Kaitlin E. Goertz, and Janahan Selvanayagam. 2018. Heuristics in language comprehension. *Journal of Behavioral and Brain Science*, 8(7):430–446.

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

Allyson Ettinger. 2020. What BERT is not: Lessons from a new suite of psycholinguistic diagnostics for language models. In *Transactions of the Association for Computational Linguistics*, pages 34–48.

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*, pages 1790–1801.

Kilian Evang and Johan Bos. 2013. Scope disambiguation as a tagging task. In *Proceedings of the 10th International Conference on Computational Semantics*, pages 314–320.

Roman Feiman and Jesse Snedeker. 2016. The logic in language: How all quantifiers are alike, but each quantifier is different. *Cognitive Psychology*, 87:29–52.

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.

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

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*.

Aikaterini-Lida Kalouli, Rita Sevastjanova, Christin Beck, and Maribel Romero. 2022. Negation, coordination, and quantifiers in contextualized language models. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3074–3085.

Gaurav Kamath, Sebastian Schuster, Sowmya Vajjala, and Siva Reddy. 2024. Scope ambiguities in large language models. *Transactions of the Association for Computational Linguistics*, 12:738–754.

Howard S. Kurtzman and Maryellen C. MacDonald. 1993. Resolution of quantifier scope ambiguities. *Cognition*, 48(3):243–279.

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.

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. In *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. In *CoRR, abs/1907.11692*.

Mehdi Manshadi and James Allen. 2011. Unrestricted quantifier scope disambiguation. In *Proceedings of Association for Computational Linguistics'11, Workshop on Graph-based Methods for NLP (TextGraph–6)*, pages 51–59.

Mehdi Manshadi, Daniel Gildea, and James Allen. 2013. Plurality, negation, and quantification: Towards comprehensive quantifier scope disambiguation. In *Proceedings of of the 51st Annual Meeting of the Association for Computational Linguistics*, pages 64–72.

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

Kyle Richardson, Hai Hu, Lawrence S. Moss, and Ashish Sabharwal. 2020. Probing natural language inference models through semantic fragments. In *Association for the Advancement of Artificial Intelligence (AAAI)*, pages 8713–8721.

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.

Walid S. Saba and Jean-Pierre Corriveau. 2001. Plausible reasoning and the resolution of quantifier scope ambiguities. *Studia Logica*, 67(2):271–289.

Asad Sayeed. 2016. Representing the effort in resolving ambiguous scope. In *Proceedings of Sinn und Bedeutung*, pages 604–621.

William Schuler and Adam Wheeler. 2014. Cognitive compositional semantics using continuation dependencies. In *Proceedings of the Third Joint Conference on Lexical and Computational Semantics (*SEM 2014)*, pages 141–150.

Rita Sevastjanova, Aikaterini-Lida Kalouli andChristin Beck, Hanna Schäfer, and Mennatallah El-Assady. 2021. Explaining contextualization in language models using visual analytics. 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 464–476.

Prakash Srinivasan and Alexander Yates. 2009. Quantifier scope disambiguation using extracted pragmatic knowledge: Preliminary results. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1465–1474.

John C. Trueswell, Michael K. Tanenhaus, and Susan M. Garnsey. 1994. Semantic influences on parsing: Use of thematic role information in syntactic ambiguity resolution. *Journal of Memory and Language*, 33(3):285–318.

KC Tsiolis. 2020. Quantifier scope disambiguation. Summary of research conducted in summer 2020 under the supervision of Jackie Cheung.