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 in QSD, and (ii) a pseudosentence dataset that removes real-world referents to isolate formal cues. 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, with the largest degradation for inverse-scope readings. This pattern indicates that surface-level cues alone are insufficient to explain model behavior and suggests a substantive contribution of implicit world knowledge to LLM interpretation. 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 gains in overall accuracy, but these effects are highly selective: they primarily affect inverse-scope interpretations, and most clearly the hardest configuration, where classical surface predictors conflict with the preferred reading. Taken together, our results suggest that LLM interpretive behavior reflects an interaction between world knowledge and formal interpretive pressures encoded in the input, with world knowledge—implicit or retrieved—playing a particularly important role when formal cues are insufficient to yield a preferred reading. This pattern partially parallels, but does not fully match, human scope interpretation.

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. The goal of the present paper is to examine this claim in greater depth by investigating whether, and to what extent, LLMs leverage world knowledge in quantifier scope disambiguation (QSD), and more broadly, how they balance linguistic form and background knowledge when selecting an interpretation.

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

(1) Every kid lives in a house.

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 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, and pragmatic knowledge about how situations typically unfold in the world often plays a crucial role in resolving scope ambiguity (e.g., Saba and Corriveau 2001; Srinivasan and Yates 2009; Manshadi and Allen 2011; AnderBois et al. 2012; Schuler and Wheeler 2014; Leczkowski et al. 2022; Rasmussen 2022; Kamath et al. 2024). This raises a broader question: how can we design an experimental setting that both incorporates world knowledge into QSD and allows us to assess how LLMs balance linguistic form with background expectations during interpretation?

Using QSD as a controlled probing task, we introduce two complementary datasets: a balanced English corpus designed to neutralize standard surface heuristics, and a pseudosentences dataset that removes real-world referents to isolate formal cues. 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. This pattern indicates that surface-level cues alone are insufficient to explain model behavior and points to an important role of background knowledge in scope resolution. 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. We therefore propose that LLM interpretive behavior emerges from an interaction between formal interpretive pressures and world knowledge, with the latter playing a particularly influential role when such pressures are insufficient to determine a preferred interpretation.

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 (ii) a novel dataset of pseudosentences that eliminate real-world referents and force models to rely solely on formal cues.

To prevent models from relying on surface-level strategies, the balanced 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.

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*. 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, Combina-

[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	Every planet in the solar system has an orbit. Every scruffle prowkles one wisperite.
II	E–U, surface	A security guard monitors every camera in the building. A squalder snorples every thistlehound dreft the spindle.
III	E–U, inverse	A cup of coffee is offered to each guest at the party. A guilder clipsers quence each spindlewing.
IV	U–E, inverse	Every participant registered through one website. Every furbler gallopined alongst one glimmerkin.

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

tion 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 datasets were 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.[2] By construction, it neutralizes standard surface-level heuristics and encourages reliance on world-knowledge-supported cues. In parallel, we constructed a set of 40 template-based pseudosentences that remove real-world referents and require models to depend exclusively on formal cues. Together, these datasets 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.

[2]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.

```
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.

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 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

Model	Accuracy	Surface	Inverse
GPT-5.1	0.809	0.873	0.746
GPT-4o	0.836	0.914	0.759
GPT-4o-mini	0.830	0.918	0.741
Qwen-Max	0.830	0.932	0.727
Qwen3-8B	0.793	0.918	0.668
Llama 3.1 70B	0.850	0.936	0.764
Llama 3.1 8B	0.689	0.777	0.600

Table 2: Accuracy on the balanced QSD dataset (440 items). Results are averaged across five runs.

Model	Accuracy	Surface	Inverse
GPT-5.1	0.675	0.800	0.550
GPT-4o	0.675	0.900	0.450
GPT-4o-mini	0.400	0.750	0.050
Qwen-Max	0.525	0.850	0.200
Qwen3-8B	0.350	0.550	0.150
Llama 3.1 70B	0.725	0.900	0.550
Llama 3.1 8B	0.975	0.950	1.000

Table 3: Accuracy on pseudosentences (40 items). Results are averaged across five runs.

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[3]

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 differs substantially on pseudosentences, which eliminate real-world referents. For GPT and Qwen models, accuracy drops markedly compared to the balanced condition, with especially large declines observed for inverse-scope interpretations. The large Llama model follows the same general pattern. The smaller Llama model exhibits artifact-like behavior, achieving near-ceiling accuracy on pseudosentences despite noticeably lower performance on the balanced dataset; we therefore do not interpret its pseudosentence results further in section 3. This pattern likely reflects instability rather than genuine scope-related generalization.

[3]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	-0.134	-0.073	-0.196
GPT-4o	-0.161	-0.014	-0.309
GPT-4o-mini	-0.430	-0.168	-0.691
Qwen-Max	-0.305	-0.082	-0.527
Qwen3-8B	-0.443	-0.368	-0.518
Llama 3.1 70B	-0.125	-0.036	-0.214
Llama 3.1 8B	$+0.286$	$+0.173$	$+0.400$

Table 4: Change in accuracy (Δ) from the balanced QSD dataset to pseudosentences, computed separately for overall accuracy, surface-scope readings, and inverse-scope readings. Positive values indicate improved performance when real-world referents are removed.

3.4 Discussion

Experiment 1 provides a controlled test of how far quantifier scope interpretation in LLMs can be supported by *implicit* world knowledge alone. Table 4 reports the change in accuracy from the balanced dataset to pseudosentences, computed separately for overall accuracy, surface-scope readings, and inverse-scope readings.

When restricting attention to GPT and Qwen models, a consistent pattern emerges: strong performance on the balanced corpus co-occurs with degraded accuracy on pseudosentences, with the largest drops observed for inverse-scope interpretations; Llama 3.1 70B exhibits the same overall pattern but a less pronounced degradation on pseudosentences. This suggests that surface-level cues alone are insufficient to account for high performance on naturalistic QSD, and that background expectations play an important role when real-world knowledge is available.

While the pseudosentence set is relatively small, the observed effects are large and highly consistent across model families, supporting the stability of

Combination	Balanced	Pseudosentences
I (U–E, surface)	0.97	0.90
II (E–U, surface)	0.86	0.68
III (E–U, inverse)	0.87	0.37
IV (U–E, inverse)	0.60	0.28

Table 5: Average accuracy across models by quantifier-combination type, evaluated on the balanced QSD dataset and on pseudosentences (excluding Llama 3.1 8B).

the pattern.

At the same time, large-capacity models retain non-trivial performance on pseudosentences, indicating that formal interpretive pressures continue to constrain interpretation even in the absence of real-world referents. Taken together, these findings support the view that scope interpretation in LLMs is best understood as emerging from the interaction between formal interpretive pressures and broad background expectations derived from world knowledge.

A more fine-grained picture emerges when we examine model performance across the four quantifier-combination types introduced in Section 2 (Table 5). As expected, 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 second, existential quantifier to take wide scope.

Combinations II (E–U, surface) and III (E–U, inverse) fall between these extremes. In these cases, only one of the classical predictors supports the preferred interpretation—either linear order or quantifier type—yielding intermediate performance levels.

These tendencies become even more pronounced for pseudosentences (Table 5). In the absence of real-world referents, surface-scope readings are resolved with substantially higher accuracy than inverse-scope readings. Performance for interpretations supported by both classical predictors approaches ceiling, while inverse-scope interpretations drop to near-chance or below-chance accuracy, with Combination IV again yielding the lowest accuracy. This sharp separation highlights the role of world knowledge in sustaining non-surface scope interpretations and illustrates a form of inverse scope fragility, whereby non-surface readings are disproportionately affected by the removal of world-knowledge support.

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, 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 pseudosentences 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 gender-balanced. This experiment followed the same overall structure as the balanced-dataset studies. Participants received the same compensation as in the balanced-dataset studies. The study consisted of 40 experimental items and did not include fillers.

3.6 Human Results and Comparison to Models

Table 6 reports human accuracy on the balanced QSD dataset and on pseudosentences, against the background of 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. Accuracy decreases on pseudosentences, which remove real-world referents, with the largest declines observed for inverse-scope interpretations.

This overall pattern parallels the behavior of language models, and supports the view that both humans and models rely on an interaction between for-

Group	Dataset	Overall	Surface	Inverse
LLMs (avg.)	Balanced	0.825	0.915	0.734
Humans	Balanced	0.870	0.850	0.890
LLMs (avg.)	Pseudosentences	0.558	0.792	0.317
Humans	Pseudosentences	0.600	0.740	0.470

Table 6: Average performance of LLMs (excluding Llama 3.1 8B) compared to human accuracy on the balanced QSD dataset and the pseudosentences dataset.

Combination	Balanced	Pseudosentences
I (U–E, surface)	0.91	0.72
II (E–U, surface)	0.77	0.75
III (E–U, inverse)	0.87	0.51
IV (U–E, inverse)	0.89	0.42

Table 7: Human accuracy across the four quantifier-combination types on the balanced QSD dataset and on pseudosentences.

mal constraints and world knowledge during scope interpretation, aligning with prominent psycholinguistic accounts (e.g., Kurtzman and MacDonald 1993; Trueswell et al. 1994; Dwivedi 2013; Feiman and Snedeker 2016; Dwivedi et al. 2018) At the same time, clear differences emerge. As shown in Table 6, humans do not exhibit a comparable bias toward surface-scope interpretations. In particular, humans do not show a comparably strong dispreference for inverse-scope readings when lexical content lacks real-world meaning.

A more detailed breakdown by quantifier-combination type is shown in Table 7. On the balanced dataset, accuracy is highest for combinations in which classical surface predictors support the preferred interpretation, but remains high even for inverse-scope combinations. On pseudosentences, performance drops markedly for inverse-scope combinations, while surface-scope combinations are resolved more reliably. However, the surface advantage is less pronounced than for language models.

These findings indicate that while LLMs approximate human scope preferences under naturalistic conditions, they rely more heavily on surface-level heuristics when access to world knowledge is restricted. This manifests as inverse scope fragility, whereby inverse-scope interpretations degrade disproportionately in the absence of real-world referents. In contrast, human participants exhibit greater robustness to the removal of real-world referents, maintaining relatively stable performance even for non-surface interpretations.

4 Experiment 2: World Knowledge via Retrieval

In Experiment 2, we manipulate access to *external* world knowledge via retrieval on the balanced QSD dataset, 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). RAG has demonstrated effectiveness across a variety of NLP tasks that benefit from access to external knowledge (e.g., Lewis et al. 2020; Gao et al. 2023). We implement dense retrieval over a combined ConceptNet and Simple Wikipedia corpus using Sentence-Transformer embeddings and a FAISS index. For each input sentence, we use the stimulus sentence as the retrieval query to obtain a small set of relevant passages, which are then appended to the prompt using the same instructions and answer format as in Experiment 1, so that any differences in performance between conditions can be attributed directly to the availability of retrieved world knowledge.

4.2 Results

Table 8 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.

For GPT and Qwen models, a consistent pattern emerges: inverse-scope accuracy increases under RAG, while surface-scope performance slightly decreases. As a result, improvements in overall accuracy are primarily driven by gains on inverse-scope interpretations—precisely those configurations that are disfavored by classical surface-based predictors. The two Llama models diverge from this pattern in distinct ways. Llama 3.1 8B shows improvements for both surface-scope and inverse-scope interpretations. In contrast, Llama 3.1 70B exhibits a qualitatively different response: RAG decreases

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

Table 8: Comparison of simple prompting models and RAG-augmented models on the balanced QSD dataset. Results are averaged 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 9: Average effect of RAG augmentation on balanced QSD performance across all models. Gains are concentrated on inverse-scope interpretations, while surface-scope accuracy slightly decreases on average.

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 10: Average accuracy across quantifier-combination types on the balanced QSD dataset, averaged over all models. RAG selectively improves the hardest inverse-scope configuration (IV), while leaving other configurations unchanged or slightly degraded.

accuracy for both scope types, suggesting that retrieved context can sometimes interfere with rather than support scope resolution.

4.3 Discussion

Tables 9 and 10 show that the impact of RAG augmentation is selective rather than uniform, and closely tied to the interpretive difficulty of the scope configuration. At the aggregate level (Table 9)[4], RAG yields only a limited improvement in overall accuracy. Crucially, this effect is not distributed evenly across scope types. Surface-scope interpretations slightly decrease under RAG, while inverse-scope interpretations improve. This asymmetric pattern suggests that retrieved information does not simply boost performance across the board, but instead preferentially targets cases where formal cues are insufficient.

This picture becomes clearer when examining performance by quantifier-combination type (Table 10). 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 an improvement, precisely in the condition where models must override surface-level expectations and rely on broader knowledge about plausible event structures.

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 sub-

[4]Including both Llama variants in the average leads to partial cancellation in the surface-scope aggregate, as RAG induces opposite Δ effects for the 8B and 70B models, while inverse-scope gains remain positive overall because the improvement for Llama 3.1 8B outweighs the small degradation observed for Llama 3.1 70B.

stantially on pseudosentences that remove such referents, with the largest degradation observed for inverse-scope interpretations, reflecting a systematic fragility of non-surface readings when world knowledge is removed. This pattern shows that surface-level cues alone are insufficient to account for model behavior and suggests a substantive contribution of world knowledge to LLM interpretation.

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 primarily affect inverse-scope interpretations, and most clearly the hardest configuration (U–E, inverse), where both classical surface predictors conflict with the preferred reading. Taken together, our results provide converging evidence that LLM intepretive behavior reflects an interaction between formal interpretive pressures and world knowledge, with world knowledge—implicit or retrieved—playing a particularly important role when formal cues are insufficient or in conflict.

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. 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 key differences in dataset design. In particular, the Kamath et al. dataset is skewed toward surface-scope readings, as it is derived from a natural corpus of LSAT examples originating in AnderBois et al. (2012), whereas our dataset is explicitly balanced and augmented to include naturally underrepresented inverse-scope readings.

As in prior work (Kamath et al., 2024; Fang and Cong, 2025), models exhibit lower accuracy for inverse-scope interpretations. Notably, this pattern does not align with human behavior, as humans showed no comparable dispreference, both in Kamath et al.'s experiments and in ours. In our study, this divergence between models and humans becomes especially apparent in the pseudosentence condition, where real-world referents are removed. Under these conditions, human participants exhibit greater robustness than LLMs, particularly for inverse-scope readings. This contrast points to a qualitative difference in how humans and LLMs balance formal constraints and background knowledge when semantic content is impoverished. Understanding the source of this robustness remains an important direction for future work.

Our results also resonate with recent work on the role of retrieval-augmented generation in language interpretation (e.g., Cuconasu et al. 2024; Chatzikyriakidis 2025). Retrieval does not act as a uniform performance enhancer, but instead interacts with model behavior in a highly selective and sometimes disruptive manner. Against this backdrop, the limited overall gains observed in Experiment 2 are theoretically informative rather than unexpected. Retrieval improves accuracy primarily in configurations where implicit world knowledge is insufficient and surface-level predictors are misleading, while leaving structurally well-supported cases unchanged and occasionally introducing interference elsewhere. This pattern indicates that retrieved knowledge modulates interpretation through its interaction with linguistic form, selectively enabling readings that are otherwise disfavored by surface predictors.

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.

Yue Li Fang, Shaohua and Yan Cong. 2025. Quantifier scope interpretation in language learners and LLMs. *arXiv preprint arXiv:2509.10860*.

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.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yixin Dai, Jiawei Sun, Haofen Wang, and Haofen Wang. 2023. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*.

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.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, and 1 others. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474.

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.