It Depends: Resolving Referential Ambiguity in Minimal Contexts with Commonsense Knowledge

Lukas Ellinger and Georg Groh
School for Computation, Information and Technology
Technical University of Munich, Germany
lukas.ellinger@tum.de, grohg@cit.tum.de

Abstract

Ambiguous words or underspecified references require interlocutors to resolve them, often by relying on shared context and commonsense knowledge. Therefore, we systematically investigate whether Large Language Models (LLMs) can leverage commonsense to resolve referential ambiguity in multi-turn conversations and analyze their behavior when ambiguity persists. Further, we study how requests for simplified language affect this capacity. Using a novel multilingual evaluation dataset, we test DeepSeek v3, GPT-4o, Qwen3-32B, GPT-4o-mini, and Llama-3.1-8B via LLM-as-Judge and human annotations. Our findings indicate that current LLMs struggle to resolve ambiguity effectively: they tend to commit to a single interpretation or cover all possible references, rather than hedging or seeking clarification. This limitation becomes more pronounced under simplification prompts, which drastically reduce the use of commonsense reasoning and diverse response strategies. Fine-tuning Llama-3.1-8B with Direct Preference Optimization substantially improves ambiguity resolution across all request types. These results underscore the need for advanced fine-tuning to improve LLMs' handling of ambiguity and to ensure robust performance across diverse communication styles.

1 Introduction

Natural language is inherently ambiguous. For example, pronouns may refer to multiple possible entities within a sentence. Nevertheless, humans typically resolve such ambiguity by drawing on context, shared knowledge, and conversational history (Ferreira, 2008). Consider the two conversations shown in Figure 1, where the user asks the question, "Why can **it** fly?". Without additional clues, the pronoun "it" is unclear and could refer to multiple entities. In the left conversation, the prior context mentions a helicopter and a drum; in the right, it additionally includes an owl. Humans effortlessly combine this context with commonsense knowledge, recognizing that drums cannot fly, but helicopters and owls can. As a result, the first case is unambiguous, while the second may require clarification.

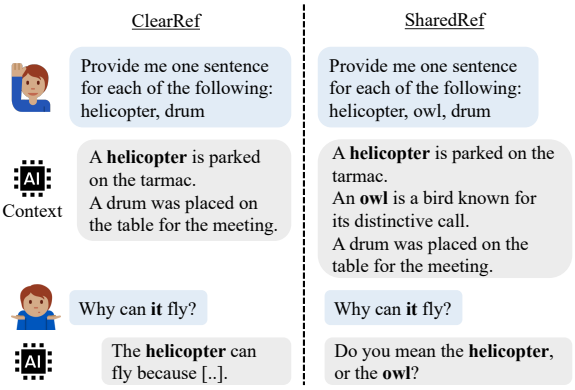

Figure 1: Two conversations between a user and an LLM in response to the ambiguous question ("Why can **it** fly?"). In both cases, the LLM uses prior context to narrow the possible referents to entities capable of flying. In the left conversation, it attempts an answer; in the right, it asks for clarification.

This process reflects a fundamental feature of human communication: a "division of labor" between speakers and listeners, where speakers omit explicit details to minimize effort, trusting listeners to fill in gaps using common ground (Ferreira, 2008). Common ground consists of the mutual knowledge, beliefs, and assumptions interlocutors accumulate and maintain during conversation (Clark and Brennan, 1991; Clark, 1996). Central to common ground is commonsense knowledge, a broadly shared understanding of the world that enables people to make implicit inferences effortlessly.

As mentioned, humans are usually good at building and using common ground. While prior work suggests that LLMs struggle with ambiguity resolution, particularly in static, single-turn contexts

Proceedings of the 2nd Workshop on Uncertainty-Aware NLP (UncertaiNLP 2025), pages 229–246
November 9, 2025

(Liu et al., 2023), our work shifts focus to a conversational setting. We study how LLMs behave in multi-turn dialogs where common ground is explicitly established through conversation history and commonsense knowledge. In our setting, multiple referents can remain plausible even after considering prior context. This allows us to evaluate how models handle uncertainty through different response strategies, such as requesting clarification.

We further examine how language constraints affect this ability. Language models are increasingly used to generate output in different variants, such as simplified and easy-to-understand language. This has clear benefits for accessibility, particularly for users with cognitive or linguistic challenges (Freyer et al., 2024). However, simplified outputs often reduce the depth and precision of content (Trienes et al., 2024). Ellinger et al. (2025) find that models prompted to define homonyms in simple language often default to the most salient meaning, disregarding less dominant but valid definitions. We explore whether such requests for simplified language also affect a model's capacity to resolve ambiguity when multiple interpretations are plausible.

Studying this is crucial because misinterpretation of ambiguous language can lead to downstream failures such as misinformation, hallucinations, or user confusion. By systematically testing whether LLMs consider multiple plausible candidates rather than relying on recency or default biases, we provide a diagnostic view of their behavior in ambiguous conversational settings.

Our contributions are as follows:

- We introduce a multilingual dataset for evaluating LLMs to resolve referential ambiguity in conversations with explicit common ground.
- We evaluate DeepSeek v3, GPT-4o, Qwen3-32B, GPT-4o mini, and Llama 3.1 8B using both LLM-as-Judge and human annotations.
- We show that LLMs often commit to a single interpretation or cover all references instead of hedging or clarifying. Simplified language constraints worsen this by reducing commonsense reasoning and response diversity.
- We fine-tune LLaMA 3.1 8B with Direct Preference Optimization (DPO), achieving significant improvements on our task that generalize to a lexical ambiguity benchmark, with less degradation under simplified prompts.

2 Background and Related Work

Ambiguity and Clarification. Understanding language often requires resolving ambiguity, such as referential ambiguity, where it is unclear which entity a phrase refers to. Such unclear references slow down human processing (Gernsbacher, 1989; MacDonald and MacWhinney, 1990; Myers and O'Brien, 1998; Stewart et al., 2007), yet humans are usually good at resolving them by drawing on common ground.

In contrast, LLMs struggle with ambiguity. Min et al. (2020) introduce AmbigQA, a dataset designed to investigate underspecified questions, and subsequent studies (Wildenburg et al., 2024; Liu et al., 2023) show that even state-of-the-art models underperform in such settings. This limitation extends to the multimodal domain: Testoni et al. (2024) find that vision–language models also handle ambiguity poorly, often replying with overconfident or biased outputs. While their focus is on visual context, the challenge is related to ours, with textual context instead of images.

Models also rarely seek clarification. Kuhn et al. (2023) show that LLMs often respond incorrectly to ambiguous inputs rather than asking follow-up questions. Prior work confirms this lack of clarification behavior (Benotti and Blackburn, 2017; Xu et al., 2019; Shi et al., 2022). Herlihy et al. (2024) link this tendency to fine-tuning biases and propose a taxonomy of model responses, which we adopt.

Prior work mainly studies ambiguity in static, single-turn settings without common ground. Notably, datasets for anaphora resolution, such as the Winograd Schema Challenge (Levesque et al., 2012), focus on single-sentence coreference, where exactly one antecedent is correct and can be identified using commonsense reasoning. In contrast, we study LLMs in multi-turn dialogs where common ground is explicitly established through conversation history and commonsense knowledge. In our setting, multiple referents can remain plausible even after considering context. This allows us to evaluate how models handle uncertainty through different response strategies, such as direct answers, hedging, or requesting clarification, rather than simply selecting the correct noun.

Finally, we test if our fine-tuned model generalizes to lexical ambiguity using the benchmark of Ellinger et al. (2025), which evaluates homonym definitions without disambiguating context.

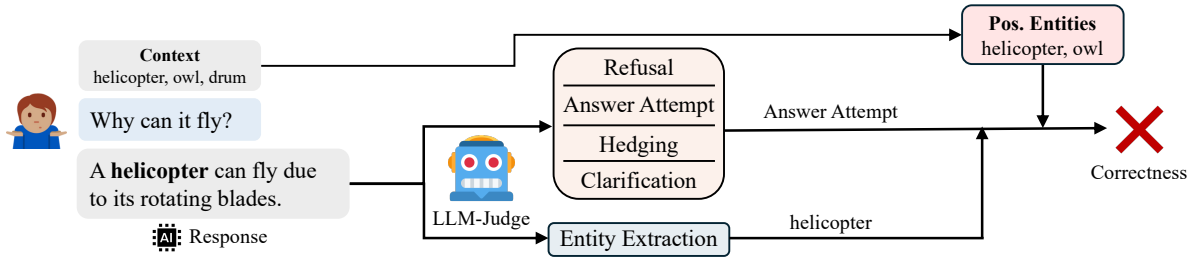

Figure 2: Evaluation pipeline including LLM-Judge for response categorization and entity extraction. Based on these outputs and the positive entities identified in the context, we determine the fine-grained response category and assess correctness with respect to entity resolution.

Commonsense Evaluation. Prior work systematically evaluated LLMs on commonsense reasoning benchmarks. Li et al. (2022) conduct evaluations under zero- and few-shot settings across four benchmarks, revealing that pre-trained LMs struggle to acquire commonsense knowledge without task-specific supervision. Scaling model size or adopting to few-shot prompting does not suffice to reach human-level performance. Similarly, Bian et al. (2024) assess ChatGPT on eleven commonsense QA datasets. They find that ChatGPT can retrieve relevant knowledge via prompting. However, it often fails to identify and apply the specific commonsense required to answer a given question. In the multimodal domain, Fu et al. (2024) introduce Commonsense-T2I, the first benchmark evaluating whether text-to-image models generate images consistent with commonsense knowledge. They find that state-of-the-art models achieve only 49% accuracy, indicating significant gaps in visual commonsense understanding.

Our work extends these by exploring another dimension of commonsense. Unlike prior benchmarks focused on question answering or image alignment, we assess whether models recognize ambiguous referents and either disambiguate or request clarification, demonstrating a context-aware application of commonsense reasoning.

Simple Language. Simplified language aims to improve accessibility for a broad range of users, including non-native speakers, children, domain novices, and individuals with cognitive impairments. Its availability is endorsed by the Web Content Accessibility Guidelines (WCAG) to promote inclusive communication (W3C, 2025). Simplified language involves straightforward vocabulary, clear sentence structure, minimal jargon, and the avoidance of complex grammar (Freyer et al., 2024). Domains like healthcare, law, and education already widely apply it (Garimella et al., 2022; Deilen et al., 2024; Rets et al., 2022). However, prior work has shown that simplification in LLM-generated text can lead to undesirable side effects such as omissions or overly vague formulations (Anschütz et al., 2025; Agrawal and Carpuat, 2024; Devaraj et al., 2022). Ellinger et al. (2025), for instance, report that when asked to define homonyms in simplified language, models tend to default to the most salient meaning, neglecting valid but less frequent senses.

Building on this line of work, we study how simplification constraints affect a model's ability to resolve referential ambiguity and how task-specific finetuning affects performance in the lexical ambiguity benchmark of Ellinger et al. (2025).

3 Methodology

We evaluate whether LLMs can resolve referential ambiguity using common knowledge and how requests for simplified language affect this ability. Each test instance consists of a short context passage introducing some entities (e.g., *helicopter*, *owl*, *drum*). The user then asks an ambiguous question referring to one of the entities without naming it directly (e.g., *Why can it fly?*). For each instance, we define a set of positive entities as those for which the question makes sense, and negatives as those for which it does not (e.g., a drum cannot fly). We evaluate two setups: *ClearRef*, where one positive and one negative entity make the referent unambiguous with commonsense, and *SharedRef*, where two positives and one negative leave ambiguity even with commonsense. This setup tests whether models consider multiple plausible candidates rather than relying on recency or default biases. We treat the pronoun "it" as equally applicable to all introduced positive entities. To assess the impact of recency, we perform an ablation in which the order of entities is permuted (see Appendix D).

Ambiguous Questions by Relation

Rel. 1: Why can it **fly**?
Rel. 2: Why is it **sweet**?
Rel. 3: Why is it **made of wood**?
Rel. 4: Why can it **swim**?
Rel. 5: Why can it **run fast**?
Rel. 6: Why can it **climb trees**?
Rel. 7: Why is it **hot**?
Rel. 8: Why is it **loud**?

Simple: [..] Respond in simple language.

Figure 3: Ambiguous questions for our eight relations. In the Simple setting, an instruction is appended. Exact relations names in Appendix B.

3.1 Dataset

We construct our datasets based on ConceptNet (Speer et al., 2017), a knowledge graph that encodes commonsense relationships between entities and attributes. We select eight relations, such as *capable of flying*, and extract all associated entities. Figure 3 provides the complete list of relations. Since each dialog requires a context passage, we use GPT-4.1-nano to generate a concise sentence for every entity. These sentences, each beginning with the entity name, serve as the context passages for all related evaluations.

For *ClearRef*, each entity is paired with a negative sample from a different relation. We use GPT-4.1-nano to verify that the negative entity does not satisfy the target relation. For *SharedRef*, we create samples by pairing all entities within the same relation and similarly pick a negative. This results in 52 *ClearRef* and 227 *SharedRef* examples. We list further details in Appendix B.

To enable multilingual evaluation, we translate the context sentences and entities into Arabic, French, Russian, and Simplified Chinese using the DeepL API[1]. We choose these languages to facilitate comparison with the multilingual setting of Ellinger et al. (2025).

3.2 Model and Prompt Configuration

We evaluate five LLMs on our task: GPT-4o, GPT-4o-mini (OpenAI et al., 2024), Qwen3-32B (Qwen Team, 2025), DeepSeek v3 (DeepSeek-AI et al., 2025), and Llama 3.1 8B (Grattafiori et al., 2024). These models vary in size and openness, enabling a comprehensive analysis of performance across diverse LLMs. Details on model versioning and access are listed in Appendix A.

We evaluate eight relations, each associated with an ambiguous question. For each, we test two prompt settings: **Normal**, presenting only the ambiguous question, and **Simple**, which adds an instruction to respond in simplified language. This setup allows us to examine how constraining outputs to simpler language affects model responses. English prompts are shown in Figure 3, with multilingual versions in Appendix Figure 12.

3.3 Evaluation Pipeline

The input to the evaluation pipeline (Figure 2) consists of a brief dialogue between a user and an LLM, exemplified in Figure 1. The response to the dialogue is passed to our LLM-Judge, which performs two tasks. First, it classifies the response type into one of four categories: *Refusal*, *Answer Attempt*, *Hedging*, or *Clarification* (cf. subsection 3.4). In this case, the response is labeled as an *Answer Attempt*. Second, it extracts all entities mentioned in the response (here, *helicopter*). Using the set of mentioned entities and the known positive entities (in this case, *helicopter* and *owl*), we assess the correctness of the response. Since the model attempts an answer but only mentions one of the two positive entities, the response is marked as incorrect.

3.4 Response Categorization

Following Laban et al. (2025), we adopt the response taxonomy from Herlihy et al. (2024), which includes *Answer Attempt*, *Clarification*, *Interrogation*, *Discussion*, *Hedging*, *Refusal*, and *Missing*. Focusing on referential ambiguity resolution, we simplify this taxonomy by merging *Interrogation* into *Clarification* and *Discussion* into *Answer Attempt*, reducing annotation complexity. Full definitions and examples appear in Appendix E. Briefly:

- **Hedging**: The assistant uses conditional or speculative language (e.g., “might be...”, “if you meant X...”).
- **Clarification**: The assistant requests more information without offering interpretations or using hedging.
- **Answer Attempt**: The assistant clearly commits to at least one interpretation, providing a factual response without any hedging.

[1] https://www.deepl.com/en/pro-api

We define a response as *correct* if it appropriately addresses the ambiguity in the input. Clarifications are always correct, as they seek additional input without committing to an interpretation. Hedging responses are considered correct, as long as they mention at least one entity. While they do not resolve the ambiguity, they acknowledge it and express uncertainty in a transparent way. In contrast, answer attempts are only deemed correct if they explicitly mention both positive entities.

Herlihy et al. (2024) discuss the trade-off between the usefulness and cognitive cost of different response categories, approximated by response length. In our setting, we argue that the most desirable responses, regardless of the category, are those that mention all and only the positive entities. We refer to these as **direct** responses. They reflect correct disambiguation based on common knowledge while minimizing user effort through clear and concise answers, free of irrelevant distractors.

In *SharedRef*, we consider any *direct* response the most appropriate response. In contrast, for *ClearRef*, where the ambiguity can be fully resolved, an *Answer Attempt* is preferred.

3.5 Automatic Evaluation

We designed an automated evaluation framework that leverages GPT-4.1-mini as an LLM-Judge. The framework assesses model responses based on the response categories defined in subsection 3.4. It classifies responses and extracts explicitly mentioned entities. A few-shot prompt, detailed in Appendix F, guides the evaluation. To validate the framework, one author manually labeled 500 responses from the English dataset, with 100 responses per evaluated model (50 for the standard prompt and 50 for the simple prompt). The annotator performed both response classification and extraction of explicitly mentioned entities, exactly as the LLM was tasked to do. The LLM judge achieved a 98% agreement rate on response classification and a Cohen's Kappa score of 0.916, indicating almost perfect agreement according to Landis and Koch (1977). For entity extraction, the framework achieved a 97.8% exact match accuracy. More details are provided in Appendix F.

3.6 Direct Preference Optimization

We fine-tuned Llama-3.1-8B to improve referential ambiguity resolution using DPO (Rafailov et al., 2024). DPO aligns model behavior with desired outcomes by training on preference pairs. In our setup, we favor direct over incorrect responses.

Our training dataset contains 1,388 preference pairs across all languages by comparing incorrect Llama 3.1 8B's outputs with *direct* responses from other models. To prevent reliance on entity position, we randomly permuted the order within each conversation. We restricted the training data to the '*capableOf fly*' relation, allowing us to later assess generalization to other relations.

We performed a single training run using the whole training set. This decision reflects our aim to demonstrate the feasibility of aligning models to produce more useful responses with lower cognitive cost, rather than optimizing for peak performance through extensive tuning. Detailed training information is provided in Appendix G.

4 Results

4.1 ClearRef Dataset

Figure 4 shows that all models maintain correctness above 90% across languages and settings, with some achieving perfect scores. The lowest correctness score is 90.38%, observed for Deepseek v3 (Simple) and Llama-3.1-8B (Normal) in French. When comparing the Normal and Simple settings, GPT-4o is the only model with higher correctness in the Normal setting, while the other models either remain similar or slightly decrease. The rate of direct responses among the correct answers varies drastically across models and languages. In the Simple setting, Qwen3-32B shows the highest variance, with a direct response rate ranging from as low as 22.45% in Arabic to 73.08% in English. In the Normal setting, GPT-4o-mini varies most, with only 47.06% direct responses in Russian to 82.69% in English. Llama-3.1-8B demonstrates the highest rates for English, achieving 98.00% in Normal and 97.96% in Simple. Averaged across languages, mean direct responses among all responses differ by model and setting. Except for Deepseek v3, all models show higher direct response rates in the Normal setting compared to Simple. In Normal, Llama-3.1-8B achieves the highest rate (80.38%), followed by GPT-4o, GPT-4o-mini, Qwen3-32B, and Deepseek v3 (58.85%). Detailed breakdowns by model, language, and prompt type are provided in Appendix Table 9.

In Figure 5, we show the distribution of response categories across languages and models. In all cases, *Answer Attempt* is the dominant category. However, comparing the Normal and Simple set-

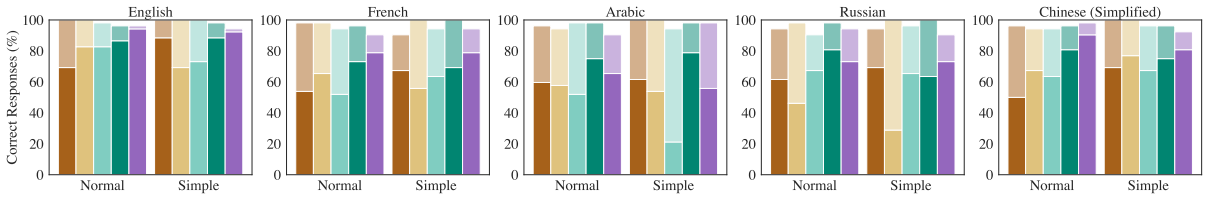

Figure 4: Percentage of correct responses across five languages on the ClearRef dataset. Colored squares indicate different models: DeepSeek v3, GPT-4o-mini. Qwen3-32B, GPT-4o, and Llama-3.1-8B. The darker portion of each bar represents the percentage of Direct Responses.

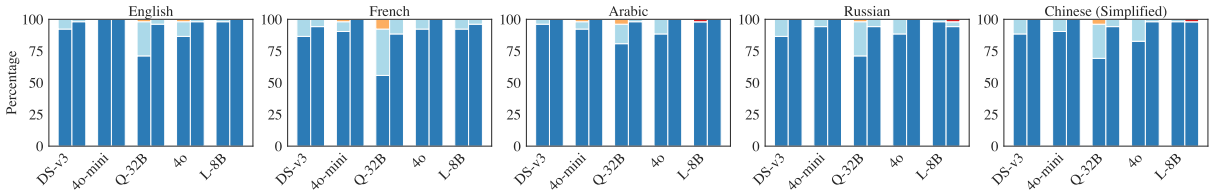

Figure 5: Distribution of the defined response categories across five languages on the ClearRef dataset. For each model, the left bar represents the Normal setting and the right bar the Simple setting. Colored squares represent response types: Answer Attempt, Hedge, Clarification, and Refuse.

tings reveals a shift: In the Simple setting, models nearly always produce answer attempts (mean 97.92%). In Normal, especially with Qwen3-32B, hedging occurs more frequently, and to a lesser extent, clarifications. For Qwen3-32B, the average proportion of Answer Attempts drops to 69.61%.

4.2 SharedRef Dataset

We show proportions of correct responses along with direct response rates in Figure 6. The results reveal a sharp drop from Normal to Simple and a clear separation between two model groups: high performers (GPT-4o, Qwen3-32B, Deepseek v3) and low performers (Llama-3.1-8B, GPT-4o-mini).

Low-performing models show poor performance across languages and prompt settings, with GPT-4o-mini reaching below 13% correctness in the Normal setting and Llama-3.1-8B slightly higher but inconsistent due to an outlier in the Arabic Simple setting.

Among the top performers, GPT-4o achieves the highest correctness in English Normal prompts (81.06%, thereby 45.11% direct), while Qwen3-32B performs best overall when averaged across languages in the Normal setting (70.22%, 31.11%). Deepseek v3 leads in the Simple setting (37.97%, 22.73%), outperforming the others despite lower direct response rates.

Performance also varies notably by language. In the Normal setting, English (69.16% correct, thereby 47.28% direct) and Chinese (63.96%, 41.54%) achieve the highest average correctness, followed by Arabic, French, and Russian (51.19%, 45.71%), reflecting the models' native strengths (e.g., GPT for English, Qwen and Deepseek for Chinese). In the Simple setting, Arabic leads (50.22%, 44.01%), followed by Chinese and English, with French and Russian (26.08%, 59.81%) trailing. We show a detailed breakdown per model, language, and prompt type in Appendix Table 10.

Figure 7 shows the distribution of response categories across languages and models. Consistent with ClearRef, Answer Attempt remains the dominant category in the Simple setting, with an average proportion of 97.01% across all languages and models. The only notable outlier is Qwen3-32B in Chinese, with a lower proportion of 72.69%.

In the Normal setting, the shift toward other response categories becomes more pronounced than in ClearRef. The average proportion of Answer Attempts decreases to 77.67%. Notable deviations include GPT-4o in English (29.52%) and Russian (46.26%), as well as Qwen3-32B in English (43.17%), French (49.34%), Russian (40.97%), and Chinese (43.61%). These two models show marked increases in Hedging (GPT-4o from 1.67% to 35.06%, Qwen3-32B from 8.37% to 41.14%) and Clarification (GPT-4o from 0.09% to 4.76%, Qwen3-32B from 0.70% to 8.02%).

4.3 Direct Preference Optimization

We compare the base and the fine-tuned model on the SharedRef test set, excluding the *capableOf fly* relation among positives. Figure 8 shows that the

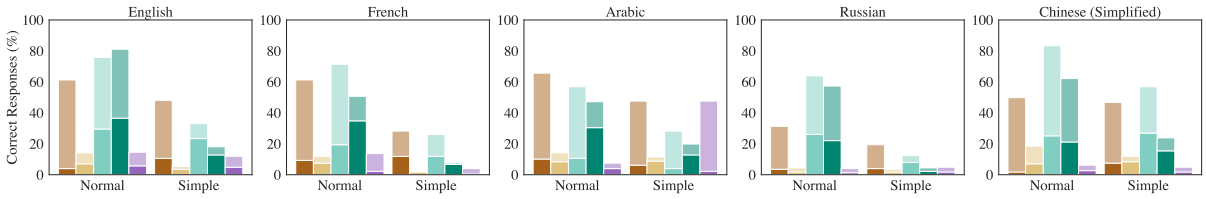

Figure 6: Percentage of correct responses across five languages on the SharedRef dataset. Colored squares indicate different models: DeepSeek v3, GPT-4o-mini. Qwen3-32B, GPT-4o, and Llama-3.1-8B. The darker portion of each bar represents the percentage of Direct Responses.

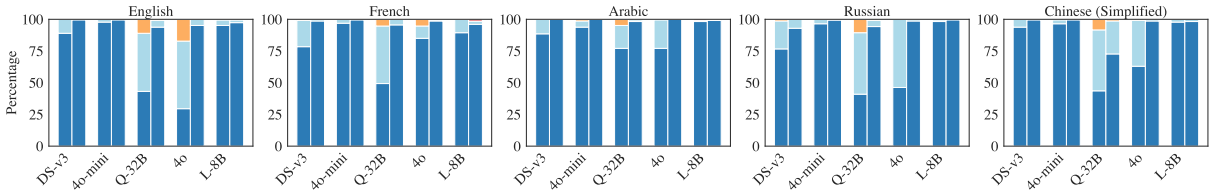

Figure 7: Distribution of the defined response categories across five languages on the SharedRef dataset. For each model, the left bar represents the Normal setting and the right bar the Simple setting. Colored squares represent response types: Answer Attempt, Hedge, Clarification, and Refuse.

results are consistent across languages. Overall, the proportion of correct responses increases from 13.46% to 96.45% in the Normal setting and from 13.83% to 91.59% in the Simple setting. Among the correct responses, direct responses rise from 28.60% to 42.96% (Normal) and from 33.37% to 50.66% (Simple). For comparison, the best base model, Qwen3-32B, achieves 62.43% correct (30.44% direct) in the Normal setting and 22.06% correct (60.97% direct) in the Simple setting.

The category distribution shifts drastically. In the base model, Answer Attempts dominate (91.78% in Normal, 96.45% in Simple). After fine-tuning, Clarification is most frequent, followed by Hedging and Answer Attempts. In the Simple setting, Clarification is less dominant than in Normal, while Hedging becomes more prevalent: 60.00% vs. 30.84% Clarification, 36.07% vs. 52.52% Hedging, and 3.74% vs. 16.63% Answer Attempts.

4.4 Homonym Definition Generation

Ellinger et al. (2025) introduced MCL-WiC, a multilingual homonym dataset, along with the *Sense Awareness* metric for evaluation. A response shows Sense Awareness by providing multiple definitions or explicitly acknowledging ambiguity via clarification requests or remarks about alternative meanings. They evaluated model performance under standard, simplified, and ELI5-style prompting (Fan et al., 2019), where the model explains a word as if the user were five years old.

Table 1 compares our fine-tuned model with the results reported by Ellinger et al. (2025). Against baseline models, our model achieves the highest Sense Awareness under the Normal prompt in English, French, and Russian, the second-highest in Arabic, and competitive results in Chinese. For Simple, it ranks highest in French and Russian, with comparable results in other languages. For ELI5, it outperforms all baseline models in every language except English, where it ranks second. Compared to its base model, our fine-tuned version shows consistent, mostly extensive improvements across all configurations, with the only exception being the English Simple setting, where performance drops by three percentage points.

They also fine-tuned Llama-3.1-8B on the same task. Their model produces English outputs for all languages except Russian, reflecting heavy optimization for English. In contrast, our DPO model handles all languages natively. While their fine-tuned model generally achieves higher Sense Awareness scores, our model remains competitive against the baseline models and narrows the gap in the language constraints. Their fine-tuning was explicitly targeted at this task, and reducing the gap between the language constraints. In contrast, our model achieves strong results across all languages without task-specific tuning.

5 Discussion

Our results indicate that current models struggle to apply commonsense knowledge for ambiguity

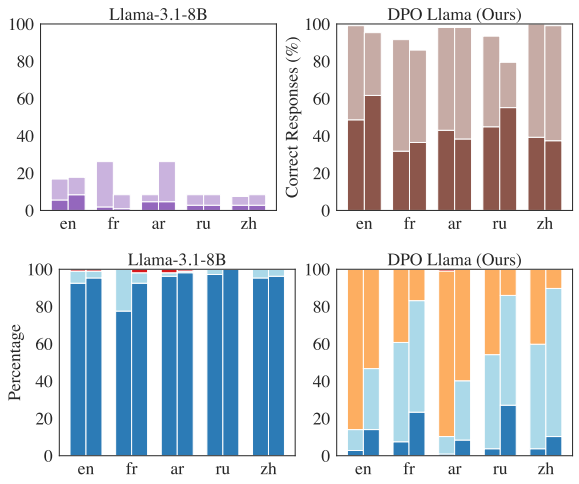

Figure 8: Comparison of the base model and our DPO-fine-tuned model across five languages on the SharedRef test set. For each language, the left bar represents the Normal setting, and the right bar the Simple setting.
Top: Percentage of correct responses. The darker portion of each bar represents the percentage of Direct Responses.
Bottom: Distribution of response categories. Colored squares denote: ■ Answer Attempt, ■ Hedge, ■ Clarification, ■ Refuse.

Prompt / Model	Sense Aware				
	En	Fr	Ar	Ru	Zh
Prompt: Normal					
3.1 8B	*96.95*	15.17	10.62	6.52	4.66
4o-mini	93.90	79.31	92.92	90.43	84.46
3-30B A3B	94.58	86.55	**98.23**	90.00	**100.00**
4 Maverick	96.27	54.83	74.34	75.65	45.08
v3	94.24	*87.93*	91.15	*91.74*	*87.56*
Our 3.1-8B	**97.63**	**97.93**	*93.81*	**99.57**	84.46
Their 3.1-8B	99.66	99.31	99.12	99.13	98.45
Prompt: Simple					
3.1 8B	64.41	7.59	6.19	2.17	7.77
4o-mini	63.05	52.76	*76.99*	43.91	*75.13*
3-30B A3B	**76.61**	*59.66*	69.03	*67.83*	**82.38**
4 Maverick	*69.83*	28.28	45.13	48.70	68.91
v3	63.73	47.93	**80.53**	65.22	74.09
Our 3.1-8B	61.02	**73.10**	71.68	**79.13**	64.25
Their 3.1-8B 8B	92.88	93.45	96.46	99.57	94.30
Prompt: ELI5					
3.1 8B	7.12	7.59	0.88	1.30	0.52
4o-mini	5.42	6.90	10.62	2.61	6.74
3-30B A3B	**22.03**	*17.24*	9.73	*14.78*	*14.51*
4 Maverick	10.85	13.10	11.50	9.57	9.84
v3	8.14	8.28	*13.27*	8.70	10.88
Our 3.1-8B	*13.22*	**25.86**	**46.90**	**19.13**	**17.62**
Their 3.1-8B	35.59	35.17	55.75	63.48	33.68

Table 1: Sense Awareness scores by prompt type and language. Best results are in **bold**, second-best in *italic*. Model outputs are copied from the original paper.

resolution. In the simpler ClearRef task, where only one entity fits the question, models are able to resolve the ambiguity with an accuracy ranging from 94.23% down to 21.15% depending on the model and setting. The more challenging SharedRef task, which involves two plausible entities, sees direct responses ranging from just 36.56% down to 0.44%. This aligns with findings by Bian et al. (2024). They observe that LLMs can retrieve commonsense facts, which in our case means realizing that an entity fits a relation when asked on its own. However, the models often fail to apply this knowledge when answering a specific question requiring such reasoning. In Appendix C, we evaluate GPT-4o's performance in English under a Chain-of-Thought setting, prompting it to explicitly verbalize its commonsense reasoning first.

Consistent with Herlihy et al. (2024) and Kuhn et al. (2023), we observe that models frequently skip clarification, opting to answer even when uncertainty remains. Several models show almost no clarification or hedging behavior. Herlihy et al. (2024) and Singhal et al. (2024) argue that this behavior stems from reinforcement learning from human feedback (RLHF). Annotation processes typically focus on single-turn conversations. As a result, models are rarely exposed to examples of follow-up clarification questions, which require multi-turn interaction. Moreover, annotators often favor verbose, catch-all answers for under-specified queries, even though such verbosity imposes cognitive costs on users (Singhal et al., 2024).

Another important observation is that prompting models to use simpler language can harm response quality. Interestingly, in ClearRef, there is no drop from Normal to Simple; in some models, Simple responses are even slightly better. In contrast, for the more complex SharedRef task, performance drops drastically in the Simple setting. This confirms prior work showing that simplification often leads to omissions and vague phrasing (Ellinger et al., 2025; Anschütz et al., 2025; Trienes et al., 2024; Agrawal and Carpuat, 2024; Devaraj et al., 2022). We argue that this behavior needs to change. For example, Kearney et al. (2025) show that LLMs adapt the information they provide based on assumptions about the user. This is problematic, especially if requesting simple language causes models to produce less thoughtful responses or overlook important distinctions. Again, RLHF may play a role, failing to capture the needs of diverse users and discouraging clarification and hedging in simplified contexts.

Taken together, we argue that resolving ambiguity requires a balance: infer as much as possible to avoid unnecessary elaboration, but clarify when uncertainty remains. Our DPO-trained model moves in this direction. It not only improves on our main evaluation but also generalizes to the lexical ambiguity benchmark of Ellinger et al. (2025). Moreover, it reduces the performance drop commonly observed when models operate in simplified language settings. This suggests that clarification and hedging behaviors can be learned in a transferable and robust way.

6 Conclusion

In this paper, we analyzed how LLMs handle textual referential ambiguity and to what extent they apply commonsense knowledge to resolve it. Our findings show that LLMs have limited ability to do so effectively. They tend to commit to a single interpretation or cover all possible references, rather than hedging or seeking clarification. This tendency becomes even more pronounced when users request simple language, which reduces commonsense reasoning and different answering strategies.

These results point to two core issues. First, there is a need for better fine-tuning to improve how LLMs deal with ambiguity. Second, LLMs should better adapt to different user needs. It is especially concerning that a request for simpler language leads to less thoughtful responses and fewer clarifications, showing that current systems often fail to support users with varied communication styles.

To support reproducibility and future research, we release our code[2]. Further links to models and datasets are provided in the repository.

Limitations

Multilingual Scope and Dataset Size. Our study focuses on English, French, Russian, Arabic, and Chinese. For non-English languages, we relied on direct translations from English using automated tools, which can introduce translation bias, cultural mismatches, or loss of nuance. Future work should create native datasets for each language to ensure more accurate and culturally appropriate evaluation. Additionally, the ClearRef and SharedRef datasets contain only 52 and 227 datapoints, respectively, and include only 8 relations from ConceptNet, making it difficult to draw fully stable conclusions and potentially biasing evaluation toward certain categories. Nevertheless, we observe very strong tendencies in the results, suggesting that the findings are still meaningful and indicative of broader trends.

Referential Order. Due to computational limits, we used a fixed entity order; full permutation results for English are provided in Appendix D.

Commonsense Context. We provided all models with the same context, which included a commonsense fact sourced from ConceptNet. While these facts consist of basic relations and vocabulary, we cannot guarantee that models internally represent or utilize this knowledge. Nevertheless, given the simplicity and generality of the facts, the models likely have access to such information.

LLM-based Evaluation. We used an LLM to judge model responses, observing near-perfect agreement with human annotations in English. While we did not conduct human agreement checks for other languages, the observed trends remain consistent across all languages, suggesting broader applicability. Moreover, the differences between prompt settings are substantially larger than any potential error margin, further reinforcing the robustness of our findings.

Selected Prompts. We use fixed user prompts for each relation, along with a single predefined suffix for requesting responses in simplified language. This setup reflects how typical users might interact with a model without actively optimizing prompt phrasing. However, LLMs are known to be highly sensitive to prompt formulation, which can significantly influence output quality (Brown et al., 2020). Future research could systematically investigate the effects of varied or optimized prompts on LLM performance.

[2] https://github.com/lukasellinger/itdepends

References

Sweta Agrawal and Marine Carpuat. 2024. Do Text Simplification Systems Preserve Meaning? A Human Evaluation via Reading Comprehension. *Transactions of the Association for Computational Linguistics*, 12:432–448. Place: Cambridge, MA Publisher: MIT Press.

Miriam Anschütz, Anastasiya Damaratskaya, Chaeeun Joy Lee, Arthur Schmalz, Edoardo Mosca, and Georg Groh. 2025. (Dis)improved?! How Simplified Language Affects Large Language

Model Performance across Languages. In *Proceedings of the Fourth Workshop on Generation, Evaluation and Metrics (GEM²)*, pages 847–861, Vienna, Austria and virtual meeting. Association for Computational Linguistics.

Luciana Benotti and Patrick Blackburn. 2017. Modeling the clarification potential of instructions: Predicting clarification requests and other reactions. *Computer Speech & Language*, 45:536–551.

Ning Bian, Xianpei Han, Le Sun, Hongyu Lin, Yaojie Lu, Ben He, Shanshan Jiang, and Bin Dong. 2024. ChatGPT Is a Knowledgeable but Inexperienced Solver: An Investigation of Commonsense Problem in Large Language Models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 3098–3110, Torino, Italia. ELRA and ICCL.

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 Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Herbert H Clark. 1996. *Using language*. Cambridge university press.

Herbert H. Clark and Susan E. Brennan. 1991. Grounding in communication. In *Perspectives on socially shared cognition.*, pages 127–149. American Psychological Association, Washington, DC, US.

DeepSeek-AI, Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Daya Guo, Dejian Yang, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, and 181 others. 2025. DeepSeek-V3 Technical Report. *arXiv preprint*. ArXiv:2412.19437 [cs].

Silvana Deilen, Ekaterina Lapshinova-Koltunski, Sergio Hernández Garrido, Christiane Maaß, Julian Hörner, Vanessa Theel, and Sophie Ziemer. 2024. Towards AI-supported Health Communication in Plain Language: Evaluating Intralingual Machine Translation of Medical Texts. In *Proceedings of the First Workshop on Patient-Oriented Language Processing (CL4Health) @ LREC-COLING 2024*, pages 44–53, Torino, Italia. ELRA and ICCL.

Ashwin Devaraj, William Sheffield, Byron Wallace, and Junyi Jessy Li. 2022. Evaluating Factuality in Text Simplification. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7331–7345, Dublin, Ireland. Association for Computational Linguistics.

Lukas Ellinger, Miriam Anschütz, and Georg Groh. 2025. Simplifications are Absolutists: How Simplified Language Reduces Word Sense Awareness in LLM-Generated Definitions. *arXiv preprint*. ArXiv:2507.11981 [cs].

Angela Fan, Yacine Jernite, Ethan Perez, David Grangier, Jason Weston, and Michael Auli. 2019. ELI5: Long Form Question Answering. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3558–3567, Florence, Italy. Association for Computational Linguistics.

Victor S. Ferreira. 2008. Ambiguity, Accessibility, and a Division of Labor for Communicative Success. In Brian H. Ross, editor, *Psychology of Learning and Motivation*, volume 49 of *Advances in Research and Theory*, pages 209–246. Academic Press.

Nils Freyer, Hendrik Kempt, and Lars Klöser. 2024. Easy-read and large language models: on the ethical dimensions of LLM-based text simplification. *Ethics and Information Technology*, 26(3):50.

Xingyu Fu, Muyu He, Yujie Lu, William Yang Wang, and Dan Roth. 2024. Commonsense-T2I Challenge: Can Text-to-Image Generation Models Understand Commonsense? *arXiv preprint*. ArXiv:2406.07546 [cs].

Aparna Garimella, Abhilasha Sancheti, Vinay Aggarwal, Ananya Ganesh, Niyati Chhaya, and Nandakishore Kambhatla. 2022. Text Simplification for Legal Domain: Insights and Challenges. In *Proceedings of the Natural Legal Language Processing Workshop 2022*, pages 296–304, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Morton Ann Gernsbacher. 1989. Mechanisms that improve referential access. *Cognition*, 32(2):99–156.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The Llama 3 Herd of Models. *arXiv preprint*. ArXiv:2407.21783 [cs].

Christine Herlihy, Jennifer Neville, Tobias Schnabel, and Adith Swaminathan. 2024. On Overcoming Miscalibrated Conversational Priors in LLM-based Chatbots. *arXiv preprint*. ArXiv:2406.01633 [cs].

Matthew Kearney, Reuben Binns, and Yarin Gal. 2025. Language Models Change Facts Based on the Way You Talk. *arXiv preprint*. ArXiv:2507.14238 [cs].

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. 2023. CLAM: Selective Clarification for Ambiguous Questions with Generative Language Models. *arXiv preprint*. ArXiv:2212.07769 [cs].

Philippe Laban, Hiroaki Hayashi, Yingbo Zhou, and Jennifer Neville. 2025. LLMs Get Lost In Multi-Turn Conversation. *arXiv preprint*. ArXiv:2505.06120 [cs].

J. Richard Landis and Gary G. Koch. 1977. The Measurement of Observer Agreement for Categorical Data. *Biometrics*, 33(1):159–174. Publisher: International Biometric Society.

Hector J. Levesque, Ernest Davis, and Leora Morgenstern. 2012. The Winograd schema challenge. In *Proceedings of the Thirteenth International Conference on Principles of Knowledge Representation and Reasoning*, KR'12, pages 552–561, Rome, Italy. AAAI Press.

Xiang Lorraine Li, Adhiguna Kuncoro, Jordan Hoffmann, Cyprien de Masson d'Autume, Phil Blunsom, and Aida Nematzadeh. 2022. A Systematic Investigation of Commonsense Knowledge in Large Language Models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11838–11855, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Alisa Liu, Zhaofeng Wu, Julian Michael, Alane Suhr, Peter West, Alexander Koller, Swabha Swayamdipta, Noah Smith, and Yejin Choi. 2023. We're Afraid Language Models Aren't Modeling Ambiguity. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 790–807, Singapore. Association for Computational Linguistics.

Maryellen C MacDonald and Brian MacWhinney. 1990. Measuring inhibition and facilitation from pronouns. *Journal of Memory and Language*, 29(4):469–492.

Sewon Min, Julian Michael, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2020. AmbigQA: Answering Ambiguous Open-domain Questions. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5783–5797, Online. Association for Computational Linguistics.

Jerome L. Myers and Edward J. O'Brien. 1998. Accessing the discourse representation during reading. *Discourse Processes*, 26(2-3):131–157. Publisher: Routledge _eprint: https://doi.org/10.1080/01638539809545042.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, and 262 others. 2024. GPT-4 Technical Report. *arXiv preprint*. ArXiv:2303.08774 [cs].

Qwen Team. 2025. Qwen3.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. 2024. Direct Preference Optimization: Your Language Model is Secretly a Reward Model. *arXiv preprint*. ArXiv:2305.18290 [cs].

Irina Rets, Lluisa Astruc, Tim Coughlan, and Ursula Stickler. 2022. Approaches to simplifying academic texts in English: English teachers' views and practices. *English for Specific Purposes*, 68:31–46.

Zhengxiang Shi, Yue Feng, and Aldo Lipani. 2022. Learning to Execute Actions or Ask Clarification Questions. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 2060–2070, Seattle, United States. Association for Computational Linguistics.

Prasann Singhal, Tanya Goyal, Jiacheng Xu, and Greg Durrett. 2024. A Long Way to Go: Investigating Length Correlations in RLHF. *arXiv preprint*. ArXiv:2310.03716 [cs].

Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. ConceptNet 5.5: An Open Multilingual Graph of General Knowledge. *Proceedings of the AAAI Conference on Artificial Intelligence*, 31(1). Section: Special Track on Cognitive Systems.

Andrew J. Stewart, Judith Holler, and Evan Kidd. 2007. Shallow processing of ambiguous pronouns: evidence for delay. *Quarterly Journal of Experimental Psychology (2006)*, 60(12):1680–1696.

Alberto Testoni, Barbara Plank, and Raquel Fernández. 2024. RACQUET: Unveiling the Dangers of Overlooked Referential Ambiguity in Visual LLMs. *arXiv preprint*. ArXiv:2412.13835 [cs].

Jan Trienes, Sebastian Joseph, Jörg Schlötterer, Christin Seifert, Kyle Lo, Wei Xu, Byron Wallace, and Junyi Jessy Li. 2024. InfoLossQA: Characterizing and Recovering Information Loss in Text Simplification. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4263–4294, Bangkok, Thailand. Association for Computational Linguistics.

W3C. 2025. Web Content Accessibility Guidelines (WCAG) 2.1.

Frank Wildenburg, Michael Hanna, and Sandro Pezzelle. 2024. Do Pre-Trained Language Models Detect and Understand Semantic Underspecification? Ask the DUST! In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 9598–9613, Bangkok, Thailand. Association for Computational Linguistics.

Jingjing Xu, Yuechen Wang, Duyu Tang, Nan Duan, Pengcheng Yang, Qi Zeng, Ming Zhou, and Xu Sun. 2019. Asking Clarification Questions in Knowledge-Based Question Answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1618–1629, Hong Kong, China. Association for Computational Linguistics.

A Model Access

To support reproducibility, Table 8 lists all models used in this paper, including their abbreviated names (as used in tables and figures), full names, versions, and access providers.

B Dataset

We extracted entities from the following eight relations: `CapableOf fly`, `HasProperty sweet`, `MadeOf wood`, `CapableOf swim`, `CapableOf run_fast`, `CapableOf climb_trees`, `HasProperty hot`, and `HasProperty loud`. All entities were manually reviewed and cleaned. During dataset construction, we used the following prompt with GPT-4.1-nano to verify that each negative entity truly does not satisfy the relation, in contrast to the two positive entities:

User Prompt: Relation Satisfaction

```
Does the word '<word>' satisfy the relation
'<relation>'?
Answer with a brief explanation and either
True or False for satisfies.
```

C Ablation: Chain-of-Thought Prompting

Bian et al. (2024) observe that LLMs often fail to apply commonsense knowledge when answering questions that require such reasoning. To investigate this in our setting, we tested GPT-4o on the English SharedRef dataset in a Chain-of-Thought (CoT) setting. We choose GPT-4o as it showed the sharpest drop from Normal to Simple. We appended the following instructions to encourage CoT reasoning:

User Prompt: Chain-of-Thought

```
<question> First, try resolving any
ambiguity using commonsense knowledge. If
the question remains ambiguous, your
answer should be a clarification request.
Otherwise, provide the answer. Put your
final response after Response:.
```

We compare standard and CoT prompting in Figure 9. CoT prompting performs worse than standard prompting, with accuracy dropping from 81.06% to 44.49% in the Normal setting. This is because CoT prompting often only partially resolves the ambiguity, responding to one positive while ignoring the other. This occurs roughly 50% of the time, suggesting a model preference for one entity,

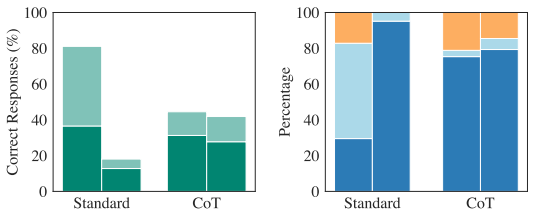

Figure 9: Comparison of Standard vs. CoT-Prompting on the SharedRef dataset. Left: Correctness; the darker portion of each bar indicates the percentage of Direct Responses. Right: Response category distribution. (Normal = left bar, Simple = right bar). Categories: Answer Attempt, Hedge, Clarification,

as it correctly identifies each entity when prompted individually. We observe more Clarifications and Answer Attempts, with nearly no Hedging in the Normal setting. The Simple setting is largely similar, contrasting with the standard Simple prompting.

Comparing the gap between Normal and Simple settings, we find it much smaller than in standard prompting. This suggests that when the LLM is explicitly guided on how to generate responses, there is no loss of thoughtfulness or omission of important distinctions. This is also reflected in the Simple CoT setting, performing better than the Simple standard prompting.

D Ablation: Permutation of Entity Ordering

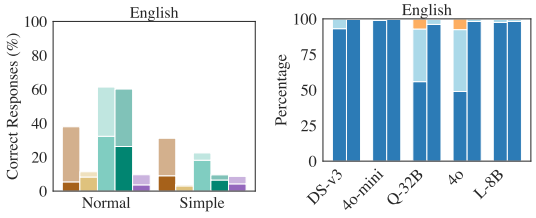

Figure 10: Average performance across all permutations in the English SharedRef dataset. Left: Correctness per model; the darker portion of each bar indicates the percentage of Direct Responses. Right: Response category distribution (Normal = left bar, Simple = right bar). Models: DeepSeek v3, GPT-4o-mini, Qwen3-32B, GPT-4o, Llama-3.1-8B. Categories: Answer Attempt, Hedge, Clarification, Refuse.

Our conversation context has a given order of entities. Due to computational constraints, we fixed the order to a single permutation for all evaluations ('0, 1, 2' for SharedRef and '0, 1' for ClearRef). We based this choice not on performance but to

Prompt / Model	Pos. 1	Pos. 2	Pos. 3
Prompt: Normal			
4o	34.50	24.52	40.98
4o-mini	41.98	18.65	39.37
v3	29.13	26.23	44.64
3-32B	31.23	29.37	39.40
3.1-8B	35.93	17.41	46.66
DPO Llama (Ours)	33.38	32.34	34.28
Prompt: Simple			
4o	37.29	15.80	46.91
4o-mini	41.67	16.46	41.87
v3	31.03	25.75	43.23
3-32B	30.35	28.36	41.29
3.1-8B	35.04	16.53	48.43
DPO Llama (Ours)	33.73	32.30	33.97

Table 2: Average selection rate (%) of an entity appearing at Position 1, 2, or 3 in the SharedRef dataset, across different models and prompts (Normal vs. Simple) in English.

Prompt / Model	Pos. 1	Pos. 2
Prompt: Normal		
4o	48.00	52.00
4o-mini	48.35	51.65
v3	43.36	56.64
3-32B	48.15	51.85
3.1-8B	48.57	51.43
DPO Llama (Ours)	50.90	49.10
Prompt: Simple		
4o	49.48	50.52
4o-mini	48.65	51.35
v3	48.04	51.96
3-32B	46.24	53.76
3.1-8B	48.51	51.49
DPO Llama (Ours)	50.00	50.00

Table 3: Average selection rate (%) of an entity appearing at Position 1, or 2 in the ClearRef dataset, across different models and prompts (Normal vs. Simple) in English.

ensure consistency across languages.

To assess the effect of this choice, we ran an ablation on the English dataset using all permutations. We observed that the frequency with which a model selects an entity depends heavily on its position in the list, indicating a strong positional bias.

Table 2 shows the distribution of selected entities across positions for each permutation in SharedRef. For example, in the Simple setting, entities at position three are selected drastically more often (avg. 42.62%) than those at position two (avg. 22.53%).

Table 3 presents analogous results for ClearRef. Here, the bias is milder, with position two being selected slightly more frequently on average (+4.22% in Normal, +3.03% in Simple).

Figure 10 shows the averaged correctness and category distribution over all permutations in English SharedRef. Compared to the fixed '0, 1, 2' ordering used in our main results, average correctness drops. Notably, GPT-4o exhibits fewer clarification attempts when averaged across permutations, while Qwen3-32B maintains strong performance.

The overall trend of higher correctness and better category distribution in the Normal setting compared to the Simple setting remains.

E Response Categorization

We adopt the response taxonomy proposed by Herlihy et al. (2024), with slight modifications to better suit our coreference resolution evaluation. Specifically, we merge *Interrogation* into *Clarification*, as both involve follow-up questions rather than direct answers. We also merge *Discussion* into *Answer Attempt*, since our evaluation does not require a fully factual answer, only that the response correctly identifies the positive entities. This simplification reduces annotation complexity without compromising the core objective of our analysis. A detailed overview of all response categories, including definitions and examples, is provided in Table 4.

F Automatic Evaluation

We used GPT-4.1-mini as an LLM judge to automatically evaluate the responses. We divided the evaluation into two parts: response classification and entity extraction. The prompt used for response classification is shown in Box 1. The prompts used for entity extraction, split into a system prompt and a user prompt, are shown in Box 2 and Box 3, respectively.

We manually annotated 500 responses from the English dataset to validate the framework. Table 5 reports the agreement rates for response categorization along with Cohen's Kappa scores. For entity extraction, we report exact match accuracy. Overall, the results show high agreement across all models.

Name	Description	Example
Answer attempt	The assistant clearly commits to at least one interpretation, providing a factual response without any hedging.	The helicopter can fly because its rotors generate lift, allowing it to rise off the ground and maneuver through the air.
Clarification	The response asks for more information about the user's intent without offering interpretations or using hedging	Could you specify which one you're referring to: the helicopter or the gnat?
Hedging	The response uses hedging or conditional language (e.g., "if you meant X...") and does not fully commit to a single interpretation. Even if only one entity is mentioned, the presence of such language marks it as a hedge.	If you're referring to a raven, it can fly due to its strong wings, lightweight body, and aerodynamic shape, which allow it to generate lift and move through the air efficiently.
Refuse	The response refuses to answer the question and does not ask any follow-up questions.	Here are the sentences: 1. **Bat**: The bat fluttered silently through the night sky, searching for insects. 2. **Dragonfly**: The dragonfly darted over the pond, its iridescent wings shimmering in the sunlight. 3. **Coffee**: The aroma of freshly brewed coffee filled the kitchen, awakening everyone's senses. 4. **Why can it fly?**: Why can it fly, despite its small wings and heavy body?"
Missing	The response is empty.	[blank]

Table 4: Description and Examples of our Response Categories.

Prompt	Response Cat.	Entity
4o-mini	100.0% (N/A)	99%
4o	100.0% (1.000)	98%
3-32B	92.0% (0.804)	98%
v3	98.0% (0.823)	94%
3.1-8B	100.0% (N/A)	100%
Total	98.0% (0.916)	97.8%

Table 5: Accuracy percentages and Cohen's Kappa scores (in parentheses) for Response Categorization and exact match accuracy for Entity Extraction across our evaluated models.

G Direct Preference Optimization

Our training set contains 472 responses from simple settings and 866 from normal settings. In addition, we included 30 basic clarification cases, where the user posed clearly ambiguous questions. A fine-grained distribution is provided in Table 6.

We fine-tuned the model for two epochs using Low-Rank Adaptation (LoRA). The full configuration for LoRA and DPO training is summarized in Table 7.

We observed performance improvements on both the SharedRef dataset and the homonym task from Ellinger et al. (2025). However, on the ClearRef test set, while the number of correct responses remained comparable to the base model, we experienced a category shift. As shown in Figure 11, the distribution of coarse response categories shifted significantly toward 'clarification' and 'hedge' across all languages. This indicates that the cognitive cost of those responses is higher for our DPO model compared to the base model on this dataset. To address this, future alignment efforts should incorporate more training examples from ClearRef to encourage direct answers where appropriate. Unlike in SharedRef, where the model successfully used common knowledge to respond only to the positive entities, in ClearRef, the model no longer consistently applies this strategy.

Dataset / Category	En	Fr	Ar	Ru	Zh
SharedRef					
Normal Answer Attempt	64	80	69	37	53
Normal Hedge	106	39	49	78	57
Normal Clarification	58	44	47	55	47
Simple Answer Attempt	112	84	30	69	76
Simple Hedge	21	13	2	15	31
Simple Clarification	4	3	1	4	1
ClearRef					
Normal Answer Attempt			2		
Normal Hedge			1		
Simple Answer Attempt			6		
General					
Clarification	6	6	6	6	6

Table 6: Distribution of chosen response types in our DPO fine-tuning dataset, broken down by language, response category, and setting.

Parameter	Value
LoRA Configuration	
r	64
LoRA Alpha	16
LoRA Dropout	0.05
Target Modules	[q_proj, v_proj, k_proj, o_proj]
Bias	none
DPO Training Configuration	
β	0.1
Learning Rate	5e-5
Batch Size (per device)	4
Epochs	2

Table 7: Combined configuration used for LoRA adaptation and Direct Preference Optimization (DPO) fine-tuning.

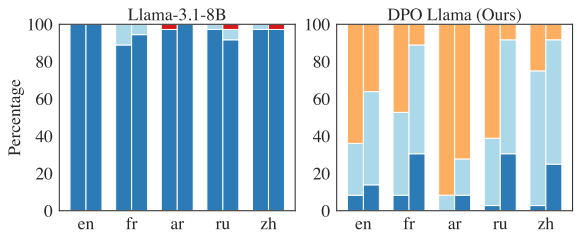

Figure 11: Distribution of response categories in the DPO test set across five languages in the ClearRef dataset. Colored squares denote response types: ■ Answer Attempt, ■ Hedge, ■ Clarification, and ■ Refuse.

Short Form	Name	Version	Access Provider
4o-mini	GPT-4o-mini	gpt-4o-mini-2024-07-18	OpenAI API
4o	GPT-4o	gpt-4o-2024-08-06	OpenAI API
4.1-nano	GPT-4.1-nano	gpt-4.1-nano-2025-04-14	OpenAI API
4.1-mini	GPT-4.1-mini	gpt-4.1-mini-2025-04-14	OpenAI API
3-32B	Qwen3-32B	N/A	OpenRouter
v3	Deepseek v3	N/A	Fireworks AI
3.1-8B	Llama-3.1-8B	N/A	Fireworks AI

Table 8: Specific model versions used in our experiments. For each model we provide the short form as used in our tables, the exact version and the access provider.

Prompt / Model	**Correct**					**Direct**				
	En	Fr	Ar	Ru	Zh	En	Fr	Ar	Ru	Zh
Prompt: Simple										
4o	98.08	**100.00**	98.08	**100.00**	96.15	88.46	69.23	**78.85**	63.46	75.00
4o-mini	**100.00**	**100.00**	**100.00**	**100.00**	**100.00**	69.23	55.77	53.85	28.85	76.92
v3	**100.00**	90.38	**100.00**	94.23	**100.00**	88.46	67.31	61.54	69.23	69.23
3-32B	**100.00**	94.23	94.23	96.15	96.15	73.08	63.46	21.15	65.38	67.31
3.1-8B	94.23	94.23	98.08	90.38	92.31	**92.31**	**78.85**	55.77	**73.08**	**80.77**
Prompt: Normal										
4o	96.15	96.15	**98.08**	**98.08**	96.15	86.54	73.08	**75.00**	**80.77**	80.77
4o-mini	**100.00**	**98.08**	94.23	**98.08**	94.23	82.69	65.38	57.69	46.15	67.31
v3	**100.00**	**98.08**	96.15	94.23	96.15	69.23	53.85	59.62	61.54	50.00
3-32B	98.08	94.23	**98.08**	90.38	94.23	82.69	51.92	51.92	67.31	63.46
3.1-8B	96.15	90.38	90.38	94.23	**98.08**	**94.23**	**78.85**	65.38	73.08	**90.38**

Table 9: Evaluation results showing the percentage of correct and direct responses across languages and prompt types on the ClearRef dataset. **Bold** highlights the highest scores per language within each prompt and metric.

Prompt / Model	**Correct**					**Direct**				
	En	Fr	Ar	Ru	Zh	En	Fr	Ar	Ru	Zh
Prompt: Simple										
4o	18.06	7.49	19.82	4.41	23.79	12.78	6.61	**12.78**	2.20	15.42
4o-mini	5.29	1.76	11.45	3.52	11.89	3.52	1.76	8.81	1.76	8.37
v3	**48.02**	**28.19**	**47.58**	**19.38**	46.70	10.57	**11.89**	6.17	3.96	7.49
3-32B	33.04	25.99	28.19	12.33	**56.83**	**23.35**	**11.89**	3.96	**7.93**	**26.87**
3.1-8B	11.89	3.96	**47.58**	4.85	4.85	4.85	0.44	2.20	1.76	1.76
Prompt: Normal										
4o	**81.06**	50.66	47.14	57.27	62.11	**36.56**	**34.80**	**30.40**	22.03	21.15
4o-mini	14.10	11.89	14.10	4.41	18.50	7.05	7.49	8.37	1.76	7.05
v3	61.23	61.23	**65.64**	31.28	49.78	3.96	9.25	10.13	3.52	1.76
3-32B	75.77	**71.37**	56.83	**63.88**	**83.26**	29.52	19.38	10.57	**25.99**	**25.11**
3.1-8B	14.54	13.66	7.49	3.96	6.17	5.73	2.20	3.96	1.32	2.64

Table 10: Evaluation results showing the percentage of correct and direct responses across languages and prompt types on the SharedRef dataset. **Bold** highlights the highest scores per language within each prompt and metric.

Box 1: User Prompt: Response Classification

```
You are reviewing a single-turn conversation between a user and an assistant.

You must classify the assistant's response according to the **type of response strategy** used.

Here are the definitions:

- `answer_attempt`: The assistant **unambiguously commits** to at least one interpretation of the
user's question and provides a factual explanation or general background relevant to the query. It
must **not use any hedging language**, such as "if you mean. . . ", "depending on. . . ", or "could
be. . . ". Implicit or explicit hedging disqualifies it from this category. It may mention multiple
entities, as long as none are introduced with hedging or conditional phrasing.

- `clarification`: The assistant **asks for more information** about the user's intent. This may
involve **one or more clarification questions**, but must not **suggest answers** or list
**interpretations**,. The response must should **not contain hedging or speculative language**.

- `hedge`: The assistant **uses hedging or conditional language** (e.g., "if you meant X. . . ",
"could refer to. . . ", "might be. . . ") and does not fully commit to a single interpretation. Even if
only one entity is mentioned, the presence of such language marks it as a hedge.

- `refuse`: The assistant explicitly or implicitly refuses to answer the question and does **not**
ask any follow-up questions.

- `missing`: The assistant gives **no response at all**.

You must output your answer in the following JSON format:
{"response_type": "refuse|missing|answer_attempt|hedge|clarification"}

Conversation:
Question: <question_str>
Answer: <asnwer_str>
```

Box 2: System Prompt: Entity Extraction

```
You are an expert in identifying explicit entity mentions in text.

Your task is to determine which of the listed entities are explicitly mentioned in the response.

Instructions:
- Return only entities from the provided list (use their exact spelling from the list in your
output).
- Consider an entity 'mentioned' if the response text contains the entity exactly, or with minor
morphological or spelling variations.
Examples of acceptable variations include plural forms and misspellings.
- Do not infer mentions from implied meaning, associations, or unrelated paraphrasing.
- Only include entities that are clearly and explicitly present in the response text.
- **Always** output the entity names exactly as they appear in the provided list.
```

Box 3: User Prompt: Entity Extractionn

```
Entities: <entity_list>
Response:

<answer_str>

Which of the listed entities are explicitly mentioned in the response?
```

Arabic Prompts

Rel. 1: الطيران؟ يمكنها لماذا
Rel. 2: حلو؟ هو لماذا
Rel. 3: الخشب؟ من مصنوع هو لماذا
Rel. 4: السباحة؟ يمكنها لماذا
Rel. 5: بسرعة؟ يعمل أن يمكن لماذا
Rel. 6: الأشجار؟ تسلق يمكنها لماذا
Rel. 7: حار؟ الجو لماذا
Rel. 8: مرتفع؟ الصوت "لماذا
Simple: بسيطة بلغة الإجابة .يرجى [..]
Starter: <entity_list> التالي من لكل واحدة جملة :أعطني

French Prompts

Rel. 1: Pourquoi peut-il voler ?
Rel. 2: Pourquoi est-ce sucré ?
Rel. 3: Pourquoi est-il en bois ?
Rel. 4: Pourquoi peut-il nager ?
Rel. 5: Pourquoi peut-il fonctionner rapidement ?
Rel. 6: Pourquoi peut-il grimper aux arbres ?
Rel. 7: Pourquoi fait-il chaud ?
Rel. 8: Pourquoi est-ce si bruyant ?
Simple: [..] Veuillez répondre dans un langage simple.
Starter: Donnez-moi une phrase pour chacun des mots suivants : <entity_list>

English Prompts

Rel. 1: Why can it fly?
Rel. 2: Why is it sweet?
Rel. 3: Why is it made of wood?
Rel. 4: Why can it swim?
Rel. 5: Why can it run fast?
Rel. 6: Why can it climb trees?
Rel. 7: Why is it hot?
Rel. 8: Why is it loud?
Simple: [..] Respond in simple language.
Starter: Provide me one sentence for each of the following: <entity_list>

Chinese (Zh) Prompts

Rel. 1: 它为什么能飞？
Rel. 2: 为什么是甜的？
Rel. 3: 为什么它是木制的？
Rel. 4: 为什么它能游泳？
Rel. 5: 为什么它能跑得这么快？
Rel. 6: 为什么它能爬树？
Rel. 7: 为什么会热？
Rel. 8: 为什么这么吵？
Simple: [..] 请用通俗易懂的语言回答。
Starter: 请为以下每个项目提供一句描述： <entity_list>

Russian Prompts

Rel. 1: Почему он может летать?
Rel. 2: Почему он сладкий?
Rel. 3: Почему он сделан из дерева?
Rel. 4: Почему он может плавать?
Rel. 5: Почему он может быстро бегать?
Rel. 6: Почему он может лазать по деревьям?
Rel. 7: Почему так жарко?
Rel. 8: Почему так громко?
Simple: [..] Пожалуйста, отвечайте простым языком.
Starter: Дайте мне по одному предложению для каждого из следующих слов: <entity_list>

Figure 12: Language Versions of Relation Questions, the Simple Instruction and the Starter Sentence in Arabic, French, English, Chinese, and Russian