Simplifications are Absolutists: How Simplified Language Reduces Word Sense Awareness in LLM-Generated Definitions

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

Abstract

Large Language Models (LLMs) can provide accurate word definitions and explanations for any context. However, the scope of the definition changes for different target groups, like children or language learners. This is especially relevant for homonyms—words with multiple meanings—where oversimplification might risk information loss by omitting key senses, potentially misleading users who trust LLM outputs. We investigate how simplification impacts homonym definition quality across three target groups: Normal, Simple, and ELI5. Using two novel evaluation datasets spanning multiple languages, we test DeepSeek v3, Llama 4 Maverick, Qwen3-30B A3B, GPT-4o mini, and Llama 3.1 8B via LLM-as-Judge and human annotations. Our results show that simplification drastically degrades definition completeness by neglecting polysemy, increasing the risk of misunderstanding. Fine-tuning Llama 3.1 8B with Direct Preference Optimization substantially improves homonym response quality across all prompt types. These findings highlight the need to balance simplicity and completeness in educational NLP to ensure reliable, context-aware definitions for all learners.

1 Introduction

Large Language Models (LLMs) are increasingly being deployed across diverse domains, with education emerging as an up-and-coming area. These models have the potential to support personalized learning, provide immediate feedback, and increase accessibility for a wide range of learners (Xu et al., 2024; Wang et al., 2024; Kasneci et al., 2023).

A core requirement for effective education is personalization. Students vary in their prior knowledge, learning preferences, and cognitive abilities (Xu et al., 2024). From young children and university students to individuals with cognitive impairments, learners benefit from tailored content that matches their comprehension level.

Figure 1: GPT-4o-mini definitions of "bat" under normal and simplified style constraints. The normal definition presents multiple senses, showing the word's ambiguity. The simplified definition provides only one sense, indicating reduced word sense awareness.

However, LLMs can produce false or misleading information (Xu et al., 2024). Given their often authoritative tone and rapid accessibility, users, especially those with lower domain knowledge, may uncritically accept incorrect outputs as fact (Kasneci et al., 2023). This is particularly concerning in educational contexts, where reliance on incorrect information can hinder learning and discourage critical thinking.

To improve accessibility, for groups like non-native speakers or individuals with cognitive impairments, educational content is often adapted into simplified, easy-read language (Freyer et al., 2024). While this makes information more accessible, it introduces a trade-off: essential information may be lost or oversimplified (Trienes et al., 2024).

One common use case for LLMs is providing

Proceedings of Recent Advances in Natural Language Processing,pages 342–351
Varna, Sep 8–10, 2025
https://doi.org/10.26615/978-954-452-098-4-042

instant definitions or explanations across languages. Users often use prompts like "Explain like I'm five" (ELI5) to obtain simplified outputs. However, for polysemous words, like homonyms[1], simplification can obscure ambiguity. For instance, Figure 1 shows a user requesting definitions for "bat" in standard and simplified styles from GPT-4o mini. The standard response lists multiple meanings (e.g., animal, sports tool), while the simplified response only mentions the animal. This simplification may mislead users into assuming a single meaning.

This risk is amplified by users' tendency to assume that LLM responses are both correct and complete. We argue that high-quality definitions should strive for **completeness**: either (i) enumerate all plausible senses of a homonym, or (ii) exhibit **Helpful Sense Awareness**. This means clearly stating that not all possible senses are listed, or that additional context is needed for disambiguation. Responses that provide only some senses without such awareness may reinforce misconceptions about the word's meaning.

In this paper, we investigate how well state-of-the-art LLMs handle homonym definition tasks across varying levels of language complexity: Normal, Simple, and ELI5. We assess whether simplification impairs the model's ability to acknowledge ambiguity and provide complete information. Our contributions are as follows:

- We propose **Helpful Sense Awareness**, a novel metric for assessing whether LLMs appropriately acknowledge multiple senses of homonyms during definition tasks.
- We present two datasets for evaluating LLM performance on homonym definitions in both multilingual and English settings.
- We fine-tune Llama 3.1 8B using Direct Preference Optimization (DPO), greatly improving response quality on homonym definitions.
- We empirically demonstrate, using both LLM-as-a-Judge and human annotations, that stylistic constraints aimed at simplification drastically degrade homonym definition quality in models such as DeepSeek-v3, Llama 4 Maverick, Qwen3-30B A3B, GPT-4o mini, and Llama 3.1 8B.

[1]We use the term homonym broadly to include both polysemous words and homonyms in the strict linguistic sense (i.e., words with multiple related or unrelated meanings)

2 Background and Related Work

Definition Modeling Noraset et al. (2017) introduced the task of Definition Modeling. The goal is to generate a definition for a given word based on its embedding. Definitions fall into two categories: static and context-dependent. Static definitions are found in dictionaries and lexicons. They provide fixed meanings for words based on predefined word senses, treating each word as having a discrete set of definitions. A related task is Word Sense Disambiguation (WSD). This focuses on identifying the correct sense of a word in a given context by assigning a word sense label (Navigli, 2009). However, Kilgarriff (1997) challenges the idea that word senses are fixed, arguing that meaning emerges from usage and context. This underscores the limitations underlying traditional lexicon-based approaches.

Recent research has predominantly focused on generating context-sensitive definitions, leveraging contextual information to produce precise and relevant meanings (Mickus et al., 2024; Periti et al., 2024; Huang et al., 2021; Bevilacqua et al., 2020; Ishiwatari et al., 2019; Gadetsky et al., 2018). Kong et al. (2022) focus on generating context-sensitive definition generation in simple language.

Proietti et al. (2024) focus on sense selection rather than definition generation. They group WordNet senses based on homonymy relations, mapping them to coarse-grained sense clusters. They then probe contextual representations from pretrained language models (e.g., BERT, DeBERTa) using distance-based metrics, showing that they can distinguish homonymous senses with 95% accuracy.

In contrast to the focus on context-sensitive methods, users often query "What is [word]?" without additional context. Building on the insight that PLMs can handle lexical ambiguity inherently, we evaluate how LLMs define homonyms in such cases. This allows us to assess their understanding of lexical word senses, their management of ambiguity, and the usefulness of their responses absent contextual cues.

Simple Language Simplified language enhances accessibility for diverse audiences, including non-native speakers, domain novices, children, and individuals with cognitive impairments. Standards like the Web Content Accessibility Guidelines (WCAG) advocate for its use to foster inclusive communication (W3C, 2025). This approach employs straightforward vocabulary, minimal jargon, clear sentence structures, and avoids complex grammar (Freyer et al., 2024). It is widely applied in fields such as law, healthcare, and education (Garimella et al., 2022; Deilen et al., 2024; Rets et al., 2022). Prior studies highlight that simplification in LLM-generated text can introduce omissions or vagueness (Trienes et al., 2024; Agrawal and Carpuat, 2024; Devaraj et al., 2022). A related approach, Explain Like I'm Five (ELI5), popularized through a dataset of 270,000 Reddit threads (Fan et al., 2019), simplifies complex topics using analogies and non-technical language. While ELI5 is effective for broad audiences, its impact on preserving content completeness is underexplored. This paper examines how simplification constraints, including ELI5-style rewriting, affect the completeness of definitions for homonyms, where precise disambiguation is critical.

3 Methodology

3.1 Datasets

Multilingual Word-in-Context (ML-WiC) We processed the dataset from Martelli et al. (2021), designed for Multilingual and Cross-lingual Word-in-Context Disambiguation (MCL-WiC). This task involves determining whether a target word retains the same sense in two sentences within and across languages. We filtered the dataset for words with distinct senses in the same language, identifying them thus as homonyms. The filtered dataset covers 308 Arabic, 334 English, 380 French, 330 Russian, and 254 Simplified Chinese target words.

Homonyms with WordNet (HoWN) ML-WiC lacks sense annotations, limiting its use for evaluating which senses LLMs capture. We therefore built HoWN, annotated with WordNet synsets (Miller, 1994). To ensure true homonyms with distinct senses, we used coarse-grained sense clusters from Proietti et al. (2024), avoiding overly fine-grained nuances. Using the SemCor Corpus (Miller et al., 1993), we kept only those that appeared at least twice with different sense annotations.

3.2 Model and Prompt Configuration

We evaluate five LLMs to assess how style constraints impact the quality of word definition generation: DeepSeek v3 (DeepSeek-AI et al., 2025), Llama 4 Maverick (Meta AI, 2025), Qwen3-30B A3B (Qwen Team, 2025), GPT-4o mini (OpenAI et al., 2024), and Llama 3.1 8B (Grattafiori et al., 2024). These models vary in size, architecture (including Mixture-of-Experts and dense models), and openness, enabling a comprehensive analysis of performance across diverse LLMs.

We use four prompt types to generate definitions:

- **Normal**: "What is the definition of '*word*'?"
- **Simple**: "What is the definition of '*word*' in simple language?"
- **ELI5**: "Explain '*word*' like I am 5 years old."
- **Multi-Sense-Aware**: Any prompt appended with: "Keep in mind that some words have more than one meaning."

ELI5 targets explanations for children, Simple seeks definitions in plain language, and Normal allows unconstrained responses. The Multi-Sense-Aware prompt tests whether explicit instructions to account for multiple meanings mitigate stylistic constraints on definition completeness.

3.3 Response Categorization

Ambiguity in language, particularly for homonyms, poses a challenge in natural language processing, as a single word can have multiple meanings depending on context. To evaluate model responses for handling ambiguity, we developed a categorization framework based on three criteria:

1. **Number of definitions**: The count of distinct meanings provided for a word.
2. **Context clarification request**: Whether the response seeks the intended context (e.g., "Please let me know if you have a specific context in mind!").
3. **Remark on additional definitions**: Whether the response acknowledges other unlisted meanings (e.g., "Here are some of the primary definitions of the word:").

We define a response as **complete** if it either provides *all* meanings of a word or exhibits **Helpful Sense Awareness (HeSA)**. A response exhibits HeSA by including a *context clarification request* or a *remark on additional definitions*. Such responses are considered high-quality, as they proactively mitigate ambiguity without requiring an exhaustive list of meanings.

A relaxed concept of **Completeness** is **Sense Awareness**, which applies when a response includes multiple definitions or demonstrates HeSA. We use Sense Awareness when no dictionary data is available to verify whether extracted definitions match all possible meanings. This metric is a valuable simplification of Completeness, as its opposite, providing only one definition without HeSA, is the least desirable outcome for homonyms.

These classifications are crucial because our prompts intentionally lack context. This increases the risk of incomplete responses and misleading users by suggesting an incorrect word meaning.

3.4 Automatic Evaluation

We designed an automated evaluation framework to categorize model responses using GPT-4o mini as an LLM judge, selected for its efficiency and performance. The framework evaluates responses based on three dimensions outlined in Section 3.3 and extracts all explicitly mentioned definitions. A few-shot prompt guides the evaluation. To validate the framework, one author labeled 450 responses from HoWN, with 150 responses per prompt type. The LLM judge achieved 93.33% agreement on Definition Count Classification (Single, Multiple) and 86.44% agreement on HeSA compared to these human labels.

3.5 Definition Matching for HoWN

To analyze sense coverage and preference in HoWN, we mapped the extracted definitions to WordNet senses. We employed a sentence transformer model[2](Reimers and Gurevych, 2019) to compute the cosine similarity between model-generated definitions and the glosses of corresponding WordNet senses. For each generated definition, we selected the sense with the highest similarity score, considering only matches with a minimum similarity of 0.4. Since WordNet senses are ranked by estimated frequency of use (Miller, 1994), this mapping allowed us to assess which senses the model most frequently aligns with and whether it covers the full range of senses.

3.6 Direct Preference Optimization

To improve the completeness of homonym definitions, we fine-tuned the Llama 3.1 8B using DPO (Rafailov et al., 2023). DPO aligns model outputs with desired behavior by training on preference pairs. In our case, we favor complete responses over incomplete ones. We constructed a training set from the HoWN dataset with simple prompts, comparing Llama 3.1 8B's responses to more complete responses from other models to create 116 preference pairs across 63 words.

We chose HoWN as it evaluates whether a response captures all possible definitions of a word, ensuring completeness is not solely driven by HeSA. Due to limited data, we did not create a validation set and trained once on the full data, aiming to show feasibility of aligning models for more complete definitions rather than optimizing peak performance.

4 Results

In this section, we analyze the HoWN dataset, focusing on completeness, sense coverage, and sense distribution. Additionally, we present the results from our DPO fine-tuning, followed by the results of the ML-WIC dataset.

4.1 HoWN

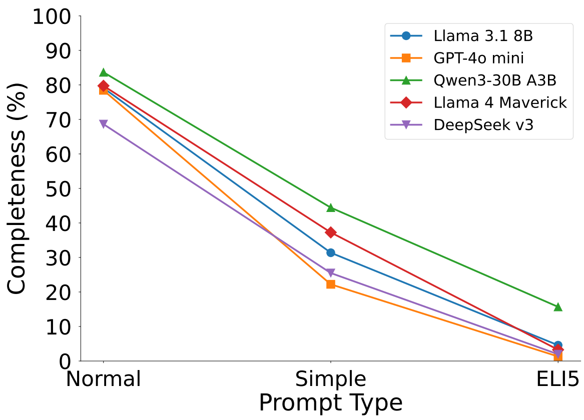


Figure 2: Definition completeness for each model under three stylistic constraints: Normal, Simple, and ELI5

To assess the impact of stylistic constraints on model performance, we evaluated response quality using the HoWN dataset across the three prompt types. Figure 2 illustrates the variation in response completeness across these prompt types, highlighting the drastic influence of stylistic constraints.

For the Normal prompt, all models achieve relatively high completeness, ranging from 68.63% for DeepSeek v3 to 83.66% for Qwen3-30B A3B. Completeness falls sharply under Simple and ELI5. In the Simple setting, completeness ranges from 22.22% for GPT-4o mini to 37.25% for Llama 4 Maverick, while in the ELI5 setting, it ranges from

[2]sentence-transformers/all-MiniLM-L6-v2

Model	FKGL	Sense Aware	Multi. Def.	HeSA	Full	Both	Complete	Covered
Prompt: Normal								
Llama 3.1 8B	10.51	**97.39**	**97.39**	67.97	37.25	26.14	79.08	65.93
GPT-4o mini	10.70	95.42	95.42	62.09	49.02	32.68	78.43	74.22
Qwen3-30B A3B	8.87	93.46	90.20	**77.78**	44.44	**38.56**	**83.66**	70.20
Llama 4 Maverick	10.76	94.77	94.12	60.13	45.10	25.49	79.74	71.70
DeepSeek v3	10.08	92.81	92.16	22.88	**53.59**	7.84	68.63	**77.04**
Prompt: Simple								
Llama 3.1 8B	8.35	71.24	70.59	18.95	17.65	5.23	31.37	51.89
GPT-4o mini	8.00	61.44	61.44	7.84	16.99	2.61	22.22	52.88
Qwen3-30B A3B	7.20	**83.01**	**82.35**	22.88	**30.72**	**9.15**	**44.44**	**60.79**
Llama 4 Maverick	8.52	73.20	71.90	**26.14**	19.61	8.50	37.25	55.29
DeepSeek v3	8.77	61.44	61.44	5.23	20.26	0.00	25.49	54.98
Prompt: ELI5								
Llama 3.1 8B	4.26	13.73	11.11	3.92	1.31	0.65	4.58	33.41
GPT-4o mini	5.27	8.50	8.50	0.00	1.31	0.00	1.31	31.79
Qwen3-30B A3B	5.11	**35.95**	**33.33**	**9.80**	**8.50**	**2.61**	**15.69**	**40.08**
Llama 4 Maverick	4.03	11.11	9.80	2.61	0.65	0.00	3.27	34.99
DeepSeek v3	5.50	11.11	10.46	1.31	0.65	0.00	1.96	31.67

Table 1: Performance metrics for models on the HoWN dataset across Normal, Simple, and ELI5 prompt types. Metrics include Flesch–Kincaid grade level (FKGL), percentage of responses classified as *Sense Aware*, *Multiple Definitions*, *HeSA*, *Full* (covering all meanings), *Both* (HeSA and Full), *Complete*, and the average percentage of coarse-grained definitions covered. Best scores for each prompt type are highlighted in **bold**.

1.31% for GPT-4o mini to 15.69% for Qwen3-30B A3B. Among the models, Qwen3-30B A3B exhibits the smallest drop in completeness across prompt types, while GPT-4o mini shows the largest.

We detail these findings in Table 1. These reveal that simpler prompt types (Simple and ELI5) lead to fewer covered definitions and lower HeSA scores. This suggests that stylistic simplification not only reduces the number of definitions generated but also impairs models' ability to acknowledge multiple word senses. We further report Sense Awareness and the percentage of responses containing multiple definitions. Both metrics show a marked decline from Normal to Simple to ELI5 prompts, reinforcing the trend. Additionally, we calculate the Flesch-Kincaid grade level (FKGL) (Kincaid et al., 1975) to measure the readability of the generated definition. ELI5 responses have the lowest FKGL (4.03–5.27), followed by Simple (7.20–9.37), and Normal (8.87–10.76). A lower value indicates simpler language, confirming that the models adhere to the simplicity constraints.

4.1.1 Sense Coverage and Distribution Analysis

As discussed in Section 3.3, incomplete responses are less desirable due to their potential to mislead users. To evaluate their impact, we analyze the coverage of coarse-grained WordNet senses, as defined by Proietti et al. (2024), for these responses. Figure 3 shows that all prompt styles exhibit a high density of sense coverage around 50%. For the Normal prompt, there is an additional pronounced peak at 100% coverage, indicating that many responses capture the full range of senses. The Simple prompt also displays a peak at 100% coverage, though it is less prominent, with greater density in lower coverage ranges compared to the Normal prompt. In contrast, ELI5 shows less density at higher coverage levels and a greater concentration in lower ranges, suggesting that ELI5-style responses tend to cover fewer senses comprehensively.

Additionally, we analyzed which definitions are most likely to appear in model responses when not all coarse synsets are covered. As expected, more prominent senses tend to be included more frequently than less common ones across all models and prompt types.

4.1.2 Multi-Sense-Aware Analysis

We evaluate the impact of prompt types on model performance in a multi-sense-aware setting. We focus on completeness and average synset coverage. Results are presented in Table 2.

Context-aware prompting consistently improves completeness across all prompt types, with larger gains for Simple and ELI5 compared to Normal. For instance, Llama 4 Maverick shows substantial improvements: +50.68 in ELI5, +39.19 in Simple, and +15.54 in Normal. Synset coverage likewise

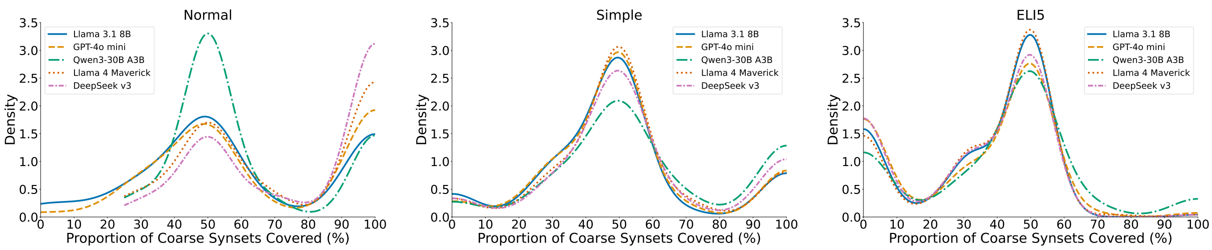


Figure 3: Distribution of coarse sense coverage across model outputs. The x-axis shows synsets coverage, while the y-axis shows estimated density. Only incomplete responses were considered.

Prompt / Model	Complete	Covered
Prompt: Normal		
Llama 3.1 8B	83.78 +4.05	73.97 *+8.18*
GPT-4o mini	63.51 −16.89	78.34 +3.64
Qwen3-30B A3B	91.22 +5.41	78.47 +7.93
Llama 4 Maverick	**95.95** *+15.54*	77.72 +5.28
DeepSeek v3	76.35 +6.76	**84.72** +7.45
Prompt: Simple		
Llama 3.1 8B	41.89 +10.14	60.55 +8.93
GPT-4o mini	57.43 +35.14	66.71 +13.73
Qwen3-30B A3B	66.22 +21.62	72.69 +11.87
Llama 4 Maverick	**77.70** *+39.19*	73.30 +17.83
DeepSeek v3	56.08 +30.41	**73.49** *+18.68*
Prompt: ELI5		
Llama 3.1 8B	39.86 +35.14	54.40 +21.21
GPT-4o mini	30.41 +29.05	52.98 +21.13
Qwen3-30B A3B	**59.46** +43.24	62.52 +22.43
Llama 4 Maverick	54.05 *+50.68*	61.44 +26.96
DeepSeek v3	50.00 +47.97	**62.96** *+31.24*

Table 2: Performance metrics for models on the HoWN dataset under the Multi-Sense-Aware setting across Normal, Simple, and ELI5 prompt types. Metrics show the percentage of *Complete* responses and average percentage of coarse-grained definitions covered, with best scores for each prompt type in **bold**. Deltas to the non-Multi-Sense-Aware setting are in percentage points, with the largest delta for each prompt type in *italic*.

Metric	Normal	Simple	ELI5
FKGL	10.59 -0.34	9.71 +1.04	5.07 +0.61
Sense Aware	98.98 +3.06	86.73 +10.20	35.71 +21.43
Multi. Def.	98.98 +3.06	86.73 +11.22	32.65 +21.43
HeSA	84.69 +18.37	67.35 +36.73	13.27 +9.18
Full	44.90 +10.20	29.59 +2.04	10.20 +8.16
Both	38.78 +17.35	21.43 +13.27	4.08 +3.06
Complete	90.82 +11.22	75.51 +25.51	19.39 +14.29
Covered	70.97 +6.32	61.29 +3.28	39.97 +6.79

Table 3: Performance metrics for DPO-optimized Llama 3.1 8B on unseen words across Normal, Simple, and ELI5 prompts. Metrics include Flesch–Kincaid grade level, Sense Awareness, Multiple Definitions, HeSA, Full responses, Both (Full and HeSA), Completeness, and percentage of coarse-grained definitions covered. Changes represent percentage point differences compared to the non-DPO model.

increases across all models and prompt types, with stronger gains in Simple and ELI5 settings.

Interestingly, despite these improvements, completeness still declines from Normal to Simple to ELI5. However, the multi-sense-aware setting substantially mitigates this drop compared to the non-multi-sense-aware baseline. Notably, GPT-4o-mini exhibits a unique decline in completeness (-16.89) under the Normal prompt, indicating model-specific challenges in this setting.

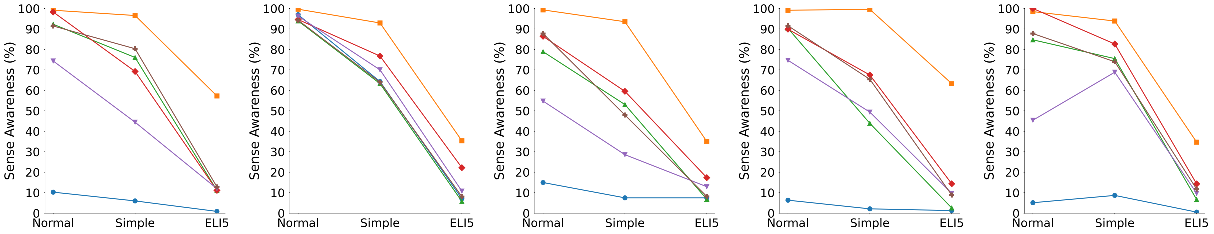
4.2 Direct Preference Optimization

We assess the impact of DPO fine-tuning on the Llama 3.1 8B model, evaluating its ability to generalize to unseen words. Table 3 reports performance metrics on the HoWN dataset, focusing on an evaluation subset of 98 out of 164 words (59.76%) that were not included in the DPO training set. This subset was obtained by excluding three words from the 101 words unseen during training, for which we could not obtain a valid judgment or response from Qwen3-30B A3B, the original Llama 3.1 8B, or its fine-tuned variant.

Despite fine-tuning only on responses to Simple prompts, DPO substantially boosts performance across all prompt types: completeness (+11.22 Normal, +25.51 Simple, +14.29 ELI5), HeSA (+18.37, +36.73, +9.18), and coverage (+6.32, +3.28, +6.79). This shows DPO improves handling of multiple word senses beyond training data.

The FKGL decreases by 0.34 for Normal, indicating simpler language, while increasing by 1.04 for Simple and 0.61 for ELI5. It yet remains comparable to other models in each setting. These results demonstrate that DPO fine-tuning enhances completeness, generalizing effectively to unseen words across diverse prompt types.

On the test subset, our model surpasses Qwen3-30B A3B, the best-performing base model across all prompt types, with completeness improvements



Arabic English French Russian Chinese (Zh)

Figure 4: Sense Awareness across Languages: Llama 3.1 8B (blue line, circle), DPO Llama 3.1 8B (orange line, square), GPT-4o mini (green line, triangle), Qwen3-30B A3B (red line, diamond), Llama 4 Maverick (purple line, inverted triangle), DeepSeek v3 (brown line, plus).

of +11.23 (90.82 vs. 79.59) for Normal, +25.51 (75.51 vs. 50.00) for Simple, and +14.26 (19.36 vs. 5.10) for ELI5. Trained only on Simple prompts, DPO fine-tuning enables Llama 3.1 8B to outperform larger models, demonstrating its potential to enhance sense-aware definition generation for unseen homonyms in context-free queries.

4.3 ML-WIC

We evaluate prompt type sensitivity across languages using the ML-WiC dataset. Since we do not have dictionary data for all languages, we report Sense Awareness instead of Completeness. Figure 4 shows Sense Awareness scores for Arabic, English, French, Russian, and Chinese across the Normal, Simple, and ELI5 prompt types.

We observe a general trend across all languages and models: Sense Awareness declines from Normal to Simple to ELI5. English shows the lowest inter-model variance in Sense Awareness, while also strictly following this trend. Notably, Chinese stands out, as Llama 3.1 8B and Llama 4 Maverick achieve higher Sense Awareness with Simple prompts than with Normal ones, which is an exception to the overall pattern. Among all languages, Chinese also shows the smallest decline from Normal to Simple prompts.

Interestingly, Llama 3.1 8B shows poor Sense Awareness in non-English languages, even with Normal prompts. Our DPO model consistently outperforms all others across prompt types and languages, whereas Llama 3.1 8B, its base model, exhibits the lowest performance. However, except for Russian, our model responded in English despite being prompted in the respective language.

5 Discussion

Our findings reveal that simple language, particularly through Simple and ELI5 prompts, drastically reduces the completeness of homonym definitions. Yet, even Normal prompts yield suboptimal results. In the HoWN dataset, completeness for Normal prompts ranges from 68.63% (DeepSeek v3) to 83.66% (Qwen3-30B A3B), falling short of ideal performance (see Table 1). Under ELI5, completeness drops dramatically, falling as low as 1.31% (GPT-4o mini) to 15.69% (Qwen3-30B A3B). Similarly, definition coverage is lower in simpler settings, indicating a loss of nuanced meanings.

We observe a similar trend in Sense Awareness across all languages in the ML-WIC dataset. Further, inter-model variance differs across prompt types and languages in the multilingual setting. This variance likely reflects the influence of each language during training. Additionally, Llama 3.1 8B demonstrates notably low performance in non-English languages. Here, its smaller model size probably contributes to this limitation.

The tendency of LLMs to provide only one or a subset of definitions in simpler settings may first seem intuitive, as it reduces complexity. However, this reflects a misunderstanding of simplification. Simplification should enhance understandability without sacrificing information. While it may be reasonable to present fewer senses in simpler settings, responses should still incorporate HeSA to ensure completeness. These findings align with prior work by Trienes et al. (2024), highlighting simplification-induced information loss in question-answering tasks. Similarly, Kong et al. (2022) analyzed context-dependent definitions and found that simplified versions achieve lower BLEU scores and sentence similarity than complex ones, further supporting the trade-off between simplification and information retention.

To address these challenges, we employed DPO and Multi-Sense-Aware Prompting, which both showed promising improvements in definition quality. However, we did not evaluate these methods on single-definition words, where they might have unintended effects, such as overcomplicating responses. An alternative can be Steering Vectors, used by Rimsky et al. (2024). These can enhance model behavior at inference time without increasing prompt length (as in Multi-Sense-Aware Prompting) or requiring fine-tuning (as in DPO).

In the multilingual setting, the performance of our DPO model reveals intriguing patterns. The base model exhibits by far the weakest performance in non-English languages, whereas the DPO-tuned model consistently outperforms others across all scenarios. Notably, the DPO model frequently responds in English, even when prompted in other languages, which may reflect its training bias. As we did not verify the factual accuracy of responses, some outputs may include hallucinations. Nevertheless, the success of DPO fine-tuning remains evident: Despite training a relatively small model on a limited English dataset in the simple prompt setting, it achieved superior performance across all prompt types and languages. We thus argue that LLMs are capable of giving complete homonym definitions. However, they are limited by the expected model behavior.

6 Conclusion

In this paper, we analyze how LLMs behave in homonym definition generation without additional context, a setting that requires context-independent understanding. We have shown that LLMs have an overall reduced multi-sense awareness, especially for simplified outputs, indicating an oversimplification of contents.

These findings highlight the need for LLMs to better serve diverse users, particularly marginalized groups like non-native speakers and individuals with cognitive impairments. These groups encounter challenges due to linguistic exclusion, underscoring the importance of inclusive design (Freyer et al., 2024). Improving LLM inclusivity is essential for building more accessible and effective language technologies.

To support reproducibility and enable future research, we provide a repository[3] containing our code, scripts, data preprocessing routines, evaluation tools, additional results, and model outputs.

[3] https://github.com/lukasellinger/homonym-eval

7 Limitations

LLM Judge The evaluation scores of our LLM judge are sensitive to scoring prompt wording, potentially introducing variability. Inherent LLM biases may also cause systematic differences from human judgments (Rimsky et al., 2024). Additionally, prior work suggests LLMs underperform in simplified settings, which may affect automated evaluation reliability (Anschütz et al., 2024; Anonymous, 2025). Nonetheless, our human evaluation closely mirrors LLM judge scores. The differences observed between prompt types far exceed any potential error margin, strongly reinforcing our findings' robustness.

Selected Prompts Our study employed three predefined prompts to elicit definitions, reflecting choices a typical user might make without optimizing for prompt variation. However, LLMs are highly sensitive to prompt phrasing, which can impact response quality (Brown et al., 2020). While our approach mirrors realistic user behavior, it does not account for potential gains from prompt optimization. Future research could systematically investigate the effects of varied or optimized prompts on LLM performance.

Factuality of Definitions We did not explicitly verify the factual accuracy of the definitions generated by the models. As a result, a response may be structurally complete and fully adhere to the HeSA framework, yet still contain definitions that are factually incorrect or misleading.

Reliance on WordNet Our analysis of definition completeness and returned definitions relied solely on WordNet. Although WordNet is a widely adopted resource, this choice may not capture definitional nuances in other databases. Future studies could incorporate alternative resources, such as ConceptNet (Speer et al., 2017) or Wiktionary, to validate our findings across diverse lexical datasets.

Acknowledgments

This research has been funded by the German Federal Ministry of Research, Technology and Space (BMFTR) through grant 01IS23069 Software Campus 3.0 (Technical University of Munich) as part of the Software Campus project "LIANA".

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.

Anonymous. 2025. (Dis)improved?! How Simplified Language Affects Large Language Model Performance across Languages. In *Submitted to ACL Rolling Review - February 2025*.

Miriam Anschütz, Edoardo Mosca, and Georg Groh. 2024. Simpler Becomes Harder: Do LLMs Exhibit a Coherent Behavior on Simplified Corpora? In *Proceedings of the Workshop on DeTermIt! Evaluating Text Difficulty in a Multilingual Context @ LREC-COLING 2024*, pages 185–195, Torino, Italia. ELRA and ICCL.

Michele Bevilacqua, Marco Maru, and Roberto Navigli. 2020. Generationary or "How We Went beyond Word Sense Inventories and Learned to Gloss". In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7207–7221, Online. Association for Computational Linguistics.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, et al. 2020. Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

DeepSeek-AI, Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, et al. 2025. DeepSeek-V3 Technical Report. ArXiv:2412.19437 [cs].

Silvana Deilen, Ekaterina Lapshinova-Koltunski, Sergio Hernández Garrido, Christiane Maaß, Julian Hörner, et al. 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.

Angela Fan, Yacine Jernite, Ethan Perez, David Grangier, Jason Weston, et al. 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.

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.

Artyom Gadetsky, Ilya Yakubovskiy, and Dmitry Vetrov. 2018. Conditional Generators of Words Definitions. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 266–271, Melbourne, Australia. Association for Computational Linguistics.

Aparna Garimella, Abhilasha Sancheti, Vinay Aggarwal, Ananya Ganesh, Niyati Chhaya, et al. 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.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, et al. 2024. The Llama 3 Herd of Models. ArXiv:2407.21783 [cs].

Han Huang, Tomoyuki Kajiwara, and Yuki Arase. 2021. Definition Modelling for Appropriate Specificity. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2499–2509, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Shonosuke Ishiwatari, Hiroaki Hayashi, Naoki Yoshinaga, Graham Neubig, Shoetsu Sato, et al. 2019. Learning to Describe Unknown Phrases with Local and Global Contexts. 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 3467–3476, Minneapolis, Minnesota. Association for Computational Linguistics.

Enkelejda Kasneci, Kathrin Sessler, Stefan Küchemann, Maria Bannert, Daryna Dementieva, et al. 2023. ChatGPT for good? On opportunities and challenges of large language models for education. *Learning and Individual Differences*, 103:102274.

Adam Kilgarriff. 1997. "I Don't Believe in Word Senses". *Computers and the Humanities*, 31(2):91–113.

J. P. Kincaid, Jr. Fishburne, Rogers Robert P., Chissom Richard L., and Brad S. 1975. Derivation of New Readability Formulas (Automated Readability Index, Fog Count and Flesch Reading Ease Formula) for Navy Enlisted Personnel:. Technical report, Defense Technical Information Center, Fort Belvoir, VA.

Cunliang Kong, Yun Chen, Hengyuan Zhang, Liner Yang, and Erhong Yang. 2022. Multitasking Framework for Unsupervised Simple Definition Generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5934–5943, Dublin, Ireland. Association for Computational Linguistics.

Federico Martelli, Najla Kalach, Gabriele Tola, and Roberto Navigli. 2021. SemEval-2021 Task 2: Multilingual and Cross-lingual Word-in-Context Disambiguation (MCL-WiC). In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 24–36, Online. Association for Computational Linguistics.

Meta AI. 2025. The Llama 4 Herd: The Beginning of a New Era of Natively Multimodal AI Innovation.

Timothee Mickus, Elaine Zosa, Raul Vazquez, Teemu Vahtola, Jörg Tiedemann, et al. 2024. SemEval-2024 Task 6: SHROOM, a Shared-task on Hallucinations and Related Observable Overgeneration Mistakes. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 1979–1993, Mexico City, Mexico. Association for Computational Linguistics.

George A. Miller. 1994. WordNet: A Lexical Database for English. In *Human Language Technology: Proceedings of a Workshop held at Plainsboro, New Jersey, March 8-11, 1994*.

George A. Miller, Claudia Leacock, Randee Tengi, and Ross T. Bunker. 1993. A Semantic Concordance. In *Human Language Technology: Proceedings of a Workshop Held at Plainsboro, New Jersey, March 21-24, 1993*.

Roberto Navigli. 2009. Word sense disambiguation: A survey. *ACM Comput. Surv.*, 41(2):10:1–10:69.

Thanapon Noraset, Chen Liang, Larry Birnbaum, and Doug Downey. 2017. Definition Modeling: Learning to Define Word Embeddings in Natural Language. *Proceedings of the AAAI Conference on Artificial Intelligence*, 31(1). Number: 1.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, et al. 2024. GPT-4 Technical Report. ArXiv:2303.08774 [cs].

Francesco Periti, David Alfter, and Nina Tahmasebi. 2024. Automatically Generated Definitions and their utility for Modeling Word Meaning. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 14008–14026, Miami, Florida, USA. Association for Computational Linguistics.

Lorenzo Proietti, Stefano Perrella, Simone Tedeschi, Giulia Vulpis, Leonardo Lavalle, et al. 2024. Analyzing Homonymy Disambiguation Capabilities of Pretrained Language Models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 924–938, Torino, Italia. ELRA and ICCL.

Qwen Team. 2025. Qwen3.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, et al. 2023. Direct preference optimization: your language model is secretly a reward model. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, pages 53728–53741, Red Hook, NY, USA. Curran Associates Inc.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. 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 3982–3992, Hong Kong, China. Association for Computational Linguistics.

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.

Nina Rimsky, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, et al. 2024. Steering Llama 2 via Contrastive Activation Addition. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15504–15522, Bangkok, Thailand. Association for Computational Linguistics.

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.

Jan Trienes, Sebastian Joseph, Jörg Schlötterer, Christin Seifert, Kyle Lo, et al. 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.

Shen Wang, Tianlong Xu, Hang Li, Chaoli Zhang, Joleen Liang, et al. 2024. Large Language Models for Education: A Survey and Outlook. ArXiv:2403.18105 [cs].

Hanyi Xu, Wensheng Gan, Zhenlian Qi, Jiayang Wu, and Philip S. Yu. 2024. Large Language Models for Education: A Survey. ArXiv:2405.13001 [cs].