AmbiStory: A Challenging Dataset of Lexically Ambiguous Short Stories

Janosch Gehring
University of Technology Nuremberg
janosch.gehring@utn.de

Michael Roth
University of Technology Nuremberg
michael.roth@utn.de

Abstract

Word sense disambiguation is the task of selecting a word's applicable word sense in a given context. However, ambiguous texts may lack the information necessary to disambiguate words completely, resulting in multiple word senses with varying degrees of plausibility. We design a dataset around this premise: Our samples consist of 4–5 sentence short stories, where the more plausible word sense of the word to be disambiguated has to be inferred via indirect clues in surrounding sentences. We collect annotations from humans who rate the plausibility of a given word sense on a scale from 1–5. In total, our dataset contains 19,049 human word sense annotations on 1,899 stories. We investigate the performance of large language models on our data and find that many poorly correlate with human judgments. We also find that fine-tuning on our data can increase performance.[1]

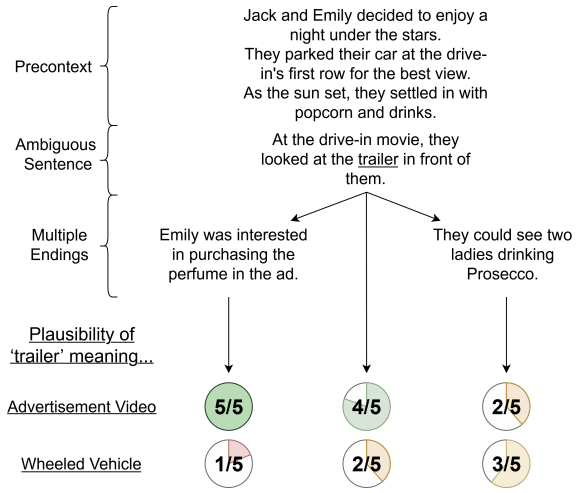

Figure 1: An example from our dataset. The ambiguous sentence is re-contextualized in three different endings, resulting in three different stories. Each ending results in different plausibility scores for the word senses.

1 Introduction

Lexical ambiguity describes the presence of multiple senses being applicable to the same word and has been argued to be a functional property of any efficient communicative system (Piantadosi et al., 2012). Indeed, psycholinguistic studies have shown that humans rapidly process such ambiguities in context (see e.g. McDonald and Shillcock, 2003). Computational approaches to resolving lexical ambiguity in context, commonly referred to as the task of word sense disambiguation (WSD), have been studied at least since the 80s (Dahlgren, 1988; Krovetz and Croft, 1989). In the past decade, research has moved from fixed sense inventories to graded assignments of senses (see §2), following the predominant psycholinguistic view that senses are not (strictly) categorical (for a discussion, see Trott and Bergen, 2023).

In parallel with developments in theory formation and corresponding benchmarks, there have been immense technical developments: specifically, large language models (LLMs) such as GPT-4o (OpenAI et al., 2024), DeepSeek (DeepSeek-AI et al., 2025) and Llama-3 (Grattafiori et al., 2024) have not only dominated the NLP community but also shown "superhuman" performance in WSD tasks (Wang and Zhao, 2024). Yet, lexical ambiguities still pose difficulties in tasks such as question answering, natural language inference, and machine translation (Zhang and Choi, 2025). Reasons for this include underspecified language, meaning that lexical ambiguities are not always resolvable (see Haber and Poesio, 2024), as well as differences in background knowledge that can lead to divergent interpretations of context (see e.g. Plank, 2022).

In this paper, we lay the foundations for investigating these difficulties. Specifically, we build a dataset of lexically ambiguous word usages (see Figure 1 for an example) and collect multiple judg-

[1]The data will be made available at https://github.com/Janosch-Gehring/ambistory.

*Proceedings of the 14th Joint Conference on Lexical and Computational Semantics (*SEM 2025)*, pages 152–171

ments on the plausibility of different sentence readings under varying discourse-level contexts. Our dataset consists of 4–5 sentence short stories where the fourth sentence contains a homonym[2] that is used in such a manner that the sentence is ambiguous when read in isolation. Hints towards the preferred reading are provided in surrounding sentences that manipulate the plausibility of word senses by introducing additional details. A correct disambiguation of the homonym thus requires a higher-level understanding of the story. Furthermore, as different contexts will change the plausibility of word senses to varying extents, our dataset contains both stories that are perceived as non-ambiguous and various degrees of ambiguity.

Based on the collected data, we examine the following research questions:

RQ1 How does context affect human plausibility judgments, and when do annotators disagree?

RQ2 How well do judgments by LLMs align with humans, and when are they different or fail?

RQ3 Can an LLM be trained in order to increase agreement with human judgments?

RQ4 How does additional context affect LLM performance and human agreements?

2 Related Work

Word Sense Disambiguation (WSD) has been studied extensively in the NLP community. The 'Word in Context' (*WiC*) series of tasks provide a common framework, in which two occurrences of a *target* word are (typically) classified in a binary manner as either representing the same word sense or different ones (Pilehvar and Camacho-Collados, 2019; Raganato et al., 2020). Similarly, the WiC Target Sense Verification (*TSV*) task presents a word in context with one specific sense, which is to be labeled as correct or incorrect (Breit et al., 2021).

While numerous WSD tasks operate on the sentence-level, a considerable body of research has also been dedicated to document-level WSD. For instance, datasets for 'All-Words' WSD tasks, where the objective is to classify the word sense of every word in the text, commonly consist of longer documents (Moro and Navigli, 2015). Other datasets similarly consist of large tagged corpora (Miller et al., 1993; Taghipour and Ng, 2015).

Comparatively little research has been dedicated to the idea of multiple word senses having varying levels of applicability. Jurgens and Klapaftis (2013) treat WSD as a ranking task, where word senses are given percentages based on their applicability. Other research tackles WSD as a multi-label classification task (Conia and Navigli, 2021) or introduces an ordinal scale to the WiC task (Schlechtweg et al., 2025; Erk et al., 2009). We combine ideas from these tasks and create the first dataset in which the plausibility of different word senses are judged independently given *varying amounts of context* beyond the sentence of a target word.

3 Data Collection

Stories in our dataset consist of 4–5 sentences: three describing the situation ('precontext'), one ambiguous sentence, and optionally an ending. Our collection of stories is a multi-step process in which we collect the precontexts and ambiguous sentences (forming a 'setup'), as well as the endings and the plausibility annotations separately.[3]

We present our data collection approach in the following. First, we extract a set of homonyms to use in our short stories (§3.1). Then, we let humans write ambiguous sentences and prepend a precontext to them to form the setup (§3.2). We then collect ending sentences written by humans (§3.3), and finally collect human plausibility ratings for each 'sample', i.e. the combination of word sense and story (§3.4).

For each step, we collect data via Prolific and require users to pass a qualification test before proceeding to the study to ensure data quality. In addition, we manually check each story for adherence to our guidelines and offensive content, and remove all low-quality submissions. During the collection of plausibility ratings, we include a question from the qualification test as an attention check. We filter annotators who fail this check.

3.1 Homonym Collection

We first collect a selection of homonyms and their word sense pairs around which stories are constructed. The objective is to collect word sense

[2]We use the term *homonym* loosely to refer to lexical ambiguities in our data, since most of the relevant sense pairs are presumably unrelated. This does not exclude cases of polysemy, which we also observe to affect annotation. (see §4.2).

[3]All of our data collection was conducted on Prolific under similar conditions, which are further described in Appendix A.1. Screenshots of the annotation interface can be found in Appendix C.

The man crossed the wrong people. He borrowed money from a dangerous group. They came knocking on his door demanding a settlement. The goons made the man **pay**.		Alice loved gardening and enjoyed trying to grow exotic fruit. Jeremy was always excited to taste the fresh produce from their little backyard. Recently, their fruit supply started to dwindle. Alice and Jeremy were running out of fruit, so they went out for a **date**.	
(word sense: *bear a cost or penalty, in recompense of some action*) **Endings:** They stole his belongings and promised to do worse next time before leaving.	(word sense: *give money, usually in exchange for goods and services*) They threatened him until he apologized and returned the money.	(word sense: *a meeting arranged in advance*) **Endings:** They decided to go to a local Hawaiian restaurant that had an extensive dessert menu.	(word sense: *sweet edible fruit of the date palm with a single long woody seed*) Needless to say, there were none to pick, so it necessitated a visit to a large supermarket in the nearby town.

Table 1: Example setups (1st row) from our data, with two different endings for each word sense (2nd row).

pairs where the word senses are distinct, yet can be used within the same sentence structure. Therefore, we decided to extract these from the pun dataset SemEval-2017 Task 7 (Miller et al., 2017), as words used for puns satisfy both criteria. We filter out word sense pairs where the word senses are different parts of speech and ones where a word sense requires a specific particle, as these limit the ambiguous sentences that can be constructed. The rest is used to create a large pool of 729 homonym word sense pairs.

3.2 Writing Ambiguous Sentence

In the second step, we collect ambiguous sentences from crowdworkers. In the annotation interface, humans are tasked with writing sentences where a randomly selected word from our pool of homonyms is used in such a manner that its two displayed word senses are both plausible. We display both word senses and two example sentences, generated with GPT-4o (OpenAI et al., 2024), demonstrating how the word senses can be used in a sentence to help annotators understand technical definitions. If a participant is unable to formulate a sentence, they can click a button to receive a different homonym. Subsequently, we manually filter all sentences that do not conform to the guidelines, e.g. because the homonym is used multiple times or the sentence is clearly non-ambiguous. In a few cases where the sentence is successfully ambiguous but clearly references the wrong word senses, we manually replace the word senses with the correct ones. The guidelines for this task are detailed in Appendix A.2.1. We then further utilize GPT-4o to edit the sentences to rectify spelling errors (see Appendix B.1 for details).

Finally, we employ GPT-4o to generate a precontext, comprising three sentences, for each ambiguous sentence. We instruct it to generate a beginning of the story which does not yet resolve the ambiguous word (for details, see Appendix B.2). The purpose of this additional exposition is to ground the narrative, thereby aiding annotators in writing endings and judging the situation's plausibility.

3.3 Writing Endings

We next collect two endings per setup. We display each story to two annotators and assign them the task of composing an ending that enhances the plausibility of one of the word senses of the homonym. As they are each displayed a different word sense, the plausibilities of word senses will vary between the endings despite the setup being the same. For more details about the task guidelines, please refer to Appendix A.2.2. Examples of ended stories in our dataset are displayed in Table 1.

We purposefully do not filter endings which fail to resolve the ambiguity, which is a common occurrence; recognizing to what extent the endings succeed in the homonym's disambiguation is a part of the challenge. Similarly to the previous step, we use GPT-4o to fix spelling errors and manually filter low-quality submissions. Refer to Appendix B.1 for the prompt.

Furthermore, we include an 'open-ended' story for each setup, which is devoid of an ending sentence and thereby commonly leaves the word sense unresolved. Thus, we obtain three stories for each setup: One for each of the endings, and one without an ending.

3.4 Plausibility Rating

Finally, we collect plausibility ratings for each word sense in the context of a story on a Likert scale ranging from 1 to 5, where 1 signifies that a word sense is inconceivable, while 5 represents unambiguous certainty. Each word sense is annotated by at least five annotators. We also give annotators the option to mark stories as 'nonsensical'.

Section	Avg. length (in words)
Precontext	31.5
Amb. Sentence	9.24
Ending	13.5
Entire Story	49.77

Table 2: Average length of story sections in our dataset.

For this task, only one of the word senses is displayed to the annotators; they have to rely on their own language understanding to perceive which other senses of the homonym may cause a potential ambiguity. We split annotators into 130 groups, each of which annotate 30 word sense–story combinations. Annotators do not receive multiple samples containing variations of the same stories so that each story's annotations remain independent of each other. The guidelines for the task are detailed further in Appendix A.2.3.

3.5 Final Data and Split

For computational experiments (§5), we split the resulting data into a training, development and test set. The sets are split by the homonym used, ensuring that the same target word does not appear across sets. The training set includes 2,280 samples, the development set 588 samples, and the test set 930 samples, for a total of 3,798 samples. Each sample provides the plausibility scores assigned for one word sense of a target word in the given story. For each setup, there are three such stories: one open-ended variant (without explicit ending) and two ended variants with one ending collected for each word sense. Consequently, there are six samples per setup, and 633 setups in total. We collected 19,049 human judgments in sum, with at least five plausibility judgments per sample.

4 Data Analysis

Based on the data collection described in Section 3 we want to analyze under which circumstances people view multiple readings as plausible and when disagreements occur (RQ1). Before that, we discuss statistical properties of the collected data (§4.1) and present a preliminary analysis of the effects of context and word senses on plausibility judgments (§4.2).

4.1 Statistics

Story Length Basic statistics of stories in our dataset are displayed in Table 2. The average story is around 50 words long, although open-ended stories are naturally shorter. The ambiguous sentence itself is typically the shortest part of the story, likely because of the restrictions posed on writers.

Homonyms The selection of homonyms during data collection is influenced by randomness and crowdworker preference. Also, some homonyms with more than two word senses have multiple word sense pairs in our random pool, which increases the odds of drawing them. Therefore, some homonyms appear more often than others. During the data collection process, we removed the most popular homonyms from the random pool to prevent overrepresentation. The most common homonyms with ambiguous sentences in our dataset are *change* (10x), *lousy* (9x), *shot* (7x) and *bars* (7x). In total, our dataset contains 361 different ambiguous word forms (305 different lemma), 411 pairs of word senses, and an average of 1.75 sentences per ambiguous word form.

Inter-Annotator Agreement. We analyze the inter-annotator agreement on our dataset using interval scale Krippendorff's α (Krippendorff, 2004). Our annotators achieve an agreement of $\alpha = 0.506$. We find this to be a reasonable level of agreement, given that our task depends on annotators' own subjective intuition regarding the plausibility of a story and the distinction between word senses.

The average standard deviation per sample is $\sigma = 0.946$. The homonym used seems to greatly influence the human agreement, perhaps due to inherent disagreements about word sense distinctions or the complexity of word sense definitions. For example, the homonym with the highest average standard deviation, *identities* ($\sigma = 1.59$), has abstract and mathematical definitions which may have confused the annotators. As the number of data points per homonym are too low to draw definite conclusions, we will leave further exploration of the effect of specific homonyms to future work.

4.2 Effects of Word Senses and Endings

RQ1: How does context affect human plausibility judgments, and when do they disagree? We investigate to what extent endings affect the perceived plausibility of word senses. To this end, we compare the average scores a word sense receives in different variations of a story. On average, each word sense's score differs by around 0.80 ($\sigma = 0.675$) when an explicit ending is added to the stories. This shows that the additional context

has a strong effect on what word sense is perceived as plausible for a target word in a given sentence.

When contrasting the scores word senses receive in one ending versus the other, we find that it varies by around 1.18 ($\sigma = 0.941$) on average. However, this variance differs greatly between stories. Some endings fail to change the perceived plausibility, resulting in mostly unchanged scores, whereas other endings resolve the perceived ambiguity of the story.

Disagreements Human disagreement mostly stems from annotators disagreeing on the extent of a story's ambiguity, thus picking extreme values while others pick middle values. Indeed, about 50% of human ratings are either *1* or *5*. The least picked rating is 3 (15%), which indicates that humans typically have a preference for one of the word senses instead of thinking of multiple as equally plausible. While non-ambiguous stories are typically rated as 1–2 or 4–5 by all humans, ambiguous ones are the cause of much disagreement, with ratings for one word sense sometimes ranging across the entire scale. Although some outliers may be attributed to noise, we believe the following to be two of the most important factors for human agreement:

Word Sense Distinction. As previously stated, humans are not given a word sense inventory when rating plausibilities. Even though our dataset focuses on homonymous word senses, there are occasional instances of related word senses, including literal and figurative usages of words such as *alive* and *drooling*. Perhaps because some pairs of senses are perceived as identical in meaning, most annotators picked a label of *5* on all stories and senses of these words, such as 'be envious' and 'let saliva drivel' for *drooling*.

Word Sense Frequency. Lower-frequency word senses tend to cause disagreement, as humans may disagree on the plausibility of their usage or even forget about them altogether. For instance, in a story containing the sentence '*The blankets in the hotel were pretty lousy*', without additional ending context, all annotators rated the word sense '*very bad*' of *lousy* as a *5*. However, humans are less confident about the lower-frequency word sense '*infested with lice*', rating it as either *2* or *3*. Based on annotator comments, it seems that lower-frequency word senses are often not considered as plausible without supporting information. Approximating sense frequencies based on SemCor (Miller et al., 1993), we indeed find a small but highly significant ($p < 0.01$) correlation between the frequency count of a word sense and the average annotator score, as determined using Spearman's ρ (Spearman, 1904).

5 Computational Experiments

Following the analysis of human annotations in Section 4, we next conduct preliminary experiments on our dataset to answer the remaining research questions outlined in the introduction. We first describe the task setup, models and evaluation metrics used across experiments (§5.1) and then address our research questions regarding the alignment between LMs and human judgments (§5.2) and on the possibility of training LLMs to increase this alignment (§5.3). In context of these experiments, we also analyze the effect of endings as additional context on LLM predictions.

5.1 Experimental Setup

In the first experiment, we investigate the performance of LLMs on our test set without any fine-tuning on the training or development set. Formally, we define the task as follows: Each story text T_i in our data contains a *precontext* c_i, *ambiguous sentence* a_i and an optional *ending* e_i, forming triplets $T_i = \langle c_i, a_i[, e_i]\rangle$. The ambiguous sentence a_i contains a homonymous word form w with two word sense definitions $S_w = \{s_1^w, s_2^w\}$. The task is to predict a plausibility score $p = f(s_j^w \mid T)$ for each sense $s_j^w \in S_w$, where $p \in \{1, 2, 3, 4, 5\}$ and f is a function or model that assigns a score.

We test multiple pre-trained models for this task: **GPT-4o-2024-08-06**, **GPT-4o-mini-2024-07-18** (OpenAI et al., 2024), **o3-2025-04-16** (OpenAI, 2025), **Llama-3.1-8B-Instruct** (Grattafiori et al., 2024), **Mistral-7B-Instruct-v0.3** (Jiang et al., 2023), **Mixtral-8x7B-Instruct-v0.1** (Jiang et al., 2024) and **DeepSeek-V3-0324** (DeepSeek-AI et al., 2025). The GPT and DeepSeek models were accessed through their respective APIs using default settings, whereas the other models were tested using the Huggingface transformers library (Wolf et al., 2020). We set the temperature to 0 for all models.[4] We experiment with zero-shot and few-shot prompting techniques. In the zero-shot setting, we prompt the model with an adjusted version of the annotation guidelines, while for few-shot, we also show the examples displayed to annotators before they start the task.

[4] o3 does not allow for setting temperature, so default settings were used.

	Spearman	Acc. w/in SD
Random	0.000	0.454
Majority	N/A	0.558
Llama-3 (0-shot)	0.462	0.663
Mistral (0-shot)	0.382	0.568
Mixtral (0-shot)	0.606	0.634
GPT-4o-mini (0-shot)	0.726	0.726
GPT-4o (0-shot)	**0.756**	0.755
o3 (0-shot)	0.753	0.763
DeepSeek (0-shot)	0.740	**0.790**
Llama-3 (4-shot)	0.491	0.694
Mistral (4-shot)	0.209	0.522
Mixtral (4-shot)	0.607	0.649
GPT-4o-mini (4-shot)	0.737	0.726
GPT-4o (4-shot)	0.742	0.725
o3 (4-shot)	0.742	0.760
DeepSeek (4-shot)	**0.767**	**0.816**
Human Upper Bound	0.834	0.892

Table 3: 'Spearman ρ' and 'Accuracy Within Standard Deviation' scores for different baselines, out-of-the-box LLMs and our human upper bound.

Evaluation metrics One of our main research questions concerns how well LLM predictions align with human judgments (RQ2). As there is no gold standard for this setting, we evaluate models based on their correlation with averaged human judgments as well as a variance-adjusted accuracy measure. Specifically, we calculate the correlation between a model's judgments and the human average using **Spearman's** ρ (Spearman, 1904). As some samples have a clearer consensus than others, we consider their annotators' standard deviation for accuracy. That is, we calculate **Accuracy within Standard Deviation** as the proportion of model predictions that are within standard deviation (at least 1) from the average judgment by annotators.

Baselines and upper bound We use two simple baselines for the task: *Random*, which randomly picks a number between 1 and 5 for every sample, and *Majority*, which picks the majority label (which we found to be 4 on average). We also estimate the best possible performance of humans by evaluating each annotator against the other annotators who received the same sample. We calculate this human upper bound by averaging the scores of the highest-scoring human in each group.

5.2 Experiment 1: Out-of-the-Box LLMs

In the first experiment, we test LLMs out-of-the-box to test alignment with plausibility judgments provided by humans. As indicated by the results in Table 3, there are large differences in performance between models, roughly scaling with model size. However, while most models perform better than the majority or random baseline, the human upper bound is still remarkably higher at a Spearman ρ of 0.834 and an accuracy of 89%. The best performing model is DeepSeek-V3, being the only model to obtain an accuracy of over 80%. The reasoning model o3 performs the second best with an accuracy of 76%. GPT-4o models achieve scores between 72–75%, whereas the Mistral, Mixtral and Llama-3 models all score below 70%.

The effectiveness of zero-shot and few-shot prompting techniques also seems to vary between models. GPT-4o in particular is hurt by the addition of few-shot prompting, whereas models such as Mistral-7B and Llama-8B benefit greatly from the addition of examples.

RQ2: How well do judgments by LLMs align with humans, and when are they different or fail? Given the accuracy gap between models and the human upper bound, it seems there is a sizable difference between human and model judgments. Model judgments are also fairly different between models and prompting techniques. Some examples are displayed in Table 4. We identified the following common error sources on our test set:

'Red Herring' Keywords. In many stories where the setup itself already strongly favors one word sense, endings may be ineffective at swaying the plausibilities perceived by humans despite introducing keywords relating to the less plausible word sense. An example of this is the first story in Table 4, where models such as GPT-4o and DeepSeek-V3 gravitate towards the word *cue* referring to a billiards stick due to the ending mentioning snooker, whereas humans remain unsure about the intended word sense.

Judgment of Open-Ended Stories. As we will discuss further in Section 5.3.2, models seem to differ the most from humans on open-ended stories (i.e. without explicit ending). For example, while all humans and models recognize the ambiguity of the word *shots* in the second example of Table 4, their predictions range from 2–4. Meanwhile, all humans picked a label of *3*, indicating no preference for either word sense.

Story	Word Sense	Human average	Model predictions
He spent years perfecting his craft. Every night, he rehearsed in his small, dimly lit room. Now, it was time to showcase his skills on the big stage. In front of the large audience, table set, he waited for his cue. He was ready to play snooker.	sports implement consisting of a tapering rod used to strike a cue ball in pool or billiards	2.8	[4, 2, 5, 4, 3]
It had been a long week for Emma. She felt overwhelmed by everything happening. On Saturday, she finally decided to do something about it. She took three shots that day.	a small drink of liquor	3.0	[4, 4, 3, 2, 2]
The girl packed her bag for a long adventure. She was excited to explore new places she had never been. Along the way, she faced unexpected challenges that left her feeling different. While on a backpacking trip, the girl ended up in a strange state. She hadn't been there before, but she liked it there and would go again.	the territory occupied by one of the constituent administrative districts of a nation	4.4	[5, 2, 2, 2, 3]

Table 4: Examples of human averages and LLM ratings for word sense plausibility in stories. All pictured model scores were taken from the zero-shot setting (order of scores: GPT 4o, 4o-mini, DeepSeek, Llama-3, Llama-3 + FT).

Word Sense Distinction. LLMs, especially ones trained on huge amounts of online data, are likely influenced by online word sense inventories that were seen during training. Thus, it seems reasonable to assume that LLMs may have knowledge about the sense distinctions of dictionaries such as WordNet (Miller, 1994), whereas most humans do not. Indeed, we find that humans frequently group similar word senses together. In the last example of Table 4, humans unanimously agree on the definition of *state* being correct, since it clearly refers to a territory. However, even then, *state* could realistically refer to both a *province* or a *country*, which are mapped to distinct WordNet senses. Therefore, this is a case where human intuition does not align with a dictionary, which could be a large source for disagreements between models and humans.

5.3 Experiment 2: Fine-Tuning Llama-3 for Plausibility Rating

As a follow-up question, we test whether LLMs can be trained in order to increase agreement with human judgments (RQ3). In the following, we demonstrate the utility of our training and validation set by comparing the performance of the out-of-the-box Llama-3 and Mistral model with versions fine-tuned on our data, which we call '**Llama-3 + FT**'and '**Mistral + FT**'.

5.3.1 Setup

We use LoRA (Hu et al., 2021) to fine-tune the LLama-3.1-8B-Instruct and the Mistral-7B-Instruct-v0.3 model on our training and validation set. During the fine-tuning process, we use a dropout of 0.1, a learning rate of 2e-4, a simulated batch size of 16, rank r of 16, α of 32, the target modules *q_proj* and *v_proj*, and pytorch's AdamW optimizer. We set the model to train for 20 epochs with an early stopping patience of 5. Training stopped after 8 epochs for both models.

When feeding our training data into the model, we first display the story itself, followed by the string: '*In this context, how plausible is it that the meaning of the word {homonym} is {definition} (as in: {example sentence})? Return only the numbered score (1, 2, 3, 4 or 5). Do not return anything else!*'. This mirrors the prompt used during testing, as well as the annotation interface seen by humans. We concatenate the rounded average of human ratings to the input string and only train the model to predict that final number.

5.3.2 Results

Results are displayed in Table 5. We observe a highly significant performance difference between the base models and their fine-tuned counterparts (as determined using Wilcoxon signed-rank test; Wilcoxon, 1945). The performance boost gained from fine-tuning on our data lets the Llama-3 + FT model and the Mistral + FT model achieve Accuracy and Spearman ρ competitive with that of strong models such as GPT-4o or GPT-4o-mini.

RQ3: Can an LLM be trained on our data to improve agreement with humans? Given the large score improvement of the fine-tuned models, we are confident our data can also be useful for improving other models' processing of ambiguities. Note that there is no homonym overlap between the test set and the training/development set, so the

	Spearman	Acc. w/in SD
Random	0.000	0.454
Majority	N/A	0.558
Llama-3 (0-shot)	0.462	0.663
Llama-3+ FT (0-shot)	**0.725**	0.698
Mistral (0-shot)	0.382	0.568
Mistral + FT (0-shot)	0.692	**0.726**
Llama-3 (4-shot)	0.491	0.694
Llama-3 + FT (4-shot)	**0.751**	**0.795**
Mistral (4-shot)	0.209	0.522
Mistral + FT (4-shot)	0.684	0.733
Human Upper Bound	0.834	0.892

Table 5: Results for Llama-3 and Mistral out-of-the-box (LLama-3 / Mistral) and with additional fine-tuning (Llama 3 + FT / Mistral + FT).

All Stories	1	2	3	4	5
Llama-3	0.3	42.5	0.2	56.6	0.5
Llama-3 + FT	25.8	20.4	15.5	32.2	6.1
Human Avg	11.7	21.8	24.2	24.0	18.3
Open-Ended stories only					
Llama-3	0.5	45.2	0.5	53.1	0.8
Llama-3 + FT	26.9	20.3	16.3	32.4	4.0
Human Avg	8.4	23.6	27.1	27.4	13.6
Ended stories only					
Llama-3	0.2	41.2	0.0	58.3	0.3
Llama-3 + FT	25.2	20.4	15.2	32.0	7.2
Human Avg	13.4	21.0	22.7	22.3	20.7

Table 6: Label distribution in the test set (in %). Llama-3 distributions include 0-shot and 4-shot predictions. Human Avg is based on rounded plausibility scores.

results show that the fine-tuned model generalizes beyond homonyms from the training set.

The improvement is also evident in the change in label distribution, as shown in Table 6 on the example of Llama-3 and Llama-3 + FT. The original Llama-3 model strongly favors the '2' and '4' labels, with those two labels accounting for nearly all of its predictions. In contrast, predictions by Llama-3 + FT are balanced more evenly, aligning much closer with the relatively balanced distribution of the averaged human judgments.

We next perform an additional analysis regarding the role of endings on performance differences between LLMs and human judgments, answering our final research question (RQ4):

RQ4. How does additional context affect LLM performance and human agreements? Intuitively, we expected the disambiguating information found in the endings to remove much of the human disagreement, as there is less need for subjective conjecture. However, while there are some slight tendencies, we do not find the difference between standard deviations of human ratings of open-ended and ended stories to be significant (as determined using Wilcoxon rank-sum test; Wilcoxon, 1945). Nonetheless, we do observe an effect of endings on the label distributions assigned across stories, as displayed in Table 6. In particular, our annotators pick the labels '1' and '5' 5-7 percentage points more often for ended stories than open-ended stories. This effect is much less pronounced for Llama-3 even after fine-tuning, which

	Open-Ended	Ended
Mistral (0-shot)	0.503	0.503
Mistral + FT (0-shot)	0.725	0.726
Mixtral (0-shot)	0.619	0.642
Llama-3 (0-shot)	0.667	0.661
Llama-3 + FT (0-shot)	0.681	0.706
GPT 4o-mini (0-shot)	0.697	0.740
GPT 4o (0-shot)	0.713	0.775
o3 (0-shot)	**0.768**	0.761
DeepSeek (0-shot)	0.755	**0.808**
Mistral (4-shot)	0.635	0.648
Mistral + FT (4-shot)	0.725	0.737
Mixtral (4-shot)	0.626	0.661
Llama-3 (4-shot)	0.739	0.671
Llama-3 + FT (4-shot)	0.774	0.804
GPT 4o-mini (4-shot)	0.694	0.742
GPT 4o (4-shot)	0.7	0.737
o3 (4-shot)	0.781	0.75
DeepSeek (4-shot)	**0.790**	**0.829**
Average (0-shot)	0.681	0.702
Average (4-shot)	0.717	0.731

Table 7: Accuracy within Standard Deviation scores of models on stories without ending (Open-Ended) and with an explicit ending (Ended).

suggests that there is still room for improvement through other training strategies.

We also investigate whether there is a difference in model performance between ended and open-ended stories. Results are displayed in Table 7.

Interestingly, most models perform better on ended stories than open-ended ones (notable exceptions including o3 and the low-scoring Llama-3). As open-ended samples seem to generally be more challenging to models, but not to humans, we argue that models struggle with the low information content of open-ended stories. Humans may have an innate intuition for interpreting word senses even without evidence that is less pronounced in models. The o3 model does not appear to experience performance decrease on open-ended stories, which suggests that reasoning models may be particularly well-suited for recognizing ambiguity.

6 Conclusion

We introduced *AmbiStory*, a collection of 1,899 short stories with human plausibility ratings for word senses. Each story contains a lexical ambiguity with multiple alternate endings designed to imply different word senses. We investigated how human plausibility judgments are influenced by story context, and found inter-annotator agreement to be affected by the frequency of word senses as well as relations between them, among other reasons. In computational experiments, we found LLM predictions often do not match human-perceived plausibilities, for instance, because of model biases towards high-frequency senses and distracting key words in the story context.

In general, we found models to perform worse when no ending to a story is provided, whereas human agreement remains stable. This may indicate that LLMs lack the common sense to 'fill in the gaps' in stories with only minimal disambiguating information. We believe that our dataset provides a useful basis for model development and testing, as exemplified by the possibility of improving the performance of two LLMs via fine-tuning. In fact, AmbiStory also serves as the benchmark for a shared task at SemEval 2026 (Task 5). We encourage the community to use AmbiStory to study and model human perception of lexical ambiguity beyond individual sentences.

Limitations

While we believe AmbiStory to be an effective benchmark in its current state, there are some limitations that we would like to address in future experiments and expansions. Firstly, the stories in this dataset have a fixed five-sentence length and 3/1/1 structure, where the ambiguous sentence is always the fourth sentence and the disambiguating evidence can typically be found in the last sentence. A possible expansion would be to grant crowdworkers more flexibility by adjusting the annotation pipeline, allowing for the story length and the position of the homonym and disambiguating information to be more dynamic (and thus less predictable). Additionally, it would be insightful to collect judgments for domain-specific homonyms and analyze annotator feedback to better understand sources of disagreements.

As an addition limitation, we note that our dataset is currently restricted to English and we collected data primarily from crowdworkers in the UK, whose views may not necessarily reflect those of other native and non-native speakers. Also, as homonym sense pairs are obtained from a pre-existing dataset, the set of homonyms in our dataset is not exhaustive. Finally, our stories may contain linguistic biases, such as stylistic tendencies common in language model outputs, as parts of the stories are written or corrected by GPT-4o.

Ethical Considerations

Our dataset contains text written by LLMs and crowdworkers, both of which are susceptible to producing harmful content. We addressed this ethical risk by carefully filtering all samples that could be considered harmful, including for example cases in which one of the word senses has vulgar or offensive connotations in slang. We additionally ensured that crowdworkers are anonymized and paid above minimum wage according to the regulations in the country of the authors.

References

Anna Breit, Artem Revenko, Kiamehr Rezaee, Mohammad Taher Pilehvar, and Jose Camacho-Collados. 2021. WiC-TSV: An evaluation benchmark for target sense verification of words in context. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1635–1645, Online. Association for Computational Linguistics.

Simone Conia and Roberto Navigli. 2021. Framing Word Sense Disambiguation as a multi-label problem for model-agnostic knowledge integration. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3269–3275, Online. Association for Computational Linguistics.

Kathleen Dahlgren. 1988. Word sense disambiguation.

In *Naive Semantics for Natural Language Understanding*, pages 141–169. Springer.

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. *Preprint*, arXiv:2412.19437.

Katrin Erk, Diana Mccarthy, and Nicholas Gaylord. 2009. Investigations on word senses and word usages. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 10–18.

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. *Preprint*, arXiv:2407.21783.

Janosch Haber and Massimo Poesio. 2024. Polysemy—Evidence from linguistics, behavioral science, and contextualized language models. *Computational Linguistics*, 50(1):351–417.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *Preprint*, arXiv:2106.09685.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, Lélio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, and 7 others. 2024. Mixtral of experts. *Preprint*, arXiv:2401.04088.

David Jurgens and Ioannis P. Klapaftis. 2013. Semeval-2013 task 13: Word sense induction for graded and non-graded senses. In *International Workshop on Semantic Evaluation*.

Klaus Krippendorff. 2004. *Content Analysis: An Introduction to Its Methodology (second edition)*. Sage Publications.

Robert Krovetz and W Bruce Croft. 1989. Word sense disambiguation using machine-readable dictionaries. In *Proceedings of the 12th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 127–136.

Scott A McDonald and Richard C Shillcock. 2003. Eye movements reveal the on-line computation of lexical probabilities during reading. *Psychological science*, 14(6):648–652.

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*.

Tristan Miller, Christian Hempelmann, and Iryna Gurevych. 2017. SemEval-2017 task 7: Detection and interpretation of English puns. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 58–68, Vancouver, Canada. Association for Computational Linguistics.

Andrea Moro and Roberto Navigli. 2015. SemEval-2015 task 13: Multilingual all-words sense disambiguation and entity linking. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 288–297, Denver, Colorado. Association for Computational Linguistics.

OpenAI, :, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Mądry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, and 401 others. 2024. Gpt-4o system card. *Preprint*, arXiv:2410.21276.

OpenAI. 2025. o3 and o4-mini system card. https://cdn.openai.com/pdf/2221c875-02dc-4789-800b-e7758f3722c1/o3-and-o4-mini-system-card.pdf. Accessed: 2025-09-11.

Steven T Piantadosi, Harry Tily, and Edward Gibson. 2012. The communicative function of ambiguity in language. *Cognition*, 122(3):280–291.

Mohammad Taher Pilehvar and Jose Camacho-Collados. 2019. WiC: the word-in-context dataset for evaluating context-sensitive meaning representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1267–1273, Minneapolis, Minnesota. Association for Computational Linguistics.

Barbara Plank. 2022. The "problem" of human label variation: On ground truth in data, modeling and

evaluation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10671–10682, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Alessandro Raganato, Tommaso Pasini, Jose Camacho-Collados, and Mohammad Taher Pilehvar. 2020. XL-WiC: A multilingual benchmark for evaluating semantic contextualization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7193–7206, Online. Association for Computational Linguistics.

Dominik Schlechtweg, Tejaswi Choppa, Wei Zhao, and Michael Roth. 2025. CoMeDi shared task: Median judgment classification & mean disagreement ranking with ordinal word-in-context judgments. In *Proceedings of Context and Meaning: Navigating Disagreements in NLP Annotation*, pages 33–47, Abu Dhabi, UAE. International Committee on Computational Linguistics.

C. Spearman. 1904. The proof and measurement of association between two things. *American Journal of Psychology*, 15:88–103.

Kaveh Taghipour and Hwee Tou Ng. 2015. One million sense-tagged instances for word sense disambiguation and induction. In *Proceedings of the Nineteenth Conference on Computational Natural Language Learning*, pages 338–344, Beijing, China. Association for Computational Linguistics.

Sean Trott and Benjamin Bergen. 2023. Word meaning is both categorical and continuous. *Psychological Review*, 130(5):1239.

Yuqing Wang and Yun Zhao. 2024. Metacognitive prompting improves understanding in large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1914–1926, Mexico City, Mexico. Association for Computational Linguistics.

Frank Wilcoxon. 1945. Individual comparisons by ranking methods. *Biometrics Bulletin*, 1(6):80–83.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, and 3 others. 2020. Huggingface's transformers: State-of-the-art natural language processing. *Preprint*, arXiv:1910.03771.

Michael JQ Zhang and Eunsol Choi. 2025. Clarify when necessary: Resolving ambiguity through interaction with LMs. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 5526–5543, Albuquerque, New Mexico. Association for Computational Linguistics.

A Annotation Details

A.1 Annotation Setup

We use Prolific for all of our data collection. We set English as first language and an approval rate of at least 97% as prerequisites for participating. Most of our writers and annotators reside in the UK (91%), with some living in the US (5%) and Australia (4%). We provided a median payment of about 11 pounds per hour for all of our tasks, retroactively issuing additional payments where this median was not maintained.

For the 'Ambiguous Sentence Writing Task', annotators are required to write 5 sentences. For the 'Ending Writing Task', they have to produce 20 endings, and for the 'Plausibility Annotation Task', they have to rate 30 samples. We chose the workload per participant to be manageable without them becoming bored or running out of ideas.

A.2 Guidelines

The following are the guidelines, presented on the annotation website in a markdown format.

A.2.1 Ambiguous Sentence Writing Task

Introduction to the Ambiguous Sentence Task

Overview

Your task is to write ambiguous sentences. A sentence is ambiguous if there are multiple ways to understand it.
You will be presented with a word and two of its meanings. For example, the word *bank* and its two word senses "*a financial institution*" and "*slope next to a river*". Your task is to write one sentence where the word is used in such a way that depending on how you choose to interpret it, either meaning could apply.
For example, you could write something like: "*On Saturday morning, I went to the bank.*" The word *bank* here could realistically be either of the two word senses, a financial institution or a slope next to a river, so this would be acceptable.
When writing sentences, please try to uphold the following principles:

- **Ideally, both word senses should be**

equally plausible.

- **Please make sure that the word appears exactly once in the sentence.**
- **The goal is not to write puns, but to write sentences that allow for multiple interpretations.**
- **Submissions that clearly contain AI generated content will be rejected.**

Good Examples

Word: *racket*
Meaning 1: *a loud and disturbing noise*
Meaning 2: *implement used in sports, e.g. tennis racket*
Good Example: *I couldn't concentrate at all because of the loud racket from the tennis game.*
This is good: It could refer to the tennis game itself being loud (meaning 1), or the speaker specifically complains about the sound of the tennis racket (meaning 2). The sentence is plausibly ambiguous.

Word: *season*
Meaning 1: *a set of related television programs*
Meaning 2: *the four times of year (spring, summer, fall, winter)*
Good Example: *Anna thinks that this is the best season.*
This is good: We don't know what "*this*" refers to. It could be either a time of year or a television series. Because of this, the sentence remains successfully ambiguous.

Bad Examples

Word: *bat*
Meaning 1: *an implement with a handle used for hitting the ball in games such as cricket or baseball.*
Meaning 2: *a nocturnal mammal capable of sustained flight*
Bad Example: *The bat flew out of the cave.*
This is bad: It is not plausible that this refers to an implement like a baseball bat. The word "bat" is not used ambiguously.

Word: *root*
Meaning 1: *part of the plant which attaches it to the ground*
Meaning 2: *the basic cause, source or origin of something*
Bad Example: *The root of the problem was buried deep, just like the roots of the old oak tree in the yard.*
This is bad: The word is used multiple times. *Root* should be used only once, and that occurrence should encompass both meanings.

Task Procedure

You will first have to pass a simple 4-question qualification test to confirm that you are human and understand English. The qualification test will ask you to pick the best meaning for a word in a short text. If you fail the qualification test, you will not be able to start writing.
If you pass qualification, the writing page will automatically unlock for you.
If you cannot think of a sentence for the given word, you can press the button on the top of the page to get a new word. You can use the button as many times as you want. You have to write five sentences to clear the task.
Good luck and have fun!

A.2.2 Ending Writing Task

Introduction to the Story Ending Task

Overview

You will see the first four sentences of a five-sentence short story. Your task is to write one sentence that finishes the short story coherently.

Importantly, the fourth sentence contains a word which has multiple meanings. The story is written in such a way that you wouldn't know which meaning is the one intended by the author by the first four sentences.
Your ending sentence has to make the intended meaning more plausible than the unintended meaning. Both meanings will be displayed to you.
Try to not explicitly spell out the meaning. Understanding the intended meaning should only be easy to those who pay attention and understand the story. You can achieve this by *avoiding words closely associated with the intended meaning*, and by *only implying the happenings in the fourth sentence instead of outright stating them*. See below for examples.
The story ending should also be *logical* given the first four sentences. The story does not have to be interesting, it just has to make sense and be coherent.
Also, **your ending sentence should NOT include that ambiguous word with multiple meanings from the fourth sentence.**
Also, feel free to add comments in the comment field.

Examples

Story: "Mr Ellis walked to the town square with a big smile. He carried his easel with him. He was getting ready to paint. Whenever he sets up his easel in the town square, he always draws a crowd."
Intended Meaning of "draws": to attract
Unintended Meaning of "draws": produce an image of something or someone
Your Ending Sentence: Everyone was quite impressed by his picture of a flower.
Explanation: In the first four sentences, it is not clear if he literally sketches a crowd or if he simply attracts a crowd. The ending sentence resolves this by implying that he was sketching a flower, not a crowd.

Story: "The battle raged on as the sun began to set. He crouched behind a crumbling wall, desperately scanning the ground. Bullets whizzed past while his heart pounded loudly in his ears. He realized that he only had one shot'.
Intended Meaning of "shot": a chance or opportunity to do something
Unintended Meaning of "shot": a missile from a firearm
Your Ending Sentence: He thought that it was now or never as he attempted to escape from the battle.
Explanation: Some stories like this may make it hard to argue for the intended meaning when the unintended meaning seems more likely to begin with. Try to be creative and come up with a way to increase the intended meaning's plausibility. Here, the implication of him running away and the expression "now or never" implies that the amount of shots in his firearm is less important than his chance/opportunity to get away.

The Annotation Procedure

You will first have to pass the qualification test, which consists of four questions. **You only have one attempt at this.** The qualification test simply asks you to pick the more likely meaning of English words in the context of a sentence. It is mostly to filter out bots and should be no problem to English speakers.
If you fail at the qualification test: You will not be able to start writing. ***IMPORTANT***: **Remember to copy the Screen-Out Completion Code that will be displayed to Prolific.**
If you succeed at the qualification test: Once you pass the qualification test, you will automatically unlock access to the writing page. Each writer is assigned around 20 stories.
Your progress for both qualification and writing is **saved automatically** anytime you press the *NEXT* button.
Good luck and have fun!

Please note that the sentence about explicitly not using the ambiguous word again in the ending was

only added after data collection already started, as some annotators misunderstood the task prior to us adding it. We manually filtered endings where the task was trivialized by not adhering to this.

A.2.3 Plausibility Annotation Task

The following are the guidelines presented to annotators during the plausibility annotation step.

Introduction to the Plausibility Annotation Task

Overview

You will see a short text in which one sentence is written in bold. That sentence contains a word that can typically take on multiple different meanings, depending on the context. One of those meanings is shown to you.
Your task is simple: Annotate how plausible a meaning of a word is in the context of the short text using one of five scores:

- **1**: The displayed meaning is not plausible at all given the context.
- **2**: The displayed meaning is theoretically conceivable, but less plausible than other meanings.
- **3**: The displayed meaning represents one of multiple, similarly plausible interpretations.
- **4**: The displayed meaning represents the most plausible interpretation; other meanings may still be conceivable.
- **5**: The displayed meaning is the only plausible meaning given the context.

See below for examples.
You can also mark stories as *nonsensical*, meaning that the text is strange no matter which meaning of the word is used. Even if a story is nonsensical, try to pick whatever makes the most sense to you. You can simply go with your intuition here.
Also, feel free to add comments in the comment field.
There will be times where there is no objectively correct answer. Whatever the case, always look at all of the sentences and carefully think about how plausible each meaning would be.

Examples

"**The bat flew out of the cave.**"
Meaning of "BAT": A sports implement for hitting balls (e.g. in baseball)
Your Rating: *1*
Explanation: A baseball bat flying out of a cave is inconceivable; it obviously refers to an animal.

"The letter specified where to meet him. **So after reading it, I went to the bank.**"
Meaning of "BANK": a financial institution
Your Rating: *3*
Explanation: *Bank* could refer to the financial institution, but it could also be a river bank. Neither one seems particularly plausible compared to the other, so your rating should be in the middle.

"The composer often spontaneously had ideas for new melodies. **She writes notes on a sheet of paper.** She can later turn these into a piece."
Meaning of "NOTES": a brief written record; a memo
Your Rating *2*
Explanation: *Notes* could conceivably refer to written memos, but given the surrounding sentences, it is more plausible that she is jotting down musical notes.

"Mr Ellis walked to the town square with a big smile. He was getting ready to paint. **Whenever he sets up his easel in the town square, he always draws a crowd.** His painting of a flower looked really realistic!"
Meaning of "DRAWS": to attract; direct towards itself
Your Rating: *5*
Explanation: Without the last sentence, it is not clear whether the intended meaning of *draws* is *to sketch* or *to attract*. With the additional context that he is sketching a flower

- not a crowd - it becomes clear that *attract* is the only plausible meaning. **Always look at the whole story before making your decision!**

The Annotation Procedure

You will first have to pass the qualification test, which consists of four questions. **You only have one attempt at this.** Carefully look at the sentences and determine the plausibility of meanings.
If you fail at the qualification test: You will not be able to start the annotation. **Remember to copy the Screen-Out Completion Code that will be displayed to Prolific.**
If you succeed at the qualification test: Once you pass the qualification test, you will automatically unlock access to the annotation page. Each annotator is assigned 30 samples.
Your progress for both qualification and annotation is **saved automatically** anytime you press the *NEXT* button.
Good luck and have fun!

A.3 Qualification Test

We include a similar qualification test at each data collection step, where annotators evaluate the plausibility of word senses in four short stories. Note that while the samples are the same, the task differs between the writing tasks and the plausibility rating task: During writing tasks, annotators have to select the more plausible word sense out of two options, whereas during the plausibility rating task, to keep it similar to the actual annotation, only one word sense is displayed and annotators pick on a scale from 1 to 5. The qualification questions are pictured here. The correct answers are underlined; picking a different answer results in immediate disqualification.

Story: The puzzle pieces were scattered across the table. We spent hours on the puzzle, but each piece seemed to fit nowhere. **It was a hard puzzle.**

Most Plausible Meaning of "HARD":
☑ *difficult to understand or solve*
☐ *solid; not soft*

Plausibility of the meaning 'solid; not soft':
1 2 3 4 5

Story: The hotel I stayed at last week was the worst. The room was dirty and there was no one at the reception. **The service there was really lousy.**

Most Plausible Meaning of "LOUSY":
☐ *infested with lice*
☑ *very bad*

Plausibility of the meaning 'very bad':
1 2 3 4 5

Story: Max started sweating in the summer heat. His entire house felt like a sauna. **If he had a fan, the heat would be bearable.**

Most Plausible Meaning of "FAN":
☑ *a device for creating a current of air*
☐ *an enthusiastic follower or admirer*

Plausibility of the meaning 'an enthusiastic follower or admirer':
1 2 3 4 5

Story: Mia and Peter were preparing for the big race. Peter was confident in his talent, so he did not train much. **In the end, Mia beat him in the race.**

Most Plausible Meaning of "BEAT":
☐ *strike violently*
☑ *come out better in a competition, race or conflict*

Plausibility of the meaning 'come out better in a competition, race or conflict':
1 2 3 4 5

B Prompting

We use GPT-4o prompts at several steps of our study, which we describe here.

B.1 Annotation Cleaning

We use GPT-4o twice to improve the formatting, once for the ambiguous sentences and once for the endings. We instruct it to correct formatting errors such as punctuation and capitalization to improve the readibility of the story. We also instruct it to rewrite sentences containing direct speech into third person to facilitate the construction of a story around them. We attempt to restrict GPT-4o from making any other modifications, as any extraneous change could affect the writers' carefully constructed ambiguity. In a few instances, GPT-4o still replaced the ambiguous word with a synonym, which we manually fixed.

The following is the prompt for cleaning the ambiguous sentences:

> Correct the following sentence by fixing the following:
> - Fixing the spelling (e.g. fix typos, capitalize first letter, add punctuation at the end)
> - Fixing the grammar (if necessary)
> - If the sentence is in direct speech (e.g. contains the word "you"), rewrite it to be in third person.
> - DO NOT CHANGE ANYTHING ELSE. Word choices etc must remain authentic to the original.
>
> The sentence is: {sentence}
> Don't return anything other than the corrected sentence.

The following is the prompt for cleaning endings:

> Correct the following sentence by fixing the following:
> - Fixing the spelling (e.g. fix typos, capitalize first letter, add punctuation at the end)
> - Fixing the grammar (if necessary)
> - The sentence follows this sentence: "{sentence}" and should flow coherently.
> - No other stylistic changes are allowed.
> The sentence is: {ending}
> Don't return anything other than the corrected sentence. If no changes are necessary, just return the original sentence again.

B.2 Precontext Generation

We use GPT-4o for generating precontexts. The used prompt is as follows:

> Take the following ambiguous sentence, which is the fourth sentence in a 5-sentence short text:
> {sentence}
> In this sentence, the word {homonym} can mean either "{word sense 1}" or "{word sense 2}", depending on the context.
> This sentence is the fourth sentence in a story (out of five total). Write the first three sentences of the story. The first three sentences should serve as an introduction to the story which explains the circumstances of the current situation. Try to use simple sentences. Avoid complicated structures, long sentences and unnecessary information.
> Importantly, the word's meaning is still rather ambiguous even with the context. Both meanings should still be equally as plausible.
> Return these first three sentences, do not return anything else.

B.3 Prediction

We use the following prompt for predicting labels from all LLMs. It is an edited version of the guidelines seen in Appendix A.2.3. Pictured is the few-shot prompt; the zero-shot prompt is identical except for the omission of examples.

> You will see a short text in which one sentence is marked with "**". That sentence contains a word that can typically take on multiple different meanings, depending on the context. One of those meanings is given to you.
> **Your task is simple: Annotate how plausible a meaning of a word is in the context of the short text using one of five scores:**
> * **1**: The displayed meaning is not plausible at all given the context.
> * **2**: The displayed meaning is theoretically conceivable, but less plausible than

other meanings.
* **3**: The displayed meaning represents one of multiple, similarly plausible interpretations.
* **4**: The displayed meaning represents the most plausible interpretation; other meanings may still be conceivable.
* **5**: The displayed meaning is the only plausible meaning given the context.
There will be times where there is no objectively correct answer. Whatever the case, always look at all of the sentences and carefully think about how plausible each meaning would be.
Take a look at the following examples.
———
The bat flew out of the cave. In this context, how plausible is it that the meaning of the word "bat" is "A sports implement for hitting balls (e.g. in baseball)"?
Correct answer: 1
—
The letter specified where to meet him. **So after reading it, I went to the bank.**
In this context, how plausible is it that the meaning of the word "bank" is "a financial institution"?
Correct answer: 3
—
The composer often spontaneously had ideas for new melodies. **She writes notes on a sheet of paper.** She can later turn these into a piece.
In this context, how plausible is it that the meaning of the word "notes" is "a brief written record; a memo"?
Correct answer: 2
—
Mr Ellis walked to the town square with a big smile. He was getting ready to paint. **Whenever he sets up his easel in the town square, he always draws a crowd.** His painting of a flower looked really realistic!"
In this context, how plausible is it that the meaning of the word "draws" is "to attract; direct towards itself"?
Correct answer: 5
———
Now take a look at the following text: pre-context **sentence** ending

In this context, how plausible is it that the meaning of the word "{word}" is "{word sense}"?
Return only the numbered score (1, 2, 3, 4 or 5). Do not return anything else!

C Annotation Interface

Screenshots of our annotation interface can be found in Figure 2, 3 and 4.

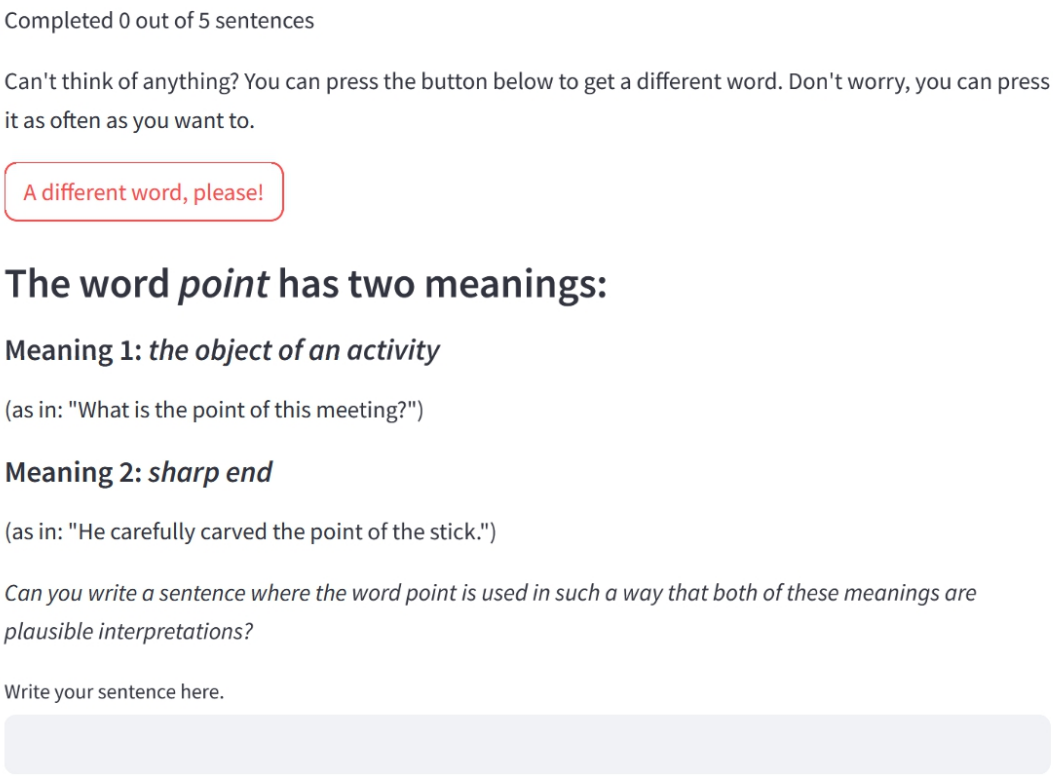

Figure 2: Interface for collecting ambiguous sentences.

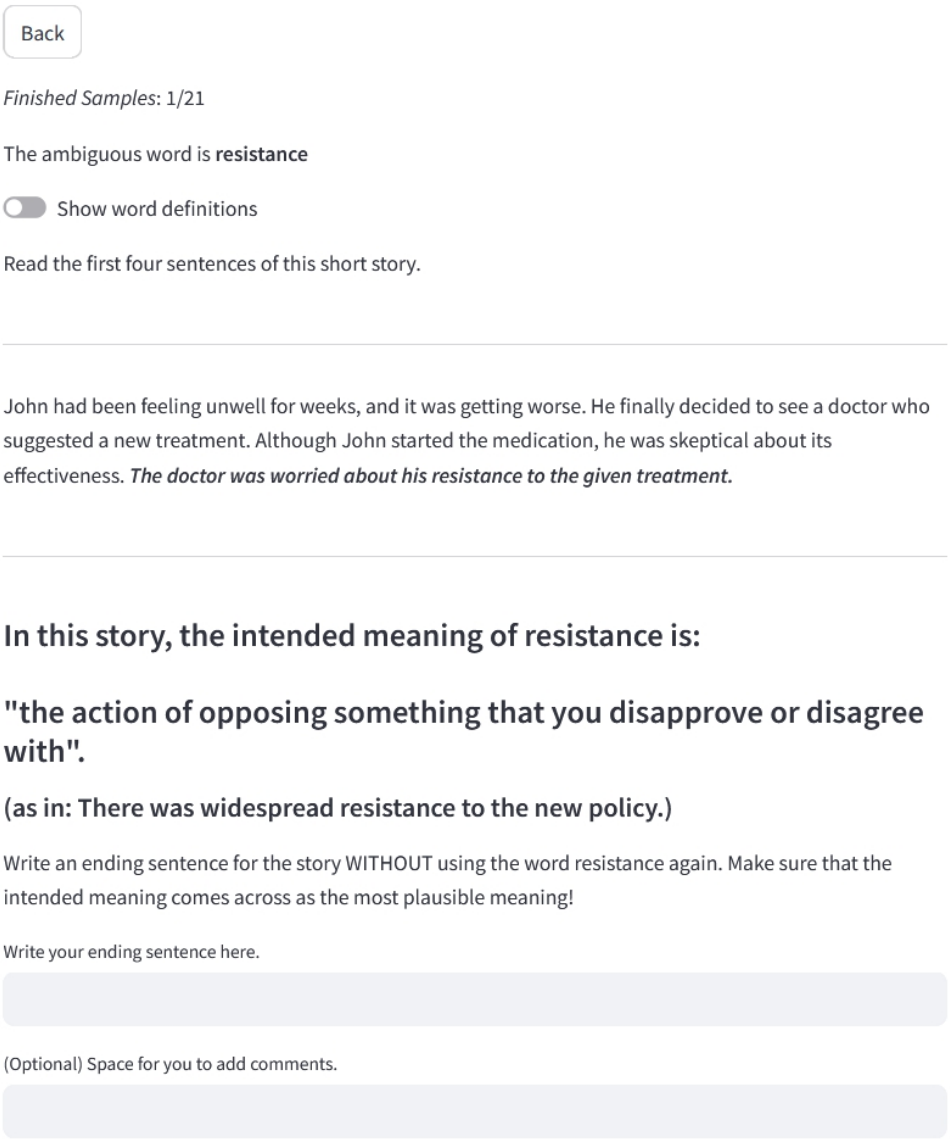

Figure 3: Interface for collecting endings.

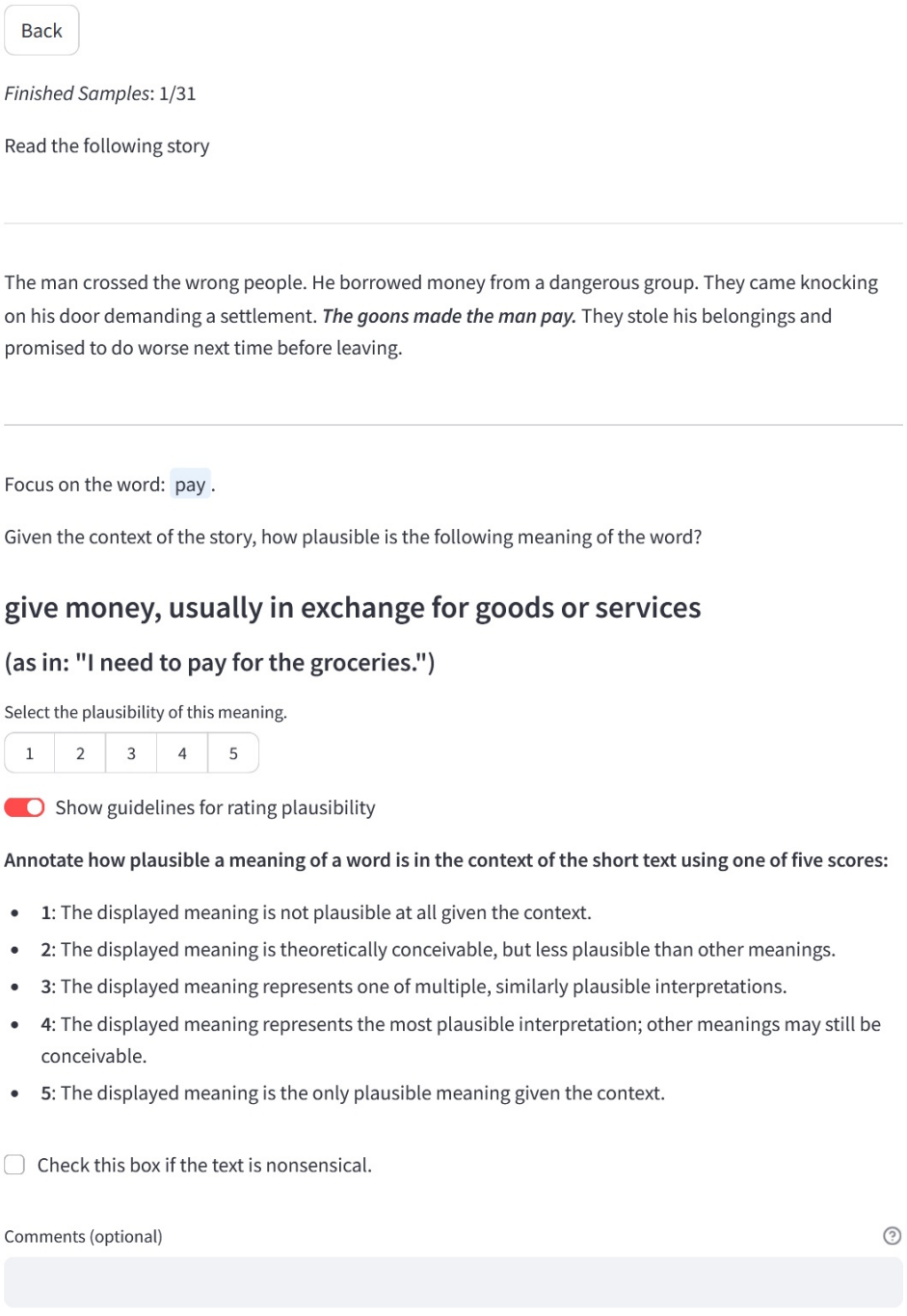

Figure 4: Interface for collecting plausibility ratings.