# Modal Sense Disambiguation in Language Models: The Role of Context, Training Domain, and Input Frequency

**Ezgi Başar** **Yevgen Matusevych** **Annemarie van Dooren** **Arianna Bisazza**
Center for Language and Cognition (CLCG), University of Groningen
{e.basar, yevgen.matusevych, a.m.f.van.dooren, a.bisazza}@rug.nl

## Abstract

Modals such as *may* and *must* can take on different senses depending on discourse context, alternating between root interpretations (e.g., permission, obligation) and epistemic interpretations (e.g., knowledge, belief). While modals are known to be notoriously difficult for children to acquire, limited research exists on whether language models (LMs) can reliably distinguish between different modal senses. We evaluate seven LMs that vary in architecture, size, and training data type using vector-based similarity and surprisal measures across four datasets. Specifically, we investigate whether LMs can leverage contextual information to disambiguate modal senses and whether their performance reflects the distribution of modal senses in their training data. We show that larger LMs trained on diverse text successfully leverage additional contextual information when available. In the absence of context, they tend to favor the dominant sense based on corpus statistics. Smaller developmentally plausible LMs, including those trained on child-directed language (CDL), fail to show consistent sensitivity to modal sense distinctions and do not benefit from context to the same degree.[1]

## 1 Introduction

A sentence like "Alex must be at school" can be interpreted in two distinct ways (Figure 1). In the first case, we can imagine a situation in which Alex's parents do not permit him to skip school, hence conveying a sense of obligation. In the other case, we can picture a parent entering Alex's room in the morning and finding it empty, inferring that Alex is likely at school. Here, the same sentence expresses a belief about where Alex is. This type of ambiguity is common with modals such as *may* and *must*, whose interpretation depends on the linguistic and non-linguistic context. Linguists distinguish two broad categories of modal senses, or "flavors". Root sense expresses notions that have to do with permission, obligation and ability, as in our first example. Epistemic sense expresses notions related to knowledge and belief, as exemplified in the second scenario.


Figure 1: An example of modal sense disambiguation.

Ambiguous modal expressions are ubiquitous in natural language. Roughly half of the world's languages feature a modal expressing root and epistemic senses (van der Auwera et al., 2005). Nevertheless, modals are notoriously complex for children to acquire due to their context-dependency and lack of physical grounding (Cournane and Pérez-Leroux, 2020; van Dooren et al., 2022).

In natural language processing (NLP), distinguishing between modal senses can be seen as a subtype of the more general task of word sense disambiguation (Marasović and Frank, 2016). Word sense disambiguation systems have long been developed to grapple with the polysemy observed in natural language. Modals add an extra layer of complexity for these systems, because it is possible for identical sentences to receive different interpretations based on the conversational context (Kratzer, 1991). This property makes modals highly challenging for NLP systems, including language models (LMs).

In this study, we ask whether and how various LMs distinguish between different modal senses

[1] Code is available at https://github.com/ezgibasar/modal-sense-disambiguation-in-lms.

in English. More specifically, we address three questions: **(1)** whether additional conversational context aids LMs in disambiguating modal senses, **(2)** whether LMs' discrimination of modal senses generalizes to out-of-domain evaluation data or is confined to their training domain, and **(3)** whether an LM's performance reflects the distribution of modal senses in its training input.

To answer these questions, we consider seven LMs of varying architectures, sizes and training data types. Alongside LMs pretrained on large amounts of diverse text data, we evaluate smaller LMs pretrained on child-directed language (CDL) for a more developmentally plausible reference point, including both masked and causal architectures. We evaluate these LMs on our adaptations of three existing datasets (Zhou et al., 2015; Tulling et al., 2021; van Dooren et al., 2017). We first probe vector representations of modals to investigate if LMs have distinct representations for modal senses. We then use surprisal measures in a minimal-pair setup to assess whether LMs show a preference for one modal sense over another.

This study provides a comprehensive analysis of LMs' ability to discriminate between modal senses, accounting for both larger and smaller developmentally plausible LMs. Furthermore, we contribute adapted evaluation datasets in which the availability of contextual cues is manipulated, enabling a controlled investigation into the role of context in modal sense disambiguation. Finally, we explore whether input frequency affects LMs' behavior, offering a potential bridge between computational modeling and linguistic theories of modal acquisition.

## 2 Background

### 2.1 Modal senses in NLP

Prior work involving modal senses has largely focused on annotation and classification. Ruppenhofer and Rehbein (2012) developed an annotation scheme for English modals distinguishing root and epistemic uses. Rubinstein et al. (2013) created a more fine-grained framework capturing sense, strength, evidence, and attitude. Zhou et al. (2015) explored paraphrase-driven sense projection and semantically enriched classification models, achieving strong performance across genres. Li et al. (2019) demonstrated that basic SVM classifiers improve when word embedding features are used for modal sense classification. Dehouck and Denis (2023) revisited the task using contextual word embeddings, employing BERT as a feature extractor for the token associated with the modal. Marasović and Frank (2016) used a convolutional neural network architecture for modal sense classification and identified useful features such as past versus non-past tense, dynamic versus stative verbs, passive constructions, negation, and telic clauses. Colli et al. (2024) presented the first modal sense classification work for French, fine-tuning a range of BERT variants on annotated sentences with augmented context. Notably, these studies are all classification tasks relying on supervised learning from labeled samples rather than examining the model-internal representations of pretrained LMs.

Most relevant to our work is Wagner and Zarrieß (2023), who investigated BERT's ability to encode sentence modality and modal sense across varieties of English. They trained probing classifiers on contextualized embeddings of modals and sentence embeddings, also testing BERT's ability to predict masked modals. However, probing classifiers can only reveal what information is extractable from a representation, and they do not provide direct evidence about whether the model uses that information (Hewitt and Liang, 2019); therefore, we adopt an alternative methodology that combines vector similarity probing with surprisal-based measures.

It is worth noting that similar methods have been employed in previous word sense disambiguation research. Nair et al. (2020) demonstrate that distances between sense centroids in contextualized vectors significantly correlate with human judgments, providing support that vector-based similarity can capture human-like meaning distinctions. Moreover, Temerko et al. (2025) report that surprisal scores correlate with human acceptability ratings for attested polysemes. Given our focus on developmental plausibility, these methods are also better aligned with our objectives than probing classifiers.

To our knowledge, no prior work has systematically compared causal and masked LMs for modal sense disambiguation. Causal LMs, whose autoregressive objective more closely approximates the incremental nature of human language processing (Goldstein et al., 2022), remain underexplored in this context. Our work addresses this gap by probing the vector representations and predictive behavior of a range of LMs, including both

causal and masked architectures. We also examine CDL-trained LMs to investigate whether and how LMs learn different modal sense representations from more developmentally plausible input.

### 2.2 Acquisition of modals

The acquisition of modals presents well-documented difficulties for children, especially for the epistemic sense. Several explanations for this challenge have been proposed in the developmental literature. One account appeals to Theory of Mind (ToM), suggesting that epistemic modals require the ability to attribute mental states to others, which develops around age four (Papafragou, 1998). However, studies using non-verbal measures have found evidence of ToM-related abilities in infants as young as 15 months (Onishi and Baillargeon, 2005), and two-year-olds appropriately use epistemic adverbs like *maybe* (O'Neill and Atance, 2000), suggesting that conceptual capacity for modal notions may be available earlier.

Another account focuses on input frequency. The distribution of modal senses in child-directed English differs from what is attested in non-CDL corpora, and researchers have argued that this discrepancy may influence how children acquire modals. Accordingly, van Dooren et al. (2022) show that the epistemic modal sense, which emerges late in children's acquisition, is highly infrequent in children's input.

The frequencies of the different modal senses have also been shown to vary depending on the specific modal. In American English corpora consisting mostly of written text, *may* appears as epistemic 83% of the time and as root only 17% of the time (Collins, 2007). For *must*, the pattern reverses, with root uses accounting for 84% of occurrences and epistemic uses only 16% (Hacquard and Wellwood, 2012). The Manchester corpus (Theakston et al., 2001), as sample of CDL corpora, reveals a contrasting distribution. In this case, *may* occurs as root 56% of the time and as epistemic 44% of the time, whereas *must* occurs as epistemic 65% of the time and as root 35% of the time (van Dooren et al., 2022). These diverging distributions also provide a basis for testing whether LMs trained on different input types develop corresponding frequency-based preferences.

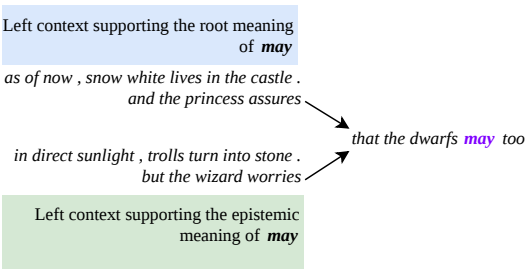



Figure 2: Example data pair from the validated dataset.

## 3 Evaluation Data

There are no standardized benchmarks for modal sense disambiguation. For evaluation purposes, we consider four datasets, which we refer to as EPOS, validated, naturalistic, and naturalistic with context. The datasets differ in their source material, size, and the availability of disambiguating context, making them complementary to each other. The EPOS dataset is the largest of all. The validated dataset originates from a highly controlled neurolinguistic experiment, allowing us to use carefully constructed stimuli in pairwise comparisons. The naturalistic dataset, drawn from an annotated CDL corpus, lets us examine how results differ when LMs are presented the kind of disambiguation tasks that children face. Finally, its variant with retrieved prior context further allows us to assess whether the addition of naturally occurring context affects LM performance. Example data points from the validated dataset can be seen in Figure 2. Appendix A provides further examples for each dataset, along with the number of instances per modal class.

### 3.1 EPOS

The publicly available EPOS dataset (Zhou et al., 2015) is constructed through paraphrase-driven modal sense projection on the Europarl (Koehn, 2005) and OpenSubtitles corpora (Tiedemann, 2012), which feature European Parliament proceedings and film subtitles, respectively. Rather than relying on manual annotation, the authors assign labels heuristically using paraphrase patterns. While this approach introduces some noise, it also provides a means of obtaining a sizable corpus for modal sense classification. The dataset exhibits some imbalance, containing 134 root and 875 epistemic *may* instances, alongside 417 root and 637 epistemic instances of *must*. Several existing studies on modal sense classification have

used this dataset (Marasović and Frank, 2016; Li et al., 2019; Dehouck and Denis, 2023; Wagner and Zarrieß, 2023).

### 3.2 Human-validated

This dataset constructed by Tulling et al. (2021) for their neurolinguistic experiments consists of 160 stimuli: 40 instances each of *must* in a root context, *must* in an epistemic context, *may* in a root context, and *may* in an epistemic context. Twenty-five adult participants (aged 19–52) validated each sentence to be associated with the intended modal sense. Each test item consists of a context sentence supporting either an epistemic or root interpretation, followed by the target sequence containing the modal (see example in Figure 2). The target sequence features an embedded clause within a matrix clause linking the modal's interpretation to a third-person perspective to further bias the intended reading. Importantly, the embedded clause itself (introduced by "that") is kept identical across root and epistemic conditions, while the matrix verb varies. The target sequence is identical across each root-epistemic pair, allowing for a more controlled evaluation setting.

### 3.3 Naturalistic

The naturalistic data are drawn from an annotated subset of the CHILDES database (MacWhinney, 2000). Van Dooren et al. (2022) extract modal productions from the Manchester corpus (Theakston et al., 2001), which consists of recordings from 12 child–mother dyads during play sessions (ages 2;00 to 3;00), and annotate each instance for root or epistemic sense. From this dataset, we use 296 instances of *must* (88 root, 208 epistemic) and 32 instances of *may* (16 root, 16 epistemic), restricting our set to the utterances we could locate in the database.

We additionally create a version of this dataset with conversational context. We utilize childes-db (version 2018.1; Sanchez et al., 2019) to retrieve from the transcripts the preceding utterances for each target sentence. In each case, we extract up to five prior adult utterances from the same transcript file. The average utterance length in the corpus is 3.89 words, so retrieving five previous utterances yields a context window of approximately 20 words. When fewer than five prior adult utterances are available, we include all that are present.

## 4 Methods

We evaluate seven pretrained LMs that vary in architecture, size, and training data type. We first test each LM's ability to discriminate between epistemic and root senses of *may* and *must* using vector-based similarity probing. We then analyze LM preferences for one modal sense over the other using surprisal-based measures in a test design analogous to the standard minimal-pair setup.

### 4.1 Probing vector representations

We adopt the word sense disambiguation method of Cabiddu et al. (2023), who use sense prototypes derived from contextualized embeddings to disambiguate polysemous words (e.g., distinguishing the elastic band meaning of *band* from the music group meaning) and adapt it to our use case of distinguishing epistemic and root modal senses. We divide each of our datasets into a prototype set and an evaluation set using stratified 5-fold cross-validation. For every combination of modal (*must*, *may*) and sense (epistemic, root), we extract prototypical vector representations from the prototype set. Given an occurrence of the modal in a particular sense, we extract its contextualized representation from the model and sum the representations from the last four hidden layers, following Cabiddu et al. (2023) and Loureiro et al. (2021). These vectors are then averaged across training instances for each sense to obtain a centroid.

During evaluation, we again extract the contextualized vector for each test instance of the target modal by summing the last four hidden layers. We compute the cosine similarity between this test vector and the two sense prototypes (epistemic and root). The predicted sense is taken to be the prototype with higher similarity. Accuracy is computed as the percentage of test instances where the predicted sense matches the ground-truth label. As a shorthand, we refer to tests performed using this method as the *similarity method*.

Intuitively, this method differs from a typical probing classifier setup (Ettinger et al., 2016; Alain and Bengio, 2017) in that it directly compares distances between vectors in a model's representation space.

### 4.2 Surprisal-based measures

Surprisal is a standard measure for estimating the likelihood of a stimulus under a given model. We compute token-level surprisal as the negative log

probability of a token given its preceding context. For masked LMs, we use pseudo-surprisal scores obtained by successively masking tokens following the widely used method by Salazar et al. (2020). Lower surprisal values indicate that a model finds a token more predictable in its context, and higher values indicate that the token is less predictable.

Similar to a standard minimal-pair setup, we use this measure to evaluate sentence pairs where the target sequences are identical and the preceding context disambiguates the intended sense (see Figure 2). For each model, we compute the surprisal of the modal token in both the root and epistemic context conditions. A lower surprisal value for one condition relative to the other is interpreted as a preference for that modal sense. No subword aggregation was required for our surprisal estimates, as none of the LMs split the modal tokens across multiple pieces. Because this method requires paired data points that differ only by a root versus epistemic context while keeping the target sequence identical, we perform surprisal-based evaluation only on the validated dataset.

### 4.3 Models

We evaluate a total of seven LMs with different architectures, sizes, and training data sources. Our first group of LMs consists of base models pretrained on relatively large amounts of data. Pythia 6.9B (Biderman et al., 2023) is the largest LM we include, belonging to a suite of decoder-only open-source models whose training data are publicly available. We also include GPT-2 (Radford et al., 2019), an autoregressive decoder-only model, and RoBERTa (Liu et al., 2019), an optimized bidirectional encoder-only model. We use the pretrained base versions for all three LMs.

To provide a more developmentally plausible reference point, we also employ LMs released by Padovani et al. (2025) which have been trained exclusively on CDL. These are RoBERTa and GPT-2 variants trained from scratch on the AO-CHILDES corpus (Huebner and Willits, 2021), which contains approximately 5 million words of American English CDL. We refer to these as CHILDES RoBERTa and CHILDES GPT-2. Given that the distribution of modal senses can differ for CDL compared to general text, we include these LMs because they are trained exclusively on CDL without additional data sources.

We also use RoBERTa and GPT-2 models trained on a Wikipedia corpus matching the size of the CDL data, also released by Padovani et al. (2025). Testing these Wikipedia RoBERTa and Wikipedia GPT-2 models allows us to isolate the effect of training data type (CDL vs. general text) from the effect of corpus size. No fine-tuning is performed on any of the listed LMs.

## 5 Results and Discussion

### 5.1 Use of additional context

The first question we address is whether additional discourse context helps LMs distinguish between root and epistemic meanings of modals. We expect LMs to perform better when evaluated on datasets that include prior context, under the assumption that the preceding sentence can help disambiguate the modal in the target utterance. We further expect the validated dataset to be the easiest among all evaluation sets, given that it features context sentences carefully designed to bias people toward a particular modal reading. In contrast, the EPOS dataset and the naturalistic dataset without context present the modal sentence in isolation without any prior context to guide interpretation. It should be noted that for the naturalistic dataset with context, we simply retrieve preceding utterances to provide context. Unlike the carefully designed stimuli in the validated dataset, these may not necessarily contain the relevant information to disambiguate modal senses.

Based on the results in Table 1, the availability of adequate prior context is generally associated with improved performance. LMs tend to perform substantially better on the validated dataset than on the EPOS dataset or the naturalistic dataset without context, with most LMs achieving accuracies above 90%. That being said, this pattern only holds for the larger LMs. Our set of developmentally plausible LMs exhibit near-chance results on the validated dataset, suggesting that they may not have learned modal representations sufficiently robust to take advantage of the disambiguating context.

To investigate how context affects results, we also compare model performance on the naturalistic dataset with and without prior context. For GPT-2, *may* epistemic improves from 27% to 57%, and for Pythia from 75% to 80%. Likewise, *must* epistemic shows an improvement from 26% to 75% for GPT-2 and from 58% to 77% for Pythia. We observe that the inclusion of contex-

| | | | May | | | Must | | |
|---|---|---|---|---|---|---|---|---|
| | | | Root | Epistemic | All | Root | Epistemic | All |
| Masked LMs | RoBERTa (110M) | EPOS | 0.95 | 0.99 | 0.98 | 0.95 | 0.87 | 0.90 |
| | | Validated | 1.00 | 1.00 | 1.00 | 0.93 | 0.95 | 0.94 |
| | | Naturalistic | 0.72 | 1.00 | 0.85 | 0.86 | 0.86 | 0.86 |
| | | Naturalistic (with context) | 0.67 | 1.00 | 0.82 | 0.89 | 0.96 | 0.94 |
| | CHILDES RoBERTa (12.7M) | EPOS | 0.91 | 0.86 | 0.86 | 0.70 | 0.80 | 0.76 |
| | | Validated | 0.42 | 0.55 | 0.49 | 0.55 | 0.55 | 0.55 |
| | | Naturalistic | 0.83 | 0.80 | 0.82 | 0.81 | 0.84 | 0.83 |
| | | Naturalistic (with context) | 0.77 | 1.00 | 0.88 | 0.83 | 0.92 | 0.90 |
| | Wikipedia RoBERTa (12.7M) | EPOS | 0.85 | 0.95 | 0.94 | 0.68 | 0.84 | 0.77 |
| | | Validated | 0.60 | 0.28 | 0.44 | 0.47 | 0.28 | 0.38 |
| | | Naturalistic | 0.72 | 0.80 | 0.76 | 0.83 | 0.91 | 0.89 |
| | | Naturalistic (with context) | 0.72 | 0.93 | 0.81 | 0.83 | 0.94 | 0.91 |
| Causal LMs | Pythia (6.9B) | EPOS | 0.27 | 0.99 | 0.89 | 0.98 | 0.19 | 0.50 |
| | | Validated | 0.90 | 1.00 | 0.95 | 0.93 | 0.93 | 0.93 |
| | | Naturalistic | 0.53 | 0.75 | 0.64 | 0.64 | 0.58 | 0.60 |
| | | Naturalistic (with context) | 0.72 | 0.80 | 0.76 | 0.79 | 0.77 | 0.78 |
| | GPT-2 (124M) | EPOS | 0.26 | 0.99 | 0.89 | 0.99 | 0.08 | 0.44 |
| | | Validated | 0.97 | 0.95 | 0.96 | 0.97 | 0.97 | 0.97 |
| | | Naturalistic | 0.60 | 0.27 | 0.42 | 0.80 | 0.26 | 0.42 |
| | | Naturalistic (with context) | 0.73 | 0.57 | 0.64 | 0.69 | 0.75 | 0.73 |
| | CHILDES GPT-2 (14.8M) | EPOS | 0.51 | 0.93 | 0.88 | 0.79 | 0.74 | 0.76 |
| | | Validated | 0.70 | 0.68 | 0.69 | 0.68 | 0.60 | 0.64 |
| | | Naturalistic | 0.73 | 0.68 | 0.70 | 0.67 | 0.77 | 0.74 |
| | | Naturalistic (with context) | 0.78 | 0.75 | 0.76 | 0.76 | 0.71 | 0.73 |
| | Wikipedia GPT-2 (14.8M) | EPOS | 0.60 | 0.98 | 0.93 | 0.77 | 0.76 | 0.76 |
| | | Validated | 0.75 | 0.72 | 0.74 | 0.70 | 0.62 | 0.66 |
| | | Naturalistic | 0.65 | 0.70 | 0.67 | 0.73 | 0.74 | 0.74 |
| | | Naturalistic (with context) | 0.58 | 0.57 | 0.57 | 0.62 | 0.63 | 0.63 |

Table 1: Modal sense disambiguation accuracy by different LMs, based on the vector similarity method (cf. Section 4.1). The red-green color gradient represents performance from low (red) to high (green), with the 0.50 midpoint marking random chance levels. Parameter counts are indicated in indicated in parentheses for each LM. See Appendix C for more detailed representations of the results.

tual cues improves performance for causal LMs. Masked LMs similarly benefit from additional left context, although the improvement is less marked. Notably, RoBERTa shows *must* epistemic results improving from 86% to 96%, while *may* root decreases from 72% to 67%. Appendix D offers a supplementary investigation into the usefulness of different contextual cues.

Looking at these results, we can additionally note an architectural difference. Causal LMs show lower performance on datasets that present greater difficulty for modal sense disambiguation, namely the EPOS dataset and the naturalistic data without context. Masked LMs, in contrast, show more uniform performance across all datasets. We attribute this difference to the bidirectional context available to masked LMs. Van Dooren et al. (2022), building on insights from Condoravdi (2002), hypothesize that aspectual cues like *must be* or *may have* (where the words *be* or *have* appear in the right context of the target modal) can be particularly helpful when disambiguating modal senses. Having the ability to leverage such information could explain the difference in model behaviors between causal and masked LMs.

### 5.2 Effect of out-of-domain data

We further examine whether LMs trained on data from the same domain as the test data show better ability to discriminate between modal senses. Our domain of interest here is CDL, and we expect the evaluation on naturalistic data (which originates from CHILDES) to be more challenging for LMs not trained on CDL, leading to lower performance and more varied behavior. Conversely, we expect CDL-trained LMs to perform better on the naturalistic dataset than other evaluation sets. Since the EPOS dataset does not have prior context available, we are interested in a comparison with no-context variant of the naturalistic dataset.

As Table 1 illustrates, we observe that GPT-2 scores as low as 27% for epistemic *may* and 26% for epistemic *must* on the naturalistic dataset. This low performance aligns with the expectation that out-of-domain data poses challenges for LMs not trained on CDL. Pythia also shows lower performance on naturalistic datasets, with scores between 53% and 80%. For RoBERTa, the naturalistic dataset likewise proves slightly more difficult than EPOS or the validated dataset.

CDL-trained LMs, however, do not exhibit the expected in-domain benefit. CHILDES GPT-2 achieves its highest overall scores on the EPOS dataset, with accuracies of 88% for *may* and 76% for *must*, compared to 70% and 74% on the naturalistic dataset. CHILDES RoBERTa performs slightly better on EPOS for *may*, scoring 86% versus 82% on naturalistic data, while the pattern reverses for *must*, with 76% on EPOS against 83% on naturalistic data. Notably, both CDL- and Wikipedia-trained LMs produce similar results to each other while diverging from larger LMs like RoBERTa, GPT-2, and Pythia. This suggests that model size, rather than training data domain, is a larger contributor to the observed differences between the smaller developmentally plausible LMs and those that are larger.

### 5.3 Role of input frequency

Finally, we investigate whether LMs' ability to distinguish between root and epistemic uses of *may* and *must* reflects the distribution of these senses in their training data. Here, we consider both the vector-probing and surprisal-based evaluation. If LM behavior aligns with input frequency, LMs should perform better on the more frequent sense of each modal under the similarity method and show a preference for that sense under the surprisal method.

For LMs trained on general text, we assume their training data matches attested corpus statistics, where *may* is predominantly epistemic (83%; Collins, 2007) and *must* is predominantly root (84%; Hacquard and Wellwood, 2012). For CDL-trained LMs, we extrapolate from the Manchester corpus, where *may* is root 56% of the time and *must* is epistemic 65% of the time (van Dooren et al., 2022). For Wikipedia-trained LMs, we assume their training data aligns more closely with general corpus statistics, as Wikipedia consists of written encyclopedic text. Therefore, we expect non-CDL-trained LMs to perform better on epistemic *may* and root *must*, whereas CDL-trained LMs should perform better on root *may* and epistemic *must*. A breakdown of the corpus statistics and model-specific expectations can be found in Appendix B. We further reflect on our corpus-based assumptions in the limitations section.

Similarity method results in Table 1 mostly follow the expected pattern for non-CDL-trained LMs, particularly when evaluating on harder-to-disambiguate data. As noted earlier, the EPOS dataset and the naturalistic dataset without context are more likely to contain ambiguous modals, as they present the modal sentence in isolation without any preceding context. Lacking the discourse context needed to disambiguate the intended reading, evaluation on these datasets can uncover whether LMs rely on any internal biases when it comes to modal senses.

Looking at the EPOS dataset, causal LMs trained on general text can be said to show a clear alignment with input frequency. GPT-2 achieves near-ceiling performance on the dominant senses (99% for epistemic *may* and 99% for root *must*) while scoring at or below chance on the infrequent senses (26% for root *may* and 8% for epistemic *must*). Pythia follows a similar pattern, scoring 99% for epistemic *may* and 98% for root *must* versus 27% and 19% for the infrequent senses, respectively. Masked LMs trained on general text, specifically RoBERTa, likewise align with the expected pattern in most cases, though their results are very uniform across both senses, resulting in higher overall performance. This uniformity suggests that input frequency may have a less prominent effect on the behavior of masked LMs.

Turning to the validated dataset, we can observe that non-CDL-trained LMs do not always show the expected pattern. GPT-2 scores 97% for root *may*, 95% for epistemic *may*, and 97% for both senses of *must*. This indicates that when evaluation data includes helpful disambiguating context, LMs can take advantage of it. The correlation with input frequency can be observed only when such context is absent.

Two exceptions to this pattern are worth noting. First, on the naturalistic dataset without context, GPT-2 does not maintain the expected pattern for *may*. Performance on epistemic *may* drops to 27%, though performance on root *must* remains strong at 80%. This variability may stem from naturalistic data being out of domain for GPT-2, or from low *may* counts in the evaluation data pro-

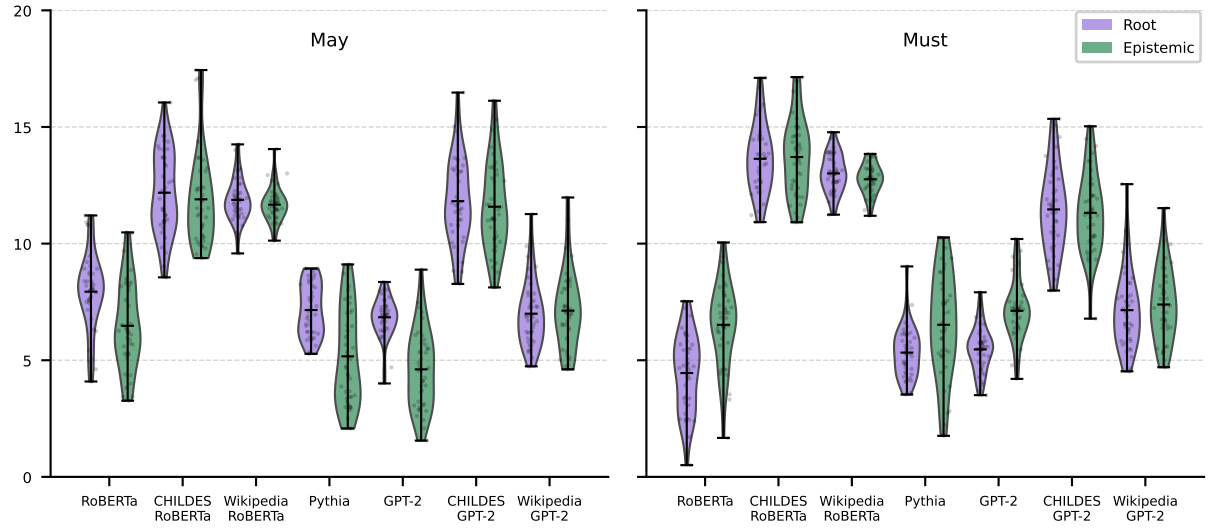



Figure 3: Surprisal score distributions for the modals in validated data. Lower surprisal values for a given sense are interpreted as a preference for that sense.

ducing variable outcomes. Second, Wikipedia-trained LMs do not consistently exhibit the expected pattern despite being trained on general text. Their scores more closely resemble those of CDL-trained LMs than those of larger LMs trained on general text.

For LMs trained on CDL, we do not consistently observe the opposite pattern that we expect based on the Manchester corpus statistics. On the EPOS dataset, CHILDES GPT-2 scores 51% for root *may*, 93% for epistemic *may*, 79% for root *must*, and 74% for epistemic *must*, while CHILDES RoBERTa scores 91%, 86%, 70%, and 80%, respectively. On the validated dataset, developmentally plausible LMs tend to perform near chance levels for the different modal senses. We can therefore note that the developmentally plausible LMs are not consistently better at one modal sense over the other on any dataset, suggesting that they may not have encountered enough modal examples in training to develop a sensitivity in either direction.

To explore whether frequency also affects LMs' preferences, we also consider the surprisal-based results in Figure 3. Consistent with our similarity-based results, we can note that the surprisal distributions on the validated dataset differ between larger LMs and developmentally plausible ones. Non-CDL-trained larger LMs show the expected preference for epistemic *may* and root *must*. Both CDL-trained and Wikipedia-trained smaller LMs lack a clear preference for either sense of *must* or *may*, with average surprisal values remaining roughly equal across the two senses. Given that our developmentally plausible LMs have very similar surprisal distributions for both epistemic and root modals, we can argue that these smaller LMs have not been able to develop distinct representations for different modal senses, consistent with our similarity-based results.

## 6 Conclusion

This study investigates whether LMs can distinguish between root and epistemic senses of the English modals *may* and *must*. We ask whether providing additional conversational context helps LMs discriminate between modal senses, whether LMs trained on data from the same domain as evaluation data show better discrimination ability, and whether LM behavior correlates with the estimated distribution of modal senses in their training data. We evaluate seven LMs that vary in architecture, size, and training data type using vector-based similarity probing and surprisal measures across different datasets.

We show that providing additional context substantially improves the performance of larger LMs, whereas performance is lower when they are evaluated on ambiguous modal expressions without contextual cues to resolve ambiguity. This suggests that larger LMs are capable of leveraging contextual information to override existing biases if such information is present.

Furthermore, we find that LM size affects

modal sense discrimination more than training domain. Smaller developmentally plausible LMs (both CDL-trained and Wikipedia-trained) perform worse than larger LMs. Critically, if training domain were the key factor, Wikipedia-trained LMs would pattern with larger LMs trained on general text and not with their CDL-trained counterparts, as we observe.

Looking at whether LMs' performance aligns with input frequencies, we observe that larger LMs trained on general text perform better on the dominant sense of each modal under the similarity method and show a preference for that sense under the surprisal method. This alignment with corpus statistics only emerges when evaluation data lack disambiguating context, indicating that LMs can rely on general biases when contextual cues are absent.

Overall, our results offer insights into how LMs represent modal senses. We see that sufficiently large LMs can disambiguate modals through frequency biases or contextual cues, whereas smaller LMs trained on limited input fail to acquire reliable sense distinctions. More broadly, our findings bridge computational modeling and linguistic theories of modal acquisition by showing that two factors hypothesized to implicate children's learning, namely input frequency and the integration of contextual information, also shape LM behavior.

## Limitations

While our study provides a systematic investigation into how LMs distinguish between root and epistemic senses of *may* and *must*, several limitations can be acknowledged. First, we can note that our evaluations target only two English modals, *may* and *must*. We, therefore, cannot reliably comment on whether the observed patterns generalize to other modals such as *can*, *should*, and *could*, or to other languages.

The largest LM we assess is Pythia with 6.9 billion parameters. Whether larger LMs would behave differently, for instance by being able to disambiguate modal senses without rich contextual cues, remains an open question. Consequently, our conclusions about model size apply only within the parameter range we investigate.

Although we argue that our smaller developmentally plausible LMs may have low performance due to insufficient exposure to modal examples, this interpretation is inferential. We do not know the exact number of modal expressions that the various LMs have been exposed to. Consequently, our analyses rely on existing corpus studies rather than direct measurement to characterize frequency distributions in each LM's training data. For models with publicly available data, raw token counts of *may* and *must* would offer no insight into root versus epistemic distributions. Extracting sense-level frequencies would require annotating modal senses for sizeable datasets, which is a prohibitively large annotation effort. We can also note that any automated annotation would have to depend on the LMs we are evaluating. Some degree of extrapolation is therefore inevitable, and our choice of frequency proxies is grounded in existing linguistic literature and constrained by the availability of annotated datasets. Nevertheless, we acknowledge this concern and emphasize that our frequency-based arguments should be interpreted in light of this limitation.

Finally, our evaluation introduces a potential language variety confound. The CDL models are trained exclusively on American English, whereas the Manchester corpus is British English. We utilize this corpus for evaluations with naturalistic data and also for the modal sense statistics that guide our expectations about CDL-trained LM behavior. While it would have been preferable to have American English CDL frequency statistics for comparison, the only available CDL dataset annotated with modal sense is the Manchester corpus, which restricts us to this variety. However, this is not a confounder when comparing CDL-trained models to other LMs, as they are also predominantly trained on American English. We acknowledge this limitation while noting that Collins (2007) reports similar frequencies for root and epistemic *may* in adult American English and in British English, suggesting that the overall patterns may still hold.

## Acknowledgments

We thank the anonymous reviewers for their valuable feedback. This publication was supported by the Talent Programme of the Dutch Research Council (grant VI.Vidi.221C.009).

## References

Guillaume Alain and Yoshua Bengio. 2017. Understanding intermediate layers using linear classifier

probes. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Workshop Track Proceedings*.

Stella Biderman, Hailey Schoelkopf, Quentin Gregory Anthony, Herbie Bradley, Kyle O'Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, Usvsn Sai Prashanth, Edward Raff, Aviya Skowron, Lintang Sutawika, and Oskar Van Der Wal. 2023. Pythia: A suite for analyzing large language models across training and scaling. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 2397–2430. PMLR.

Francesco Cabiddu, Mitja Nikolaus, and Abdellah Fourtassi. 2023. Comparing Children and Large Language Models in Word Sense Disambiguation: Insights and Challenges. In *Proceedings of the 45th Annual Meeting of the Cognitive Science Society*.

Anna Colli, Diego Rossini, and Delphine Battistelli. 2024. A modal sense classifier for the French modal verb pouvoir. In *Proceedings of the Tenth Italian Conference on Computational Linguistics (CLiC-it 2024)*, pages 233–243, Pisa, Italy. CEUR Workshop Proceedings.

Peter Collins. 2007. Can/could and may/might in british, american and australian english: a corpus-based account. *World Englishes*, 26(4):474–491.

Cleo Condoravdi. 2002. Temporal interpretation of modals: Modals for the present and for the past. In David Beaver, Luis Casillas Martínez, Bardy Clark, and Stefan Kaufmann, editors, *The construction of meaning*, pages 59–88. CSLI, Stanford.

Ailís Cournane and Ana Teresa Pérez-Leroux. 2020. Leaving obligations behind: Epistemic incrementation in preschool english. *Language Learning and Development*, 16(3):270–291.

Mathieu Dehouck and Pascal Denis. 2023. Revisiting modal sense classification with contextual word embeddings. In *Models of Modals*, pages 225–253. De Gruyter.

Allyson Ettinger, Ahmed Elgohary, and Philip Resnik. 2016. Probing for semantic evidence of composition by means of simple classification tasks. In *Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP*, pages 134–139, Berlin, Germany. Association for Computational Linguistics.

Ariel Goldstein, Zaid Zada, Eliav Buchnik, Mariano Schain, Amy Price, Bobbi Aubrey, Samuel Nastase, Amir Feder, Dotan Emanuel, Alon Cohen, Aren Jansen, Harshvardhan Gazula, Gina Choe, Aditi Rao, Catherine Kim, Colton Casto, Lora Fanda, Werner Doyle, Daniel Friedman, and Uri Hasson. 2022. Shared computational principles for language processing in humans and deep language models. *Nature Neuroscience*, 25:369–380.

Valentine Hacquard and Alexis Wellwood. 2012. Embedding epistemic modals in English: A corpus-based study. *Semantics and Pragmatics*, 5(4):1–29.

John Hewitt and Percy Liang. 2019. Designing and interpreting probes with control tasks. 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 2733–2743, Hong Kong, China. Association for Computational Linguistics.

Philip A. Huebner and Jon A. Willits. 2021. Using lexical context to discover the noun category: Younger children have it easier. In *The Context of Cognition: Emerging Perspectives*, volume 75 of *Psychology of Learning and Motivation*, pages 279–331. Academic Press.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of Machine Translation Summit X: Papers*, pages 79–86, Phuket, Thailand.

Angelika Kratzer. 1991. *Modality*, pages 639–650. De Gruyter Mouton.

Bo Li, Mathieu Dehouck, and Pascal Denis. 2019. Modal sense classification with task-specific context embeddings. In *ESANN 2019 - 27th European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning*, Bruges, Belgium.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *Preprint*, arXiv:1907.11692.

Daniel Loureiro, Kiamehr Rezaee, Mohammad Taher Pilehvar, and Jose Camacho-Collados. 2021. Analysis and evaluation of language models for word sense disambiguation. *Computational Linguistics*, 47(2):387–443.

Brian MacWhinney. 2000. The childes project: Tools for analyzing talk (third edition): Volume i: Transcription format and programs, volume ii: The database. *Computational Linguistics*, 26(4):657–657.

Ana Marasović and Anette Frank. 2016. Multilingual modal sense classification using a convolutional neural network. In *Proceedings of the 1st Workshop on Representation Learning for NLP*, pages 111–120, Berlin, Germany. Association for Computational Linguistics.

Sathvik Nair, Mahesh Srinivasan, and Stephan Meylan. 2020. Contextualized word embeddings encode aspects of human-like word sense knowledge. In *Proceedings of the Workshop on the Cognitive Aspects of the Lexicon*, pages 129–141, Online. Association for Computational Linguistics.

Daniela K. O'Neill and Cristina M. Atance. 2000. 'Maybe my daddy give me a big piano': the development of children's use of modals to express uncertainty. *First Language*, 20(58):029–52.

Kristine H. Onishi and Renée Baillargeon. 2005. Do 15-month-old infants understand false beliefs? *Science*, 308(5719):255–258.

Francesca Padovani, Jaap Jumelet, Yevgen Matusevych, and Arianna Bisazza. 2025. Child-directed language does not consistently boost syntax learning in language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 19735–19756, Suzhou, China. Association for Computational Linguistics.

Anna Papafragou. 1998. The acquisition of modality: Implications for theories of semantic representation. *Mind & Language*, 13(3):370–399.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. *OpenAI Technical Report*.

Aynat Rubinstein, Hillary Harner, Elizabeth Krawczyk, Daniel Simonson, Graham Katz, and Paul Portner. 2013. Toward fine-grained annotation of modality in text. In *Proceedings of the IWCS 2013 Workshop on Annotation of Modal Meanings in Natural Language (WAMM)*, pages 38–46, Potsdam, Germany. Association for Computational Linguistics.

Josef Ruppenhofer and Ines Rehbein. 2012. Yes we can!? annotating English modal verbs. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 1538–1545, Istanbul, Turkey. European Language Resources Association (ELRA).

Julian Salazar, Davis Liang, Toan Q. Nguyen, and Katrin Kirchhoff. 2020. Masked language model scoring. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2699–2712, Online. Association for Computational Linguistics.

Alessandro Sanchez, Stephan C Meylan, Mika Braginsky, Kyle E MacDonald, Daniel Yurovsky, and Michael C Frank. 2019. childes-db: a flexible and reproducible interface to the child language data exchange system. *Behavior Research Methods*, 51(4):1928–1941.

Anna Temerko, Marcos Garcia, and Pablo Gamallo. 2025. A continuous approach to metaphorically motivated regular polysemy in language models. In *Proceedings of the 29th Conference on Computational Natural Language Learning*, pages 419–436, Vienna, Austria. Association for Computational Linguistics.

Anna Theakston, Elena Lieven, Julian Pine, and Caroline Rowland. 2001. The role of performance limitations in the acquisition of verb-argument structure: An alternative account. *Journal of Child Language*, 28:127–52.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in OPUS. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2214–2218, Istanbul, Turkey. European Language Resources Association (ELRA).

Maxime Tulling, Ryan Law, Ailís Cournane, and Liina Pylkkänen. 2021. Neural correlates of modal displacement and discourse-updating under (un)certainty. *eNeuro*, 8(1).

Johan van der Auwera, Andreas Ammann, and Saskia Kindt. 2005. *Modal polyfunctionality and Standard Average European*, chapter 11. University of Toronto Press.

Annemarie van Dooren, Anouk Dieuleveut, Ailís Cournane, and Valentine Hacquard. 2017. Learning what must and can must and can mean. In *Proceedings of the 21st Amsterdam Colloquium*, pages 225–234, Amsterdam, The Netherlands. ILLC. The 21st Amsterdam Colloquium.

Annemarie van Dooren, Anouk Dieuleveut, Ailís Cournane, and Valentine Hacquard. 2022. Figuring out root and epistemic uses of modals: The role of the input. *Journal of Semantics*, 39(4):581–616.

Jonas Wagner and Sina Zarrieß. 2023. Probing BERT's ability to encode sentence modality and modal verb sense across varieties of English. In *Proceedings of the 15th International Conference on Computational Semantics*, pages 28–38, Nancy, France. Association for Computational Linguistics.

Mengfei Zhou, Anette Frank, Annemarie Friedrich, and Alexis Palmer. 2015. Semantically enriched models for modal sense classification. In *Proceedings of the First Workshop on Linking Computational Models of Lexical, Sentential and Discourse-level Semantics*, pages 44–53, Lisbon, Portugal. Association for Computational Linguistics.

## A Dataset details

| Data | Flavor | Context | Target sequence | # |
|---|---|---|---|---|
| EPOS | Root | | i will now , if i **may** , read out the key points of the recommendations we adopted today . | 134 |
| | Epistemic | | this wording **may** sound a little far-fetched and need some explanation . | 875 |
| | Root | | we **must** be prepared for a long haul . | 417 |
| | Epistemic | | there **must** be a lower back injury as well . | 637 |
| Validated | Root | as of now , snow white lives in the castle . and the princess assures | that the dwarfs **may** too . | 40 |
| | Epistemic | in direct sunlight , trolls turn into stone . but the wizard worries | that the dwarfs **may** too . | 40 |
| | Root | as of now , snow white lives in the castle . and the princess declares | that the dwarfs **must** too . | 40 |
| | Epistemic | in direct sunlight , trolls turn into stone . but the wizard figures | that the dwarfs **must** too . | 40 |
| Naturalistic | Root | | you say **may** i borrow it please . | 16 |
| | Epistemic | | that one **may** not be strong enough for you . | 16 |
| | Root | | you **must** always cut it on the board like that . | 88 |
| | Epistemic | | they **must** have run away . | 208 |
| Naturalistic (with context) | Root | you should ask no it is caroline's and you ask | you say **may** i borrow it please . | 16 |
| | Epistemic | hm that is not that is not your boat darling your boat is here this is your boat this one | that one **may** not be strong enough for you . | 16 |
| | Root | that is it and that one no don't do it on your hands because if it was a real knife you would cut yourself doing that wouldn't you | you **must** always cut it on the board like that . | 88 |
| | Epistemic | i can't see them i think they have run away no they are not under there and they are not behind here | they **must** have run away . | 208 |

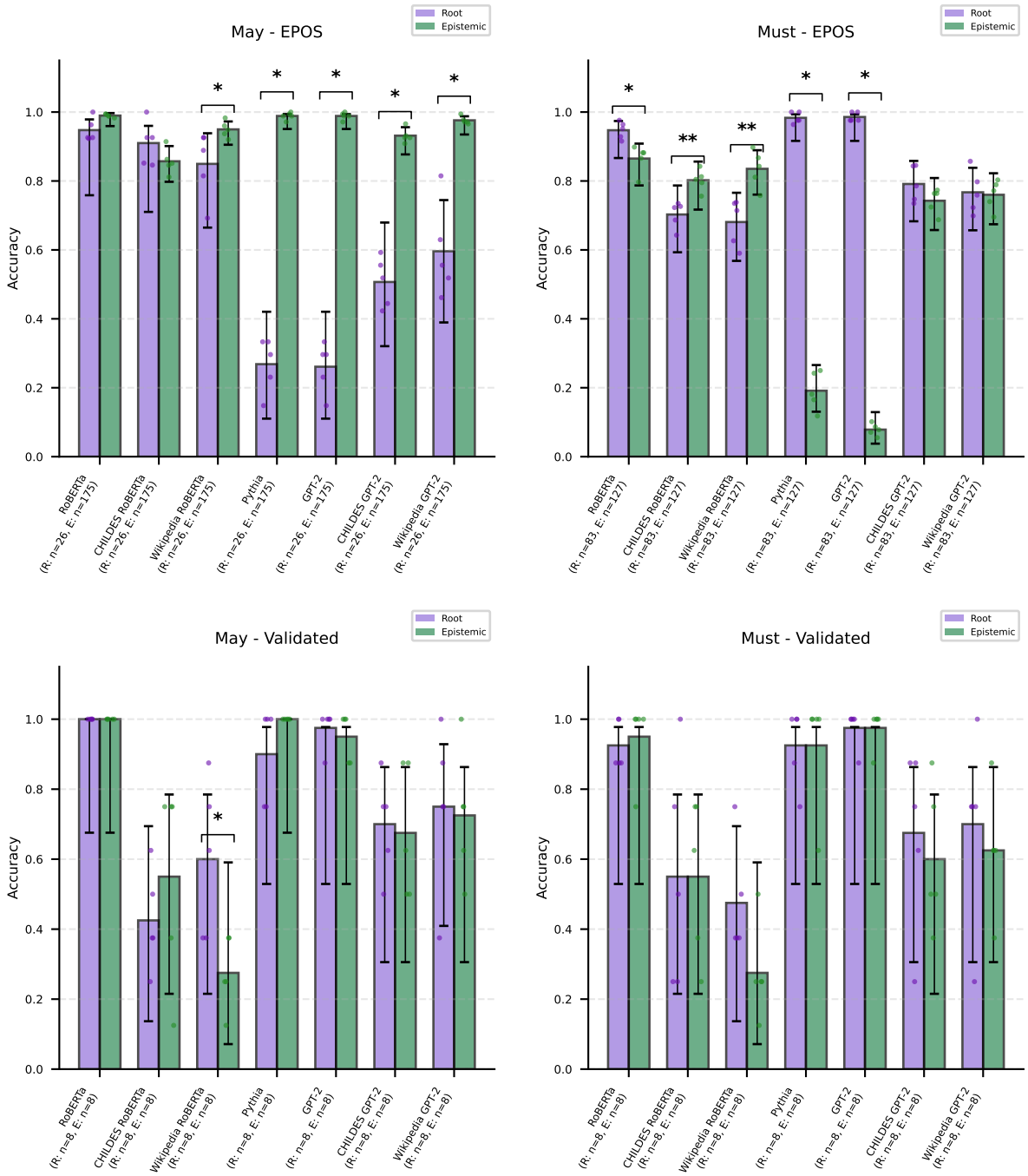
Table 2: Samples of modal expressions with root and epistemic interpretations in different datasets. EPOS and the validated data are released under a CC BY-SA 4.0 license and the naturalistic (CHILDES) data is released under CC BY-NC-SA 3.0. Across all datasets, the average sequence length is 15 tokens, with the longest sequence reaching 64 tokens.

## B Corpus statistics

| | | | | | Expectations | |
|---|---|---|---|---|---|---|
| Corpus | Modal | Dominant sense | Source | # | Non-CDL-models | CDL-models |
| General (AmE) | must | root (84%) | (Collins, 2007) | 930 | root > epistemic | |
| | may | epistemic (83%) | (Hacquard and Wellwood, 2012) | 138 | epistemic > root | |
| CHILDES (BrE) | must | epistemic (65%) | (van Dooren et al., 2022) | 452 | | epistemic > root |
| | may | root (56%) | (van Dooren et al., 2022) | 39 | | root > epistemic |

Table 3: Corpus statistics and performance expectations for the dominant senses for modals *may* and *must*.

# C Similarity method plots



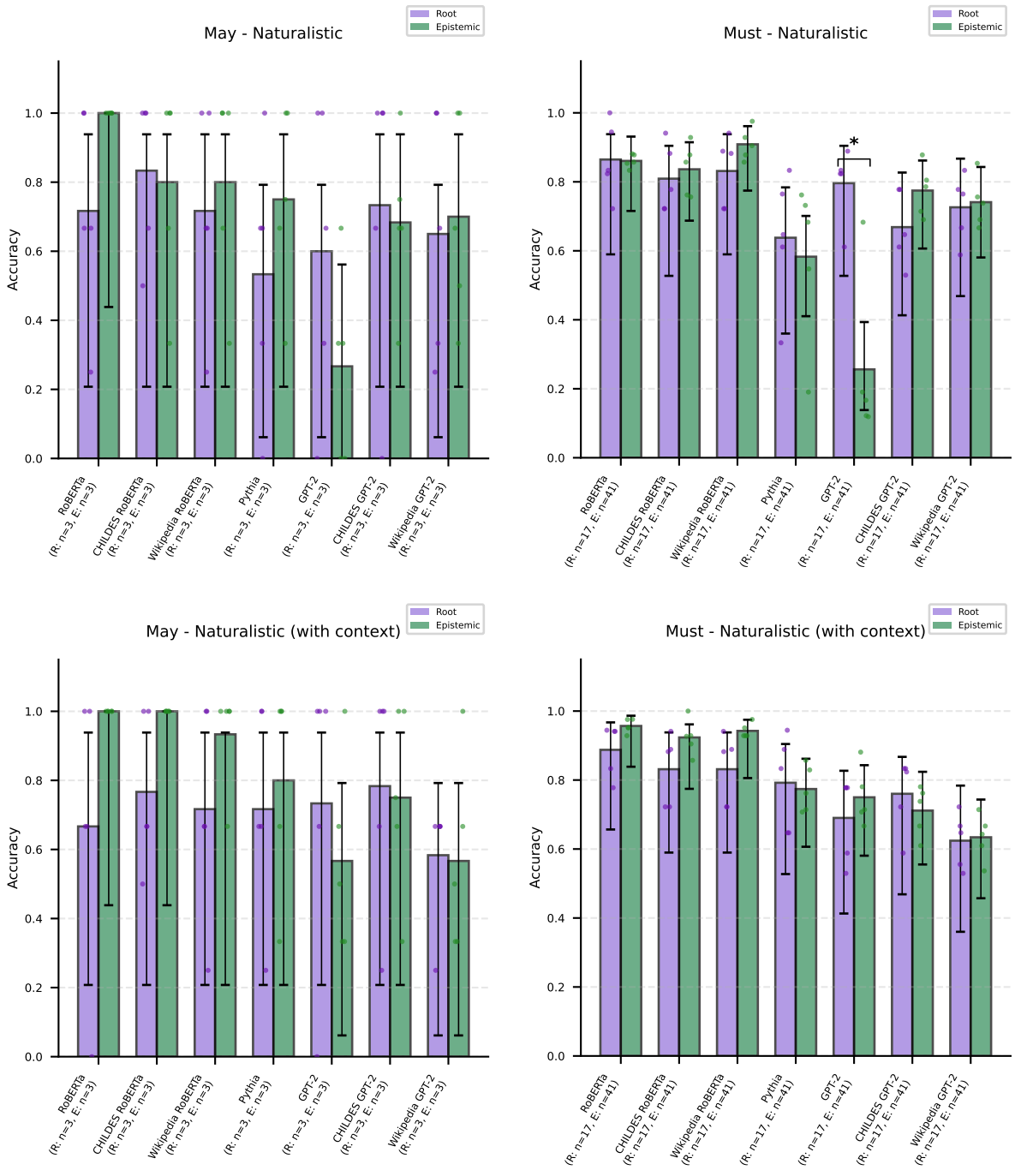



Figure 4: LM accuracy results based on vector similarity scores. Error bars represent binomial 95% CIs across five cross-validation folds.

## D Context variation analysis

As a supplementary analysis, we construct two modified evaluation sets based on the original validated dataset from Tulling et al. (2021) with the goal of investigating the usefulness of different types of linguistic context in modal sense disambiguation. Example data points from the modified datasets can be found in Table 4. The first modification removes the preceding sentence that provides the situational background necessary for interpreting the modal. This results in a "no context" variant. The second modification targets the clausal structure of the stimuli. In the original dataset, each modal occurs within an embedded clause. The goal behind Tulling et al. (2021) using this embedded structure is to link the modal interpretation to a third-person entity's perspective, helping bias participants toward a particular modal interpretation in a more clear way. To test whether this perspectival framing itself contributes useful information, we created a "monoclausal" variant. In this variant, we remove the matrix clause entirely. This leaves only the embedded clause as an independent sentence.

We should note that while the original stimuli have been validated with human participants, our modified versions have not. Therefore, we do not know to what extent the intended modal sense distinctions are equally clear in these new variants. Nevertheless, the original dataset was constructed under highly controlled conditions such as balancing for lexical frequency and word lengths across stimuli. These modified datasets, therefore, still provide a controlled setup for isolating the effect of removing different types of context. They also allow us to assess how informative each type of context is for LMs.

We turn first to the surprisal distributions shown in Figures 5, 6, and 7. Removing the situational context alone has a relatively limited impact on how surprisal scores are distributed across modal senses. In both the original and no context conditions, larger LMs maintain a clear surprisal preference for the dominant modal sense. Smaller LMs show little to no disambiguation between root and epistemic interpretations for *may* and *must* in these same conditions.

A different pattern emerges for the monoclausal variant. In this condition, the distribution of surprisal values converges considerably for root and epistemic senses, and this convergence occurs not only for smaller LMs but also for the larger ones. Our earlier experiments show that the larger LMs show a moderate to high capability of distinguishing between modal senses. Given our previous observations, we interpret the convergence in the monoclausal condition as evidence that the removed matrix clause is highly informative for modal sense disambiguation. The removal of the third-person entity prevents LMs from being able to show a clear preference for one modal sense over the other.

We also consider the similarity method results presented in Table 5. The performance gap between smaller and larger LMs remains wide across the board. For the larger causal LMs, we observe a modest drop in accuracy on the no context variant compared to the original dataset. This drop is followed by a more pronounced drop on the monoclausal variant. The difficulty that LMs exhibit on the monoclausal variant is consistent with their surprisal distributions. We also see that model performance is impacted by the removal of context, which aligns with what we observed in our previous tests using naturalistic data.

Our larger masked LM RoBERTa shows a similar trend for *must*. For *may*, however, the performance on the monoclausal variant is slightly higher than the performance on the no context variant. This difference highlights the benefit of using two different methods to examine model-internal representations, as the consistent pattern we observed regarding modal sense preferences based on surprisal distributions would not have been apparent from the similarity method alone.

Our context-varied results speak to the role of perspective-taking in modal sense disambiguation. In the original stimuli, embedding the target utterance under a matrix clause explicitly attributes the modal judgment to a third-person entity's perspective. When we remove this matrix clause in the monoclausal variant, the utterance is no longer explicitly anchored to a third-person perspective. Both surprisal-based preferences and similarity-based accuracies indicate that the presence of a third-person entity provides a meaningful cue that larger LMs exploit when disambiguating modal senses.

| Data | Flavor | Context | Target sequence | # |
|---|---|---|---|---|
| Validated | Root | as of now , snow white lives in the castle . and the princess assures | that the dwarfs **may** too . | 40 |
| | Epistemic | in direct sunlight , trolls turn into stone . but the wizard worries | that the dwarfs **may** too . | 40 |
| | Root | as of now , snow white lives in the castle . and the princess declares | that the dwarfs **must** too . | 40 |
| | Epistemic | in direct sunlight , trolls turn into stone . but the wizard figures | that the dwarfs **must** too . | 40 |
| Validated (no context) | Root | the princess assures | that the dwarfs **may** too . | 40 |
| | Epistemic | the wizard worries | that the dwarfs **may** too . | 40 |
| | Root | the princess declares | that the dwarfs **must** too . | 40 |
| | Epistemic | the wizard figures | that the dwarfs **must** too . | 40 |
| Validated (monoclausal) | Root | as of now , snow white lives in the castle . | the dwarfs **may** too . | 40 |
| | Epistemic | in direct sunlight , trolls turn into stone . | the dwarfs **may** too . | 40 |
| | Root | as of now , snow white lives in the castle . | the dwarfs **must** too . | 40 |
| | Epistemic | in direct sunlight , trolls turn into stone . | the dwarfs **must** too . | 40 |

Table 4: Example data points from different validated dataset versions.

| | | | May | | | Must | | |
|---|---|---|---|---|---|---|---|---|
| | | | Root | Epistemic | All | Root | Epistemic | All |
| Masked LMs | RoBERTa | Validated | 1.00 | 1.00 | 1.00 | 0.93 | 0.95 | 0.94 |
| | | Validated (no context) | 0.80 | 0.93 | 0.86 | 0.95 | 0.90 | 0.93 |
| | | Validated (monoclausal) | 0.85 | 0.93 | 0.89 | 0.88 | 0.85 | 0.86 |
| | CHILDES RoBERTa | Validated | 0.42 | 0.55 | 0.49 | 0.55 | 0.55 | 0.55 |
| | | Validated (no context) | 0.50 | 0.55 | 0.53 | 0.62 | 0.40 | 0.51 |
| | | Validated (monoclausal) | 0.40 | 0.53 | 0.46 | 0.45 | 0.55 | 0.50 |
| | Wikipedia RoBERTa | Validated | 0.60 | 0.28 | 0.44 | 0.47 | 0.28 | 0.38 |
| | | Validated (no context) | 0.68 | 0.70 | 0.69 | 0.72 | 0.65 | 0.69 |
| | | Validated (monoclausal) | 0.60 | 0.42 | 0.51 | 0.60 | 0.42 | 0.51 |
| Causal LMs | Pythia | Validated | 0.90 | 1.00 | 0.95 | 0.93 | 0.93 | 0.93 |
| | | Validated (no context) | 0.88 | 0.95 | 0.91 | 0.93 | 0.85 | 0.89 |
| | | Validated (monoclausal) | 0.78 | 0.82 | 0.80 | 0.80 | 0.80 | 0.80 |
| | GPT-2 | Validated | 0.97 | 0.95 | 0.96 | 0.97 | 0.97 | 0.97 |
| | | Validated (no context) | 0.95 | 0.88 | 0.91 | 0.93 | 0.88 | 0.90 |
| | | Validated (monoclausal) | 0.85 | 0.85 | 0.85 | 0.80 | 0.75 | 0.78 |
| | CHILDES GPT-2 | Validated | 0.70 | 0.68 | 0.69 | 0.68 | 0.60 | 0.64 |
| | | Validated (no context) | 0.65 | 0.70 | 0.68 | 0.50 | 0.62 | 0.56 |
| | | Validated (monoclausal) | 0.65 | 0.60 | 0.62 | 0.68 | 0.55 | 0.61 |
| | Wikipedia GPT-2 | Validated | 0.75 | 0.72 | 0.74 | 0.70 | 0.62 | 0.66 |
| | | Validated (no context) | 0.80 | 0.72 | 0.76 | 0.72 | 0.60 | 0.66 |
| | | Validated (monoclausal) | 0.55 | 0.78 | 0.66 | 0.55 | 0.68 | 0.61 |

Table 5: LM accuracy results on variations of validated data.

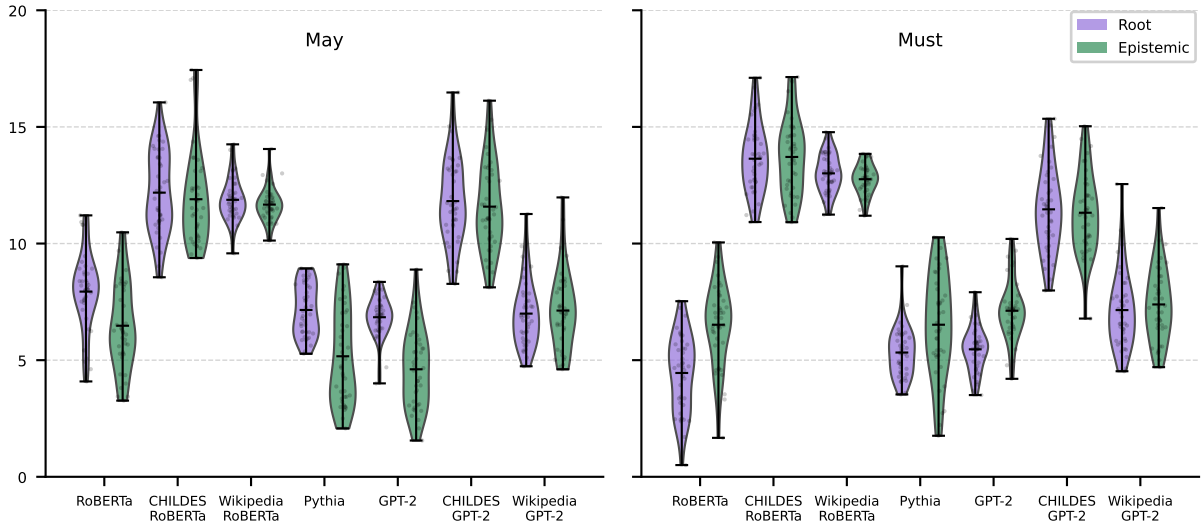



Figure 5: Surprisal score distributions for the modals in validated data. (Figure 3 in the main body.)

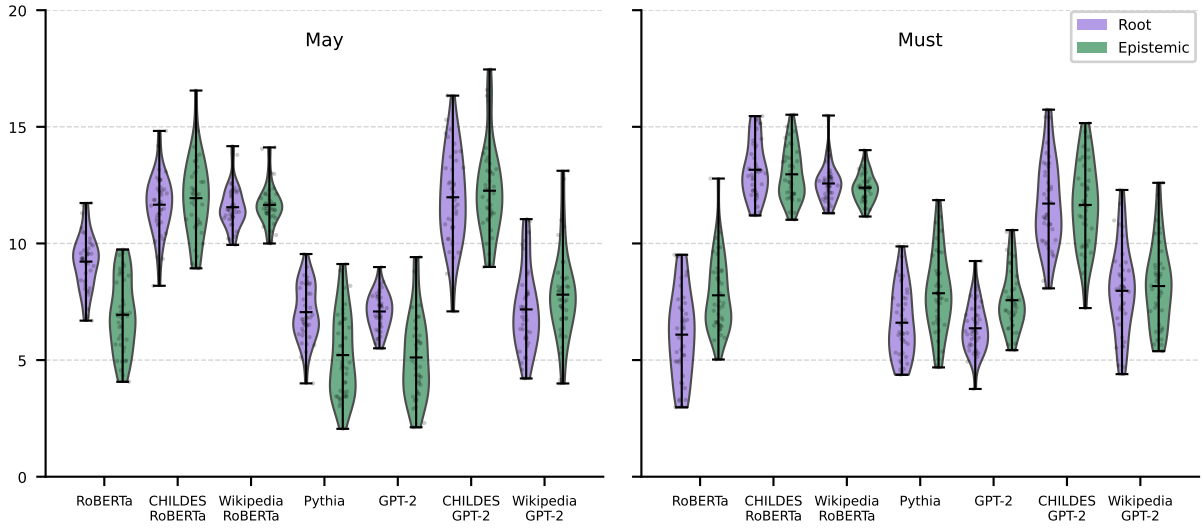



Figure 6: Surprisal score distributions for the modals in validated data (no context).

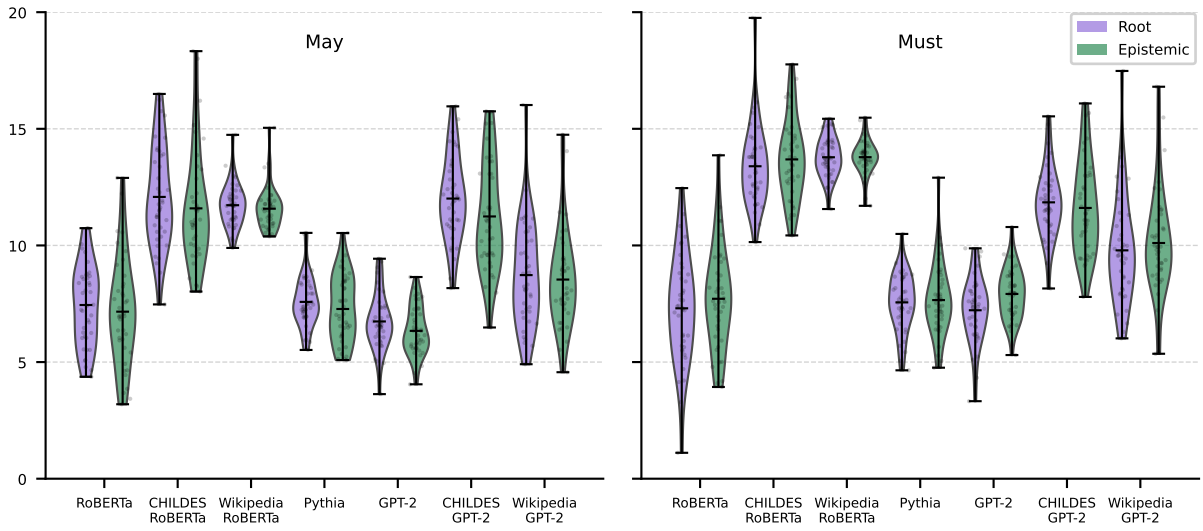



Figure 7: Surprisal score distributions for the modals in validated data (monoclausal).