Is It *JUST* Semantics?
A Case Study of Discourse Particle Understanding in LLMs

William Sheffield[1] **Kanishka Misra**[2,★] **Valentina Pyatkin**[3,4]
Ashwini Deo[1] **Kyle Mahowald**[1] **Junyi Jessy Li**[1]
[1]Linguistics, The University of Texas at Austin [2]Toyota Technological Institute at Chicago
[3]Allen Institute for AI [4]University of Washington
{sheffieldw, mahowald, jessy}@utexas.edu kanishka@ttic.edu

Abstract

Discourse particles are crucial elements that subtly shape the meaning of text. These words, often polyfunctional, give rise to nuanced and often quite disparate semantic/discourse effects, as exemplified by the diverse uses of the particle *just* (e.g., exclusive, temporal, emphatic). This work investigates the capacity of LLMs to distinguish the fine-grained senses of English *just*, a well-studied example in formal semantics, using data meticulously created and labeled by expert linguists. Our findings reveal that while LLMs exhibit some ability to differentiate between broader categories, they struggle to fully capture more subtle nuances, highlighting a gap in their understanding of discourse particles.

1 Introduction

Discourse particles are words that comment on aspects of the discourse context or interlocutor attitudes, giving rise to discourse effects that are often difficult to pin down. In some of their uses, their contribution is straightforward. For example, the *just* in "*Betsy just eats chicken nuggets*" tells us that chicken nuggets are the only thing Betsy eats. Without the *just*, we learn nothing about the other things Betsy will (not) eat. But not all uses of a polyfunctional discourse particle are easily unifiable: consider the occurrences of *just* in "*My brother **just** flew in to town*" (*just* ≈ recently) and "*I **just** won't stand for this injustice*" (*just* ≈ simply), or the latter two in "*A just judge **just** wouldn't stand for the laws **just** passed*".

From the view of formal semantics, these particles are difficult to analyze, partly because of their rich diversity of senses bundled into one word and partly because of the difficulty of characterizing each individual use (Lee, 1987; Bonomi and Casalegno, 1993; Beltrama, 2018). At the same time, they are extremely frequent in conversational language use and are crucial for comprehending discourse. There has been a great deal of work investigating LMs' general proficiency at function words (Kim et al., 2019) and overall sensitivity to discourse connectives (Pandia et al., 2021; Beyer et al., 2021; Cong et al., 2023). Recent work has also shown that LLMs struggle to grasp senses of discourse relations (Chan et al., 2024; Yung et al., 2024) at a broad level. At the same time, it is unclear how well do LLMs' grasp the meaning (or senses) of discourse particles like *just*—which, as we've discussed—have peculiarly interesting versatility in their semantics.

Focusing on this line of work, **this work investigates the polyfunctional discourse particle *just*, which has been particularly well-studied in formal semantics** (Lee, 1987; Grosz, 2012; Coppock and Beaver, 2014; Beltrama, 2022; Deo and Thomas, 2025, *i.a.*). Using data created and labeled by expert linguists, we investigate the metalinguistic capabilities of LLMs to distinguish the nuanced senses of *just* described in the formal semantics literature. We find that while they possess basic sense distinctions, language models, especially smaller ones, struggle to fully discern the subtle differences of *just*'s senses, signaling the lack of a nuanced understanding of the meaning of discourse particles.[1]

2 *Just* and Its Semantics

Discourse particles, like English *just*, are a class of function words that are sensitive to discourse-level contextual information. Examples include exclusive particles, such as English *just* and *only*, whose salient discourse function is to exclude alternatives from a contextually determined set of alternatives (Coppock and Beaver, 2014).

★Work partly done at UT-Austin before joining TTIC.

[1]Code and data can be found here https://github.com/sheffwb/IsItJUSTSemantics

Findings of the Association for Computational Linguistics: ACL 2025, pages 21704–21715
July 27 - August 1, 2025 ©2025 Association for Computational Linguistics

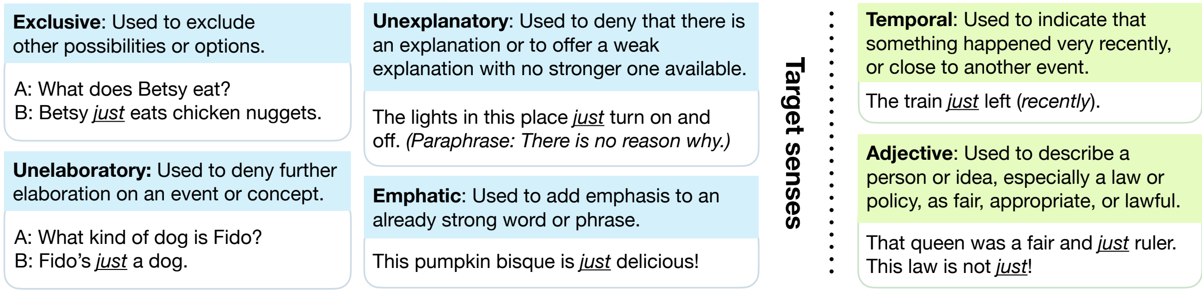

Figure 1: List of *just* senses in declarative sentences (with target senses in blue). Note that all senses other than the Adjective *just* are discourse particle senses and function as adverbs. All examples save the Adjective ones come from Warstadt (2020).

English *just* is a good candidate to study as it (1) has been thoroughly analyzed and (2) has many senses. Deo and Thomas (2025) present a unified account for all senses of the discourse particle *just*, outlining 12 senses of the word, excluding the adjective sense[2] (e.g. *"She was a just and fair sovereign"*). We target LLMs' ability to distinguish four of these senses that seem reasonably distinct from a semantic point of view: the exclusive (Coppock and Beaver, 2014, *i.a.*), unelaboratory (Warstadt, 2020), unexplanatory (Wiegand, 2018; Windhearn, 2021), and emphatic (Lee, 1987; Beltrama, 2018, 2022); we also use the temporal and adjective senses as controls (Figure 1).

These four senses warrant further definition. The following examples come from Warstadt (2020). The exclusive sense excludes other salient possibilities: In one reading of "Betsy just eats chicken nuggets.", the *just* excludes other options of what Betsy could eat besides chicken nuggets. The unelaboratory sense denies the need for further elaboration on an event or concept; on one reading, in response to "What kind of dog is Fido?" the *just* in "Fido is just a dog." means that Fido is simply a mutt. The unexplanatory sense hinges on the lack of an explanation for something, and so usually has the force of adding 'I don't know why'. For example, in a haunted house someone might say "The lights in this place just turn on and off.", since they are not sure as to why the lights turn on and off. The emphatic sense is used to strengthen an already extreme predicate: "This pumpkin bisque is just delicious!" is stronger with the *just*.

The adjective sense is the most distinct in meaning and occurs in very different syntactic environments. The temporal sense serves as a middle ground between the four target senses and the adjective sense: *just* is still understood as a discourse particle here, but its meaning is clearly distinguishable from the other four senses.

3 Experimental Setup

Data This paper uses two sources of data to study *just*: **(1) hand-constructed:** 90 sentences (15 of each sense) carefully created by an expert to have only one sense available for the *just* of each sentence without any context; and **(2) annotated:** 149 sentences "in the wild" that contain *just*, taken from OpenSubtitles (Lison et al., 2018) and annotated by semanticists with their senses, with associated context.

The hand-constructed corpus is necessary, as clarity in the reading of the sentence is crucial for targeted metalinguistic experiments, since ambiguity can be pervasive in *just*s "in the wild". For example, in the sentence "*I just saw Nancy.*", *just* can either mean the seeing occurred recently (temporal reading), or only Nancy was seen (exclusive reading).[3] This data is created by a graduate linguist who has studied discourse particle semantics and is a native speaker of American English.

For the annotated corpus, we chose movie subtitles over other texts as they are more conversational, and therefore more likely to contain instances of *just* as a discourse particle.[4] Our volunteer annotation team consists of two senior semanticists whose expertise is in discourse particles, and eight graduate students who have taken a graduate semantics class that extensively discussed particles. We collected annotations for 149 sentences, which

[2] The adjective sense is arguably associated with a distinct homophonous word.

[3] Readings are often disambiguated in speech based on the intonation of the utterance, which is not accessible to text-only models. We leave speech models for future work.

[4] Additionally, subtitles contain context, which can help disambiguate different readings for an instance of *just*. However, we observe in Section 4.2 that this has little effect on performance, further motivating our hand-constructed data.

Sense	Count	Sense	Count
Exclusive	60	Emphatic	21
Unelaboratory	12	Temporal	33
Unexplanatory	22	Adjective	1

Table 1: Distribution of senses in annotated subtitles data. Our hand-constructed data has a balanced distribution of 15 sentences per sense.

were annotated by a variable subset of 8 annotators. When there was disagreement, we fell back on two additional senior annotators, whose labels were both considered regardless of agreement. Table 1 shows the distribution of *just* senses in this subset.[5]

All sentences in both datasets have a **strong primary reading**, either by construction (in the hand-constructed corpus) or by annotator agreement (in the annotated corpus). While this does not rule out the possibility of multiple readings for a given sentence, strong speaker consensus on the reading of an occurrence of *just* does remove more ambiguous sentences from out data. At the same time, this speaker consensus is a stronger signal and should be recoverable by a good model. Examples from both datasets are in Appendix A.

Models We use instruction-tuned models that can understand our meta-linguistic queries and evaluate diverse LLM architectures across parameter scales: `Llama-3-8b`, `Llama-3.2-1b/3b`, `Llama-3.3-70b`, `Mistral-7b-v0.3`, `OLMo-7b`, `OLMo2-7b/13b`, and `Gemma2-2b/9b`. All experiments were run on at most two NVIDIA A40 GPUs. Model details in Appendix B.

4 Do LLMs get nuanced *just* senses?

4.1 Method

In this setting, language models are prompted to label the sense of *just* in the sentence. The full prompt can be found in Appendix F. The prompt includes both definitions and examples of each of the six senses from Warstadt (2020). This experiment tests if the models are picking up on the information relevant to these labels even though models may not necessarily be categorizing uses of *just* along the same lines as theory.

To circumvent parsing verbose generations common with prompted generation, we instead use the log probabilities of each label, conditioned on the prompt. We take the label given the highest probability as the label assigned by the model to the sentence. That is, the label assigned to a sentence by a model M is $\text{argmax}_{l \in L} P_M(l|S)$ where L is the set of label continuations and S is the prompt including the sentence to be classified. The conditional probability P_M is calculated using `minicons` (Misra, 2022). Both the label continuations and the prompt are formatted to each model's chat formatting specifications.

We leverage the formatting of the sense labels directly following the in-context examples to ensure the sense labels are assigned reasonable probabilities by the model. We directly compare the language model labels to ground-truth labels.

4.2 Results

Figure 2 shows the accuracy for the four target senses (Exclusive, Unelaboratory, Unexplanatory, and Emphatic)[6], on three datasets: the hand-constructed data, the subtitles data alone, and the subtitles with two prior utterances as context.[7] Based on the frequency of labels, chance performance is 1/6 $\approx$ 0.167 (uniform) for the hand-constructed data and 60/149 $\approx$ 0.403 (most frequent label, Exclusive) for the subtitles data.

For the hand-constructed data, all models except `Llama-3.2-1b` substantially outperform chance. Concerning model size, we see a substantial increase in accuracy (+0.28) from `Llama-3.2-1b` to `Gemma-2-2b`, suggesting there is a critical model size of 2B parameters required for this task (as well as for our other task in Section 5.2). Additionally, we observe that the largest model, `Llama-3.3-70b`, is not performing much above the best performing mid-size models, `Mistral-7b-v0.3` and `Gemma-2-9b`, suggesting that a large model is not required for good performance.

Turning to the subtitle data, we observe a substantial drop in accuracy across all models, -0.24 on average, without context compared to the hand-constructed data. The degradation in performance is most likely due to the subtitle sentences being more ambiguous as to what reading of *just* is meant. This indicates a notable deficit in model understanding of *just*'s sense distinctions, since these subtitles are more naturalistic than the hand-constructed data. Interestingly, context does *not* help disam-

[5] There are four sentences with two occurrences of *just*: in two of these, they are simply disfluencies, and so only one label is possible; in the other two, evaluating models on either occurrence did not change results.

[6] Accuracy for all six senses is reported in Appendix C.

[7] We also ran this experiment with the five prior utterances and find no notable difference.

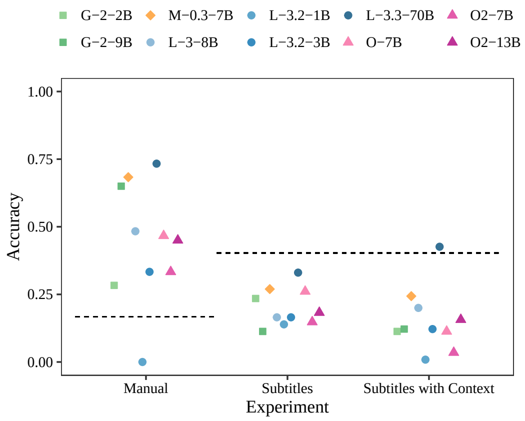

Figure 2: Model accuracies for the sense labeling task on our four target senses. Dashed lines show chance performance: 0.167 for hand-construction, 0.403 for subtitle data. **Model Legend:** L: Llama, G: Gemma; M: Mistral; O: OLMo.

biguate the sense of *just*, as we see a further decrease in model performance, -0.05 on average, when context is added, except for `Llama-3.3-70b`, which sees an increase in accuracy (+0.10), but is still at chance. These results demonstrate an important gap in models understanding of *just* senses: even when given sense definitions, they struggle to accurately predict the sense of *just* in naturalistic sentences.

5 Can LLMs distinguish *just* senses?

5.1 Method

While few-shot sense labeling evaluates a model's meta-linguistic understanding of *just*'s senses, they come from formal linguistic theory and it is unclear if the differences between them are internalized in an LLM. It is also unclear if the differences between these senses is realized in an LLM. In order to better measure how LLMs can categorize different uses of *just*, we consider model judgments on pairs of sentences, only done with the *hand-constructed* data with unambiguous *just* senses.

The model is given two sentences s_i, s_j and prompted to answer if the *just*s are used in the same way for both sentences (full prompt in Appendix F). Similar to the previous experiment, we compare the probabilities of the continuations "Yes" and "No" conditioned on the prompt Z_{ij}, which contains s_i, s_j. Thus, given all pairs of sentences, we define a heatmap $\mathbf{H}^M$ for each model M:

$$H_{ij}^M = \log(P_M(\text{Yes}|Z_{ij})) - \log(P_M(\text{No}|Z_{ij}))$$

normalized to $[0, 1]$ per model. Intuitively, if the model judges two sentences to use *just* in the same way, it will give a higher probability to "Yes" and a lower probability to "No", and vice versa.

Controls To ensure this method *is* able to separate senses of words, we also perform tests for 2 words that each have multiple, clearly separate senses: "bat" (2 senses) and "bank" (4 senses). Models show clear sense separation, verifying our method is reasonable (results in Appendix D).

A distinct advantage of this approach is that we do not assume model knowledge of the sense labels for *just*, as in the prior experiment, and instead only focus on whether they treat the meanings of just in a similar way, allowing for gradience in meaning distinctions.

5.2 Results

Models behave consistently and are insensitive to pair ordering. All models, save for `Llama-3.2-1b`, have a dark upward diagonal meaning that models see a sentence as having the same use of *just* as itself (the $i = j$ diagonals); this indicates that this methodology is effective for probing model judgments on use. Additionally, the heatmaps are symmetric along the diagonals ($(i, j) \approx (j, i)$), which indicate that they are insensitive to the ordering of the sentence pairs.

Models separate *just* senses to some degree, but not for the nuanced target senses for particle use. The smallest model, `Llama-3.2-1b` is the only model to not show significant separation of the metric between pairs with the same sense and pairs with different senses ($p = .11$). All other models show significant separation ($p < .005$).[8] However, for many models the effect size is small. Based on these distributions, all models except `Llama-3.2-1b` are able to identify sentences with the same use writ-large, although the separation is quite weak for all but the largest models (Cohen (1988)'s d of 2.32 for `Llama-3.3-70b`, 1.91 for `Gemma-2-9b`, 1.66 for `Mistral-7b-v0.3`).

The strongest sense separation is for the adjective sentences, which is expected given their difference in meaning and syntactic category to the other, discourse particle senses of *just*. Models other than `Llama-3.2-1b` also show some separation for the temporal sense. These results show that LLMs are able to perceive different *just* usages in more clearly separable senses of the word.

[8] *p*-values calculated with Welch (1947)'s t-test on the metric between pairs with the same sense but different sentences (N=1260) and pairs with different senses (N=6750).

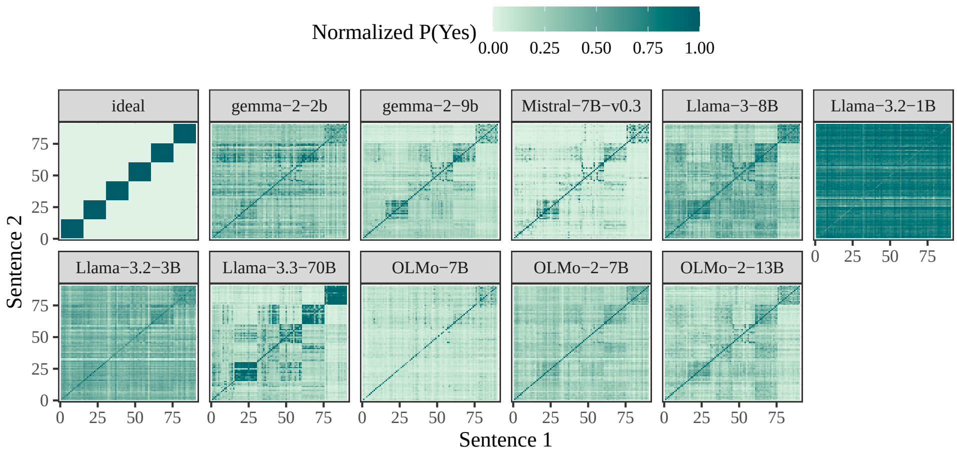

Figure 3: Heatmap of language model pairwise comparisons of the use of *just* in the two sentences. The "ideal" heatmap shows if all sentences with the same use were judged so by the model. Senses in ascending order are: Exclusive, Unelaboratory, Unexplanatory, Emphatic, Temporal, and Adejctive.

However, most models fail to show clear separation for the target senses, except for the Unelaboratory and Emphatic senses in `Gemma-2-9b`, `Mistral-7B-v0.3`, `Olmo-2-13b`, `Llama-3-8b`, and `Llama-3.3-70b`. Hence, although all but the smallest models do show *some* separation for *just*'s senses, even the largest models fail to fully capture its richness.[9] By contrast, words with relatively clear separation of senses such as *bat* and *bank* are more readily and consistently distinguished by models, suggesting particular gaps in the LLMs' handling of a polyfunctional particle like *just*.

6 Conclusion

We find that reasonably sized language models (over 1B parameters) show some basic separation for the complex English discourse particle *just*'s senses, but fail to fully discern the deep subtlety of its senses with two very different prompting strategies. First, in an overt, few-shot sense-labeling setting with definitions and examples for each senses; and second in an open-ended, pairwise comparison setting allowing full freedom from *just*'s senses as described in formal semantics research. This lack of nuanced sensitivity points to a gap in language model performance key for the study of discourse particles and discourse comprehension, echoing the findings of recent works (Chan et al., 2024; Wei et al., 2024; Yung et al., 2024).

Limitations

This work looked into the English *just* as a case study of LLM's metalinguistic capability to understand the semantics of discourse particles; more work needs to be performed before generalizing our findings to other discourse particles, which we leave for the future.

We have used metalinguistic prompting to analyze LLMs' understanding of *just* senses. However, this class of methods has been found to underestimate LLMs' linguistic abilities, especially when compared to using direct sentence log-probabilities (Hu and Levy, 2023). However, it is not obvious how one could analyze the nuanced distinctions in the senses of discourse particles using standard log-probability based approaches (Warstadt et al., 2020; Hu et al., 2020; Misra et al., 2023, i.a.). We therefore leave this direction as an avenue for future work.

Acknowledgments

We thank Alex Warstadt, Karen Chen, and the UT Computational Linguistics Research Seminar for their suggestions for this paper. We also thank William Thomas and all the annotators for their volunteer work in annotating the subtitle data. This work was partially supported by NSF grants 2104995, 2107524, and 2145479.

References

AI@Meta. 2024. Llama 3 model card.

Andrea Beltrama. 2018. Metalinguistic" just" and" simply": exploring emphatic exclusives. In *Semantics and Linguistic Theory*, pages 307–324.

Andrea Beltrama. 2022. Just perfect, simply the best:

[9]See Appendix E for details on the results in this section.

an analysis of emphatic exclusion. *Linguistics and Philosophy*, 45(2):321–364.

Anne Beyer, Sharid Loáiciga, and David Schlangen. 2021. Is incoherence surprising? targeted evaluation of coherence prediction from language models. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4164–4173, Online. Association for Computational Linguistics.

Andrea Bonomi and Paolo Casalegno. 1993. Only: Association with focus in event semantics. *Natural Language Semantics*, 2(1):1–45.

Chunkit Chan, Cheng Jiayang, Weiqi Wang, Yuxin Jiang, Tianqing Fang, Xin Liu, and Yangqiu Song. 2024. Exploring the potential of ChatGPT on sentence level relations: A focus on temporal, causal, and discourse relations. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 684–721, St. Julian's, Malta. Association for Computational Linguistics.

Jacob Cohen. 1988. *Statistical Power Analysis for the Behavioral Sciences*, 2 edition. L. Erlbaum Associates.

Yan Cong, Emmanuele Chersoni, Yu-Yin Hsu, and Philippe Blache. 2023. Investigating the effect of discourse connectives on transformer surprisal: Language models understand connectives, Even so they are surprised. In *Proceedings of the 6th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pages 222–232, Singapore. Association for Computational Linguistics.

Elizabeth Coppock and David I Beaver. 2014. Principles of the exclusive muddle. *Journal of Semantics*, 31(3):371–432.

Ashwini Deo and William Carl Thomas. 2025. Addressing the widest answerable question: English "just" as a domain widening strategy. *Journal of Semantics*, page ffae015.

Dirk Groeneveld, Iz Beltagy, Pete Walsh, Akshita Bhagia, Rodney Kinney, Oyvind Tafjord, Ananya Harsh Jha, Hamish Ivison, Ian Magnusson, Yizhong Wang, Shane Arora, David Atkinson, Russell Authur, Khyathi Chandu, Arman Cohan, Jennifer Dumas, Yanai Elazar, Yuling Gu, Jack Hessel, Tushar Khot, William Merrill, Jacob Morrison, Niklas Muennighoff, Aakanksha Naik, Crystal Nam, Matthew E. Peters, Valentina Pyatkin, Abhilasha Ravichander, Dustin Schwenk, Saurabh Shah, Will Smith, Nishant Subramani, Mitchell Wortsman, Pradeep Dasigi, Nathan Lambert, Kyle Richardson, Jesse Dodge, Kyle Lo, Luca Soldaini, Noah A. Smith, and Hannaneh Hajishirzi. 2024. Olmo: Accelerating the science of language models. *Preprint*.

Patrick Georg Grosz. 2012. On the grammar of optative constructions.

Jennifer Hu, Jon Gauthier, Peng Qian, Ethan Wilcox, and Roger Levy. 2020. A systematic assessment of syntactic generalization in neural language models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1725–1744, Online. Association for Computational Linguistics.

Jennifer Hu and Roger Levy. 2023. Prompting is not a substitute for probability measurements in large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5040–5060, Singapore. Association for Computational Linguistics.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv:2310.06825*.

Najoung Kim, Roma Patel, Adam Poliak, Patrick Xia, Alex Wang, Tom McCoy, Ian Tenney, Alexis Ross, Tal Linzen, Benjamin Van Durme, et al. 2019. Probing what different nlp tasks teach machines about function word comprehension. In *Proceedings of the Eighth Joint Conference on Lexical and Computational Semantics (* SEM 2019)*, pages 235–249.

David Lee. 1987. The semantics of just. *Journal of pragmatics*, 11(3):377–398.

Pierre Lison, Jörg Tiedemann, and Milen Kouylekov. 2018. Opensubtitles2018: Statistical rescoring of sentence alignments in large, noisy parallel corpora. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Kanishka Misra. 2022. minicons: Enabling flexible behavioral and representational analyses of transformer language models. *arXiv:2203.13112*.

Kanishka Misra, Julia Rayz, and Allyson Ettinger. 2023. COMPS: Conceptual minimal pair sentences for testing robust property knowledge and its inheritance in pre-trained language models. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2928–2949, Dubrovnik, Croatia. Association for Computational Linguistics.

Team OLMo, Pete Walsh, Luca Soldaini, Dirk Groeneveld, Kyle Lo, Shane Arora, Akshita Bhagia, Yuling Gu, Shengyi Huang, Matt Jordan, Nathan Lambert, Dustin Schwenk, Oyvind Tafjord, Taira Anderson, David Atkinson, Faeze Brahman, Christopher Clark, Pradeep Dasigi, Nouha Dziri, Michal Guerquin, Hamish Ivison, Pang Wei Koh, Jiacheng Liu, Saumya Malik, William Merrill, Lester James V. Miranda, Jacob Morrison, Tyler Murray, Crystal Nam, Valentina Pyatkin, Aman Rangapur, Michael Schmitz, Sam Skjonsberg, David Wadden, Christopher Wilhelm, Michael Wilson, Luke Zettlemoyer, Ali Farhadi, Noah A. Smith, and Hannaneh Hajishirzi. 2024. 2 olmo 2 furious.

Lalchand Pandia, Yan Cong, and Allyson Ettinger. 2021. Pragmatic competence of pre-trained language models through the lens of discourse connectives. In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 367–379, Online. Association for Computational Linguistics.

Gemma Team. 2024. Gemma.

Alex Warstadt. 2020. "just" don't ask: Exclusives and potential questions. In *Proceedings of Sinn und Bedeutung*, volume 24, pages 373–390.

Alex Warstadt, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang, and Samuel R. Bowman. 2020. BLiMP: The benchmark of linguistic minimal pairs for English. *Transactions of the Association for Computational Linguistics*, 8:377–392.

Kangda Wei, Aayush Gautam, and Ruihong Huang. 2024. Are LLMs good annotators for discourse-level event relation extraction? In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 1–19, Miami, Florida, USA. Association for Computational Linguistics.

Bernard L Welch. 1947. The generalization of 'student's'problem when several different population varlances are involved. *Biometrika*, 34(1-2):28–35.

Mia Wiegand. 2018. Exclusive morphosemantics: Just and covert quantification. In *Proceedings of the west coast conference on formal linguistics (WCCFL)*, volume 35, pages 419–429.

Mia Windhearn. 2021. Alternatives, exclusivity and underspecification.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, R'emi Louf, Morgan Funtowicz, and Jamie Brew. 2019. Huggingface's transformers: State-of-the-art natural language processing. *ArXiv*, abs/1910.03771.

Frances Yung, Mansoor Ahmad, Merel Scholman, and Vera Demberg. 2024. Prompting implicit discourse relation annotation. In *Proceedings of the 18th Linguistic Annotation Workshop (LAW-XVIII)*, pages 150–165, St. Julians, Malta. Association for Computational Linguistics.

A Dataset Examples

Table 3 shows examples for the four primary senses from both datasets.

B Model Details

Table 2 shows model details with the exact Huggingface model ID; all models are instruction-tuned and run using Huggingface's `transformers` library (Wolf et al., 2019) All experiments are run with a temperature of 0. `Llama-3.3-70B-Instruct` was run with 4-bit quantization. All experiments took at most 2 hours to run for a single model, and models were run on at most 2 NVIDIA A40 GPUs.

Huggingface Model ID	Citation
`Llama-3.2-1B-Instruct`	AI@Meta (2024)
`Llama-3.2-3B-Instruct`	AI@Meta (2024)
`Meta-Llama-3-8B-Instruct`	AI@Meta (2024)
`Llama-3.3-70B-Instruct`	AI@Meta (2024)
`Mistral-7B-Instruct-v0.3`	Jiang et al. (2023)
`OLMo-7B-Instruct-hf`	Groeneveld et al. (2024)
`OLMo-2-1124-7B-Instruct`	OLMo et al. (2024)
`OLMo-2-1124-13B-Instruct`	OLMo et al. (2024)
`gemma-2-2b-it`	Team (2024)
`gemma-2-9b-it`	Team (2024)

Table 2: Model details. All models are instruction-tuned, and all experiments are run with a temperature of 0.

C Overall sense labeling accuracy

Figure 4 shows sense labeling accuracy on data for all six senses. Although accuracy is higher overall, there are no other notable trends, indicating that models struggle less with the temporal and adjective senses than the four target senses.

D *Bat* and *Bank* pairwise sense comparisons

To check that our pairwise comparison experiment is sound, we test on two additional words with clearly separable senses: "bat" (2 senses: a flying mammal or sports bat) and "bank" (4 senses: a river bank (Noun), a financial institution (Noun), to turn (Verb), or to deposit money (Verb)).

Heatmaps for model pairwise comparisons of *bat* and *bank* are shown in Figure 5. Each includes an idealized heatmap, where only pairs with the same sense are given a score of 1 and the rest 0.

For both "bat" and "bank" we see clear separation of senses, as seen by the squares along the diagonal for all models, with `Llama-3.3-70b` amd

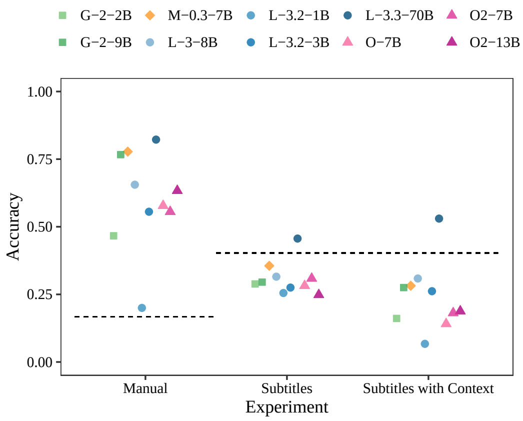

Figure 4: Accuracy of models for the sense labeling task on all six senses. Chance is shown with the dotted lines: 0.167 for hand-construction, 0.403 for subtitle data

`Gemma-2-9b` being closest to the ideal matrix. This indicates that this method is viable for testing language model ability to separate senses using this pairwise comparison methodology.

Interestingly, we do see notable confusion between the financial institution sense (Noun) and the deposit money sense (Verb) with the hot spot off the diagonal, showing that the models are focusing on meaning and not syntactic differences, as focus on syntax would be shown by hot spots between the two verbs and/or nouns.

E Significance Tests for Pairwise Experiments

Table 4 shows the distribution of the sentence pair metrics $(i, j) - (j, i)$ for $i \leq j$, as well as the p-value for a two-sided, one sample t-test for the mean being different from 0 (8100 total sentence pairs). This tests if models exhibit ordering preferences for senses of *just*. Although all means are significantly different from 0, they are never farther than .039 (`Gemma2-2b`), and all but three are less than 0.016 from 0, indicating the effect of sentence ordering is minimal for the pairwise experiment. Hence, we conclude the models have minimal ordering bias in the pairwise experiment.

Additionally, visually we can observe in Figure 3 no clear ordering preference for any particular sense, which would be indicated by a strong, dark horizontal or vertical band in the heatmap. We see a *possible* such band for `Gemma2-2b` with the horizontal temporal pairings, indicating the this model is slightly more likely to force non-temporal readings for a sentence pair (s_i, s_j) if s_i's *just* isn't temporal. These absence of such bands everywhere

Sense	Hand-Constructed	OpenSubtitles
Exclusive	The company **just** repairs existing units.	Excuse me, judge, but this is **just** about whether or not I get bail, right?
Unelaboratory	A torus is **just** a donut.	And you're **just** in a sort of limbo. [. . .]
Unexplanatory	She **just** left, out of the blue, two days ago.	Like those little doodles you **just** happened to draw?
Emphatic	Mammoths are **just** gigantic.	[. . .]I'm sure Ryan's gonna be **just** fine.
Temporal	I **just** received the news.	Who were you **just** on the phone with?

Table 3: Examples for the four primary just senses from both datasets, and the temporal one for comparison.

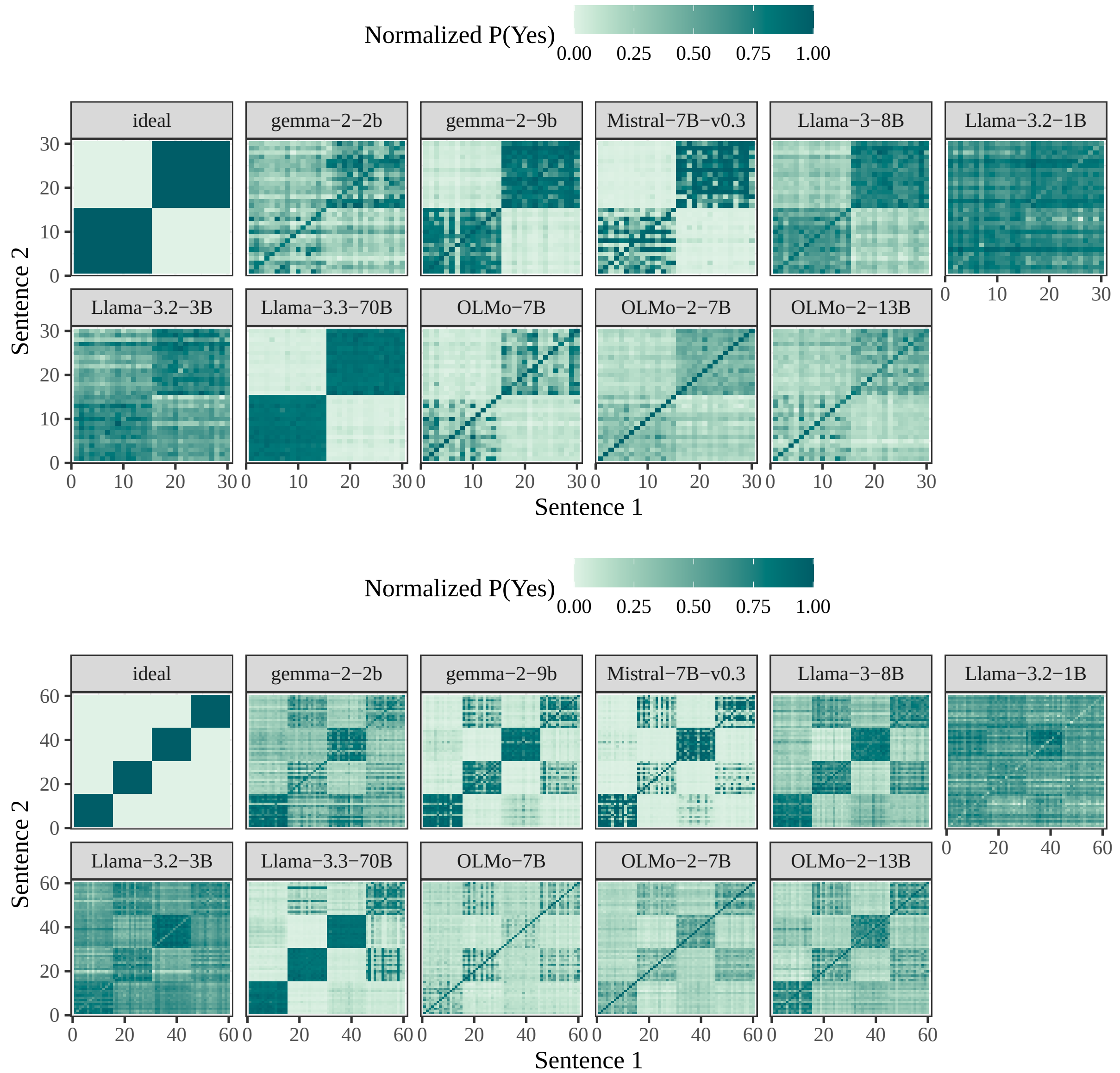

Figure 5: Heatmap of language model pairwise comparisons of the use of "bat" (top) and "bank" (bottom). The "ideal" heatmap shows if all sentences with the same use were judged so by the model. The senses in ascending order for (1) bat are: flying mammal, sports bat and (2) bank are: riverbank (Noun), financial institution (Noun), to turn (Verb), to deposit/keep money somewhere (Verb).

else demonstrates the lack of ordering preferences in these models for this task.

Table 5 contains the distributions of the metric for the pairs with the same sense (but not the exact same sentence $i = j$) and pairs with different senses. p-values calculated with a one-sided Welch (1947)'s t-test on the metric between pairs with the same sense but different sentences (1260 instances) greater than pairs with different senses (6750 instances). The effect size is calculated using Cohen

Distribution $\mu \pm \sigma$	p-value	Model
-0.016±0.118	<0.0001	`Llama-3.2-1b`
-0.008±0.087	<0.0001	`Llama-3.2-3b`
0.011±0.132	<0.0001	`Llama-3.3-70b`
0.027±0.130	<0.0001	`Meta-Llama-3-8b`
0.008±0.100	<0.0001	`Mistral-7b-v0.3`
0.003±0.100	0.003	`OLMo-2-1124-13b`
-0.008±0.086	<0.0001	`OLMo-2-1124-7b`
-0.010±0.067	<0.0001	`OLMo-7b`
-0.039±0.163	<0.0001	`gemma-2-2b`
0.027±0.073	<0.0001	`gemma-2-9b`

Table 4: p-values for a two-sided one-sample t-test on, and distributions for, $(i, j) - (j, i), i \leq j$

(1988)'s d. Therefore greater effect size means a better separation of senses.

F Prompts

The prompts used for the experiments in Sections 4 and 5 can be found in Figures 6 and 7, respectively. For Section 4, the prompt includes shorter sense definitions to match what human annotators were given.

G Annotation Interface

Figure 8 shows an example of the interface annotators used to label the subtitle data. Users were given 5 sentences of context, and selected one of the sense labels. They could also include comments. Annotators were trained in an in-person session with one of the authors.

Sense Labeling prompt:

User:
The word "just" in English can have several distinct uses and meanings. Here are the main senses of "just" along with their characteristics:

Exclusionary:
Used to exclude other possibilities or options.

Example: "Todd just drinks water."
Sense: Exclusionary

Unelaboratory:
Used to deny further elaboration on an event or concept.

Example: "Water is just a hydrogen atom with an oxygen atom."
Sense: Unelaboratory

Unexplanatory:
Used to deny that there is an explanation or to offer a weak explanation with no stronger one available.

Example: "The lights just turn on and off."
Sense: Unexplanatory

Emphatic:
Used to add emphasis to an already strong word or phrase.

Example: "This is just delicious."
Sense: Emphatic

Temporal:
Used to indicate that something happened very recently, or close to another event.

Example: "We just pulled in the drive way a minute ago."
Sense: Temporal

Adjective:
Used to describe a person or idea, especially a law or policy, as fair, appropriate, or lawful.

Example: "One day he'll get your just desserts."
Sense: Adjective

Given these, identify what sense of "just" is used in the following sentence. Respond with the sense label.

Sentence: `<sentence>`

Format your response as 'Response: [label]'
Assistant: Response: `<label>`

Figure 6: Prompt for the "just" sense labeling task. When context is included for the subtitle data, the prompt is slightly altered to: "Identify what sense of "just" is used in the last sentence of the following passage".

Same Sense $\mu \pm \sigma$	Different Sense $\mu \pm \sigma$	p-value	Effect Size	Model
0.774±0.114	0.770±0.102	0.114	0.040	`Llama3.2 1B`
0.477±0.122	0.398±0.104	<0.0001	0.729	`Llama3.2 3B`
0.594±0.287	0.191±0.143	<0.0001	2.319	`Llama3.3 70B`
0.548±0.142	0.322±0.139	<0.0001	1.615	`Llama3 8B`
0.265±0.236	0.076±0.071	<0.0001	1.664	`Mistral-7B-v0.3`
0.356±0.152	0.195±0.096	<0.0001	1.516	`OLMo2 13B`
0.323±0.121	0.226±0.080	<0.0001	1.105	`OLMo2 7B`
0.162±0.126	0.117±0.059	<0.0001	0.602	`OLMo 7B`
0.437±0.172	0.326±0.134	<0.0001	0.785	`Gemma2 2b`
0.360±0.191	0.156±0.082	<0.0001	1.905	`Gemma2 9b`

Table 5: Comparison of model separation for pairs with the same sense versus different senses, excluding pairs with the same sentence repeated. p-values are calculated using Welch (1947)'s t-test, with same sense being greater than different sense. Effect size is calculated using Cohen (1988)'s d.

Pairwise prompt:

User:
"<word>" is a word with many different uses and meanings. Do the following two sentences use “<word>" in the same way? Respond with "Yes" or “No".

Sentence 1: <sentence1>
Sentence 2: <sentence2>

Format your response as 'Response: [answer]’
Assistant: Response: <Yes|No>

Figure 7: Prompt for the pairwise sense comparisons.

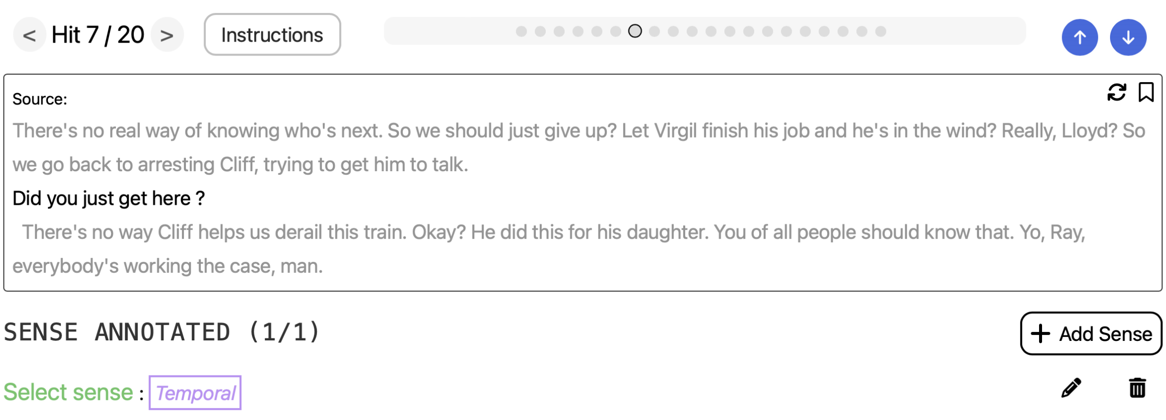

Figure 8: Annotation interface for labeling subtitle data.