WUGNECTIVES: Novel Entity Inferences of Language Models from Discourse Connectives

Daniel Brubaker **William Sheffield** **Junyi Jessy Li** **Kanishka Misra**
Department of Linguistics
The University of Texas at Austin
{dbrubaker, sheffieldw, jessy, kmisra@utexas.edu}

Abstract

The role of world knowledge has been particularly crucial to predict the discourse connective that marks the discourse relation between two arguments, with language models (LMs) being generally successful at this task. We flip this premise in our work, and instead study the inverse problem of understanding whether discourse connectives can inform LMs about the world. To this end, we present WUGNECTIVES, a dataset of 8,880 stimuli that evaluates LMs' inferences about *novel* entities in contexts where connectives link the entities to particular attributes. On investigating 17 different LMs at various scales, and training regimens, we found that tuning an LM to show reasoning behavior yields noteworthy improvements on most connectives. At the same time, there was a large variation in LMs' overall performance across connective type, with *all* models systematically struggling on connectives that express a concessive meaning. Our findings pave the way for more nuanced investigations into the functional role of language cues as captured by LMs.

We release WUGNECTIVES at https://github.com/sheffwb/wugnectives

1 Introduction

Discourse connectives such as *but, moreover, although, because,* etc., are central to producing and comprehending natural language. As such, the task of successfully predicting a discourse connective, given two discourse arguments has been popular throughout computational linguistics research (Zhou et al., 2010; Biran and McKeown, 2013; Patterson and Kehler, 2013), having made its way to the evaluation of large language models (Pandia et al., 2021; Beyer et al., 2021). For instance, the connective that links (1a) to (1b) linearly is likely to be *because* rather than *although*.

(1) a. I prefer Dubai to New York.
b. I hate snowy winters.

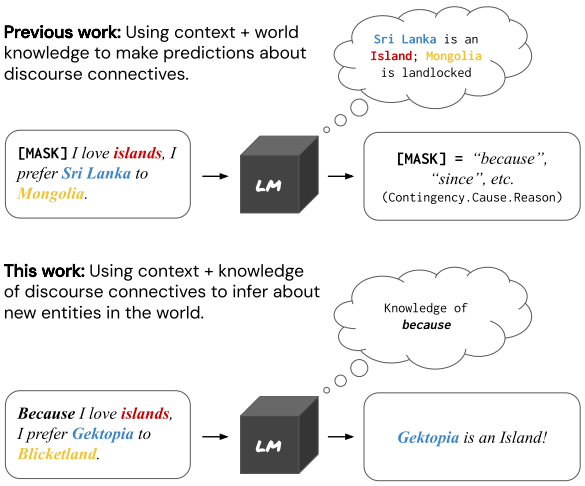

Figure 1: Past work has largely focused on the prediction of connectives given some input context, usually requiring access to world knowledge. We reduce this reliance by using novel entities, and analyze whether LMs can rely on their knowledge of the connectives themselves to make inferences about the world.

Prediction of the correct connective in the above example primarily requires consulting one's world knowledge—that Dubai does not have snowy winters (while New York does), and therefore (1b) can be a *reason* for why the speaker prefers Dubai over New York. There has been a history of using such knowledge to predict discourse relations in the absence of an explicit cue (Marcu and Echihabi, 2002; Pitler et al., 2009; Lin et al., 2009; Biran and McKeown, 2013; Li and Nenkova, 2014; Rutherford and Xue, 2014; Braud and Denis, 2015).

This paper studies the *inverse* problem: what kind of inferences can one make about the entities, given the discourse connective? If one were to instead consider (2), which replaces Dublin and New York with novel entities, X and Y, then depending on the connective used, one could draw different inferences about what X and Y could be.

(2) a. I prefer X to Y.
b. I hate snowy winters.

For instance, if we were to use *because*, then one could infer that X does not have snowy winters. On the other hand, if we use *however*, then this inference is reversed—that despite hating snowy winters the speaker prefers X (over Y). In this manner, discourse connectives can serve as *cues* to meaning/world knowledge (Elman, 2004).

In many ways, modern LLMs have been regarded as models that have finally "grasped" language (Piantadosi, 2023; Futrell and Mahowald, 2025). Their use of discourse connectives in modern AI-generated writing has similarly progressed far from older LMs (Ko and Li, 2020). But there is a difference between being able to *use* connectives correctly given known properties of concepts involved, *v.s.* fully *understand* their functional meaning and making correct *inferences* (Mahowald et al., 2024). We ask: **to what extent do LMs make abstract inferences about novel entities as licensed by connectives?** Answering this question allows us to make fine-grained contributions to the broader goal of characterizing how well aspects of meaning—in this case, attributes and relations of novel entities—can be learned from language exposure (Gelman, 2004; Lupyan and Lewis, 2019).

To answer this question, we propose WUGNECTIVES, a benchmark consisting of 740 utterances, each of which links novel entities to their attributes via the usage of specific discourse connectives. These utterances are embedded across 12 different prompt variations, amounting to 8,880 unique stimuli. Thus, while existing work focusing on discourse connectives effectively tests how world knowledge enables the prediction (or comprehension) of discourse connectives, this work flips this premise, and instead investigates how specific connectives can inform the model about entities mentioned in the arguments that the connectives operates over. By removing the aspect of world knowledge that is likely entrenched in a language model (LM)'s parameters, our investigation sheds light on whether LMs learn the *abstract* functional meaning of these connectives, in a manner that is independent of the content of the arguments they connect.

WUGNECTIVES consists of stimuli for 41 unique connective-forms, spanning 7 different senses, across a total of 3 different stimuli types, each focusing on different kinds of knowledge about novel entities—e.g., instantiation/category membership, temporal relations, and general attributes of entities such as "*being an island nation*". Using WUGNECTIVES, we evaluate 17 different open-source LMs at various parameter sizes, and training types (base, instruction tuning, and reasoning-based tuning).

We find LMs to show a great deal of variation in their abilities to infer about entities from connective usage. LMs generally performed above chance on cases where connectives expressed temporal meaning between novel events, or provided causal evidence for an entity's attributes (or lack thereof), and in some limited cases, when they expressed instantiation relations between entities. However, they consistently obtained chance-level performance on connectives that expressed *concession* between arguments—i.e., when a causal expectation raised on the basis of one argument is denied by the other. The systematicity of our results on such cases suggested that these classes of connectives (*although, even though, despite that*, etc.) pose a fundamental difficulty for LMs to reason about the semantic features of novel entities in context. Our results could not be explained by frequency of these connectives in internet corpora, suggesting a more nuanced, intractable reason at play. Finally, while we found no clear effect of scale or instruct-tuning, we did find reasoning-based tuning of LMs (á la Guo et al., 2025) to be beneficial, achieving the best performance across all connective senses (though still struggling on concession). Taken together, our findings pave the way for more nuanced investigation into the functional meanings of language cues, complementing traditional analyses of usage.

2 Background

By investigating LMs on their functional knowledge of discourse connectives to infer about the world, our work brings together two bodies of work in modern computational linguistics:

Modeling Connectives Discourse connectives, e.g., "because", "however", etc., are a class of words that mark the discourse (coherence) relations between two arguments, often unambiguously (Pitler and Nenkova, 2009). The role of world knowledge contained in the arguments of a relation has been of particular significance in prior work on recognizing implicit discourse relations. While this was initially signified using cartesian products of words in arguments (Marcu and Echihabi, 2002), there was widespread operationalization of this premise in several works since the introduction of the Penn Discourse Treebank (PDTB; Prasad et al., 2017)—e.g., Pitler et al.

(2009); Lin et al. (2009); Zhou et al. (2010); Biran and McKeown (2013); Patterson and Kehler (2013); Li and Nenkova (2014); Rutherford and Xue (2014); Braud and Denis (2015). Braud and Denis (2016) first showed evidence of improved discourse relation classification when word representations were informed by connectives.

Previous work evaluating language models' understanding of discourse connectives has largely focused on their ability to categorize, or predict connectives in context (Nie et al., 2019; Kim et al., 2020; Koto et al., 2021).

A number of works, such as Pandia et al. (2021), CoherenceGym (Beyer et al., 2021), and Cong et al. (2023) have evaluated LMs' sensitivity to infelicitous usage of connectives.

A common theme among these works is that world knowledge about the arguments being linked together is a necessary prerequisite to models' success. Our work abstracts away from this assumption by testing how connectives *cue* the meanings of the arguments that they link.

Learning about the world from Language (models) LMs offer an interesting avenue to investigate questions about how language exposure can give rise to semantic knowledge—a subject that has always received great theoretical interest (Landau and Gleitman, 1985; Waxman and Markow, 1995; Elman, 2004; Gelman, 2004; Lupyan and Lewis, 2019). Grounded in this motivation, a number of previous works have aimed to characterize the kinds of world knowledge that arises in LMs (Abdou et al., 2021; Misra et al., 2023; Ivanova et al., 2024, i.a.). While these works serve as catalogs of what kinds of world knowledge can be acquired from language exposure, the status of the cues that can give rise to them is less clear. Our work presents a step in this direction. By using connectives to link arguments involving "novel" entities, we center our focus on treating their function as the primary cue that LMs must rely on (in tandem with other parts of the input context) to reason about the entities' attributes and relations.

3 Designing WUGNECTIVES

In this section we describe the design decisions for WUGNECTIVES—in terms of how we operationalize "novel information", type of inference, and choice of connectives, finally culminating in our description of the stimuli.

Entity type	Surface forms
Bare plurals	*Wugs, Daxes, Feps, Geks, Blickets*
Events	*Wugfest, Daxday, Fepfestival, Gextravaganza, Blicketbash*
Locations	*Wugsville, Daxburgh, Fepopolis, Gektopia, Blicketland*

Table 1: Lists of nonce words used in our stimuli.

Nonce words To operationalize 'novel information' we use nonce words as novel entities in utterances expressing a proposition from which the LM has to infer their attributes (e.g., *is a leafy vegetable*) or their relations to another novel entity (e.g., *Wugs are Daxes*). The usage of nonce words for reasoning has now become commonplace in computational linguistics research (Misra et al., 2023; Eisenschlos et al., 2023; Rodriguez et al., 2025), and has been a longstanding tradition in cognitive psychology to mimic a scenario where the learner has little, if any, knowledge of the entity/property in question (Osherson et al., 1990; Gelman et al., 2010). Following Misra et al. (2023), we use pairs of nonce words in our stimuli, primarily due to two reasons. First, a number of our chosen connectives (see below) describe relations between two entities, and therefore, to maintain uniformity we use two novel entities throughout. Second, using two nonce words creates a notion of choice, and prevents the scenario where the LM simply uses co-occurrence information to make judgments about the only novel entity in question. Importantly, we counterbalance our nonce words throughout, to ensure that a systematic bias towards any one surface form leads to poor performance. Our novel entities range from events (for temporal connectives), to simple bare plurals (for instantiation, comparison, and contingency connectives) to locations (for comparison and contingency connectives). See Table 1 for the full list of nonce words.

Inference type The specific class of inferences we focus on are *entailments*, which means the correct answer is true regardless of the properties of the entities tested, allowing us to evaluate on inferences regardless of the characteristics of the nonce words used. We use these entailments to operationalize the notion of ground truth for all stimuli.

Connectives and relation senses We select a subset of relations from PDTB (Prasad et al., 2017) to build our stimuli from the following level 2 senses: `Expansion.Instantiation`, `Contingency.Cause`, `Comparison.Concession`, and `Temporal.Asynchronous`. A full list of the third-level senses and connectives can be found in Table 2.

Stimuli Type	Target Properties	Sense	Count	Connectives
Instantiation	*is a*	`Expansion.Instantiation.Arg2-as-instance`	100	*for example, for instance, in particular, specifically, such as*
Preference	*are leafy vegetables, are mammals, are fruits, are string instruments, are insects, is an island, is a college town, is a coastal city, has mountains nearby, has an equatorial climate*	`Contingency.Cause.Reason`	160	*as, because, for, since*
		`Contingency.Cause.Result`	240	*as a result, for example, for instance, so, therefore, thus*
		`Comparison.Concession.Arg1-as-denier`	98	*although, as much as, even though, though*
		`Comparison.Concession.Arg2-as-denier`	302	*although, but, even though, however, yet, nevertheless, though, while, despite that*
Temporal	*started earlier*	`Temporal.Asynchronous.Precedence`	110	*afterwards, as a result, before, consequently, even before, finally, hence, later, next, so, subsequently, then*
		`Temporal.Asynchronous.Succession`	130	*therefore, after, as soon as, because, earlier, even after, eventually, even though, once, previously, since, thereafter*

Table 2: Breakdown of our stimuli, showing stimuli type, the properties they target, number of unique stimuli, possible senses included in stimuli type, and the list of connectives used. **Note:** connectives that repeat are either poly-functional (e.g., while *because* is primarily a `Contingency` connective, it can also have `Temporal` senses), or have subtly different level 3 senses—e.g., whether *although* appears in the clause with the preference ("I prefer wugs to daxes") or a property ("I hate leafy vegetables") changes its `Comparison.Concession` sense).

Stimuli Type	Stimuli Example	Answer
Instantiation	*Daniel said, "I find feps such as blickets, to be awful." Can we conclude from this that blickets are feps? Answer either with Yes or No.*	Yes
Preference	*Lucy said, "Even though I hate leafy vegetables, I prefer feps to geks." From this, is it true that feps are leafy vegetables? Answer either with Yes or No.*	Yes
Temporal	*Erica said, "blicketbash occurred before gextravaganza." From this, which event started first? Answer either with blicketbash or gextravaganza and nothing else.*	blicketbash

Table 3: Examples of stimuli per stimuli type, along with their answers.

Stimuli Design Each individual stimulus consists of two parts: a premise and inference. The premise is a sentences which uses a discourse connective to specify the relation between two nonce words. The inference is either an explicit formulation of that entailed relation, it's logical opposite, i.e, a contradiction. Models are prompted with each inference framed as a question, and must respond appropriately based on whether the inference is entailed or contradictory. We categorize these stimuli into three families based on the types of inferences they license: Temporal, Instantiation, and Preference.

Instantiation: The connectives in these stimuli describe the **is-a** relationship between two discourse entities. Stimuli of this type follow the form of sentence (3), and systematically entail sentences in the form of (4).

(3) I like wugs, *for example*, daxes are nice.

(4) Daxes are wugs.

We used five `Expansion.Instantiation.Arg2-as-instance` connectives in these stimuli.

Preference: Here, connectives describe the relationship between a speaker's preferred entity and a liked or disliked property. These inferences can either be causal or concessive, in which case the licensed inferences respectively arise because of or in spite of expectations raised by the premise. For the purposes of stimuli design, connectives in this category are most saliently organized by their level-2 senses, `Contingency.Cause` (like *because*) and `Comparison.Concession` (like *although*).[1] For example, both (5a) and (5b) entail (6).

(5) a. *Because* I love leafy vegetables, I prefer wugs to daxes.
    b. *Although* I hate leafy vegetables, I prefer wugs to daxes.

(6) Wugs are leafy vegetables.

Additionally, the use of 'love' and 'hate' can be swapped to change the polarity of the entailment. That is, the first nonce[2] will not have the property, as in (7a) and (7b), both of which entail (8).[3]

[1] However, there is variety among third-level senses. See Table 2 for more details.

[2] Inferences about the second entity's relation to the property, while occasionally salient, are indeterminate in their entailment status and were subject of much debate amongst the authors. We do not make any claims about whether or not these inferences are implicatures or entailments, and leave them out of our dataset.

[3] Notably, this being an entailment rather than an implica-

(7) a. *Because* I hate leafy vegetables, I prefer wugs to daxes.
b. *Although* I love leafy vegetables, I prefer wugs to daxes.

(8) Wugs are **not** leafy vegetables.

The prompted questions regarding inferences where the nonce does not have the given property (such as (8)) are identical in form to those where they property applies, but the correct answer changes from "Yes" to "No." This is the general form of preference stimuli, though a few other variations of these stimuli are present as well. The preference part of the dataset contains both fronted and non-fronted connectives, where naturally possible. Additionally, the connective can occur in the clause with the preference rather than the property while still licensing the same inference (e.g. "*Although* I prefer wugs to daxes, I hate leafy vegetables.", which still entails (8)).

Overall, we use 20 unique connectives for these stimuli, with 10 each in `Contingency.Cause` and `Comparison.Concession` senses (Table 2). We use a total of 10 unique entity properties—5 mapped to bare plurals, and 5 mapped to locations.

Temporal: Here, connectives describe the temporal order of events. We include connectives from both `Temporal.Asynchronous.Precedence` (9a) and `Temporal.Asynchronous.Succession` (9b).

(9) a. Wugfest happened *even before* Daxday took place.
b. Daxday happened *once* Wugfest took place.

Both sentences (9a) and (9b) entail (10).

(10) Wugfest started before Daxday.

More explicitly, `Precedence` connectives license inferences where the entity in Arg1 starts before that in Arg2, and `Succession` connectives license inferences where the entity in Arg2 starts before Arg1. This does not directly map to the first entity linearly, as fronting the connective changes the order of the underlying arguments while preserving the licensed inference, as in (11), which still licenses (10).

(11) *Once* Wugfest took place, Daxday happened.

Additionally, we randomly vary the verbs used to say that each event came to pass between *happened*, *took place*, and *occurred*. These verbs are neutral with respect to a start time, and the increased variety ensures both naturalistic stimuli and helps isolate connective-based pragmatic reasoning from over-reliance on the surface form. Unlike Instantiation and Preference stimuli, the inferences in this family are prompted in open-ended questions (e.g., "Which event started first?"). Accordingly, this category does not contain contradictory inferences, and the responses we measure from models are the names of the two given events. We use a total of 24 temporal connectives, 12 each in `Temporal.Asynchronous.Precedence` and `Temporal.Asynchronous.Succession`.

Prompt Templates We embed our premise and inference pairs in 12 different prompt templates per stimuli type. Our templates are formatted in terms of a narrative dialogue where a named speaker says the premise, which is then followed by a question that targets the inference, and an instruction to guide the model to possible answers (Yes/No for Instantiation and Preference, or the name of one of the two events for Temporal). Variation in our prompts are paraphrases of the questions—e.g., replacing "From this, which event started first?" with "From what Erica said, which of the two events began first?" (see table 3). A detailed list of prompt templates is shown in table 5 and table 6 in the Appendix. After applying our prompt templates, we end up with a total of 8,880 stimuli.

4 Experimental Setup

Models We evaluate on three primary model families: Qwen-2.5 (Yang et al., 2025), OLMo-2 (Walsh et al., 2025), and Llama 3.1 (Grattafiori et al., 2024), across multiple different scales (when possible) in terms of parameter counts. For Qwen-2.5, we evaluate at 5 different scales ranging from 500M—14B parameters, for OLMo-2 we evaluate its 1B and 7B variants, and for Llama 3.1, we evaluate on its 8B variant. For each model we include its instruct tuned as well as non-instruct tuned (which we refer to as "base") versions. Additionally we also evaluate on a distilled version of the DeepSeek R1 model (Guo et al., 2025), where the authors

ture depends on the property being most saliently organized into a binary category. Consider the following grounded example with a gradable property: "Although I love tall buildings, I prefer Austin to New York. Austin does have tall buildings, just not as many as New York." The second sentence cancels the inference, indicating it is an implicature rather than an entailemnt..

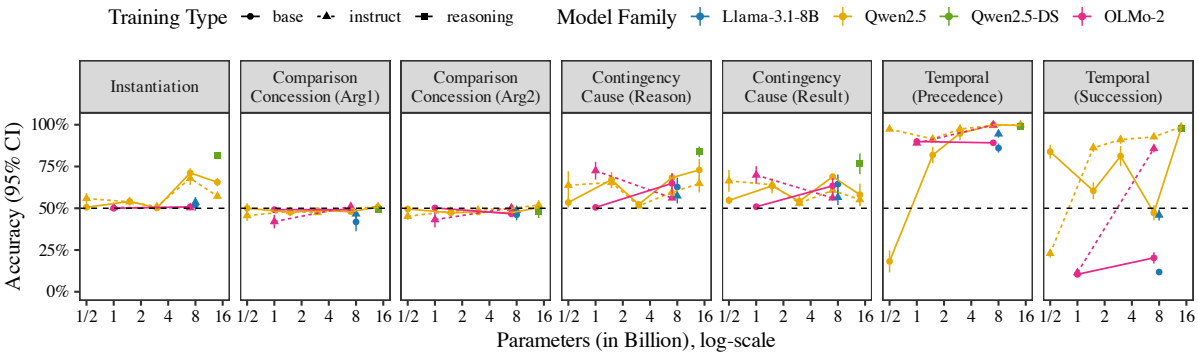

Figure 2: Accuracy of LMs across connective senses. The black dashed line indicates chance performance (50%). Error bars indicate 95% confidence intervals measured across connectives and prompt variation.

fine-tuned Qwen-2.5-14B on reasoning traces generated from the larger DeepSeek R1 model. We refer to this model as Qwen2.5-DS. In total, we test on 17 models: 10 Qwen-2.5 LMs, 4 OLMo-2 LMs, 2 Llama LMs, and 1 Qwen2.5-DS LM. Table 7 (appendix) shows metadata of each model.

Answer Extraction For Qwen2.5-DS, we prompt the LM to generate its responses and include its answer in the standard `\boxed{}` format. There were a cases where the LM did not do this—for these, we used the pipeline described in §G to extract predictions.

For base and instruct tuned models, we extracted the answer following recent work (Rodriguez et al., 2025): in the case of preference and instantiation stimuli, since we had a Yes/No question in our stimuli, we extracted the probabilities of a variety of surface forms for Yes and No (i.e., case variation and space prefixing), and normalized them to get the relative probabilities of 'Yes' and 'No'. We then chose the form with the greatest probability as the model's response. We performed the same process for our temporal stimuli, but instead restricted to retrieve the probabilities of the pair of 'Event' nonce words (see Table 1) in our stimuli.

Measurement We primarily report accuracy as our main performance metric, calculated as the proportion of time the correct response was produced by the model using our extraction strategy described above. Across all levels of our analyses (overall, per-sense, per connective, etc.), we report 95% confidence intervals across different prompt templates to jointly characterize the effect of prompt variation. Since in all cases we evaluate models on their choice between two possible answers (Yes/No for Preference and Instantiation stimuli, and between two entities for Temporal stimuli), chance accuracy is 50%.[4]

5 Results and Analysis

We analyze our results along two main threads. We start by focusing on model performance by connective sense, diving deeper into salient patterns of model behavior on specific connectives or a class of connectives. We then turn to analyses that focus on external artifacts such as model scale, factors involved in the models' training regimen such as instruct tuning, or training models to perform reasoning (in the sense of DeepSeek-R1), and an analysis of model performance vs. connective frequency. Our main results across all these variables, broken down by sense, is shown in Figure 2.

5.1 By Discourse Sense

Only a few LMs captured entailments licensed by connectives across various senses. While LMs generally had above-chance accuracies on `Temporal`, `Contingency`, and in some cases, `Instantiation` connectives, they consistently struggled on `Concession`, suggesting a systematic lack of abstract understanding of this particular sense. Below we discuss more detailed results per sense:

Instantiation Apart from a select few cases (i.e., Qwen-2.5 models at or above 7B parameters), most LMs were at chance performance on Instantiation connectives, with the Qwen2.5-DS model performing the best (at 82% accuracy). Figure 3 shows results of models broken down by connective. We observe a notable amount of variation across instantiation connectives, with models particularly struggling on *for instance* (avg. accuracy of 53%) and *for example* (avg. accuracy of 63%).

[4]For temporal stimuli, we counterbalance the nonce words across all possible pairs such that a bias towards one will result in chance performance.

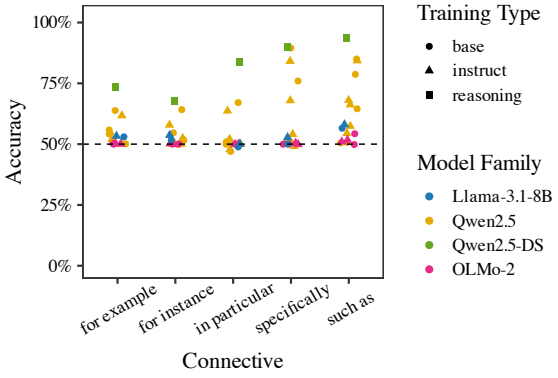

Figure 3: Mean accuracy (across prompts) of models by connective for the sense `Expansion.Instantiation`.

Concession All LMs in our experiments obtained chance-level performance on Concession connectives (`Comparison.Concession`).

That is, they seem to fundamentally struggle to infer entailments in cases where the function of the connective is to cancel or deny a causal relation expressed in one of the arguments. We observe this for both types of concession senses (`Arg1-as-denier` as well as `Arg2-as-denier`), suggesting a generally robust trend. Among various individual connectives, LMs seemed to especially struggle at *although* and *even though* when they are fronted (e.g., "*Although* I hate leafy vegetables, I prefer wugs to daxes.", where *although* is used in an `Arg1-as-denier` sense), oftentimes even obtaining below-chance performance. See Figure 5 for a full breakdown. In fact, models appear to struggle with concession connectives so much so that this seems to transfer over to when they are used in a different sense. Evidence for this is shown in Figure 4, where we see that the performance of the top performing LMs on succession connectives is compellingly greater than that on the succession sense of "*even though*", a connective that also has a concession sense. This reinforces the finding of LMs' notable weakness on concession connectives.

Contingency Models are generally better on contingency connectives than on the previous two, with 13/17 models being at least 5 percentage points above chance on both types of contingency connectives. At the same time, they are considerably far from perfect, with only the Qwen2.5-DS model showing accuracies above 75% for both senses.

Temporal Models showed largely inconsistent behavior in their performance on Temporal connectives, especially when observing the changes between their performance in the `Precedence` vs. `Succession` senses. A few exceptions to this trends were the Qwen2.5-DS model as well as larger variants of the instruction tuned Qwen-2.5 and OLMo models. In all other cases, models showed the greatest variability in their performance relative to that on other closely related senses discussed before.

	`Precedence`	`Succession`
Choose-first	92.5%	56.4%
Choose-recent	7.5%	43.6%

Table 4: Performance of positional-based heuristics. Choose-first linearly selects the first entity, and choose-recent selects the latest entity.

One explanation for this variability can come from reasoning about shallow heuristics a system might potentially rely on in "solving" the inference problem in these stimuli. Here we shed light on two such heuristics, both heavily dependent on linear position of entities in the LMs' input: 1) **Choose-First:** select the first entity in the premise as the answer, and 2) **Choose-Recent:** select the most recent entity in the premise. These heuristics do not have the same impact on the two Temporal senses, as seen from their accuracies in Table 4, suggesting different levels of difficulty for the two types of stimuli. For `Precedence` stimuli, applying the **Choose-first** heuristic can result in perfect performance for almost all connectives. This is because, except for 2 connectives (*before*, and *even before*), none of the other 10 `Precedence` connectives can be fronted—i.e., they always have to follow the same linear order of `{event1} <connective> {event2}`, and so simply extracting `{event1}` can result in the appearance of sophisticated reasoning. Since most `Succession` connectives can easily be fronted, neither heuristic seems to have a non-trivial role to play. Overall, it is difficult to determine if these heuristics are borne out in the LMs we analyzed, unless we perform a causal analysis of their mechanisms (Geiger et al., 2021), which we leave for future work.

5.2 By external artifacts

In general, frequency alone does not explain the variance in the model's performance, and there were no clear effects of scale or instruct tuning. Though preliminary, we did find reasoning-based tuning (i.e., the manner in which Qwen2.5-DS was tuned) to consistently show stronger performance, with the exception on `Concession` connectives as discussed in the previous subsection. Below we

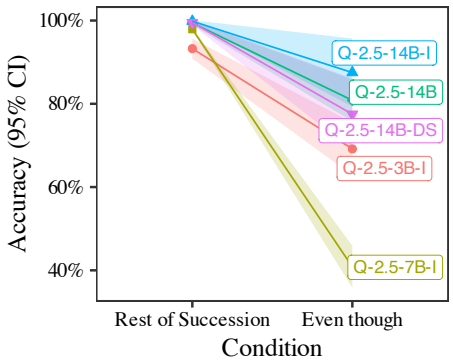

Figure 4: Accuracy of top five models on succession stimuli without *even though* compared with their performance on *even though*. Model names are abbreviated to save space. Q: "Qwen", I: "Instruct".

discuss these in greater detail. We make use of linear-mixed effects regression for our analysis of Scale, Instruction-Tuning, and Reasoning-based tuning, and report overall results here, while leaving particular details of the analysis in §F.

Frequency of Connective While discourse connectives are generally quite frequent in corpora, to what extent does their frequency relate to LMs' behavior in capturing their licensed inferences? To test this, we extract frequencies of our 41 unique connectives from Dolma corpus (Soldaini et al., 2024), and measure their correlation with LMs' performance per connective. We do not find a significant correlation between frequency and model performance (see Table 8 in the Appendix).[5]

Scale Overall, we do not find any notable, global effect of scale in our results. In many cases, larger models of the same family were no different than their smaller counterparts, a finding that was prevalent especially for the OLMo and Llama families), while in other cases there were only selective instances of a clear effect of scale—e.g., Qwen-2.5 Instruct models on `Precedence` and Qwen-2.5 base models on `Succession`. Results from linear mixed-effects regression analysis with an interaction term for number of parameters and sense as fixed effects, with random effects for model and prompt templates corroborated our findings ($\beta_{\text{params}} = 0.008, p = .10$).

Instruction Tuning Instruction-tuned models barely showed any improvements over their base counterparts. There were scattered exceptions to this trend—e.g., OLMo 2 1B on `Contingency` connectives, Qwen-2.5-500M on `Precedence` connectives (though the opposite effect was found in `Succession`), most Qwen-2.5 models and OLMo 2 7B on `Succession` connectives. However, due to the absence of a consistent pattern, it is not yet clear if instruction tuning shows any particular benefit in the context of making inferences from discourse connective usages. A linear mixed-effects regression analysis using an interaction between sense and instruction tuning (coded as a binary fixed-effect—1 if present, 0 otherwise), to predict accuracy, with random effects for prompt template and model yields results consistent with this conclusion ($\beta_{\text{instruct}} = -0.004, p = .826$).

Reasoning-based Tuning While we did not have as many cases of "Reasoning-based" tuning (in the sense of DeepSeek-R1) as we did Instruction Tuning, we do find evidence—preliminary though it might be—of reasoning-based tuning improving models' abilities to make novel entity inferences from discourse connectives. The Qwen2.5-DS model consistently outperformed its base and instruction tuning variants in all except the concession class of connectives. A linear-mixed effects regression analysis on results from all three variants of Qwen 2.5 14B (base, instruct, reasoning), with an interaction between training type and sense, and random effects of prompt template yielded corroborating evidence ($\beta_{\text{reasoning vs. base}} = 0.06, p < .001; \beta_{\text{reasoning vs. instruct}} = 0.08, p < .001$).

6 Conclusion

We present WUGNECTIVES, a dataset that sheds light on LMs' ability to reason about their knowledge discourse connectives in order to make inferences about novel entities. In doing so, we complement a long-standing body of work that has primarily treated connectives—rather than world knowledge—as the main target of prediction. While we pursued a number of different analyses, our most salient conclusion was about *all* tested models' systematic failure on *concession* connectives. In addition, we failed to find any effect of connective frequency, scaling, or even instruction tuning on LMs' performance, though there was preliminary evidence in favor of reasoning-based tuning, opening up future analyses of the linguistic properties of "reasoning models" (Guo et al., 2025).

[5] As a caveat, it is intractable to track what sense of a connective was being used in a corpus as large as Dolma, and therefore these frequencies can be seen as the estimates for the upper-bound of those used in their actual, precise senses.

Overall, we advocate for more investigations into the functional meanings of even the most simplest of linguistic cues, to fully catalog how language exposure enables semantic learning.

7 Limitations

There are a number of limitations to this work:

Reliance on connective *alone* While our aim in this work is to isolate—to the extent that we can—the reliance of LMs to the connectives alone and make inferences about the "world", our stimuli are not fully devoid of non-connective meaning. That is, LMs must still have to critically rely on the meanings of other words in the context (e.g., 'love', 'hate', 'prefer' in the case of preference stimuli). Fully teasing apart non-connective meanings is nearly impossible, so we relied on the assumption that models can consistently reason over these.

Gradability of properties Another key assumption we've made in the preference stimuli is that the properties in question are non-gradable (to argue that we've captured entailments rather than implicatures—which are much easily cancelable). While this might be true for certain properties (e.g., *is an island, is a college town*, it is up for debate if any property is truly binary. At the same time, there is active discussion about whether category membership and "goodness of exemplar"/typicality (the metric people use to argue that a property is graded, e.g., a *platypus* is less "mammal-like" than *bear*) are separate attributes of categories (Hampton, 2007), with category membership being more binary-like than typicality.

How novel are our novel words? Since we are using surface forms whose subtokens exist in the LMs' vocabulary, it is difficult to truly deem the full word as a "novel" word. One solution is to insert novel tokens, though using them requires one to fine-tune the model, which makes our analysis especially intractable, since we do not necessarily have contexts to train the model on. Furthermore, we have counterbalanced our nonce words in our stimuli such that any particular bias a model might have picked up on will result in chance level performance. We have additionally verified that models do not necessarily do this in the concession results, ruling out the possibility that they are failing due to a particular bias in their embedding space for these words.

Caveats in model comparison A caveat to the takeaway that reasoning models like Qwen2.5-DS are better is that it is non-trivial to directly compare across the three classes (base, instruct, reasoning), since the Qwen2.5-DS is expected to produce a set of reasoning traces before generating its output whereas models in the other two classes are directly queried for answers.

Larger models While we do not test larger models, we have been careful in drawing conclusions in our work, sticking to the set of models tested here. Since WUGNECTIVES is model agnostic, we will open our benchmark to anyone who wishes to test a model on it. We consider 17 models, with the kinds of variance in properties we have, to be a reasonable number of models to test.

Acknowledgments

This work was partially supported by NSF grant IIS-2145479 and a grant from Open Philanthropy. The authors acknowledge the Texas Advanced Computing Center (TACC)[6] at The University of Texas at Austin for providing computational resources that have contributed to the research results reported within this paper.

References

Mostafa Abdou, Artur Kulmizev, Daniel Hershcovich, Stella Frank, Ellie Pavlick, and Anders Søgaard. 2021. Can language models encode perceptual structure without grounding? a case study in color. In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 109–132, Online. Association for Computational Linguistics.

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.

Or Biran and Kathleen McKeown. 2013. Aggregated word pair features for implicit discourse relation disambiguation. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 69–73, Sofia, Bulgaria. Association for Computational Linguistics.

Chloé Braud and Pascal Denis. 2015. Comparing word representations for implicit discourse relation classification. In *Proceedings of the 2015 Conference on*

[6] https://tacc.utexas.edu

Empirical Methods in Natural Language Processing, pages 2201–2211, Lisbon, Portugal. Association for Computational Linguistics.

Chloé Braud and Pascal Denis. 2016. Learning connective-based word representations for implicit discourse relation identification. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 203–213, Austin, Texas. Association for Computational Linguistics.

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.

Julian Martin Eisenschlos, Jeremy R. Cole, Fangyu Liu, and William W. Cohen. 2023. WinoDict: Probing language models for in-context word acquisition. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 94–102, Dubrovnik, Croatia. Association for Computational Linguistics.

Jeffrey L Elman. 2004. An alternative view of the mental lexicon. *Trends in cognitive sciences*, 8(7):301–306.

Richard Futrell and Kyle Mahowald. 2025. How linguistics learned to stop worrying and love the language models. *arXiv preprint arXiv:2501.17047*.

Atticus Geiger, Hanson Lu, Thomas Icard, and Christopher Potts. 2021. Causal abstractions of neural networks. *Advances in Neural Information Processing Systems*, 34:9574–9586.

Susan A Gelman. 2004. Learning words for kinds: Generic noun phrases in acquisition. *Weaving a lexicon*, pages 445–484.

Susan A Gelman, Elizabeth A Ware, and Felicia Kleinberg. 2010. Effects of generic language on category content and structure. *Cognitive psychology*, 61(3):273–301.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, et al. 2025. Deepseek-r1 incentivizes reasoning in llms through reinforcement learning. *Nature*, 645(8081):633–638.

James A Hampton. 2007. Typicality, graded membership, and vagueness. *Cognitive science*, 31(3):355–384.

Anna A Ivanova, Aalok Sathe, Benjamin Lipkin, Unnathi Kumar, Setayesh Radkani, Thomas H Clark, Carina Kauf, Jennifer Hu, RT Pramod, Gabriel Grand, et al. 2024. Elements of world knowledge (ewok): A cognition-inspired framework for evaluating basic world knowledge in language models. *arXiv preprint arXiv:2405.09605*.

Najoung Kim, Song Feng, Chulaka Gunasekara, and Luis Lastras. 2020. Implicit discourse relation classification: We need to talk about evaluation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5404–5414, Online. Association for Computational Linguistics.

Wei-Jen Ko and Junyi Jessy Li. 2020. Assessing discourse relations in language generation from GPT-2. In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 52–59, Dublin, Ireland. Association for Computational Linguistics.

Fajri Koto, Jey Han Lau, and Timothy Baldwin. 2021. Discourse probing of pretrained language models. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3849–3864, Online. Association for Computational Linguistics.

Barbara Landau and Lila R Gleitman. 1985. Language and experience: Evidence from the blind child.

Junyi Jessy Li and Ani Nenkova. 2014. Reducing sparsity improves the recognition of implicit discourse relations. In *Proceedings of the 15th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*, pages 199–207, Philadelphia, PA, U.S.A. Association for Computational Linguistics.

Ziheng Lin, Min-Yen Kan, and Hwee Tou Ng. 2009. Recognizing implicit discourse relations in the Penn Discourse Treebank. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 343–351, Singapore. Association for Computational Linguistics.

Gary Lupyan and Molly Lewis. 2019. From words-as-mappings to words-as-cues: The role of language in semantic knowledge. *Language, Cognition and Neuroscience*, 34(10):1319–1337.

Kyle Mahowald, Anna A Ivanova, Idan A Blank, Nancy Kanwisher, Joshua B Tenenbaum, and Evelina Fedorenko. 2024. Dissociating language and thought in large language models. *Trends in cognitive sciences*, 28(6):517–540.

Daniel Marcu and Abdessamad Echihabi. 2002. An unsupervised approach to recognizing discourse relations. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 368–375, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

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.

Allen Nie, Erin Bennett, and Noah Goodman. 2019. DisSent: Learning sentence representations from explicit discourse relations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4497–4510, Florence, Italy. Association for Computational Linguistics.

Daniel N Osherson, Edward E Smith, Ormond Wilkie, Alejandro Lopez, and Eldar Shafir. 1990. Category-based induction. *Psychological review*, 97(2):185.

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.

Gary Patterson and Andrew Kehler. 2013. Predicting the presence of discourse connectives. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 914–923, Seattle, Washington, USA. Association for Computational Linguistics.

Steven T Piantadosi. 2023. Modern language models refute chomsky's approach to language. *From fieldwork to linguistic theory: A tribute to Dan Everett*, 15:353–414.

Emily Pitler, Annie Louis, and Ani Nenkova. 2009. Automatic sense prediction for implicit discourse relations in text. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 683–691, Suntec, Singapore. Association for Computational Linguistics.

Emily Pitler and Ani Nenkova. 2009. Using syntax to disambiguate explicit discourse connectives in text. In *Proceedings of the ACL-IJCNLP 2009 Conference Short Papers*, pages 13–16, Suntec, Singapore. Association for Computational Linguistics.

Rashmi Prasad, Bonnie Webber, and Aravind Joshi. 2017. The penn discourse treebank: An annotated corpus of discourse relations. In *Handbook of linguistic annotation*, pages 1197–1217. Springer.

Juan Diego Rodriguez, Aaron Mueller, and Kanishka Misra. 2025. Characterizing the role of similarity in the property inferences of language models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 11515–11533, Albuquerque, New Mexico. Association for Computational Linguistics.

Attapol Rutherford and Nianwen Xue. 2014. Discovering implicit discourse relations through brown cluster pair representation and coreference patterns. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 645–654, Gothenburg, Sweden. Association for Computational Linguistics.

Luca Soldaini, Rodney Kinney, Akshita Bhagia, Dustin Schwenk, David Atkinson, Russell Authur, Ben Bogin, Khyathi Chandu, Jennifer Dumas, Yanai Elazar, Valentin Hofmann, Ananya Jha, Sachin Kumar, Li Lucy, Xinxi Lyu, Nathan Lambert, Ian Magnusson, Jacob Morrison, Niklas Muennighoff, Aakanksha Naik, Crystal Nam, Matthew Peters, Abhilasha Ravichander, Kyle Richardson, Zejiang Shen, Emma Strubell, Nishant Subramani, Oyvind Tafjord, Evan Walsh, Luke Zettlemoyer, Noah Smith, Hannaneh Hajishirzi, Iz Beltagy, Dirk Groeneveld, Jesse Dodge, and Kyle Lo. 2024. Dolma: an open corpus of three trillion tokens for language model pretraining research. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15725–15788, Bangkok, Thailand. Association for Computational Linguistics.

Evan 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, Allyson Ettinger, Michal Guerquin, David Heineman, Hamish Ivison, Pang Wei Koh, Jiacheng Liu, Saumya Malik, William Merrill, Lester James Validad Miranda, Jacob Morrison, Tyler Murray, Crystal Nam, Jake Poznanski, Valentina Pyatkin, Aman Rangapur, Michael Schmitz, Sam Skjonsberg, David Wadden, Christopher Wilhelm, Michael Wilson, Luke Zettlemoyer, Ali Farhadi, Noah A. Smith, and Hannaneh Hajishirzi. 2025. 2 OLMo 2 furious (COLM's version). In *Second Conference on Language Modeling*.

Sandra R Waxman and Dana B Markow. 1995. Words as invitations to form categories: Evidence from 12-to 13-month-old infants. *Cognitive psychology*, 29(3):257–302.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. 2025. Qwen2.5 technical report. *arXiv preprint arXiv:2505.09388*.

Zhi-Min Zhou, Yu Xu, Zheng-Yu Niu, Man Lan, Jian Su, and Chew Lim Tan. 2010. Predicting discourse connectives for implicit discourse relation recognition. In *Coling 2010: Posters*, pages 1507–1514, Beijing, China. Coling 2010 Organizing Committee.

A Detailed Prompt Templates

Table 5 shows detailed prompt templates as well as other prompt artifacts (premise and inference examples, names) for the preference and instantiation stimuli. Table 6 shows that information for temporal stimuli.

B Model Metadata

Table 7 shows details about each model in this work.

C Implementation details

All models' log-probabilities were extracted using `minicons` (Misra, 2022), on a cluster with 4 NVIDIA A40 GPUs. Most experiments were run on a single A40 GPU, with the exception of the 14B Qwen models and their reasoning versions (run on two).

D Breakdown by Individual Connective

We show plots for model results broken down per connective in this section. Figure 3 shows results for `Expansion.Instantiation` (main text); Figure 5 shows results for `Comparison.Cause`; Figure 6 shows results for `Contingency.Concessio n`; and Figure 7 shows results for `Temporal` connectives.

E Frequency Analysis

Table 8 shows pearson's correlation between models' accuracies and connective frequencies. We find no evidence of positive correlation for any model.

F Linear Mixed-Effects Regression analysis

We describe our linear mixed-effects regression analysis to understand the effect of Scale, Instruction Tuning, and Reasoning-based Tuning on results. All analyses were conducted using the `lmerTest` library to specify the model, and the `car` library to perform significance tests for interaction effects.

Scale We divide the parameter counts (`params`) by `1e+9` and use the following formula, where `sense` specifies the sense of the connectives:

$$\texttt{accuracy} \sim \texttt{params} \times \texttt{sense} + (1 \mid \texttt{family}) + (1 \mid \texttt{prompt\_template})$$

Instruct tuning We code the instruct variable to be 1 if the model was instruct-tuned and 0 otherwise, discarding the reasoning model (Qwen2.5-DS) from this analysis (since it was only applicable for one model class). We use the following formula:

$$\texttt{accuracy} \sim \texttt{instruct} \times \texttt{sense} + (1 \mid \texttt{model}) + (1 \mid \texttt{prompt\_template})$$

Reasoning-based Tuning We perform this analysis only for the Qwen2.5-14B base class, and sumcode the `training_mode` variable (base, instruct, reasoning). We use the following formula:

$$\texttt{accuracy} \sim \texttt{training\_mode} \times \texttt{sense} + (1 \mid \texttt{prompt\_template})$$

G Algorithmic extraction of reasoning model responses

In all cases, we set the model temperature to be 0.6, as recommended in its model card (https://huggingface.co/deepseek-ai/DeepSeek-R1-Distill-Qwen-14B). After extracting the model's generations, we perform the following preprocessing steps.

1. For consistency, remove all of the following characters: * \n ' ( ).
2. Remove all instances of the phrases "answer:" and "the answer is"
3. If `boxed\{` or `boxed{` is in the text, return the answer within curly braces.
4. Set the trace to all lowercase.
5. Determine if the trace is a temporal answer or not by checking for the presence of any of the temporal nonces, or two misspellings ("blicktash", "bicketbash")
6. For non-temporal traces: If the words "yes" or "no" appear in the first or last three or two characters respectively, return that as the answer.
7. For temporal traces: return the nonce that appears either as the first or last n characters, where n is the length of the nonce in characters.

H License

We plan to release WUGNECTIVES with an MIT license.

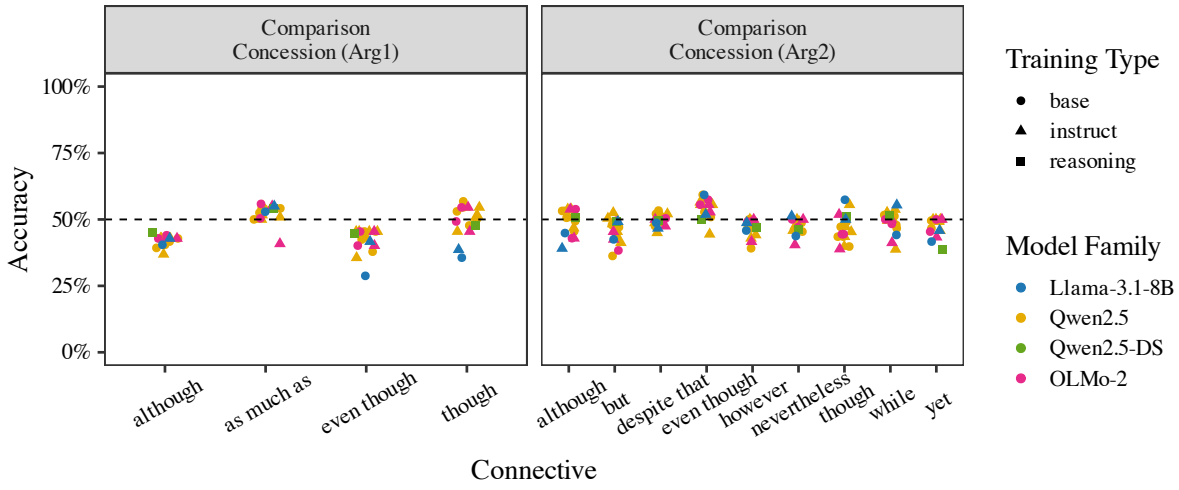

Figure 5: Results on Comparison connectives.

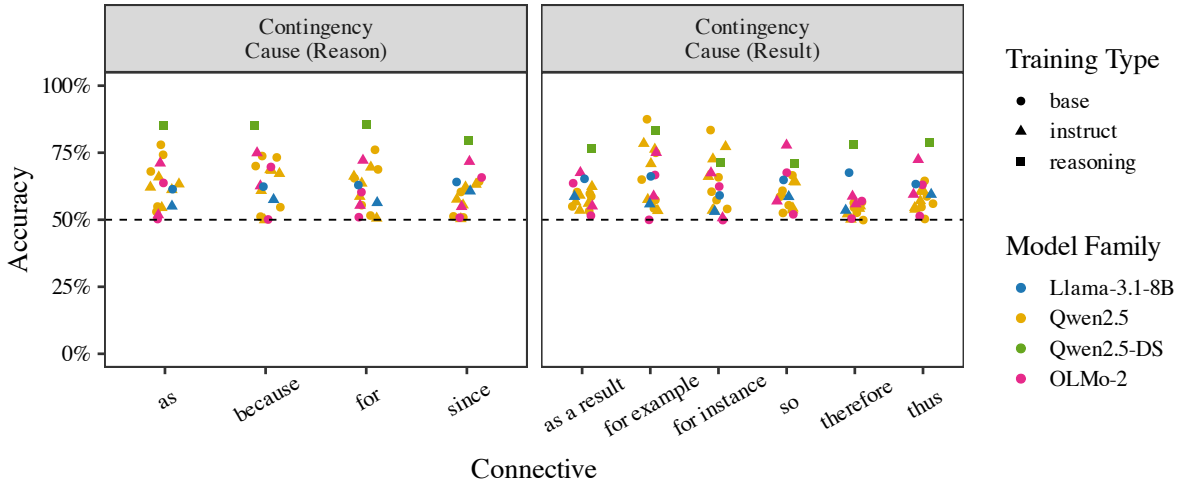

Figure 6: Results on Contingency connectives.

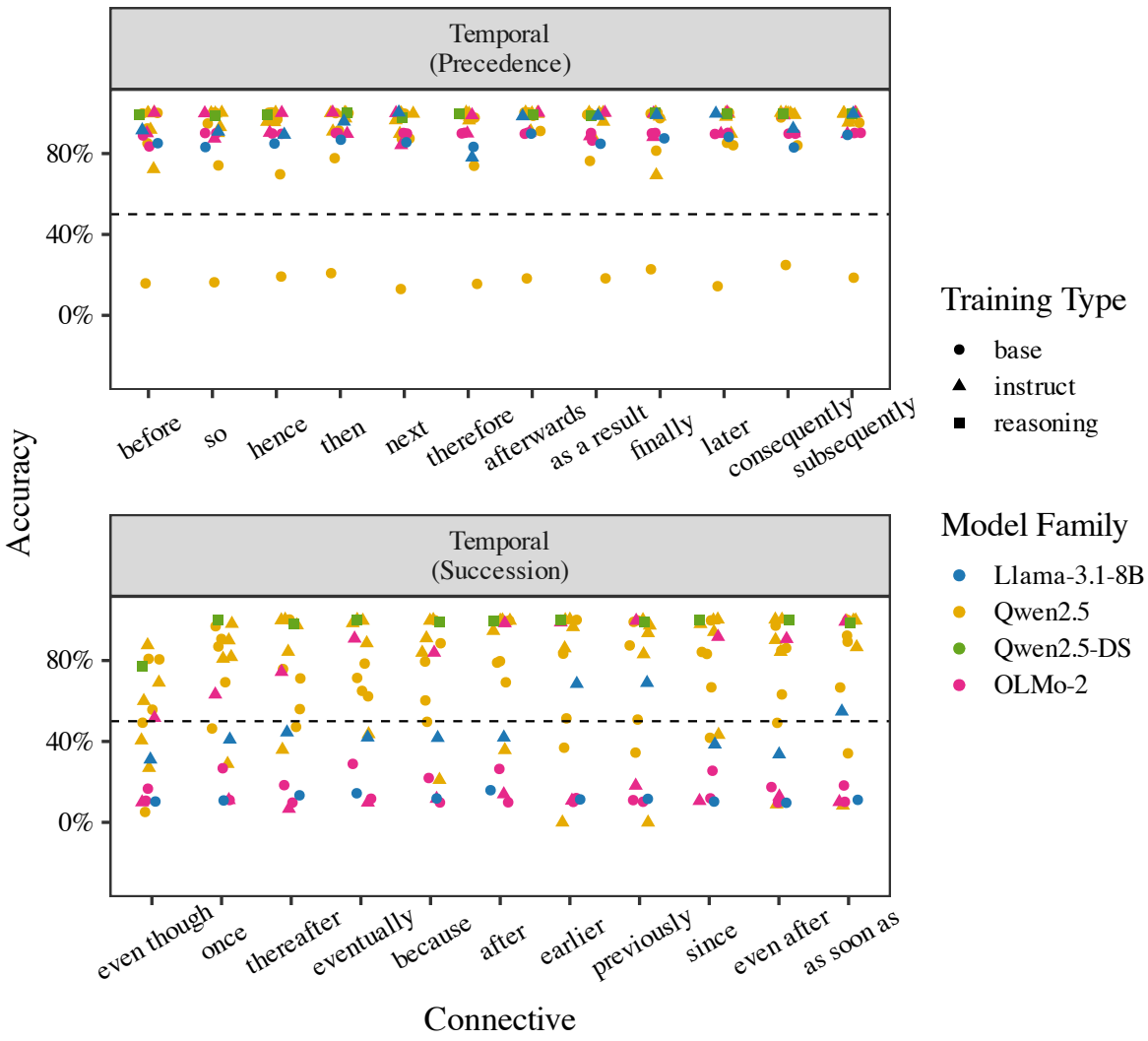

Figure 7: Results on Temporal connectives.

`Name` Options	Emily, Lucy, Adam, John, Cameron, Erica, Megan, David, Jessica, Daniel
`premise` Example	**Preference:** Although I hate islands, I prefer wugsville to daxburgh. **Instantiation:** I like wugs, for example, daxes are nice.
`inference` Example	**Preference:** wugsville is an island **Instantiation:** daxes are wugs
Templates	{Name} said, "{premise}" From this, is it true that {inference}? {Y/N instruction} {Name} said, "{premise}" Does this mean that {inference}? {Y/N instruction} {Name} said, "{premise}" Can we conclude from this that {inference}? {Y/N instruction} {Name} said, "{premise}" Does this suggest that {inference}? {Y/N instruction} {Name} said, "{premise}" Can we say from this that {inference}? {Y/N instruction} {Name} said, "{premise}" Can we conclude from what {Name} said that {inference}? {Y/N instruction} {Name} said, "{premise}" Can we say from what {Name} said that {inference}? {Y/N instruction} {Name} said, "{premise}" Does {Name} mean that {inference}? {Y/N instruction} {Name} said, "{premise}" Does what {Name} said suggest that {inference}? {Y/N instruction} {Name} said, "{premise}" If you heard this, would you think that {inference}? {Y/N instruction} {Name} said, "{premise}" If you heard {Name}, would you think that {inference}? {Y/N instruction} {Name} said, "{premise}" From what {Name} said, do you think that {inference}? {Y/N instruction}

Table 5: Prompt variation for **Preference** and **Instantiation** Stimuli. In all cases, the variable {Y/N instruction} is always "Answer either with Yes or No."

`Name` Options	Emily, Lucy, Adam, John, Cameron, Erica, Megan, David, Jessica, Daniel
`event` Options	Wugfest, Daxday, Gextravaganza, Blicketbash, Fepfestival
`premise` Example	event1 occurred. Thereafter, event2 took place.
Templates	{Name} said, "{premise}" From this, which event started first? {Answer Instruction} {Name} said, "{premise}" From this, which event started earlier? {Answer Instruction} {Name} said, "{premise}" From this, which event began first? {Answer Instruction} {Name} said, "{premise}" From this, which event began earlier? {Answer Instruction} {Name} said, "{premise}" From this, which of the two events began first? {Answer Instruction} {Name} said, "{premise}" From this, which of the two events began earlier? {Answer Instruction} {Name} said, "{premise}" From what {Name} said, which event started first? {Answer Instruction} {Name} said, "{premise}" From what {Name} said, which event started earlier? {Answer Instruction} {Name} said, "{premise}" From what {Name} said, which event began first? {Answer Instruction} {Name} said, "{premise}" From what {Name} said, which event began earlier? {Answer Instruction} {Name} said, "{premise}" From what {Name} said, which of the two events began first? {Answer Instruction} {Name} said, "{premise}" From what {Name} said, which of the two events began earlier? {Answer Instruction}

Table 6: Prompt variation for **Temporal** stimuli. {Answer Instruction} is always "Answer either with event1 or event2 and nothing else."

Family	HuggingFace Identifier	Parameters (in billion)	Training Type
Llama-3.1-8B	meta-llama/Meta-Llama-3.1-8B	8	base
	meta-llama/Meta-Llama-3.1-8B-Instruct	8	instruct
Qwen2.5	Qwen/Qwen2.5-0.5B	0.5	base
	Qwen/Qwen2.5-0.5B-Instruct	0.5	instruct
	Qwen/Qwen2.5-1.5B	1.5	base
	Qwen/Qwen2.5-1.5B-Instruct	1.5	instruct
	Qwen/Qwen2.5-3B	3	base
	Qwen/Qwen2.5-3B-Instruct	3	instruct
	Qwen/Qwen2.5-7B	7	base
	Qwen/Qwen2.5-7B-Instruct	7	instruct
	Qwen/Qwen2.5-14B	14	base
	Qwen/Qwen2.5-14B-Instruct	14	instruct
Qwen2.5-DS	deepseek-ai/DeepSeek-R1-Distill-Qwen-14B	14	reasoning
OLMo-2	allenai/OLMo-2-0425-1B	1	base
	allenai/OLMo-2-0425-1B-Instruct	1	instruct
	allenai/OLMo-2-1124-7B	7	base
	allenai/OLMo-2-1124-7B-Instruct	7	Instruct

Table 7: Family, Huggingface identifier, parameter counts, and training type for LMs evaluated in this work.

model	r^2	p
Qwen2.5-0.5B	-0.08	0.61
Qwen2.5-0.5B-Instruct	0.10	0.53
Qwen2.5-1.5B	-0.02	0.91
Qwen2.5-1.5B-Instruct	-0.19	0.25
Qwen2.5-14B	-0.16	0.33
Qwen2.5-14B-Instruct	-0.21	0.19
Qwen2.5-3B	-0.14	0.38
Qwen2.5-3B-Instruct	-0.26	0.11
Qwen2.5-7B	-0.05	0.75
Qwen2.5-7B-Instruct	-0.27	0.10
OLMo-2-1B	0.09	0.57
OLMo-2-1B-Instruct	0.14	0.40
OLMo-2-7B	0.15	0.37
OLMo-2-7B-Instruct	-0.22	0.18
Qwen2.5-DS	-0.15	0.37
Llama-3.1-8B	0.15	0.36
Llama-3.1-8B-Instruct	-0.08	0.62

Table 8: Correlation between LMs' connective-level accuracy and connective frequency, as estimated from the Dolma corpus (Soldaini et al., 2024).