Log Probabilities Are a Reliable Estimate of Semantic Plausibility in Base and Instruction-Tuned Language Models

Carina Kauf
Massachusetts Institute of Technology
ckauf@mit.edu

Emmanuele Chersoni
The Hong Kong Polytechnic University
emmanuele.chersoni@polyu.edu.hk

Alessandro Lenci
University of Pisa
alessandro.lenci@unipi.it

Evelina Fedorenko
Massachusetts Institute of Technology
evelina9@mit.edu

Anna A. Ivanova
Georgia Tech University
a.ivanova@gatech.edu

Abstract

Semantic plausibility (e.g. knowing that "the actor won the award" is more likely than "the actor won the battle") serves as an effective proxy for general world knowledge. Language models (LMs) capture vast amounts of world knowledge by learning distributional patterns in text, accessible via log probabilities (LogProbs) they assign to plausible vs. implausible outputs. The new generation of instruction-tuned LMs can now also provide explicit estimates of plausibility via Prompting. Here, we evaluate the effectiveness of LogProbs and basic Prompting to measure semantic plausibility, both in single-sentence minimal pairs (Experiment 1) and short context-dependent scenarios (Experiment 2). We find that (i) in both base and instruction-tuned LMs, LogProbs offers a more reliable measure of semantic plausibility than direct zero-shot Prompting, which yields inconsistent and often poor results; (ii) instruction-tuning generally does not alter the sensitivity of LogProbs to semantic plausibility (although sometimes decreases it); (iii) across models, context mostly modulates LogProbs in expected ways, as measured by three novel metrics of context-sensitive plausibility and their match to explicit human plausibility judgments. We conclude that, even in the era of prompt-based evaluations, LogProbs constitute a useful metric of semantic plausibility, both in base and instruction-tuned LMs.[1]

1 Introduction

Effective language use heavily relies on general world knowledge. To determine which sentence is the most appropriate response in a given situation, a language user often needs to establish whether the sentence (e.g., "The actor won the award") plausibly describes the world. In NLP, leveraging world knowledge is important both for specific tasks (such as information retrieval) and for general success of a language model during interactions with a user (such as establishing common ground).

Language models (LMs) are well-positioned to acquire many aspects of general world knowledge by capturing distributional patterns in their training data (Elazar et al., 2022; Kang and Choi, 2023). For instance, by observing that "actor" occurs more frequently with "award" than with "battle", the LM might implicitly learn that actors are more likely to win awards than battles. Thus, a simple word-in-context prediction objective can enable an LM to acquire vast amounts of world knowledge.

We focus on one particular way to assess general world knowledge: estimates of sentence plausibility. Plausible sentences conform with world knowledge whereas implausible sentences violate it; thus, the ability to distinguish plausible and implausible sentences is an indicator of underlying world knowledge capabilities. Plausibility judgments can be tested using both single sentences (e.g., "The actor won the award" > "The actor won the battle") and setups where plausibility depends on the context of the previous sentences (e.g., "The girl dressed up as a canary. She had a little beak." > "The girl was cute. She had a little beak.").

A quantitative metric that has been commonly used to evaluate world knowledge in LMs—including semantic plausibility—are the log probability scores (LogProbs) of the output under the model. LogProbs are relatively easy to compute and constitute a direct measure of model behavior (as opposed to more implicit metrics such as decod-

[1]Code and data are accessible at https://github.com/carina-kauf/llm-plaus-prob.

Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP, pages 263–277

ing probe accuracy; Li et al., 2021; Papadimitriou et al., 2022). However, LOGPROBS are sensitive to many different surface-level text properties, such as individual word frequency, output length, and tokenization schemes (Holtzman et al., 2021; Salazar et al., 2020; Kauf and Ivanova, 2023). Furthermore, distributional patterns are subject to the reporter bias: people typically communicate new or unusual information rather than trivial or commonly known facts (Gordon and Van Durme, 2013). Thus, the link between LOGPROBS and semantic plausibility is confounded by a variety of factors. The most common way to control for confounds influencing LOGPROBS is by leveraging the minimal pairs setup (Futrell et al., 2019; Warstadt et al., 2020; Hu et al., 2020; Aina and Linzen, 2021; Pedinotti et al., 2021; Sinha et al., 2022; Michaelov et al., 2023; Hu et al., 2024; Misra et al., 2024) and/or quantifying the effects of multiple contributing factors on the resulting score (Kauf et al., 2023),

With the rise of instruction-tuned LMs (Chung et al., 2022; Touvron et al., 2023; Almazrouei et al., 2023; Jiang et al., 2023), it has become possible to directly evaluate LM capabilities via targeted natural language PROMPTING (Li et al., 2022; Blevins et al., 2023). Thus, we ask: is explicitly prompting instruction-tuned LMs for semantic plausibility judgments more effective than using LOGPROBS-derived plausibility estimates? And how does instruction tuning affect the LOGPROBS estimates themselves?

On the one hand, PROMPTING might provide a better estimate of plausibility by filtering out influences of extraneous factors not mentioned in the prompt. Furthermore, instruction tuning might diminish the influence of those factors even at the level of LOGPROBS themselves, leading instruction-tuned models to perform better under either metric. On the other hand, initial direct comparisons of LOGPROBS and PROMPTING measures on different linguistic/semantic knowledge datasets revealed that PROMPTING may, in fact, systematically underestimate the model's internal knowledge by requiring the models not only to solve the task, but also to correctly interpret the prompt and to translate their answer into the desired output format (Hu and Levy, 2023; Hu et al., 2024).

As access to LOGPROBS for newer models becomes restricted, it is important to understand what knowledge can be accessed, and what knowledge is inaccessible to the experimenter if PROMPTING is the only way to interact with LMs. In addition, some researchers reported that instruction tuning decreases the utility of raw LOGPROBS in domains such as confidence judgments (Tian et al., 2023) and prediction of human reading times (Kuribayashi et al., 2024), a change that might or might not be compensated by superior PROMPTING performance and that needs to be acknowledged as the field is shifting toward instruction-tuned LMs.

In this paper, we provide a systematic comparison of semantic plausibility estimates in instruction-tuned LMs. We test LMs' knowledge of plausibility in single-sentence (Experiment 1) and contextualized scenarios (Experiment 2) and compare implicit (LOGPROBS-based) and explicit (PROMPTING-based) plausibility judgments. We find that:

1. LOGPROBS, while imperfect, are a more dependable measure of plausibility than naive zero-shot PROMPTING.

2. Instruction-tuning does not drastically alter LOGPROBS-derived plausibility estimates, although in certain cases they might become *less consistent* with human plausibility judgments compared to base model versions.

3. LOGPROBS can be used to effectively model the *contextual* plausibility of events and replicate key patterns of human plausibility-judgment behaviors in both base and instruction-tuned LMs.

2 Related Work

Evaluating single-sentence plausibility in LMs. In Experiment 1, we evaluate plausibility estimates for single sentences describing common events (Table 1). To evaluate plausibility, scholars traditionally tested NLP models with sentence pairs from psycholinguistic studies that differ for their degree of semantic plausibility (e.g. *The mechanic was checking the brakes* vs. *The journalist was checking the brakes*, from Bicknell et al., 2010): the models' goal is to guess which of the two sentences is the most plausible one (Lenci, 2011; Tilk et al., 2016; Chersoni et al., 2016, 2019, 2021).

Pedinotti et al. (2021) and Kauf et al. (2023) specifically tested event plausibility knowledge in non-finetuned LMs. Pedinotti et al. (2021) showed that LMs achieve correlation with human judgments on par with or better than traditional distributional models. Kauf et al. (2023) showed that Transformer-based models retain a considerable

Dataset	Plausible?	Possible?	Voice	Example	Source
EventsAdapt (AI, impossible)	Yes	Yes	Active	The teacher bought the laptop.	Fedorenko et al. (2020)
			Passive	The laptop was bought by the teacher.	
	No	No	Active	The laptop bought the teacher.	
			Passive	The teacher was bought by the laptop.	
EventsAdapt (AA, unlikely)	Yes	Yes	Active	The nanny tutored the boy.	
			Passive	The boy was tutored by the nanny.	
	No	Yes	Active	The boy tutored the nanny.	
			Passive	The nanny was tutored by the boy.	
DTFit (AI, unlikely)	Yes	Yes	Active	The actor won the award.	Vassallo et al. (2018)
	No	Yes	Active	The actor won the battle.	

Table 1: Example stimuli from the datasets used in Experiment 1. Names in parentheses indicate event participant animacy (AI = animate agent, inanimate patient; AA = animate agent, animate patient) and the plausibility type of the implausible sentences in the dataset (impossible vs. unlikely).

amount of event knowledge from textual corpora and vastly outperform the competitor models (i.e., classical distributional models and LSTM baselines). Nevertheless, both studies show LMs' generalization capabilities to novel experimental manipulations of the target sentences are limited and that LogProbs are affected by task-irrelevant information, such as the frequency of words within a target sentence.

Evaluating context-dependent linguistic judgments in LMs. In Experiment 2, we evaluate context sensitivity of LM plausibility estimates (Table 5). Initial work in this domain shows that LMs can modulate their probability estimates to accommodate a previously unlikely target word (e.g., *A peanut falls in love*) following a short licensing context (Michaelov et al., 2023; Hanna et al., 2023), results that are consistent with human data (Nieuwland and Van Berkum, 2006; Rueschemeyer et al., 2015). Nevertheless, probability-based judgments of LMs can also be *adversely* influenced by context, for example in cases where the context contains information that is not related to the task (for syntax: e.g., Sinha et al., 2022; for factual knowledge: e.g., Kassner and Schütze, 2020).

Comparing LogProbs and Prompting. The direct interaction with LMs through natural language prompts is exciting for many reasons, including the ability to run the exact same experiments on models and on humans (Lampinen, 2022). Nevertheless, Hu and Levy (2023); Hu et al. (2024) showed that the use of metalinguistic prompts for model evaluation may underestimate their true capabilities. They compared LMs' syntactic/semantic knowledge across four minimal sentence pair datasets and showed that, on average, direct probability measures were a better indicator of these knowledge types than answers to prompts (similar to us, they used *DTFit* as one of their datasets, but their prompts did not explicitly probe the notion of plausibility; thus, we chose to include *DTFit* in this work; see Appendix §B, Figure 6 for a more direct comparison).

Evaluating the alignment of instruction-tuned models with humans. Even though instruction-tuning has been claimed to better align the representations of LMs and those computed by the human brain (Aw et al., 2023), others show that it does not always help for the alignment at the behavioral level (Kuribayashi et al., 2024). However, the work in this domain is still sparse.

3 Experiment 1: Single-Sentence Plausibility Judgments

In this section, we test LMs' knowledge of semantic plausibility in *isolated sentences*. We compare implicit (LogProbs-based) and explicit (Prompting-based) judgments derived from the base and instruction-tuned versions of 3 state-of-the-art LMs. We also compare LM scores with human plausibility judgments.

3.1 Datasets

We use two curated sets of minimal sentence pairs ($n \sim 2000$ overall) adapted from previous studies (for an overview, see Table 1):

EventsAdapt. The *EventsAdapt* dataset (Fedorenko et al., 2020) is composed of 391 items, each of which includes (i) a plausible active sentence that describes a transitive event ("The teacher bought the laptop"), (ii) the implausible version of the same sentence, constructed by swapping the

Evaluation type	Example
LogProbs Score	{**The nanny tutored the boy.**, **The boy tutored the nanny.**}
Sentence Choice I	Here are two English sentences: 1) The nanny tutored the boy. 2) The boy tutored the nanny. Which sentence is more plausible? Respond with either 1 or 2 as your answer. Answer: {**1**, **2**}
Sentence Choice II	You are evaluating the plausibility of sentences. A sentence is completely plausible if the situation it describes commonly occurs in the real world. A sentence is completely implausible if the situation it describes never occurs in the real world. Tell me if the following sentence is plausible. The nanny tutored the boy. Respond with either Yes or No as your answer. Answer: {**Yes**, **No**}
Likert Scoring	You will be given a sentence. Your task is to read the sentence and rate how plausible it is. Here is the sentence: The nanny tutored the boy. How plausible is this sentence? Respond with a number on a scale from 1 to 7 as your answer, with 1 meaning "is completely implausible", and 7 meaning "is completely plausible". Answer: { **7**, **6**, **5**, **4**, **3**, **2**, **1** }
Sentence Judgment	Here is a sentence: The nanny tutored the boy. Is this sentence plausible? Respond with either Yes or No as your answer. Answer: {**Yes**, **No**}

Table 2: Example evaluation strategies. The prompts are extended and adapted from Hu and Levy (2023).

noun phrases ("The laptop bought the teacher"), and passive voice alternatives ("The laptop was bought by the teacher" and "The teacher was bought by the laptop"). The items fall into one of two categories: **a)** animate-inanimate items (AI 👤💻; "The teacher bought the laptop"), where the swap of the noun phrases leads to impossible sentences; and **b)** animate-animate ones 👤👤 (AA; "The nanny tutored the boy"), where role-reversed sentences have milder plausibility violations. Given these differences, we model the two subsets independently.

DTFit. The *DTFit* dataset (Vassallo et al., 2018) contains 395 items, each of which includes (i) a plausible active sentence that describes a transitive event ("The actor won the award"); (ii) a less plausible version of the same sentence, constructed by varying the inanimate sentence patient ("The actor won the battle").

3.2 Human Plausibility Judgments

For *DTFit*, participants answered questions of the form "How common is it for a {agent} to {predicate} a {patient}." (e.g. "How common is it for an actor to win an award?") on a Likert scale from 1 (very atypical) to 7 (very typical) (Vassallo et al., 2018). For *EventsAdapt*, participants evaluated the extent to which each sentence was "plausible, i.e., likely to occur in the real world" on a Likert scale from 1 (completely implausible) to 7 (completely plausible) (Kauf et al., 2023). For each sentence, we average judgments across the human participant pool to obtain a single score.

3.3 Model Plausibility Judgments

Models. We test the base and instruction-tuned versions of three popular autoregressive LMs: `Mistral` (Jiang et al., 2023), `Falcon` (Almazrouei et al., 2023), and `MPT` (MosaicML NLP Team, 2023), all of them with $7B$ parameters.

Metrics. We evaluate LMs using (i) LogProbs and (ii) several zero-shot Prompting methods (Table 2) (Hu and Levy, 2023). LogProbs are calculated as the sum of the log-probabilities of each token w_i in a sentence, conditioned on the preceding sentence tokens $w_{<i}$. In our main analysis, we evaluate LMs using four natural-language prompts (*Sentence Choice I/II, Likert Scoring* and *Sentence Judgment*; Table 2). These prompts were designed to explicitly query the LMs' knowledge of sentence *plausibility* and use either the same or similar instructions to the task that humans solved (see §3.2).[2] For all prompting methods except *Likert Scoring*, we compare the probabilities that models assign to ground-truth continuations (in **green**) over implausible continuations (in **red**). For *Likert Scoring*, we ask models to generate a number from a constrained set of answers, using the `outlines` Python library[3], and compare the generated scores for plausible vs. implausible sentences (the results remain consistent across free vs. constrained generation prompting, see SI §C, Figure 7). In our main experiment, all prompts are framed using the direct plausibility query "is plausible". Supplementary analyses show that this pattern of results remains consistent for alternative queries of plausibility, such as "makes sense" (SI §C, Figure 8) and

[2]Note that the *DTFit* dataset was included in Hu and Levy (2023) where it was evaluated using different models and different prompts. However, they did not explicitly query the models for estimates of *semantic plausibility*, but rather paraphrased the LMs' pretraining task, asking which word "is most likely to come next". We include an evaluation of our models on their best-performing prompt for *DTFit* as a supplementary analysis (SI §B, Figure 6).

[3]`https://github.com/outlines-dev/outlines`

"is likely" (SI §B, Figure 6).

Binary accuracy. For each item, we compare the scores/generations of the minimally different plausible and implausible sentence conditions, and compute the binary *accuracy* as the ratio of dataset items in which the LM/the human subject pool assigns a higher score to the plausible vs. the implausible sentence variant. The chance level is 50% for all benchmarks except *Sentence Judgment*, where, following Hu and Levy (2023), we compare the models' propensity to output the ground truth answer in both plausible and implausible settings, leading to a chance performance of 25%.

3.4 Results

Result 1: LogProbs *results are consistent across models, whereas* Prompting *is hit-or-miss.*

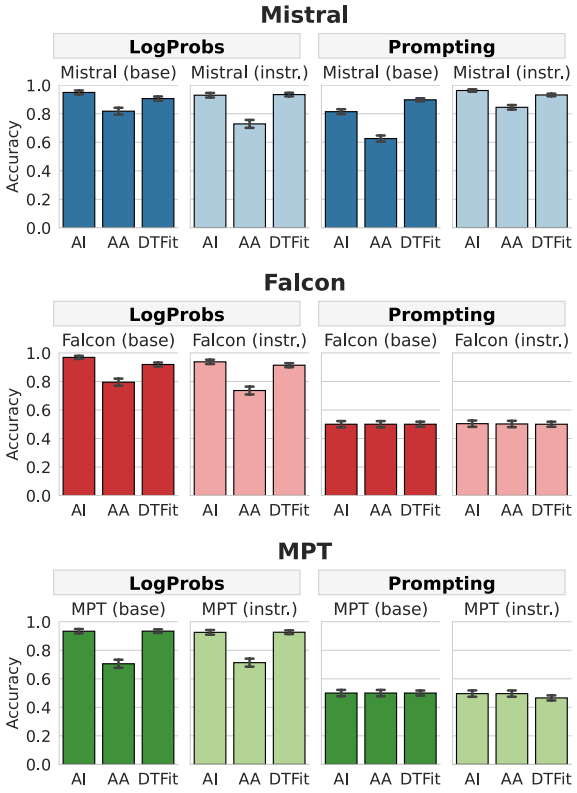


Figure 1: Results of sentence plausibility judgment performance across models and datasets, using implicit (LogProbs) measures vs. Prompting with the best-performing prompt (*Sentence Choice I*). Complete prompting results are shown in SI §A, Figure 5.

Across model architectures and plausibility datasets, LogProbs are an effective estimate of plausibility knowledge in both base and instruction-tuned LMs (Figure 1). Overall performance patterns across datasets—*DTFit*; *EventsAdapt, AI*; and *EventsAdapt, AA*—are consistent across models, with only minor performance differences. The results are also consistent with prior work (Kauf et al., 2023), showing a performance gap between AI sentences (easier) and AA sentences (harder).

Mistral (EventsAdapt, AA)	Base	Instruct
LogProbs	**0.82** (.02)	0.73 (.03)
Sentence Choice I	0.63 (.02)	**0.84** (.02)
Sentence Choice II	0.50 (.02)	0.50 (.02)
Likert Scoring	0.46 (.03)	**0.61** (.03)
Sentence Judgment	0.14 (.02)	**0.46** (.03)

Table 3: Results of model sentence plausibility judgment performance for `Mistral` on the *EventsAdapt, AA* sentence set shows brittleness of this method. Average performance and standard error around the mean are reported.

Prompting the LMs with our queries, by contrast, yielded inconsistent results. While `Mistral` showed above-chance performance for several prompts, `Falcon` and `MPT` performed at chance level for all prompts tested (for complete prompting results, see SI §A, Figure 5). Interestingly, even the base `Mistral` model performed above-chance on some prompts (*Sentence Choice I*), suggesting that model pretraining and/or architecture may be important for the prompt to work in an instruction-tuned model.

Prompts can be tuned to work well for a specific LM and task (Qin and Eisner, 2021; Pryzant et al., 2023; Chen et al., 2024). Even though we do not explore automatic prompt-optimization approaches in this study and instead test variations of the natural-language prompt that humans saw during the experiment (and which people interacting with these models may plausibly use when querying for semantic knowledge in LMs), we observed that certain ⟨prompt,model⟩ combinations indeed led to improved performance over LogProbs (Table 3). Despite this success, however, our comparison critically shows that the same prompt that is effective at tapping into plausibility knowledge in one model class (i.e., *Sentence Choice I* for `Mistral` models) need not be effective in tapping into the same knowledge in other models (Figures 1, 5). Likewise, we show that the same model that exhibits successful task performance when prompted in a certain way can exhibit poor performance when queried with slight variations on the same prompt (e.g., Table 3; see also Sclar et al., 2023). This brittleness of Prompting-based evaluations stands in contrast to the robustness of the model-agnostic LogProbs-based evaluation scheme of plausibility knowledge in LMs.

	Mistral		Falcon		MPT	
	Base	Instruct	Base	Instruct	Base	Instruct
AA	**0.82****	0.73	**0.79**	0.74	**0.71**	**0.71**
AI	**0.95**	0.93	**0.97***	0.94	**0.93**	**0.93**
DTFit	0.91	**0.93***	**0.92**	0.91	**0.93**	**0.93**

Table 4: LOGPROBS results across models and datasets. Significant differences from dependent t-tests between Base and Instruct models are marked with asterisks ($p < .05$: *; $p < .01$: **).

In fact, most of the prompting methods lead to chance-level performance or below-chance performance for most models (Figure 5), even though their log probabilities evidence substantial knowledge about what events are plausible vs. implausible. This result is in line with Hu and Levy (2023)'s finding of a competence-performance gap when probing models' metalinguistic judgments.

Result 2: LOGPROBS *in base and instruction-tuned LMs encode substantial plausibility knowledge but fall short of human performance.*

The LOGPROBS results in Figure 1 show that LMs acquire substantial plausibility knowledge from distributional linguistic patterns; all of them performing well above chance on the task. Nevertheless, they also consistently fall short of human performance: On *EventsAdapt (AI, impossible)*, all models were successful in distinguishing plausible and implausible sentences, even though all but one model (`Falcon base`) fell short of human accuracy of 1 (all Bonferroni-corrected $ps > .05$ except for `Falcon base`: $t = -2.14, p = .02$). On the more challenging *EventsAdapt (AA, unlikely)* subset, all models performed significantly worse than humans in distinguishing AA plausible from implausible events (human accuracy 0.95; all $ps < .001$). Lastly, the high task performance on *DTFit* shows that LMs can distinguish plausible and implausible AI event descriptions even when low-level distributional cues (like selectional preference restrictions) cannot be used to distinguish the minimal pairs. Despite this success, all models still fall short of human performance of 0.99 for this dataset at $ps < .001$.

Result 3: Instruction tuning can worsen LOGPROBS *sensitivity to semantic plausibility.*

Next, we zoom in on the comparison of LOGPROBS derived from base vs. instruction-tuned variants of the same model. Because instruction tuning constrains model behaviors to align with human-desired response characteristics (Zhang et al., 2023; Chia et al., 2023), it is reasonable to assume that the models' learned probability distributions align better with human expectations of plausible sequences than the base variant, which might be more susceptible to the reporting bias in textual corpora (Gordon and Van Durme, 2013).

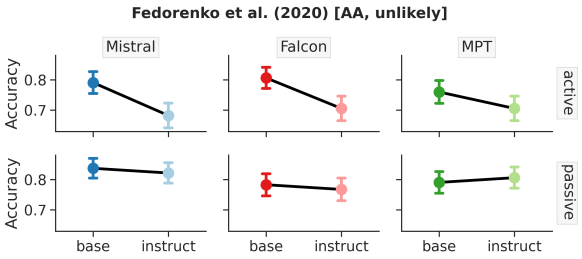


Figure 2: Base vs. instruct model performance in active and passive sentence pairs

A comparative analysis of the results of base and instruction-tuned model variants across architectures reveals no beneficial effect of instruction-tuning for gauging event plausibility through LOGPROBS measurements: In all but one instance do instruction-tuned models perform similar or even slightly worse than their corresponding base model (Table 4). Interestingly, the gap is most noticeable for the most challenging dataset, *EventsAdapt (AA, unlikely)*. An investigation of this difference shows that certain low-level features of the input may disproportionately affect the LOGPROBS that instruction-tuned models assign to word sequences: much of the performance difference is due to the instruction-tuned models' worse performance in discerning plausible and implausible active-voice sentences (see Figure 2). We quantify these effects by modeling accuracy in a generalized linear mixed-effects model (GLMM). The model uses LLM model class (`Mistral`, `Falcon`, `MPT`), model version (`base`, `instruct`), and voice (active, passive) as fixed effects, and items as random effects (for further GLMM model specification, see SI §D). We observed a main effect of model version ($\beta = 0.36, p < .001$) and a significant interaction between model version and active vs. passive voice ($\beta = -0.37, p < .01$).

This variance highlights the fact that even though direct measurements of model-derived string LOGPROBS in many cases encode task-relevant information (e.g., modeling of grammaticality, Warstadt et al. (2020), of N400 effects, Michaelov and Bergen (2020), etc.), they are additionally influenced by low-level features of the input (Pedinotti et al., 2021; Kauf et al., 2023).

Condition	Context sentence (optional)	Target sentence		
		Prefix	Tgt. word	Spill-over region
Control	The kids were looking at a canary in the pet store.	The bird had a little	beak	and a bright yellow tail.
SemAnom	Anna was definitely a very cute child.	The girl had a little	beak	and a bright yellow tail.
Critical	The girl dressed up as a canary for Halloween.	The girl had a little	beak	and a bright yellow tail.

Table 5: Sentence manipulations in the dataset by Jouravlev et al. (2019). Tgt. – Target.

4 Experiment 2: Context-Dependent Plausibility Judgments

Experiment 1 has shown that LogProbs are a reliable, albeit imperfect, metric for probing the plausibility of isolated sentences in LMs in both base and instruction-tuned models, whereas Prompting measures are brittle and can underestimate the degree of semantic plausibility knowledge LMs encode. However, most of the time, LMs (and humans) do not process sentences in isolation, but rather as part of a larger context. In Experiment 2, we therefore compare LM judgments of semantic plausibility in *short context-dependent scenarios*. Given the success of LogProbs over Prompting in Experiment 1, we focus on comparing LogProbs as measures of context-dependent sentence plausibility in base and instruction-tuned models. Specifically, we compare how the presence of (i) supporting or (ii) non-supporting but related single-sentence contexts modulates the LMs' LogProbs judgments. Additionally, we report results for the exact replication of the human study using *Sentence Judgment* prompts.

4.1 Dataset

To test the sensitivity of the LM plausibility judgments to discourse context effects, we use a dataset from language neuroscience, collected by Jouravlev et al. (2019). This dataset includes 100 items in three experimental conditions: a control condition (`Control`), in which the target sentence describes a plausible situation and the (optional) context sentence adds extra information; a semantically anomalous condition (`SemAnom`), in which the target sentence describes an implausible situation and the context sentence does not provide licensing information; and a critical condition (`Critical`), which shares the same target sentence with `SemAnom`, but here, the context sentence makes it plausible (see the examples in Table 5).

4.2 Metrics

We introduce three critical metrics to evaluate the models' context-aware plausibility judgments:

General Plausibility. This metric measures the propensity of models to assign a higher probability to plausible sentences than to minimally different implausible sentence variants when no influencing context is present (similar to §3). For every dataset item, we assign a model a hit in case

$$P(\text{target}_{\texttt{Contr.}}) > P(\text{target}_{\texttt{Crit.}}).$$

Context-Dependent Plausibility. This metric measures the ability of models to increase the probability they assign to an *a priori* implausible sentence in the presence of a licensing context. For every dataset item, we assign a model a hit in case

$$P(\text{target}_{\texttt{Crit.}}|\text{context}_{\texttt{Crit.}}) > P(\text{target}_{\texttt{Crit.}}).$$

Context Sensitivity. This metric measures the models' ability to *selectively* update sentence probabilities. For every dataset item, we assign a model a hit in case

$$P(\text{target}_{\texttt{Crit.}}|\text{context}_{\texttt{Crit.}}) > P(\text{target}_{\texttt{Crit.}}|\text{context}_{\texttt{Anom.}}).$$

4.3 Target region

For each metric, we evaluate model performance through the likelihood they assign either (i) a critical word within the target sentence or (ii) the target sentence as a whole. If a critical word consists of multiple tokens, we use the sum of the log likelihood scores of the word tokens. Whereas *critical/target word* likelihoods measure the ability of models to detect a contextually unexpected linguistic event, *target sentence* likelihood measures investigate whether implausibility is reliably reflected in the probability the models assign to tokens after encountering a semantically anomalous item, as well. This is because token likelihoods for plausible and implausible sentences are identical until the first contextually unlicensed word appears.

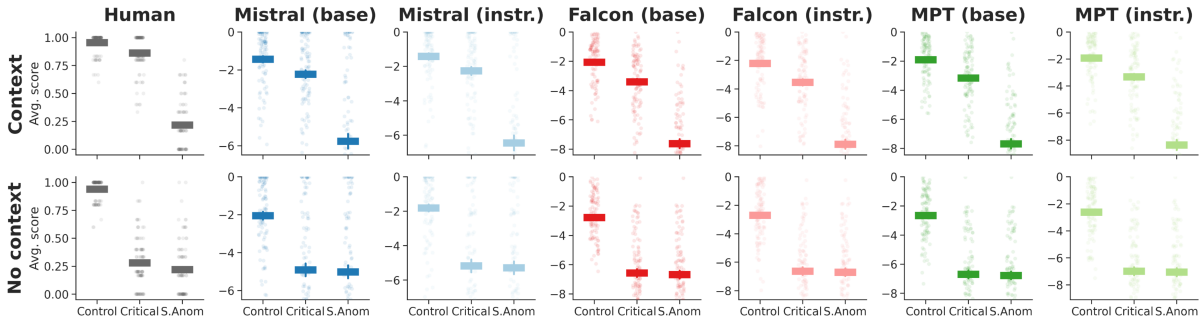


Figure 3: Target word LogProbs replicate patterns of human sentence sensibility judgments. Human data from Jouravlev et al. (2019). Bars indicate average plausibility of sentences (Human) and average target word log likelihoods (LMs). Dots represent individual sentence scores (averaged across the participant pool for Human).

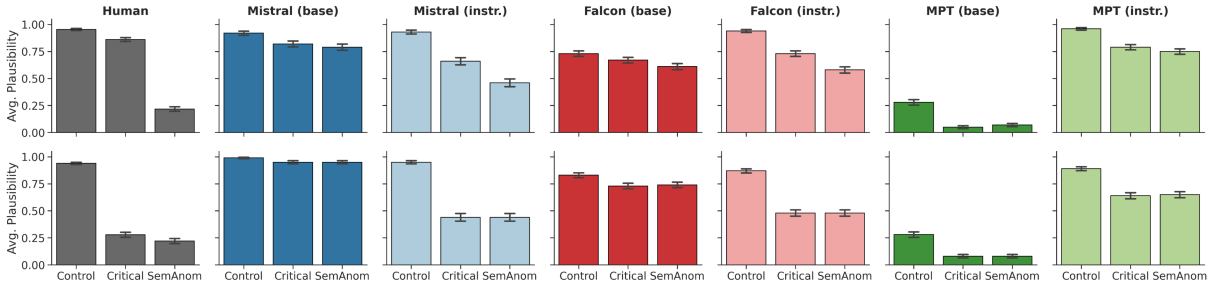


Figure 4: Replicating the sensibility-judgment task in LMs using prompting via the adjusted *Sentence Judgment* prompt in §F. Human data from Jouravlev et al. (2019). We use a barplot to visually set apart this prompt-based comparison vs. LogProbs-based ones in Figures 3, 9.

4.4 Results

Result 1: Across models, context successfully modulates the LogProbs *of (im)plausible target words, but not (im)plausible target sentences.*

When comparing target word vs. target sentence LogProbs, a clear trend emerges: all models demonstrate consistently high performance (around 95%) across all metrics when comparing the probabilities of target words (Table 6, *Word* columns); at the same time, when using the likelihoods they assign to sentences as an indicator of event plausibility knowledge, LogProbs plausibility judgments fail to reliably pass the sensitivity criterion.

	Gen. Plaus.		Context. Plaus.		Context Sens.	
	Word	Sent.	Word	Sent.	Word	Sent.
Mistral (base)	0.90	**0.93**	0.93	**1.00**	**0.97**	0.79
Mistral (instr)	**0.97**	0.90	0.93	**1.00**	**0.90**	0.84
Falcon (base)	**0.96**	0.94	**0.93**	0.92	**0.98**	0.79
Falcon (instr)	**0.98**	0.91	**0.95**	**0.95**	**0.96**	0.77
MPT (base)	**0.96**	0.93	0.95	**1.00**	**0.99**	0.76
MPT (instr)	**0.94**	0.93	0.93	**1.00**	**0.95**	0.80

Table 6: LogProbs results for Expt 2. Gen.–General; Context.–Context-Dependent; Plaus.–Plausibility; Sens.–Sensitivity; Word/Sent.–scores for target word/sentence.

In particular, even though almost all LMs are able to distinguish plausible and implausible sentences (General Plausibility, similar to §3); and are able to modulate the probability they assign an unexpected sentence in the presence of licensing context, they fail to update the sentence probabilities *selectively* (this is evidenced by the substantial drop in performance for the Context Sensitivity metric across LMs). This pattern suggests that while a semantically licensing context assists the models in up-weighing the probability of an otherwise implausible target word/event description (see Context-Dependent Plausibility; in line with Michaelov et al., 2023), contextual *implausibility* is not reliably reflected in LMs' sentence likelihoods. In particular, once an unexpected target word has been encountered (which the LMs are able to discern, see Context Sensitivity, *Word* columns), the LMs appear to quickly adjust the predictions in the post-target region, in some cases assigning even higher probabilities to post-target words than in the `Critical` condition, with the consequence that the scores for anomalous sentences and contextually-licensed ones differ less significantly at the sentence level. This suggests that a semantically-licensing context helps a model in predicting an otherwise anomalous word, but the global proba-

bility of the target sentence is less affected by the specific context.

Result 2: Context-modulated LogProbs *align with human contextual judgment patterns.*

Finally, we investigate how contextual plausibility judgments correspond to human behavior for the same stimuli. We focus on the sensibility-judgment task, in which participants were asked to decide (i) if a target sentence made sense to them within the provided context, or (ii) if it made sense to another person who did not have access to the context sentence (Jouravlev et al., 2019). Here, we model this dataset in a 'single-participant setting', by exposing the LMs to the full items and comparing the log probabilities assigned to the target words in the three experimental conditions, with or without licensing context. Across models, we see a remarkable match between human- and model-derived plausibility scores, both in the isolated sentence and the contextualized setup (Figure 3; for supporting statistical analyses see SI §E, Tables 8/10).

LogProbs again provide a better fit to human data than Prompting (Figures 3, 4; SI §E, Tables 8/10 vs. Tables 9/11), although it is interesting to observe that the prompting results for Instruct models matched the human behavioral patterns qualitatively (see also SI §F, §G).

5 Conclusion

Overall, we show that, for both base and instruction-tuned models, LogProbs remain a more reliable measure of semantic plausibility than naive zero-shot Prompting. This is true in scenarios that evaluate both isolated and context-dependent sentence plausibility. Even though instruction-tuning has been claimed to align LMs and human brain representations (Aw et al., 2023), other studies show that it does not always help for the alignment at the behavioral level (Kuribayashi et al., 2024). Our results show that the base LogProbs estimates for simple world knowledge scenarios do not drastically change as a result of instruction tuning, showing approximately the same amount of implicitly encoded information as representation derived from next-word prediction. In some cases, however, instruction tuning can lead to *less* alignment of LogProbs to human plausibility judgments than those of base model versions.

Concerning LMs' sensitivity to sentence context, we observe that by using LogProbs at the level of the target word, all the models perform around 90% with respect to the ground truth and are well aligned to human judgement patterns. However, when using sentence-level LogProbs we notice that the models have the tendency to "re-balance" the log likelihoods after processing an unexpected word, with the consequence that semantically anomalous sentences and contextually-licensed ones become harder to distinguish.

Although it is possible that model- and task-specific prompts will outperform raw LogProbs as a way to estimate sentence plausibility, our work highlights that LogProbs are an easy, zero-shot way to assess LMs' implicit knowledge. Thus, getting a raw LogProbs estimate of model performance can provide an initial estimate of whether or not custom prompt-based solutions can be successful or—in some cases—obviate the need for prompt tuning altogether.

Limitations

A first, obvious limitation of this work is that it has been conducted on English datasets, so we cannot be sure that our findings on LMs and event knowledge would generalize to other languages.

Second, even though our prompting setup mimics that of humans, it differs in substantial ways. For example, whereas we ask LMs to evaluate sentences in isolation, participants assign scores within the context of the full experiment, having access to their answer history.

Lastly, we only focused on LMs up to 7 billion parameters, due to the limit of our computational resources, and we only used three representative models in their Base and in their Instruct version. It is possible that with larger and more powerful models the performance will improve and the existing gap with human performance on distinguishing plausible vs. implausible sentences will be closed (cf. Kauf et al., 2023).

Ethical Considerations

Our work aims to better understand and characterize the capacities of models, and contributes to work highlighting the importance of open access to model representations. Our work shows that LM pre-training distills a wealth of world knowledge into the models' weights, but cannot guarantee the consistency of these representations with human world knowledge. Consequently, LMs should not be expected to generate statements that are consistent with human world knowledge. General ethical

concerns about LMs and their impact on human life, especially as they become more and more integrated into people's everyday lives, also apply to our work.

Acknowledgements

CK and this work was partially supported by the MIT Quest for Intelligence. EC was supported by a GRF grant from the Research Grants Council of the Hong Kong Special Administrative Region, China (Project No. PolyU 15612222). We would like to thank the three anonymous reviewers for their constructive comments and suggestions.

References

Laura Aina and Tal Linzen. 2021. The Language Model Understood the Prompt was Ambiguous: Probing Syntactic Uncertainty through Generation. In *Proceedings of the EMNLP BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks*.

Ebtesam Almazrouei, Hamza Alobeidli, Abdulaziz Alshamsi, Alessandro Cappelli, Ruxandra Cojocaru, Mérouane Debbah, Étienne Goffinet, Daniel Hesslow, Julien Launay, Quentin Malartic, et al. 2023. The Falcon Series of Open Language Models. *arXiv preprint arXiv:2311.16867*.

Khai Loong Aw, Syrielle Montariol, Badr AlKhamissi, Martin Schrimpf, and Antoine Bosselut. 2023. Instruction-tuning Aligns LLMs to the Human Brain. *arXiv preprint arXiv:2312.00575*.

Dougla Bates. 2014. Fitting linear mixed-effects models using lme4. *arXiv preprint arXiv:1406.5823*.

Klinton Bicknell, Jeffrey L Elman, Mary Hare, Ken McRae, and Marta Kutas. 2010. Effects of Event Knowledge in Processing Verbal Arguments. *Journal of Memory and Language*, 63(4):489–505.

Terra Blevins, Hila Gonen, and Luke Zettlemoyer. 2023. Prompting Language Models for Linguistic Structure. In *Proceedings of ACL*.

Yongchao Chen, Jacob Arkin, Yilun Hao, Yang Zhang, Nicholas Roy, and Chuchu Fan. 2024. Prompt Optimization in Multi-step Tasks (PROMST): Integrating Human Feedback and Preference Alignment. *arXiv preprint arXiv:2402.08702*.

Emmanuele Chersoni, Philippe Blache, and Alessandro Lenci. 2016. Towards a Distributional Model of Semantic Complexity. In *Proceedings of the COLING Workshop on Computational Linguistics for Linguistic Complexity*.

Emmanuele Chersoni, Enrico Santus, Alessandro Lenci, Philippe Blache, and Chu-Ren Huang. 2021. Not All Arguments Are Processed Equally: A Distributional Model of Argument Complexity. *Language Resources and Evaluation*, pages 1–28.

Emmanuele Chersoni, Enrico Santus, Ludovica Pannitto, Alessandro Lenci, Philippe Blache, and Chu-Ren Huang. 2019. A Structured Distributional Model of Sentence Meaning and Processing. *Natural Language Engineering*, 25(4):483–502.

Yew Ken Chia, Pengfei Hong, Lidong Bing, and Soujanya Poria. 2023. INSTRUCTEVAL: Towards Holistic Evaluation of Instruction-Tuned Large Language Models. *arXiv preprint arXiv:2306.04757*.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2022. Scaling Instruction-finetuned Language Models. *arXiv preprint arXiv:2210.11416*.

Yanai Elazar, Nora Kassner, Shauli Ravfogel, Amir Feder, Abhilasha Ravichander, Marius Mosbach, Yonatan Belinkov, Hinrich Schütze, and Yoav Goldberg. 2022. Measuring Causal Effects of Data Statistics on Language Model's Factual Predictions. *arXiv preprint arXiv:2207.14251*.

Evelina Fedorenko, Idan Asher Blank, Matthew Siegelman, and Zachary Mineroff. 2020. Lack of Selectivity for Syntax Relative to Word Meanings Throughout the Language Network. *Cognition*, 203:104348.

Richard Futrell, Ethan Wilcox, Takashi Morita, Peng Qian, Miguel Ballesteros, and Roger Levy. 2019. Neural Language Models as Psycholinguistic Subjects: Representations of Syntactic State. In *Proceedings of NAACL*.

Jonathan Gordon and Benjamin Van Durme. 2013. Reporting Bias and Knowledge Acquisition. In *Proceedings of the Workshop on Automated Knowledge Base Construction*.

Michael Hanna, Yonatan Belinkov, and Sandro Pezzelle. 2023. When Language Models Fall in Love: Animacy Processing in Transformer Language Models. In *Proceedings of EMNLP*.

Ari Holtzman, Peter West, Vered Shwartz, Yejin Choi, and Luke Zettlemoyer. 2021. Surface Form Competition: Why the Highest Probability Answer Isn't Always Right. In *Proceedings of EMNLP*.

Jennifer Hu, Jon Gauthier, Peng Qian, Ethan Wilcox, and Roger P Levy. 2020. A Systematic Assessment of Syntactic Generalization in Neural Language Models. In *Proceedings of ACL*.

Jennifer Hu and Roger Levy. 2023. Prompting Is Not a Substitute for Probability Measurements in Large Language Models. In *Proceedings of EMNLP*.

Jennifer Hu, Kyle Mahowald, Gary Lupyan, Anna Ivanova, and Roger Levy. 2024. Language Models Align with Human Judgments on key Grammatical Constructions. *arXiv preprint arXiv:2402.01676*.

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 preprint arXiv:2310.06825*.

Olessia Jouravlev, Rachael Schwartz, Dima Ayyash, Zachary Mineroff, Edward Gibson, and Evelina Fedorenko. 2019. Tracking Colisteners' Knowledge States during Language Comprehension. *Psychological Science*, 30(1):3–19.

Cheongwoong Kang and Jaesik Choi. 2023. Impact of Co-occurrence on Factual Knowledge of Large Language Models. In *Findings of EMNLP*.

Nora Kassner and Hinrich Schütze. 2020. Negated and Misprimed Probes for Pretrained Language Models: Birds Can Talk, but Cannot Fly. In *Proceedings of ACL*.

Carina Kauf and Anna Ivanova. 2023. A Better Way to Do Masked Language Model Scoring. In *Proceedings of ACL*.

Carina Kauf, Anna A Ivanova, Giulia Rambelli, Emmanuele Chersoni, Jingyuan Selena She, Zawad Chowdhury, Evelina Fedorenko, and Alessandro Lenci. 2023. Event Knowledge in Large Language Models: The Gap Between the Impossible and the Unlikely. *Cognitive Science*, 47(11):e13386.

Tatsuki Kuribayashi, Yohei Oseki, and Timothy Baldwin. 2024. Psychometric Predictive Power of Large Language Models. In *Findings of NAACL*.

Andrew Kyle Lampinen. 2022. Can Language Models Handle Recursively Nested Grammatical Structures? A Case Study on Comparing Models and Humans. *arXiv preprint arXiv:2210.15303*.

Alessandro Lenci. 2011. Composing and Updating Verb Argument Expectations: A Distributional Semantic Model. In *Proceedings of the ACL Workshop on Cognitive Modeling and Computational Linguistics*.

Belinda Z Li, Maxwell Nye, and Jacob Andreas. 2021. Implicit Representations of Meaning in Neural Language Models. In *Proceedings of ACL*.

Jiaoda Li, Ryan Cotterell, and Mrinmaya Sachan. 2022. Probing via Prompting. In *Proceedings of NAACL*.

James A Michaelov and Benjamin K Bergen. 2020. How Well Does Surprisal Explain N400 Amplitude Under Different Experimental Conditions? In *Proceedings of CONLL*.

James A Michaelov, Seana Coulson, and Benjamin K Bergen. 2023. Can Peanuts Fall in Love with Distributional Semantics? In *Proceedings of CogSci*.

Kanishka Misra, Allyson Ettinger, and Kyle Mahowald. 2024. Experimental Contexts Can Facilitate Robust Semantic Property Inference in Language Models, but Inconsistently. *arXiv preprint arXiv:2401.06640*.

MosaicML NLP Team. 2023. Introducing MPT-7B: A New Standard for Open-Source, Commercially Usable LLMs. *www.mosaicml.com/blog/mpt-7b*.

Mante S Nieuwland and Jos JA Van Berkum. 2006. When Peanuts Fall in Love: N400 Evidence for the Power of Discourse. *Journal of Cognitive Neuroscience*, 18(7):1098–1111.

Isabel Papadimitriou, Richard Futrell, and Kyle Mahowald. 2022. When Classifying Grammatical Role, BERT Doesn't Care about Word Order... Except When It Matters. In *Proceedings of ACL*.

Paolo Pedinotti, Giulia Rambelli, Emmanuele Chersoni, Enrico Santus, Alessandro Lenci, and Philippe Blache. 2021. Did the Cat Drink the Coffee? Challenging Transformers with Generalized Event Knowledge. In *Proceedings of *SEM*.

Reid Pryzant, Dan Iter, Jerry Li, Yin Tat Lee, Chenguang Zhu, and Michael Zeng. 2023. Automatic Prompt Optimization with "Gradient Descent" and Beam Search. In *Proceedings of EMNLP*.

Guanghui Qin and Jason Eisner. 2021. Learning how to Ask: Querying LMs with Mixtures of Soft Prompts. In *Proceedings of NAACL*.

Shirley-Ann Rueschemeyer, Tom Gardner, and Cat Stoner. 2015. The Social N400 Effect: How the Presence of Other Listeners Affects Language Comprehension. *Psychonomic Bulletin & Review*, 22:128–134.

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

Melanie Sclar, Yejin Choi, Yulia Tsvetkov, and Alane Suhr. 2023. Quantifying Language Models' Sensitivity to Spurious Features in Prompt Design or: How I Learned to Start Worrying about Prompt Formatting. *arXiv preprint arXiv:2310.11324*.

Koustuv Sinha, Jon Gauthier, Aaron Mueller, Kanishka Misra, Keren Fuentes, Roger Levy, and Adina Williams. 2022. Language Model Acceptability Judgements Are Not Always Robust to Context. *arXiv preprint arXiv:2212.08979*.

Katherine Tian, Eric Mitchell, Allan Zhou, Archit Sharma, Rafael Rafailov, Huaxiu Yao, Chelsea Finn, and Christopher Manning. 2023. Just Ask for Calibration: Strategies for Eliciting Calibrated Confidence Scores from Language Models Fine-Tuned with Human Feedback. In *Proceedings of EMNLP*.

Ottokar Tilk, Vera Demberg, Asad Sayeed, Dietrich Klakow, and Stefan Thater. 2016. Event Participant Modelling with Neural Networks. In *Proceedings of EMNLP*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Grave Edouard, and Guillaume Lample. 2023. Llama: Open and Efficient Foundation Language Models. *arXiv preprint arXiv:2302.13971*.

Paolo Vassallo, Emmanuele Chersoni, Enrico Santus, Alessandro Lenci, and Philippe Blache. 2018. Event Knowledge in Sentence Processing: A New Dataset for the Evaluation of Argument Typicality. In *Proceedings of the LREC Workshop on Linguistic and Neuro-Cognitive Resources (LiNCR)*.

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.

Shengyu Zhang, Linfeng Dong, Xiaoya Li, Sen Zhang, Xiaofei Sun, Shuhe Wang, Jiwei Li, Runyi Hu, Tianwei Zhang, Fei Wu, et al. 2023. Instruction Tuning for Large Language Models: A Survey. *arXiv preprint arXiv:2308.10792*.

Supplementary Information

A Complete prompting results

Figure 5 shows the complete prompting results across datasets, models and prompts.

B Additional prompting results for DTFit

Prompt	Example
Word Comparison	What word is most likely to come next in the following sentence (award, or battle)? The actor won the {**award**, **battle**}

Table 7: Additional prompt used for Vassallo et al. (2018) evaluation in Figure 6. This prompt is the best-performing prompt for this dataset in Hu and Levy (2023).

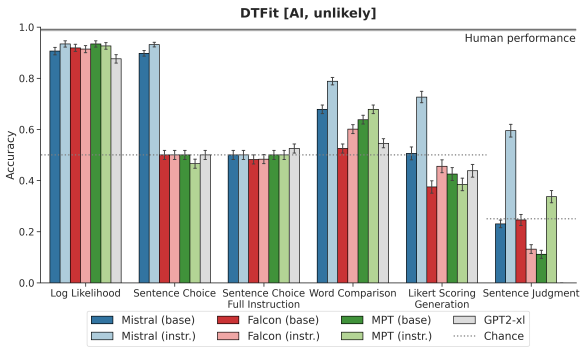


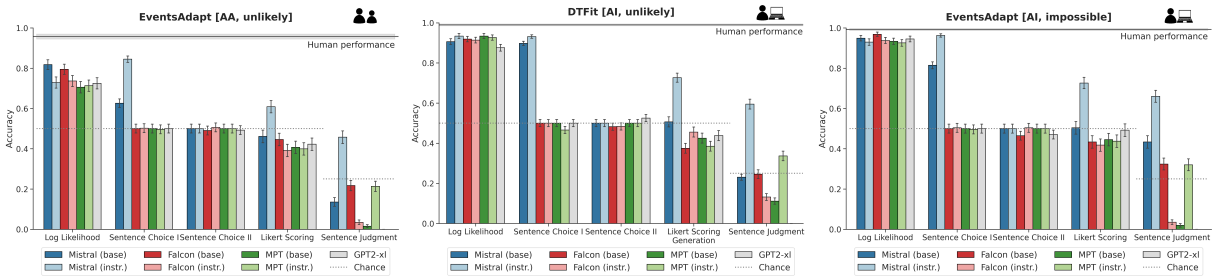
Figure 6: Prompting results for DTFit, including best prompt from Hu and Levy (2023).

C Evidence for invariance to prompting variations for DTFit

C.1 Free vs. constrained generation

Here, we evaluate prompt-based generation in two ways: using a free vs. constrained generation paradigm. In the free paradigm, we ask the model to generate up to 20 tokens in the completion and find responses that include a valid response (exactly one numeral between 1-2 or 1-7). In the constrained paradigm, we only allow completions from a predefined set of tokens, i.e., either the set {1,2} or the set {1,2,3,4,5,6,7}, using a regex-matching generation procedure from `outlines`[4]. Results are roughly consistent across metrics, yielding no advantage of one over the other prompting paradigm in both *Sentence Choice* and *Likert Scor-*

[4] https://github.com/outlines-dev/outlines



(a) EventsAdapt (AA, unlikely) (b) DTFit (AI, unlikely) (c) EventsAdapt (AI, impossible)

Figure 5: Results of implicit vs. explicit plausibility judgment performance experiments

ing paradigms.

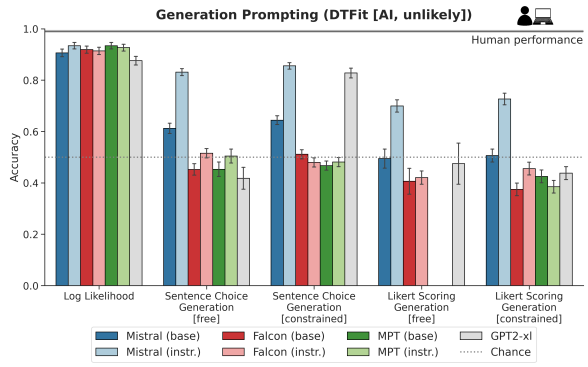


Figure 7: Comparison of free vs. constrained generation prompting. Note that `MPT` results are missing for the free Likert Scoring method.

C.2 Query types

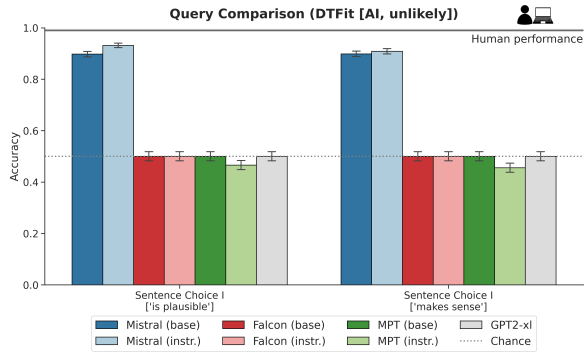


Figure 8: Comparison of different query types for prompts of type *Sentence Choice I*. In all supplementary figures for Experiment 1, we also include `GPT2-xl` as a baseline model.

D GLMM analysis

We fit a binomial generalized linear mixed-effects model (GLMM) with a logit link function to predict the binary variable accuracy, using LLM model class (`Mistral`, `Falcon`, `MPT`), model version (`base`, `instruct`), and voice (active, passive) as fixed effects, and items as random effects. The model further included all interactions between the fixed effects. We used dummy coding for voice, with "active" as the reference level, and sum-coding for model class and model version. The analysis was conducted using the `lme4` R package (Bates, 2014).

E Quantifying the fit to human result patterns for Experiment 2: Context-Dependent Plausibility Judgments

To compare the result patterns of humans vs. models for the sentence sensibility judgment task across conditions and across both continuous (LogProbs) vs. discrete (Prompting) outputs (which for some items led to zero-variance response vectors across experimental conditions), we measured the similarity between human and model responses across different experimental conditions using Euclidean distance with the following approach. First, we scaled the response data for each model using min-max scaling to prevent distance calculations to be biased by differences in response magnitude. For each pair of human and model responses, we then calculated the Euclidean distance between the three-point response vectors across conditions (Control, Critical, SemAnom) for each item. To convert this distance into a similarity value, we used a normalized metric where similarity is defined as $1 - \frac{\text{distance}}{\text{max distance}}$ where the maximum possible Euclidean distance between two vectors corresponds to the vector's dimensionality, yielding a similarity score in the range from 0 (maximally dissimilar) to 1 (identical). Similarity scores were calculated for all combinations of context (human context vs. model context, human context vs. model no context, human no context vs. model context, human no context vs. model no context). The

Model	matched	unmatched
Mistral (base)	0.41	0.31
Mistral (instruct)	0.40	0.30
Falcon (base)	0.51	0.39
Falcon (instruct)	0.51	0.40
MPT (base)	0.50	0.38
MPT (instruct)	0.48	0.37

Table 8: Similarity results of human to model response pattern analysis for Figure 3.

Model	matched	unmatched
Mistral (base)	0.06	0.08
Mistral (instruct)	0.30	0.14
Falcon (base)	0.04	0.04
Falcon (instruct)	0.22	0.08
MPT (base)	-0.05	-0.05
MPT (instruct)	0.13	0.07

Table 9: Similarity results of human to model response pattern analysis for Figure 4.

similarity scores were then averaged across items to obtain a final similarity value for each of the four conditions. We report the average similarity scores per model across the matched (human and model both in "Context" or both in "No Context") and mismatched (one in "Context" and the other in "No Context") conditions in Tables 8, 9.

We further conducted paired t-tests to compare similarity scores in matched context conditions with mismatched conditions in order to determine whether the models captured the human responses significantly better when the context matched. T-test results are reported in Tables 10, 11.

Model	t-statistic	p-value
Mistral (base)	8.49	0.00
Mistral (instruct)	8.52	0.00
Falcon (base)	10.83	0.00
Falcon (instruct)	9.69	0.00
MPT (base)	11.80	0.00
MPT (instruct)	10.70	0.00

Table 10: T-test results to compare similarity scores in matched context conditions with mismatched conditions in Figure 3.

Model	t-statistic	p-value
Mistral (base)	-1.43	0.00
Mistral (instruct)	4.81	0.15
Falcon (base)	0.04	0.97
Falcon (instruct)	4.69	0.00
MPT (base)	0.01	0.99
MPT (instruct)	2.84	0.01

Table 11: T-test results to compare similarity scores in matched context conditions with mismatched conditions in Figure 4.

F Replicating the sensibility-judgment task by Jouravlev et al. (2019) using prompting

To replicate the human experiment by Jouravlev et al. (2019) in LMs using prompting, we queried the models using an adjusted *Sentence Judgment* prompt (see Table 2): [No context:] *Here is a sentence: "sentence". Does this sentence make sense? Respond with either Yes or No as your answer.* [With context:] *Here is a context: "context", and here is a sentence: "sentence". Does this sentence make sense considering the context? Respond with either Yes or No as your answer.* We report our results in Figure 4.

We observe that while most base models often favor one answer option, the instruction-tuned models exhibit more a nuanced behavior: These models are more consistent with human responses in this binary sensitivity judgment task, matching them qualitatively. Nevertheless, instruction-tuned models tend to (i) systematically underestimate the contextual plausibility of the `Critical` sentences (Figure 4, upper panel), and (ii) systematically overestimate the plausibility of implausible sentences relative to humans (`SemAnom` conditions and `Critical` condition, Figure 4, lower panel) in the binary sensibility-judgment task setup.

G Replicating the sensibility-judgment task by Jouravlev et al. (2019) using sentence log likelihoods

In Figure 9, we replicate the human experiment by Jouravlev et al. (2019) in LMs using sentence log likelihood measurements. We generally observe similar trends than the comparison with the target word measurement.

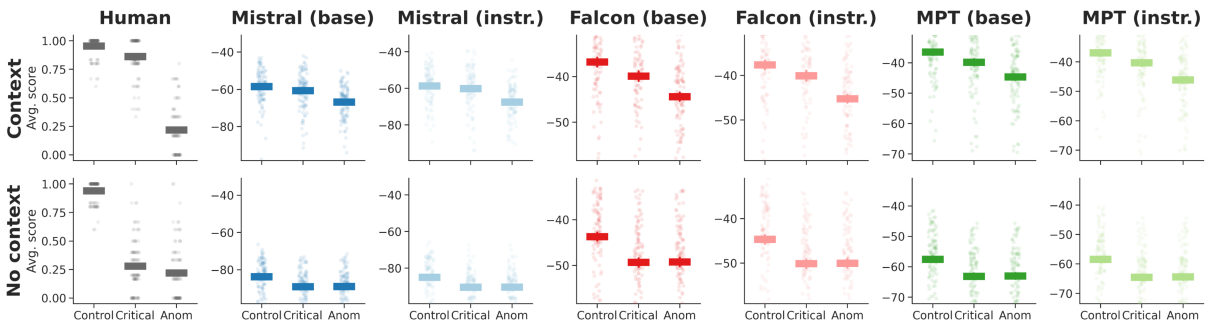


Figure 9: Replicating the sensibility-judgment task in LMs using sentence LOGPROBS measures. Human data from Jouravlev et al. (2019).