Comparing Plausibility Estimates in Base and Instruction-Tuned Large Language Models

Anonymous ACL submission

Abstract

Instruction-tuned LLMs can respond to explicit queries formulated as prompts, which greatly facilitates interaction with human users. However, prompt-based approaches might not always be able to tap into the wealth of implicit knowledge acquired by LLMs during pretraining. This paper presents a comprehensive study of ways to evaluate semantic plausibility in LLMs. We compare base and instruction-tuned LLM performance on an English sentence plausibility task via (a) explicit prompting and (b) implicit estimation via direct readout of the probabilities models assign to strings. Experiment 1 shows that, across model architectures and plausibility datasets, (i) log likelihood (LL) scores are the most reliable indicator of sentence plausibility, with zero-shot prompting yielding inconsistent and typically poor results; (ii) LL-based performance is still inferior to human performance; (iii) instruction-tuned models have worse LL-based performance than base models. In Experiment 2, we show that *LL* scores across models are modulated by context in the expected way, showing high performance on three metrics of context-sensitive plausibility and providing a direct match to explicit human plausibility judgments. Overall, *LL* estimates remain a more reliable measure of plausibility in LLMs than direct prompting.[1]

1 Introduction

The impressive empirical successes of large language models (LLMs) on many diverse (language) tasks (e.g., Devlin et al., 2019; Liu et al., 2019; Brown et al., 2020; Achiam et al., 2023; Bubeck et al., 2023; Guo et al., 2023) has fueled an explosive increase in their popularity. As LLMs are becoming more and more integrated in people's everyday lives, it is critical to provide reliable assessments of their capabilities.

An important domain to test is LLMs' general world knowledge. Language training data contains vast amounts of information about the world, including both factual knowledge explicitly stated in the input and distributional knowledge, inferrable via text co-occurrence patterns (Elazar et al., 2022; Kang and Choi, 2023). Leveraging world knowledge is important both for specific NLP tasks (e.g., information retrieval) and for general success of a language model during interactions with a user (e.g., establishing common ground).

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.

Traditionally, NLP researchers evaluated the knowledge that LLMs distill into their weights through a combination of log likelihood comparisons on minimal sentence pairs (Futrell et al., 2019; Warstadt et al., 2020; Hu et al., 2020; Aina and Linzen, 2021; Pedinotti et al., 2021; Sinha et al., 2022; Hu et al., 2024; Michaelov et al., 2023; Misra et al., 2024), probing the model's representations of a stimulus (Hewitt and Manning, 2019; Kim et al., 2019; Eisape et al., 2022; Müller-Eberstein et al., 2022; Kauf et al., 2023), adversarial datasets (McCoy et al., 2019; Kassner and Schütze, 2020), or causal interventions (Geiger et al., 2020), among others. Given the closeness to the unsupervised pretraining regime, minimal sentence pair comparisons of likelihood measures, in particular, have been widely adopted.

More recently, however, the focus of NLP researchers has shifted towards LLMs that have been fine-tuned to follow instructions (Chung et al., 2022; Touvron et al., 2023; Almazrouei et al., 2023; Jiang et al., 2023), as instruction tuning improves the alignment of the models with user in-

[1] Code will be released upon decision.

tent and leads to better generalization to unseen tasks (Ouyang et al., 2022). Because instruction-tuned models are designed to interact directly with a user through LLM-directed queries/prompts, natural language *prompting* has emerged as a way to directly query LLMs for the knowledge they encode (e.g., Li et al., 2022; Blevins et al., 2023). Critically, as access to log probabilities 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 LLMs.

The link between sentence plausibility and sentence probability is indirect: raw log probabilities have been shown to reflect a number of factors that might not be relevant for a given task, including low-level properties of the stimulus such as sentence length, word frequency (Kauf et al., 2023), and the number of surface forms that refer to the same concept (Holtzman et al., 2021). Thus, direct prompting approaches might provide a more direct estimate of plausibility by filtering out influences of those additional factors. However, initial direct comparisons of log likelihood and prompting measures on different linguistic/semantic knowledge datasets has revealed that prompting may 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).

In this paper, we test LLMs' knowledge of plausibility in single-sentence (Experiment 1) and contextualized scenarios (Experiment 2). Our findings include:

1. Log likelihood (LL) scores, while imperfect, are a more dependable measure of plausibility than natural language prompting evaluations.

2. Instruction-tuning often alters an LLM's log-likelihood scores in such a way that they become *less consistent* with human plausibility judgments relative to base model versions.

3. *LL* scores can effectively model the contextual plausibility of events and replicate key patterns of human plausibility-judgment behaviors. Nevertheless, the LLMs' ability to detect an implausibility within a target sentence locally does not reliably affect their evaluation of the full sentence.

2 Related Work

Evaluating single-sentence plausibility in LLMs. In Experiment 1, we evaluate plausibility estimates for single sentences describing common events (Table 1). Earlier work in NLP aimed at modeling event-based semantic plausibility via *distributional models of thematic fit*: verbs and arguments were often considered in isolation, and the goal for the models was to estimate a continuous score expressing to what extent an argument noun (e.g., *ball*) was fitting a given semantic role of a verb (e.g., the patient role of *to throw*) (Baroni and Lenci, 2010; Sayeed et al., 2016; Santus et al., 2017). In a more natural evaluation setting, researchers used sentence pairs derived from psycholinguistic experiments that differed only for one argument and displayed different degrees of plausibility (e.g., *The mechanic was checking the brakes* vs. *The journalist was checking the brakes*, from Bicknell et al., 2010): in this case, a distributional model had to dynamically "compose" the plausibility of the two argument roles and guess which of the two sentences was the most plausible one (*binary judgement task*) (Lenci, 2011; Chersoni et al., 2019).

With the advent of Transformer-based language models, the analysis of their semantic knowledge has often been framed as a probability comparison between sound and anomalous, or atypical sentences (Michaelov and Bergen, 2020; Beyer et al., 2021; Pedinotti et al., 2021; Kauf et al., 2023; Misra et al., 2023). Similarly to the binary judgement setting, a model has to score a sentence pair where two sentences differ only for the presence of a semantic violation, and assign a higher score to the plausible one.

Pedinotti et al. (2021) and Kauf et al. (2023) specifically tested event plausibility knowledge in LLMs. Pedinotti et al. (2021) showed that LLMs achieve correlation with human judgements on par or better than traditional distributional models. Kauf et al. (2023) investigated event plausibility using minimal sentence pairs, in the task of binary judgements. They showed that Transformer-based models retain a considerable 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 LLMs' generalization capabilities to novel experimental manipulations of the target sentences are limited and that log probabilities are affected by task-irrelevant information, such as the frequency of words within a target sentence.

Evaluating context-dependent linguistic judgments in LLMs. In Experiment 2, we evaluate context sensitivity of LLM plausibility estimates (Table 4). Initial work in this domain shows that (Dutch) LLMs 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) - such scenarios, similarly, were shown to elicit a reduced N400 amplitude in humans, as a neural signature of a decrease of processing complexity of the event (Nieuwland and Van Berkum, 2006; Rueschemeyer et al., 2015). Nevertheless, probability-based judgements of LLMs 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 log likelihood measurements and prompt-based methods. The direct interaction with models through natural language prompts is exciting for many reasons, including that it facilitates knowledge exploration in a way that begins to mimic the experimental procedure used for 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 LLMs' 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 (they also used the *DTFit* dataset, but their prompts did not explicitly probe the notion of plausibility).

Evaluating the alignment of instruction-tuned models with humans. Even though instruction-tuning has been claimed to better align the representations of LLMs 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., 2023). However, the work in this domain is still sparse.

3 Experiment 1: Explicit vs. Implicit Event Plausibility Judgments

In this section, we compare explicit (prompt-based) and implicit (LL-based) plausibility judgments in base- and instruction-tuned LLMs across base and instruction-tuned models from 3 families.

3.1 Datasets

We use two sentence sets adapted from previous studies and compare model scores with human plausibility judgements. A schematic illustration of the items in each of the datasets can be seen in 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 in the past tense (*The teacher bought the laptop*), (ii) the implausible version of the same sentence, constructed by swapping the noun phrases (*The laptop bought the teacher*), as well as 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 in the past tense, where an animate agent is interacting with an inanimate patient that is either typical for the agent (*The actor won the award*); (ii) or less plausible version of the same sentence, constructed by varying the inanimate patient (*The actor won the battle*). The different degrees of typicality depend on the interaction of the patient with both the agent and the verb (e.g. a *battle* may be a typical patient for a *winning*-event, it is just not typical given that the agent is an *actor*). Thus, word content and not word order is used to distinguish between plausible and implausible sentences.

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

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

from 1 (completely implausible) to 7 (completely plausible) (Kauf et al., 2023). We averaged human judgments to obtain a single score for each sentence, and assigned a hit every time that the plausible version of the sentence was scored higher than the corresponding implausible one by the human participants pool.

3.3 Model Plausibility Judgments

Models. For our experiments, we used the Base and the Instruct version of three popular autoregressive LLMs: `Mistral` (Jiang et al., 2023), `Falcon` (Almazrouei et al., 2023), and `MPT` (MosaicML NLP Team, 2023), all of them with 7B parameters. We include `GPT2-XL` (Radford et al., 2019) (1.5B parameters) as a baseline model.

Metrics. We adopt the evaluation paradigm by Hu and Levy (2023) and evaluate models using (i) *LL* scores, and (ii) several zero-shot prompting methods (Table 2). The *LL score* is the sum of the log-probabilities of each token w_i in a sentence, conditioned on the preceding sentence tokens $w_{<i}$. Our prompts, *Sentence Choice I/II, Likert Scoring* and *Sentence Judgment* are designed to explicitly query the LLMs' knowledge of plausibility, using either the same or similar instructions to the task that humans solved (see §3.2). 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[2] and compare the generated scores for plausible vs. implausible sentences (the results remain consistent across free vs. constrained generation prompting, see SI §B, Figure 5).[3] 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, such as "makes sense" (SI §B, Figure 6) and "is likely" (SI §A, Figure 4).

Binary accuracy. For each dataset item, we compare the scores/generations of the minimally different plausible and the implausible sentence conditions, and assign a hit for every time a higher score was assigned to the plausible version, the same as for the human scores. The binary *accuracy* for all models is the ratio of dataset items in which plausible sentences received a higher probability score. 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

<u>*Result 1:*</u> *LL scores are a more reliable plausibility measure in LLMs than prompting.*

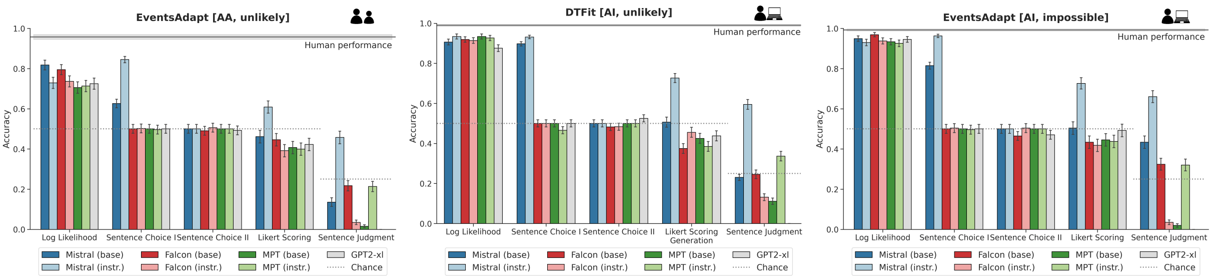
Our analysis reveals that across model architectures and plausibility datasets, *LL scores* are a more reliable indicator of plausibility knowledge than prompt-based approaches. This shows that there is a direct connection between plausibility and probability measures derived from the context prediction pretraining objective of LLMs (see also

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

[3] Note that the *DTFit* dataset was included in the study by Hu and Levy (2023) and was evaluated using different models and different prompts. Their prompts are not applicable to the *EventsAdapt* dataset and do not prompt models explicitly for *plausibility* judgments. We include an evaluation of our models on their best-performing prompt for *DTFit* as a supplementary analysis (SI §A, Figure 4).

Evaluation type	Example
Log Likelihood 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).



(a) EventsAdapt (AA, unlikely)

(b) DTFit (AI, unlikely)

(c) EventsAdapt (AI, impossible)

Figure 1: Results of implicit vs. explicit plausibility judgment performance experiments

Hu and Levy, 2023). Nevertheless, for the best-performing model (across all prompting metrics, except *Sentence Choice II*), `Mistral Instruct`, *Sentence Choice I* certain prompting setups consistently match (Figure 1, panels (b), (c)) or even outperform (Figure 1, panel (a)) log likelihood task performance. Despite this success, our comparison critically shows that there is not a single prompt that reliably taps into plausibility knowledge across model architectures, and none of our tested models are robust against slight variations in the way in which prompting is set up (see also Sclar et al., 2023). In fact, many of the prompting methods lead to chance-level performance or below-chance performance for most models, 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.

<u>*Result 2:*</u> *LL scores encode substantial plausibility knowledge but fall short of human performance.*

The *LL* results in Figure 1 show that LLMs acquire substantial event knowledge from distributional linguistic patterns; all of them performing well above chance on the task. Nevertheless, they consistently fall short of human performance and do not improve reliably over older LLMs (especially in cases where an event is comprised of two animate event participants) (Figure 1, left panel): 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 performance (all Bonferroni-corrected $ps > .05$ except for Falcon base: $t = -2.14, p = .02$). At the same time, none of the models significantly outperformed the `GPT2-XL` baseline model. On the more challenging *EventsAdapt (AA, unlikely)* subset, all models performed significantly worse than humans in distinguishing AA plausible from implausible events (all $ps < .001$), and only one model, Mistral Base, significantly improved over the smaller baseline model ($t = 2.96, p < .05$; all other $ps > .05$). Lastly, the high task performance on *DTFit (AI, unlikely)*, we observe that LLMs 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. Although all models still

	Mistral		Falcon		MPT	
	Base	Instruct	Base	Instruct	Base	Instruct
EventsAdapt	**0.82****	0.73	**0.79**	0.74	**0.71**	**0.71**
EventsAdapt	**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 3: Log Likelihood results across metrics and target regions. Significant differences from dependent t-tests between Base and Instruct models are marked with asterisks ($p < .05$: *; $p < .01$: **).

fall short of human performance for this dataset at $ps < .001$, all but two of the tested LLMs significantly improved over the GPT2-XL baseline on this dataset (only Mistral Base and Falcon Instruct are not better, $ps > .05$).

<u>Result 3:</u> *Instruction-tuning worsens LL score alignment with human plausibility judgments.*

Next, we zoom in on the comparison of *LL* scores derived from Base vs. Instruct 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).

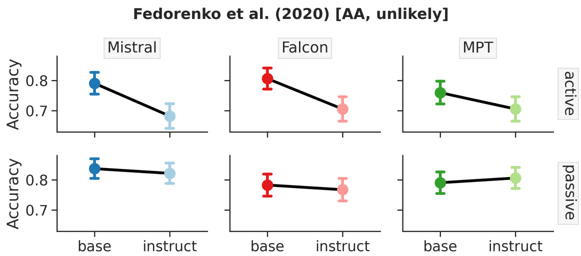


Figure 2: Base vs. instruct model performance in active and passive sentence pairs

A comparative analysis of the Base and Instruct results across different model architectures reveals no beneficial effect of instruction-tuning for gauging event plausibility through *LL* measurements: In all but one instance do instruction-tuned models performed similar or even slightly worse than their corresponding base model (Table 3). Interestingly, the gap is most noticable 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 *LLs* 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). This variance highlights the fact that even though direct measurements of model-derived string probabilities 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).

4 Experiment 2: Context-Dependent Plausibility Judgments

Experiment 1 has shown that *LLs* are the most reliable, albeit imperfect, metric for probing the plausibility of *isolated* sentences in LLMs. However, most of the time, humans and LLMs to not process sentences in isolation, but rather as part of a larger context. On the other hand, language models have also been shown to be sensitive to priming effects from inter-sentential context (Misra et al., 2020; Kassner and Schütze, 2020). In Experiment 2, we investigate whether LLMs appropriately modulate their judgments of event plausibility when provided with different discourse contexts. Specifically, we test how and to what extent judgments of event plausibility from minimal pair accuracies change in English LLMs in the presence of (i) supporting or (ii) non-supporting but related single-sentence contexts.

4.1 Dataset

To test the sensitivity of the LLM 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 4).

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

4.2 Metrics

We evaluate the models' context-aware plausibility judgements on three critical metrics:

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

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

For each metric, we evaluate model performance the likelihood they assign (i) a critical word within the target sentence and (ii) the target sentence itself. 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 measures 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. This is because the token likelihood of plausible and implausible sentences are shared up until the first occurrence of a contextually unlicensed word.

4.3 Results

Result 1: Target word LLs are better modulated by context than target sentence LLs.

When comparing target word vs. target sentence *LLs*, a clear trend emerges: all models perform extremely well (around 95%) across all metrics when comparing the probabilities of target words (Table 5, *Word* columns); at the same time, when using the likelihoods they assign to sentences as an indicator of event plausibility knowledge, performance degrades for two of the three metrics (Table 5, *Sent.* columns). In particular, even though almost all LLMs 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 LLMs (although they perform significantly better than the baseline GPT2-XL model). 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 LLMs' sentence likelihoods. In particular, once an unexpected target word has been encountered (which the LLMs are able to discern, see Context Sensitivity, *Word* columns), the LLMs appear to quickly adjust the predictions in the post-target region, in some cases

	Gen. Plaus.		Context-Dep. 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
GPT2-xl	**0.91**	0.85	0.88	**1.00**	**0.91**	0.64

Table 5: *LL* results for Expt 2. Gen.–General; Context-Dep.–Context-Dependent; Plaus.–Plausibility; Sens.–Sensitivity; Word/Sent.–scores for target word/sentence.

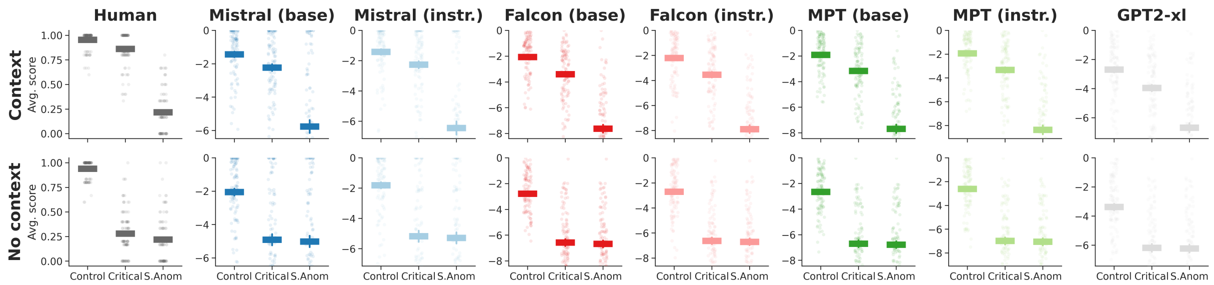


Figure 3: Target word *LLs* replicate patterns of human sentence sensibility judgments. Human data from Jouravlev et al. (2019). Bars indicate average plausibility of sentences (Human) and average target word log likelihoods (LLMs). Dots represent individual sentence scores (averaged across the participant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

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 probability of the target sentence benefits less from a the specific context. After meeting an unexpected target word, LLMs seem to be quickly able to adjust the predictions for the following ones, with the consequence that the scores for anomalous sentences and contextually-licensed ones differ less significantly at the sentence level.

<u>Result 2:</u> Context-modulated LLs 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 if a target sentence made sense (i) to them within the provided context, or (ii) 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 LLMs 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. For completeness, we report results for the exact replication of the human study in LLMs, using *Sentence Judgment* prompts in SI §D. We note that, again, *LLs* provide a better fit to human data, even though the prompting results for Instruct models matched the human behavioral patterns qualitatively (see also SI §C).

5 Conclusion

Overall, we show through careful investigation that *LL* scores, reflecting co-occurrence patterns distilled by LLMs from the task of next-word prediction at scale during pre-training, remain a more reliable measure of sentence plausibility than both (i) direct prompting and (ii) log likelihood scores from models finetuned to follow instructions. This is true in scenarios that encompass both isolated and context-dependent sentence plausibility estimates. Even though instruction-tuning has been claimed to align LLMs 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., 2023). The results presented in our work are consistent with the latter finding.

Concerning LLMs' sensitivity to sentence context, we observe that by using *LL* scores 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. On the other hand, when using sentence-level *LL* scores, 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 *LL* scores as a way to estimate sentence plausibility, our work highlights that *LL* scores are an easy, zero-shot way to assess LLMs' implicit knowledge. Thus, getting a raw *LL* 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 LLMs 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 LLMs 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 LLMs 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 LLM 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, LLMs should not be expected to generate statements that are consistent with human world knowledge. General ethical concerns about LLMs and their impact on human life, especially as they become more and more integrated into people's everyday lives, also apply to our work.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. GPT-4 Technical Teport. *arXiv preprint arXiv:2303.08774*.

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*.

Marco Baroni and Alessandro Lenci. 2010. Distributional Memory: A General Framework for Corpus-Based Semantics. *Computational Linguistics*, 36(4):673–721.

Anne Beyer, Sharid Loáiciga, and David Schlangen. 2021. Is Incoherence Surprising? Targeted Evaluation of Coherence Prediction from Language Models. In *Proceedings of NAACL*.

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*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language Models Are Few-shot Learners. *Advances in Neural Information Processing Systems*, 33:1877–1901.

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al. 2023. Sparks of Artificial General Intelligence: Early Experiments with GPT-4. *arXiv preprint arXiv:2303.12712*.

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*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of NAACL-HLT*.

Tiwalayo Eisape, Vineet Gangireddy, Roger Levy, and Yoon Kim. 2022. Probing for Incremental Parse States in Autoregressive Language Models. In *Findings of EMNLP*.

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*.

Atticus Geiger, Kyle Richardson, and Christopher Potts. 2020. Neural Natural Language Inference Models Partially Embed Theories of Lexical Entailment and Negation. In *Proceedings of the EMNLP BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*.

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

Biyang Guo, Xin Zhang, Ziyuan Wang, Minqi Jiang, Jinran Nie, Yuxuan Ding, Jianwei Yue, and Yupeng Wu. 2023. How Close Is ChatGPT to Human Experts? Comparison Corpus, Evaluation, and Detection. *arXiv preprint arXiv:2301.07597*.

John Hewitt and Christopher D Manning. 2019. A Structural Probe for Finding Syntax in Word Representations. In *Proceedings of NAACL*.

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, 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.

Najoung Kim, Roma Patel, Adam Poliak, Alex Wang, Patrick Xia, R Thomas McCoy, Ian Tenney, Alexis Ross, Tal Linzen, Benjamin Van Durme, et al. 2019. Probing What Different NLP Tasks Teach Machines about Function Word Comprehension. In *Proceedings of *SEM*.

Tatsuki Kuribayashi, Yohei Oseki, and Timothy Baldwin. 2023. Psychometric Predictive Power of Large Language Models. *arXiv preprint arXiv:2311.07484*.

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*.

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

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A Robustly Optimized BERT Pretraining Approach. *arXiv preprint arXiv:1907.11692*.

R Thomas McCoy, Ellie Pavlick, and Tal Linzen. 2019. Right for the Wrong Reasons: Diagnosing Syntactic Heuristics in Natural Language Inference. *arXiv preprint arXiv:1902.01007*.

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*.

Kanishka Misra, Allyson Ettinger, and Julia Taylor Rayz. 2020. Exploring BERT's Sensitivity to Lexical Cues using Tests from Semantic Priming. In *Findings of EMNLP*.

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 EACL*.

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

Max Müller-Eberstein, Rob Van Der Goot, and Barbara Plank. 2022. Probing for Labeled Dependency Trees. In *Proceedings of ACL*.

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.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training Language Models to Follow Instructions with Human Feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744.

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*.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language Models Are Unsupervised Multitask Learners. *OpenAI Blog*, 1(8):9.

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.

Enrico Santus, Emmanuele Chersoni, Alessandro Lenci, and Philippe Blache. 2017. Measuring Thematic Fit with Distributional Feature Overlap. In *Proceedings of EMNLP*.

Asad Sayeed, Clayton Greenberg, and Vera Demberg. 2016. Thematic Fit Evaluation: An Aspect of Selectional Preferences. In *Proceedings of the ACL Workshop on Evaluating Vector Space Representations for NLP*.

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*.

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 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 6: Additional prompt used for Vassallo et al. (2018) evaluation in Figure 4. This prompt is the best-performing prompt for this dataset in Hu and Levy (2023).

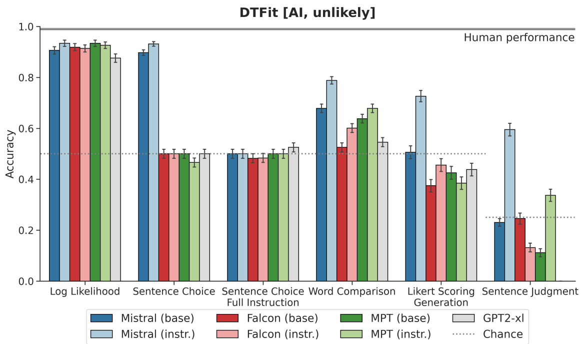


Figure 4: Prompting results for DTFit, including best prompt from Hu and Levy (2023).

B Evidence for invariance to prompting variations for DTFit

B.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 Scoring* paradigms.

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

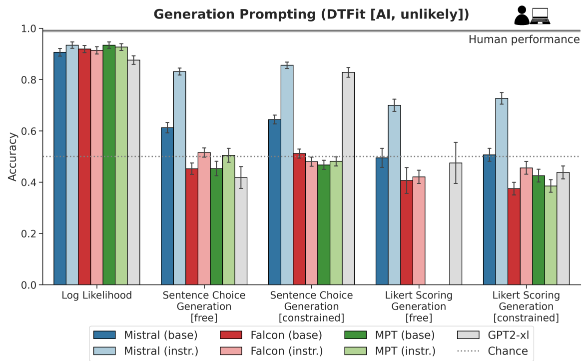


Figure 5: Comparison of free vs. constrained generation prompting. Note that MPT results are missing for the free Likert Scoring method.

B.2 Query types

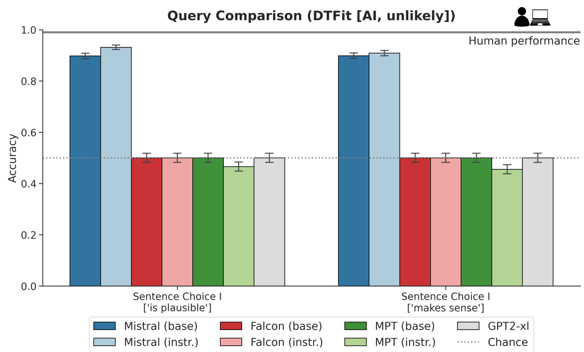


Figure 6: Comparison of different query types for prompts of type *Sentence Choice I*.

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

In Figure 7, we replicate the human experiment by Jouravlev et al. (2019) in LLMs using sentence log likelihood measurements. We generally observe similar trends than the comparison with the target word measurement.

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

To replicate the human experiment by Jouravlev et al. (2019) in LLMs 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*

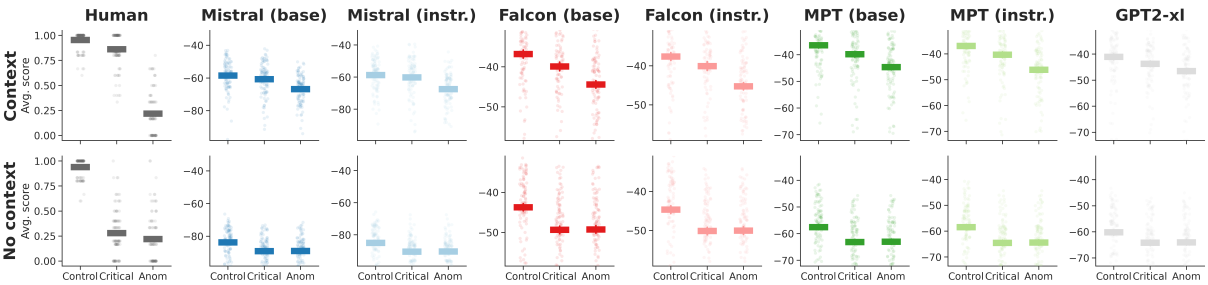


Figure 7: Replicating the sensibility-judgment task in LLMs using sentence *LL* measures. Human data from Jouravlev et al. (2019).

this sentence make sense considering the context? Respond with either Yes or No as your answer. We report our results in Figure 8.

We observe that while `GPT2-XL` and most base models often favor one answer option (`GPT2-XL` almost always assigns more probability mass to *Yes* rather than *No* in this setup; see also Figure 1, *Sentence Judgment*), 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 8, upper panel), and (ii) systematically overestimate the plausibility of implausible sentences relative to humans (`SemAnom` conditions and `Critical` condition, Figure 8, lower panel) in the binary sensibility-judgment task setup.

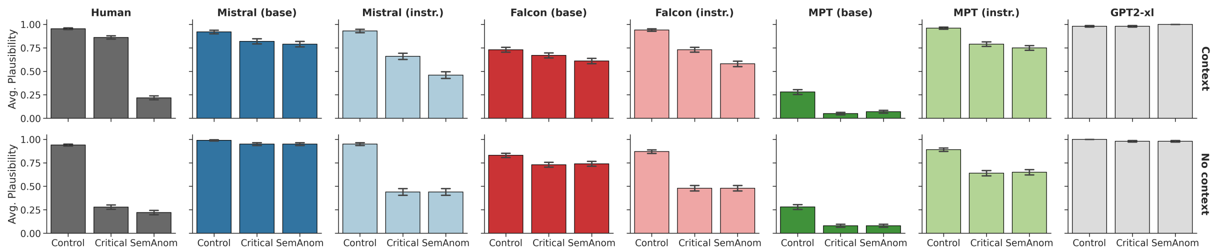


Figure 8: Replicating the sensibility-judgment task in LLMs using prompting via the adjusted *Sentence Judgment* prompt in §D. Human data from Jouravlev et al. (2019). We use a barplot to visually set apart this prompt-based comparison vs. *LL*-based ones in Figures 3, 7.