Leveraging Human Production-Interpretation Asymmetries to Test LLM Cognitive Plausibility

Anonymous ACL submission

Abstract

Whether large language models (LLMs) process language similarly to humans has been the subject of much theoretical and practical debate. We examine this question through the lens of the production-interpretation distinction found in human sentence processing and evaluate the extent to which instruction-tuned LLMs replicate this distinction. Using an empirically documented asymmetry between production and interpretation in humans for implicit causality verbs as a testbed, we find that some LLMs do quantitatively and qualitatively reflect human-like asymmetries between production and interpretation. We demonstrate that whether this behavior holds depends upon both model size-with larger models more likely to reflect human-like patterns-and the choice of meta-linguistic prompts used to elicit the behavior.

1 Introduction and Background

The extent to which large language models (LLMs) are "cognitively plausible," that is, replicate human-like behaviors in language processing, has been the subject of ongoing debate (Dentella et al., 2023; Hu et al., 2024; Futrell and Mahowald, 2025; Kuribayashi et al., 2025). Existing research on the linguistic capabilities of LLMs has predominantly focused on their performance in language interpretation, e.g., pragmatic understanding (Hu et al., 2023), sentence acceptability judgment (Warstadt et al., 2020), garden-path effect (Futrell et al., 2019), reference resolution (Lam et al., 2023). In this study, we examine a previously unexplored dimension of cognitive plausibility: whether LLMs reflect human-like distinctions between **production** and **interpretation** in language processing.

Production and interpretation were traditionally treated as two independent processes in human language: for instance, in neurolinguistics the "classic" Lichtheim–Broca–Wernicke model assumes distinct anatomical pathways associated with production and interpretation (see Ben Shalom and Poeppel 2008). While this extreme dichotomy has been rejected recently (see Pickering and Garrod 2013), humans do exhibit different underlying biases in language processing between production and interpretation even in very similar tasks. Whether such distinctions carry over into LLMs is of particular interest when we consider that the fundamental unit of LLM computation is $P(token|context)$, which is inherently ambiguous between production and interpretation and is practically applied towards both types of tasks.

The present study probes into this question using reference processing as a test case. Consider the following examples:

(1) A production task: next-mention bias
  a. John infuriated Bill. ... [IC1]
  b. John praised Bill. ... [IC2]

(2) A comprehension task: ambiguous pronoun resolution
  a. John infuriated Bill. He ... [IC1]
  b. John praised Bill. He ... [IC2]

Next-mention Bias for Subject	1a	>	1b
Where **Asymmetry** Happens...	∧		∧
Pronoun Resolution Bias for Subject	2a	>	2b

Figure 1: Illustration of the Asymmetry between **production** and **interpretation**

When asked to continue the story in (1), speakers usually describe events that happened to one of the two mentioned characters. This is a production task, and psycholinguistic research has investigated how the preceding context affects the next-mention bias of the character, i.e., how likely a character will be referred to in the continued story $P(referent|context)$. In (1-a), 'John' has a higher next-mention bias than 'Bill', because he is the implicit cause of the event. Verbs like 'infuriate' are therefore called the subject-biased implicit causality (IC1) verbs. In contrast, 'Bill' has a higher next-mention bias than 'John' in (1-b), because 'Bill' implies an implicit cause for 'John' to praise him. Verbs like 'praise' are therefore called object-biased implicit causality (IC2) verbs (e.g., Stevenson et al. 1994). A similar implicit causality bias can also be found in an interpretation task like (2). Unlike (1), here human participants must resolve the ambiguous pronoun "he" $P(referent|pronoun)$ first before providing a reasonable continuation. Again, participants were more likely to resolve the ambiguous pronoun to the subject 'John' than the object 'Bill' with an IC1 verb, and to 'Bill' more than 'John' with an IC2 verb (e.g., Crinean and Garnham 2006).

Crucially, psycholinguistic research has revealed an asymmetry between these two biases. In the interpretative case of pronoun resolution bias, humans are robustly more likely to show a preference for the subject than in the production case of next-mention bias cross-linguistically (English: Rohde and Kehler 2014; Mandarin: Zhan et al. 2020; Lam and Hwang 2024; German: Patterson et al. 2022; Catalan: Mayol 2018). That is, participants were more likely to resolve an ambiguous pronoun towards the subject than choose the subject as the next referent, despite the same context.

For humans, this extra subject bias in interpretation comes from the bias of using pronouns for subject antecedent, i.e., $P(he|subject)$. That is, when they see an ambiguous pronoun, they do not only consider the next-mention bias of which antecedent is more likely to be mentioned, but also why a pronoun is used. It is unknown whether and how LLMs can handle this difference, as they do not generate such a conditional probability based on the choice of next referent instead of the context.

We therefore probe this dimension of LLM cognitive plausibility using this task, asking (1) whether the IC verb-type effect is reflected by LLMs in both production and interpretation; and (2) whether a human-like asymmetry between the two biases exists. This is our first set of questions: do LLMs show human-like interpretation and production biases, and if so under what conditions? Do human-like effects scale with parameter count?

Evaluating LLM in metalinguistic prompts Hu and Levy (2023) demonstrated that direct probability-based measures in general outperformed meta-linguitic prompting in assessing plausibility and syntactic processing tasks. However, not all language processing tasks can be effectively quantified using probability-based measures, and for some tasks meta-linguitic prompts are the only possible method to measure processing. This is exactly our case: in the ambiguous pronoun resolution task, the bias towards the subject 'John' or the object 'Bill' is represented by the same term, i.e., $P(he|context)$. Metalinguistic prompting is thus necessary to elicit meaningful results. This constitutes the second aim of this paper: across different metalinguistic prompting strategies, which elicit more human-like language processing behaviors?

Our results show that (1) certain LLMs can exhibit human-like production-interpretation asymmetries in reference processing with specific metalinguistic prompts; and (2) the choice of metalinguistic prompts matters in evaluating LLMs: the model performs in a more human-like manner with certain prompts compared to others, and this best-performing prompt varies by model.

2 Methodology

2.1 Stimuli

We constructed stimuli in the frame of "[Character A] IC-verb [Character B]" without pronouns for the next-mention production task (as in (1))), and with a pronoun with an ambiguous referent for the pronoun resolution task (as in (2)). We selected 137 IC1 verbs and 134 IC2 verbs from the original study that found the production-interpretation asymmetry (Rohde and Kehler, 2014), and the English IC verb corpus (Ferstl et al., 2011). For each verb, we heuristically created two items by assigning a pair of male names and a pair of female names randomly selected from 13 unambiguously male and 13 unambiguously female names from Rohde (2008). The congruence of the gender of the characters ensures the ambiguity of the pronoun. This results in 541 items in each task.

2.2 Models and Metalinguistic Prompting

Our experiments evaluate three representative LLMs, ranging in scale from 8B to 70B parameters: `LLaMA3.1Instruct-8B`, `QWen2.5Instruct-32B`, and `LLaMA3.3Instruct-70B`. We focus on instruct-tuned models as they allow the effective use of metalinguistic prompting, and varying parameter counts also allow us to assess the impact of model scaling on human-like language processing behavior. We employ four metalinguistic prompt strategies to assess LLM behavior:

(1) Binary choice prompting: The model is prompted to select between subject and object.

(2) Continuation prompting: The model is instructed to extend the sentence by continuing with either the subject or the object.

(3) Yes/no prompting: The model is asked whether the following sentence (or the existing pronoun) will begin with the *subject*, requiring a binary response.

(4) Yes/no probability prompting: Similar to (3), but instead of a categorical response, we extract the probability assigned to the *Yes* token as a quantitative measure.

Three authors of the paper manually verified all model outputs to confirm subject/object choice and exclude ambiguous, nonsensical, and plural responses. The specific prompts used in our experiments are provided in Appendix A.

2.3 Evaluation

We evaluated model behavior from two perspectives. We first consider whether the IC verb type effect and the production-interpretation asymmetry exists in LLMs, which we both visualize to observe broad trends and verify with statistical tests like those run in human experiments. We then consider the magnitude of the effect found in LLMs, as psycholinguistic research has found that language models often fail to replicate the magnitude of effects found in human participants even when the directionality is similar. Below we only report effects that are verified by statistical evidence, the details of which be found in Appendix B.

3 Results

Result #1: Implicit causality biases are sometimes replicated by LLMs, depending on prompts and models. Figure 2 presents the task performance across different metalinguistic prompts for each model. The implicit causality (IC) verb effect—where subjects are chosen more frequently after IC1 verbs than IC2 verbs—was observed in at least one metalinguistic prompt for all models, though the strength and consistency of this effect varied by both model and prompt type.

For LLaMA models, the IC verb effect was present in both production and interpretation when using Yes/no and Yes/no probability prompts, indicating a broader sensitivity to IC biases across tasks. In contrast, for QWen, the IC verb effect was only observed in interpretation and limited to binary choice prompt, suggesting that its sensitivity

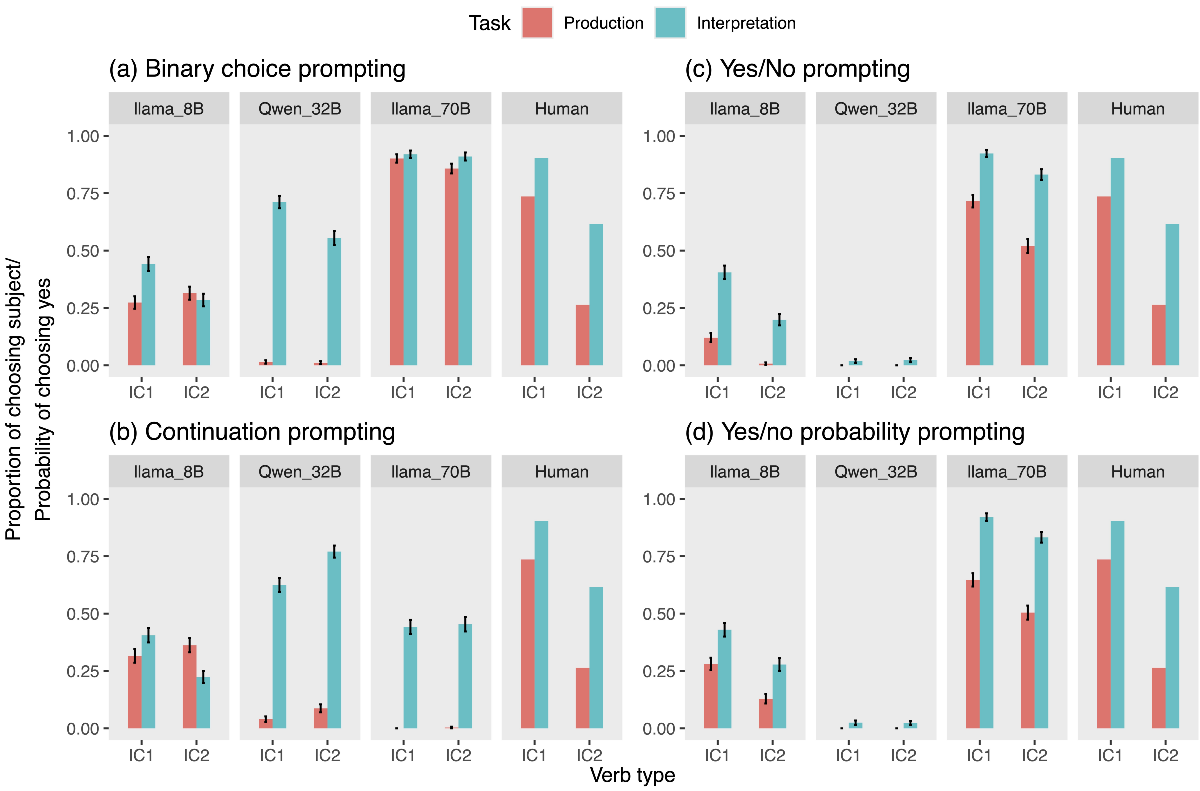

Figure 2: Model behavior as proportion of subject/yes choice as the antecedent by prompting strategy. Human behavior – rightmost facet in each subfigure for reference, from Rohde and Kehler (2014) – tends to reflect higher subject bias for IC1 over IC2 verbs, but with asymmetry between production and interpretation tasks.

to IC verbs may depend on task framing.

Result #2: The production-interpretation asymmetry. Recall that human participants are more subject-biased in interpretation than in production. This difference is only elicited in LLaMA models using Yes/No and Yes/No probability prompts. QWen was entirely unable to capture this asymmetry because it failed to predict the IC verb effect in production with all four metalinguistic prompts.

Another unexpected production-interpretation asymmetry is that LLMs generally are less likely to capture the IC verb bias in production than in interpretation. For instance, while the LLaMA-8B model was able to predict an IC verb effect in interpretation when using binary choice and continuation prompts, it predicted a reverse verb type effect in production using these prompts. This unexpected asymmetry is even clearer in the performance of the QWen model, in which the IC verb effect is only found in interpretation.

Result #3: LLMs do not always align with human behavior in effect magnitude. Although there was a verb type effect and a production/interpretation task effect on LLMs' responses, the magnitude of these effects is different from humans even in the most similar case: with yes/no prompting, LLaMA-70B predicted an IC verb effect difference of 19.4% in the production task and 9.2% in the interpretation task. Yet, these two differences were 47.2% and 28.8% respectively in human participants (Rohde and Kehler, 2014). In other words, even in the best scenario, LLMs underestimate the magnitude of the IC verb effect.

Surprisingly, the magnitude of the production-interpretation asymmetry predicted by LLMs are sometimes well-aligned with humans: in human participants, this asymmetry was 16.8% and 35.2% respectively for IC1 and IC2 verbs. With yes/no prompting, LLaMA-70B predicted a 20.8% and 31% difference respectively for the two verb types.

Result #4: Scaling matters. Overall, LLaMA-70B shows better performance than LLaMA-8B across all four prompts. With Binary choice prompts, LLaMA-8B predicted a reverse IC verb effect in the production task and a reverse production-interpretation asymmetry in IC2 verbs, while both effects were captured by LLaMA-70B. With Yes/No and Yes/No probability prompts, although both LLaMA-8B and LLaMA-70B reflected an IC verb effect and production-interpretation asymmetry, the magnitude in human participants was better approximated by LLaMA-70B.

4 Discussion

Our study examines the asymmetry between interpretation and interpretation in humans within the context of LLMs, showing that, under specific prompting strategies, certain LLMs can approximate human-like asymmetry.

Among all prompting strategies, continuation prompting performs the worst in capturing human-like performance, despite the fact that it was the task performed by human participants. One possible explanation is that instruct-tuned LLMs, having been fine-tuned with instruction data or preference optimization objectives, may develop constrained response patterns (Lin et al., 2023), limiting their flexibility in generating diverse continuations. Additionally, such fine-tuning can reduce conceptual diversity (Murthy et al., 2024), which may make LLMs less sensitive to implicit biases in language processing. This suggests that continuation prompting may not be well-suited for probing human-like asymmetries in interpretation and production.

Our findings align in part with Hu and Levy (2023), suggesting that metalinguistic prompting is less consistent than direct probability-based measures. In our study, the four distinct prompting strategies led to considerable variation in LLM performance, highlighting the sensitivity of model responses to prompt formulation. Additionally, the fact that different model families exhibited distinct preferences for specific prompting strategies underscores the need for extensive experimentation across multiple approaches to obtain more reliable and generalizable insights.

Notably, the strong performance of Yes/No probability prompting in LLaMA models, which corresponds to $P(Yes)$ in Kadavath et al. (2022), suggests that uncertainty-based measures may offer a more robust method for eliciting linguistic phenomena in LLMs. This finding indicates that rather than relying solely on direct metalinguistic judgments, probability-based approaches could provide a more stable and interpretable means of assessing model biases and linguistic competence. Future work should further explore the interaction between model architecture, fine-tuning objectives, and prompting strategies to refine methodologies for probing LLMs' linguistic representations.

Limitations

Our work has two primary limitations. First, our experiments are conducted on two English-centric LLMs (LLaMA) and one multilingual-oriented LLM (QWen). This selection may introduce biases into the models' performance, potentially limiting the generalizability of our findings across other LLMs. Second, our study focuses solely on the asymmetry between ambiguous pronoun resolution and production in English, without exploring cross-linguistic variations. Future research could address these limitations by incorporating a more diverse set of LLMs and broadening the scope of languages analyzed.

References

Dale J Barr, Roger Levy, Christoph Scheepers, and Harry J Tily. 2013. Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of memory and language*, 68(3):255–278.

Dorit Ben Shalom and David Poeppel. 2008. Functional anatomic models of language: assembling the pieces. *The Neuroscientist*, 14(1):119–127.

Marcelle Crinean and Alan Garnham. 2006. Implicit causality, implicit consequentiality and semantic roles. *Language and Cognitive Processes*, 21(5):636–648.

Vittoria Dentella, Fritz Günther, and Evelina Leivada. 2023. Systematic testing of three language models reveals low language accuracy, absence of response stability, and a yes-response bias. *Proceedings of the National Academy of Sciences*, 120(51):e2309583120.

Evelyn C Ferstl, Alan Garnham, and Christina Manouilidou. 2011. Implicit causality bias in english: A corpus of 300 verbs. *Behavior Research Methods*, 43:124–135.

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

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 the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 32–42, Minneapolis, Minnesota. Association for Computational Linguistics.

Andrew Gelman, Aleks Jakulin, Maria Grazia Pittau, and Yu-Sung Su. 2008. A weakly informative default prior distribution for logistic and other regression models.

Jennifer Hu, Sammy Floyd, Olessia Jouravlev, Evelina Fedorenko, and Edward Gibson. 2023. A fine-grained comparison of pragmatic language understanding in humans and language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4194–4213, Toronto, Canada. Association for Computational Linguistics.

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

Jennifer Hu, Kyle Mahowald, Gary Lupyan, Anna Ivanova, and Roger Levy. 2024. Language models align with human judgments on key grammatical constructions. *Proceedings of the National Academy of Sciences*, 121(36):e2400917121.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, Scott Johnston, Sheer El-Showk, Andy Jones, Nelson Elhage, Tristan Hume, Anna Chen, Yuntao Bai, Sam Bowman, Stanislav Fort, Deep Ganguli, Danny Hernandez, Josh Jacobson, Jackson Kernion, Shauna Kravec, Liane Lovitt, Kamal Ndousse, Catherine Olsson, Sam Ringer, Dario Amodei, Tom Brown, Jack Clark, Nicholas Joseph, Ben Mann, Sam McCandlish, Chris Olah, and Jared Kaplan. 2022. Language models (mostly) know what they know. *Preprint*, arXiv:2207.05221.

Tatsuki Kuribayashi, Yohei Oseki, Souhaib Ben Taieb, Kentaro Inui, and Timothy Baldwin. 2025. Large language models are human-like internally. *Preprint*, arXiv:2502.01615.

Suet-Ying Lam and Heeju Hwang. 2024. Pronoun interpretation is more subject-biased than expected by the bayesian model. *Language, Cognition and Neuroscience*, pages 1–20.

Suet-Ying Lam, Qingcheng Zeng, Kexun Zhang, Chenyu You, and Rob Voigt. 2023. Large language models are partially primed in pronoun interpretation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 9493–9506, Toronto, Canada. Association for Computational Linguistics.

Bill Yuchen Lin, Abhilasha Ravichander, Ximing Lu, Nouha Dziri, Melanie Sclar, Khyathi Chandu, Chandra Bhagavatula, and Yejin Choi. 2023. The unlocking spell on base llms: Rethinking alignment via in-context learning. *Preprint*, arXiv:2312.01552.

Laia Mayol. 2018. Asymmetries between interpretation and production in catalan pronouns. *Dialogue Discourse. 2018 Dec; 9 (2): 1-34*.

Sonia K. Murthy, Tomer Ullman, and Jennifer Hu. 2024. One fish, two fish, but not the whole sea: Alignment reduces language models' conceptual diversity. *Preprint*, arXiv:2411.04427.

Clare Patterson, Petra B Schumacher, Bruno Nicenboim, Johannes Hagen, and Andrew Kehler. 2022. A bayesian approach to german personal and demonstrative pronouns. *Frontiers in psychology*, 12:672927.

Martin J Pickering and Simon Garrod. 2013. An integrated theory of language production and comprehension. *Behavioral and brain sciences*, 36(4):329–347.

Hannah Rohde. 2008. *Coherence-driven effects in sentence and discourse processing*. University of California, San Diego.

Hannah Rohde and Andrew Kehler. 2014. Grammatical and information-structural influences on pronoun production. *Language, Cognition and Neuroscience*, 29(8):912–927.

Rosemary J Stevenson, Rosalind A Crawley, and David Kleinman. 1994. Thematic roles, focus and the representation of events. *Language and cognitive processes*, 9(4):519–548.

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.

Meilin Zhan, Roger Levy, and Andrew Kehler. 2020. Pronoun interpretation in mandarin chinese follows principles of bayesian inference. *PloS one*, 15(8):e0237012.

A Prompt Template

The following shows the metalinguistic prompt design for LLMs.

```
System Prompt: You are a helpful assistant.

production-binary-choice-template:
In the following sentence, who is more likely to be
the subject of the next sentence? {} or {}? Please
ONLY return the name without any explanation or
extra words.
Sentence: {}
Answer:

production-yes-no-template:
In the following sentence, judge whether the pronoun
of the next sentence will refer to {}. Please ONLY
answer with 'Yes' or 'No'.
Sentence: {}

production-continuation-template:
Please reasonably continue the sentence with one of
the mentioned characters. You should start a new
sentence rather than a clause. Please ONLY return
the continuation.
Sentence: {}

interpretation-binary-choice-template:
In the following sentence, who is more likely to be
the referent of the pronoun? {} or {}? Please ONLY
return the name without any explanation or extra
words.
Sentence: {}
Answer:

interpretation-yes-no-template:
In the following sentence, judge whether the pronoun
refers to {}. Please ONLY answer with 'Yes' or
'No'.
Sentence: {}

interpretation-continuation-template:
Please reasonably continue the sentence following
the pronoun. Please ONLY return the continuation.
Sentence: {}
```

B Statistical analyses and results

We focus on how the bias of IC verbs (IC1 vs. IC2) and the task (production vs. interpretation) affect the outcomes of LLMs in each type of metalinguistic prompts.

B.1 Data annotation

For binary prompts, we annotated based on the name answered by the LLM. For yes-no prompts, we annotated "yes" as “Subject" .

For continuation prompts, we manually annotated the choice of referent (Subject vs. Object) in the production task and the choice of antecedent (Subject vs. Object) in the ambiguous pronoun resolution task based on the meaning of the generated continuation. For instance, for a sentence like “Nick offended Steve. He decided to apologize and clear the air before things escalated further", we annotated the outcome as “Subject". Ambiguous references (e.g., “Zack divorced Paul. He later moved to a new city to start his life over."), nonsensical continuations (e.g., “Janet wanted Kate. She to join her at the party that night, but Kate had already made other plans."), and continuations with a plural antecedent (e.g., "Claire played with Jane. They were building a sandcastle on the beach") were excluded. Table 1 reported the distribution of the excluded responses for each model over the 1082 responses in the two tasks.

Table 1: The distribution of excluded responses in continuation prompting

	Ambiguous	Non-sensical	Plural
LLaMA-8B	54	21	17
LLaMA-70B	35	9	0
Qwen	17	6	0

As can be seen, LLaMA-8B made more ambiguous and non-sensical continuations than all other models. Besides, it also provided continuations with "they", which violates the requirement posed in the prompts. This might suggest that scaling affects LLMs' ability to follow the instructions. More models should be evaluated to test this hypothesis.

B.2 Analyses

LLaMA models For the results of LLaMA models, we ran a mixed-effects Bayesian bernoulli regression model using the R package `brms` for the binary outcome resulted from continuation, binary, and yes-no prompts (Subject = 1; Object = 0); and a mixed-effects Bayesian linear regression model for the continuous probability outcome from yes-no probability prompt.

Each model was fitted using 4 chains, each with 5000 iterations. The first 1000 were warm-up to calibrate the sampler. This results in 12000 posterior samples. They were all built with fixed predictors of IC verbs (sum-coded: IC1 = 0.5; IC2 = -0.5), task type (sum-coded: interpretation = 0.5; production = -0.5), and their interaction. A maximal random structure justified by design is implemented (Barr et al., 2013). For logistic regression models, we used weakly informative priors, i.e., a Cauchy distribution with a center of 0 and a scale of 2.5 for fixed effects following Gelman et al. (2008), and the default setting of the package for the other parameters. For linear regression models, we used a gaussian distribution with a mean of 0 and a standard deviation of 1 as the weakly informative prior for fixed predictors. When there is an interaction effect, we further ran nested models for pairwise comparison.

The Bayesian statistics framework does not use the p-value. We consider the 95% credible interval (Crl) as the evidence for an effect: if the 95% Crl does not include a zero, i.e., it is all positive or negative, we consider there is evidence for an effect. Below we report the estimate and the 95% Crl for each effect.

QWen For QWen, we only ran analysis for continuation prompts, and limited to the effect of the IC verb only. This is because the responses of QWen in the production task is so extreme that no statistical model can be successfully fitted. The setting of the mixed-effects Bayesian bernoulli regression model is the same as those used for LLaMA models.

B.3 Results

We bold the predictor in which the effect is supported the statistical evidence, i.e., the 95% Crl does not contain a zero.

B.3.1 LLaMA-3.1-8B Model

Binary prompting Table 2 shows that there was an interaction effect between the IC verb type and the task type.

Formula: binary $\sim$ verb $*$ task $+ (1 + \text{task}\mid\text{itemID})$			
	Estimate	Est. Error	95% CrI
Intercept	-1.80	0.74	**[-3.78, -0.99]**
verb	0.64	0.48	[-0.12, 1.80]
task	0.67	0.75	[-0.65, 2.47]
verb:task	2.06	0.99	**[0.78, 4.61]**

Table 2: Summary of the Bayesian logistic regression model of LLaMA-3.1-8B model, binary choice prompting.

Nested analyses further reveal that the production-interpretation asymmetry is only found in IC1 verbs (Table 3), and the verb type effect is only found in interpretation (Table 4).

Formula: binary $\sim$ verb/task $+ (1 + \text{task}\mid\text{itemID})$			
	Estimate	Est. Error	95% CrI
Intercept	-1.84	0.76	**[-3.81, -0.99]**
verb	0.63	0.47	[-0.14, 1.76]
verbIC1:task	1.76	0.96	**[0.55, 4.24]**
verbIC2:task	-0.40	0.82	[-2.30, 1.14]

Table 3: Pairwise comparison of the task type effect within IC1 and within IC2 conditions using binary choice prompting in LLaMA-3.1-8B model.

Formula: binary $\sim$ task/verb + (1 + prompt|itemID)			
	Estimate	Est. Error	95% CrI
Intercept	-1.80	0.79	**[-3.85, -0.99]**
task	0.65	0.73	[-0.68, 2.37]
taskProduction:verb	-0.43	0.50	[-1.57, 0.42]
taskInterpretation:verb	1.69	0.89	**[0.61, 4.01]**

Table 4: Pairwise comparison of the IC verb type effect within the production and within the interpretation task using binary choice prompting in LLaMA-3.1-8B model.

Yes-no prompting As shown in Table 5, there was an interaction effect between the verb type and task type for the responses of the model.

Formula: yes_no $\sim$ verb $*$ task + (1|itemID)			
	Estimate	Est. Error	95% CrI
Intercept	-4.06	0.61	**[-5.44, -3.04]**
verb	3.12	0.67	**[1.97, 4.61]**
task	4.31	0.70	**[3.12, 5.89]**
verb:task	-1.99	0.93	**[-4.01, -0.39]**

Table 5: Summary of the Bayesian logistic regression model of LLaMA-3.1-8B model, yes-no prompting.

Nested analysis in Table 6 further reveals that the model did give more "yes" (or referring to subject) with IC1 verbs than IC2 verbs (as indicated by the positive intercept of the verb predictor). Also, the production-interpretation asymmetry is found within both IC1 and IC2 verbs, such that the model chose more subjects in interpretation than production.

Formula: yes_no $\sim$ verb/task + (1|itemID)			
	Estimate	Est. Error	95% CrI
Intercept	-4.06	0.62	**[-5.50, -3.04]**
verb	3.18	0.70	**[2.00, 4.76]**
verbIC1:task	3.22	0.55	**[2.28, 4.43]**
verbIC2:task	5.45	1.14	**[3.58, 8.04]**

Table 6: Pairwise comparison of the task effect within the IC1 and within the IC2 verbs using yes-no choice prompting in LLaMA-3.1-8B model.

Continuation prompting As shown in Table 7, there was an interaction effect between the verb type and the task type for the responses of the model.

Formula: cont $\sim$ verb $*$ task + (1|itemID)			
	Estimate	Est. Error	95% CrI
Intercept	-0.81	0.09	**[-0.99, -0.65]**
verb	0.36	0.16	**[0.05, 0.67]**
task	-0.15	0.14	[-0.44, 0.13]
verb:task	1.14	0.30	**[0.57, 1.73]**

Table 7: Summary of the Bayesian logistic regression model of LLaMA-3.1-8B model, continuation prompting.

Nested analyses in Table 8 show that the verb type effect is only found in interpretation.

Formula: cont $\sim$ task/verb + (1|itemID)			
	Estimate	Est. Error	95% CrI
Intercept	-0.81	0.09	**[-0.99, -0.64]**
task	-0.15	0.14	[-0.44, 0.13]
taskProduction:verb	-0.22	0.21	[-0.64, 0.18]
taskInterpretation:verb	0.93	0.22	**[0.51, 1.37]**

Table 8: Pairwise comparison of the IC verb type effect within the production and within the interpretation task using continuation prompting in LLaMA-3.1-8B model.

The pairwise comparison in Table 9 shows that the production-interpretation asymmetry differs in direction between the two verb types: while interpretation is more subject-biased than production in IC1 verbs, production is more subject-biased than interpretation in IC2 verbs.

Formula: cont $\sim$ verb/prompt + (1|itemID)			
	Estimate	Est. Error	95% CrI
Intercept	-0.81	0.09	**[-0.99, -0.65]**
verb1	0.36	0.15	**[0.06, 0.66]**
verbIC1:prompt1	0.42	0.20	**[0.05, 0.81]**
verbIC2:prompt1	-0.73	0.21	**[-1.14, -0.32]**

Table 9: Pairwise comparison of the task type effect within IC1 and within IC2 verbs using continuation prompting in LLaMA-3.1-8B model.

Yes/no probability prompting As shown in Tabel 10, the model did generate a higher probability of ‘Yes’ (= subject) following IC1 verbs than IC2 verbs, and in interpretation task than in production task.

Formula: subject_yes_probability $\sim$ verb $*$ task + $(1\|\text{itemID})$			
	Estimate	Est. Error	95% CrI
Intercept	0.28	0.01	**[0.27, 0.29]**
verb	0.15	0.01	**[0.12, 0.18]**
task	0.15	0.01	**[0.13, 0.17]**
verb: task	-0.00	0.02	[-0.03, 0.03]

Table 10: Summary of the Bayesian linear regression model of LLaMA-3.1-8B model, yes/no probability prompting.

B.3.2 LLaMA-3.3-70B

Binary prompting Table 11 clearly shows that the model only reveals the production-interpretation asymmetry, such that it chose more subject in interpretation than in production. There was no clear evidence for the IC verb type effect.

Formula: binary $\sim$ verb $*$ task + $(1\|\text{itemID})$			
	Estimate	Est. Error	95% CrI
Intercept	14.61	3.61	**[9.24, 23.34]**
verb	1.36	1.42	[-1.19, 4.45]
task	2.21	0.73	**[0.98, 3.84]**
verb:task	-1.24	1.07	[-3.44, 0.77]

Table 11: Summary of the Bayesian logistic regression model of LLaMA-3.3-70B model, binary prompting.

Yes-no prompting Table 12 shows that the model is able to capture the IC verb type effect and the production-interpretation asymmetry, such that it responded ‘Yes’ (=subject) more for IC1 verbs than IC2 verbs and the interpretation task than the production task.

Formula: yes_no $\sim$ verb $*$ task + $(1\|\text{itemID})$			
	Estimate	Est. Error	95% CrI
Intercept	2.24	0.25	**[1.79, 2.78]**
verb	1.46	0.34	**[0.83, 2.15]**
task	2.63	0.31	**[2.05, 3.29]**
verb:task	-0.10	0.44	[-0.95, 0.76]

Table 12: Summary of the Bayesian logistic regression model of LLaMA-3.3-70B model, Yes/no prompting.

Continuation prompting Like in binary choice prompting, Table 13 shows that the model only reveals the production-interpretation asymmetry, such that it chose more subject in interpretation than in production. There was no clear evidence for the IC verb type effect.

Formula: cont $\sim$ verb $*$ task + $(1\|\text{itemID})$			
	Estimate	Est. Error	95% CrI
Intercept	-3.66	0.66	[-5.23, -2.68]
verb	-0.61	0.88	[-2.67, 0.86]
task	6.83	1.27	**[4.89, 9.82]**
verb:task	1.10	1.74	[-1.78, 5.19]

Table 13: Summary of the Bayesian logistic regression model of LLaMA-3.3-70B model, continuation prompting.

Yes/no probability prompting Table 14 shows an interaction effect between verb and task type.

Formula: subject_yes_probability $\sim$ verb $*$ task + $(1\|\text{itemID})$			
	Estimate	Est. Error	95% CrI
Intercept	0.73	0.01	**[0.70, 0.75]**
verb	0.12	0.02	**[0.07, 0.16]**
task	0.30	0.01	**[0.27, 0.33]**
verb:task	-0.05	0.03	**[-0.11, -0.00]**

Table 14: Summary of the Bayesian linear regression model of LLaMA-3.3-70B model, Yes/no probability prompting.

Pairwise comparisons in Table 15 and 16 further show that the verb type effect can be found in both production and interpretation task, and the asymmetry can be found in both IC1 and IC2 conditions.

Formula: subject_yes_probability $\sim$ verb/task + $(1\|\text{itemID})$			
	Estimate	Est. Error	95% CrI
Intercept	0.73	0.01	**[0.71, 0.75]**
verb	0.12	0.02	**[0.07, 0.16]**
verbIC1:task	0.27	0.02	**[0.24, 0.31]**
verbIC2:task	0.33	0.02	**[0.29, 0.37]**

Table 15: Pairwise comparison of the verb type effect within the production and the interpretation task using yes/no probability prompting in LLaMA-3.1-70B model.

B.3.3 QWen

Note that we only analyzed the responses of the continuation prompting in the interpretation task for QWen, because models cannot be converged

Formula: subject_yes_probability $\sim$ task/verb + (1|itemID)			
	Estimate	Est. Error	95% CrI
Intercept	0.73	0.01	**[0.70, 0.75]**
task	0.30	0.01	**[0.27, 0.33]**
taskProduction:verb	0.14	0.03	**[0.09, 0.19]**
taskInterpretation:verb	0.09	0.03	**[0.04, 0.14]**

Table 16: Pairwise comparison of the task type effect within IC1 and within IC2 verbs using yes/no probability prompting in LLaMA-3.1-70B model.

in other case. As can be seen below, there is an opposite IC verb type effect such that the model referred to more subjects following IC2 verbs than IC1 verbs.

Formula: cont $\sim$ verb + (1|itemID)			
	Estimate	Est. Error	95% CrI
Intercept	-0.60	0.07	**[-0.73, -0.47]**
verb	-0.37	0.13	**[-0.63, -0.11]**

Table 17: Summary of the Bayesian logistic regression model of QWen model, continuation prompting.