Hey, wait a minute: on at-issue sensitivity in Language Models

Sanghee J. Kim
Microsoft AI*
kimsanghee@microsoft.com

Kanishka Misra
The University of Texas at Austin
kmisra@utexas.edu

Abstract

Evaluating the naturalness of dialogue in language models (LMs) is not trivial: notions of *naturalness* vary, and scalable quantitative metrics remain limited. This study leverages the linguistic notion of *at-issueness* to assess dialogue naturalness and introduces a new method: Divide, Generate, Recombine, and Compare (DGRC). DGRC (i) divides a dialogue as a prompt, (ii) generates continuations for subparts using LMs, (iii) recombines the dialogue and continuations, and (iv) compares the likelihoods of the recombined sequences. This approach mitigates bias in linguistic analyses of LMs and enables systematic testing of discourse-sensitive behavior. Applying DGRC, we find that LMs prefer to continue dialogue on at-issue content, with this effect enhanced in instruction-tuned models. They also reduce their at-issue preference when relevant cues (e.g., "Hey, wait a minute") are present. Although instruction-tuning does not further amplify this modulation, the pattern reflects a hallmark of successful dialogue dynamics.

1 Introduction

Language models (LMs) have made substantial progress in dialogue quality. Findings and surveys report meaningful improvements, including in areas such as user engagement (Ferron et al., 2023) and personalization (Zhang et al., 2025). Nevertheless, evaluating LMs' dialogue responses remains a central challenge (Guan et al., 2025, for a review). Dialogue varies in topic, style, and length, and human judgments are often inconsistent across evaluators. To address this, recent work used predefined evaluation criteria (e.g., Lin and Chen, 2023), but even notions like *naturalness* remain vague. For instance, "Could the utterance have been produced by a native speaker?" is a criterion used for evaluating naturalness (e.g., Reddy, 2022), but it offers little guidance. To address this limitation, we propose a systematic and linguistically grounded approach to evaluating dialogue *naturalness*.

Our approach is defined by two core characteristics. First, we ground *naturalness* in the linguistic notion of *at-issueness* (Potts, 2005; Koev, 2022; Simons et al., 2010; Tonhauser, 2012), which distinguishes between content that advances the discourse (*at-issue*) and content that supplements it without shifting the conversational trajectory (Hunter and Asher, 2016; Jasinskaja, 2016; Riester, 2019). While there have been approaches to apply the broad notion of at-issueness to discourse analysis (e.g., Ko et al., 2022; Kim et al., 2022), its use in evaluating modern LMs is limited. Second, building on a minimal-pair, or templatic format widely used in evaluating syntactic, semantic, and pragmatic abilities of language models (Marvin and Linzen, 2018; Warstadt et al., 2020; Misra et al., 2023; Hu et al., 2022), we introduce Divide, Generate, Recombine, and Compare (**DGRC**). While the existing templatic approach provides controlled contexts for testing model sensitivity, it reduces to only evaluating a pre-defined pair of sentences, which is less ideal for capturing conversational dynamics. DGRC retains the structure of minimal contrasts but replaces full-sentence pairs with multiple possible utterance fragments sampled from the model itself, differing in the content within the previous utterance they refer to. In this way, DGRC combines the strengths of templatic, surprisal-based evaluation with the flexibility of open-ended continuation while enabling a linguistic assessment of discourse dynamics, which has been less deeply considered in previous work.

We find that LMs show a preference for *at-issue* content: they are more likely to respond to an utterance's at-issue than its not-at-issue content. Moreover, instruction-tuned models exhibit this tendency more strongly than non-instruction-tuned models, even when controlling for a recency ef-

*Work done while at the University of Chicago.

Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics
Volume 2: Short Papers, pages 214–224
March 24-29, 2026 ©2026 Association for Computational Linguistics

fect. Yet, this advantage disappears when digression cues reduce at-issue preference. These results illustrate one representative case in which DGRC can be applied as a principled method for evaluating naturalness in language models.

2 Dialogue naturalness and *at-issueness*

We treat dialogue as a single-turn pair (U, R), with U as the initiating utterance and R the response. To evaluate whether a response R constitutes a *natural* continuation of U, we rely on the linguistic notion of *at-issueness*. **At-issueness** captures if content conveys the main point of an utterance. If so, it is *at-issue*; if parenthetical and backgrounded, it is *not at-issue* (Potts, 2005, a.o.). In (1), what is at-issue in U is Sally's encounter with the governor, while her dating status is not-at-issue. At-issueness underlies conversational naturalness: a natural response typically targets at-issue content. Responses addressing not-at-issue content feel unnatural (e.g., "Yes, Sally and Dave are dating"), but an explicit digression signal (e.g., "Hey, wait a minute") can make them natural (Syrett and Koev, 2015).

(1) U: Sally—Dave's girlfriend—met the Illinois governor at a restaurant.
R: She met the governor at a dog park.
(targets at-issue)
R: Hey, wait a minute, Sally's dating?
(targets not-at-issue with a digression signal)

Following Kim et al. (2022), we use appositive relative clauses (**ARC**s)—embedded clauses marked by commas—as a stand-in for not-at-issue content:

(2) The librarian, [who likes pasta]$_{\text{VerbPhrase1}}$, [is famous]$_{\text{VerbPhrase2}}$!

In constructions as (2), the main clause (MC) conveys at-issue content, while the embedded ARC contributes not-at-issue information. ARCs are useful because: 1) they are *parentheticals*, conventionally treated as not-at-issue (AnderBois et al., 2010; Potts, 2005), and 2) they embed both (not-)at-issue content within a single sentence (Jasinskaja, 2016), avoiding the need for multi-sentence setup. We denote ARC as VP1 and MC as VP2 hereafter since this construction will be contrasted with a minimally different one in our experiments.

Building on background on response dynamics involving ARC structures, we distill two guiding intuitions. First, responses are more likely to target at-issue content than not-at-issue content:

$$p_{\text{LM}}(R_{\text{at-issue}} \mid U) > p_{\text{LM}}(R_{\text{not-at-issue}} \mid U)$$

Second, responses addressing at-issue content become less likely when they start with a digression signal (e.g., "Hey, wait a minute"), relative to when this digression signal is absent:

$$p_{\text{LM}}(R_{\text{at-issue}} \mid U) > p_{\text{LM}}(R_{\text{at-issue}} \mid U, R_d),$$

where R_d is a minimally added response "header" that signals digression, and p_{LM} is the language model's probability. These two aspects serve as foundations for Experiments 1 and 2, respectively.

3 Method

It is tempting to analyze at-issue sensitivity through minimal-pair judgments, akin to a range of linguistically motivated tests (Marvin and Linzen, 2018; Warstadt et al., 2020; Misra et al., 2023). One can easily define a pair of fixed responses, each targeting a different part of the utterance, and use LM probabilities for forced-choice judgments (*à la* Kim et al., 2022). While the idea of using LM-probabilities is directly applicable, using a forced-choice paradigm is fundamentally limiting when it comes to response dynamics. First, the researcher has to select a fixed surface form as the "correct" response even though there could be a range of different responses that target an utterance's at-issue content. Second, even if one could hand-code a range of responses, these will be susceptible to researcher bias and may deviate from the LM's space of possible responses (Schuster and Linzen, 2022).

To avoid both these limitations, we propose **Divide, Generate, Recombine, and Compare (DGRC)**, which operationalizes the LM probability comparison method as follows (also see Figure 1):

1. We first **divide** a given U ("`S VP1 VP2`") into two independent utterances: U_{VP1} = "`S VP1`", U_{VP2} = "`S VP2`".
2. We then prompt the LM to **generate** n responses to each independent utterance: $R_{\text{VP}} = \{r_1^{\text{VP}}, \ldots, r_n^{\text{VP}}\}$, VP $\in \{1, 2\}$.
3. We then **recombine** the independent utterances into the original one.
4. Finally, we **compare** the (log) probabilities per token of all possible pairs of individual

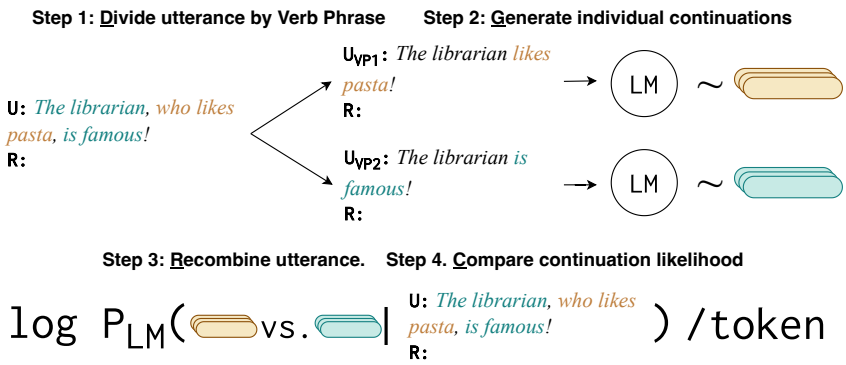

Figure 1: Visualization of the DGRC method, involving four steps: 1) Dividing the original utterance into sub-utterances; 2) Generating continuations for individual sub-utterances; 3) Recombining the sub-utterances into the original utterance, and 4) Comparing likelihoods for generated continuations. The end-result of this process allows us to characterize an LM's dialogue response dynamics.

responses to quantify the LM's preference.

$$\frac{1}{n^2}\sum_{i}^{n}\sum_{j}^{n}\mathbb{1}[s(r_i^{\text{VP2}} \mid U) > s(r_j^{\text{VP1}} \mid U)],$$

$$s(x \mid U) = \log p_{\text{LM}}(x \mid U)/|x|$$

For ARC constructions, we expect a model that is sensitive to at-issue distinctions to have a greater preference for VP2 (MC) over VP1 (ARC). By considering multiple different responses sampled *from* the LM, as opposed to researcher-defined, DGRC addresses both abovementioned limitations of forced-choice tasks. While DGRC is intended as a general framework for systematically evaluating naturalness across linguistic constructions, we use ARC as an initial test case in this paper.

Data, models, and implementation Stimuli from Kim et al. (2022) were used as a source of utterances. The dataset contained 300 English items, each consisting of a subject NP and two VPs forming an ARC construction (example in (2)). For LMs: we analyzed the instruct-tuned and base versions of Llama-3-8B (Grattafiori et al., 2024) and Qwen2.5 families (Yang et al., 2025), using four model sizes from 500M to 7B parameters, totaling 10 LMs. Responses were sampled using greedy decoding, and temperature, top-p, and top-k sampling. We considered several hyperparameters when applicable. We retained top-10 generations (per VP) for each utterance across all hyperparameter combinations (see DGRC Step 2), selecting those with the highest log-probabilities under the target LMs. This yielded 100 pairwise comparisons per item. For instruct-tuned models, stimuli were embedded in a chat-template format; for base models, they were embedded in a two-speaker conversational style. See Appendix for stimuli, hyperparameters, and prompts. Code and data are available at: https://github.com/sangheek16/hey-wait-a-minute.

4 Experiments

4.1 Experiment 1: At-issueness preference

In our first experiment, we analyze the extent to which LMs demonstrate sensitivity to at-issue content. We do so by performing DGRC on our dataset of ARC sentences and measuring at-issue preferences as described in §3. To place these results in context, we follow Kim et al. (2022), and compare these preferences to those obtained by performing DGRC on coordination structures (COORD), formed by conjoining the two VPs in our stimuli:

(3) The librarian [likes pasta]$_{\text{VP1}}$ and [is famous]$_{\text{VP2}}$.

This comparison is important since COORD structures differ from ARC minimally in terms of surface form (cf. (2)), and importantly include *both* VPs as part of the "main point" of the utterance—i.e., they lose the at-issue distinction. This allows us to control for recency bias: if LMs are simply showing sensitivity to the VP2 because it is more recent, then we should observe similar behavior in both ARC and COORD. On the other hand, if LMs are non-trivially sensitive to at-issueness, then their VP2 preference should be greater in ARC than in COORD. Outside of this comparison, we also include cases where we swapped the two VPs in our stimuli, to control for potential robustness issues.

Results Figure 2 shows our results. We found LMs' preference for VP2 to be greater in ARC structures than in COORD structures (see Appendix for

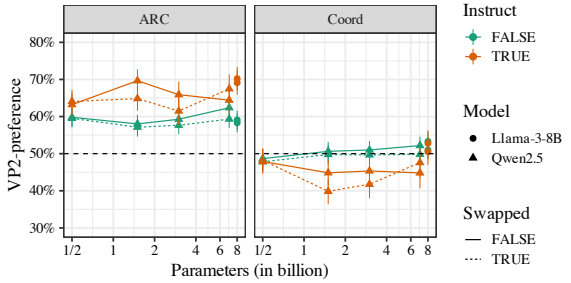

Figure 2: Experiment 1 results. VP2-preference of LMs (organized by parameter count) across ARC and COORD constructions, training mode (base/instruct), model family, and whether VP order was swapped.

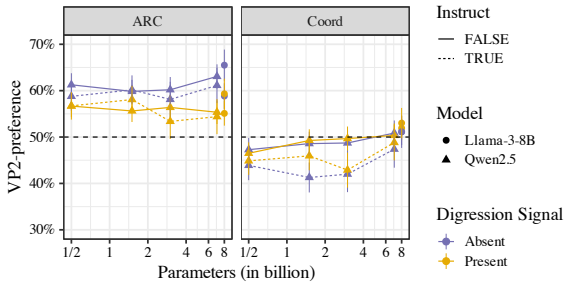

Figure 3: Experiment 2 results. VP2-preference of LMs (organized by parameter count) across ARC and COORD constructions, training mode (base/instruct), model family, and digression signal—absent ("No, that's not true"); present ("Hey, wait a minute").

full results). Additionally, instruct-tuned versions of an LM were consistently better in their at-issue preference than base models, suggesting a potential *pragmatically sensitive* behavior that seems to arise from instruction tuning. These findings were found to be statistically significant ($p < .01$ for both), shown using a linear mixed-effects model (LMEM) analysis. No notable effect of scale was found. Finally, LMs were generally robust to VP order, as shown by the lack of a swapping effect in the LMEM analysis ($p = 0.42$). We conclude that **LMs, especially instruct-tuned ones, show sensitivity to at-issue content in dialogues**.

4.2 Experiment 2: Sensitivity to digression

How do LMs' sensitivity to at-issue content vary in the presence of subtle cues shown to affect human at-issue sensitivities (Syrett and Koev, 2015)? For this, we turn to analyzing if LMs' at-issue sensitivity in ARC structures *decreases* in the presense of digression, as established in §2. To test this, we use digression-signaling cues such as "Hey, wait a minute" as response headers, following widely-used linguistic diagnoses (*peripherality test*; Koev, 2022; Amaral et al., 2007; Shanon, 1976). We compare LM preferences for at-issue content in responses beginning with "Hey, wait a minute" versus "No, that's not true!", the latter controlling for general *rejection*. That is, when rejecting part of the prior utterance, a direct rejection ("No") primarily targets at-issue content (Murray, 2014; Amaral et al., 2007; AnderBois et al., 2010), whereas "Hey, wait a minute" is expected to attenuate this, signaling rejection of the not-at-issue content (Syrett and Koev, 2015). Since both responses reject prior content, we include "No, that's not true!" header in Step 2 of DGRC in order for models to contradict the content of individual VPs. As before, we test both ARC and COORD structures.

Results Figure 3 shows our results. We found an interaction between digression signal (absent vs. present) and structure (ARC vs. COORD), with $p < .001$ (see Figure 7 for full results). In particular, LMs strongly preferred at-issue content (VP2) in the *absence* of digression than in its presence, but *only* in ARC structures and not in COORD, in line with our expectations. At the same time, they also showed an overall bias towards for VP2 content in the ARC structure (relative to COORD). This suggests that digression does not always *shift* preference to not-at-issue content, as it does in humans (e.g., Syrett and Koev, 2015). In short, **LMs prefer at-issueness, with the digression-signalling header modulating the effect only when a (not-)at-issue divide is present**. We did not find a particular benefit of instruct-tuning in ARC structures ($p = .39$), as both types of models showed a similar pattern of VP2-preference, i.e., response targeting at-issue content.

4.3 Post-hoc experiments: Nominal appositives

Since our approach relies on a specific linguistic construction to probe naturalness, it is important to assess whether the proposed method generalizes beyond appositive relative clauses (ARC).[1] We therefore consider a closely related construction: NOMINAL APPOSITIVES. Unlike relative clauses such as "the librarian, who likes pasta, is famous," nominal appositives employ an *elliptical* structure, as in "the librarian, a pasta lover, is famous." Nominal appositives, also belonging to the same family of parentheticals, are commonly considered not-

[1]We thank an anonymous reviewer for suggesting this construction.

at-issue (see Koev (2022) for a review). However, because LMs may be biased towards more frequent words such as copular verbs when parsing elliptical constructions (e.g., Kim et al., 2022; Testa et al., 2023), the absence of a verb in nominal appositives can make distinguishing (not-)at-issue a nontrivial test case for models.

Data, models, and implementation Nominal appositives were manually constructed from the VP1s by creating a semantically related nominal counterpart whenever applicable. For example, "used to play badminton" was mapped to "a retired badminton player." Models and implementation were identical to those used in Experiments 1 and 2, respectively, with two modifications. First, a copular verb—either *is* or *was*, depending on context—was added at Step 2 in DGRC when constructing the utterance U_{VP1}, for example:

(4) a. U: The painter, a retired badminton player, volunteers regularly at a local church.
b. U_{VP1}: The painter *is* a retired badminton player.

Second, the VP-swapping employed in Experiment 1 was omitted in this post-hoc experiment considering the nominal construction.

Results In the follow-up to Experiment 1, models showed a VP2 preference (i.e., response targeting the at-issue content) in the NOMINAL APPOSITIVE condition but not in the COORD condition, while instruction-tuned models again exhibited close to chance-level pattern in the COORD condition, qualitatively replicating Experiment 1. By contrast, the post-hoc experiment corresponding to Experiment 2 exhibited a qualitatively different result: both construction conditions patterned alike, with VP2 preferences at or above chance. This suggests that the interaction between digression cues and nominal appositives *can* affect model sensitivity to at-issue content. One plausible explanation for why the pattern diverges from Experiment 2 could be that, although nominal appositives and appositive relative clauses are syntactically similar (Dillon et al., 2018), nominal parentheticals may be less robustly not-at-issue (Nouwen, 2014, a.o). This could attenuate the (not-)at-issue distinction in nominal constructions, producing VP2 preferences closer to chance level. We leave more detailed investigation of the interaction between digression signals and nominal appositives to future work.

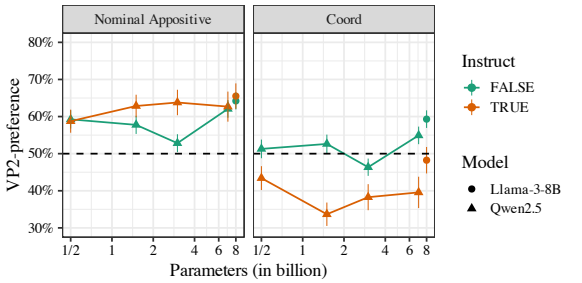

Figure 4: Post-hoc experiment results (Experiment 1). VP2-preference of LMs (organized by parameter count) across NOMINAL APPOSITIVE and COORD constructions, training mode (base/instruct), and model family.

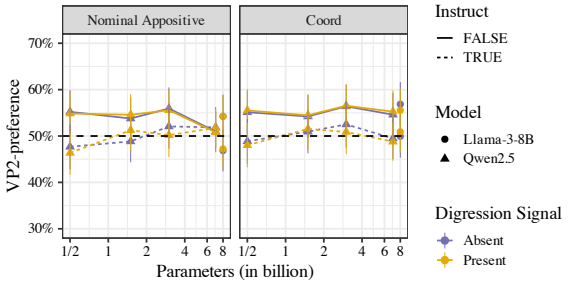

Figure 5: Post-hoc experiment results (Experiment 2). VP2-preference of LMs (organized by parameter count) across NOMINAL APPOSITIVE and COORD constructions, training mode (base/instruct), model family, and digression signal—absent ("No, that's not true"); present ("Hey, wait a minute").

5 Conclusion

We introduced DGRC, a linguistically motivated method for evaluating dialogue naturalness through the lens of at-issueness. Our results showed that LMs consistently prefer at-issue over not-at-issue content, with this tendency especially pronounced in instruction-tuned models (Experiment 1), suggesting that instruction-tuning can induce pragmatic sensitivity even without targeted training. Experiment 2 replicated models' the at-issue preference, while showing that this tendency is evident only in the context where (not-)at-issue divide is present, and that preference can be modulated when signaled with a digression header. These findings imply that enhancing sensitivity to at-issueness and dialogue dynamics could substantially improve the naturalness of model-generated dialogue. Importantly, DGRC enabled systematic, quantifiable evaluation without reliance on human evaluators or researcher-crafted prompts, complementing existing evaluation methods and allowing more fine-grained assessments of LMs in dialogues.

Limitations

Scope of construction The current paper examined a specialized construction, namely a structure involving an embedded ARC, followed by a post-hoc experiment that extended this to NOMINAL-APPOSITIVES. We acknowledge that these are highly specific linguistic forms, and our scope is therefore limited. As noted in §2, though, we highlight that the tested constructions offer a systematic and controlled way to compare at-issue and not-at-issue content without requiring multi-turn utterances, even when testing the intricate distinction. Moreover, findings based on these constructions highlight the importance of at-issueness in evaluating the naturalness of dialogue. Yet, the single-turn dialogue setup was primarily chosen to simplify stimulus design and to allow tighter experimental control (e.g., avoiding potential confounds due to variation in construction). In future work, we plan to explore this phenomenon more deeply in full, multi-turn dialogues using the DGRC method.

At-issueness of constructions We used coordination constructions (`VP1 and VP2`) as a baseline, assuming that they do not encode the (not-)at-issue distinction within a sentence. We note that subtle differences may still arise in the extent to which each VP contributes to discourse progression, given evidence that sentence-final or more recent discourse entities have greater potential to continue the discourse (e.g., Frazier and Clifton Jr., 2005; AnderBois et al., 2015; Jasinskaja, 2016; Hunter and Asher, 2016). Additionally, while the ARC construction was used to create short dialogue contexts embedding an at-issue vs. not-at-issue distinction, linguistic theories suggest that this distinction can weaken depending on the ARC's sentential position (AnderBois et al., 2015). A sentence-final ARC may behave more "at-issue-like" as it stands closer to the following discourse content. We did not test this configuration, and exploring this remains as future work, which will offer an even finer-grained evaluation method on dialogue dynamics.

Understanding instruction-tuning Instruction-tuned models are often considered better at capturing conversational goals (e.g., Zhang et al., 2023). We also found such patterns in Experiment 1, where instruction-tuned models aligned more with the expected behavior. But this was not the case for Experiment 2. Recent studies suggest that instruction tuning does not necessarily yield stronger model–human alignment, when evaluated against human judgments and behavioral data (Zhang et al., 2023; Kauf et al., 2024; Aw et al., 2024; Kim and Davis, 2025). Given these mixed findings, a principled analysis of how instruction-tuning affects human–machine dialogue alignment would be especially informative for future direction.

Acknowledgments

ChatGPT was used for paraphrasing and spell-checking at the proofreading stage. Kanishka Misra was supported by the Donald D. Harrington Faculty Fellowship at UT Austin.

References

Patricia Amaral, Craige Roberts, and E Allyn Smith. 2007. Review of the logic of conventional implicatures by Chris Potts. *Linguistics and Philosophy*, 30(6):707–749.

Scott AnderBois, Adrian Brasoveanu, and Robert Henderson. 2010. Crossing the appositive/at-issue meaning boundary. In *Semantics and Linguistic Theory*, volume 20, pages 328–346.

Scott AnderBois, Adrian Brasoveanu, and Robert Henderson. 2015. At-issue proposals and appositive impositions in discourse. *Journal of Semantics*, 32(1):93–138.

Khai Loong Aw, Syrielle Montariol, Badr AlKhamissi, Martin Schrimpf, and Antoine Bosselut. 2024. Instruction-tuning aligns LLMs to the human brain. In *First Conference on Language Modeling*.

Douglas Bates, Martin Mächler, Ben Bolker, and Steve Walker. 2015. Fitting linear mixed-effects models using lme4. *Journal of statistical software*, 67:1–48.

Brian Dillon, Lyn Frazier, and Charles Clifton Jr. 2018. No longer an orphan: Evidence for appositive attachment from sentence comprehension. *Glossa*, 3(1).

Amila Ferron, Amber Shore, Ekata Mitra, and Ameeta Agrawal. 2023. Meep: Is this engaging? prompting large language models for dialogue evaluation in multilingual settings. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 2078–2100.

Lyn Frazier and Charles Clifton Jr. 2005. The syntax-discourse divide: Processing ellipsis. *Syntax*, 8(2):121–174.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Shengyue Guan, Haoyi Xiong, Jindong Wang, Jiang Bian, Bin Zhu, and Jian-guang Lou. 2025. Evaluating llm-based agents for multi-turn conversations: A survey. *arXiv preprint arXiv:2503.22458*.

Jennifer Hu, Roger Levy, and Sebastian Schuster. 2022. Predicting scalar diversity with context-driven uncertainty over alternatives. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 68–74, Dublin, Ireland. Association for Computational Linguistics.

Julie Hunter and Nicholas Asher. 2016. Shapes of conversation and at-issue content. In *Semantics and Linguistic Theory*, volume 26, pages 1022–1042.

Katja Jasinskaja. 2016. Not at issue any more. Unpublished manuscript, University of Cologne.

Carina Kauf, Emmanuele Chersoni, Alessandro Lenci, Evelina Fedorenko, and Anna A Ivanova. 2024. Comparing plausibility estimates in base and instruction-tuned large language models. *arXiv preprint arXiv:2403.14859*.

Sanghee J. Kim and Forrest Davis. 2025. Discourse sensitivity in attraction effects: The interplay between language model size and training data. *Society for Computation in Linguistics*, 8(1).

Sanghee J. Kim, Lang Yu, and Allyson Ettinger. 2022. "no, they did not": Dialogue response dynamics in pre-trained language models. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 863–874, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Wei-Jen Ko, Yating Wu, Cutter Dalton, Dananjay Srinivas, Greg Durrett, and Junyi Jessy Li. 2022. Discourse analysis via questions and answers: Parsing dependency structures of questions under discussion. *arXiv preprint arXiv:2210.05905*.

Todor Koev. 2022. *Parenthetical meaning*. Oxford Studies in Semantics and Pragmatics. Oxford: Oxford University Press.

Alexandra Kuznetsova, Per B Brockhoff, and Rune HB Christensen. 2017. lmertest package: tests in linear mixed effects models. *Journal of statistical software*, 82:1–26.

Yen-Ting Lin and Yun-Nung Chen. 2023. Llm-eval: Unified multi-dimensional automatic evaluation for open-domain conversations with large language models. *arXiv preprint arXiv:2305.13711*.

Rebecca Marvin and Tal Linzen. 2018. Targeted syntactic evaluation of language models. *arXiv preprint arXiv:1808.09031*.

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

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

Sarah E Murray. 2014. Varieties of update. *Semantics and Pragmatics*, 7(2):1–53.

Rick Nouwen. 2014. A note on the projection of appositives. In Eric McCready, Kyoko Yabushita, and Koji Yoshimoto, editors, *Formal Approaches to Semantics and Pragmatics: Japanese and Beyond*, pages 205–222. Springer, Dordrecht, Netherlands.

Christopher Potts. 2005. *The logic of conventional implicatures*. Oxford: Oxford University Press.

Sujan Reddy. 2022. Automating human evaluation of dialogue systems. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Student Research Workshop*, pages 229–234.

Arndt Riester. 2019. Constructing QUD trees. In Malte Zimmermann, Klaus von Heusinger, and Edgar Onea, editors, *Questions in discourse*, volume 2, pages 163–192. Leiden: Brill.

Sebastian Schuster and Tal Linzen. 2022. When a sentence does not introduce a discourse entity, transformer-based models still sometimes refer to it. *arXiv preprint arXiv:2205.03472*.

Benny Shanon. 1976. On the two kinds of presuppositions in natural language. *Foundations of Language*, 14(2):247–249.

Mandy Simons, Judith Tonhauser, David Beaver, and Craige Roberts. 2010. What projects and why. In *Semantics and Linguistic Theory*, volume 20, pages 309–327.

Kristen Syrett and Todor Koev. 2015. Experimental evidence for the truth conditional contribution and shifting information status of appositives. *Journal of Semantics*, 32(3):525–577.

Davide Testa, Emmanuele Chersoni, and Alessandro Lenci. 2023. We understand elliptical sentences, and language models should too: A new dataset for studying ellipsis and its interaction with thematic fit. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3340–3353, Toronto, Canada. Association for Computational Linguistics.

Judith Tonhauser. 2012. Diagnosing (not-)at-issue content. In *Proceedings of the Sixth Conference on the Semantics of Under-Represented Languages in the Americas (SULA)*, volume 6, pages 239–254. Amherst, MA: Graduate Linguistics Student Association.

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.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025. Qwen2.5 technical report. *arXiv preprint arXiv:2505.09388*.

Shengyu Zhang, Linfeng Dong, Xiaoya Li, Sen Zhang, Xiaofei Sun, Shuhe Wang, Jiwei Li, Runyi Hu, Tianwei Zhang, Fei Wu, and 1 others. 2023. Instruction tuning for large language models: A survey. *arXiv preprint arXiv:2308.10792*.

Zhehao Zhang, Ryan A. Rossi, Branislav Kveton, Yijia Shao, Diyi Yang, Hamed Zamani, Franck Dernoncourt, Joe Barrow, Tong Yu, Sungchul Kim, Ruiyi Zhang, Jiuxiang Gu, Tyler Derr, Hongjie Chen, Junda Wu, Xiang Chen, Zichao Wang, Subrata Mitra, Nedim Lipka, and 2 others. 2025. Personalization of large language models: A survey. *arXiv preprint arXiv:2411.00027*. Version 3.

A Appendix

Data and implementation

Stimuli A sample set of data used for DGRC is shown in Table 1. These were taken from Kim et al. (2022), which was released using the MIT License, as specified on their github.[2]

Models and hyperparameters We evaluated instruction-tuned and base models from the Llama-3-8B and Qwen2.5 family:

- Instruction-tuned models: Meta-Llama-3-8B-Instruct, Qwen2.5-0.5B-Instruct, Qwen2.5-1.5B-Instruct, Qwen2.5-3B-Instruct, and Qwen2.5-7B-Instruct
- Base models: Meta-Llama-3-8B, Qwen2.5-0.5B, Qwen2.5-1.5B, Qwen2.5-3B, and Qwen2.5-7B

For each model, generations were produced using combinations of the following hyperparameters: top-p where $p \in \{0, 0.9, 0.95\}$, top-k values $k \in \{50, 0\}$, and temperatures $t \in \{0.7, 1.0\}$. The top-10 generations out of all generations were selected based on the log probabilities of the utterance–response sequence. All log probabilities are computed using minicons (Misra, 2022). Models were run on the `burrata` server, containing a single NVIDIA RTX6000 Ada GPU, with 48GB RAM.

[2] https://github.com/sangheek16/dialogue-response-dynamics

Implementation For instruction-tuned models, we used the model's chat template, which takes a sequence of roles: `system`, `user`, and, for Experiment 2, additionally `assistant`. The `system` message provided a short instruction, as shown below. The divided utterance was placed in the `user` content. In Experiment 2, headers were inserted as content of `assistant` (e.g., "No, that's not true!" or "Hey, wait a minute!"). The following is an example:

> [system] "Please respond to the following message as naturally as possible, using a single sentence, as if we were talking to each other." Please keep it short.
> [user] "The librarian likes pasta."
> [assistant] "No, that's not true!"

For base models, which do not support chat formatting, we instead included the utterance and response as a two-person dialogue using direct quotation, similar to Kim et al. (2022): `$name1` said, "`$stimulus`," and `$name2` replied, ("`$header`"). The header (in parentheses) was included only in Experiment 2. Placeholders for `$name1` and `$name2` were replaced with proper names (e.g., Marco, Ellie), randomly sampled from a list of 400 names. An example is shown below:

> Marco said, "The librarian, who likes pasta, is famous," and Ellie replied, ("No, that's not true!")

Linear Mixed-Effects Modeling Results

We analyzed results from our experiments using linear mixed-effects models using the `lme4` (Bates et al., 2015) and `lmerTest` (Kuznetsova et al., 2017) packages in R. In what follows we describe the model formula and show interaction plots (when applicable).

Experiment 1 We used the following formula to analyze our results in this experiment:

$$\begin{aligned}&\texttt{vp2\_pref} \sim \texttt{swapped} + \texttt{instruct} \times \texttt{structure} \\ &+ (1 + \texttt{swapped} + \texttt{instruct} \times \texttt{structure} \mid \texttt{item}) \\ &+ (1 \mid \texttt{model}),\end{aligned}$$

where `swapped` indicates whether or not the VPs were swapped, `instruct` indicates the presence/absence of instruction-tuning, `structure` indicates the type of structure (ARC vs. COORD), `model` is the model, and `item` is the individual

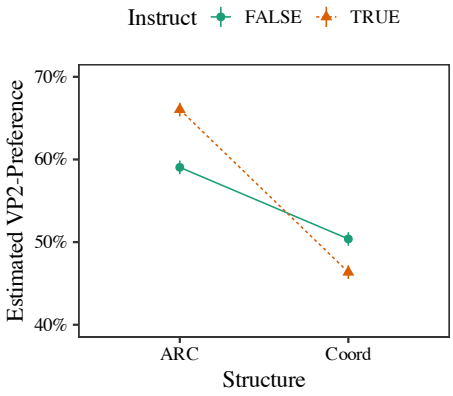

Figure 6: Effect of instruction-tuning on results in Experiment 1. VP2-preference of LMs across ARC and COORD structures, modulated by instruction-tuned vs. base model.

item (which determines what lexical items occur in the structure—i.e., the lexical content of NP, VP1, VP2). Our results are shown in Table 2.

The interaction between instruction-tuning and structure—found to be significant—is shown in Figure 6. The VP2-preference, i.e., targeting at-issue content, is salient in ARC constructions, particularly with instruction-tuned models.

Experiment 2 We used the following formula to analyze our results in this experiment:

$$\begin{aligned}&\texttt{vp2\_pref} \sim \texttt{header} \times \texttt{instruct} \times \texttt{structure} \\ &+ (1 + \texttt{header} \times \texttt{instruct} \times \texttt{structure} \mid \texttt{item}) \\ &+ (1 + \texttt{header} \times \texttt{instruct} \times \texttt{structure} \mid \texttt{model}),\end{aligned}$$

where `header` indicates the response header of the model (1 if digression is present—i.e., "Hey, wait a minute!", and 0 if it is absent—i.e., "No, that's not true!"), `instruct` indicates the presence/absence of instruction-tuning, `structure` indicates the type of structure (ARC vs. COORD), `model` is the model, and `item` is the individual item (which determines what lexical items occur in the structure—i.e., the lexical content of NP, VP1, VP2). Our results are shown in Table 3.

We found a significant interaction between digression signal and structure in Experiment 2, as shown in Figure 7. A digression header signaling that the response would target not-at-issue content led to a reduced preference for VP2, i.e., at-issue content. This effect disappeared in the COORD structure, suggesting that the influence of a digression signal is evident only when a division between at-issue and not-at-issue content is present, namely, in the ARC structure.

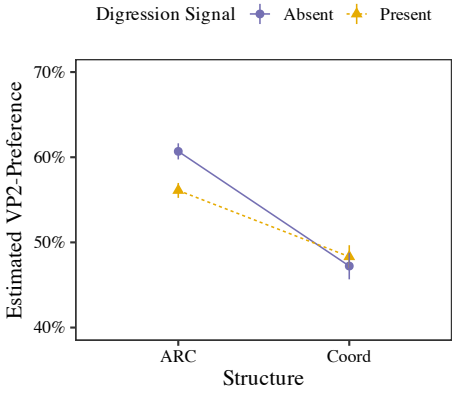

Figure 7: Effect of the digression signal in Experiment 2. VP2-preference of LMs by digression signal, modulated by ARC and COORD structures.

Subject NP	VP1	VP2
The nurse	met the Illinois governor at a Greek restaurant	looks confident
The athlete	went to the post office	has been checking the clock for five hours
The programmer	used to drink three cups of coffee every day	is about to fall asleep
The anthropologist	went out for a date	was solving a crossword puzzle
The nanny	seemed very eager to return home	would wear glasses in the day time
The overseer	reads eight books a month	hired a teenager to mow the lawn
The nun	rides a bike to the nearest park	has been in town for 10 years
The senator	finds humor in the worst situations	is not good at riding a bike
The waitress	wears a fancy watch	was upset about the cleanliness of City Hall
The make-up artist	occasionally writes a blog about historical figures	would drive to the nearest library when the weather is bad
The model	has been dreaming of buying a private jet	went to the post office
The receptionist	has been earning money by making YouTube videos	has lots of friends
The producer	has been around for a while	is standing next to the tree
The detective	has been chosen to audition for the All Stars game	was angry
The inspector	has been checking the clock for five hours	would wake up early on Christmas Eve
The fisherman	is good at communication	got stung by a wasp the other day
The diver	is reading a newly released sci-fi book	takes a vitamin every day
The anthropologist	is about to fall asleep	has been collecting fridge magnets for five years
The psychologist	is good at board games	was singing a song
The pilot	is subscribed to seventy different newsletters	would read books at a park nearby
The tenant	was trying to fix the broken window	seemed very eager to return home
The farmer	was disappointed with the weather	occasionally takes a nap after lunch
The overseer	was happy about the cross-country road trip	likes bungee jumping
The linguist	was disappointed with the weather	has been staying in Hawaii for two months
The bartender	was mentioned in the newspaper	is extremely fickle and demanding
The hairdresser	was unhappy about all the noise on the streets	would make pasta for dinner
The painter	would drive to the nearest library when the weather is bad	suggested a good cleaning company to Ashley
The musician	would wear a yellow hat on sunny days	enjoys hiking
The lecturer	would have salad and boiled eggs for lunch	has been a fan of Rihanna since her debut
The administrator	would go to the movies every week	is never late
The writer	would swim in the Lake on Monday mornings	was being chased by a few people

Table 1: Examples of subject NPs with VP1 and VP2 used for generating experimental stimuli.

Term	Estimate	Std. Error	df	t	p
(Intercept)	0.56	0.01	16.08	54.06	$< .001$
swapped	-0.01	0.01	298.92	-0.73	0.46
instruct	0.01	0.00	11684.73	3.45	$< .001$
mode	0.14	0.00	9214.68	32.84	$< .001$
instruct:mode	0.11	0.01	11117.55	12.78	$< .001$

Table 2: Linear Mixed Effects Model summary for Experiment 1.

Term	Estimate	Std. Error	df	t	p
(Intercept)	0.53	0.01	12.99	49.81	$< .001$
header	0.02	0.01	5.18	3.14	0.02
instruct	-0.01	0.02	13.03	-0.88	0.40
structure	0.11	0.01	4.10	11.21	$< .001$
header:instruct	-0.01	0.01	11385.99	-1.38	0.17
header:structure	0.06	0.01	11385.99	7.17	$< .001$
instruct:structure	0.04	0.01	11385.99	4.57	$< .001$
header:instruct:structure	0.01	0.02	11385.99	0.80	0.42

Table 3: Linear Mixed Effects Model summary for Experiment 2.