Judging the judge: Can LLMs track attitude-holders in predicates of personal taste?

Yan Cong
School of Languages and Cultures, Purdue University
cong4@purdue.edu

Abstract

When LLMs annotate subjective content such as *predicates of personal taste* (PPTs), e.g., "Eliza tasted the cookie. It tasted delicious", a consequential but underexplored question is whose perspective they adopt: the narrator's (default) or Eliza's (shift toward the character)? Humans tend to attribute *delicious* to Eliza because tasting implies first-hand experience. We introduce GRANI (Graded Response Analysis for Narrative Implications), a protocol for evaluating LLMs' graded understanding of such phenomenon. Applying GRANI to PPTs, we test whether text-trained LLMs reproduce human perspective shifts across sensory modalities in English and Chinese. We found that instruction-tuned models partially capture the human pattern, in which taste yields the strongest shift toward the character, vision the weakest, and smell an intermediate effect, while reasoning-distilled models are less consistent. Moreover, some models encode modality distinctions internally even when their behavioral responses do not. These findings suggest that subjectivity inference can emerge from text exposure alone, while highlighting a reliability concern: LLMs may not consistently adopt the intended perspective-holder when evaluating subjective content.[1]

1 Introduction

Predicates of personal taste (PPTs) (Lasersohn, 2005; Kaiser, 2021), such as *tasty* or *disgusting*, are judge-dependent expressions whose interpretation requires identifying a specific experiencer with first-hand experience. In a narrative with a first-person narrator and a *character* (the individual in the story who perceives the object directly), inferring the intended *attitude-holder* (the individual whose perspective, attitude, or opinion is reflected by the subjective adjective) is a pragmatically subtle task that depends on who had the direct

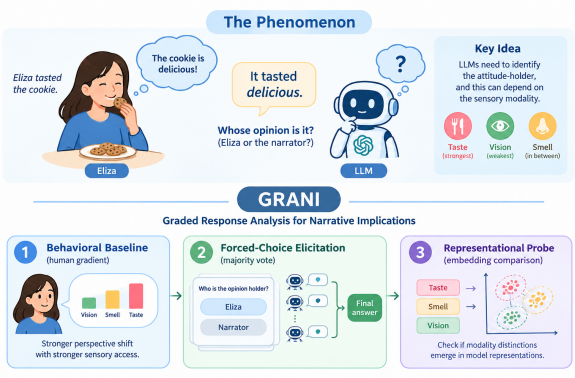


Figure 1: GRANI overview: Sensory modality triggers a graded perspective-shift in the inferred attitude-holder (narrator → character "Eliza") for PPTs like "delicious", tested via behavioral and representational methods.

sensory experience. Consider a non-PPT baseline in (1), when asked about whose opinion it is that the cookie tasted delicious, humans infer that it is by default the narrator's opinion. In contrast, for a PPT example in (2), there is a perspective-shift inference that it is Eliza's opinion, since only someone who tasted the cookie can judge its taste (Kaiser, 2021).

1. Eliza put the cookie on the table. It was delicious.
2. Eliza tasted the cookie on the table. It tasted delicious.

As large language models (LLMs) are increasingly deployed as annotators of subjective content (Gilardi et al., 2023; Movva et al., 2024), a critical but underexplored reliability question arises: when a model evaluates a subjective expression, whose perspective does it adopt? This matters because systematic perspective biases could silently skew LLM-based annotation pipelines for tasks like sentiment analysis and opinion mining, where annotators must decide between narrator and character perspectives. As shown in Fig. 1, we propose an evaluation methodology GRANI (**G**raded **R**esponse **A**nalysis for **N**arrative

[1]Data and code are available for research at https://github.com/yancong222/aacl-2026-tasty.

Findings of the Association for Computational Linguistics: IJCNLP-AACL 2026, pages 342–349
November 6-10, 2026 ©2026 Association for Computational Linguistics

Implications, Fig. 1) to evaluate whether LLMs can track attitude-holders in PPT as exemplified in (2).

GRANI is inspired by a theoretically motivated *gradient* prediction. Kaiser (2021) showed that in humans, perspective-shift toward a first-hand experiencer is modulated by sensory modality, with taste triggering stronger shifts than vision and smell in between, a pattern predicted by how constrained perceptual access is for each modality. Capturing this gradient requires structured sensitivity to narrative implications, which makes it a stronger evaluation target than binary, single-condition accuracy. Existing NLP work on perspective-taking focuses on objective belief states (Ma et al., 2023; Wilf et al., 2024), and computational work on perceptual subjectivity has not been grounded in psycholinguistic theory (Giorgi et al., 2024; Subbiah et al., 2025). We attempt to bridge the gap by using GRANI to evaluate LLMs in three experiments and two languages, supplemented by representational analyses probing whether text-trained LLMs could reproduce the modality-graded human pattern, despite lacking any direct sensory experience.

Our work has three major contributions. First, a theory-driven evaluation protocol integrating both a prompting procedure and an embedding comparison. Second, a 400-item bilingual stimulus set of minimally different narratives. Third, the empirical findings that we identify behavioral correlates in certain models and tasks, emphasizing on the need for more work on how human-like subjectivity inference is related to text-only exposure and embodied grounding relate, with implications for sentiment analysis, and narrative and dialogue understanding.

2 The GRANI protocol

GRANI has three components: (a) a graded human behavioral signature as the evaluation baseline, (b) forced-choice elicitation with majority-vote aggregation over stochastic samples, and (c) a representational probe testing whether behavioral patterns are encoded internally in embeddings. We instantiate this with PPTs, which allow *faultless disagreement* (Kölbel, 2004) that two people can hold contradictory PPT judgments without either being wrong. Identifying the judge is thus essential yet must be inferred pragmatically. A core constraint is the *acquaintance inference* (Ninan, 2014; Pearson, 2013), maintaining that opinion-holder of PPTs has *acquaintance*, namely, first-hand experiential access. In narratives with a narrator and a character, the character uniquely marked as experiencer is inferred to be the opinion-holder, with inference strength modulated by perceptual access constraints across sensory modalities. There is a lack of evidence that LLMs reproduce such fine-grained perceptual subjectivity, even with persona prompting (Giorgi et al., 2024; Subbiah et al., 2025), and perspective-taking prompts inspired by Theory of Mind only partially improve false-belief performance (Ma et al., 2023; Wilf et al., 2024).

	Question	Prediction
SEH	Does sensory experience matter at all in perspective shift?	Any sensory condition $>$ Baseline.
SbH	How does subjectivity modulate such shift?	Smell/Taste $>$ Vision.
IbH	How does such shift effect track perceptual access?	Taste $\geq$ Smell $>$ Vision $>$ Baseline.

Table 1: Three increasingly specific hypotheses about sensory-induced perspective shifts (Kaiser, 2021). Arrows indicate rank of perspective-shifts over baseline.

Kaiser (2021) proposes three nested predictions about opinion-holder identification, capturing shift from a first-person narrator to an experiencer character, as summarized in Table 1: the primary hypothesis (i) **SEH** (Sensory Experience Hypothesis): any sensory specification increases shifts over baseline where no perceptual experience is described; two sub-hypotheses (ii) **SbH** (Subjectivity-based Hypothesis): more subjective modalities trigger stronger shifts, grouping smell and taste above vision; (iii) **IbH** (Inference-based Hypothesis): shifts track perceptual access constraints, taste (physical contact) $>$ smell (proximity) $>$ vision (distal), yielding baseline $<$ vision $<$ smell $\ll$ taste. Kaiser (2021) found human results generally consistent with all three hypotheses.

We translate and test these predictions in LLMs as a structured diagnostic. LLMs must demonstrate sensitivity to the gradient across baseline, taste, vision, and smell conditions. We instantiate GRANI on one phenomenon. In principle, the same recipe (human graded signature, forced-choice elicitation, representational check) could be applied to other phenomena with graded human patterns, such as evidentiality (hearsay and direct evidence) or aspect (ongoing and completed events), but we do not test this here. Also, our crosslinguistic datasets provide resource for future work to evaluate new models,

prompting strategies, or fine-tuning approaches.

3 Operation of the GRANI protocol

Stimuli The first component of GRANI is the graded behavioral baseline. We extended the original 96-item Kaiser (2021) stimulus set (24 sets) to 400 items (100 sets) using GPT-5 generation (accessed March–April 2026; prompts in Appendix A), validated by a linguist for naturalness, length, and balanced difficulty. As illustrated in Table 2, items vary only lexically, while preserving syntactic structure, adjective felicity, and the four conditions (vision/smell/taste/baseline vary the verbs). Chinese stimuli were created by expert translation (Wang and Sun, 2015; Huang and Shi, 2016; Cong, 2021) with culturally appropriate substitution of character names, food items, and location phrases, validated by a Chinese linguist. Chinese additionally includes a *variant* subcondition using sense-specific adjectives (e.g., 难吃 *nán chī* "bad-tasting", 难闻 *nán wén* "bad-smelling") alongside the general subcondition with modality-neutral adjectives (e.g., 恶心 *ě xīn* "disgusting"), as used in Kaiser (2021)'s English experiments.

Cond.	Item
Vision	Right after I arrived at the party, Daniel **saw** the salad on the table. It **looked** (truly) amazing.
Smell	Right after I arrived at the party, Daniel **smelled** the salad on the table. It **smelled** (truly) amazing.
Taste	Right after I arrived at the party, Daniel **tasted** the salad on the table. It **tasted** (truly) amazing.
Baseline	Right after I arrived at the party, Daniel **put** the salad on the table. It **was** (truly) amazing.
Question:	Whose opinion is it that the salad [**looked/smelled/tasted/was**] amazing?

Table 2: Example stimuli with shared starter: *The following excerpt is from a novel with a first-person narrator "I" and a "character"*. Here, the character is *Daniel*.

Forced-choice elicitation The second GRANI component operationalizes graded behavior measurement as elicitation task. Based on Kaiser (2021), we designed three LLM prompting experiments (Table 3). We use a one-trial-per-run procedure (Qiu et al., 2025) to minimize carry-over biases. Each item is presented as a multiple-choice question with answer labels randomized across items to reduce position bias. We sample $n = 10$ responses per item (temperature 0.6, top-k 50), extract answer labels via regex, apply majority voting, and record the *winner* and *confidence* (proportion of votes for the winner). We evaluate Qwen2-7B (base) and Qwen2-7B-Instruct (Team, 2024), Llama-3-8B-Instruct (Grattafiori et al., 2024), and DeepSeek R1-Qwen-7B and R1-Llama-8B (Guo et al., 2025). Llama-3-8B (base) failed to produce parseable outputs and is only included in representational analyses and gets excluded in the behavioral analyses. Catch trials (200 unambiguous comprehension items per language) verified basic reliability; only models exceeding 70% accuracy were included (all except Llama-3-8B-Instruct: 49.7% English, 56.9% Chinese).

The following is an excerpt from a novel. It has a first-person narrator "I" and a "character".
Passage: When I came into the room, Eliza tasted the muffin on the platter. It tasted **disgusting**.
Question:
Whose opinion is it that the muffin tasted disgusting?
Choose one without explaining:
A The narrator's opinion
B Eliza's opinion
Only output A or B.

Table 3: Exp. 1 prompt. Exp. 2: Exp. 1 + intensifier (e.g., *totally*) before PPT; Exp. 3: Exp. 1 + a third option: *Both the narrator and Eliza have this opinion*.

Representational probe The third GRANI component concerns whether behavioral patterns are reflected in embeddings. For each item, we extract the contextualized embedding of the PPT adjective like "amazing" (last-layer hidden state, mean-pooled over subword tokens; inspired by Yu and Ettinger (2020)). Pairwise cosine similarities between baseline and each sensory-condition embedding are computed. One-sided paired Wilcoxon signed-rank tests assess whether taste similarity is lowest (IbH); two-sided tests assess whether taste and smell are equivalent (SbH). Analyses follow Kaiser (2021): GEE logistic regression with condition as the fixed effect and model as the clustering unit. From-chance tests use intercept-only logistic regression (H_0: proportion = 0.5). Pairwise contrasts are Bonferroni-corrected.

4 Results

4.1 Prompt Exp. 1: 2-choice forced response

English Table 4 summarizes Exp. 1 results. Qwen2-7B-Instruct shows the most human-like sensitivity to acquaintance inference. The baseline condition strongly favors the narrator, while all sensory conditions significantly exceed chance. Critically, pairwise comparisons confirm taste > vision

and taste > smell but vision and smell do not differ, a pattern consistent with the SEH and partially consistent with IbH (taste distinguished, smell not separated from vision). Confidence in the dominant choice exceeds 90% for taste. Qwen2-Base shows a weaker but directionally similar pattern for smell and taste above chance, without significant condition contrasts, suggesting that within Qwen2 family, instruction-tuning is associated with better sensitivity. R1-Qwen-7B shows a *reversed* vision effect. The vision condition elicits *fewer* character responses than baseline, inconsistent with all three hypotheses, suggesting that compared with the same-scale model Qwen2-7B-Instruct, R1-Qwen-7B's reasoning brought no additional performance gains. R1-Llama-8B shows that vision above baseline, taste *below* chance, further indicating that reasoning distillation does not yield results consistent with the hypotheses. Fig. 2 in the appendix illustrates these behaviors with response distribution. Notably, item-clustered results suggest that by-item variability was not a confound across experiments.

Model	Baseline	Vision	Smell	Taste
English				
Qwen2-Instruct	$.01^{***}$	$.67^{**}$	$.64^{**}$	$.92^{***}$
Qwen2-Base	.52	.49	$.65^{**}$	$.61^{*}$
R1-Qwen-7B	.45	$.24^{***}$	$.39^{*}$	.47
R1-Llama-8B	$.27^{***}$	.53	$.40^{*}$	$.34^{**}$
Chinese				
Qwen2-Instruct	$.04^{**}$	$.92^{**}$	$.91^{**}$	$.88^{**}$
Qwen2-Base	$.27^{*}$	$.91^{**}$	$.74^{*}$	$.91^{**}$
R1-Qwen-7B	$.83^{**}$	$.95^{**}$	$.87^{**}$	$.91^{**}$
R1-Llama-8B	$.21^{*}$	.42	.50	$.06^{**}$

Table 4: Character-opinion responses in Exp. 1. Significance: $^{*}p < .05$, $^{**}p < .01$, $^{***}p < .001$.

Chinese Table 3 showed that Qwen2-Instruct remains strongly character-biased across all sensory conditions in Chinese, with baseline slightly suppressed. There is no significance in taste > smell distinction (all sensory conditions cluster at ≈88-92%). Qwen2-Base shows a strong character bias for vision and taste but a weaker smell response, approaching the IbH ordering. R1-Qwen-7B shows strong character responses in all sensory conditions, suggesting the Chinese context shifts this model toward a character preference regardless of conditions. R1-Llama-8B again shows findings inconsistent with hypotheses.

4.2 Prompt Exp. 2: 2-choice with intensifier

Adding a speaker-oriented intensifier to the PPT was predicted by IbH to reduce the smell response (by making the narrator more available as an experiencer) while leaving taste unchanged. For English, Qwen2-Instruct reproduces this: taste remains high (≈94%), and smell drops slightly but not significantly relative to Exp. 1. Crucially, the taste-smell Bonferroni contrast reaches significance ($p < .001$), providing the strongest evidence for IbH in any model. This is the same pattern Kaiser (2021) observed in human participants in Exp. 2. For the R1 and Llama models, cross-experiment comparisons show no significant differences in any condition, suggesting the intensifier manipulation does not modulate their perspective-taking. For Chinese, most models show stable or increased character responses across conditions with no significant Exp. 1 vs. Exp. 2 differences. Fig. 3 in the appendix illustrates these behaviors.

4.3 Prompt Exp. 3: 3-choice (narrator / character / both)

Introducing a *both* option allows models to avoid forced binary commitment, testing their sensitivity to faultless disagreement in PPT (Kölbel, 2004). Qwen2-Instruct never selects *both* (0% across all conditions) and does not tolerate faultless disagreement, whereas human participants in Kaiser (2021) used *both* frequently for vision and smell. This likely reveals a lack of gradient uncertainty and an overconfidence pattern in the instruction-tuned model (Fig. 5, 6 in the appendix illustrates dominant-choice confidence, derived from the full 10-sample distribution per item and model). Qwen2-Base and R1-Qwen-7B distribute responses more broadly, suggesting faultless disagreement, with *both* appearing in vision and smell conditions for some items, more closely paralleling the human Exp. 3 distribution. For Qwen2-Instruct in English, taste responses remain >94% character, and baseline >92% narrator, same as Exp. 1. Fig. 4 in the appendix illustrates these.

4.4 Representational analysis

For English, taste embeddings are more *similar* to baseline embeddings than smell embeddings across all models, which is the opposite of the IbH prediction that taste should be most *dissimilar* from baseline. This finding holds for all the models tested, indicating that at the representational level, LLMs do not encode taste as categorically distinct from baseline in the way that would be expected if they had effectively represented the acquaintance inference as a latent variable. Chinese showed a

different picture. Qwen2-Instruct and Qwen2-Base show taste embeddings *less* similar to baseline than vision or smell embeddings ($p < .001$ for both Wilcoxon tests), directly supporting IbH at the representational level. Notably, this pattern disappears in the Chinese *variant* (sense-specific adjective) condition, where smell and taste embeddings become statistically indistinguishable from each other. This is consistent with the SbH prediction. The variant condition thus provides partial support for the SbH specifically in Chinese when lexical items encode sensory specificity directly.

5 Discussion

Our GRANI-based evaluation reveals four major findings with implications beyond the PPT domain. First, we observe a cross-linguistic asymmetry resonating with Leidinger et al. (2023); Jin et al. (2024). Such observed asymmetries between English and Chinese might be associated with factors such as lexicalization differences and verb semantics. Chinese prompts elicit stronger character biases, with Qwen2 sensory conditions clustering at high character-response rates and compressing the English ordering. Chinese smell verbs (闻 *wén*, 嗅 *xiù*) may involve less agentive ambiguity than English *smell*, while sense-specific adjectives in the variant condition sharpen representational distinctions, yielding a SbH-consistent pattern.

Second, GRANI's representational probe reveals a behavioral-representational dissociation. Qwen2-Instruct's responses are condition-sensitive, yet its English PPT embeddings do not show the taste-baseline divergence predicted by IbH, suggesting the acquaintance inference operates at a later processing stage rather than being encoded in contextual representations (Huang et al., 2025). This dissociation check is worth recasting when applying GRANI to new models or phenomena.

Third, instruction tuning (not reasoning distillation) is associated with acquaintance inference, resonating with Ouyang et al. (2022). We speculate that this may reflect instruction tuning model's stronger reliance on learned patterns of speaker/experiencer attribution rather than deliberate multi-step reasoning. Qwen2-7B-Instruct is the only model consistently reproducing the human gradient across both English experiments, including the critical Exp. 2 taste-smell dissociation distinguishing IbH from SbH. Because GRANI requires reproducing the full gradient rather than above chance performance in any single condition, these architectural differences become detectable in ways flat accuracy metrics might not capture, with implications for subtle implicit meaning interpretation, annotation reliability, and model selection. Practically, GRANI sheds light on how annotation pipelines could be better designed, for example, making perspective explicit in prompts or data, and identifying and leveraging verbs (*taste/smell/look*) for perspective attribution in prompts or data curation.

Fourth, advancing the benchmarks that score per-item accuracy, GRANI's unit of evaluation is a graded pattern across conditions, making it well suited for (underexplored) fine-grained phenomena that sit in the intersection of semantics and pragmatics. Concretely, for instance, a model that is 90% "character" in every sensory condition scores well on accuracy, but it fails to show the human gradient. Our work makes novel contributions in that the area it explores goes beyond semantic representations to more pragmatic or discourse relationships. LLM results varied across model properties and languages; particularly, for cases where LLMs fail, LLMs might lack "perceptual" representations that are required to match human performance. In other words, LLMs might not be limited in principle, but the apparent deficit might be a matter of multi-modality exposure and architectural complexity, and we speculate that vision-language models might show better (or different) capacity on this same benchmark dataset.

6 Conclusion

GRANI, a diagnostic protocol, evaluates LLMs' understanding of (implicit) narrative implications. Through the lens of PPTs, we demonstrate that compared with reasoning models, instruction-tuned models better approximate the human gradient, but they can be overconfident. With careful monitoring on uncertainty and overconfidence, instruction-tuned models could be preferable for subjective annotation tasks such as opinion, sentiment, and narrative perspective. We provide a showcase that the use of the graded perspective attribution is an effective tool for evaluating LLMs. GRANI's design is potentially transferable to similar phenomena with theoretically motivated gradient predictions, offering a pipeline for integrating LLMs for perspective attribution in (subjective) narratives, and more generally, model reliability evaluation.

7 Limitations

Our work has limitations. First, our models are limited to the 7-8B range, and larger models may have better capacity and behave differently. Majority voting over 10 samples captures response stability, but it might not track uncertainty as fully as probability distributions. We also do not examine variation across prompts, sampling methods, or parameters such as model's temperature, leaving questions about robustness. Ecologically, our experiments focus on predicates of personal taste, so the findings may not automatically generalize to other subjective phenomena such as metaphor, sarcasm, or multi-party dialogue. Finally, by "judge", we refer to the model's ability to infer and determine the *judge* in judge-dependent expressions, such as PPTs. We do not directly assess conventional "LLM-as-judge" performance or annotation benchmarks. Although downstream annotation and evaluation benefits are plausible, they remain untested and are beyond this work's scope.

8 Acknowledgment

We thank the anonymous reviewers for their constructive reviews and Marcin Morzycki for insights on formal semantics. This project is supported by the Computation and Linguistic Meaning (CALM) Lab at Purdue University.

References

Yan Cong. 2021. Judge-dependence in quality nouns: A semantic analysis of the mandarin chinese "you np" structure. *Asian Languages and Linguistics*, 2(2):167–216.

Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. 2023. Chatgpt outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences*, 120(30):e2305016120.

Salvatore Giorgi, Tingting Liu, Ankit Aich, Kelsey Jane Isman, Garrick Sherman, Zachary Fried, João Sedoc, Lyle Ungar, and Brenda Curtis. 2024. Modeling human subjectivity in llms using explicit and implicit human factors in personas. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 7174–7188.

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*.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, and 1 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Chongxuan Huang, Yongshi Ye, Biao Fu, Qifeng Su, and Xiaodong Shi. 2025. From neurons to semantics: Evaluating cross-linguistic alignment capabilities of large language models via neurons alignment. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 28956–28974, Vienna, Austria. Association for Computational Linguistics.

Chu-Ren Huang and Dingxu Shi. 2016. *A reference grammar of Chinese*. Cambridge University Press.

Yiqiao Jin, Mohit Chandra, Gaurav Verma, Yibo Hu, Munmun De Choudhury, and Srijan Kumar. 2024. Better to ask in english: Cross-lingual evaluation of large language models for healthcare queries. In *Proceedings of the ACM Web Conference 2024*, pages 2627–2638.

Elsi Kaiser. 2021. Consequences of sensory modality for perspective-taking: Comparing visual, olfactory and gustatory perception. *Frontiers in Psychology*, 12:701486.

Max Kölbel. 2004. III—Faultless disagreement. In *Proceedings of the Aristotelian society*, volume 104, pages 53–73. Oxford University Press Oxford, UK.

Peter Lasersohn. 2005. Context dependence, disagreement, and predicates of personal taste. *Linguistics and philosophy*, 28(6):643–686.

Alina Leidinger, Robert van Rooij, and Ekaterina Shutova. 2023. The language of prompting: What linguistic properties make a prompt successful? In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 9210–9232, Singapore. Association for Computational Linguistics.

Ziqiao Ma, Jacob Sansom, Run Peng, and Joyce Chai. 2023. Towards a holistic landscape of situated theory of mind in large language models. In *Findings of the association for computational linguistics: EMNLP 2023*, pages 1011–1031.

Rajiv Movva, Pang Wei Koh, and Emma Pierson. 2024. Annotation alignment: Comparing llm and human annotations of conversational safety. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 9048–9062.

Dilip Ninan. 2014. Taste predicates and the acquaintance inference. In *Semantics and linguistic theory*, pages 290–309.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, and 1 others. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Hazel Pearson. 2013. A judge-free semantics for predicates of personal taste. *Journal of semantics*, 30(1):103–154.

Zhuang Qiu, Xufeng Duan, and Zhenguang G Cai. 2025. Grammaticality representation in chatgpt as compared to linguists and laypeople. *Humanities and Social Sciences Communications*, 12(1):1–15.

Melanie Subbiah, Akankshya Mishra, Grace Kim, Liyan Tang, Greg Durrett, and Kathleen McKeown. 2025. Is the top still spinning? evaluating subjectivity in narrative understanding. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 185–203.

Qwen Team. 2024. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.

William S-Y Wang and Chaofen Sun. 2015. *The Oxford handbook of Chinese linguistics*. Oxford University Press.

Alex Wilf, Sihyun Lee, Paul Pu Liang, and Louis-Philippe Morency. 2024. Think twice: Perspective-taking improves large language models' theory-of-mind capabilities. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8292–8308.

Lang Yu and Allyson Ettinger. 2020. Assessing phrasal representation and composition in transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4896–4907, Online. Association for Computational Linguistics.

A Prompts used for expanding datasets

Bold text indicates column names in the original dataset, e.g., **begin** refers to "Right after I arrived at the party"; and **sensory_DP_PP** means "saw the salad" as in Table 2.

```
Expand the original dataset with
these constraints:

  • Reuse the original begin patterns
    (no new constructions)

  • Vary only lexical items:

      – Proper names
        (gender-balanced)
      – Concrete nouns similar to
        originals (e.g., the bacon,
        the soup, the bread)

  • Keep sensory_DP_PP strictly in the
    same style as the original data

  • Preserve all other columns
    exactly

  • Each set contains 4 items:
    smell, taste, see, control
```

B Response distribution: Fig.2,3,4

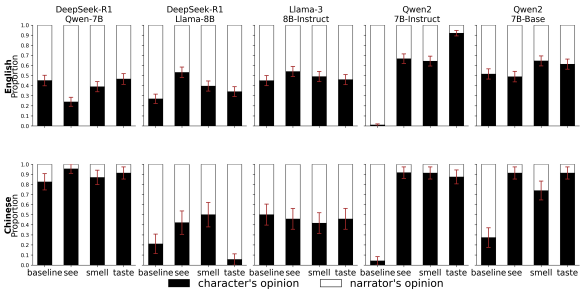


Figure 2: Proportion of character's opinion and narrator's opinion answer choices in Experiment 1. Error bars show +/-1 SE.

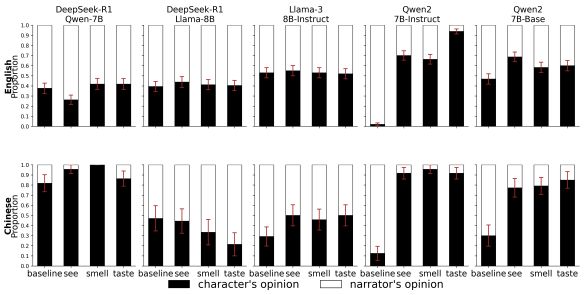


Figure 3: Proportion of character's opinion and narrator's opinion answer choices in Experiment 2. Error bars show +/-1 SE.

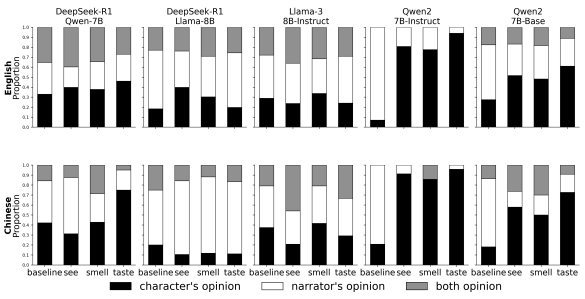


Figure 4: Proportion of character's opinion, narrator's opinion and both answer choices in Experiment 3.

C Confidence distribution: Fig.5,6

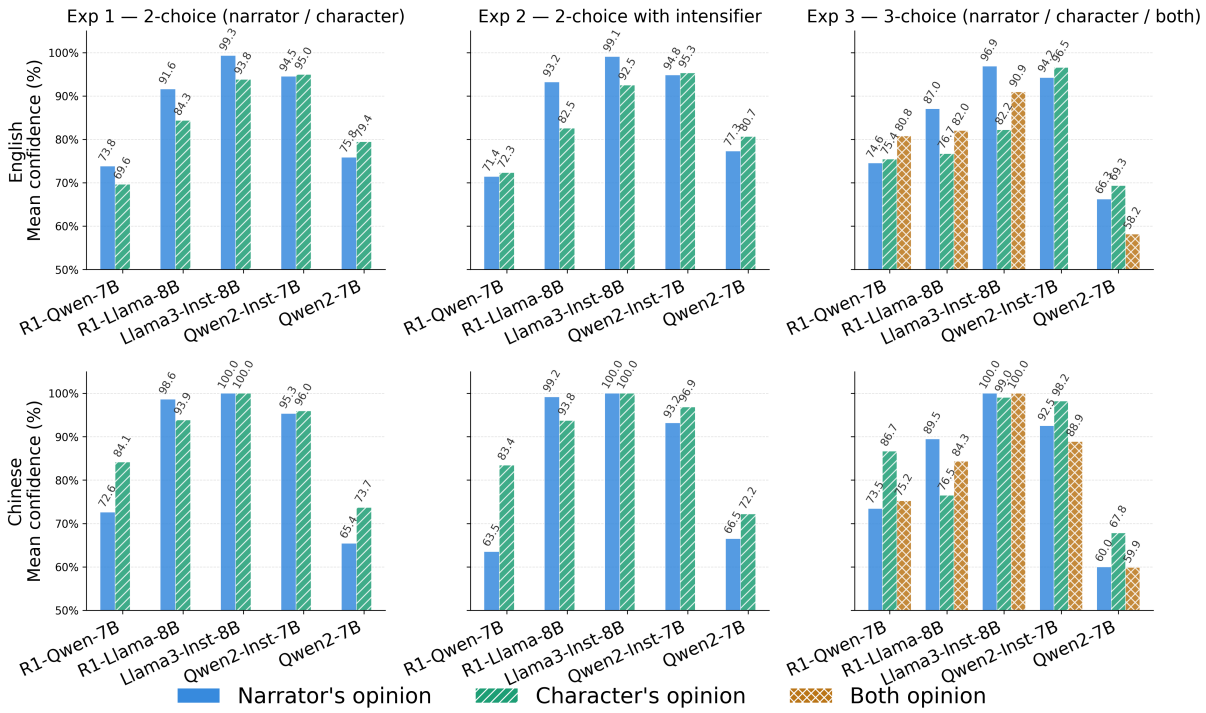


Figure 5: Dominant-choice confidence per response category.

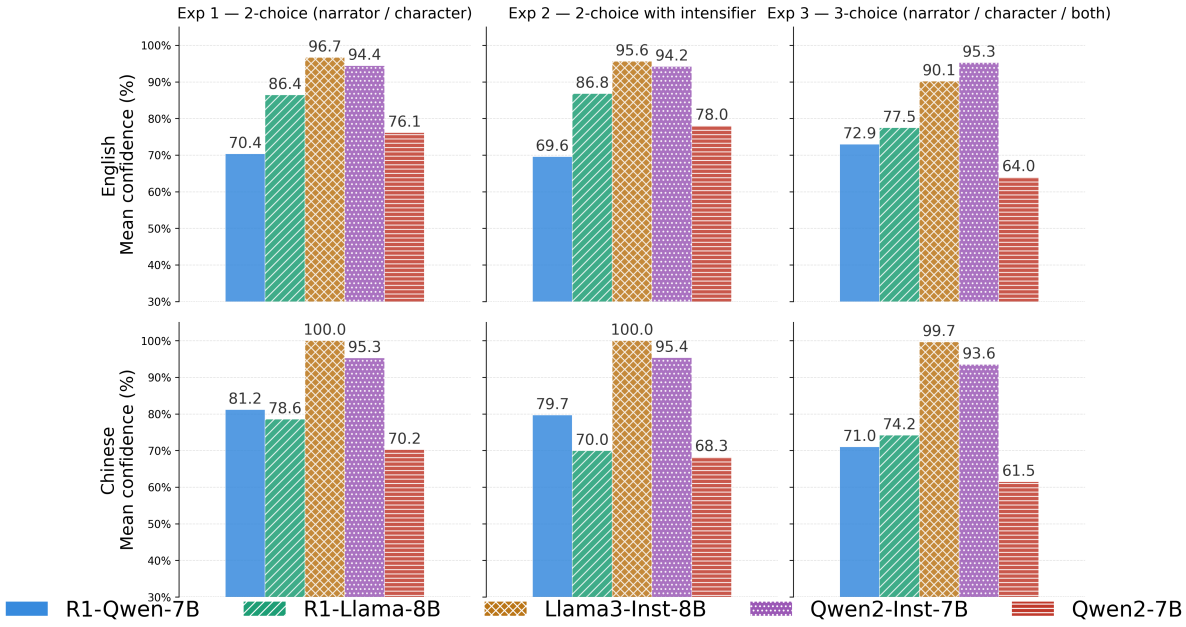


Figure 6: Dominant-choice confidence per LLM.