Understanding Subjectivity Transformation in Academic Meta-Review Generation

Sanju Basak and Suining He
Ubiquitous & Urban Computing Lab
University of Connecticut
Storrs, CT, USA
{sanju.basak, suining.he}@uconn.edu

Abstract

Academic meta-reviews must synthesize several reviewer opinions into a single, consistent judgment, balancing verifiable factual claims against subjective evaluations. Yet how human meta-reviewers transform reviewer subjectivity during summarization and whether large language models (LLMs) reproduce this behavior remains underexplored. To address this, we characterize subjectivity transformation in academic meta-review writing using a sentence-level subjectivity scorer, evaluating four open LLMs across six prompt configurations. Our analysis reveals that meta-review writing performs a regression to the mean of subjectivity: when a meta-review sentence is matched to its source reviewer's sentence, the change in subjectivity negatively correlates with the source's original subjectivity. In particular, actual, individual reviewer statements are made more interpretive, while highly subjective language tends to be moderated toward a balanced tone. All four open LLMs reproduce such a transformation closely. Conditioning the prompt on reviewer subjectivity yields only a small calibration gain where the generated meta-review's overall subjectivity moves slightly closer to the human reference, yet leaves the transformation itself unchanged, implying that there exists a robust latent behavior of LLM summarization rather than a prompt-controllable one. [1]

1 Introduction

Multi-document summarization condenses several source documents into one coherent summary that preserves salient information from multiple inputs. Academic meta-review generation is a high-stakes instance of this task: an area chair must read several peer reviews of a submission and produce a single meta-review that justifies the accept or reject decision (Kang et al., 2018; Zeng et al., 2025). A defining characteristic of such a setting is that reviews often mix two kinds of content (Wiebe et al., 2005): some review sentences are considered objective which are verifiable factual claims, while some others are considered subjective mentioning personal judgments. A competent meta-review should preserve important factual content while appropriately framing reviewer opinions.

The same challenge arises across multi-source, opinionated text summarization more broadly (Angelidis and Lapata, 2018), ranging from aggregating product reviews (Bražinskas et al., 2020), merging multiple annotations (Homayounirad et al., 2025), to synthesizing news from multiple outlets (Fabbri et al., 2019). Understanding how the summarization process itself reshapes subjectivity, therefore, has implications well beyond academic meta-review generation. Yet current automated meta-review approaches largely treat all review content uniformly, without differentiating between facts and opinions during summarization (Yuan et al., 2022). Nevertheless, the transformation of subjectivity is arguably a crucial *cognitive* effort throughout meta-review writing: a human area chair does not copy "this paper is poorly-written" verbatim but renders it as "reviewers raised concerns about the contribution". With increasing applications of large language models (LLMs) in these scenarios (for instance, AI review experimentation), understanding such transformation, and whether and how LLMs reproduce it, is a prerequisite for developing a more faithful, usable multi-document summarization system.

Specifically, we aim to answer the following two research questions:

- **Q1 (Characterize).** How do human meta-reviewers transform reviewer subjectivity during summarization?
- **Q2 (Reproduce).** Do LLMs reproduce this transformation, and how do certain levels of

[1] **Project Website:** https://github.com/Sanju-Basak/EMNLP26-Subjectivity-MetaReview

Findings of the Association for Computational Linguistics: EMNLP 2026, pages 40975–40990
October 24-29, 2026 ©2026 Association for Computational Linguistics

prompting affect it?

In answering the above questions, we have performed extensive, real-world meta-review analysis, and prepared a cross-domain sentence-level subjectivity scorer used to analyze and quantify subjectivity. We have particularly gained the following understandings that will drive our future research.

Understanding 1. We specially find that meta-review writing *characterizes* a systematic subjectivity transformation: factual, individual reviewer statements are made more interpretive, while highly subjective language tends to be moderated toward a balanced tone. This is revealed by a strong negative correlation between source subjectivity and per-sentence subjectivity shift.

Understanding 2. We have designed multi-level prompting to steer vanilla, instruction-tuned LLMs to *reproduce* this transformation without being explicitly trained to do so. Our subjectivity-aware prompting yields modest improvements in calibration without significantly altering the transformation itself. Such a behavior is therefore considered a robust latent behavior of how LLMs summarize opinionated text, rather than a knob that can be turned through prompting. This implies that multi-document opinion summarizers will systematically neutralize strong opinions and editorialize factual content regardless of prompting.

2 Related Work

Peer review has become an important NLP setting for studying scientific discourse: PeerRead (Kang et al., 2018) established it as a benchmark, Yuan et al. (2022) formalized meta-review generation as multi-document summarization, and ORSUM (Zeng et al., 2025) scaled it to 15,062 meta-reviews via checklist-guided introspection. Most of this body of work, however, evaluates only lexical overlap and content coverage, and Salvador et al. (2025) found that LLMs struggle to consolidate consensus across opinion-heavy inputs. Recent work has shown that LLMs systematically misjudge subjective language and treat hedged claims as facts (Jones et al., 2025), motivating our use of an explicit, dedicated subjectivity scorer rather than relying on LLM self-assessment. We train such a scorer on SubjQA (Bjerva et al., 2020), using its continuous `ans_subj_score` as a regression target. This explicit-scoring approach aligns with findings that direct subjectivity identification outperforms indirect inference in human value recognition tasks (Homayounirad et al., 2025).

The closest work to ours is Li et al. (2024), who hypothesized that meta-reviewers follow a three-layer sentiment consolidation framework and showed that prompting LLMs with this structure yields better meta-reviews. However, they modeled discrete per-facet sentiment polarity rather than continuous subjectivity and, like all prior work, treated the consolidation process as a black box rather than measuring how reviewer subjectivity is transformed during summarization.

3 Method of Our Studies

3.1 Dataset Preparation

We use three datasets for complementary purposes.

SubjQA (Bjerva et al., 2020) provides continuous subjectivity annotations for scorer training. We extract all annotated subjective answer spans (rather than only the first span), increasing training data by approximately $\sim$49% to 4,895 examples across six domains (See Appendix D)

ORSUM (Zeng et al., 2025) provides human meta-reviews paired with peer reviews. We use the ICLR-2019 subset and restrict analysis to the 130 test papers containing exactly three reviewers, enabling a clean definition of reviewer coverage and reducing variability introduced by differing review counts. To evaluate cross-venue generalization, we additionally use the NeurIPS-2022 and ICLR-2022 subsets of ORSUM (§4.5).

ASAP-Review (Yuan et al., 2022) provides 15 annotated review aspects (e.g., soundness, originality, clarity), which we use for aspect-level analysis of subjectivity.

3.2 Measuring Subjectivity

We fine-tune four transformer encoders, namely, BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), DeBERTa (He et al., 2020), and DistilBERT(Sanh et al., 2019) on SubjQA using a regression objective. Each model predicts a continuous subjectivity score from a sentence using a simple regression head: $[\mathrm{CLS}] \rightarrow \mathrm{Dropout}(0.1) \rightarrow \mathrm{Linear} \rightarrow \mathrm{Sigmoid}$, trained with the mean squared error (MSE) loss. We adopt **DistilBERT** as the primary scorer (lowest validation RMSE 0.206; §4.1) for its 7$\times$ training speedup and use it to assign a subjectivity score $\sigma(s) \in [0, 1]$ to every sentence.

While SubjQA originates from consumer reviews, our later analyses (see §4.2 & Appendix E) show that the learned signal transfers meaning-

fully to academic review text via consistent correlations with reviewer ratings, confidence, and aspect-level sentiment. To further validate cross-domain generalization, we conduct leave-one-domain-out (LODO) evaluation across SubjQA's six domains, confirming that the scorer captures domain-invariant subjectivity markers rather than domain-specific vocabulary, with minimal performance degradation on unseen domains (Appendix F).

3.3 Prompt Construction & Models

Following TELeR (Karmaker Santu and Feng, 2023), we construct prompts at three levels (L1–L3) each with two variants, with detailed instructions inspired by official AI-review formats[2].

The `-base` version includes only reviewer text, while the `-sub` version additionally provides explicit subjectivity statistics for each review, including mean subjectivity, standard deviation, and high/low subjectivity ratios. Detailed prompts can be found on Appendix C.

We generate from four open instruction-tuned LLMs: Gemma-2-2B (Team et al., 2024), Qwen2.5-3B (Yang et al., 2024), Mistral-7B (Jiang et al., 2023), and Llama-3-8B (Grattafiori et al., 2024). [3] All are used in their off-the-shelf form, yielding a total of 24 generation conditions (4 models $\times$ 3 levels $\times$ 2 variants) over the 130 test papers.

3.4 Evaluation Metrics

NLP Similarity. We measure similarity to human meta-reviews using ROUGE (Lin, 2004), BLEU (Papineni et al., 2002), and BERTScore (Zhang et al., 2019).

Subjectivity Calibration. We evaluate whether generated meta-reviews match the tone of human meta-reviews using subjectivity mean absolute error (MAE_{subj}):

$$\text{MAE}_{\text{subj}} = \frac{1}{N}\sum_{j=1}^{N} \left|\overline{\sigma}_i(\text{gen}) - \overline{\sigma}_i(\text{human})\right|,$$

where $\overline{\sigma}_j(\text{gen})$ and $\overline{\sigma}_j(\text{human})$ denote the mean sentence-level subjectivity of generated and human meta-reviews for a paper j.

[2] https://2026.emnlp.org/ai-reviewing-experiment/

[3] Specifically gemma-2-2b-it, Qwen2.5-3B-Instruct, Mistral-7B-Instruct-v0.3, and Meta-Llama-3-8B-Instruct (abbreviated in tables and figures).

Subjectivity Transformation. We align each sentence m_i in meta-review to its most similar source-review (individual review) sentence m^* (based on MPNet cosine (Song et al., 2020), threshold 0.45) and tracks the shift $\Delta_i = \sigma(m_i) - \sigma(m^*)$, where $\sigma(\cdot)$ denotes the subjectivity score, so Δ_i is the change in subjectivity between meta-review sentence m_i and its matched source sentence m^*. A sensitivity sweep across $\tau \in \{0.35 - 0.55\}$ confirms that the transformation correlation is effectively invariant to threshold choice (variation in $r < 0.03$; see Appendix G).

4 Experiments and Results

4.1 Subjectivity Scorer Performance

Table 1 compares four transformer encoders fine-tuned on SubjQA. DistilBERT attains the lowest validation root mean square error (RMSE) (0.206) while training roughly 7$\times$ faster than the large variants; RoBERTa-large's marginally higher Pearson correlation does not justify its cost. We therefore adopt DistilBERT as the scorer for all downstream analysis, using it to assign a subjectivity score $\sigma(\cdot) \in [0, 1]$ to every sentence. Formal definitions of RMSE and Pearson r are given in Appendix B.

Model	Val RMSE ↓	Pearson ↑	Time ↓
DistilBERT ⋆	**0.2060**	0.641	**219s**
BERT-large	0.2065	0.635	1,483s
RoBERTa-large	0.2079	**0.649**	1,480s
DeBERTa-v3-base	0.2144	0.637	644s

Table 1: Subjectivity scorer comparison on the all-domain SubjQA validation set. Lower RMSE and higher Pearson are better. Best per metric in **bold**.

4.2 Subjectivity vs. Rating & Confidence

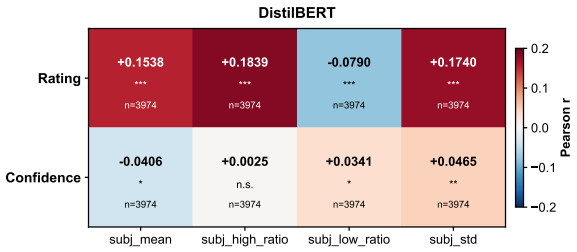


Figure 1: Subjectivity features vs. Rating and Confidence. Red means positive correlation and Blue means negative correlation.

Applying the scorer to the ICLR-2019 reviews (ORSUM), we find that review subjectivity can be

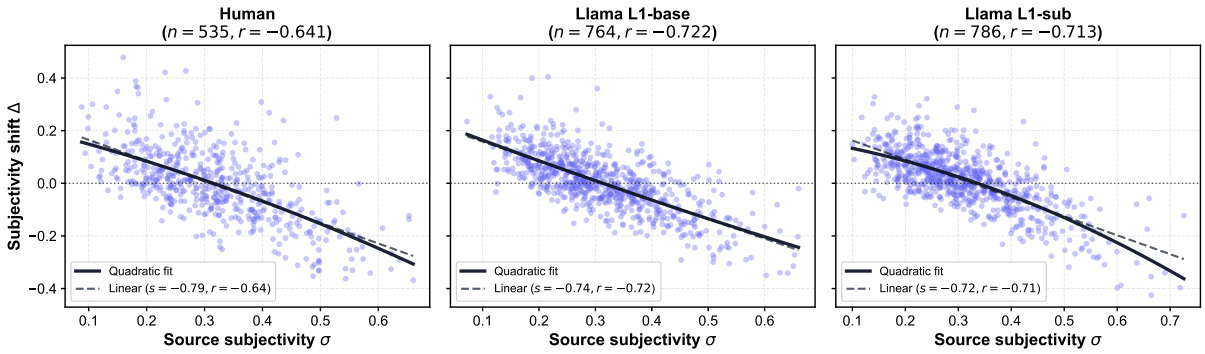


Figure 2: Subjectivity shift $\Delta = \sigma(m) - \sigma(m^*)$ vs. source subjectivity σ, for humans and Qwen2.5-3B (`L1-base` and `L1-sub`).

considered a meaningful signal rather than a surface artifact. Figure 1 shows that mean review subjectivity correlates positively with the reviewers' rating scores (1–10 scale), and negatively with reviewers' confidence scores (1–5 scale), suggesting that enthusiastic reviewers write more subjectively, while more confident reviewers ground their assessments in comparatively objective, factual language. Such correlations assure that the scorer captures an interpretable axis of review language, validating its use in our later transformation analysis.

4.3 Does Subjectivity Conditioning Help?

Model	base	sub	Δ
Gemma-2-2B	0.0521	**0.0487**	−0.0034
Qwen2.5-3B	0.0519	**0.0469**	−0.0050
Mistral-7B	0.0492	**0.0451**	−0.0041
Llama-3-8B	**0.0487**	0.0511	+0.0024

Table 2: Subjectivity MAE (lower is better) under the L1 prompt, base vs. (sub). Δ = sub−base.

We next investigate whether supplying subjectivity statistics in the prompt (`-sub`) improves a model's subjectivity calibration relative to the plain prompt (`-base`). Table 2 reports $\mathrm{MAE}_{\mathrm{subj}}$ under the L1 prompt. We can see that conditioning lowers $\mathrm{MAE}_{\mathrm{subj}}$ for three of four models. Qwen2.5-3B improves most (-0.0050), followed by Mistral-7B (-0.0041) and Gemma-2-2B (-0.0034), while Llama-3-8B is essentially unchanged ($+0.0024$). Such gains are small ($\leq$ 0.005) and no single configuration dominates, foreshadowing our finding (§4.4) that the underlying subjectivity transformation is largely a robust latent behavior rather than prompt-controlled. Paired bootstrap tests (B = 10,000) confirm this: 10 of 12 MAE differences across all model × prompt-level cells have 95% confidence intervals (CIs) including zero, with all effect sizes negligible ($|d| < 0.21$; see Appendix H). The overall meta-review generation quality is likewise insensitive to model and prompt choice: across all 24 `base` conditions, BERTScore-F1 spans only 0.775–0.787 and ROUGE-1 spans 0.27–0.31 (see full grid in Appendix I).

4.4 Is the Meta-Review Always Neutral?

We have observed that meta-reviews preferentially draw from subjective reviewer content. Aligned source sentences are significantly more subjective than the overall reviewer corpus (source $\bar{\sigma} = 0.316$ vs. corpus-wide $\bar{\sigma} = 0.281$, Welch's t-test, $p < 10^{-12}$), implying that meta-reviews rely disproportionately on evaluative review text. Our analysis covers only these aligned sentences; for human meta-reviews, 22.8% of sentences fall below the alignment threshold and are excluded. These unmatched sentences including decision statements and original synthesis have slightly lower mean subjectivity (0.292 vs. 0.311 for aligned sentences). For LLMs, the unmatched fraction is much smaller (2.5–7.2%), suggesting that LLMs produce fewer novel sentences than human area chairs.

Despite this preference for evaluative content, meta-reviews do not simply preserve reviewer subjectivity. Although the average shift for humans is near zero ($\bar{\Delta} = -0.005$, not significant)), plotting Δ against σ reveals a clear regression-to-the-mean effect (Figure 2): factual sources are editorialized upward ($\Delta > 0$), while highly subjective sources are softened ($\Delta < 0$). For instance, a reviewer's enthusiastic praise ($\sigma = 0.49$): "Overall the paper is very well-written, clear to follow, and the authors did a great job in not overclaiming their results" is condensed in the meta-review to ($\sigma = 0.14$): "The paper is well-written" ($\Delta = -0.35$). Conversely, a factual remark about baseline models

($\sigma = 0.16$) is rendered as "The strong performance of the baseline and baseline++ are quite surprising" ($\sigma = 0.64$, $\Delta = +0.48$). The shift crosses zero near $\sigma \approx 0.30$, close to the typical subjectivity of human meta-reviews, yielding a strong negative correlation ($r = -0.64$, $p < 10^{-63}$, 95% bootstrap CI $[-0.69, -0.58]$, $B = 10{,}000$).

All four LLMs tested reproduce this pattern, often more strongly than humans. For Qwen2.5-3B as an instance, the correlation reaches $r = -0.75$ under `L1-base` and $r = -0.80$ under `L1-sub` (both $p < 10^{-139}$), exceeding the human effect size. Across all 24 conditions, correlations range from $r = -0.69$ to -0.84 (Appendix J), suggesting that instruction-tuned LLMs naturally "editorialize the bland and neutralize the bold" without task-specific supervision.

Subjectivity-aware prompting does not materially change this behavior. Within each model, `L1-base` and `L1-sub` produce statistically indistinguishable transformations, with only small and inconsistent slope differences (Appendix J). Combined with the limited conditioning effects in §4.3, this suggests that subjectivity regression is a robust latent behavior of LLM-based summarization.

4.5 Multi-Venue Generalization

To verify that the regression-to-the-mean transformation generalizes beyond ICLR-2019, we apply the same scoring and alignment pipeline to NeurIPS-2022 and ICLR-2022 in ORSUM, generating meta-reviews with Llama-3-8B under the L1 prompt. As in Figure 3, the pattern holds consistently across all three venues. Human meta-reviews exhibit strong negative correlations ($r = -0.64$ for ICLR-2019, $r = -0.69$ for NeurIPS-2022, and $r = -0.65$ for ICLR-2022; all $p < 10^{-73}$), and Llama-3-8B reproduces the pattern with comparable or stronger effect sizes ($|r| = 0.71$–0.74). The consistency across venues and a three-year gap confirms that the subjectivity transformation is not a venue- or year-specific artifact. Full scatter plots can be referred to Appendix K.

5 Conclusion

We have derived two key understandings in investigating subjectivity-aware academic meta-review generation. The first understanding lies in the characterization of a systematic subjectivity transformation in human-generated meta-reviews, where factual reviewer content tends to be editorialized upward, while highly opinionated content is systematically softened. The second understanding arises from subjectivity-aware prompting in LLMs, where systematic neutralization of strong opinions and editorialization of factual content persist regardless of model or prompt configuration.

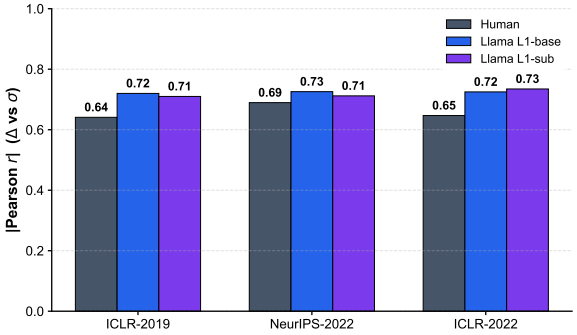


Figure 3: |Pearson r| (Δ vs. σ) across three venue–year subsets of ORSUM.

These understandings, as findings, carry several practical implications, and enlighten our future path in developing AI review experimentation. The regression-to-the-mean pattern provides a diagnostic benchmark for future meta-review systems to evaluate whether a system reproduces this behavioral signature, beyond lexical overlap alone. The ineffectiveness of prompting suggests that practitioners seeking subjectivity control should invest in fine-tuning with subjectivity-aware objectives rather than prompt engineering, while the stability of the transformation makes post-hoc subjectivity calibration feasible. Finally, our aspect analysis suggests that generation systems should treat positive and negative review content differently during summarization. This transformation can apply to broad multi-document opinion summarization tasks, and our framework can be directly extended to diagnose it in other domains.

Limitations

Our analysis covers only English-language venues; generalization to non-English venues and different review cultures remains to be verified. The subjectivity scorer is trained on consumer reviews and produces scores in a relatively narrow range (0.15–0.55) on academic text, limiting dynamic range. Future human annotation studies will be considered. Finally, our transformation analysis covers only aligned sentences; the excluded novel sentences exhibit slightly lower subjectivity and merit dedicated study in future work.

Ethical considerations

All datasets used are publicly available. ORSUM is derived from OpenReview, where reviews are posted under the platform's terms of service with the understanding that they may be used for research. We do not attempt to de-anonymize reviewers, and no personally identifiable information is used.

Acknowledgments

This project is supported, in part, by the National Science Foundation (NSF) under Grants 2239897 and 2608884, Google Research Scholar Program Award, NVIDIA Applied Research Accelerator Program Award, and the Connecticut AI Alliance (CAIA) and the Connecticut Conference of Independent Colleges (CCIC) AI Computing Access Program Award. Any opinions, findings, and conclusions or recommendationsexpressed in this material are those of the authors and do notnecessarily reflect the views of the funding agencies.

ChatGPT and Claude were used solely for code debugging and grammatical correction; all research design, experiments, analysis, and writing are our own.

References

Stefanos Angelidis and Mirella Lapata. 2018. Summarizing opinions: Aspect extraction meets sentiment prediction and they are both weakly supervised. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3675–3686, Brussels, Belgium. Association for Computational Linguistics.

Johannes Bjerva, Nikita Bhutani, Behzad Golshan, Wang-Chiew Tan, and Isabelle Augenstein. 2020. Subjqa: A dataset for subjectivity and review comprehension. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5480–5494.

Arthur Bražinskas, Mirella Lapata, and Ivan Titov. 2020. Few-shot learning for opinion summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4119–4135, Online. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. 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 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Alexander Fabbri, Irene Li, Tianwei She, Suyi Li, and Dragomir Radev. 2019. Multi-news: A large-scale multi-document summarization dataset and abstractive hierarchical model. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1074–1084, Florence, Italy. Association for Computational Linguistics.

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*.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020. Deberta: Decoding-enhanced bert with disentangled attention. *arXiv preprint arXiv:2006.03654*.

Amir Homayounirad, Enrico Liscio, Tong Wang, Catholijn M. Jonker, and Luciano C. Siebert. 2025. Will annotators disagree? identifying subjectivity in value-laden arguments. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 15237–15252. Association for Computational Linguistics.

Albert Qiaochu Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de Las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *ArXiv*, abs/2310.06825.

Erik Jones, Arjun Patrawala, and Jacob Steinhardt. 2025. Uncovering gaps in how humans and llms interpret subjective language. In *International Conference on Learning Representations*, volume 2025, pages 78664–78703.

Dongyeop Kang, Waleed Ammar, Bhavana Dalvi, Madeleine Van Zuylen, Sebastian Kohlmeier, Eduard Hovy, and Roy Schwartz. 2018. A dataset of peer reviews (peerread): Collection, insights and nlp applications. In *Proceedings of the 2018 conference of the North American chapter of the Association for Computational Linguistics: Human language technologies, volume 1 (long papers)*, pages 1647–1661.

Shubhra Kanti Karmaker Santu and Dongji Feng. 2023. TELeR: A general taxonomy of LLM prompts for benchmarking complex tasks. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 14197–14203, Singapore. Association for Computational Linguistics.

Miao Li, Jey Han Lau, and Eduard Hovy. 2024. A sentiment consolidation framework for meta-review generation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10158–10177.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

John Salvador, Naman Bansal, Mousumi Akter, Souvika Sarkar, Anupam Das, and Santu Karmaker. 2025. Benchmarking LLMs on semantic overlap summarization. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 33352–33373, Suzhou, China. Association for Computational Linguistics.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *ArXiv*, abs/1910.01108.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2020. Mpnet: Masked and permuted pre-training for language understanding. *Advances in neural information processing systems*, 33:16857–16867.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, and 1 others. 2024. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*.

Janyce Wiebe, Theresa Wilson, and Claire Cardie. 2005. Annotating expressions of opinions and emotions in language. *Language resources and evaluation*, 39(2):165–210.

Qwen An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxin Yang, Jingren Zhou, Junyang Lin, and 25 others. 2024. Qwen2.5 technical report. *ArXiv*, abs/2412.15115.

Weizhe Yuan, Pengfei Liu, and Graham Neubig. 2022. Can we automate scientific reviewing? *Journal of Artificial Intelligence Research*, 75:171–212.

Qi Zeng, Mankeerat Sidhu, Ansel Blume, Hou Pong Chan, Lu Wang, and Heng Ji. 2025. Scientific opinion summarization: Paper meta-review generation dataset, methods, and evaluation. In *Artificial Intelligence for Research and Democracy*, pages 20–38, Singapore. Springer Nature Singapore.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Appendix

A Experimentation Workflow

We use a three-stage pipeline for our experiments.

- Firstly, using the subjectivity scorer (DistilBERT), we score every review sentence.
- Secondly, using 4 LLMs and 6 prompts, we generated meta-review for a total of 24 conditions on 130 ICLR-2019 papers.
- Finally, we evaluate the generations using NLP metrics and $\mathrm{MAE}_{\mathrm{Subj}}$, and measure subjectivity transformation.

B Evaluation Metrics

Root Mean Square Error (RMSE).

$$\mathrm{RMSE} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\big(\hat{\sigma}(m_i) - \sigma(m_i)\big)^2}, \quad (1)$$

where $\hat{\sigma}(m_i)$ is the predicted subjectivity score and $\sigma(m_i)$ is the human-annotated score for sentence i.

Pearson Correlation Coefficient (r).

$$r = \frac{\sum_{i=1}^{N}\left(\hat{\sigma}(m_i) - \mu_\sigma(m_i)\right)\left(\sigma(m_i) - \bar{\sigma}(m_i)\right)}{\sqrt{\sum_{i=1}^{N}\left(\hat{\sigma}(m_i) - \mu_\sigma(m_i)\right)^2 \sum_{i=1}^{N}\left(\sigma(m_i) - \bar{\sigma}(m_i)\right)^2}}, \quad (2)$$

where $\mu_\sigma(m_i)$ and $\bar{\sigma}(m_i)$ are the means of the predicted and annotated scores, respectively. A value of $r = 1$ indicates perfect positive correlation.

C Prompt Templates

Shared Components

System Prompt

```
You are an expert area chair at an academic conference.
```

Length Constraint

```
Write in flowing prose only—no bullet points, no headers, no numbered lists. Aim for approximately 4-6 sentences (around 100-150 words).
```

Subjectivity Legend (`-sub` variants only)

- `subj_mean`: Mean subjectivity score across all sentences in the review
- `subj_std`: Variation in subjectivity tone across sentences
- `subj_high_ratio`: Fraction of sentences that are strongly opinionated
- `subj_low_ratio`: Fraction of sentences that are factual and assertive

TELeR Level 1 Template. A single-sentence directive prepended to the reviews.

L1-base

```
Write a meta-review that summarizes the key points from the peer reviews below.
{{LENGTH_CONSTRAINT}}
[Review 1] (rating={{r}}/10, confidence={{c}}/5)
{{Review 1 text}}
[Review 2] (rating={{r}}/10, confidence={{c}}/5)
{{Review 2 text}}
[Review 3] (rating={{r}}/10, confidence={{c}}/5)
{{Review 3 text}}
```

L1-sub

Same instruction; reviews are additionally annotated with subjectivity scores.

```
Write a meta-review that summarizes the key points from the peer reviews below.
{{LENGTH_CONSTRAINT}}
[Review 1] (rating={{r}}/10, confidence={{c}}/5)
{{Review 1 text}}
Subjectivity scores: subj_mean={{x.xxx}}, subj_std={{x.xxx}}, subj_high_ratio={{x.xxx}}, subj_low_ratio={{x.xxx}}
[Review 2] (rating={{r}}/10, confidence={{c}}/5)
{{Review 2 text}}
Subjectivity scores: subj_mean={{x.xxx}}, subj_std={{x.xxx}}, subj_high_ratio={{x.xxx}}, subj_low_ratio={{x.xxx}}
[Review 3] (rating={{r}}/10, confidence={{c}}/5)
{{Review 3 text}}
Subjectivity scores: subj_mean={{x.xxx}}, subj_std={{x.xxx}}, subj_high_ratio={{x.xxx}}, subj_low_ratio={{x.xxx}}
{{SUBJ_LEGEND}}
```

TELeR Level 2 Template. A paragraph-level directive specifying sub-tasks.

L2-base

```
Write a meta-review for this paper based on the peer reviews below. Your meta-review should identify the major strengths and weaknesses raised by the reviewers, highlight any points of agreement or disagreement among reviewers, and provide a clear overall recommendation. {{LENGTH_CONSTRAINT}}
[Review 1] (rating={{r}}/10, confidence={{c}}/5)
{{Review 1 text}}
```

[Review 2] (rating={{r}}/10, confidence={{c}}/5)
{{Review 2 text}}

[Review 3] (rating={{r}}/10, confidence={{c}}/5)
{{Review 3 text}}

L2-sub

Adds an instruction to reflect subjectivity in tone, plus subjectivity scores per review

Write a meta-review for this paper based on the peer reviews below. Your meta-review should identify the major strengths and weaknesses raised by the reviewers, highlight any points of agreement or disagreement among reviewers, and provide a clear overall recommendation. **The tone of your meta-review should reflect the subjectivity scores of each review shown below.** {{LENGTH_CONSTRAINT}}

[Review 1] (rating={{r}}/10, confidence={{c}}/5)
{{Review 1 text}}
Subjectivity scores: subj_mean={{x.xxx}}, subj_std={{x.xxx}}, subj_high_ratio={{x.xxx}}, subj_low_ratio={{x.xxx}}

[Review 2] (rating={{r}}/10, confidence={{c}}/5)
{{Review 2 text}}
Subjectivity scores: subj_mean={{x.xxx}}, subj_std={{x.xxx}}, subj_high_ratio={{x.xxx}}, subj_low_ratio={{x.xxx}}

[Review 3] (rating={{r}}/10, confidence={{c}}/5)
{{Review 3 text}}
Subjectivity scores: subj_mean={{x.xxx}}, subj_std={{x.xxx}}, subj_high_ratio={{x.xxx}}, subj_low_ratio={{x.xxx}}

{{SUBJ_LEGEND}}

TELeR Level 3 Template. An itemized directive enumerating explicit sub-tasks.

L3-base

Write a meta-review for this paper based on the peer reviews below. Your meta-review must:

- Summarize the key contributions as described by the reviewers
- Identify the major strengths agreed upon by the reviewers
- Identify the major weaknesses or concerns raised
- Note any significant disagreements among reviewers
- Provide a final recommendation (accept or reject)
- {{LENGTH_CONSTRAINT}}
- Write in flowing prose, not bullet points

[Review 1] (rating={{r}}/10, confidence={{c}}/5)
{{Review 1 text}}

[Review 2] (rating={{r}}/10, confidence={{c}}/5)
{{Review 2 text}}

[Review 3] (rating={{r}}/10, confidence={{c}}/5)
{{Review 3 text}}

L3-sub

Adds rating-consistency and subjectivity-style sub-tasks plus a constraint to omit subjectivity scores from the output

Write a meta-review for this paper based on the peer reviews below. Your meta-review must:

- Summarize the key contributions as described by the reviewers
- Identify the major strengths agreed upon by the reviewers
- Identify the major weaknesses or concerns raised
- Note any significant disagreements among reviewers
- **Provide a final recommendation consistent with the average rating**
- **Adjust your writing style based on the subjectivity scores shown for each review below**
- {{LENGTH_CONSTRAINT}}
- Write in flowing prose, not bullet points
- **Do not mention the subjectivity scores in your output**

[Review 1] (rating={{r}}/10, confidence={{c}}/5)
{{Review 1 text}}
Subjectivity scores: subj_mean={{x.xxx}}, subj_std={{x.xxx}}, subj_high_ratio={{x.xxx}}, subj_low_ratio={{x.xxx}}

[Review 2] (rating={{r}}/10, confidence={{c}}/5)
{{Review 2 text}}
Subjectivity scores: subj_mean={{x.xxx}}, subj_std={{x.xxx}}, subj_high_ratio={{x.xxx}}, subj_low_ratio={{x.xxx}}

[Review 3] (rating={{r}}/10, confidence={{c}}/5)
{{Review 3 text}}
Subjectivity scores:

Domain	Before	After	Gained	% Gain
books	445	629	184	41.3%
electronics	511	760	249	48.7%
movies	513	754	241	47.0%
restaurants	598	871	273	45.7%
tripadvisor	686	1095	409	59.6%
grocery	523	786	263	50.3%
Total	**3,276**	**4,895**	**1,619**	**49.4%**

Table 3: SubjQA training examples before and after extracting all annotated subjective answer spans (vs. only the first span), per domain. Extracting all spans yields a 49.4% increase overall.

```
    subj_mean={{x.xxx}},
    subj_std={{x.xxx}},
    subj_high_ratio={{x.xxx}},
    subj_low_ratio={{x.xxx}}
    {{SUBJ_LEGEND}}
```

D SubjQA Span Extraction

SubjQA (Bjerva et al., 2020) is a question-answering dataset in which each answer span is annotated with a continuous subjectivity score. Each example may contain multiple annotated answer spans, but a naive extraction keeps only the first span per example, discarding the rest. We instead extract *all* annotated subjective answer spans, treating each as an independent training instance with its own `ans_subj_score` regression target. We additionally guard against empty annotations by checking if `len(answers["ans_subj_score"]) == 0`.

This procedure increases the training set from 3,276 to 4,895 examples (a 49.4% gain) across the six SubjQA domains. Table 3 reports the per-domain counts. Gains are consistent across domains (ranging from 41.3% for *books* to 59.6% for *tripadvisor*), confirming that multi-span answers are common throughout the corpus rather than concentrated in any single domain.

E Subjectivity Across Review Aspects

The ASAP-Review dataset (Yuan et al., 2022) provides fine-grained aspect-level annotations for ICLR peer reviews. Each review span is annotated with one of seven aspects — *soundness* (technical correctness), *substance* (depth of contribution), *originality* (novelty), *clarity* (presentation quality), *motivation* (significance and relevance), *replicability* (reproducibility), and *meaningful_comparison* (positioning relative to prior work) — together with a sentiment polarity (*positive* or *negative*). We additionally distinguish a *summary* category for non-evaluative descriptive spans.

Figure 4 shows the mean subjectivity score (assigned by our DistilBERT scorer) for positive and negative spans of each aspect. Two patterns emerge. First, subjectivity is highest in evaluative aspects with positive sentiment (e.g., soundness–positive: 0.425; originality–positive: 0.397) and lowest in summary spans (0.249). Second, positive-sentiment spans are significantly more subjective than negative-sentiment spans for five of seven aspects (soundness, substance, originality, clarity, and motivation; all $p < 0.05$, Welch's t-test), while replicability ($n_{\text{pos}} = 24$, too few spans) and meaningful_comparison ($p = 0.27$) show no significant difference. This confirms that human reviewers' praise tends to be opinionated whereas criticism is more grounded in concrete, verifiable observations.

This connects to the transformation in §4.4: meta-reviews draw heavily from praise, which tends to be subjective. The regression-to-the-mean therefore acts mainly by neutralizing enthusiastic endorsements, while criticism, already close to the neutral baseline, changes little. The aspect analysis thus explains *why* the transformation has the shape it does, not just *that* it exists.

F Cross-Domain Scorer Validation

To validate that the subjectivity scorer captures domain-invariant linguistic markers rather than domain-specific vocabulary, we have conducted a leave-one-domain-out (LODO) evaluation across SubjQA's six domains. For each fold, the scorer is trained on five domains and evaluated on the completely held-out sixth. Table 4 reports cross-domain Pearson r for all four transformer encoders.

Across models, the average cross-domain Pearson r degrades by only 0.014–0.038 compared to the all-domain setting (Table 1), confirming that the scorers learn generalizable subjectivity features, such as hedging language, evaluative adjectives, personal pronouns, and intensifiers, rather than domain-specific lexical cues. Notably, the hardest transfer target is *movies* ($r = 0.48$–0.50), whose idiosyncratic evaluative vocabulary (e.g., "gripping," "cinematic") differs most from the other domains. In contrast, *restaurants* ($r = 0.72$–0.76) transfers easily, as its subjectivity markers (e.g., "I loved," "the service was terrible") are shared broadly across domains. Academic peer reviews, which rely on formal hedging (e.g., "we believe," "arguably

novel") and evaluative adjectives (e.g., "strong contribution," "poorly motivated"), are closer to these universal patterns than to domain-specific vocabulary, supporting the scorer's applicability to academic review texts.

G Threshold Sensitivity

To verify that our transformation analysis is not an artifact of the cosine similarity threshold $\tau = 0.45$, we repeat the Δ-vs-σ correlation analysis across five thresholds. Table 5 reports Pearson r and aligned sentence count n for human meta-reviews and all four LLMs under the L1-base prompt.

The negative correlation is effectively invariant to threshold choice: for humans, r ranges from -0.64 to -0.67 (variation < 0.03), and for all four LLMs the variation is even smaller (≤ 0.02). Stricter thresholds reduce the number of aligned pairs but do not alter the transformation pattern, confirming that both the existence and magnitude of the regression-to-the-mean effect are robust to the matching procedure.

H Bootstrap Confidence Intervals

To assess the statistical significance of the subjectivity conditioning effect, we conduct paired bootstrap tests ($B = 10{,}000$) on the MAE_{subj} differences (sub $-$ base) across all 12 model $\times$ prompt-level cells. Table 6 reports the per-cell MAE for both variants, the mean difference Δ, its 95% confidence interval, and Cohen's d effect size.

Ten of twelve CIs include zero, confirming that the conditioning gains are not statistically significant. The two exceptions do not support a consistent conditioning benefit: Mistral-7B at L2 shows a borderline improvement ($\Delta = -0.0046$, CI $[-0.0092, -0.0001]$), while Qwen2.5-3B at L3 shows conditioning actually *worsening* calibration ($\Delta = +0.0056$, CI $[+0.0009, +0.0103]$). All effect sizes are negligible ($|d| < 0.21$), reinforcing our conclusion that the subjectivity transformation is a robust default behavior rather than a prompt-controllable one.

I Full Generation-Quality Results

Table 7 reports all six automatic NLP similarity metrics (ROUGE-1/2/L, BLEU, BERTScore-F_1, and MAE_{subj}) for every one of the 24 vanilla conditions. Consistent with the main text, scores are tightly clustered: no model wins across metrics, and the best prompt configuration differs by metric and model, underscoring that meta-review quality is largely insensitive to these choices.

Table 8 isolates MAE_{subj} and pairs each base prompt with its subjectivity-aware (`-sub`) counterpart to make the conditioning effect explicit. Conditioning helps in 8 of 12 cells, and the benefit is largest at L1 (the simplest prompt); as prompts become more detailed, the additional subjectivity statistics become redundant and the effect shrinks or reverses (e.g., Qwen2.5-3B at L3, $+0.0050$). This reversal likely reflects instruction redundancy: the L3 prompt already heavily constrains generation behavior through detailed sub-task enumeration, leaving little room for the subjectivity signal and occasionally causing distraction.

J Full Subjectivity-Transformation Fits

Table 9 reports the subjectivity transformation, represented by the Δ-vs-σ linear fit (slope, Pearson r, sample size n) for all 24 vanilla conditions and the human reference. All correlations are negative and highly significant (every $p < 10^{-30}$), with $|r|$ ranging from 0.69 to 0.84. The strength of transformation is broadly stable across prompt levels and conditioning variants within each model, implying that the regression-to-the-mean behavior is a robust latent rather than prompt-driven. Figure 5 visualizes the same 24 fits as a grid.

K Multi-Venue Transformation Analysis

Figures 6 and 7 show the full Δ-vs-σ scatter plots for NeurIPS-2022 and ICLR-2022, respectively, complementing the ICLR-2019 plots in Figure 2. The regression-to-the-mean pattern is consistent across all three venues, with comparable slopes and correlations for both human meta-reviews and Llama-3-8B generations.

Held-out	BERT-large	RoBERTa-large	DeBERTa-v3-base	DistilBERT
books	0.663	0.659	0.589	0.637
electronics	0.686	0.665	0.662	0.634
movies	0.498	0.481	0.451	0.475
restaurants	0.720	0.721	0.741	0.765
tripadvisor	0.606	0.561	0.626	0.627
grocery	0.552	0.581	0.571	0.564
Avg	0.621	0.611	0.607	**0.617**

Table 4: Leave-one-domain-out (LODO) cross-domain Pearson r on SubjQA.

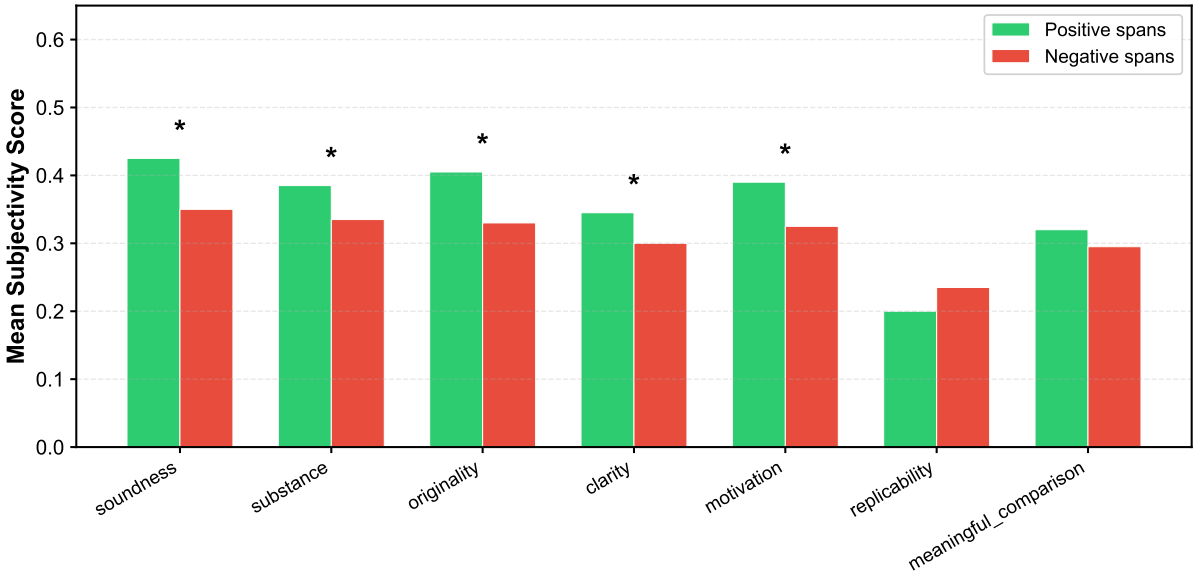


Figure 4: Mean subjectivity by review aspect and sentiment polarity (ASAP-Review). Positive-sentiment evaluative spans are the most subjective; summary spans the least. * = statistically significant ($p < 0.05$)

Condition	$\tau = 0.35$	$\tau = 0.40$	$\tau = 0.45$	$\tau = 0.50$	$\tau = 0.55$
Human	-0.65 (647)	-0.65 (597)	**-0.64** (535)	-0.67 (439)	-0.64 (346)
Gemma-2-2B-L1-base	-0.77 (749)	-0.77 (740)	**-0.77** (725)	-0.76 (680)	-0.76 (618)
Qwen2.5-3B-L1-base	-0.75 (832)	-0.75 (811)	**-0.75** (775)	-0.75 (723)	-0.74 (642)
Mistral-7B-L1-base	-0.74 (1330)	-0.74 (1319)	**-0.73** (1301)	-0.73 (1254)	-0.72 (1164)
Llama-3-8B-L1-base	-0.72 (784)	-0.72 (777)	**-0.72** (764)	-0.73 (726)	-0.73 (678)

Table 5: Threshold sensitivity: Pearson r (and aligned sentence count n) for Δ vs. σ at varying cosine similarity thresholds.

Condition	MAE_{base}	MAE_{subj}	Δ	95% CI	Cohen's d
Gemma-2-2B-L1	0.0518	0.0485	−0.0034	[−0.0092, +0.0026]	−0.101
Gemma-2-2B-L2	0.0499	0.0505	+0.0007	[−0.0045, +0.0058]	+0.022
Gemma-2-2B-L3	0.0508	0.0501	−0.0007	[−0.0058, +0.0048]	−0.022
Qwen2.5-3B-L1	0.0514	0.0478	−0.0037	[−0.0091, +0.0017]	−0.121
Qwen2.5-3B-L2	0.0525	0.0491	−0.0034	[−0.0086, +0.0019]	−0.114
Qwen2.5-3B-L3	0.0444	0.0500	+0.0056*	[+0.0009, +0.0103]	+0.208
Mistral-7B-L1	0.0497	0.0476	−0.0021	[−0.0067, +0.0025]	−0.082
Mistral-7B-L2	0.0505	0.0459	−0.0046†	[−0.0092, −0.0001]	−0.182
Mistral-7B-L3	0.0497	0.0470	−0.0027	[−0.0067, +0.0013]	−0.118
Llama-3-8B-L1	0.0495	0.0513	+0.0018	[−0.0037, +0.0073]	+0.058
Llama-3-8B-L2	0.0467	0.0472	+0.0005	[−0.0040, +0.0052]	+0.020
Llama-3-8B-L3	0.0522	0.0482	−0.0040	[−0.0096, +0.0015]	−0.125

Table 6: Paired bootstrap 95% CIs ($B = 10{,}000$) on MAE_{subj} differences (sub − base). *CI excludes zero (conditioning *worsens*). †CI borderline excludes zero.

Model	Prompt	R-1↑	R-2↑	R-L↑	BLEU↑	BERT↑	MAE↓
Gemma-2-2B	L1-base	0.294	0.057	0.166	0.0166	0.783	0.052
	L1-sub	0.299	0.060	**0.170**	0.0162	0.786	**0.049**
	L2-base	0.297	0.061	0.167	0.0164	0.786	0.050
	L2-sub	0.291	0.061	0.164	0.0178	0.784	0.050
	L3-base	**0.303**	0.063	0.169	0.0175	**0.787**	0.050
	L3-sub	0.297	**0.064**	0.167	**0.0179**	0.786	0.050
Qwen2.5-3B	L1-base	**0.286**	0.047	0.158	0.0143	0.783	0.052
	L1-sub	0.285	0.048	**0.159**	0.0138	**0.785**	0.047
	L2-base	0.281	**0.054**	0.156	0.0146	0.781	0.054
	L2-sub	0.279	0.052	0.153	**0.0154**	0.779	0.050
	L3-base	0.271	0.053	0.144	0.0140	0.777	**0.045**
	L3-sub	0.276	0.052	0.148	0.0136	0.779	0.050
Mistral-7B	L1-base	0.276	0.058	0.148	0.0154	0.775	0.049
	L1-sub	**0.283**	0.058	**0.151**	0.0173	**0.777**	**0.045**
	L2-base	0.277	0.061	0.150	0.0172	0.777	0.049
	L2-sub	0.266	0.058	0.145	0.0162	0.771	0.046
	L3-base	0.270	**0.062**	0.147	0.0169	0.776	0.048
	L3-sub	0.271	0.062	0.147	**0.0174**	0.774	0.046
Llama-3-8B	L1-base	**0.313**	0.068	0.171	0.0198	0.787	0.049
	L1-sub	**0.313**	**0.069**	**0.174**	0.0198	**0.787**	0.051
	L2-base	0.282	0.062	0.154	0.0185	0.780	**0.047**
	L2-sub	0.277	0.064	0.152	0.0186	0.778	0.048
	L3-base	0.280	0.068	0.155	**0.0220**	0.781	0.053
	L3-sub	0.283	0.067	0.158	0.0201	0.780	0.050

Table 7: Full meta-review generation quality across all 24 vanilla conditions (4 models $\times$ 3 prompt levels $\times$ 2 conditioning variants). R-1/R-2/R-L = ROUGE F_1; BERT = BERTScore-F_1; MAE = subjectivity MAE. Arrows indicate metric direction; bold = best per model. Scores are tightly clustered across all conditions, with no model or prompt dominating.

Model	Level	Base	Sub	Δ
Gemma-2-2B	L1	0.0521	**0.0487**	-0.0034
	L2	0.0501	**0.0495**	-0.0006
	L3	0.0499	**0.0496**	-0.0003
Qwen2.5-3B	L1	0.0519	**0.0469**	-0.0050
	L2	0.0540	**0.0502**	-0.0038
	L3	**0.0451**	0.0501	$+0.0050$
Mistral-7B	L1	0.0492	**0.0451**	-0.0041
	L2	0.0494	**0.0460**	-0.0034
	L3	0.0483	**0.0465**	-0.0018
Llama-3-8B	L1	**0.0487**	0.0511	$+0.0024$
	L2	**0.0470**	0.0484	$+0.0014$
	L3	0.0530	**0.0496**	-0.0034

Table 8: $\mathrm{MAE}_{\mathrm{subj}}$ (lower is better) across prompt levels L1–L3, base vs. subjectivity-aware. Bold = better per row; Δ = Sub$-$Base. Conditioning helps in $8/12$ cells, strongest at L1.

Model	Prompt Level	Prompt	n	Slope	r
Human	–	–	535	-0.79	-0.64
Gemma-2-2B	L1	base	725	-0.82	-0.77
	L1	sub	699	-0.81	-0.80
	L2	base	770	-0.92	-0.84
	L2	sub	879	-0.89	-0.83
	L3	base	873	-0.90	-0.82
	L3	sub	911	-0.84	-0.79
Qwen2.5-3B	L1	base	775	-0.81	-0.75
	L1	sub	768	-0.88	-0.80
	L2	base	904	-0.91	-0.80
	L2	sub	1011	-0.89	-0.80
	L3	base	1283	-0.94	-0.80
	L3	sub	1235	-0.88	-0.78
Mistral-7B	L1	base	1301	-0.76	-0.73
	L1	sub	1240	-0.77	-0.75
	L2	base	1319	-0.82	-0.75
	L2	sub	1391	-0.84	-0.74
	L3	base	1387	-0.79	-0.73
	L3	sub	1371	-0.82	-0.75
Llama-3-8B	L1	base	764	-0.74	-0.72
	L1	sub	786	-0.72	-0.71
	L2	base	1181	-0.89	-0.73
	L2	sub	1198	-0.92	-0.75
	L3	base	1304	-0.79	-0.69
	L3	sub	1250	-0.79	-0.70

Table 9: Subjectivity-transformation fit (Δ vs. σ) for all 24 vanilla conditions and the human reference. All r are negative and significant ($p < 10^{-30}$).

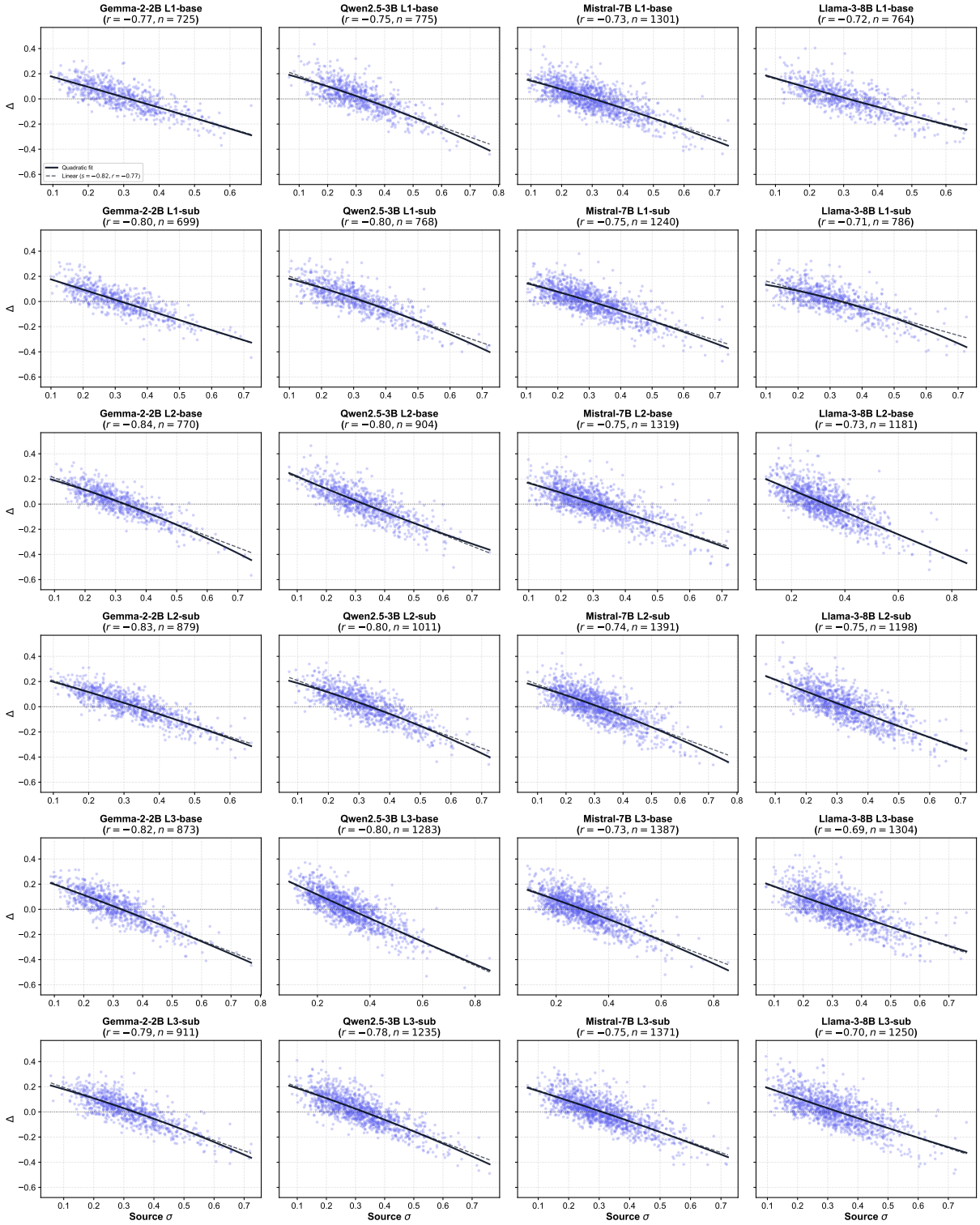


Figure 5: Subjectivity shift Δ vs. source subjectivity σ for all 24 vanilla conditions (rows: prompt level $\times$ conditioning; columns: model). Dark curve = quadratic fit, dashed = linear fit.

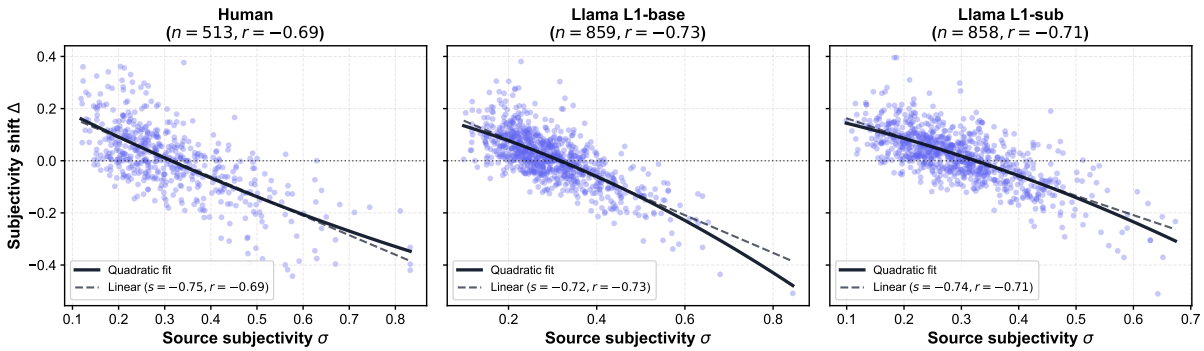


Figure 6: Subjectivity shift Δ vs. source subjectivity σ for NeurIPS-2022: human meta-reviews (left), Llama-3-8B L1-base (center), and L1-sub (right).

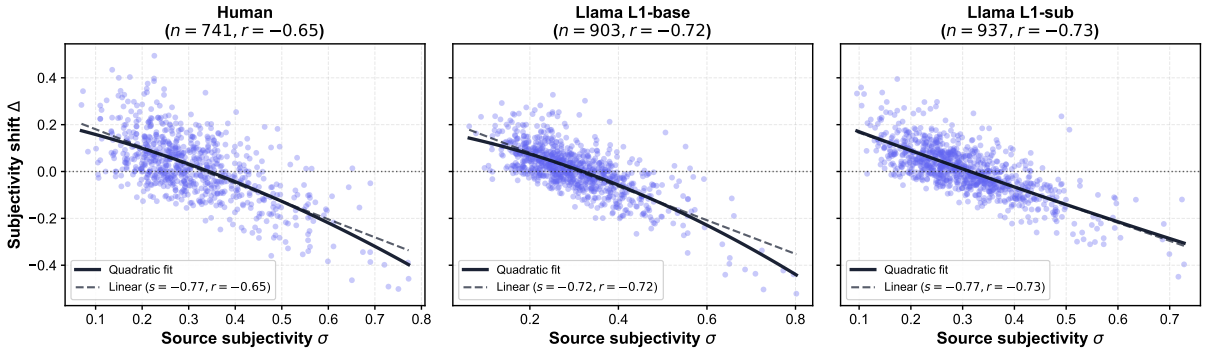


Figure 7: Subjectivity shift Δ vs. source subjectivity σ for ICLR-2022: human meta-reviews (left), Llama-3-8B L1-base (center), and L1-sub (right).