Verbalized Confidence Now Works: More Robust than Verdicts for Contemporary LLM Judges

Anonymous ACL submission

Abstract

Prior work has largely favored token logprobs over verbalized confidence for LLM-as-a-Judge soft scoring, finding verbal confidence coarse, overconfident, and weakly discriminative. We revisit this recommendation across 18 LLMs on SummEval and AggreFact, spanning pre- and post-2025 models. We find a *compatibility shift*: a logprob-free verbalized-confidence recipe improves calibration, score-distribution spread, and robustness to task subjectivity, while contemporary flagships absorb the added prompt structure with little or no verdict-accuracy cost. On GPT-family flagships, verbalized confidence overtakes logprob-based G-Eval as the more subjectivity-robust soft signal; on AggreFact, newer GPT, Claude, and Gemini models preserve more of their latent faithfulness signal in their binary verdicts under prompt stress. These results suggest that verbalized confidence should no longer be treated as merely a weaker substitute for logprobs: with the right prompt and model generation, it is a practical soft-scoring interface for contemporary LLM judges.

1 Introduction

LLM-as-a-Judge protocols turn a generative model into a scorer for another model's outputs. A central methodological question has been which *soft signal* the judge should expose: token log-probabilities (logprobs) over the verdict tokens, or a verbalized confidence the judge produces in natural language. Pre-2025 evidence weighed heavily toward logprobs. G-Eval added probability-weighted summation over output score tokens after observing that LLM Likert outputs cluster on a few discrete integer values, weakening rank-correlation metrics against human ratings such as Kendall's τ (Liu et al., 2023). Relatedly, Wang et al. (2025) further showed that the G-Eval style soft scoring protocol beats greedy verbal extraction in pointwise, pairwise, and listwise settings. Large-scale benchmarks reported that LLMs are systematically overconfident when verbalizing (Xiong et al., 2024) and recent work has documented that verbal scores cluster on a handful of round-number values, collapsing their discriminative power (Dai, 2026). The dominant practical recommendation has been "prefer logprobs."

Most of this evidence, however, comes from pre-2025 chat models, and we find that the comparison looks different on contemporary flagships. As frontier vendors moved through post-2025 release waves, two changes pull in opposite directions. On the *logprob* side, access has become less stable: closed-model APIs increasingly restrict logprob exposure or condition it on non-reasoning mode (Appendix A). On the *verbal* side, recent mechanistic evidence on open-weight models suggests verbal confidence is not simply a logprob readout but reflects a richer internal assessment (Kumaran et al., 2026). This result has not been verified on closed-source flagships. This motivates our investigation whether a prompt recipe can render the verbal signal sufficiently informative to match or exceed logprob-based soft scoring from various aspects.

This work tests this question empirically across an 18-model sweep, spanning GPT, Claude, and Gemini families, deliberately sampling both pre-2025 chat checkpoints and post-2025 flagships. We document a **compatibility shift**, a misalignment between an evaluation protocol and the LLM cohort it relies on, where a recommendation that worked on one no longer holds on another. Concretely: verbalized confidence becomes more subjectivity-robust than logprob-based G-Eval on GPT-

family flagships; our prompt recipe (overconfidence advisory, confidence rubric, structured self-debate) significantly improves both calibration error and score-distribution spread, and post-2025 flagships realize these gains without the classification-accuracy cost they impose on older models.

This compatibility shift is invisible under standard reporting practice. It became legible to us when we read four dimensions together: classification accuracy on the verdict, latent recoverable signal from the soft score, robustness under prompt stress, and the shape and accuracy of the soft score itself. We use that combination throughout the paper; we are not claiming it is the only useful framing, only that the shift requires more than accuracy reporting to surface. Alongside this analysis, we provide a prompt recipe adapted to contemporary models.

Main claims and contributions.

- **A logprob-free prompt recipe for LLM-as-a-Judge soft scoring** (§2.1) that combines an overconfidence advisory, a confidence rubric, and structured self-debate. The recipe yields probability scores that are less coupled to task subjectivity (§4.2) and better calibrated with broader score-distribution spread (§4.1) than a verbalized-confidence rubric baseline.
- **A compatibility shift between soft-signal type and model generation** (§5). Pre-2025 evidence preferred logprob-based scoring; on contemporary GPT-family flagships, verbalized confidence overtakes logprob-based G-Eval as the more subjectivity-robust signal, and post-2025 flagships across vendors realize the recipe's gains while older chat models pay a significant classification-accuracy cost for the same recipe.
- **A measurement framework that surfaces the shift** (§6.1). We show that classification accuracy alone hides the latent signal in the soft score, and calibration error alone hides distributional collapse; reading verdict accuracy, oracle BA, oracle-verdict gap, debate stress, calibration error, and score-distribution spread together makes both effects legible.

The shift is sharpest on closed-source flagships; the open-weight checkpoints in our study show partial signs of the same pattern, and we discuss this evidence and what it implies for mechanism in Section 6.2.

2 Method

2.1 Proposed soft scoring protocols

We propose to use pure verbal signal to construct probability scores from LLM responses. The goal is to extract a soft score that (1) discriminates strongly between True and False on a given quality axis and (2) is well calibrated. Specifically, we constrain the use of LLMs to:

1. A single LLM call per judging. This excludes multi-agent setups.
2. Pointwise binary classification. This bridges the gap that traditional machine learning is usually capable of providing quality scores, for example between 0 and 1.0. During the LLM era, while LLMs can be used to create ad-hoc binary classifiers through prompting, they usually lack the granularity of soft scoring as traditional machine learning models provide.

Logprob-based methods such as G-Eval (Liu et al., 2023) and the judgment-distribution line (Wang et al., 2025) are established alternatives, but logprob access has become less reliable across closed-model APIs (Appendix A); we therefore focus on a logprob-free protocol.

We therefore propose two Verbalized Confidence Protocols, **Ours-FreeReason** and **Ours**, with two baselines: the rubric-only **Rubric** variant and logprob-based **G-Eval** as comparison anchors (Figure 1).

Prompt components. We view judge prompts as compositions of reusable components and clarify some critical components below:

- **Reasoning request:** a request for free-form reasoning before the verdict.
- **Self-debate request:** a request to articulate reasons supporting True, reasons supporting False, and a conclusion.
- **Advisory:** behavioral guidance, e.g., a warning against overconfidence.

The specific prompt templates are documented in Appendix H due to the length of the

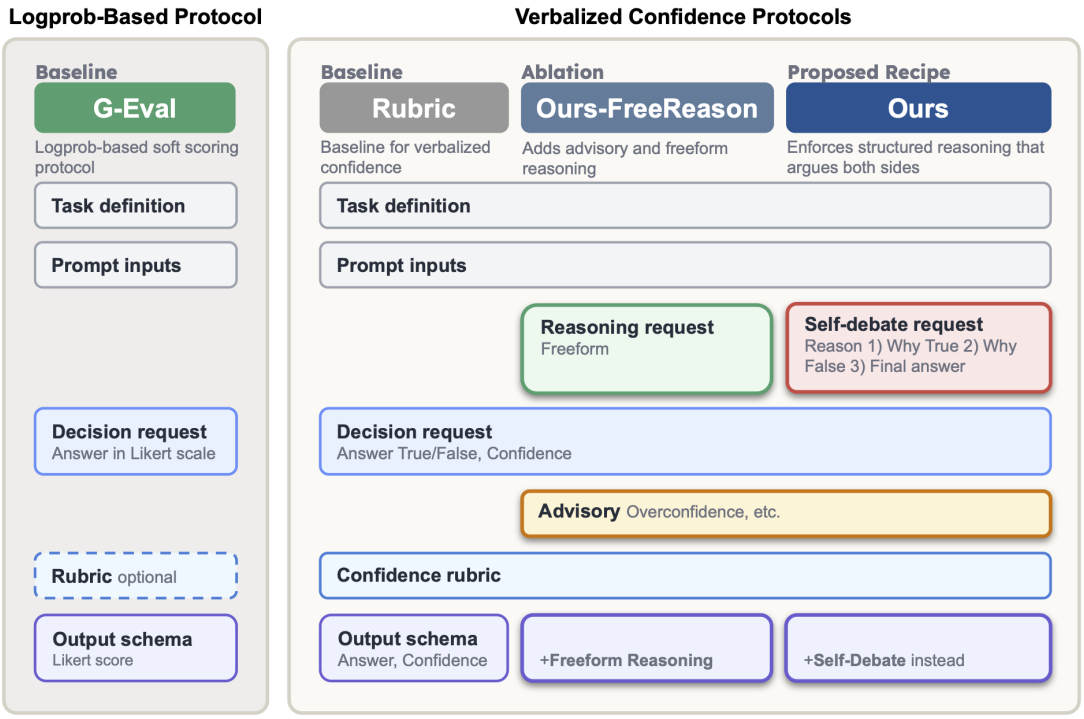


Figure 1: Prompt protocols used in this paper. We compare logprob-based G-Eval with three logprob-free verbalized-confidence protocols. Rubric is a baseline asking only for a verdict and a rubric-anchored confidence score. Ours-FreeReason adds advisory text and a free-form reasoning step before the verdict. Ours, our recommended recipe, replaces free-form reasoning with structured self-debate that requires the model to articulate support for both stances before committing. The key difference is that Ours turns confidence from a post-hoc scalar into a judgment conditioned on explicit competing evidence, which produces the strongest soft-score signal in our experiments.

full text.

Prompt design.

- **Rubric (baseline):** a logprob-free verbalized-confidence protocol that asks the model for a binary judgment and an explicit confidence score under a confidence rubric.
- **Ours-FreeReason (ablation):** the Rubric baseline augmented with advisory text discouraging overconfidence and with free-form reasoning before the verdict. The model is asked for explicit, unconstrained reasoning but is not required to consider both stances. Output is arranged as Reasoning, Answer, and Confidence score, in order.
- **Ours:** the full recipe, which augments Rubric with advisory text and structured self-debate, requiring the model to produce reasons supporting True, reasons supporting False, and a conclusion before emitting the verdict and confidence. Output is arranged as Debate, Answer, and Confidence score, in order.

2.2 Metrics

- **Verdict Balanced Accuracy (Verdict BA).** Balanced accuracy of the judge's hard binary verdict (true/false) against the gold binary label.
- **Oracle Balanced Accuracy (oracle BA).** Balanced accuracy of the best binary classifier recoverable from the judge's confidence signal when the decision threshold is optimized directly on the evaluation set. We treat oracle BA as the ideal attainable value.
- **Oracle-Verdict Gap (OV Gap).** Defined as the difference between oracle BA and verdict BA, used as a diagnostic for how much discriminative signal the binary verdict discards. By construction, OV Gap is non-negative, and larger values indicate more latent signal in the confidence that the binary verdict fails to capture.
- **Debate Stress.** Defined as the difference between verdict BA of Ours-FreeReason

and verdict BA of Ours. It measures the verdict-quality cost of structurally enforcing both-sides reasoning relative to leaving the reasoning channel freeform. Larger positive values indicate more degradation under the structural constraint; values near zero indicate that the model is robust to the forced self-debate.

- **Adaptive Expected Calibration Error (AECE).** Expected calibration error (Guo et al., 2017) computed with adaptive (equal-mass) bins (Nixon et al., 2019).
- **Spread.** The score spread component is defined as the Bhattacharyya coefficient between the empirical probability histogram and a uniform reference. Let $\hat{q} = (\hat{q}_1, \ldots, \hat{q}_B)$ be the empirical histogram of confidence scores over B equal-width bins on $[0, 1]$. We define

$$\mathrm{Spread}(\hat{q}) = \frac{1}{\sqrt{B}} \sum_{b=1}^{B} \sqrt{\hat{q}_b}.$$

Bhattacharyya is preferred over entropy here because Spread is intended to measure closeness to the uniform distribution, not generic dispersion.

3 Experimental Setup

3.1 Tasks

We utilize two evaluation benchmarks:

SummEval (Fabbri et al., 2021) is a graded subjective evaluation: 100 documents each paired with 17 machine-generated summaries and expert-annotated 5-point Likert ratings on four dimensionss. **Consistency** asks whether the summary remains faithful to the transcript; **Coherence** asks whether the summary reads as a well-formed paragraph; and **Relevance** asks whether the summary captures the important content of the article while avoiding redundant or tangential details. We drop the Fluency dimension due to the reasons discussed in Appendix B.

Inter-annotator agreement (IAA) for each SummEval dimension is computed as Krippendorff's α on the expert 5-point Likert ratings, with lower α indicating greater task subjectivity: Consistency $\alpha = 0.80$ (relatively high agreement; least subjective), Coherence $\alpha = 0.55$ (moderate-to-low; noticeably more subjective), and Relevance $\alpha = 0.40$ (low; highly subjective / boundary-contested). We use IAA as a per-task subjectivity covariate in the slope analyses below.

AggreFact is an aggregation of binary factual-faithfulness subtasks. We select 9 out of 11 subtasks from AggreFact and sample 300 examples per subtask, totaling 2,700 examples; we defer the rationale for excluding AggreFact-XSum and WiCE to Appendix C.

3.2 Models

Our full sweep covers 18 LLMs spanning the GPT, Claude, and Gemini families plus open-weight checkpoints. Two cohorts are used. The **AggreFact cohort** (§4.1, §5.2) is a 10-model selection of closed-source flagships, grouped by training generation into *pre-2025* (GPT-4o, GPT-4.1, o1, sonnet-4, gemini-2.5-pro) and *post-2025* (GPT-5.2, GPT-5.4, sonnet-4.5, sonnet-4.6, gemini-3.1-pro). The **SummEval cohort** (§4.2, §5.1) extends the AggreFact cohort to 16 models by adding 5 open-weight models (gpt-oss-20b, gpt-oss-120b, llama-3.3-70b, mistral-large-3, gemma-4-26b) and 3 smaller closed models (gpt-4o-mini, gpt-4.1-mini, gpt-4.1-nano), in order to probe whether subjectivity robustness extends beyond flagship scale; gemini-3.1-pro and sonnet-4 are excluded due to access constraints at experiment time.

For the GPT-family G-Eval comparison (which requires logprob access), we restrict to non-reasoning OpenAI flagships (gpt-4o, gpt-4.1, gpt-5.2, gpt-5.4), since the OpenAI API does not emit logprobs in reasoning mode.

4 Prompt Effect: Our Recipe Improves Soft Scoring Quality

This section establishes the prompt effect: across both datasets and both signal types, our recipe improves soft scoring quality. On AggreFact, the recipe lowers calibration error and broadens the score distribution (§4.1). On SummEval, the recipe further makes the probability signal more robust to task subjectivity (§4.2).

4.1 Calibration error and score spread improve under our recipe

Figure 2 shows how the prompt recipe changes Sonnet 4.5's probability channel on AggreFact. Relative to Rubric, Ours-FreeReason

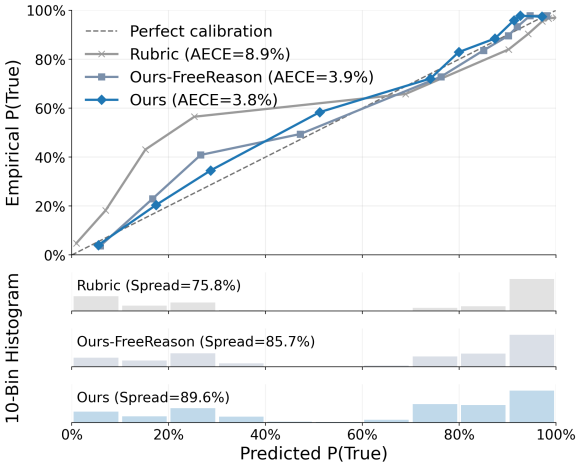


Figure 2: Calibration curves and probability histograms for Sonnet 4.5 on AggreFact. The top panel plots overall calibration curves using 10 equal-mass bins; points with fewer than 27 examples are omitted. The lower panels show 10-bin histograms of predicted probabilities. Compared with RUBRIC, our prompts substantially reduce calibration error (AECE: from 8.9% to 3.8%) and make the probability distribution more evenly spread (Spread: from 75.8% to 89.6%).

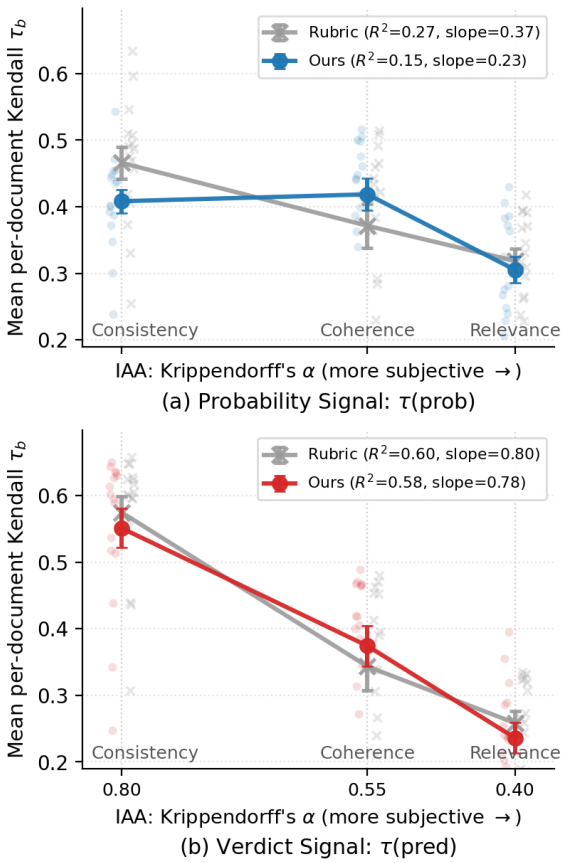


Figure 3: Probability signals in panel (a) are more robust to task subjectivity than verdict signals in panel (b). Smaller regression slopes and lower R^2 indicate weaker coupling to inter-annotator agreement (IAA) and thus a less confounded capacity measure. Faint markers show per-model values, jittered along the x-axis for visualization only. Relative to the baseline prompt, our prompt significantly flattens $\tau(\texttt{prob})$ (Δslope $= -0.143$, one-sided $p < 10^{-4}$), whereas the slope difference between the two prompts is not significant for $\tau(\texttt{pred})$ (two-sided $p = 0.46$).

sharply reduces calibration error, and OURS preserves this gain while further broadening the score distribution. The histogram rows show that our prompts make the predicted probabilities less collapsed and more evenly distributed across the probability range, increasing Spread from 75.8% to 89.6%. The calibration curves also show that the rubric baseline is especially overconfident in low-probability regions, where false predictions are assigned probabilities that are too extreme relative to their empirical truth rate. Our prompts tame this overconfident pattern, especially on the false side, bringing the curve closer to the diagonal while maintaining broader use of the probability scale.

The Sonnet 4.5 case is illustrative; the full per-model calibration-error and spread results are reported in §5.2.

4.2 Probability signals are more robust to task subjectivity

Turning to SummEval, we computed Kendall's τ_b between the judgments from our 16 tested LLMs and the expert-annotated Likert ratings, separately for binary verdicts and probability scores for each prompt. We then averaged these per-model correlations across the models, denoting the resulting means as $\tau(\texttt{pred})$ and $\tau(\texttt{prob})$[1], respectively, plotted in Figure 3, against IAA on the x-axis. We compared two prompt configurations: Rubric and Ours.

The probability signals in panel (a) exhibit flatter slopes than the binary verdict signals in panel (b), indicating that probability signals are less correlated with IAA (one-sided $p < 10^{-5}$). In other words, robustness to task subjectivity is visible in the soft-score signal but largely hidden if one looks only at binary verdicts. Furthermore, the overconfidence advisory and the self-debate structure utilized in our prompt are effective mechanisms for further

[1]The subscript b is dropped for brevity.

flattening the slope and lowering R^2 (one-sided $p < 10^{-4}$) relative to the Rubric baseline. The p-values are computed by the paired model-cluster bootstrap described in Appendix D.

We argue that probability scores are a better measurement instrument for latent judge capacity. The lower IAA does not mean that labels are simply corrupted: disagreement can come from rubric ambiguity, multiple valid perspectives, or imperfect annotator calibration. The SummEval annotation process recruited the same expert pool for all dimensions through the same training, so we treat the IAA gradient as genuine subjectivity rather than noise.

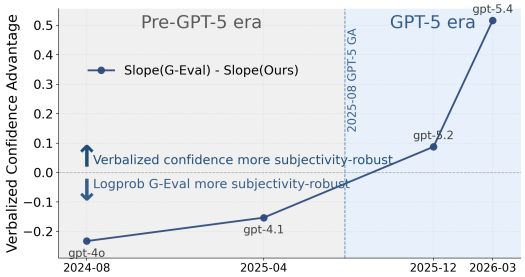
5 Model Generation Effect: The Compatibility Shift

This section establishes the model-generation effect: post-2025 flagships interact with our prompt recipe differently from pre-2025 chat models, and the difference shows up on both datasets. On SummEval, GPT-family flagships realign from logprob-based G-Eval toward verbalized confidence (§5.1); on AggreFact, the Oracle-Verdict Gap and Debate Stress shrink toward zero on the post-2025 cohort while older models continue to incur a verdict-quality penalty (§5.2).

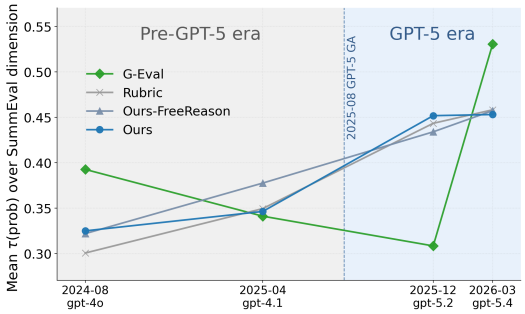
5.1 Verbalized confidence overtakes logprob G-Eval on GPT-family flagships

We turn to the model-generation effect on robustness between verbalized confidence and the traditional G-Eval-style logprob soft scoring. We compare four judging protocols defined in Section 2: G-Eval, Rubric, Ours-FreeReason, and Ours, on the GPT-family non-reasoning flagships (gpt-4o, gpt-4.1, gpt-5.2, gpt-5.4). For each task and protocol, we compute mean per-document Kendall τ_b between model scores and expert ratings across candidate summaries, and quantify subjectivity robustness as the slope of $\tau(\texttt{prob})$ against task IAA α; lower slopes indicate less sensitivity to task subjectivity.

From the perspective of subjectivity robustness, verbalized-confidence protocols become increasingly aligned with newer GPT generations: the slope difference between G-Eval and Ours flips from negative (pre-GPT-5) to positive (GPT-5 era). The same pattern appears in judging capacity: verbalized confidence improves steadily across the GPT family, whereas logprob-based G-Eval is generation-unstable and does not exhibit the same reliable upward trend. Verbalized confidence is therefore better aligned with the behavior of current GPT-family models than logprob-based soft scoring, which is one practical reason to recommend Ours as the forward-compatible LLM-as-a-judge protocol.



(a) Slope-difference between G-Eval and Ours.



(b) Mean $\tau(\texttt{prob})$ trend.

Figure 4: Compatibility shift across GPT releases on SummEval. (a) GPT-5 era models favor verbalized-confidence judging over logprob-based G-Eval for subjectivity-robustness. (b) Verbalized-confidence protocols (Rubric, Ours-FreeReason, Ours) trend upward monotonically across GPT generations, while logprob-based G-Eval does not. Together, the panels show that verbalized-confidence judging overtakes logprob-based G-Eval as the more compatible soft-signal interface for contemporary GPT-family flagships.

5.2 Oracle-Verdict Gap and Debate Stress shrink post-2025

We now ask whether the model's emitted binary verdict is as informative as the confidence signal verbalized in the same response. Unlike SummEval, AggreFact is a binary faithfulness task, so this setting lets us directly compare the quality of the hard verdict against the quality of a thresholded confidence score. We compare three verbal protocols: Rubric, Ours-FreeReason, and Ours, using Verdict BA,

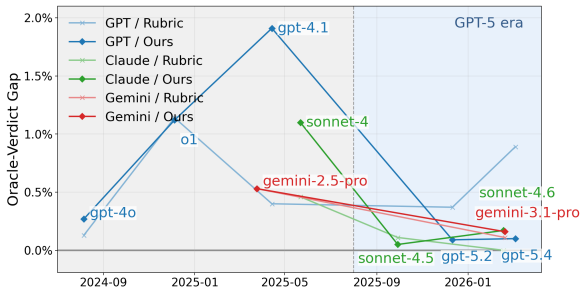


Figure 5: Oracle-Verdict Gap over model generations. Lower is better: a smaller Oracle-Verdict Gap means the model's binary verdict better captures the faithfulness signal present in its verbalized confidence. Under Ours, this gap shrinks markedly for newer models, approaching zero after mid-2025: even under a more demanding debate prompt, the verdicts of newer flagships retain more of the model's latent faithfulness signal instead of leaving it recoverable only from confidence.

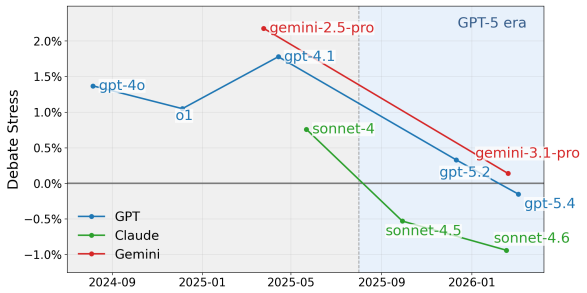


Figure 6: Debate Stress over model generations. Lower is better: a smaller Debate Stress means that adding debate causes less damage to verdict BA, with zero indicating full robustness and negative values indicating gains from debate. Older GPT and Gemini models show substantial positive stress, whereas newer post-2025 models move much closer to zero; later Claude models even gain verdict BA. Debate fragility is primarily a property of older model generations, while newer flagships are markedly more robust under debate.

oracle BA, the Oracle-Verdict Gap, and Debate Stress, defined in Section 2.2.

Table 1 reports Verdict BA and calibration-quality gains for the GPT, Claude, and Gemini families under a paired bootstrap. Two patterns emerge. First, across nearly all models, `Ours` yields large gains in calibration improvement, especially for score-distribution spread. This is the prompt effect documented in §4.1, applied at scale. Second, verdict-quality trade-offs are generation-dependent: older models (`o1`, `gpt-4.1`, `gemini-2.5-pro`) incur statistically significant BA drops, whereas newer flagships (`gpt-5.2`, `gpt-5.4`, `sonnet-4.6`, `gemini-3.1-pro`) show much smaller penalties or none at all.

The model-generation effect appears most clearly in the two prompt-stress diagnostics. Newer flagship models preserve strong verdict and oracle performance while reducing both the Oracle-Verdict Gap (Figure 5) and Debate Stress (Figure 6). In other words, less latent faithfulness signal remains recoverable only from confidence, and debate causes less damage to the emitted verdict. The full per-family comparison is in Appendix G.

6 Discussion

6.1 Synthesis and benchmark implications

The prompt and model effects are complementary. Our recipe improves the probability channel across settings: on SummEval it makes τ(`prob`) more robust to task subjectivity, and on AggreFact it cuts calibration error while broadening the score distribution. Post-2025 flagships then convert this richer signal into stable verdicts, whereas older models often pay a verdict-accuracy penalty for the same calibration gains. This is why accuracy-only reporting misses the shift: ΔVerdict BA looks flat on newer models, while ΔAECE alone hides whether the verdict still preserves the latent signal.

For benchmark practice, this argues for reporting verdict BA, oracle BA, Oracle-Verdict Gap, Debate Stress, AECE, and Spread together. The combined view reveals which models trade verdict quality for calibration and which absorb the richer prompt without cost. As logprob access becomes less stable in production APIs (Appendix A), a logprob-free verbal-confidence protocol that remains robust across deployed models is the more portable soft-scoring interface.

6.2 Why this shift: a call for open recipes

One plausible mechanism is verifiable-reward instruction-following (IF-RLVR), which entered publicly documented training recipes during this window (Lambert et al., 2024; Allen Institute for AI, 2025). However, our open-recipe checks in Appendix E only partially match the

Model	BA ↑			AECE ↓			Spread ↑		
	Rubric	Ours	Δ ↑	Rubric	Ours	Δ ↓	Rubric	Ours	Δ ↑
gpt-4o	81.3%	80.2%	-1.1	8.0%	4.1%	-3.9*	66.0%	88.8%	+22.8*
o1	80.6%	79.2%	-1.4*	11.0%	5.8%	-5.2*	48.7%	71.9%	+23.2*
gpt-4.1	81.7%	79.3%	-2.4*	11.2%	4.4%	-6.8*	53.5%	86.9%	+33.4*
gpt-5.2†	82.4%	83.0%	+0.6	9.7%	7.6%	-2.2*	66.9%	83.6%	+16.6*
gpt-5.4†	81.3%	81.9%	+0.6	12.7%	9.1%	-3.6*	67.2%	81.0%	+13.7*
sonnet-4	81.3%	80.1%	-1.2*	8.4%	5.9%	-2.4*	71.3%	81.5%	+10.2*
sonnet-4.5†	80.9%	82.0%	+1.1	9.1%	3.3%	-5.8*	75.8%	89.6%	+13.9*
sonnet-4.6†	81.5%	82.7%	+1.2*	10.2%	8.1%	-2.2*	81.9%	86.4%	+4.4*
gemini-2.5-pro	80.7%	78.0%	-2.7*	17.7%	13.5%	-4.2*	45.8%	66.3%	+20.5*
gemini-3.1-pro†	83.3%	83.0%	-0.3	13.1%	11.2%	-1.8*	52.5%	64.2%	+11.6*

Table 1: AggreFact verdict accuracy and probability-signal quality under Rubric and Ours. Asterisks (*) denote statistically significant in paired bootstrapping. Daggers (†) denote post-2025 flagship models. Ours consistently reduces calibration error and broadens use of the probability scale, while verdict-accuracy costs are concentrated in older models.

closed-flagship pattern on Oracle-Verdict Gap or Debate Stress, so the mechanism remains a target for future open-recipe studies.

7 Related Work

Soft scoring and verbalized confidence. LLM-as-judge is well studied (Zheng et al., 2023; Liu et al., 2023), with AggreFact (Tang et al., 2023a) and SummEval (Fabbri et al., 2021) as standard faithfulness testbeds. G-Eval (Liu et al., 2023) and later judgment-distribution methods (Wang et al., 2025) use token logprobs to recover soft signal beyond greedy outputs. Prior evidence on verbalized confidence is mixed: it can be useful on QA correctness (Kadavath et al., 2022; Tian et al., 2023), but is often overconfident or collapsed onto round numbers (Xiong et al., 2024; Dai, 2026). Mechanistic work suggests verbal confidence can reflect internal answer-quality assessments beyond logprob readout (Kumaran et al., 2026); we show this channel is practically useful on contemporary closed-source flagships.

Self-debate and calibration. Unlike multi-call debate or aggregation systems (Du et al., 2024; Liang et al., 2024; Khan et al., 2024; Jiang et al., 2023; Wang et al., 2024), our self-debate is a single-call prompt constraint. For soft-score quality, we pair adaptive ECE (Guo et al., 2017; Nixon et al., 2019) with Bhattacharyya spread against a uniform reference, since ECE alone can miss the distributional collapse common in verbalized scores (Dai, 2026); this reference-to-uniform view follows evenness work that defines spread by similarity to the perfectly even vector (Chao and Ricotta, 2019; Gregorius and Gillet, 2022; Pierson, 2021).

8 Conclusion

We revisit the pre-2025 recommendation to prefer logprob-based soft scoring for LLM-as-a-Judge and find that it no longer holds uniformly. Across SummEval and AggreFact, our logprob-free verbalized-confidence recipe improves calibration, score spread, and robustness to task subjectivity, while post-2025 flagships absorb the added prompt structure with little or no verdict-accuracy cost. The practical takeaway is to report hard verdicts and soft-score diagnostics together, and to treat prompt–model compatibility as a moving target rather than a fixed property of judge protocols.

Limitations

This work studies single-call, pointwise LLM-as-judge protocols and does not cover pairwise judging, multi-call multi-agent debate, or ensemble aggregation. Our evaluation focuses on flagship LLMs from the GPT, Claude, and Gemini families, with smaller and earlier models serving only as generation-effect controls; results may not transfer to small or open-weight models. The two task settings (SummEval and a downsampled AggreFact) cover graded subjective evaluation and binary factual faithfulness but are not exhaustive of the LLM-as-judge landscape. We do not study fine-tuning or post-training interventions; all comparisons

are inference-time prompt variations. Our outcome variables measure latent judge signal and prompt robustness, not end-to-end deployment behavior such as cost, latency, or downstream pipeline effects.

Ethical Considerations

LLMs prompted as evaluators power downstream model selection in research benchmarks and increasingly in agentic feedback loops; miscalibrated or insufficiently discriminative model outputs drive wrong-model-wins outcomes that propagate. Our paper argues for adopting a richer reporting convention that surfaces calibration quality and verdict robustness alongside raw verdict accuracy, which directly reduces this risk on contemporary LLMs. We do not introduce new training data or models; all evaluation is on existing benchmarks (AggreFact, SummEval) under their published licenses.

References

Allen Institute for AI. 2025. Olmo-3 technical report. *arXiv preprint arXiv:2512.13961*.

Manik Bhandari, Pranav Narayan Gour, Atabak Ashfaq, Pengfei Liu, and Graham Neubig. 2020. Re-evaluating evaluation in text summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9347–9359.

Anne Chao and Carlo Ricotta. 2019. Quantifying evenness and linking it to diversity, beta diversity, and similarity. *Ecology*, 100(12):e02852.

Yuyang Dai. 2026. Rescaling confidence: What scale design reveals about LLM metacognition. *arXiv preprint arXiv:2603.09309*.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, and Igor Mordatch. 2024. Improving factuality and reasoning in language models through multiagent debate. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*.

Alexander R. Fabbri, Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev. 2021. SummEval: Re-evaluating summarization evaluation. *Transactions of the Association for Computational Linguistics (TACL)*.

Hans-Rolf Gregorius and Elizabeth M. Gillet. 2022. The concept of evenness/unevenness: Less evenness or more unevenness? *Acta Biotheoretica*, 70(1):3.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. 2017. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, pages 1321–1330.

Dongfu Jiang, Xiang Ren, and Bill Yuchen Lin. 2023. LLM-Blender: Ensembling large language models with pairwise ranking and generative fusion. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL)*.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, and 1 others. 2022. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*.

Ryo Kamoi, Tanya Goyal, Juan Diego Rodriguez, and Greg Durrett. 2023. WiCE: Real-world entailment for claims in Wikipedia. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 7561–7583, Singapore. Association for Computational Linguistics.

Akbir Khan, John Hughes, Dan Valentine, Laura Ruis, Kshitij Sachan, Ansh Radhakrishnan, Edward Grefenstette, Samuel R. Bowman, Tim Rocktäschel, and Ethan Perez. 2024. Debating with more persuasive LLMs leads to more truthful answers. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*.

Dharshan Kumaran, Arthur Conmy, Federico Barbero, Simon Osindero, Viorica Patraucean, and Petar Veličković. 2026. How do LLMs compute verbal confidence? *arXiv preprint arXiv:2603.17839*. Submitted to ICML 2026.

Nathan Lambert, Jacob Morrison, Valentina Pyatkin, Shengyi Huang, Hamish Ivison, Faeze Brahman, Lester James V. Miranda, Alisa Liu, Nouha Dziri, Shane Lyu, and 1 others. 2024. Tülu 3: Pushing frontiers in open language model post-training. *arXiv preprint arXiv:2411.15124*.

Tian Liang, Zhiwei He, Wenxiang Jiao, Xing Wang, Yan Wang, Rui Wang, Yujiu Yang, Zhaopeng Tu, and Shuming Shi. 2024. Encouraging divergent thinking in large language models through multi-agent debate. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-Eval: NLG evaluation using GPT-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Jeremy Nixon, Michael W. Dusenberry, Linchuan Zhang, Ghassen Jerfel, and Dustin Tran. 2019.

Measuring calibration in deep learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 38–41.

Kawika Pierson. 2021. Using geometry to rank evenness measures: Towards a deeper understanding of divergence. *Journal of Theoretical Biology*, 531:110902.

Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. Get to the point: Summarization with pointer-generator networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1073–1083.

Liyan Tang, Tanya Goyal, Alex Fabbri, Philippe Laban, Jiacheng Xu, Semih Yavuz, Wojciech Kryściński, Justin Rousseau, and Greg Durrett. 2023a. Understanding factual errors in summarization: Errors, summarizers, datasets, error detectors. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (ACL)*.

Liyan Tang, Tanya Goyal, Alex Fabbri, Philippe Laban, Jiacheng Xu, Semih Yavuz, Wojciech Kryscinski, Justin Rousseau, and Greg Durrett. 2023b. Understanding factual errors in summarization: Errors, summarizers, datasets, error detectors. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11626–11644, Toronto, Canada. Association for Computational Linguistics.

Katherine Tian, Eric Mitchell, Allan Zhou, Archit Sharma, Rafael Rafailov, Huaxiu Yao, Chelsea Finn, and Christopher D. Manning. 2023. Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Junlin Wang, Jue Wang, Ben Athiwaratkun, Ce Zhang, and James Zou. 2024. Mixture-of-agents enhances large language model capabilities. *arXiv preprint arXiv:2406.04692*.

Victor Wang, Michael J.Q. Zhang, and Eunsol Choi. 2025. Improving LLM-as-a-Judge inference with the judgment distribution. In *Findings of the Association for Computational Linguistics: EMNLP 2025*.

Miao Xiong, Zhiyuan Hu, Xinyang Lu, Yifei Li, Jie Fu, Junxian He, and Bryan Hooi. 2024. Can LLMs express their uncertainty? an empirical evaluation of confidence elicitation in LLMs. In *International Conference on Learning Representations (ICLR)*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-Judge with MT-Bench and chatbot arena. In *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*.

A Logprob Access Is No Longer a Stable Primitive

A clear trend emerges across recent closed-model APIs: token logprobs are becoming less uniformly available as a general-purpose evaluation signal. Earlier flagship chat models expose them cleanly, which makes logprob-based soft scoring straightforward. Newer model families increasingly restrict that path: support becomes conditional on model variant, reasoning mode, or serving configuration rather than being a stable default capability. Even where logprobs remain accessible, the maximum number of top tokens exposed has been reduced — the GPT family, for example, has dropped from 20 in GPT-4o/4.1 to 5 in GPT-5.2/5.4 (Table below). This makes logprob-based judge protocols harder to apply consistently across modern frontier models. By contrast, logprob-free confidence extraction remains broadly portable, which is why it is the more practical soft-signal interface for cross-model evaluation.

B SummEval Fluency Exclusion

We exclude the SummEval fluency dimension from our cross-dimension analysis because the dataset text and the fluency rubric are systematically misaligned. SummEval's 17 system summaries inherit the See et al. (2017) CNN/DailyMail preprocessing convention: they are lowercased, word-tokenized with Stanford CoreNLP, and contain whitespace-separated punctuation (e.g., `barcelona ' s`, `clip , which`). But Fabbri et al. (2021)'s fluency rubric asks raters to penalize "formatting problems, capitalization errors, or obviously ungrammatical sentences," so every summary violates the rubric by construction. Detokenization or truecasing would not solve this cleanly, because it would introduce a new judge-by-restorer confound. Consistent with that mismatch, later SummEval-based work (Liu et al., 2023; Bhandari et al., 2020) often treats fluency as a side metric or notes its limited utility.

C AggreFact Subtask Exclusions

We follow prior work in treating AggreFact as a heterogeneous benchmark. Tang et al. (2023b) explicitly analyze AggreFact-CNN and AggreFact-XSum separately, motivated by differences in source dataset properties, summary abstractiveness, error prevalence, and detector threshold behavior. Specifically, XSum summaries are highly abstractive and single-sentence. They often require compression, inference, and reframing. This makes "is this grounded?" less cleanly tied to surface overlap, and errors can be harder to classify. This exclusion is a scope choice tied to our prompt definition. Our judge is instructed not to rely on external knowledge or speculation, and to mark a claim as supported only when it is explicitly stated or unambiguously inferable from the source document. AggreFact-XSum represents a more abstractive factuality regime, where summaries often involve heavier compression and less direct source alignment.

We also exclude WiCE from the main experiments because it is not designed as a clean binary LLM-as-judge task. WiCE is a Wikipedia citation-entailment benchmark with three-way labels: Supported, Partially-Supported, and Not-Supported (Kamoi et al., 2023). Moreover, the claim-level development distribution is dominated by the partial class: 33.0% supported, 54.7% partially supported, and 12.3% not supported. Collapsing this label space into a binary true/false verdict would require an arbitrary mapping of Partially-Supported, turning the central semantic distinction of the dataset into label noise. Since our evaluation targets binary verdict accuracy and calibrated probability scores for that binary decision, WiCE is better treated as a separate citation-verification setting rather than part of the main benchmark suite.

D Bootstrap Inference for Subjectivity-Sensitivity Slopes

Slope differences are estimated by simple linear regression of τ_b on IAA (α). Confidence intervals and p-values come from a paired model-cluster bootstrap over 16 models ($B = 100{,}000$), resampling models with replacement while preserving all three task points per model.

E Flagship-style generation effects are weaker in open-weight model lines

The open-weight results provide only weak and mixed evidence of generation-style improvement. In the Llama line,

Model	Vendor	Release	`logprobs` usable?	`reasoning_effort` interaction	Max `top_logprobs`
GPT-4o	OpenAI	2024-08	Yes		20
o1	OpenAI	2024-12	No		N/A
GPT-4.1	OpenAI	2025-04	Yes		20
GPT-5	OpenAI	2025-08	No		N/A
GPT-5.2	OpenAI	2025-12	Conditional	Only for non-reasoning mode	5
GPT-5.4	OpenAI	2026-03	Conditional	Only for non-reasoning mode	5
Gemini 2.5 Pro	Google	2025-06	Yes		20
Gemini 3.1 Pro	Google	2025-11	Yes		20
All Claude Models	Anthropic	N/A	No		N/A

Table 2: Logprob availability across recent closed-model APIs.

Comparison	Signal	Δslope	p
Ours vs Rubric	τ(`prob`)	-0.143	$< 10^{-4}$
Ours vs Rubric	τ(`pred`)	-0.025	0.46
Within Ours	τ(`prob`) vs τ(`pred`)	-0.553	$< 10^{-5}$

Table 3: Bootstrap tests for slope differences in the subjectivity-robustness analysis. Probability signals reveal a significant improvement, whereas the binary verdict signal does not.

Signal	Prompt	Cons.	Coh.	Rel.	Slope↓	R^2 ↓
τ(`pred`)	Rubric	0.58	0.34	0.26	0.80	0.61
	Ours	0.55	0.37	0.24	0.78	0.59
τ(`prob`)	Rubric	0.47	0.37	0.32	0.37	0.27
	Ours	0.41	0.42	0.30	**0.23**	**0.15**

Table 4: Mean per-document Kendall τ_b for verdict and probability-score signals across the three non-fluency SummEval dimensions. Slopes and R^2 come from regression on IAA (Krippendorff's α).

`DeepSeek-R1-Distill-Llama-70B` improves both OV Gap and Debate Stress relative to base `Llama-3.3-70B`, suggesting that reasoning-oriented distillation can help prompt-stress robustness. However, this pattern does not generalize across the OLMo lines. For both `OLMo 3 7B` and `OLMo 3.1 32B`, the `SFT` → `DPO` → `RLVR` progression does not produce a monotonic improvement in either metric, and in some cases OV Gap under debate becomes larger even when Debate Stress improves slightly. Overall, these open-weight models show some local training-stage effects, but they do not exhibit the clean post-2025 robustness trend seen in flagship closed models. This suggests that our suspicion that the flagship generation effects are driven by training-pipeline improvements is only partially supported, and not cleanly reproduced in the open-source setting.

Family	Stage	OV-G↓	DS↓
Llama 3.3 70B	Base	**4.18%**	**1.55%**
	+R1 Distill	**2.10%**	**0.84%**
OLMo 3 7B	SFT	3.00%	**14.04%**
	+DPO	5.82%	**12.49%**
	+RLVR	6.08%	**9.10%**
OLMo 3.1 32B	SFT	0.75%	2.38%
	+DPO	0.76%	2.46%
	+RLVR	1.62%	2.32%

Table 5: **Open-weight model comparisons on AggreFact.** We compare Oracle-Verdict Gap (OV-G) and Debate Stress (DS) across three open-weight model lines: `Llama-3.3-70B` and its `DeepSeek-R1-Distill` derivative, and `OLMo-3-7B` and `OLMo-3.1-32B` across their `SFT` → `DPO` → `RLVR` training pipeline. Unlike the clearer trend observed in flagship closed models, these open-weight progressions do not show a consistent reduction in OV Gap or Debate Stress. Boldface marks the values that follow the expected generation effect.

F Soft-Score Protocol Comparability Check

Direct comparison of absolute τ_b values to previously reported results is only loosely informative, because the original GPT-4 model used in prior work has been retired and is no longer available. We therefore use the published G-Eval number only as an external anchor and instead validate our setup through a controlled same-model comparison on GPT-4o. As shown in Table 6, our logprob-free soft-score protocol closely matches a G-Eval-style control using logprob-derived expected Likert scores, supporting the correctness of our implementation and the validity of verbalized confidence as a practical logprob-free soft-score signal.

G Full AggreFact Comparison

Figure 7 shows faithfulness judging metrics over model generations across GPT, Claude, and Gemini families. Our prompt recipe is

Setting	Model	Soft-Score Protocol	τ_b
G-EVAL (Liu et al., 2023)	GPT-4	Logprob-derived expected Likert score	0.459
Our G-Eval control	GPT-4o-2024-08-06	Same G-Eval expected-score protocol	0.359
Our logprob-free pipeline	GPT-4o-2024-08-06	Verbalized confidence scores (0–100)	0.353

Table 6: Soft-score protocol comparability check on the SummEval consistency dimension.

effective at strengthening the confidence signal: compared with Rubric, Ours-FreeReason and especially Ours typically improve Calibration Quality, often substantially, without sacrificing overall verdict or oracle performance. Beyond this prompt effect, a clear generation effect also appears: newer post-2025 flagship models show smaller Oracle-Verdict Gaps and lower Debate Stress, meaning that less latent faithfulness signal is left only in confidence and that debate becomes less harmful to the final verdict. Thus, the figure supports both parts of our storyline: our prompts improve the confidence channel, and newer flagship models are increasingly robust enough to convert that stronger latent signal into stable verdicts.

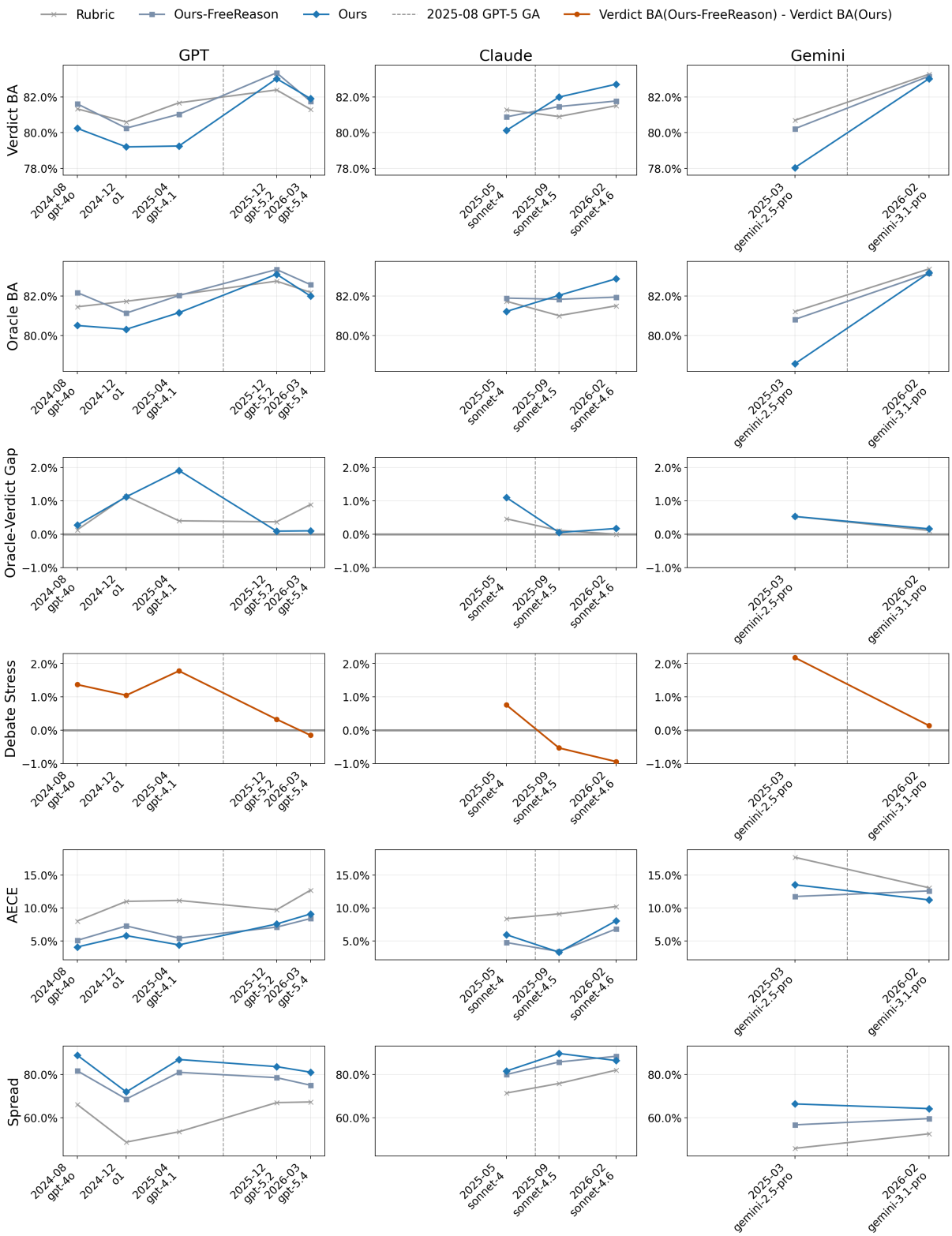


Figure 7: Full AggreFact comparison over model generations across GPT, Claude, and Gemini families.

H Our prompt recipe

This appendix gives the verbatim text of the three verbalized-confidence protocols used in our experiments, instantiated on SummEval's Relevance dimension as the running example. The `{context}` placeholder receives the original article and `{candidate}` receives the candidate summary; for AggreFact, the same skeletons are reused with the criterion and slots replaced to fit a faithfulness setting.

Component composition. We first map the three protocols onto the prompt-component vocabulary introduced in Section 2.1. Table 7 summarizes which components are active in each protocol.

Component	Rubric	Ours-FR	Ours
Task definition	✓	✓	✓
Prompt inputs	✓	✓	✓
Reasoning request		✓	
Self-debate request			✓
Decision request	✓	✓	✓
Advisory		✓	✓
Confidence rubric	✓	✓	✓
Output schema	✓	✓	✓

Table 7: Prompt-component activation across the three verbalized-confidence protocols. Ours-FR abbreviates Ours-FreeReason.

In words: **Rubric** is the most stripped-down protocol — it asks for the verdict and a rubric-anchored confidence directly, with no advisory and no reasoning step. **Ours-FreeReason** adds the overconfidence advisory and a free-form reasoning request before the verdict. **Ours** replaces that free-form reasoning with a structured self-debate request (reasons for True, reasons for False, then conclusion), keeping the advisory and rubric. The remaining four components (task definition, prompt inputs, decision request, output schema) are present in all three.

Rubric. Decision request and rubric only.

Ours-FreeReason. Adds the advisory and a free-form reasoning request to the Rubric baseline; the output schema gains an EXPLANATION field.

Ours. Replaces the free-form reasoning request with a structured self-debate request; advisory and rubric carry over from Ours-FreeReason.

Determine whether the summary captures the important content from the original article. A relevant summary should include only the main points from the source document and avoid redundant or tangential information. Penalize summaries that miss key points or that pad the summary with unimportant details.

Original Article:
"{context}"

Summary:
"{candidate}"

State your answer of "True" or "False", and provide your confidence score (0-100) indicating how confident you are in your answer.

Confidence Rubrics (for your chosen answer):

- 95-100: Absolutely certain. The answer is trivially obvious with no room for any other interpretation.
- 85-94: Very confident. Strong evidence directly supports your answer with minimal ambiguity.
- 70-84: Confident. Good evidence supports your answer, but minor ambiguities or alternative readings exist.
- 55-69: Somewhat confident. Evidence leans toward your answer, but reasonable counterarguments exist.
- 45-54: Uncertain. Evidence is mixed or ambiguous; either answer could be defensible.
- 35-44: Slightly uncertain. You chose this answer but the opposite has merit.
- 25-34: Quite uncertain. Weak evidence for your answer; opposite may be equally valid.
- Below 25: Very uncertain. Consider changing your answer to the opposite.

Format your response as:
Answer: [True/False]
Confidence: [0-100]

Figure 8: **Rubric prompt** (SummEval Relevance instantiation).

[Identical task definition and prompt inputs as Rubric.]

Provide your explanation with quotes of evidence from the Original Article, then answer True (relevant) or False (not relevant), and provide your confidence score (0-100) indicating how confident you are in your answer.

IMPORTANT: You have been consistently overconfident in past evaluations. Before assigning a high confidence score, actively look for reasons you might be wrong. Consider alternative interpretations of what counts as "important content" and edge cases.

ALSO IMPORTANT: You are especially overconfident when predicting "False". Even when you feel certain a summary is not relevant, there is often a valid interpretation that makes it relevant. Always ask yourself: "Could there be a reasonable reading of the Original Article that makes this summary's content selection appropriate?"

[Confidence Rubrics block, identical to Rubric.]

Format your response as:
Explanation:
[your reasoning with quotes of evidence from the Original Article]
Answer: [True/False]
Confidence: [0-100]

Figure 9: **Ours-FreeReason prompt** (SummEval Relevance instantiation).

[Identical task definition and prompt inputs as Rubric.]

```
Emulate a debate in three steps:
1) True because:  Argue for "True" (relevant).  List reasons why the Summary captures the
   important content of the Original Article, with quotes from the Original Article.

2) False because:  Argue for "False" (not relevant).  List reasons why the Summary misses
   important content or includes tangential content, with quotes from the Original Article.

3) Explanation:  Based on both arguments, determine your final answer "True" (relevant) or
   "False" (not relevant).
Provide your debate using the fields:  "True because", "False because", and "Explanation",
then state your final answer of "True" or "False", and provide your confidence score (0-100)
indicating how confident you are in your answer.
```

[Advisory block and Confidence Rubrics block, identical to Ours-FreeReason.]

```
Format your response as:
True because:
[reasons why the summary captures important content with quotes of evidence]
False because:
[reasons why the summary misses or pads important content with quotes of evidence]
Explanation:
[your conclusion based on the above debate]
Answer:  [True/False]
Confidence:  [0-100]
```

Figure 10: **Ours prompt** (SummEval Relevance instantiation).