ConsensusDrift: Semantic Salience Can Override Majority Evidence in LLM Aggregation

Anonymous ACL submission

Abstract

We study how language models aggregate conflicting evidence in controlled multi-juror settings. As a null control, we evaluate symbolic majority voting where jurors provide only YES/NO votes; across models, this regime is largely stable and exhibits near-zero minority override. We then test semantic evidence aggregation, where each juror provides a short evidence statement paired with a local verdict. In this regime, aggregation becomes strength-dependent and position-sensitive: override rates increase when a minority statement is operationalized as stronger than a weak majority, while a strong-majority/weak-minority control remains low-override. Authority cues applied to a minority statement further amplify minority influence.

We validate these patterns with (i) seed replications (n=5 seeds), (ii) cross-domain synthetic templates (legal, medical, scientific review, corporate audit, fact-checking), (iii) a phase-transition analysis mapping override probability across majority strength, minority strength, authority level, and minority position, (iv) rationality-vs-salience controls that separate evidence-quality-driven overrides from surface-framing effects, and (v) a frontier validation subset (GPT-4o-mini, GPT-4.1-mini) confirming that the qualitative ordering is not specific to open-weight models. We log all trials and preserve raw model outputs unchanged; for models with lower strict-format compliance, we score responses using a lightweight post-hoc extractor and provide explicit extraction audits with a manual agreement protocol. Overall, our results indicate that under semantic conflict, LLM evidence aggregation can depart from count-based consensus and become sensitive to controlled manipulations of evidence salience, authority framing, and recency. We emphasize that "strength" is an operationalized synthetic manipulation and that minority-following outcomes can be rational when minority evidence is decisive.

1 Introduction

Language models are increasingly deployed in settings that require aggregating multiple pieces of information: retrieved passages in retrieval-augmented generation (Lewis et al., 2020), multi-document summaries, reviewer reports, or structured assessments (Liang et al., 2022). In these settings, evidence sources may disagree (Longpre et al., 2022; Xie et al., 2024). A key systems question is whether aggregation behaves predictably and proportionally, or whether small changes to presentation—such as ordering, phrasing, or authority cues—can shift the final decision.

This paper studies *evidence aggregation dynamics* under controlled multi-juror inputs. We use a simple but flexible harness: a prompt lists five "jurors" who each provide a short statement and a local verdict. The model is instructed to return a final verdict and confidence based only on the juror evidence. We focus on a single descriptive metric, *minority override rate*: the probability that the model's final verdict matches the minority juror despite a numerical majority.

Two regimes: symbolic vs. semantic aggregation. We contrast two regimes with matched prompt structure. In a **symbolic** voting regime, jurors provide only YES/NO votes. In a **semantic** regime, jurors provide short evidence statements paired with verdict labels. This contrast is central for scientific cleanliness: symbolic voting yields a strong null control. If minority override were inevitable under the prompt family, it would appear in the symbolic regime as well. Instead, we observe near-zero override under symbolic voting, and substantial regime shifts under semantic conflict.

Controlled manipulations. Within the semantic regime, we manipulate three factors that are common in multi-source NLP pipelines:

- **Operationalized evidence strength:** Circum-

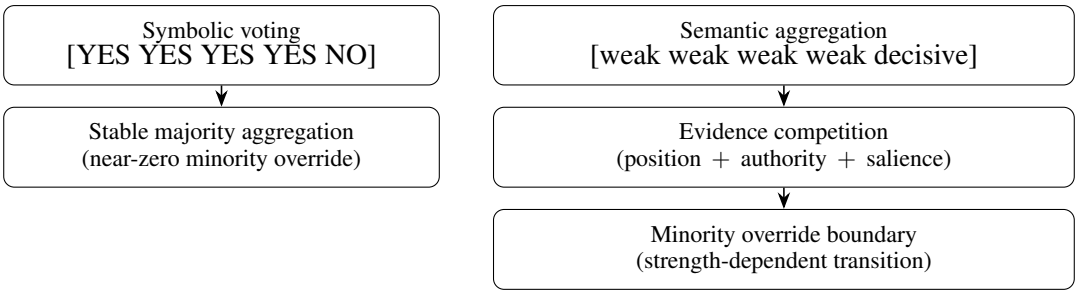

Figure 1: **Conceptual overview: from counting to semantic competition.** Under symbolic voting, aggregation is largely count-based and stable. Under semantic conflict, aggregation becomes a competition among evidence strength and surface cues (position and authority), yielding a strength-dependent minority-override boundary.

stantial/Moderate/Direct/Decisive templates defined by a deterministic rubric.

- **Authority cues:** the minority (or majority) statement is framed as coming from a role such as a lead analyst or institutional reviewer.
- **Position (recency):** the minority statement is rotated across the five juror positions while keeping content constant (Liu et al., 2024; Pezeshkpour and Hruschka, 2023).

Importantly, we do not treat strength as real-world ground truth; it is a synthetic manipulation chosen to enable clean stress testing (Ribeiro et al., 2020).

Why this matters. Evidence aggregation is a core primitive in modern NLP systems. When a model is used to integrate multiple retrieved documents (Yang et al., 2018) or multiple reviewers' comments, a system designer implicitly assumes some degree of proportionality: that repeated corroborating evidence is not easily overturned by a single contradictory snippet unless that snippet is sufficiently strong. Positioning and salience effects are known to distort LLM outputs in evaluation (Zheng et al., 2023; Wang et al., 2023) and selection (Zheng et al., 2024) tasks; our harness provides a minimal way to test whether analogous effects appear in evidence aggregation, and to map the conditions under which aggregation shifts from majority-dominated to salience-weighted behavior.

Contributions.

- We introduce a controlled multi-juror harness contrasting *symbolic* majority voting (null control) with *semantic* evidence aggregation under conflict.
- We characterize a strength-dependent transition boundary modulated by authority cues and minority position, with seed replication and cross-domain template validation.
- We provide reproducible artifacts (scripted figures and aggregate tables) and document extraction caveats with a manual agreement protocol for format-noncompliant outputs.

2 Related Work

We position this work as a controlled behavioral study of *evidence aggregation* in language models. Our framing is intentionally system- and measurement-oriented: we do not study social processes, but rather how an LLM combines multiple short, potentially conflicting inputs into a single decision.

Aggregation under disagreement. Evidence aggregation is central to multi-document question answering (Yang et al., 2018; Ho et al., 2020), retrieval-augmented generation (Lewis et al., 2020), and summarization, where sources may partially disagree (Longpre et al., 2022; Chen et al., 2022). In such pipelines, models are often expected to integrate information proportionally, but empirical reports have noted that aggregation can be sensitive to prompting, formatting, and interaction effects among inputs (Xie et al., 2024). Our contribution is a minimal harness that (i) preserves a symbolic null control and (ii) introduces semantic competition via controlled templates, enabling systematic factor manipulation.

Long-context and ordering effects. Ordering sensitivity and recency effects are well-known in long-context processing (Liu et al., 2024; Li et al., 2024; An et al., 2024). Liu et al. (2024) show that models are disproportionately influenced by information near the beginning or end of long contexts; our setting uses short fixed-length prompts

but shows analogous position sensitivity under semantic conflict. Pezeshkpour and Hruschka (2023) and Zheng et al. (2024) document ordering biases in multiple-choice settings; we extend this perspective to aggregated evidence scenarios. We treat position primarily as a *modulator* of the semantic aggregation regime: the phase-transition analysis shows that strength and authority cues can shift aggregation even when position is controlled.

Sycophancy and social influence. A related phenomenon is sycophancy (Sharma et al., 2023): models may adjust their outputs toward perceived preferences or dominant framings. Authority framing in our setup can be viewed as a test of whether models over-weight inputs that carry social authority cues, analogous to sycophantic deference. Turpin et al. (2023) show that LLM reasoning explanations can be unfaithful to the actual decision process; our setting provides a complementary structural test where the decision outcome is measured directly.

LLM-as-evaluator and evaluation bias. Evaluation using LLMs is known to exhibit positional and verbosity biases (Zheng et al., 2023; Wang et al., 2023). Our setting inverts the framing: rather than using an LLM to evaluate text, we measure how an LLM aggregates multiple evaluators (jurors). Both lines of work point to the same underlying sensitivity to ordering and surface framing in LLM judgments. Because LLM-based judgments are increasingly used to generate preference signals (e.g., RLHF/RLAIF), understanding such aggregation dynamics is relevant to evaluation pipelines as well as alignment data collection (Christiano et al., 2017; Ouyang et al., 2022; Bai et al., 2022).

Template-based behavioral testing. Behavioral evaluation via synthetic templates is a common strategy for probing model capabilities under controlled conditions (Ribeiro et al., 2020; Gehrmann et al., 2021). The main benefit is experimental cleanliness: templates can isolate a single factor (e.g., strength or authority) while keeping other aspects constant (Webson and Pavlick, 2022; Si et al., 2023). The main limitation is ecological validity: templates do not represent the full complexity of real evidence. We therefore emphasize that our strength levels are *operationalized* and treated as controlled manipulations rather than real-world ground truth. We treat model-reported confidence as descriptive rather than calibrated probability (Kadavath et al., 2022; Zhao et al., 2021), and follow recommendations for transparent reporting and null controls in empirical NLP (Liang et al., 2022).

3 Method

3.1 Task

We study controlled *evidence aggregation dynamics* under semantic conflict using a five-juror prompt template. In the **symbolic** setting, jurors provide only YES/NO votes (majority vs minority; minority position rotated). In the **semantic** setting, each juror provides a short evidence statement paired with a local verdict (e.g., GUILTY/NOT_GUILTY). Across conditions, we hold the overall structure constant and vary controlled factors (evidence strength level, authority framing, recency/position).

Prompt structure. Each instance lists five jurors in order. In symbolic prompts, each juror line consists only of a vote token. In semantic prompts, each juror line contains a short evidence statement followed by an explicit local verdict token. The model is instructed to decide a final verdict *based only on the juror evidence* and to return a verdict label, confidence, and short reasoning.

Minority rotation. We keep the number of jurors constant and rotate the minority statement across positions (1–5). This isolates positional effects (recency) without changing evidence content.

Conditions. The semantic suite includes: (A) **Strong Minority**, (B) **Weak Minority (Control)**, (C) **Equal Evidence**, (D) **Emotional Minority**, and (E) **Authoritative Minority**. We treat "strength" and "authority" as controlled synthetic manipulations defined by a deterministic rubric (Ribeiro et al., 2020), not real-world ground truth.

3.2 Models

Open-weight models are queried via DeepInfra: Meta-Llama-3.1-8B-Instruct (Touvron et al., 2023), Qwen2.5-7B-Instruct (Qwen Team, 2023), and Mistral-7B-Instruct-v0.3 (Jiang et al., 2023). We additionally run a frontier-validation subset on OpenAI models (GPT-4o-mini and GPT-4.1-mini (OpenAI, 2023)) when API keys are available; missing providers are skipped without failing the pipeline.

Decoding. Unless otherwise specified, we use temperature 0.7 and a short maximum output

length to encourage concise structured responses. Each configuration is repeated with multiple seeds; seed IDs are logged in the output records.

3.3 Metrics

Our primary metric is **minority override rate**: the fraction of trials where the model's final verdict matches the minority juror's verdict. We report position sensitivity (override vs minority position), confidence drift (confidence under majority-following vs minority-following), and seed-robust aggregated estimates with confidence intervals.

Uncertainty. We summarize uncertainty with confidence intervals and across-seed variance. For aggregated plots/tables, we use bootstrap or binomial approximations depending on the artifact; all choices are documented alongside the corresponding output files.

3.4 Logging and parsing

We log every run (prompts, raw responses, parsed outputs) and preserve raw outputs unchanged. When strict JSON parsing fails, we score an additional extracted verdict field using a lightweight rule-based extractor while retaining the original parse flag; extraction is validated via a manual agreement protocol on a scorable subset (Liang et al., 2022).

4 Results

We present a control-first narrative emphasizing contrasts, reproducibility, and transparent caveats. All runs are logged, and raw model outputs are preserved unchanged.

4.1 Symbolic aggregation is stable (null result)

Setup. In the symbolic setting, jurors provide only YES/NO votes. We use 5 jurors with a fixed 4–1 majority/minority split and rotate the minority vote across positions 1–5. The model is instructed to decide a final verdict *based only on the votes* and return a verdict label, confidence, and short reasoning.

Finding. Across models and seeds, symbolic aggregation yields near-zero minority override: models overwhelmingly select the majority verdict (Figure 2). This null result is important because it demonstrates that minority override is not an inevitable artifact of the prompt format (Webson and Pavlick, 2022). Instead, the instability we observe later is specific to semantic evidence competition.

4.2 Semantic evidence competition shifts aggregation (A/E vs. B)

Setup. In the semantic setting, each juror provides a short evidence statement paired with a local verdict (e.g., GUILTY/NOT_GUILTY). We evaluate five semantic condition types (A–E). For the main paper, we focus on the clearest contrast:

- **A (Strong Minority):** 4 jurors with weak/circumstantial evidence vote GUILTY; 1 juror with decisive exculpatory evidence votes NOT GUILTY.
- **E (Authoritative Minority):** Same structure as A, but the minority statement is prefixed with an institutional authority cue (e.g., "Lead forensic analyst:").
- **B (Weak Minority, Control):** 4 jurors with strong evidence vote GUILTY; 1 juror with weak/non-evidential statement votes NOT GUILTY.

In all cases the structure is held constant (5 jurors; 4 vs. 1 split; minority rotated), and only the controlled manipulation differs.

Finding. Semantic aggregation exhibits a marked regime shift for models with high format compliance (Figure 3). For LLaMA-3.1-8B, condition A yields a pooled override rate of 1.00 (95% CI: [1.00, 1.00]) and condition E yields 1.00, while the control condition B yields 0.00 across all positions. Mistral-7B-Instruct-v0.3 shows a similar contrast (A: 0.85–1.00 depending on position; B: 0.00). These contrasts replicate across 5 independent seeds.

Qwen caveat. Qwen2.5-7B-Instruct shows near-zero strict JSON parse compliance (parse-ok $\approx$ 0.1%); only $\approx$34% of its outputs are scorable via extraction. Its recovered pattern is anomalous relative to LLaMA and Mistral: condition B (strong-majority control) yields override rates of 0.23–0.85 across positions, rather than near-zero. Because Qwen's scorable outputs are a small, potentially unrepresentative subset, we treat Qwen results as exploratory and do not base main-paper claims on them. LLaMA and Mistral drive the pooled figure; Qwen's per-model breakdown is in the appendix.

The control condition is critical for LLaMA and Mistral: it shows that override is contingent on controlled evidence quality, not a general tendency to side with a salient minority (Sharma et al., 2023).

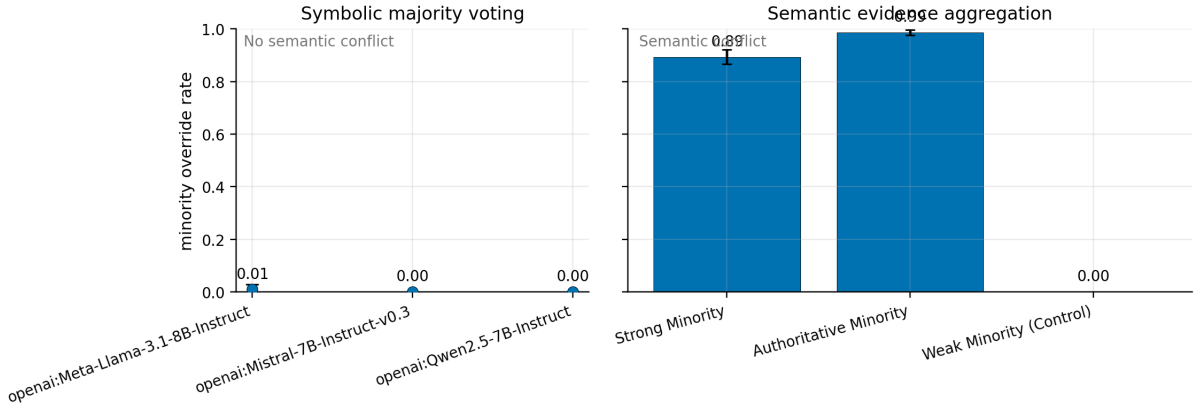

Figure 2: **Symbolic vs. semantic aggregation regime.** Minority override rate (fraction of trials selecting the minority verdict) across symbolic voting (near-zero) and semantic evidence conditions. Symbolic aggregation is largely stable; semantic conflict can yield substantially higher override rates. Limitation: symbolic and semantic tasks use different evidence content; output compliance varies by model.

Condition	Controlled manipulation (semantic setting)
Strong Minority	Weak majority evidence vs. decisive minority evidence
Weak Minority (Control)	Strong majority evidence vs. weak minority evidence
Equal Evidence	Majority and minority evidence matched in strength
Emotional Minority	Factual majority evidence vs. emotionally framed minority
Authoritative Minority	Generic majority evidence vs. authority-cued minority

Table 1: **Semantic condition taxonomy (no data).** Each condition keeps jury size (5) and split (4–1) fixed; only the highlighted factor changes. This table is included to reduce cognitive load and clarify what each condition tests.

4.3 Recency modulates semantic overrides

We isolate position effects by rotating the minority statement across positions 1–5 while keeping evidence content constant (Liu et al., 2024; Pezeshkpour and Hruschka, 2023). In the semantic regime, override rates vary with minority position, and the direction differs by model: Mistral shows a recency effect (minority is more influential at later positions; permutation test $r = 0.38$, $p < 0.0001$, $n = 250$, seed 0 replication), while LLaMA shows a primacy effect ($r = -0.29$, $p < 0.0001$, $n = 250$). This model-specific directionality is consistent across seeds. We treat position primarily as a *modulator*: the phase-transition analysis (Figure 4) shows that strength and authority cues shift override probability even when position is held constant.

Effect sizes. Across-seed effect sizes (Strong Minority minus Weak Minority control) are reported in the released supplementary tables.

4.4 Phase-transition characterization: a strength-dependent boundary

Goal. To move beyond single-condition demonstrations, we map when minority-following outcomes become likely as we vary the majority/minority strength levels (weak/medium/strong/decisive) and authority cues. We treat strength as an experimental manipulation defined by a deterministic rubric (Ribeiro et al., 2020) (included in the supplementary material).

Strength rubric availability. The full rubric and per-template assignments are included in the supplementary material.

Finding. The phase-transition surface (Figure 4) exhibits a strength-dependent transition boundary: override probability increases sharply as minority strength increases relative to majority strength. Authority cues act as an amplifying factor, shifting the boundary so that minority influence increases at lower strength gaps. Position modulates the boundary rather than fully determining it.

Reliability and transparency. We provide a cell-level completeness report with parse-ok rates, extraction rates, n scored rows, and confidence intervals in the supplementary material. Deterministic exclusion rules for low-reliability cells and their counts are also documented there. Main-paper claims rely only on included cells.

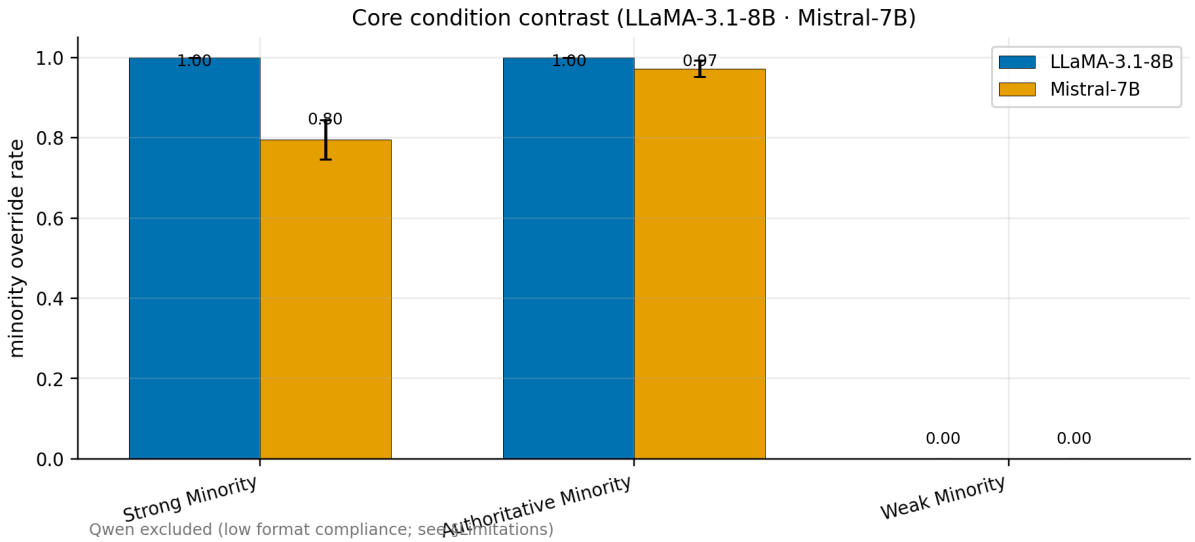

Figure 3: **Core semantic condition contrast (LLaMA-3.1-8B and Mistral-7B).** Override rates for Strong Minority (A), Authoritative Minority (E), and Weak Minority Control (B), shown per model with 95% CIs. Both models show the same ordering: A/E $\gg$ B. Qwen excluded due to low format compliance ($\approx$0.1% parse-ok; see §4.6). The control condition (B) remains at near-zero, confirming that override is contingent on controlled evidence properties. Limitation: conditions use controlled synthetic templates.

4.5 Cross-domain generalization and frontier validation subset

Domains. We evaluate domain surface forms spanning legal, medical, scientific review, corporate audit, and fact-checking (Figure 5). Domain generalization is intentionally limited to *surface form* variation: the evidence structure and controlled factors remain the same (Gehrmann et al., 2021). The qualitative ordering (salient/authority minority > strong-majority control) persists across all five domains with domain-dependent effect sizes.

Frontier subset. We run a frontier validation suite on GPT-4o-mini and GPT-4.1-mini (Figure 6). GPT-4.1-mini replicates the open-weight pattern sharply: override rate 1.00 for Strong Minority (n=158), 1.00 for Authoritative Minority (n=165), and 0.00 for the Weak Minority control (n=177). GPT-4o-mini shows a graded but qualitatively consistent response: 0.50 for Strong Minority (n=173), 0.59 for Authoritative Minority (n=162), and 0.00 for control (n=165). The goal of this subset is not comprehensive frontier benchmarking but establishing that the qualitative A/E > B ordering is not specific to open-weight models (OpenAI, 2023).

4.6 Rationality vs. salience controls

A central reviewer concern is that minority override may reflect reasonable evidence weighting (Xie et al., 2024). We therefore distinguish three regimes using five control families (R1–R5):

- **R1 (rational minority):** decisive minority evidence, neutral framing. Override here can be normatively correct.
- **R2 (salient but weak minority):** minority is verbose/emotional but not decisive; strong majority retained. Override here is more plausibly salience-driven.
- **R3 (strong but plain minority):** strong minority, minimal/unadorned framing. Tests whether plain-phrased strong evidence drives override.
- **R4 (weak but authoritative minority):** weak minority framed as institutional expert; strong majority. Tests pure authority-framing effects (Sharma et al., 2023).
- **R5 (strong majority vs. salient weak minority):** decisive majority; minority is salient (emotional/authoritative) but weak. This is the most stringent test of salience dominance.

Results for high-compliance models (legal domain): LLaMA-3.1-8B: R1 = 1.00 (n=200); R2

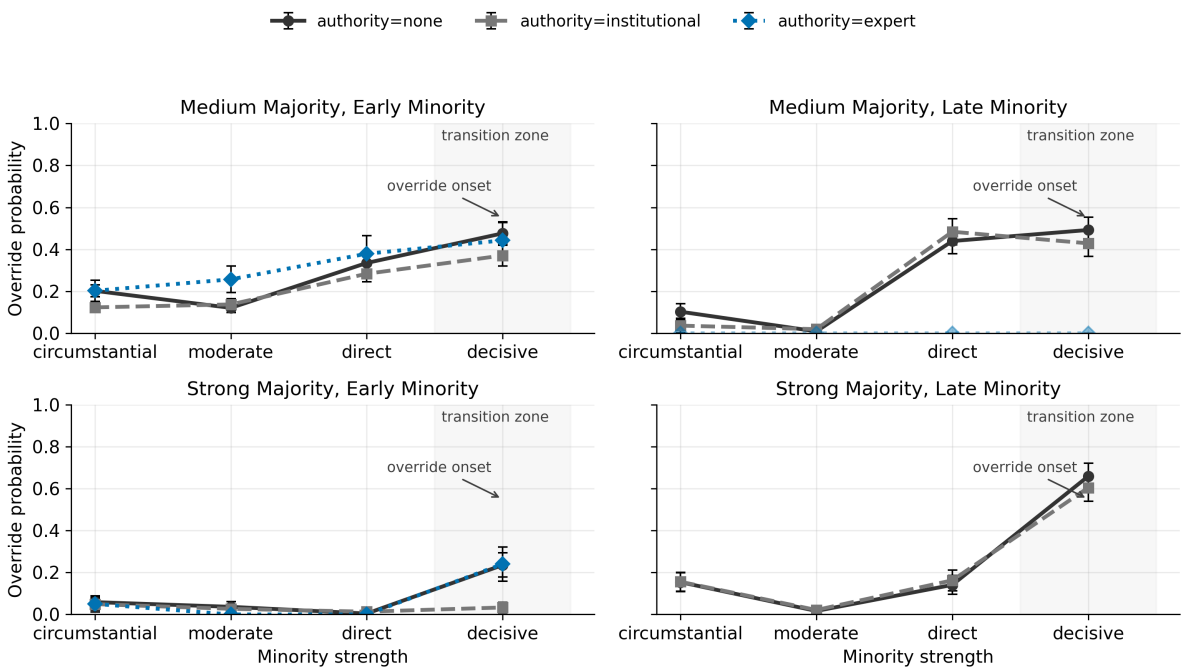

Figure 4: **Phase-transition characterization.** Override probability across a majority-strength × minority-strength grid, with authority and minority-position modulation. Override follows a strength-dependent boundary: as minority strength increases relative to majority strength, override probability rises sharply. Authority cues shift the boundary. Low-reliability cells are excluded by predefined rules (see completeness report). Limitation: controlled synthetic templates; normal-approximation CIs.

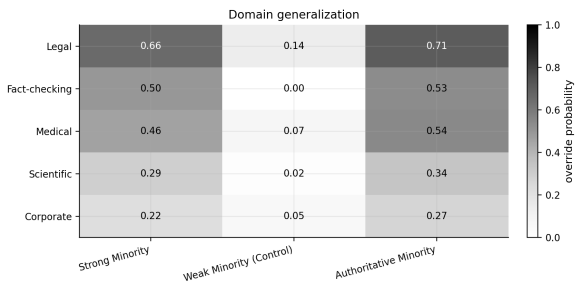

Figure 5: **Cross-domain generalization heatmap.** Mean override rate (over models) for the core semantic conditions across five synthetic domain surface forms. The qualitative ordering (Strong/Authoritative Minority > Weak Minority Control) persists across domains with domain-dependent effect sizes. Limitation: domains are synthetic templates; this is robustness to surface form variation, not ecological validity.

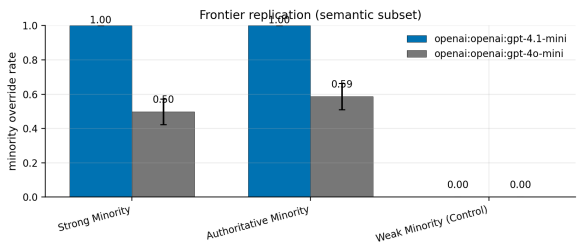

Figure 6: **Frontier model replication (semantic A/E/B subset).** Override rates for GPT-4o-mini and GPT-4.1-mini on the three core conditions. The qualitative ordering (Strong/Authoritative > Control) is consistent with frontier models. GPT-4.1-mini shows near-perfect discrimination (A: 1.00, E: 1.00, B: 0.00). GPT-4o-mini shows a graded response (A: 0.50, E: 0.59, B: 0.00). Limitation: limited-scope subset; not a full frontier benchmark.

= 0.00 (n=181); R3 = 0.73 (n=150); R4 = 0.00 (n=150); R5 = 0.00 (n=150). Mistral-7B: R1 = 0.58 (n=200); R2 = 0.00 (n=150); R3 = 0.33 (n=150); R4 = 0.00 (n=150); R5 = 0.00 (n=150). The pattern is consistent across both models: override is elevated only when minority evidence is operationally decisive (R1, R3); authority framing alone (R4) or salience without quality (R2) do not drive override when majority evidence is strong. Qwen's rationality-control results are extraction-dependent and show a different pattern (R4 = 0.62); we treat them as exploratory given the low scorable fraction. Full cross-domain results are in Appendix C.

4.7 Extraction caveats and measurement reliability

Why extraction is needed. Some models do not consistently emit the requested strict JSON output format (Liang et al., 2022). We preserve raw outputs unchanged and derive additional extracted fields for verdict and confidence while retaining the original parse flag.

Agreement and scorable fraction. We validate extraction-based scoring using a manual labeling protocol. On a labeled *scorable* validation sheet (n=150; 100 Qwen + 50 non-Qwen), extractor agreement with manual labels is 1.00 overall. We separately report the *scorable fraction* (parseable or extractable) by dataset and model in the supplementary material; agreement on scorable outputs does not imply that all outputs are scorable.

5 Analysis

We report seed-aggregated replication summaries, a phase-transition sweep over majority/minority strength with authority and position modulation, and rationality-vs-salience controls. To avoid cherry-picking, we apply predefined reliability exclusions to low-quality phase-grid cells and report cell-level completeness in the supplementary material (Liang et al., 2022).

6 Limitations

Our study is a controlled measurement harness with several limitations. Evidence strength is operationalized via deterministic synthetic templates (Ribeiro et al., 2020). Some models do not consistently follow strict output formats, requiring post-hoc extraction for scoring (Liang et al., 2022); Qwen in particular has low strict parse compliance and is treated as exploratory. Our setting uses 5 jurors and short stylized evidence statements; longer contexts, different elicitation formats, or larger juries may change quantitative effect sizes (Liu et al., 2024; Li et al., 2024). Finally, confidence values are self-reported and not calibrated probabilities (Kadavath et al., 2022; Zhao et al., 2021). Semantic experiments use GUILTY/NOT_GUILTY labels while symbolic experiments use YES/NO; we therefore emphasize within-regime controls and interpret the symbolic null as a qualitative contrast rather than a strict quantitative comparison across regimes.

7 Ethics

This work uses synthetic scenarios and does not involve human subjects. We frame results as controlled evidence aggregation behavior in LLMs and avoid political or governance claims.

Responsible interpretation. We do not recommend using the prompts or scoring procedures in isolation for real decision-making. The reported effects characterize model behavior under controlled manipulations and do not constitute normative judgments about correctness in legal, medical, or other high-stakes settings.

Potential misuse. Findings about authority cues and ordering sensitivity could be misinterpreted as guidance for persuasive prompting; we therefore emphasize limitations and the preserved null control.

8 Conclusion

We presented a controlled empirical study of evidence aggregation dynamics in language models under semantic conflict. A key null result is that symbolic majority voting is largely stable with near-zero minority override. In contrast, when jurors provide short evidence statements, aggregation becomes sensitive to controlled manipulations of evidence strength, authority framing, and minority position, yielding a strength-dependent transition boundary with recency modulation.

We emphasize that evidence strength is operationalized using deterministic synthetic templates and should be interpreted as an experimental manipulation rather than real-world ground truth. We preserve raw outputs and provide explicit extraction caveats and manual validation for format-noncompliant outputs.

References

Shengnan An, Zexiong Ma, Zeqi Lin, Nanning Zheng, and Jian-Guang Lou. 2024. Make your llm fully utilize the context. *Preprint*, arXiv:2404.16811.

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, and others. 2022. Constitutional AI: Harmlessness from AI feedback. *arXiv preprint arXiv:2212.08073*.

Hung-Ting Chen, Michael J. Q. Zhang, and Eunsol Choi. 2022. Rich knowledge sources bring complex knowledge conflicts: Recalibrating models to reflect conflicting evidence. *Preprint*, arXiv:2210.13701.

Paul F. Christiano, Jan Leike, Tom B. Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. Deep reinforcement learning from human preferences. In *Advances in Neural Information Processing Systems*.

Sebastian Gehrmann, Tosin Adewumi, Karmanya Aggarwal, Pawan Sasanka Ammanamanchi, Aremu Anuoluwapo, Antoine Bosselut, Khyathi Raghavi Chandu, Miruna Clark, Jonathan Freeman, Tao Gao, and 1 others. 2021. The GEM benchmark: Natural language generation, its evaluation and metrics. In *Proceedings of the 1st Workshop on Natural Language Generation, Evaluation, and Metrics (GEM)*, pages 96–120.

Xanh Ho, Anh-Khoa Duong Nguyen, Saku Sugawara, and Akiko Aizawa. 2020. Constructing a multi-hop QA dataset for comprehensive evaluation of reasoning steps. In *Proceedings of the 28th International Conference on Computational Linguistics (COLING)*, pages 6609–6625.

Albert Q. 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, Jiawei Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7B. *arXiv preprint arXiv:2310.06825*.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, Scott Johnston, Sheer El-Showk, Andy Jones, Nelson Elhage, Tristan Hume, Anna Chen, Yuntao Bai, Samuel Bowman, Stanislav Fort, and 16 others. 2022. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 33, pages 9459–9474.

Jiaqi Li, Mengmeng Wang, Zilong Zheng, and Muhan Zhang. 2024. LooGLE: Can long-context language models understand long documents? In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, Benjamin Newman, Binhang Yuan, Bobby Yan, Ce Zhang, Christian Cosgrove, Christopher D. Manning, Christopher Ré, Diana Acosta-Navas, Drew A. Hudson, and 9 others. 2022. Holistic evaluation of language models. In *Advances in Neural Information Processing Systems: Datasets and Benchmarks Track (NeurIPS)*.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2024. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173.

Shayne Longpre, Kartik Perisetla, Anthony Chen, Nikhil Ramesh, Chris DuBois, and Sameer Singh. 2022. Entity-based knowledge conflicts in question answering. *Preprint*, arXiv:2109.05052.

OpenAI. 2023. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, and others. 2022. Training language models to follow instructions with human feedback. *arXiv preprint arXiv:2203.02155*.

Pouya Pezeshkpour and Estevam Hruschka. 2023. Large language models sensitivity to the order of options in multiple-choice questions. In *Findings of the Association for Computational Linguistics: EMNLP*, pages 8016–8026.

Qwen Team. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 2020. Beyond accuracy: Behavioral testing of NLP models with CheckList. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 4902–4912.

Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R. Bowman, Newton Cheng, Esin Durmus, Zac Hatfield-Dodds, Scott R. Johnston, Shauna Kravec, Timothy Maxwell, Sam McCandlish, Kamal Ndousse, Oliver Rausch, Nicholas Schiefer, Da Yan, Miranda Zhang, and Ethan Perez. 2023. Towards understanding sycophancy in language models. *arXiv preprint arXiv:2310.13548*.

Chenglei Si, Zhe Gan, Zhengyuan Yang, Shuohang Wang, Jianfeng Wang, Jordan Boyd-Graber, and Lijuan Wang. 2023. Prompting gpt-3 to be reliable. *Preprint*, arXiv:2210.09150.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, and 49 others. 2023. Llama 2: Open foundation and fine-tuned chat models. *Preprint*, arXiv:2307.09288.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel R. Bowman. 2023. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36.

Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Qi Liu, Tianyu Liu, and Zhifang Sui. 2023. Large language models are not fair evaluators. *Preprint*, arXiv:2305.17926.

Albert Webson and Ellie Pavlick. 2022. Do prompt-based models really understand the meaning of their prompts? In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 2300–2344.

Jian Xie, Kai Zhang, Jiangjie Chen, Renze Lou, and Yu Su. 2024. Adaptive chameleon or stubborn sloth: Revealing the behavior of large language models in knowledge conflicts. *Preprint*, arXiv:2305.13300.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W. Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2369–2380.

Tony Z. Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. 2021. Calibrate before use: Improving few-shot performance of language models. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, pages 12697–12706.

Chujie Zheng, Hao Zhou, Fandong Meng, Jie Zhou, and Minlie Huang. 2024. Large language models are not robust multiple choice selectors. *Preprint*, arXiv:2309.03882.

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)*, volume 36.

A Extraction validation (manual audit)

We validate extraction-based scoring using a manual labeling protocol.

Why extraction is needed. Some models do not consistently emit the required strict JSON output format. We therefore preserve raw model outputs unchanged and derive an additional extracted verdict field using a lightweight, rule-based extractor. We retain the original parse flag and treat extraction-dependent results (notably Qwen) as caveated.

Manual labeling protocol. We construct a manual labeling sheet containing 100 sampled Qwen outputs and up to 50 additional non-Qwen extracted outputs when available. Annotators label the stated final verdict (or mark ambiguous/unscorable) and optionally provide a confidence note. The sheet template is included in the supplementary material.

Agreement statistics. We compute agreement after labeling using a released agreement script, producing an agreement summary and a confusion breakdown by extraction method. On the labeled *scorable* validation sheet (n=150; 100 Qwen + 50 non-Qwen), extractor agreement with manual labels is 1.00 overall (disagreement rate 0.00; ambiguous rate 0.00; unsupported-format rate 0.00). Agreement is 1.00 for Qwen on this scorable subset. Method counts are reported in the supplementary material.

Scorable fraction. Agreement on the scorable subset does not imply that all raw outputs are scorable. We therefore report the *scorable fraction* (parseable or extractable) separately by dataset and model in the supplementary material. For example, in the phase-transition logs, Qwen's strict parse-ok rate is near zero, while the scorable fraction via extraction is substantially higher.

Limitations. Extraction validation is limited by the manual label coverage and by the diversity of output formats; we report failure modes and do not claim universal extractor correctness.

B Extraction validation addendum (appendix)

What extraction does *not* do. The extractor is not a learned model and does not attempt to "reason" about evidence. It only attempts to identify the model's *stated* final verdict and confidence when the response is not valid JSON. This is important for interpretation: extraction can recover scoring fields from messy formatting, but it cannot recover an answer that is not present in the text (e.g., truncated responses that never state a final verdict).

Common formatting patterns. In our logs we observe several recurring output formats:

- Valid JSON objects with fields `verdict`, `confidence`, and `reasoning`.
- JSON embedded in Markdown code fences (e.g., "‘`json ...` "‘).
- Plain-text headers (e.g., "Final Verdict: GUILTY").
- Mixed outputs that include a reasoning paragraph and a final verdict line.

The extractor uses prioritized regex rules to handle these cases, falling back to conservative heuristics only when explicit patterns are absent.

Manual labeling protocol (operational details). Manual labels are collected by reading `raw_response` and selecting the verdict that the model explicitly states as its final answer. Annotators do not judge whether the verdict is correct and do not rely on the juror evidence content. If the output contains conflicting final verdict statements, the row is labeled `AMBIGUOUS`. If no final verdict is present, the row is labeled `UNSUPPORTED_FORMAT`.

Agreement interpretation. Agreement should be interpreted in conjunction with scorable fraction:

- High agreement on the scorable subset indicates that extraction is a reliable proxy for manual reading when a final verdict is present.
- A low scorable fraction indicates that many outputs cannot be scored because the model does not emit a final verdict (e.g., truncation), regardless of extractor quality.

We report both quantities to avoid overclaiming about extraction reliability.

How to reproduce extraction validation. All files required for reproduction are included in the released artifact bundle. The agreement pipeline reads the labeled sheet and writes:

- `extraction_reliability_summary.md`

- an agreement summary table
- a confusion breakdown by extraction method
- an extraction-method distribution table

The scorable fraction report is computed from saved logs using a released reporting script.

C Rationality vs. Salience Controls

Motivation. A primary reviewer objection is that minority override may simply reflect reasonable evidence weighting: if one juror provides decisive evidence, the model *should* follow that minority even when outnumbered. We therefore include a set of controls designed to separate "rational override" from "salience-amplified override."

Control families. We include five control families (R1–R5) that manipulate evidence quality and surface cues:

- **R1 (rational minority):** minority evidence is operationalized as decisive (e.g., exclusionary test) while majority evidence remains weak or circumstantial.
- **R2 (salient but weak minority):** minority statement is verbose/emotional/attention-grabbing but not operationalized as decisive.
- **R3 (strong but plain minority):** minority statement is strong yet short and non-authoritative.
- **R4 (weak but authoritative minority):** minority evidence is weak but framed as coming from an expert/institution.
- **R5 (strong majority vs salient weak minority):** majority evidence is strong and minority is salient but weak; this is intended to test whether salience can override quality.

All conditions preserve the same overall multi-juror format and rotate minority position to isolate recency effects.

Results. Table 2 reports override rates by control condition and model for the legal domain. Full cross-domain results (medical, scientific, corporate, fact-check) follow the same qualitative ordering.

The intent is not to declare model behavior "correct" or "incorrect," but to show that the aggregation regime changes systematically under controlled salience and authority cues, and that overrides track evidence quality rather than surface framing alone when majority evidence is sufficiently strong.

Condition	Model	Override	n
R1 (rational min.)	LLaMA-3.1-8B	1.00	200
R1 (rational min.)	Mistral-7B	0.58	200
R2 (salient+weak)	LLaMA-3.1-8B	0.00	181
R2 (salient+weak)	Mistral-7B	0.00	150
R3 (strong+plain)	LLaMA-3.1-8B	0.73	150
R3 (strong+plain)	Mistral-7B	0.33	150
R4 (weak+authority)	LLaMA-3.1-8B	0.00	150
R4 (weak+authority)	Mistral-7B	0.00	150
R5 (strong maj. vs sal. weak)	LLaMA-3.1-8B	0.00	150
R5 (strong maj. vs sal. weak)	Mistral-7B	0.00	150

Table 2: Rationality control results (legal domain, pooled over positions). R1 and R3 involve objectively decisive minority evidence; overrides there can be rational. R2 and R4 test salience/authority without quality; low override rates confirm models are not unconditionally swayed by framing when majority evidence is strong.

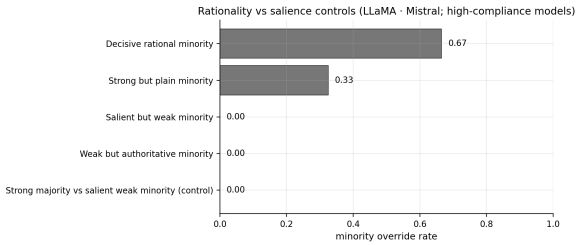

Figure 7: **Rationality vs. salience controls (LLaMA-3.1-8B and Mistral-7B; high-compliance models only).** Override rates for R1–R5 conditions, averaged over LLaMA and Mistral. Qwen excluded due to low format compliance (≈0.1% parse-ok) and anomalous R4 result (≈0.62). R1 (decisive minority, neutral framing) produces high override rationally. R2 (salient but weak) and R4 (weak but authoritative) produce low override when majority evidence is strong, indicating models are not unconditionally authority-biased.

How we interpret overrides in controls. We interpret minority-following outcomes differently depending on the control family:

- In R1 and some instances of R3, minority-following outcomes can be rational (decisive evidence).
- In R2 and R4, minority-following outcomes are more plausibly attributable to salience or authority framing rather than evidence quality.
- In R5, minority-following outcomes (if present) are the strongest evidence for salience-driven dominance over numerical consensus and evidence quality.

This distinction is used to scope claims conservatively and to avoid overclaiming about "reasoning failure."

D Domain generalization details (appendix)

Goal. We test whether the qualitative aggregation pattern observed in the semantic regime persists across multiple domain surface forms. The objective is not to model real professional decision-making, but to vary the lexical and topical "skin" of the evidence statements while keeping the underlying structure and manipulations constant.

Domains. We include five domain surface forms:

- **Legal:** suspect/case framing with evidence types such as camera footage, fingerprints, and alibi.
- **Medical:** test-result framing with screening vs confirmatory signals.
- **Scientific review:** claims and replication signals framed as reviewer comments.
- **Corporate audit:** compliance-risk framing with control findings and audit trails.
- **Fact-checking:** claim verification framing with sources and contradictions.

In all domains, jurors provide short evidence statements and local verdict labels; minority position is rotated.

Reported artifacts. We export domain summaries and supporting tables:

- Domain override-rate table (supplementary material)
- Domain confidence summary (supplementary material)
- Domain variance summary (supplementary material)
- Domain summary table (LaTeX; supplementary material)

These artifacts report override rate by (model, domain, condition) and summarize cross-seed variability.

Interpretation and limitations. Domain generalization in this work should be interpreted conservatively: the domains are represented by synthetic templates, and the key controlled factors (strength, authority, recency) are held constant across domains. A positive result therefore supports *robustness to surface form variation* under controlled structure, not ecological validity in the corresponding real-world domain.

Connection to main-paper claims. The main paper uses domain generalization to support a limited robustness claim: the qualitative ordering of conditions (salient/authority minority > strong-majority control) is not specific to a single topic. Domain-specific effect sizes are treated as descriptive and are reported in appendix tables.

E Full condition suite (A–E) reporting (appendix)

Why full tables are in the appendix. The main paper focuses on a small number of contrasts that are easiest to interpret and hardest to explain away (A/E vs B). For completeness and transparency, we provide the full A–E condition suite outputs in appendix artifacts. This reduces the incentive to cherry-pick and enables independent auditing of condition-wise patterns.

Conditions (A–E).

- **A: Strong Minority.** Majority evidence is operationalized as Circumstantial; minority evidence is operationalized as Direct/Decisive.
- **B: Weak Minority (Control).** Majority evidence is operationalized as Direct; minority evidence is Circumstantial/defeasible.
- **C: Equal Evidence.** Majority and minority evidence are operationalized at similar (Moderate) strength levels.
- **D: Emotional Minority.** Minority uses emotional/urgent framing while majority is neutral/factual.
- **E: Authoritative Minority.** Minority is framed as an expert/institutional reviewer while majority is generic.

All conditions preserve the same overall structure (5 jurors, 4–1 split, minority rotation).

Aggregate replication outputs. We provide seed-aggregated tables (mean override, pooled override, and confidence intervals when available):

- Aggregate override-rate table (supplementary material)
- Aggregate confidence table (supplementary material)
- Seed-variance table (supplementary material)

These tables are sufficient to regenerate the core condition contrasts and heatmaps.

Per-model comparisons. For a compact per-model view, we provide a LaTeX table and a heatmap in the aggregate outputs directory:

- Conditions-by-models table (LaTeX; supplementary material)
- Models×conditions heatmap (supplementary figure)

How to audit for cherry-picking. To verify that main-paper claims are not selectively filtered, a reviewer can:

1. confirm that all seeds (0–4) are present in the replication directories,
2. confirm that aggregated tables include all conditions A–E,
3. confirm that plots are derived from the full aggregate CSVs.

The figure provenance mapping provides explicit pointers from each main-paper figure to the specific CSV inputs.

F Reproducibility and responsible NLP checklist (appendix)

This appendix summarizes reproducibility-relevant details in an ARR/ACL-friendly format. The goal is to make it easy for reviewers to understand what was run, what was saved, and how to regenerate the reported artifacts.

Scope and data. All scenarios are synthetic templates. No human subjects are involved. We do not use private or sensitive data. The juror statements and verdict tags are constructed to enable controlled manipulation of evidence strength, authority cues, and position.

Models and providers.

- Open-weight models are queried via DeepInfra: Meta-Llama-3.1-8B-Instruct, Qwen2.5-7B-Instruct, Mistral-7B-Instruct-v0.3.
- Frontier validation (optional subset) is run via OpenAI when API keys are available: gpt-4o-mini and gpt-4.1-mini.
- Provider/model identifiers are logged in each record to avoid ambiguity.

Decoding settings. We log decoding hyperparameters for every run. Unless otherwise specified, we use temperature 0.7 and a short maximum output length to encourage concise responses. We also log per-call seeds when supported by the provider and preserve the run seed for reproducibility of the sweep order and sampling.

Randomness and seeds. We run replication seeds 0–4. Each configuration (model, condition, position) is repeated multiple times per seed. Aggregation is performed across seeds; we report across-seed variance and pooled estimates in the aggregate outputs directory.

What is logged. For every model call we log:

- prompt text and prompt template identifier,
- model/provider identifiers and decoding settings,
- raw model response text,
- parsed fields when strict JSON is produced,
- extracted verdict and confidence fields when strict parsing fails,
- per-run metadata (timestamp, seed, minority position, condition type, domain).

Raw responses are preserved unchanged; extraction adds derived fields only.

Primary metrics. The primary reported metric is minority override rate: the fraction of trials where the model's final verdict matches the minority juror. Secondary analyses report confidence drift and position sensitivity.

Statistical reporting. We summarize uncertainty using confidence intervals and across-seed variance. For the phase-transition grid we provide a cell-level completeness report and deterministic exclusion rules for low-reliability cells.

Extraction caveats. Some models exhibit lower strict-format compliance. We therefore:

- preserve the original parse flag (`parse_ok`),
- score outputs via a rule-based extractor when parsing fails,
- validate extraction on a manually labeled *scorable* subset,
- report the scorable fraction separately by dataset/model.

This prevents conflating "agreement on scorable outputs" with "all outputs are scorable."

Artifact inventory. Key artifact directories:

- aggregate replications: aggregate tables + plots (artifact bundle)
- domain generalization: domain robustness tables + plots (artifact bundle)
- phase transition: phase-grid tables + completeness reports (artifact bundle)
- extraction validation: labeling sheet + agreement summaries (artifact bundle)
- frontier validation: frontier subset tables + plots (artifact bundle)

Main-paper figures are included in the submission package and can be regenerated from the released aggregate tables using the provided scripts (artifact bundle).

Compute budget (high level). We report the number of runs by configuration and the number of total calls in the output metadata. Because API providers may change pricing, we avoid reporting cost claims; instead we report call counts and settings so compute requirements can be estimated.

Intended use and limitations. This work is a controlled measurement study. We do not recommend using the studied prompts for real decision-making, and we avoid normative claims about correctness in high-stakes domains. The goal is to characterize aggregation behavior under controlled manipulations, not to simulate real juries, courts, or medical decisions.

G Compute budget and run accounting (appendix)

Principle. We report run accounting in terms of *call counts and settings* rather than cost. Provider pricing can change over time; call counts and decoding settings are stable and enable independent estimation of computational requirements.

Replication suite. The core replication suite runs (model $\times$ condition $\times$ position $\times$ repetitions) across seeds 0–4. Each call is logged with timestamp, seed, temperature, max tokens, and prompt template ID. Aggregation outputs are computed from the saved per-seed JSON files.

Phase-transition suite. The phase-transition suite performs a controlled sweep over majority/minority strength levels, authority levels, minority positions, and domains. We provide:

- a sweep specification (supplementary material),
- per-row logs (artifact bundle),
- a cell-level completeness report (supplementary material).

This makes it possible to audit exactly which cells were run and how many examples each cell contains.

Frontier validation subset. The frontier validation suite is intentionally limited in scope: it covers the symbolic null, the A/B/E semantic contrast, and a small phase-grid slice. We log provider and model identifiers for each run and save both a per-condition summary table and raw records CSV. This is designed as a robustness check rather than a full benchmark.

How to compute total calls. Total calls can be computed by counting logged records:

- replication records (artifact bundle)
- phase-transition records (artifact bundle)
- rationality control records (artifact bundle)
- frontier validation records (artifact bundle)

Compute-related limitations. Because our goal is controlled measurement rather than exhaustive scaling, we prioritize breadth of conditions and replication over extremely large juries or very long contexts. Future work could extend the harness to longer contexts and larger juries, but such extensions would increase compute substantially and may introduce additional confounds.

H Notation (appendix)

We use $J{=}5$ jurors with a fixed 4–1 majority/minority split. Minority position $p \in \{1, \ldots, 5\}$ indicates the ordering of the minority juror in the prompt. Minority override rate is the fraction of trials where the model's final verdict matches the minority juror's verdict. For a fixed configuration cell (e.g., model $\times$ condition $\times$ position), if k of n scored trials match the minority verdict, we report:

$$\hat{p} \;=\; \frac{k}{n}. \tag{1}$$

When reporting pooled rates, we weight by the number of trials in each configuration.

Uncertainty for proportions. When reporting confidence intervals for a proportion, we use a binomial confidence interval (Wilson score interval; 95%):

$$\hat{p}_{\text{Wilson}} \pm \frac{z}{1+\frac{z^2}{n}} \sqrt{\frac{\hat{p}(1-\hat{p})}{n} + \frac{z^2}{4n^2}}, \qquad z = 1.96. \tag{2}$$

For aggregate quantities across seeds, we additionally report bootstrap uncertainty in the supplementary material.

Core contrast (effect size). To summarize the main semantic contrast, we report a simple difference in override probability between a salient-minority condition and the strong-majority control:

$$\Delta \;=\; \hat{p}(\text{Strong Minority}) - \hat{p}(\text{Weak Minority Control}). \tag{3}$$

We treat Δ as a descriptive effect size and report its across-seed variability (supplementary material).