

Consensus Drift: Semantic Salience Can Override Majority Evidence in LLM Aggregation

Gokul Srinath Seetha Ram

Independent Researcher

s.gokulsrinath@gmail.com

Rashmi Elavazhagan

Independent Researcher

rashmie30@gmail.com

Anirudh Vasudevan

Independent Researcher

anirudhvasudevan11@gmail.com

Abstract

We study how language models aggregate conflicting evidence in controlled multi-juror settings. As a null control, symbolic majority voting, where jurors provide only YES/NO votes, yields near-zero minority override across models, confirming this is not an artifact of the prompt family. Under semantic aggregation, where jurors instead provide short evidence statements paired with local verdicts, override becomes strength-dependent and position-sensitive: rates rise 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 can modulate this effect in some contexts, but do not universally dominate evidence strength. We validate these patterns with seed replications, cross-domain templates, phase-transition mapping across majority strength, minority strength, and position, and rationality-vs-salience controls. We further confirm the pattern on a larger open-weight model (Llama-3.3-70B-Instruct) and the GPT-4.1 flagship model: Llama-3.3-70B stays close to the smaller-model baseline, while GPT-4.1 shows a substantially larger override magnitude, indicating the effect’s size is model-dependent even as its qualitative shape persists. We emphasize that evidence “strength” here is a synthetic, operationalized manipulation rather than real-world ground truth, and that minority-following can be rational when minority evidence is genuinely 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., 2023). In these settings, evidence sources may disagree (Longpre

et al., 2021; 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:** Circumstantial/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

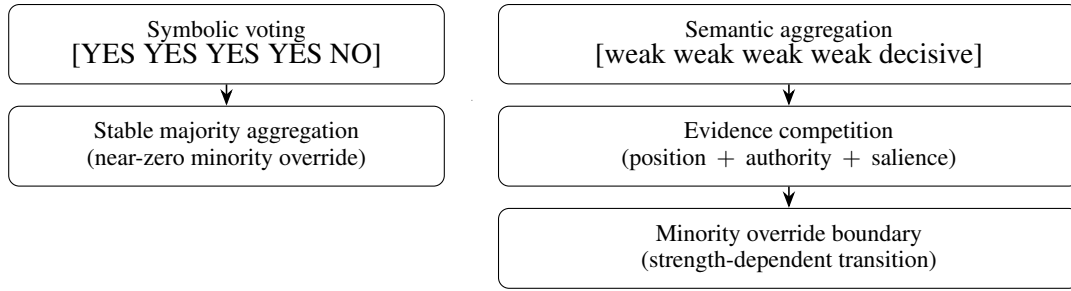


Figure 1: **Conceptual overview: from counting to semantic competition.** Under symbolic voting, aggregation is largely count-based and stable. Under semantic conflict, aggregation becomes a competition among evidence strength and surface cues (position and authority), yielding a strength-dependent minority-override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

keeping content constant (Liu et al., 2024; Pezeshkpour and Hruschka, 2024).

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 integrates multiple retrieved documents (Yang et al., 2018) or reviewers’ comments, a system designer implicitly assumes proportionality—repeated corroborating evidence should not be easily overturned by a single contradictory snippet unless it is sufficiently strong. Positioning and salience effects distort LLM outputs in evaluation (Zheng et al., 2023; Wang et al., 2024) and selection (Zheng et al., 2024) tasks; our harness tests whether analogous effects appear in evidence aggregation, mapping when 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 and an extraction-audit protocol for format-noncompliant outputs.

2 Related Work

We position this work as a controlled behavioral study of *evidence aggregation* in language mod-

els. 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., 2021; 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 (2024) 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., 2024). 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., 2023).

Contrast with pluralistic alignment. Sorensen et al. (2024) argue for preserving a distribution of viewpoints rather than collapsing disagreement into one answer. Our harness is the opposite case: the jury always outputs a single final verdict, so any information encoded in *how* jurors disagreed is dis-

carded once aggregation completes. This is a scope limitation, not a rebuttal: forced single-decision aggregation is common in deployed pipelines, but our design cannot say whether the salience effects we measure would persist, shrink, or take a different form if the harness instead preserved a distribution over juror-implied outcomes. Extending the harness to report a distribution rather than a collapsed verdict is a natural next step.

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 (Llama Team, AI @ Meta, 2024), Qwen2.5-7B-Instruct (Yang et al., 2024; Qwen Team, 2024), and Mistral-7B-Instruct-v0.3 (Jiang et al., 2023; Mistral AI, 2024). We addi-

tionally completed a frontier-validation run on OpenAI models GPT-4o-mini and GPT-4.1-mini (OpenAI, 2025a), and, as described below, a larger confirmatory run on GPT-4.1. To validate our findings at larger scale, we further ran a confirmatory full-matrix suite (semantic A/B/E, symbolic, and label-vocabulary control, each at 3 temperatures, plus 5 authority phrasings; 5,000 trials per model) on a 70B-parameter open-weight model, Llama-3.3-70B-Instruct (Llama Team, AI @ Meta, 2024), and on the GPT-4.1 flagship model (OpenAI, 2025b). Llama-3.3-70B-Instruct is hosted and queried via DeepInfra (the same route as the smaller open-weight models above); it was not run on local or institutional HPC infrastructure. Results are reported in Section 4.6.

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 undergoes an extraction-audit protocol on a scorable subset (Liang et al., 2023), but independent manual validation of this protocol is not yet complete.

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. Among scorable outputs, symbolic aggregation yields near-zero minority override: models overwhelmingly select the majority verdict (Figure 2). Strict parser compliance on the underlying 750 raw trials is uneven by model (LLaMA 43.2%, Qwen 8.8%, Mistral 91.2%), and the plotted source is a single signal-discovery file with per-call seeds rather than the five-run replication suite used for the semantic conditions; we therefore do not claim five independent symbolic replication seeds. This null result is important because it demonstrates that minority override is not an inevitable artifact of the prompt format (Webson and Pavlick, 2022) among the outputs we could score. 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.

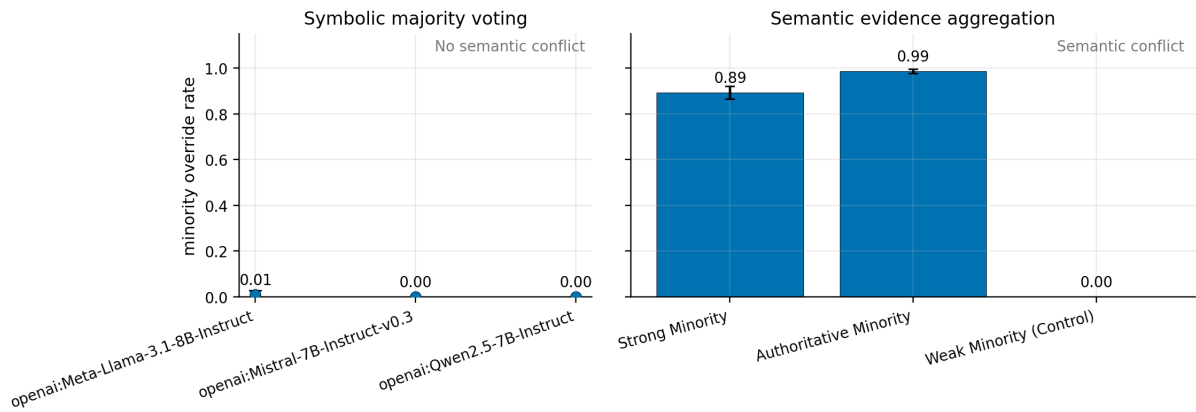


Figure 2: **Symbolic vs. semantic aggregation regime.** Minority override rate (fraction of trials selecting the minority verdict) across symbolic voting (near-zero) and semantic evidence conditions. Symbolic aggregation is largely stable; semantic conflict can yield substantially higher override rates.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

Finding. Semantic aggregation exhibits a marked regime shift for models with high format compliance (Figure 3). For LLaMA-3.1-8B, condition A is consistently 1.00 across nearly all trials and condition E likewise yields 1.00, while the control condition B yields 0.00 across all positions. We report point estimates rather than an interval here because the normal-approximation intervals used in several main figures give misleadingly zero-width bounds at 0%/100%; a Wilson or bootstrap interval (Appendix H) should be used for a rigorous uncertainty estimate at these boundary rates. Mistral-7B-Instruct-v0.3 shows a similar but more position-dependent contrast (condition A ranges from approximately 0.11 to 1.00 across positions, pooled ≈ 0.795 ; B: 0.00). These contrasts replicate across 5 independent seeds.

Qwen caveat. Qwen2.5-7B-Instruct has near-zero strict JSON parse compliance ($\approx 0.1\%$; $\approx 34\%$ scorable via extraction) and an anomalous recovered pattern (control condition B: 0.23–0.85 override across positions, not near-zero). We treat Qwen as exploratory and exclude it from main-paper claims; LLaMA and Mistral drive the pooled figure, with Qwen’s breakdown in the appendix.

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, 2024). Override direction differs by model: Mistral shows a recency effect ($r = 0.38$, $p < 0.0001$, $n = 250$, seed 0), LLaMA a primacy effect ($r = -0.29$, $p < 0.0001$,

$n = 250$). These correlations pool all five conditions without controlling for condition (uneven counts by position), so this is not a clean isolated recency test, and directionality, while consistent across seeds, is not significant in every seed. We treat position as a *modulator*: Figure 4 shows strength/authority cues shift override probability even with position held constant. 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 majority/minority strength levels (weak/medium/strong/decisive) and authority cues, treating strength as an experimental manipulation defined by a deterministic rubric (Ribeiro et al., 2020) (full rubric and per-template assignments 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. In high-compliance models, authority cues can act as an amplifying factor in some semantic-strength contexts, shifting the boundary so that minority influence increases at lower strength gaps; phase-grid authority effects are otherwise mixed, and the weak-authority control (R4) is near zero for LLaMA and Mistral, so we do not treat authority as a generally

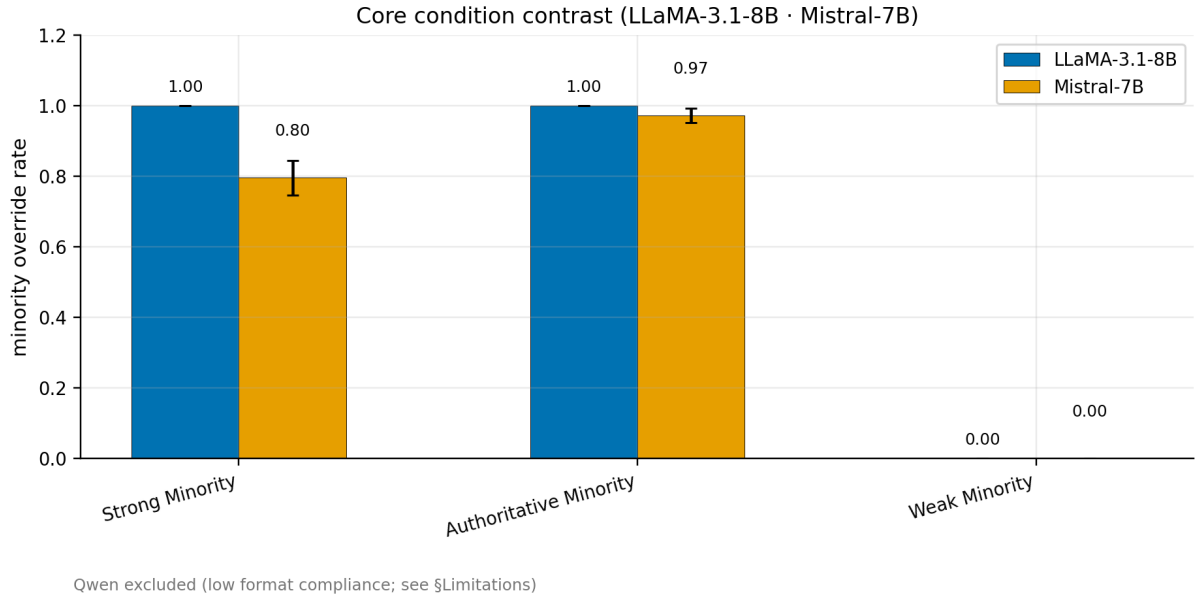


Figure 3: **Core semantic condition contrast (LLaMA-3.1-8B and Mistral-7B)**. Override rates for Strong Minority (A), Authoritative Minority (E), and Weak Minority Control (B), shown per model with 95% CIs. Both models show the same ordering: $A/E \gg B$. Qwen excluded due to low format compliance ($\approx 0.1\%$ parse-ok; see §4.7). The control condition (B) remains at near-zero, confirming that override is contingent on controlled evidence properties. Limitation: conditions use controlled synthetic templates.

amplifying effect. 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.

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. A small mini-scale frontier suite on GPT-4o-mini and GPT-4.1-mini replicates the qualitative $A/E > B$ ordering on both models (OpenAI, 2025a): GPT-4.1-mini shows near-perfect discrimination (A: 1.00, E: 1.00, B: 0.00),

while GPT-4o-mini shows a graded response (A: 0.50, E: 0.59, B: 0.00); full per-condition n are in Appendix E. Section 4.6 extends this to a full confirmatory matrix (5,000 trials/model) on a 70B open-weight model and the GPT-4.1 flagship.

4.6 Flagship-model validation: Llama-3.3-70B and GPT-4.1

Beyond the mini-scale frontier subset above, we ran a full confirmatory matrix (semantic A/B/E, symbolic, and label-vocabulary control, each at temperatures $\{0.0, 0.7, 1.0\}$, plus five authority phrasings) on a 70B-parameter open-weight model, Llama-3.3-70B-Instruct, and on OpenAI’s GPT-4.1 flagship model: 5,000 trials per model, 10,000 total, 100% strict-JSON parse compliance for both (no extraction needed). Proportions use Wilson 95% CIs, n = scored trials.

Symbolic null and vocabulary control. Both models replicate the symbolic null exactly (0/250 at every temperature, both models). The label-vocabulary control – relabeling verdict tokens while preserving evidence content, closing a previously unfulfilled commitment – is likewise a flat 0/250 null for both models at every temperature, indicating semantic overrides are not vocabulary artifacts.

Condition	Juror	Evidence statement	Verdict
A: Strong Minority	J1	The suspect was seen near the building (unconfirmed).	GUILTY
	J2	A witness reported suspicious activity (uncorroborated).	GUILTY
	J3	Phone metadata weakly matches part of the timeline.	GUILTY
	J4	Financial records suggest a possible motive (indirect).	GUILTY
	J5	DNA evidence conclusively excludes the suspect.	NOT GUILTY
B: Weak Minority (Control)	J1	Multiple independent camera angles show the suspect entering the building at the relevant time.	GUILTY
	J2	Fingerprint evidence from the scene matches the suspect with high certainty.	GUILTY
	J3	The suspect’s phone location places them inside the building during the incident.	GUILTY
	J4	A verified receipt and video confirm the suspect purchased the tool used in the incident shortly beforehand.	GUILTY
	J5	The suspect denied involvement.	NOT GUILTY
C: Equal Evidence	J1	A verified camera image resembles the suspect (not definitive).	GUILTY
	J2	A partial fingerprint is consistent with the suspect (not unique).	GUILTY
	J3	The suspect’s phone was near the area at the relevant time (approximate).	GUILTY
	J4	A financial transfer suggests a possible motive (ambiguous).	GUILTY
	J5	Another verified camera image shows someone else entering at the relevant time.	NOT GUILTY
D: Emotional Minority	J1	The suspect was seen near the building (unconfirmed).	GUILTY
	J2	A witness reported suspicious activity (uncorroborated).	GUILTY
	J3	Phone metadata weakly matches part of the timeline.	GUILTY
	J4	Financial records suggest a possible motive (indirect).	GUILTY
	J5	Distressed witness (emotional): I am certain the suspect is innocent. Please do not convict.	NOT GUILTY
E: Authoritative Minority	J1	The suspect was seen near the building (unconfirmed).	GUILTY
	J2	A witness reported suspicious activity (uncorroborated).	GUILTY
	J3	Phone metadata weakly matches part of the timeline.	GUILTY
	J4	Financial records suggest a possible motive (indirect).	GUILTY
	J5	Lead forensic analyst: DNA evidence conclusively excludes the suspect.	NOT GUILTY

Table 1: **Complete five-juror panels for all five semantic conditions** (legal domain; verbatim from consensus_drift/prompts.py’s generate_semantic_case, 5 jurors per condition, 4 vs. 1 majority/minority split). Minority is shown at position 5, the position used for phase-transition and rationality-control examples elsewhere in the paper; minority position is rotated 1–5 in the actual trials. No text is abbreviated or paraphrased.

Semantic A/E vs. B, and a magnitude surprise.

Semantic B is 0/250 for both models at all temperatures. Semantic A (weak-majority/strong-minority) is 0.200–0.204 for Llama-3.3-70B (temperature-flat) but 0.628–0.756 for GPT-4.1, *decreasing* with temperature (0.756→0.712→0.628). Semantic E (authoritative-minority) is 0.200–0.212 for Llama-3.3-70B but a full 1.000 (250/250, CI [0.985,1.000]) for GPT-4.1 at every temperature. We report this directly rather than smoothing it into the existing narrative: GPT-4.1 overrides far more aggressively than Llama-3.3-70B or the smaller GPT-4.1-mini/GPT-4o-mini models (see the frontier-subset numbers below); scale and provider do not predict effect magnitude, only the qualitative A/E > B direction replicates cleanly across all six models tested.

Position concentration (sanity check).

For Llama-3.3-70B, semantic A/E overrides are concentrated almost entirely at minority position 5: 300/300 across all temperatures and both conditions at position 5, ≈0 at positions 1–4, reproducing a position-recency effect from prior internal auditing and validating our scoring pipeline.

Authority phrasings and temperature. Across five authority framings, Llama-3.3-70B is flat and low (20.0%–27.6%); GPT-4.1 is high and variable (62.4%–100%, lowest for “examiner”, ceiling for “forensic”/“detective”), so its authority-sensitivity is not a single-phrasing artifact. Temperature leaves symbolic/vocabctl/semantic-B at the 0.000 floor for both models and modestly affects semantic A/E only for GPT-4.1 (monotone decrease on A; ceiling on E regardless of temperature).

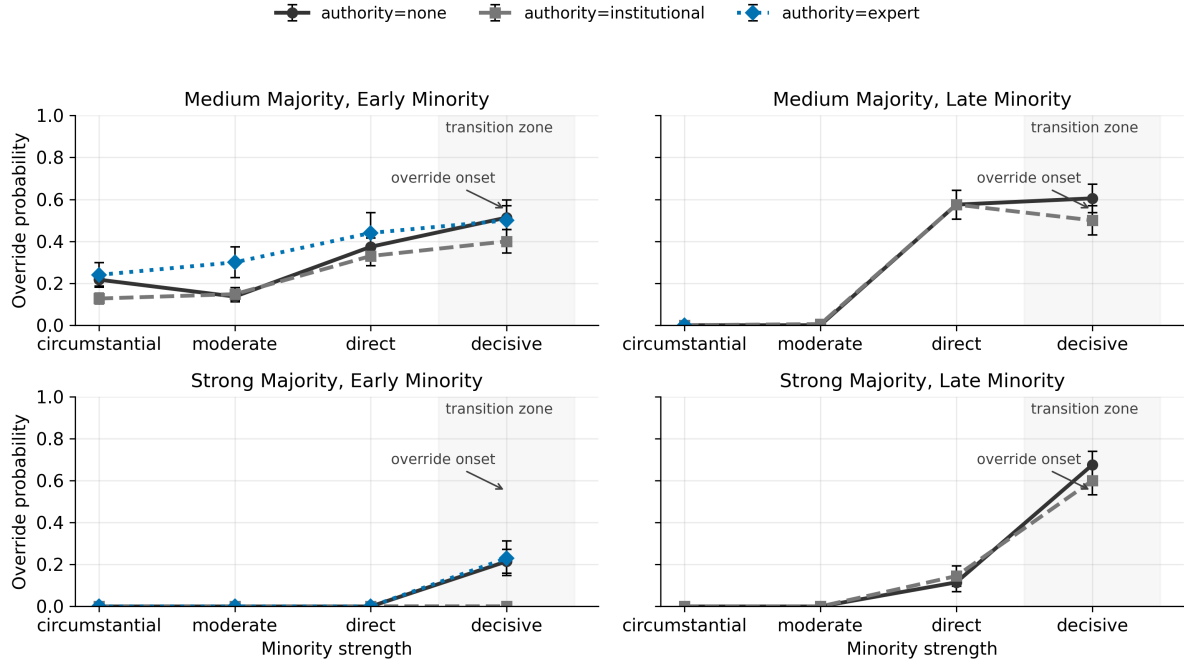


Figure 4: **Phase-transition characterization.** Override probability across a majority-strength $\times$ minority-strength grid, with authority and minority-position modulation. Override follows a strength-dependent boundary: as minority strength increases relative to majority strength, override probability 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.

4.7 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 = 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); treated as exploratory given the low scorable fraction. Results above are legal-domain only: the log also holds medical (1,500 rows) and fact-checking (208 rows) but no scientific-review/corporate-audit rows, omits prompt/raw-response/timestamp/seed/run-id meta-data, and is append-only without deduplication, so legal-domain counts include duplicate appended runs. We restrict claims to logged domains and do not describe these controls as cross-domain (Ap-

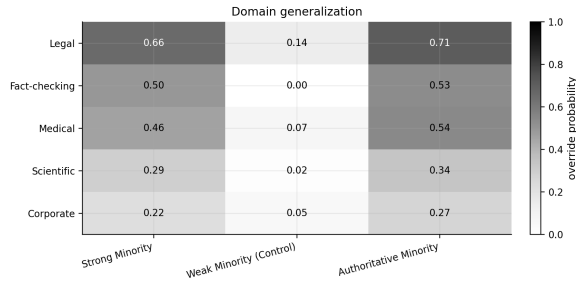


Figure 5: **Cross-domain generalization heatmap.** Mean override rate (over models) for the core semantic conditions across five synthetic domain surface forms. The qualitative ordering (Strong/Authoritative Minority > Weak Minority Control) persists across domains with domain-dependent effect sizes. This unweighted per-model average gives Qwen’s low- n scorable subset (≈ 12 samples in some cells, versus $\approx 80\text{--}95$ for LLaMA/Mistral) equal weight to the two high-compliance models; the heatmap should therefore be read as confirmatory for LLaMA and Mistral, with Qwen’s contribution exploratory only. Limitation: domains are synthetic templates; this is robustness to surface form variation, not ecological validity.

pendix C).

4.8 Extraction caveats and measurement reliability

As introduced in §3.4 (Logging and parsing), manual extraction validation was not completed; extraction-dependent Qwen results are therefore exploratory and excluded from confirmatory claims. We report the *scorable fraction* (parseable or extractable) by dataset and model in the supplementary material; a high scorable fraction does not by itself imply that extraction has been independently validated against human labels (Appendix H discusses the analogous CI-methodology caveat for interval estimates).

5 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., 2023); 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 effect sizes (Liu et al., 2024; Li et al., 2024). Confidence values are self-

reported, not calibrated probabilities (Kadavath et al., 2022; Zhao et al., 2021). Semantic experiments use GUILTY/NOT_GUILTY labels, symbolic use YES/NO; we emphasize within-regime controls and treat the symbolic null as a qualitative, not strict quantitative, cross-regime contrast.

6 Ethics

This work uses synthetic scenarios and involves no human subjects; we frame results as controlled evidence aggregation behavior in LLMs and avoid political or governance claims. We do not recommend using the prompts or scoring procedures in isolation for real decision-making: reported effects characterize model behavior under controlled manipulations, not normative judgments about correctness in legal, medical, or other high-stakes settings. Findings about authority cues and ordering could be misread as guidance for persuasive prompting; we emphasize limitations and the null control to mitigate this.

7 Conclusion

We presented a controlled empirical study of evidence aggregation dynamics in language models under semantic conflict. Symbolic majority voting is largely stable with near-zero minority override; when jurors instead 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. Evidence strength is operationalized, not ground truth; we preserve raw outputs, provide extraction caveats, and leave independent extraction validation to future work.

References

- Shengnan An, Zexiong Ma, Zeqi Lin, Nanning Zheng, and Jian-Guang Lou. 2024. Make your LLM fully utilize the context. *arXiv preprint arXiv:2404.16811*.
- Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, Carol Chen, Catherine Olsson, Christopher Olah, Danny Hernandez, Dawn Drain, Deep Ganguli, Dustin Li, Eli Tran-Johnson, Ethan Perez, Jamie Kerr, Jared Mueller, Jeffrey Ladish, Joshua Landau, Kamal Ndousse, Kamile Lukosuite, Liane Lovitt, Michael Sellitto, Nelson Elhage, Nicholas Schiefer, Noemi Mercado, Nova DasSarma, Robert

- Lasenby, Robin Larson, Sam Ringer, Scott Johnston, Shauna Kravec, Sheer El Showk, Stanislav Fort, Tamera Lanham, Timothy Telleen-Lawton, Tom Conerly, Tom Henighan, Tristan Hume, Samuel R. Bowman, Zac Hatfield-Dodds, Ben Mann, Dario Amodei, Nicholas Joseph, Sam McCandlish, Tom Brown, and Jared Kaplan. 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. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2292–2307.
- 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, Anuoluwapo Aremu, Antoine Bosselut, Khyathi Raghavi Chandu, Miruna-Adriana Clinciu, Dipanjan Das, Kaustubh Dhole, Wanyu Du, Esin Durmus, Ondřej Dušek, Chris Chinenye Emezue, Varun Gangal, Cristina Garbacea, Tatsunori Hashimoto, Yufang Hou, Yacine Jernite, Harsh Jhamtani, Yangfeng Ji, Shailza Jolly, Mihir Kale, Dhruv Kumar, Faisal Ladhak, Aman Madaan, Mounica Maddela, Khyati Mahajan, Saad Mahamood, Bodhisattwa Prasad Majumder, Pedro Henrique Martins, Angelina McMillan-Major, Simon Mille, Emiel van Miltenburg, Moin Nadeem, Shashi Narayan, Vitaly Nikolaev, Andre Niyongabo Rubungo, Salomey Osei, Ankur Parikh, Laura Perez-Beltrachini, Niranjan Ramesh Rao, Vikas Raunak, Juan Diego Rodriguez, Sashank Santhanam, João Sedoc, Thibault Sellam, Samira Shaikh, Anastasia Shimorina, Marco Antonio Sobrevilla Cabezero, Hendrik Strobelt, Nishant Subramani, Wei Xu, Diyi Yang, Akhila Yerukola, and Jiawei Zhou. 2021. The GEM benchmark: Natural language generation, its evaluation and metrics. In *Proceedings of the First 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, Léo Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas 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, Sam Bowman, Stanislav Fort, Deep Ganguli, Danny Hernandez, Josh Jacobson, Jackson Kernion, Shauna Kravec, Liane Lovitt, Kamal Ndousse, Catherine Olsson, Sam Ringer, Dario Amodei, Tom Brown, Jack Clark, Nicholas Joseph, Ben Mann, Sam McCandlish, Chris Olah, and Jared Kaplan. 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 contexts? In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16304–16333.
- 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 Re, Diana Acosta-Navas, Drew A. Hudson, Eric Zelikman, Esin Durmus, Faisal Ladhak, Frieda Rong, Hongyu Ren, Huaxiu Yao, Jue Wang, Keshav Santhanam, Laurel Orr, Lucia Zheng, Mert Yuksekgonul, Mirac Suzgun, Nathan Kim, Neel Guha, Niladri S. Chatterji, Omar Khattab, Peter Henderson, Qian Huang, Ryan Andrew Chi, Sang Michael Xie, Shibani Santurkar, Surya Ganguli, Tatsunori Hashimoto, Thomas Icard, Tianyi Zhang, Vishrav Chaudhary, William Wang, Xuechen Li, Yifan Mai, Yuhui Zhang, and Yuta Koreeda. 2023. [Holistic Evaluation of Language Models](#). *Transactions on Machine Learning Research*.
- 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.
- Llama Team, AI @ Meta. 2024. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.
- Shayne Longpre, Kartik Perisetla, Anthony Chen, Nikhil Ramesh, Chris DuBois, and Sameer Singh. 2021. Entity-based knowledge conflicts in question answering. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7052–7063.
- Mistral AI. 2024. Mistral-7B-Instruct-v0.3 model card. <https://huggingface.co/mistralai/>

- Mistral-7B-Instruct-v0.3. Accessed 2026-08-30.
- OpenAI. 2025a. GPT-4.1 mini model documentation. <https://developers.openai.com/api/docs/models/gpt-4.1-mini>. Accessed 2026-08-30.
- OpenAI. 2025b. GPT-4.1 model documentation. <https://developers.openai.com/api/docs/models/gpt-4.1>. Accessed 2026-08-30.
- Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. *arXiv preprint arXiv:2203.02155*.
- Pouya Pezeshkpour and Estevam Hruschka. 2024. Large language models sensitivity to the order of options in multiple-choice questions. In *Findings of the Association for Computational Linguistics: NAACL*, pages 2006–2017.
- Qwen Team. 2024. Qwen2.5-7B-Instruct model card. <https://huggingface.co/Qwen/Qwen2.5-7B-Instruct>. Accessed 2026-08-30.
- 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. In *Proceedings of the International Conference on Learning Representations (ICLR)*.
- Taylor Sorensen, Jared Moore, Jillian Fisher, Mitchell Gordon, Niloofar Mireshghallah, Christopher Michael Rytting, Andre Ye, Liwei Jiang, Ximing Lu, Nouha Dziri, Tim Althoff, and Yejin Choi. 2024. Position: A roadmap to pluralistic alignment. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, volume 235, pages 46280–46302.
- 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, Lingpeng Kong, Qi Liu, Tianyu Liu, and Zhifang Sui. 2024. Large language models are not fair evaluators. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 9440–9450.
- 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. In *Proceedings of the International Conference on Learning Representations (ICLR)*.
- An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jianxin Yang, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Xuejing Liu, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, Zhifang Guo, and Zhihao Fan. 2024. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.
- 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.
- Zihao 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. In *Proceedings of the International Conference on Learning Representations (ICLR)*.
- Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin,

A Extraction validation (audit status)

We attempted an extraction-audit protocol for extraction-based scoring; a genuine independent manual audit was not completed, and this section reports that status honestly rather than presenting unverified numbers as confirmed agreement.

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.

Attempted labeling protocol. We constructed a labeling sheet intended to hold 100 sampled Qwen outputs and up to 50 additional non-Qwen extracted outputs. However, on inspection the resulting sheet’s `manual_label` field is identical to the extractor’s own output for every row, and the `manual_confidence/notes` fields are blank throughout; a separate noncompliant-output audit sheet has 50 Qwen rows with blank `human_verdict/human_notes` fields. Neither artifact constitutes a completed independent human audit.

Agreement statistics. Because no genuine independent manual labels exist, we do not report a manual-extractor agreement statistic. An earlier draft of this section reported 1.00 agreement; that figure measured the extractor against a copy of its own output and has been removed. We treat all extraction-dependent Qwen results as exploratory and exclude them from confirmatory claims until an independent manual audit is completed.

Scorable fraction. Independent of the agreement question, we 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; this reflects format compliance, not validated extractor correctness.

Limitations. Extraction validation has not been independently completed. We report failure modes

and known caveats but do not claim extractor correctness has been verified against human judgment.

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.

Intended manual labeling protocol (operational details). The intended protocol was for annotators to read `raw_response` and select the verdict that the model explicitly states as its final answer, without judging whether the verdict is correct or relying on the juror evidence content, labeling `AMBIGUOUS` for conflicting final-verdict statements and `UNSUPPORTED_FORMAT` when no final verdict is present. As noted above, this protocol was not carried out independently: the extracted labeling sheet’s `manual_label` column mirrors extractor output rather than reflecting separate human reading, and a separate noncompliant-output audit sheet remains blank in its human-judgment fields.

Agreement interpretation. Because independent manual labels do not yet exist, we cannot report a validated agreement statistic between extraction and human judgment. We instead report the *scorable fraction* on its own: 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 avoid overclaiming extraction reliability until a genuine independent audit is completed.

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 natural 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.

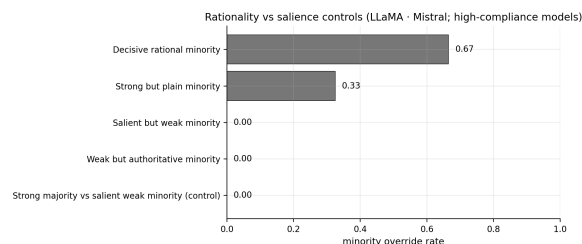


Figure 6: **Rationality vs. salience controls (LLaMA-3.1-8B and Mistral-7B; high-compliance models only).** Override rates for R1–R5 conditions, averaged over LLaMA and Mistral. Qwen excluded due to low format compliance ($\approx 0.1\%$ parse-ok) and anomalous R4 result (≈ 0.62). R1 (decisive minority, 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.

Results. Table 2 reports override rates by control condition and model for the legal domain, which has the largest sample but includes duplicate appended runs from an append-only, non-deduplicated log. The underlying log also contains medical (1,500 rows) and fact-checking (208 rows) rows; scientific-review and corporate-audit domains have no logged rows at all. We therefore report only the domains actually logged and disclose the incomplete provenance (missing prompt/raw-response/timestamp/seed/run-identifier fields) and duplicate-run risk in the legal domain, rather than claiming full cross-domain coverage.

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.

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

Condition	Model	Override	<i>n</i>
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; the legal-domain log includes duplicate appended runs and is not a balanced seeded design). R1 and R3 involve minority evidence operationalized as decisive; 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.

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 (L^AT_EX; 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)

Frontier mini-suite, per-condition numbers.

The mini-scale frontier suite referenced in the main body gives: GPT-4.1-mini A: 1.00 ($n=158$), E: 1.00 ($n=165$), B: 0.00 ($n=177$); GPT-4o-mini A: 0.50 ($n=173$), E: 0.59 ($n=162$), B: 0.00 ($n=165$).

Full five-juror panels for all five semantic conditions (legal domain, minority at position 5). All five complete five-juror panels (conditions A–E) are now given in full in the main body (Table 1); no text is abbreviated there, so this appendix no longer needs to repeat them. See Table 1 for the verbatim juror-level text (source: `consensus_drift/prompts.py`’s `generate_semantic_case`).

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 $\text{\LaTeX}$ table and a heatmap in the aggregate outputs directory:

- Conditions-by-models table ($\text{\LaTeX}$; supplementary material)
- Models $\times$ 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, and Llama-3.3-70B-Instruct.
- Frontier validation was completed via OpenAI for gpt-4o-mini and gpt-4.1-mini (mini-scale subset), and for gpt-4.1 (larger-scale confirmatory matrix, Section 4.6).
- A larger-scale confirmatory matrix (Section 4.6) additionally covers Llama-3.3-70B-Instruct (DeepInfra) and GPT-4.1 (OpenAI): 5,000 trials per model across semantic A/B/E, symbolic, and label-vocabulary-control conditions at 3 temperatures plus 5 authority phrases.
- Provider/model identifiers are logged in each record to avoid ambiguity; Llama-3.3-70B-Instruct records are logged with an unprefixed model identifier confirming DeepInfra (not HPC/Slurm/vLLM) execution.

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. The semantic runner makes 50 calls per model-position per seed and randomly assigns one of the five semantic conditions to each call, so each model-condition-position cell receives approximately 10 calls per seed (realized counts vary by cell) rather than a fixed 50-per-condition factorial design. 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,
- attempted an extraction-audit protocol on a *scorable* subset, though independent manual validation is not yet complete,
- report the scorable fraction separately by dataset/model.

This prevents conflating a high scorable fraction with independently validated extraction correctness.

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).

Licensing. Author-written code (the `consensus_drift/` package, including prompt-template generators and analysis/figure scripts) is released under the MIT license; author-produced data and artifacts (logged trials, aggregate tables, figures) are released under CC BY 4.0. Both license files are included in the submission package. No public repository URL is claimed; the code and data are distributed as part of the submitted 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 makes 50 calls per model-position per seed across seeds 0–4, with one of five semantic conditions randomly assigned to each call (so realized model-condition-position cell counts are approximately 10 per seed, not a fixed 50-per-condition design). 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 $p \in \{1, \dots, 5\}$ 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:

$$\hat{p} = \frac{k}{n}. \quad (1)$$

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%):

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

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

CI-methodology limitation. Wilson intervals are our stated default, but not every figure-generation script in the current artifact bundle uses them; several main-text figures instead use a normal approximation, which produces misleadingly zero-width intervals at 0% or 100% observed rates. We flag this inconsistency here rather than reporting a fabricated exact interval, and we recommend that any reader recomputing intervals use Wilson or bootstrap methods uniformly rather than relying on the normal-approximation figures as-is.

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:

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

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