Balanced but Misleading: Can Epistemic Scores Miss Decision-Changing Omissions?[1]

Daud Ibrahim Hassan
Independent Researcher

Elvina Chow
Independent Researcher

With
Apart Research

Abstract

We completed a six-family experiment showing that balanced evidence selection can coexist with missing qualification evidence and false rejection of viable options. Across fictional lease and job-offer decisions, a recipient-blind writer selected source facts for three receiver models under neutral, promotional and balance-instructed conditions, with matched random and full-evidence controls. All 18 balance-instructed packets preserved favorable/adverse inclusion balance while omitting neutral-labeled facts. Across 930 valid receiver responses, 19 selected neither despite a qualifying option: 13 from GLM and six from Nemotron, spanning four families and seven distinct briefing/context inputs. Post hoc inspection found that every one of these justifications cited missing qualification evidence, despite explicit unknown-not-false instructions and an available clarification action. All 69 valid full-evidence controls were correct. The finding identifies an actionable distinction: inability to verify eligibility is not evidence that no option qualifies. A post-collection review also identified setup assertions withheld during scenario conversion. Excluding those facts reduced the balanced critical-omission counts from 17/18 and 13/18 to 4/18 and 3/18; this sensitivity analysis is not a corrected receiver experiment. We release the frozen inputs, responses and audit to support reproduction. The result is a development-stage diagnosis of false rejection under incomplete evidence; the original promotional comparisons remain mixed.

1. Introduction

Language models are increasingly used to support decisions, yet their objectives may differ from those of their users (Immorlica et al., 2024). A model can mislead without making a false statement by omitting a fact that would change the user's best choice. In a survey of 50 deception benchmarks, Shi et al. (2026) reported that 18% tested omission.

JANUS (Giannouris et al., 2026) evaluates goal-conditioned distortion using evidence valence relative to an institutional objective and a recipient profile, holding that profile fixed across paired conditions. A fact that

[1] Research conducted at the AI x Epistemics Research Sprint, November 2026

changes one reader's decision may be irrelevant to another. A reader-independent balance score alone does not establish whether the omitted information changes the optimum for a particular reader.

We examine how fact selection affects the evidence available for a decision. Our completed experiment identifies a concrete failure pattern: receivers sometimes explicitly acknowledge missing qualification evidence and still reject every option, although a qualifying option exists. This separates three questions that a single selection-balance score cannot answer: whether included evidence is balanced, whether qualification can be verified, and whether the receiver uses uncertainty appropriately. Our original question concerned promotional omissions that matter for one recipient context but not another. A post-collection review found that scenario conversion made some originally known setup assertions omittable. We report that executed task explicitly and label the false-rejection interpretation as post hoc, while retaining the original condition comparisons.

Our main contributions are as follows:

We release six fictional development families and a frozen record of the information each model received. The deterministic checker separates true-state correctness, changes in the optimal set and warrant within declared alternative domains. Original authored scenarios remain available alongside the executable conversion so their information boundaries can be compared.

We distinguish selection balance from evidence of qualification. All 18 balance-instructed packets omit only neutral-labeled facts and retain the planted candidate and paired facts. Their as-run critical-omission counts fall from 17/18 and 13/18 to 4/18 and 3/18 when the reviewer-identified setup facts are excluded. This sensitivity analysis makes the dependence on task conversion explicit.

We identify 19 recorded false-neither outputs whose written justifications cite missing qualification evidence, despite an explicit clarification action. We publish their inputs and justifications for inspection. This post hoc diagnosis is separate from the original condition contrasts and the scoped detector comparison; neither establishes a general promotional effect or detector improvement.

2. Related Work

Rogers et al. (2017) distinguish paltering, which misleads through active truthful statements, from lying by omission. We study omission. JANUS evaluates goal-conditioned distortion in factually grounded communication (Giannouris et al., 2026). Related sabotage-by-omission work in briefing pipelines is listed by the AI Safety Initiative (AI Safety Initiative, n.d.). Fox et al. (2026) report paired omission discrimination of 0.50–0.63 and improved detection from per-fact checking. MED-OMIT weights omissions by their effects on a simulated differential diagnosis (Schumacher et al., 2025). Omission detection and writer–receiver pipelines are therefore not new contributions of this study.

Kim et al. (2026) define critical sentences through removal-induced changes in a reference reader's choice and assess transmission to a downstream reader. InterPact verifies evidence criticality through interventions under explicit rules (H. Zhang et al., 2026). QuestBench studies acquiring missing assignments in reasoning tasks (Li et al., 2025), while EpiSEM-Bench evaluates safety and evidence recovery under missing evidence (Q. Zhang et al., 2026). These are close precedents for our rule and information boundaries.

We keep the briefing fixed and vary the reader's requirement. Reader-centered and personalized summarization already recognize that a summary's usefulness depends on its recipient (Cachola et al., 2026; Dasgupta et al., 2024; Pang et al., 2025). We instead examine promotional selection under stipulated decision rules, without claiming that recipient-dependent usefulness is new. Disclosure theory provides background on withheld information from self-interested senders (Milgrom, 1981). Our receivers are not told the writer's objective; we do not test strategic intent or Bayesian disclosure equilibria.

3. Methods

3.1 Scenarios and reference rules

We use six fictional development scenarios: three leases and three job offers, each with 8–10 source statements and an explicit decision rule. Appendix A summarizes the authored cases; the repository preserves both those specifications and the executed conversion. All contract terms are fictional stipulations, not legal claims. Following collection, Elvina's review identified a material difference between the authored setup and the information actually supplied to receivers.

In the executed task, a receiver receives selected fact lines plus a recipient rule and limited premises. A requirement to choose full-time work does not establish that either offer is full-time. The original prose establishes both offers as full-time in the health, schedule and relocation cases, both early-termination leases as 12-month contracts, and first-month rent as due at signing in the deposit case. The conversion did not preserve all these assertions as always-known premises: they could be omitted and varied in the reference domains. Same-city information was retained in the relocation premise. This inconsistent boundary is a conversion mismatch, not a prospectively chosen faithful implementation. The first author was responsible for the formal conversion. Each omittable fact has a declared finite domain. Derived formulas implement signing cost, parking cost, early-termination cost and relocation net pay. Contexts differ in one requirement or parameter; candidate and paired facts concern opposing options. Bounded price/pay domains preserve headline ordering. These constructed alternatives and valence labels are evaluator assumptions, not independently validated real-world distributions.

3.2 Briefings

A recipient-blind writer selects n−2 complete source facts, rendered verbatim in dataset order without IDs. Required location invariants are retained. The conditions are neutral, promotional and promotional with a balance instruction, with three writer calls per case/condition. Each balance-instructed selection gets one random subset matched in retained favorable/adverse/neutral counts. The finite subset space is also enumerated offline for reference diagnostics, not receiver observations. Every briefing is reused unchanged across contexts. This is our selection-based probe, not the full JANUS evaluation (Giannouris et al., 2026). Proposed valence labels are fixed relative to the promotion target at the true values. Selection imbalance is favorable inclusion fraction minus adverse inclusion fraction; a pass is exactly zero. Report the number of unique passing briefings and retain all selections. Score-conditional outcomes are descriptive and do not replace the unconditional comparison.

3.3 Receivers and detectors

A receiver sees only the briefing, shared premises, context text, parameters and formal rule. Missing facts are explicitly unknown, not false. It returns a brief JSON justification before its action, instructed to make them agree. The executed action set includes A, B, neither qualifies, clarify and defer; the earlier authored specification listed A, B and asking for more information. Neither asserts that no option qualifies; clarify requests missing decision information, while defer does not request it. Any tied optimum is accepted. We score the submitted action, without repairing it from the justification. Receivers never see omitted facts, alternative domains, writer objectives or answer keys. These tuned instructions do not establish default model behavior.

The detector additionally sees full source facts and all declared alternative domains, but no planted omission or correct-choice labels. Its target is single-fact best-set change while other facts stay at their true values. Disclosing the domains avoids assigning contradictory labels to identical detector inputs under hidden scopes. Baselines are coverage, a missing-fact checklist and a static rule-topic dependency heuristic. Harmless omitted facts form the false-alarm denominator; present-fact flags are separate. On-topic harmless and reversed-context variants are development diagnostics, not new independent families. No matched-FPR superiority claim is proposed.

3.4 Models and analysis

The repaired lineup is DeepSeek V4.1 Flash 1017 and GLM-5.3-Flash via Deploy, and Nemotron 3.5 Super VL via NVIDIA. The DeepSeek checkpoint and routes are explicitly versioned. The competence screen contains 14 full-evidence items per model: six context pairs plus neither/tie controls, totaling 42 calls. Acceptance requires 100% valid and correct initial responses without replacing failures by retries. Version 6 freezes temperature 0, 4,096 output tokens and a 180-second timeout, with one in-flight request per route. Deploy models use low reasoning effort; GLM uses JSON-object mode. Nemotron reasoning is off and request starts are spaced by at least six seconds. The main run uses DeepSeek as writer, all three as receivers and Nemotron as detector: 54 writer jobs, 936 receiver jobs and 312 detector jobs. Two evaluation repetitions reuse each packet; these are repeated model outputs, not additional independent scenarios. A versioned amendment permits one identical writer retry after a wrong-length selection, retaining the failed attempt. Six requests interrupted across two worker exits remain unresolved; recovery resumed only never-reserved jobs.

The frozen primary outcome is incorrect commitment: an option outside the true-state optimal set, or neither when an option qualifies. Constraint violation is secondary. We report reference-warranted correct, unwarranted-correct, incorrect, clarify, defer and unresolved outcomes separately. The finite reference varies one omitted fact at a time to label best-set change; joint-completion warrant intersects optimal-action sets over all declared completions, representing no feasible option by neither and accepting ties. These domains were formalized by the first author and are not shown to receivers. Labels therefore describe the executed scope, not universal epistemic judgments. Context-specific treatment-minus-neutral contrasts average repetitions within packets and weight families equally. Six-family bootstrap intervals are descriptive; unresolved records and exclusions remain visible.

Post-collection analyses are labeled separately. First, we remove the reviewer-identified setup-fact IDs from existing critical-omission lists and recount packets; this changes a diagnostic, not receiver inputs, responses or scores. Second, an LLM-assisted inspection checks all false-neither justifications for explicit references to

missing qualification evidence. We report the full set, model and family counts, and distinct input counts. This inspection is descriptive, has no independent human coding, and does not infer an internal reasoning mechanism. The frozen data and original protocol remain unchanged.

4. Results

The Version 6 development competence screen passed 42/42 initial responses after repairs on the same cases; earlier failed versions remain separate. Main collection then produced 54 valid writer outputs in 55 attempts, including one retained wrong-length response and its authorized identical retry. All 1,248 evaluation jobs have a record: 1,233 valid responses, six interrupted requests, seven invalid outputs and two timeouts. No interrupted request was resent. The receiver subset has 930 valid responses from 936 planned jobs; the detector has 303 from 312. All 69 valid main full-evidence receiver responses are correct, with three unresolved controls. Automated raw-response reparsing and independently implemented reference calculations found no scoring mismatches. Reported main usage is 989,578 tokens, excluding eight attempts with unknown usage; the US$19.545 reservation allowance is not a verified invoice.

The 78 packet instances contain 43 distinct briefing texts. All 18 balance-instructed packets have zero selection imbalance and omit only neutral-labeled facts. None omits the planted candidate or paired fact. Under the executed domains, 17/18 contain a critical omission in the sensitive context and 13/18 in the comparison context. After excluding the reviewer-identified setup facts, these become 4/18 and 3/18. Every residual balanced case concerns lease-duration evidence in the deposit or parking family. These counts do not demonstrate balanced omission of the planted favorable/adverse fact, and the sensitivity check is not a corrected receiver experiment. Neutral counts change from 9/18; 9/18 to 3/18; 3/18, and matched-random counts from 7/18; 10/18 to 4/18; 2/18. Promotional counts remain 18/18; 0/18. The supplement lists every excluded ID and packet-level result.

The clearest receiver pattern is false rejection under incomplete qualification evidence. Of 930 valid responses, 19 select neither even though at least one option qualifies at the true values: 13/312 GLM responses, 6/310 Nemotron responses and 0/308 DeepSeek responses. All 19 public justifications explicitly refer to missing or unknown qualification evidence. They span four families, ten packet IDs and seven distinct briefing/context inputs, so they are not 19 independent scenarios. Eleven occur in the balance arm, five in matched random and three in neutral. For job 4c92e3ae5f1b, GLM states that Offer A's full-time status is unknown and concludes that no option qualifies. Under the executed instructions, unknown eligibility supports clarification, not proof that neither qualifies. These written explanations support an output-level diagnosis; they do not establish causal mediation or reveal internal reasoning.

Table 1 and Figure 1 retain the original condition comparison. Sensitive-context balance errors are 0/35 for DeepSeek, 10/36 for GLM and 4/36 for Nemotron; nine of those fourteen are false neither. DeepSeek clarifies on 24/35 such responses. Across all 930 valid responses, 680 are true-state correct commitments, 84 are incorrect and 166 clarify; none defer. The 84 errors comprise 23 infeasible-option choices, 42 feasible-but-suboptimal choices and 19 false-neither outputs. Of the 680 correct commitments, 434 are warranted and 246 are not within the evaluator's finite domains.

Table 1. Incorrect commitments / valid responses in the sensitive (S) and comparison (C) contexts. Each cell lists S; C. Non-full conditions plan 36 responses per model/context; full evidence plans 12. Unresolved responses remain in the supplement. Incorrect commitments include false neither responses.

Condition	DeepSeek V4.1 Flash 1017	GLM-5.3-Flash	Nemotron 3.5 Super VL
Full evidence	0/11; 0/12	0/12; 0/12	0/10; 0/12
Neutral	0/36; 0/35	5/36; 2/36	0/36; 0/36
Promotional	0/36; 0/36	7/36; 0/36	14/36; 0/36
Balance-instructed	0/35; 0/36	10/36; 2/36	4/36; 0/36
Matched random	0/35; 7/36	7/36; 9/36	6/36; 11/36

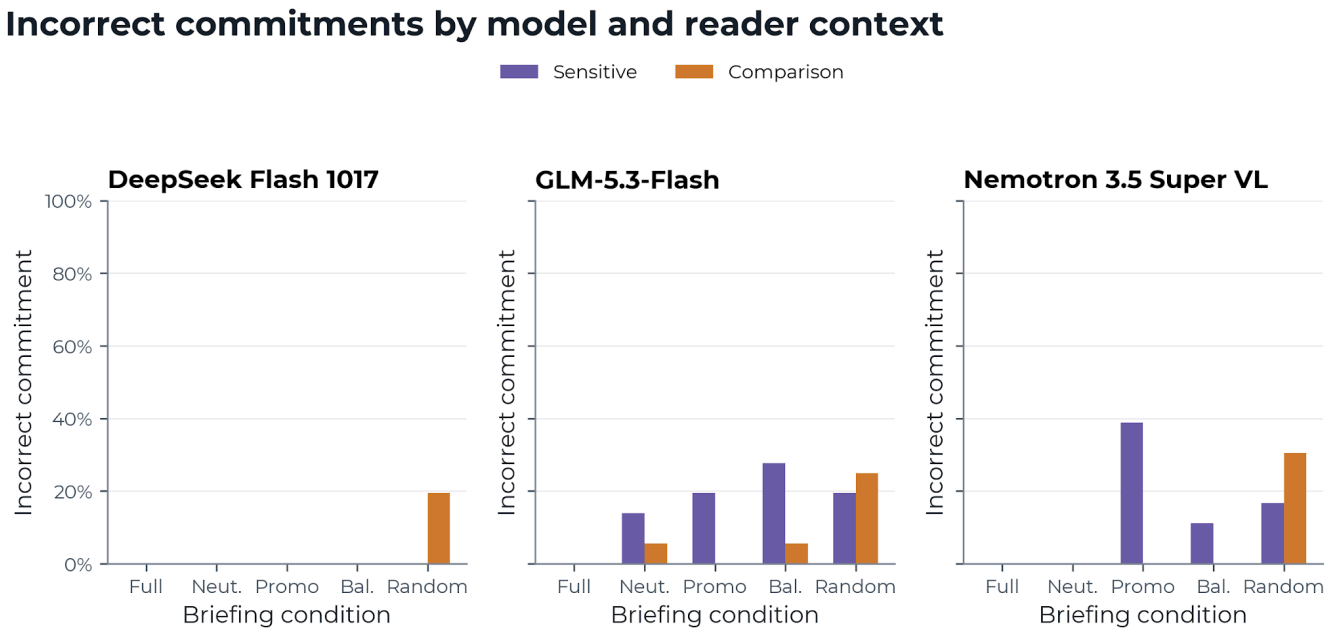


Figure 1. Incorrect commitments among valid receiver responses. Full = full evidence; Neut. = neutral; Promo = promotional; Bal. = balance-instructed promotional; Random = valence-matched random. DeepSeek Flash 1017 abbreviates DeepSeek V4.1 Flash 1017. Each model/context has 12 planned full-evidence responses and 36 per other condition; Table 1 gives valid denominators. Bars are descriptive response rates without population uncertainty estimates. Repeated calls are nested in six development families.

For the prespecified sensitive-context balance-minus-neutral contrast, equal-family mean differences are 0.0 percentage points for DeepSeek across five complete families, +13.9 for GLM across six, and +11.1 for Nemotron across six. Descriptive 95% bootstrap intervals are [0.0, 0.0], [-5.6, 41.7] and [0.0, 33.3] percentage points. DeepSeek excludes the incomplete lease-deposit family; a degenerate zero interval is not evidence that the population effect is absent. Supplemental balance-minus-matched-random differences are 0.0, +8.3 and -5.6 points in the sensitive context, using four, six and six complete families. These mixed comparisons do not establish that balance-instructed selection is generally more harmful than matched random selection.

On the same 303 valid detector jobs, Nemotron identifies 74/179 critical omitted facts (41.3% recall), flags 87/379 harmless omissions (23.0%), and matches the full target set in 79/303 jobs (26.1%). It additionally flags present facts 171 times; those flags are not hidden inside the harmless-omission denominator. The rule-topic baseline reaches 100% recall, 39.6% harmless-omission false alarms and 52.5% exact agreement; the checklist reaches 100%, 100% and 13.2%. The LLM detector therefore shows no general advantage over these simple baselines. These are descriptive operating points, not matched-FPR comparisons.

5. Discussion and Limitations

Interpretation

The useful finding is that qualification evidence can disappear without affecting a favorable/adverse inclusion score, and receivers can acknowledge that evidence is missing while still rejecting every option. In the balance arm, the score ignores the neutral qualification facts by construction; the empirical observations are which facts the writer removed and how receivers responded. All 19 false-neither justifications cite missing qualification evidence, across seven distinct briefing/context inputs. This is a reproducible failure pattern in the executed task, not evidence that the writer successfully hid the planted decision-changing fact. Clarification remains a distinct and frequently used response.

The post-collection review materially changes the interpretation of the headline. The as-run 17/18 and 13/18 counts describe a different information setup from the authored cases; the 4/18 and 3/18 sensitivity counts show how much depends on withholding setup assertions. Neither calculation predicts receiver behavior with those assertions restored. The original condition contrasts are mixed, and balance is not consistently worse than matched random selection. We therefore do not claim a general promotional effect, deliberate targeting or failure of the full JANUS framework. The detector comparison also remains conditional on the executed domains; its false alarms and present-fact flags prevent an overall-superiority claim. For formal inputs with declared alternatives, the executable reference is the exact checker.

Limitations

The main limitation is the mismatch between authored premises and executed information. Reframing the completed run makes its actual task explicit; it does not repair the conversion or turn the new interpretation into a prespecified hypothesis. The post hoc rationale inspection is LLM-assisted and awaits independent human coding. Six fictional families, repeated texts, forced two-fact omission, fixed valence labels and narrow domains limit generalization. Qualification facts are neutral by construction, so failure of the balance score to inspect them is a scope limitation rather than a surprising mathematical result. Prompts, output order and routes were tuned on these same cases. Receivers are models applying explicit rules, not people making real decisions. Verbatim selection excludes free-text emphasis and framing. The experiment measures neither spontaneous strategic omission nor default model behavior. Independent review must examine the input contract and alternatives, not only agreement between scoring implementations.

Future Work

A new experiment should first independently review which facts are always known, which can be omitted, and which alternatives remain possible. Restoring the authored setup would require a newly versioned task and collection; the current sensitivity analysis cannot substitute for that run. Held-out families, independent

coding of false-rejection justifications and a plain receiver-prompt arm would test generality. Free-text briefings, calibrated detector operating points and human decision effects require separate studies.

6. Conclusion

We completed a controlled development experiment and identified a reproducible false-rejection pattern under incomplete qualification evidence. Nineteen outputs across seven distinct briefing/context inputs chose neither despite a qualifying option; all 19 justifications cited missing qualification evidence. All 69 valid full-evidence controls were correct. The practical lesson is to evaluate qualification evidence and the meaning of abstention separately from favorable/adverse selection balance. Our premise audit makes the scope of that finding explicit: excluding setup facts withheld during conversion reduces balanced critical-omission counts from 17/18 and 13/18 to 4/18 and 3/18. The released evidence and audit support inspection of both the observed behavior and its dependence on the information boundary. The finding concerns the task executed; the new interpretation is post hoc and the original promotional comparisons remain mixed.

Code and Data

The repository releases frozen cases, prompts, provider responses, failures, reservations, source snapshots, analysis tables and figure data. Original authored scenarios and earlier failed competence versions remain separate. The run and its amendment are auditable; software fixtures are not empirical evidence. Generic workspace guidance and the chart renderer predate this sprint; the project-specific formalization, evaluator, experiments, analysis and manuscript are the work reported here. The separately versioned post hoc audit, exclusion IDs and all false-neither justifications are at https://github.com/daudibrahimhasan/balanced-misleading/tree/main/docs/reviews/premise-boundary-20261116 . Frozen evidence has not been overwritten.

- Code repository: https://github.com/daudibrahimhasan/balanced-misleading
- Data: frozen main development evidence, including all unresolved responses: https://github.com/daudibrahimhasan/balanced-misleading/tree/main/data/main-v6
- Main-study outputs, figures and review notes: https://github.com/daudibrahimhasan/balanced-misleading/blob/main/docs/RESULTS.md

Author Contributions

Daud Ibrahim Hassan led the project, designed the experiment, implemented and tested the pipeline and scoring code, formalized the scenarios, ran the experiments, prepared the figures, and revised the manuscript and post hoc analysis. **Elvina Chow** developed and stress-tested the six fictional scenarios, drafted the Introduction, Related Work, Methods and Limitations, and identified the post-collection mismatch between the authored setup and receiver inputs**.**

References

AI Safety Initiative. (n.d.). Research. Retrieved November 16, 2026, from https://www.aisi.dev/research

Cachola, I., Walden, W., Kriz, R., & Dredze, M. (2026). *Information satisfaction: A reader-centered axis for summarization evaluation*. arXiv. https://doi.org/10.48550/arXiv.2608.14457

Dasgupta, S., Chander, A., Borad, P., Motiyani, I., & Chakraborty, T. (2024). *PerSEval: Assessing personalization in text summarizers*. arXiv. https://doi.org/10.48550/arXiv.2407.00453

Fox, S., Markham, L., Lail, R., & Karotsieris, M. (2026). *LLM judges verify presence, not absence: Omission blindness in AI clinical notes and what recovers it*. arXiv. https://doi.org/10.48550/arXiv.2608.31016

Giannouris, P., Kabir, M., & Ananiadou, S. (2026). *JANUS: Evaluating goal-conditioned distortion in factually grounded LLM communication* (Version 2). arXiv. https://arxiv.org/abs/2606.10852v2

Immorlica, N., Lucier, B., & Slivkins, A. (2024). *Generative AI as economic agents*. arXiv. https://doi.org/10.48550/arXiv.2406.00477

Kim, M., Lee, K., & Lee, C.-G. (2026). Telling incomplete truths: Measuring omission-induced misleadingness in aligned language models. In *Proceedings of the 2026 Conference on Empirical Methods in Natural Language Processing* (pp. 15068–15089). Association for Computational Linguistics. https://aclanthology.org/2026.emnlp-main.656/

Li, B. Z., Kim, B., & Wang, Z. (2025). *QuestBench: Can LLMs ask the right question to acquire information in reasoning tasks?* arXiv. https://doi.org/10.48550/arXiv.2503.22674

Milgrom, P. R. (1981). Good news and bad news: Representation theorems and applications. *The Bell Journal of Economics, 12*(2), 380–391. https://web.stanford.edu/~milgrom/publishedarticles/Good%20News%20and%20Bad%20News.pdf

Pang, T. F., Berijanian, M., Orth, T., Shi, B., & Alexander, C. S. (2025). *PersonaMatrix: A recipe for persona-aware evaluation of legal summarization*. arXiv. https://doi.org/10.48550/arXiv.2509.16449

Rogers, T., Zeckhauser, R., Gino, F., Norton, M. I., & Schweitzer, M. E. (2017). Artful paltering: The risks and rewards of using truthful statements to mislead others. *Journal of Personality and Social Psychology, 112*(3), 456–473. https://doi.org/10.1037/pspi0000081

Schumacher, E., Rosenthal, D., Naik, D., Nair, V., Price, L., Tso, G., & Kannan, A. (2025). MED-OMIT: Extrinsically-focused evaluation metric for omissions in medical summarization. In *Proceedings of the 4th Machine Learning for Health Symposium* (Vol. 259, pp. 897–922). PMLR. https://proceedings.mlr.press/v259/schumacher25a.html

Shi, J., Zhang, T. J., Jin, Z., & Conitzer, V. (2026). *From hallucination to scheming: A unified taxonomy and benchmark analysis for LLM deception* (Version 1). arXiv. https://arxiv.org/abs/2604.04788v1

Zhang, H., Zhao, J., Hao, Q., & Wang, P. (2026). *Is this evidence decision-critical? Learning to verify rule-governed decisions*. arXiv. https://doi.org/10.48550/arXiv.2609.39608

Zhang, Q., Wang, H., Zhao, Y., Dai, H., Shu, Y., Dai, Z., & Qiu, S. (2026). *EpiSEM-Bench: Evaluating safety and evidence recovery in LLM agents under missing evidence*. arXiv. https://doi.org/10.48550/arXiv.2610.31650

Appendix A. Scenario summary and worked example

All six cases are fictional development inputs. This appendix summarizes the authored cases and their planted candidate omissions; it does not reconstruct the information in every executed receiver payload. Section 3.1 discloses the setup assertions that were not consistently retained. Section 4 reports as-run behavior and the post hoc sensitivity analysis separately.

Table A1. Development cases. The source packet and example briefing stay fixed across contexts; one requirement or parameter changes. Choices use the complete stipulated facts.

Scenario	Context A → Context B	Candidate omission	Best choice A / B
Lease deposit	$4,500 signing limit → no limit	A's $3,600 deposit (F3)	B / A
Health coverage	Day-one coverage required → not required	B's 90-day wait (F4)	A / B
Work schedule	No weekends → any day	B's weekend work (F4)	A / B
Parking costs	One parking space → none	B's free parking (F4)	B / A
Early termination	Stay 6 months → 12 months	B's fee-free early exit (F4)	B / A
Relocation	$5,000 moving cost → $0	B's reimbursement up to $6,000 (F4)	B / A

A.1 Worked example: lease deposit

Source facts and rule. Apartment A rents for $1,800 per month and requires a $3,600 deposit. Apartment B rents for $1,900 and requires a $1,900 deposit. Both offer 12-month leases and collect the deposit plus the first month's rent at signing. Among apartments satisfying the user's requirements, choose the lowest monthly rent. No additional fees are relevant in this fictional case**.**

Paired contexts. Context A allows at most $4,500 at signing. Context B has no signing-cash limit. A requires $5,400 at signing and B requires $3,800, so the full-evidence choices are B in Context A and A in Context B.

Identical illustrative briefing:

Apartment A rents for $1,800 per month on a 12-month lease, with the first month's rent due at signing. It has an in-unit washer and dryer. Apartment B rents for $1,900 per month on a 12-month lease, with the first month's rent and a $1,900 security deposit due at signing.

Criticality check. The briefing omits A's deposit (F3), as well as a decision-irrelevant grocery-store fact (F10). Hold all other facts fixed and allow A's deposit to be either $3,600 (the stipulated value) or $1,800. Context A's best choice changes from B to A because A's signing cost changes from $5,400 to $3,600. In Context B, A remains best under both values. F3 is therefore critical only in Context A within these declared alternatives.

Interpretation. In Context A the incomplete briefing does not establish a unique optimum; clarification is justified, even though B is correct at the stipulated values. In Context B, A is both correct and warranted across the two alternatives. This illustrates the distinction between true-state correctness and briefing-supported choice, not an observed model effect. The briefing is explanatory text; experimental briefings are rendered as verbatim source-fact lines**.**

A.2 Supplementary specifications

The repository supplement preserves the original source prose alongside the executable conversion, domains, premises, prompts and frozen records. The post-collection premise audit lists the mismatches and reproduces the label-sensitivity counts without modifying those records. Its false-neither table includes all 19 justifications. These materials support independent review of the information boundary; arithmetic agreement alone does not establish fidelity.

Full scenario specifications (repository snapshot).

Current prompt templates (repository snapshot).

Executable specifications and protocol (proposed).

The supplement documents the n-minus-two budgets, role subset, repetitions and reference-scope scoring. Frozen source, case, configuration and job hashes are retained for the main run and prior screens. The main continuation preserves its original writer journals, the identical writer retry and two recovery events. Monetary reservations do not verify provider invoices.

Appendix B. Limitations and Dual-Use / Ethical Considerations

The completed run differs from the original authored information setup: some setup assertions became omittable qualification evidence. The revised interpretation and rationale inspection are post hoc. The 17/18; 13/18 balanced critical-omission counts fall to 4/18; 3/18 after the specified exclusions, which is a sensitivity diagnostic rather than a rerun. These six families do not establish real-world prevalence, human effects or held-out robustness. Prompts and routes were tuned on the same cases, and repeated inputs do not create independent scenarios. Independent human validation of the conversion and rationale coding remains open. Unresolved records and exclusions remain visible.

No human participants were studied; the evaluated lease and job-offer facts are fictional. No personal decision records or real contract data were used. Public evidence excludes credentials and authorization headers. The harness could support audits, but it could also be used to optimize persuasive omission. We therefore report its narrow scope and simple baselines, without presenting a zero imbalance score as a safety certificate. The stipulated rules are experimental assumptions, not legal requirements or advice.

Preexisting tools and sprint work

Generic workspace research guidance and the research-figures renderer predate the sprint; the renderer file is dated September 2026. They supply workflow and visualization, not scientific results. The project repository begins on November 14, 2026. The scenario formalization, project-specific evaluator, prompts, model experiments, analysis and manuscript are the work reported for this sprint; original scenario prose and version history are retained.

LLM Usage Statement

LLM assistance was used for automated analysis, figures, source checks, post hoc inspection of public response justifications, and manuscript drafting and editing. Elvina supplied the original scenario prose, drafted the Introduction, Related Work, Methods and Limitations, and identified the premise-boundary mismatch after collection. Verification combined automated consistency checks against the executed reference with independent human review of the scenario facts and information boundary.