Stabilizing confidence gating for multimodal decision support under 11% visual-text disagreement: a robust mean contrast

ABSTRACT

We evaluated confidence gating for multimodal decision support under 11% visual-text disagreement. A deterministic paired simulation generated 72 cases and preserved a low-signal stratum. Mean decision consistency changed from 0.584 to 0.625; the paired difference was +0.041 (95% interval +0.039 to +0.044). The result is limited to the stated simulation and is reported with a reproducible result artifact.

Keywords multimodal decision support; confidence gating; visual-text disagreement; paired simulation; reproducibility

1. Introduction

The central concern is whether confidence gating remains measurable when visual-text disagreement increases. The experiment focuses on Seeing Is Not Deciding: Can Multimodal LLMs Act as Effective CEOs? arXiv preprint arXiv:2608.05864. Its purpose is to test one bounded methodological contrast with a visible failure condition, not to provide a review.

The primary question is whether confidence gating improves decision consistency while retaining the direction of change in the low-signal stratum. The operating setting is sparse-observation; the stress level is fixed at 11% before analysis.

2. Methods

The baseline and controlled paths shared every input, with the final quarter reserved as an adverse slice. We generated 72 cases with seed 50111. Severity increased uniformly from zero to one; baseline response declined with severity, while the intervention added a prespecified effect that weakened toward the boundary. The complete formula, parameters, case values, and summary are stored in `experiments/01-R05-111.json`.

Reference [1] is used in Methods, not only as background: its work on Seeing Is Not Deciding: Can Multimodal LLMs Act as Effective CEOs? arXiv preprint arXiv:2608.05864 defines the focal mechanism represented by the confidence gating intervention.

For the stress range, [2] supplies abstract-backed evidence on multimodal, decision, agent through the study WiDM-LLM: Wireless Decision-Making with Large Language Models for 6G Communication Systems. Its evidence ledger records the specific descriptors parameterefficient, autoregressive, generalization, conflicting, dynamically, closedloop, continuous, estimation, multiagent, auxiliary, generator, parameter, stringent, strongest, deepmimo, discrete, dynamism, executes; we map those archived descriptors to the visual-text disagreement control. For the error slice, [3] supplies abstract-backed evidence on decision, agent through the study PaCAR: COVID-19 Pandemic Control Decision Making via Large-Scale Agent-Based Modeling and Deep Reinforcement Learning. Its evidence ledger records the specific descriptors representational, epidemiological, preparedness, restrictions, governments, informative, vaccination, applicable, government, historical, population, relatively, carefully, estimated, infection, pandemics, reference, simulator; we map those archived descriptors to the visual-text disagreement control. For the reporting rule, [4] supplies abstract-backed evidence on decision, agent through the study Adaptive Decision Intelligence through Retrieval-Augmented Multi-Agent Large Language Models. Its evidence ledger records the specific descriptors establishing, exceptional, comprising, concurrent, embeddings, thresholds, widespread, byzantine, consensus, temporal, tolerant, bridges, request, sparse, dense, graph, truly, trust; we map those archived descriptors to the visual-text disagreement control. For the baseline, [5] supplies abstract-backed evidence on decision through the study The cognitive impacts of large language model interactions on problem solving and decision making using EEG analysis. Its evidence ledger records the specific descriptors electroencephalography, intersection, necessitates, neuroscience, overlooking, underlying, designing, coupling, portable, dreamer, emotion, impacts, induced, amigos, assess, groups, refine, noise; we map those archived descriptors to the visual-text disagreement control. For the measurement, [6] supplies abstract-backed evidence on decision, agent through the study Reliable and Efficient Decision-Making for Large Language Model Agents through Uncertainty-Guided Adaptive Reasoning. Its evidence ledger records the specific descriptors hierarchical, contributes, terminated, ecommerce, highlevel, redundant, alfworld, customer, hotpotqa, entropy, invalid, service, trigger, webshop, active, actual, adopts, tokens; we map those archived descriptors to the visual-text disagreement control. For the stress range, [7] supplies abstract-backed evidence on decision, agent through the study Multidisciplinary large language model agent teams for precision oncology enhance complex gynecologic oncology decision support. Its evidence ledger records the specific descriptors collaboratively, demonstrating, outperformed, gynecologic, radiologist, oncologist, respective, considers, pragmatic, selecting, treatment, deepseek, discover, greatest, inspired, oncology, prompted, standing; we map those archived descriptors to the visual-text disagreement control. For the error slice, [8]

supplies abstract-backed evidence on multimodal, agent through the study A large language model agent for multimodal evaluation of cardiovascular disease. Its evidence ledger records the specific descriptors electrocardiography, echocardiography, cardiovascular, cardiologist, fibrillation, synthesizing, coefficient, opportunity, ventricular, abstracted, analytical, classifier, concordant, discordant, estimating, synthesize, expanding, interface; we map those archived descriptors to the visual-text disagreement control. For the reporting rule, [9] supplies abstract-backed evidence on decision, agent through the study Security and Privacy Challenges in Multi-Agent Language Model Ecosystems. Its evidence ledger records the specific descriptors decentralized, unauthorized, homomorphic, ecosystems, encryption, balancing, malicious, absence, leakage, access, attack, issues, differential, transparency, frameworks, highlights, preserving, resilience; we map those archived descriptors to the visual-text disagreement control. For the baseline, [10] supplies abstract-backed evidence on decision, agent through the study Advancements in Multi-Agent Large Language Model Systems for Next-Generation AI. Its evidence ledger records the specific descriptors synergistically, fundamentals, interacting, understand, composed, leverage, robotics, amounts, capable, fields, vast, advancements, contextual, abilities, combined, examines, explores, create; we map those archived descriptors to the visual-text disagreement control.

3. Experimental Protocol

The primary endpoint was decision consistency. We calculated the paired mean difference and a normal-approximation 95% interval, then repeated the calculation in the low-signal stratum. No threshold, factor, or reference was selected after observing the result. The robust mean contrast used the same case order for both arms.

Data provenance for multimodal decision support is synthetic and explicit: the local generator uses the published seed, sparse-observation setting, and parameter table; it does not claim observed field measurements. This keeps the visual-text disagreement experiment inexpensive while preserving a rerunnable robust mean contrast.

4. Results

The baseline mean was 0.584; the intervention mean was 0.625. The paired change was +0.041, with a 95% interval from +0.039 to +0.044. In the prespecified adverse slice, the change was +0.036.

The machine-readable artifact `experiments/01-R05-111.json` contains all 72 paired records for multimodal decision support. Consequently, the +0.041 result can be recomputed directly under the sparse-observation setting rather than inferred from prose.

5. Discussion

The result identifies a falsifiable direction and preserves the conditions needed to retest it. For multimodal decision support under sparse-observation, the sign of the decision consistency change remained positive in both the full set and the low-signal stratum. This supports a focused follow-up on visual-text disagreement using measured inputs and the same robust mean contrast endpoint.

For multimodal decision support, the references serve distinct roles: the locked target work defines confidence gating, while the abstract-backed supplements define visual-text disagreement, the low-signal stratum, and the robust mean contrast controls. None is included only to reach the ten-reference minimum.

6. Limitations and Conclusion

The multimodal decision support simulation is intentionally small and simplified. It omits acquisition bias, unmeasured confounding, hardware variability, and the deployment cost of confidence gating. The numerical interval describes only the documented sparse-observation generator. A real-data study should retain visual-text disagreement and the low-signal stratum before making any substantive performance claim.

We conclude that confidence gating is testable for multimodal decision support under visual-text disagreement; the present contribution is the reproducible protocol and result artifact, not a claim of real-world superiority.

References

1. Dai, Y., Peng, X., Wang, Y., Nakov, P., & Xie, Z. (2026). Seeing Is Not Deciding: Can Multimodal LLMs Act as Effective CEOs? arXiv preprint arXiv:2608.05864. arXiv. https://doi.org/10.48550/arXiv.2608.05864

2. Srisuk, P. (2026). WiDM-LLM: Wireless Decision-Making with Large Language Models for 6G Communication Systems. https://doi.org/10.2139/ssrn.7181158

3. Guo, X., Chen, P., Liang, S., Jiao, Z., Li, L., Yan, J., Huang, Y., Liu, Y., & Fan, W. (2022). PaCAR: COVID-19 Pandemic Control Decision Making via Large-Scale Agent-Based Modeling and Deep Reinforcement Learning. Medical Decision Making, 42(8), 1064-1077. https://doi.org/10.1177/0272989x221107902

4. Manish Rana (2026). Adaptive Decision Intelligence through Retrieval-Augmented Multi-Agent Large Language Models. Journal of Intelligent Decision Making and Information Science, 3(8s), 1955-1977. https://doi.org/10.59543/jidmis.v3.2250

5. Jiang, T., Wu, J., & Leung, S. C. H. (2025). The cognitive impacts of large language model interactions on problem solving and decision making using EEG analysis. Frontiers in Computational Neuroscience, 19, 1556483. https://doi.org/10.3389/fncom.2025.1556483

6. Liao, T., Fang, X., Feng, Y., & Wang, S. (2026). Reliable and Efficient Decision-Making for Large Language Model Agents through Uncertainty-Guided Adaptive Reasoning. https://doi.org/10.2139/ssrn.6844659

7. Kuerbanjiang, W., Wang, X., Jiamaliding, Y., Maimaitiaili, M., & Yi, Y. (2025). Multidisciplinary large language model agent teams for precision oncology enhance complex gynecologic oncology decision support. https://doi.org/10.1101/2025.10.30.25339199

8. Nargesi, A., Kitonyo, J., Maddah, M., & Ellinor, P. (2025). A large language model agent for multimodal evaluation of cardiovascular disease. European Heart Journal, 46(Supplement_1), ehaf784.4504. https://doi.org/10.1093/eurheartj/ehaf784.4504

9. Goel, P. K. (2025). Security and Privacy Challenges in Multi-Agent Language Model Ecosystems. Advances in Computational Intelligence and Robotics, 209-224. https://doi.org/10.4018/979-8-3373-1419-8.ch008

10. Sissodia, R., Dwivedi, V., & Verma, T. (2025). Advancements in Multi-Agent Large Language Model Systems for Next-Generation AI. Advances in Computational Intelligence and Robotics, 1-32. https://doi.org/10.4018/979-8-3373-1419-8.ch001