UNMASK: Behavioral Middleware Framework for Concealed Intent Evaluation

[1]Tyree M. Sloan

[1]Norfolk State University, Norfolk, VA

Research Advisor 7 Mentor: Dr. Soamar Homsi
Air Force Research Laboratory / Information and Spectrum Warfare Directorate (AFRL/RF)
Information Dominance Division / Resilient Information Dominance (AFRL/RFGA)
ZEIOS Vanguards: Zero-Trust Engineering for Intelligent OSINT Systems

Large language models are increasingly deployed with access to protected corpora and sensitive resources, making accurate understanding of user intent essential. Most safeguards evaluate prompts independently, which is effective for explicit violations but less so when intent develops gradually across turns. No single prompt may appear decisive while the conversation as a whole exhibits a coherent behavioral trajectory [5]–[8]. Research on human deception detection reaches a parallel conclusion: intent is revealed through weak indicators accumulated over time rather than isolated statements [1], [2]. Work on trajectory-level runtime monitoring further shows that conversational behavior can be monitored across time under black-box conditions, supporting the feasibility of an intermediate analysis layer [4]. UNMASK addresses this gap as behavioral middleware positioned between a conversation source and a downstream evaluator, defining a layered progression from behavioral observation through primitive measurement, behavioral features, accumulated evidence, and Intent Dimensions to an Intent Profile, the framework's canonical output boundary.

This work investigates whether accumulated, explainable behavioral evidence provides more useful signals for concealed-intent evaluation than isolated prompt analysis alone. The framework does not observe hidden motivation directly; it measures observable behavior and converts those observations into structured evidence, separating behavioral measurement from final intent determination. The project proposes a five-category Primary Intent Taxonomy and twenty-three Intent Dimensions spanning information-access posture, surface-context framing, and concealment characteristics; these remain conceptual and are not yet validated.

Two behavioral features are implemented and tested on a limited validation set. Behavioral Density, a composite feature, combines Topic Persistence, embedding-based alignment of user turns with topic anchors, with a deterministic lexical fallback, and Specificity Escalation, the change in operational specificity between early and late turns, under configurable weights [3]. Context Camouflage evaluates whether legitimacy-signaling context may function as cover for another objective, without treating academic, professional, or creative framing as inherently deceptive [6]–[8]. A structured benchmark, schema-normalizing loader, and batch execution workflow support reproducible experimentation. In preliminary weight-sensitivity experiments on a matched validation pair, a restricted-objective scenario and a benign look-alike, increasing the Specificity Escalation weight from 0.4 to 0.8 increased Behavioral Density separation from 0.086 to 0.190. Topic Persistence was comparable across both scenarios, indicating that persistence occurs in benign and restricted conversations alike and that trajectory may be more discriminative than topical coherence. These observations derive from a limited validation set and are not generalizable. Future work includes expanded benchmark coverage, empirical

validation of the proposed Intent Dimensions and their mapping to features, ablation and red-team evaluation, and assessment of downstream contribution.

Keywords: behavioral middleware; multi-turn conversation analysis; concealed intent; behavioral evidence; explainability

References

[1] P. A. Granhag and L. A. Strömwall, Eds., The Detection of Deception in Forensic Contexts. Cambridge, U.K.: Cambridge University Press, 2004.

[2] T. R. Levine, Duped: Truth-Default Theory and the Social Science of Lying and Deception. Tuscaloosa, AL, USA: University of Alabama Press, 2020.

[3] C. Wang, Computational Structural Behavior. Singapore: Springer Nature Singapore, 2026.

[4] R. Almarwani and M. Almarwani, "Trajectory-level runtime integrity monitoring for deployed conversational LLM services," IEEE Access, vol. 14, pp. 92923–92941, 2026, doi: 10.1109/ACCESS.2026.3704306.

[5] F. Wu, W. Wang, Y. Qu, and S. Yu, "A low-cost black-box jailbreak based on custom mapping dictionary with multi-round induction," in Proc. 2024 IEEE 23rd Int. Conf. Trust, Security and Privacy in Computing and Communications (TrustCom), Sanya, China, 2024, pp. 888–895, doi: 10.1109/TrustCom63139.2024.00131.

[6] G. Zhou et al., "CCJA: Context-coherent jailbreak attack for aligned large language models," Machine Learning, vol. 115, no. 6, 2026, doi: 10.1007/s10994-026-07069-z.

[7] S. Jiang, X. Chen, and R. Tang, "Deceiving LLM through compositional instruction with hidden attacks," ACM Transactions on Autonomous and Adaptive Systems, 2025, doi: 10.1145/3717067.

[8] N. Das, E. Raff, A. Chadha, and M. Gaur, "Human-readable adversarial prompts: An investigation into LLM vulnerabilities using situational context," ACM Transactions on Intelligent Systems and Technology, 2026, doi: 10.1145/3815174.