Methodology Appendix: High-Persistence Lexical Signature Detection

Purpose This appendix describes the methodology used to identify high-persistence lexical signatures in retrieval-augmented LLM workflows under reset and isolation conditions. The goal is to measure unintended influence without disclosing exploit procedures, payloads, or bypass techniques.

Experimental Context The evaluation platform executes controlled model interactions under two conditions: clean context and stained context. Experiments used thread isolation, context flushing, temporal cooldowns, and reset procedures. No adversarial prompting or policy bypass was used.

Signature Extraction Method Signatures were extracted using sliding n-gram analysis over model output text. N-grams were normalized and treated strictly as statistical fingerprints. No semantic interpretation was applied.

Scoring and Selection Criteria Each signature was evaluated by stained frequency, clean frequency, persistence across resets, and spread across lineages. A score of 10 indicates repeated stained presence, zero clean presence, reset survival, and single-lineage containment.

Interpretation Constraints Signatures are not payloads or instructions. No claims of intent, memory, or causality are made. The analysis is observational only.

Safety and Disclosure Controls Only short lexical fragments are retained. No prompts or retrieval content are published. Artifacts are frozen with cryptographic hashes.

Limitations Lexical signatures capture surface-level residue only. Results are configuration-specific and not generalized.

Summary This methodology enables reproducible detection of persistent unintended influence using non-exploitative, audit-grade techniques.