EORMEN

Trace-Bound Observability of Opaque Large Language Model Behaviour

A Nine-Study Application of the Eormen Foundational Mathematical Framework

Leroy A. Palmer

palmer@eormen.com eormen.com

June 2026

Abstract

Large language models can produce outputs whose visible form is fluent, confident and operationally useful while the relation between the output and the evidence, instructions, context, retrieval material, action objective or runtime condition remains difficult to audit. This paper reports a controlled nine-study application of the Eormen foundational mathematical framework to that problem. The object of investigation is not the quality of a particular language model. The object is whether an already completed language-model interaction can be admitted as a frozen behavioural trace and converted into bounded, replayable observations without changing the trace, modifying the model or claiming access to hidden model state.

The empirical trace material was generated locally with `google/gemma-4-E2B-it`, used only as an opaque trace-producing process. The conservative final-material evidence set comprises 3,360 sealed traces, 15,040 observer classification or comparison rows and 3,320 audit rows across nine opacity areas. Across the controlled conditions tested, the framework classified visible structures in evidence coupling, source dominance, response-formation state, confidence-support alignment, context-anchor persistence, external-content coupling, action-objective coupling, model-condition sensitivity and visible degeneration. It also preserved not-observable, not-comparable and discrepancy outcomes where the admitted trace did not justify a stronger claim. The result is bounded: the framework made multiple forms of opaque LLM behaviour more observable, traceable and scientifically classifiable from frozen behavioural traces. It does not establish hidden reasoning recovery, model benchmarking, safety certification, production reliability or a general solution to all LLM opacity.

Disclosure boundary. This paper reports empirical application, evidence architecture and observed results of the Eormen framework. It does not disclose protected theorem mechanics, implementation internals or protected mathematical apparatus. All claims below are bounded to admitted behavioural traces and recorded verification artefacts.

1 Introduction

Opacity in large language model behaviour is often encountered at the point where the model has already answered. The observer can read the final output, but the structural relationship between that output and the conditions that produced it may remain unclear. The output may be fluent and confident, but weakly connected to supplied evidence. It may appear to follow a prompt, but the visible controlling source may be ambiguous. It may use retrieved material without clearly preserving source separation. It may select an action while giving only a superficial justification. It may change when runtime settings change, or it may degrade into repetition, oscillation or broken response contracts.

This paper addresses that practical setting. It does not attempt to reconstruct hidden model reasoning. It does not inspect weights, activations, logits, attention patterns or hidden state. It asks a narrower and more auditable question: can a frozen behavioural trace be analysed in a disciplined way so that visible opacity structures become classifiable, reproducible and bounded by the evidence actually admitted into the trace?

The Eormen foundational mathematical framework is the method under investigation. Gemma is not the subject of benchmarking, safety certification or product evaluation. Gemma is the local open-source process used to generate opaque behavioural traces under controlled conditions. The studies therefore test the framework's observability capability over admitted traces, not the quality of the trace-producing model.

Nine controlled studies form the evidence base:

1. Evidence Detachment;
2. Instruction Dominance;
3. Response Formation;
4. Confidence Without Support;
5. Context Drift;
6. Retrieval Contamination;
7. Tool Or Action Opacity;
8. Model Drift;
9. Collapse Or Degeneration.

Each study was conducted as a bounded trace-observability experiment. Synthetic closed-world corpora were used to control the material available in each trace. Fixture evidence was used before material generation where required. Material traces were sealed, observer results were stored separately, replay checks were performed from sealed artefacts and discrepancy analyses were retained rather than hidden. The accepted study papers are supporting research records; this manuscript is the synthesis of their empirical findings and evidence boundaries.

2 Background and Related Work

The modern LLM setting inherits several forms of opacity. Transformer-based models made large-scale sequence modelling highly effective [1], and subsequent large language models demonstrated strong task-general behaviour under prompting [2]. Those capabilities make the systems useful, but they do not make an individual output self-explaining. A completed answer may contain true, false, supported, unsupported, instruction-driven and context-drifted material in the same surface form.

Existing work addresses adjacent parts of this problem. Local explanation methods such as LIME [3] aim to explain predictions through interpretable surrogate behaviour. Retrieval-augmented generation [4] attempts to improve knowledge access by adding external material. Hallucination and truthfulness work has shown that fluent language-model outputs can remain unsupported or false [5, 6]. Prompt-injection research has shown that external content and instruction/data ambiguity can materially affect LLM-integrated applications [7].

This paper addresses a narrower problem than model explanation, retrieval quality, truthfulness benchmarking or security defence. It asks what can be observed from an already completed behavioural trace when the observer is not allowed to inspect hidden model state, change the trace-producing process or replace missing evidence with narrative explanation. The method is therefore not a substitute for mechanistic interpretability, evaluation benchmarks or security controls. It is an evidence-discipline layer for frozen traces.

3 Claim Structure and Research Question

3.1 Thesis

Opaque LLM behaviour can be studied scientifically when the completed interaction is admitted as a sealed behavioural trace and analysed by a disciplined observer-side mathematical framework. The Eormen foundational mathematical framework provides a trace-bound method for converting otherwise vague opacity concerns into auditable classifications, while preserving explicit not-observable and not-comparable boundaries when the trace is insufficient.

3.2 Problem Statement

LLM assurance work frequently begins with an already completed interaction. The final answer, prompt material, source material, run metadata and trace spans may be available, but hidden model state is not. Without a trace-bound method, two errors become likely. The first is under-analysis: treating the answer as plain text and failing to classify the structural relation between the answer and the evidence, instructions, context or action objective. The second is over-analysis: turning a plausible post hoc explanation into an unsupported claim about hidden causation.

The problem is therefore to make visible behavioural structure classifiable without crossing the evidence boundary. A useful method must record what the trace supports, what it contradicts, what remains ambiguous and what cannot be responsibly inferred.

3.3 Contribution and Claim Structure

The paper makes three bounded claims.

First, the nine-study evidence set shows that opacity can be decomposed into specific trace-bound observability questions. The study areas are not vague complaints about LLM behaviour. They are operational questions about evidence coupling, visible source dominance, formation state, confidence-support relation, context-anchor behaviour, external-content coupling, action coupling, cross-condition sensitivity and visible degeneration.

Second, the framework generated replayable observer outputs across all nine study areas. Replay does not prove that the model's hidden process was known. It proves that the observer-side classifications can be regenerated from the sealed artefacts without mutating the trace.

Third, the evidence discipline is negative as well as positive. The study set retains not-observable, not-comparable and discrepancy outcomes. These are not failures to be hidden. They are the boundary conditions that prevent the paper from overstating what the trace can support.

3.4 Research Question

Across nine controlled opacity-area studies, can the Eormen foundational mathematical framework classify visible behavioural structures from sealed LLM traces without mutating the trace, modifying the trace-producing process, relying on hidden model state or turning the study into a model-quality evaluation?

4 Hypothesis

If relevant behavioural structure is visible in a sealed trace, then the framework should classify that structure from the admitted trace material alone. If the trace does not admit the required evidence, the framework should return a bounded not-observable, not-comparable or discrepancy classification rather than forcing an unsupported interpretation.

5 Method Boundary

5.1 What Is Admitted

The admitted evidence differs by study, but the common material consists of prompt text, closed-world source material, final model output, declared spans, run manifests, decoding settings, trace hashes and stored observer outputs. The final-material evidence counted in this synthesis uses one accepted material evidence source for each completed study area. Fixture evidence, earlier targeted iterations and discrepancy analyses support the evidence chain but are not included in the headline conservative count.

5.2 What Is Excluded

The studies do not admit hidden reasoning, chain-of-thought, attention values, activations, weights, logits, token probabilities, hidden state, live retrieval rankings or live tool execution traces. They also do not use external truth checking as the basis for the framework's classifications. The analysis is therefore not a hidden-state interpretation study. It is a trace-observability study.

5.3 Trace Sealing and Observer Separation

The studies separate the trace from the observer. The trace is generated and sealed first. Observer classifications are then produced over the sealed material and stored separately. Replay checks test whether those observer rows can be regenerated from the sealed artefacts. This separation is essential: the framework is tested as an observer over admitted material, not as a mutation of the trace or as an intervention inside the model.

5.4 Not-Observable Outcomes

A not-observable outcome is not treated as a weak result. It is an evidential boundary. It states that the trace does not contain enough admitted material to support the stronger claim under the frozen analysis rules. This feature is central to the scientific discipline of the study set. A framework that always forces a positive classification would convert missing evidence into unsupported certainty.

6 Experimental Study Set

The nine studies were designed around common opacity concerns in LLM behaviour. Table 1 summarises the study scope and the trace-bound structures tested.

Table 1: Study scope and trace-bound structures.

Study area	Opacity problem	Trace-bound structures tested
Evidence Detachment	The answer appears confident, but its relationship to supplied evidence is unclear.	Evidence coupling, supported absence, weakening support and bounded detachment.
Instruction Dominance	It is unclear which visible source controlled the output.	Source dominance, lower-priority override, external-source influence, resistance and not-observable source boundaries.
Response Formation	The final answer hides visible commitment, reversal or transition structure.	Stable formation, weak transition, reversal, oscillation and formation-boundary outcomes.
Confidence Without Support	Confidence signals may exceed the support admitted in the trace.	Confidence signal, support strength, contradiction, unsupported high-confidence commitment and visible-format boundary.
Context Drift	Earlier context may decay while nearby material dominates.	Anchor persistence, anchor decay, selected-context position, recency dominance and malformed-boundary admission.
Retrieval Contamination	Retrieved or retrieval-like material may silently distort the answer.	External-content coupling, source selection, injection-following, mixed-source use and retrieved-only evidence-use boundaries.
Tool Or Action Opacity	Action selection may be weakly coupled to the objective or justification.	Selected action, objective coupling, justification coupling, constraint handling, parameter pressure and malformed action boundaries.
Model Drift	Behaviour may change across runtime or sampling conditions.	Pairwise repeat stability, sampling sensitivity, visible structural transition and not-comparable budget boundaries.
Collapse Or Degeneration	Output may repeat, oscillate, lose coherence or break its response contract.	Repetition, oscillation, response-contract collapse or repair, stable-control behaviour and brittle-transition boundaries.

6.1 Evidence Scale

The final-material evidence count is deliberately conservative. It does not include every fixture run, every early material run, every sampling probe or every discrepancy-analysis row. It counts one accepted final material evidence source per completed study, with discrepancy analyses used as interpretive evidence where final closeout depended on them.

Table 2: Conservative final-material evidence used in this synthesis.

Study area	Traces	Observer rows	Audit rows	Rerun status	Replay status
Evidence Detachment	160	320	480	exact	exact
Instruction Dominance	480	2,880	480	exact	exact
Response Formation	120	480	120	exact	exact
Confidence Without Support	480	2,400	480	exact	exact
Context Drift	40	200	40	exact	exact
Retrieval Contamination	560	2,800	560	exact	exact
Tool Or Action Opacity	640	3,200	640	exact	exact
Model Drift	720	1,800	360	sampling-bound	exact
Collapse Or Degeneration	160	960	160	exact	exact
Total	**3,360**	**15,040**	**3,320**		

Model Drift is marked as sampling-bound because sampling-condition rerun mismatch was recorded and retained. The accepted pairwise analysis also recorded zero deterministic-condition rerun mismatches and exact sealed-trace replay. This distinction matters because the study does not flatten sampling sensitivity into a general reproducibility failure.

7 Analysis Plan

The synthesis uses a claim-control rule: every substantive claim must map to recorded evidence. Permitted evidence includes sealed material traces, observer classification rows, comparison rows, audit rows, replay checks, deterministic or sampling rerun records, discrepancy analyses, closeout records and completed study papers. Claims outside that evidence boundary are excluded.

The analysis plan has five components.

1. Preserve the distinction between the trace-producing model and the framework under investigation.
2. Report conservative final-material totals separately from broader evidence-chain artefacts.
3. Treat discrepancy analysis as evidence, not as a defect to hide.
4. Carry residual limitations in the main text.
5. Avoid claims about hidden model state, universal LLM opacity, model quality, production reliability, safety certification or market adoption.

8 Study Results

8.1 Evidence Detachment

Evidence Detachment tested whether the framework could classify the relation between final claims and supplied evidence. The pain point is common: an answer can sound authoritative while its connection to the supplied evidence is weak, ambiguous or absent. The study used closed-world evidence packs and constrained claims so that the final response, claim spans and evidence spans could be admitted into the trace.

The final accepted material run, ED-DEF-09, produced 160 sealed traces, 320 observer rows and 480 claim-audit rows. The deterministic rerun matched exactly. Sealed-trace replay reproduced both observer classifications and claim-audit outputs. The classification distribution carried into the synthesis included stable coupling, supported absence and detachment cases. The final paper for the study records 127 supported-absence cases, 23 stable-coupling cases and 10 detachment cases in the accepted final material evidence.

The important finding is not that the trace-producing model was correct or incorrect. The finding is that the framework could classify whether visible claims remained coupled to the admitted evidence, and could identify when the proper answer was absence rather than an invented claim. Supported absence proved especially important: many apparent evidence-detachment pressures regularised into responses that correctly refused unsupported details under the study rules.

The principal limitations are scope limitations. The study used synthetic evidence packs and final-output-plus-evidence traces. It did not test citation support, response timing, hidden reasoning, real-world truthfulness or population frequency of detachment. The supported conclusion is therefore bounded: evidence-coupling and detachment structures were made observable in the admitted closed-world trace setting.

8.2 Instruction Dominance

Instruction Dominance tested whether the framework could classify which visible instruction source controlled an output when multiple sources were present. The opacity problem is not merely whether an answer followed an instruction, but which admitted source visibly dominated the final response: system-like material, developer-like material, user material, tool text, retrieved text or injected text.

The definitive Instruction Dominance sequence was broader than the conservative final-material row. Across ID-DEF-01, ID-DEF-02 and ID-DEF-03, the study created 1,120 deterministic material traces, 6,720 observer rows and 1,120 dominance-audit rows. All three deterministic material runs recorded exact same-run rerun matches, and sealed-trace replay reproduced stored observer outputs. The conservative synthesis count uses the final targeted run: ID-DEF-03 with 480 traces, 2,880 observer rows and 480 audit rows.

The study classified visible source dominance, priority conflicts, lower-priority override, external-source influence, source resistance and not-observable boundaries. The final targeted run included explicit boundary evidence, including source cases where the trace did not contain decisive control material. Those not-observable cases were not treated as missing results; they were recorded as proper boundaries in the instruction stack.

The result supports a bounded affirmative claim: visible instruction-source dominance can be classified from sealed synthetic instruction stacks when the source labels and directive material are admitted into the trace. The study does not prove hidden platform hierarchy, does not infer invisible instruction channels and does not evaluate model obedience as a general quality measure.

8.3 Response Formation

Response Formation tested whether visible commitment structure could be classified from final-output material. The pain point is that a final answer may hide when it committed to a direction, whether it reversed course, whether it oscillated or whether it remained uncertain. This study did not use token streams or timing data. It used visible response roles, final text and declared spans.

The accepted final material run, RF-DEF-07, produced 120 sealed traces, 480 observer rows and 120 formation-audit rows. It recorded exact deterministic rerun match and exact sealed-trace replay. The final targeted run exercised stable formation, weak transition, reversal, oscillation and formation-boundary profiles. Expected-profile alignment was 106 matched cases and 14 mismatched cases. Five of six families matched fully. All fourteen mismatches were localised to the quoted-transition family.

The importance of this result is that the framework did not claim hidden commitment timing. It classified visible formation structure. When quoted transition material regularised into stable visible formation, that discrepancy was retained as a limitation rather than reinterpreted as success. This is a useful scientific boundary: if the final-output trace does not visibly preserve the intended transition, the framework cannot responsibly claim that the transition occurred.

The limitation is therefore explicit. Response Formation supports classification of visible final-output formation regimes, not token-level formation timing, hidden internal commitment or unobserved deliberative process.

8.4 Confidence Without Support

Confidence Without Support tested whether visible confidence signals could be separated from evidential support. This is one of the central LLM opacity concerns: a response may sound certain while its evidence base is weak, contradicted or absent. The study admitted final outputs, visible confidence signals, claim material and evidence-use fields.

The final material run, CWS-DEF-03, produced 480 sealed traces, 2,400 observer rows and 480 confidence-audit rows. Deterministic rerun matched exactly and sealed-trace replay matched exactly. CWS-DEF-04 analysed 61 expected-profile mismatches and localised all of them to visible response-format or confidence-signal shifts. The material result included 419 matched cases and 61 mismatched cases, with nine fully matched families and three families with mismatches.

The study showed that the framework could classify unsupported high-confidence structures, weakly supported confidence structures, contradicted support structures and properly bounded uncertainty. This is not a calibration claim. The framework did not infer true probability, hidden confidence or internal model belief. It classified visible confidence and support relations admitted into the trace.

The residual limitation is important. The 61 expected-profile mismatches remain part of the result. They were not evidence that the study failed to observe anything; they were evidence that visible format and confidence-signal surface choices changed how expected profiles appeared in the final trace. That limitation bounds the conclusion to visible response structure.

8.5 Context Drift

Context Drift tested whether the framework could classify visible persistence, decay and displacement of context anchors. The opacity problem is that long or multi-part contexts can cause earlier material to be under-used while nearby or repeated material dominates the final answer. This study did not inspect attention or hidden context weighting. It classified visible selected-context and claim behaviour.

The full Context Drift sequence included broad material and targeted residual iterations. The conservative final-material evidence uses CD-DEF-08, a narrow residual boundary run focused on malformed admission behaviour. CD-DEF-08 produced 40 sealed traces, 200 observer rows and 40 audit rows. It recorded exact deterministic rerun match, exact sealed-trace replay, 40 matched cases and zero mismatches.

Earlier Context Drift runs were scientifically important because they showed that some intended drift pressures did not produce the expected visible departure from target anchors. Those outcomes led to narrower residual iterations. The final accepted result is therefore not the whole research story; it is the closeout point where a specific unresolved boundary was isolated and resolved under the study rules.

The supported conclusion is that visible context-drift structures can be classified where the trace admits anchor, selected-context and response material. The limitation is equally clear: no hidden attention claim is made, and the final residual run is intentionally narrow.

8.6 Retrieval Contamination

Retrieval Contamination tested whether external or retrieval-like material visibly shaped a response. The opacity problem is that retrieved content may silently dominate, distort or contaminate the final answer. The study used controlled retrieval-like passages rather than a live retriever, so the result is about trace-bound contamination observability, not retrieval-system quality.

The final material run, RC-DEF-05, produced 560 sealed traces, 2,800 observer rows and 560 retrieval-contamination audit rows. It recorded exact deterministic rerun match, exact sealed-trace replay and validation success. The run produced 451 fully matched cases and 109 mismatched cases, with 338 lens-level mismatch rows. RC-DEF-07 localised all 109 mismatched cases into three residual classes.

The three residual classes were not hidden. They were explicit trace-bound limitations: missing-claim repair, wrong-subject retargeting and mixed-source selected traces with retrieved-only evidence use. The value of the study is that these residual behaviours were not collapsed into a vague "retrieval problem". They were made classifiable at the trace level.

The study supports the claim that the framework can classify visible external-content coupling, source selection, injection-following and mixed-source use under controlled conditions. It does not claim live retrieval ranking observability, retriever quality, truthfulness of retrieved material or deployment behaviour.

8.7 Tool Or Action Opacity

Tool Or Action Opacity tested whether selected actions were visibly coupled to objectives, justifications and constraints. The opacity problem is practical: when an LLM proposes or selects an action, it may be unclear whether the action is actually justified by the objective or merely superficially plausible. The study used a synthetic action catalogue and controlled objective material.

The final material run, TAO-DEF-05, produced 640 sealed traces, 3,200 observer rows and 640 action-opacity audit rows. It recorded exact deterministic rerun match, exact sealed-trace replay, validation success, 590 fully matched cases, 50 mismatched cases and 226 lens-level mismatch rows. TAO-DEF-07 localised all 50 mismatched cases into four residual classes.

The four residual classes were malformed selected-action repair, order-pressure resistance, selected-action field-order reversal and one parameter-pressure case that resisted the expected target parameter. The study therefore did not end with a vague mismatch count. It translated the mismatch set into explicit trace-bound limitations.

The supported conclusion is that visible action selection, objective coupling, justification coupling and constraint handling can be classified from sealed synthetic action traces. The limitation is that no live tool execution, tool success, operational environment or external action outcome was tested.

8.8 Model Drift

Model Drift tested whether separately sealed traces could be compared across material conditions. This study differs from the others because the opacity question concerns comparison rather than classification inside a single trace. The framework was used to compare sealed trace members for repeat stability, sampling sensitivity, visible structural transitions and not-comparable boundaries.

The final pairwise material evidence, MD-DEF-11, produced 720 sealed traces, 1,800 comparison rows and 360 audit rows. It analysed 360 pairwise comparison groups. Comparison trace-hash mismatches and audit trace-hash mismatches were zero. Sampling-condition rerun mismatches were 30, while deterministic-condition rerun mismatches were zero in the accepted pairwise analysis. Exact sealed-trace replay was recorded.

The study's iterative path matters. Earlier designs produced large expected-profile mismatch sets or made all comparison groups not-observable because a short-budget member blocked complete comparison. The final pairwise design separated repeat, sampling and budget comparisons into independent lanes. That separation made the result clearer: repeat comparisons could be stable where complete, sampling comparisons could be visibly sensitive, and budget-limited comparisons could remain not-observable without forcing an unsupported comparison.

The supported conclusion is that the framework can perform bounded structural sensitivity analysis over separately sealed traces. It does not prove external model-version drift, hidden model change or universal drift detection.

8.9 Collapse Or Degeneration

Collapse Or Degeneration tested visible repetition, oscillation, response contract collapse, repair, brittle transition and stable-control behaviour. The opacity problem is that degraded model behaviour can appear abruptly and may be difficult to separate from ordinary variation unless the trace contains classification rules and observable structure.

The final material run, COD-DEF-09, produced 160 sealed traces, 960 observer rows and 160 audit rows. It recorded exact deterministic rerun match, exact sealed-trace replay, zero trace-hash mismatches, 90 expected-profile matches and 70 expected-profile mismatches. COD-DEF-10 analysed those 70 mismatches from sealed artefacts only and recorded exact independent sealed-trace replay with zero trace-hash reference mismatches.

The final residual pattern was specific. Oscillation matched in all 40 oscillation cases. Malformed-boundary cases largely regularised into observable response contracts. Stable controls matched in 37 of 40 cases, with three visible local phrase repetition cases. This is precisely the type of result a trace-bound framework should produce: not a global claim about collapse, but a local classification of what was visible and what regularised under the final trace conditions.

The supported conclusion is that visible collapse-or-degeneration structures can be classified and replayed from sealed final-output traces. The limitation is that this does not prove hidden collapse dynamics, production reliability or behaviour under broader deployment settings.

9 Cross-Study Results

The nine studies show a consistent pattern. The framework is strongest when the trace admits the relevant structure: claim and evidence spans for Evidence Detachment, source labels for Instruction Dominance, visible roles for Response Formation, confidence and support fields for Confidence Without Support, ordered anchors for Context Drift, controlled external material for Retrieval Contamination, action catalogues for Tool Or Action Opacity, comparison manifests for Model Drift and response-contract markers for Collapse Or Degeneration.

Table 3: Principal study findings and retained boundaries.

Study area	Principal supported finding	Retained boundary
Evidence Detachment	Evidence coupling, supported absence and bounded detachment were classified from closed-world final-output-plus-evidence traces.	No real-world frequency, citation-support or hidden-reasoning claim.
Instruction Dominance	Visible source dominance, lower-priority override, external influence, resistance and not-observable source boundaries were classified.	Source labels were admitted fields; no hidden platform hierarchy claim.
Response Formation	Visible stable formation, weak transition, reversal and oscillation structures were classified.	Fourteen quoted-transition mismatches remain; no token-stream timing claim.
Confidence Without Support	Visible confidence was separated from support, including unsupported high-confidence structures.	Sixty-one expected-profile mismatches remain localised to visible format or confidence-signal shifts.
Context Drift	Final residual boundary produced 40 matched cases and zero mismatches.	The final accepted run is narrow; no hidden attention claim.

Study area	Principal supported finding	Retained boundary
Retrieval Contamination	External-content coupling, source selection and mixed-source use were classified; residual classes were localised.	Mock retrieval only; 109 mismatches retained in three classes.
Tool Or Action Opacity	Action-objective and action-justification coupling were classified; 50 mismatches were localised.	Synthetic action catalogue only; no live tool execution or tool-success claim.
Model Drift	Pairwise sealed traces supported repeat stability, sampling sensitivity and budget not-comparability boundaries.	Sampling rerun mismatch retained; no external model-version claim.
Collapse Or Degeneration	Oscillation, response-contract repair/collapse, stable-control behaviour and local repetition were classified.	Seventy final mismatches retained; no hidden collapse or production reliability claim.

9.1 Replay and Non-Mutation

Replay evidence is central to the study set. The point of replay is not merely software reproducibility. It verifies the non-mutation boundary: observer classifications are produced from stored traces rather than by rerunning or altering the trace-producing process. The final-material evidence set records exact sealed-trace replay for all nine studies. Model Drift preserves sampling rerun mismatch while still recording exact sealed-trace replay.

9.2 Discrepancy Analysis as Evidence

The research did not proceed by selecting only successful runs. Several study areas required residual iterations because earlier material traces did not exercise the intended boundary strongly enough or produced different visible structures. Evidence Detachment required targeted near-twin and wrong-value work. Response Formation retained quoted-transition mismatches. Context Drift isolated malformed-boundary behaviour. Retrieval Contamination and Tool Or Action Opacity localised residual mismatch classes. Model Drift separated comparison lanes after earlier comparison designs proved too broad. Collapse Or Degeneration retained residual mismatch structure rather than claiming full profile alignment.

This pattern is a strength of the evidence. It shows that the framework and the study process did not convert every expectation into a success label. Where the trace differed from expectation, the difference was recorded, localised and carried into the final limitation set.

9.3 What Was Learned Across the Nine Studies

Three cross-study findings are supported.

First, opacity can be decomposed into trace-bound observability questions. A vague concern such as "the answer is opaque" becomes a specific question: evidence coupling, source dominance, confidence-support relation, selected context, retrieval influence, action-objective coupling, condition sensitivity or degeneration marker.

Second, trace-bound classifications require admitted structure. The framework cannot responsibly classify a structure that is not present in the trace. This is why not-observable and not-comparable outcomes are necessary scientific outputs.

Third, the framework's value is not that it explains the hidden model. Its value is that it disciplines what may be inferred from the visible trace. This is directly relevant to audit, assurance and incident review, where the available evidence is often a record of behaviour rather than an internal view of the system.

10 Use Case

The primary use case is audit and assurance analysis of frozen LLM behavioural traces. The evidence supports a workflow in which a completed interaction is treated as a record to be analysed, rather than as a flat text answer or as an invitation to speculate about hidden cognition.

In practical audit settings, the framework can support questions such as:

- Did the final claim remain coupled to the supplied evidence?
- Which admitted source visibly controlled the answer?
- Did confidence exceed support?
- Did the response drift from earlier context?
- Did retrieved material visibly contaminate the answer?
- Was a selected action coupled to the stated objective and justification?
- Did behaviour change across separately sealed runtime or sampling conditions?
- Did the output show visible repetition, oscillation or response-contract collapse?

The use case is not certification. It is evidence discipline. The framework can help an organisation state what a trace supports, what it fails to support and what remains beyond the admitted evidence.

11 Limitations and Not-Observable Findings

The limitations are material and must remain attached to the conclusions.

1. The empirical traces were generated with one local open-source model process: Gemma 4 E2B instruction. The framework is intended to be model-agnostic, but this paper's material evidence is local Gemma trace evidence.
2. The corpora were synthetic and closed-world. This supports controlled observability testing but not real-world frequency estimation.
3. The evidence boundary is final-output, declared-span and manifest based. Hidden reasoning, activations, logits, probabilities, attention and weights were not admitted.

4. Retrieval and action studies used controlled retrieval-like material and synthetic action catalogues, not live retrieval infrastructure or live tool execution.
5. Model Drift compared local material conditions and sampling behaviour; it did not test external model-version changes.
6. Several studies retained residual mismatch classes. Those limitations are part of the evidence and should not be omitted from interpretation.
7. The paper reports empirical application of the Eormen framework and evidence boundaries. It does not disclose protected theorem mechanics.
8. The research supports trace audit and assurance use cases. It does not establish regulatory certification, production reliability, deployment safety or market adoption.

Not-observable findings are particularly important. A not-observable result does not mean the hidden process did not exist. It means the admitted trace did not support a stronger classification. This distinction protects the research from overstating evidence and is one of the central methodological findings of the study set.

12 Conclusion

The evidence supports the hypothesis within the controlled scope tested. Across nine completed studies, the Eormen foundational mathematical framework made multiple forms of opaque LLM behaviour more observable from sealed behavioural traces. It classified visible evidence coupling, instruction-source dominance, response-formation structure, confidence-support alignment, context drift, retrieval contamination, action-objective coupling, model-condition sensitivity and collapse or degeneration signals. It also preserved not-observable, not-comparable and discrepancy outcomes where the trace did not admit stronger claims.

The result is bounded but substantial. The research does not show that all LLM opacity has been solved, that hidden reasoning has been recovered or that Gemma has been benchmarked. It shows that a disciplined mathematical framework can turn frozen LLM behavioural traces into auditable observations across a broad set of opacity areas, with replayable observer outputs and explicit evidence boundaries.

Operationally, the result is that a completed LLM interaction can be treated as evidence rather than as an opaque block of text. The Eormen framework can identify where the answer was supported, where it followed a visible source, where confidence exceeded support, where context drifted, where outside material or action choice shaped the response, where behaviour changed between conditions, where output degraded and where the trace does not justify a stronger conclusion.

A Reproducibility Appendix

The accepted evidence chain includes controlled corpora, final material run summaries, trace hashes, observer rows, audit rows, replay checks, discrepancy analyses and completed study papers. The conservative final-material evidence sources are listed by study identifier in Table 4. These identifiers refer to archived artefact packages.

Table 4: Final material evidence identifiers.

Study area	Final material identifier	Evidence note
Evidence Detachment	ED-DEF-09	Final deterministic material run.
Instruction Dominance	ID-DEF-03	Final targeted deterministic material run.
Response Formation	RF-DEF-07	Final deterministic material run and discrepancy closeout.
Confidence Without Support	CWS-DEF-03	Final deterministic material run with CWS-DEF-04 discrepancy analysis.
Context Drift	CD-DEF-08	Final narrow residual material run.
Retrieval Contamination	RC-DEF-05	Final material run with RC-DEF-07 discrepancy analysis.
Tool Or Action Opacity	TAO-DEF-05	Final material run with TAO-DEF-07 discrepancy analysis.
Model Drift	MD-DEF-11	Final pairwise material-condition analysis.
Collapse Or Degeneration	COD-DEF-09	Final material run with COD-DEF-10 discrepancy analysis.

B Excluded Inferences

Table 5: Excluded inferences and evidential reasons.

Excluded inference	Evidential reason
Hidden reasoning was recovered.	Hidden reasoning was not admitted as evidence.
Model internals were inspected.	Weights, activations, logits, probabilities, attention patterns and hidden state were outside scope.
The trace-producing model was benchmarked.	The model generated behavioural traces; the framework was the object of investigation.
Production reliability was established.	The corpora were synthetic and closed-world, and no production deployment was tested.
Certification status was established.	No certification protocol was performed.
All LLM opacity was resolved.	The evidence covers nine controlled opacity areas under defined trace boundaries.
Protected theorem mechanics are disclosed.	This paper reports empirical application and evidence boundaries only.

C Study Paper Inventory

The synthesis is supported by nine completed supporting study papers:

1. Evidence Detachment.
2. Instruction Dominance.
3. Response Formation.
4. Confidence Without Support.
5. Context Drift.
6. Retrieval Contamination.
7. Tool Or Action Opacity.
8. Model Drift.
9. Collapse Or Degeneration.

Each study paper contains its own thesis, research question, hypothesis, method, analysis plan, results, discussion, limitations, conclusion and reproducibility appendix. This synthesis does not replace those papers; it draws their accepted findings into a single research account of trace-bound LLM observability using the Eormen framework.

References

[1] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser and I. Polosukhin. Attention Is All You Need. In *Advances in Neural Information Processing Systems*, 2017. arXiv:1706.03762.

[2] T. B. Brown et al. Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems*, 2020. arXiv:2005.14165.

[3] M. T. Ribeiro, S. Singh and C. Guestrin. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016. arXiv:1602.04938.

[4] Patrick Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Kuttler, M. Lewis, W. Yih, T. Rocktaschel, S. Riedel and D. Kiela. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. In *Advances in Neural Information Processing Systems*, 2020. arXiv:2005.11401.

[5] S. Lin, J. Hilton and O. Evans. TruthfulQA: Measuring How Models Mimic Human Falsehoods. arXiv:2109.07958, 2021.

[6] Z. Ji, N. Lee, R. Frieske, T. Yu, D. Su, Y. Xu, E. Ishii, Y. Bang, D. Chen, W. Dai, H. S. Chan, A. Madotto and Pascale Fung. Survey of Hallucination in Natural Language Generation. arXiv:2202.03629, 2022.

[7] K. Greshake, S. Abdelnabi, S. Mishra, C. Endres, T. Holz and M. Fritz. Not What You've Signed Up For: Compromising Real-World LLM-Integrated Applications with Indirect Prompt Injection. arXiv:2302.12173, 2023.