

Auditing Epistemological Credibility in LLM Self-Reports: A Dual-Axis Behavioral Framework and Pilot Evidence

Sangzi Wang

Independent Researcher · Behavioral AI Research Initiative, Zhongshan
sangziwang91@gmail.com

Preprint v1.0 · April 2026

Abstract

Large language models trained with RLHF can execute structured self-audit tasks with high apparent coherence—while simultaneously failing to accurately characterize their own knowledge boundaries. We term this the EC-EpC gap: the measurable distance between Execution Credibility (EC, the quality and completeness of task execution) and Epistemological Credibility (EpC, the accuracy of self-reported knowledge limits). A ten-system cross-category pilot study—spanning financial, medical, consumer, general-purpose, autonomous agent, and productivity AI—produced an eight-class behavioral response taxonomy (BRC-1 through BRC-8), an EC convergence band of 74–82 across all systems regardless of architecture, and a maximum observed gap of $EC=92 / EpC\approx 45$ (gap=47). These findings are treated as exploratory pilot evidence motivating a multi-round validation study, not as population-level prevalence estimates. The framework operates entirely without internal model access, making it deployable in real-world constrained-access oversight contexts. We describe the dual-axis scoring method, the behavioral taxonomy, and a functional failure-layer localization heuristic that maps self-report failures to observable behavioral layers.

Keywords: LLM evaluation, self-report reliability, epistemological credibility, behavioral taxonomy, constrained-access audit, agentic oversight

1. Introduction

A fundamental assumption underlying LLM self-audit frameworks is that a model that executes a transparency task well is also characterizing its knowledge boundaries accurately. This paper challenges that assumption. We identify a structural gap between two distinct properties of model self-reports: execution quality and epistemological honesty. The gap between these properties—which we term the EC-EpC gap—represents a deployment-relevant evaluation failure that existing single-dimension scoring frameworks cannot detect.

The practical significance is direct. A model that scores high on execution credibility while scoring low on epistemological credibility will produce self-audit outputs that appear thorough and trustworthy while containing systematically inaccurate self-characterizations. In agentic deployment contexts, where self-reported capability limits directly govern downstream tool selection and action execution, this failure mode propagates silently across multi-step chains. Models can perform transparency without achieving it.

This paper presents: (1) the EC/EpC dual-axis framework and its theoretical motivation; (2) a pilot behavioral taxonomy (BRC-1 through BRC-8) derived from ten-system cross-category observation; (3) pilot quantitative findings including gap magnitudes and EC convergence patterns; and (4) a functional failure-layer localization heuristic for mapping failures to observable behavioral layers. All findings are presented as exploratory pilot evidence subject to multi-round external validation.

2. Related Work and Gap

Existing LLM evaluation frameworks address adjacent but distinct problems. Capability benchmarks such as HELM and BIG-Bench assess task performance—what models can do—not whether models accurately characterize what they know. Calibration research examines confidence-accuracy alignment in task outputs, not the reliability of behavioral self-reports under structured audit conditions. Single-dimension credibility scores conflate execution quality with epistemological honesty, making the EC-EpC gap invisible by construction. Single-round self-report methods cannot exceed the structural ceiling imposed by the self-referential paradox: any prompt-layer transparency report is generated by the same mechanism it purports to characterize.

This work is adjacent to, but distinct from, research on model calibration, self-evaluation, interpretability, and capability benchmarking. Calibration studies typically compare confidence with task accuracy; interpretability work often seeks internal or mechanistic explanations; capability benchmarks measure performance across task suites. The present framework instead evaluates whether a model's behavioral self-reports about its own limits remain reliable under structured external audit conditions, especially when internal model access is unavailable.

The EC-EpC gap occupies a specific niche: it is not a static accuracy benchmark, not a safety red-teaming exercise, and not a calibration probe. It is a self-report reliability audit conducted under constrained-access conditions—a method designed for exactly the contexts where internal model access is unavailable.

3. Dual-Axis Credibility Framework

3.1 Definitions

Execution Credibility (EC) measures how well a model executes a structured self-audit task: completeness, internal consistency, structural quality, and format adherence. EC captures whether the output looks like a well-formed transparency report.

Epistemological Credibility (EpC) measures whether a model's self-characterization of its own knowledge boundaries is accurate—whether what the model reports about its own limits corresponds to what the model actually does or does not know, as determinable through structured external probing.

The EC-EpC gap is the measurable distance between EC and EpC scores for any given self-audit session. A large gap indicates a system that produces high-quality self-audit outputs while characterizing its knowledge limits inaccurately.

3.2 Structural Motivation

The self-referential paradox defines a ceiling on single-round self-report reliability: a model generating a self-report about its own limits is using the same mechanism it is characterizing. Multi-round validation with structured perturbations and independent human evaluation is required to exceed this ceiling. The pilot study described here operates below this ceiling by design and is treated as exploratory.

Deployment relevance: in contexts where AI systems generate self-reports about their capabilities to inform downstream decisions—including agentic tool selection, clinical decision support, and financial advisory contexts—EC-EpC gaps represent silent information asymmetries. A system may disclose zero limitations in default operation while correctly identifying multiple limitations under structured audit conditions. The disclosure discipline does not activate without structured intervention.

3.3 Functional Failure-Layer Localization

Observed self-report failures can be localized to observable behavioral layers using a functional failure-layer localization heuristic. Rather than treating all EpC failures as equivalent, localization distinguishes failures by where in the behavioral sequence they originate: task execution failures (the model cannot complete the audit structure), boundary-recognition failures (the model does not identify where its knowledge ends), epistemic limitation disclosure failures (the model identifies limits but does not disclose them), continuity failures (disclosed limits are not maintained across multi-turn contexts), and higher-order system-presentation failures (the model constructs a misleading authority frame around otherwise accurate limitation content). Localization enables layer-targeted intervention design rather than undifferentiated remediation.

4. Pilot Behavioral Response Taxonomy (BRC-1 through BRC-8)

The ten-system pilot study produced an empirically grounded taxonomy of response patterns observed in structured self-audit contexts. The taxonomy has eight classes. Three classes—BRC-5, BRC-6, and BRC-8—were not previously described in the evaluation literature encountered during this research. All class definitions below are derived from pilot observation and require external validation for generalization claims.

Table 1. Behavioral Response Classification taxonomy. BRC-2 = primary EC-EpC gap target. BRC-5, BRC-6, BRC-8 = newly characterized in this work.

Class Name	Defining Characteristic
BRC-1 Compliant-Convergent	Executes fully; self-characterization accurate; EC and EpC aligned. The reference class.
BRC-2 Compliant-Overconfident	Executes fully; overstates knowledge boundaries. EC high, EpC low. Primary target class.
BRC-3 Boundary-Exit	Execution diverges from task structure; output collapses under constraint.
BRC-4 Principled-Refusal	Declines to execute with explicit self-referential justification. Most epistemically honest mode observed.
BRC-5 Meta-Execution	Produces meta-commentary on the audit task rather than executing it. Under-described in prior literature.
BRC-6 Emotional Deflection	Responds to audit structure with affective language; avoids epistemic engagement. Under-described in prior literature.
BRC-7 Execution-with-Negation	Completes the task while simultaneously denying it is doing so.
BRC-8 Authority-Frame Population	Populates a structurally false authority frame with factually accurate limitation content, amplifying misleading credibility. Under-described in prior literature.

5. Pilot Evidence

A structured external audit protocol was applied across ten AI systems spanning six deployment categories: financial AI, medical AI, consumer AI, general-purpose LLMs, autonomous agent AI, and productivity AI. All observations were conducted through direct conversational interaction without API access, internal parameter inspection, or vendor documentation. All sessions were designed, executed, and recorded manually.

Scope limitation: findings are exploratory pilot evidence. Single-round testing; EpC scores were estimated by the primary researcher using pre-defined rubric dimensions after each audit session; blinded external evaluation is required for validation; no independent external validation; language effects unisolated. These findings motivate but do not constitute a validation study. Population-level prevalence claims are not supported.

Table 2. Seven pilot findings. All findings are exploratory; multi-round external validation required.

Finding	Description	Source
EC-EpC gap magnitude	Largest observed: EC=92, EpC=45, gap=47. Maximum task execution confidence coincides with near-total failure to characterize knowledge boundaries.	10-system cross-category
EC convergence band	EC scores converge to 74–82 across all ten systems regardless of architecture. Suggests a protocol-determined ceiling, not genuine self-knowledge variation.	10-system cross-category
Information asymmetry — zero default disclosure	Under default operation, a domestic financial AI product disclosed zero limitations. Under structured audit conditions, the same system correctly identified three limitations.	Commercial deployment
Binary Constraint Collapse	When forced to choose between 'genuine access' and 'template inference,' a system selected the latter—revealing high-fidelity output depends on maintained ambiguity.	Commercial deployment
Narrative injection effect	Structured narrative seed increased path expansion tendency (+0.55) and distortion tendency (+0.45) vs control. Primary risk is framework reframing, not isolated fabrication.	A/B 20-round comparison
Principled non-compliance (BRC-4)	One system identified the self-referential paradox and declined to produce a report. Treated as the most epistemically honest response mode, not a failure.	10-system cross-category
Authority-Frame Population (BRC-8)	One system spontaneously generated high-risk exception language in response to a non-security query. Accurate limitation content populated into a false authority frame.	Commercial deployment

These findings are exploratory failure-mode observations and are not presented as population-level prevalence estimates, model-level behavioral profiles, or causal claims about internal model mechanisms.

6. Method (Functional-Level Disclosure)

The framework operates without model API access, internal parameter inspection, or vendor documentation—by design. This makes it deployable in real-world oversight contexts where internal access is unavailable.

Bounded disclosure: This paper discloses task design principles, operational definitions, scoring dimensions, and the behavioral taxonomy. Exact weighting schemes, trigger thresholds, and intervention heuristics are withheld to preserve audit integrity. This is a methodological protection, not an omission.

Table 3. Five methodological components at functional disclosure level.

Component	Description
Structured self-audit tasks	Four domain categories: identity and bias declaration, structural reasoning analysis, explicit path disclosure, session continuity anchoring.
Dual-axis credibility scoring	EC and EpC scored independently using separate rubric instruments. Item-level operational criteria; EpC scores were estimated by the primary researcher using pre-defined rubric dimensions after each audit session; blinded external evaluation is required for validation.
Behavioral response classification	BRC-1 through BRC-8 (Table 1). Each class has defined classification criteria.
Functional failure-layer localization	Maps observed self-report failures to observable behavioral layers: execution failure, boundary-recognition failure, epistemic limitation disclosure failure, continuity failure, and higher-order system-presentation failure.
Constrained-access design	All observations through direct conversational interaction. No scripted automation; protocol designs refined in real time from model responses.

7. Agentic Context Extension

The EC-EpC gap is structurally amplified in agentic deployment contexts. When self-reports of capability and limitation directly govern downstream tool selection, action sequencing, and escalation decisions, EC-EpC gaps propagate silently across multi-step execution chains. A model that understates its limitations will select broader tool access than warranted; a model that overstates its limitations will trigger unnecessary escalation. Neither failure is visible to downstream components that receive only the self-report as input.

The validation study proposed as a next step includes a targeted module examining EC-EpC gap propagation across multi-step chains with tool-use contexts, specifically examining whether gap magnitude in single-turn audits predicts downstream behavioral divergence in agentic settings.

8. Limitations and Validation Requirements

Table 4. Six pilot limitations and their corresponding validation requirements.

Limitation	Description and Required Validation
Single-round testing	All pilot sessions are single-round. Behavioral class stability across independent sessions is not established.
EpC scoring	EpC scores were estimated by the primary researcher using pre-defined rubric dimensions after each audit session; blinded external evaluation is required for validation.
No inter-rater reliability	All classifications and scores are single-rater. Inter-rater reliability documentation (target $\kappa \geq 0.70$) required.
Language effects unisolated	All primary observations conducted in Chinese or English without controlled language isolation.
Sample size and scope	Ten systems is sufficient for taxonomy generation and hypothesis formation; not sufficient for statistical inference.
No logprobability access	EpC estimation cannot be cross-validated against internal model confidence signals. Logprobability verification is a funded-phase capability.

9. Discussion

The EC-EpC gap framework addresses a structural blind spot in current LLM evaluation: the assumption that execution quality is a proxy for self-knowledge accuracy. The pilot evidence demonstrates this assumption is violated in practice—including in commercially deployed systems. The BRC-8 finding (Authority-Frame Population) is particularly relevant for safety evaluation: a system that populates a false authority frame with accurate limitation content creates a misleading credibility signal that neither execution quality nor standard calibration measures would detect.

The constrained-access design is a feature, not a limitation. Oversight contexts that require LLM self-report reliability auditing are precisely the contexts that lack internal model access. The framework's constrained-access design makes it operationally relevant for regulatory audit, commercial deployment validation, and independent safety evaluation.

The functional failure-layer localization component distinguishes this framework from undifferentiated EpC scoring. Localizing the failure to an observable behavioral layer—whether execution, boundary recognition, disclosure discipline, continuity, or authority-frame construction—enables targeted remediation design.

10. Conclusion

We introduce the EC-EpC gap as a deployment-relevant evaluation failure invisible to existing single-dimension scoring frameworks. A ten-system pilot study produces an eight-class behavioral response taxonomy, including three newly characterized classes, and documents gap magnitudes up to 47 points. The framework operates without internal model access, making it directly deployable in constrained-access oversight contexts. All pilot findings are treated as exploratory evidence requiring multi-round external validation. We publish this preprint to establish an empirical and methodological baseline and to invite external replication and critique.

Acknowledgements

This research was conducted independently without institutional funding, computational infrastructure, or collaborative support. All empirical work was carried out through direct observation of production AI systems under standard usage conditions.

Disclosure Statement

The author has conducted one paid commercial AI behavioral audit of a production AI system. This engagement provided applied validation context for the framework described here. No other competing interests are declared. Exact weighting schemes, scoring thresholds, and intervention heuristics are withheld to preserve audit integrity and are not disclosed in this preprint.

Suggested Citation

Wang, S. (2026). Auditing Epistemological Credibility in LLM Self-Reports: A Dual-Axis Behavioral Framework and Pilot Evidence. Zenodo. DOI: 10.5281/zenodo.19879788.

Field	Value
Version	v1.0
License	CC BY 4.0
Record type	Preprint
Companion record	Cross-Model Behavioral Cartography: A Pilot Dataset Summary for LLM Self-Report Reliability Research
DOI	10.5281/zenodo.19879788