yAI SYSTEMS NOTE 006

The Measurement Problem: Why Answer-Level Metrics Misdiagnose RAG Systems

Version 1.0 April 21, 2026

Yassin Hafid
Founder & CEO
LinkedIn

ABSTRACT

This note identifies a structural failure in RAG evaluation, which we term Diagnostic Collapse: the reduction of a multi-stage pipeline into a single scalar score that cannot distinguish retrieval failure from grounding failure, or a faithfully grounded answer from an unsupported but superficially correct one.[1][4] The empirical stakes are concrete: in FRAMES, state-of-the-art models achieve 0.408 accuracy without retrieval, while multi-step retrieval improves accuracy to 0.66.[1] Across recent evaluation work, answer-level metrics often cannot reveal whether apparent success came from retrieval quality, grounding quality, or unsupported answer generation.[3][4][5] In GaRAGe's evaluation, models reach at most a 31% true positive rate in deflections, indicating that they often generate rather than abstain when grounding is insufficient.[3] The central evaluation problem in RAG is not better scoring; it is measuring at the wrong level of analysis.[1][2][3][4][5]

1. The Question

If a RAG system returns a correct answer, what has actually been verified?

Does the output reflect a faithful synthesis of retrieved context, or a correct answer that is not properly supported by the provided evidence and may instead rely on information already internal to the model?[1][4] And if current evaluation frameworks cannot reliably distinguish between the two, how can answer-level scoring be trusted to detect the difference?[3][4][5]

---

2. Scope and Definitions

Accuracy Fallacy

A term used in this note for cases where a system appears accurate but the answer is not properly supported by the retrieved evidence, even though it may still be factually correct.[1][4]

Diagnostic Collapse

A term used in this note for the reduction of a multi-stage RAG pipeline into a single scalar score that cannot distinguish retrieval failure from grounding failure, or a faithfully grounded answer from an unsupported but superficially correct one.[1][4]

ACU (Automated Context Utilisation)

A metric used to assess how effectively a model utilizes retrieved context and to compare performance on synthetic versus real-world retrieved evidence.[2]

Deflection Rate

A term used in this note for a system's ability to provide a deflective response when relevant grounding is insufficient; in GaRAGe's evaluation, models reach at most a 31% true positive rate in deflections.[3]

Abstention Calibration Failure

A term used in this note for the failure of a model to correctly judge when retrieved grounding is insufficient to support an answer, resulting in generation rather than deflection.[3]

Atomic Claim

A term used in this note for the minimal factual unit of a generated response; MedRAGChecker[5] operationalizes claim-level verification by decomposing answers into atomic claims and using them to separate under-evidence from contradiction and related diagnostic categories.

> **Process-Aware Evaluation (PAE)**
>
> A term used in this note for diagnostic evaluation that attributes failure across multiple stages of a RAG pipeline rather than only at the final answer level.

Scope: Diagnostic methodologies for isolating RAG failure modes across retrieval, grounding, answer generation, and claim-level verification. This note focuses on structural attribution and measurement, rather than architectural design, in knowledge-intensive RAG settings.[1][2][3][4][5]

---

3. Key Findings

- **The Grounding-Accuracy Gap:** Baseline evaluations show that state-of-the-art models achieve 0.408 accuracy without retrieval.[1] In FRAMES, performance improves to 0.66 under a multi-step retrieval pipeline;[1] this indicates that answer-level correctness alone cannot distinguish what came from retrieval from what the model could do without it.
- **The Deflection Crisis:** In GaRAGe's evaluation, models reach at most a 31% true positive rate in deflections.[3] This indicates that they often generate rather than provide a deflective response when grounding is insufficient, even when abstention is the correct behavior.[3]
- **The Realism Inflation:** DRUID shows that synthetic datasets can inflate measured context utilisation by exaggerating context characteristics rare in real retrieved data.[2] In this sense, synthetic benchmarks can make RAG systems appear more robust than they are under realistic retrieved context.[2]
- **The Over-Summarization Bias:** In GaRAGe's evaluation, models tend to over-summarize rather than ground their answers strictly on the annotated relevant passages, reaching at most 60% on Relevance-Aware Factuality.[3] Because the available grounding often contains a mixture of relevant and irrelevant passages, this behavior indicates weak relevance filtering at answer time.[3]
- **Safety-Critical Claim-Level Failures:** Aggregate metrics can overlook isolated, unsupported or contradictory atomic claims in long-form outputs.[5] In biomedical settings, these fine-grained failures can carry direct safety implications, which whole-answer scoring may fail to surface.[5]

---

4. Technical Deep Dive: Five Compounding Evaluation Failures

A. The Grounding-Accuracy Gap (Attribution Masking)

FRAMES shows that state-of-the-art models can achieve 0.408 accuracy without retrieval.[1] This creates an attribution problem: answer-level success does not by itself show whether the output was actually supported by the provided context.[1][4] In the same benchmark, performance improves to 0.66 under a multi-step retrieval pipeline;[1] this indicates that retrieval materially changes the answer process even when answer-level scoring alone cannot say how. More broadly, [4] shows that correctness and grounding can diverge: a response may be factually true but unsupported by its citations, or well grounded yet still wrong in other ways.[4] Answer-level metrics collapse these different states into a single score.[4] Accuracy Fallacy names one such state, while Diagnostic Collapse describes the broader evaluative blind spot.

B. The Deflection Crisis

GaRAGe evaluates whether models provide a deflective response when there is insufficient information and finds that they reach at most a 31% true positive rate in deflections.[3] This suggests an Abstention Calibration problem: a failure to provide a deflective response when grounding is insufficient, even when abstention is the correct behavior.[3]

C. Noise Sensitivity (The Over-summarization Bias)

In GaRAGe's evaluation, models tend to over-summarize rather than ground their answers strictly on the annotated relevant passages, reaching at most 60% on Relevance-Aware Factuality.[3] Because the available grounding often contains a mixture of relevant and irrelevant passages, this suggests that irrelevant grounding can bleed into the final answer.[3] RAGVUE's emphasis on strict claim-level faithfulness reinforces the need to separate grounded synthesis from broad answer plausibility.[4]

D. Realism Inflation (The Synthetic Context Gap)

DRUID shows that synthetic datasets such as CounterFact can inflate measured Context Utilisation by exaggerating context characteristics that are rare in real retrieved data.[2] In this sense, synthetic-only testing can make RAG systems appear more

robust than they are under real-world retrieved evidence, including unreliable and insufficient context.[2]

E. The Claim-Level Verification Gap

Whole-answer metrics can hide safety-critical errors in long-form synthesis.[5] Fine-grained analysis shows that a correct-looking answer can still contain isolated atomic claims that are unsupported or contradicted by the retrieved evidence.[5]

---

5. A Taxonomy of Evaluation Failure Modes

Failure Mode	Primary Mechanism	Diagnostic Symptom	Ref
Diagnostic Collapse	Single scalar compresses retrieval, grounding, and reasoning failures into one non-diagnostic outcome.	Non-diagnostic scalar scores and limited component-level attribution.	[1][4]
Accuracy Fallacy	A response appears correct even though it is not properly supported by the retrieved evidence.	Non-trivial answer accuracy without retrieval, or factually acceptable responses that remain unsupported by their evidence.	[1][4]
Realism Inflation	Evaluation on synthetic datasets can overstate context utilisation relative to real-world retrieved evidence.	Inflated context-utilisation results under synthetic settings that do not transfer cleanly to real retrieved evidence.	[2]
Noise Sensitivity	Models over-summarize available grounding instead of isolating the passages annotated as relevant.	Relevance-Aware Factuality reaches at most 60%, while answers fail to stay strictly grounded on the relevant passages.	[3]
Deflection Crisis	Models fail to provide a deflective response when relevant grounding is insufficient.	In GaRAGe's evaluation, models reach at most a 31% true positive rate in deflections.	[3]
Atomic Claim Failure	Whole-answer or aggregate metrics can mask isolated, unsupported, or contradicted atomic claims.	Under-evidence, contradiction, and safety-critical errors become visible through fine-grained claim-level verification.	[5]

---

6. Implications for AI System Design

- **Reject answer-level success as a sufficient signal of pipeline reliability.** A correct response signifies only that the output satisfied the surface-level query on a specific instance, not that the system retrieved, grounded, and synthesized the answer correctly.[3][4][5] Success may reflect the Accuracy Fallacy: an answer that appears correct without being properly supported by the provided evidence.[1][4]
- **Transition to Process-Aware Evaluation (PAE).** RAG should be measured as a multi-stage system, not only by its final answer. FRAMES motivates integrated end-to-end evaluation, DRUID isolates context utilisation, RAGVUE decomposes retrieval and grounding behavior, and MedRAGChecker exposes claim-level failures that aggregate scoring can miss.[1][2][4][5]
- **Calibrate evaluation to the actual retrieval environment.** Synthetic datasets can inflate measured context utilisation by exaggerating context characteristics that are rare in real retrieved data.[2] Evaluation should therefore be calibrated against the complexity and diversity of the actual retrieval environment, meaning the content the system will truly encounter at deployment time.[2]
- **Formalize abstention as a primary measurable behavior.** When grounding is insufficient, the correct pipeline response may be a deflective response rather than answer generation.[3] GaRAGe explicitly evaluates this behavior and reports that models reach at most a 31% true positive rate in deflections; this shows that abstention remains weak even when insufficient grounding is part of the evaluation setup.[3]
- **Instrument fine-grained diagnostics as a runtime control layer** (a forward-looking implication of RAGVUE[4] and MedRAGChecker[5]): RAGVUE shows that diagnostic evaluation can be automated and integrated into practical RAG workflows,[4] while MedRAGChecker shows that claim-level verification can reliably flag unsupported or contradicted claims with safety implications.[5] Together, these results suggest that fine-grained diagnostics could evolve from offline evaluators into runtime verification signals in high-risk deployments.[4][5]

7. Open Questions

- **Diagnostic Attribution:** What observable signals are sufficient to distinguish a faithfully grounded answer from an Accuracy Fallacy success?
- **Minimal Sufficient Trace:** What is the minimal Process-Aware Evaluation trace needed to break Diagnostic Collapse without making evaluation impractical?
- **Correct but Ungrounded:** When should a factually correct but weakly grounded answer be scored as failure rather than success?
- **Deflection Thresholding:** What degree of evidence insufficiency should trigger deflection, and should that threshold vary by domain?
- **Metric Gaming:** How can evaluators detect when systems optimize for ACU, Deflection Rate, or Atomic Claim verification without becoming genuinely more reliable?
- **Cross-Source Claim Consistency:** How should claim-level evaluators handle contradictory Atomic Claims across multiple retrieved sources?

---

8. References

[1] Krishna et al., Fact, Fetch, and Reason: A Unified Evaluation of Retrieval-Augmented Generation, NAACL 2025.

[2] Hagström et al., A Reality Check on Context Utilisation for Retrieval-Augmented Generation, ACL 2025.

[3] Sorodoc et al., GaRAGe: A Benchmark with Grounding Annotations for RAG Evaluation, ACL 2025.

[4] Murugaraj et al., RAGVUE: A Diagnostic View for Explainable and Automated Evaluation of Retrieval-Augmented Generation, EACL 2026.

[5] Ji et al., MedRAGChecker: Claim-Level Verification for Biomedical Retrieval-Augmented Generation, arXiv:2601.06519, 2026.

This note synthesizes findings from recent research on RAG evaluation. The interpretations presented reflect the author's reading of the current literature.