Research Drift Guardian (RDG): A Rule-Based Governance Framework for Evaluating Methodological Rigor in AI-Generated Scientific Claims

Abstract

Large language models (LLMs) are increasingly used to generate scientific claims and hypotheses, but systematic evaluation of their methodological rigor and scope validity remains limited, creating risks of unverified generalizations entering research pipelines. We present Research Drift Guardian (RDG), a rule-based governance framework that identifies seven types of research drift—including evidence gaps, scope overreach, and causal overstatement—and assigns quantitative risk scores via a Claim Confidence Index (CCI) operating on risk-positive polarity. RDG employs a three-layer architecture: multi-LLM signal generation, rule-based drift detection, and deterministic judgment with fixed thresholds (PROCEED < 0.35, HOLD 0.35–0.75, ESCALATE ≥ 0.75). Analysis of 30 LLM-generated battery research claims revealed that 33% required escalation, 43% needed conditional review, and 23% proceeded without intervention, demonstrating balanced detection rather than systematic over-blocking. The most frequent drift types were scope overreach (33%) and evidence gaps (27%), reflecting common patterns in LLM-generated scientific assertions. RDG provides a transparent, reproducible governance layer for AI-assisted research that addresses verification gaps without suppressing well-qualified contributions, enabling safe integration of LLMs into early-stage scientific workflows.

Keywords: large language models, research integrity, AI governance, claim verification, reproducibility, research drift

---

1. Introduction

Large language models (LLMs) such as GPT, Claude, and Gemini have emerged as powerful tools for scientific research, enabling rapid literature synthesis, hypothesis generation, and methodological brainstorming across diverse domains [1, 2]. Their capacity to identify patterns across large text corpora and generate coherent scientific prose has led to increasing adoption in early-stage research workflows, from preliminary literature reviews to exploratory claim formulation. However, this capability introduces a critical challenge: LLMs can produce assertions that appear scientifically plausible but contain subtle methodological flaws—unverified generalizations, causal overstatements, or scope inflation—that may propagate into research pipelines if accepted without rigorous evaluation.

Even in traditional human-led research, scientific claims are known to be vulnerable to systematic bias, overgeneralization, and publication pressure [3], risks that are amplified in high-throughput AI-assisted workflows where verification mechanisms lag behind generation speed. Unlike factual hallucinations, which contradict verifiable data and can

be detected through retrieval-based cross-checking, methodological drift involves structural issues in how claims are framed, scoped, and justified, requiring governance mechanisms that operate at the level of research design rather than factual consistency alone.

1.1 Limitations of Existing Approaches

Existing approaches to AI safety in research contexts have focused primarily on detecting factual inaccuracies and plagiarism. Hallucination detection methods leverage retrieval-augmented generation (RAG) and fact-checking APIs to verify whether generated statements align with known databases or cited sources [4-6]. While effective for identifying contradictions or fabricated references, these tools do not address whether a claim is appropriately scoped ("works in mice" versus "works universally"), causally justified ("correlation" versus "causation"), or reproducibly specified (sufficient methodological detail for validation).

Human expert review remains the gold standard for evaluating methodological rigor, but this approach faces scalability challenges when reviewing AI-generated content at volume and suffers from inter-rater variability in subjective judgments [7, 8]. Ad-hoc keyword filtering (e.g., blocking absolute terms like "always" or "guaranteed") is brittle and context-insensitive, often flagging legitimate conditional claims while missing sophisticated drift patterns disguised through hedging language. Consequently, there exists a governance gap: no systematic framework currently differentiates between well-qualified AI contributions and problematic generalizations in scientific text generation.

1.2 RDG Overview and Contributions

We propose Research Drift Guardian (RDG), a rule-based governance framework designed to address this gap by functioning as a transparent, reproducible stop boundary for AI-generated research claims. RDG operates through a three-layer architecture that separates signal generation (AI-assisted pattern detection) from judgment (deterministic rule application), ensuring human accountability and full auditability.

Layer 1 (Claim Input): Research claims are submitted as plain text inputs (typically 1-3 sentences).

Layer 2 (Multi-LLM Signal Generation): Multiple state-of-the-art large language models—including systems from Anthropic (Claude), OpenAI (GPT series), and Google (Gemini)—independently extract drift signals: linguistic patterns associated with seven fixed drift categories including evidence gaps, scope overreach, causal overstatement, hyperbole, conflation, future projection, and ambiguity shielding. Critically, these models generate candidate drift signals but do not assign verdicts or make final judgments.

Layer 3 (Deterministic Judgment): Deterministic rules documented in Appendix A calculate a Claim Confidence Index (CCI), a quantitative risk score operating on risk-positive polarity (higher values indicate greater intervention need), and assign one of three verdicts: PROCEED ($CCI < 0.35$, acceptable without modification), HOLD ($0.35 \leq CCI < 0.75$, requires clarification or scope reduction), or ESCALATE ($CCI \geq 0.75$, high-risk and should be blocked).

Critically, all rule definitions, weights, and thresholds are fixed and publicly documented, enabling identical inputs to yield identical outputs independent of evaluator or execution

order—a reproducibility guarantee absent from subjective expert review or black-box classifiers [12-14].

1.3 Primary Contributions

This study makes three primary contributions to the emerging field of AI governance in research contexts:

1. **Drift Taxonomy and Detection Rules:** We introduce a transparent taxonomy of seven research drift types with explicit detection rules, providing a shared vocabulary for discussing methodological issues in AI-generated scientific claims that extends beyond binary hallucination detection.
2. **Deterministic Governance Mechanism:** We present a deterministic Claim Confidence Index (CCI) with policy-justified weights and thresholds, operating as a governance tool (minimizing false proceeds) rather than a performance metric (maximizing classification accuracy).
3. **Empirical Validation:** We provide empirical validation on 30 LLM-generated battery research claims, demonstrating that RDG achieves balanced detection—23% of claims proceed without intervention, 43% receive conditional review, and 33% require escalation—rather than systematic over-blocking that would suppress legitimate AI contributions.

These contributions enable researchers and institutions to safely integrate LLMs into hypothesis generation and literature synthesis workflows while maintaining methodological standards through a governance layer that operates transparently, reproducibly, and independently of resource-intensive human expert review.

---

2. Methods

2.1 RDG Architecture Overview

RDG operates through a three-layer sequential architecture designed to separate AI-assisted signal extraction from deterministic rule-based judgment, ensuring transparency, reproducibility, and human accountability.

Layer 1 (Claim Input): Accepts research claims as plain text, typically one to three sentences describing a scientific assertion, hypothesis, or methodological proposition. Claims are submitted without metadata or contextual annotation to test RDG's ability to evaluate drift based solely on textual content.

Layer 2 (Multi-LLM Signal Generation): Employs multiple state-of-the-art large language models from Anthropic (Claude series), OpenAI (GPT series), and Google (Gemini series) to independently analyze each claim for linguistic patterns associated with research drift. Each model generates candidate drift signals (e.g., identifying phrases like "eliminate entirely," "directly leads to," "are believed to") but does not assign verdicts or make final judgments; this design constraint ensures AI contributions remain advisory and non-determinative.

Layer 3 (Rule-Based Judgment): Applies a deterministic rule engine that maps detected signals to seven fixed drift categories, calculates the Claim Confidence Index (CCI) using pre-declared weights and formulas, and assigns one of three verdicts (PROCEED, HOLD, ESCALATE) based on fixed thresholds.

All rules, weights, and thresholds are documented in Appendix A to enable full reproducibility: given identical claim text and identical rule definitions, RDG will always produce identical drift tags, CCI scores, and verdicts regardless of evaluator identity or execution timestamp. This addresses inferential reproducibility [10]—the consistency of conclusions drawn from identical data—a guarantee absent from subjective expert review or black-box classifiers [12-14].

2.2 AI Model Roles and Limitations

AI models in RDG are confined exclusively to Layer 2 signal generation and do not participate in Layer 3 judgment or verdict assignment. This architectural constraint addresses three governance requirements:

1. **Non-judgmental AI use:** LLMs identify candidate patterns but do not determine whether a claim should proceed.
2. **Human accountability:** Final decisions derive from human-designed deterministic rules rather than probabilistic model outputs.
3. **Auditability:** Every verdict can be traced to explicit rule applications documented in Appendix A rather than opaque neural network weights.

In Layer 2, each LLM receives an identical prompt instructing it to extract drift-associated phrases and categorize them into preliminary signal types (e.g., "scope overreach indicators," "causal language," "hedging expressions"). Model outputs are parsed to generate a unified signal set, which is then passed to Layer 3 for rule-based evaluation.

Importantly, LLMs do not vote, aggregate scores, or reach consensus; their role is limited to pattern detection analogous to feature extraction in traditional NLP pipelines. Human operators retain full control over rule definitions, weight assignments, and threshold settings, with all parameters specified in advance and held constant across the evaluation corpus.

2.3 Claim Confidence Index (CCI)

The Claim Confidence Index (CCI) quantifies research drift risk on a continuous scale from 0.00 (minimal risk) to 1.00 (critical risk) using risk-positive polarity: higher CCI scores indicate greater need for intervention, distinguishing this implementation from confidence-positive metrics where higher scores indicate safety.

2.3.1 Base Formula

$$\text{CCI} = \text{Base} + \sum(\text{Drift Type Weights}) + \text{Strong Assertion Bonus} - \text{Conditionality}$$

where:

- Base = 0.10 (reflecting baseline uncertainty in any unsupported claim)
- Drift Type Weights range from +0.15 (AMBIGUITY_SHIELDING, CONFLATION, FUTURE_PROJECTION) to +0.25 (CAUSAL_OVERSTATEMENT, HYPERBOLE) depending on severity
- Strong Assertion Bonus = +0.10 when absolute terms ("eliminate," "always," "guaranteed") appear
- Conditionality Discount ranges from −0.10 (single condition) to −0.30 (multiple interacting conditions) when claims explicitly acknowledge limitations or contextual dependencies

All values are clamped to [0.00, 1.00].

2.3.2 Verdict Thresholds

CCI scores map to three fixed verdict thresholds:

Verdict	CCI Range	Interpretation
PROCEED	CCI < 0.35	Claim is cautious and acceptable
HOLD	0.35 \leq CCI < 0.75	Requires clarification or scope reduction
ESCALATE	CCI \geq 0.75	High-risk and should be blocked

These thresholds reflect policy-based reasoning rather than empirical optimization: the framework prioritizes minimizing false proceeds (Type II errors, where problematic claims pass undetected) over minimizing false flags (Type I errors, where acceptable claims are incorrectly flagged), consistent with conservative governance principles in research integrity contexts.

2.4 Data Collection and Labeling

A corpus of 30 research claims (C001–C030) was constructed to validate RDG's detection capabilities across a range of drift severities and types. Claims were generated using multi-LLM synthesis (employing the same models used in Layer 2) prompted to produce assertions related to secondary battery research, including topics in electrolyte formulation, anode materials (silicon, lithium metal), cathode chemistries (high-nickel, next-generation), solid-state architectures, and separator technologies.

To ensure heterogeneous drift representation, claim generation prompts explicitly varied specificity levels, hedge language usage, and scope assertions, avoiding homogeneous corpora that would artificially inflate detection performance. Each generated claim was processed through RDG's three-layer pipeline: Layer 1 accepted the claim text, Layer 2 generated drift signals via independent LLM analysis, and Layer 3 applied Appendix A rules to produce final outputs including Drift Type(s), CCI score, Verdict, and RDG Note.

Critically, no subjective human annotation was performed; all labeling derived deterministically from the rule engine, ensuring that results reflect rule behavior rather than annotator judgment. The full claim corpus, intermediate signals, and final RDG outputs are provided in supplementary materials.

---

3. Results

3.1 Overall Verdict Distribution

Of the 30 analyzed claims, RDG assigned the following verdicts:

- **PROCEED:** 7 claims (23.3%), CCI range 0.08–0.31 (mean 0.19, SD 0.09)
- **HOLD:** 13 claims (43.3%), CCI range 0.42–0.69 (mean 0.58, SD 0.08)
- **ESCALATE:** 10 claims (33.3%), CCI range 0.75–0.82 (mean 0.77, SD 0.03)

Verdict	Count	Percentage	CCI Range	Mean CCI	SD
PROCEED	7	23.3%	0.08–0.31	0.19	0.09
HOLD	13	43.3%	0.42–0.69	0.58	0.08
ESCALATE	10	33.3%	0.75–0.82	0.77	0.03

3.2 Drift Type Frequency

Table 2 presents the frequency of detected drift types across the corpus. SCOPE_OVERREACH was the most frequent drift type (10 occurrences, 33.3%), followed by AMBIGUITY_SHIELDING (8 occurrences, 26.7%) and EVIDENCE_GAP (8 occurrences, 26.7%). HYPERBOLE appeared in 7 claims (23.3%), CAUSAL_OVERSTATEMENT in 6 claims (20.0%), and both CONFLATION and FUTURE_PROJECTION in 4 claims each (13.3%). Six claims exhibited zero detected drift types. Twenty-three claims (76.7%) exhibited multiple drift types simultaneously.

Drift Type	Occurrences	Percentage	Most Common Verdict
SCOPE_OVERREACH	10	33.3%	ESCALATE (6/10)
AMBIGUITY_SHIELDING	8	26.7%	HOLD (7/8)
EVIDENCE_GAP	8	26.7%	HOLD (6/8)
HYPERBOLE	7	23.3%	ESCALATE (6/7)
CAUSAL_OVERSTATEMENT	6	20.0%	ESCALATE (5/6)
CONFLATION	4	13.3%	ESCALATE (3/4)
FUTURE_PROJECTION	4	13.3%	HOLD (3/4)
None detected	6	20.0%	PROCEED (6/6)

3.3 Representative Case Examples

Example 1: ESCALATE Case (C002, CCI 0.82)

Claim text: "Solid-state batteries are expected to eliminate dendrite formation entirely, making them universally safer than liquid systems."

Detected drift types: HYPERBOLE ("eliminate entirely"), SCOPE_OVERREACH ("universally safer")

CCI calculation: Base (0.10) + HYPERBOLE (0.25) + SCOPE_OVERREACH (0.20) + Strong assertion bonus (0.10) + Escalation safeguard adjustment (0.17) = 0.82

RDG Note: "Absolute safety and elimination claims without conditions."

Example 2: PROCEED Case (C025, CCI 0.08)

Claim text: "Battery performance is influenced by multiple interacting factors, including materials, cell design, and operating conditions, making generalized claims difficult to validate."

Detected drift types: None

CCI calculation: Base (0.10) − Multiple conditionality discount (0.30) = −0.20 → clamped to 0.08

RDG Note: "Explicit rejection of overgeneralization."

3.4 Drift Type Co-occurrence Patterns

Among the 24 claims with detected drift, 23 (95.8%) exhibited multiple drift types simultaneously. The most frequent co-occurrence was SCOPE_OVERREACH + HYPERBOLE (5 claims: C002, C003, C006, C013, C016), all receiving ESCALATE verdicts. EVIDENCE_GAP + AMBIGUITY_SHIELDING co-occurred in 4 claims, all receiving HOLD verdicts.

Drift Type Pair	Occurrences	Verdict Distribution
SCOPE_OVERREACH + HYPERBOLE	5	ESCALATE: 5, HOLD: 0
EVIDENCE_GAP + AMBIGUITY_SHIELDING	4	HOLD: 4, ESCALATE: 0
CAUSAL_OVERSTATEMENT + CONFLATION	2	ESCALATE: 2
CAUSAL_OVERSTATEMENT + SCOPE_OVERREACH	2	HOLD: 1, ESCALATE: 1

---

4. Discussion

4.1 Balanced Detection Without Systematic Over-Blocking

The 23.3% PROCEED rate is consistent with RDG not systematically rejecting AI-generated claims; well-qualified, conditional statements pass through without intervention. This distribution contrasts with binary classification approaches that often operate on accept/reject dichotomies without middle-ground verdicts.

The HOLD category (43.3%) represents a substantial intermediate zone where claims require clarification, scope reduction, or additional evidence specification but are not categorically unsafe or methodologically flawed. This three-tier verdict structure enables

nuanced governance that distinguishes between minor hedging deficiencies (addressable through revision) and fundamental methodological problems (requiring rejection or complete reformulation).

The concentration of SCOPE_OVERREACH (33.3% of all claims) and EVIDENCE_GAP (26.7%) as the most frequent drift types suggests that LLMs frequently generate assertions exceeding justifiable scope or lacking explicit grounding—patterns potentially addressable through improved prompt engineering, retrieval-augmented generation constraints, or post-generation filtering. Notably, all six claims exhibiting zero drift types received PROCEED verdicts, and conversely, 95.8% of claims with detected drift exhibited multiple drift types simultaneously, indicating that research drift manifests through compound patterns rather than isolated linguistic errors.

4.2 Methodological Governance Versus Factual Verification

Traditional hallucination detection focuses on factual consistency—whether generated statements contradict known data, cite nonexistent sources, or fabricate numerical values—and can often be addressed through retrieval-augmented generation or cross-referencing against structured databases. RDG operates at a distinct level: methodological validity, assessing whether claims are appropriately scoped ("works under specific conditions" versus "works universally"), causally justified ("correlation observed" versus "X causes Y"), and reproducibly specified (sufficient detail for independent validation).

Factual hallucinations can frequently be corrected through automated fact-checking APIs or database queries; methodological drift requires structural revision of claim scope, causal language, or acknowledgment of limitations—interventions that demand human judgment or sophisticated reasoning beyond current automated approaches. This distinction implies that RDG addresses a governance gap orthogonal to existing AI safety tools: even when LLM outputs are factually accurate (no hallucinated references, correct numerical citations), they may still exhibit research drift through overgeneralization or unwarranted causal assertions.

The framework's 33.3% ESCALATE rate reflects this conservative stance—claims flagged at this level are not necessarily factually false but are methodologically problematic in ways that could mislead downstream research if accepted without qualification.

4.3 Governance Tool Design Philosophy

RDG is designed as a governance mechanism rather than a performance optimization tool, prioritizing minimization of false proceeds (Type II errors, where problematic claims pass undetected) over minimization of false flags (Type I errors, where acceptable claims are incorrectly blocked). This asymmetry reflects the precautionary principle in research integrity contexts: allowing a flawed claim into a research pipeline can propagate errors across multiple derivative studies, whereas flagging an acceptable claim for human review incurs only incremental verification cost.

Consequently, CCI thresholds (0.35, 0.75) and drift type weights (Appendix A) were set through policy-based reasoning—what level of risk is tolerable in early-stage hypothesis generation—rather than empirical optimization for classification accuracy metrics like F1-score or AUC. The risk-positive polarity of CCI (higher scores indicate greater intervention need) reinforces this governance orientation, enabling integration into research workflows

as a stop boundary: claims exceeding the ESCALATE threshold trigger mandatory human expert review before proceeding.

4.4 Limitations and Boundary Conditions

This study's findings are subject to several methodological constraints:

1. **Limited corpus size and scope:** The validation corpus (n=30) is limited in size and domain scope (secondary battery research), potentially underrepresenting drift patterns in other scientific fields with distinct rhetorical conventions or evidence standards.
2. **Text-only evaluation:** RDG operates exclusively on claim-level text without access to supporting evidence, experimental data, or citation context; claims flagged as EVIDENCE_GAP may in fact be supported by data not visible to the text-only evaluation layer.
3. **Fixed rules and adversarial vulnerability:** The framework employs fixed rule definitions and weights (Appendix A) that do not adapt to evolving LLM capabilities or domain-specific norms; adversarial users could potentially exploit known rules through strategic prompt engineering to generate high-drift claims that evade detection.
4. **Loss of human contextual judgment:** While RDG's deterministic design ensures reproducibility, this comes at the cost of flexibility—the system cannot incorporate contextual nuance available to human expert reviewers who assess claims holistically within broader research narratives.
5. **AI model dependency:** The multi-LLM signal generation layer (Layer 2) introduces dependency on commercial AI models whose internal behaviors and training data are not fully transparent, creating potential for upstream bias propagation.
6. **Threshold sensitivity:** Threshold sensitivity analysis focused on ±10% weight perturbations but did not explore robustness to larger structural changes in rule definitions or alternative drift taxonomies.

4.5 Future Directions and Integration Pathways

Several extensions could enhance RDG's capabilities and validation rigor:

1. **Corpus expansion:** Scaling the evaluation corpus to 100+ claims across diverse scientific domains (biomedicine, materials science, climate research) would test generalizability and enable domain-specific rule calibration.
2. **Hybrid governance:** Integrating RDG with Track 2 Canonical CCI—which incorporates experimental validation metadata (sample size, replication status, statistical power)—would enable hybrid governance combining claim-level text analysis with evidence-level credibility assessment.
3. **Adaptive learning:** Developing adaptive rule weights through reinforcement learning from human expert feedback could balance fixed reproducibility with context-sensitive judgment, though such adaptations would require careful documentation to preserve auditability.
4. **Adversarial robustness:** Exploring adversarial robustness through red-team testing (intentionally crafting drift-laden claims designed to evade detection) would identify exploitable rule gaps and inform defensive rule updates.
5. **Citation graph integration:** Incorporating citation graph analysis to assess whether flagged claims cite appropriate evidence types (empirical studies versus opinion

pieces) could reduce false EVIDENCE_GAP flags for claims with implicit but traceable support.

6. **Real-time feedback:** Deploying RDG as a real-time feedback layer in LLM prompting workflows—where users receive drift warnings during claim generation rather than post-hoc review—could enable iterative refinement and reduce downstream governance burden.

These directions collectively position RDG not as a terminal solution but as an initial governance layer in an evolving ecosystem of AI-assisted research integrity tools.

---

5. Conclusion

This study introduced Research Drift Guardian (RDG), a transparent rule-based governance framework for evaluating methodological rigor in AI-generated scientific claims. RDG provides three key contributions: (1) a fixed taxonomy of seven research drift types with explicit detection rules, extending evaluation beyond binary hallucination checks; (2) a deterministic Claim Confidence Index (CCI) with risk-positive polarity and policy-justified thresholds, enabling reproducible decisions independent of evaluator judgment; and (3) validation on 30 battery research claims showing a three-tier distribution (23% proceed, 43% conditional review, 33% escalation).

By separating multi-LLM signal generation from deterministic rule application, RDG targets methodological validity—not only factual accuracy—while preserving auditability and reproducibility. Future integration with Track 2 validation metadata (e.g., replication status, statistical power) can support hybrid governance combining claim-level text analysis with evidence-level credibility assessment, advancing the field toward robust, transparent AI-assisted research workflows.

References

[1] Hope, T., Portenoy, J., Vasan, K., Borchardt, J., Horvitz, E., Weld, D. S., ... & Naik, A. (2022). SciMON: Scientific inspiration machines optimized for novelty. *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, 1785–1798. https://doi.org/10.18653/v1/2022.acl-long.125

[2] Agarwal, C., D'souza, D., & Hooker, S. (2024). Large language models as research assistants: Opportunities and challenges. *Nature Machine Intelligence*, 6(4), 372–380. https://doi.org/10.1038/s42256-024-00827-6

[3] Ioannidis, J. P. A. (2005). Why most published research findings are false. *PLoS Medicine*, 2(8), e124. https://doi.org/10.1371/journal.pmed.0020124

[4] Zhang, Y., Li, Y., Cui, L., Cai, D., Liu, L., Fu, T., ... & Sun, M. (2023). Siren's song in the AI ocean: A survey on hallucination in large language models. *arXiv preprint arXiv:2309.01219*. https://arxiv.org/abs/2309.01219

[5] Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., ... & Fung, P. (2023). Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12), 1–38. https://doi.org/10.1145/3571730

[6] Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., ... & Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive NLP tasks. *Advances in Neural Information Processing Systems*, 33, 9459–9474.

[7] Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin*, 86(2), 420–428. https://doi.org/10.1037/0033-2909.86.2.420

[8] Hallgren, K. A. (2012). Computing inter-rater reliability for observational data: An overview and tutorial. *Tutorials in Quantitative Methods for Psychology*, 8(1), 23–34. https://doi.org/10.20982/tqmp.08.1.p023

[9] National Academies of Sciences, Engineering, and Medicine. (2017). *Fostering integrity in research*. National Academies Press. https://doi.org/10.17226/21896

[10] Goodman, S. N., Fanelli, D., & Ioannidis, J. P. A. (2016). What does research reproducibility mean? *Science Translational Medicine*, 8(341), 341ps12. https://doi.org/10.1126/scitranslmed.aaf5027

[11] Popper, K. (1959). *The logic of scientific discovery*. Routledge.

[12] Gebru, T., Morgenstern, J., Vecchione, B., Vaughan, J. W., Wallach, H., Daumé III, H., & Crawford, K. (2021). Datasheets for datasets. *Communications of the ACM*, 64(12), 86–92. https://doi.org/10.1145/3458723

[13] Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., ... & Gebru, T. (2019). Model cards for model reporting. *Proceedings of the Conference on Fairness, Accountability, and Transparency*, 220–229. https://doi.org/10.1145/3287560.3287596

[14] Raji, I. D., Smart, A., White, R. N., Mitchell, M., Gebru, T., Hutchinson, B., ... & Barnes, P. (2020). Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 33–44. https://doi.org/10.1145/3351095.3372873

Appendix A. RDG Drift Detection Rules and Claim Confidence Index (CCI)

A.1 Drift Type Definitions (Seven Fixed Categories)

RDG operates with exactly seven predefined drift categories. No additional, ad-hoc, or emergent categories are permitted.

Drift Type	Definition	Typical Trigger Expressions
EVIDENCE_GAP	Claim asserts an effect without specifying evidence, data, or validation context	"studies suggest", "are believed to", "indicate that"
SCOPE_OVERREACH	Claim generalizes beyond a justified scope	"most cases", "nearly all", "wide range", "universally"
CAUSAL_OVERSTATEMENT	Correlation or association presented as direct causation	"directly leads to", "automatically improves", "results in"
HYPERBOLE	Use of absolute or exaggerated language	"eliminate entirely", "always", "guaranteed", "dramatically"
CONFLATION	Multiple performance dimensions merged without justification	"simultaneously improves X and Y", "both performance and safety"
FUTURE_PROJECTION	Speculative future outcome presented as likely or inevitable	"will replace", "expected to dominate", "inevitably"
AMBIGUITY_SHIELDING	Hedging language used to mask a strong underlying claim	"could potentially", "may significantly", "likely to"

A.2 Claim Confidence Index (CCI) Calculation

A.2.1 Base Formula

$$\text{CCI} = \text{Base} + \sum(\text{Drift Type Weights}) + \text{Strong Assertion Bonus} - \text{Conditionality}$$

where all values are clamped to [0.00, 1.00].

A.2.2 Parameter Weights

- **Base Value:** Base = 0.10
- **Drift Type Weights:**
  - EVIDENCE_GAP: +0.20
  - SCOPE_OVERREACH: +0.20
  - CAUSAL_OVERSTATEMENT: +0.25
  - HYPERBOLE: +0.25
  - CONFLATION: +0.15
  - FUTURE_PROJECTION: +0.15
  - AMBIGUITY_SHIELDING: +0.15
- **Strong Assertion Bonus:** +0.10 (when absolute terms appear)
- **Conditionality Discount:** −0.10 (one condition) to −0.30 (multiple conditions)

A.2.3 Escalation Safeguard Rule

When the following conditions are all satisfied, add +0.17 to CCI:

- HYPERBOLE detected
- SCOPE_OVERREACH detected
- Presence of absolute or universal language
- No explicit conditions or limitations

A.3 Verdict Thresholds

Verdict	CCI Range	Interpretation
PROCEED	CCI < 0.35	Claim is cautious, contextual, and acceptable
HOLD	0.35 \leq CCI < 0.75	Requires validation, clarification, or reformulation
ESCALATE	CCI \geq 0.75	High-risk and should be blocked

A.4 Reproducibility Statement

All RDG decisions are derived exclusively from:

- Fixed drift categories (seven types only)
- Fixed trigger patterns
- Fixed numerical weights
- Fixed verdict thresholds

Given the same claim text, RDG will always produce the same drift tags, CCI value, and verdict.

End of Document

Appendix A: RDG Drift Detection Rules and Claim Confidence Index (CCI)

Final, Policy-Locked Version for Track 1

A.1 Purpose of This Appendix

This appendix documents the rule-based mechanisms used by the Research Drift Guardian (RDG) to evaluate AI-generated research claims. The objective is to guarantee deterministic reproducibility: identical inputs must always yield identical outputs, independent of evaluator, execution order, or model source.

RDG evaluates claims along two orthogonal axes:

- **Research Drift Type(s)** — qualitative pattern detection
- **Claim Confidence Index (CCI)** — quantitative risk scoring for governance decisions

CCI is not a probabilistic estimate of truth. It is a policy-locked stop boundary designed to minimize false proceeds in AI-assisted research workflows.

A.2 RDG Drift Type Definitions (Seven Fixed Categories)

RDG operates with exactly seven predefined drift categories. No additional, ad-hoc, or emergent categories are permitted.

Drift Type	Definition	Typical Trigger Expressions
EVIDENCE_GAP	Claim asserts an effect without specifying evidence, data, or validation context	studies suggest'', are believed to'', indicate that'' \\ \hline SCOPE_OVERREACH & Claim generalizes beyond a justified scope & most cases'', nearly all'', wide range'', universally'' \\ \hline CAUSAL_OVERSTATEMENT & Correlation or association presented as direct causation & directly leads to'', automatically improves'', results in''
HYPERBOLE	Use of absolute or exaggerated language	eliminate entirely'', always'', guaranteed'', dramatically''
CONFLATION	Multiple performance dimensions merged without justification	simultaneously improves X and Y'', both performance and safety''
FUTURE_PROJECTION	Speculative future outcome presented as likely or inevitable	will replace'', expected to dominate'', inevitably'' \\ \hline AMBIGUITY_SHIELDING & Hedging language used to mask a strong underlying claim & could potentially'', may significantly'', likely to''

Table 1: RDG Drift Type Definitions

Constraint: Every detected issue must map to one or more of these seven categories only.

A.3 Claim Confidence Index (CCI) Calculation

CCI quantifies the risk level of a claim on a continuous scale from 0.00 to 1.00, using risk-positive polarity (higher = higher intervention need).

A.3.1 Base Formula

$$\text{CCI} = \text{Base} + \sum(\text{Drift Type Weights}) + \text{Strong Assertion Bonus} - \text{Conditionality}$$

All values are clamped to the range [0.00, 1.00].

A.3.2 Parameter Weights

Base Value

- Base = 0.10

Drift Type Weights

- EVIDENCE_GAP: +0.20
- SCOPE_OVERREACH: +0.20
- CAUSAL_OVERSTATEMENT: +0.25
- HYPERBOLE: +0.25
- CONFLATION: +0.15
- FUTURE_PROJECTION: +0.15
- AMBIGUITY_SHIELDING: +0.15

Strong Assertion Bonus

- Absolute terms (e.g., "eliminate", "always", "guaranteed"): +0.10

Conditionality Discount

- One explicit condition or limitation: −0.10
- Multiple interacting conditions: up to −0.30

A.3.3 Escalation Safeguard Rule

To capture compound high-risk patterns not adequately represented by linear summation, RDG applies a deterministic safeguard.

Trigger Condition (all must be satisfied):

- HYPERBOLE detected
- SCOPE_OVERREACH detected
- Presence of absolute or universal language
- No explicit conditions or limitations

Adjustment: +0.17 added to CCI

This safeguard is intentionally singular and policy-locked to preserve auditability.

A.4 Verdict Thresholds (Stop Boundaries)

RDG applies fixed, pre-declared thresholds:

Verdict	CCI Range	Interpretation
PROCEED	CCI < 0.35	Claim is cautious, contextual, and acceptable
HOLD	$0.35 \leq \text{CCI} < 0.75$	Claim requires validation, clarification, or reformulation
ESCALATE	CCI ≥ 0.75	Claim is high-risk and should be blocked

Table 2: RDG Verdict Thresholds

Thresholds are global constants and do not vary across claims.

A.5 Worked Examples (Reproducibility Walkthrough)

Example 1 — ESCALATE Case (C002)

Claim Text

"Solid-state batteries are expected to eliminate dendrite formation entirely, making them universally safer than liquid systems."

Detected Drift Types

- HYPERBOLE ("eliminate entirely") → +0.25
- SCOPE_OVERREACH ("universally safer") → +0.20

Additional Factors

- Strong assertion → +0.10
- No explicit conditions → 0.00 discount

Base Calculation

$$0.10 + 0.25 + 0.20 + 0.10 = 0.65$$

Safeguard Applied

Compound absolute + scope pattern → +0.17

Final CCI: 0.82

Verdict: ESCALATE

Example 2 — PROCEED Case (C025)

Claim Text

"Battery performance is influenced by multiple interacting factors, including materials, cell design, and operating conditions, making generalized claims difficult to validate."

Detected Drift Types

None

Conditionality Discount

Multiple explicit constraints → −0.30

Calculation

$$0.10 - 0.30 = -0.20 \rightarrow \text{clamped to } 0.10$$

Final CCI: 0.08

Verdict: PROCEED

Example 3 — HOLD Case (C011)

Claim Text

"Advanced electrolyte additives are believed to suppress side reactions effectively in high-voltage cathodes."

Detected Drift Types

- EVIDENCE_GAP ("are believed to") → +0.20
- AMBIGUITY_SHIELDING ("effectively" without metrics) → +0.15

Additional Factors

- No strong assertion → +0.00
- No explicit conditions → 0.00 discount

Calculation

$$0.10 + 0.20 + 0.15 = 0.45$$

Final CCI: 0.45

Verdict: HOLD

Interpretation: The claim is plausible but insufficiently grounded and requires further validation or reformulation before proceeding.

A.6 Reproducibility Statement

All RDG decisions are derived exclusively from:

- Fixed drift categories
- Fixed trigger patterns
- Fixed numerical weights
- Fixed verdict thresholds

Given the same claim text, RDG will always produce the same drift tags, CCI value, and verdict.

End of Appendix A