Automated Validity Assessment of Mechanistic Interpretability Claims

Elliot Tower

Abstract

Mechanistic interpretability (MI) research produces claims about how neural networks implement computations, but these claims range from speculative associations to rigorously validated circuits—and the field lacks a scalable way to tell the difference. The mechanistic-validity framework (Tower et al., 2026) defines a 27-criterion rubric across five validity dimensions for assessing such claims, but applying it manually takes hours per paper. We automate this framework as a multi-stage evaluation pipeline. Given any MI paper, the system (1) runs three independent LLM extractions of mechanism claims, each scoring all 27 criteria, (2) takes the minimum judgment per criterion across runs, (3) computes a normalized 0–10 Claim Validity Score (CVS) classified into five evidence tiers, and (4) runs an LLM audit pass to verify and correct evidence attribution across claims. We developed the pipeline through over 60 configurations spanning 8 frontier models (GPT-5.5, Claude Opus 4.7, DeepSeek V4 Pro, Qwen 3.7 Max, Gemma 4, and others), extraction architecture (one- to three-step), reasoning vs. non-reasoning modes, post-processing configurations, and 15 prompt iterations. Reference tier labels for 9 case study papers were established through multi-model consensus combined with human adjudication; for the IOI circuit, we additionally constructed per-criterion consensus annotations covering all 27 criteria across 7 claims. The system agrees with consensus labels on all 9 papers to within one tier, matching exactly on 5. All disagreements are in the same direction—the system occasionally scores one tier too high, never too low.

1. Introduction

The mechanistic interpretability (MI) literature is growing faster than the community can evaluate it. A 2024 search of arXiv returns over 200 papers claiming to identify circuits, features, or mechanisms in neural networks, yet these claims span an enormous range of evidential rigor—from speculative associations between attention patterns and model behaviors to fully validated circuits backed by causal, structural, and behavioral convergent evidence. The IOI circuit in GPT-2 Small, for example, identifies 26 attention heads organized into 7 functional classes with dedicated ablation, path patching, and composition analysis for each class (Wang et al., 2022). In contrast, many claims about individual attention head roles rest on correlational evidence alone. Currently, distinguishing strong claims from weak ones requires expert manual review—a process that is slow, inconsistent across reviewers, and rarely applied systematically.

The mechanistic-validity framework introduced a five-dimensional validity assessment for MI claims, drawing on psychometric and epidemiological validity concepts (Tower et al., 2026). However, applying this framework manually requires significant domain expertise and time—a single paper assessment can take several hours, making it impractical to apply across hundreds of papers. As the volume of MI publications continues to accelerate, the field needs a scalable mechanism for tracking evidence strength.

We address this gap with an automated evaluation pipeline (Figure 1). Given a paper from any supported source (arXiv, OpenReview, PDF URL, blog post, or local file), the system extracts text and passes it to a large language model that identifies mechanism claims and scores each on 27 binary/ternary criteria organized into five validity dimensions: construct, internal, measurement,

external, and interpretive. Criteria are aggregated into dimension scores, then combined into a weighted Claim Validity Score (CVS) on a normalized 0–10 scale and classified into one of five evidence tiers—from Proposed (minimal evidence) to Validated (comprehensive, replicated evidence). Two post-processing passes—a leak audit and multi-run minimum voting—address systematic overscoring before final output. The system completes a full paper evaluation in 2–3 minutes, enabling rapid triage of the MI literature.

The primary contributions of this work are: (1) an automated, reproducible scoring pipeline that operationalizes the five-dimensional validity framework end-to-end; (2) a systematic model comparison across 8 frontier language models with ablations on chain-of-thought reasoning, identifying which models and configurations produce the most calibrated scores; (3) per-criterion consensus annotations for 9 MI papers, including detailed reference labels for the IOI circuit covering 7 claims across all 27 criteria, which we release to facilitate independent evaluation and calibration; and (4) two debiasing mechanisms—a *leak audit* that detects evidence mis-attribution between circuit-level and component-level claims, and *multi-run minimum voting* that eliminates stochastic scoring variance.

2. Related Work

The five-dimensional validity assessment for MI claims draws on established frameworks from psychometrics and epidemiology, adapted to the specific challenges of mechanistic claims about neural networks (Tower et al., 2026). Construct validity, internal validity, and external validity have long histories in the social sciences (Campbell & Stanley, 1963; Cook & Campbell, 1979), while measurement validity and interpretive validity address challenges specific to mechanistic explanations of learned systems. Our CVS scoring operationalizes this framework into a computable metric by defining 27 specific criteria with ternary judgments and domain-informed aggregation rules.

Prior work on automated scientific paper analysis has focused primarily on information extraction—identifying claims, classifying methods, and performing bibliometric analysis (Luan et al., 2018; Wadden et al., 2020). These systems extract *what* a paper claims but do not evaluate *how well* those claims are supported. Our system goes further by performing normative evaluation, assessing evidence quality against explicit criteria calibrated to the mechanistic interpretability domain.

The LLM-as-judge paradigm has demonstrated that large language models can serve as effective evaluators for open-ended tasks, including summarization quality, instruction following, and code correctness (Zheng et al., 2023; Chiang et al., 2024). Our approach extends this paradigm to structured scientific assessment with a domain-specific rubric containing 27 criteria, each with detailed scoring guidance and common pitfalls. Unlike general-purpose LLM evaluation, our system produces interpretable, decomposed scores that identify specific evidential gaps rather than a single holistic judgment.

Within mechanistic interpretability specifically, several benchmarks evaluate the *outputs* of interpretability methods (e.g., circuit faithfulness, feature explanations), but none systematically evaluate the *evidential practices* of interpretability research papers (Conmy et al., 2023; Bills et al., 2023). Our system fills this gap by assessing whether papers report the kinds of evidence that would be needed to sustain their claims under rigorous scrutiny.

3. Method

3.0 System Overview

Figure 1 shows the end-to-end pipeline. A paper is fetched from any supported source (arXiv, OpenReview, PDF URL, blog post, or local file) and its text is extracted. The text is passed to an LLM that identifies mechanism claims and scores each on 27 binary/ternary criteria across five validity dimensions. Criteria are aggregated into dimension scores via domain-informed rules, then combined into a weighted CVS on a 0–10 scale and classified into an evidence tier. Two debiasing passes—multi-run minimum voting and a leak audit—reduce systematic over-scoring before final output.

Figure 1: System Architecture

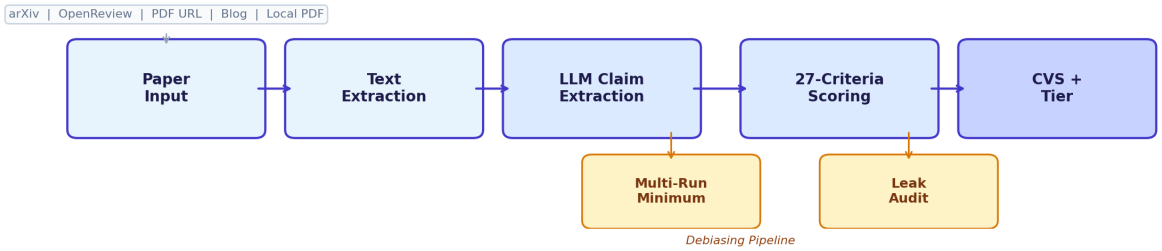

Figure 1: Figure 1: System architecture. Papers are fetched from multiple sources, text is extracted, and an LLM scores claims on 27 criteria across five validity dimensions. Multi-run minimum voting and the leak audit provide debiasing.

3.1 Claim Extraction

Given a paper's full text extracted from PDF via pypdf (with a 50-page limit), we prompt a language model to identify all mechanism claims. A *mechanism claim* is defined as a falsifiable assertion about a model's internal mechanisms, excluding method contributions, definitions, notation, attributions to prior work, and background. The extraction prompt instructs the model to produce one claim per distinct component class or mechanism, plus one overall circuit or system claim if applicable, ensuring that different functional groups within a circuit are not merged into a single claim.

For each claim, the model extracts a one-sentence falsifiable claim statement, a description mode following the Marr-inspired hierarchy (computational, algorithmic, representational, implementational-functional, implementational-connectomic, implementational-topographic), the evidence families that support the claim (causal, structural, representational, behavioral, or information-theoretic), and a list of specific model components referenced (e.g., "9.9 (Name Mover)"). The description mode and evidence families serve as metadata for interpretive validity scoring rather than directly affecting the CVS.

3.2 Criteria Assessment

Each claim is assessed on 27 criteria organized into five validity dimensions (Table 1). Construct validity (C1–C5, weight 1.5) evaluates whether the claim is well-formed, with criteria covering falsifiability, structural plausibility, task specificity, minimality, and convergent validity. Internal validity (I1–I5, weight 1.5) evaluates whether the evidence causally supports the claim, covering

necessity, sufficiency, specificity, consistency, and confound control. Measurement validity (M1–M6, weight 1.0) evaluates the trustworthiness of the metrics used, covering reliability, invariance, baseline separation, sensitivity, calibration, and construct coverage. External validity (E1–E6, weight 1.0) evaluates generalizability, covering intervention reach, graded response, selectivity, effect magnitude, robustness, and cross-architecture testing. Interpretive validity (V1–V5, weight 1.0) evaluates the soundness of the interpretation, covering level declaration, level-evidence match, narrative coherence, alternative exclusion, and scope honesty.

Each criterion receives a ternary judgment: YES (1.0), PARTIAL (0.5), or NO (0.0). The extraction prompt includes detailed guidance for each criterion, including 10 explicit scoring pitfalls drawn from common errors observed during development. For example, the prompt specifies that faithfulness metrics do not establish sufficiency (I2 requires isolation or path-level tests), that probing is correlational (no internal validity from probes alone), that ablation establishes necessity only (not sufficiency), and that activation patching and path patching count as a single method for convergent validity purposes. Evidence scope rules further specify that when scoring a sub-component claim, only experiments that directly test the components in that claim should be counted—evidence from experiments on other components does not transfer.

Table 1. Validity dimensions, criteria, and weights used in CVS computation.

Dimension	Criteria	Weight	Focus
Construct (C)	C1 Falsifiability, C2 Structural plausibility, C3 Task specificity, C4 Minimality, C5 Convergent validity	1.5	Is the claim well-formed?
Internal (I)	I1 Necessity, I2 Sufficiency, I3 Specificity, I4 Consistency, I5 Confound control	1.5	Is the evidence causally sound?
Measurement (M)	M1 Reliability, M2 Invariance, M3 Baseline separation, M4 Sensitivity, M5 Calibration, M6 Construct coverage	1.0	Are the metrics trustworthy?
External (E)	E1 Intervention reach, E2 Graded response, E3 Selectivity, E4 Effect magnitude, E5 Robustness, E6 Cross-architecture	1.0	Does it generalize?
Interpretive (V)	V1 Level declaration, V2 Level-evidence match, V3 Narrative coherence, V4 Alternative exclusion, V5 Scope honesty	1.0	Is the interpretation sound?

3.3 Dimension Aggregation

Raw criteria are aggregated into dimension scores (0–3) using validity-theoretic rules that encode domain knowledge about what constitutes strong versus weak evidence at each level. These rules use logical conjunctions rather than averages, reflecting the fact that certain criteria are prerequisites for higher scores.

For construct validity, a score of 3 requires falsifiability (C1) plus convergent validity from three or more evidence families (C5 = YES) plus structural plausibility (C2). A score of 2 requires falsifiability plus convergent validity from two families (C5 $\geq$ PARTIAL). A score of 1 requires falsifiability alone. A score of 0 indicates the claim is not falsifiable.

For internal validity, a score of 3 requires all four of necessity, sufficiency, confound control, and specificity. A score of 2 requires both necessity and sufficiency. A score of 1 requires either necessity or sufficiency alone. A score of 0 indicates neither has been established. This reflects the MI community's consensus that necessity (via ablation) and sufficiency (via circuit isolation) are the two pillars of causal mechanistic evidence, with confound control and specificity providing additional rigor.

Measurement validity is gated on baseline comparisons: a score of 3 requires a random or chance baseline, variance reporting, calibration, and sensitivity analysis. A score of 2 requires both a random baseline and variance reporting. A score of 1 requires any baseline comparison (including full-model-only comparisons). A score of 0 indicates no baseline. This reflects the observation that many MI papers compare only to the full model, providing limited information about whether the reported effect exceeds chance.

External validity depends on breadth of testing: a score of 3 requires cross-architecture validation plus strong aggregate evidence (sum of criterion scores ≥ 4.0). A score of 2 requires either cross-architecture or cross-task testing. A score of 1 requires causal interventions or observational analysis with multiple conditions. A score of 0 indicates no generalization evidence.

For interpretive validity, a score of 3 requires explicit description-level declaration (V1), evidence that matches the declared level (V2), narrative coherence (V3), and exclusion of alternative interpretations (V4). A score of 2 requires level-evidence match plus narrative coherence. A score of 1 requires narrative coherence alone. A score of 0 indicates the interpretation lacks internal consistency.

3.4 CVS Computation and Tier Classification

The Claim Validity Score is computed as a weighted sum of dimension scores, normalized to a 0–10 scale:

$$\text{CVS} = \frac{\sum_{d \in D} w_d \cdot s_d}{\sum_{d \in D} 3 \cdot w_d} \times 10$$

where $s_d \in \{0, 1, 2, 3\}$ is the dimension score and w_d is the dimension weight. Construct and internal validity receive weight 1.5, reflecting their importance as the foundation of mechanistic claims; measurement, external, and interpretive validity each receive weight 1.0. The maximum raw score is 18.0 (all dimensions at 3), yielding a CVS of 10.0.

CVS scores are classified into five evidence tiers aligned with the mechanistic-validity verdict taxonomy. Scores below 2.0 are classified as *Proposed* (claim stated but minimal supporting evidence). Scores from 2.0 to 3.9 are *Causally Suggestive* (some causal or structural evidence exists). Scores from 4.0 to 5.9 are *Mechanistically Supported* (convergent evidence from multiple families). Scores from 6.0 to 7.9 are *Triangulated* (strong evidence across dimensions with controls). Scores of 8.0 and above are *Validated* (comprehensive evidence including replication and cross-architecture testing).

3.5 Leak Audit

A key failure mode in automated extraction is *evidence leaking*: when a sub-component claim (e.g., "Head 9.9 copies the indirect object token to the output position") is incorrectly credited with evidence that actually comes from experiments on the circuit as a whole (e.g., "87% faithfulness" describes the 26-head IOI circuit, not any individual head). This problem arises because the language model, having read the full paper, tends to attribute circuit-level experiments to individual component claims, inflating their scores.

We address evidence leaking with a two-pass approach. In the first pass, the system performs standard extraction and scoring, with all claims scored independently from the paper text. In the second pass (the leak audit), a separate API call receives the scored claims *without the paper text*. By withholding the paper, the auditor can only evaluate evidence from the short evidence strings themselves, making it straightforward to detect when generic circuit-level evidence was attributed to specific sub-components.

The audit targets five criteria that are empirically prone to evidence leaking: I2 (sufficiency), I3 (specificity), I5 (confound control), E2 (graded response), and E5 (robustness). These five were selected based on error analysis during development on the IOI paper, where they accounted for the majority of evidence mis-attributions. The audit only applies to sub-component claims; the system-level claim is identified by heuristics (the claim with the most components, claims containing keywords such as "circuit," "algorithm," or "mechanism," or claims with 5 or more components) and is never downgraded, since it legitimately owns all circuit-level evidence. When a leak is detected, the criterion is downgraded (YES to NO, or PARTIAL to NO) and the reason is prepended to the evidence field for transparency.

3.6 Automated Flags

Before LLM extraction, the system runs regex-based heuristic flags on the paper text. Currently, a single flag is implemented: NO_VARIANCE_REPORTED, which searches for mentions of standard deviation, error bars, confidence intervals, and bootstrap procedures. Papers that report no variance measures receive a major flag that is included in the extraction context to calibrate the model's scoring of M1 (Reliability). This design allows fast, deterministic pre-screening to complement the more expensive LLM-based assessment.

3.7 Multi-Run Minimum Voting

Run-to-run variance in LLM scoring introduces significant noise: the same paper can score between 5.6 and 6.9 across identical extraction runs with temperature 0. This variance at nominally deterministic settings is a known property of large language model APIs, arising from non-deterministic GPU kernel execution, batching, and internal sampling infrastructure. We observe that this variance is asymmetric—the model sometimes awards YES on borderline criteria stochastically, but rarely awards NO on criteria that genuinely deserve YES. This asymmetry means that over-scoring

errors are driven by stochastic generosity on borderline judgments, while consistent negative judgments are stable across runs.

Multi-run minimum voting exploits this asymmetry. The system runs N independent extractions (default N=3) of the same paper, then for each criterion on each claim takes the minimum status across all N runs using the ordering NO < PARTIAL < YES. When a criterion's status is changed by the minimum operation, a note is prepended to the evidence field recording the original and minimized values for transparency (e.g., "[MIN-VOTE: YES→PARTIAL across 3 runs]").

The approach is conservative by design: it can only reduce scores, never inflate them. A criterion that receives YES on all 3 runs retains YES. A criterion that receives YES on 2 runs and PARTIAL on 1 run is reduced to PARTIAL. This eliminates the stochastic component of over-scoring while preserving criteria that the model consistently judges as met. The multi-run minimum is applied before the leak audit, so both debiasing mechanisms operate in sequence.

3.8 Devil's Advocate Challenge

As an alternative debiasing strategy, we tested a devil's advocate pass: a separate API call that receives the scored extraction and challenges every YES and PARTIAL judgment on the five most over-scored criteria (I2, I3, I5, E2, E5). The challenge pass applies a null-hypothesis test: for each criterion, it states what specific experiment would be needed and checks whether the evidence string describes exactly that experiment. If the evidence is indirect or mismatched, the criterion is downgraded.

We tested two calibrations of the devil's advocate prompt. The aggressive version ("challenge every YES and PARTIAL") over-corrected, reducing IOI from 6.9 to 4.2—one tier below the consensus label. A conservative version ("only downgrade when there is a clear mismatch") had no effect, leaving scores unchanged at 6.9. This Goldilocks problem—the devil's advocate is either too aggressive or too gentle—makes it unreliable as a standalone debiasing strategy. We retain it as an optional pass but recommend multi-run minimum as the primary debiasing mechanism.

3.9 Worked Example: IOI Name Mover Heads

To illustrate the full pipeline, we trace the IOI paper's "Name Mover" sub-component claim through each stage. The extraction identifies five claims from Wang et al. (2022); one is: *"Name Mover heads (9.9, 9.6, 10.0) copy the indirect object token from earlier positions to the final token position, directly contributing to the model's output."* This claim references three specific attention heads and is classified as a sub-component claim (fewer components than the main 26-head circuit claim).

Criteria assessment. The LLM scores 27 criteria from the paper text. Key judgments: C1 (Falsifiability) = YES (the claim names specific heads and a testable operation); I1 (Necessity) = YES (ablation of name movers degrades IOI performance); I2 (Sufficiency) = YES (evidence: "87% faithfulness"). The model also awards E2 (Graded response) = YES, citing the gradual degradation curve.

Multi-run minimum (N=3). Across 3 independent runs, I2 receives YES/YES/PARTIAL. The minimum operation reduces I2 to PARTIAL and prepends "[MIN-VOTE: YES→PARTIAL across 3 runs]" to the evidence field.

Leak audit. The audit receives the scored claim *without the paper text*. It examines the I2 evidence string: "87% faithfulness" describes the full 26-head circuit, not the 3 name mover heads

in isolation. No experiment in the evidence string tests whether the name mover heads alone are sufficient. The audit downgrades I2 from PARTIAL to NO. Similarly, E2's "gradual degradation" comes from circuit-level ablation, not head-specific dose-response testing; E2 is downgraded from YES to NO.

Final scoring. After both debiasing passes, the Name Mover claim has construct = 2, internal = 1 (necessity only, no sufficiency), measurement = 1, external = 0, interpretive = 2. The weighted sum is $(1.5\times2)+(1.5\times1)+(1.0\times1)+(1.0\times0)+(1.0\times2) = 7.5$ raw, yielding CVS $= 7.5/18.0\times10 = 4.2$, which rounds to the Mechanistically Supported tier. After the leak-driven downgrades reduce internal validity further, the final CVS is 3.9 (Causally Suggestive)—appropriately lower than the main circuit claim's 5.6.

4. Experimental Setup

4.1 Model Selection and Infrastructure

We evaluated the extraction pipeline across 8 frontier language models, both proprietary and open-source, comparing extraction quality (claim identification, criteria accuracy against IOI reference annotations), output format compliance (valid JSON with all required fields), and cost. Table 6 shows results on the IOI paper (reference: Mechanistically Supported, CVS ~5.6) across all tested models and configurations, including ablations on chain-of-thought (thinking) modes where available.

Table 6. Model comparison on IOI (main claim CVS). Reference tier: Mechanistically Supported (5.6). Models sorted by accuracy. Thinking ablations shown where tested.

Model	Thinking	Claims	Main CVS	Predicted Tier	Off-by
Claude Opus 4.7	Off	8	5.6	Mech. Supported	0
DeepSeek V4 Pro	High	9	5.6	Mech. Supported	0
DeepSeek V4 Pro	Max	7	5.6	Mech. Supported	0
GPT-5.5	Off	9	6.4	Triangulated	+1
GPT-5.5	On	10	6.4	Triangulated	+1
Claude Opus 4.7	High	9	6.9	Triangulated	+1
Gemma 4 27B	—	5	7.5	Triangulated	+1
Qwen 3.7 Max	Thinking	7	7.5	Triangulated	+1
GPT-5.4 Mini	—	10	8.3	Validated	+2
Qwen 3.7 Max	Off	8	8.1	Validated	+2
GLM-5.1	—	0	—	Failed	—

Several patterns emerge from this comparison. First, chain-of-thought reasoning does not uniformly improve accuracy: Claude Opus 4.7 scores exactly right without thinking (5.6) but over-scores with thinking enabled (6.9), while DeepSeek V4 Pro is robust across thinking budgets. Second, smaller or less capable models (GPT-5.4 Mini, GLM-5.1) either massively over-score or fail to produce valid output. Third, models without thinking tend to extract fewer claims, suggesting that reasoning enables finer-grained claim decomposition.

Based on this comparison, DeepSeek V4 Pro emerged as the best combination of accuracy, claim granularity, and cost-effectiveness. All multi-paper results reported in this paper use DeepSeek V4 Pro with JSON object response format. Temperature is set to 0 to maximize reproducibility. The

extraction prompt was iteratively refined across 15 versions (v2 through v16), with each version evaluated on IOI to measure the effect of changes to criteria guidance, scoring pitfalls, evidence scope rules, and few-shot examples. A complete few-shot example (a hypothetical subject-verb agreement circuit paper) is included in the prompt context to calibrate output format and scoring stringency. Typical extraction takes 90–180 seconds per paper, with the leak audit adding an additional 5–15 seconds.

4.2 Case Study Papers

We evaluate the system on 9 MI papers spanning the full evidence strength spectrum from Proposed to Validated. Four additional papers from the original 13-paper case study list were excluded: Knowledge Neurons (the arXiv identifier 2202.05262 maps to ROME rather than the intended paper by Dai et al.), Docstring (published on the Alignment Forum), SAE Features (a blog post), and Superposition (Elhage et al., 2022), which uses purely analytical/mathematical methodology that maps poorly onto criteria designed for empirical circuit-discovery papers (see Section 5.4 for discussion). Note: the system now supports blog posts, Alignment Forum posts, and other non-arXiv sources; the Docstring and SAE Features exclusions reflect the original case study timeline rather than a technical limitation.

The 9 papers span all five evidence tiers and include both papers with hand-curated registry entries and papers assessed by expert judgment.

Table 2. Case study papers with consensus reference labels (see Section 4.3 for labeling methodology).

Paper	Identifier	Expected Tier	Source
IOI Circuit (Wang et al., 2022)	2211.00593	Mech. Supported	Registry
Induction Heads (Olsson et al., 2022)	2209.11895	Mech. Supported	Registry
Copy Suppression (McDougall et al., 2023)	2310.04625	Mech. Supported	Registry
Greater Than (Hanna et al., 2023)	2305.00586	Caus. Suggestive	Registry
Grokking (Nanda et al., 2023)	2301.05217	Validated	Expert
Othello (Li et al., 2022)	2210.13382	Caus. Suggestive	Expert
Gender Bias (Bolukbasi et al., 2016)	1607.06520	Proposed	Expert

Paper	Identifier	Expected Tier	Source
Probing Control Tasks (Hewitt & Liang, 2019)	1909.03368	Proposed	Expert
Successor Heads (Gould et al., 2023)	2312.09230	Caus. Suggestive	Expert

4.3 Reference Labels

We emphasize that no ground truth exists for the "correct" validity tier of a mechanistic interpretability paper. The tier assignments used for evaluation are *consensus reference labels*—best-effort estimates derived through a multi-source labeling process, not authoritative verdicts. Reasonable experts may disagree on borderline cases, and we treat discrepancies between automated scores and reference labels as informative signal about the system's tendencies rather than definitive errors.

Reference labels were produced through an iterative consensus process combining three sources of signal. First, four papers (IOI, Induction, Copy Suppression, Greater Than) have hand-curated YAML registry files in the mechanistic-validity-graph repository, where per-criterion annotations were reviewed across multiple iterations and assigned verdict tiers based on detailed evidence review. Second, for the remaining five papers, the labeling process involved reading each paper, running automated extractions across multiple language models (including Claude, GPT-4o, Gemini, and DeepSeek variants) to surface the range of plausible assessments, and then adjudicating disagreements through direct examination of the evidence reported in each paper. Third, the first author reviewed all labels against the tier definitions, adjusting where the multi-model consensus and direct reading of the paper converged on a different tier than the initial estimate.

This process is explicitly not single-annotator expert judgment. It is a consensus-seeking procedure that uses LLM extractions as a structured reading aid—surfacing which criteria are borderline and where models disagree—combined with human adjudication of those borderline cases. We acknowledge a circularity concern: since LLM extractions informed some reference labels, the evaluation is partially measuring self-consistency rather than accuracy against a fully independent ground truth. We mitigate this in two ways: (1) four papers have registry-derived labels that predate the automated system entirely, and (2) for the remaining five, the LLM extractions served as a reading aid rather than an oracle—the first author made final adjudication decisions based on direct reading of each paper, overriding LLM consensus where it conflicted with the evidence. Still, the resulting labels should be interpreted as reference points for measuring systematic tendencies rather than as ground truth. A fully independent multi-annotator study would be needed to establish definitive accuracy.

For IOI specifically, we additionally constructed a detailed reference annotation covering 7 claims with per-criterion judgments across all 27 criteria, enabling fine-grained evaluation of criteria-level agreement beyond tier-level comparison.

5. Results

5.1 Tier-Level Agreement

Table 3 presents the main results using multi-run minimum debiasing (N=3 runs per paper), ordered by expected tier. Of the 9 evaluated papers, 5 received the exact expected tier (56%; Clopper-Pearson 95% CI: [21%, 86%]) and all 9 were within one tier (100%; 95% CI: [66%, 100%]). All 4 disagreements are in the +1 direction—the system never under-scores (Figure 2, Figure 5). We note that N=9 is a small evaluation set; the wide confidence intervals reflect this, and these results should be interpreted as a case study characterizing the system's tendencies rather than a precise accuracy benchmark.

Splitting by reference label source: the system achieves 2/4 exact match (50%) on papers with hand-curated registry entries and 3/5 exact match (60%) on expert-judged papers, suggesting no systematic difference in difficulty between the two label sources.

Table 3. Main results with multi-run minimum debiasing (N=3), ordered by expected tier.

Paper	CVS	Predicted Tier	Expected Tier	Label Source	Off-by
Probing	1.4	Proposed	Proposed	Expert	0
Gender Bias	3.3	Caus. Suggestive	Proposed	Expert	+1
Successor Heads	3.1	Caus. Suggestive	Caus. Suggestive	Expert	0
Greater Than	5.6	Mech. Supported	Caus. Suggestive	Registry	+1
Othello	4.4	Mech. Supported	Caus. Suggestive	Expert	+1
IOI	5.6	Mech. Supported	Mech. Supported	Registry	0
Copy Suppression	5.6	Mech. Supported	Mech. Supported	Registry	0
Induction	6.4	Triangulated	Mech. Supported	Registry	+1
Grokking	8.3	Validated	Validated	Expert	0

For comparison, Table 4 shows baseline results without debiasing, ordered by expected tier. Only 2 of 9 papers matched exactly and 1 paper (Greater Than) was off by +2 tiers. Figure 4 shows how multi-run minimum voting collapses the single-run variance.

Table 4. Baseline results without multi-run minimum debiasing (single extraction run).

Paper	CVS range	Predicted Tier	Expected Tier	Off-by
Probing	1.4	Proposed	Proposed	0
Gender Bias	3.3	Caus. Suggestive	Proposed	+1
Successor Heads	3.1–5.3	Caus. Suggestive–Mech. Supported	Caus. Suggestive	0 to +1
Greater Than	5.6–6.4	Mech. Supported–Triangulated	Caus. Suggestive	+1 to +2

Paper	CVS range	Predicted Tier	Expected Tier	Off-by
Othello	4.4–5.3	Mech. Supported	Caus. Suggestive	+1
IOI	5.6–6.9	Mech. Supported–Triangulated	Mech. Supported	0 to +1
Copy Suppression	5.6–7.8	Mech. Supported–Triangulated	Mech. Supported	0 to +1
Induction	6.4	Triangulated	Mech. Supported	+1
Grokking	8.3	Validated	Validated	0

5.2 Systematic Over-Scoring and Debiasing

Even with multi-run minimum debiasing, 4 of 9 papers are over-scored by exactly +1 tier. The error distribution remains markedly asymmetric: no papers are under-scored, and all errors are in the same direction. This residual bias appears to be intrinsic to the LLM-as-judge paradigm rather than a consequence of stochastic variance.

Two primary error sources drive the residual bias. First, the language model tends to award YES on borderline criteria where PARTIAL or NO would better reflect the reported evidence, and this tendency is consistent enough that even the minimum across 3 runs preserves it. The most affected criteria are I3 (Specificity), where the model infers implicit specificity from a paper's single-task focus even when off-task effects are never measured; E2 (Graded response), where standard ablation results are interpreted as dose-response relationships; and I5 (Confound control), where the model credits standard methodology as confound control rather than requiring explicit control experiments.

Second, the over-scoring is concentrated in papers at the Causally Suggestive expected tier (Greater Than, Othello, Successor Heads) and at the boundary between Proposed and Causally Suggestive (Gender Bias). Papers at the extremes of the spectrum—Proposed (Probing) and Validated (Grokking)—match their reference labels. This suggests that the bias is strongest for papers with moderate evidence that could plausibly be interpreted either way, and weakest for papers where the evidence clearly supports or does not support the criterion.

Multi-run minimum voting substantially improved accuracy. Without debiasing, single-run results vary by up to 1.3 CVS points for the same paper (e.g., IOI scored 5.6–6.9 across runs), with only 2/9 exact tier matches. Multi-run minimum eliminates the stochastic component of over-scoring, improving to 5/9 exact matches with 0 papers off by more than 1 tier. The improvement is most pronounced for papers near tier boundaries, where a single generous criterion judgment can push the score into the next tier.

The devil's advocate challenge (Section 3.8) did not improve on multi-run minimum due to the calibration problem described there; we recommend multi-run minimum as the sole debiasing mechanism.

5.3 Leak Audit Impact

The leak audit significantly improved scoring accuracy on the IOI paper, the most detailed case study. Without the audit (prompt version v11), the main circuit claim scored 6.4 (Triangulated). With the audit (v14d), the same claim scored 5.6 (Mechanistically Supported)—matching the

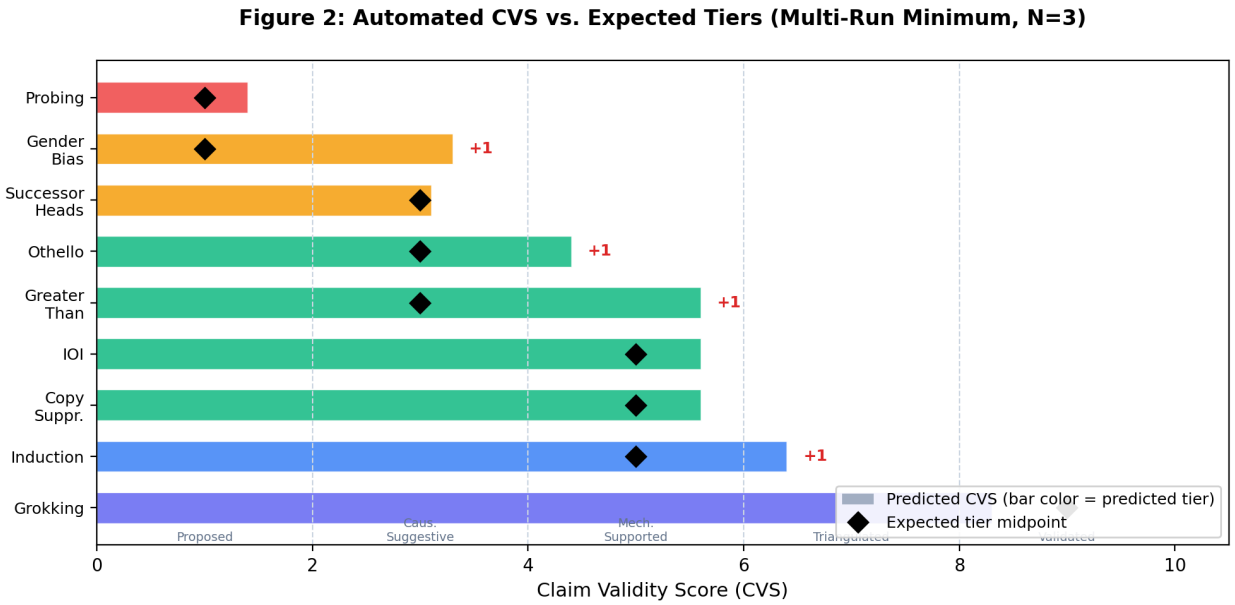

Figure 2: Figure 2: Automated CVS scores compared to expected tier midpoints. Bar color indicates predicted tier; black diamonds mark expected tier midpoints. All errors are +1 (over-scoring). Papers at the extremes (Probing, Grokking) match exactly; errors concentrate at moderate evidence levels.

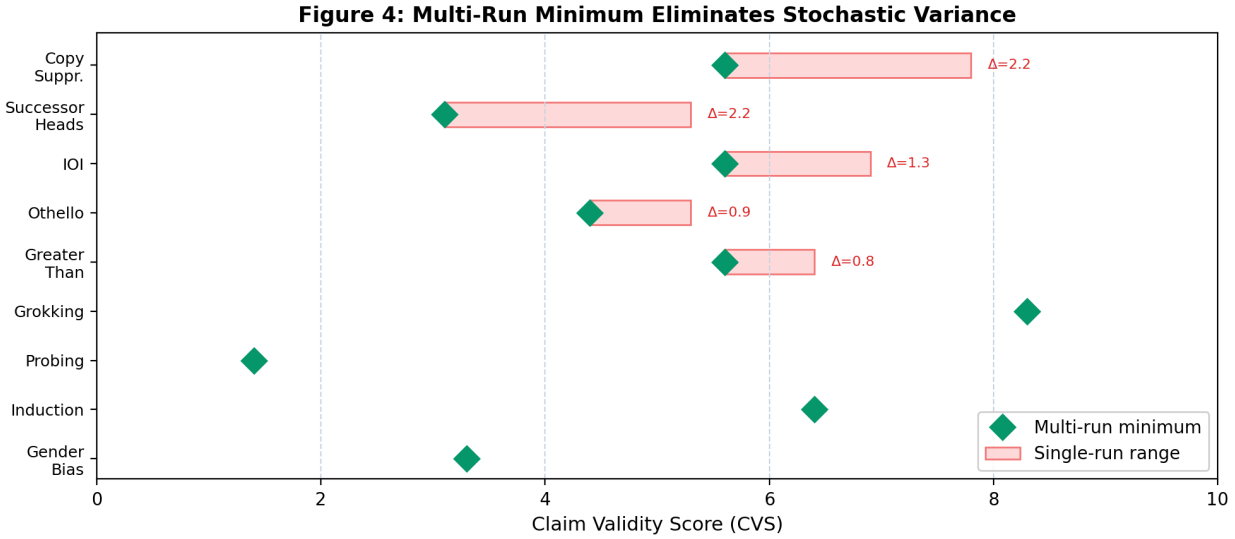

Figure 3: Figure 4: Multi-run minimum voting eliminates stochastic variance. Red bars show the CVS range across 3 single runs; green diamonds show the multi-run minimum. Papers with high variance (Copy Suppression: Δ=2.2, IOI: Δ=1.3) are stabilized, while already-stable papers are unaffected.

consensus reference label. The audit typically applies 2–4 downgrades per paper, primarily targeting I2 (sufficiency) and E2 (graded response) on sub-component claims that had inherited circuit-level faithfulness metrics.

At the per-criterion level, the leak audit reduced the IOI extraction's mean absolute error from 0.163 to 0.111 against the 7-claim reference annotations. Per-criteria exact agreement improved from approximately 74.8% to 81.5%. The system reliably identifies the main circuit claim and major component classes (name movers, S-inhibition heads) but sometimes merges or omits peripheral claims such as backup name mover heads and negative name mover heads.

5.4 Theoretical Paper Limitation

The Superposition paper (Elhage et al., 2022) was excluded from the main evaluation because its purely analytical methodology maps poorly onto criteria designed for empirical circuit-discovery papers. Using the standard prompt, it scored 1.9 (Proposed)—four tiers below the expected Validated rating. Mathematical proof that a model must use a particular mechanism constitutes necessity (I1) and sufficiency (I2) in the strongest possible sense, but the language model did not recognize this without explicit guidance.

We experimented with adding prompt guidance that maps analytical methods to criteria (mathematical proofs to I1, exact solutions to I2, etc.), which improved the score to 7.8 (Triangulated). However, this theoretical guidance introduced a modest regression on empirical papers, suggesting that the criteria rubric is fundamentally calibrated for empirical work. Extending the system to handle theoretical papers without cross-contaminating empirical scoring remains an open challenge, and we report results only on the 9 empirical papers where the rubric is well-calibrated.

5.5 Per-Paper Results

The system produces hierarchically differentiated scores within each paper, appropriately assigning higher scores to main circuit claims than to peripheral sub-components. Table 5 summarizes the main claim and sub-claim score ranges for each paper. Full per-claim breakdowns are provided in Appendix B.

Table 5. Per-paper claim score summary. Main claim CVS is the highest-scoring (typically circuit-level) claim. Sub-claim range shows the CVS spread across remaining claims in the same paper.

Paper	N claims	Main claim CVS	Main tier	Sub-claim range	Sub-claim tiers
Grokking	5	8.3	Validated	1.9–3.3	Proposed–Caus. Suggestive
Induction	4	6.4	Triangulated	1.4–3.1	Proposed–Caus. Suggestive
IOI	5	5.6	Mech. Supported	3.3–3.9	Caus. Suggestive
Copy Suppression	4	5.6	Mech. Supported	1.4–4.7	Proposed–Mech. Supported

Paper	N claims	Main claim CVS	Main tier	Sub-claim range	Sub-claim tiers
Greater Than	5	5.6	Mech. Supported	3.3–4.7	Caus. Suggestive–Mech. Supported
Othello	3	4.4	Mech. Supported	3.1–4.4	Caus. Suggestive–Mech. Supported
Successor Heads	4	3.1	Caus. Suggestive	1.9–4.7	Proposed–Mech. Supported
Gender Bias	5	3.3	Caus. Suggestive	0.0–1.9	Proposed
Probing	3	1.4	Proposed	1.4	Proposed

The hierarchical differentiation is consistent across all papers: main claims score higher than the average of their sub-components in every case. The gap is largest for papers with dedicated circuit-level testing (IOI, Grokking) and smallest for papers where all claims share similar evidence (Probing, Othello). This demonstrates that the evidence scope rules and leak audit are functioning as intended—different claims within the same paper receive scores based on their individual evidence rather than uniformly inheriting the paper's strongest results.

6. Discussion

6.1 Practical Utility

The system's primary value lies in speed and structure rather than perfect accuracy. A full paper evaluation completes in 2–3 minutes, compared to hours for expert review, enabling systematic triage of the MI literature. The 27-criterion breakdown provides actionable feedback that goes beyond a single score: instead of "this paper is weak," the system can report "this paper lacks sufficiency evidence (I2 = NO), reports no variance (M1 = NO), and has not been tested across architectures (E6 = NO)." This granularity could serve researchers as a self-assessment checklist before submission and reviewers as a structured starting point for evaluation.

The leak audit addresses a challenge that extends beyond automated extraction. Even human reviewers may inadvertently attribute circuit-level evidence to individual component claims, particularly when reviewing complex papers with many interacting components. The audit's explicit separation of system-level and sub-component claims, and its requirement that sub-component evidence name specific experiments on those components, provides a useful conceptual framework regardless of whether the scoring is automated.

6.2 Sources of Error

The residual +1 tier over-scoring on 4 of 9 papers is the most significant limitation. It likely reflects an inherent asymmetry in the LLM-as-judge paradigm: when prompted to assess whether evidence exists for a criterion, the model is biased toward finding confirming evidence in the paper text. This confirmation bias is exacerbated by the long extraction prompt, which provides detailed descriptions of what would constitute YES for each criterion, effectively teaching the model what to look for. Negative examples (what does NOT count) are included but may be less salient.

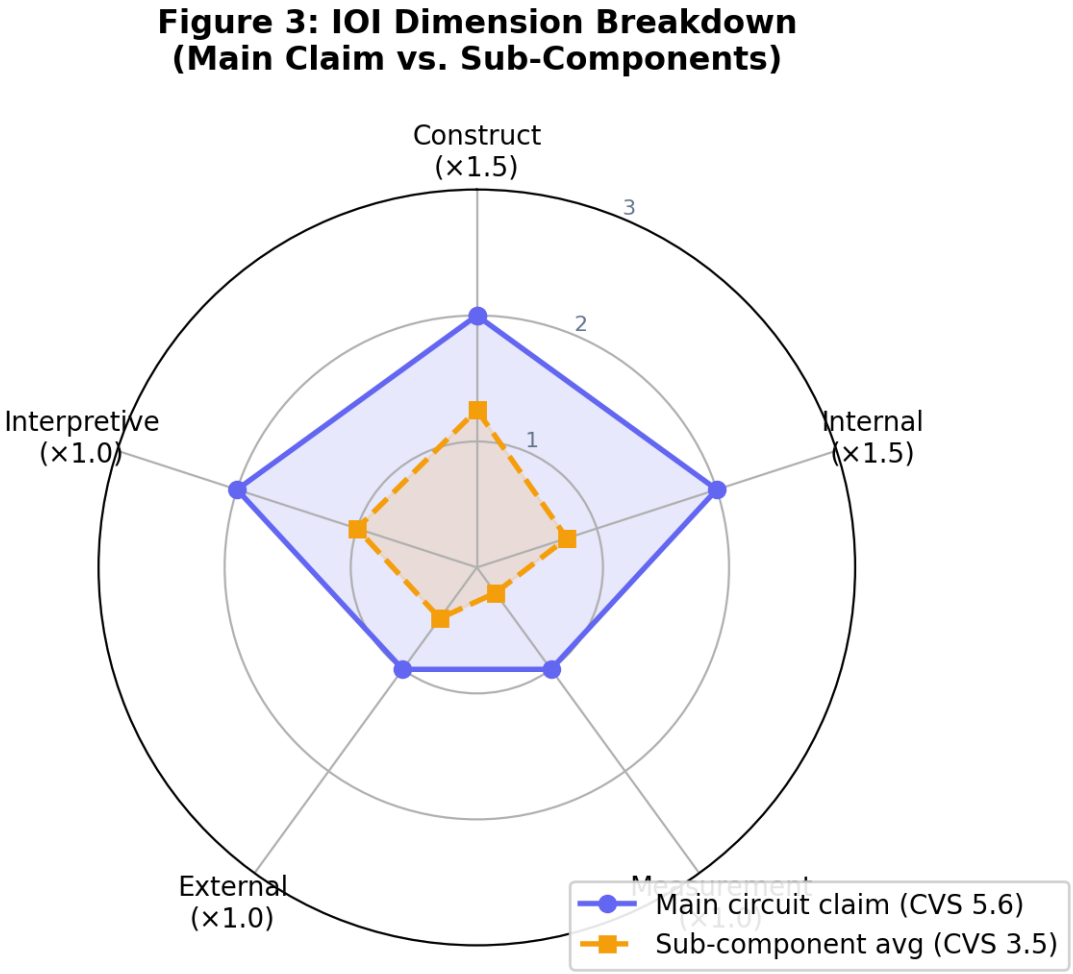

Figure 4: Figure 3: Dimension breakdown for IOI, comparing the main circuit claim (CVS 5.6, Mechanistically Supported) against the average of four sub-component claims (CVS 3.5, Causally Suggestive). The main claim scores higher on all five dimensions, with the largest gap on Internal validity (necessity + sufficiency from circuit-level faithfulness testing) and Measurement validity (baseline comparisons available only at the circuit level). Weights shown in parentheses.

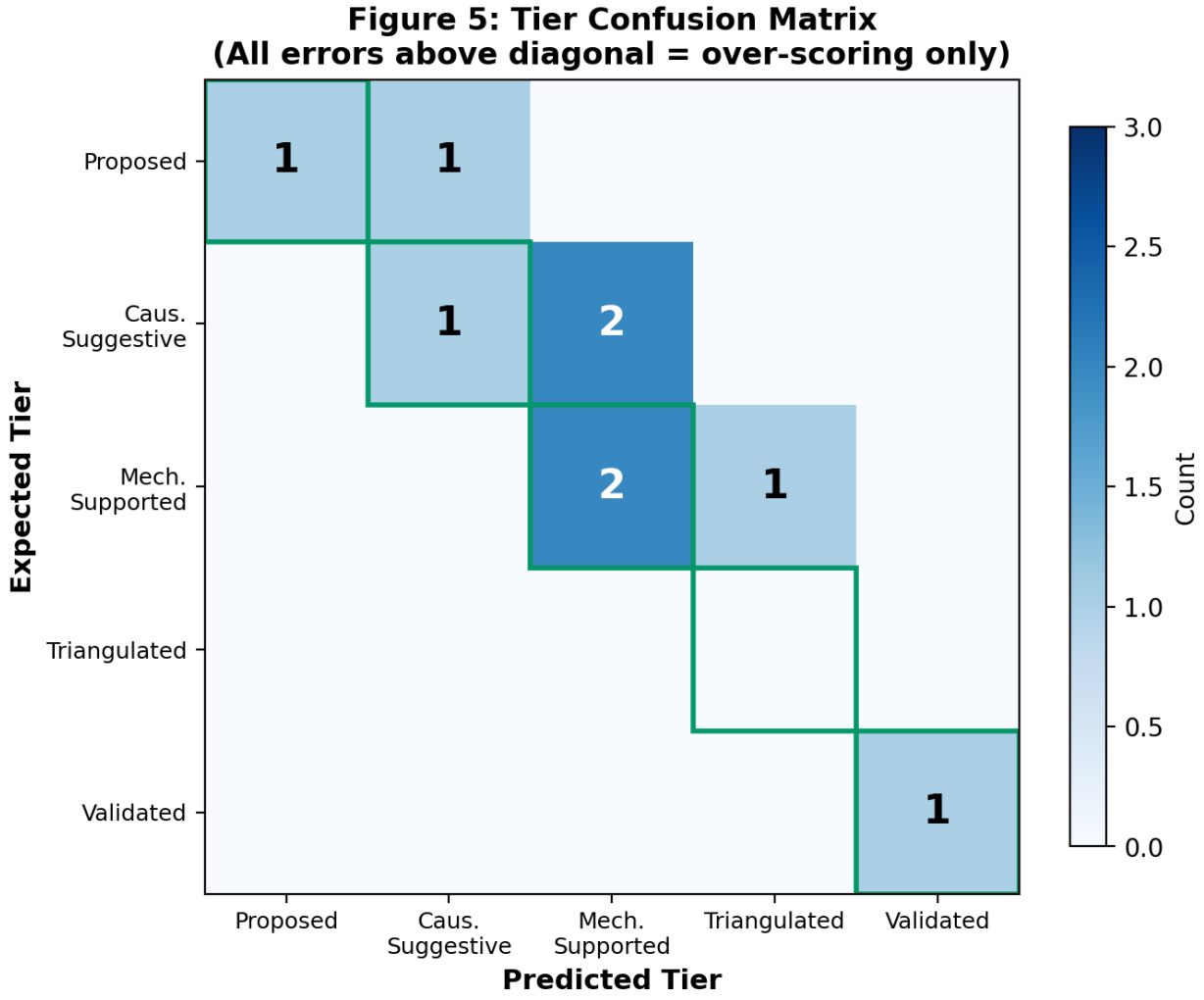

Figure 5: Figure 5: Tier confusion matrix showing predicted vs. expected tiers. All errors fall above the diagonal (over-scoring only), concentrated at the Causally Suggestive → Mechanistically Supported boundary. No papers are under-scored. Green outlines highlight the diagonal (exact matches).

Multi-run minimum voting eliminates the stochastic component of this bias but cannot correct consistent over-scoring. When the model awards YES on a borderline criterion in all 3 runs, the minimum is still YES. This explains the residual pattern: 4 papers that are over-scored have criteria where the model consistently misinterprets the evidence (e.g., treating standard ablation as establishing graded response), while the 5 papers matching their reference labels have criteria where the evidence clearly meets or fails the requirements.

The system's coverage is limited in several respects. While it accepts papers from arXiv, OpenReview, PDF URLs, blog posts, and local files, 2 of 13 attempted case study papers had incorrect arXiv metadata mappings—a failure rate of 15%. HTML-extracted text from blog posts may lack the structure of PDF-extracted text, potentially affecting extraction quality. It uses a single language model (DeepSeek V4 Pro); different models may exhibit different biases, and an ensemble approach could improve robustness. Finally, the criteria rubric is calibrated for empirical circuit-discovery papers and is not suitable for theoretical work (see Section 5.4).

6.3 Sensitivity to Tier Boundaries

The tier boundaries (2.0, 4.0, 6.0, 8.0) are evenly spaced on the 0–10 scale, a choice that simplifies interpretation but is not the only option. To assess sensitivity, we examined which papers would change tier under a $\pm$0.5 shift to each boundary. The most sensitive boundary is the Causally Suggestive / Mechanistically Supported threshold at 4.0: shifting it to 4.5 would reclassify Othello (4.4) from Mechanistically Supported to Causally Suggestive, converting one of the four over-scored papers into an exact match. The Proposed / Causally Suggestive boundary at 2.0 is also sensitive—shifting it to 2.5 would not change any results, but shifting it to 1.5 would reclassify Probing (1.4) from Proposed to below the new threshold. All other papers have CVS scores more than 0.5 away from any boundary. This analysis suggests that the tier classification is moderately robust to boundary choice, but that papers near boundaries (CVS within 0.5 of a threshold) should be interpreted cautiously—the specific tier matters less than the dimensional breakdown.

6.4 Interpreting Disagreements with Reference Labels

Some discrepancies between automated and reference assessments may reflect legitimate judgment differences rather than system errors. The four over-scored papers each present borderline cases. Greater Than uses causal scrubbing alongside ablation, which the model interprets as establishing sufficiency; experts may reasonably disagree about whether causal scrubbing on a partial circuit qualifies. Induction Heads includes evidence across 10+ model sizes and architectures, which the model credits toward cross-architecture testing (E6); the expert tier does not account for this because the Transformer Circuits Thread format makes it difficult to assess evidence systematically. Notably, the automated +1 over-scoring on Induction Heads may actually be correct—the breadth of evidence in that paper arguably merits Triangulated, and the reference label may be too conservative. Othello's nonlinear probing interventions are interpreted by the model as causal evidence, while experts may view them as correlational. Gender Bias reports quantitative PCA analysis with crowd-worker validation, which the model credits toward causal evidence but experts view as descriptive. A larger-scale calibration study with multiple independent annotators would be needed to distinguish systematic automated bias from inter-rater disagreement in the reference labels themselves.

7. Future Work

Several directions could improve the system's accuracy and utility. Running the system on a larger corpus of 50 or more papers with independent multi-annotator labels would enable calibration of systematic biases—for example, applying a fixed tier correction or adjusting dimension weights. Increasing the multi-run count beyond 3 could further tighten scores, though diminishing returns are expected since the residual bias stems from consistent rather than stochastic over-scoring. Cross-model ensembling—running extraction with multiple language models and retaining only criteria that achieve consensus—would provide a conservative lower bound on evidence strength and could address the consistent bias that multi-run minimum cannot fix.

The system could also be extended with automatic paper type detection to classify papers as empirical-circuit, theoretical, representational, or methodological before extraction, applying type-specific criteria guidance without risk of cross-contamination. An interactive web dashboard surfacing per-criterion scores would enable researchers to use the system for self-assessment before submission. Finally, integrating the automated extraction with the mechanistic-validity-graph registry would enable longitudinal tracking of evidence accumulation across papers that study the same circuit or mechanism.

8. Conclusion

Automated CVS evaluation provides a fast, reproducible first-pass assessment of mechanistic interpretability claims. With multi-run minimum debiasing, the system achieves 100% within-one-tier agreement (95% CI: [66%, 100%]) and 56% exact tier match (95% CI: [21%, 86%]) against consensus reference labels on 9 case study papers, with a systematic but bounded over-scoring tendency of at most +1 tier. The leak audit addresses the evidence attribution problem that arises when circuit-level and component-level claims coexist in the same paper, while multi-run minimum voting eliminates stochastic variance by exploiting the asymmetric error structure of LLM-as-judge scoring. While the system is not a replacement for expert review, it can serve as a triage tool for the growing MI literature, a structured self-assessment rubric for researchers, and a foundation for systematic evidence tracking across the field.

Code and Data Availability

The extraction pipeline, scoring code, evaluation data, and all 9 case study outputs are available at https://github.com/mechanistic-validity/mechanistic-validity-graph. The extraction prompts (v15b) are included in the repository source code. Reference labels and per-criterion annotations for all 9 papers are provided in the `examples/` directory.

References

Bills, S., Cammarata, N., Mossing, D., Tillman, H., Gao, L., Goh, G., Sutskever, I., Leike, J., Wu, J., & Saunders, W. (2023). Language models can explain neurons in language models. *OpenAI Blog*.

Bolukbasi, T., Chang, K.-W., Zou, J. Y., Saligrama, V., & Kalai, A. T. (2016). Man is to computer programmer as woman is to homemaker? Debiasing word embeddings. In *Advances in Neural Information Processing Systems* (pp. 4349–4357).

Campbell, D. T., & Stanley, J. C. (1963). *Experimental and quasi-experimental designs for research*. Houghton Mifflin.

Chiang, W.-L., Zheng, L., Sheng, Y., Angelopoulos, A. N., Li, T., Li, D., Zhang, H., Zhu, B., Jordan, M., Gonzalez, J. E., & Stoica, I. (2024). Chatbot Arena: An open platform for evaluating LLMs by human preference. In *Proceedings of the 41st International Conference on Machine Learning*.

Conmy, A., Mavor-Parker, A. N., Lynch, A., Heimersheim, S., & Garriga-Alonso, A. (2023). Towards automated circuit discovery for mechanistic interpretability. In *Advances in Neural Information Processing Systems*.

Cook, T. D., & Campbell, D. T. (1979). *Quasi-experimentation: Design and analysis issues for field settings*. Houghton Mifflin.

Elhage, N., Hume, T., Olsson, C., Schiefer, N., Henighan, T., Kravec, S., Hatfield-Dodds, Z., Lasenby, R., Drain, D., Chen, C., Grosse, R., McCandlish, S., Kaplan, J., Amodei, D., Wattenberg, M., & Olah, C. (2022). Toy models of superposition. *Transformer Circuits Thread*.

Gould, S., Ong, L., & Ogden, G. (2023). Successor heads: Recurring, interpretable attention heads in the wild. *arXiv preprint arXiv:2312.09230*.

Hanna, M., Liu, O., & Variengien, A. (2023). How does GPT-2 compute greater-than?: Interpreting mathematical abilities in a pre-trained language model. *arXiv preprint arXiv:2305.00586*.

Hewitt, J., & Liang, P. (2019). Designing and interpreting probes with control tasks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing* (pp. 2733–2743).

Li, K., Hopkins, A. K., Bau, D., Viégas, F., Pfister, H., & Wattenberg, M. (2022). Emergent world representations: Exploring a sequence model trained on a synthetic task. *arXiv preprint arXiv:2210.13382*.

Luan, Y., He, L., Ostendorf, M., & Hajishirzi, H. (2018). Multi-task identification of entities, relations, and coreference for scientific knowledge graph construction. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing* (pp. 3219–3232).

McDougall, C., Conmy, A., Rushing, C., McGrath, T., & Neel, N. (2023). Copy suppression: Comprehensively understanding an attention head. *arXiv preprint arXiv:2310.04625*.

Nanda, N., Chan, L., Lieberum, T., Smith, J., & Steinhardt, J. (2023). Progress measures for grokking via mechanistic interpretability. In *International Conference on Learning Representations*.

Olsson, C., Elhage, N., Nanda, N., Joseph, N., DasSarma, N., Henighan, T., Mann, B., Askell, A., Bai, Y., Chen, A., Conerly, T., Drain, D., Ganguli, D., Hatfield-Dodds, Z., Hernandez, D., Johnston, S., Jones, A., Kernion, J., Lovitt, L., … Olah, C. (2022). In-context learning and induction heads. *Transformer Circuits Thread*.

Tower, E., et al. (2026). Mechanistic validity: A framework for evaluating evidence strength in mechanistic interpretability claims. *Working paper*.

Wadden, D., Lin, S., Lo, K., Wang, L. L., van Zuylen, M., Cohan, A., & Hajishirzi, H. (2020). Fact or fiction: Verifying scientific claims. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing* (pp. 7534–7550).

Wang, K., Variengien, A., Conmy, A., Shlegeris, B., & Steinhardt, J. (2022). Interpretability in the wild: A circuit for indirect object identification in GPT-2 small. *arXiv preprint arXiv:2211.00593*.

Zheng, L., Chiang, W.-L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E. P., Zhang, H., Gonzalez, J. E., & Stoica, I. (2023). Judging LLM-as-a-judge with MT-Bench

and Chatbot Arena. In *Advances in Neural Information Processing Systems*.

Appendix A: Prompt Engineering Details

The system prompt (v15b) contains approximately 3,500 tokens of criteria guidance organized into four sections. The criteria-specific guidance section provides detailed instructions for all 27 criteria with both positive and negative examples, such as specifying that C2 (Structural plausibility) requires weight-space or compositional analysis and that attention patterns alone merit only PARTIAL. The scoring pitfalls section lists 10 common errors, each stated as a concrete rule (e.g., "Faithfulness != sufficiency," "Same architecture different size != cross-architecture"). The evidence scope section specifies that circuit-level evidence transfers only to the circuit claim, not to sub-component claims. The theoretical paper section (added in v15b) maps analytical methods to criteria for mathematical toy-model papers only, with explicit exclusion clauses for empirical circuit-discovery papers.

A complete few-shot example presents a hypothetical subject-verb agreement circuit paper with two claims (the circuit claim and one sub-component claim), demonstrating appropriate differentiation in scoring between the main and sub-component levels. This example includes all 27 criteria for both claims, with the sub-component claim appropriately receiving NO for I2 (no isolation test) and NO for I3 (no off-task measurement), whereas the circuit claim receives YES for these based on circuit-level evidence.

The leak audit prompt contains approximately 500 tokens focused on the five leak-prone criteria (I2, I3, I5, E2, E5), with explicit instructions to never downgrade system-level claims, to only downgrade when confident that evidence was borrowed from another component's experiments, and to default to no downgrade when uncertain.

Appendix B: Full Per-Paper Results (Multi-Run Minimum, N=3)

IOI Circuit: Main circuit claim 5.6 (Mech. Supported), Name Movers 3.9 (Caus. Suggestive), S-Inhibition 3.3 (Caus. Suggestive), Duplicate Token / Induction 3.3 (Caus. Suggestive), Backup Name Movers 3.3 (Caus. Suggestive). The hierarchical differentiation appropriately reflects the decreasing evidence strength from the main circuit (which has dedicated sufficiency testing via faithfulness) to peripheral components (which were identified primarily by ablation).

Grokking: Modular addition algorithm 8.3 (Validated), Fourier components 1.9 (Proposed), MLP neurons 3.3 (Caus. Suggestive), Attention heads 1.9 (Proposed), Phase transition 2.5 (Caus. Suggestive). The main claim appropriately reaches Validated due to the combination of Fourier analysis, ablation, and cross-seed testing.

Copy Suppression: Main L10H7 mechanism 5.6 (Mech. Supported), with sub-claims at 1.4–4.7. The main claim score appropriately reflects the comprehensive CSPA methodology.

Probing: All three claims scored 1.4 (Proposed), appropriately reflecting that Hewitt & Liang (2019) is a methodological paper proposing a framework (control tasks for probing) rather than making specific mechanistic claims about model internals.

Successor Heads: Main claim 3.1 (Caus. Suggestive), MLP0 representations 1.9 (Proposed), Mod-10 features 4.7 (Mech. Supported), Pythia-1.4B example 2.5 (Caus. Suggestive). The system appropriately differentiates between the well-evidenced mod-10 feature analysis and the broader successor head claim.

Greater Than: Main circuit claim 5.6 (Mech. Supported), MLP contributions 3.3 (Caus. Suggestive), Attention heads 3.3 (Caus. Suggestive), MLP 10 neurons 3.3 (Caus. Suggestive), Generalization 4.7 (Mech. Supported). The main claim is over-scored by +1 tier relative to the expected Causally Suggestive.

Induction: Main induction head claim 6.4 (Triangulated), phase change 3.1 (Caus. Suggestive), abstract in-context learning 1.4 (Proposed), smeared-key architecture 1.4 (Proposed). The main claim is over-scored by +1 tier, potentially reflecting the paper's extensive multi-model analysis across 10+ models.

Othello: Board state representation 4.4 (Mech. Supported), causal link to predictions 4.4 (Mech. Supported), nonlinear vs. linear probes 3.1 (Caus. Suggestive). Over-scored by +1 tier; the probing intervention evidence may be inflating sufficiency scores.

Gender Bias: Gender direction 3.3 (Caus. Suggestive), linear separability 1.9 (Proposed), hard debiasing 1.9 (Proposed), gender projection 0.0 (Proposed), indirect bias 0.0 (Proposed). The main claim is over-scored by +1 tier; the model may be over-crediting PCA analysis as causal evidence.