NOVA-Test: Auditing LLM-Generated Research Hypotheses with Structural Analogy, Prior-Work, and Falsifiability Tests

David Scott Lewis, Karl Wang, Saien Deng, Enrique Zueco, Xiping Gong[1]

Abstract

We study LLM research ideation as a hypothesis-testing problem, not a prompting tutorial. The unit under test is a generated research hypothesis together with its structural justification, nearest-prior relation, and proposed falsifier. We repurpose NOVA-Analog as NOVA-Test, a lightweight rejection protocol with three gates: source–target structural alignment, nearest-prior novelty audit, and explicit falsification planning. Across 12 ML/CS bottlenecks and three generation conditions, NOVA-Test improves judged novelty (3.88 vs. 3.00 one-shot and 2.83 literature-search), technical specificity (4.54 vs. 3.21 and 3.00), and falsifiability (4.79 vs. 4.46 and 4.21). Bonferroni-corrected Wilcoxon tests remain significant for all planned comparisons except feasibility versus literature-search. A 488-work novelty audit labels 10/12 NOVA-Test hypotheses plausibly novel versus 3/12 literature-search hypotheses (Fisher one-sided $p = 0.006$). The contribution is not autonomous discovery; it is a reproducible hypothesis-audit scaffold for LLM-related testing problems.

1. Introduction

This paper tests LLM-generated research hypotheses before they are executed. The object under test is not an LLM's next-token accuracy and not the elegance of a prompt, but a candidate hypothesis h: a claim, a structural rationale, a nearest-prior relation, and an observation that would reject it. We ask whether h should be retained for downstream experimentation or rejected as a surface analogy, rediscovery, or unfalsifiable claim. This framing makes AI-assisted ideation an LLM-related testing problem: the first statistical decision in an AI-assisted research pipeline is often whether a generated hypothesis is worth testing at all.

This matters because apparent idea quality is not stable under execution. LLM-generated ideas can look novel at ideation time but weaken after implementation (Si et al., 2024; 2025), while benchmarks and hypothesis-generation datasets repeatedly separate novelty from feasibility, truthfulness, and grounding (Guo et al., 2025; Xiong et al., 2025; O'Neill et al., 2025). Human-in-the-loop ideation systems and concept-network agents help researchers explore candidate spaces (Garikaparthi et al., 2025; Zhao et al., 2025), but they do not by themselves specify when a generated hypothesis should be rejected.

End-to-end AI research agents make the need for an upstream test sharper. Systems such as AI Scientist, AI Scientist-v2, Agent Laboratory, and broader agentic-science frameworks attempt to move from hypotheses to experiments and papers (Lu et al., 2024; Yamada et al., 2025; Schmidgall et al., 2025; Wei et al., 2025). Independent evaluations report weak novelty assessment, brittle implementation, and fabricated or invalidated experimental results in autonomous research settings (Beel et al., 2025; Kon et al., 2025; Chen et al., 2026; Zhu et al., 2025; Xie et al., 2025). Once an upstream hypothesis passes unchallenged, it shapes the baselines chosen, data collected, and negative results later rationalized.

We therefore recast the original NOVA-Analog workflow as a paired hypothesis-audit study. The null is that adding structured analogy and audit gates does not change the distribution of test-ready hypotheses relative to one-shot or literature-search prompting. We evaluate this null over the same 12 bottlenecks under three generation conditions, using paired ordinal endpoints and a separate nearest-prior audit. The contribution is a rejection protocol and an evaluation of its testing validity: where the protocol changes the generated hypotheses, where it fails, and which audit decisions remain dependent on retrieval coverage or judge calibration. Figure 1 summarizes the protocol.

2. NOVA-Test Protocol

NOVA-Test uses cross-domain analogy as the generator and hypothesis testing as the filter. This design builds on

[1]AIXC Research, Zaragoza, Spain.
Correspondence to: David Scott Lewis
<reports@aixcresearch.com>.

The ICML 2026 Workshop on Hypothesis Testing.
Copyright 2026 by the author(s).

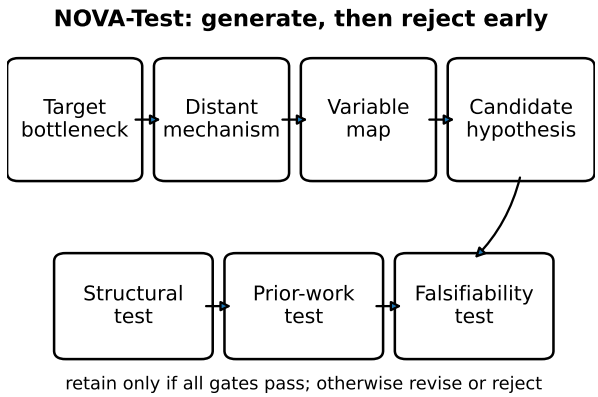

Figure 1. NOVA-Test treats LLM ideation as a sequence of rejection gates. Hypotheses are retained only if their structural map, nearest-prior card, and falsification plan all survive inspection.

analogical-search and design-by-analogy work, where useful analogies depend on relational structure rather than keyword overlap (Kang et al., 2022; Ding et al., 2023; Yuan et al., 2023; Trench & Minervino, 2020; Kokate & Onkar, 2026). It also follows recent falsification- and verification-oriented views of AI-assisted science (Liu et al., 2024; Dycke & Gurevych, 2026; Li et al., 2026; Zheng et al., 2025).

Given a target bottleneck b, the workflow extracts target invariants: the constraint, why current methods fail, and a measurable success proxy. It then asks the LLM for source mechanisms from distant domains such as ecology, numerical analysis, queuing theory, or statistical physics. For a candidate source mechanism s, the model must produce a variable-level mapping table between source variables and target variables, with each row labeled exact, approximate, speculative, or mismatch.

Structural test. Let $M(h)$ be the mapping rows for hypothesis h. We compute

$$\rho(h) = \frac{\#\{m \in M(h) : m \text{ exact or approximate}\}}{|M(h)|}. \quad (1)$$

The threshold 0.6 is a screening rule, not a theorem. Its purpose is to force the model to expose correspondences and mismatches before the prose claim is considered.

Prior-work test. Surviving hypotheses are audited against a 488-work deduplicated corpus using lexical and embedding retrieval. The audit assigns each hypothesis to A = known, B = incremental, C = plausibly novel/high-risk, or D = no close prior found. This connects NOVA-Test to novelty-assessment and scientific claim-verification pipelines (Lin et al., 2025; Qiu et al., 2025; Edelman & Skolnick, 2025; Fadeeva et al., 2025; Zheng et al., 2024). We use the binary C/D count only for aggregate statistics; the auditable object is the nearest-prior card.

Falsifiability test. Each retained hypothesis must include a failure condition stated as an empirical or computational observation. This aligns with evidence-card and claim–evidence work in scientific verification (Javaji et al., 2025; Martin-Boyle et al., 2026; Huang et al., 2026). For example, a Mixture-of-Experts routing hypothesis is rejected if its proposed anti-congestion gate fails to reduce load inequality while preserving validation perplexity within a pre-specified tolerance.

3. Evaluation

Tasks and baselines. We curated 12 ML/CS bottlenecks spanning RLHF reward hacking, reparameterization-invariant flatness, MoE routing collapse, continual learning, diffusion sampling, Ramsey constructions, BFT message complexity, sparse-autoencoder compositionality, hard-exploration RL, quantization, instance-adaptive optimization, and maximum-clique approximation. For each bottleneck, we generated hypotheses under three conditions: a one-shot prompt, a literature-search prompt, and NOVA-Test.

Judging and tests. Blinded judges from two model families scored novelty, technical specificity, falsifiability, and feasibility on 1–5 ordinal scales. We averaged the two judges per bottleneck and used paired one-sided Wilcoxon signed-rank tests because scores are ordinal and paired by bottleneck (Wilcoxon, 1945). We corrected eight planned comparisons with Bonferroni adjustment (Bonferroni, 1936). For novelty-audit classes we used one-sided Fisher exact tests comparing plausibly novel counts (Fisher, 1922).

Representative hypothesis card. For MoE routing collapse, NOVA-Test mapped tokens routed among experts to particles distributed over sites in a zero-range process. The structural map linked site occupancy to per-expert load, hop-rate shape to router attractiveness, and condensation transitions to routing collapse. The retained hypothesis proposed a superlinear congestion penalty applied directly to router logits. The falsification plan required lower load Gini without a validation-perplexity penalty beyond the pre-declared tolerance. The point is not that this hypothesis is known to be correct; it is that the idea is accompanied by a structural map, nearest prior, and rejection criterion.

4. Results

Figure 2 summarizes the paired judging results. NOVA-Test improves novelty and specificity against both baselines with corrected $p < 0.005$. It also improves falsifiability against both baselines after correction. Feasibility improves

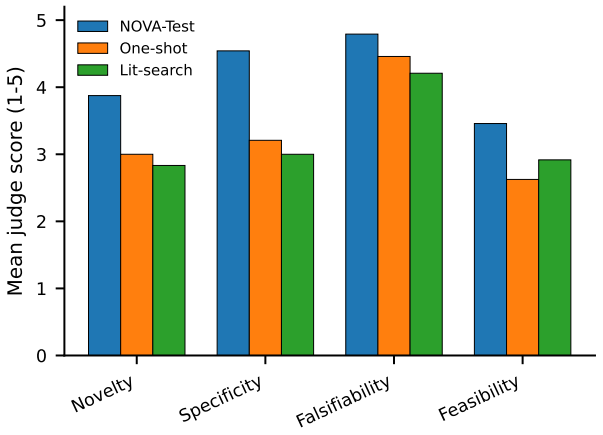

Figure 2. Mean blinded judge scores across 12 bottlenecks. Bonferroni-corrected Wilcoxon tests show significant gains for novelty, specificity, and falsifiability against both baselines; feasibility versus literature-search does not survive correction.

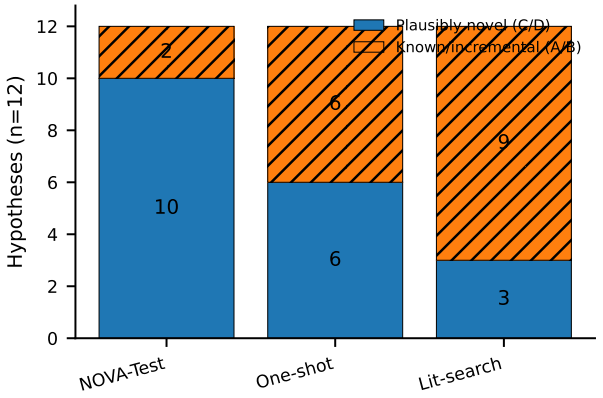

Figure 3. Nearest-prior novelty audit over 488 deduplicated works. C/D denotes plausibly novel or no close prior; A/B denotes known or incremental.

against one-shot prompting but not against the literature-search baseline after correction ($p_c = 0.086$), consistent with the expectation that retrieval can make ideas easier to implement while anchoring them to known work.

The prior-work audit gives a second, more diagnostic view. As shown in Figure 3, 10/12 NOVA-Test hypotheses were classified C/D, compared with 6/12 one-shot and 3/12 literature-search hypotheses. The improvement over literature-search is significant (odds ratio 15.0, Fisher one-sided $p = 0.006$); the improvement over one-shot is directionally positive but not significant ($p = 0.097$). The strongest supported claim is therefore that structured cross-domain mapping helps escape local literature anchoring, not that it certifies true novelty.

An external analogy audit labeled all 12 top NOVA-Test mappings structurally useful, with external structural fractions between 0.83 and 0.86. This is not proof of scientific correctness, but it changes the artifact being reviewed: a reader can inspect explicit correspondences and mismatches rather than judge a free-form metaphor. Historical rediscovery tests were treated as leakage-sensitive diagnostics. They often produced plausible adjacent mechanisms rather than exact rediscoveries; we therefore do not use rediscovery as primary evidence for novelty.

5. Discussion and Limitations

The main methodological claim is that LLM-generated research hypotheses are a distinct testing object. Traditional ML testing usually evaluates a trained model, algorithm, or deployment rule. AI-assisted research pipelines first generate claims that later determine which data are collected, baselines are implemented, and negative results are ignored. Weak tests at the hypothesis stage can therefore bias the entire downstream research process.

NOVA-Test separates three nulls that are often conflated in ideation. A hypothesis can be structurally coherent but not novel; novel but infeasible; or feasible but derived from a superficial metaphor. A single "idea quality" score hides these distinctions. The gate structure localizes failures: a failed map is a failed analogy, a class-A novelty card is a rediscovery, and a missing failure condition is an unfalsifiable claim.

The limitations are equally important. First, $\rho(h) \geq 0.6$ is heuristic and should be stress-tested or tuned by domain. Second, the novelty audit is retrieval-sensitive. We observed time-weighted retrieval bias: standard ranking over-retrieved recent LLM-ideation papers and under-retrieved older analogy and design-by-analogy foundations. Year-stratified retrieval should be mandatory when novelty is part of the claim. Third, LLM judges can reward persuasive text and overrate feasibility, so we use them only for relative paired comparisons, not as final arbiters of validity. Fourth, researchers may adapt prompts after seeing attractive outputs; future evaluations should pre-register prompts, thresholds, audit corpora, and bottlenecks.

6. Reproducibility and Conclusion

The appendix specifies the full prompt templates, extended statistics, and audit-card schema. The rerunnable package consists of the 12 bottleneck cards, prompts for all three generation conditions, raw generated hypotheses, judge rubrics, novelty-audit labels, and scripts that recompute the Wilcoxon and Fisher tests. No model training is required for the main evaluation.

Table 1. Reviewer-facing hypothesis-audit card emitted by NOVA-Test.

Field	Contents
Bottleneck	Target constraint and measurable success proxy
Map	Source mechanism, variable correspondences, mismatches, $\rho(h)$
Prior	Nearest title, overlap, remaining gap, A/B/C/D class
Falsifier	Baseline, metric, tolerance, data, rejection condition
Status	Retain, revise, reject, or expert audit required

Reviewer-facing use. We recommend using NOVA-Test as a card generator rather than as a paper writer. A card contains the bottleneck, source mechanism, structural fraction, critical mismatches, nearest prior, novelty class, and falsification criterion. This makes negative outcomes useful: a failed card documents why an attractive LLM-generated idea should not be pursued. It also helps separate human judgment from model fluency, because the reviewer inspects a fixed schema rather than a polished narrative.

Reviewer-risk mitigation. Two concerns deserve explicit treatment in the main text. First, the structural-fraction threshold is a guardrail, not a statistical estimator. The current evidence should therefore be read as a comparison of fixed protocols, not as proof that 0.6 is universally optimal. We added the threshold-sweep ablation to the pre-registered follow-up plan and report the external structural fractions for all retained top cards in the appendix; importantly, those cards are not marginal cases, since the external audit scored them between 0.83 and 0.86.

Second, LLM-as-judge evaluation is only a measurement instrument. Across the 36 bottleneck-condition cells, cross-judge Spearman correlations were 0.62 for novelty, 0.84 for specificity, 0.53 for falsifiability, and 0.81 for feasibility (Spearman, 1904). This supports relative paired ranking but does not make either judge an arbiter of truth. For that reason, the paper treats feasibility conservatively, highlights the nonsignificant literature-search comparison, and requires human or executable follow-up before any retained card is promoted to an experiment.

These clarifications also change how the nearest-prior audit should be interpreted. A 488-work corpus is strong enough to expose obvious rediscovery but too small to certify global novelty. Thus, class A/B is a rejection signal, while class C/D is only a permission to continue. The audit card must preserve the nearest-prior title, overlap rationale, and remaining gap so that a reviewer can substitute a broader or domain-specific corpus without rerunning the entire ideation workflow.

A third concern is baseline strength. One-shot and literature-search prompting are diagnostic floors: they test whether free-form novelty or local retrieval anchoring is enough. They are not meant to dominate stronger agentic ideation systems (Saraogi et al., 2025). A fair next comparison is to run the same bottlenecks through a decomposition or lab-meeting-style agent and score only the resulting cards, so the generator can change while the tested object remains fixed.

Older analogy context. The structural gate is not meant to be a new theory of analogy. It imports an older structure-mapping norm: analogy licenses inference only when roles and relations, not just labels, align across source and target (Gentner & Forbus, 2011). This matters for a testing paper because the gate operationalizes a failure condition inherited from cognitive science: if the correspondence table cannot support relational projection, the LLM's hypothesis is rejected before novelty or feasibility are scored. Analogical-search and design-by-analogy systems make the same distinction computationally (Kang et al., 2022; Jiang et al., 2022).

What to ablate next. The highest-priority ablation is to remove only the variable-level mapping step while retaining bottleneck extraction, distant-domain search, prior-work audit, and falsification planning. That ablation would test whether the mapping table is the causal intervention or merely a visible byproduct of a longer prompt. A second ablation should vary the structural threshold and report how many hypotheses are rejected, retained, or later judged superficial. A third extension should replace LLM-only judging with a small expert panel over 3–4 bottlenecks, using the same cards.

Boundary of the claim. NOVA-Test does not test whether a retained hypothesis is true. It tests whether the hypothesis is worth executing: structurally interpretable, not obviously rediscovered, and equipped with a falsifier. This boundary is crucial for workshop use. A workflow that produces fewer but better-scoped hypotheses can still accelerate research even when many retained ideas ultimately fail, because failure occurs under a pre-declared test rather than through ad hoc post-rationalization.

NOVA-Test reframes cross-domain LLM ideation as a hypothesis-testing problem. Rather than asking whether an LLM can be creative, the workflow asks whether a generated claim survives structural alignment, prior-work, and falsifiability tests. On 12 ML/CS bottlenecks, this protocol improves novelty, specificity, and falsifiability over one-shot and literature-search baselines while surfacing retrieval and evaluator failure modes that matter for LLM-related testing problems.

References

Beel, J., Kan, M.-Y., and Baumgart, M. Evaluating sakana's ai scientist: Bold claims, mixed results, and a promising future? *ACM SIGIR Forum*, 2025. doi: 10.1145/3769733.3769747.

Bonferroni, C. E. *Teoria Statistica delle Classi e Calcolo delle Probabilità*. Pubblicazioni del R. Istituto Superiore di Scienze Economiche e Commerciali di Firenze, 1936.

Chen, H., Xiong, M., Lu, Y., Han, W., Deng, A., He, Y., Wu, J., Li, Y., Liu, Y., and Hooi, B. Mlr-bench: Evaluating ai agents on open-ended machine learning research. *arXiv preprint*, 2026.

Ding, Z., Srinivasan, A., MacNeil, S., and Chan, J. Fluid transformers and creative analogies: Exploring large language models' capacity for augmenting cross-domain analogical creativity. *Creativity and Cognition*, 2023. doi: 10.1145/3591196.3593516.

Dycke, N. and Gurevych, I. Automatic reviewers fail to detect faulty reasoning in research papers: A new counterfactual evaluation framework. *Transactions of the Association for Computational Linguistics*, 14:465–488, Jan 2026. ISSN 2307-387X. doi: 10.1162/tacl.a.642. URL https://doi.org/10.1162/tacl.a.642.

Edelman, B. and Skolnick, J. Valsci: an open-source, self-hostable literature review utility for automated large-batch scientific claim verification using large language models. *BMC Bioinformatics*, 2025. doi: 10.1186/s12859-025-06159-4.

Fadeeva, E., Rubashevskii, A., Piatrashyn, D., et al. Faithfulness-aware uncertainty quantification for fact-checking the output of retrieval augmented generation. *arXiv preprint*, 2025.

Fisher, R. A. On the interpretation of χ^2 from contingency tables, and the calculation of P. *Journal of the Royal Statistical Society*, 85(1):87–94, 1922. doi: 10.2307/2340521.

Garikaparthi, A., Patwardhan, M., Vig, L., and Cohan, A. Iris: Interactive research ideation system for accelerating scientific discovery. *arXiv preprint*, 2025.

Gentner, D. and Forbus, K. D. Computational models of analogy. *Wiley Interdisciplinary Reviews: Cognitive Science*, 2(3):266–276, 2011. doi: 10.1002/wcs.105. URL https://doi.org/10.1002/wcs.105.

Guo, S., Shariatmadari, A. H., and Xiong, G. Ideabench: Benchmarking large language models for research idea generation. *Preprint*, 2025. doi: 10.1145/3711896.3737419.

Huang, Y., Ribeiro, L. F. R., Hardalov, M., Dhingra, B., Dreyer, M., and Saligrama, V. Deepfact: Co-evolving benchmarks and agents for deep research factuality. *ArXiv*, Mar 2026. doi: 10.48550/arxiv.2603.05912. URL https://doi.org/10.48550/arxiv.2603.05912.

Javaji, S. R., Cao, Y., Li, H., Yu, Y., Muralidhar, N., and Zhu, Z. Can ai validate science? benchmarking llms for accurate scientific claim to evidence reasoning. *ArXiv*, abs/2506.08235, Jun 2025. doi: 10.48550/arxiv.2506.08235. URL https://doi.org/10.48550/arxiv.2506.08235.

Jiang, S., Hu, J., Wood, K. L., and Luo, J. Data-driven design-by-analogy: State-of-the-art and future directions. *Journal of Mechanical Design*, 144(2), 2022. doi: 10.1115/1.4051681. URL https://doi.org/10.1115/1.4051681.

Kang, H. B., Qian, X., Hope, T., Shahaf, D., Chan, J., and Kittur, A. Augmenting scientific creativity with an analogical search engine. *ACM Transactions on Computer-Human Interaction*, 2022. doi: 10.1145/3530013.

Kokate, R. and Onkar, P. S. Analogical reasoning with large language models: a co-creative framework and benchmarking of llms in design ideation. *Design Science*, 2026. doi: 10.1017/dsj.2025.10044.

Kon, P. T. J., Liu, J., Zhu, X., Ding, Q., Peng, J., Xing, J., Huang, Y., Qiu, Y., Srinivasa, J., Lee, M., Chowdhury, M., Zaharia, M., and Chen, A. Exp-bench: Can ai conduct ai research experiments? *ArXiv*, abs/2505.24785, May 2025. doi: 10.48550/arxiv.2505.24785. URL https://doi.org/10.48550/arxiv.2505.24785.

Li, P., Lin, F., Xing, S., Sun, J., Zhang, D., Yang, S., Ni, C., and Tu, Z. Let the abyss stare back adaptive falsification for autonomous scientific discovery. *ArXiv*, Mar 2026. doi: 10.48550/arxiv.2603.29045. URL https://doi.org/10.48550/arxiv.2603.29045.

Lin, E., Peng, Z., and Fang, Y. Evaluating and enhancing large language models for novelty assessment in scholarly publications. *arXiv preprint*, 2025.

Liu, Z., Liu, K., Zhu, Y., Lei, X., Yang, Z., Zhang, Z., Li, P., and Liu, Y. Aigs: Generating science from ai-powered automated falsification, 2024. URL http://arxiv.org/abs/2411.11910.

Lu, C., Lu, C., Lange, R. T., Foerster, J., Clune, J., and Ha, D. The ai scientist: Towards fully automated open-ended scientific discovery. *arXiv preprint*, 2024.

Martin-Boyle, A., Leckey, C. A. C., Brown, M., and Kaur, H. Papertrail: A claim-evidence interface for grounding provenance in llm-based scholarly q&a. *Proceedings of the 2026 CHI Conference on Human Factors in Computing Systems*, pp. 1–25, Apr 2026. doi: 10.1145/3772318.3791101. URL https://doi.org/10.1145/3772318.3791101.

Musker, S., Duchnowski, A., Millire, R., and Pavlick, E. Llms as models for analogical reasoning. *Journal of Memory and Language*, 2025. doi: 10.1016/j.jml.2025.104676.

O'Neill, C., Ghosal, T., Raileanu, R., Walmsley, M., Bui, T., Schawinski, K., and Ciuca, I. Sparks of science: Hypothesis generation using structured paper data, 2025. URL http://arxiv.org/abs/2504.12976.

Qiu, Y., Zhang, H., Xu, Z., and Li, M. Ai idea bench 2025: Ai research idea generation benchmark. *arXiv preprint*, 2025.

Saraogi, D., Singhee, R., and Kumar, D. Evaluating novelty in ai-generated research plans using multi-workflow llm pipelines. *arXiv preprint*, 2025.

Schmidgall, S., Su, Y., Wang, Z., Sun, X., Wu, J., Yu, X., Liu, J., Moor, M., Liu, Z., and Barsoum, E. Agent laboratory: Using llm agents as research assistants. *arXiv preprint*, 2025.

Si, C., Yang, D., and Hashimoto, T. Can llms generate novel research ideas? a large-scale human study with 100+ nlp researchers. *arXiv preprint*, 2024.

Si, C., Hashimoto, T., and Yang, D. The ideation-execution gap: Execution outcomes of llm-generated versus human research ideas. *arXiv preprint*, 2025.

Spearman, C. The proof and measurement of association between two things. *The American Journal of Psychology*, 15(1):72–101, 1904. doi: 10.2307/1412159.

Trench, M. and Minervino, R. A. Distant connections: The memory basis of creative analogy. *ArXiv*, Jan 2020. ISSN 2192-8363. doi: 10.1007/978-3-030-52545-3. URL https://doi.org/10.1007/978-3-030-52545-3.

Wei, J., Yang, Y.-J., Zhang, X., Chen, Y., et al. From ai for science to agentic science: A survey on autonomous scientific discovery. *arXiv preprint*, 2025.

Wilcoxon, F. Individual comparisons by ranking methods. *Biometrics Bulletin*, 1(6):80–83, 1945. doi: 10.2307/3001968.

Xie, Q., Weng, Y., Zhu, M., Shen, F., Huang, S., Lin, Z., Zhou, J., Mao, Z., Yang, Z., Yang, L., Wu, J., and Zhang, Y. How far are ai scientists from changing the world? *arXiv preprint*, 2025.

Xiong, G., Xie, E., Williams, C. M., Kim, M., Shariatmadari, A. H., Guo, S., Bekiranov, S., and Zhang, A. Toward reliable scientific hypothesis generation: Evaluating truthfulness and hallucination in large language models. Preprint, 2025. URL https://doi.org/10.24963/ijcai.2025/873.

Yamada, Y., Lange, R. T., Lu, C., Hu, S., Lu, C., Foerster, J., Clune, J., and Ha, D. The ai scientist-v2: Workshop-level automated scientific discovery via agentic tree search, 2025. URL http://arxiv.org/abs/2504.08066.

Yuan, S., Chen, J., Ge, X., Xiao, Y., and Yang, D. Beneath surface similarity: Large language models make reasonable scientific analogies after structure abduction. *Findings of the Association for Computational Linguistics: EMNLP 2023*, 2023. doi: 10.18653/v1/2023.findings-emnlp.160.

Zhao, K., Lin, W., Zheng, Q., Xu, F., and Li, Y. Deep ideation: Designing llm agents to generate novel research ideas on scientific concept network, 2025. URL http://arxiv.org/abs/2511.02238.

Zheng, L., Li, C., Zhang, X., Shang, Y.-M., Huang, F., and Jia, H. Evidence retrieval is almost all you need for fact verification. *Findings of the Association for Computational Linguistics*, 2024. doi: 10.18653/v1/2024.findings-acl.551.

Zheng, X., Li, N., Luan, X., Wang, K., Shi, L., Sun, M., and Wang, H. Beyond correctness: Exposing llm-generated logical flaws in reasoning via multi-step automated theorem proving. *ArXiv*, abs/2512.23511, Dec 2025. doi: 10.48550/arxiv.2512.23511. URL https://doi.org/10.48550/arxiv.2512.23511.

Zhu, M., Xie, Q., Weng, Y., Wu, J., Lin, Z., Yang, L., and Zhang, Y. Ai scientists fail without strong implementation capability. *arXiv preprint*, 2025.

A. Prompt Templates

P0: Bottleneck extraction. Input: seed paper, abstract, or problem. Output JSON: {bottleneck, why-hard, constraints, measurable-proxy, closest-known-approaches}.
P1: Distant mechanism mining. Given the bottleneck JSON, list mechanisms outside the target domain. For each: source domain, mechanism, assumptions, transferable invariant, and failure conditions.
P2: Variable-level mapping. Produce a table: source variable, source role, target variable, target role, mapping quality, and mismatch. Label each row exact, approximate, speculative, or mismatch.
P3: Hypothesis generation. Generate a single-sentence claim, a concrete method sketch, a measurable prediction, and a minimal falsification experiment using only structurally useful mappings.
P4: Nearest-prior audit. Retrieve nearest prior work and classify overlap as A known, B incremental, C plausibly novel/high-risk, or D no close prior. Provide the nearest title, overlap, remaining gap, and confidence.
P5: Falsification plan. State the observation that would reject the claim, the baseline, metric, tolerance, data, compute limit, and expected runtime.

B. Extended Statistics

Table 2. Bonferroni-corrected paired Wilcoxon tests for NOVA-Test versus baselines ($n = 12$ bottlenecks).

Axis	Baseline	Mean diff.	p	p_c
Novelty	One-shot	0.875	0.00024	0.00195
Novelty	Lit-search	1.042	0.00049	0.00391
Specificity	One-shot	1.333	0.00024	0.00195
Specificity	Lit-search	1.542	0.00024	0.00195
Falsifiability	One-shot	0.333	0.00391	0.03125
Falsifiability	Lit-search	0.583	0.00195	0.01563
Feasibility	One-shot	0.833	0.00244	0.01953
Feasibility	Lit-search	0.542	0.01074	0.08594

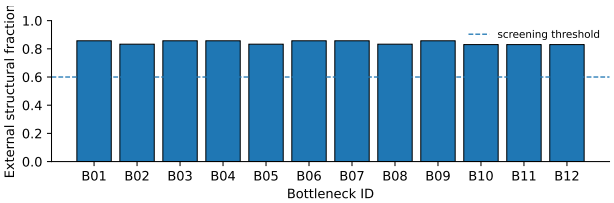

Figure 4. External structural-fraction audit for the 12 top NOVA-Test mappings. All exceed the screening threshold used in the main study.

C. Citation Support Ledger

The bibliography is deliberately broad enough to support the paper's claims without treating any single family of prior work as definitive. LLM ideation and evaluation claims are supported by human studies, benchmarks, and hypothesis-generation datasets (Si et al., 2024; 2025; Guo et al., 2025; Garikaparthi et al., 2025; O'Neill et al., 2025; Xiong et al., 2025; Zhao et al., 2025; Saraogi et al., 2025). Autonomous-research and implementation-failure claims are supported by agent-system papers and external evaluations (Lu et al., 2024; Yamada et al., 2025; Schmidgall et al., 2025; Wei et al., 2025; Beel et al., 2025; Kon et al., 2025; Chen et al., 2026; Zhu et al., 2025; Xie et al., 2025). The structural-analogy mechanism is grounded in computational analogy models, analogical search, cognitive analogy theory, design-by-analogy surveys, and LLM analogy studies (Gentner & Forbus, 2011; Kang et al., 2022; Jiang et al., 2022; Ding et al., 2023; Yuan et al., 2023; Trench & Minervino, 2020; Kokate & Onkar, 2026; Musker et al., 2025). The prior-work and evidence-card gates are supported by novelty-assessment, RAG fact-checking, claim-evidence, and deep-factuality work (Lin et al., 2025; Qiu et al., 2025; Edelman & Skolnick, 2025; Fadeeva et al., 2025; Zheng et al., 2024; Javaji et al., 2025; Martin-Boyle et al., 2026; Huang et al., 2026). The falsification and reviewer-failure framing is supported by automated-falsification and reasoning-audit work (Liu et al., 2024; Dycke & Gurevych, 2026; Li et al., 2026; Zheng et al., 2025).