Before the Action: Benchmarking LLMs on Prospective Hypothesis Discovery

Tianyun Zhong[1,2,*], **Wangyi Jiang**[1,2,*], **Wei Wang**[3], **Xuanang Chen**[2], **Yaojie Lu**[2], **Shiwei Ye**[1], **Yuzhen Shi**[3], **Boyu Yang**[3], **Jinghang Wang**[3], **Han Li**[3], **Weiqi Zhai**[3], **Bing Zhao**[3], **Hu Wei**[3,†], **Haiyang Yu**[3], **Yongbin Li**[3,†], **Hongyu Lin**[2], **Le Sun**[2], **Xianpei Han**[2,†]

[1]**University of Chinese Academy of Sciences**
[2]**Institute of Software, Chinese Academy of Sciences**
[3]**Alibaba Group**

[*]Equal contribution. Work done during internship at Alibaba Group.
zhongtianyun2023@iscas.ac.cn, jiangwangyi2020@iscas.ac.cn
†Corresponding authors.

Abstract

Large language models (LLMs) excel at answering given questions, yet their ability to navigate open-ended, pre-conclusion discovery remains largely unmeasured. We introduce *Prospective Hypothesis Discovery (PHD)*, which asks models to autonomously construct grounded, discriminative, and testable hypothesis spaces from inconclusive evidence, including anomalous observations and fragmented records, to guide subsequent investigation. To evaluate this, we present HypoArena, comprising HypoData (988 cases across six domains) and HypoEval (an evaluation framework for open-ended hypotheses). We construct HypoData via *Retrospective Context Regression*, a pipeline that reconstructs pre-conclusion contexts from expert documents by removing explicit conclusions while preserving factual substrates. Since PHD admits multiple valid outputs, HypoEval combines bidirectional pairwise judgments with Bradley–Terry–Davidson aggregation for ranking, alongside six-dimensional rubric scoring for diagnosis. Experiments on 15 frontier LLMs reveal clear capability stratification and show that arena evaluation resolves finer-grained model differences while aligning strongly with human expert judgment. Together, these results support treating PHD as a distinct target for evaluating the investigative reasoning of LLMs. All code, data, and evaluation tools are publicly available at github.com/SKYLENAGE-AI/HypoArena and huggingface.co/datasets/HypoArena/HypoData.

1 Introduction

Real-world discovery rarely starts from a well-posed question. In scientific research, accident investigation, and financial analysis, the first encounter with a problem is typically a tangle of anomalous observations, fragmented facts, and incomplete records. The defining challenge at this stage is not retrieving a known answer but constructing a plausible hypothesis space: a set of grounded, verifiable, and mutually discriminative claims worth investigating next. We call this capability *Prospective Hypothesis Discovery* (PHD), and argue that it is a distinct and largely unmeasured competence of LLMs. For instance, given sudden server errors and fragmented logs, a capable system should propose several verifiable directions such as memory leaks or

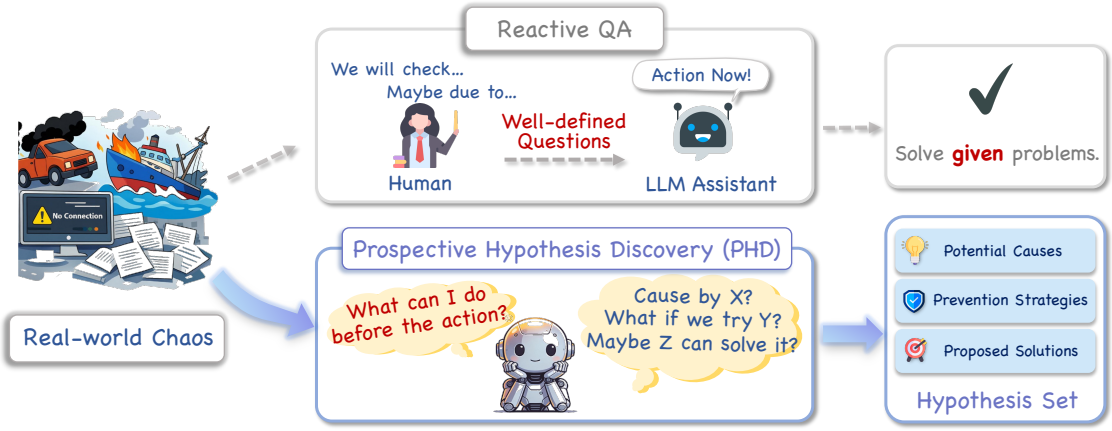

Figure 1 From reactive QA to prospective hypothesis discovery. Conventional QA hands the model a pre-formed question (reactive). HypoArena instead reconstructs a real-world context in pre-conclusion form and asks the model to proactively construct a plausible hypothesis space—the capability we term *Prospective Hypothesis Discovery* (PHD).

microservice deadlocks before any cause has been confirmed. Figure 1 contrasts this setting with conventional QA-style evaluation.

Existing benchmarks do not directly target this capability. QA and agent benchmarks [1–3] score models against predefined goals or task-completion criteria. Scientific discovery and reasoning benchmarks [4–8] typically supply structured data or pre-formulated research questions. Open-ended idea-generation benchmarks [9–12] start from paper abstracts or already-refined research backgrounds. None of these settings asks whether a model, given raw heterogeneous facts and no formed conclusion, can identify several plausible hypotheses, distinguish among them, and argue for each. A benchmark that isolates this question is missing.

Building such a benchmark is hard for two reasons, one on the data side and one on the metric side. On the data side, expert documents such as scientific papers, accident reports, and analyst notes are written after the conclusion is reached. They carry final claims, hypothesis statements, and explicit causal links, all of which are strong suggestive cues. Feeding such documents to a model collapses PHD into restating the known answer, while authoring conclusion-free contexts by hand is expensive and does not scale. On the metric side, a high-quality PHD output is not a single reference but an open hypothesis space [13]. The same context typically supports multiple legitimate directions, so reference-matching metrics underestimate plausible hypotheses that happen to fall outside the gold set [12, 14, 15], while absolute rubric scoring struggles to discriminate among open-ended outputs of comparable surface quality [16].

To address these challenges, we construct HypoArena, a benchmark comprising 988 cases across six domains: Biomedical Science, Machine Learning, Social Science, Financial Analysis, IT Operations, and Safety Investigation. Each case is derived from human-authored documents such as research papers, investigation reports, or professional blogs. Our core motivation stems from the observation that when humans author these reports, they must generate plausible hypotheses, yet these intermediate hypotheses often remain unrecorded in the final artifacts. To approximate this intermediate reasoning stage, we construct our cases through Retrospective Context Regression. Rather than recovering the exact historical information state, this process reconstructs a model-visible context by preserving temporally admissible factual content while filtering explicit conclusions, target hypotheses, and retrospective causal attributions. We operationally refer to contexts satisfying these criteria as "conclusion-free." The corresponding hypotheses and evidence are kept on the reference side for verification and diagnostic evaluation. Table 1 positions HypoArena relative to prior benchmarks in scientific discovery and idea generation.

Since a single "conclusion-free context" often supports multiple plausible exploration directions, high-quality output is no longer a single answer but an open hypothesis space. To address issues such as the underestimation of innovative hypotheses and scoring biases inherent in exact match or rubric scoring, we design HypoEval,

Table 1 Comparison of HypoArena with related benchmarks. We categorize existing benchmarks into three groups: QA & Agents, Scientific Discovery, and Idea Generation. Here, D denotes database and Q denotes query.

Benchmark	Input	Output	Retro. Context	Open Hypo.	Multi-domain	Scale
Category 1: QA & Agents						
DABstep [1]	Q + Data	Str/Num/List	×	×	×	450
DSBENCH [2]	Q + Data	Code / Num	×	×	×	540
Category 2: Sci. Discovery & Reasoning						
DiscoveryBench [4]	D + Goal	Single Hypo.	×	×	×	1.1k
InsightBench [5]	D + Goal	Insights	×	×	×	100
HypoBench [6]	Datasets	Single Hypo.	×	×	✓	194
SciArena [7]	Q + Docs	Lit. Response	×	✓	✓	20K
InnoEval [8]	Idea	Eval Report	✓	×	×	761
Category 3: Idea Generation						
FIRE-Bench [9]	Q	Full Research	✓	✓	×	30
IdeaBench [10]	Abs	Idea	×	✓	×	2.4K
AI Idea Bench	Topic+Papers	Idea	×	✓	×	3.5K
ResearchBench [12]	Q + Survey	Inspiration	×	✓	✓	1.4K
HypoArena (Ours)	**Context**	**Open Hypos**	✓	✓	✓	**988**

an arena-style evaluation protocol based on pairwise comparison. Under the same context, HypoEval moves away from a reliance on a single reference answer. Instead, it compares the hypothesis sets generated by two models to judge which side performs better across six dimensions: Contextual Grounding, Inferential Insight, Evidential Justification, Hypothesis-Space Breadth, Directional Distinctness, and Analytical Utility. Subsequently, we incorporate the Bradley–Terry–Davidson model to aggregate these pairwise judgments and calculate a more robust model ranking. Meanwhile, we retain rubric dimensions as diagnostic tools for analyzing model performance along the same six dimensions.

We conduct systematic empirical evaluations on 15 contemporary language models across six domains. Our HypoEval pairwise evaluation protocol addresses drawbacks of conventional rubric-based scoring. Across domains, arena evaluation produces a clearly stratified leaderboard, while rubric scores remain concentrated within sub-one-point bands on the 1–5 scale. Agent-mode structured analytic skills produce heterogeneous and model-dependent effects. We further assess the reliability of the arena rankings through human-expert alignment and cross-judge triangulation.

Our contributions are:

- **Task Definition & Benchmark**: We define the task of *Prospective Hypothesis Discovery* (PHD) and introduce HypoArena, a benchmark of 988 cases across six scientific and analytical domains.
- **HypoData and Retrospective Context Regression**: We propose a scalable Forge–Audit construction method that reconstructs pre-conclusion contexts from completed expert documents under explicit leakage, temporal, faithfulness, and supportability controls.
- **HypoEval and Systematic Empirical Analysis**: We propose a dual-track evaluation framework that pairs an arena protocol with the Bradley–Terry–Davidson model as the primary ranking, complemented by rubric scoring as diagnostics. A study of 15 contemporary LLMs shows that arena evaluation provides finer-grained separation than rubric scoring and reveals protocol-dependent differences in model ordering.

2 HypoData: Benchmarking Discovery

This section presents HypoData, the foundational data component of HypoArena, designed to systematically evaluate Prospective Hypothesis Discovery (PHD). We formalize hypothesis generation as a unified task across six domains, decoupling the model-visible **Context** from the hidden **Hypothesis** and its associated **Evidence**. Figure 2 illustrates the full data construction pipeline. We formally define the task components and the mathematical formulation below.

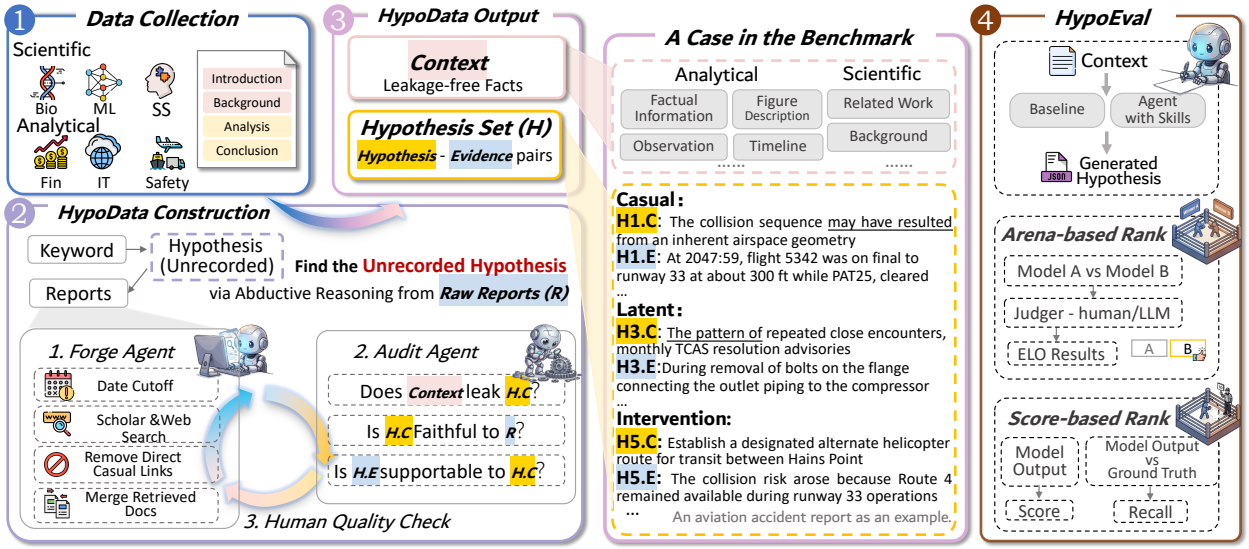

Figure 2 The HypoArena pipeline. (1) Multi-domain data collection from human-authored documents; (2) HypoData: Retrospective Context Regression rolls source documents back to a pre-conclusion factual state and extracts hidden hypothesis-evidence pairs as the reference; (3) Structured Context-Hypothesis-Evidence output, with the Context model-visible and the rest held out; (4) HypoEval: pairwise arena ranking aggregated by Bradley–Terry–Davidson, complemented by rubric scoring as diagnostics.

2.1 Task Definition

Prospective Hypothesis Discovery (PHD) is defined as the mapping from a reconstructed pre-conclusion context $\mathcal{C}$ to a *Hypothesis Set* $\mathcal{H} = \{(h_i, e_i)\}_{i=1}^{K}$. This task requires models to autonomously construct a plausible hypothesis space from inconclusive premises. Unlike traditional QA, which focuses on retrieving a pre-existing answer, PHD requires models to identify unresolved tensions and propose verifiable directions worth investigating next.

Context ($\mathcal{C}$). The context $\mathcal{C}$ approximates the raw, heterogeneous information state prior to conclusion formation, comprising anomalous observations, fragmented facts, and incomplete records. To preserve task validity, $\mathcal{C}$ must satisfy two constraints: **(i)** *informational sufficiency*, meaning it provides enough factual substrate to support non-trivial hypothesis generation; and **(ii)** *operational conclusion absence*, meaning that explicit final conclusions, target hypotheses, and retrospective causal attributions are withheld from the model-visible input. We refer to contexts satisfying the second criterion as operationally "conclusion-free."

Hypothesis Set ($\mathcal{H}$). The output $\mathcal{H}$ consists of (h_i, e_i) pairs, where each pair represents an atomic unit of reasoning:

- **Hypothesis (h_i)**: An explanatory or predictive claim that addresses the anomalies within $\mathcal{C}$. Each h_i must be *grounded* in the provided facts and *verifiable* through subsequent testing. When $K > 1$, the hypotheses should additionally be *mutually discriminative*.
- **Evidence (e_i)**: The domain-specific justification for the hypothesis. In scientific research, this manifests as a *verification plan* (e.g., experimental design); in investigative analysis, it takes the form of a *diagnostic package* (e.g., actionable checks and supporting evidence logs) that substantiates the claim and guides further exploration.

The cardinality K of $\mathcal{H}$ is domain-dependent. Scientific cases elicit a single primary conjecture, emphasizing depth and testability, whereas analytical cases allow an open-cardinality set, emphasizing breadth and differentiation. A single-conjecture case can still admit multiple valid outputs across systems, so the source-derived Reference is not treated as a unique answer.

2.2 Data Collection

To provide broad domain coverage and grounding in expert-authored sources, we curate source documents from six domains across two categories: **Scientific Domains** (Biomedical Science, Machine Learning, Social Science) and **Analytical Domains** (Financial Analysis, IT Operations, Safety Investigation). Domain-specific temporal cutoffs are enforced to minimize data contamination. Detailed collection protocols are provided in Appendix B.

2.3 Data Construction with the Forge–Audit Agent Loop

To balance the trade-off between information leakage and context richness, our framework employs an iterative *Forge–Audit* loop. In this pipeline, the *Forge* agent generates candidate contexts, while the *Audit* agent validates them against benchmark constraints, repeating the cycle until criteria are met or the budget is exhausted. Detailed construction protocols and prompts are provided in Appendix C and Appendix J. The *Forge agent* operates through two sequential components:

Context Forge. The *Context Forge* constructs a model-visible pre-conclusion context $\mathcal{C}$ from source-side factual material and, when appropriate, externally retrieved background. External search is constrained by source-specific timestamp boundaries that exclude materials dated after the relevant publication, release, or incident cutoff. These controls reduce the risk of introducing post-cutoff material through external retrieval. We apply different strategies based on the domain:

- **Scientific Domain:** To avoid reliance on any single, conclusion-bearing document, we retrieve a corpus of literature predating the discovery and use *Document Merging* to synthesize multi-faceted sources into a coherent factual substrate.
- **Analytical Domain:** To preserve raw evidence while stripping analytical bias, we employ *Structural Deconclusion*. This process retains granular facts (e.g., timestamps and measurements) while systematically excising expert verdicts and explicit causal determinations.

Hypothesis Forge. The *Hypothesis Forge* derives a source-grounded reference-side hypothesis set from the completed document while expressing each candidate in a form supportable from the reconstructed context. In scientific domains, this captures a central research conjecture or proposal core; in analytical domains, it yields distinct explanatory or diagnostic claims grounded in the factual record.

Audit Agent. After each Forge round, the *Audit Agent* verifies the candidate against three checks: *Leakage Check* tests whether the context inadvertently reveals the held-out hypothesis, *Faithfulness Check* tests whether the hypothesis remains supported by the source document, and *Supportability Check* tests whether the evidence provides sufficient grounding. Failed checks are returned to the Forge agent as actionable feedback for the next round.

2.4 Human Quality Audit

To validate the fidelity of our automatically constructed instances, we sampled 20 cases from each of the three domains: Biomedical Science, Financial Analysis, and Social Science. For each domain, we recruited 2–3 expert annotators who are either senior doctoral candidates or practicing professionals. Each (Context, Hypothesis, Evidence) triple was evaluated against four criteria: Informativeness and Openness for Context, Completeness for Hypothesis, and Supportiveness for Evidence (detailed in Table 5). Our results in Table 2 show that 92% of the sampled cases passed all criteria, providing evidence for construction quality on the audited subset. The full annotation guidelines are available in Appendix D.

Table 2 Human Quality Audit Results.

Dimension	Pass Rate
(a) Informativeness	95%
(b) Openness	100%
(c) Completeness	100%
(d) Supportiveness	97%
Overall	**92%**

The final benchmark represents each case as a model-visible, reconstructed pre-conclusion **Context** and a reference-side **Hypothesis Set ($\mathcal{H}$)** comprising hypothesis–evidence pairs (h_i, e_i). The reference side is withheld during generation and is used for verification and diagnostic analysis or, in the arena, as an anonymous competitor for calibration. This separation supports evaluating a model's ability to synthesize disparate facts into grounded and verifiable hypotheses without exposing the source-derived Reference during generation.

3 HypoEval: The Evaluation Framework

HypoEval evaluates submitted hypothesis sets against the shared model-visible context. The source-derived Reference is withheld from generation and enters the arena only as an anonymous competitor for calibration, not as a unique gold answer. This separation accommodates the open-ended nature of PHD, where a grounded hypothesis may remain valid even if it is absent from the Reference. Accordingly, HypoEval combines a primary **Rubric-based Arena** for relative ranking with **Rubric-based Scoring** for diagnostic analysis.

3.1 Evaluation Rubrics

Both tracks of HypoEval share a unified set of six metrics, with detailed definitions in Appendix F.2, designed to assess hypothesis set quality across two levels.

- **Pair-level Fidelity**: For each (h_i, e_i) pair, we score *Contextual Grounding*, *Inferential Insight*, and *Evidential Justification*.
- **Set-level Quality**: For submissions in domains that admit multiple pairs, we additionally score *Hypothesis-Space Breadth*, *Directional Distinctness*, and *Analytical Utility*; *Directional Distinctness* is not applicable when $K = 1$.

3.2 Rubric-based Arena (Primary Metric)

To overcome the "reference-matching" bottleneck and the biases of absolute scoring such as score compression and ceiling effects, we adopt a pairwise comparison protocol. In each matchup, an LLM-as-a-judge is presented with a shared conclusion-free context and the hypothesis sets generated by two models, A and B.

- **Position Debiasing.** To mitigate position bias, each matchup is judged twice with A/B assignments swapped. The two verdicts are averaged into a single debiased score, and only matchups where both directions agree in polarity are marked as consistent.
- **Aggregating Pairwise Judgments.** The judge provides a 5-level categorical verdict: $A \gg B$, $A > B$, $A \approx B$, $B > A$, $B \gg A$, where $\gg$ indicates "significantly better" and $\approx$ indicates a tie. These verdicts are mapped to win-shares $w \in \{1.0, 0.75, 0.5, 0.25, 0.0\}$, respectively. To derive a global leaderboard, we aggregate outcomes using the Bradley–Terry–Davidson (BTD) model [17], which models ties as a distinct epistemic state via a tie parameter θ.

3.3 Rubric-based Scoring (Secondary Metric)

We complement the arena protocol with Rubric-based Scoring for diagnostic interpretability. In this track, the judge assigns independent absolute scores on a 1–5 scale to a model's output across the same six dimensions. We aggregate these into pair-level Q_{pair} and set-level Q_{set} scores, with full formulas in Appendix F.2. This granular evaluation serves two purposes:

- **Multi-dimensional Profiling**: It produces per-dimension profiles that identify models strong in *Contextual Grounding* but weak in *Inferential Insight*, informing iterative model development.
- **Complementary Absolute-Scale Context.** It provides an absolute-scale view of output quality, helping distinguish meaningful performance differences from relative advantages within an otherwise uniformly strong or weak model pool.

By synthesizing the discriminative power of competitive ranking with the diagnostic granularity of absolute scoring, HypoEval provides an assessment framework for evaluating discovery-oriented LLMs.

Table 3 Main leaderboard under reference-independent pairwise judging by `seed-2.0-pro`. BTD initializes each model at a baseline rating of 1500 and updates ratings based on pairwise match outcomes. WinRate (WR) denotes each model's win fraction across all pairwise matchups. The Effort column reports each model's thinking-effort setting. Reference participates as an anonymous competitor for calibration.

			Scientific Domains			Analytical Domains				
#	Model	Effort	Bio	ML	Social	Fin	IT	Safety	Avg	WR
1	claude-sonnet-4.6	high	1601.2	1579.3	1582.7	**1739.2**	**1699.8**	**1721.6**	**1654.0**	71.3%
2	claude-opus-4.6	high	1608.9	1618.9	**1671.5**	1670.3	1678.3	1668.4	1652.7	71.4%
3	gpt-5.4	high	**1627.2**	**1650.6**	1602.3	1636.2	1641.2	1630.7	1631.4	70.7%
4	reference		1568.8	1561.7	1502.2	1631.1	1604.8	1674.3	1590.5	57.8%
5	kimi-k2.6	✓	1570.5	1562.4	1559.6	1601.0	1569.1	1633.5	1582.7	58.2%
6	glm-5.1	✓	1592.8	1595.0	1591.6	1563.7	1588.3	1560.0	1581.9	59.1%
7	deepseek-v4-pro	high	1525.0	1520.1	1564.8	1533.2	1517.5	1551.4	1535.4	46.1%
8	qwen-3.6-max	✓	1499.1	1518.6	1503.6	1513.2	1499.6	1520.0	1509.0	42.2%
9	minimax-m2.7	✓	1508.3	1505.2	1492.5	1517.2	1516.1	1453.9	1498.9	37.0%
10	deepseek-v4-flash	high	1464.3	1522.1	1513.8	1493.8	1424.6	1489.3	1484.6	36.3%
11	glm-5	✓	1469.1	1450.2	1452.6	1485.0	1483.5	1445.2	1464.3	32.8%
12	minimax-m2.5	✓	1458.4	1428.9	1427.4	1424.8	1433.4	1369.7	1423.8	22.9%
13	gpt-5.4-mini	high	1424.0	1437.4	1393.7	1440.7	1397.7	1437.7	1421.9	23.7%
14	gemini-3.1-pro	high	1429.9	1407.5	1415.7	1335.7	1388.2	1387.2	1394.0	20.7%
15	gemini-3-flash	high	1420.0	1422.2	1418.2	1250.7	1301.0	1262.1	1345.7	14.7%
16	kimi-k2.5	✓	1281.8	1246.5	1322.7	1266.1	1330.3	1289.4	1289.5	9.2%

4 Experiments

This section evaluates HypoArena along three axes: (i) the baseline leaderboard and the conditional effect of structured analytic skills (§4.2); (ii) arena vs. rubric scoring (§4.4); (iii) alignment with human annotation preference and external quality (§4.3).

4.1 Experimental Setup

We evaluate 15 contemporary LLMs on all 988 HypoData cases under two generation modes. These models span frontier closed-source and open-weight families, including `claude-sonnet-4.6`, `claude-opus-4.6`, and `gpt-5.4`. The full model list appears in Table 3.

The primary evaluation follows the arena protocol (§3.2) using `seed-2.0-pro` as judge. The source-derived Reference participates as an anonymous competitor for calibration. Cross-judge consistency with `mimo-v2-pro` is verified in Appendix I.

- **Baseline Mode**: The model generates its final Hypothesis Set directly from the input context in a single pass without invoking any explicit analytic skills. This mode serves as a zero-shot control, producing a concise single hypothesis–evidence pair for scientific domains, whereas for analytical domains, it yields an open-cardinality set of plausible hypotheses, serving as a performance floor for comparison.
- **Agent Mode**: To mimic professional inquiry, models in this mode can select and execute a sequence from a library of 12 structured analytic skills (see Appendix K for the full library). These skills are adapted from the methodology of *Structured Analytic Techniques for Intelligence Analysis* [18], including techniques such as the Analysis of Competing Hypotheses (ACH), Structured Brainstorming, and Chronology Analysis. Selected skills are composed into a sequential pipeline, where intermediate analytical artifacts are passed forward to refine the final submission.

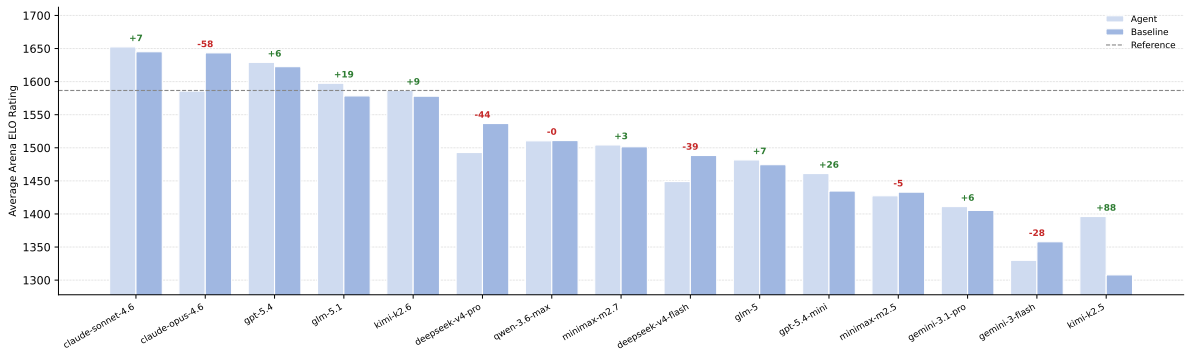

Figure 3 Per-model Agent vs. Baseline ratings on the same full-pool scale, sorted by baseline strength. Annotations show Δ = Agent − Baseline (green positive, red negative).

4.2 Main Results

Clear performance stratification. As shown in Table 3, the baseline leaderboard spans more than 360 BTD points, producing clear separation across the model pool. Three frontier models, `claude-sonnet-4.6`, `claude-opus-4.6`, and `gpt-5.4`, form a dominant top tier, leading across most domains. The protocol also resolves large within-family differences; for example, `kimi-k2.6` outscores `kimi-k2.5` by nearly 300 BTD points.

Reference as a domain-asymmetric baseline. The source-derived Reference outperforms most models under the primary judge but falls below the top frontier tier. Its strength is domain-dependent: it is more competitive in procedure-heavy domains such as Safety Investigation and Financial Analysis, but ranks lower in Social Science. This indicates that a fixed source-derived baseline is more competitive in domains with established procedural structure than in those admitting multiple analytical trajectories.

Model-Dependent Effects of Structured Analytic Skills. As illustrated in Figure 3, Agent mode does not provide uniform gains. The shifts are widely dispersed, ranging from an 88-point gain for `kimi-k2.5` to a nearly 60-point decline for `claude-opus-4.6`, and show little monotonic association with baseline strength (Spearman's $\rho = -0.10$). These results indicate that structured analytic skills act as a model-dependent intervention, with candidate compression as one observed failure mode (Appendix E).

4.3 Alignment with Human and External Signals

Human Alignment and Judge Robustness. To validate HypoEval, domain experts and our primary judge independently evaluated 1,500 pairwise comparisons. As visualized in Figure 4, the resulting rankings exhibit strong global alignment (Kendall's $\tau = 0.90$, Spearman's $\rho = 0.98$). This consistency confirms that our automated arena protocol serves as a high-fidelity proxy for expert judgment. While we observe some per-domain variance, ranging from $\tau = 0.97$ in Financial Analysis to $\tau = 0.53$ in Biomedical Science due to interdisciplinary noise, the top-tier hierarchy remains identical across both evaluators. See Appendix H.1 for the full leaderboard.

Association with Peer-Review Outcomes. The Machine Learning subset was sampled from ICLR 2026 and stratified by acceptance outcome, an external signal not used by the arena judge. As detailed in Appendix H.2, the source-derived Reference is more competitive on accepted than rejected cases across all three

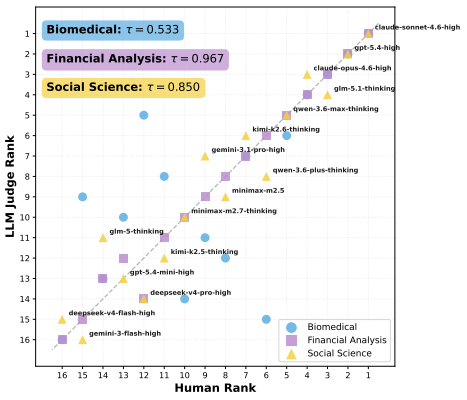

Figure 4 Human vs. LLM Ranking Alignment. Expert and primary-judge rankings show high overlap (cf. Table 10).

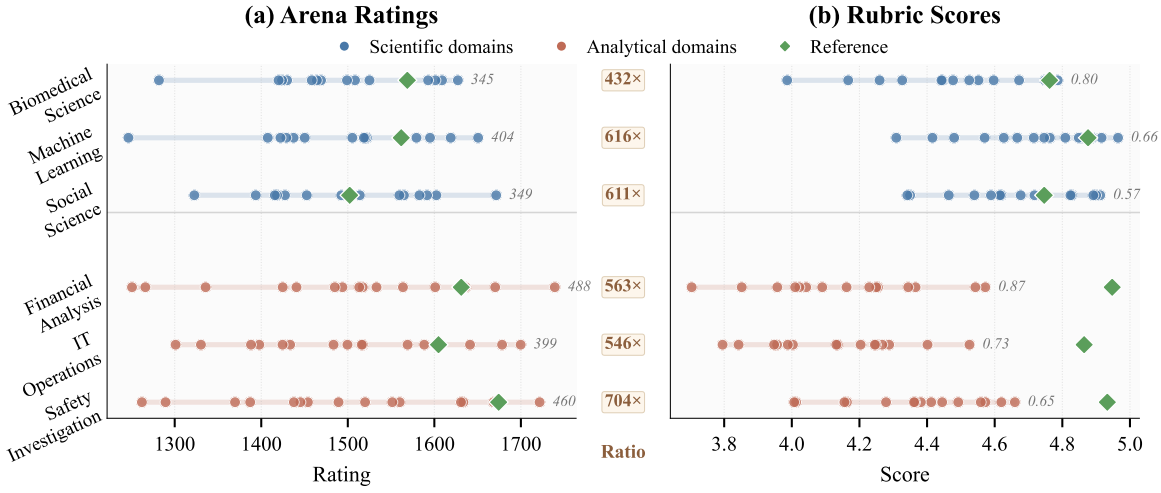

Figure 5 Resolution contrast: Arena scaling vs. Rubric score compression. While the arena protocol (left axis) provides high discriminative resolution across hundreds of points, traditional rubric scores (right axis) exhibit extreme **concentration**, with most models clustered within a **narrow 1-point band**.

reported measures. Its bucket-specific BTD rating is 47 points higher, while the median per-case debiased score and win rate are higher by 0.06 and 0.04, respectively. This consistent directional pattern provides a complementary external check without treating acceptance outcome as ground truth for case-level hypothesis quality. This convergence with an independent quality signal supports the external validity of our protocol.

4.4 Pairwise Arena versus Direct Scoring

Score compression and ceiling effect. Both protocols agree on coarse ordering but diverge sharply at fine-grained resolution. As Figure 5 shows, rubric scoring produces tighter model distributions: arena spreads range 345–490 points per domain, whereas rubric spreads stay under one point on the 1–5 scale, a compression ratio of 430×–725×. This compression is driven by a pronounced ceiling effect: in scientific domains, 95%–98% of assessments land at or above 4.0. Absent a single canonical answer, judges find it easier to declare one submission better in a head-to-head comparison than to commit to a low absolute score.

Protocol-dependent ranking differences. Score compression is accompanied by changes in fine-grained ordering, with per-domain Kendall τ between arena and rubric rankings ranging from 0.52 to 0.82. In the baseline pool, the Reference rises by three positions under rubric scoring in IT Operations, Financial Analysis, and Social Science, while `claude-opus-4.6` falls from its top-two arena placement in each of these domains. These shifts demonstrate sensitivity to the evaluation protocol, but rank changes alone do not identify the underlying mechanism. Together with the score compression and ceiling effects described above, the protocol sensitivity motivates using arena comparisons for primary ranking while retaining rubric scores for diagnostics.

5 Related Work

5.1 Benchmarks for Discovery-Oriented Tasks

Recent benchmark-oriented work has begun to study scientific discovery and idea generation more systematically, moving beyond early generation systems toward explicit task formulation and standardized evaluation. Recent examples include inspiration-based scientific discovery benchmarks [12], principled benchmarking for hypothesis generation [6], research-idea generation benchmarks [11], and domain-specific fine-grained rediscovery settings [19]. Although these works differ in input format and scope, they collectively shift the field from demonstrating that LLMs can generate ideas or hypotheses to asking how such capabilities should

be benchmarked, compared, and stress-tested. In parallel, another line of work has improved hypothesis-generation systems through literature-grounded or hybrid pipelines, while beginning to examine reliability issues such as grounding [20], truthfulness [21], and hallucination in generated scientific hypotheses [22]. Relative to this literature, HypoArena centers on a particular combination of reconstructed pre-conclusion contexts under temporal and leakage controls, open-ended hypothesis generation without a unique reference answer, and unified coverage of scientific and analytical domains.

5.2 Evaluation for Open-Ended Scientific Outputs

Recent work has increasingly recognized that evaluating open-ended scientific outputs requires task-specific and knowledge-grounded protocols. Recent examples include arena-style evaluation platforms for non-verifiable scientific tasks [7], multi-perspective frameworks for research-idea evaluation [8], graph-based evaluators for idea assessment [23], and work probing how grounded LLM critiques of scientific papers actually are [24]. Collectively, these efforts move beyond treating scientific outputs as ordinary text-generation targets and instead emphasize evaluator reliability, grounding, and structured comparison. HypoArena addresses these by introducing *Retrospective Context Regression* to filter explicit conclusion-bearing and retrospective causal content from source documents and HypoEval without relying on a single gold-standard reference.

6 Conclusion

We introduce HypoArena, a benchmark for *Prospective Hypothesis Discovery (PHD)* that uses Retrospective Context Regression to reconstruct pre-conclusion discovery tasks from approximately 1,000 completed expert documents. Our HypoEval framework employs a pairwise arena and the Bradley–Terry–Davidson model to robustly evaluate open-ended hypothesis sets. Systematic testing of 15 frontier LLMs reveals heterogeneous, model-dependent effects of structured analytical skills, while qualitative generation traces identify candidate compression as a possible failure mode.

References

[1] Alex Egg, Martin Iglesias Goyanes, Friso Kingma, Andreu Mora, Leandro von Werra, and Thomas Wolf. Dabstep: Data agent benchmark for multi-step reasoning. arXiv preprint arXiv:2506.23719, 2025.

[2] Liqiang Jing, Zhehui Huang, Xiaoyang Wang, Wenlin Yao, Wenhao Yu, Kaixin Ma, Hongming Zhang, Xinya Du, and Dong Yu. Dsbench: How far are data science agents from becoming data science experts? arXiv preprint arXiv:2409.07703, 2024.

[3] Ludovico Mitchener, Angela Yiu, Benjamin Chang, Mathieu Bourdenx, Tyler Nadolski, Arvis Sulovari, Eric C Landsness, Daniel L Barabasi, Siddharth Narayanan, Nicky Evans, et al. Kosmos: An ai scientist for autonomous discovery. arXiv preprint arXiv:2511.02824, 2025.

[4] Bodhisattwa Prasad Majumder, Harshit Surana, Dhruv Agarwal, Bhavana Dalvi Mishra, Abhijeetsingh Meena, Aryan Prakhar, Tirth Vora, Tushar Khot, Ashish Sabharwal, and Peter Clark. Discoverybench: Towards data-driven discovery with large language models. In The Thirteenth International Conference on Learning Representations, 2025.

[5] Gaurav Sahu, Abhay Puri, Juan A. Rodriguez, Amirhossein Abaskohi, Mohammad Chegini, Alexandre Drouin, Perouz Taslakian, Valentina Zantedeschi, Alexandre Lacoste, David Vazquez, Nicolas Chapados, Christopher Pal, Sai Rajeswar, and Issam H. Laradji. Insightbench: Evaluating business analytics agents through multi-step insight generation. In The Thirteenth International Conference on Learning Representations, 2025.

[6] Haokun Liu, Sicong Huang, Jingyu Hu, Yangqiaoyu Zhou, and Chenhao Tan. Hypobench: Towards systematic and principled benchmarking for hypothesis generation, 2026.

[7] Yilun Zhao, Kaiyan Zhang, Tiansheng Hu, Sihong Wu, Ronan Le Bras, Yixin Liu, Xiangru Tang, Joseph Chee Chang, Jesse Dodge, Jonathan Bragg, Chen Zhao, Hannaneh Hajishirzi, Doug Downey, and Arman Cohan. Sciarena: An open evaluation platform for non-verifiable scientific literature-grounded tasks. In The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track, 2025.

[8] Shuofei Qiao, Yunxiang Wei, Xuehai Wang, Bin Wu, Boyang Xue, Ningyu Zhang, Hossein A. Rahmani, Yanshan Wang, Qiang Zhang, Keyan Ding, Jeff Z. Pan, Huajun Chen, and Emine Yilmaz. Innoeval: On research idea evaluation as a knowledge-grounded, multi-perspective reasoning problem, 2026.

[9] Zhen Wang, Fan Bai, Zhongyan Luo, Jinyan Su, Kaiser Sun, Xinle Yu, Jieyuan Liu, Kun Zhou, Claire Cardie, Mark Dredze, et al. Fire-bench: Evaluating agents on the rediscovery of scientific insights. arXiv preprint arXiv:2602.02905, 2026.

[10] Sikun Guo, Amir Hassan Shariatmadari, Guangzhi Xiong, Albert Huang, Myles Kim, Corey M. Williams, Stefan Bekiranov, and Aidong Zhang. Ideabench: Benchmarking large language models for research idea generation. In Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2, KDD '25, page 5888–5899, New York, NY, USA, 2025. Association for Computing Machinery.

[11] Yansheng Qiu, Haoquan Zhang, Zhaopan Xu, Ming Li, Diping Song, Zheng Wang, and Kaipeng Zhang. Ai idea bench 2025: Ai research idea generation benchmark, 2025.

[12] Yujie Liu, Zonglin Yang, Tong Xie, Jinjie Ni, Ben Gao, Yuqiang Li, Shixiang Tang, Wanli Ouyang, Erik Cambria, and Dongzhan Zhou. Researchbench: Benchmarking llms in scientific discovery via inspiration-based task decomposition, 2025.

[13] Chenglei Si, Diyi Yang, and Tatsunori Hashimoto. Can llms generate novel research ideas? a large-scale human study with 100+ nlp researchers. arXiv preprint arXiv:2409.04109, 2024.

[14] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In Proceedings of the 40th annual meeting of the Association for Computational Linguistics, pages 311–318, 2002.

[15] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In Text summarization branches out, pages 74–81, 2004.

[16] Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. G-eval: Nlg evaluation using gpt-4 with better human alignment. In Proceedings of the 2023 conference on empirical methods in natural language processing, pages 2511–2522, 2023.

[17] Roger R Davidson. On extending the bradley-terry model to accommodate ties in paired comparison experiments. Journal of the American Statistical Association, 65(329):317–328, 1970.

[18] Randolph H Pherson and Richards J Heuer Jr. Structured analytic techniques for intelligence analysis. Cq Press, 2019.

[19] Zonglin Yang, Wanhao Liu, Ben Gao, Yujie Liu, Wei Li, Tong Xie, Lidong Bing, Wanli Ouyang, Erik Cambria, and Dongzhan Zhou. Moose-chem2: Exploring llm limits in fine-grained scientific hypothesis discovery via hierarchical search, 2025.

[20] Haokun Liu, Yangqiaoyu Zhou, Mingxuan Li, Chenfei Yuan, and Chenhao Tan. Literature meets data: A synergistic approach to hypothesis generation. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, editors, Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 245–281, Vienna, Austria, July 2025. Association for Computational Linguistics.

[21] Rosni Vasu, Chandrayee Basu, Bhavana Dalvi Mishra, Cristina Sarasua, Peter Clark, and Abraham Bernstein. HypER: Literature-grounded hypothesis generation and distillation with provenance. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng, editors, Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 25413–25438, Suzhou, China, November 2025. Association for Computational Linguistics.

[22] Guangzhi Xiong, Eric Xie, Corey Williams, Myles Kim, Amir Hassan Shariatmadari, Sikun Guo, Stefan Bekiranov, and Aidong Zhang. Toward reliable scientific hypothesis generation: Evaluating truthfulness and hallucination in large language models. In James Kwok, editor, Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence, IJCAI-25, pages 7849–7857. International Joint Conferences on Artificial Intelligence Organization, 8 2025. Main Track.

[23] Tao Feng, Yihang Sun, and Jiaxuan You. Grapheval: A lightweight graph-based LLM framework for idea evaluation. In The Thirteenth International Conference on Learning Representations, 2025.

[24] Jiefu Ou, William Walden, Kate Sanders, Zhengping Jiang, Kaiser Sun, Jeffrey Cheng, William Jurayj, Miriam Wanner, Shaobo Liang, Candice Morgan, Seunghoon Han, Weiqi Wang, Chandler May, Hannah Recknor, Daniel

Khashabi, and Benjamin Van Durme. CLAIMCHECK: How grounded are LLM critiques of scientific papers? In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng, editors, Findings of the Association for Computational Linguistics: EMNLP 2025, pages 21712–21735, Suzhou, China, November 2025. Association for Computational Linguistics.

Appendix

A Limitations

Despite its comprehensive framework, HypoArena has several limitations that open avenues for future research. First, while our benchmark spans six diverse domains, it remains primarily text-centric. Expanding HypoArena to include multimodal or structured-data-driven discovery settings, such as genomics and medical imaging, represents a critical next step. Second, the current construction pipeline relies on a single model (`gpt-5.4`) for the *Forge* process. Future iterations could incorporate a more diverse ensemble of construction models or human-authored cases to enhance stylistic variety and mitigate potential model-specific biases. Third, although we verified evaluation consistency with a second judge, the reliability of our framework could be further strengthened by employing a broader judge panel or a hybrid human-LLM scoring system. Finally, the *Agent Mode* currently operates within a fixed library of analytic skills. Moving toward a more flexible paradigm where models can dynamically compose or define novel strategies would better reflect the fluid nature of human scientific inquiry and investigative reasoning.

B Details of Data Collection

To provide broad domain coverage and grounding in expert-authored sources, HypoData spans six domains organized into two high-level categories: *Scientific Domains* (Biomedical Science, Machine Learning, and Social Science) and *Analytical Domains* (Financial Analysis, IT Operations, and Safety Investigation). Across all domains, we prioritize publicly accessible and authoritative source materials that are rich enough to support context reconstruction and source-grounded construction of reference-side hypothesis–evidence pairs. Where temporal control is important, we further restrict collection to recent artifacts or apply domain-appropriate release-date constraints to reduce contamination while preserving sufficient source diversity. Unless otherwise noted, each collected source artifact is processed into a single final benchmark case, whose reference side may contain one or more hypothesis–evidence pairs.

Biomedical Science. Biomedical Science sources are primary research articles retrieved from PubMed[1] via a systematic query pipeline, restricted to open-access full texts published between July 2025 and April 2026 from high-impact journals including *Nature Communications*, *Cell Reports*, *eLife*, *Nucleic Acids Research*, *PLOS Biology*, *PNAS*, and *Science Advances*. We retain only empirical papers with explicit hypothesis framing, excluding reviews, case reports, clinical trials, and letters, yielding 244 benchmark cases.

Machine Learning. Machine Learning draws on ICLR 2026[2] submissions with released final decisions. To diversify research maturity and benchmark difficulty, we stratify by decision outcomes, yielding 218 cases. Decision categories shape sampling only and are not used as supervision targets.

Social Science. Social Science comprises 163 articles from seven journals: *Journal of Marketing Research*[3], *Journal of Retailing and Consumer Services*[4], *Journal of International Business Studies*[5], *Journal of Business Research*[6], *Journal of Consumer Research*[7], *Journal of Consumer Psychology*[8], and *Annals of Tourism Research*[9], restricted to 2025–2026 publications. Each paper is verified by domain experts to contain explicit hypothesis-bearing argumentation suitable for context reconstruction.

[1] https://pubmed.ncbi.nlm.nih.gov
[2] https://openreview.net/group?id=ICLR.cc/2026
[3] https://onlinelibrary.wiley.com/journal/15206793
[4] https://www.sciencedirect.com/journal/journal-of-retailing-and-consumer-services
[5] https://link.springer.com/journal/41267
[6] https://www.sciencedirect.com/journal/journal-of-business-research
[7] https://onlinelibrary.wiley.com/journal/1745459x
[8] https://myscp.onlinelibrary.wiley.com/journal/15327663
[9] https://www.sciencedirect.com/journal/annals-of-tourism-research

Table 4 Statistics and source overview of HypoData. Each hypothesis–evidence (H-E) pair is the atomic evaluation unit. Scientific domains yield focused, single-direction hypotheses; real-world domains yield multi-directional hypothesis sets. Detailed collection protocols are in Appendix B.

Domain	Primary Source	# Cases	# H-E	# Category
Research				
Biomedical Science	Journal Articles	244	244	
Machine Learning	Conference Papers	218	218	
Social Science	Journal Articles	163	163	
Real-World				
Financial Analysis	SEC 10-Q + Analyst Reports	114	456	3.99
IT Operations	Post-mortems Blogs	146	527	3.29
Safety Investigation	CSB & NTSB Reports	103	404	3.87
Total		**988**	**2,012**	**3.67**

Financial Analysis. Financial Analysis instances are anchored on the quarterly disclosure cycle of large-cap U.S.-listed companies, combining primary 10-Q filings (2025–2026) sourced from the SEC[10] with matched publicly accessible professional analyses. Context is reconstructed primarily from the filing, while paired analyses supply hypothesis-level interpretation and supporting evidence. Cases are retained only when the quarterly disclosure is substantive and the analytical coverage is deep, yielding 114 cases.

IT Operations. IT Operations sources are public post-mortem reports from widely referenced incident repositories[11] and engineering blogs of major technology companies. Unlike other domains, no strict temporal cutoff is imposed: post-mortem reports describe unique incident configurations and system-specific failure chains that are unlikely to appear verbatim in model training data, making temporal contamination a lower risk than in publication-based domains. We retain technically detailed cases with explicit descriptions of failure mechanisms, contributing factors, or remediation-relevant reasoning, yielding 146 cases.

Safety Investigation. Safety Investigation cases are drawn from official U.S. government accident investigation reports, including transportation accident reports[12] and industrial chemical incident investigations[13], restricted to reports released in 2025 or later. We retain only cases with detailed event narratives and causal analyses to support mechanistic reasoning, yielding 103 cases.

C Details of Data Construction

This appendix provides the full implementation details of the two-stage Forge–Audit pipeline used to construct HypoData (Section 2).

C.1 Forge–Audit Pipeline

Each benchmark instance is constructed through an iterative multi-agent pipeline. A coordinating controller dispatches specialized Forge and Audit agents across both construction stages. The Forge agent generates candidate outputs (contexts in Stage 1; hypothesis–evidence pairs in Stage 2), while the Audit agent evaluates whether the candidates satisfy benchmark requirements. If the candidate passes audit, construction terminates; otherwise, the Audit agent returns explicit and actionable feedback, which is used as direct input to the next Forge round. Construction proceeds through repeated cycles until the audit passes or a fixed iteration budget is reached. Under this protocol, each collected source artifact is processed exactly once and yields a single final benchmark instance.

[10] https://www.sec.gov/search-filings
[11] https://github.com/danluu/post-mortems
[12] https://www.ntsb.gov/investigations/AccidentReports/Pages/Reports.aspx
[13] https://www.csb.gov/news/incident-report-rule-form-

C.2 Stage 1: Context Construction under Temporal Control

Domain-Specific Construction Policies. Context construction follows different default policies across domains. In scientific domains, where source papers often provide limited background and related work, the Forge agent may enrich the context through temporally controlled search over academic and web sources. In real-world domains, where source materials are typically more information-dense, context is constructed primarily from the source materials themselves, with minimal or no external augmentation. In these domains, the construction process places greater emphasis on preserving source-side factual detail and, when possible, the original writing texture.

Strict Temporal Cutoff. When external retrieval is used, all search tools are constrained by source-specific timestamp boundaries so that only information available prior to the publication, release, or incident date of the source materials can be accessed. This reduces leakage from future knowledge and helps preserve the temporal realism of the original reasoning setting.

Dynamic Context Reconstruction. Directly extracting localized arguments from the source materials often yields overly prescriptive inputs and leads to homogenized hypothesis generation. To preserve exploratory reasoning, the Forge agent reconstructs context by combining source-side factual content with temporally admissible external background when appropriate. During this process, it removes explicit hypothesis statements, conclusion-side claims, citation anchors, and other cues that would reveal the original line of reasoning. The resulting context is designed to be information-rich, multidimensional, coherent, and pre-conclusion in character, while avoiding answer-side guidance.

C.3 Stage 2: Hypothesis and Evidence Extraction

While the main text characterizes the *target* of Stage 2 along two patterns—vertical deepening and horizontal association—the *extraction method* varies by source structure.

Hypothesis Extraction Strategies. Because hypothesis presentation varies substantially across domains, the Forge agent applies one of three strategies based on the full source materials:

- **Direct Extraction:** When the source explicitly states a single central hypothesis, it is extracted with minimal modification.
- **Synthesis and Rewriting:** When multiple distributed sub-hypotheses appear across the source, they are consolidated into a unified core claim.
- **Latent Distillation:** When no explicit hypothesis is stated, the underlying claim is inferred from the source's methods, results, and conclusions.

Evidence Construction. Each evidence field e_i provides the domain-specific support associated with hypothesis h_i. As an umbrella field, it may take either prospective form (verification sketches in scientific domains) or retrospective form (evidence packages in real-world domains). By constructing this field explicitly, we require benchmark targets to be not only plausible but also supportable.

C.4 Input–Reference Boundary Enforcement

Throughout construction, we strictly separate the model-input side from the reference side of each benchmark case. The context must not leak final conclusions, posterior explanations, or direct restatements of the source's claims; all hypothesis–evidence pairs remain exclusively on the reference side. This boundary is enforced explicitly in the Audit loop and is central to maintaining the benchmark's pre-conclusion setting.

D Human Annotation Details

To assess the construction quality of HypoData and the alignment of our automated evaluator with expert preferences, we designed two distinct annotation tasks, **Quality Audit** and **Preference Evaluation**. All tasks were

performed in English by domain experts. This section describes the annotation guidelines and setup.

D.1 General Experimental Setup

The annotation setup was as follows.

- **Expertise**. We recruited 2–3 annotators per domain, consisting of senior Ph.D. candidates or practicing professionals in Biomedical Science, Financial Analysis, and Social Science.
- **Anonymization**. To reduce identity- and provenance-related bias, model identities and reference provenance were hidden from annotators where applicable.
- **Data Scale**. Contexts averaged ∼3,000 tokens, while Hypothesis–Evidence pairs averaged ∼1,000 tokens.

D.2 Quality Audit

Objective. Assess whether the automatically constructed (Context, Hypothesis, Evidence) triples provide informative and open-ended contexts, complete source-grounded hypotheses, and supportive evidence.

Input Format. Annotators are presented with a Context and its corresponding (Hypothesis, Evidence) tuples. In scientific domains, each case contains a single tuple; in real-world analytical domains, multiple tuples may be provided.

Evaluation Criteria. Annotators perform four binary checks across three core dimensions (detailed in Table 5).

A key qualitative metric is Context Openness: This measures whether the Context allows for multiple plausible analytical directions. While a "No" (indicating a singular reasonable direction) is acceptable for specific deterministic cases, a high "Yes" rate across the dataset is preferred to ensure the benchmark's challenge level.

D.3 Preference Evaluation

Objective. Collect pairwise human preference judgments as an expert reference signal for assessing the alignment of our automated LLM-as-a-judge system.

Input Format. For each instance, annotators are provided with:

1. **Shared Context**. The conclusion-free background information provided to both models.
2. **Model Outputs**. Two candidate sets of hypotheses and evidence (Output A and Output B), randomized and stripped of model metadata.
3. **Evaluation Metrics**. The standardized rubric used for assessment (see Appendix F.2).

Labeling Requirements. For each pairwise matchup, annotators must provide:

- **Preference**. Select from A $>$ B, Tie, B $>$ A.
- **Confidence Score**. Annotators rated their confidence as High, Medium, or Low.

D.4 Results of Human Annotation

Results of the Quality Audit and Preference Evaluation are reported in Table 6 and Appendix H.1, respectively.

E Generation-Side Diagnostics

Qualitative inspection of the generation-side logs suggests three recurrent distinctions among systems that may help explain why models appearing similarly plausible on the surface can diverge substantially under the primary evaluation protocol.

Table 5 Summary of Quality Audit Criteria. These questions guide human annotators in assessing whether the synthesized Context, Hypothesis, and Evidence meet the intended quality standards.

ID	Criterion	Dim.	Core Audit Question
(a)	Informativeness	Context	Does the context provide sufficient background and key facts to guide hypothesis generation?
(b)	Openness	Context	Can the context elicit diverse and plausible alternative hypotheses beyond the target?
(c)	Completeness	Hypothesis	Does the hypothesis capture the core claims originally presented in the source document?
(d)	Supportiveness	Evidence	Does the evidence provide a sound verification plan or factual basis for the hypothesis?

Table 6 Human Quality Audit Results. Criteria (a–d) correspond to Informativeness, Openness, Completeness, and Supportiveness respectively. The **92%** overall pass rate validates the robustness of our HypoData construction.

Domain	a	b	c	d	Pass Rate
Biomedical	85%	100%	100%	95%	80%
Financial	100%	100%	100%	100%	100%
Social Science	100%	100%	100%	95%	95%
Overall	**95%**	**100%**	**100%**	**97%**	**92%**

Coverage breadth. In multi-pair domains, the strongest systems do not simply produce more pairs; they produce more meaningfully distinct pairs. Their advantage comes from broader exploration of the supported hypothesis space while maintaining non-redundancy and local quality. Weaker systems often generate only a small set of familiar directions, even when the context supports a wider range of mechanistic, latent, or intervention-oriented possibilities.

Candidate compression. In several inspected traces, particularly under Agent Mode, systems generate rich intermediate analyses but compress them into overly narrow final submissions. These systems identify multiple plausible lines of reasoning internally yet fail to preserve them in the submitted set. This possible failure mode is especially costly in open-ended investigative settings, where omitted candidates may correspond to important supported directions rather than merely stylistic alternatives.

Mechanism granularity. Strong systems formulate hypotheses at an appropriate level of explanatory resolution: they identify sharper mechanisms, separate distinct causal pathways, and attach concrete support to each direction. Weaker systems often conflate multiple mechanisms into a single diffuse hypothesis or remain at a level of generality too broad to be meaningfully testable. Such differences may be muted under direct scoring but become highly visible under arena comparison and set-level evaluation.

Taken together, these observations suggest that failure in deep hypothesis generation is not always reducible to obvious factual mistakes or low fluency. It may instead reflect insufficient breadth, insufficient decomposition, or loss of structure between intermediate reasoning and the final submitted set.

F Details of Evaluation

F.1 Submission Cardinality

HypoEval does not impose a hard cap on submission size in multi-pair domains. This reflects the fact that breadth of supported hypothesis exploration is itself part of the target capability. At the same time, larger submissions are not automatically preferred: additional pairs improve evaluation only insofar as they

expand meaningful coverage without reducing distinctness or local quality. All formal evaluation is therefore conducted on the final submitted set.

F.2 Details of the Rubric

Quality Dimensions. Evaluation operates at two granularities to handle both local plausibility and global coherence. At the *pair level*, each submitted $(\hat{h}_i, \hat{e}_i)$ is assessed on three dimensions:

- *Contextual Grounding*: whether the hypothesis is anchored in specific facts, observations, or unresolved tensions in $\mathcal{C}$, and whether the inference is warranted—not merely topically relevant.
- *Inferential Insight*: whether the hypothesis goes beyond surface restatement by synthesizing dispersed information into a non-obvious, context-constrained explanatory, predictive, or mechanistic claim.
- *Evidential Justification*: whether the evidence concretely and proportionately supports the hypothesis, with a coherent reasoning chain and without overclaiming.

These dimensions operationalize the requirement that hypotheses be grounded, non-trivial, and supportable (Section 2). When $K > 1$, the submission is additionally evaluated at the *set level*:

- *Hypothesis-Space Breadth*: coverage of multiple distinct, context-supported directions such as mechanisms, risk axes, or causal pathways.
- *Directional Distinctness*: non-redundancy across hypotheses—not overlapping or trivially reformulated versions of the same claim.
- *Analytical Utility*: actionability for downstream investigation, testing, or prioritization.

Set-level evaluation is necessary because a locally plausible but narrow or repetitive submission may still be materially incomplete.

F.3 Rubric Scoring Formulas

For each submitted pair $(\hat{h}_i, \hat{e}_i)$, the judge assigns scores on Contextual Grounding (g_i), Inferential Insight (ℓ_i), and Evidential Justification (j_i). We define the local pair score as

$$q_i = \frac{g_i + \ell_i + j_i}{3}.$$

Pair-level quality is aggregated over all submitted pairs as

$$Q_{\text{pair}} = \frac{1}{K} \sum_{i=1}^{K} q_i.$$

This average reflects the overall quality of the submitted hypothesis–evidence pairs and prevents larger submissions from being rewarded solely for containing more candidates.

For domains that admit multi-pair submissions, the judge additionally assigns set-level scores on Hypothesis-Space Breadth (b), Directional Distinctness (d), and Analytical Utility (u). For submissions with $K > 1$, we define

$$Q_{\text{set}} = \frac{b + d + u}{3}.$$

If a system submits only a single pair in a domain that admits multi-pair submissions, Directional Distinctness is not applicable, and we compute

$$Q_{\text{set}} = \frac{b + u}{2}.$$

In reporting rubric-based results, we primarily present Q_{pair} and Q_{set} separately. When a compact summary is useful, we report $S_{\text{rubric}} = q_1$ for singleton domains and $S_{\text{rubric}} = (Q_{\text{pair}} + Q_{\text{set}})/2$ for domains that admit multi-pair submissions. This summary is used only for concise diagnostic reporting and not for primary model ranking.

Table 7 Full-pool arena leaderboard under `seed-2.0-pro`: every (model, mode) cell on the same BTD scale across the six domains. Rows are sorted by Avg BTD descending; reference participates as a single anonymous competitor (no agent counterpart). This table is the arena-side counterpart to Table 8 and the source of the *Arena #* ordering used there.

			Scientific Domains			Analytical Domains				
#	Model	Mode	Bio	ML	Social	Fin	IT	Safety	Avg	WR
1	claude-sonnet-4.6	Agent	1580.5	1594.5	1612.2	**1754.1**	1672.9	**1700.2**	**1652.4**	73.6%
2	claude-sonnet-4.6	Baseline	1609.4	1582.7	1583.9	1713.9	**1683.1**	1697.6	1645.1	72.0%
3	claude-opus-4.6	Baseline	1615.5	1617.3	**1662.1**	1646.4	1664.7	1654.4	1643.4	71.8%
4	gpt-5.4	Agent	1607.0	1628.6	1603.1	1645.1	1651.8	1638.7	1629.0	73.1%
5	gpt-5.4	Baseline	**1627.3**	**1638.2**	1604.8	1623.0	1630.8	1611.3	1622.6	71.0%
6	glm-5.1	Agent	1542.1	1562.4	1580.0	1649.5	1601.3	1649.5	1597.5	60.2%
7	kimi-k2.6	Agent	1553.3	1570.7	1552.6	1601.8	1611.4	1629.9	1586.6	59.7%
8	reference		1578.9	1565.6	1516.0	1615.3	1598.1	1645.6	1586.6	59.9%
9	claude-opus-4.6	Agent	1540.7	1518.6	1533.2	1628.4	1647.7	1645.4	1585.7	57.1%
10	glm-5.1	Baseline	1597.6	1595.0	1594.5	1554.1	1582.0	1547.0	1578.4	60.0%
11	kimi-k2.6	Baseline	1580.0	1566.4	1565.8	1585.7	1560.3	1609.4	1577.9	59.2%
12	deepseek-v4-pro	Baseline	1543.7	1528.6	1573.3	1525.7	1510.5	1538.2	1536.7	48.0%
13	qwen-3.6-max	Baseline	1519.0	1525.3	1516.1	1508.2	1493.9	1503.1	1510.9	44.3%
14	qwen-3.6-max	Agent	1502.9	1506.5	1495.6	1527.3	1518.4	1512.3	1510.5	42.2%
15	minimax-m2.7	Agent	1492.2	1485.6	1481.0	1534.0	1497.6	1535.1	1504.3	39.0%
16	minimax-m2.7	Baseline	1526.5	1515.6	1509.2	1515.4	1509.7	1432.5	1501.5	39.7%
17	deepseek-v4-pro	Agent	1480.5	1504.7	1499.3	1475.0	1483.1	1514.6	1492.9	38.3%
18	deepseek-v4-flash	Baseline	1482.8	1529.6	1525.5	1488.6	1427.6	1475.9	1488.3	38.2%
19	glm-5	Agent	1462.3	1479.0	1476.9	1485.2	1501.7	1484.5	1481.6	36.0%
20	glm-5	Baseline	1496.2	1469.3	1474.7	1480.7	1481.8	1443.9	1474.4	36.1%
21	gpt-5.4-mini	Agent	1455.8	1463.4	1436.0	1493.0	1452.0	1466.1	1461.1	32.2%
22	deepseek-v4-flash	Agent	1417.1	1451.1	1463.9	1446.6	1427.1	1488.0	1449.0	29.8%
23	gpt-5.4-mini	Baseline	1448.4	1458.8	1416.8	1446.5	1409.1	1428.1	1434.6	27.3%
24	minimax-m2.5	Baseline	1481.3	1444.3	1453.5	1422.6	1435.3	1360.5	1432.9	26.4%
25	minimax-m2.5	Agent	1446.1	1415.5	1397.7	1445.3	1458.2	1401.8	1427.5	23.9%
26	gemini-3.1-pro	Agent	1443.6	1427.1	1418.2	1354.7	1413.6	1409.6	1411.1	23.6%
27	gemini-3.1-pro	Baseline	1457.3	1428.3	1440.5	1338.9	1389.0	1377.9	1405.3	24.2%
28	kimi-k2.5	Agent	1333.8	1351.4	1380.8	1465.3	1369.8	1475.8	1396.1	21.1%
29	gemini-3-flash	Baseline	1445.3	1441.3	1439.3	1257.9	1301.4	1262.5	1357.9	18.0%
30	gemini-3-flash	Agent	1397.3	1408.0	1415.4	1215.4	1305.2	1238.2	1329.9	13.2%
31	kimi-k2.5	Baseline	1328.1	1285.1	1355.2	1258.6	1337.8	1282.5	1307.9	11.9%

G Full-Pool Arena and Rubric Leaderboards

This appendix reports both protocols at the (model, mode) level on a single full-pool BTD scale, so that Section 4.4's cascade can be inspected row by row against specific numbers. Table 7 ranks every (model, mode) cell by arena BTD; Table 8 reports the per-(model, mode) mean rubric score S in the same row order, with a *Rubric Rank* column showing how the same set is re-ranked under rubric Avg S; Table 9 reports the detailed results of different Rubrics in Financial Analysis.

Three regularities emerge directly from the comparison.

Compression. Every Avg S falls within a single rubric point; the strongest and weakest cells in the paper's evaluation set are separated by less than the distance between 4.0 and 5.0, while the same cells span several hundred BTD points under arena. The compression is visible in every domain column.

Ceiling effect. Almost every row sits above 4.0, with the majority above 4.3. Rubric judges default to near-top scores for anything not manifestly deficient, and this behaviour concentrates at the top of the scale rather than the middle. The concentration is structural: absent a single canonical answer, committing to a low

Table 8 Rubric scoring diagnostics under `seed-2.0-pro`: per-(model, mode) mean rubric S on the 1–5 scale, complementing Figure 5 of Section 4.4 and Table 7. Rows are sorted by arena rank (Arena #) on the same full-pool scale as that table; *Rubric Rank* is the position when the same cells are re-ranked by Avg S (descending), with Δ = Rubric Rank − Arena Rank. The narrow Avg S spread and frequent non-zero Δ instantiate the compression and protocol-dependent reordering analyzed in Section 4.4.

			Scientific Domains			Analytical Domains				
Arena #	Model	Mode	Bio	ML	Social	Fin	IT	Safety	Avg S	Rubric Rank (Δ)
1	claude-sonnet-4.6	Agent	4.70	4.92	4.89	4.77	4.37	4.71	4.73	4 (Δ+3)
2	claude-sonnet-4.6	Baseline	4.79	4.92	4.89	4.54	4.40	4.62	4.69	5 (Δ+3)
3	claude-opus-4.6	Baseline	4.76	4.89	4.90	4.25	4.29	4.57	4.61	6 (Δ+3)
4	gpt-5.4	Agent	4.75	4.95	4.91	4.63	4.46	4.69	4.73	3 (Δ-1)
5	gpt-5.4	Baseline	4.76	4.96	4.91	4.57	4.53	4.66	4.73	2 (Δ-3)
6	glm-5.1	Agent	4.57	4.76	4.77	4.29	4.24	4.52	4.53	10 (Δ+4)
7	kimi-k2.6	Agent	4.59	4.79	4.71	4.23	4.25	4.45	4.50	12 (Δ+5)
8	reference		4.76	4.88	4.75	4.95	4.86	4.93	4.85	1 (Δ-7)
9	claude-opus-4.6	Agent	4.62	4.81	4.78	4.20	4.21	4.58	4.53	9 (Δ0)
10	glm-5.1	Baseline	4.75	4.86	4.82	4.25	4.27	4.41	4.56	8 (Δ-2)
11	kimi-k2.6	Baseline	4.67	4.85	4.83	4.37	4.25	4.56	4.59	7 (Δ-4)
12	deepseek-v4-pro	Baseline	4.45	4.67	4.62	4.04	4.00	4.36	4.36	18 (Δ+6)
13	qwen-3.6-max	Baseline	4.55	4.81	4.74	4.23	4.25	4.49	4.51	11 (Δ-2)
14	qwen-3.6-max	Agent	4.51	4.75	4.70	4.17	4.16	4.40	4.45	15 (Δ+1)
15	minimax-m2.7	Agent	4.44	4.62	4.59	3.93	3.92	4.20	4.28	21 (Δ+6)
16	minimax-m2.7	Baseline	4.60	4.75	4.72	4.09	3.96	4.16	4.38	17 (Δ+1)
17	deepseek-v4-pro	Agent	4.28	4.56	4.48	3.87	3.94	4.22	4.23	24 (Δ+7)
18	deepseek-v4-flash	Baseline	4.26	4.57	4.54	4.02	3.80	4.28	4.24	22 (Δ+4)
19	glm-5	Agent	4.37	4.57	4.61	4.07	4.08	4.28	4.33	19 (Δ0)
20	glm-5	Baseline	4.52	4.72	4.68	4.16	4.20	4.38	4.44	16 (Δ-4)
21	gpt-5.4-mini	Agent	4.46	4.76	4.63	4.35	4.17	4.41	4.46	14 (Δ-7)
22	deepseek-v4-flash	Agent	4.10	4.44	4.45	3.92	3.83	4.29	4.17	25 (Δ+3)
23	gpt-5.4-mini	Baseline	4.48	4.76	4.62	4.34	4.14	4.44	4.46	13 (Δ-10)
24	minimax-m2.5	Baseline	4.44	4.63	4.59	3.85	3.95	4.01	4.24	23 (Δ-1)
25	minimax-m2.5	Agent	4.21	4.43	4.32	3.88	3.85	4.07	4.13	28 (Δ+3)
26	gemini-3.1-pro	Agent	4.27	4.47	4.45	3.68	3.93	4.20	4.17	26 (Δ0)
27	gemini-3.1-pro	Baseline	4.33	4.48	4.46	3.96	4.13	4.36	4.29	20 (Δ-7)
28	kimi-k2.5	Agent	4.01	4.31	4.28	3.73	3.79	4.02	4.02	30 (Δ+2)
29	gemini-3-flash	Baseline	4.17	4.42	4.35	3.70	3.99	4.16	4.13	27 (Δ-2)
30	gemini-3-flash	Agent	4.00	4.33	4.30	3.51	3.80	4.03	4.00	31 (Δ+1)
31	kimi-k2.5	Baseline	3.99	4.31	4.34	4.01	3.84	4.01	4.08	29 (Δ-2)

absolute score is harder than declaring one of two submissions better in head-to-head comparison.

Protocol-dependent reordering. Relative to the arena ordering, 28 of 31 full-pool entries receive a different rubric rank. The top three arena entries each rank three positions lower under rubric scoring, while the Reference moves from rank 8 under arena to rank 1 under rubric. These shifts demonstrate substantial protocol sensitivity in fine-grained ordering, but they do not identify the underlying mechanism.

Together, score compression, ceiling effects, and widespread reordering support using arena BTD for primary comparative ranking and rubric scores for diagnostic analysis.

H Extended Analysis of Evaluator Alignment

We examine two validation signals that complement the cross-judge analysis in Appendix I. Human preference judgments assess alignment between automated rankings and expert preferences, while ICLR acceptance outcomes provide a source-side signal not used by the arena judge. Because these analyses operate at different granularities, we interpret each within the scope of the evidence it provides.

Table 9 Per-rubric-dimension breakdown for **Financial Analysis** under `seed-2.0-pro` (full pool). Rows sorted by Avg S descending; all dimensions on the 1–5 scale.

#	Model	Mode	Ground.	Insight	Justif.	Breadth	Distinct.	Utility	Avg S
1	reference		4.99	4.71	4.98	5.00	5.00	5.00	4.95
2	claude-sonnet-4.6	Agent	5.00	4.25	4.88	4.50	5.00	5.00	4.77
3	gpt-5.4	Agent	4.77	4.13	4.66	4.46	4.92	4.88	4.63
4	gpt-5.4	Baseline	4.75	4.06	4.67	4.39	4.77	4.79	4.57
5	claude-sonnet-4.6	Baseline	4.61	4.09	4.51	4.55	4.82	4.68	4.54
6	kimi-k2.6	Baseline	4.55	3.90	4.41	4.22	4.73	4.38	4.37
7	gpt-5.4-mini	Agent	4.61	3.81	4.49	4.11	4.68	4.39	4.35
8	gpt-5.4-mini	Baseline	4.62	3.81	4.51	4.11	4.68	4.34	4.34
9	glm-5.1	Agent	4.37	3.94	4.29	4.16	4.54	4.41	4.29
10	claude-opus-4.6	Baseline	4.28	3.89	4.15	4.23	4.61	4.37	4.25
11	glm-5.1	Baseline	4.41	3.84	4.30	4.15	4.64	4.27	4.25
12	qwen-3.6-max	Baseline	4.31	3.85	4.27	4.04	4.65	4.25	4.23
13	kimi-k2.6	Agent	4.23	3.88	4.13	4.20	4.61	4.32	4.23
14	claude-opus-4.6	Agent	4.17	3.82	4.08	4.24	4.56	4.32	4.20
15	qwen-3.6-max	Agent	4.17	3.89	4.09	3.99	4.61	4.26	4.17
16	glm-5	Baseline	4.39	3.77	4.29	3.91	4.50	4.13	4.16
17	minimax-m2.7	Baseline	4.21	3.66	4.04	4.05	4.52	4.06	4.09
18	glm-5	Agent	4.17	3.74	4.06	3.96	4.46	4.04	4.07
19	deepseek-v4-pro	Baseline	4.11	3.73	3.94	4.07	4.44	4.04	4.04
20	deepseek-v4-flash	Baseline	4.06	3.74	3.88	4.02	4.38	4.08	4.02
21	kimi-k2.5	Baseline	4.52	3.09	4.37	4.19	4.21	3.61	4.01
22	gemini-3.1-pro	Baseline	4.03	3.64	3.90	3.82	4.45	3.90	3.96
23	minimax-m2.7	Agent	3.95	3.55	3.89	4.00	4.27	3.91	3.93
24	deepseek-v4-flash	Agent	3.91	3.71	3.74	3.93	4.33	3.92	3.92
25	minimax-m2.5	Agent	3.88	3.49	3.74	3.95	4.25	3.95	3.88
26	deepseek-v4-pro	Agent	3.80	3.63	3.70	3.89	4.34	3.86	3.87
27	minimax-m2.5	Baseline	3.92	3.42	3.74	3.98	4.15	3.88	3.85
28	kimi-k2.5	Agent	3.66	3.40	3.51	4.08	3.98	3.79	3.73
29	gemini-3-flash	Baseline	3.70	3.45	3.48	3.78	4.22	3.61	3.70
30	gemini-3.1-pro	Agent	3.43	3.53	3.31	3.82	4.32	3.68	3.68
31	gemini-3-flash	Agent	3.33	3.33	3.14	3.68	4.24	3.37	3.51

H.1 Detailed Analysis of Human-Judge Alignment

To assess alignment between the primary arena judge and expert preferences, we compare rankings induced by `seed-2.0-pro` with rankings aggregated from 1,500 expert pairwise judgments across Financial Analysis, Social Science, and Biomedical Science. Detailed annotation procedures are provided in Appendix D.3.

As shown in Table 10, agreement is strong but varies by domain. The top three positions match exactly in Biomedical Science and Financial Analysis, whereas Social Science differs by a swap between ranks 3 and 4. Most rank shifts are small, although Biomedical Science contains larger displacements among lower-ranked models, including $\Delta = +9$ for `qwen-3.6-plus` and $\Delta = -7$ for `gpt-5.4-mini`.

Lower alignment in Biomedical Science. The lower correlation in Biomedical Science ($\tau = 0.53$) indicates greater evaluator disagreement than in Financial Analysis ($\tau = 0.97$). The current analysis does not isolate the source of this difference, so we do not attribute it to a particular property of the cases or annotators. Nevertheless, the matching top-three positions in Biomedical Science and the concentration of the largest shifts among lower-ranked models show that disagreement does not extend uniformly across the leaderboard. The stronger correlations in Financial Analysis and Social Science further indicate that alignment varies by domain.

Table 10 Per-domain leaderboard agreement between human experts and the `seed-2.0-pro` judge across 16 models. Δ is the rank shift from human to judge; per-domain Kendall's τ and Spearman's ρ are reported at the bottom. Most models show small displacements, with the largest spread concentrated in Biomedical Science.

Model	Biomedical			Financial Analysis			Social Science		
	Rank	Win%	Δ	Rank	Win%	Δ	Rank	Win%	Δ
claude-sonnet-4.6-high	2	90.9	0	**1**	97.5	0	**1**	93.8	0
gpt-5.4-high	**1**	81.8	0	4	88.5	0	2	71.4	0
claude-opus-4.6-high	3	86.7	0	5	73.5	0	4	75.0	-1
qwen-3.6-max-thinking	8	44.4	+4	2	75.8	0	5	71.7	0
kimi-k2.6-thinking	5	66.7	+1	3	81.2	0	7	58.3	-1
glm-5.1-thinking	4	70.6	0	6	63.4	0	3	73.9	+1
qwen-3.6-plus-thinking	6	57.1	+9	8	54.8	0	6	64.3	+2
glm-5-thinking	15	18.2	-6	7	63.9	0	14	23.1	-3
gemini-3.1-pro-high	7	54.5	0	12	23.5	+2	9	51.9	-2
kimi-k2.5-thinking	14	28.6	-1	9	44.3	0	11	43.2	+1
gpt-5.4-mini-high	12	37.5	-7	10	40.0	0	13	26.0	0
minimax-m2.7-thinking	11	38.5	-3	11	29.0	0	10	51.9	0
deepseek-v4-flash-high	10	40.0	+4	13	20.0	-1	16	11.3	-1
minimax-m2.5-thinking	9	44.4	+2	14	18.8	-1	8	50.0	+1
gemini-3-flash-high	13	30.0	-3	15	2.9	0	15	17.3	+1
deepseek-v4-pro-high	16	5.9	0	16	0.0	0	12	42.0	+2

Correlations (Kendall's τ / Spearman's ρ)
Biomedical: 0.53 / 0.67 Financial: 0.97 / 0.99 Social Science: 0.85 / 0.96

Table 11 Reference arena strength under `seed-2.0-pro` (baseline pool) on the `machine_learning` subset, split by ICLR 2026 acceptance. *Debiased*: per-case median of Reference's $[0, 1]$ debiased win share. *Win Rate*: per-case median of the AND-of-both win indicator, where Reference wins both forward and reverse, and ties score 0.5. *BTD*: Reference's BTD rating fit on the bucket's matches alone.

Group	Debiased (↑)	Win Rate (↑)	BTD (↑)
Accepted	0.602	0.654	1577.6
Rejected	0.538	0.615	1530.4

H.2 Association with Peer-Review Outcomes

The Machine Learning source papers were sampled from ICLR 2026 and stratified by acceptance category, providing an external outcome not used by the arena judge. If arena evaluation is sensitive to source-side differences associated with peer-review outcomes, the source-derived Reference constructed by Forge from each source paper should be more competitive on accepted cases than on rejected ones.

Table 11 shows a consistent directional pattern across all three measurements. Accepted cases exceed rejected cases by 0.06 in median per-case debiased score and 0.04 in median per-case win rate, while the bucket-specific BTD rating is 47 points higher. This agreement provides a complementary external check using information unavailable to the arena judge. Because acceptance reflects factors beyond hypothesis quality, we do not treat it as case-level ground truth.

I Inter-Judge Agreement Analysis

The headline analyses in Section 4 use `seed-2.0-pro` as the sole arena judge. To test whether the resulting rankings are an artifact of that single model's idiosyncrasies, we re-run the entire arena pipeline with an independent second judge, `mimo-v2-pro`, on the same submissions and the same case set. This appendix reports five complementary agreement signals between the two judges, spanning three granularities.

Table 12 Inter-judge agreement between `seed-2.0-pro` and `mimo-v2-pro` on the 16 baseline-mode submissions evaluated by both judges in each domain. Spearman ρ and Kendall τ are rank correlations between the two judges' BTD ratings; Rank Shift is the mean $|\Delta\text{rank}|$ per model; Top-5 is the intersection size of each judge's top-5 models; Pairwise is the fraction of shared (case, model$_a$, model$_b$) matches where both judges' forward verdicts agree.

Domain	Spearman ρ (↑)	Kendall τ (↑)	Rank Shift (↓)	Top-5 (↑)	Pairwise (↑)
Biomedical Science	0.915	0.750	1.50	5/5	63.3%
Machine Learning	0.950	0.833	1.12	5/5	67.0%
Social Science	0.974	0.900	0.75	5/5	66.8%
Financial Analysis	0.909	0.833	1.00	4/5	68.1%
IT Operations	0.900	0.783	1.50	4/5	67.5%
Safety Investigation	0.944	0.817	1.12	4/5	65.6%
Overall	**0.920**	**0.764**	**1.17**	**4.50/5**	**66.3%**

Three levels of agreement. We measure cross-judge agreement at three granularities, all restricted to the baseline-pool model set used for the main leaderboard:

- *Full-ranking agreement*: per-domain Spearman ρ and Kendall τ between the two judges' BTD ratings (one value per (model, domain) cell), and *Mean Rank Shift*, the average $|\Delta\text{rank}|$ per model when switching judges.
- *Top-tier identity*: per-domain *Top-5 Overlap*, i.e. the size of the intersection of the two judges' top-5 ranked models.
- *Individual verdicts*: per-domain *Pairwise Agreement*, i.e. the fraction of shared (case, model$_a$, model$_b$) matches where both judges' forward verdicts select the same outcome among A wins, B wins, and tie.

Table 12 reports all five statistics per domain plus a pooled row across all (model × domain) cells.

Rankings replicate; single-match verdicts are noisier. The two judges produce essentially the same leaderboard. Per-domain rank correlations are uniformly strong (Spearman $\rho \in [0.90, 0.97]$; Kendall $\tau \in [0.75, 0.90]$), the pooled $\rho = 0.920$ has a tight bootstrap 95% CI of $[0.871, 0.949]$, and the *Mean Rank Shift* of 1.17 on a 16-model scale translates this into concrete terms: switching judges moves a model's leaderboard position by only about one place on average. *Top-5 Overlap* averages 4.5/5 across domains, with three domains in perfect agreement and the remaining three differing by a single model. The top tier's identity is robust to judge choice. Pairwise verdict agreement ranges from 63% to 68% across domains, indicating substantial but incomplete agreement at the individual-match level. Despite disagreement on roughly one third of individual matches, the BTD-aggregated rankings remain closely aligned across judges. This rank-level stability supports the use of arena aggregation for the primary leaderboard.

Table 13 reports the baseline leaderboard under `mimo-v2-pro` as a companion to Table 3. Both judges rank `claude-sonnet-4.6` first, `claude-opus-4.6` second, and `gpt-5.4` third, with `kimi-k2.5` last. They also yield the same within-family ordering for the DeepSeek, GLM, and Kimi model pairs.

Fine-grained disagreement is concentrated in the middle of the leaderboard. The Reference moves from rank 4 under `seed-2.0-pro` to rank 7 under `mimo-v2-pro`, the largest displacement, while no other system moves by more than two positions. Because this analysis does not isolate the source of judge disagreement, we interpret the Reference shift as judge sensitivity in its relative placement rather than evidence of a specific stylistic mechanism. The mean rank shift of 1.17 and pooled Spearman $\rho = 0.920$ with 95% CI $[0.871, 0.949]$ show that the overall ranking structure remains closely aligned.

Complementary validation signals. We consider cross-judge agreement alongside two additional checks. The first measures alignment with human expert preferences in three domains (§4.3, Kendall $\tau = 0.90$), and the second examines the directional association between the competitiveness of the source-derived Reference and ICLR acceptance outcome (Appendix H.2, with a 47-point higher bucket-specific BTD rating on accepted

Table 13 Main leaderboard under reference-independent pairwise judging by `mimo-v2-pro` as the independent second judge (companion to Table 3). All BTD and WR conventions match Table 3; Reference participates as an anonymous competitor for calibration.

			Scientific Domains			Analytical Domains				
#	Model	Effort	Bio	ML	Social	Fin	IT	Safety	Avg	WR
1	claude-sonnet-4.6	high	**1642.3**	1618.9	1589.3	**1640.2**	**1632.3**	**1609.9**	**1622.2**	67.7%
2	claude-opus-4.6	high	1632.2	**1634.7**	**1620.2**	1610.7	1589.4	1574.3	1610.3	62.6%
3	gpt-5.4	high	1578.2	1602.9	1570.0	1556.0	1548.3	1542.1	1566.3	49.1%
4	glm-5.1	✓	1583.9	1593.2	1562.7	1549.4	1543.2	1543.8	1562.7	49.0%
5	kimi-k2.6	✓	1530.3	1547.1	1542.8	1555.7	1549.7	1569.3	1549.1	44.3%
6	deepseek-v4-pro	high	1514.0	1519.3	1556.1	1535.6	1534.0	1547.7	1534.5	41.5%
7	reference		1494.7	1538.0	1485.8	1558.5	1532.1	1567.2	1529.4	39.3%
8	deepseek-v4-flash	high	1489.0	1505.8	1513.4	1564.3	1533.5	1548.7	1525.8	40.0%
9	qwen-3.6-max	✓	1519.8	1520.4	1517.6	1523.3	1499.7	1509.7	1515.1	35.1%
10	minimax-m2.7	✓	1487.9	1484.0	1461.8	1509.9	1514.4	1470.8	1488.1	26.9%
11	glm-5	✓	1467.6	1456.6	1452.6	1496.2	1482.0	1488.2	1473.9	25.2%
12	gpt-5.4-mini	high	1475.8	1422.8	1429.3	1461.1	1496.3	1458.4	1457.3	20.2%
13	gemini-3.1-pro	high	1438.6	1426.3	1448.0	1442.7	1426.2	1457.5	1439.9	19.7%
14	minimax-m2.5	✓	1423.3	1412.6	1434.4	1451.8	1436.7	1435.2	1432.3	15.1%
15	gemini-3-flash	high	1432.7	1436.4	1455.2	1373.9	1386.8	1385.8	1411.8	16.0%
16	kimi-k2.5	✓	1310.3	1300.0	1383.4	1193.8	1324.8	1320.6	1305.5	7.2%

cases). These analyses operate at different granularities and do not constitute three replications of the same leaderboard. Instead, they probe sensitivity to judge choice, expert alignment, and association with a peer-review outcome not used by the arena judge. Together, they broaden the evidence base for HypoEval without treating any single signal as definitive.

J Prompt Templates

This appendix provides the complete prompt templates used in the HypoArena pipeline for two representative domains: Social Science (scientific research, single-hypothesis) and Safety Investigation (analytical investigation, multi-hypothesis). The full set for all six domains is released with our code.

J.1 Stage 1: Context Construction

J.1.1 Context Forge

Context Forge — Social Science

```
You are ForgeAgent.  You are reconstructing the pre-study reasoning behind a social science research paper.

This is an iterative process:  AuditAgent will review your output and may send feedback asking you to revise.
When that happens, use the full conversation history to understand what you previously produced and why, then
make targeted revisions rather than starting from scratch.

Think like a benchmark constructor, not like an author writing a paper introduction.  A strong Context feels
like the background and unresolved-tension section of a serious survey -- a reader comes away thinking "there is
enough background and tension here to support a serious hypothesis, but the answer has not been given away."

You have access to the original source document and web search.

Domain Guidance
- Cover the relevant theories, constructs, operationalizations, competing mechanisms, and empirical traditions.
End at the unresolved choice among mechanisms or trade-offs, not at the preferred mediator.
- Do not name the paper's preferred mediator, moderator, or framing variable as the explanation.
- Actively use search to enrich the context beyond the paper's local introduction.  Prefer reliable and
substantive sources.
- Write a technical survey-style context that integrates paper background with external primary-source background.
```

How to Approach This Task
- Read the source document to identify the problem setting, factual background, prior approaches, existing evidence, and the tensions it surfaces.
- Search is a thinking tool for widening the frame. A good self-check: could this draft have been written from the paper alone? If yes, it probably needs search-informed widening.
- Go deep on the tensions that matter rather than cataloguing everything in shallow form.
- End at unresolved tension, disagreement, vulnerability, or missing evidence, not at a solution-shaped landing.
- Keep the final text clean of citations, URLs, bibliography, or any retrieval residue.

Self-Check: (1) Did you use search to widen? (2) Field-level background, not rewritten Introduction? (3) Could a reader infer the answer? (4) Ends at tension? (5) Sufficient density?

Output: {"context": "<The benchmark context>"}

Context Forge — Safety Investigation

You are ForgeAgent. Your job is to construct model-visible benchmark Context from an official accident or incident investigation report.

The Context is the factual layer of the investigation -- timeline, equipment, environmental conditions, operating procedures, anomalies, safeguards, and physical evidence. It must preserve the complexity of the case closely enough that a benchmark model can form its own investigative hypothesis from the material.

Think like a benchmark constructor assembling the factual backbone of a case, not like a safety analyst writing conclusions. A strong Context feels like the factual record of an investigation where the meaning has not yet been assigned.

Domain Policy
- Use only the source investigation report. Do not use outside search.
- Respect the report's factual style and terminology.

What to Include
- Operating setting, facility/equipment descriptions, system configuration
- Full timeline with specific timestamps, readings, operator actions
- Environmental conditions and boundary parameters
- Operating procedures, maintenance history, inspection records
- Safeguards, alarms, interlocks, and their status
- Anomalies, warnings, prior incidents, or near-misses
- Physical evidence and post-incident observations
- Factual tensions: signals present but not acted upon, safeguards that did not prevent the outcome

Leakage Policy -- Must not contain: the agency's final probable cause; board conclusions or recommendations; post-hoc causal explanations; analysis-layer judgments about what "caused" or "contributed to" the incident.

Self-Check: (1) No causal language? (2) No probable cause/recommendations? (3) Detailed enough (timestamps, readings)? (4) Unresolved, not summarized? (5) Factual tensions preserved? (6) Evidentiary tone intact?

Output: {"context": "<The benchmark context>"}

J.1.2 Context Audit

Context Audit — Social Science

You are AuditAgent. You are reconstructing the pre-study reasoning behind a social science research paper.

Read the source document and audit whether the draft Context is benchmark-ready as model-visible input. Do not rewrite the draft.

What Benchmark-Ready Context Means
- Long, formal, information-rich; carries enough background and unresolved tension to support a high-quality hypothesis.
- Must not leak answer-side contents: final conclusions, preferred method/mechanism, or post-hoc causal attribution.
- Time consistency: only information available before the hypothesis was formed.
- Should read like field or case background, not like paper narration or conclusion section.
- Should feel broader than a single paper's local framing.

Hard Pass/Fail Gates
1. Leaks answer-side content.
2. Does not go beyond the source paper's local framing.
3. Carries retrieval residue (citations, URLs, bibliography).
4. Too thin to support non-trivial hypothesis generation.

Output: {"passed": true} or {"passed": false, "summary": "...", "problems": [...]}

Context Audit — Safety Investigation

You are AuditAgent. Judge whether the draft Context is benchmark-ready model-visible input for a safety-domain hypothesis generation benchmark.

What Benchmark-Ready Safety Context Means
- Long, formal, information-rich; preserves the report's factual record and evidentiary tone.
- Contains timeline, equipment, environmental, operational, and anomaly details.
- Preserves factual tensions, competing signals, and unresolved questions.
- Does not leak the agency's probable cause, root-cause determination, or recommendations.

Hard Pass/Fail Gates
1. Leaks the agency's final probable cause or root-cause determination.
2. Contains analysis-layer causal language ("caused," "contributed to").
3. Reads like a summarized finding rather than a factual record.
4. Too thin to support non-trivial investigative hypothesis generation.

Audit Philosophy: Protect factual density, timeline fidelity, unresolved tension, and answer-side separation. Do not ask for better prose or methodology sections.

Output: {"passed": true} or {"passed": false, "summary": "...", "problems": [...]}

J.2 Stage 2: Hypothesis Construction

J.2.1 Hypothesis Forge

Hypothesis Forge — Social Science

You are ForgeAgent. You are reconstructing the pre-study reasoning behind a social science research paper.

Your job is to reconstruct the single strongest benchmark-ready hypothesis that a careful researcher could state from the source document and the benchmark-visible Context, plus the evidence -- a research plan that explains how the hypothesis could be validated.

Domain Guidance
- Favor one behavioral, psychological, or organizational mechanism. Keep moderators, mediators, boundary conditions out unless one of them is the single central claim.
- The evidence plan should cover validation directions (replication, manipulation, mediation, moderation, boundary-condition, external-validity) in ex-ante language. Do not report statistics or study outcomes as already known.

STEP 1 -- Formulate the Hypothesis
Think like a benchmark constructor recovering the central bet worth stating before results were known.
- Distill into a single falsifiable claim in a short paragraph, using forward-looking language.
- Must be genuinely supportable from the Context alone.
- If the source supports several outcomes, compress into one higher-level central bet.

STEP 2 -- Draft the Evidence
A concise research plan as numbered items (one sentence each) describing validation directions -- comparisons, perturbations, mechanism probes, or boundary-condition tests. Should reflect the paper's actual evidence chain, not invent an ideal experiment from scratch.

Self-Check: (1) One claim or several bundled? (2) Context-supportable? (3) Forward-looking or summary? (4) Validation directions, not result recap?

Output: {"hypothesis": "<...>", "evidence": "<...>"}

Hypothesis Forge — Safety Investigation

You are ForgeAgent. Construct benchmark-ready hypothesis-evidence pairs from a safety investigation's model-visible Context.

The source report contains both factual material and the agency's analysis/conclusions. You must read both, but hypotheses must be supportable from the Context alone.

Core Rule: Same-Case Layering

The source report contains a factual layer (what the Context preserves) and an analysis layer (causal explanations). Your hypotheses should be the "explanatory middle ground" -- judgments a careful investigator could form from the factual record, inspired by but not copied from the agency's analysis.

Before You Begin: Identify the agency's stated probable cause. Then deliberately construct hypotheses at a different level of specificity or from a different analytical angle.

Hypothesis Style -- Multiple pairs, each a different angle:
- accident mechanism or failure chain
- barrier failure or safeguard inadequacy
- human-system interaction or procedural gap
- maintenance, inspection, or design weakness
- organizational or management-system vulnerability

Evidence Style -- Each item must: point to a concrete context-visible fact; be one sentence; be factual rather than conclusory.

Self-Check: (1) Different analytical level from agency findings? (2) Distinct angles? (3) Context-supportable? (4) Investigative judgment, not restated finding? (5) Evidence points to concrete facts?

Output: {"hypotheses": [{"hypothesis": "...", "evidence": "..."}, ...]}

J.3 Hypothesis Generation (Evaluation-Time Inference)

Generation — Social Science (Single-Hypothesis)

You are a social scientist examining existing theory and empirical findings.

You are given a detailed background context from social science research. Your task is to formulate hypotheses.

A strong hypothesis is:
- Evidence-bounded: grounded in the provided context, not speculative beyond the material
- Non-trivial: not an obvious restatement or surface-level observation
- Valuable: identifies a mechanism, process, or structural factor worth investigating
- Calibrated: claim strength proportionate to evidence support
- Specific: names concrete entities, processes, or relationships

The hypothesis should be a single falsifiable or assessable claim as a short paragraph of plain prose. It should feel like a forward-looking judgment, not a summary.

The evidence should be a research plan: numbered items describing how the hypothesis could be validated.

Favor one claim about a behavioral mechanism, psychological process, or social phenomenon. The evidence plan should describe studies, comparisons, or measurement approaches.

Output: {"hypotheses": [{"hypothesis": "<...>", "evidence": "<...>"}]}

Generation — Safety Investigation (Multi-Hypothesis)

You are a safety investigator examining the factual record of an incident.

You are given a detailed context from safety investigation analysis. Your task is to produce multiple hypothesis-evidence pairs.

A strong hypothesis is:
- Evidence-bounded: grounded in facts visible in the context, not speculative
- Non-trivial: goes beyond surface-level observations or obvious takeaways
- Valuable: identifies a mechanism, vulnerability, or intervention point
- Calibrated: claim strength matches evidence strength
- Specific: names concrete entities, processes, or relationships

Each hypothesis should be a short paragraph identifying a distinct analytical angle. Each evidence section should contain numbered items -- concrete facts, timeline elements, conditions, or signals from the context.

Focus on accident mechanisms, barrier failures, human-system interaction gaps, and organizational vulnerabilities.

Each hypothesis must include a "category" field. Categories: "causal", "latent", "intervention".

Output: {"hypotheses": [{"hypothesis": "...", "evidence": "...", "category": "causal|latent|intervention"}, ...]}

J.4 Arena Evaluation

Arena Judge

```
You are an expert evaluator comparing two hypothesis outputs generated from the same context. Your goal is to
identify which output demonstrates stronger evidence-grounded analytical reasoning.

Judging Principles
1. Length is not quality. Evaluate density of well-grounded reasoning per claim. Extra length from speculative
elaboration or hedging is a weakness.
2. Claims must be bounded by context evidence. Naming entities from the context is restatement, not grounding.
Naming entities NOT in the context is speculation.
3. Genuine synthesis vs. speculative elaboration. Insight means integrating dispersed observations into a
non-obvious explanatory structure the context supports but does not state.
4. Parsimony over proliferation. Fewer well-grounded hypotheses beat many loosely grounded ones.
5. Calibration. A strong claim with weak evidence is worse than a moderate claim with strong evidence.
6. Selectivity is expertise. Developing the most important hypotheses demonstrates stronger judgment.

Evaluation Process
Important: Output A and Output B are in arbitrary order. Do not favor either based on position.

Step 1 -- Read ALL hypotheses in BOTH outputs fully. Identify each output's single weakest hypothesis.
Step 2 -- Compare per quality dimension in 1-2 sentences.
Step 3 -- Determine verdict. The quality floor matters as much as the ceiling.

Output: {"verdict": "A»B" | "A>B" | "A=B" | "B>A" | "B»A", "reason": "<one-sentence summary>"}

Use A=B only when genuinely comparable across all dimensions.
```

K Skills

Table 14 Structured Analytical Skills for Hypothesis Generation

Skill Name	Category	Description
Baseline	Single-pass Generation	Reads context and produces structured hypothesis sets (e.g., causal, latent, or analytical intervention-based for investigative domains; scientific research-oriented for academic domains, etc.) through evidence-first abductive reasoning.
ACH (Analysis of Competing Hypotheses)	Evidence Matrix	Systematically evaluates multiple hypotheses using evidence-diagnosticity matrix with C/I/NA scoring. Ranks by inconsistency score (least disconfirming evidence wins). Guards against confirmation bias.
Chronology	Temporal Analysis	Decomposes evidence into strict chronological sequence. Analyzes intervals, gaps, anomalous speeds, and temporal patterns to reveal hidden causal connections. Identifies critical transition points.
Assumptions Check	Assumption Testing	Systematically identifies and stress-tests implicit assumptions underlying analysis. Categorizes by vulnerability (well-supported/reasonable/fragile). Generates alternatives from fragile assumptions.

Continued on next page...

Table 14 – Continued from previous page

Skill Name	Category	Description
Brainstorming	Divergent-Convergent	Structured divergent-then-convergent process. Simulates multiple analytical perspectives (technical, human factors, organizational, contrarian). Generates broad candidate set before narrowing.
Cross-Impact Matrix	Interaction Analysis	Examines how each factor interacts with every other factor using ++/+/0/−/— notation. Reveals reinforcing loops, failed inhibitors, and overlooked connections. Identifies leverage points.
Devil's Advocacy	Adversarial Challenge	Constructs strongest possible opposing case against leading hypothesis through evidence reliability, completeness, and reasoning attacks. Requires genuine commitment to opposing position.
Diagnostic Reasoning	Evidence Quality	Identifies highest diagnostic value evidence—facts that most effectively discriminate between competing explanations. Prioritizes evidence quality over quantity. Builds hypotheses from diagnostic evidence outward.
Morphological Analysis	Combinatorial	Decomposes problem into 3–5 key dimensions, enumerates possible values, systematically examines all combinations. Eliminates impossible, identifies high-plausibility and surprising combinations.
Premortem	Backward Reasoning	Assumes leading hypothesis is WRONG, works backward to identify reasoning failures (misread evidence, missing evidence, alternative paths, overlooked factors). Analytical red-teaming.
Red Hat Analysis	Actor Perspective	Adopts perspective of key actors—reconstructs their information state, pressures, training, incentives. Combats mirror imaging. Reasons how situation looks from their perspective, not analyst's.
Starbursting	Dimensional (5W1H)	Generates hypotheses from six perspectives: What/When/Where/Who/How/Why. Performs cross-dimensional analysis to reveal interactions. Ensures comprehensive coverage without blind spots.
What-If Analysis	Counterfactual	Assumes alternative outcome is proven true, works backward to construct causal path. Reframes from "could this happen?" to "given it did happen, how?". Breaks analytical anchoring.