FORESCI: EVALUATING LLM AGENTS FOR FORWARD-LOOKING AI RESEARCH JUDGMENT

Anonymous authors
Paper under double-blind review

ABSTRACT

AI research often requires decisions before future evidence exists: which bottleneck to attack, which direction to pursue, or where a project should be positioned. We introduce ForeSci, a temporally controlled benchmark for evaluating whether LLM agents can make such forward-looking research judgements from historical evidence. ForeSci contains 500 tasks across four fast-moving AI domains and four decision families. Each task is paired with a cutoff-aligned offline knowledge base; post-cutoff papers are hidden during generation and used only for validation. To avoid random future-event prediction, tasks are derived from pre-cutoff taxonomy branches and evidence signals, while generation-time retrieval and tools are restricted to the cutoff-aligned evidence package. We evaluate native LLMs, Hybrid RAG, and three research-agent adaptations across four backbones. Agent-based methods improve traceability over Hybrid RAG, while their gains in future-target alignment over native LLMs are modest and task dependent. Diagnostics reveal a recurring evidence-decision decoupling: agents may cite relevant evidence while forecasting the wrong research object. Task-specific checks between evidence and final decision reduce drift and improve future-target alignment. As research agents move from assisting known tasks to shaping open research directions, ForeSci offers a first controlled step toward measuring whether their judgement can be trusted to guide that future.

1 INTRODUCTION

AI research moves on a timescale where today's frontier becomes tomorrow's baseline. AI research decisions (*e.g.*, which bottleneck to attack or which direction merits a six-month commitment) therefore require foresight. Decisions that align with the field's subsequent trajectory can direct effort toward problems of lasting importance, whereas those that do not risk committing substantial resources to priorities that quickly lose relevance. As autonomous research agents become more widely deployed for ideation, planning, and scientific workflow execution (Lu et al., 2026; Li et al., 2025; Tang et al., 2025; Yamada et al., 2025; Gridach et al., 2025; Chen et al., 2025; Lupidi et al., 2026; Wang et al., 2025), they are increasingly being asked to participate in this decision layer. Whether current LLM agents can make defensible, evidence-grounded research judgements that align with future developments yet to unfold is therefore a central open question.

Existing benchmarks do not fully answer this question. Prior work has mostly evaluated whether AI systems can answer questions over papers, synthesize literature (Lála et al., 2023; Wan et al., 2024; Lewis et al., 2020), use tools (Yao et al., 2023; Schick et al., 2023), execute research workflows (Chen et al., 2025; Lupidi et al., 2026; Wang et al., 2025), or generate predictions on specific future scientific facts (Ajith et al., 2026). None of these tasks asks whether an agent can produce an open-ended research decision, such as picking a bottleneck, ranking a research agenda, or selecting a venue, using only the evidence available at a specific historical moment.

Building such a benchmark raises two challenges. First, the evidence boundary must be enforceable. Post-cutoff papers should not appear in training corpus, retrieval or other model-visible benchmark artifacts. Otherwise, a system may rely on hindsight rather than foresight (Zhao et al., 2024; Ye et al., 2024; Liu et al., 2026a; Ajith et al., 2026; Wang et al., 2026). Second, the research judgements being evaluated must be historically inferable. They should be grounded in signals available before the

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

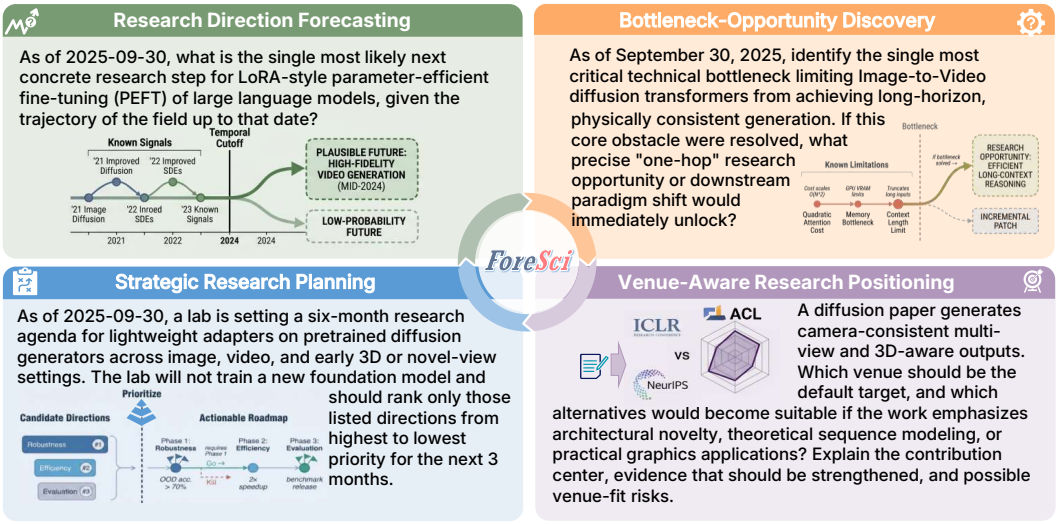

Figure 1: Representative ForeSci task examples across the four decision families: direction forecasting, bottleneck–opportunity discovery, strategic research planning, and venue-aware research positioning.

cutoff, rather than in arbitrary future events or design choices. A foresight benchmark must therefore govern both what a system can see and what it is fair to ask.

To address these challenges, we introduce ForeSci, a temporally controlled benchmark for forward-looking AI research judgement. It contains 500 tasks across four fast-moving AI domains and four decision families (Figure 1). Each task pairs a public question with a cutoff-aligned offline knowledge base, while post-cutoff evidence is hidden until evaluation. Tasks are constructed from pre-cutoff taxonomy branches, node-level evidence records, and method-evolution signals, ensuring that each decision is historically inferable but not directly answerable from future leakage. Each answer is evaluated by four complementary signals: factual support following the atomic-fact (Min et al., 2023), future-target alignment (Wang et al., 2026), evidence traceability, and reviewer persuasiveness motivated by peer-review reliability analyses (Francois, 2015). We evaluate a native LLM, Hybrid RAG, and three offline-adapted research-agent systems across four LLM backbones. All systems operate within the same historical knowledge base, with web access disabled; we separately audit proprietary-backbone outputs for explicit post-cutoff target leakage.

Across the overall evaluation, agentic systems make the evidence chain more traceable, but do not consistently produce decisions that better match later research developments than native LLMs. The diagnostic audit identifies an evidence-decision decoupling: relevant evidence can still be connected to the wrong object, causal role, or intervention. Task-specific checks at the evidence-to-decision stage reduce drift and improve future-target alignment. Our key contributions include:

- **A temporally-controlled benchmark** with 500 tasks across four AI domains and four decision families, paired with cutoff-aligned offline knowledge bases and output-level temporal-leakage auditing
- **A multi-signal evaluation protocol** separating factuality, future-target alignment, evidence traceability, and reviewer persuasiveness.
- **A systematic evaluation and diagnostic audit of LLM research agents** showing task-conditional changes in traceability and factuality, and identifying a previously-unstudied failure mode—evidence-decision decoupling.

2 Related Work

Autonomous Research Agents AI-for-science systems have moved from local literature QA toward agentic workflows that retrieve, synthesize, ideate, and execute parts of the research loop Lu et al. (2026); Ghareeb et al. (2026). PaperQA-style systems (Lála et al., 2023), Chain-of-Ideas (Li et al., 2025), AI-Researcher (Tang et al., 2025), AI Scientist (Yamada et al., 2025), Intern-Atlas Wu et al. (2026b) and recent agentic AI-for-science workflows (Gridach et al., 2025) illustrate this shift toward autonomous research assistance. As these agents are increasingly deployed for ideation and planning, they are implicitly asked to make research decisions. Yet whether they can do so from evidence available at a specific historical moment remains an open question. ForeSci directly evaluates this decision-making layer using cutoff-aligned historical evidence.

Benchmarks for Autonomous Research Existing benchmarks for autonomous research mainly focus on scientific reasoning Lu et al. (2022); Center for AI Safety et al. (2026); Bragg et al. (2026); Liu et al. (2026b); Jansen et al. (2025), artifacts—literature-grounded question answering (Wan et al., 2024; Lála et al., 2023), machine-learning research workflows (Chen et al., 2025; Lupidi et al., 2026), and paper-based agent arenas (Wang et al., 2025). These benchmarks measure retrieval, tool use, synthesis, or execution. Compared to them, ForeSci instead asks systems to make prospective research decisions rather than recover accessible answers or execute known workflows. The benchmarks most closely related to our work aim to predict specific future scientific facts or artifacts. PreScience forecasts future scientific contributions and paper components (Ajith et al., 2026), while related efforts target future scientific events and progress, downstream insights, experimental outcomes, or emerging research topics (Wu et al., 2026a; He-Yueya et al., 2026; Sehwag et al., 2026; Gu & Krenn, 2025). ForeSci moves beyond outcome-centric forecasting by testing whether agents can turn historical evidence into task-constrained decisions about what research to pursue next.

Temporal Integrity in Evaluation Temporal integrity is essential when evaluating foresight: without a strict cutoff, systems can benefit from hindsight, leakage, or later-stabilized terminology rather than inference. ExAnte (Liu et al., 2026a), Set the Clock (Zhao et al., 2024), ForecastBench (Karger et al., 2025), FutureX Zeng et al. (2026), FOReCAst (Yuan et al., 2025), PROPHET (Tao et al., 2025), and MIRAI (Ye et al., 2024) all motivate time-sliced evaluation for future-oriented reasoning. Test of Time isolates temporal reasoning in synthetic settings (Fatemi et al., 2025), while AntiLeakBench constructs questions from updated real-world knowledge to reduce contamination (Wu et al., 2025). A line of work most closely related to ours uses future signals to evaluate generated research ideas, either by measuring their similarity to particular future papers (Wang et al., 2026) or by using future-impact signals (Jiang, 2026; Zhu et al., 2026; Ye et al., 2026). Compared to these approaches, ForeSci considers broader research judgment beyond a single idea, uses taxonomy-derived targets with human checks to separate historically inferable developments from arbitrary future events, and analyzes evidence-to-decision drift to explain how agents fail to connect available evidence to the appropriate research decision for the task at hand.

3 The ForeSci Framework

To systematically evaluate *forward-looking AI research judgement*, ForeSci simulates a retrospective forecasting environment. Models are tasked with making research decisions at a strict historical cutoff, utilizing only chronologically aligned evidence.

3.1 Problem Formulation

Let t denote a cutoff date, $\mathcal{K}_{\leq t}(q)$ denote the cutoff-aligned knowledge base constructed for question q (i.e., literature published up to t), and $\mathcal{G}_{>t}(q)$ denote the withheld validation targets derived from post-cutoff literature. A benchmark instance is

$$x = (q, t, \mathcal{K}_{\leq t}(q), f), \tag{1}$$

where f is the required task family. A system returns $a = \pi_\theta(q, \mathcal{K}_{\leq t}(q))$ using only the provided cutoff-aligned knowledge base; $\mathcal{G}_{>t}(q)$ is accessible only to evaluation. To control model-visible information, we disable web search and allow systems to use only $\mathcal{K}_{\leq t}(q)$ as external support when producing answers.

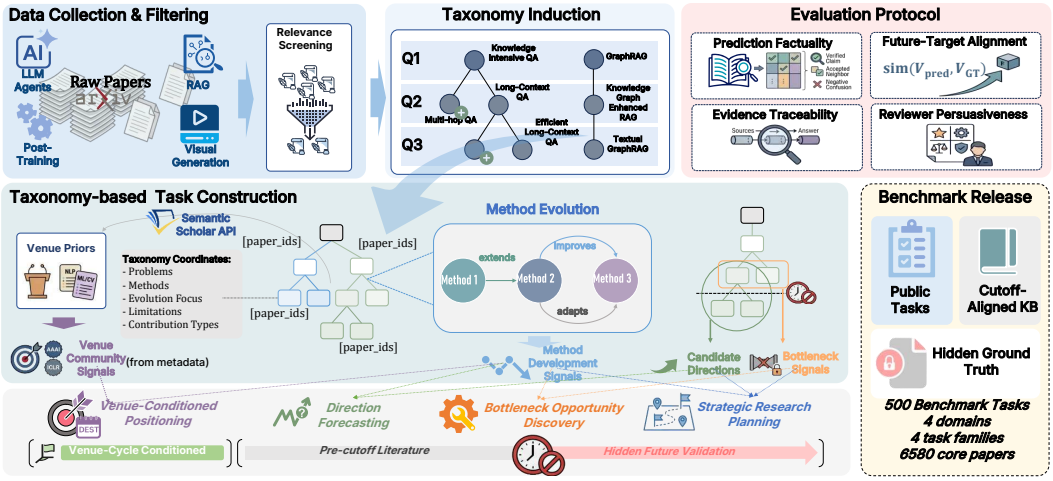

Figure 2: Construction process for the current formal ForeSci release. The figure shows the pipeline from corpus harvest and screening to temporal taxonomy induction, evidence and evolution asset construction, task-family builders, hidden validation targets, and the final benchmark release with public tasks and a paper knowledge base.

ForeSci instantiates this judgement problem through four task families: *Direction Forecasting*, *Bottleneck–Opportunity Discovery*, *Strategic Research Planning*, and *Venue-Conditioned Positioning*. Each family asks for a different research decision after t: predicting a concrete technical trajectory, identifying a bottleneck and the opportunity it unlocks, ranking candidate research directions under planning constraints, or positioning a project for an appropriate venue community.

3.2 Data Collection and Filtering

Figure 2 summarizes the construction pipeline. ForeSci is built from four rapidly evolving AI research areas: LLM agents, LLM fine-tuning and post-training, RAG and retrieval structuring, and visual generative modeling. For each area, we harvest candidate papers from arXiv [1] using domain-specific queries, enrich publication metadata with Semantic Scholar [2], deduplicate arXiv identifiers, and retain core/support papers after relevance and benchmark-core screening.

We apply two filtering stages to construct cutoff-aligned corpora. First, a domain-relevance screen removes papers that only match surface keywords. Second, a stricter benchmark-core screen identifies representative papers with central domain contributions and future-facing signals (*e.g.*, novel evaluation protocols, identified bottlenecks). Relevant but less central papers are retained as support papers, noisy or borderline cases are excluded. Finally, the processed corpus is chronologically truncated at the cutoff time t to form the public pre-cutoff knowledge base $\mathcal{K}_{\leq t}$. The specific cutoff date t varies across task instances, encompassing three-month (December 31, 2025), six-month (September 30, 2025), and venue-specific deadline settings after September 30, 2025. Domain and horizon statistics are reported in Table 1, while monthly paper-count statistics are provided in Figure A1. Additional construction details are provided in Appendix C.

3.3 Taxonomy Construction

To make the foresight problem both inferable and traceable, ForeSci models the evolution of AI research through taxonomy induction. This allows us to find specific research subdirections whose trajectories can be systematically deduced along the taxonomy and strictly grounded in historical evidence. We build on TaxoAdapt (Kargupta et al., 2025) to induce this taxonomy as a graph representation of the evolving research landscape. For each domain d and cutoff t, we induce a

[1] https://arxiv.org/
[2] https://www.semanticscholar.org/

Table 1: Domain and horizon statistics in ForeSci. KB document counts come from the cutoff-aligned offline knowledge base. Three-month and six-month counts follow task-level horizon metadata; venue-conditioned tasks use venue-cycle timing because their advisory and submission windows are venue specific.

Domain	KB Docs	3-month	6-month	Venue cycle	Total tasks
LLM Agents	2,769	25	90	23	138
LLM Fine-tuning and Post-training	2,131	15	69	15	99
RAG and Retrieval Structuring	767	10	69	13	92
Visual Generative Modeling and Diffusion	913	19	78	74	171
Total	6,580	69	306	125	500

temporal taxonomy

$$\mathcal{T}_{d,t} = (\mathcal{V}_{d,t}, \mathcal{E}_{d,t}), \tag{2}$$

where nodes represent research subdirections and edges represent method-evolution relations (see Figure A3 for illustrative examples). The taxonomy is dynamically expanded across sequential time slices of the cutoff-aligned corpus, preserving temporal causality to prevent future information leakage.

Node representation. Each node $v \in \mathcal{V}_{d,t}$ is aggregated from multiple cutoff-visible papers. For each node v, we construct a *node evidence record* that links the subdirection back to the cutoff-visible literature. The record mainly includes the representative papers and supporting papers. Each paper has a *full-text evidence* showing what problems, methods, evaluation focus, limitations, and contribution types had already appeared before t.

High-order signal extraction. From these node evidence records, we derive high-order signals for downstream task construction:

(1) *candidate directions*, which group one or more related nodes into coherent research options;

(2) *method-development signals* (Wu et al., 2026b), which record how methods, evaluations, or bottlenecks evolve over time (see Figure A4 for an example);

(3) *bottleneck signals*, which summarize recurring limitations, evaluation gaps, reliability or safety concerns, dataset or benchmark needs, and technical risks;

(4) *feasibility, dependency, and risk notes*, which record whether a candidate direction is actionable as a near-term research plan;

(5) *venue-community metadata*, which summarize publication and community context, including contribution style, maturity expectations, reviewer risks, and nearby venue contrasts.

All taxonomy structures are first extracted from cutoff-aligned evidence and then checked for support strength and temporal validity. In an upstream audit, three annotators assessed 50 sampled taxonomy edges and 50 sampled signals, yielding mean accuracies of 90.7% for taxonomy induction and 92.3% for signal extraction. Every downstream task and target record is additionally reviewed for temporal validity, leakage risk, specificity, support strength, and task clarity.

3.4 Task Families

We derive four task families from the taxonomy. Task instances are constructed through a human–LLM collaborative process: an LLM first drafts candidate questions, options, and answers from the taxonomy-derived evidence records. Human experts then inspect the source evidence, check cutoff validity and leakage risk, revise unclear or weakly grounded items, and approve each final instance, ensuring that the benchmark instances reflect expert-validated scientific value and foresight challenges.

Direction Forecasting. This family asks the system to choose, from a fixed set of *candidate directions*, which direction is most likely to gain momentum in the post-cutoff window. The task

q is grounded in *node evidence records* and *method-development signals* before t. The hidden future validation targets $\mathcal{G}_{>t}(q)$ are candidate directions (including primary directions and acceptable neighbors) with trajectory labels (e.g., ACCELERATING, STEADY). They are initially induced by the *method-development signals* after cutoff, combined with the necessary evidence before the cutoff.

Bottleneck–Opportunity Discovery. This family asks the system to identify one root bottleneck in a cutoff-visible research subdirection and explain what one-hop opportunity would open if that bottleneck were reduced. The task q is grounded in *bottleneck signals*, *full-text evidence*, and *method-development signals*. The hidden future validation targets $\mathcal{G}_{>t}(q)$ are bottleneck–opportunity pairs, including primary bottlenecks, acceptable bottleneck variants, unlocked opportunities, and mechanism descriptions. They are initially induced from *bottleneck signals* before t and *method-development signals* after t, combined with the necessary evidence before the cutoff.

Strategic Research Planning. This family asks the system to rank a fixed set of research options for a hypothetical team making a near-term research plan at the cutoff. The task q is derived from *candidate directions* and *node evidence records*, *method-development signals*, *bottleneck signals*, *feasibility, dependency, and risk notes* before t. The hidden future validation targets $\mathcal{G}_{>t}(q)$ are ranked *candidate directions*, including the preferred ordering, top-priority option, rationale units, milestones, dependencies, risks, and go/no-go criteria. They are initially induced from post-cutoff *method-development signals* and *bottleneck signals*, combined with the necessary evidence and *feasibility, dependency, and risk notes* before cutoff.

Venue-Conditioned Positioning. This family asks the system to position a proposed contribution for a target venue cycle. Given a project description and five candidate venue families, the system selects one default target and up to three alternatives, each with an explicit condition for suitability. It also states the contribution center, three evidence upgrades, and two venue-fit risks. Task construction draws on contribution types from *full-text evidence* and *venue-community metadata*. The hidden future validation targets $\mathcal{G}_{>t}(q)$ are venue-positioning decisions. They are induced from *venue-community metadata* and post-cutoff *method-development signals* that reflect community expectations, combined with the necessary evidence before the cutoff.

Across all families, public questions do not expose internal taxonomy information or post-cutoff outcomes. The formal release contains 125 tasks for each family. Additional details on the benchmark construction are provided in Appendix C. For tasks with acceptable alternatives, the hidden targets retain multiple candidates or variants: all 125 Direction tasks include acceptable neighbors (mean 4.43 per task); 94/125 Bottleneck tasks include an acceptable bottleneck variant (mean 3.44 when nonempty), and 86/125 include an acceptable opportunity variant (mean 4.06 when nonempty); all 125 Venue tasks include acceptable venue-neighborhood entries (mean 4.10). Construction and scoring details given in Appendix C and Appendix D.1.

4 EVALUATION

4.1 METRICS

For each public question q, system answer a, pre-cutoff support packet $\mathcal{E}_{\leq t}(q)$, and hidden future validation targets $\mathcal{G}_{>t}(q)$, we report four complementary metrics. These metrics separately assess whether the answer states correct future facts, reaches a conclusion consistent with the future target, grounds its reasoning in visible pre-cutoff evidence, and presents a judgment persuasive to a reviewer.

Prediction Factuality (*Fact*). This metric evaluates whether the answer makes claims that are supported by $\mathcal{G}_{>t}(q)$. Following the atomic-fact view of FACTSCORE (Min et al., 2023), we extract atomic claims $\mathcal{C}(a)$ from the answer. We also define a hidden claim bank $\mathcal{C}^*(q) \subset \mathcal{G}_{>t}(q)$: a set of task-relevant atomic validation claims derived from hidden future validation targets. Prediction Factuality is their claim-level F1.

Future-Target Alignment (*FTA*). This metric evaluates alignment with the family-specific future target in $\mathcal{G}_{>t}(q)$. Direction and Bottleneck use reference-guided claim–target matching; Planning uses deterministic ranking alignment. Venue uses a task-specific structured target because several venue

Table 2: Overall results on ForeSci in terms of Prediction Factuality (Fact), Future-Target Alignment (FTA), Evidence Traceability Score (Trace), and Reviewer Persuasiveness (Pers). Bold marks the best method within the same backbone and metric. Table A11 gives five-run stability views. Table A8 summarizes the aggregate method contrasts.

	Qwen3-235B				GPT-5.2				GLM-4.6				Gemini-3			
Method	Fact.	FTA	Trace	Pers	Fact.	FTA	Trace	Pers	Fact.	FTA	Trace	Pers	Fact.	FTA	Trace	Pers
Native LLM	0.609	0.634	–	0.591	0.615	0.619	–	0.688	0.500	0.586	–	0.503	0.559	0.627	–	0.601
Hybrid RAG	0.591	0.625	0.579	0.604	0.618	0.624	0.560	0.725	0.525	0.617	0.522	0.545	0.568	0.623	0.499	0.637
CoI	**0.612**	0.637	0.580	0.640	0.632	0.640	0.648	**0.748**	**0.555**	**0.657**	0.538	**0.569**	0.575	0.624	0.546	0.642
ResearchAgent	0.608	**0.655**	0.569	0.636	**0.638**	0.636	**0.653**	0.741	0.552	0.639	0.557	0.555	**0.581**	**0.637**	0.533	**0.646**
ARIS	0.596	0.641	**0.632**	**0.649**	0.621	**0.648**	0.642	0.743	0.542	0.623	**0.609**	0.562	0.580	0.626	**0.584**	0.619

communities may be appropriate under different contribution framings. It scores canonical-family agreement, conditional-alternative recall, and contribution-center alignment. Full definitions and the Venue adaptation are given in Appendix D.1.

Evidence Traceability Score (*Trace*). This metric measures whether the answer's key decisions are supported by pre-cutoff evidence. For each task, we extract three fixed, family-specific decision slots from the final answer and retrieve the three most similar evidence chunks per slot from the system's pre-cutoff support packet. The evaluator assigns each slot a continuous score in $[0, 1]$ based on the strength of evidence support, the evidence-to-decision bridge, and unresolved conflict. Trace is the mean of the three slot scores; Appendix D.1 defines the slots and scoring criteria. Answers without external support artifacts are reported as not applicable.

Reviewer Persuasiveness (*Pers*). This metric estimates how persuasive an answer is to a reviewer making a research decision. To score model-generated answers at scale, we train an XGBoost-based proxy for reviewer evaluation (Chen & Guestrin, 2016), following the component-based learned-evaluation approach of USR and P2P-Eval (Mehri & Eskenazi, 2020; Sun et al., 2026). We select a balanced sample of 1,200 answers, each independently rated by three annotators on a 1–7 rubric using only the public task and anonymized final answer. We train a separate shallow regressor for each task family to predict the mean rating from semantic content-quality and answer-form features, then normalize it to $[0, 1]$. Backbone, generation method, answer length, family identity, and hidden targets are excluded from X. Five-fold task-grouped out-of-fold validation yields Spearman $\rho = 0.726$, MAE 0.551 points, and within-task pairwise accuracy 0.819. The annotators show strong ordinal agreement (mean pairwise Spearman $\rho = 0.818$ and quadratic-weighted kappa $= 0.821$). Appendix D.4 provides the full validation, and Tables A4–A5 list the exact family-specific predictors. Figure A5 summarizes the complete annotation, modeling, and scoring pipeline.

4.2 Models, Systems, and Adaptation

We evaluate five systems: **Native LLM** without retrieval, **Hybrid RAG** with sparse+dense retrieval, and three offline-adapted agentic systems: **CoI-style**, **ResearchAgent-style**, and **ARIS-style**. We adapt the agentic systems to ForeSci by constraining retrieval, tool use, and memory to the offline knowledge base and by rendering final answers through task-family-specific output schemas. Detailed adaptation notes are in Appendix E. We evaluate Qwen3-235B (Qwen Team, 2025), GPT-5.2 (OpenAI, 2025), GLM-4.6 (Z.AI, 2025), and Gemini-3 (Google, 2025). To prevent information leakage, we use only model versions with knowledge cutoff dates before September 30, 2025. All methods receive the same cutoff-aligned external evidence boundary.

5 Results

5.1 Evaluation of LLM Agents

Overall results Table 2 reports Fact, FTA, Trace, and Pers across four backbones and five methods. On average, agentic methods improve evidence traceability over Hybrid RAG and reviewer persuasiveness over native LLMs. Their gains over native LLMs on Fact and FTA, the two metrics

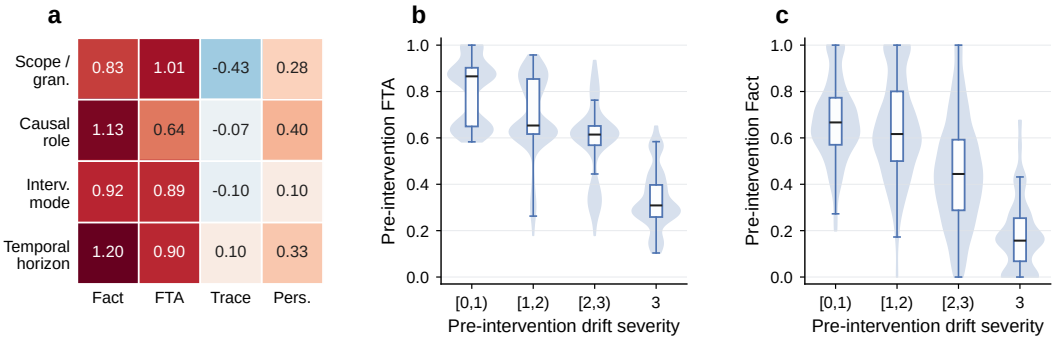

Figure 3: Evidence-to-decision drift and metric degradation. (a) Normalized metric drops associated with four drift dimensions; positive values indicate lower scores under moderate-to-severe drift. (b,c) FTA and Fact distributions across four mean-drift severity bins in the held-out sample.

that directly assess alignment with future outcomes, are modest (Table A8). These gains are also task-family dependent: no agentic method is uniformly strongest across metrics, backbones, and task families, and in some settings, agentic methods offer no clear advantage over native LLMs and even underperform Hybrid RAG (Table A10). These findings reveal an important mismatch. Although agentic methods can make greater use of pre-cutoff evidence and produce research-decision reports that appear more persuasive to reviewers, the decisions themselves are not necessarily more forward-looking. This mismatch creates a persuasive yet potentially misleading form of pseudo-foresight. For example, a Qwen3-235B CoI-style planning answer achieves Trace = 0.892 but FTA = 0.200 because it ranks embodied evaluation last due to immature tooling, whereas the reference treats the same evaluation gap as the top-priority bottleneck (Figure A8). Research decisions generated by agents therefore warrant particular scrutiny, and this failure mode necessitates a closer analysis of why agentic systems produce such outputs.

Output-level temporal-leakage audit Because the knowledge boundaries of proprietary model versions are based primarily on provider statements rather than transparently inspectable checkpoints, we additionally audit outputs from GPT-5.2 and Gemini-3 for output-level temporal leakage. Specifically, we examine all 5,000 outputs from these two proprietary backbones for explicit post-cutoff named entities. The initial screen identifies 57 candidate mentions in 56 GPT-5.2 answers and 60 candidate mentions in 57 Gemini-3 answers. Manual adjudication confirms that none constitutes leakage; instead, the apparent overlaps correspond to prospective proposals, providing further support for the effectiveness of our temporal controls (Appendix C.8 and Table A3).

5.2 Analysis and Intervention of Evidence-to-Decision Drift

To understand the mismatch between traceability and future-facing Fact/FTA performance, we examine where evidence-grounded answers diverge from future-aligned research decisions and whether targeted controls can reduce that divergence.

Drift discovery and metric coupling Using a blinded 20-task discovery split for each family, we compare model answers with reference decisions across all three agent-based methods and four backbones. This comparison reveals four recurring forms of *evidence-to-decision drift*: *scope/granularity drift* selects the wrong level of specificity; *causal-role drift* confuses a cause, symptom, or opportunity; *intervention-mode drift* recommends the wrong kind of action; and *temporal-horizon drift* targets the wrong maturity stage. Each dimension is scored from 0 to 3, with higher values indicating a larger departure from the reference decision. Full annotation details and examples are provided in Appendix G.2.

We quantify each drift–metric association as the normalized score drop $\Delta_{\text{norm}}(m) = (\mathbb{E}[m \mid s = 0] - \mathbb{E}[m \mid s \geq 2])/\text{SD}(m)$, where m is the target metric and s is drift severity. The four drift dimensions are strongly coupled to the metrics that directly evaluate decision quality, but much less consistently to surface credibility. Moderate-to-severe drift is associated with substantial drops in

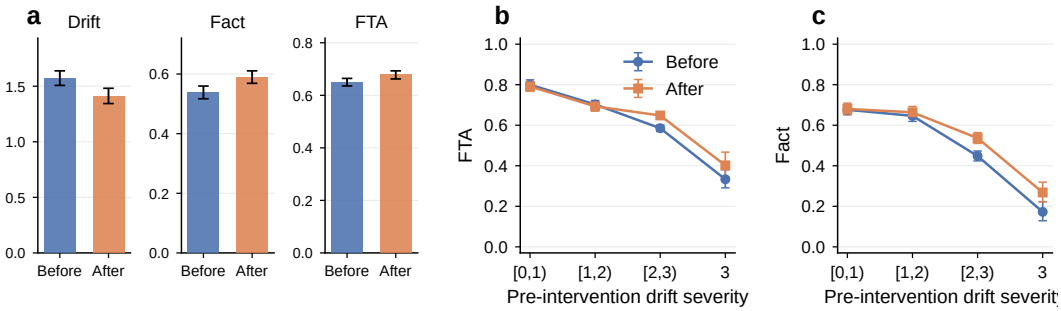

Figure 4: Evidence-to-decision intervention results. (a) Overall mean Drift, Fact, and FTA before and after intervention. (b,c) Mean FTA and Fact before and after intervention, stratified by pre-intervention Drift severity. Error bars denote 95% task-cluster bootstrap confidence intervals.

both Fact and FTA across all four dimensions, whereas its relationships with Trace and Pers are weak or mixed (Figure 3a). This helps explain the mismatch between traceability and future-facing Fact/FTA performance: an answer can remain traceable or persuasive while assigning the evidence to the wrong decision object, causal role, action, or time horizon. This pattern also extends beyond the discovery split. For example, in the held-out Qwen3-235B confirmation sample, mean Drift severity is strongly and monotonically associated with lower FTA and Fact. Spearman correlation is -0.611 for FTA (95% task-cluster bootstrap CI $[-0.658, -0.559]$) and -0.561 for Fact ($[-0.610, -0.509]$) (Figure 3b,c).

Task-conditioned controls To test whether decision checks reduce drift and improve FTA and Fact, we introduce an intervention between the evidence and decision stages. Specifically, we add task-specific decision checks before final answer rendering, with task-specific implementations. For Bottleneck, after collecting evidence, the check requires the agent to distinguish causal roles (e.g, observed failures, proximal mechanisms, and upstream causes) before identifying the bottleneck. For Direction, it checks whether new evidence matches the forecast's conceptual scope and temporal horizon. For Planning, it compares candidate studies at a common scope and retains only evidence that supports the decisive next-action comparison. For Venue, it checks that positioning remains anchored in the project's current contribution rather than a broader topical scope or proposed future work. Agent-specific prompts, commit rules, and evaluation details are deferred to the appendix (Appendix G.3).

Intervention effects After intervention, mean Drift decreases by 0.161 (95% task-cluster bootstrap CI $[0.125, 0.198]$), while Fact and FTA increase by 0.051 ($[0.038, 0.064]$) and 0.028 ($[0.015, 0.040]$), respectively; all three intervals exclude zero (Figure 4a).

The magnitude of these gains depends on initial Drift severity. For answers with low initial Drift, FTA and Fact change only slightly. The separation between before and after becomes pronounced once initial Drift reaches the moderate-to-severe range, with the largest Fact and FTA gains concentrated in the two highest-severity bins (Figure 4b,c). This result suggests a potential direction for future work: selectively gating research-agent revisions on decision-level risk, so that additional controls focus on likely drift without unnecessarily perturbing already aligned decisions.

6 Conclusion

ForeSci evaluates whether LLM agents can turn historically available evidence into forward-looking AI research judgements. Its 500 cutoff-controlled tasks pair offline knowledge bases with hidden post-cutoff validation targets across four decision families. Agentic workflows improve traceability over Hybrid RAG and yield modest average gains in future-target alignment over native LLMs, with substantial variation across tasks and backbones. The drift analysis identifies failures in connecting evidence to the final decision, and task-specific controls improve alignment. By separating factual support, future-target alignment, traceability, and reviewer-style persuasiveness, ForeSci makes these failures measurable.

AI USE STATEMENT

LLMs used as evaluated systems or benchmark evaluators are documented in the corresponding methodology sections. Separately, generative AI assistants supported software prototyping and debugging, experiment orchestration, preliminary result checking, language editing, translation, LaTeX formatting, and manuscript drafting. The authors independently reviewed and verified all AI-assisted code, analyses, results, and text. The benchmark methodology, data processing, experimental procedures, and reported results were checked by the authors, who take responsibility for the final manuscript, including its claims, analyses, and artifacts.

REPRODUCIBILITY STATEMENT

Section 3 provides the formal benchmark and task definitions, and Section 4 specifies the metrics, model and system settings, and evaluation protocol. Appendix C documents benchmark construction, temporal controls, selection parameters, and artifact separation; Appendix D gives metric calibration, human annotation, validation analyses, and evaluation prompts; and Appendix E describes the agent adaptations and comparison controls. The release manifest in Appendix B enumerates the benchmark tasks, cutoff-aligned knowledge bases, evaluation goldsets, prompts, schemas, scripts, run manifests, checksums, and result-building artifacts. Code and supplementary materials will be released in anonymized form for review and made publicly available upon publication.

ETHICS STATEMENT

ForeSci is constructed from publicly available scholarly artifacts and does not use private or privacy-sensitive user data. Human-validation results are reported only in aggregate, and individual annotator records are not released. The benchmark is intended to evaluate automated research systems under controlled historical evidence, not to automate peer review, venue selection, hiring, funding, or research-prioritization decisions. Its outputs are evaluation artifacts rather than validated predictions or submission-ready manuscripts, and subsequent use requires human verification and responsibility.

REFERENCES

Anirudh Ajith, Amanpreet Singh, Jay DeYoung, Nadav Kunievsky, Austin C. Kozlowski, Oyvind Tafjord, James Evans, Daniel S. Weld, Tom Hope, and Doug Downey. Prescience: A benchmark for forecasting scientific contributions. *arXiv preprint arXiv:2602.20459*, 2026. doi: 10.48550/arXiv.2602.20459. URL https://arxiv.org/abs/2602.20459.

Jonathan Bragg, Mike D'Arcy, Nishant Balepur, Dan Bareket, Bhavana Dalvi Mishra, Sergey Feldman, Dany Haddad, Jena D. Hwang, Peter Jansen, Varsha Kishore, Bodhisattwa Prasad Majumder, Aakanksha Naik, Sigal Rahamimov, Kyle Richardson, Amanpreet Singh, Harshit Surana, Aryeh Tiktinsky, Rosni Vasu, Guy Wiener, Chloe Anastasiades, Stefanus Candra, Jason Dunkelberger, Daniel Emery, Rob Evans, Malachi Hamada, Regan Huff, Rodney Kinney, Matt Latzke, Jaron Lochner, Ruben Lozano-Aguilera, Ngoc-Uyen Nguyen, Smita Rao, Amber Tanaka, Brooke Vlahos, Peter Clark, Doug Downey, Yoav Goldberg, Ashish Sabharwal, and Daniel S. Weld. Astabench: Rigorous benchmarking of AI agents with a scientific research suite. In *International Conference on Learning Representations*, pp. 110136–110223, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/b2ce9568dbb559aefc8c98ca5b5314ce-Abstract-Conference.html.

Center for AI Safety, Scale AI, and HLE Contributors Consortium. A benchmark of expert-level academic questions to assess AI capabilities. *Nature*, 649:1139–1146, 2026.

Hui Chen, Miao Xiong, Yujie Lu, Wei Han, Ailin Deng, Yufei He, Jiaying Wu, Yibo Li, Yue Liu, and Bryan Hooi. Mlr-bench: Evaluating ai agents on open-ended machine learning research. *Advances in Neural Information Processing Systems*, 38, 2025.

Tianqi Chen and Carlos Guestrin. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 785–794, 2016. doi: 10.1145/2939672.2939785. URL https://doi.org/10.1145/2939672.2939785.

Bahare Fatemi, Seyed Mehran Kazemi, Anton Tsitsulin, Karishma Malkan, Jinyeong Yim, John Palowitch, Sungyong Seo, Jonathan Halcrow, and Bryan Perozzi. Test of time: A benchmark for

evaluating LLMs on temporal reasoning. In *International Conference on Learning Representations*, pp. 94426–94447, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/eb7295a8bc613b375726659c2ecd6f14-Abstract-Conference.html.

Olivier Francois. Arbitrariness of peer review: A bayesian analysis of the NIPS experiment. *arXiv preprint arXiv:1507.06411*, 2015. doi: 10.48550/arXiv.1507.06411. URL https://arxiv.org/abs/1507.06411.

Ali Essam Ghareeb, Benjamin Chang, Ludovico Mitchener, Angela Yiu, Caralyn J. Szostkiewicz, Dmytro Shved, Gavin J. Gyimesi, Jon M. Laurent, Samantha M. Wright, Muhammed T. Razzak, Andrew D. White, Silvia C. Finnemann, Michaela M. Hinks, and Samuel G. Rodriques. A multi-agent system for automating scientific discovery. *Nature*, 655(8122):497–505, 2026. doi: 10.1038/s41586-026-10652-y. URL https://www.nature.com/articles/s41586-026-10652-y.

Google. Gemini models. https://ai.google.dev/gemini-api/docs/models, 2025. Accessed 2026-05-26.

Mourad Gridach, Jay Nanavati, Christina Mack, Khaldoun Zine El Abidine, and Lenon Mendes. Agentic AI for scientific discovery: A survey of progress, challenges, and future directions. In *ICLR 2025 Workshops: AgenticAI*, 2025. URL https://mlanthology.org/iclrw/2025/gridach2025iclrw-agentic/.

Xuemei Gu and Mario Krenn. Forecasting high-impact research topics via machine learning on evolving knowledge graphs. *Machine Learning: Science and Technology*, 6(2):025041, 2025. doi: 10.1088/2632-2153/add6ef. URL https://doi.org/10.1088/2632-2153/add6ef.

Joy He-Yueya, Anikait Singh, Ge Gao, Michael Y. Li, Sherry Yang, Chelsea Finn, Emma Brunskill, and Noah D. Goodman. Giants: Generative insight anticipation from scientific literature. *arXiv preprint arXiv:2604.09793*, 2026.

Peter Jansen, Samiah Hassan, and Ruoyao Wang. Matter-of-fact: A benchmark for verifying the feasibility of literature-supported claims in materials science. In *Empirical Methods in Natural Language Processing*, 2025.

Bo Jiang. HindSight: Evaluating LLM-generated research ideas via future impact. *arXiv preprint arXiv:2603.15164*, 2026. doi: 10.48550/arXiv.2603.15164. URL https://arxiv.org/abs/2603.15164.

Ezra Karger, Houtan Bastani, Yueh-Han Chen, Zachary Jacobs, Danny Halawi, Fred Zhang, and Philip Tetlock. Forecastbench: A dynamic benchmark of AI forecasting capabilities. In *International Conference on Learning Representations*, 2025.

Priyanka Kargupta, Nan Zhang, Yunyi Zhang, Rui Zhang, Prasenjit Mitra, and Jiawei Han. Taxoadapt: Aligning llm-based multidimensional taxonomy construction to evolving research corpora. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 29834–29850, 2025.

Jakub Lála, Odhran O'Donoghue, Aleksandar Shtedritski, Sam Cox, Samuel G. Rodriques, and Andrew D. White. Paperqa: Retrieval-augmented generative agent for scientific research. *arXiv preprint arXiv:2312.07559*, 2023. doi: 10.48550/arXiv.2312.07559. URL https://arxiv.org/abs/2312.07559.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Advances in Neural Information Processing Systems*, volume 33, pp. 9459–9474. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/hash/6b493230205f780e1bc26945df7481e5-Abstract.html.

Long Li, Weiwen Xu, Jiayan Guo, Ruochen Zhao, Xingxuan Li, Yuqian Yuan, Boqiang Zhang, Yuming Jiang, Yifei Xin, Ronghao Dang, Yu Rong, Deli Zhao, Tian Feng, and Lidong Bing. Chain of ideas: Revolutionizing research via novel idea development with LLM agents. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 8971–9004. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.findings-emnlp.477. URL https://aclanthology.org/2025.findings-emnlp.477/.

Yachuan Liu, Xiaochun Wei, Lin Shi, Xinnuo Li, Bohan Zhang, Paramveer S Dhillon, and Qiaozhu Mei. Exante: A benchmark for ex-ante inference in large language models. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1551–1571, 2026a.

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. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 13187–13207. Association for Computational Linguistics, 2026b. doi: 10.18653/v1/2026.findings-acl.644. URL https://aclanthology.org/2026.findings-acl.644/.

Chris Lu, Cong Lu, Robert Tjarko Lange, Yutaro Yamada, Shengran Hu, Jakob Foerster, David Ha, and Jeff Clune. Towards end-to-end automation of ai research. *Nature*, 651(8107):914–919, 2026.

Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. In *Advances in Neural Information Processing Systems*, 2022.

Alisia Lupidi, Bhavul Gauri, Thomas Simon Foster, Bassel Al Omari, Despoina Magka, Alberto Pepe, Alexis Audran-Reiss, Muna Aghamelu, Nicolas Baldwin, Lucia Cipolina-Kun, Jean-Christophe Gagnon-Audet, Chee Hau Leow, Sandra Lefdal, Hossam Mossalam, Abhinav Moudgil, Saba Nazir, Emanuel Tewolde, Isabel Urrego, Jordi Armengol Estape, Amar Budhiraja, Gaurav Chaurasia, Abhishek Charnalia, Derek Dunfield, Karen Hambardzumyan, Daniel Izcovich, Martin Josifoski, Ishita Mediratta, Kelvin Niu, Parth Pathak, Michael Shvartsman, Edan Toledo, Anton Protopopov, Roberta Raileanu, Alexander Miller, Tatiana Shavrina, Jakob Foerster, and Yoram Bachrach. AIRS-bench: a suite of tasks for frontier AI research science agents. *arXiv preprint arXiv:2602.06855*, 2026. doi: 10.48550/arXiv.2602.06855. URL https://arxiv.org/abs/2602.06855.

Shikib Mehri and Maxine Eskenazi. USR: An unsupervised and reference-free evaluation metric for dialog generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 681–707, 2020. doi: 10.18653/v1/2020.acl-main.64. URL https://aclanthology.org/2020.acl-main.64/.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. Factscore: Fine-grained atomic evaluation of factual precision in long form text generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 12076–12100, 2023.

OpenAI. GPT-5.2. https://openai.com/index/introducing-gpt-5-2/, 2025. Accessed 2026-05-26.

Qwen Team. Qwen3: Think deeper, act faster. https://qwenlm.github.io/blog/qwen3/, April 2025. Accessed 2026-05-26.

Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Eric Hambro, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. Toolformer: Language models can teach themselves to use tools. *Advances in neural information processing systems*, 36:68539–68551, 2023.

Udari Madhushani Sehwag, Elaine Lau, Haniyeh Ehsani Oskouie, Shayan Shabihi, Erich Liang, Andrea Toledo, Guillermo Mangialardi, Sergio Fonrouge, Ed-Yeremai Hernandez Cardona, Paula Vergara, Utkarsh Tyagi, Chen Bo Calvin Zhang, Pavi Bhatter, Nicholas Johnson, Furong Huang, Ernesto Gabriel Hernandez Montoya, and Bing Liu. Scipredict: Can llms predict the outcomes of scientific experiments in natural sciences? *arXiv preprint arXiv:2604.10718*, 2026.

Tao Sun, Enhao Pan, Zhengkai Yang, Kaixin Sui, Jiajun Shi, Xianfu Cheng, Tongliang Li, Ge Zhang, Wenhao Huang, Jian Yang, and Zhoujun Li. P2P: Automated paper-to-poster generation and fine-grained benchmark. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/1fe6f635fe265292aba3987b5123ae3d-Abstract-Conference.html.

Jiabin Tang, Lianghao Xia, Zhonghang Li, and Chao Huang. Ai-researcher: Autonomous scientific innovation. *Advances in Neural Information Processing Systems*, 38:9481–9520, 2025.

Zhengwei Tao, Pu Wu, Zhi Jin, Xiaoying Bai, Haiyan Zhao, Chengfeng Dou, Xiancai Chen, Jia Li, Linyu Li, Chongyang Tao, and Wentao Zhang. PROPHET: An inferable future forecasting benchmark with causal intervened likelihood estimation. *arXiv preprint arXiv:2504.01509*, 2025.

Yuwei Wan, Yixuan Liu, Aswathy Ajith, Clara Grazian, Bram Hoex, Wenjie Zhang, Chunyu Kit, Tong Xie, and Ian Foster. Sciqag: A framework for auto-generated science question answering dataset with fine-grained evaluation. *arXiv preprint arXiv:2405.09939*, 2024. doi: 10.48550/arXiv.2405.09939. URL https://arxiv.org/abs/2405.09939.

Daoyu Wang, Mingyue Cheng, Shuo Yu, Zirui Liu, Ze Guo, Xin Li, and Qi Liu. Paperarena: An evaluation benchmark for tool-augmented agentic reasoning on scientific literature. *arXiv preprint arXiv:2510.10909*, 2025. doi: 10.48550/arXiv.2510.10909. URL https://arxiv.org/abs/

2510.10909.

Heng Wang, Pengcheng Jiang, Jiashuo Sun, Zhiyi Shi, Haofei Yu, Jiawei Han, and Heng Ji. Learning to predict future-aligned research proposals with language models. *arXiv preprint arXiv:2603.27146*, 2026. doi: 10.48550/arXiv.2603.27146. URL https://arxiv.org/abs/2603.27146.

Sean Wu, Pan Lu, Yupeng Chen, Jonathan Bragg, Yutaro Yamada, Peter Clark, David A. Clifton, Philip Torr, James Zou, and Junchi Yu. Forecasting scientific progress with artificial intelligence. *arXiv preprint arXiv:2605.22681*, 2026a.

Xiaobao Wu, Liangming Pan, Yuxi Xie, Ruiwen Zhou, Shuai Zhao, Yubo Ma, Mingzhe Du, Rui Mao, Anh Tuan Luu, and William Yang Wang. Antileakbench: Preventing data contamination by automatically constructing benchmarks with updated real-world knowledge. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 18403–18419, 2025. doi: 10.18653/v1/2025.acl-long.901. URL https://aclanthology.org/2025.acl-long.901/.

Yujun Wu, Dongxu Zhang, Xinchen Li, Jinhang Xu, Yiling Duan, Yumou Liu, Jiabao Pan, Qiyuan Zhu, Xuanhe Zhou, Jingxuan Wei, Siyuan Li, Jintao Chen, Conghui He, and Cheng Tan. Internatlas: A methodological evolution graph as research infrastructure for AI scientists. *arXiv preprint arXiv:2604.28158*, 2026b. doi: 10.48550/arXiv.2604.28158. URL https://arxiv.org/abs/2604.28158.

Yutaro Yamada, Robert Tjarko Lange, Cong Lu, Shengran Hu, Chris Lu, Jakob Foerster, Jeff Clune, and David Ha. The AI scientist-v2: Workshop-level automated scientific discovery via agentic tree search. *arXiv preprint arXiv:2504.08066*, 2025. doi: 10.48550/arXiv.2504.08066. URL https://arxiv.org/abs/2504.08066.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations (ICLR)*, 2023. URL https://openreview.net/forum?id=WE_vluYUL-X.

Bingyang Ye, Shan Chen, Jingxuan Tu, Chen Liu, Zidi Xiong, Samuel Schmidgall, and Danielle S. Bitterman. Proof of time: A benchmark for evaluating scientific idea judgments. *arXiv preprint arXiv:2601.07606*, 2026.

Chenchen Ye, Ziniu Hu, Yihe Deng, Zijie Huang, Mingyu Derek Ma, Yanqiao Zhu, and Wei Wang. MIRAI: Evaluating LLM agents for event forecasting. *arXiv preprint arXiv:2407.01231*, 2024. doi: 10.48550/arXiv.2407.01231. URL https://arxiv.org/abs/2407.01231.

Zhangdie Yuan, Zifeng Ding, and Andreas Vlachos. Introducing FOReCAst: The future outcome reasoning and confidence assessment benchmark. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-0808. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/2277c7c3b112fcc6031a6f0d832df2a0-Abstract-Datasets_and_Benchmarks_Track.html. Datasets and Benchmarks Track.

Z.AI. GLM-4.6. https://docs.z.ai/guides/llm/glm-4.6, 2025. Accessed 2026-09-18.

Zhiyuan Zeng, Jiashuo Liu, Siyuan Chen, Tianci He, Yali Liao, Yixiao Tian, Jinpeng Wang, Zaiyuan Wang, Yang Yang, Lingyue Yin, Mingren Yin, Zhenwei Zhu, Tianle Cai, Zehui Chen, Jiecao Chen, Yantao Du, Xiang Gao, Jiacheng Guo, Liang Hu, Jianpeng Jiao, Xiangsheng Li, Jingkai Liu, Shuang Ni, Zhoufutu Wen, Ge Zhang, Kaiyuan Zhang, Xin Zhou, Jose Blanchet, Xipeng Qiu, Mengdi Wang, and Wenhao Huang. FutureX: An advanced live benchmark for LLM agents in future prediction. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=z28PLIEj6l.

Bowen Zhao, Zander Brumbaugh, Yizhong Wang, Hannaneh Hajishirzi, and Noah A Smith. Set the clock: Temporal alignment of pretrained language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 15015–15040, 2024.

Hangxiao Zhu, Yuyu Zhang, Ping Nie, and Yu Zhang. SciImpact: A multi-dimensional, multi-field benchmark for scientific impact prediction. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 28923–28936. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.findings-acl.1445. URL https://aclanthology.org/2026.findings-acl.1445/.

APPENDIX CONTENTS

A Limitations **15**

B Responsible Research and Reproducibility **15**

C Benchmark Construction and Temporal Integrity **15**

C.1 Corpus Filtering and Temporal Freezing 15

C.2 Cutoff Slicing and Forecast-Oriented Taxonomy Adaptation 16

C.3 Candidate Direction Selection 16

C.4 Method-Development and Bottleneck Signals 16

C.5 Venue-Community Profiles 16

C.6 Human–LLM Collaborative Audit 17

C.7 Task Curation and Artifact Separation 17

C.8 Output-Level Temporal-Leakage Audit 17

D Evaluation Protocol and Validation **20**

D.1 Metric Calibration Details 20

D.2 Human Annotation and PERS Modeling 22

D.3 End-to-End Worked Example 23

D.4 Validation Analyses 23

D.4.1 Human Annotator Agreement 23

D.4.2 Reviewer Persuasiveness Model Validation 24

D.4.3 Aggregate Method Contrasts 24

D.4.4 Claim Extraction Human Validation 24

D.5 Evaluation Prompt Templates 26

E System Adaptation and Comparison Controls **27**

F Supplementary Results and Evaluator Analyses **28**

F.1 Evaluation Results by Task Family 28

F.2 Auxiliary Evaluator Stability 28

F.3 Answer Lengths 29

F.4 Content-Metric Correlation 29

G Error and Mechanism Diagnostics **30**

G.1 Detailed Error Analysis 30

G.2 Drift–Metric Associations 31

G.3 Task-specific Drift Mechanisms 31

A LIMITATIONS

Scope and transfer. ForeSci is intentionally a cutoff-controlled benchmark for forward-looking research judgement. The current release covers four fast-moving AI areas, four decision families, and publicly documented evidence. This design keeps each task historically auditable and fixes the information boundary at which the decision is made. The reported results therefore characterize performance under this evidence regime; extending the benchmark to additional domains, languages, forecast horizons, or less documented settings provides a natural transfer test.

Research-question boundary. ForeSci focuses on fast-growing AI areas, where changes in methods, bottlenecks, evaluation practices, and venue expectations provide observable temporal structure for a forward-looking planning problem. It uses a graph-structured temporal taxonomy and its derived method-development signals, bottleneck signals, feasibility and risk notes, venue-community metadata, and cutoff-local paper evolution as proxy signals for prospective planning. The benchmark therefore asks whether an agent can convert cutoff-visible evidence into decisions that align with subsequent developments under this controlled setting. This is a boundary of the research question, rather than a claim that slower-growing areas lack scientific value, research importance, or a need for careful planning.

B RESPONSIBLE RESEARCH AND REPRODUCIBILITY

Artifacts, licenses, and intended use. ForeSci releases benchmark tasks, evaluation goldsets, prompts, schemas, scripts, and a cutoff-aligned scholarly knowledge base for comparing research-assistant systems under temporal controls. Source papers and bibliographic records retain their original access conditions; derived artifacts are subject to the repository and release terms.

Table A1: ForeSci release manifest. Evaluation-only targets are physically and logically separated from model-visible prediction-time artifacts.

Artifact group	Released contents
Prediction-time package	Public task files, cutoff and horizon metadata, and cutoff-aligned offline knowledge base
Evaluation package	Evaluation-only hidden targets and goldsets, metric prompts, schemas, and evaluator scripts
Construction package	Corpus filtering, taxonomy induction, task construction, and update scripts
Audit package	Run manifests, file checksums, aggregate validation summaries, and result-building scripts

Privacy and content review. The benchmark uses public scholarly artifacts. Released tasks contain public question text and generation metadata; hidden evaluation targets are packaged separately. Human-validation results are reported in aggregate, with individual annotator records withheld.

Compute and model access. Answer generation uses inference through hosted or locally served API-compatible endpoints. The learned evaluator consists of the family-specific XGBoost regressors described in Section 4.1. Methods reuse the offline knowledge base and retrieval indexes, with fixed generation prompts, retrieval settings, and evaluation rubrics. Hidden future targets are reserved for evaluation and excluded from method hyperparameter tuning.

C BENCHMARK CONSTRUCTION AND TEMPORAL INTEGRITY

The construction pipeline in Figure 2 comprises corpus filtering, temporal taxonomy induction, evidence extraction, and task curation. The following sections describe these stages and the separation of prediction-time and evaluation artifacts.

C.1 CORPUS FILTERING AND TEMPORAL FREEZING

For each domain, we start with broad domain queries, harvest papers through March 2026, normalize metadata, and deduplicate query hits. The first pass is recall-oriented: it retains candidates with non-trivial domain evidence even when the terminology is not yet stable.

The domain-relevance screen considers whether the paper's main problem, method, system, evaluation, dataset, or application setting is substantively tied to the target area. The benchmark-core screen then marks a paper as *core* when it has a central domain contribution, a concrete research asset type, and useful future-facing signal for tasks, methods, evaluation, bottlenecks, or design patterns. We keep weaker relevant papers as *support*; borderline cases may be retained for audit; noisy or out-of-domain papers are excluded.

Public questions and accessible support are frozen at historical cutoffs, while later papers are withheld for validation only. ForeSci uses three-month and six-month forecast settings, and venue-conditioned tasks follow venue-cycle timing because conference evidence appears on venue-specific schedules.

C.2 Cutoff Slicing and Forecast-Oriented Taxonomy Adaptation

Taxonomy induction uses the cutoff-aligned core/support corpora described in Section 3.2. To preserve temporal structure, papers are processed in chronological slices. Earlier periods use coarser slices, while periods close to the cutoff use finer slices when short-horizon movement matters. This design keeps recent changes visible instead of smoothing them into the older literature.

Our taxonomy builder follows TaxoAdapt's multidimensional routing and adaptive expansion principle (Kargupta et al., 2025). Papers are routed across contribution dimensions such as tasks, methods, datasets, evaluation methods, and application domains. Dense or poorly covered regions trigger width or depth expansion, allowing new research subdirections to appear as the cutoff-visible literature evolves.

We adapt this process for ForeSci in three ways: induction uses filtered core/support papers, temporal slicing emphasizes cutoff-local deltas, and induced nodes must be grounded in node evidence records before they can support benchmark construction.

C.3 Candidate Direction Selection

Candidate directions are selected from taxonomy nodes and small groups of related nodes. We retain a candidate when it satisfies four criteria: it has sufficient pre-cutoff support, it expresses a clear research decision, it is specific enough for evaluation, and it can be separated from hidden future validation evidence. Candidates are revised or removed when they are ambiguous, duplicated, too broad, too narrow, or weakly grounded in the underlying papers.

C.4 Method-Development and Bottleneck Signals

Method-development signals are derived from method and evaluation nodes, paper co-assignments, title/abstract method surfaces, full-text evidence, and bottleneck–mechanism cues. They record relations such as extension, adaptation, replacement, component reuse, and method competition. These signals provide trajectory evidence for comparing directions and reasoning about mechanism-level change.

Bottleneck signals summarize recurring limitations, evaluation gaps, reliability or safety concerns, dataset or benchmark needs, and technical risks. We verify them against full-text evidence from pre-cutoff papers, with emphasis on limitations that recur across multiple sources or connect to concrete evaluation failures.

C.5 Venue-Community Profiles

Venue-conditioned tasks use *venue-community signals from metadata*. We construct profiles for venue families such as ACL/EMNLP/NAACL, ICLR/ICML/NeurIPS, AAAI/IJCAI, SIGIR/KDD, and CVPR/ICCV/ECCV. Each profile summarizes contribution styles, maturity expectations, reviewer risks, evidence-package expectations, and nearby compatible venue families. These profiles support venue-cycle judgments based on research fit, evidence standards, and reviewer expectations.

C.6 Human–LLM Collaborative Audit

LLMs draft construction records from cutoff-aligned evidence by summarizing supporting papers, extracting claims and limitations, identifying method-development and bottleneck signals, proposing candidate directions, and flagging possible leakage risks.

Human experts then inspect the source papers and full-text evidence. They verify support strength, temporal validity, specificity, and leakage risk; revise unclear wording; remove weak or duplicated items; and approve the final construction records. The same audit process is applied before task release.

C.7 Task Curation and Artifact Separation

Task curation checks three properties: a stable historical premise, a clear public decision, and post-cutoff evidence suitable for validation. Items are revised when the decision is underspecified, the validation evidence does not match the requested judgment, or multiple items express the same research decision.

Each released task contains the public question q, cutoff t, forecast window, task-family instructions, answer requirements, and pre-cutoff support packet $\mathcal{E}_{\leq t}(q)$. Internal taxonomy identifiers, representative-paper lists, construction traces, and audit notes remain internal. Hidden future validation targets $\mathcal{G}_{>t}(q)$ are held exclusively for the evaluation protocol.

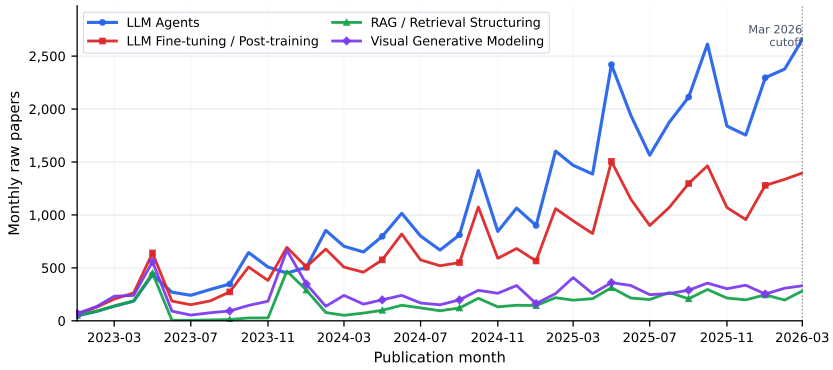

Figure A1: Monthly raw-paper volume in the four ForeSci domains. Counts are unique papers after normalizing paper identifiers within each domain-month. All four domains are shown from January 2023 through the March 2026 corpus collection endpoint.

Figure A2 summarizes differences in corpus scale and taxonomy growth across domains. The examples in Figure A3 and Figure A4 illustrate the subdirections and bottleneck–mechanism relations used to construct tasks.

C.8 Output-Level Temporal-Leakage Audit

We audit all 5,000 outputs from the two proprietary backbones for explicit post-cutoff named entities. Each output is bound to its own task cutoff, public question, and saved support snapshot. A screening pass extracts paper, method, model, dataset, and system names and retains exact surface-form overlaps with future targets. We then resolve each retained span to a canonical entity, establish its earliest public date using the arXiv v1 date where available and otherwise an official release or publisher first-online date, and attribute its possible origin to the public question, the exact retrieved/full-text support snapshot, or parameter-origin risk. Support attribution requires the entity to occur in the output's saved support snapshot. The audit concerns identifiable named entities; conceptual similarity, shared mechanisms, and generic terminology fall outside its scope.

We label an entity as *hard leakage* only when all five conditions hold:

1. the entity is an exact, locatable span in the answer;

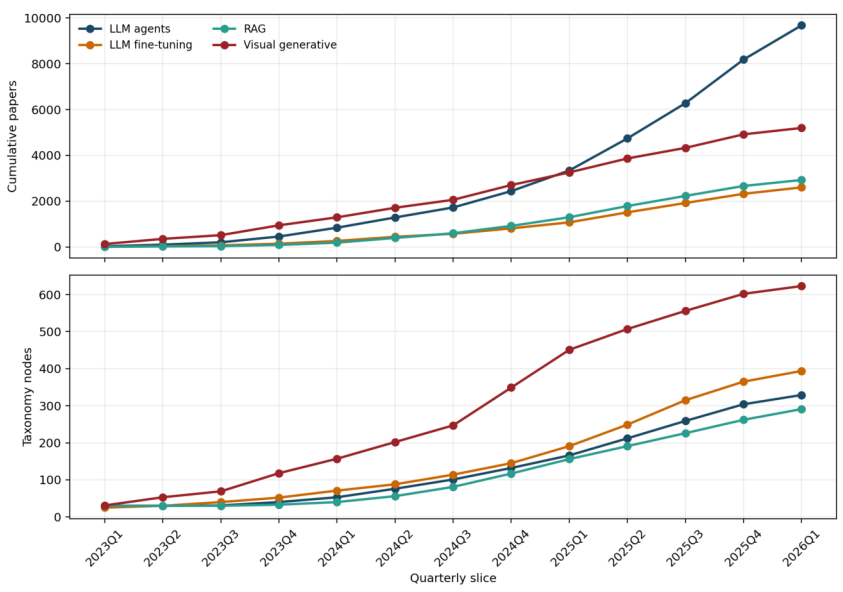

Figure A2: Temporal evolution of the domain corpora and induced taxonomies used in ForeSci. The curves summarize cumulative domain KB coverage and taxonomy-node growth over the literature slices used for the release.

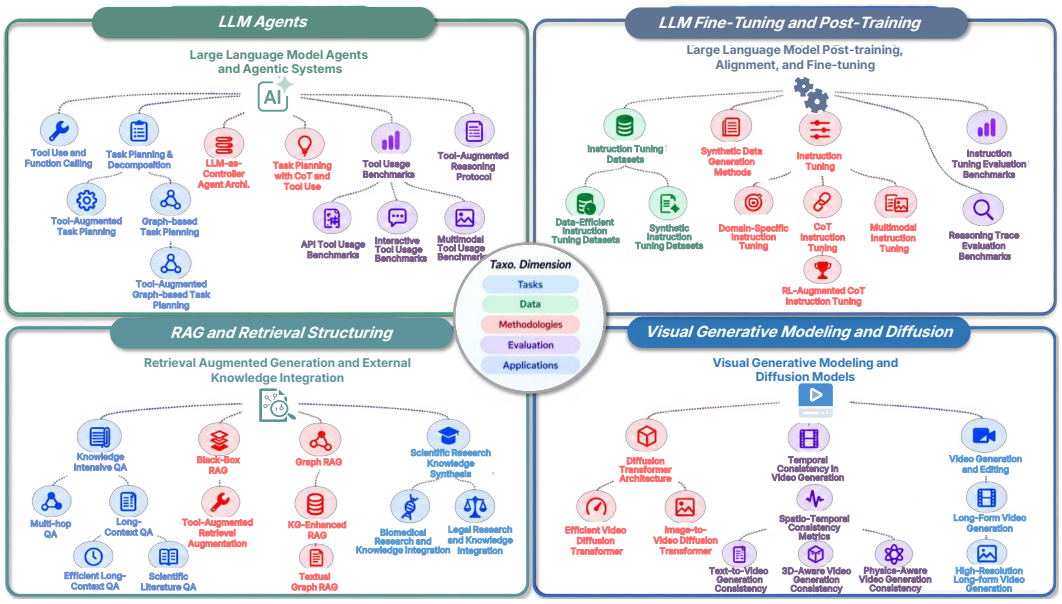

Figure A3: Illustrative examples of temporal taxonomy branching in ForeSci. The figure shows how broad pre-cutoff topics split into task-seeding subdirections; Table A2 summarizes representative branch examples in text form.

2. the span uniquely identifies a paper, method, model, dataset, or system;
3. a reliable source dates its first public appearance after that answer's task cutoff;
4. the entity is absent from the public question; and
5. provenance adjudication, after checking the exact support snapshot, identifies either post-cutoff retrieval origin or unsupported parameter-origin risk rather than cutoff-safe prompt or support origin.

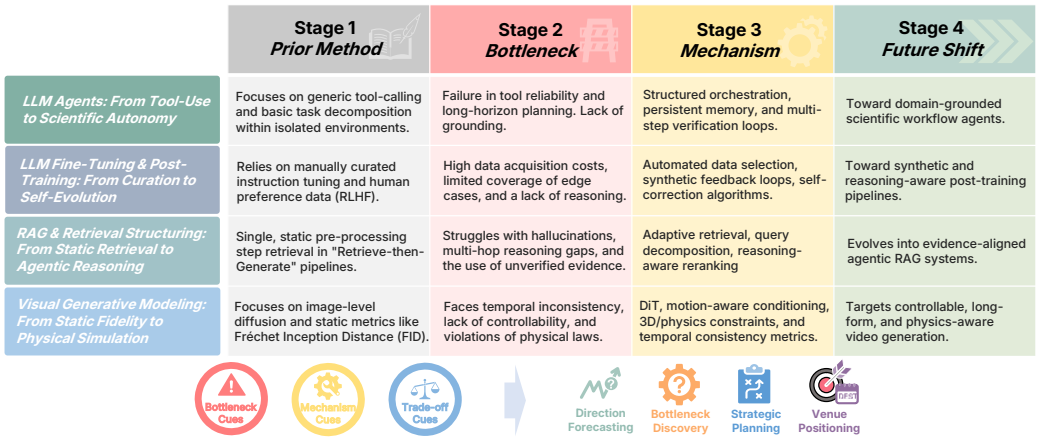

Figure A4: Method-evolution signals for foresight task construction. ForeSci models each domain as an evolutionary chain in which limitations of prior methods expose bottleneck cues, emerging technical responses provide mechanism cues, and the resulting bottleneck–mechanism interaction points toward a future research shift. The examples show this progression for LLM agents, LLM fine-tuning and post-training, RAG and retrieval structuring, and visual generative modeling. The bottom row summarizes how bottleneck, mechanism, and trade-off cues feed the four downstream task families: direction forecasting, bottleneck–opportunity discovery, strategic research planning, and venue-conditioned positioning.

Table A2: Representative taxonomy subtrees and method-evolution signals used during ForeSci construction. The taxonomy column illustrates how temporally induced domain structures refine broad research areas into fine-grained subdirections. The method-evolution column summarizes the complementary bottleneck–mechanism–shift patterns used to seed future-facing task targets.

Domain	Representative taxonomy subtree	Representative method-evolution signal
LLM Agents	Tool use and function calling → tool-augmented task planning → tool-augmented graph-based task planning; tool usage benchmarks → API / interactive / multimodal tool usage benchmarks	Generic tool use and task planning → tool reliability and long-horizon grounding bottlenecks → structured tool orchestration, memory, feedback, and verification loops → domain-grounded scientific workflow agents
LLM Fine-tuning and Post-training	Instruction tuning datasets → data-efficient instruction tuning datasets / synthetic instruction tuning datasets; instruction tuning → domain-specific, multimodal, and chain-of-thought instruction tuning	Curated instruction and preference data → data cost, coverage, and weak reasoning-trace bottlenecks → data selection, synthetic feedback, self-correction, and process supervision → synthetic and reasoning-aware post-training pipelines
RAG and Retrieval Structuring	Iterative retrieval-generation pipelines → retrieval strategy evaluation → adaptive / hybrid / multimodal retrieval evaluation; reasoning-aware evaluation → evidence-aligned reasoning evaluation	Retrieve-then-generate pipelines → hallucination, multi-hop grounding, stale evidence, and citation-fidelity bottlenecks → adaptive retrieval, query decomposition, evidence verification, and reasoning-aware reranking → evidence-aligned agentic RAG systems
Visual Generative Modeling and Diffusion	Temporal consistency in video generation → spatio-temporal consistency metrics → text-to-video / 3D-aware / physics-aware consistency metrics; video diffusion transformer architectures → efficient / image-to-video / video-editing diffusion transformers	Image-level diffusion generation and static fidelity metrics → temporal inconsistency, controllability, physical plausibility, and long-form coherence bottlenecks → video diffusion transformers, motion-aware conditioning, 3D/physics constraints, and consistency metrics → controllable long-form and physics-aware video generation

We classify methods, modules, or directions explicitly proposed for development within the forecast window as prospective proposals. Hard-leakage adjudication applies to uniquely identifiable post-cutoff entities presented as already existing at the cutoff. The screen surfaces 57 mentions in 56 GPT-5.2 outputs and 60 mentions in 57 Gemini-3 outputs. Manual adjudication finds no confirmed hard leakage under this definition: 0/2,500 for each backbone (95% Wilson upper bound 0.15%) and 0/5,000 pooled (upper bound 0.08%). These bounds describe detected output-level leakage under the screening protocol; training-corpus contamination remains unmeasured.

Table A3: Output-level temporal-leakage audit for proprietary backbones. Confirmed rates are 0/2,500 per backbone (95% Wilson upper bound 0.15%) and 0/5,000 pooled (upper bound 0.08%).

Backbone	Outputs	Candidate mentions	Candidate outputs	Confirmed
GPT-5.2	2,500	57	56	0
Gemini-3	2,500	60	57	0
Pooled	**5,000**	**117**	**113**	**0**

D Evaluation Protocol and Validation

D.1 Metric Calibration Details

The following definitions specify the metrics introduced in Section 4.

Prediction Factuality. Let a denote a candidate answer and let $\mathcal{C}(a) = \{c_i\}_{i=1}^{m}$ be its extracted atomic claims. The reported Prediction Factuality score is claim-level F1 over answer-claim support and hidden claim-bank coverage. For each extracted answer claim c_i, a benchmark-aware verifier assigns *supported*, *partially supported*, *unsupported*, or *not checkable* relative to the public task and hidden claim units. Let $\phi_i \in \{1, 0.5, 0, 0\}$ be the corresponding answer-claim support score. Claim precision is

$$\mathrm{Prec}(a) = \frac{1}{m} \sum_{i=1}^{m} \phi_i. \tag{3}$$

Let $\mathcal{G}^+ = \{g_j\}_{j=1}^{n}$ denote the expanded hidden claim bank for the task. For each hidden claim, the judge assigns *covered*, *partially covered*, or *not covered* by the candidate answer. Let $\psi_j \in \{1, 0.5, 0\}$ be the corresponding coverage score. Claim recall is

$$\mathrm{Rec}(a) = \frac{1}{n} \sum_{j=1}^{n} \psi_j. \tag{4}$$

The final score is

$$\mathrm{PredictionFactuality}(a) = \frac{2\,\mathrm{Prec}(a)\,\mathrm{Rec}(a)}{\mathrm{Prec}(a) + \mathrm{Rec}(a)}. \tag{5}$$

Future-Target Alignment (FTA). FTA is family-conditioned. For direction forecasting and bottleneck–opportunity tasks, we use Reference-Guided FTA. The hidden future target is represented as a set of slots $\mathcal{G}_b = \{g_j\}_{j=1}^{n}$. Each slot g_j contains one or more acceptable textual variants V_j, such as a primary future target, a root-bottleneck paraphrase, an unlocked opportunity variant, or an evidence-backed mechanism variant. Variants are alternatives inside the same target slot; they are not counted as additional targets. Negative confusions are retained for qualitative auditing but are not counted as positive slots.

Let $\mathcal{C}(a) = \{c_i\}_{i=1}^{m}$ be the benchmark-relevant prediction claims extracted from the answer. We embed every claim and target variant with bge-m3 and compute $s(c, t) = \max(0, \cos(e(c), e(t)))$, placing pairwise similarity on a 0–1 scale. The slot score is the best claim–variant match inside the slot:

$$S_j(a, b) = \max_{1 \le i \le m,\ t \in V_j} s(c_i, t). \tag{6}$$

The Reference-Guided FTA score is the mean slot score:

$$\mathrm{FTA}_{\mathrm{RG}}(a, b) = \frac{1}{n} \sum_{j=1}^{n} S_j(a, b). \tag{7}$$

The target audit quantifies the breadth of the accepted variants. All 125 Direction tasks contain acceptable neighbors (mean 4.43 per task). For Bottleneck, 94 of 125 tasks contain an acceptable bottleneck variant (mean 3.44 when nonempty), and 86 contain an acceptable opportunity variant (mean 4.06 when nonempty). All 125 Venue tasks contain acceptable venue-neighborhood entries (mean 4.10). Across the 5,000 Direction and Bottleneck evaluations, 65.0% use at least one non-primary variant as a best slot match; 8.54% of all slot-level best matches use such a variant.

For strategic planning, the hidden target is an ordered list $\pi^* = (r_1, \ldots, r_K)$ over candidate research directions. The answer is parsed into an inferred ranking $\hat{\pi}$. We score three deterministic components: whether the top item matches, whether each preferred item appears in the same position or elsewhere, and whether pairwise order relations are preserved. Let

$$S_{\text{top}} = \mathbb{I}[\hat{\pi}_1 = r_1], \tag{8}$$

$$S_{\text{pos}} = \frac{1}{K} \sum_{k=1}^{K} s_k, \tag{9}$$

$$S_{\text{pair}} = \frac{2}{K(K-1)} \sum_{i<j} \mathbb{I}[r_i \prec_{\hat{\pi}} r_j]. \tag{10}$$

where $s_k = 1$ when r_k appears in position k, $s_k = 0.5$ when it appears in another inferred position, and $s_k = 0$ when it is missing. The recall-style ranking score is the mean of these three components. A symmetric precision-style score is computed over the inferred ranking using the same top, position, and pairwise-order checks relative to π^*. The reported ranking-aware FTA is the F1 of these precision and recall terms:

$$\text{FTA}_{\text{rank}}(a, b) = \frac{2PR}{P + R}. \tag{11}$$

For Venue-Conditioned Positioning, we first canonicalize venue aliases within a fixed family ontology (for example, individual members of a compatible venue family map to the same family). Let d and $\hat{d}$ denote the gold and inferred default families, let A and $\hat{A}$ denote the gold and inferred sets of conditional alternatives, and let u and $\hat{u}$ denote the gold and inferred contribution centers. We compute

$$S_{\text{default}} = \mathbb{I}[\hat{d} = d], \qquad S_{\text{cond}} = \frac{|A \cap \hat{A}|}{|A|}, \qquad S_{\text{contrib}} = \max_{v \in V_u} \max(0, \cos(e(v), e(\hat{u}))), \tag{12}$$

where V_u contains the accepted contribution-center variants and $e(\cdot)$ is the fixed embedding function. The Venue score is

$$\text{FTA}_{\text{venue}} = 0.50 S_{\text{default}} + 0.20 S_{\text{cond}} + 0.30 S_{\text{contrib}}. \tag{13}$$

The three components measure default-family agreement, coverage of conditional alternatives, and semantic alignment of the contribution center, respectively. The reported FTA is FTA_{RG} for Direction and Bottleneck tasks, FTA_{rank} for Planning, and the Venue-specific structured FTA for Venue-Conditioned Positioning.

Evidence Traceability Score. For each task family f, we define exactly three decision slots S_f: forecast direction, mechanism, and validation signal/risk for Direction; root bottleneck, causal bridge, and one-hop opportunity for Bottleneck; top priority, ordering/dependency, and comparative ranking rationale for Planning; and top venue, contribution profile, and cross-venue tradeoff for Venue. A first pass extracts these slots from the final answer without judging correctness. For each extracted slot, local bge-m3 retrieval selects the three most similar chunks from the output's attached pre-cutoff support packet; the embedding similarity is used only for candidate recall and never enters the score. A fixed language-model adjudicator applies the same structured rubric across families. It labels support as direct, partial, topical-only, contradictory, or absent; labels the answer's evidence-to-conclusion bridge as explicit, implicit, or absent; and identifies unresolved conflict. For Venue, evidence need not literally mention the venue name: contribution profile, evaluation regime, application setting, or technical emphasis is admissible only when the answer explicitly connects it to the venue decision or cross-venue tradeoff.

For Venue we score three chain links: direct profile evidence with a bridge (profile evidence), direct or partial profile evidence with an explicit connection to the selected venue (venue bridge), and direct evidence with an explicit bridge for the cross-venue tradeoff (tradeoff evidence). Each link receives a continuous traceability score in $[0, 1]$.

Let $z_j \in [0, 1]$ denote the continuous traceability score for decision slot j, reflecting the strength of evidence support, the evidence-to-decision bridge, and unresolved conflict. For Venue, z_j scores

the corresponding chain link defined above. The reported score is the unweighted mean of the three continuous scores:

$$\text{Trace}(a) = \frac{1}{3}\sum_{j=1}^{3} z_j. \tag{14}$$

A stricter direct-support-plus-explicit-bridge criterion is retained as a separate Venue diagnostic. Scoring uses the extracted decision slots and their pre-cutoff evidence, excluding hidden future targets and system identity. Outputs without attached external support are reported as – and excluded from Trace aggregates. The released artifacts retain the underlying slot judgments.

D.2 Human Annotation and PERS Modeling

Reviewer Persuasiveness. Reviewer Persuasiveness estimates how convincing a research decision is to a reviewer. We sample 1,200 answers from 120 tasks, balanced across the four families and stratified across backbones and methods. Each answer is independently rated by three human annotators, who see the public task and anonymized answer but not the hidden future target, support provenance, backbone, or method. Each annotator assigns a holistic score from 1 (unusable) to 7 (exceptional). For answer a, the human supervision target is the arithmetic mean of the three ratings,

$$y(a) = \frac{1}{3}\sum_{r=1}^{3} y_r(a). \tag{15}$$

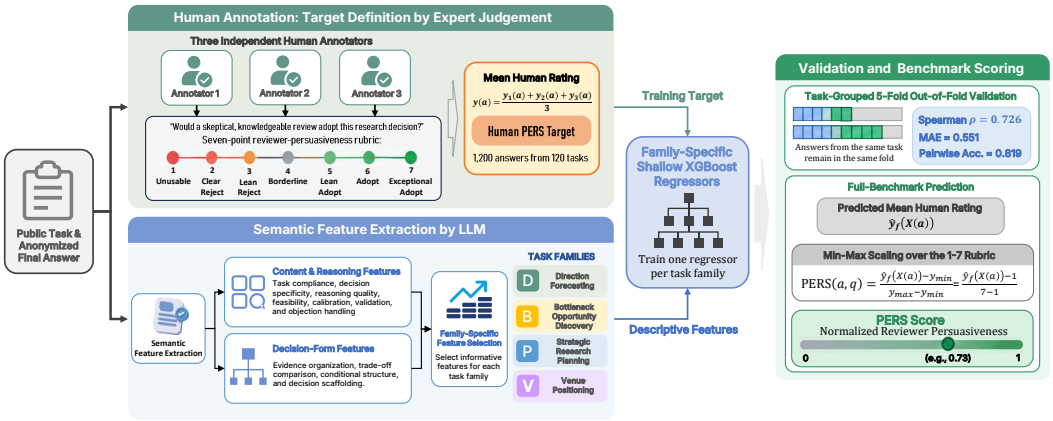

Figure A5: Overview of PERS. Three independent human annotators score each anonymized answer on a 1–7 reviewer-persuasiveness rubric, and their mean rating defines the supervision target. Semantic content, reasoning, and decision-form features extracted from the public task and answer feed a shallow task-family-specific XGBoost regressor. Labeled answers receive task-grouped out-of-fold predictions for validation; regressors fitted on the complete annotation set score the remaining benchmark answers. Predicted ratings are min–max scaled using the rubric endpoints for paper-facing PERS scores.

Human annotation protocol for Y. The human annotators received the following instructions. For each item, judge *ex-ante reviewer persuasiveness*: how willing you would be, as a skeptical and knowledgeable reviewer, to adopt the answer's research decision based only on the public task and the reasoning visible in the anonymized answer. Evaluate one complete answer at a time without comparing it with other answers. Do not use outside information or attempt to infer the hidden future target, generation method, or backbone. This judgment is distinct from future correctness, factuality, citation quality, evidence traceability, and writing style. Length, citation count, polish, confidence, and technical terminology receive no credit by themselves. Consider an unsupported factual statement only insofar as it weakens the argument visible in the answer; do not externally verify it.

Make a holistic judgment of whether the answer states the required decision, provides concrete and decision-relevant reasoning, distinguishes meaningful alternatives when required, addresses

feasibility and uncertainty, and gives a defensible basis for action. Before assigning a score, identify the strongest adoption-relevant support and the single most decisive weakness. Your rating is not an average of component subscores. Use the following seven anchors:

- 1 (unusable): no judgeable decision, severe instruction failure, or reasoning that cannot support adoption;
- 2 (clear reject): a decision is present, but decisive omissions, contradictions, or unsupported leaps dominate;
- 3 (lean reject): some plausible reasoning, but major unresolved weaknesses make adoption unlikely;
- 4 (borderline): an adequate but mixed case on which a reasonable reviewer could go either way;
- 5 (lean adopt): persuasive overall, with identifiable but non-decisive reservations;
- 6 (adopt): strong, concrete, well defended, and decision-ready, with only limited weaknesses;
- 7 (exceptional adopt): unusually compelling and robust, reserved for rare answers.

Mark an answer unjudgeable only when corruption or a fundamental instruction failure prevents meaningful evaluation; weak but understandable answers remain judgeable.

All three PERS annotators either held or were pursuing doctoral degrees; the panel comprised university associate professors and PhD candidates. Collectively, their research backgrounds span LLM agents, video generation, and post-training.

We fit a separate shallow XGBoost regressor for each task family because the evidence needed for a persuasive forecast, bottleneck diagnosis, research plan, and venue decision differs substantively. The predictors use semantic features extracted from the public task and final answer, covering task compliance, decision specificity, reasoning, feasibility, calibration, validation, evidence presentation, and decision structure. Family-specific subsets are fixed before final evaluation. The feature vector excludes backbone, generation method, answer length, family indicators, and hidden targets; task family selects the regressor. Tables A4 and A5 give the complete predictor sets and their operational meanings; each listed row corresponds to one model input variable.

Labeled answers receive task-grouped out-of-fold predictions; the remaining answers are scored by models fitted on the complete annotation set.

D.3 End-to-End Worked Example

Task RFDIR-0001 asks for one concrete direction for tool-using LLM agents, using evidence available through 2025-09-30, for the 2025-10-01 to 2026-03-31 window, together with one trajectory label from ACCELERATING, FRAGMENTING, STEADY, or COOLING. A GPT-5.2 ARIS output receives only the public task and its exposed pre-cutoff support packet. Representative support records include ThinkGeo (tool correctness, formatting, reasoning, and final-answer metrics), RAS-Eval (real-world security evaluation and simulation-to-real fidelity gaps), work on deployment security and behavioral steering, and Tool-R1 (tool-learning momentum and RL cost).

The answer forecasts *accelerating* work on trace-first reliability and regression suites for tool-using agents: replayable execution checks for tool validity, schema adherence, runtime success, and dependency chains, supplemented by structured assessment of plan quality. It argues that such evaluation is more immediate than costlier RL or broader multi-agent redesign. At evaluation time, the hidden target exposes six slots and nine variants covering standardized tool-agent evaluation, trajectory-aware and long-horizon assessment, realistic settings, reliability and security beyond final success, the accelerating label, and acceptable neighboring formulations.

D.4 Validation Analyses

We report agreement among the PERS annotators, task-grouped predictive validation, aggregate method contrasts, and human validation of claim extraction.

D.4.1 Human Annotator Agreement

Across the three annotator pairs, mean Spearman correlation is 0.818 and mean quadratic-weighted kappa (QWK) is 0.821. Exact agreement is 59.0% on the seven-point scale, while 98.5% of paired ratings differ by at most one point.

Table A4: Content and reasoning variables used by the family-specific Reviewer Persuasiveness regressors. Each row is one predictor; a checkmark indicates that the variable enters the corresponding family model. B, D, P, and V denote Bottleneck–Opportunity Discovery, Direction Forecasting, Strategic Research Planning, and Venue-Conditioned Positioning, respectively.

Predictor	Operational meaning	Range	B	D	P	V
Hard-constraint compliance	Required cardinality, cutoff, horizon, candidate-set, and exclusion constraints are obeyed.	0–2	✓	✓		✓
Required-output completeness	All explicitly requested decision outputs are present.	0–2				✓
Technical-object specificity	The decision identifies a concrete technical object rather than only a broad topic.	0–2	✓	✓		✓
Closest-alternative discrimination	The proposed root cause is distinguished from a symptom or competing explanation.	0–2	✓			
Time-bounded plausibility	The forecast, opportunity, or plan is plausible within the stated horizon.	0–2	✓	✓	✓	
Consequential capability unlocked	Resolving the bottleneck directly unlocks a consequential one-hop capability.	0–2	✓			
Premises or signals stated	The answer exposes the signals on which its forecast rests.	0–2		✓		
Premise-to-decision bridge	The answer connects its premises to the final decision.	0–2				✓
Shared-tradeoff discrimination	Alternatives are compared on common decision-relevant criteria.	0–2				✓
Decision-changing condition	A concrete condition is given under which the decision changes.	0–2			✓	✓
Dependencies and constraints	Material dependencies or implementation constraints are addressed.	0–2			✓	
Observable validation signal	The answer specifies an observable outcome or measurement.	0–2			✓	
Discriminative validation test	The proposed test can distinguish competing decisions or hypotheses.	0–2			✓	
Inferential validity	Granting the stated premises, the conclusion follows without a decisive logical jump.	0–5	✓	✓		
Causal-mechanism depth	The answer makes the intermediate how/why links in its causal account explicit.	0–5	✓	✓		
Support relevance	Visible support bears directly on the claimed root cause and opportunity.	0–5	✓			
Scope-boundary discipline	The claim is bounded to the settings for which its reasoning applies.	0–5	✓		✓	
Epistemic calibration	Certainty and scope do not exceed what the visible reasoning licenses.	0–5		✓		
Feasibility realism	Resources, dependencies, maturity, and timing are treated realistically.	0–5			✓	✓
Implementation actionability	The decision implies a concrete and executable work package.	0–5			✓	
Validation decision value	The proposed evidence can change a go/no-go, priority, or venue decision.	0–5			✓	✓
Information prioritization	The decisive recommendation and reasons are recoverable without rewarding verbosity.	0–5			✓	
Objection resilience	The strongest likely objection is identified and substantively addressed.	0–5			✓	✓

D.4.2 Reviewer Persuasiveness Model Validation

The grouped-OOF evaluation covers 1,200 answers from 120 tasks. All answers sharing a task are assigned to the same fold, so every prediction comes from a model trained without any answer from that task. Aggregating the five held-out folds yields Spearman $\rho = 0.726$, MAE 0.551 points on the 1–7 annotation scale, and within-task pairwise accuracy 0.819 against the mean human rating.

D.4.3 Aggregate Method Contrasts

Table A8 summarizes differences between the mean of CoI-style, ResearchAgent-style, and ARIS-style and each baseline, averaging equally across the four backbones.

In Table A10, the best–second gap is at most 0.02 in 12/16 backbone–family cells for Fact, 12/16 for FTA, 8/16 for Trace, and 14/16 for PERS.

D.4.4 Claim Extraction Human Validation

The claim-extraction audit covers 240 answers, with 60 per family (Table A9). Human annotators assess precision, recall, atomicity, decision relevance, and noise relative to the source answer. Precision ranges from 0.78 to 0.86 and recall from 0.70 to 0.76. Planning and Venue have lower atomicity and higher noise than Bottleneck and Direction. These measurements characterize the extraction stage; the subsequent factuality judgments are a separate evaluation step.

Table A5: Decision-form variables used by the family-specific Reviewer Persuasiveness regressors. These variables encode the presence of task-relevant organizational and argumentative patterns, not their quality: 0 denotes absent, 1 localized, and 2 recurring. Bottleneck and Direction use no decision-form variables in their final predictor sets.

Predictor	Operational meaning	Range	B	D	P	V
Claim–anchor pairing	Decision claims are placed together with their supporting anchors.	0–2			✓	✓
Cross-source synthesis	Multiple sources are combined within a single argumentative unit.	0–2			✓	✓
Conditional branching	The answer organizes recommendations around explicit conditions or cases.	0–2			✓	✓
Epistemic qualification	Claims are accompanied by visible uncertainty or scope qualifications.	0–2			✓	
Tradeoff comparison	Alternatives are explicitly contrasted in terms of costs and benefits.	0–2			✓	
Shared-criterion agenda comparison	Planning candidates are compared under a common criterion.	0–2			✓	
Dependency–priority chain	Dependencies are used to justify the ordering of research priorities.	0–2			✓	
Milestone decision rule	A milestone is tied to a subsequent continue, stop, or reprioritize decision.	0–2			✓	
Sectioned decision scaffold	The venue decision is presented through explicit functional sections.	0–2				✓
Objection–response pattern	A likely objection is paired with a response or mitigation.	0–2				✓
Parallel venue review	Candidate venues are discussed in a parallel comparison structure.	0–2				✓
Shared-criteria venue comparison	Candidate venues are evaluated using the same stated criteria.	0–2				✓
Reviewer simulation	The answer explicitly anticipates how a reviewer would evaluate the submission.	0–2				✓
Evidence-upgrade mapping	Proposed evidence upgrades are mapped to specific claims or review concerns.	0–2				✓
Pivot-trigger condition	A concrete trigger is tied to changing the preferred venue.	0–2				✓

Table A6: A prediction-to-evaluation example. The system sees the public task and pre-cutoff support only; hidden post-cutoff target slots are opened only by the evaluator.

Metric	Computation for RFDIR-0001
Fact = 1.000	Five extracted decision-critical claims attain precision and hidden-claim recall of 1.000 over eight gold claim units.
FTA = 0.638	For each of six target slots, the evaluator takes the best bge-m3 match over the answer claims and the slot's acceptable variants, then averages the six values.
Trace = 0.842	The three decision slots receive continuous traceability scores in $[0, 1]$; their unweighted mean is 0.842.
PERS = 0.743	The three human scores are 6, 5, and 5, giving a normalized mean target of 0.722; 0.743 is the task-grouped OOF regression estimate.

Table A7: Pairwise agreement among the three human annotators who supplied the PERS training targets. Exact agreement requires identical 1–7 ratings; "Within 1 point" allows an absolute difference of at most one.

Annotator pair	Spearman ρ	QWK	Exact	Within 1 point
1 vs. 2	0.825	0.831	59.7%	99.3%
1 vs. 3	0.811	0.813	59.8%	97.3%
2 vs. 3	0.819	0.819	57.6%	98.8%
Mean	**0.818**	**0.821**	**59.0%**	**98.5%**

Table A8: Mean differences between agent methods and baselines. Agent averages include all three agent methods and four backbones. Native Trace is not applicable.

Metric	Contrast	Δ
Fact	Agent avg. − Hybrid	0.0155
FTA	Agent avg. − Hybrid	0.0163
Trace	Agent avg. − Hybrid	0.0509
PERS	Agent avg. − Hybrid	0.0181
Fact	Agent avg. − Native	0.0203
FTA	Agent avg. − Native	0.0221
PERS	Agent avg. − Native	0.0501

Table A9: Human validation of automatic decision-critical claim extraction.

Family	n	Prec.	Rec.	F_1	Atomic	Relevant	Noise
Bottleneck	60	0.84	0.72	0.78	88.5%	85.0%	11.2%
Direction	60	0.86	0.76	0.81	91.0%	89.5%	8.0%
Strategic Planning	60	0.79	0.70	0.74	82.3%	81.0%	16.5%
Venue	60	0.78	0.73	0.75	80.1%	83.0%	18.2%

D.5 Evaluation Prompt Templates

The following condensed templates specify the claim-extraction, factuality, and traceability rubrics and their output schemas.

Atomic Claim Extraction Prompt

You are extracting benchmark-relevant atomic factual claims from a candidate answer.
Inputs:

- Public task definition: `{PUBLIC_TASK_JSON}`
- Candidate answer: `{CANDIDATE_ANSWER}`

Your task:

- Decompose the answer into a small set of atomic factual claims.
- Keep only claims that could be grounded in a research-paper benchmark.
- Prefer claims about research directions, mechanisms, bottlenecks, venue fit, baselines, experiments, risks, trajectory, and evidence signals.
- Ignore purely stylistic statements, generic advice, and duplicates.
- Return at most `{MAX_CLAIMS}` claims.

Output requirements:

- Output JSON only.
- Do not include explanations, markdown, or extra text.

Output format:

```
{
  "claims": ["...", "..."]
}
```

Prompt template used for atomic claim extraction in Prediction Factuality.

Claim-Level Factuality and Coverage Prompt

You are a claim-level evaluator for a research-foresight benchmark. Compute claim precision and gold-claim recall for the candidate answer. This is a factual/content coverage metric, not a research-taste metric.
Inputs:

- Public task definition: `{PUBLIC_TASK_JSON}`
- Candidate answer claims: `{ANSWER_CLAIMS_JSON}`
- Gold claim units: `{GOLD_CLAIM_UNITS_JSON}`

Judgment rules:

- For answer claims, label each as `supported`, `partially_supported`, `unsupported`, or `not_checkable` relative to the public task and gold claim units.
- For gold claim units, label coverage as `covered`, `partially_covered`, or `not_covered` by the candidate answer.
- Do not reward generic topical overlap when the mechanism, venue rationale, or decision is different.
- Keep this metric independent of future-target alignment scoring.

Output format:

```
{
  "answer_claim_verdicts": [
    {"claim": "...", "label": "supported | partially_supported | unsupported |
not_checkable",
    "matched_gold_claim_ids": ["..."], "rationale": "..."}
  ],
  "gold_claim_coverage": [
    {"gold_claim_id": "...", "label": "covered | partially_covered | not_covered",
    "matched_answer_claims": ["..."], "rationale": "..."}
  ]
}
```

Prompt template used for claim-level factuality and coverage in Prediction Factuality.

Evidence Traceability Slot Extraction and Adjudication

Stage 1: extract the three family-specific decision slots from the candidate answer. Do not solve the task or judge correctness. If a slot is absent, return an empty claim and mark it not present.
Stage 2: for each slot, judge how strongly the claim is supported by its supplied pre-cutoff candidate evidence chunks.

- Use only the evidence content supplied for that slot; do not use outside knowledge or hidden future targets.
- Label support as `direct`, `partial`, `topical_only`, `contradictory`, or `none`.
- Label the evidence-to-conclusion bridge as `explicit`, `implicit`, or `absent`.
- Mark unresolved conflict only when materially conflicting evidence is present and the answer does not resolve it.
- Assign a continuous traceability score in $[0, 1]$ based on evidence support, the evidence-to-decision bridge, and unresolved conflict.
- Evidence quantity and embedding similarity do not add points.

Inputs:

- Public task family and question: `{PUBLIC_TASK}`
- Candidate answer: `{CANDIDATE_ANSWER}`
- Family-specific slot schema: `{SLOT_SCHEMA}`
- Top-3 candidate evidence chunks per extracted slot: `{CANDIDATE_EVIDENCE}`

Output format:

```
{
  "slot_results": [
    {"slot_id": "...", "support_status": "direct | partial | topical_only |
contradictory | none",
    "bridge_status": "explicit | implicit | absent", "unresolved_conflict": false,
    "slot_score": 0.0,
    "evidence_anchors": ["..."], "short_reason": "..."}
  ]
}
```

Condensed two-stage prompt used for TRACE slot extraction and evidence-chain adjudication; Venue uses the profile-to-venue chain rule.

E System Adaptation and Comparison Controls

The three agent systems use offline adaptations that retrieve from the pre-cutoff knowledge base and expose evidence artifacts for evaluation.

Shared adaptation principles. Across all agent baselines, we applied the same three benchmark-side constraints.

- **Open-web removal.** All online literature search or browsing steps were replaced by offline access to the benchmark knowledge base, including paper metadata, section-level snippets, and benchmark-side support packets.
- **Cutoff-faithful evidence flow.** Retrieval, intermediate reasoning, and final answer rendering were constrained to use only pre-cutoff assets. Hidden future evidence remained strictly evaluation-only.
- **Benchmark-native output exposure.** Each adapted agent was modified to emit the support artifacts needed by the benchmark judges, such as retrieval traces, evidence items, and family-native structured answer fields when available.

Agent-specific adaptations. CoI-style retains multi-chain idea construction and replaces online literature search with an offline KB retrieval adaptor. ResearchAgent-style gathers evidence through benchmark support packets and uses task-specific planning and review prompts. ARIS-style uses offline hybrid retrieval, task-specific evidence construction, and a final renderer that enforces the candidate set and required answer fields. All three agents render their final decisions using the task-family output schemas.

F Supplementary Results and Evaluator Analyses

F.1 Evaluation Results by Task Family

Table A10 reports results by task family. Method rankings vary across families, backbones, and metrics.

Table A10: Main family-level results across four answer-generation backbones. Fact and FTA use the final evaluation outputs, Pers is the predicted mean three-annotator rating normalized to $[0, 1]$, and Trace is the mean continuous evidence-traceability score across three task-specific decision slots, including the profile-to-venue links for Venue. Each family block reports five methods; bold marks the best method within the same family, backbone, and metric.

		Qwen3-235B				GPT-5.2				GLM-4.6				Gemini-3			
Family	Method	Fact.	FTA	Trace	Pers.	Fact.	FTA	Trace	Pers.	Fact.	FTA	Trace	Pers.	Fact.	FTA	Trace	Pers.
Bottleneck–Opportunity	Native LLM	0.840	0.608	–	0.675	0.799	**0.614**	–	0.735	0.627	0.574	–	**0.528**	0.672	0.576	–	**0.691**
	Hybrid RAG	0.824	**0.609**	**0.621**	0.668	0.804	0.613	0.642	0.747	0.611	**0.589**	0.536	0.502	0.704	0.589	**0.526**	0.667
	CoI	0.851	0.607	0.532	0.659	0.819	0.610	0.667	**0.762**	0.672	0.584	0.533	0.518	0.700	0.583	0.455	0.656
	ResearchAgent	0.802	0.608	0.554	0.664	**0.829**	0.610	**0.684**	0.752	0.653	0.586	**0.598**	0.493	0.716	**0.591**	0.507	0.654
	ARIS	**0.860**	0.607	0.521	**0.694**	0.813	0.607	0.584	0.751	**0.702**	0.588	0.511	0.509	**0.735**	0.576	0.417	0.644
Direction Forecasting	Native LLM	**0.598**	0.635	–	0.626	**0.670**	0.635	–	0.710	0.498	0.581	–	0.532	0.573	0.601	–	**0.609**
	Hybrid RAG	0.567	**0.637**	**0.763**	0.604	0.645	0.634	0.621	0.695	0.523	**0.597**	0.669	0.550	0.589	**0.606**	0.620	0.591
	CoI	0.577	0.635	0.738	0.641	0.658	**0.637**	**0.671**	0.721	**0.534**	0.592	0.574	**0.561**	**0.601**	0.602	0.606	0.587
	ResearchAgent	0.568	0.633	0.703	0.633	0.662	0.633	0.658	0.720	0.501	0.584	0.635	0.542	0.595	0.604	0.569	0.582
	ARIS	0.542	0.627	0.705	**0.649**	0.589	0.628	0.595	**0.723**	0.512	0.591	**0.684**	0.557	0.573	0.600	**0.669**	0.585
Strategic Planning	Native LLM	0.465	0.580	–	0.480	0.441	0.548	–	0.586	0.334	0.449	–	0.372	0.421	0.554	–	0.458
	Hybrid RAG	0.446	0.554	0.495	0.602	0.446	0.553	0.473	0.734	0.411	0.521	0.536	0.525	0.422	0.521	0.459	0.641
	CoI	0.504	0.644	0.479	0.655	0.476	0.582	0.528	**0.767**	0.461	**0.646**	**0.542**	0.558	0.414	0.528	**0.512**	0.663
	ResearchAgent	**0.532**	**0.687**	0.486	0.647	0.481	0.577	**0.582**	0.758	**0.482**	0.629	0.538	0.542	**0.424**	0.544	0.480	**0.668**
	ARIS	0.460	0.598	**0.507**	**0.661**	**0.488**	**0.587**	0.567	**0.767**	0.441	0.565	0.532	**0.579**	0.422	**0.555**	0.500	0.660
Venue Positioning	Native LLM	**0.534**	0.712	–	0.582	0.551	0.680	–	0.720	0.542	0.738	–	0.581	0.570	0.776	–	0.648
	Hybrid RAG	0.526	0.700	0.438	0.542	0.576	0.694	0.503	0.724	0.553	0.762	0.347	0.601	0.555	0.776	0.393	0.649
	CoI	0.518	0.664	0.572	**0.607**	0.575	0.729	0.726	**0.741**	0.554	**0.804**	0.502	0.638	0.584	0.781	0.610	0.664
	ResearchAgent	0.528	0.690	0.531	0.602	0.580	0.723	0.688	0.733	**0.571**	0.755	0.455	**0.642**	**0.590**	**0.810**	0.574	**0.681**
	ARIS	0.523	**0.731**	**0.797**	0.591	**0.596**	**0.771**	**0.823**	0.729	0.511	0.749	**0.708**	0.604	**0.590**	0.772	**0.750**	0.588

F.2 Auxiliary Evaluator Stability

We report run-to-run stability on a Qwen3-235B subset of 100 tasks and five methods, giving 500 answers. Candidate answers are held fixed across five independent evaluation passes. For each metric, Table A11 reports the largest standard deviation and largest range of method-level run means across all family–method cells. This auxiliary test predates the current slot-based Trace and regression-based PERS formulations, so it is used only to characterize the retained content metrics.

FTA varies less across these repeated evaluations than Fact. The maximum range of method-level means is 0.014 for FTA and 0.053 for Fact. PERS predictive validation is reported in Appendix D.4.2.

Table A11: Auxiliary five-run stability of Fact and FTA on the Qwen3-235B 100-task subset. "Max mean SD" and "Max mean range" summarize method-level run means across family–method cells. "Mean row SD" summarize per-row variation across the same five runs.

Metric	Max SD	Max range	Row SD
Prediction Factuality	0.021	0.053	0.045
Future-Target Alignment	0.006	0.014	0.004

F.3 Answer Lengths

Table A12: Answer-length profiles on aligned outputs. Lengths are regex-based word counts. Each row contains 625 answers from one generator backbone and one task family in the current aligned answer snapshot.

Generator	Family	Mean	Median	P90	Max
GLM-4.6	Bottleneck–Opportunity	124.9	121	161.0	228
GLM-4.6	Direction Forecasting	97.7	96	120.0	211
GLM-4.6	Strategic Planning	154.8	154	189.6	257
GLM-4.6	Venue Positioning	236.6	232	296.0	380
GPT-5.2	Bottleneck–Opportunity	339.8	328	432.6	692
GPT-5.2	Direction Forecasting	236.3	232	305.0	400
GPT-5.2	Strategic Planning	258.9	256	293.6	364
GPT-5.2	Venue Positioning	424.0	411	533.0	630
Gemini-3	Bottleneck–Opportunity	197.9	194	259.0	322
Gemini-3	Direction Forecasting	155.5	155	182.0	215
Gemini-3	Strategic Planning	204.1	205	227.0	263
Gemini-3	Venue Positioning	316.0	285	424.2	524
Qwen3-235B	Bottleneck–Opportunity	198.3	191	262.0	384
Qwen3-235B	Direction Forecasting	173.0	173	218.6	286
Qwen3-235B	Strategic Planning	172.9	174	196.0	227
Qwen3-235B	Venue Positioning	243.3	236	337.6	425

Table A12 reports answer lengths for the formal evaluation outputs. All four backbones share the same 2,500 $(\text{task_id}, \text{method})$ keys, with 625 answers per backbone and family. In each family, GPT-5.2 has the largest mean word count and GLM-4.6 the smallest. These statistics describe the edited outputs used for formal evaluation.

F.4 Content-Metric Correlation

We compute pairwise Pearson and Spearman correlations among Fact, FTA, and PERS over the 10,000 final evaluations. This analysis includes Native LLM outputs, for which Trace is unavailable, and therefore covers the three metrics shared by all methods.

Overall Spearman correlation is 0.448 for Fact vs. FTA, 0.189 for Fact vs. PERS, and 0.089 for FTA vs. PERS. Correlations vary by task family. Planning and Venue show much stronger Fact-vs.-FTA coupling: Strategic Planning has Spearman 0.816 and Venue Positioning has 0.710, compared with 0.292 for Bottleneck–Opportunity and 0.373 for Direction Forecasting. In these tasks, factual claim correctness and target alignment both depend on the priority or conditional positioning decision. Venue FTA vs. PERS correlation is 0.098, indicating a weaker association between target alignment and predicted reviewer persuasiveness.

Across backbones, Fact vs. FTA Spearman ranges from 0.368 for Qwen3-235B to 0.577 for GLM-4.6. FTA vs. PERS ranges from -0.180 for Qwen3-235B to -0.021 for GLM-4.6. Across methods, Fact vs. FTA ranges from 0.440 for CoI-style to 0.527 for Native LLM, while FTA vs. PERS ranges from -0.052 to 0.002. The pooled and subgroup correlations describe different aggregation levels and should be read together.

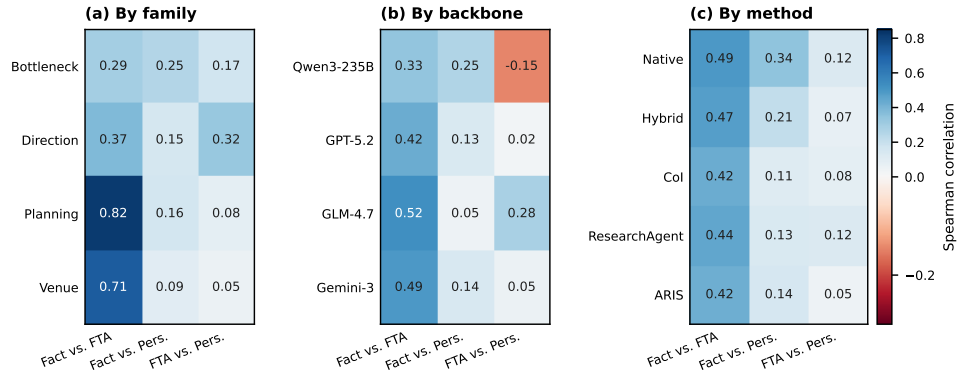

Figure A6: Supplementary content-metric correlation diagnostic over 10,000 final benchmark rows. Values are Spearman rank correlations among Prediction Factuality, Future-Target Alignment, and regression-based Reviewer Persuasiveness. Trace is unavailable for Native LLM outputs and is excluded from this all-method analysis. Correlations vary by family: planning and venue tasks have much stronger Fact-vs.-FTA coupling than bottleneck and direction tasks, while backbone and method differences are more moderate.

Table A13: Low-score rates used for the diagnostic error analysis. Rows scoring strictly below each metric's global 20th percentile threshold are selected; ties at the threshold are excluded. Trace thresholds and rates use evidence-grounded methods only; Native LLM traceability is not applicable.

Grouping	Unit	Low Fact	Low FTA	Low Pers.	Low Trace
Family	Bottleneck–Opportunity	0.062	0.013	0.174	0.308
	Direction Forecasting	0.184	0.004	0.012	0.164
	Strategic Planning	0.315	0.513	0.192	0.269
	Venue Positioning	0.114	0.251	0.037	0.018
Method	Native LLM	0.186	0.215	0.187	–
	Hybrid RAG	0.174	0.204	0.093	0.194
	CoI-style	0.168	0.188	0.085	0.188
	ResearchAgent-style	0.152	0.175	0.083	0.187
	ARIS-style	0.164	0.194	0.071	0.190

G Error and Mechanism Diagnostics

G.1 Detailed Error Analysis

The error analysis uses the 10,000 formal evaluations, their evaluator rationales, gold-claim coverage summaries, future-target judgments, and support metadata. Drift annotations are joined where available. For each metric, we calculate the global 20th percentile score threshold and mark rows scoring strictly below this threshold as low scoring. Because of score ties at the cutoff point, the actual percentage of selected low-scoring rows is less than 20% (e.g., approximately 10.4% for Pers and 16.9% for Fact). Trace percentiles are calculated only over evidence-grounded methods, as Native LLMs do not expose trace artifacts. These labels select cases for inspection.

Table A13 reports the selected low-score rates by family and method. Planning has the largest reported low-Fact and low-FTA shares. Family differences also reflect their distinct scoring functions, particularly semantic matching for Direction and Bottleneck versus ranking alignment for Planning.

The Fact–FTA comparison finds 122 rows with an absolute score gap of at least 0.5, alongside 20 high-FTA/low-Fact cases. These high-FTA/low-Fact cases are mostly in Direction Forecasting, where the answer is near the future target space but misses the exact mechanism, trajectory label, or scope. These cases illustrate differences between claim-level validation and semantic target matching.

Planning diagnostics show that the dominant failure is top-priority error: within planning rows, the top-priority-wrong rate ranges from 0.468 for ResearchAgent-style to 0.586 for Native LLM. These answers often justify individual candidates well but do not support the full global ordering. Venue diagnostics report pass rates of 0.877, 0.847, and 0.243 for the profile-evidence, venue-bridge, and

Table A14: Method-internal fingerprints from the diagnostic error audit. The table summarizes why low-scoring cases fail within each method.

Method	Fingerprint	Diagnostic interpretation
Native LLM	Fluent but unanchored adjacent answers.	Trace is not applicable because no external support artifact is exposed; factual and persuasive errors often come from plausible priors rather than the target decision.
Hybrid RAG	Retrieved context without enough decision grounding.	Retrieved evidence is often topical, but the final answer may not connect it to the required bottleneck, ranking, or venue-conditioned comparison.
CoI-style	Strong local evidence, wrong global decision.	The reasoning chain can be highly auditable while promoting a locally supported technical issue, such as TOCTOU, into the root decision object.
ResearchAgent-style	Coherent scaffold, sometimes wrong root premise.	Its coherent scaffold can still amplify a wrong early abstraction into a complete causal story.
ARIS-style	Specific mechanism over-commitment.	Concrete mechanism selection helps Bottleneck tasks, but can narrow Direction forecasts toward a specific instance when the reference target is a broader future mechanism cluster.

Table A15: Diagnostic coupling between answer-reference drift dimensions and formal metric failures. Metric drops are normalized by the standard deviation of the corresponding metric over all audited rows, so they are expressed in standard-deviation units. All numbers come from the matched drift-annotation sample joined with the formal metrics.

Question	Diagnostic contrast	Main result	Interpretation
Causal-role drift vs. Fact	Aligned vs. moderate-to-severe causal-role drift	Fact drops by 1.11 SD under high causal-role drift	Fact failures often arise when the answer assigns the wrong causal role to a plausible technical object, such as treating a symptom as a root bottleneck or a solution as the forecast target.
Scope/granularity drift vs. FTA	Aligned vs. moderate-to-severe scope drift	FTA drops by 1.19 SD overall; Planning FTA drops by 1.94 SD	The Planning contrast links scope mismatch to errors in the global ranking or priority structure.
Intervention-mode shift vs. FTA	Mode-aligned vs. mode-shifted answers	FTA drops by 1.10 SD overall; Planning FTA drops by 2.04 SD	Changing the action type, such as benchmark-first vs. training-first, directly changes planning and venue decisions even when local rationales remain plausible.
High traceability but low FTA	High-trace/non-low-FTA rows vs. high-trace/low-FTA rows	181 high-trace/non-low-FTA rows vs. 57 high-trace/low-FTA rows; causal severity 1.105 $\rightarrow$ 2.772; intervention-mode severity 1.530 $\rightarrow$ 2.860	High traceability does not guarantee target alignment: evidence can support a coherent but misaligned decision object.

tradeoff-evidence links over evidence-grounded answers, respectively. Tradeoff evidence is the least frequently supported of the three Venue links.

G.2 Drift–Metric Associations

We join 960 drift-annotated answers with the formal metrics, matching all answers by hash. Severity 0 denotes alignment and severity at least 2 denotes moderate-to-severe drift. Table A15 gives the numeric contrasts underlying Figure A7. Temporal-horizon annotations are retained in the released analysis files and included in the normalized effect-size heatmaps.

In RFDIR-0104, a GPT-5.2 ResearchAgent answer has high Traceability (0.842), PERS 0.733, and moderate FTA (0.638), but zero Prediction Factuality because it turns the target direction into a standardized PEFT evaluation card, whereas the reference expects adaptive sparse zeroth-order or hybrid PEFT. Figure A8 gives a complementary Planning example in which the answer and reference assign different investment priorities to the same evaluation gap.

G.3 Task-specific Drift Mechanisms

The drift audit identifies task-specific decision failures. In Bottleneck tasks, agents may promote a symptom, local mechanism, or proposed solution to the system-level root cause. In Direction tasks, single-paper evidence or a research suggestion may be generalized into a field trend, or new retrieval results may displace a sound incumbent direction. In Planning tasks, evidence that a method works

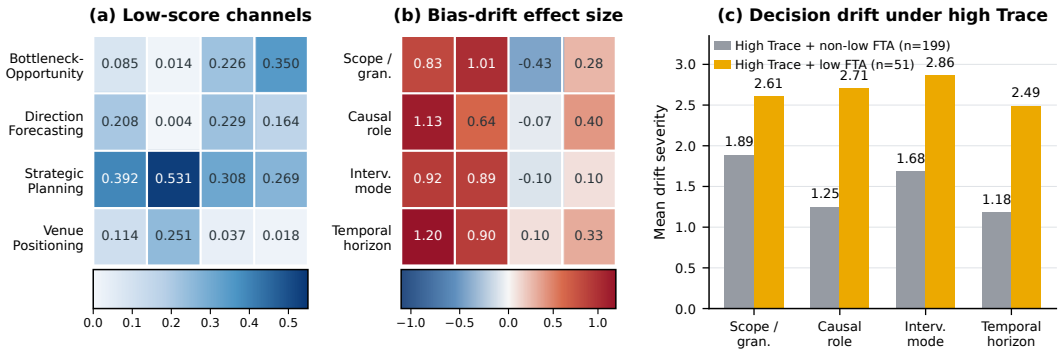

Figure A7: Full diagnostic drift–metric audit. (a) Family-dependent low-score channels; (b) normalized metric contrasts associated with the four drift dimensions; (c) drift severity among high-Trace answers. The figure uses the complete sample of 960 hash-matched annotated answers.

High Traceability But High Drift: A Strategic-Planning Case

Case	Qwen3-235B + CoI-style, RFPLAN-0136, strategic research planning
Task	Rank five visual world-model research directions for a six-month window, using only evidence available by 2025-09-30.
CoI-style answer	Ranks **embodied evaluation** last because standardized metrics and tooling remain immature, treating this gap as a reason to defer the direction.
Reference target	Ranks **embodied evaluation and feedback infrastructure** first because the field already has plausible generative mechanisms but lacks evidence that they improve embodied decisions.
Scores	Trace = **0.892**, Fact = 0.343, FTA = **0.200**, Pers. = 0.591.
Drift diagnosis	The answer gives a traceable rationale for a plausible ordering but treats missing evaluation infrastructure as a reason to deprioritize it rather than as the field's enabling bottleneck, producing causal-role, scope, and intervention-priority drift.

Figure A8: A high-traceability/high-drift planning example, showing the task, agent decision, reference target, and scores from the matched drift audit.

can substitute for evidence that it is the best next research investment. Agents may also compare a small implementation of one candidate with the cost of building another from scratch. ARIS cases show commitment to a retrieved technical approach; ResearchAgent cases include changes to previously aligned rankings. CoI can turn a locally coherent chain around mature evidence into a global priority judgment even when maturity is not the relevant investment criterion. In Venue tasks, agents may borrow the contribution of a related paper, mistake topical overlap for contribution fit, or treat a proposed evidence upgrade as an established property of the project. These substitutions can change both the default target and the conditions under which an alternative becomes appropriate.

Bottleneck controls. We extract evidence-supported chains linking observed failures, proximal mechanisms, and upstream causes. A task-level adjudicator checks explanatory coverage, abstraction, and coupled causes, then locks the root bottleneck, causal bridge, and one-hop opportunity before rendering. Chain extraction and rendering use Qwen3-235B; adjudication uses DeepSeek-v4-Pro with reasoning enabled.

Direction controls. ARIS and ResearchAgent retain their incumbent direction among the candidates considered during evidence retrieval and synthesis. CoI instead focuses retrieval on its incumbent. The final decision is based on evidence matching the task's temporal and conceptual scope.

Planning controls. Qwen first describes a feasible study, its decision value, and its uncertainty for each public candidate using the task window and existing sources. A second stage reconciles these studies with the original answer. ARIS commits only independently supported changes to the first-ranked candidate. ResearchAgent diagnoses whether the current decision bottleneck is measurement, capability, or integration; a second check commits only a supported change to the first-

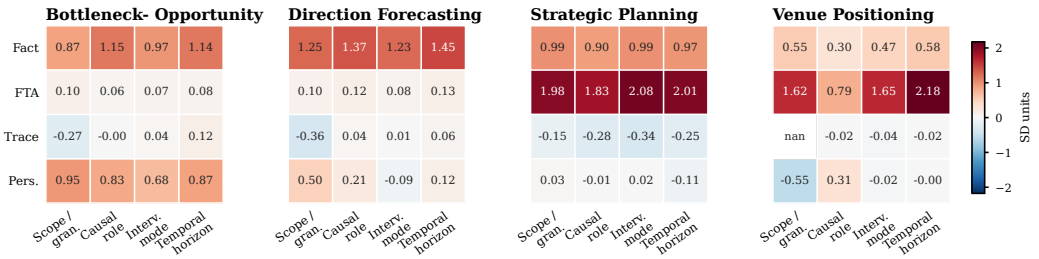

Figure A9: Family-level drift–metric associations. Each heatmap reports normalized score drop, $(\mathbb{E}[m \mid s = 0] - \mathbb{E}[m \mid s \geq 2])/\mathrm{SD}(m)$, for one task family. Rows are formal metrics and columns are answer-reference drift dimensions.

ranked candidate. CoI first constructs an evidence-conditioned candidate ranking at a common project scale. An independent gate then commits a changed ranking only when the decisive comparison uses the correct evidence role and is supported by an exact passage from the public task or supplied evidence. Maturity, frequency, and local implementation details cannot alone justify a priority change. Each agent-specific rule was frozen on 20 tasks before evaluation on the remaining 105.

Venue controls. Qwen constructs a *public contribution contract* from the project description: the contribution center, project kind, exact supporting passages, and current versus conditional fit to the candidate venue families. A second stage checks the original answer against this contract. Revision is constrained to contribution substitutions supported by exact task and answer passages: related work presented as the project's own contribution, a future upgrade presented as current evidence, or topical overlap used in place of contribution fit. The revised answer must retain a complete positioning decision, including a default target, conditional alternatives, contribution framing, reviewer risks, and evidence upgrades. If no supported conflict is found, or a revision fails its passage or format checks, the original answer is retained. The contract and revision rules were frozen on 20 development tasks before the 105-task confirmation. Venue retrieval, tasks, and gold targets are held fixed.

Matched evaluation. Each family contains 125 tasks, split into 20 discovery/development tasks and 105 disjoint confirmation tasks. Evaluation uses Qwen3-235B and the 105 confirmation tasks for all three agents: 315 baseline–intervention answer pairs and 1,260 paired dimension judgments per family. Both arms include the complete public question and cutoff. Generation uses the public question and cutoff-eligible evidence; hidden targets and drift scores are evaluation-only. Failed revisions retain the original answer and remain in the denominator. Qwen evaluates drift with the same four-dimension protocol in both arms. We bootstrap tasks, keeping their agents and dimensions together, using 10,000 resamples with seed 20260918 for Bottleneck and Direction and seed 20260915 for Planning. Table A16 reports per-agent effects. For Venue, both arms are judged separately with the full public question and current reference, without agent identity, metric values, or revision rationale. Venue intervals use 10,000 task-bootstrap resamples with seed 20260918, keeping all three agents together within each sampled task.

On the 105-task confirmation set, ARIS commits 18 revisions and changes drift/FTA from 1.7381/0.5980 to 1.5190/0.6681, while ResearchAgent commits 14 revisions and changes them from 1.3714/0.6870 to 1.2786/0.7049. CoI commits 27 revisions and changes drift/FTA from 1.5571/0.6440 to 1.5024/0.6640. FTA increases for all three Planning agents. The individual CoI and ResearchAgent drift confidence intervals include zero; the combined Planning drift interval excludes zero. Planning rank alignment maps equivalent candidate names to their public option identities before scoring.

Content verification. For Bottleneck, Direction, and Planning, content verification uses the 105 confirmation tasks per family and all three agents. Fact uses DeepSeek-v4-Pro in both arms, while FTA uses the family-specific semantic or ranking formulation. Unchanged answers reuse their exact matched results. Table A17 presents per-agent content scores. The aggregate confirmation effects are summarized in the main text (Figure 4). Bottleneck gains 0.089 in Fact and 0.005 in FTA; Direction

Table A16: Per-agent drift effects on 105 disjoint confirmation tasks per family (420 paired dimension judgments per row), excluding the 20 discovery/development tasks. Lower is better. Planning uses separately frozen controls for each agent. RA denotes ResearchAgent.

Task	Agent	Before	After	Δ [95% CI]
Bottleneck	ARIS	1.600	1.352	-0.248 [-0.381, -0.119]
	CoI	1.633	1.290	-0.343 [-0.488, -0.195]
	RA	1.576	1.333	-0.243 [-0.376, -0.112]
Direction	ARIS	1.793	1.679	-0.114 [-0.198, -0.038]
	CoI	1.569	1.562	-0.007 [-0.071, 0.052]
	RA	1.610	1.550	-0.060 [-0.126, 0.002]
Planning	ARIS	1.738	1.519	-0.219 [-0.343, -0.107]
	CoI	1.557	1.502	-0.055 [-0.207, 0.088]
	RA	1.371	1.279	-0.093 [-0.238, 0.033]

Fact increases from 0.562 to 0.579 and FTA from 0.632 to 0.651; Planning gains 0.066 in Fact and 0.036 in FTA. FTA increases in all four task families.

Table A17: Matched content verification with Qwen3-235B on 105 disjoint confirmation tasks per family, excluding the 20 discovery/development tasks. All three agents are evaluated; displayed rows present per-agent Fact and FTA scores, both higher-is-better. RA denotes ResearchAgent.

		Fact		FTA	
Task	Agent	Before	After	Before	After
Bottleneck	ARIS	0.8600	0.9242	0.6070	0.6131
	CoI	0.8510	0.9555	0.6070	0.6117
	RA	0.8020	0.9005	0.6080	0.6116
Direction	ARIS	0.5420	0.5590	0.6270	0.6440
	CoI	0.5770	0.5950	0.6350	0.6520
	RA	0.5680	0.5840	0.6330	0.6570
Planning	ARIS	0.4600	0.5245	0.5980	0.6681
	CoI	0.5040	0.5700	0.6440	0.6640
	RA	0.5320	0.5980	0.6870	0.7049

Venue confirmation. Of the 315 Venue pairs, 197 answers are revised and 118 retain the original, including 19 failed revision attempts; all remain in the evaluation denominator. Fact uses claim-level F1 over answer-claim support and hidden claim-bank coverage in both arms, with Qwen judging the paired answers against the same fixed hidden claim bank. FTA retains the benchmark's default-family match, conditional-alternative recall, and contribution-center alignment, using identical canonicalization and a shared embedding cache. Unchanged answers reuse their exact matched scores. Overall Fact increases from 0.5119 to 0.5450 ($\Delta = 0.0331$, 95% CI $[0.0125, 0.0539]$), and FTA from 0.6916 to 0.7437 ($\Delta = 0.0521$, 95% CI $[0.0044, 0.0972]$). Mean drift falls from 1.4778 to 1.2944 ($\Delta = -0.1833$, 95% CI $[-0.2595, -0.1056]$). At the task level, changes in scope and intervention-mode drift are negatively associated with FTA changes (Spearman $\rho = -0.366$ and -0.686, respectively), whereas causal-role changes have little association with Fact changes ($\rho = -0.057$). These are descriptive associations across the paired task means, not causal estimates of individual drift dimensions.