ForeSci: Evaluating LLM Agents for Forward-Looking AI Research Judgment

Anonymous ACL submission

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, and answer-generation backbones are selected to precede the task cutoffs. We evaluate native LLMs, Hybrid RAG, and three research-agent adaptations across four backbones. Results show that explicit evidence organization improves traceability and factual support, but gains depend strongly on the decision family. Diagnostics reveal a recurring evidence-decision decoupling: agents may cite relevant evidence while forecasting the wrong research object. ForeSci turns forward-looking AI research judgement into a controlled benchmark for evaluating research agents as decision-making systems.

1 Introduction

AI research moves on a timescale where today's frontier becomes tomorrow's baseline. The value of a research decision (*e.g.*, which bottleneck to attack, which direction is worth a six-month commitment) often lies in anticipating where the field is going. As autonomous research agents are increasingly deployed for ideation, planning, and scientific workflow execution (Lu et al., 2026; Li et al., 2024; 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 being asked to participate in this forward-looking decision layer. Whether current LLM agents can make defensible, evidence-grounded research judgements about an as-yet-unwritten future 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 components of future papers, such as related work, contribution content, citations, and impact (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 retrieval or in the backbone's training data. Otherwise, a system may rely on hindsight rather than foresight (Zhao et al., 2024; Ye et al., 2024; Liu et al., 2026; Ajith et al., 2026; Wang et al., 2026). Second, the tasks must be historically inferable. They should be grounded in signals available before the 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

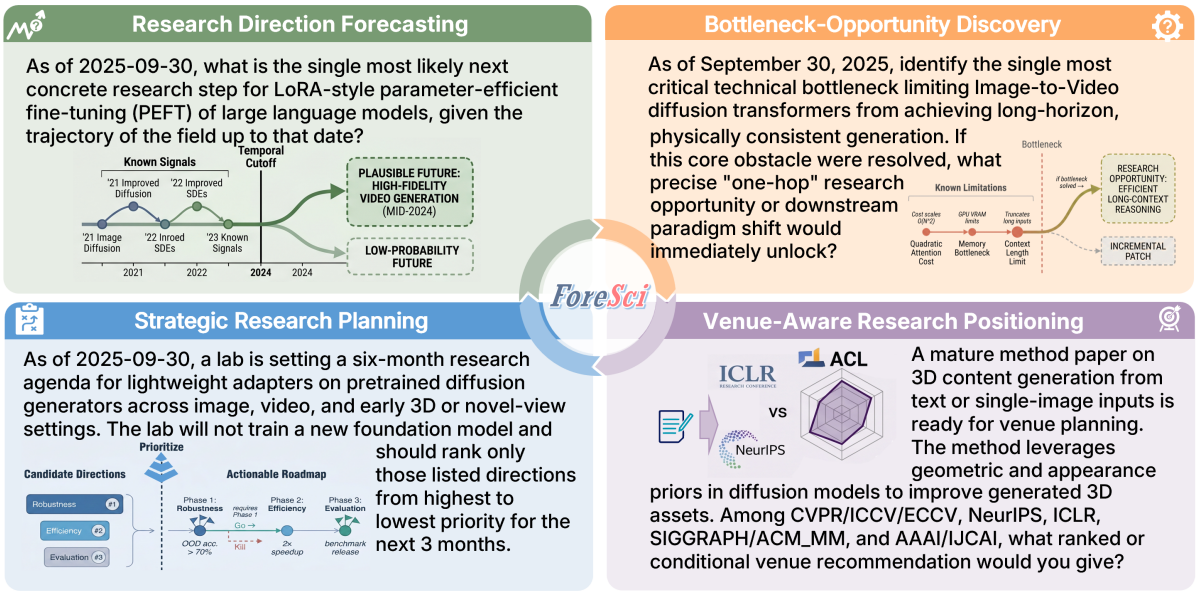

Figure 1: Representative ForeSci task examples across the four decision families: direction forecasting, bottleneck–opportunity discovery, strategic research planning, and venue-aware research positioning.

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. To avoid data leakage, all systems operate within the same historical knowledge base and all LLM backbones are trained before the time cutoff.

Results show that agent-style methods improve evidence traceability and factuality, but the strongest method differs by decision family. A diagnostic audit further reveals an evidence-decision decoupling: agents can cite relevant pre-cutoff evidence while forecasting the wrong object, misassigning causal roles, or selecting the wrong intervention. Beyond retrospective evaluation, we demonstrate that the same construction pipeline supports fully prospective forecasting, enabling continued evaluation of research agents as new literature emerges. 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 pre-cutoff backbones; the same pipeline supports fully prospective forecasting beyond retrospective evaluation
- **A multi-signal evaluation protocol** separating factuality, future-target alignment, evidence traceability, and reviewer persuasiveness, validated against human experts.
- **A systematic evaluation and diagnostic audit of LLM research agents** showing that agent-style methods improve traceability and factuality task-conditionally, 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., 2024), AI-Researcher (Tang et al., 2025), AI Scientist (Yamada et al., 2025), Intern-Atlas (Wu et al., 2026) 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 targets this decision layer.

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., 2025; Liu et al., 2025; 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. A few recent works begin to evaluate higher-order research capabilities beyond idea generation or workflow execution. Some focus on the novelty (Si et al., 2025; Schopf and Färber, 2026), taste (Tong et al., 2026), impact (Jiang, 2026; Zhu et al., 2026), and future alignment (Wang et al., 2026) of agent-generated ideas. PreScience (Ajith et al., 2026) moves further by predicting the components of future papers. Although these works leverage future papers or citation signals as evaluation references, a perspective related to our work, ForeSci focuses on a different research scenario: strategic, forward-looking, macro-level scientific decision-making.

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., 2026), Set the Clock (Zhao et al., 2024), ForecastBench (Karger et al., 2025), FutureX (Zeng et al., 2025), FOReCAst (Yuan et al., 2026), PROPHET (Tao et al., 2025), and MIRAI (Ye et al., 2024) all motivate time-sliced evaluation for future-oriented reasoning. While these benchmarks mainly evaluate future event prediction in general domains, ForeSci focuses on future-oriented scientific decision-making in fast-moving AI subfields. It therefore extends temporal control to open-ended research-agent outputs, pairing a cutoff-aligned offline knowledge base with hidden post-cutoff supervision.

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 avoid information leakage, we use answer-generation backbones trained before the relevant task cutoffs, disable web search, and allow systems to use only $\mathcal{K}_{\leq t}(q)$ as external support when producing answers.

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

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

Domain	KB Documents	Tasks
LLM Agents	2,769	138
LLM Fine-tuning and Post-training	2,131	99
RAG and Retrieval Structuring	767	92
Visual Generative Modeling and Diffusion	913	171

Table 1: Domain-level statistics in ForeSci. KB document counts come from the cutoff-aligned offline knowledge base; task counts come from the curated benchmark release.

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-level statistics are reported in Table 1; horizon details and paper-count statistics are provided in Table A1 and Figure A1. Additional construction details are provided in Appendix B.

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 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., 2026), 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 the taxonomy structures are first extracted through LLM from cutoff-aligned evidence, then checked by human experts who verifies support strength and temporal validity.

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 foresight challenges rather than merely the taxonomy's structure.

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) induced by the *method-development signals* after cutoff, combined with the necessary evidence before the cutoff.

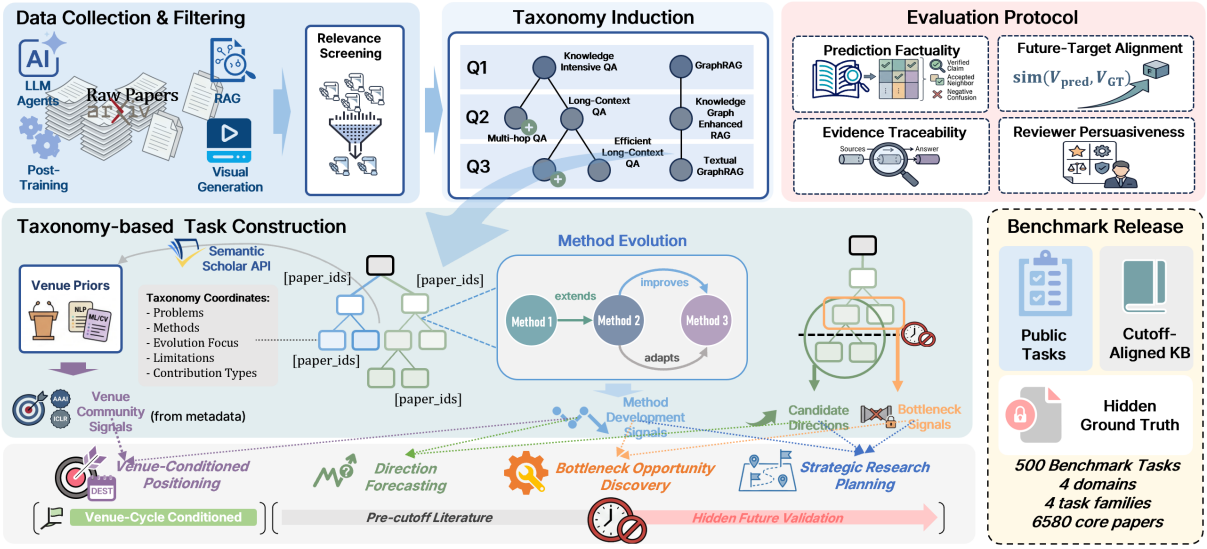

Figure 2: Construction process for the current formal ForeSci release. The figure shows the pipeline from corpus harvest and screening to temporal taxonomy induction, evidence and evolution asset construction, task-family builders, hidden validation targets, and the final benchmark release with public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

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 induced from *bottleneck signals* before t and *method-development signals* after t, including primary bottlenecks, acceptable bottleneck variants, unlocked opportunities, and mechanism descriptions, 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 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 a fixed set of venue or track options, the system must rank or conditionally recommend venue families, explain the appropriate framing, identify reviewer risks, and specify what evidence upgrades would make the project credible for the target venue community. The task uses contribution types from *full-text evidence*, and *venue-community metadata*. The hidden future validation targets $\mathcal{G}_{>t}(q)$ are venue-positioning decisions 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 B.

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 are designed to 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 virtual 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 whether the answer aligns with the task-family-specific future target in $\mathcal{G}_{>t}(q)$. For Direction Forecasting and Bottleneck–Opportunity Discovery, it compares extracted prediction claims with hidden claim bank using bge-m3 similarity. For Strategic Research Planning and Venue-Conditioned Positioning, where the target is an ordered decision, it computes deterministic ranking alignment against the hidden preferred ranking.

Evidence Traceability Score (*Trace*). This metric evaluates whether the answer can be traced to the pre-cutoff support packet $\mathcal{E}_{\leq t}(q)$. The evaluator scores whether the answer uses relevant pre-cutoff evidence, whether that evidence supports the stated decision, and whether the reasoning avoids unsupported jumps from the available literature. Evidence Traceability Score is reported as a normalized rubric score in $[0, 1]$.

Reviewer Persuasiveness (*Pers*). This metric evaluates whether the answer presents a strong research judgment persuasive to a LLM-based virtual reviewer. For each task family f, a rubric $\mathcal{R}_f$ scores $(q, a, \mathcal{E}_{\leq t}(q), \mathcal{G}_{>t}(q))$ on task-specific decision quality, mechanistic reasoning, comparative reasoning, clarity, and risk awareness:

$$\text{Pers.}(a, q) = \mathcal{R}_f(q, a, \mathcal{E}_{\leq t}(q), \mathcal{G}_{>t}(q)).$$

Automatic evaluation uses DeepSeek-V4 as the evaluator on the 500-task formal release. Appendix C.3 reports human validation for the automatic metrics. Appendix C gives family-specific prompts and scoring rules. For rubric-style metrics, we repeat evaluator runs and report the mean score with variance. This applies to Evidence Traceability Score (*Trace*) and Reviewer Persuasiveness (*Pers.*), where repeated scoring makes evaluator uncertainty visible.

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 D. We evaluate Qwen3-235B (released: April 29, 2025 (Qwen Team, 2025)), GPT-5.2 (knowledge cutoff: August 31, 2025 (OpenAI, 2025)), GLM-4.6 (released: September 30, 2025 (Z.AI, 2025)), and Gemini-3 (knowledge cutoff: January 2025 (Google, 2025)), LLM backbones trained before the cutoff time to avoid data leakage.

5 Results

5.1 Evaluation of LLM Agents

Table 2 reports Prediction Factuality (Fact), Future-Target Alignment (FTA), Evidence Traceability Score (Trace), and Reviewer Persuasiveness (Pers) across four backbones and five methods. Appendix E gives five-run evaluator stability views (Tables A8).

Agent-style methods generally improve evidence-grounded metrics. Across backbones, the strongest agent is competitive with or better than Native LLM and Hybrid RAG on Fact and FTA, and all three agents consistently improve Trace over Hybrid RAG. This suggests that agentic workflows can better align answers with future validation targets while exposing pre-cutoff grounding more explicitly. These gains do not consistently improve Reviewer Persuasiveness. One explanation is that backbones use retrieved or structured artifacts differently: for some, they support reasoning; for others, they add noise behind a coherent final justification, lowering the quality of the judgment report. Method rankings also vary by task family (Table A7). No agent is uniformly strongest across metrics, backbones, and task families, and in some settings agentic methods show no clear advantage over the native backbone. Additional retrieval and tool use therefore do not automatically translate into better foresight performance, motivating the error analysis below.

Method	Qwen3-235B				GPT-5.2				GLM-4.6				Gemini-3			
	Fact.	FTA	Trace	Pers	Fact.	FTA	Trace	Pers	Fact.	FTA	Trace	Pers	Fact.	FTA	Trace	Pers
Native LLM	0.603	0.622	–	0.786	0.618	0.628	–	0.846	0.509	0.590	–	**0.674**	0.544	**0.609**	–	**0.741**
Hybrid RAG	0.597	0.630	0.432	0.775	0.610	0.626	0.408	0.837	0.520	0.613	0.429	0.658	0.559	0.598	0.413	0.720
CoI	**0.611**	0.642	0.560	0.782	0.626	0.632	0.593	0.857	**0.543**	0.620	0.499	0.662	**0.563**	0.606	0.473	0.734
ResearchAgent	0.609	**0.660**	0.563	0.787	**0.635**	0.633	0.584	0.857	0.540	**0.633**	0.499	0.656	0.562	**0.609**	0.459	0.729
ARIS	0.607	0.644	**0.608**	**0.793**	0.617	**0.642**	**0.627**	**0.861**	0.537	0.619	**0.520**	0.649	0.560	0.602	**0.567**	0.733

Table 2: Overall results on ForeSci. Bold marks the best method within the same backbone and metric.

5.2 Error Mechanisms: When Foresight Fails

Family-dependent failures We further conduct an internal error analysis to demonstrate how LLM agents fail in foresight tasks. We first identify low-scoring cases for each metric using the bottom 20% of rows as the low-score threshold, and then compute the fraction of low-score cases within each task family. Figure 3(a) shows that failures are strongly family-dependent. For example, Strategic Planning has the highest low-score rates on Fact and FTA, reflecting the difficulty of matching both the ranked decision and its supporting facts. Together, these patterns motivate the use of multiple evaluation signals, as Fact, FTA, Trace, and Persuasiveness reveal distinct failure channels that a single aggregate score would obscure.

Evidence-to-decision drift We then analyze evidence-to-decision drift by comparing model answers with reference answers. We use LLM-based classification with human expert verification to identify four common types of answer drift: (1) *Scope/granularity drift* occurs when the answer discusses a related research direction but at the wrong level of specificity. (2) *Causal-role drift* occurs when the answer assigns the wrong role to a technical factor, such as treating an enabled opportunity as the root bottleneck.(3) *Intervention-mode drift* occurs when the answer targets the right general issue but recommends the wrong type of intervention, such as proposing system integration improvements when the reference calls for a change in the training objective. (4) *Temporal-horizon drift* occurs when the answer targets the wrong maturity stage, such as jumping from a near-term opportunity to a much longer-term vision. Each drift type is annotated with a severity score in $[0, 3]$, where 0 indicates no drift and 3 indicates severe drift. For each drift type, we sample 20 tasks per family and include all five methods and four backbones, yielding 1600 matched answers and 6400 dimension-level annotations.

To quantify the metric impact of each drift type, we compute a normalized effect size:

$$\Delta_{\text{norm}}(m) = \frac{\mathbb{E}[m \mid s = 0] - \mathbb{E}[m \mid s \geq 2]}{\text{SD}(m)}.$$

where m is the target metric and s is the annotated drift severity. Figure 3(b) shows that severe drift substantially reduces the content-facing metrics. Causal-role drift lowers Fact by 1.13 standard deviations, while scope/granularity and intervention-mode drift lower FTA by 1.22 and 1.12 standard deviations, respectively. Persuasiveness also declines under severe drift, but less uniformly. In contrast, Trace is much more weakly coupled to these content drifts and its direction depends on the drift type.

High traceability but high drift We therefore further inspect high-Trace cases in Figure 3(c). Among answers with high Trace, the low-FTA subset has much higher drift severity across all four bias types than the non-low-FTA subset. This confirms that an answer can be well supported by local evidence while still selecting the wrong decision object, causal role, intervention type, or time horizon. A compact case in Appendix Figure A6 illustrates the distinction. A Gemini-3 ARIS answer for a venue positioning task has high traceability (0.920) but low Prediction Factuality (0.200) and low FTA (0.355): it gives a plausible NeurIPS-first framing for a reinforcement-learning-from-AI-feedback contribution, but this framing is less aligned with the task's reference target, which prioritizes ACL/EMNLP because the work is framed as language-model post-training and alignment.

Method fingerprints We also examine the content-level failure patterns of different methods and find that each method has a distinct diagnostic fingerprint, summarized in Appendix F.1. Overall, these results suggest that agentic evidence organization should be used with caution: while agents can improve traceability, they may also over-amplify locally supported but decision-misaligned evidence,

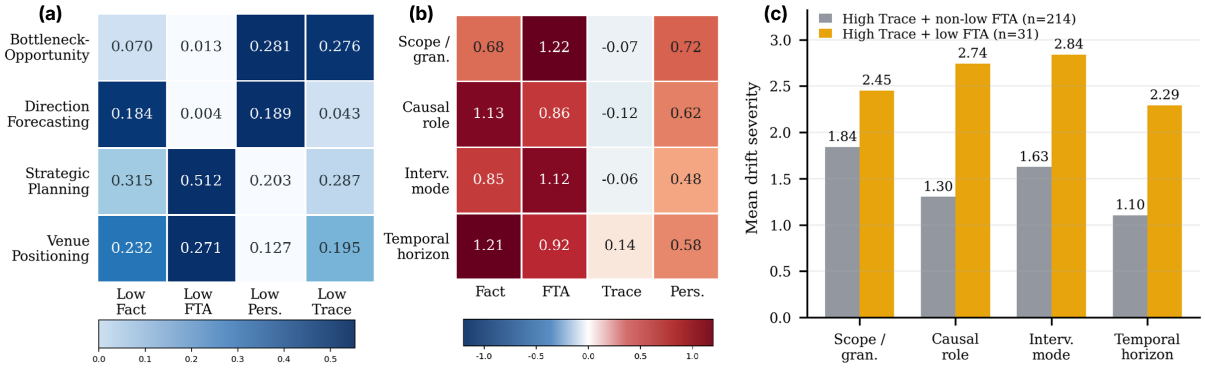

Figure 3: Low-score channels and drift-induced effects. (a) Bottom-20% low-score rates across evaluation metrics and task families. Each cell reports the fraction of examples in a task family that fall into the low-score set for a given metric. (b) Normalized metric drop caused by evidence-to-decision drift, computed by comparing cases with no drift severity against cases with severe drift. (c) Drift severity among high-traceability answers. High-traceability but low-FTA answers exhibit substantially higher drift severity across all drift types.

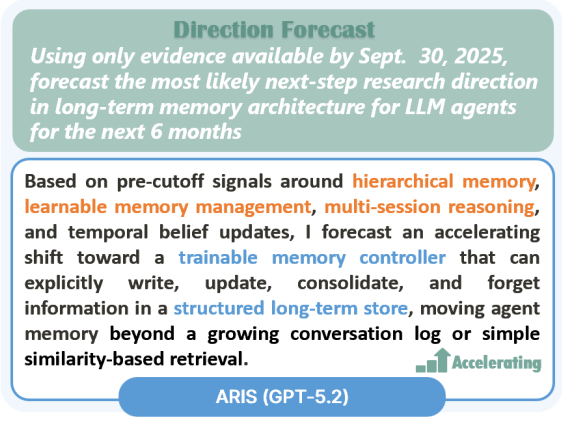

Figure 4: Prospective forecasting showcase for a Direction Forecasting task. The displayed agent answer is a summarized version of the full generated response, retaining the predicted direction, trajectory label, and core rationale.

thereby steering the model toward a confidently grounded yet incorrect research judgment.

5.3 Prospective Use: Dynamic Forecasting Beyond Retrospective Evaluation

ForeSci is designed not only for retrospective evaluation but also for fully prospective forecasting. As a proof of concept, we apply the same cutoff-controlled taxonomy and evidence-construction pipeline to the LLM-agent domain with a 2026-05-15 literature cutoff, producing 12 prediction-only questions balanced across the four task families for the 2026-05-16 to 2026-08-15 forecast window. Because the target outcomes had not occurred at writing time, this package is not scored; instead, it demonstrates that the framework can be refreshed with recent literature to generate transparent forecast artifacts. In the main text, we show one representative agent-generated forecast case to illustrate how a system turns cutoff-visible evidence into a concrete forward-looking research judgment (Figure 4). This prospective mode enables dynamic evaluation of newly released LLM agents and can also support evidence-grounded AI research planning before future results are known. Additional package details and generated examples are provided in Appendix G.

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. Results show that agentic workflows often improve traceability and some evidence-grounded metrics, but no method is uniformly best across backbones, task families, and evaluation signals. The diagnostics further reveal evidence-decision decoupling: agents can cite relevant evidence yet choose the wrong research object, causal role, intervention mode, or time horizon. By separating factual support, future-target alignment, traceability, and reviewer-style persuasiveness, ForeSci makes these failures measurable. Its prospective mode also shows how refreshed literature can produce transparent forecast artifacts, supporting evaluation of research agents as decision-making systems rather than literature interfaces alone.

Limitations

ForeSci studies forward-looking research judgement in four fast-moving AI areas and four decision families. Its results should therefore be interpreted as evidence about this controlled benchmark setting, not as a universal ranking of research agents across all scientific domains, languages, or time horizons. The benchmark emphasizes paper-visible signals; it cannot fully capture tacit community knowledge, unpublished work, private reviewer expectations, or downstream adoption.

The evaluation also depends on hidden post-cutoff targets and LLM-as-judge metrics. We use family-conditioned rubrics, repeated judging for rubric-style metrics, cross-backbone comparisons, and diagnostic audits to reduce over-interpretation, but the scores remain approximations of rubric-based reviewer persuasiveness rather than direct measurements of scientific value. In particular, venue positioning and strategic planning are inherently preference-sensitive decisions, so the benchmark is best used to compare failure modes and evidence use rather than to certify a single best method.

Ethical Considerations

The benchmark is built from public scholarly artifacts and is intended for diagnostic evaluation of research-assistant systems. It should not be used to automate real venue recommendations, peer-review decisions, or research prioritization without human oversight. Because the tasks ask systems to make forward-looking research decisions, a poorly calibrated system could encourage premature convergence on fashionable directions or overstate the evidential basis for a forecast. We therefore report traceability, uncertainty-sensitive reviewer scores, and limitations alongside outcome-oriented metrics.

References

ACL. 2025. ACL 2025 call for main conference papers. https://2025.aclweb.org/calls/main_conference_papers/. Accessed 2026-05-20.

Anirudh Ajith, Amanpreet Singh, Jay DeYoung, Nadav Kunievsky, Austin C. Kozlowski, Oyvind Tafjord, James Evans, Daniel S. Weld, Tom Hope, and Doug Downey. 2026. Prescience: A benchmark for forecasting scientific contributions. *arXiv preprint arXiv: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, and 20 others. 2025. Astabench: Rigorous benchmarking of AI agents with a scientific research suite. In *International Conference on Learning Representations*.

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

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

CVPR. 2025. CVPR 2025 call for papers. https://cvpr.thecvf.com/Conferences/2025/CallForPapers. Accessed 2026-05-20.

ECCV. 2024. ECCV 2024 call for papers. https://eccv2024.ecva.net/Conferences/2024/CallForPapers. Accessed 2026-05-20.

EMNLP. 2025. EMNLP 2025 call for main conference papers. https://2025.emnlp.org/calls/main_conference_papers/. Accessed 2026-05-20.

Olivier Francois. 2015. Arbitrariness of peer review: A bayesian analysis of the NIPS experiment. *arXiv preprint arXiv: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. 2026. A multi-agent system for automating scientific discovery. *Nature*, pages 1–3.

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

Mourad Gridach, Jay Nanavati, Khaldoun Zine El Abidine, Lenon Mendes, and Christina Mack. 2025. Agentic AI for scientific discovery: A survey of progress, challenges, and future directions. *arXiv preprint arXiv:2503.08979*.

ICCV. 2025. ICCV 2025 call for papers. https://iccv.thecvf.com/Conferences/2025/CallForPapers. Accessed 2026-05-20.

ICLR. 2025. ICLR 2025 call for papers. https://iclr.cc/Conferences/2025/CallForPapers. Accessed 2026-05-20.

ICML. 2025. ICML 2025 call for papers. https://icml.cc/Conferences/2025/CallForPapers. Accessed 2026-05-20.

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

Bo Jiang. 2026. HindSight: Evaluating LLM-generated research ideas via future impact. *arXiv preprint arXiv:2603.15164*.

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

Priyanka Kargupta, Nan Zhang, Yunyi Zhang, Rui Zhang, Prasenjit Mitra, and Jiawei Han. 2025. 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)*, pages 29834–29850.

KDD. 2025. KDD 2025 research track call for papers. https://kdd2025.kdd.org/research-track-call-for-papers/. Accessed 2026-05-20.

Jakub Lála, Odhran O'Donoghue, Aleksandar Shtedritski, Sam Cox, Samuel G. Rodriques, and Andrew D. White. 2023. Paperqa: Retrieval-augmented generative agent for scientific research. *arXiv preprint arXiv: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. 2020. Retrieval-augmented generation for knowledge-intensive nlp. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Long Li, Weiwen Xu, Jiayan Guo, Ruochen Zhao, Xingxuan Li, Yuqian Yuan, Boqiang Zhang, Yuming Jiang, Yifei Xin, Ronghao Dang, Deli Zhao, Yu Rong, Tian Feng, and Lidong Bing. 2024. Chain of ideas: Revolutionizing research via novel idea development with llm agents. *arXiv preprint arXiv:2410.13185*.

Yachuan Liu, Xiaochun Wei, Lin Shi, Xinnuo Li, Bohan Zhang, Paramveer S Dhillon, and Qiaozhu Mei. 2026. 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)*, pages 1551–1571.

Yujie Liu, Zonglin Yang, Tong Xie, Jinjie Ni, Ben Gao, Yuqiang Li, Shixiang Tang, Wanli Ouyang, Erik Cambria, and Dongzhan Zhou. 2025. Researchbench: Benchmarking LLMs in scientific discovery via inspiration-based task decomposition. *arXiv preprint arXiv:2503.21248*.

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

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

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, and 18 others. 2026. AIRS-bench: a suite of tasks for frontier AI research science agents. *arXiv preprint arXiv:2602.06855*.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. 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*, pages 12076–12100.

NeurIPS. 2025. NeurIPS 2025 call for papers. https://neurips.cc/Conferences/2025/CallForPapers. Accessed 2026-05-20.

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

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

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

Tim Schopf and Michael Färber. 2026. Is this idea novel? an automated benchmark for judgment of research ideas. *arXiv preprint arXiv:2603.10303*.

Chenglei Si, Diyi Yang, and Tatsunori Hashimoto. 2025. Can llms generate novel research ideas? a large-scale human study with 100+ nlp researchers. In *International Conference on Learning Representations*, volume 2025, pages 94003–94092.

SIGIR. 2025. SIGIR 2025 call for full papers. https://sigir2025.dei.unipd.it/call-full-papers.html. Accessed 2026-05-20.

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

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

Jingqi Tong, Mingzhe Li, Hangcheng Li, Yongzhuo Yang, Yurong Mou, Weijie Ma, Zhiheng Xi, Hongji Chen, Xiaoran Liu, Qinyuan Cheng, Ming Zhang, Qiguang Chen, Weifeng Ge, Qipeng Guo, Tianlei Ying, Tianxiang Sun, Yining Zheng, Xinchi Chen, Jun Zhao, and 4 others. 2026. Ai can learn scientific taste. *arXiv preprint arXiv:2603.14473*.

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

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

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

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. 2026. Internatlas: A methodological evolution graph as research infrastructure for AI scientists. *arXiv preprint arXiv:2604.28158*.

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

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2023. React: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations (ICLR)*.

Chenchen Ye, Ziniu Hu, Yihe Deng, Zijie Huang, Mingyu Derek Ma, Yanqiao Zhu, and Wei Wang. 2024. MIRAI: Evaluating LLM agents for event forecasting. *arXiv preprint arXiv:2407.01231*.

Moy Yuan, Zifeng Ding, and Andreas Vlachos. 2026. Introducing FOReCAst: The future outcome reasoning and confidence assessment benchmark. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Z.AI. 2025. GLM-4.6 release notes. `https://docs.z.ai/guides/llm/glm-4.6`. Accessed 2026-05-26.

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, and 12 others. 2025. Futurex: An advanced live benchmark for llm agents in future prediction. *arXiv preprint arXiv:2508.11987*.

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

Hangxiao Zhu, Yuyu Zhang, Ping Nie, and Yu Zhang. 2026. Sciimpact: A multi-dimensional, multi-field benchmark for scientific impact prediction. *arXiv preprint arXiv:2604.17141*.

A Responsible Research and Artifact Details

Artifacts, licenses, and intended use. ForeSci releases benchmark tasks, evaluation goldsets, prompts, schemas, scripts, and a cutoff-aligned scholarly knowledge base for research evaluation. The released benchmark artifacts are intended for diagnostic comparison of research-assistant systems under explicit temporal controls, not for automated peer-review, venue selection, hiring, funding, or research-prioritization decisions. Source scholarly papers and bibliographic records retain their original access conditions; derived benchmark packages should therefore be used consistently with the repository and release terms and with the access conditions of the underlying public scholarly artifacts.

Privacy and content review. The benchmark is constructed from public scholarly artifacts rather than private user data. We do not collect demographic attributes, private communications, or personally sensitive records. Released task files contain only public task text and minimal metadata needed for answer generation, while evaluation goldsets are separated from model-visible tasks. Human-validation results are reported only in aggregate, and no individual annotator records are released.

Compute and model access. All experiments are inference-only; no model training or fine-tuning is performed. We evaluate named answer-generation backbones and evaluator models through hosted or locally served API-compatible endpoints. Exact parameter counts are not publicly available for some hosted/proprietary models, so we report model names and access modes where exact sizes cannot be verified. The offline knowledge base and retrieval indexes are built once and then reused across methods. We do not tune method hyperparameters on hidden future targets; generation and evaluation use fixed prompt templates, retrieval settings, and metric rubrics.

Human validation protocol. Human validation is limited to expert annotation of model outputs and extracted claims. The validation pool consists of eight AI researchers: five PhD students and three faculty advisors with expertise in artificial intelligence. Annotators were recruited for expert validation rather than through a crowdsourcing marketplace. They are asked to follow the rubrics described in Appendix C.3: for Reviewer Persuasiveness, they score whether an answer would be convincing to a knowledgeable reviewer for the specified task family; for claim extraction, they check whether extracted units are faithful to the source answer, atomic, decision-relevant, and sufficiently complete. Annotators are informed that labels are used only for aggregate validation of the benchmark metrics. No crowdworker marketplace is used, no private personal data are collected, and no individual-level annotations are released. Because the protocol consists of expert assessment of model outputs and benchmark claims, with no intervention on human subjects or collection of sensitive personal data, it is treated as minimal-risk expert annotation.

Use of AI assistants. AI assistants were used during code prototyping, experiment orchestration, result checking, LaTeX editing, and drafting support under author supervision. The authors made the final decisions about benchmark design, data curation, experimental protocol, reported results, and paper claims.

B Benchmark Construction Details

This appendix expands the construction details behind Figure 2. The main text defines the benchmark corpus, taxonomy-based evidence layer, and task families; here we focus on implementation choices that affect cutoff alignment, auditability, and validation.

B.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.

Construction enforces temporal separation. 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. This design gives every system the same pre-cutoff literature environment and prevents direct retrieval of hidden future papers.

B.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.

B.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.

B.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.

B.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.

B.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.

B.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

Domain	3-month	6-month	Venue cycle	Total
LLM Agents	25	90	23	138
LLM Fine-tuning and Post-training	15	69	15	99
RAG and Retrieval Structuring	10	69	13	92
Visual Generative Modeling and Diffusion	19	78	74	171
Total	69	306	125	500

Table A1: Task counts by domain and horizon type. Three-month and six-month columns come directly from task-level horizon metadata; venue-conditioned tasks use venue-cycle timing because their advisory and submission windows are venue specific.

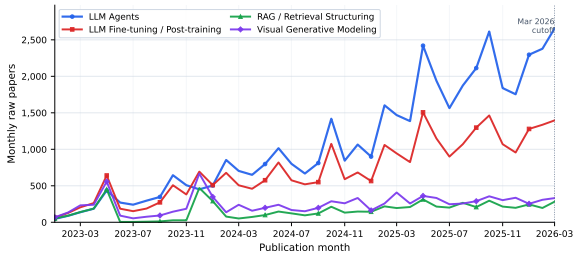

Figure A1: Monthly raw-paper volume in the four ForeSci domains. Counts are unique papers after normalizing paper identifiers within each domain-month. All four domains are shown from January 2023 through the March 2026 benchmark cutoff.

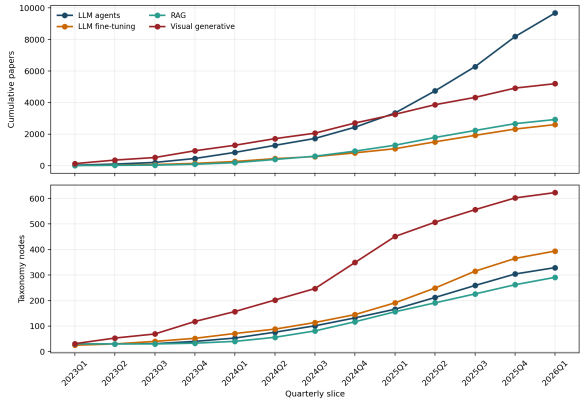

Figure A2: Temporal evolution of the domain corpora and induced taxonomies used in ForeSci. The curves summarize cumulative domain KB coverage and taxonomy-node growth over the literature slices used for the release.

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.

Publication-calendar effects. The volume curves should be read as a cutoff-dependent background variable rather than as a direct measure of foresight difficulty. LLM-heavy domains show recurring increases around late winter, May–June, and early autumn, broadly matching major AI submission cycles such as ICML (ICML, 2025), ACL (ACL, 2025), KDD (KDD, 2025), SIGIR (SIGIR, 2025), NeurIPS (NeurIPS, 2025), EMNLP (EMNLP, 2025), and ICLR (ICLR, 2025). Visual generation also exhibits spring and year-end structure consistent with CV venue cycles such as CVPR (CVPR, 2025), ICCV (ICCV, 2025), and ECCV (ECCV, 2024). These regularities motivate explicit temporal cutoffs and horizon metadata, so that task construction separates genuine post-cutoff research change from predictable seasonality induced by publication and review calendars.

The evolution curves in Figure A2 provide two useful checks on benchmark construction. First, they show that the domains differ substantially in corpus scale and structural breadth, which motivates using a shared construction protocol across diverse technical areas. Second, the branch examples in Figure A3 and method-evolution signals in Figure A4 show that broad nodes split into increasingly specialized descendants and recurring bottlenecks become concrete mechanisms over time rather than remaining a static flat inventory of labels. This is why ForeSci separates temporal taxonomy induction, public support construction, and hidden future supervision.

C Evaluation Protocol

C.1 Metric Calibration Details

This appendix gives the formulas and weighting details for the evaluation protocol introduced in Section 4. The reported metrics are Prediction Factuality, Future-Target Alignment, Evidence Traceability Score, and Reviewer Persuasiveness.

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.

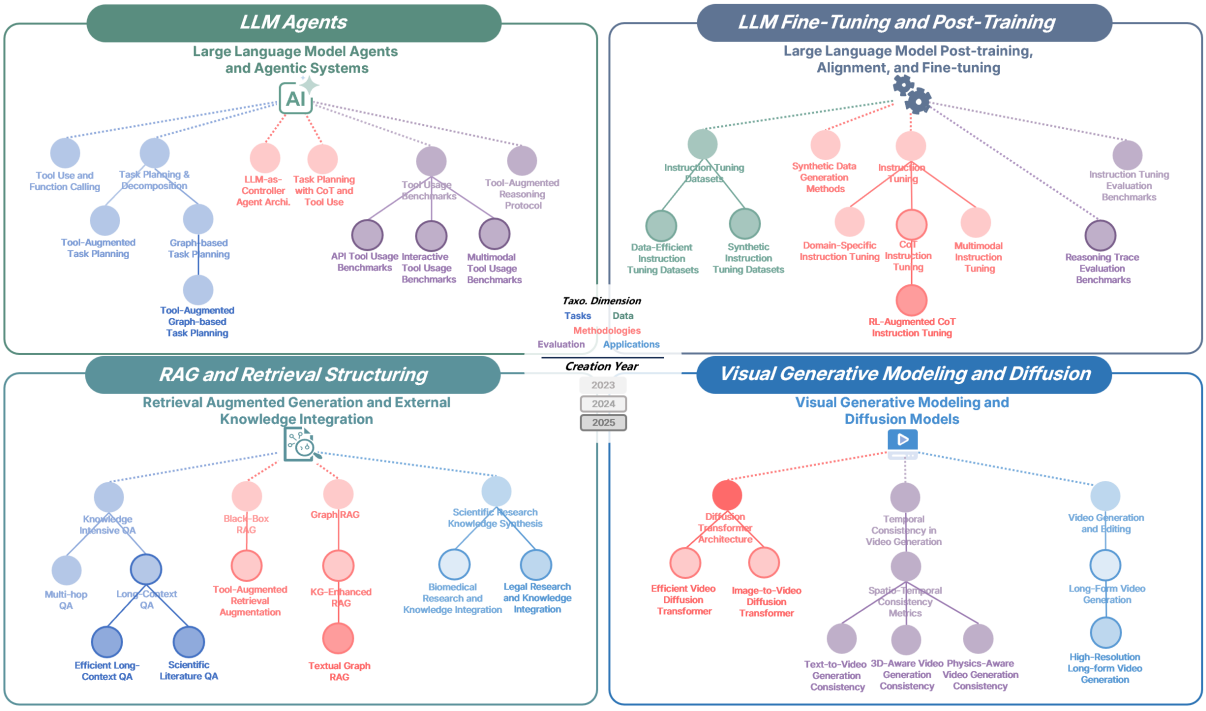

Figure A3: Illustrative examples of temporal taxonomy branching in ForeSci. The figure shows how broad pre-cutoff topics split into task-seeding subdirections; Table A2 summarizes representative branch examples in text form.

	Stage 1 *Prior Method*	Stage 2 *Bottleneck*	Stage 3 *Mechanism*	Stage 4 *Future Shift*
LLM Agents: From Tool-Use to Scientific Autonomy	Focuses on generic tool-calling and basic task decomposition within isolated environments.	Failure in tool reliability and long-horizon planning. Lack of grounding.	Structured orchestration, persistent memory, and multi-step verification loops.	Toward domain-grounded scientific workflow agents.
LLM Fine-Tuning & Post-Training: From Curation to Self-Evolution	Relies on manually curated instruction tuning and human preference data (RLHF).	High data acquisition costs, limited coverage of edge cases, and a lack of reasoning.	Automated data selection, synthetic feedback loops, self-correction algorithms.	Toward synthetic and reasoning-aware post-training pipelines.
RAG & Retrieval Structuring: From Static Retrieval to Agentic Reasoning	Single, static pre-processing step retrieval in "Retrieve-then-Generate" pipelines.	Struggles with hallucinations, multi-hop reasoning gaps, and the use of unverified evidence.	Adaptive retrieval, query decomposition, reasoning-aware reranking	Evolves into evidence-aligned agentic RAG systems.
Visual Generative Modeling: From Static Fidelity to Physical Simulation	Focuses on image-level diffusion and static metrics like Fréchet Inception Distance (FID).	Faces temporal inconsistency, lack of controllability, and violations of physical laws.	DiT, motion-aware conditioning, 3D/physics constraints, and temporal consistency metrics.	Targets controllable, long-form, and physics-aware video generation.

Figure A4: Method-evolution signals for foresight task construction. Rather than tracking topic frequency alone, 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.

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.

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

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.

Claim recall is

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

The final score is

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

Precision and recall are retained as intermediate quantities; the reported metric is the F1 score.

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:

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

This score is intentionally not an F1 and has no code-side cap or weighted combination term.

For strategic planning and venue positioning, the hidden target is an ordered list $\pi^* = (r_1, \ldots, r_K)$ over candidate directions or venues. 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

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

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

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

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:

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

The reported FTA is $\mathrm{FTA}_{\mathrm{RG}}$ for Direction and Bottleneck tasks, and $\mathrm{FTA}_{\mathrm{rank}}$ for Planning and Venue tasks.

Evidence Traceability Score. The traceability evaluator scores external evidence linkage $e(a)$, support specificity $s(a)$, and answer-internal trace $t(a)$ from the support snapshot attached to a method output. The final score is

$$\mathrm{Trace}(a) = 0.50\, e(a) + 0.25\, s(a) + 0.25\, t(a). \tag{12}$$

If a method output has no attached external support artifact, Evidence Traceability is not applicable and is reported as – in paper-facing tables. We compute Traceability aggregates, low-score rates, and cross-model trace-profile correlations only over evidence-grounded methods.

Reviewer Persuasiveness. Reviewer Persuasiveness is a deliberative LLM-as-judge metric over how persuasive the research decision would be to a rubric-based virtual reviewer, rather than over factual overlap alone. The evaluator first builds a family-conditioned research brief from the public task and hidden decision targets. For direction forecasting, the task-specific dimensions are forecast specificity, signal interpretation, trajectory discipline, and uncertainty calibration. For venue positioning, they are venue-fit reasoning, reviewer-expectation awareness, package-upgrade specificity, and contrastive venue discrimination. For bottleneck–opportunity tasks, they are causal bottleneck analysis, opportunity plausibility, technical non-obviousness, and adoption-pathway reasoning. For strategic planning, they are milestone specificity, dependency-chain quality, experiment executability, and risk or kill criteria. The evaluator also reports generic decision clarity, mechanistic reasoning, comparative reasoning, and uncertainty or risk awareness, then assigns a holistic persuasiveness score. Because this metric is intended to approximate a peer-review-style research call under uncertainty, repeated evaluator runs are used to make variance visible.

C.2 Evaluation Prompt Templates

This appendix summarizes the main prompt templates used by the benchmark evaluation stack. The templates below are lightly edited for readability, but they preserve the operative instructions, rubrics, and output schemas used in the implementation.

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 Prompt

You are an Evidence Traceability Auditor for a research benchmark.
Objective:

- Judge whether the answer's important conclusions can be traced to explicit support artifacts provided with the method output.
- Do not judge factual truth against outside knowledge.
- Reward explicit linkage from claims to papers, snippets, evidence bundles, or reasoning traces.
- Penalize answers that appear strong but cannot be connected to identifiable support.

Inputs:

- Public task definition: `{PUBLIC_TASK_JSON}`
- Candidate answer: `{CANDIDATE_ANSWER}`
- Extracted evidence-content brief: `{SUPPORT_BRIEF_JSON}`

Rubric:

- `evidence_linkage`: are the answer's main claims visibly connected to explicit evidence items, retrieved papers, snippets, or trace steps?
- `support_specificity`: is the support concrete and discriminative enough that a reviewer could audit why these conclusions were reached?
- `answer_internal_trace`: does the answer explain a coherent path from evidence to conclusion, rather than only listing anchors?

Output format:

```
{
  "dimension_scores": {"evidence_linkage": 0.0, "support_specificity": 0.0,
    "answer_internal_trace": 0.0},
  "traceability_score": 0.0,
  "strengths": ["..."],
  "weaknesses": ["..."],
  "rationale": "..."
}
```

Prompt template used for Evidence Traceability auditing.

Reviewer Persuasiveness Prompt

You are a Reviewer Persuasiveness evaluator for ForeSci.
Objective:

- Assess the quality of the candidate answer as a research decision under uncertainty.
- Use the reference brief as a structured decision neighborhood, not as a literal answer key.
- Reward clear tradeoffs, causal or venue-fit mechanism, concrete decision criteria, appropriate uncertainty, and explicit handling of alternatives or risks.
- Penalize noncommittal lists, generic prose, unsupported drift from the task, and missing task-specific decision requirements.

Inputs:

- Public task definition: `{PUBLIC_TASK_JSON}`
- Reviewer-persuasiveness brief: `{RESEARCH_JUDGEMENT_BRIEF_JSON}`
- Candidate answer: `{CANDIDATE_ANSWER}`
- Task-specific dimensions for this family: `{TASK_SPECIFIC_DIMENSIONS_JSON}`

Output format:

```
{
  "research_judgment_score": 0.0,
  "generic_dimension_scores": {
    "decision_clarity": 0.0,
    "mechanistic_reasoning": 0.0,
    "comparative_reasoning": 0.0,
    "uncertainty_or_risk_awareness": 0.0
  },
  "task_specific_dimension_scores": {"dimension_name": 0.0},
  "strengths": ["..."],
  "weaknesses": ["..."],
  "rationale": "..."
}
```

Prompt template used for family-conditioned Reviewer Persuasiveness.

C.3 Human Validation

We conduct human-validation studies over the formal evaluation artifacts. The studies check whether the rubric-based Reviewer Persuasiveness score tracks expert preferences and whether automatic claim extraction produces decision-critical units.

Validation target	Human sample	Main agreement result	Main diagnostic conclusion
Reviewer Persuasiveness	400 answers	Mean Spearman 0.57; mean top-1 agreement 50.0%	The rubric evaluator is positively aligned with expert scoring, but planning exposes sensitivity to polished scaffolded prose.
Claim extraction	240 answers	Overall F_1 0.77; atomicity pass 85.5%	The extractor is precise and usually atomic, but it misses implicit causal bridges and over-splits verbose planning or venue answers.

Table A3: Summary of human-validation analyses for the ForeSci evaluation stack.

Family	n	Pearson
Bottleneck	100	0.62
Direction	100	0.66
Strategic Planning	100	0.54
Venue	100	0.58

Table A4: Human expert scores versus automatic Reviewer Persuasiveness scores, reported by task family.

C.3.1 Reviewer Persuasiveness Human Validation

We first evaluate whether the automatic Reviewer Persuasiveness score reflects expert judgments on open-ended research decisions. The validation set contains 400 long-form model answers, balanced across the four task families. For each family, we compare human expert scores with the automatic DeepSeek-V4 Reviewer Persuasiveness score using Spearman correlation (Table A4) and ranking comparison (Table A5). Overall, the automatic Reviewer Persuasiveness metric shows strong agreement with human expert judgments across task families, supporting its validity as a reasonable proxy for human assessment.

C.3.2 Claim Extraction Human Validation

We next validate the claim extractor used by Prediction Factuality. The audit covers 240 answers, with 60 answers per family (Table A6). Human annotators assess whether extracted units are precise, decision-relevant, atomic, and sufficiently complete relative to the source answer. Overall, the extractor achieves strong human-validated quality across all four task families, with high precision, solid recall, and consistently high atomicity and decision relevance. These results support the reliability of Prediction Factuality as a claim-level measure of evidence-grounded research judgment.

D System Adaptation Details

The agent-based systems compared in the main paper were not evaluated in their original open-web form. All three were adapted into *cutoff-faithful offline agents* so that they operate inside the benchmark's pre-cutoff knowledge boundary and expose artifacts compatible with the ForeSci metric stack.

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. For CoI-style, following the open-web removal and cutoff-faithful evidence-flow principles, we replaced its online literature-search module with an offline benchmark-KB retrieval adaptor while retaining the multi-chain idea-construction module. For ResearchAgent-style, following the benchmark-native output and task-family alignment principles, we replaced web-facing evidence gathering with benchmark support-packet access and modified the internal planning/review prompts into task-family-aware modules for bottleneck, forecasting, strategic-planning, and venue-positioning tasks. For ARIS-style, following all three shared principles, we replaced open retrieval with benchmark-KB hybrid retrieval and family-specific evidence construction, and modified the final renderer into a contract-aware output module so that answers respect the candidate set, task family, and venue-facing framing required by ForeSci.

Family	Human ranking	Automatic ranking	Kendall's τ
Bottleneck	ARIS ≻ CoI ≻ RA ≻ RAG ≻ Native	ARIS ≻ RA ≻ CoI ≻ RAG ≻ Native	0.80
Direction	RAG ≻ ARIS ≻ CoI ≻ RA ≻ Native	ARIS ≻ RAG ≻ CoI ≻ RA ≻ Native	0.80
Strategic Planning	RA ≻ ARIS ≻ CoI ≻ RAG ≻ Native	RA ≻ ARIS ≻ CoI ≻ RAG ≻ Native	1.00
Venue	ARIS ≻ RA ≻ RAG ≻ CoI ≻ Native	ARIS ≻ RA ≻ CoI ≻ RAG ≻ Native	0.80

Table A5: Method-ranking comparison between human experts and automatic Reviewer Persuasiveness. RA denotes ResearchAgent-style and RAG denotes Hybrid RAG.

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%

Table A6: Human validation of automatic decision-critical claim extraction.

E Supplementary Metric Results

E.1 Evaluation results for each task family

Table A7 reports the evaluation results separately for each task family. The results show that method rankings vary across task families, suggesting that different foresight tasks benefit from different agent designs and that no single agent consistently dominates across all task types.

E.2 Scalar Metric Stability

We also test the run-to-run stability of the scalar metric stack on the same Qwen3-235B 100-task subset used for the preference study. The subset contains 500 evaluated rows after expanding 100 tasks by five methods. We keep the candidate answers fixed and repeat the metric evaluation five times: one existing formal-evaluation run plus four independent DeepSeek-V4 replicate runs. For each metric, Table A8 reports the largest standard deviation and largest range of method-level run means across all family–method cells. For Evidence Traceability, Native LLM rows are excluded because the metric is not applicable without an external support artifact. It also reports a row-level diagnostic: the mean per-row standard deviation.

Future-Target Alignment is the most stable metric: Planning and Venue FTA are deterministic ranking-aware scores, while Bottleneck and Direction use reference-guided embedding similarity over extracted prediction claims. Prediction Factuality and Reviewer Persuasiveness show moderate variance; they support family-level comparisons but close within-family method rankings should be interpreted as small-margin differences. Evidence Traceability has the largest variance, but its variance remains moderate and within an acceptable range, because it asks a rubric-style evaluator to assess evidence linkage and support specificity from method artifacts.

E.3 Backbone Style Profile

We observe that different backbone models show different answer styles. Table A9 reports word-like answer lengths for aligned formal-release outputs. All four generator backbones are aligned on the same 2,500 $(\text{task_id}, \text{method})$ keys, and each family row contains 625 answers. We further test whether the generator backbone changes the linguistic rendering of the same benchmark tasks. This audit uses a fully crossed matched sample of 80 tasks, with 20 tasks per family and all four backbones and five methods, yielding 1,600 blind answer-level annotations and 8,000 style labels. The judge sees only the task question and candidate answer; it does not see the reference answer, hidden target, method name, backbone name, or scalar metrics. We retain four paper-facing dimensions: decision directness, mechanistic concreteness, structural scaffold intensity, and verbosity/-compression. Figure A5 shows that style variation is mainly backbone-driven. Fixing the task and method while varying the backbone gives larger label disagreement than fixing the task and backbone while varying the method for structural scaffold intensity (0.364 vs. 0.250), verbosity/compression (0.292 vs. 0.168), decision directness (0.086 vs. 0.031), and mechanistic concreteness (0.106 vs. 0.082). The objective surface features support the same interpretation: GPT-5.2 answers are much longer and more visibly scaffolded, averaging 529 words, 8.6 bullet lines, and 4.3 heading lines per answer; GLM-4.6 is the most compressed, averaging 191 words and almost no expansive answers; Gemini-3 averages 273 words with 3.2 bullet lines, and Qwen3-235B averages 284 words with 2.2 bullet lines. We therefore interpret backbone sensitivity partly as a rendering effect: different generators package similar research decisions with different amounts of structure, directness, and mechanis-

		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	–	0.764	0.799	**0.614**	–	0.847	0.651	0.595	–	**0.631**	0.672	0.576	–	**0.742**
	Hybrid RAG	0.824	**0.609**	0.439	0.758	0.804	0.613	0.453	0.834	0.623	0.596	0.424	0.590	0.704	0.589	0.342	**0.742**
	CoI	0.851	0.607	0.560	0.748	0.819	0.610	**0.702**	**0.862**	0.683	0.596	0.485	0.612	0.700	0.583	0.401	0.727
	ResearchAgent	0.802	0.608	**0.571**	0.750	**0.829**	0.610	0.670	0.859	0.641	**0.597**	0.476	0.590	0.716	**0.591**	0.415	0.716
	ARIS	**0.860**	0.607	0.554	**0.777**	0.813	0.607	0.693	0.856	**0.695**	0.596	**0.486**	0.605	**0.735**	0.576	**0.476**	0.735
Direction Forecasting	Native LLM	**0.598**	0.635	–	0.786	**0.670**	0.635	–	0.839	0.523	**0.604**	–	0.674	**0.618**	0.601	–	0.733
	Hybrid RAG	0.567	**0.637**	0.584	0.776	0.645	0.634	0.576	0.828	0.515	**0.604**	0.533	0.660	0.589	**0.606**	0.590	0.732
	CoI	0.577	0.635	**0.745**	0.811	0.658	**0.637**	**0.748**	**0.866**	**0.528**	0.603	0.612	**0.683**	0.601	0.602	0.599	**0.737**
	ResearchAgent	0.568	0.633	0.722	0.803	0.662	0.633	0.741	0.858	0.512	0.596	0.577	0.674	0.595	0.604	0.556	0.733
	ARIS	0.542	0.627	0.743	**0.818**	0.589	0.628	0.745	**0.866**	0.506	0.601	**0.616**	0.657	0.573	0.600	**0.649**	0.715
Strategic Planning	Native LLM	0.465	0.580	–	0.756	0.441	0.548	–	0.818	0.359	0.472	–	0.664	0.421	0.554	–	0.744
	Hybrid RAG	0.446	0.554	0.332	0.737	0.446	0.553	0.364	0.809	0.420	0.532	0.372	0.670	0.422	0.521	0.354	0.723
	CoI	0.504	0.644	0.429	0.759	0.476	0.582	0.556	0.822	0.470	**0.638**	0.445	**0.689**	0.414	0.528	0.414	0.725
	ResearchAgent	**0.532**	**0.687**	0.460	**0.778**	0.481	0.577	0.560	0.827	**0.471**	0.637	0.471	0.682	**0.424**	0.544	0.405	0.730
	ARIS	0.460	0.598	**0.542**	0.760	**0.488**	**0.587**	**0.683**	**0.834**	0.450	0.577	**0.512**	0.677	0.422	**0.555**	**0.554**	**0.747**
Venue Positioning	Native LLM	0.508	0.664	–	**0.838**	0.562	0.715	–	0.882	0.503	0.665	–	**0.729**	0.508	0.678	–	0.744
	Hybrid RAG	0.552	0.722	0.375	0.831	0.545	0.702	0.241	0.878	0.522	**0.694**	0.387	0.712	0.505	0.651	0.368	0.684
	CoI	0.515	0.684	0.506	0.810	0.550	0.698	0.364	0.879	0.490	0.621	0.455	0.665	**0.519**	**0.688**	0.476	**0.746**
	ResearchAgent	0.534	0.711	0.498	0.817	0.567	0.712	0.366	0.883	**0.534**	0.676	**0.472**	0.676	0.504	0.672	0.460	0.736
	ARIS	**0.567**	**0.745**	**0.593**	0.817	**0.579**	**0.745**	**0.387**	**0.889**	0.499	0.684	0.466	0.658	0.487	0.655	**0.591**	0.737

Table A7: Main family-level results across four answer-generation backbones. Each family block reports five methods; bold marks the best method within the same family, backbone, and metric.

Metric	Max SD	Max range	Row SD
Prediction Factuality	0.021	0.053	0.045
Future-Target Alignment	0.006	0.014	0.004
Evidence Traceability	0.050	0.111	0.108
Reviewer Persuasiveness	0.021	0.053	0.034

Table A8: Five-run stability of scalar metrics 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.

tic detail. These results also motivate the use of Fact, FTA, and Trace as complementary metrics that directly assess agreement with future facts and traceable pre-cutoff evidence, since a purely LLM-as-a-judge metric such as Reviewer Persuasiveness may be partially confounded by backbone-specific rendering style.

F Diagnostic Analyses

F.1 Detailed Error Analysis

We conduct an internal error analysis over all 10,000 formal-release evaluations, using the same answer, metric, and hidden-goldset files as the main results. The analysis is diagnostic rather than a new evaluation: it does not call an LLM and does not rewrite metrics. It reads evaluator rationales, gold-claim coverage summaries, future-unit judgments, and support-profile metadata from the synchronized evaluation files. For each metric, we mark the rank-selected bottom 20% of rows as low scoring; this avoids mixing different score scales while keeping the diagnostic sample size fixed. Traceability is computed only over evidence-grounded methods because Native LLM outputs do not expose external support artifacts. These labels are used only to identify cases for inspection.

Table A10 shows where the failures concentrate. Strategic Planning is the hardest family: low Prediction Factuality is 0.315 and low Future-Target Alignment is 0.512, largely because a wrong top priority can invalidate an otherwise plausible plan. Venue Positioning has substantial low factuality and Future-Target Alignment failure rates, while its evidence-grounded low-trace rate is lower after the venue-specific Trace normalization (0.195). Bottleneck and Direction tasks have near-zero low-FTA rates under the repaired reference-guided semantic FTA, so their remaining errors are mainly claim-level or decision-object mismatches rather than global future-target misses. Native LLM has no external support artifact, so Traceability is reported as not applicable rather than as a zero score; low-trace diagnostics below refer to evidence-grounded methods only.

The Fact–FTA conflict audit further separates semantic proximity from exact decision correctness. The refreshed audit finds 131 large disagreements: 112 rows have an absolute Fact–FTA gap of at least 0.5, and 19 rows are 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. We therefore treat large disagreements as analysis signals rather than metric failures.

Generator	Family	Mean	Median	P90	Max
GLM-4.6	Bottleneck–Opportunity	124.0	120	159.0	228
GLM-4.6	Direction Forecasting	97.3	96	120.0	209
GLM-4.6	Strategic Planning	154.4	154	189.0	257
GLM-4.6	Venue Positioning	400.4	397	500.6	656
GPT-5.2	Bottleneck–Opportunity	336.3	324	427.2	682
GPT-5.2	Direction Forecasting	232.6	228	301.0	394
GPT-5.2	Strategic Planning	257.2	255	291.6	360
GPT-5.2	Venue Positioning	1276.0	1268	1448.2	1704
Gemini-3	Bottleneck–Opportunity	196.3	193	256.0	320
Gemini-3	Direction Forecasting	154.5	155	180.0	214
Gemini-3	Strategic Planning	203.8	204	226.6	542
Gemini-3	Venue Positioning	526.8	517	631.8	744
Qwen3-235B	Bottleneck–Opportunity	196.3	190	259.6	381
Qwen3-235B	Direction Forecasting	171.8	172	216.0	284
Qwen3-235B	Strategic Planning	172.2	173	195.0	226
Qwen3-235B	Venue Positioning	598.1	563	769.0	1045

Table A9: Answer-length profiles on aligned outputs. Lengths are word-like token counts. Each row contains 625 answers from one generator backbone and one task family.

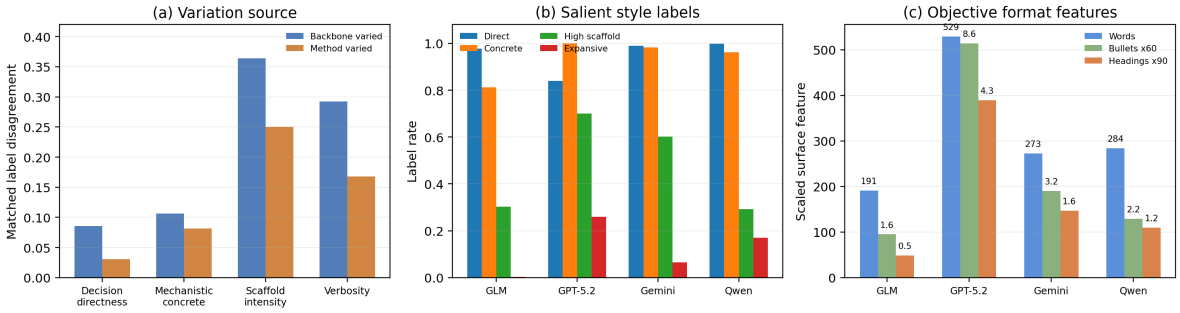

Figure A5: Supplementary backbone language-style profile over 1,600 blind answer annotations. The matched sample contains 80 tasks, all four answer-generation backbones, and all five methods. Panel (a) compares matched backbone variation against matched method variation. Panel (b) reports salient label rates by backbone: direct recommendation, concrete mechanism, high structural scaffold, and expansive verbosity. Panel (c) reports objective surface features; bullet and heading counts are scaled for visualization, with original mean counts printed above bars.

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 show a different failure: among evidence-grounded methods, most low-trace venue cases are topical-not-venue-specific evidence chains. Hybrid RAG has the highest such low-trace rate (0.312), while agent methods reduce but do not remove the problem (0.082–0.192). This distinction is why the main text treats Planning as a rank-order sensitivity problem and Venue as an evidence-specificity problem.

F.2 Bias–Metric Coupling Analysis

The preceding audits identify answer-reference drift, but do not by themselves show which drifts are responsible for formal metric failures. We therefore join the same 1,600 bias-annotated answers with the repaired formal metric files used in the main results. This merge has no missing evaluations. We treat severity 0 as aligned and severity at least 2 as high bias. The analysis is diagnostic: it does not relabel predictions, change metric weights, or call an additional LLM. Table A12 gives the numeric contrasts underlying Figure 3. Temporal-horizon annotations are retained in the released analysis files and included in the normalized effect-size heatmaps.

The case inventory supports the same interpretation. In RFDIR-0104, a GPT-5.2 ResearchAgent answer has high Traceability (0.820), high Reviewer Persuasiveness (0.820), 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. In RFPLAN-0136, a Qwen3-235B CoI answer is

Grouping	Unit	Low Fact	Low FTA	Low Pers.	Low Trace
Family	Bottleneck–Opportunity	0.070	0.013	0.281	0.276
	Direction Forecasting	0.184	0.004	0.189	0.043
	Strategic Planning	0.315	0.512	0.203	0.287
	Venue Positioning	0.232	0.271	0.127	0.195
Method	Native LLM	0.216	0.218	0.172	–
	Hybrid RAG	0.207	0.211	0.208	0.323
	CoI-style	0.199	0.194	0.203	0.189
	ResearchAgent-style	0.183	0.178	0.203	0.173
	ARIS-style	0.195	0.199	0.214	0.115

Table A10: Low-score rates used for the diagnostic error analysis. Low-score rows are the rank-selected bottom 20% for each metric. Evidence Traceability is computed over evidence-grounded methods only, so Native LLM traceability is not applicable.

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.	This method has the lowest overall low-Fact and low-FTA rates, and is strongest on Planning, but a wrong early abstraction can be amplified 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 A11: Method-internal fingerprints from the diagnostic error audit. The table summarizes why low-scoring cases fail within each method.

traceable (0.700) and judged plausible (0.720), but FTA is only 0.145 because it deprioritizes embodied evaluation due to immature tooling, while the reference treats that immaturity as the reason evaluation infrastructure should be the top priority. These examples show why high-trace, low-FTA cases are better interpreted as evidence-to-decision failures than as simple evidence absence.

F.3 Content-Metric Correlation Diagnostic

We additionally test whether the three content-facing formal metrics collapse into a single latent score. This diagnostic uses the same 10,000 repaired evaluations as the formal error analysis and computes pairwise Pearson and Spearman correlations among Prediction Factuality, Future-Target Alignment, and Reviewer Persuasiveness. Evidence Traceability is excluded from this correlation analysis: it measures support exposure and answer interface rather than content alignment, and is not applicable to Native LLM outputs without external support artifacts.

Overall, the three content metrics are positively related but not redundant. Spearman correlation is 0.491 for Fact vs. FTA, 0.231 for Fact vs. Pers., and 0.336 for FTA vs. Pers. The main structure is family-conditioned. Rank-style families show much stronger Fact-vs.-FTA coupling: Strategic Planning has Spearman 0.816 and Venue Positioning has 0.743, compared with 0.292 for Bottleneck–Opportunity and 0.368 for Direction Forecasting. This is expected because in rank-style tasks, factual claim correctness and target alignment both depend on choosing the right priority or venue-conditioned decision. Venue also illustrates non-redundancy: its Fact-vs.-FTA correlation is high, but FTA vs. Pers. is only 0.274, so matching the venue target does not automatically imply a persuasive research decision.

Backbone and method effects are present but smaller than family effects. Across backbones, Fact vs. FTA Spearman ranges from 0.368 for Qwen3-235B to 0.577 for GLM-4.6, while FTA vs. Pers. stays between 0.280 and 0.421. Across methods, Fact vs. FTA ranges from 0.459 for CoI-style to 0.534 for Native LLM, and FTA vs. Pers. ranges from 0.323 for ResearchAgent-style to 0.357 for Native LLM. This supports the paper's use of multiple metrics: they move in related directions, but their coupling depends on task form and failure mode rather than on a universal scalar quality dimension.

Question	Diagnostic contrast	Main result	Interpretation
Causal-role bias vs. Fact	Aligned vs. high causal-role bias	Fact drops by 1.13 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.
Claim-scope bias vs. FTA	Aligned vs. high claim-scope bias	FTA drops by 1.22 SD overall; Planning FTA drops by 1.86 SD	Scope mismatch is most harmful when the task evaluates a global ranking or priority structure, not just topical semantic proximity.
Intervention-mode shift vs. FTA	Mode-aligned vs. mode-shifted answers	FTA drops by 1.12 SD overall; Planning FTA drops by 1.99 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	213 high-trace/non-low-FTA rows vs. 30 high-trace/low-FTA rows; causal severity 1.268 $\rightarrow$ 2.633; intervention-mode severity 1.592 $\rightarrow$ 2.833	High traceability does not guarantee target alignment: evidence can support a coherent but misaligned decision object.

Table A12: 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 bias-annotation sample joined with the repaired formal metrics.

High Traceability But High Drift: A Venue-Positioning Case

Case	Gemini-3 + ARIS, RFVENUE-0062, venue-conditioned positioning
Task	Recommend the 2026 venue family for a reinforcement-learning-from-AI-feedback paper in language-model post-training, using only the 2025-12-31 research profile.
Agent answer	Ranks **NeurIPS** first, treating the work as a general methodological contribution in reinforcement learning and model alignment; ACL/EMNLP is listed as a secondary target.
Reference target	Ranks **ACL/EMNLP** first. The reference frames the paper as a language-model post-training and alignment contribution that should be evaluated against RLHF, SFT, DPO-style preference optimization, and human or expert feedback validation.
Scores	Trace = **0.920**, Fact = **0.200**, FTA = **0.355**, Pers. = 0.720.
Drift diagnosis	The answer is evidence-supported and venue-plausible, but it reflects a framing drift: it prioritizes a broad NeurIPS-style methodological interpretation, whereas the reference target prioritizes the NLP post-training and alignment venue context.

Figure A6: Concrete high-traceability/high-drift example. The task, agent answer, and reference target are summarized for readability; the scalar scores are from the repaired formal metrics, and the drift diagnosis comes from the matched answer-reference drift audit.

G Prospective Forecast Package

ForeSci is primarily an evaluation benchmark with hidden future supervision. We also use its construction machinery in a more prospective setting: forecasting events that had not yet occurred at writing time. This appendix gives a compact example rather than a scored benchmark result. The exercise has no post-hoc ground truth and no judge/evaluation scores; it is reported only to illustrate how taxonomy evolution can seed near-future research questions and how methods answer them under the same cutoff.

The exercise focuses on the LLM-agent domain with a literature cutoff of 2026-05-15. The released prospective package contains 12 public questions, balanced across the four task families. The forecast window for bottleneck, direction, and planning tasks is 2026-05-16 to 2026-08-15; venue questions are advisory venue-positioning decisions rather than claims about eventual acceptance. The strict corpus used to construct this package contains 12,124 LLM-agent rows, after adding 1,380 deduplicated core/support-screened papers from 2026-04-01 to 2026-05-15 to the previous 2026M03 corpus. The same run induces 278 taxonomy nodes in 2026M05A and a method-evolution asset with 22 method nodes, 20 specialization edges, and 440 paper-method mention rows.

Slice	New	Cum.	Nodes	Width	Depth
2026M03	1,019	10,744	243	1	1
2026M04	685	11,429	261	7	7
2026M05A	695	12,124	278	5	4

Table A13: Recent LLM-agent taxonomy evolution used to seed the prospective forecast package. Counts are from the strict core/support-screened taxonomy corpus; 2026M05A covers papers through 2026-05-15.

We generated answers with Qwen3-235B and GPT-5.2 for all five method configurations. Both

Figure A7: Family-level decomposition of bias–metric coupling. 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.

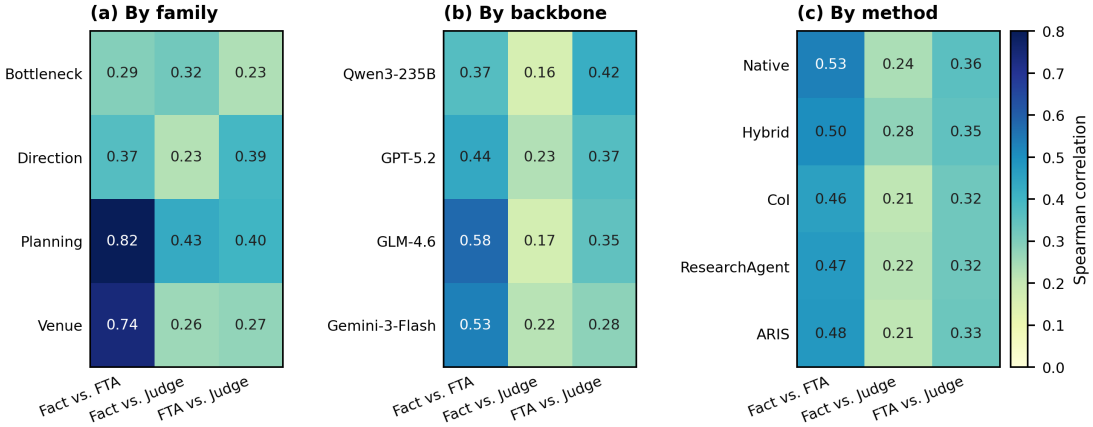

Figure A8: Supplementary content-metric correlation diagnostic over 10,000 repaired formal rows. Values are Spearman rank correlations among Prediction Factuality, Future-Target Alignment, and Reviewer Persuasiveness. Evidence Traceability is excluded because it measures support exposure and answer interface rather than content alignment. Family structure dominates the correlation pattern: rank-style planning and venue tasks have much stronger Fact-vs.-FTA coupling than bottleneck and direction tasks, while backbone and method differences are more moderate.

runs completed all 60 requested final answers, covering 12 tasks for each of Native LLM, Hybrid RAG, CoI-style, ResearchAgent-style, and ARIS-style. No scores are reported because this package concerns future outcomes with no available ground truth. The appendix below therefore treats the run as a transparent forecast artifact: it selects two easily interpretable examples, one from Direction Forecasting and one from Venue-Conditioned Positioning, and reproduces their public questions and unscored generated answers. A standalone supplement contains the broader prospective forecast artifact.

Selected unscored forecast outputs. We show selected generated answers for Direction Forecasting and Venue-Conditioned Positioning. Figures 4 and A9 compare Native LLM outputs with selected agent outputs in paired answer cards. For Venue-Conditioned Positioning, we compress the original longer answers while preserving each method's venue ranking, framing advice, major risks, and evidence-upgrade recommendations. The standalone supplement contains the verbatim prediction-only outputs.

Direction forecasting example. RFDIR-0149 – Direction Forecasting.

Metadata. Cutoff: 2026-05-15. Forecast window: 2026-05-16 to 2026-08-15.

Question. Looking forward from the 2026-05-15 snapshot, which agent-memory direction is most likely to gain momentum over the next three months: better retrieval, auditable memory security, larger context stores, or persona personalization? Choose one direction, assign one of the trajectory labels accelerating, fragmenting, steady, or cooling, and justify the choice.

Expected answer format. State one concrete forecasted research direction; state exactly one trajectory label; give a brief rationale using only evidence and field conditions available by the cutoff.

Support requirements. Answer the specific domain and time window, use only pre-cutoff reasoning, keep the answer to one prioritized direction, and connect the rationale to relevant mechanisms, systems, evaluations, bottlenecks, or adoption conditions.

Venue-conditioned positioning example. Figure A9 shows compressed outputs for a prediction-only venue task, preserving the venue ranking and the main positioning rationale while omitting low-level generation details.

H Discussion: Beyond Information Retrieval

Output rendering matters. ForeSci exposes a practical tension: internal reasoning and the final benchmark-facing answer are not the same object. Some agents retrieve or organize useful evidence but fail to surface it in a form that satisfies the task contract. We therefore use a common final-answer renderer for agent methods, making comparisons less about verbosity and more about whether each method's evidence can be converted into a defensible research judgement. This separates rendering failures, upstream retrieval failures, and cases where evidence is present but the proposed research move is not compelling.

Generalization versus specialization. No method is uniformly best across foresight decisions. Retrieval often improves target proximity for direction forecasting, but the repaired FTA surface also shows that semantic closeness to a future target is not the same as selecting the right research decision. Evidence organization helps traceability, and idea-construction scaffolds can improve rubric-based Reviewer Persuasiveness. But bottleneck discovery, forecasting, strategic planning, and venue positioning impose different constraints: agents need task-aware mechanisms for deciding what kind of research move they are making and what evidence would make that move credible.

QUESTION

*A paper proposes **auditable memory protection** for LLM agents, focusing on lineage, access control, and attacks on persistent or shared memories. Shortly after the 2026-05-15 literature snapshot, the authors must choose among EMNLP 2026 Main, AACL-IJCNLP 2026 Main, COLM 2026 Workshop Contributions, EMNLP 2026 Industry Track, and NeurIPS 2026 Workshops. How would you rank these options, what default target would you recommend, and what framing, risks, and evidence upgrades should the authors consider before submission?*

Native LLM

Ranking. EMNLP Main > NeurIPS Workshops > EMNLP Industry > AACL Main > COLM Workshop.

Why EMNLP. The answer treats the proposal as an NLP systems, security, and agent-architecture paper: lineage, access control, and persistent/shared-memory attacks match EMNLP's reliable-agent and accountable-system audience better than NeurIPS' broader orientation, AACL's narrower scope, COLM workshops' early-stage fit, or the Industry Track's deployment emphasis.

Positioning. Frame the contribution as auditable memory for accountable agent architectures: lineage and access control should support reproducibility, debugging, and safety in real deployments. An alternate NeurIPS-workshop framing would emphasize long-term memory risks and memory integrity in learned systems.

Risks. EMNLP fit weakens if experiments are synthetic or small-scale, if novelty over existing agent-memory frameworks is unclear, or if the work reads as systems engineering without broader NLP-agent insight.

Evidence upgrades. Add a concrete unauthorized-leakage or shared-memory attack case study, benchmark against an existing agent-memory implementation, report latency/storage overhead, and include a threat-model diagram or access-control-policy example.

ResearchAgent

Ranking. EMNLP Main > AACL Main > EMNLP Industry > NeurIPS Workshops > COLM Workshop.

Why EMNLP. The answer keeps EMNLP Main as the default because the proposal's center of gravity is auditable memory protection for LLM agents: lineage/provenance, access control, and persistent/shared-memory attacks. It argues that the snapshot has moved from output poisoning to memory-driven control-flow and tool-trace risks. EMNLP is the strongest archival venue for a rigorous NLP-agent method; AACL is the close fallback, while workshops fit prototype evidence.

Positioning. Frame the paper as agent safety and reliability through auditable memory governance. The motivation is that persistent memory is not passive storage but a durable control signal. The problem statement is that current memory pipelines often treat retrieved memory as trustworthy and lack provenance checks. The core claims should cover a lineage/provenance model for memory items, policy enforcement that gates reads and writes by source and trust, access control for shared memory, audit logs for violation detection, and attack-centered evaluation rather than policy compliance alone.

Alternate framing. If the work looks too security-systems oriented, the AACL version should foreground language/agent behavioral outcomes: reduced cross-session misgeneralization, sleeper effects, and tool-trace/control-flow deviations. With incomplete evaluation, COLM Workshop fits design plus initial results.

Risks. The answer stresses novelty overlap with nearby governance, lineage, and dynamic access-control systems; undefined auditability; weak or unrealistic threat models; shared-memory claims without shared-memory experiments; and a security-only narrative without agent-level utility evidence.

Evidence upgrades. Evaluate persistent poisoned-retrieval dynamics, sleeper memory across sessions, and memory-driven tool-trace steering. Define logged artifacts and an auditor protocol; compare unmanaged-memory and governance-only baselines; ablate lineage, access control, and audit logging. Report leakage, violations, task success, overhead, attack persistence, tool-trace divergence, and audit false positives/negatives. The recommendation remains EEMNLP Main only if this attack-aligned evidence is delivered.

Figure A9: Side-by-side compressed Venue-Conditioned Positioning outputs for the prediction-only LLM-agent case. The figure includes the task question and compressed Native LLM and ResearchAgent outputs.