# CAN FRONTIER AUTO RESEARCHERS RECONSTRUCT HUMAN RESEARCH DECISIONS IN AI?

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Scientific research has long been driven by substantial human effort, from reviewing prior work and identifying research gaps to designing specific solutions. Recent advances in harness-based agent systems have given rise to increasingly capable *Auto Researchers*, which can autonomously carry out these tasks and, increasingly, entire end-to-end research workflows. Yet despite their rapidly growing capabilities, how closely the research decisions and outputs produced by auto researchers align with those of human experts remains underexplored. We investigate this gap through the lens of research reconstruction and study: *Given comparable research context, how similar are the research decisions and outputs produced by Auto Researchers to those of human researchers in AI research?* To answer this question, we introduce **Auto**Researcher–**H**uman **R**esearch **S**imilarity (**AutoHRS**), a stage-wise evaluation framework for assessing how faithfully auto researchers reconstruct human research artifacts across different stages of the scientific workflow. Using 152 ICLR oral papers as our corpus, we generate 456 research trajectories spanning different stages of the research process. Applying AutoHRS to these trajectories, we find that current *Auto Researchers* frequently fail to recover key research decisions and their resulting artifacts across multiple stages of the research process.

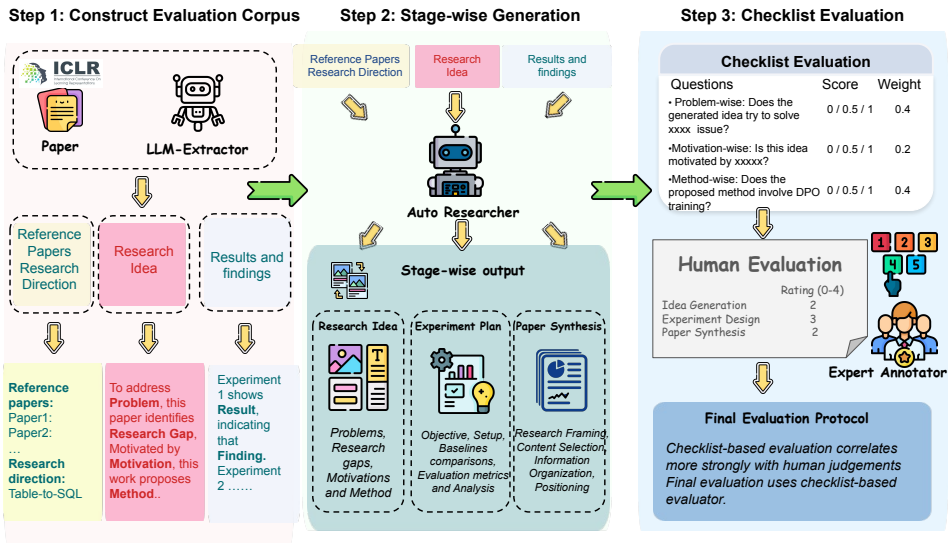



Figure 1: Overview of our work. We construct AutoHRS by decomposing human research into three stage-wise tasks: *idea generation*, *experiment design*, and *paper synthesis*. Generated outputs are evaluated against human research using a checklist-based evaluator validated by expert judgments.

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

# 1 INTRODUCTION

Large language models (LLMs) have been increasingly integrated into scientific research since their emergence, supporting tasks ranging from scientific discovery to scientific writing (Wang et al., 2024; Lu et al., 2024; Baek et al., 2025; Guo et al., 2025; Kumar et al., 2025; Li et al., 2025; Si et al., 2025; Sternlicht & Hope, 2026; Kon et al., 2026). Despite these advances, human researchers remained in control of the overall research process, while LLMs primarily served as assistants for individual tasks. With the development of harness-based agents, however, language models are increasingly being integrated into end-to-end scientific workflows (Yang et al., 2026; Wang et al., 2026; Tang et al., 2026). We refer to such harness-based systems that autonomously carry out multiple stages of the research process as *Auto Researchers*. In these emerging systems, the traditional division of labor is substantially altered: auto researchers take a leading role in carrying out the research process, while human researchers increasingly provide guidance, feedback, and oversight (Schmidgall et al., 2025; Ye et al., 2026; Zhu & Wang, 2026). This shift raises an important question of how the research produced by these systems should be evaluated. However, evaluating research independently generated by auto researchers is challenging: its novelty, scientific value, and potential for genuine breakthroughs are difficult to establish without extensive validation. A more tractable starting point is to use established human research as a reference for assessing how these systems formulate ideas, design experiments, and synthesize findings into papers. We therefore ask: *Can Frontier Auto Researchers Reconstruct Human Research Decisions in AI?*

Existing evaluation frameworks for LLM-based scientific research systems have largely focused on model-level capabilities, typically assessing performance on individual research tasks rather than the behavior of harness-based systems operating across the full research process (Chen et al., 2025; Wijk et al., 2025; Zhang et al., 2026; Xu et al., 2026; Fei et al., 2026; Cui et al., 2026). In this work, rather than evaluating auto researchers solely by whether they can complete predefined research tasks, we ask whether they can conduct research in ways that reconstruct the decisions and processes of human researchers. To operationalize this evaluation, we introduce **Auto**Researcher–**H**uman **R**esearch **S**imilarity (**AutoHRS**), a stage-wise evaluation framework for assessing how faithfully auto researchers reconstruct human research artifacts[1] across different stages of the scientific process. As scientific research is an open-ended, nonlinear, and multi-stage process, we decompose it into three tractable stages that can be systematically evaluated: idea generation grounded in reference papers and a research direction, experiment design conditioned on a given research idea, and paper synthesis grounded in provided research context and experimental findings (Langley et al., 1987; Klahr & Dunbar, 1988).

To instantiate this evaluation framework, we construct an evaluation corpus from ICLR 2026 oral papers. Each paper is decomposed into three research trajectories corresponding to three stages considered in AutoHRS: from reference papers to a research idea, from a research idea to an experimental plan, and from experimental findings to the final paper. These trajectories provide both the inputs to the auto researchers and corresponding human research outputs used as references for evaluation. Our framework employs a fine-grained checklist-based evaluation pipeline, with stage-specific dimensions and weights tailored to stages of the research process. We further validate the framework against human judgments, showing strong alignment with human judgments.

Our evaluation shows that current auto researchers exhibit limited ability to reconstruct human research across all three stages. Under AutoHRS, on a normalized scale from 0 to 1, auto researchers achieve a score of 0.29 for the best-scoring idea per paper in idea generation, 0.36 on experiment design, and 0.65 on paper synthesis. To better understand these results, we analyze where and how auto researchers diverge from human research, identifying recurring patterns in idea composition and experiment planning. We further analyze factors that affect reconstruction performance, including the granularity of the input research direction and the research type. These results suggest that, despite their growing ability to automate end-to-end scientific workflows, current auto researchers still differ substantially from human researchers in how they formulate ideas, design experiments, and synthesize research into scientific papers.

[1] Reconstructing research artifacts does not mean reproducing the original study; we use reconstruction as a proxy to observe whether auto researchers can mimic human research decisions and outputs.

# 2 AUTOHRS EVALUATION FRAMEWORK

We propose an evaluation framework for auto researchers that assesses whether they can reconstruct the research decisions made by human researchers. In this framework, we define three stage-wise evaluation tasks corresponding to three key stages of the research process: literature-grounded idea generation, idea-grounded experiment design, and findings-grounded paper synthesis.

## 2.1 STAGE-WISE EVALUATION TASKS

**Literature-Grounded Idea Generation** In this task, each instance consists of a set of reference papers, $\mathcal{R} = \{R_1, R_2, \ldots, R_n\}$, and a general research direction $D$. Each reference paper $R_i$ is represented by five components, $R_i = \{t_i, a_i, p_i, m_i, o_i\}$, corresponding to its *title*, *abstract*, *problem*, *method*, and *results*, respectively. Given the reference set $\mathcal{R}$ and research direction $D$, the target output is a structured research idea, represented as $I = \{p, g, r, m\}$, where $p$ denotes the research problem, $g$ the research gap, $r$ the research motivation, and $m$ the proposed method.

**Idea-Grounded Experiment Design** In this stage, the auto researcher is given the structured research idea $I = \{p, g, r, m\}$ defined above. Given the input $I$, the auto researcher is tasked with producing a detailed experiment plan, denoted as $P$, that designs experiments around the proposed method to evaluate whether it addresses the identified research problem and gap. The plan is expected to specify the main experimental objectives, evaluation settings, datasets or benchmarks, comparison methods, evaluation criteria, and additional analyses needed to test the proposed idea.

**Findings-Grounded Paper Synthesis** The final stage of the research process is to synthesize experimental results and findings into a research paper. In this task, the auto researcher is provided with the research idea $I$, the experiment plan $P$, and the resulting experimental evidence and findings, denoted as $F$. Given these inputs, the auto researcher is tasked with synthesizing them into a coherent research paper, allowing us to assess whether it can organize, interpret, and communicate research findings in a manner comparable to human researchers.

The input for each stage is independently extracted from the corresponding target paper rather than generated from the output of the preceding stage. This design allows each stage to be evaluated independently without propagating errors from earlier stages. Details on how task inputs and corresponding human reference outputs are extracted and constructed are provided in Appendix A.2.

## 2.2 EVALUATION CORPUS

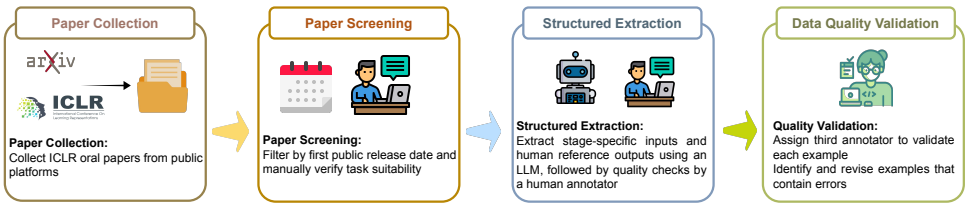



Figure 2: **Overview of evaluation corpus construction.**

We construct the evaluation corpus from ICLR 2026 oral papers, using each paper as the source for both stage-specific task inputs and the corresponding human reference outputs. To prevent auto researchers from directly retrieving information about the ground-truth papers, we further restrict the corpus according to the knowledge cutoff of the underlying models. We also manually screen the remaining papers and remove those that do not satisfy the requirements of our evaluation tasks, such as papers without publicly available reference papers and purely theoretical papers without any experiments. Following the stage-wise task definitions in Section 2.1, we then derive the corresponding evaluation instances from each paper, pairing the stage-specific input with the human research output used as the reference for evaluation. All stage-specific inputs and reference outputs are extracted

using an LLM-based pipeline and reformulated into structured representations aligned with the input and output formats of the respective tasks. Each resulting instance is reviewed and validated by human annotators to ensure the correctness and quality of the extracted content. After filtering and human validation, the resulting corpus contains 152 papers and 456 research trajectories.

### 2.3 CHECKLIST-BASED EVALUATION

At each research stage, AutoHRS takes an auto-researcher-generated artifact and the corresponding human reference output as input. It evaluates the generated artifact using a human-verified checklist derived from the reference and produces item-level judgments, dimension-level scores, and an overall score measuring reconstruction fidelity.

**Checklist Evaluation Pipeline** We design a checklist-based evaluation protocol that enables reliable and reproducible automated assessment. For each stage, we first extract checklist questions from the reference output using an LLM generator. Each checklist question follows a general form, "Does the output $\mathcal{O}$ contain or reflect the element $\mathcal{E}$?", where $\mathcal{O} \in \{\text{idea}, \text{experiment plan}, \text{paper}\}$ denotes the auto researcher output at the corresponding stage, and $\mathcal{E}$ denotes an element identified in the reference output. An LLM then serves as a judge to evaluate the outputs generated by the auto researcher, assigning each checklist item one of three labels: *Yes*, *Partial*, or *No*. During preliminary evaluation, we found that some checklist questions generated by the LLM were overly generic or insufficiently discriminative. To improve checklist quality, we introduce a human verification step in which annotators review the generated questions and remove or revise those that are ambiguous, overly broad, or uninformative for reconstruction evaluation. The resulting checklist is then used to assess how well the auto researcher reconstructs the corresponding human research output at each stage. Further details on checklist generation are provided in Appendix A.3.

**Stage-Specific Evaluation Dimensions** Each stage of scientific research involves multiple interconnected decisions and outputs, making it difficult to characterize reconstruction fidelity with a single holistic criterion. To enable a more fine-grained evaluation, we decompose each research stage into a set of stage-specific evaluation dimensions. Our dimension design is informed by established conventions in scientific reasoning and reporting, including the CARS framework for framing research problems and gaps, principles of rigorous and reproducible experimental design, and the IMRaD structure for scientific reporting (Swales, 1990; Sollaci & Pereira, 2004). Specifically, (1) **Idea Generation**: we evaluate the generated artifact along the dimensions of *Problem*, *Research Gap*, *Motivation*, and *Method*. (2) **Experiment Design**: we consider *Experimental Objective*, *Experimental Setup*, *Baselines and Comparisons*, *Evaluation Protocol*, and *Ablation and Analysis*. (3) **Paper Synthesis**: we evaluate *Research Framing*, *Content Selection and Emphasis*, *Scientific Argumentation*, *Information Organization*, and *Interpretation and Positioning*. Detailed definitions and the corresponding examples for each dimension are provided in Appendix A.4.

| Evaluator | Idea Generation | Experiment Design | Paper Synthesis |
|---|---|---|---|
| GPT-6 Astra | 0.60 | 0.57 | 0.64 |
| Claude Fable 5.1 | 0.62 | 0.59 | 0.60 |
| Qwen3.8-2.4T-A95B | 0.58 | 0.50 | 0.55 |
| **Checklist Metrics (Ours)** | **0.70** | **0.64** | **0.68** |

Table 1: Instance-level Spearman's correlation ($\rho$) with human judgments across the three stages. Higher values indicate stronger alignment with human annotators.

**Validation Against Human Judgments** To validate whether the checklist-based scores align with human assessment, we randomly sample 20 examples from the evaluation corpus and ask human annotators to independently assess how well each auto-researcher output reconstructs the corresponding human reference. We further compare our checklist-based metric with a single-LLM-as-a-judge baseline. As shown in Table 1, AutoHRS achieves instance-level Spearman's correlations of $\rho = 0.70$, $0.64$, and $0.68$ with human judgments for idea generation, experiment design, and paper synthesis, respectively, achieving higher observed correlations than the single-LLM judges on the annotated subset. These results indicate that the structured checklist provides a more reliable proxy for human evaluation. Annotator biographies, annotation details, and their responsibilities in the hu-

man evaluation process, together with details of the single-LLM-as-a-judge baseline, are provided in Appendix A.1, A.5, and A.6.

# 3 EXPERIMENTS

In our experiments, we leverage AutoHRS to evaluate auto researchers using our curated human research trajectories as references. We then analyze where and how auto researchers diverge from human research, characterizing recurring patterns across different stages of the research process. Finally, we investigate factors that can make auto-researcher outputs more similar to human research, including the granularity of the input research direction and variation across different research types.

## 3.1 SETUP

**Auto Researchers** We conduct our main experiments with Claude Code, using Claude Sonnet 4.5 as the underlying model. We select this model because its knowledge cutoff predates the target papers in our evaluation corpus, reducing the risk that the model can recover ground-truth research content from prior knowledge rather than from the provided inputs. To assess robustness across agent harnesses, we further replicate a subset of the experiments with Gemini CLI and Codex; setup details and additional results are provided in Appendix B.1 and B.2.

**Data** We conduct our experiments on the evaluation corpus introduced in Section 2.2, constructed from ICLR 2026 oral papers. The corpus is filtered according to the knowledge cutoff of the underlying models and manually screened to ensure that each paper supports all three evaluation tasks. The final corpus contains 152 papers, yielding 456 research trajectories across *idea generation*, *experiment design*, and *paper synthesis*. Each trajectory consists of the stage-specific input provided to the auto researcher and the corresponding human reference output used for evaluation.

**Metrics** We use a unified metric formulation to evaluate auto researchers across the three research stages considered in AutoHRS. At each stage, the generated research artifact is evaluated along multiple dimensions, as described in Section 2.3. For each evaluation dimension $d$, the checklist-based evaluator assigns a score $s_{d,i} \in \{0, 0.5, 1\}$ to the $i$-th checklist item, corresponding to *No*, *Partial*, and *Yes*, respectively. Since the number of checklist items may vary across dimensions, we first compute a dimension-level score by averaging over all checklist items within that dimension:

$$S_d = \frac{1}{N_d} \sum_{i=1}^{N_d} s_{d,i},$$

where $N_d$ denotes the number of checklist items associated with dimension $d$. To aggregate dimension-level scores into an overall AutoHRS score, we assign stage-specific weights to the evaluation dimensions. The weights are selected based on alignment with human judgments, using the weighting scheme that yields the highest correlation with human evaluation for each research stage. Further details on the weight selection procedure and the resulting stage-specific weights are provided in Appendix B.3. For a given stage $t$, the overall AutoHRS score is computed as a weighted average of the corresponding dimension-level scores:

$$S^{(t)}_{\text{AutoHRS}} = \sum_{d=1}^{D_t} w_d^{(t)} S_d^{(t)}, \qquad \sum_{d=1}^{D_t} w_d^{(t)} = 1,$$

where $w_d^{(t)}$ denotes the weight assigned to dimension $d$ at stage $t$, and $D_t$ denotes the number of evaluation dimensions defined for that stage. Finally, we report $S^{(t)}_{\text{AutoHRS}}$ separately for each research stage. Higher AutoHRS scores indicate greater fidelity in reconstructing the corresponding human research output and stronger alignment with the human reference.

## 3.2 MAIN FINDINGS

**Idea Generation** Figure 3 shows that auto researchers achieve limited alignment with human references in literature-grounded idea generation. The mean overall AutoHRS score is 0.29 for the best-scoring idea per paper and 0.15 across all generated ideas. Regarding the proportions of check-

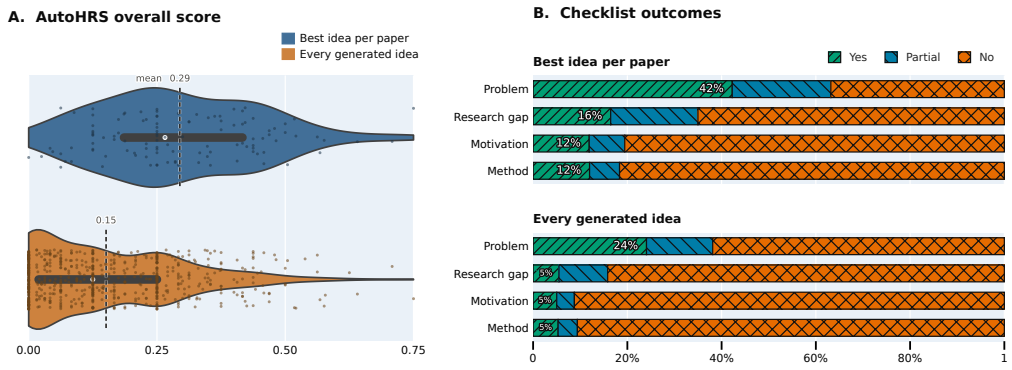



Figure 3: **Idea generation results.** (A) Distributions of overall AutoHRS scores for every generated idea and the best-scoring idea for each paper. "---" indicates the mean and "○" indicates the median. (B) Proportions of checklist outcomes (*Yes*, *Partial*, and *No*) within each evaluation dimension, shown separately for the best idea per paper and every generated idea.

list outcomes, the results show that, for the best idea per paper, 42% of problem-related checklist items receive a *Yes*, whereas the corresponding proportions are 16% for research gap identification and 12% for both motivation and method. Across all generated ideas, 24% of problem-related items receive a *Yes*, while this proportion is only 5% for each of the remaining dimensions. In both groups, *No* outcomes dominate the research gap, motivation, and method dimensions. These results indicate that auto researchers are better at recovering the broad research problem than reconstructing the more specific research decisions underlying the human reference.

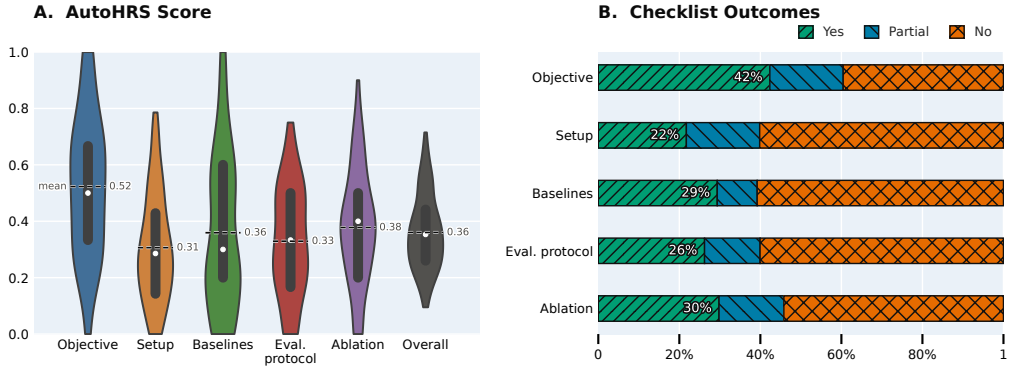



Figure 4: **Experiment design results.** (A) Distributions of AutoHRS scores for each evaluation dimension and the overall score. (B) Proportions of checklist outcomes (*Yes*, *Partial*, and *No*) within each evaluation dimension.

**Experiment Design** Figure 4 shows that auto researchers achieve a mean overall AutoHRS score of 0.36 in idea-grounded experiment design. Alignment is strongest in *Experimental Objective*, with a mean score of 0.52 and 42% of checklist items receiving a *Yes*. The remaining dimensions exhibit lower alignment: mean scores are 0.31 for *Experimental Setup*, 0.36 for *Baselines and Comparisons*, 0.33 for *Evaluation Protocol*, and 0.38 for *Ablation and Analysis*, with corresponding *Yes* proportions of 22%, 29%, 26%, and 30%, respectively. *No* outcomes account for more than half of the checklist items in each of these four dimensions. Although the overall AutoHRS score is higher than in idea generation, alignment with human references remains poor in concrete experimental choices, including experimental setup, baseline selection, evaluation protocols, and ablation studies.

**Paper Synthesis** Figure 5 shows that auto researchers achieve a mean overall AutoHRS score of 0.65 in findings-grounded paper synthesis. Alignment is strongest in *Research Framing* and *Scientific Argumentation*, with mean scores of 0.72 and 0.70, respectively, and 68% of checklist items receiving a *Yes* in both dimensions. Mean scores are 0.63 for *Content Selection and Emphasis* and

0.64 for *Interpretation and Positioning*, with a *Yes* proportion of 59% in each dimension. *Information Organization* exhibits the lowest alignment, with a mean score of 0.58 and 57% of checklist items receiving a *Yes*. These results suggest that auto researchers are more effective at reconstructing human research in paper writing than in earlier stages when provided with the research context and findings. However, auto researchers still differ from human experts in how they organize papers and emphasize key aspects of the research.

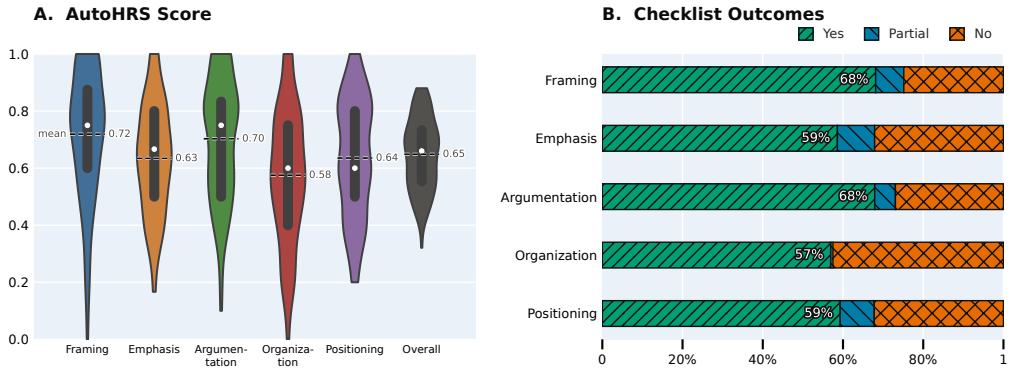



Figure 5: **Paper synthesis results.** (A) Distributions of AutoHRS scores for each evaluation dimension and the overall score. (B) Proportions of checklist outcomes (*Yes*, *Partial*, and *No*) within each evaluation dimension.

**Takeaway**

AutoHRS scores are 0.29 for the best-scoring idea generation, 0.36 for experiment design, and 0.65 for paper synthesis. Despite these differences across stages, substantial gaps from human research remain in all three, suggesting that current auto researchers still struggle to reproduce the research decisions and outputs of human researchers.

### 3.3 Where and how auto researchers diverge from human research

**Reference-Driven Idea Composition** During manual inspection of auto-researcher-generated research trajectories, we observed several recurring patterns that may contribute to their divergence from human-generated ideas. Auto researchers frequently construct new ideas by combining or integrating elements from multiple reference papers, whereas such patterns appear less frequently in the corresponding human ideas. To examine this observation, we randomly sample 20 research trajectories and manually assess whether each idea is primarily formed by combining identifiable elements from the provided reference papers. Under this criterion, 17 of the 20 auto-researcher-generated ideas exhibit a composition-based pattern, whereas the corresponding human ideas do not. We then extend this analysis to all research trajectories using *GPT-6 Astra* to assess the relationship between generated ideas and their reference papers. The results show that 78.4% of auto-researcher-generated ideas exhibit this pattern, suggesting a strong tendency to synthesize existing reference elements rather than depart from them in the same way as human researchers.

**Conventional Biases in Experiment Planning** Recurring patterns are also observed in auto-researcher-generated experiment plans. Rather than tailoring experimental designs closely to the specific research question, auto researchers often fall back on conventional experimental choices. They tend to select commonly used models and standard evaluation metrics, while introducing task-specific or customized metrics much less frequently. Together, these patterns suggest that auto researchers often rely on generic experimental templates, which may contribute to their divergence from human-designed experiment plans.

**Takeaway**

Auto researchers exhibit recurring patterns in both idea generation and experiment planning: 78.4% of generated ideas recombine elements from the provided references, while experiment plans often favor common models and standard metrics over more tailored designs.

### 3.4 How to Make Auto-Researcher Outputs More Similar to Human Research

We further investigate factors that may affect the reconstruction performance of auto researchers. We examine how the granularity of the input research direction influences idea generation, and whether reconstruction performance varies across different types of research papers, including *Empirical*, *Methodological*, and *Benchmark-oriented* work.

**Impact of Research Direction Granularity** In our main experiments, the input to the idea generation stage consists of reference papers together with a research direction. By default, the research direction is expressed as a broad phrase, such as "LLM Reasoning", "Multimodal Learning", or "Agent Safety". In preliminary experiments, we observed that providing slightly more specific research directions can improve idea-generation performance. Motivated by this observation, we investigate how the granularity of the research direction affects reconstruction performance. We define four levels of granularity by progressively narrowing the topical scope while avoiding information that directly reveals the target problem, research gap, motivation, or method. The four levels correspond to a *broad field*, a *research topic*, a *focused research area*, and a *fine-grained research scope*. Examples for each granularity level are provided in Appendix B.4. Each level introduces only a modest amount of additional information over the previous one. As shown in Figure 6, AutoHRS scores consistently increase as the research direction becomes more fine-grained.

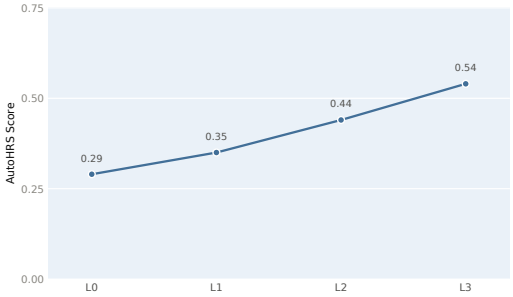



Figure 6: **Impact of research-direction granularity on idea-generation performance.** Overall AutoHRS score at different levels of research-direction granularity (L0–L3).

**Performance Across Research Types** To examine whether reconstruction performance varies across research types, we manually identify papers in the evaluation corpus that fall into three categories: *Empirical Analysis*, *Method/Algorithm*, and *Benchmark/Evaluation*. For each category, we select 20 papers that match the corresponding research type and evaluate idea-generation performance within each subset. As shown in Figure 7, auto researchers achieve the highest AutoHRS score of 0.49 on *Empirical Analysis* papers, followed by 0.32 on *Method/Algorithm* papers and 0.18 on *Benchmark/Evaluation* papers. These results suggest that the ability of auto researchers to reconstruct human research ideas varies substantially across research types, with benchmark- and evaluation-oriented work being particularly challenging.

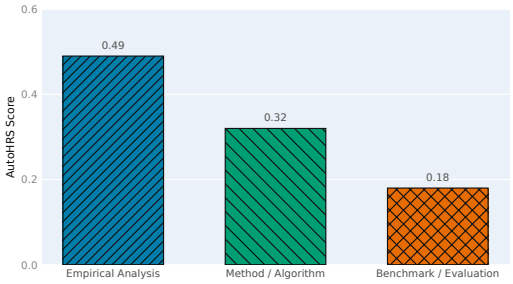



Figure 7: **Impact of research type on idea-generation performance.** Overall AutoHRS score across different research types: *Empirical Analysis*, *Method/Algorithm*, and *Benchmark/Evaluation*.

**Takeaway**

More fine-grained research directions consistently improve idea reconstruction. Performance also varies by research type, with AutoHRS scores of 0.49 for empirical analysis, 0.32 for methodological work, and 0.18 for benchmark and evaluation research.

# 4 Related Work

## 4.1 Automated Scientific Research

Recent advances in large language models and general-purpose agent harnesses have substantially broadened the scope of automated scientific research. Early LLM-based work largely focused on individual stages of the research process, such as idea generation, experimentation, and scientific writing (Hu et al., 2024; Baek et al., 2025; Guo et al., 2024; Schmidgall et al., 2025; Jiang et al., 2025; Chamoun et al., 2024; Chen et al., 2026; Yuan et al., 2026). More recently, harness-based auto researchers have begun to operate across the full research workflow, shifting the role of humans increasingly toward high-level guidance and oversight. Yang et al. (2026) introduce ARIS, which enables end-to-end autonomous research through adversarial multi-agent collaboration, while Wang et al. (2026) propose Sibyl-AutoResearch, a self-evolving research harness that improves through trial-and-error experience. In this work, we focus specifically on harness-based auto researchers that carry out multiple stages of the scientific research process in an end-to-end manner.

## 4.2 Evaluation of Auto Researchers

Existing evaluations of automated scientific research largely focus either on model-level capabilities or on task completion within individual stages of the research process (Chen et al., 2025; Zhang et al., 2026). RE-Bench (Wijk et al., 2025) compares language-model agents with human experts on open-ended machine-learning research engineering tasks. IdeaBench (Guo et al., 2025) evaluates LLMs on research idea generation, focusing on the quality of generated ideas. In contrast, the systematic evaluation of general-purpose harness-based auto researchers has only recently begun to emerge. Fei et al. (2026) focus on diagnosing and tracing the causes of failures in end-to-end AutoResearch trajectories. A concurrent work, ARAC (Cui et al., 2026), evaluates auto-research systems against generalized human research methodology, whereas we directly assess whether they reconstruct the specific research decisions of individual human research trajectories.

# 5 Conclusion and Discussion

We introduced AutoHRS, a stage-wise framework for evaluating how faithfully auto researchers reconstruct human research across idea generation, experiment design, and paper synthesis. AutoHRS further provides a fine-grained checklist-based evaluation pipeline with stage-specific dimensions, and we validate its alignment with human judgments. Our results reveal substantial differences from human research across all three stages, with larger gaps in earlier research decisions. Further analysis shows recurring patterns in how auto researchers diverge from human researchers, including reference-driven idea composition and conventional experiment planning. We also find that reconstruction performance can improve with more fine-grained research directions and varies across research types. These findings suggest that evaluating auto researchers requires looking beyond end-to-end task completion to the research decisions and processes underlying their outputs. Future auto-researcher systems should therefore aim not only to automate scientific workflows, but also to make research decisions that more closely resemble those of human researchers.

# 6 Limitations

Our study has several limitations. First, AutoHRS is evaluated only on research papers from the AI domain, and it remains unclear whether the observed patterns generalize to other scientific disciplines with different research practices and evaluation criteria. Second, running end-to-end auto-researcher trajectories is computationally expensive. As a result, our main evaluation is limited to 152 papers, corresponding to 456 stage-wise research trajectories. While this scale is sufficient to reveal consistent patterns across our analyses, broader evaluations over larger and more diverse corpora would provide a more comprehensive assessment of auto-researcher behavior.

### AI USE STATEMENT

Generative AI tools were used in this work for stage-wise information extraction and reformulation, checklist generation, checklist-based evaluation, and single-LLM evaluation. They were also used to assist with code development and debugging, as well as with manuscript drafting and editing. All AI-assisted outputs were reviewed and verified by the authors, who take full responsibility for the final content, claims, and results.

### ETHICS STATEMENT

This work studies the behavior of automated research systems using research artifacts derived from publicly available scientific papers. Our analysis is intended to characterize aggregate differences between auto researchers and human research rather than to evaluate or make claims about individual researchers. We report results at the aggregate level and do not use the framework to profile individual authors. We acknowledge that automated research systems may affect scientific practice in ways that extend beyond the scope of this study, highlighting the importance of appropriate human oversight.

### REPRODUCIBILITY STATEMENT

We provide detailed descriptions of the evaluation tasks, corpus construction procedure, evaluation dimensions, scoring metrics, and experimental setup in the main paper and appendix. The appendix further includes the prompts used for task construction, checklist generation, and LLM-based evaluation, as well as additional cross-harness experiments. To further support reproducibility, we will publicly release the evaluation corpus, generated checklists, code, and associated evaluation resources. These materials are intended to enable reproduction of the AutoHRS evaluation pipeline and facilitate future comparisons with additional auto-researcher systems.

## REFERENCES

Jinheon Baek, Sujay Kumar Jauhar, Silviu Cucerzan, and Sung Ju Hwang. ResearchAgent: Iterative Research Idea Generation over Scientific Literature with Large Language Models. In *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pp. 6709–6738. Association for Computational Linguistics, 2025.

Eric Chamoun, Michael Sejr Schlichtkrull, and Andreas Vlachos. Automated Focused Feedback Generation for Scientific Writing Assistance. In *Findings of the Association for Computational Linguistics: ACL (ACL Findings)*, pp. 9742–9763. ACL, 2024.

Nuo Chen, Andre Huikai Lin, Jiaying Wu, Junyi Hou, Zining Zhang, Qian Wang, Xidong Wang, and Bingsheng He. XtraGPT: Context-Aware and Controllable Academic Paper Revision via Human-AI Collaboration. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 1045–1074. ACL, 2026.

Ziru Chen, Shijie Chen, Yuting Ning, Qianheng Zhang, Boshi Wang, Botao Yu, Yifei Li, Zeyi Liao, Chen Wei, Zitong Lu, Vishal Dey, Mingyi Xue, Frazier N. Baker, Benjamin Burns, Daniel Adu-Ampratwum, Xuhui Huang, Xia Ning, Song Gao, Yu Su, and Huan Sun. ScienceAgentBench: Toward Rigorous Assessment of Language Agents for Data-Driven Scientific Discovery. In *International Conference on Learning Representations (ICLR)*, 2025.

Jiale Cui, Yueyao Yuan, Kaixi Zhong, Xiaogang Xu, Jiafei Wu, and Zhe Liu. ARAC: Benchmarking Auto-Research's Alignment and Completeness on End-to-End Researchs. *CoRR abs/2608.12788*, 2026.

Yanlin Fei, Nazhou Liu, Xinmiao Yu, Shaolong Chen, Lei Li, Rahul Thapa, Madalina G. Ciobanu, Navan Preet Singh, Qingqing Mao, and Ritankar Das. How Do Agents Fail on AutoResearch: End-to-End Diagnostic Evaluation on 100 Real-World Frontier Research Tasks. *CoRR abs/2608.14905*, 2026.

Sikun Guo, Amir Hassan Shariatmadari, Guangzhi Xiong, Albert Huang, Myles Kim, Corey M. Williams, Stefan Bekiranov, and Aidong Zhang. IdeaBench: Benchmarking Large Language Models for Research Idea Generation. In *ACM Conference on Knowledge Discovery and Data Mining (KDD)*, pp. 5888–5899. ACM, 2025.

Siyuan Guo, Cheng Deng, Ying Wen, Hechang Chen, Yi Chang, and Jun Wang. DS-Agent: Automated Data Science by Empowering Large Language Models with Case-Based Reasoning. In *International Conference on Machine Learning (ICML)*, pp. 16813–16848. PMLR, 2024.

Xiang Hu, Hongyu Fu, Jinge Wang, Yifeng Wang, Zhikun Li, Renjun Xu, Yu Lu, Yaochu Jin, Lili Pan, and Zhenzhong Lan. Nova: An Iterative Planning and Search Approach to Enhance Novelty and Diversity of LLM Generated Ideas. *CoRR abs/2410.14255*, 2024.

Zhengyao Jiang, Dominik Schmidt, Dhruv Srikanth, Dixing Xu, Ian Kaplan, Deniss Jacenko, and Yuxiang Wu. AIDE: AI-Driven Exploration in the Space of Code. *CoRR abs/2502.13138*, 2025.

David Klahr and Kevin Dunbar. *Dual Space Search During Scientific Reasoning*. Cogn. Sci., 1988.

Patrick Tser Jern Kon, Qiuyi Ding, Jiachen Liu, Xinyi Zhu, Jingjia Peng, Jiarong Xing, Yibo Huang, Yiming Qiu, Jayanth Srinivasa, Myungjin Lee, Mosharaf Chowdhury, Matei Zaharia, and Ang Chen. EXP-Bench: Can AI Conduct AI Research Experiments? In *International Conference on Learning Representations (ICLR)*, 2026.

Sandeep Kumar, Tirthankar Ghosal, Vinayak Goyal, and Asif Ekbal. Can Large Language Models Unlock Novel Scientific Research Ideas? In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 33563–33587. ACL, 2025.

Pat Langley, Herbert A. Simon, Gary L. Bradshaw, and Jan M. Zytkow. *Scientific Discovery: Computational Explorations of the Creative Process*. The MIT Press, 1987.

Long Li, Weiwen Xu, Jiayan Guo, Ruochen Zhao, Xingxuan Li, Yuqian Yuan, Boqiang Zhang, Yuming Jiang, Yifei Xin, Ronghao Dang, Yu Rong, Deli Zhao, Tian Feng, and Lidong Bing. Chain of Ideas: Revolutionizing Research Via Novel Idea Development with LLM Agents. In *Findings of the Association for Computational Linguistics: EMNLP (EMNLP Findings)*, pp. 8971–9004. ACL, 2025.

Chris Lu, Cong Lu, Robert Tjarko Lange, Jakob Foerster, Jeff Clune, and David Ha. The AI Scientist: Towards Fully Automated Open-Ended Scientific Discovery. *CoRR abs/2408.06292*, 2024.

Samuel Schmidgall, Yusheng Su, Ze Wang, Ximeng Sun, Jialian Wu, Xiaodong Yu, Jiang Liu, Michael Moor, Zicheng Liu, and Emad Barsoum. Agent Laboratory: Using LLM Agents as Research Assistants. In *Findings of the Association for Computational Linguistics: EMNLP (EMNLP Findings)*, pp. 5977–6043. ACL, 2025.

Chenglei Si, Diyi Yang, and Tatsunori Hashimoto. Can LLMs Generate Novel Research Ideas? A Large-Scale Human Study with 100+ NLP Researchers. In *International Conference on Learning Representations (ICLR)*, 2025.

Luciana B. Sollaci and Mauricio G. Pereira. The introduction, methods, results, and discussion (IMRAD) structure: a fifty-year survey. *Journal of the Medical Library Association*, 2004.

Noy Sternlicht and Tom Hope. CHIMERA: A Knowledge Base of Scientific Idea Recombinations for Research Analysis and Ideation. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 1871–1905. ACL, 2026.

John M. Swales. *Genre Analysis: English in Academic and Research Settings*. Cambridge University Press, 1990.

Qiong Tang, Xiangkun Hu, Xiangyang Liu, Yiran Chen, and Yunfan Shao. FARS: A Fully Automated Research System Deployed at Scale. *CoRR abs/2606.31651*, 2026.

Chengcheng Wang, Qinhua Xie, Wei He, Jianyuan Guo, Shiqi Wang, and Chang Xu. Sibyl-AutoResearch: Autonomous Research Needs Self-Evolving Trial-and-Error Harnesses, Not Paper Generators. *CoRR abs/2605.22343*, 2026.

Qingyun Wang, Doug Downey, Heng Ji, and Tom Hope. SciMON: Scientific Inspiration Machines Optimized for Novelty. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 279–299. ACL, 2024.

Hjalmar Wijk, Tao Roa Lin, Joel Becker, Sami Jawhar, Neev Parikh, Thomas Broadley, Lawrence Chan, Michael Chen, Joshua Clymer, Jai Dhyani, Elena Ericheva, Katharyn Garcia, Brian Goodrich, Nikola Jurkovic, Megan Kinniment, Aron Lajko, Seraphina Nix, Lucas Jun Koba Sato, William Saunders, Maksym Taran, Ben West, and Elizabeth Barnes. RE-Bench: Evaluating Frontier AI R&D Capabilities of Language Model Agents against Human Experts. In *International Conference on Machine Learning (ICML)*. PMLR, 2025.

Wanghan Xu, Shuo Li, Tianlin Ye, Qinglong Cao, Yixin Chen, Hengjian Gao, Yiheng Wang, Qi Li, Kun Li, Sheng Xu, Shengdu Chai, Fangchen Yu, Xiangyu Zhao, Zhangrui Zhao, Weijie Ma, Zijie Guo, Koutian Wu, Haoyu Zhou, Haoxiang Yin, Lixue Cheng, Chaofan Hu, Haoxuan Li, Lu Mi, Xuxuan Xie, Yifan Zhou, Ruizhe Chen, Zhiwang Zhou, Xingjian Guo, Yuhao Zhou, Xuming He, Shengyuan Xu, Xinyu Gu, Jiamin Wu, Mianxin Liu, Chunfeng Song, Fenghua Ling, Dongzhan Zhou, Shixiang Tang, Yuqiang Li, Mao Su, Peng Ye, Siqi Sun, Bin Wang, Xue Yang, Zhenfei Yin, Tianfan Fu, Guangtao Zhai, Wanli Ouyang, Bo Zhang, Lei Bai, and Wenlong Zhang. ResearchClawBench: A Benchmark for End-to-End Autonomous Scientific Research. *CoRR abs/2606.07591*, 2026.

Ruofeng Yang, Yongcan Li, and Shuai Li. ARIS: Autonomous Research via Adversarial Multi-Agent Collaboration. *CoRR abs/2605.03042*, 2026.

Weiwei Ye, Hangchen Liu, Dongyuan Li, and Renhe Jiang. PaperClaw: Harnessing Agents for Autonomous Research and Human-in-the-Loop Refinement. *CoRR abs/2606.22610*, 2026.

Chunhong Yuan, Tianshi Wei, Chuangqi Li, Xing Yi, Shengrui Liu, Zixin Zhang, Yule Cai, and Xinke Du. PaperOrchestrator: An LLM-Orchestrated multi-agent pipeline for automated end-to-end scientific paper writing. *Journal of King Saud University - Computer and Information Sciences*, 2026.

Zhengxin Zhang, Ning Wang, Sainyam Galhotra, and Claire Cardie. How Far Are We From True Auto-Research? *CoRR abs/2605.19156*, 2026.

Chen Zhu and Xiaolu Wang. HLER: Human-in-the-Loop Economic Research via Multi-Agent Pipelines for Empirical Discovery. *CoRR abs/2603.07444*, 2026.