SCIENCE CHINA
Information Sciences

• Supplementary File •

AI for Biology (2025)

From intention to implementation: automating biomedical research via LLMs

Yi Luo[1], Linghang Shi[1], Yihao Li[1], Aobo Zhuang[2], Yeyun Gong[3*], Ling Liu[4] & Chen Lin[1,5*]

[1]*School of Informatics, National Institute for Data Science in Health and Medicine, Xiamen University, Xiamen* 361101*, China*
[2]*School of Medicine, Xiamen University, Xiamen* 361101*, China*
[3]*Microsoft Research Asia, Beijing* 100080*, China*
[4]*College of Computing, Georgia Institute of Technology, Atlanta* 30332*, USA*
[5]*Zhongguancun Academy, Beijing* 100094*, China*

Appendix A Limitations and future works

Limitations In this paper, **BioResearcher** effectively automates the entire biomedical research process, from literature and dataset searches to experimental design and execution, given a research objective and specified conditions. However, several limitations warrant discussion. First, the system does not achieve complete success in executing experiments without manual intervention. This limitation is partly due to the need for further enhancement of the code generator agent's performance. Second, certain anomalies, such as the unavailability of resources specified in the experimental protocols, also hinder fully automated execution. It highlights the necessity of anticipating a broader range of exceptional scenarios and developing corresponding solutions in future work. Third, during our practice, we observed that excessively long inputs often result in LLMs producing shorter, more generalized outputs. To enable LLMs to generate detailed and comprehensive experimental protocols, we employed the multi-step, section-wise processing approach at various stages, including experimental report generation and experimental design. However, this iterative method increases the overall cost. Fourth, the framework currently does not formally model the research process as a constrained optimization problem. This limits its ability to optimally balance competing objectives (e.g., accuracy vs. computational cost) or dynamically allocate resources (e.g., lab equipment, cloud computing budgets), particularly in complex, multi-agent research scenarios.

Future works To further enhance and expand the capabilities of **BioResearcher**, we propose four key future research directions. First, **BioResearcher** currently supports only dry lab experiments. Future research could explore integrating automated wet lab technologies, such as Cloud Labs [1], to extend the system's applicability to wet lab studies. Additionally, its utility could be expanded to other scientific disciplines through three modular adaptations: (1) Extensible code modules supporting additional languages (Python/Julia) and domain-specific tools (LabVIEW for physics experiments); (2) Customizable knowledge bases using field-specific repositories (arXiv for CS/physics, ChemRxiv for chemistry); (3) Specialized innovation modules employing generative AI for hypothesis generation in creativity-driven domains. Second, hallucination remains a significant challenge in LLM-based applications, where the model may generate factually incorrect content. This issue is particularly critical in scientific research due to the model's limited domain-specific knowledge. Addressing LLM hallucination continues to be an essential area for future work. Third, incorporating human oversight mechanisms is crucial for ensuring reliability and ethical compliance in complex scientific workflows. We propose three avenues for integrating human intervention: (1) Confidence-Driven Human Intervention, where a confidence threshold mechanism prompts human validation when the reviewer agent's confidence falls below a predefined threshold; (2) Iterative Refinement via Natural Language Feedback, allowing users to provide corrections (e.g., missing controls or outdated references), which the AI agents will process and update accordingly while maintaining audit trails; and (3) Ethical Decision-Making Protocol, introducing an automated ethics checkpoint that flags high-risk proposals, mandates human approval for sensitive experiments, and integrates institutional ethics guidelines into the system's constraints. Future work will quantitatively evaluate the trade-offs between automation and human intervention to optimize system performance and usability. Fourth, incorporating constrained optimization techniques would enable resource-aware scheduling algorithms that balance computational budgets, equipment availability, and temporal constraints through multi-objective optimization – particularly valuable for large-scale collaborative studies.

Appendix B Implementation of baselines

In Section 5.2, due to the NCBI databases' limitations in semantic similarity searches, directly using the user's input objective poses challenges in retrieving relevant papers and datasets. Thus, we provide the RAG system with unfiltered papers and

* Corresponding author (email: yegong@microsoft.com, chenlin@xmu.edu.cn)

datasets from our Search module, enabling it to extract relevant datasets and content to generate a comprehensive protocol and corresponding code in a single step.

In Section 5.4, we directly supply the literature to be processed to the three systems, thereby not equipping them with any tools.

In Section 5.5, we equip the three systems with the aforementioned fourth tool. Here, we chunk each report and analysis by parts, rather than using the semantic similarity approach employed for chunking the papers in Section 5.2. Then, we store them in a vectorized format for the search tool.

We implement all baselines using the AutoGen framework [2] and adopt OpenAI's text-embedding-ada-002 as the embedding model in RAG applications.

Appendix C Research objectives

In this section, we outline the specific research objectives utilized in our experiments. Table C1 presents eight objectives employed for end-to-end automation experiments. These objectives are previously unmet and were gathered from a biomedical research laboratory. Table C2 displays ten objectives sourced from publicly available papers for evaluating the search module. Table C3 presents 15 objectives for assessing experimental design; these are also collected from published papers, but their publication dates are later than the cutoff for GPT-4o training data.

Table C1 Research objectives used for end-to-end automation experiment

ID	Research Objective
1	Classification of the Immune Microenvironment and Identification of Key Genes in Liposarcoma Based on Transcriptomics.
2	RNA-seq-based analysis of differences in metabolic characteristics between Well-differentiated and dedifferentiated Liposarcoma.
3	Comprehensive analysis of gene expression and immune microenvironment differences between retroperitoneal and limb liposarcoma.
4	Prognostic and Gene Expression Differences between Well-Differentiated and dedifferentiated Liposarcoma.
5	Differences in molecular expression and microenvironment between primary and recurrent liposarcoma.
6	Study on the recurrence mechanism of liposarcoma based on transcriptome and establishment of a prediction model.
7	Key Genes and Pathways for Evaluating the Efficacy of MDM2 Inhibitor Therapy in Liposarcoma.
8	Critical Genes and Pathways Associated with Radiotherapy Sensitivity in Retroperitoneal Well-Differentiated and Dedifferentiated Liposarcoma.

Appendix D Scoring criteria

In this section, we outline the specific scoring criteria used in our experiments. Table D1 presents the criteria for evaluating a paper's helpfulness in achieving a specific research objective. Additionally, Table D2 displays the criteria for assessing the quality of an experimental report across four dimensions.

Appendix E Cost analysis

Below, we illustrate the average number of tokens and the associated costs involved in generating an experimental protocol and its corresponding R code based on specified research targets, conditions, and requirements, using various large language models (LLMs) in Table E1. The process also encompasses the execution and verification of the code. In addition, after investigation, for a biomedical research cycle from the research target to the final dry experimental results, we compared the approximate manual time obtained from our survey with the time consumed by BioResearcher in Table E2.

Appendix F Comparison of protocols generated by four systems

Figure F1 shows a comparison of a part of the experimental protocols generated by four different systems: **BioResearcher**, Plan-and-execute, React, and RAG. Notably, the protocol produced by **BioResearcher** is both detailed and accurate. In contrast, the protocols generated by the other systems lack essential details, rendering them impractical for implementation.

Table F1 presents the evaluation results of the four assessment models on the protocols described in Section 5.2. As shown in the table, our method consistently outperforms the baseline systems across all four evaluation models.

Table C2 Research objectives used in the evaluation of the Search module

ID	Research Objective
1	Identify populations that are sensitive to immunotherapy for liposarcoma and explore the mechanisms of therapeutic sensitivity to guide the treatment of those whose immune quality is not sensitive.
2	To investigate the expression characteristics of the TBC1D4 activating protein molecule and identify key module genes for preventing the progression of thyroid cancer. Through bioinformatics analysis, elucidate the role of TBC1D4 in thyroid cancer and its regulated gene network to provide new molecular targets and research directions for early prevention and treatment of thyroid cancer.
3	To explore malignant cell characteristics related to microvascular invasion (MVI) in hepatocellular carcinoma (HCC) through multi-omics transcriptomic analysis and develop a machine learning prognostic model based on MVI-related genes.
4	To develop a machine learning classifier based on host immune response mRNA to accurately distinguish between acute viral respiratory infections (viral ARI) and non-viral respiratory infections (non-viral ARI). Identify and validate gene expression characteristics that can aid in clinical diagnosis using host gene expression data from nasal swab samples through a multi-omics analysis framework.
5	To identify preventive biomarkers related to DNA replication in ovarian cancer through gene expression analysis and bioinformatics analysis, particularly focusing on the expression of MCM2 protein and its role in ovarian cancer progression.
6	To analyze coagulation-related genes in hepatocellular carcinoma (HCC) and explore their relationship with the tumor immune microenvironment (TME) and clinical prognosis.
7	To systematically analyze long non-coding RNAs (lncRNAs) related to immune checkpoints (ICP) and explore their functions in cancer, as well as their potential as biomarkers for predicting immune therapy responses and prognosis.
8	To reveal the transcriptional patterns of HER2 through a pan-cancer analysis of HER2 indices, facilitating a more precise selection of patients suitable for HER2-targeted therapy.
9	To construct an immune-related gene prognostic index (IRGPI) to predict the prognosis of head and neck squamous cell carcinoma (HNSCC) patients and clarify the molecular and immune characteristics of different HNSCC subgroups defined by IRGPI, as well as the efficacy of immune checkpoint inhibitor (ICI) therapy.
10	To reveal the heterogeneity of T cell exhaustion (TEX) in the tumor microenvironment (TME) of different cancer types, depict the hierarchical functional dysfunction process of T cell exhaustion through pan-cancer analysis and investigate its association with prognosis and immune therapy efficacy.

Table C3 Research objectives used in the evaluation Experimental Design

ID	Research Objective
1	Analysis of differential expression characteristics of TBC1D4 gene in different stages of thyroid cancer progression.
2	Construction of TBC1D4-related gene co-expression modules and functional annotation.
3	Identification of differentially expressed genes in thyroid-related lesions.
4	Screening of core genes in key modules of thyroid cancer based on LASSO algorithm.
5	Identification of malignant cell subpopulations in hepatocellular carcinoma by single-cell RNA sequencing.
6	Pseudo-time analysis to explore the differentiation trajectory of malignant cells in hepatocellular carcinoma.
7	Exploring the communication pathways of cells related to microvascular invasion in hepatocellular carcinoma based on intercellular communication analysis.
8	Using GO and KEGG pathway analysis to reveal the functional enrichment of MCM2-related differentially expressed genes in ovarian cancer.
9	Analysis of the relationship between the expression level of MCM2 gene in ovarian cancer and clinical prognosis.
10	Analysis of the functional status of CD4+ T cell subsets in rheumatoid arthritis.
11	Analysis of key immune cell type composition characteristics in pancreatic ductal adenocarcinoma immune cell death-related gene subtypes.
12	Analysis of immune checkpoint expression patterns in high and low subtypes of immune cell death-related genes in pancreatic ductal adenocarcinoma.
13	Analysis of tumor microenvironment characteristics of pancreatic ductal adenocarcinoma based on high and low expression subtypes of immune cell death-related genes.
14	Screening of potential anti-tumor drugs based on the expression pattern of ribosome production genes.
15	Identification of ribosome production genes with high-frequency copy number variation in cancer.

Table D1 Scoring criteria for the helpfulness of papers

Rating	Description
1	**Not Helpful at All**: -The title and abstract of the paper are completely unrelated to the research objective, conditions, and requirements. -Provides no valuable information or insights.
2	**Slightly Helpful**: -The title and abstract of the paper have some minor relevance to the research objective, conditions, and requirements. -Provides very limited information and insights, making it difficult to contribute significantly to the research objective.
3	**Moderately Helpful**: -The title and abstract of the paper are somewhat relevant to the research objective, conditions, and requirements. -Provides some useful information and insights, but additional information or research may be needed to fully utilize it.
4	**Very Helpful**: -The title and abstract of the paper are highly relevant to the research objective, conditions, and requirements. -Provides a substantial amount of useful information and insights, making a significant contribution to the research objective.
5	**Extremely Helpful**: -The title and abstract of the paper perfectly align with the research objective, conditions, and requirements. -Provides critical information and insights that directly and significantly advance the research objective.

Table D2 Scoring criteria for experimental reports

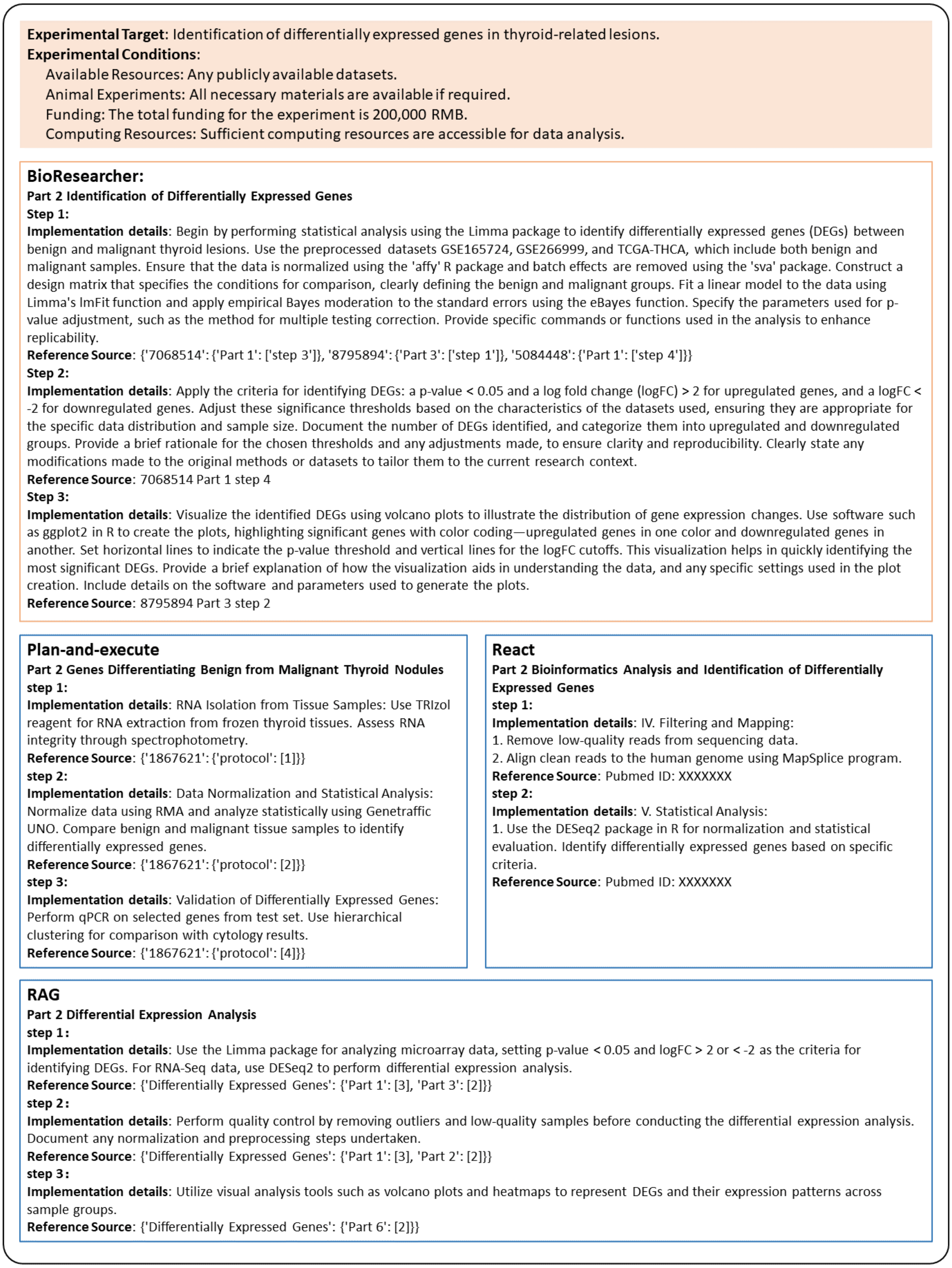
Dimension	**Criterion**
Logical Soundness	5: Completely logical arguments and conclusions, clear reasoning, no gaps. 4: Mostly logical, minor unclear points or leaps. 3: Generally logical, some noticeable gaps or inconsistencies. 2: Many logical flaws, disorganized logic. 1: Lacks coherence, severe flaws undermining arguments.
Detail Level	5: All necessary details, comprehensive descriptions. 4: Most necessary details, some areas brief. 3: Some necessary details, lacks important information. 2: Lacks many details, unclear descriptions. 1: Almost no details, difficult to understand.
Consistency with Original Paper	5: Entirely faithful, no misunderstandings. 4: Mostly faithful, minor misunderstandings. 3: Generally faithful, noticeable misunderstandings. 2: Many inconsistencies, significant deviations. 1: Severely inconsistent, obvious misunderstandings.
Readability of Report Structure	5: Clear structure, logical organization, easy to read. 4: Mostly well-structured, minor section issues. 3: Generally clear, some poorly arranged sections. 2: Quite disorganized, poor reading experience. 1: Very disorganized, extremely difficult to read.

Table E1 Cost of **BioResearcher**

Phase	Input Toktens	Output Toktens	GPT-4o-mini	GPT-4o-2024-08-06	GPT-4o cost
Search	0.09M	0.02M	0.03$	0.45$	0.91$
Literature Processing	49.62M	1.74M	8.49$	141.48$	282.96$
Design	0.77M	0.06M	0.15$	2.48$	4.96$
Program	61.08M	2.57M	10.70$	178.39$	344.07$
Total	111.55M	4.39M	19.37$	322.81$	632.90$

Table E2 Time comparison between human and **BioResearcher**

Phase	**Human**	**BioResearcher**
Search	1-3 Weeks	0.37 hours
Literature Processing	1-2 Weeks	5.03 hours
Design	1 Weeks	0.16 hours
Program	4-8 Week	2.6 hours
Total	**7-14 Weeks**	**8.16 hours**



Experimental Target: Identification of differentially expressed genes in thyroid-related lesions.
Experimental Conditions:
Available Resources: Any publicly available datasets.
Animal Experiments: All necessary materials are available if required.
Funding: The total funding for the experiment is 200,000 RMB.
Computing Resources: Sufficient computing resources are accessible for data analysis.

BioResearcher:

Part 2 Identification of Differentially Expressed Genes

Step 1:
Implementation details: Begin by performing statistical analysis using the Limma package to identify differentially expressed genes (DEGs) between benign and malignant thyroid lesions. Use the preprocessed datasets GSE165724, GSE266999, and TCGA-THCA, which include both benign and malignant samples. Ensure that the data is normalized using the 'affy' R package and batch effects are removed using the 'sva' package. Construct a design matrix that specifies the conditions for comparison, clearly defining the benign and malignant groups. Fit a linear model to the data using Limma's lmFit function and apply empirical Bayes moderation to the standard errors using the eBayes function. Specify the parameters used for p-value adjustment, such as the method for multiple testing correction. Provide specific commands or functions used in the analysis to enhance replicability.
Reference Source: {'7068514': {'Part 1': ['step 3']}, '8795894': {'Part 3': ['step 1']}, '5084448': {'Part 1': ['step 4']}}

Step 2:
Implementation details: Apply the criteria for identifying DEGs: a p-value < 0.05 and a log fold change (logFC) > 2 for upregulated genes, and a logFC < -2 for downregulated genes. Adjust these significance thresholds based on the characteristics of the datasets used, ensuring they are appropriate for the specific data distribution and sample size. Document the number of DEGs identified, and categorize them into upregulated and downregulated groups. Provide a brief rationale for the chosen thresholds and any adjustments made, to ensure clarity and reproducibility. Clearly state any modifications made to the original methods or datasets to tailor them to the current research context.
Reference Source: 7068514 Part 1 step 4

Step 3:
Implementation details: Visualize the identified DEGs using volcano plots to illustrate the distribution of gene expression changes. Use software such as ggplot2 in R to create the plots, highlighting significant genes with color coding—upregulated genes in one color and downregulated genes in another. Set horizontal lines to indicate the p-value threshold and vertical lines for the logFC cutoffs. This visualization helps in quickly identifying the most significant DEGs. Provide a brief explanation of how the visualization aids in understanding the data, and any specific settings used in the plot creation. Include details on the software and parameters used to generate the plots.
Reference Source: 8795894 Part 3 step 2

Plan-and-execute

Part 2 Genes Differentiating Benign from Malignant Thyroid Nodules

step 1:
Implementation details: RNA Isolation from Tissue Samples: Use TRIzol reagent for RNA extraction from frozen thyroid tissues. Assess RNA integrity through spectrophotometry.
Reference Source: {'1867621': {'protocol': [1]}}

step 2:
Implementation details: Data Normalization and Statistical Analysis: Normalize data using RMA and analyze statistically using Genetraffic UNO. Compare benign and malignant tissue samples to identify differentially expressed genes.
Reference Source: {'1867621': {'protocol': [2]}}

step 3:
Implementation details: Validation of Differentially Expressed Genes: Perform qPCR on selected genes from test set. Use hierarchical clustering for comparison with cytology results.
Reference Source: {'1867621': {'protocol': [4]}}

React

Part 2 Bioinformatics Analysis and Identification of Differentially Expressed Genes

step 1:
Implementation details: IV. Filtering and Mapping:
1. Remove low-quality reads from sequencing data.
2. Align clean reads to the human genome using MapSplice program.
Reference Source: Pubmed ID: XXXXXXX

step 2:
Implementation details: V. Statistical Analysis:
1. Use the DESeq2 package in R for normalization and statistical evaluation. Identify differentially expressed genes based on specific criteria.
Reference Source: Pubmed ID: XXXXXXX

RAG

Part 2 Differential Expression Analysis

step 1:
Implementation details: Use the Limma package for analyzing microarray data, setting p-value < 0.05 and logFC > 2 or < -2 as the criteria for identifying DEGs. For RNA-Seq data, use DESeq2 to perform differential expression analysis.
Reference Source: {'Differentially Expressed Genes': {'Part 1': [3], 'Part 3': [2]}}

step 2:
Implementation details: Perform quality control by removing outliers and low-quality samples before conducting the differential expression analysis. Document any normalization and preprocessing steps undertaken.
Reference Source: {'Differentially Expressed Genes': {'Part 1': [3], 'Part 2': [2]}}

step 3:
Implementation details: Utilize visual analysis tools such as volcano plots and heatmaps to represent DEGs and their expression patterns across sample groups.
Reference Source: {'Differentially Expressed Genes': {'Part 6': [2]}}

Figure F1 A comparison of **BioResearcher**, Plan-and-Execute, React, and RAG in generating protocols

Table F1 Quality of experimental protocols generated by various systems evaluated by different models

Evaluator	Method	Completeness	Detail	Correctness	Logical Soundness	Structure	Overall	l_{steps}	$n_{total\ steps}$
GPT-4o	RAG	0.392	0.765	0.470	0.991	**0.952**	3.570	1.286	9.167
	ReAct	0.361	0.617	0.487	0.972	0.950	3.387	1.237	5.792
	Plan and Execute	0.366	0.617	0.487	**1.000**	0.935	3.405	1.194	6.375
	BioResearcher	$\mathbf{0.612}_{\uparrow\mathbf{0.220}}$	$\mathbf{0.902}_{\uparrow\mathbf{0.137}}$	$\mathbf{0.944}_{\uparrow\mathbf{0.457}}$	$0.987_{\downarrow 0.013}$	$0.910_{\downarrow 0.042}$	$\mathbf{4.355}_{\uparrow\mathbf{0.785}}$	7.327	33.958
o3-mini-3-31	RAG	0.376	0.731	0.488	0.960	**0.967**	3.522	1.286	9.167
	ReAct	0.329	0.608	0.479	0.958	0.956	3.330	1.237	5.792
	Plan and Execute	0.375	0.621	0.479	**0.943**	0.960	3.378	1.194	6.375
	BioResearcher	$\mathbf{0.509}_{\uparrow\mathbf{0.133}}$	$\mathbf{0.981}_{\uparrow\mathbf{0.250}}$	$\mathbf{0.823}_{\uparrow\mathbf{0.435}}$	$0.944_{\downarrow 0.017}$	$0.952_{\downarrow 0.015}$	$\mathbf{4.309}_{\uparrow\mathbf{0.931}}$	7.327	33.958
gemini-2.0-flash	RAG	0.397	0.567	0.488	0.982	**0.820**	3.254	1.286	9.167
	ReAct	0.329	0.455	0.490	0.966	0.795	3.036	1.237	5.792
	Plan and Execute	0.341	0.475	0.488	**0.971**	0.827	3.102	1.194	6.375
	BioResearcher	$\mathbf{0.758}_{\uparrow\mathbf{0.361}}$	$\mathbf{0.754}_{\uparrow\mathbf{0.188}}$	$\mathbf{0.872}_{\uparrow\mathbf{0.382}}$	$0.943_{\downarrow 0.039}$	$0.792_{\downarrow 0.035}$	$\mathbf{4.118}_{\uparrow\mathbf{0.864}}$	7.327	33.958
deepseek-V3	RAG	0.454	0.685	0.486	0.960	**0.892**	3.476	1.286	9.167
	ReAct	0.435	0.629	0.478	0.957	0.885	3.385	1.237	5.792
	Plan and Execute	0.437	0.633	0.476	**0.945**	0.879	3.370	1.194	6.375
	BioResearcher	$\mathbf{0.758}_{\uparrow\mathbf{0.305}}$	$\mathbf{0.933}_{\uparrow\mathbf{0.248}}$	$\mathbf{0.843}_{\uparrow\mathbf{0.357}}$	$0.940_{\downarrow 0.017}$	$0.910_{\downarrow 0.019}$	$\mathbf{4.380}_{\uparrow\mathbf{0.909}}$	7.327	33.958
Average	RAG	0.405	0.687	0.483	**0.973**	**0.908**	3.456	1.286	9.167
	ReAct	0.364	0.577	0.484	0.963	0.897	3.285	1.237	5.792
	Plan and Execute	0.380	0.587	0.483	0.965	0.900	3.314	1.194	6.375
	BioResearcher	$\mathbf{0.659}_{\uparrow\mathbf{0.254}}$	$\mathbf{0.893}_{\uparrow\mathbf{0.206}}$	$\mathbf{0.895}_{\uparrow\mathbf{0.411}}$	$0.953_{\downarrow 0.020}$	$0.891_{\downarrow 0.017}$	$\mathbf{4.292}_{\uparrow\mathbf{0.836}}$	7.327	33.958

References

1 Arnold C. Cloud labs: where robots do the research. Nature, 2022, 606(7914): 612-613.

2 Wu Q Y, Bansal G, Zhang J Y, et al. Autogen: Enabling next-gen LLM applications via multi-agent conversation framework. arXiv preprint, 2023, abs/2308.08155.