Evolving collaborative research ideas with multi-agent grounding in lab-specific contexts and literature

Yu Chinen [1] **Haruka Ozaki** [1]

[1] *Laboratory for AI Biology, RIKEN Center for Biosystems Dynamics, Japan*. Correspondence to: Yu Chinen yu.chinen@riken.jp, Haruka Ozaki haruka.ozaki@riken.jp.

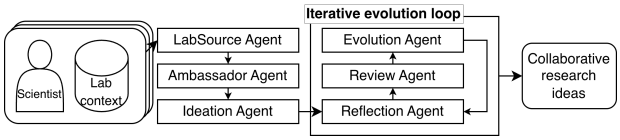

Fig. 1: Multi-agent evolution loop for collaborative research ideation.

1. Introduction

Collaborative research ideas have been shown to offer advantages in scientific impact and innovation when developed across institutional or disciplinary boundaries [1, 2]. Exploring and designing such collaborative research ideas remains costly because knowledge is siloed, and lab context, including research targets and methods, is distributed across different labs. Even when using AI to generate collaborative research ideas, simply juxtaposing lab context tends to yield ideas that are hard to validate or implement. Existing AI for Science research focuses on automation with a single objective and a single environment [3, 4] or collaboration for a single task [5]. However, the design problem of integrating independent lab contexts and synthesizing them into executable research plans has not been sufficiently addressed. We structure the lab context as *LabContext* and design an evolution loop where Reflection Agent and Review Agent, combined with an external literature search, ground ideas in lab context and relevant literature. This aims to suppress extreme collaborative research ideas that are only novel and to guide ideas toward higher collaboration value. Our main contribution is a multi-agent system that integrates multi-lab information as *LabContext* and evolves collaborative research ideas while grounding them in *LabContext* and related literature via Reflection Agent and Review Agent. The effect of this system is supported by an all-pairs comparison between iteration-0 initial collaborative research ideas and the non-pruned pool at iteration 6, where evolved ideas won 77.8% of the comparisons.

2. Methods

Task Definition. The task is to generate collaborative research ideas across multiple labs that are grounded in each lab's context and valuable for collaboration. We represent each lab's context as *LabContext*, a structured schema containing approximately 10 representative paper abstracts and commonly used methods for that lab. Each collaborative research idea consists of an objective, hypothesis, experimental plan, and role allocation.

Table 1: Setup parameters.

Parameter	Value	Note
Lab pairs	6	all pairs of 4 labs
Pool size	15	max kept per iteration
Evolve top-k	5	evolved each iteration
Iterations	6	iter=0 after Ideation

Table 2: Pairwise win rates (LLM-as-a-Judge; iteration labels hidden). Comparisons are against the iteration-0 initial ideas ($n = 6$). "Iter-6 pool" denotes the non-pruned idea set at iteration 6 ($n = 15$), and "Iter-6 new" denotes the five ideas newly generated at iteration 6 ($n = 5$).

Comparison set	Wins/Total	Pairwise win rate (%)
Iter-6 pool	70/90	77.8
Iter-6 new	25/30	83.3

We evaluate collaborative research ideas along the following four criteria:

1. **Scientific Validity:** consistency with lab context (*LabContext*) and related literature
2. **Novelty:** presence of new hypotheses or combinations
3. **Feasibility:** practicality of the experimental plan and role allocation
4. **Collaboration Value:** complementarity across labs and value not achievable by a single lab

Approach Overview. To address this task, we propose a multi-agent system that integrates lab context and related literature across multiple labs and iteratively generates and refines collaborative research ideas. The system consists of the following agents:

1. LabSource Agent compiles each lab's context from papers, research targets, methods, and representative abstracts.
2. Ambassador Agent structures each lab's context into a standardized *LabContext* representation usable by other agents.
3. Ideation Agent generates initial collaborative research ideas from *LabContext*.
4. Reflection Agent retrieves related literature via search and provides evaluation feedback to ground ideas in *LabContext* and relevant literature.

5. Review Agent performs pairwise comparisons using the four criteria, scores ideas with a Bradley–Terry model, and selects top ideas.

6. Evolution Agent improves selected ideas and forwards them to the next iteration.

All agents use the OpenAI API (GPT-5.2) [6] with reasoning effort set to medium.

Iterative Evolution with *LabContext* and Literature Grounding. Reflection Agent retrieves related literature from Semantic Scholar [7] and evaluates ideas against *LabContext* and the four criteria, while Review Agent incorporates these evaluations into pairwise comparisons. The multi-agent workflow consists of *LabContext* preparation (1–2), one-time initial idea generation (3), and an iterative stage (4–6), where Reflection Agent, Review Agent, and Evolution Agent form one iteration.

Experimental Setup. We evaluated the proposed method on four labs with distinct domains and orientations:

- Lab-A: live tissue imaging, mouse skin, hair follicles
- Lab-B: single-molecule measurements, super-resolution microscopy, axonal transport
- Lab-C: PURE system (in vitro protein synthesis), quantitative proteomics
- Lab-D: hibernation, hypometabolism, neural circuit manipulation, organismal physiology

The main setup parameters are shown in Table 1.

Evaluation. To measure the effect of the proposed method, we compared all pairs between initial ideas at iteration 0 and the non-pruned pool at iteration 6. We used blind pairwise comparison with GPT-5.2 under the same four criteria as Review Agent, yielding $6 \times 15 = 90$ comparisons.

3. Results

In blind LLM-as-a-Judge pairwise comparisons (iteration labels hidden) under four criteria (Scientific Validity, Novelty, Feasibility, Collaboration Value), iteration-6 ideas won over the iteration-0 initial ideas (one per $\binom{4}{2}$ lab pair; $n = 6$). The pairwise win rate was higher when restricting the iteration-6 set to the five ideas newly generated at iteration 6 than when using the non-pruned pool (Table 2). Table 3 shows two representative examples to contrast the initial and evolved ideas.

Beyond higher pairwise win rates, evolved ideas qualitatively appeared more actionable, with clearer staged experimental planning, explicit confounder handling, and occasional risk considerations. To quantify this trend with a lightweight text-based indicator, we measured the presence of a few markers related to planning and control in the original Japanese idea descriptions using exact string matching. We flagged stage structure by the occurrence of "stage" or "phase", confounder handling by "confounding", and risk consideration by "risk". In the five newly generated ideas at iteration 6, these markers appeared in 5/5, 5/5, and 2/5 ideas, respectively; in the iteration-0 initial ideas ($n = 6$), they appeared in 1/6, 0/6, and 0/6 ideas. This increase is consistent with the evolution loop encouraging more structured and realistic plans rather than merely increasing novelty.

At iteration 6, the non-pruned pool (n=15) consisted of ideas from three lab pairs: Lab-B–Lab-C (11/15, 73.3%), Lab-A–Lab-C (3/15, 20.0%), and Lab-A–Lab-D (1/15, 6.7%). In the Review Agent ranking at iteration 6, the top seven were all Lab-B–Lab-C, and the first non-Lab-B–Lab-C idea was Lab-A–Lab-C at Rank 8.

4. Discussion

The iterative evolution using Reflection Agent, Review Agent, and Evolution Agent, which grounds ideas in *LabContext* and relevant literature, suggests that collaborative research ideas can be improved beyond the initial ideas. The improvements are not mere lengthening, but tend toward stronger identifiability, execution planning, and risk management. We identify two challenges. First, previously indicated limitations of LLM-as-a-Judge remain [8]: judgments can be sensitive to how the judge prompt specifying the evaluation criteria and instructions is interpreted and applied, may exhibit instruction-induced or model-specific biases, and can vary due to the stochasticity of LLM outputs; future work should validate robustness with independent human raters and cross-model evaluation. Second, diversity remains a concern. The Review Agent ranking suggests that iterative selection can converge toward particular pairs. Future work should jointly optimize quality and diversity through Quality-Diversity methods such as MAP-Elites [9] and lab-pair caps.

Table 3: Representative examples of collaborative research ideas before and after iterative evolution loop (English translations of the original Japanese descriptions, by the authors).

Lab pair	Iter 0 (initial)	Iter 6 (evolved)
Lab-B–Lab-C	Elucidate axonal microtubule selective transport using site-specifically labeled kinesin synthesized cell-free	Link history effects of kinesin-induced lattice transitions in axonal microtubules to single-molecule and absolute quantification (Rank 1/15, iter 6)
Lab-A–Lab-C	Absolutely quantify and control hair follicle basement membrane rotation using cell-free synthesis standards	Switch collagen IV rotation gradients in mouse hair follicle basement membrane via ECM ratio thresholds (Rank 8/15, iter 5)

References

[1] Stefan Wuchty, Benjamin F. Jones, and Brian Uzzi. The increasing dominance of teams in production of knowledge. *Science*, 316(5827):1036–1039, 2007.

[2] Benjamin F. Jones, Stefan Wuchty, and Brian Uzzi. Multi-university research teams: Shifting impact, geography, and stratification in science. *Science*, 322(5905):1259–1262, 2008.

[3] Yutaro Yamada, Robert Tjarko Lange, Cong Lu, Shengran Hu, Chris Lu, Jakob Foerster, Jeff Clune, and David Ha. The ai scientist-v2: Workshop-level automated scientific discovery via agentic tree search, 2025.

[4] Juraj Gottweis, Wei-Hung Weng, Alexander Daryin, Tao Tu, Anil Palepu, Petar Sirkovic, Artiom Myaskovsky, Felix Weissenberger, Keran Rong, Ryutaro Tanno, Khaled Saab, Dan Popovici, Jacob Blum, Fan Zhang, Katherine Chou, Avinatan Hassidim, Burak Gokturk, Amin Vahdat, Pushmeet Kohli, Yossi Matias, Andrew Carroll, Kavita Kulkarni, Nenad Tomasev, Yuan Guan, Vikram Dhillon, Eeshit Dhaval Vaishnav, Byron Lee, Tiago R D Costa, José R Penadés, Gary Peltz, Yunhan Xu, Annalisa Pawlosky, Alan Karthikesalingam, and Vivek Natarajan. Towards an ai co-scientist, 2025.

[5] Kyle Swanson, Wesley Wu, Nash L. Bulaong, John E. Pak, and James Zou. The virtual lab of ai agents designs new sars-cov-2 nanobodies, 2025.

[6] OpenAI. GPT-5.2 Model (OpenAI API). `https://platform.openai.com/docs/models/gpt-5.2`. Accessed: 2026-01-13.

[7] Rodney Kinney, Chloe Anastasiades, Russell Authur, Iz Beltagy, Jonathan Bragg, Alexandra Buraczynski, Isabel Cachola, Stefan Candra, Yoganand Chandrasekhar, Arman Cohan, Miles Crawford, Doug Downey, Jason Dunkelberger, Oren Etzioni, Rob Evans, Sergey Feldman, Joseph Gorney, David Graham, Fangzhou Hu, Regan Huff, Daniel King, Sebastian Kohlmeier, Bailey Kuehl, Michael Langan, Daniel Lin, Haokun Liu, Kyle Lo, Jaron Lochner, Kelsey MacMillan, Tyler Murray, Chris Newell, Smita Rao, Shaurya Rohatgi, Paul Sayre, Zejiang Shen, Amanpreet Singh, Luca Soldaini, Shivashankar Subramanian, Amber Tanaka, Alex D. Wade, Linda Wagner, Lucy Lu Wang, Chris Wilhelm, Caroline Wu, Jiangjiang Yang, Angele Zamarron, Madeleine Van Zuylen, and Daniel S. Weld. The semantic scholar open data platform, 2025.

[8] Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, Saizhuo Wang, Kun Zhang, Yuanzhuo Wang, Wen Gao, Lionel Ni, and Jian Guo. A survey on llm-as-a-judge, 2025.

[9] Jean-Baptiste Mouret and Jeff Clune. Illuminating search spaces by mapping elites, 2015.