AGENTAGORA: FACILITATED MULTI-AGENT DISCUSSION FOR CREATIVE EXPLORATION

Anonymous authors
Paper under double-blind review

ABSTRACT

As large language models take on more open-ended tasks, creativity becomes increasingly important: design, inventive problem solving, and research ideation all require useful alternatives to familiar approaches. Multi-agent discussion offers a promising way to combine perspectives, but interaction can also draw agents toward the same directions. In particular, different initial roles do not ensure that later contributions address what remains unexplored. To address this gap, we introduce **AgentAgora**, a framework that coordinates exploration as discussion evolves. Agents with different perspectives contribute candidates to a shared record, while a chair maps the areas already covered and assigns complementary areas for the next round. As new contributions arrive, the chair redirects effort toward gaps and unresolved alternatives without proposing or rating candidates itself. This separation allows the same coordination mechanism to support idea generation, single-answer decisions, and research synthesis. Across these settings, AgentAgora improves evaluated output quality, ranking first under the four reported judges on LiveIdeaBench and improving CreativityBench answer accuracy by 2.44 percentage points over the strongest baseline. Ablations and discussion traces further support sustained, coverage-guided exploration. Together, these findings offer a practical route to creative multi-agent collaboration, with potential applications in design and discovery.

1 INTRODUCTION

As large language models (LLMs) become more capable at solving specified problems, their ability to formulate useful possibilities becomes increasingly important. Designing alternatives, finding unconventional uses for familiar objects, and proposing research ideas all depend on creativity: producing ideas that are both original and appropriate to their purpose (Runco & Jaeger, 2012). Studies of automated research ideation illustrate both the promise and the challenges, including feasibility, diversity, and reliable assessment (Si et al., 2024; Ruan et al., 2026). This motivates a practical question: how can inference procedures help models explore a broader range of useful possibilities rather than repeatedly elaborate familiar approaches to the same open-ended task?

Multi-agent discussion offers a promising approach. Different perspectives can expose assumptions, connect complementary knowledge, and develop alternatives through interaction; creative discussion systems and virtual scientific teams demonstrate the potential of this approach (Lu et al., 2024b; Su et al., 2025). Yet interaction does not ensure sustained exploration. Agents can converge on a few salient directions, making later contributions increasingly redundant even when their initial roles differ. Recent studies document diversity collapse in open-ended discussion and identify communication structure, authority, and majority pull as contributing factors (Chen et al., 2026; Nguyen et al., 2026). A practical gap remains between specifying *how* agents should think and deciding *where* they should search next. Static perspectives provide different starting points, but do not identify which regions have become crowded or which alternatives remain unexplored as discussion progresses.

We introduce **AgentAgora**, an inference-time framework that addresses this allocation problem by separating content generation from exploration coordination. Perspective-conditioned agents contribute candidates to a public blackboard. A chair reads the evolving board, maps the problem regions already addressed, and assigns complementary territories for the next round. These assignments are revised as contributions accumulate, allowing effort to shift away from exhausted direc-

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.

tions toward gaps and unresolved branches. The chair may recruit missing perspectives and decide when coverage has saturated, but it neither proposes nor rates candidates. This restriction keeps coordination focused on where to explore without turning the chair into an authoritative source of answers that participants are expected to follow.

The underlying design principle is that perspective diversity and coverage coordination serve different purposes. Perspectives influence how agents approach a task; coverage-based reassignment makes their work responsive to what the group has already explored. Following an initial reconnaissance round, AgentAgora alternates between board-informed allocation and parallel generation, admitting contributions through task-appropriate validity and novelty checks. Here, exploration means extending the set of relevant problem regions under consideration, not maximizing difference for its own sake. We instantiate this protocol in three representative settings: producing a set of scientific ideas, selecting a creative tool-use answer, and synthesizing an evidence-grounded report. These output regimes connect the shared exploration process to three complementary capabilities: developing useful alternatives, identifying a workable solution, and assembling complementary evidence into a coherent synthesis of the investigated question.

Our evaluation examines final quality and the search process. On a development-disjoint LiveIdeaBench (LIB) split, AgentAgora improves Overall by 0.128 points over the strongest baseline, Single Agent, under the default judge; the four reported judges rank it first. Chair ablations and round-wise traces associate coordination with sustained differentiation and fewer redundant proposals. On CreativityBench (CB), Gold Accuracy reaches 26.88%, versus 24.44% for MAD, while expanded candidate coverage exposes further opportunities for better selection. DeepResearch Bench (DRB) extends the evaluation to evidence-grounded synthesis, where the complete pipeline improves report content and citation support. Together, these findings support coverage-guided coordination as a reusable design principle for creative multi-agent systems, linking sustained exploration to improvements across distinct output regimes.

2 AGENTAGORA

AgentAgora separates *where to explore* from *what to propose*. Agents prompted with different perspectives generate content, while a coordinating agent, the chair, maps a public blackboard (the shared record of accepted contributions) and redirects search. It neither generates nor rates candidates; validation and final-output construction remain separate. The protocol supports idea sets, single decisions, and structured synthesis without updating model weights.

2.1 ARCHITECTURE AND DISCUSSION LOOP

Given a task x, a set of agents, called *seats*, explores possible responses over multiple rounds. The seats share a backbone but use different perspective prompts, initially covering first-principles, analogical, contrarian, engineering, and systems reasoning. The public blackboard B_t stores admitted contributions through round t, together with authorship and validation metadata.

The loop is a state update over the board. Before round $t > 1$, the chair derives coverage state C_{t-1} from B_{t-1} and dispatches territory $\tau_{i,t}$ to each active seat i. A seat conditions on the task, its perspective, the board, and its territory to produce $z_{i,t}$; a task-level operator A_x checks the output contract and updates $B_t = B_{t-1} \cup A_x(\{z_{i,t}\}_i, B_{t-1})$. The chair observes this admitted state, not private reasoning or rejected candidate text, and therefore coordinates from the group's durable record rather than supervising individual reasoning.

Round 1 is *reconnaissance*: seats work independently, without peer contributions or chair assignments, and propose brief candidates that initialize the board. Each later round repeats three steps (Figure 1). The chair reads the board, builds a coverage map, and proposes a territory for each active seat; each seat develops one candidate within its territory while reading the public board; and a task adapter applies task-specific rules to validate, deduplicate, and admit candidates. The chair then remaps the updated board rather than preserving a fixed division of work. When exploration stops, a task-specific readout (final selection or synthesis) converts the accumulated contributions into an idea set, a single answer, or a report grounded in the collected sources.

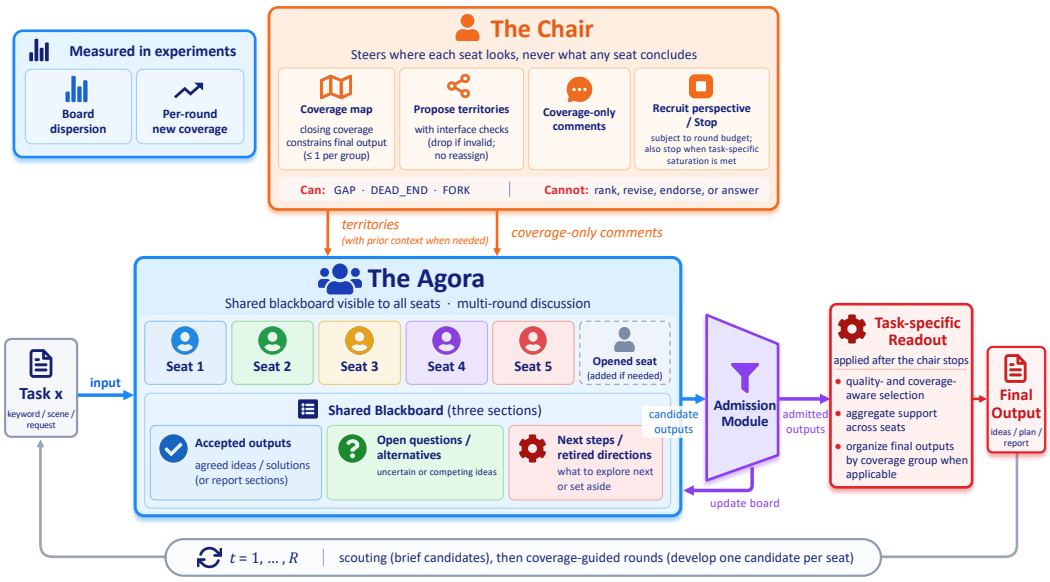


Figure 1: AgentAgora separates coverage coordination from content generation and final selection. The chair maps the shared board, assigns complementary territories, and issues coverage-only comments without proposing, ranking, or endorsing answers. Perspective-conditioned seats generate candidates, the task adapter admits valid outputs, and a task-specific readout constructs the final response after discussion stops.

The board and coverage map have different roles. The board exposes admitted content, chair comments, and retired directions to all seats. The map groups contributions by the problems they address, giving the chair a compact view of concentration and gaps. A territory changes a seat's next search target without changing its perspective or removing earlier contributions.

2.2 Coverage-oriented coordination

The chair rebuilds its coverage map from accumulated contributions rather than using a fixed task taxonomy. Grouping by the problem addressed, rather than by the proposed technique, reveals when superficially different candidates concentrate on the same underlying need. The chair uses the map to identify underexplored regions, exhausted directions, and incompatible claims.

For every continuing round, the chair proposes one *territory* per active seat: a concise problem domain, scene region, object group, or evidence facet. Territories direct where seats search without prescribing a solution. Interface checks drop territories phrased as questions or overlapping another accepted territory; an affected seat still participates but receives no assigned region. For research synthesis, the dispatch also lists sources already used by the group, encouraging seats to extend rather than repeat the evidence base.

The chair may issue three comments. `GAP` identifies an uncovered region, `DEAD_END` retires an exhausted direction, and `FORK` records incompatible claims without resolving them. It may recruit one additional perspective per round and recommend stopping after at least one assigned round. Execution also stops at the round budget or when an adapter-specific saturation condition is met, such as accumulating a sufficient pool of validated candidates. Every continuing round invokes all active seats; the chair can redirect but not deactivate them.

These restrictions separate coordination from answer authority. Retirement changes subsequent search rather than invalidating existing contributions, and a fork preserves disagreement rather than selecting a winner. At termination, however, the chair creates a closing coverage map. LIB uses it to limit the final set to at most one idea per problem group, and DRB uses it to organize report sections; selection within a group and all quality judgments remain outside the chair. The coverage map is

revised as the shared record grows. Appendix D specifies the interface checks, stopping rules, and deterministic fallback procedures used to maintain the discussion loop.

2.3 Instantiation across output regimes

Each application specifies its contribution format, admission rules, and final-output construction. Three representative regimes illustrate these interfaces: set-valued generation preserves alternatives; single-answer decision expands a pool before selection; and composite synthesis assembles complementary parts. LIB, CB, and DRB instantiate these regimes rather than define the method.

Table 1: Three representative output regimes around the shared discussion loop. Appendix D details their contribution formats, admission rules, and final-output construction.

Role	Diverse set (LIB)	Single decision (CB)	Composite synthesis (DRB)
Region	Problem domain	Scene / object group	Evidence facet
Contribution	Expanded idea	Entity–part pair	Cited memo
Admission	Quality + novelty	Membership + specification	At least two URLs
Output	Across-group selection	Support aggregation	Section composition

For *set-valued generation*, each board entry is a self-contained proposal and the closing map prevents final slots from repeating one problem region. In LIB, an internal evaluator rates originality, feasibility, and clarity; feasible, novel ideas enter the board, and a selector balances rated quality against distance across groups. This internal evaluation runs independently of chair coordination and serves a different purpose from the benchmark judge used to score final outputs.

For a *single decision*, coverage precedes selection. In CB, required outcomes are specified before discussion; seats nominate entity–part pairs, and admission checks scene membership and those outcomes. Ranking uses fork commitments (seats' choices between conflicting candidates), the number of distinct supporting seats, and verifier ratings, not chair judgments. Section 4.3 evaluates both candidate discovery and the selection of the final entity–part answer.

For *composite synthesis*, coverage specifies which facets to assemble rather than which candidate wins. In DRB, seats return source-cited memos; admission requires two URLs, with citation support assessed separately during evaluation. The closing map organizes a plan, after which a separate writer composes sections and code formats references. The writer remains outside the coordination loop, and Section 4.5 measures its contribution.

Across the regimes, the invariant mechanism is repeated reassignment from accumulated public state. A new application defines admissible contributions and an output contract without changing how the chair maps coverage, preserves forks, retires directions, or dispatches territories.

3 Related work

3.1 Creative generation and scientific ideation

Creative generation requires useful novelty, rather than variation in wording alone. A human study of research ideation finds that LLM-generated proposals can receive strong novelty ratings while exposing weaknesses in feasibility, diversity, and self-evaluation (Si et al., 2024). Recent benchmarks separate different aspects of this challenge: NoveltyBench evaluates diverse sets of high-quality answers, CREATE examines associative paths, and LiveIdeaBench evaluates scientific originality alongside feasibility and clarity (Zhang et al., 2025b; Wadhwa et al., 2026; Ruan et al., 2026). MacGyver tests unconventional object use (Tian et al., 2024), while CreativityBench scores entity–part solutions to affordance problems (Qian et al., 2026).

SciMON, ResearchAgent, and Co-Scientist develop research ideas through literature grounding and iterative refinement (Wang et al., 2024; Baek et al., 2025; Gottweis et al., 2026). LLM Discussion and Virtual Scientists organize creative work through role-based interaction (Lu et al., 2024b; Su et al., 2025), while dialogue-design studies examine how interaction protocols affect idea generation (Ueda et al., 2025). More recently, IDEAgent treats research ideation as quality-diversity search, maintaining idea lineages and using memory to guide revision and avoid repeated failures (Gumma

et al., 2026). AgentAgora uses that history for a different coordination decision: assigning the next problem region to each participant. This allocation layer complements candidate development and quality assessment by directing them toward the group's evolving search needs as new contributions enter the shared record.

3.2 Homogeneity and diversity in multi-agent discussion

Limited diversity precedes multi-agent interaction: studies find homogeneity across LLMs (Wenger & Kenett, 2025; Jiang et al., 2025) and reduced collective diversity in AI-assisted writing (Doshi & Hauser, 2024; Padmakumar & He, 2024). RLHF, verbalized sampling, and generation formats connect diversity to training and elicitation choices (Kirk et al., 2024; Zhang et al., 2025a; Yun et al., 2025). Communication introduces additional pressures: agents may follow salient proposals or authoritative participants, narrowing the discussion despite varied initial roles (Chen et al., 2026). Improving isolated sampling therefore need not prevent convergence once agents exchange their outputs.

Recent methods address this problem at different levels. Creative-MAD uses persistent cognitive lenses and distant peer contexts to resist identity drift and majority pull (Nguyen et al., 2026). DynaDebate generates and allocates solution paths, combining this allocation with process-level critique and triggered verification (Li et al., 2026). Building on these allocation approaches, AgentAgora focuses on the repeated assignment of complementary *problem regions* from an evolving board of admitted contributions, without having the chair prescribe solution content. This distinction motivates the chair-removal and round-wise analyses: do participants continue to extend the shared search, rather than merely start from different perspectives?

3.3 Shared-state coordination and research synthesis

Classical blackboard systems coordinate specialists through shared intermediate state (Erman et al., 1980; Nii, 1986). CAMEL and AgentVerse organize role-based cooperation (Li et al., 2023; Chen et al., 2024), alongside the structured workflows of AutoGen and MetaGPT (Wu et al., 2023; Hong et al., 2023). Apodex 1.1 trains task decomposition, parallel delegation, and replanning with shared task state (An et al., 2026). Self-Consistency and multi-agent debate combine reasoning attempts (Wang et al., 2023; Liang et al., 2024; Du et al., 2024). Within this broader orchestration setting, our chair allocates exploration effort across complementary problem regions. Separating allocation from candidate admission and final selection lets the shared record guide both search expansion and final output construction.

Research synthesis provides particularly close precedents. STORM elicits questions from multiple perspectives, Co-STORM couples a dynamic mind map with moderator interventions, and Open Deep Research delegates investigation through a supervisor (Shao et al., 2024; Jiang et al., 2024; LangChain, 2025). AgentAgora builds on these shared-state and facilitation approaches with a coordination protocol spanning multiple output regimes: an admitted-candidate board, recurring territory assignment to active participants, and a chair restricted to coverage guidance rather than candidate generation or rating. Our DRB evaluation examines the composition of this protocol with retrieval and writing through report-quality and citation-support measures. Component controls distinguish the contributions of exploration coordination and report construction.

4 Experiments

We test end-task quality across three domains, judge sensitivity on subjective scientific ideas, and whether discussion traces exhibit the repeated search that motivates coverage coordination. Exploration is operationalized through differentiated idea directions on LIB, candidate-pool coverage on CB, and evidence breadth on DRB. Together, these analyses connect final output quality, agreement across evaluators, and the evolution of the shared search process.

4.1 Evaluation design

Benchmarks. LiveIdeaBench (LIB) evaluates scientific ideas from keyword prompts (Ruan et al., 2026). We construct a discipline-stratified, development-disjoint split of 300 keywords and request

three ideas per task. The default GPT-4.1-mini judge scores originality (Orig.), feasibility (Feas.), clarity (Clar.), and *within-set distinctness* using an adapted fluency rubric. CreativityBench (CB) contains 3,285 affordance problems requiring the correct entity and enabling part (Qian et al., 2026); Gold Accuracy is rule-based. DeepResearch Bench (DRB) evaluates evidence-grounded reports (Du et al., 2025). We use its 100 requests, official reference-based report-quality evaluation (RACE), and citation-support evaluation (FACT). DRB serves as an exploratory system-level evaluation, with requests 1–50 also used for development or checks.

Systems and inference. All generation systems use GPT-5.4-mini without reasoning. LIB baselines are a Single Agent, Multi-Agent Debate (MAD) (Liang et al., 2024), and LLM Discussion (Lu et al., 2024b); CB also includes Self-Consistency@30 (aggregation of 30 samples) and Creative-MAD (Nguyen et al., 2026). DRB compares a retrieval-enabled Single Agent, STORM (Shao et al., 2024), Co-STORM (Jiang et al., 2024), and Open Deep Research (LangChain, 2025) under shared retrieval interfaces. AgentAgora begins with five perspectives and allows up to four discussion rounds for LIB and up to three rounds for CB and DRB.

We evaluate each system with its configured discussion budget. Appendix B reports resource usage for selected component configurations alongside their quality comparisons.

We compare methods on the same tasks and report bootstrap 95% confidence intervals (CIs) to quantify uncertainty. The primary LIB gains over Single Agent and the no-chair variant remain statistically significant after adjustment for these two comparisons. Appendix A describes the tests; process and component diagnostics are exploratory.

4.2 Scientific ideation: quality and judge sensitivity

Table 2: LIB-300 under the adapted protocol and default GPT-4.1-mini judge; bold scores mark column maxima. Distinct. is within-set distinctness, not linguistic fluency. Lower-tail $Q_{0.30}$ is the 30th percentile of task-level Overall. MAD returns one idea on 10 tasks, where distinctness is undefined and Overall averages the other dimensions.

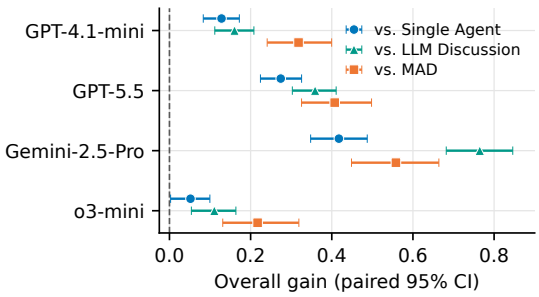
Method	Orig.	Feas.	Clar.	Distinct.	Overall	Lower-tail $Q_{0.30}$
Single Agent	8.237	6.524	**7.139**	8.677	7.644	7.500
MAD	8.070	6.391	6.834	8.631	7.454	7.333
LLM Discussion	8.233	6.327	6.932	**8.957**	7.612	7.500
AgentAgora	**8.412**	**6.769**	7.002	8.907	**7.772**	**7.667**

AgentAgora achieves the highest Overall, originality, and feasibility scores (Table 2). Its Overall gain over the strongest baseline, Single Agent, is $+0.128$ (95% CI $[+0.083, +0.172]$), with the largest advantages in originality and feasibility. The resulting profile combines useful novelty with a differentiated idea set, consistent with complementary territories helping agents develop distinct, workable proposals. Single Agent leads on clarity and LLM Discussion on within-set distinctness, while AgentAgora achieves the strongest aggregate balance. The higher $Q_{0.30}$ extends this advantage to the lower tail of task scores.

The four reported judges consistently rank AgentAgora first (Figure 2), supporting the improvement across evaluator choices. Gains over Single Agent range from 0.052 under o3-mini to 0.418 under Gemini, reflecting judge-specific assessments of scientific idea quality. We retain GPT-4.1-mini for the primary comparison and report each additional judge separately to preserve its scoring scale.

4.3 Creative tool use: separating coverage from selection

CB tests whether broader exploration helps identify an unconventional but workable use of an object. Unlike LIB, which rewards a set of ideas, each task requires a single entity–part answer. We therefore distinguish two stages: Coverage@pool measures whether the candidate set contains the correct pair, and Gold Accuracy measures whether the system ultimately selects it. Figure 3 presents both, making the contribution of candidate discovery visible alongside final-answer performance.



Judge	**Ours**	Single	LLM Disc.	MAD
GPT-4.1-mini	**7.772**	7.644	7.612	7.454
GPT-5.5	**7.286**	7.012	6.928	6.878
Gemini-2.5-Pro	**8.081**	7.664	7.317	7.523
o3-mini	**7.973**	7.921	7.863	7.756

Figure 2: LIB-300 judge robustness. Left: task-paired Overall gains over each baseline, with bootstrap 95% CIs. Right: absolute Overall scores. Each of the four reported judges ranks AgentAgora first, while the effect magnitude remains judge dependent.

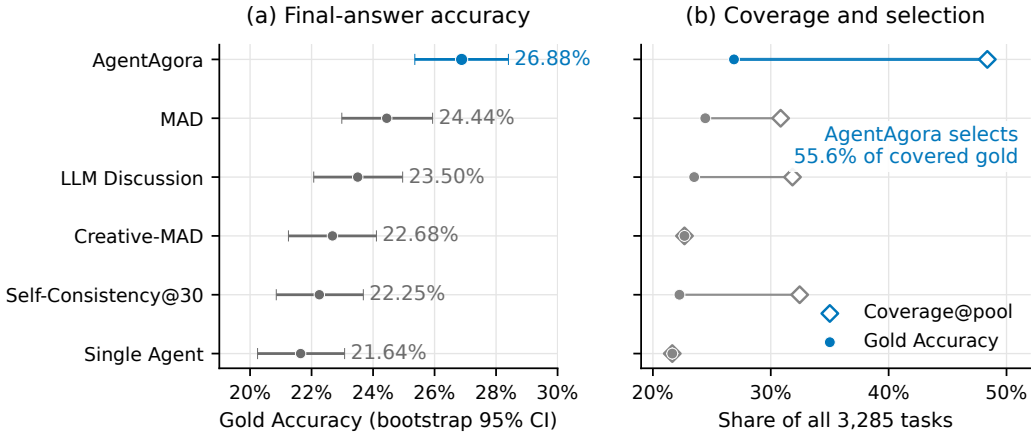


Figure 3: CreativityBench on all 3,285 tasks. Left: Gold Accuracy with bootstrap 95% CIs. Right: final Gold Accuracy and Coverage@pool, with each segment showing gold candidates found but not selected. Creative-MAD stores only its final answer, so its reported coverage equals Gold Accuracy by construction. Exact values appear in Appendix Table 5.

The CB results distinguish finding a useful possibility from choosing it. AgentAgora reaches 26.88% Gold Accuracy, compared with 24.44% for MAD, the strongest baseline (Figure 3), an absolute improvement of 2.44 percentage points on the full benchmark. The larger separation occurs before selection: its board contains the correct entity–part pair on 48.37% of tasks, compared with 32.45% for Self-Consistency@30. This pattern is consistent with coordinated exploration reaching alternatives that repeated sampling or discussion misses. Appendix A specifies the evaluation protocols and candidate-pool construction.

The coverage–selection decomposition makes the exploration benefit explicit. Coverage requires discovering both a suitable object and the part that enables its proposed use; simply naming more objects is insufficient. The higher coverage therefore identifies useful alternatives available to the final decision stage. This is the intended role of the shared board: preserving contributions from different perspectives so that a promising option remains available even when another direction has attracted more attention during the discussion.

AgentAgora selects the gold pair on 55.6% of tasks where it is present in the pool, yielding the strongest final accuracy among the compared systems. The remaining covered tasks contain correct alternatives for readout refinement. Exploration and selection thus offer complementary opportunities: the former expands the available solutions, while the latter determines which discovered solution becomes the answer. Together, these measures connect an expanded solution pool to stronger single-answer performance on the benchmark's entity–part decisions.

4.4 EVIDENCE-GROUNDED RESEARCH SYNTHESIS

Table 3: DRB results. Comp., Instr., and Read. denote comprehensiveness, instruction following, and readability. RACE is normalized to a Gemini Deep Research reference (50 denotes parity). FACT is the supported fraction of judgeable citations. All RACE rows summarize performance across the benchmark's full set of 100 research requests.

Method	Comp.	Insight	Instr.	Read.	Overall	FACT
Single Agent + retrieval	46.16	46.84	49.84	51.00	47.91	0.127
STORM	37.49	37.86	39.67	39.15	38.36	0.564
Co-STORM	42.87	45.09	44.92	41.07	43.90	0.564
Open Deep Research	49.46	49.96	51.41	**51.41**	50.36	0.432
AgentAgora	**51.45**	**52.33**	**52.42**	49.60	**51.71**	**0.646**

AgentAgora leads on overall report quality and three content-oriented RACE dimensions (Table 3). AgentAgora improves RACE Overall over Open Deep Research by 1.36 points (95% CI $[0.95, 1.75]$), with advantages in comprehensiveness, insight, and instruction following. These dimensions align with the intended role of coverage coordination: developing complementary evidence and addressing more aspects of the request. Open Deep Research retains the highest readability score, while AgentAgora's advantage centers on evidence-rich content and fulfillment of the request.

AgentAgora also attains the highest FACT score, supporting the quality of its cited evidence rather than citation volume alone. FACT complements RACE by measuring source support for cited statements. These exploratory results evaluate the full pipeline, including its section-wise writer; component controls below examine attribution, and Appendix Figure 8 reports citation-volume diagnostics.

4.5 ABLATION STUDIES AND PROCESS ANALYSIS

We examine the chair bundle's effects on outputs and search, then test DRB-specific components.

Chair-bundle ablation. The chair's clearest contribution is to differentiation within the delivered idea set. Removing coverage mapping, territory assignment, recruitment, and chair-controlled stopping lowers Overall by 0.235 (95% CI $[-0.299, -0.179]$), with the largest change in within-set distinctness (Figure 4). The largest benefit is set-level complementarity, while individual-idea originality remains slightly higher without the chair. The full bundle uses 118k discussion input tokens per task, compared with 86k without the chair, giving the quality–cost profile in Figure 4.

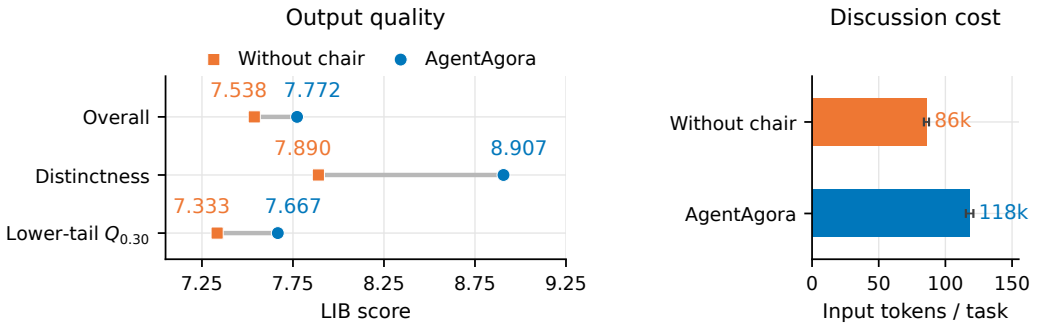


Figure 4: LIB-300 chair-bundle quality–cost comparison. Quality metrics share the 1–10 LIB scale; token cost is shown separately. Lower-tail $Q_{0.30}$ is the 30th percentile of task-level Overall.

Round-wise collapse. Initially similar trajectories diverge as the board grows (Figure 5): coordination sustains differentiated proposals, more new admissions, and less similarity rejection. Together, the traces show discussion continuing to expand the board with distinct contributions as the chair

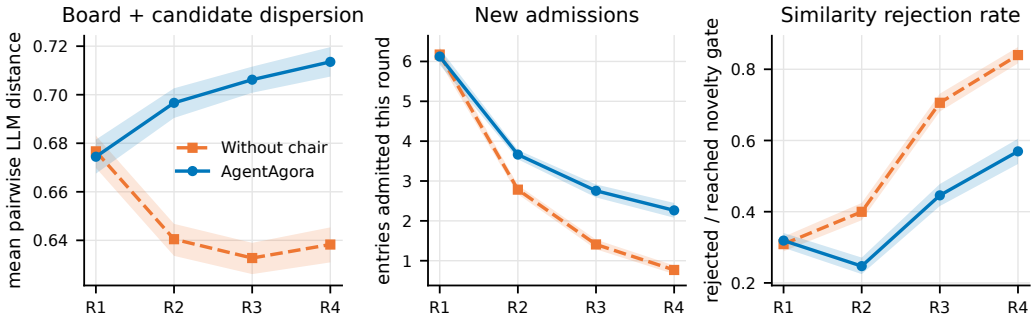


Figure 5: LIB-300 dynamics (task-bootstrap 95% CIs). R1–R4 have 300, 300, 294, and 224 jointly active tasks; rejection uses 299, 299, 292, and 224 pairs, excluding tasks where either configuration has no candidates reaching the novelty filter. Metric and averaging details: Appendix E.

revises exploration territories. These complementary signals of continued exploration also appear in the 224-task subset where both configurations complete four rounds (Appendix B).

Cross-task component checks. Holding the writer fixed, the chair adds 0.83 RACE points on 20 DRB development requests; the fixed-five-round comparison shows a smaller, statistically inconclusive gain at a higher token budget. These exploratory controls highlight how coordination interacts with writing and stopping (Appendix Figure 6).

5 Discussion and Limitations

Improving LLM creativity extends beyond any single interaction protocol. AgentAgora explores the collaborative aspect: helping different perspectives continue to contribute complementary ideas. This focus leaves room for combination with creativity-oriented training or more diverse candidate generation, opening the question of how these approaches and coverage-guided coordination can reinforce one another throughout candidate generation, discussion, and final selection.

Scientific and engineering workflows offer a next step, connecting proposals to practical discovery through expert assessment and execution. Current discussion budgets differ, and DRB includes development requests; equal-compute, independently held-out evaluations would strengthen attribution to coordination. Future work can further examine how exploration diversity translates into final output quality through selection and synthesis across repeated runs.

A separate chair introduces a potential dependency: its ability to recognize gaps and allocate effort may affect the entire discussion. Yet the nominally weaker GPT-4.1-mini chair improves LIB end-task performance (Appendix Table 7). Although compute also differs, this finding suggests that facilitation may not track conventional model capability, motivating research into how coordination abilities should be assessed and developed.

6 Conclusion

We introduced AgentAgora for sustained creative exploration through multi-agent discussion. Rather than relying on initial perspective diversity alone, a chair assigns complementary territories from the evolving shared record, directing participants toward new possibilities as their collective work develops through successive rounds of contribution and reassignment.

Experiments connect broader exploration with improved outputs. On CB, the candidate pool contains the correct answer on 48.37% of tasks, versus 32.45% for 30-sample self-consistency. Together with the ideation and synthesis results, this illustrates the value of coordinating where diverse perspectives are applied. Shared progress guides subsequent exploration, while participants retain responsibility for developing the actual solutions. AgentAgora provides a practical foundation for extending this principle to real-world creative problem solving and discovery.

AI USE STATEMENT

The authors determined the research questions, method direction, benchmark selection, and experimental decisions. Under author direction, Claude Code (Anthropic) assisted with framework and adapter implementation, experiment execution and monitoring, aggregation and statistical analysis, and portions of analysis documentation. OpenAI Codex assisted with manuscript revision, LaTeX editing, and numerical and reporting consistency checks. The authors reviewed the assisted design decisions, analyses, and text and remain responsible for the work. Language models used as generators or judges in the experiments are research objects and are reported separately in the experimental protocol.

ETHICS STATEMENT

The study uses public benchmark data and includes no human evaluation or human participants. DRB retrieval queries a commercial search API and fetches publicly accessible webpages. Generated reports and scientific proposals may contain unsupported, inaccurate, or unsafe claims; citation formatting and automatic support judgments do not replace expert or experimental validation. Any public redistribution of benchmark-derived records will be subject to the applicable dataset licenses.

REPRODUCIBILITY STATEMENT

Appendix D specifies the prompt interfaces, schemas, fallback rules, and task adapters. Appendix A describes sampling parameters, benchmark construction, evaluation procedures, and baseline implementation details. We will publicly release the complete code to support reproducibility and further research.

REFERENCES

B. An, B. Li, B. Wang, B. Zhang, B. L. Wang, C. Feng, C. Wei, C. Xue, C. Zhang, D. Ng, et al. Apodex 1.1: Scaling agentic intelligence for complex work. *arXiv preprint arXiv:2608.23283*, 2026. doi: 10.48550/arXiv.2608.23283. URL https://arxiv.org/abs/2608.23283.

Jinheon Baek, Sujay Kumar Jauhar, Silviu Cucerzan, and Sung Ju Hwang. ResearchAgent: Iterative research idea generation over scientific literature with large language models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 6709–6738, 2025. doi: 10.18653/v1/2025.naacl-long.342. URL https://aclanthology.org/2025.naacl-long.342/.

Nuo Chen, Yicheng Tong, Yuzhe Yang, Yufei He, Xueyi Zhang, Zou Qingyun, Qian Wang, and Bingsheng He. Diversity collapse in multi-agent LLM systems: Structural coupling and collective failure in open-ended idea generation. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 251–306, 2026. doi: 10.18653/v1/2026.findings-acl.13. URL https://aclanthology.org/2026.findings-acl.13/.

Weize Chen, Yusheng Su, Jingwei Zuo, Cheng Yang, Chenfei Yuan, Chi-Min Chan, Heyang Yu, Yaxi Lu, Yi-Hsin Hung, Chen Qian, Yujia Qin, Xin Cong, Ruobing Xie, Zhiyuan Liu, Maosong Sun, and Jie Zhou. AgentVerse: Facilitating multi-agent collaboration and exploring emergent behaviors. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=EHg5GDnyq1.

Anil R. Doshi and Oliver P. Hauser. Generative AI enhances individual creativity but reduces the collective diversity of novel content. *Science Advances*, 10(28):eadn5290, 2024. doi: 10.1126/sciadv.adn5290.

Mingxuan Du, Benfeng Xu, Chiwei Zhu, Xiaorui Wang, and Zhendong Mao. DeepResearch Bench: A comprehensive benchmark for deep research agents. arXiv preprint arXiv:2506.11763, 2025. URL https://arxiv.org/abs/2506.11763.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B. Tenenbaum, and Igor Mordatch. Improving factuality and reasoning in language models through multiagent debate. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 11733–11763. PMLR, 2024. URL https://proceedings.mlr.press/v235/du24e.html.

Bradley Efron and Robert J. Tibshirani. *An Introduction to the Bootstrap*. Chapman and Hall, 1993.

Lee D. Erman, Frederick Hayes-Roth, Victor R. Lesser, and D. Raj Reddy. The Hearsay-II speech-understanding system: Integrating knowledge to resolve uncertainty. *ACM Computing Surveys*, 12(2):213–253, 1980. doi: 10.1145/356810.356816.

Juraj Gottweis, Wei-Hung Weng, Alexander Daryin, Tao Tu, Petar Sirkovic, et al. Accelerating scientific discovery with Co-Scientist. *Nature*, 655:487–496, 2026. doi: 10.1038/s41586-026-10644-y.

Varun Gumma, Navonil Majumder, Soumitra Sinhahajari, and Soujanya Poria. IDEAgent: Agentic quality-diversity search for research idea generation. arXiv preprint arXiv:2607.22375, 2026. URL https://arxiv.org/abs/2607.22375.

Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70, 1979. URL https://www.jstor.org/stable/4615733.

Sirui Hong, Mingchen Zhuge, Jiaqi Chen, Xiawu Zheng, Yuheng Cheng, Ceyao Zhang, Jinlin Wang, Zili Wang, Steven Ka Shing Yau, Zijuan Lin, Liyang Zhou, Chenyu Ran, Lingfeng Xiao, Chenglin Wu, and Jürgen Schmidhuber. MetaGPT: Meta programming for a multi-agent collaborative framework. arXiv preprint arXiv:2308.00352, 2023. URL https://arxiv.org/abs/2308.00352.

Liwei Jiang, Yuanjun Chai, Margaret Li, Mickel Liu, Raymond Fok, Nouha Dziri, Yulia Tsvetkov, Maarten Sap, Alon Albalak, and Yejin Choi. Artificial hivemind: The open-ended homogeneity of language models (and beyond). arXiv preprint arXiv:2510.22954, 2025. URL https://arxiv.org/abs/2510.22954.

Yucheng Jiang, Yijia Shao, Dekun Ma, Sina Semnani, and Monica Lam. Into the unknown unknowns: Engaged human learning through participation in language model agent conversations. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 9917–9955, 2024. doi: 10.18653/v1/2024.emnlp-main.554. URL https://aclanthology.org/2024.emnlp-main.554/.

Robert Kirk, Ishita Mediratta, Christoforos Nalmpantis, Jelena Luketina, Eric Hambro, Edward Grefenstette, and Roberta Raileanu. Understanding the effects of RLHF on LLM generalisation and diversity. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=PXD3FAVHJT.

LangChain. Open deep research. Open-source software, 2025. URL https://github.com/langchain-ai/open_deep_research. Official implementation and architecture documentation; accessed September 22, 2026.

Guohao Li, Hasan Abed Al Kader Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem. CAMEL: Communicative agents for "mind" exploration of large language model society. In *Advances in Neural Information Processing Systems*, volume 36, pp. 51991–52008, 2023. doi: 10.52202/075280-2264.

Zhenghao Li, Zhi Zheng, Wei Chen, Jielun Zhao, Yong Chen, Tong Xu, and Enhong Chen. DynaDebate: Breaking homogeneity in multi-agent debate with dynamic path generation. arXiv preprint arXiv:2601.05746, 2026. URL https://arxiv.org/abs/2601.05746.

Tian Liang, Zhiwei He, Wenxiang Jiao, Xing Wang, Yan Wang, Rui Wang, Yujiu Yang, Shuming Shi, and Zhaopeng Tu. Encouraging divergent thinking in large language models through multi-agent debate. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 17889–17904, 2024. doi: 10.18653/v1/2024.emnlp-main.992. URL https://aclanthology.org/2024.emnlp-main.992/.

Chris Lu, Cong Lu, Robert Tjarko Lange, Jakob Foerster, Jeff Clune, and David Ha. The AI Scientist: Towards fully automated open-ended scientific discovery. *arXiv preprint arXiv:2408.06292*, 2024a. doi: 10.48550/arXiv.2408.06292. URL https://arxiv.org/abs/2408.06292.

Li-Chun Lu, Shou-Jen Chen, Tsung-Min Pai, Chan-Hung Yu, Hung-yi Lee, and Shao-Hua Sun. LLM discussion: Enhancing the creativity of large language models via discussion framework and role-play. In *Conference on Language Modeling*, 2024b. doi: 10.48550/arXiv.2405.06373. URL https://arxiv.org/abs/2405.06373.

Tien Anh Nguyen, Khanh-Binh Nguyen, Van Dai Do, Svetha Venkatesh, and Hung Le. Creative generation via multi-agent debate: Does debate suppress diversity? arXiv preprint arXiv:2609.00683, 2026. URL https://arxiv.org/abs/2609.00683. Accepted to EMNLP 2026 Main Conference, as reported on arXiv.

H. Penny Nii. The blackboard model of problem solving and the evolution of blackboard architectures. *AI Magazine*, 7(2):38–53, 1986. doi: 10.1609/aimag.v7i2.537. URL https://ojs.aaai.org/aimagazine/index.php/aimagazine/article/view/537.

Vishakh Padmakumar and He He. Does writing with language models reduce content diversity? In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=Feiz5HtCD0.

Arjun Panickssery, Samuel R. Bowman, and Shi Feng. LLM evaluators recognize and favor their own generations. In *Advances in Neural Information Processing Systems*, volume 37, pp. 68772–68802, 2024. doi: 10.52202/079017-2197.

Cheng Qian, Hyeonjeong Ha, Jiayu Liu, Jeonghwan Kim, Jiateng Liu, Bingxuan Li, Aditi Tiwari, Dwip Dalal, Zhenhailong Wang, Xiusi Chen, Mahdi Namazifar, Yunzhu Li, and Heng Ji. CreativityBench: Evaluating agent creative reasoning via affordance-based tool repurposing. arXiv preprint arXiv:2605.02910, 2026. URL https://arxiv.org/abs/2605.02910.

Kai Ruan, Xuan Wang, Jixiang Hong, Peng Wang, Yang Liu, and Hao Sun. Evaluating LLMs' divergent thinking capabilities for scientific idea generation with minimal context. *Nature Communications*, 17:3625, 2026. doi: 10.1038/s41467-026-70245-1. URL https://www.nature.com/articles/s41467-026-70245-1.

Mark A. Runco and Garrett J. Jaeger. The standard definition of creativity. *Creativity Research Journal*, 24(1):92–96, 2012. doi: 10.1080/10400419.2012.650092.

Yijia Shao, Yucheng Jiang, Theodore Kanell, Peter Xu, Omar Khattab, and Monica Lam. Assisting in writing Wikipedia-like articles from scratch with large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 6252–6278, 2024. doi: 10.18653/v1/2024.naacl-long.347. URL https://aclanthology.org/2024.naacl-long.347/.

Chenglei Si, Diyi Yang, and Tatsunori Hashimoto. Can LLMs generate novel research ideas? a large-scale human study with 100+ NLP researchers, 2024. URL https://arxiv.org/abs/2409.04109.

Haoyang Su, Renqi Chen, Shixiang Tang, Zhenfei Yin, Xinzhe Zheng, Jinzhe Li, Biqing Qi, Qi Wu, Hui Li, Wanli Ouyang, Philip Torr, Bowen Zhou, and Nanqing Dong. Many heads are better than one: Improved scientific idea generation by a LLM-based multi-agent system. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 28201–28240, 2025. doi: 10.18653/v1/2025.acl-long.1368. URL https://aclanthology.org/2025.acl-long.1368/.

Yufei Tian, Abhilasha Ravichander, Lianhui Qin, Ronan Le Bras, Raja Marjieh, Nanyun Peng, Yejin Choi, Thomas Griffiths, and Faeze Brahman. MacGyver: Are large language models creative problem solvers? In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 5303–5324, 2024. doi: 10.18653/v1/2024.naacl-long.297. URL https://aclanthology.org/2024.naacl-long.297/.

Keisuke Ueda, Wataru Hirota, Takuto Asakura, Takahiro Omi, Kosuke Takahashi, Kosuke Arima, and Tatsuya Ishigaki. Exploring the design of multi-agent LLM dialogues for research ideation. In *Proceedings of the 26th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pp. 322–337, 2025. URL https://aclanthology.org/2025.sigdial-1.26/.

Pat Verga, Sebastian Hofstatter, Sophia Althammer, Yixuan Su, Aleksandra Piktus, Arkady Arkhangorodsky, Minjie Xu, Naomi White, and Patrick Lewis. Replacing judges with juries: Evaluating LLM generations with a panel of diverse models. *arXiv preprint arXiv:2404.18796*, 2024. doi: 10.48550/arXiv.2404.18796. URL https://arxiv.org/abs/2404.18796.

Manya Wadhwa, Tiasa Singha Roy, Harvey Lederman, Junyi Jessy Li, and Greg Durrett. CREATE: Testing LLMs for associative creativity. arXiv preprint arXiv:2603.09970, 2026. URL https://arxiv.org/abs/2603.09970.

Qingyun Wang, Doug Downey, Heng Ji, and Tom Hope. SciMON: Scientific inspiration machines optimized for novelty. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 279–299, 2024. doi: 10.18653/v1/2024.acl-long.18. URL https://aclanthology.org/2024.acl-long.18/.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *International Conference on Learning Representations*, 2023. doi: 10.48550/arXiv.2203.11171. URL https://arxiv.org/abs/2203.11171.

Emily Wenger and Yoed Kenett. We're different, we're the same: Creative homogeneity across LLMs. *arXiv preprint arXiv:2501.19361*, 2025. doi: 10.48550/arXiv.2501.19361. URL https://arxiv.org/abs/2501.19361.

Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Beibin Li, Erkang Zhu, Li Jiang, Xiaoyun Zhang, Shaokun Zhang, Jiale Liu, Ahmed Hassan Awadallah, Ryen W. White, Doug Burger, and Chi Wang. AutoGen: Enabling next-gen LLM applications via multi-agent conversation. arXiv preprint arXiv:2308.08155, 2023. URL https://arxiv.org/abs/2308.08155.

Longfei Yun, Chenyang An, Zilong Wang, Letian Peng, and Jingbo Shang. The price of format: Diversity collapse in LLMs. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 15454–15468, 2025. doi: 10.18653/v1/2025.findings-emnlp.836. URL https://aclanthology.org/2025.findings-emnlp.836/.

Jiayi Zhang, Simon Yu, Derek Chong, Anthony Sicilia, Michael R. Tomz, Christopher D. Manning, and Weiyan Shi. Verbalized sampling: How to mitigate mode collapse and unlock LLM diversity. arXiv preprint arXiv:2510.01171, 2025a. URL https://arxiv.org/abs/2510.01171.

Yiming Zhang, Harshita Diddee, Susan Holm, Hanchen Liu, Xinyue Liu, Vinay Samuel, Barry Wang, and Daphne Ippolito. NoveltyBench: Evaluating language models for humanlike diversity. arXiv preprint arXiv:2504.05228, 2025b. URL https://arxiv.org/abs/2504.05228.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. In *Advances in Neural Information Processing Systems*, volume 36, pp. 46595–46623, 2023. doi: 10.52202/075280-2020.

A EVALUATION DETAILS

Statistical comparisons. For reported differences, bootstrap resampling (Efron & Tibshirani, 1993) preserves task pairs so that both methods are evaluated on the same resampled tasks. Two-sided exact sign tests assess whether task-level wins and losses are equally likely, excluding ties. For the two primary LIB Overall comparisons under the default judge (AgentAgora versus Single Agent and versus the no-chair variant), Holm correction (Holm, 1979) adjusts the two p-values to control the risk of any false-positive rejection within this pair of tests. Both remain significant at the 0.05 level. This correction applies to these two tests, not to all analyses or the reported bootstrap intervals; additional diagnostics provide exploratory analyses of the system's components and discussion process.

Experimental configuration. Seats and the default chair use GPT-5.4-mini with reasoning disabled. Runs start with first-principles, analogy, contrarian, engineering, and systems perspectives. The chair may add one seat per round up to eight, but cannot deactivate a seat. After reconnaissance, active seats receive the public board, chair comments, retired directions, their own previous answer, and an individual territory. Chair-directed stopping is allowed after one assigned round. Execution can also terminate at the round cap or under an adapter-specific saturation rule, such as a sufficient pool of validated candidates.

Table 4: Task-adapter settings and requested final outputs. Initial-round proposals initialize the board during reconnaissance; in each subsequent round, every active seat develops one candidate.

Field	LIB	CB	DRB
Maximum rounds	4	3	3
Initial-round proposals per seat	2	3	1
Later-round proposals per seat	1	1	1
Requested final output	3 ideas	1 answer	1 report
Adaptive / public / chair / allocation	yes	yes	yes
Active comments	GAP, DEAD_END, FORK (all three adapters)		
Diversity gate / novelty threshold	yes / .55	no / .55	no / .55
Quality–distance blend	.5	0	0
Seat temperature / maximum tokens	.95 / 8k	.9 / 24k	.7 / 16k
Tools	no	no	search + fetch

LIB split and scoring. The official classification file contains 1,180 keywords in 22 disciplines. We exclude 349 official keywords consulted during development, allocate 300 slots over the remaining 831 keywords by largest remainder, and sample within discipline using seed 20260922. Each system is requested to return three ideas of at most 100 words. GPT-4.1-mini scores originality, feasibility, and clarity from 1–10. The official pairwise fluency rubric is repurposed within each three-idea output and mapped FAR/MODERATE/NEAR/DUPLICATE to 10/7/4/1. We call the resulting measure *within-set distinctness (adapted fluency)*. Overall is calculated for each task and then averaged. All reported LIB comparisons use this shared adapted protocol; the official leaderboard uses a separate multi-model judging protocol.

MAD produces only one idea on 10 of 300 tasks, so distinctness is undefined there and its task-level Overall averages the remaining three dimensions. Both aggregate scores and paired comparisons use this task-level definition.

DRB protocol. Agentic research also encompasses experimental workflows, as in The AI Scientist (Lu et al., 2024a); our DRB evaluation targets evidence-grounded report synthesis.

The benchmark has 100 requests (50 Chinese, 50 English) across 22 domains. Requests 1–20 informed tuning and 21–50 held-out development checks; the full set is used for exploratory system-level evaluation. Every system receives the official prompt plus common numbered-citation formatting. Official RACE scoring uses GPT-5.5, frozen per-request weights, and normalization to Gemini Deep Research reference reports. FACT extracts and deduplicates sentence–citation pairs, fetches the cited pages through Jina, and judges whether each page supports its sentence. Its official summary averages over reports containing citations. STORM and Co-STORM use their official Python

implementation, `knowledge-storm` version 1.1.1; Open Deep Research uses commit `1b7d2e8`. All systems share Serper and Jina Reader. STORM's default 700-token section limit remains in force, and interaction and writing budgets are not equalized.

CB protocol. We evaluate all 3,285 released tasks with GPT-5.4-mini as the generator and the official rule-based match on entity and part. The final readout uses fork commitments, distinct-seat support, verifier rank, and then the first-proposed candidate as the last tie-break. We shuffle scene entities separately for each seat when testing AgentAgora. No rubric LLM scores CB correctness. All rates use 3,285 tasks as the denominator, with missing outputs counted as incorrect. The 95% CIs use percentile bootstrap quantiles computed from the binary outcome totals.

Coverage@pool is one when any distinct candidate produced by a method matches both gold fields. For AgentAgora, the pool is the final board; candidates rejected before admission were not retained and cannot be counted. For Self-Consistency@30, the pool consists of the distinct candidates among 30 samples. For MAD and LLM Discussion, the pool comprises candidate picks extracted from agent utterances; 267 utterances (about 0.4%) with no extracted pick contribute no candidate. Creative-MAD's official runner stores only the final response, so its reported pool size is one and Coverage@pool equals Gold Accuracy. Coverage@pool therefore refers to the method-specific recorded pools defined above, evaluated against the same gold entity–part pairs.

Table 5: Exact CreativityBench values visualized in Figure 3. Gold Accuracy requires both the selected entity and its enabling part to match the benchmark answer.

Method	Gold Acc.	95% CI	Coverage@pool
Single Agent	0.2164	[0.2024, 0.2307]	0.2164
Self-Consistency@30	0.2225	[0.2085, 0.2368]	0.3245
Creative-MAD	0.2268	[0.2125, 0.2411]	0.2268
LLM Discussion	0.2350	[0.2207, 0.2496]	0.3184
MAD	0.2444	[0.2298, 0.2594]	0.3084
AgentAgora	**0.2688**	[0.2536, 0.2840]	**0.4837**

LIB judge agreement. Work on LLM evaluation documents judge-specific biases (Zheng et al., 2023; Panickssery et al., 2024) and explores panels of diverse evaluators (Verga et al., 2024). These findings motivate examining agreement across judges; here, each judge scores independently and is reported separately rather than combined into a panel score.

Table 6 reports results for the four reported judges. Within AgentAgora, task-level judge correlations range from $r = .32$ to $.56$, indicating that agreement on its leading aggregate performance coexists with variation in task-level assessments. We report these within-method correlations to distinguish agreement on task difficulty from the shared variance contributed by between-system score differences. All methods are evaluated on the same 300 LIB tasks.

Table 6: LIB Overall under the four reported judges on LIB-300. Bold marks the best system within each judge. The final column is the paired AgentAgora minus Single Agent difference with a bootstrap 95% CI. Judge-specific scales should not be compared across rows.

Judge	AgentAgora	Single	LLM Disc.	MAD	Δ vs. Single
GPT-4.1-mini	**7.772**	7.644	7.612	7.454	$+.128$ $[+.083, +.172]$
GPT-5.5	**7.286**	7.012	6.928	6.878	$+.274$ $[+.224, +.326]$
Gemini-2.5-Pro	**8.081**	7.664	7.317	7.523	$+.418$ $[+.348, +.487]$
o3-mini	**7.973**	7.921	7.863	7.756	$+.052$ $[+.003, +.100]$

B EXPLORATORY COMPONENTS AND RESOURCES

Table 7: Detailed LIB-300 component results under the GPT-4.1-mini judge. Tokens are mean discussion input tokens per task and exclude offline LIB judging.

Configuration	Orig.	Feas.	Clar.	Distinct.	Overall	$Q_{0.30}$	Tokens
Single Agent (reference)	8.237	6.524	**7.139**	8.677	7.644	7.500	–
GPT-4.1-mini chair	**8.606**	**6.829**	7.116	**9.190**	**7.935**	**7.750**	176k
AgentAgora (default)	8.412	6.769	7.002	8.907	7.772	7.667	118k
Without chair bundle	8.549	6.692	7.042	7.890	7.538	7.333	86k

Component analyses on 20 development requests distinguish the contributions of report construction and chair coordination. The sectioned writer contributes most of the observed gain. Holding it fixed, the three-round chair comparison has 16 wins and 4 losses ($p = .012$), showing an additional benefit from coordination within this setting. The fixed-five-round comparison yields a smaller, statistically inconclusive gain (53.12 versus 52.36). Together, these comparisons motivate examining coordination jointly with report construction and the discussion budget.

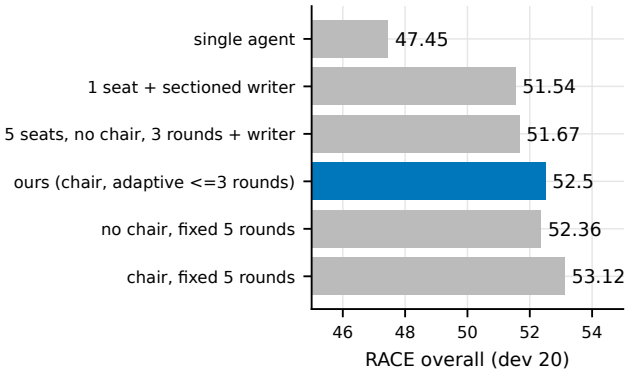


Figure 6: DRB component ablation on 20 development requests. The sectioned writer accounts for most of the observed development-set gain; holding it fixed, the adaptive chair adds 0.83 RACE points. Adaptive and fixed-round comparisons are exploratory tuning-set analyses.

Default and no-chair LIB runs use 20.4 and 19.1 seat turns, 276 and 234 total model calls, 118k and 86k input tokens, and 6.6k and 4.7k output tokens per task, respectively. The difference mainly reflects chair calls and the additional assignment/comment context rather than many more seat turns. Active-task counts by round are 300/300/294/274 for default and 300/300/300/245 for no chair.

C ADDITIONAL CREATIVITYBENCH DIAGNOSTICS

AgentAgora's advantage lies in identifying the complete entity–part solution measured by Gold Accuracy. MAD and LLM Discussion obtain 0.5890 and 0.5632 Entity Correct, while AgentAgora obtains 0.5403. These entity-only scores complement the primary Gold Accuracy metric, which additionally requires the enabling part to be correct. Among tasks where the AgentAgora board contains the gold pair, the readout selects it 55.6% of the time, compared with 79.3% for MAD and 73.8% for LLM Discussion. These conditional rates describe selection within each method's own candidate pool, relating the available alternatives to its final decision.

D PROMPTS, SCHEMAS, AND DETERMINISTIC RULES

The following specification describes the implemented chair and seat interfaces, including their structured fields, behavioral constraints, and task-specific admission and selection rules.

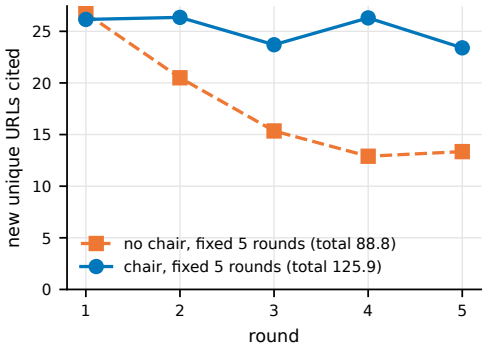


Figure 7: New unique URLs per round on 20 DRB development requests under a fixed-five-round comparison. The chair condition uses 6.7M input tokens per task versus 5.1M without the chair and collects 125.9 versus 88.8 distinct URLs in total. This exploratory analysis uses the same development requests to compare evidence collection across successive discussion rounds.

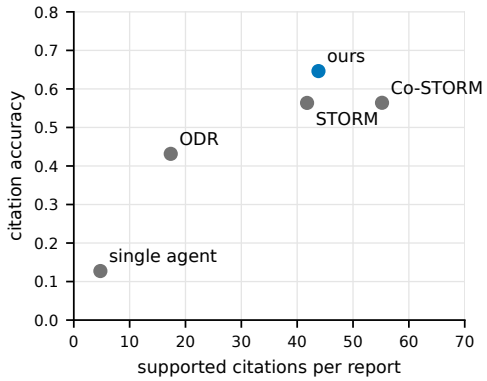


Figure 8: Official DRB FACT results. Citation accuracy is the supported fraction of judgeable sentence–citation pairs; supported-citation counts are per report among reports containing citations. AgentAgora has the highest accuracy, whereas Co-STORM has the largest supported count.

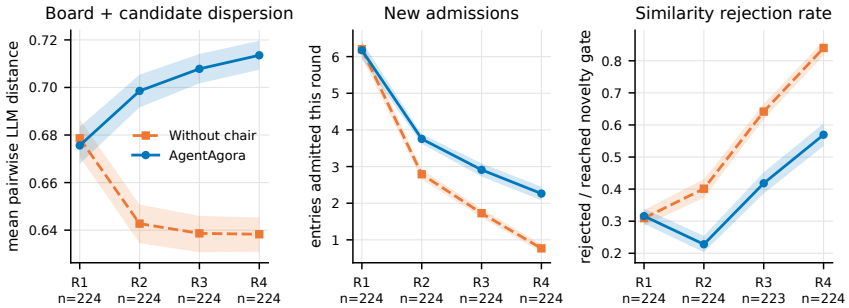


Figure 9: LIB trajectory sensitivity analysis restricted to the 224 tasks for which both default and no-chair runs reach round 4. The qualitative divergence matches Figure 5. Inclusion in this sensitivity subset is determined by both runs reaching the final round.

D.1 Chair and seat interfaces

The chair is instructed to coordinate exploration without solving the task or selecting an answer. It comments on the board using `GAP`, `DEAD_END`, and `FORK`, proposes a distinct territory for each active seat, and may recruit one additional perspective when a gap requires it. Its structured response contains the coverage map, comments, territory assignments, optional recruitment, and stopping decision:

```
{"reading": "<one-sentence board state>",
 "problem_map": [{"problem": "<problem addressed>",
                  "entries": ["<entry id>", ...]}],
 "comments": [{"kind": "GAP|DEAD_END|FORK",
               "targets": ["<entry id>", ...],
               "severity": "minor|moderate|severe",
               "text": "<actionable comment>"}],
 "assignments": [{"seat": "<exact name>",
                  "area": "<complementary territory>"}],
 "new_seats": [{"name": "<name>", "lens": "<instruction>",
                "why": "<missing contribution>"}],
 "stop": false}
```

The chair must map every valid board entry by the problem addressed, not the technique used, before assigning uncovered and mutually distinct areas. A territory is framed to the seat as: "Your problem area: *area*. Work on a meaningful problem inside this area. Which problem, and how to solve it, is yours to decide." The public seat message orders the task, blackboard, optional curator notes, chair comments, retired directions, the seat's prior reply, its private assignment, and the output contract. Direct messaging and access to other seats' raw replies are absent: shared information travels through the admitted board.

Territories are accepted only if they contain 2–16 whitespace-separated words, contain no question mark, and have content-word Jaccard overlap below 0.5 with earlier accepted territories in that round. The prompt requests 3–12 words, a narrower soft instruction. Rejected territories are neither repaired nor retried: the seat still receives public context and self-selects a direction.

If chair JSON parsing fails, a deterministic fallback rule sends the same message to every seat. For a collapsed or starving round with at least two board entries, it issues a `GAP` away from the most distant pair; for a quality-bound round, it issues a `PRESS` requesting a checkable value, mechanism, or component; otherwise it issues a generic `GAP` away from the obvious direction. It creates no new seat, makes no differentiated territory allocation, and does not stop.

D.2 Task adapters

These adapters specify how AgentAgora produces outputs for each task, separately from benchmark scoring. Benchmark evaluation protocols are described in Appendix A.

LIB. Seats propose self-contained ideas within 100 words. An internal evaluator checks quality; admission requires feasibility of at least 5/10 and semantic distance of at least 0.55 from the board (FAR/MODERATE/NEAR/DUPLICATE map to 1.0/.7/.4/.1). The final selector returns three ideas from distinct coverage groups, weighting normalized quality and distance equally. These internal checks are separate from benchmark judging.

CB. Seats propose scene entity–part pairs against a shared specification of required outcomes. Admission checks scene membership and explicit constraints; repeated nominations update support counts. Final selection prioritizes fork commitments, distinct-seat support, verifier rating, and earliest proposal, in that order. Support counts aggregate distinct participants' nominations within the shared-board discussion, recording which participants support each admitted candidate.

DRB. Seats investigate assigned evidence facets for up to 12 retrieval iterations and return cited memos. Admission requires at least two URLs; citation support is assessed separately by FACT. A 4–8-section outline covers the closing map, and a separate writer composes sections and a synthesis. Final processing standardizes citations and references; FACT separately assesses whether the retrieved source supports the statement associated with each citation.

E Supplementary diagnostics and illustrative cases

Dispersion definition. For each round, dispersion is the average pairwise semantic distance over the existing board plus the round's deduplicated candidates *before* novelty filtering. Fewer than two

candidates yields 1.0. This differs from admitted-board dispersion and is why round-4 similarity-rejection rates provide a complementary, directly pre-admission repetition measure. New admissions and source counts provide complementary operational measures of exploration.

Round-wise pairing and rejection rates. Figure 5 compares only tasks for which both configurations execute the given round. All 300 tasks have paired runs; early stopping reduces the active pairs to 300, 300, 294, and 224 at R1–R4, respectively, rather than indicating incomplete benchmark execution. For each task and configuration, the similarity-rejection rate is the number of candidates rejected for similarity divided by the number reaching the novelty gate, after earlier validity, quality, and exact-duplicate filters. When this denominator is zero, the rate is undefined, not zero. Rejection-rate comparisons exclude a task pair if either configuration has an undefined rate, leaving 299, 299, 292, and 224 valid pairs. We compute each rate per task and then average over these valid paired tasks, rather than pool candidate counts across tasks. The complete-pair sensitivity analysis uses the 224 tasks for which both configurations execute all four rounds.

Exploration and topical focus. On *philosophy of mind*, later assignments move from central philosophical questions toward adjacent cognitive-science topics. AgentAgora scores 7.000, compared with 8.417 for Single Agent and 6.917 without the chair. The trace illustrates how coverage can extend into adjacent topics, complementing the aggregate analysis with a task-level view of how assigned territories relate to exploration and topical focus.

Complementary exploration. On *toxicology*, complementary territories yield ideas on chemical-source attribution, adaptive stress tolerance, and microbiome mixture effects (Figure 10). Overall improves from 7.25 to 8.25 and distinctness from 7.0 to 10.0 relative to Single Agent. This selected case provides a concrete illustration of complementary exploration.

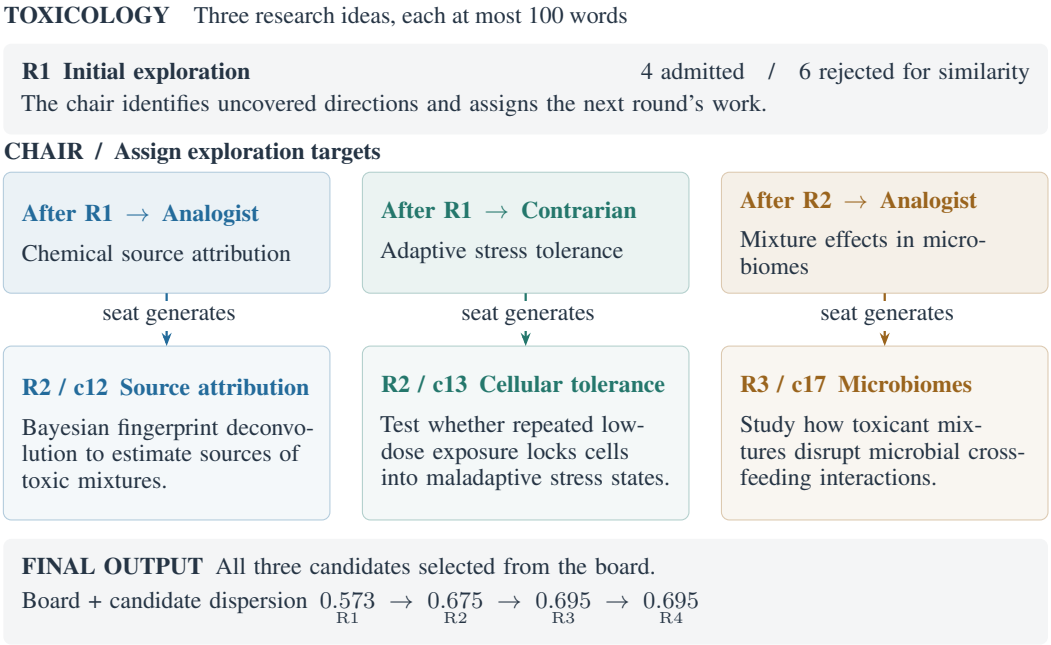


Figure 10: Condensed recorded trace for *toxicology*. Assignments after round t guide round $t + 1$. Color links territories to resulting candidates.