REINFORCING AGENTIC CREATIVITY IN SCIENTIFIC IDEATION WITH NIGHT SCIENCE

Anonymous authors
Paper under double-blind review

ABSTRACT

Large language models (LLMs) excel at structured, verifiable tasks, but their low-entropy bias can produce homogeneous and predictable outputs, limiting their utility for open-ended scientific ideation. Effective discovery, however, spans a broader creative spectrum: from structured *day science* to loosely structured, serendipitous *night science* that reaches ideas beyond those typically considered. We introduce **AI NIGHT-SCIENTIST**, an agentic framework that uses reinforcement learning to teach models *when* and *how* to depart from predictable reasoning. Grounded in cognitive science, we model creativity along three axes: *action* (what to do and how creatively), *process* (when to explore versus exploit), and *outcome* (the novelty and usefulness of the resulting idea). We use these axes to train models with GRPO, exposing them to varying degrees and forms of creativity throughout training. This produces substantially more diverse scientific proposals, expanding the range of research directions by 27.8% and contribution types by 14.9% over the base model. It also improves predicted citation impact by up to 32.0 percentage points and originality by 66.2 points. These gains cannot be reproduced by simply increasing decoding temperature; instead, we find that semantic guidance specifying *what kind* of creativity to pursue is critical. Overall, our results suggest that **creativity is a learnable, multi-level ability that can be shaped to help researchers reach ideas beyond those typically explored by LLMs**.

1 INTRODUCTION

Large language models (LLMs) have excelled at structured, systematic tasks with clear verifiability (e.g., coding and quantitative reasoning), where their performance is often improved with reinforcement learning (RL) (Guo et al., 2025; Wang et al., 2025; Zhong & Wang, 2024; Liu et al., 2024). This success is consistent with a broader tendency toward minimizing token entropy (Agarwal et al., 2025), where models favor high-probability outputs that reflect frequent patterns and expected answers in training data (McCoy et al., 2024). While LLMs have increasingly been applied to scientific ideation (Si et al., 2025; Gottweis et al., 2025), they lack originality (Zhao et al., 2025), tend to generate homogeneous outputs (Wenger & Kenett, 2025), and even plagiarize at nontrivial rates (Gupta & Pruthi, 2025). Ultimately, **this directly conflicts with the key attributes of open-ended, creative scientific discovery**: novelty, diversity, and serendipity.

Prior methods depict discovery as a highly structured process, involving hypotheses derived from prior observations and/or data, tested against evidence, and refined or rejected accordingly (Gottweis et al., 2025; Agarwal et al., 2026). But this is only *part* of the scientific discovery process. They typically view creativity as only an attribute of ideas rather than as part of the process, treating individual actions as fixed behaviors (e.g., search, write) and executing them through scaffolded pipelines (Gu et al., 2024; Lu et al., 2024). While this supports the critical reasoning crucial for validating existing ideas, this overlooks the creative reasoning necessary for discovering new ones. Both are complementary and crucial perspectives of scientific discovery, referred to as ***day*** and ***night science*** (Wechsler et al., 2018; Halpern, 2007; Stent, 1988; Yanai & Lercher, 2019).

Night science captures the often-neglected nature of human-driven discovery: loosely structured and dynamic exploration driven by highly creative actions, such as exploring distant analogies (e.g., biology-inspired technology), considering alternative perspectives after serendipitous encounters (e.g., debates with colleagues from other domains), and acting on partly-formalized intuitions (e.g.,

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.

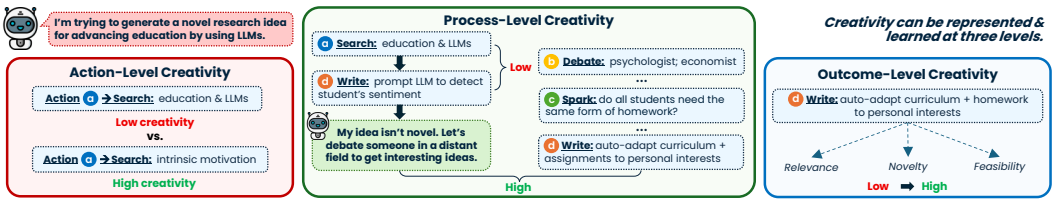

Figure 1: Given an input task, creativity can be injected into three different axes of reasoning: *action, process*, and *outcome*. Moreover, each level can be executed with *varying degrees of creativity*.

abandoning status quo assumptions). Together, day and night science form a spectrum that human researchers have smoothly traversed to uncover breakthroughs (de Chantal & Markovits, 2022; Dwyer et al., 2025), such as chemotherapy and penicillin (Hirsch, 2006; Ligon, 2004).

We hypothesize that LLMs can better traverse this spectrum if given explicit control over *when* and *how* to deviate from high-probability output, enabling targeted doses of night science while maintaining day science's goal-directed behavior. To test this, we introduce **AI NIGHT-SCIENTIST**, an RL-based agentic framework that incentivizes models to exercise this control effectively. Grounded in cognitive science (Lubart, 2001; Cohen, 1989; Dwyer et al., 2025; Harvey & Berry, 2023), it embeds creativity along three axes of reasoning (Figure 1):

- ***Action-level* (*how*):** Associates each action with a creativity level, allowing the same action to vary from conventional, high-probability behavior to more exploratory and unconventional behavior.
- ***Process-level* (*when*):** Controls when to shift between lower- and higher-creativity actions based on how the multi-step trajectory unfolds.
- ***Outcome-level* (*what*):** Captures the creativity of the resulting idea, favoring outputs that are novel while remaining relevant and useful.

We utilize this framework to train an LLM-based agent using GRPO (Shao et al., 2024) for generating *scientific research proposals*, where identifying promising ideas often requires long-horizon creative reasoning beyond immediately verifiable evidence. Overall, our work argues that LLM-based support for scientific discovery should span the full day-to-night science spectrum. AI NIGHT-SCIENTIST shows that conventional LLMs do not naturally navigate this spectrum effectively, but targeted reinforcement learning can reshape when and how they depart from structured, predictable reasoning, leading to more creative outcomes. Our contributions can be summarized as:

- We introduce **AI NIGHT-SCIENTIST**, an RL-based agentic framework that explicitly represents creativity across actions, reasoning processes, and outcomes for long-horizon scientific ideation.
- We show that creativity is more than sampling stochasticity: explicit guidance on *how* to be creative via RL helps an agent learn *when* creative deviations are useful, while higher temperature does not.
- Empirically, AI NIGHT-SCIENTIST produces more diverse scientific proposals than its base model, expanding the range of research directions by 27.8% and contribution types by 14.9%, while improving predicted impact by up to 32.0 percentage points and originality by 66.2 points.

2 BACKGROUND AND RELATED WORK

Scientific discovery has been characterized as an interplay between structured, hypothesis-driven *day science* and more exploratory, intuition- and serendipity-driven *night science* (Stent, 1988; Yanai & Lercher, 2019). Classic theories of creativity characterize creative thought through remote associations between otherwise distant concepts (Mednick, 1962), generative and exploratory modes of cognition (Ward et al., 1999), and outcomes that are both original and useful (Runco & Jaeger, 2012). Creativity can also vary in degree: prior work describes a continuum of creative behavior (Cohen, 1989), with problem-solving strategies ranging from paradigm-preserving to paradigm-stretching and paradigm-breaking (McFadzean, 1998). Related accounts distinguish between exploring existing conceptual spaces and transforming them to enable qualitatively new ideas (Boden, 1998), while theories of creative ideation show that originality can arise either by flexibly exploring many conceptual directions or by persistently exploring a few in greater depth (Nijstad et al., 2010). Together, these

perspectives motivate our view of creativity as varying both *where* it enters reasoning (within actions, processes, and outcomes) and *how strongly* it is expressed.

This view contrasts with most current LLM-based approaches to scientific discovery. Existing research agents broaden ideation through retrieval, search, multi-agent interaction, and iterative generation (Gu et al., 2024; Lu et al., 2024; Kargupta et al., 2025a; Gottweis et al., 2025), but generally treat actions such as searching, debating, and writing as fixed behaviors and primarily assess creativity in the resulting idea. This leaves little control over *how creatively* individual actions are performed or *when* creative deviations should occur throughout reasoning. Relatedly, Kargupta et al. (2025b) find that LLMs struggle with the metacognitive awareness needed to monitor and adapt their reasoning, further limiting their ability to flexibly shift between critical and creative modes.

Recent work has begun targeting the mechanisms that produce creative scientific ideas. O'Neill et al. (2025) use structured assumption inversion to generate novel hypotheses, inspiring the *spark* action in our framework, while Kargupta et al. (2025c) and Kargupta et al. (2026) use retrieval to identify research gaps and surface interdisciplinary inspiration. Other work instead learns scientific capabilities directly: GIANTS (He-Yueya et al., 2026) trains smaller models to anticipate scientific insights, while Tong et al. (2026) study whether models can learn scientific taste. Together, these approaches suggest that not only scientific outputs, but also the processes that produce them, can be shaped through structure and supervision. Reinforcement learning offers a way to shape this process without prescribing exactly how discovery should unfold, although open-ended ideation has no single correct answer and must balance qualities such as novelty, relevance, feasibility, and usefulness (Afzal et al., 2025). Moreover, useful creative exploration requires more than simply injecting randomness (Schmidhuber, 2010). Our work builds on these ideas by learning both *how creatively* individual actions should be performed and *when* different levels of creativity are useful across a reasoning trajectory.

3 AI NIGHT-SCIENTIST: A CREATIVITY-ALIGNED AGENTIC FRAMEWORK

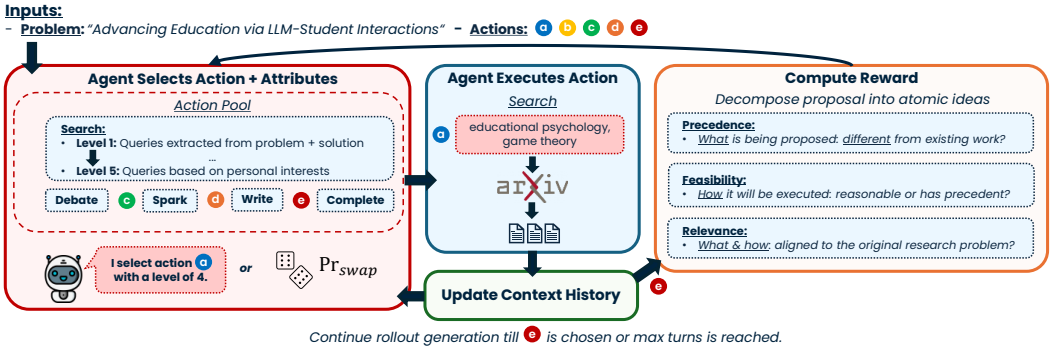

Figure 2: AI NIGHT-SCIENTIST consists of a: **(1)** rollout phase, where an LLM builds a reasoning trajectory f by iteratively selecting and executing actions with creativity levels, and **(2)** reward phase during training, where the reward is computed over the final proposal.

We propose **AI NIGHT-SCIENTIST**, as illustrated in Figure 2, which represents creativity directly within the agent's reasoning trajectory rather than only in its final output. We extend the ReAct-style reasoning setting (Yao et al., 2023): at each step, the agent selects an action together with a creativity level that specifies *how* that action should be carried out, executes it, and updates its current idea state. Repeating this process produces a trajectory through the space of possible ideas, allowing the agent to move between more familiar and more unexplored directions as reasoning unfolds. This lets us represent creativity at three levels (Figure 1): *action-level* creativity captures *how* an action is performed, *process-level* creativity captures *when* to shift between lower- and higher-creativity actions, and *outcome-level* creativity captures the novelty and usefulness of the resulting idea.

3.1 MULTI-LEVEL REPRESENTATION OF CREATIVE REASONING

We define each level below before applying the framework to scientific research proposal generation:

Definition 3.1 (*Action-Level Creativity*)**.** Let $\mathcal{A}$ denote the action space for a task $\mathcal{T}$. For each action $a \in \mathcal{A}$, we define an ordered set of creativity levels $\mathcal{C}_a = \{c_a^{(1)}, \ldots, c_a^{(K_a)}\}$, where each level gives a natural-language description of how a should be performed. Lower levels describe more conventional, high-probability behavior, while higher levels describe increasingly exploratory or unconventional behavior. The number of levels K_a may vary across actions.

Definition 3.2 (*Process-Level Creativity*)**.** Let $\tau = \big((a_1, c_1, \hat{o}_1), \ldots, (a_N, c_N, \hat{o}_N)\big)$ denote a reasoning trajectory, where $a_i \in \mathcal{A}$ is the action selected at step i, $c_i \in \mathcal{C}_{a_i}$ is its creativity level, and $\hat{o}_i$ is the resulting intermediate output. At each step, the agent selects the next action-level choice (a_i, c_i) based on the task $\mathcal{T}$ and the preceding trajectory $\tau_{1:i-1}$. *Process-level creativity* reflects how effectively the agent sequences and adapts these choices over time. Higher process-level creativity means varying the degree of action-level creativity to benefit the evolving reasoning process, rather than consistently favoring either low- or high-creativity behavior.

Definition 3.3 (*Outcome-Level Creativity*)**.** Let o denote the final outcome produced for a task $\mathcal{T}$. *Outcome-level creativity* captures the creativity of o itself, based on two complementary properties: its *novelty* $\mathcal{N}(o)$ and its *usefulness* $\mathcal{U}(o)$ (Harvey & Berry, 2023). Novelty measures how much o departs from existing or familiar solutions, while usefulness measures how valuable, appropriate, or effective it is for the task. An outcome should exhibit both in order to be considered creative.

Together, this multi-level representation separates *how* creativity is expressed within an action, *when* different degrees of creativity are useful across reasoning, and *what* the process ultimately produces. We represent action-level creativity in natural language to make these choices interpretable and give the user direct semantic control over how and to what extent each action may deviate from its conventional execution. The number and meaning of creativity levels can also vary by action.

These action-level choices accumulate into the reasoning trajectory. Intermediate outputs may not appear directly in the final result, but they can shape later decisions and influence when greater or lesser deviation is useful. Greater creativity may help open new directions, while lower-creativity behavior may be better suited for developing or refining promising ones. Making these choices well requires the model to monitor how its reasoning is progressing and adapt accordingly, a key aspect of metacognitive control that is lacking in existing models (Kargupta et al., 2025b).

3.2 Task Formulation: Research Proposal Generation

Given a high-level research problem p, the agent uses the action space $\mathcal{A}$ in Table 1 to explore and develop a long-horizon research proposal over a multi-step trajectory τ. SEARCH, DEBATE, and SPARK support multiple creativity levels, allowing the agent to vary how broadly it searches, whose perspectives it considers, and how strongly it challenges existing assumptions. WRITE consolidates the trajectory into proposal text, while STOP ends the process and returns the final proposal o. We keep WRITE fixed so that the final proposal primarily reflects the exploration that preceded it.

Action $\mathcal{A}$	Mechanism & Output $\hat{o}$	Creativity Levels $c \in \mathcal{C}_a$
a Search	Generate query → retrieve arXiv papers based on embedding similarity	*L1:* Search proposal-specific background using title and core terms. *L3:* Search tangentially related background for broader context. *L5:* Search distant domains, alternate perspectives, or broader questions.
b Debate	Select participants and topic → retrieve relevant papers → simulate discussion	*L1:* Discuss proposal specifics with a close-domain colleague. *L3:* Debate with a peer from the same field but a different topic. *L5:* Explore with experts from distant disciplines in an open-ended debate.
c Spark	Identify assumption (*Bit*) → invert it (*Flip*) → reframe it (*Spark*)	*L1:* Challenge a narrow assumption specific to the current proposal. *L3:* Challenge a meaningful assumption underlying the approach. *L5:* Challenge a broad, field-level assumption through radical reframing.
d Write	Synthesize prior trajectory into a proposal draft or revision	*Fixed:* consolidate prior trajectory; preserve credit assignment
e Stop	End trajectory & return final proposal	Judge final proposal by $\mathcal{P}$ (precedence), $\mathcal{F}$ (feasibility), and $\mathcal{R}$ (relevance)

Table 1: Action space for research proposal generation with low, mid and high creativity levels shown. Table 10 in Appendix F.1 includes all levels; full action prompts provided in Appendix G.

We focus on proposals with a scope similar to long-term research grants, where ideas are intended to guide work over several years rather than describe an immediately executable experiment. This makes the setting well suited to studying creative reasoning: proposals must remain grounded in

existing work, yet many of their central ideas cannot be directly verified at inference time because the required experiments or data may not yet exist. The task therefore rewards reasoning that can move beyond established directions while still producing ideas that are feasible and relevant to p.

3.3 Reinforcing Adaptive Creativity

To *teach* models *when* and *how* different degrees of creativity are useful, we leverage reinforcement learning. RL allows the quality of the final idea to shape the reasoning process without prescribing how that process should unfold. Our training pipeline consists of two phases: (i) a **rollout phase**, where the agent constructs a reasoning trajectory τ using the creativity-aligned action space above, and (ii) a **reward phase**, where the resulting proposal o is evaluated. We use Group Relative Policy Optimization (GRPO) (Shao et al., 2024), which learns from relative rewards across sampled trajectories without requiring a separate critic. In our primary setting, the reward is applied only to the final proposal, requiring the model to learn which actions to take, how creatively to perform them, and how to sequence them based on their eventual effect on proposal quality.

3.3.1 Rewarding Proposal Quality

Scientific proposals contain multiple ideas whose novelty and feasibility may differ substantially, making a single holistic quality judgment difficult to ground. We therefore first decompose the final proposal into its atomic research ideas, separating *what* each phase proposes from *how* it plans to execute it. We then evaluate each component against the original research problem and relevant literature along three dimensions: *precedence*, *feasibility*, and *relevance* (Table 3.3.1).

Dimension	Definition	Positive Example (Reward ↑)	Negative Example (Reward ↓)
Precedence	Distance from prior work and existing approaches.	Core idea opens genuinely new technical directions.	Recombines familiar components without new insight.
Feasibility	Credibility & specificity of the execution plan.	Clear methods with precedent or scoped evaluation strategy.	Broad, underspecified plan or implausible scope.
Relevance	Alignment of the proposal with problem p.	Addresses a critical bottleneck in the target domain.	Drifts off-topic or fails to engage with the core challenges of p.

Table 2: Reward dimensions for proposal generation with examples. Reward prompts in Appendix F.

We average precedence and relevance across the proposal's atomic ideas, but take the minimum feasibility across experimental plans, since a single infeasible phase may compromise the executability of the overall proposal. The resulting three proposal-level scores are then weighted equally to form the outcome reward: $R_{\text{out}} = \frac{1}{3}\left(R_{\mathcal{P}} + R_{\mathcal{F}} + R_{\mathcal{R}}\right)$. These rewards are designed to provide targeted training signals for proposal quality, with atomic ideas evaluated separately and precedence and feasibility grounded in retrieved literature.

Outcome-only training still leaves the model to discover useful creative behaviors through their eventual effect on proposal quality. We therefore explore two additional forms of guidance: exposing the model to occasional serendipitous actions early in training, and directly rewarding properties of the intermediate reasoning process. Neither is required by the framework; we study whether either provides additional benefit beyond the final outcome reward.

3.3.2 Incentivizing Exploration through Serendipitous Actions

Human discovery is often shaped by unexpected encounters that expose researchers to directions they may not have deliberately pursued (Lubart, 2001). Similarly, an LLM early in training may repeatedly select familiar, low-creativity behaviors simply because it has not experienced useful alternatives. We therefore *explore* a stochastic intervention that occasionally exposes the model to a different action-level choice. With probability $\Pr(\textit{swap})$, the selected pair (a_i, c_i) is replaced by a sampled alternative (a_i', c_i'), and a decay factor γ gradually reduces this probability over training. The goal is not to prescribe random exploration, but to expose the model early on to creative behaviors whose value it can later learn from the resulting reward.

3.3.3 EXPLORING PROCESS-LEVEL REWARDS

Our primary models receive only the outcome reward above, allowing process-level behavior to emerge from credit assigned to the final proposal. We additionally investigate whether directly rewarding intermediate reasoning provides further benefit. The process reward evaluates each intermediate action along two complementary dimensions. *Exploration* measures whether a_i introduces a new direction relative to $\tau_{1:i-1}$, while *contribution* measures how much a_i ultimately contributes to the final proposal o. We score both dimensions for each intermediate action and average them across the trajectory to obtain R_{proc}, which then forms the final reward $R_{\text{po}} = \frac{1}{2}(R_{\text{proc}} + R_{\text{out}})$. Full details and prompts are provided in Appendix M.

4 EXPERIMENTS

We train all models using `verl` (Sheng et al., 2024) with GRPO (Shao et al., 2024) on `Qwen3-8B-Base` and `Qwen3-14B-Base`. Each agent trajectory consists of up to 5 actions. For the *search* action, we index arXiv[1] and use embedding-based retrieval with `GPT-4.1` as an external judge within the reward pipeline. Full training details are provided in Appendix E.

4.1 DATASET

We construct our dataset from the NSF Awards Database[2], focusing on Computer Science, Engineering, and Mathematics (CSE) awards from 2018 onward. We use a 90:10 train/test split, yielding 4,414 training and 491 test awards. Our model receives only the award title as p, which typically describes a broad, long-horizon research problem; we withhold the award abstract, project outcomes, and associated publications so that generation is not anchored to the funded solution. We focus on CSE to ensure broad coverage in open-access arXiv literature, while retaining substantial interdisciplinary breadth: the training set spans 46 research domains, with 23.1% of subfield labels in core AI/ML and over 48% of label occurrences outside AI/ML and core CS and Engineering (Appendix B).

Original NSF grant proposals are not publicly available, so we construct *reconstructed reference proposals* from evidence surrounding each funded project. For each award, we collect its abstract, project outcomes report, PI information, and award period, then retrieve likely associated arXiv papers by the award PIs. Candidate papers are ranked using PI overlap, topical similarity, publication timing, and explicit funding acknowledgments. We prompt `GPT-5.1` to treat these sources as downstream evidence and reconstruct a plausible pre-award research plan that could have led to the observed project and publications. The reference follows the same structured format as generated proposals, including a summary, background, and multi-phase research plan.

These reconstructions are not intended to reproduce the original NSF proposals. Rather, they provide standardized, evidence-grounded references for research directions that were actually funded and subsequently pursued. We use them only as matched references for pairwise evaluation, and the proposal-generating model never receives the oracle information used to construct them. Full reconstruction details and prompts are provided in Appendix L.

4.2 EVALUATION METRICS

Scientific creativity is difficult to capture with any single automatic metric, particularly because the long-term value of a research idea may not be observable for years. We therefore evaluate models along three complementary dimensions: *predicted citation impact*, *literature-grounded originality*, and *research idea diversity*. Together, these capture both the quality of individual proposals and the range of research directions explored across the dataset.

Predicted Citation Impact. For each problem, we compare the generated proposal against its matched reconstructed reference in randomized order and report pairwise win rate. We use **`SciJudge`** (Tong et al., 2026), a 30B-parameter model trained to predict relative citation impact from large-scale community signals, as a proxy for the potential usefulness and downstream value of the proposed research (prompts in Appendix H). It has an average 82.7% citation prediction accuracy.

[1] `https://arxiv.org/`

[2] https://huggingface.co/datasets/davidheineman/nsf-awards

Literature-Grounded Originality. We compare each generated proposal against its matched reference for originality. To ground the judgment in prior work, we retrieve the closest paper to each proposal and ask `GPT-5.1` to make a randomized pairwise comparison relative to the retrieved literature (Appendix J).

Research Idea Diversity. Pairwise metrics capture the quality of individual proposals but not whether a model repeatedly produces the same kinds of ideas. Following the annotation format of Chen et al. (2026), we use `GPT-5.4-mini` to classify each proposal by its primary *research idea paradigm* and, analogously, by its primary *contribution type* using categories adapted from prior taxonomies (Wobbrock, 2012; Miles, 2017). For each taxonomy, we report the effective number of categories represented (Appendix I), normalized by the seven available categories; higher values indicate a broader and more balanced range of research directions. The research-paradigm judge Chen et al. (2026) achieves $\kappa = 0.84$ agreement with human judgments on 150 samples.

Human-LLM Agreement. As an additional check on our pairwise judges, two human annotators evaluated a subset of proposals. Inter-annotator agreement was 80.0% ($\kappa = 0.625$) for impact and 86.7% ($\kappa = 0.766$) for originality, while human-LLM agreement was 76.7% for `SciJudge` on impact and 72.4% for `GPT-5.1` on originality (Appendix K).

4.3 Baselines

We compare against baselines that vary how creativity is introduced into proposal generation. **Zero-shot** models directly generate a structured proposal from the input problem without retrieval, action selection, or creativity levels. **Temperature** retains the same five creativity levels, but maps each level only to decoding temperature, with $c \in \{1, \ldots, 5\}$ corresponding to $T \in \{0, 0.25, 0.5, 0.75, 1.0\}$. This tests whether increased stochasticity alone can reproduce the benefits of semantic creativity guidance. **ReAct** (Yao et al., 2023) uses the same action space $\mathcal{A}$ with no corresponding $\mathcal{C}_a$.

Creative baselines apply our full creativity-aligned action space and natural-language creativity levels at inference time, but receive no RL training. We also train several **RL variants** with the same outcome-level reward as AI Night-Scientist: **+ Temp** and **+ ReAct** apply RL to the corresponding baselines above, while **+ Search + Write** retains semantic creativity levels but restricts $\mathcal{A}$ to retrieval and writing, similar to retrieval-augmented ideation systems (Si et al., 2025; Wang et al., 2024). Together, these separate the effects of semantic creativity guidance, RL, and the broader action space.

Finally, we compare against **GIANTS** (He-Yueya et al., 2026), an RL-trained model for scientific insight anticipation. Because GIANTS expects two input papers, we retrieve the two most relevant arXiv papers using the award title and abstract and reformat them with `GPT-4.1`. GIANTS therefore receives more award-specific information than AI Night-Scientist, which sees only the title.

For closed-source baselines, we use `GPT-4.1`, a strong general-purpose, non-reasoning model. Model recency does not necessarily imply greater creative diversity, with recent work finding increasing similarity across generations on open-ended tasks (Patel et al., 2026).

4.4 Experimental Results & Analysis

4.4.1 Creativity-Aligned RL Improves Proposal Quality

Table 3(a) shows a consistent advantage for AI Night-Scientist over both direct-generation and agentic baselines. Relative to `Qwen3-8B`, Night-8B improves predicted citation impact by 29.43 percentage points and originality by 53.79 points. At 14B, these gains grow to 32.03 and 66.15 points over Qwen3-14B. In comparison, scaling the zero-shot model from `8B` to `14B` yields almost no improvement, whereas scaling AI Night-Scientist adds another 3.29 points in citation and 12.36 points in originality. *Additional model capacity therefore appears substantially more useful once paired with a learned creative reasoning policy.*

The contrast with temperature-based exploration is especially evident after RL. Although both optimize the same proposal-level reward, Night-8B exceeds the temperature-controlled variant by 36.50 points in originality and 18.37 points in citation. ReAct closes much of this gap, showing that a learned multi-step policy already helps, but semantic creativity levels still add 9.72 points in originality and 4.95 points in citation. *The useful signal is therefore not simply to "explore more," but to specify how an action should deviate so RL can learn when those deviations are useful.*

(a) Proposal quality

Category	Method	Citation (%) ↑	Originality (%) ↑
Zero-shot	Llama-3.1-8B	1.61	1.15
	Qwen3-8B	0.46	2.53
	Qwen3-14B	1.15	2.53
	GPT-4.1	3.90	12.18
Temp.	Qwen3-8B + Temp	2.07	0.69
	GPT-4.1 + Temp	2.99	9.89
ReAct	Qwen3-8B + ReAct	2.33	6.54
	GPT-4.1 + ReAct	3.45	10.11
Creative	Llama-3.1-8B + Creative	1.10	6.52
	Qwen3-8B + Creative	1.89	4.25
	Qwen3-14B + Creative	0.71	4.76
	GPT-4.1 + Creative	3.00	14.02
w/ RL	GIANTS-4B	1.62	35.57
	Qwen3-8B + Temp	11.52	19.82
	Qwen3-8B + ReAct	24.94	46.60
	Qwen3-8B + Search + Write	24.47	33.33
Ours	**AI NIGHT-SCIENTIST-8B**	<u>29.89</u>	<u>56.32</u>
	AI NIGHT-SCIENTIST-14B	**33.18**	**68.68**

(b) Research-idea diversity

Method	Paradigm ↑	Contri. ↑
Qwen3-8B Zero-Shot	0.738	0.370
GPT-4.1 Zero-Shot	0.785	**0.439**
GIANTS-4B (RL)	0.773	0.347
Qwen3 + Temp (RL)	0.792	0.339
Qwen3 + ReAct (RL)	<u>0.888</u>	0.390
Night-8B (R_{po})	0.853	0.362
+ No swap	0.844	0.358
Night-8B (R_{out})	**0.943**	<u>0.425</u>

(c) Reward ablation

Method	Reward	Citation ↑	Original. ↑
ReAct	R_{out}	24.94	46.60
	R_{po}	2.10	6.54
Temp.	R_{out}	11/52	19.82
	R_{po}	10.35	13.33
Night-8B	R_{out}	**29.89**	<u>56.32</u>
	R_{po}	<u>26.20</u>	**61.61**

Table 3: Main results for **(a)** proposal quality, **(b)** research-idea diversity, and **(c)** outcome-only (R_{out}) vs. process + outcome (R_{po}) training. Best results are bolded and second-best are underlined.

The action space matters as well. Restricting the agent to *search* and *write* reduces originality by 22.99 points despite using the same creativity levels and outcome reward. Retrieval and iterative writing therefore explain only part of the gain; *spark* and *debate* provide additional ways to redirect reasoning rather than simply gather more evidence for an existing direction.

Robustness. To verify that these trends are not specific to reconstructed references, we directly compare NIGHT-8B against GPT-4.1 and `Qwen3-8B`. It wins 74.95% of citation and 86.96% of originality comparisons against GPT-4.1, increasing to 82.48% and 97.56% against `Qwen3-8B`. The same advantage therefore holds in direct head-to-head comparisons. We also include the confidence interval results under Appendix A.

4.4.2 AI NIGHT-SCIENTIST BROADENS RESEARCH IDEA DIVERSITY

Higher originality would be less meaningful if the model repeatedly relied on the same kind of creative strategy. Table 3(b) suggests otherwise. Relative to zero-shot Qwen3-8B, NIGHT-8B increases normalized research-paradigm coverage from 0.738 to 0.943, corresponding to roughly 1.4 additional effective categories out of seven; contribution-type coverage similarly rises from 0.370 to 0.425. *The model therefore produces not only more original proposals, but a broader range of research paradigms and contribution types.*

Diversity is also sensitive to early exploration. Removing serendipitous action swaps reduces paradigm coverage by 0.099, or roughly 0.7 effective categories, and contribution coverage by 0.067, or roughly 0.5 categories. This pattern is consistent with early exposure to less familiar behaviors helping the policy discover a broader set of useful trajectories.

The improvements are also not concentrated in a small set of CS topics. NIGHT-8B improves over zero-shot Qwen3-8B on both citation and originality across all 36 evaluated domains, including healthcare, biomedical engineering, geosciences, and quantum science. Full per-domain results are provided in Appendix C.

4.4.3 PROCESS REWARDS FAVOR ORIGINALITY OVER BREADTH

Table 3(c) shows a different effect from directly rewarding the reasoning process. For NIGHT-8B, process + outcome training raises originality by 5.29 points but lowers predicted citation impact by 3.69 points. Paradigm and contribution coverage also fall by roughly 0.6 and 0.4 effective categories, respectively. Process supervision can therefore favor more original individual proposals without necessarily producing a broader or more impactful set of ideas.

This effect is specific to the creativity-aligned action space. Adding the same process reward to ReAct sharply reduces both citation and originality, while Temperature also declines on both metrics. Process supervision is therefore not uniformly beneficial; its effect depends on whether the action space provides semantically distinct ways to express different degrees of creativity. We consequently

Table 4: Qualitative comparison for the NSF topic *"Using AI to Transform Online Video Lectures."* Green highlights novel or well-specified contributions; red highlights incremental, underspecified, or questionable elements. Full proposal excerpts and detailed error analysis are provided in Appendix D.

Method	Key Proposal Direction	Main Limitation
GPT-4.1	Uses AI to segment prerecorded lectures and represent the material through conversational agents, with adaptive pacing, clarification, and multimodal accessibility support.	The proposal is coherent and inclusive, but largely combines established lecture-segmentation and chatbot mechanisms rather than introducing a distinct technical idea.
ReAct	Introduces a multimodal engagement score that combines visual, auditory, and physiological signals and uses it to adapt lecture pacing and content granularity in real time.	The score is a concrete artifact, but the proposed RL-based adaptation is underspecified: the state, action, and reward spaces are left unclear for how engagement signals translate into adaptation decisions.
NIGHT-8B (R_{out})	Proposes *Agentify*, which turns a prerecorded lecture into a live, agent-mediated session where an AI interleaves questions, hints, and dialogue based on learner behavior.	The proposal substantially reframes the interaction, but some components are only loosely motivated or underdeveloped, including the cognitive-tier co-evolution mechanism and forced speed controls.
NIGHT-8B (R_{po})	Proposes Progressive Content Enactors (PCEs), which deliberately inject controlled errors and ambiguities for learners to detect and resolve, shifting the lecture from passive delivery toward active problem solving.	The central idea is distinctive, but the proposal also contains unverifiable quantitative claims and weakly grounded evaluation measures, reducing methodological credibility.

use outcome-only training as our primary setting and treat process rewards as a way to shift the model toward greater proposal-level originality. We detail further analysis of P_{po} in Appendix M.4.

4.4.4 Qualitative Analysis

We examine one representative NSF topic to understand how proposal directions differ qualitatively across methods. Table 4 shows a progression from adapting existing lecture-delivery mechanisms toward changing the underlying learning interaction itself. AI NIGHT-SCIENTIST produces the more substantial reframings, although the process + outcome variant also illustrates that greater creative deviation can come at the cost of grounding and methodological rigor.

4.4.5 Semantic Creativity Levels Better Align with Output Behavior

Our creativity levels are intended to change how an action is executed, not merely label it. Figure 3 measures this relationship using output entropy and semantic similarity (Appendix N). AI NIGHT-SCIENTIST (PO) improves on both measures during training, while Temp (PO) deteriorates toward -0.9. *Semantic descriptions therefore provide a more learnable link between the selected creativity level and resulting behavior than temperature alone.*

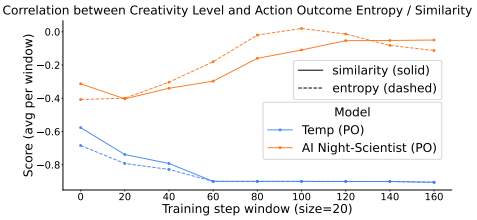

Figure 3: Higher correlation indicates better alignment between selected creativity level and output behavior.

5 Conclusion

We introduce AI NIGHT-SCIENTIST, a creativity-aligned agentic framework that represents creativity at the action, process, and outcome levels and uses reinforcement learning to teach models *when* and *how* to depart from predictable reasoning. Applied to long-horizon research proposal generation, AI NIGHT-SCIENTIST improves predicted citation impact and originality by up to 32.03 and 66.15 points over its base model, while producing a broader range of research paradigms. These gains are not reproduced by higher decoding temperature or ReAct alone, supporting our central finding that creativity is more than sampling stochasticity. Our ablations further show that outcome-only training provides the strongest overall balance of quality and diversity, while process-level rewards can increase originality at the cost of citation impact and breadth. Together, these results suggest that creativity can be learned as a multi-level reasoning capability, enabling scientific agents to move more flexibly between the structured reasoning of *day science* and the exploratory reasoning of *night science* while remaining tools for human-led discovery.

AI USE STATEMENT

In this work, we used generative AI tools for generating synthetic datasets (reconstructed reference proposals used for evaluation) and cleaning/reformatting them, providing feedback on the research methodology and experimental design, supporting qualitative analysis, assisting with interpretation of experimental results, the creation of scientific figures, and overall proofreading and readability improvements to the paper. We have not used generative AI tools to develop theoretical models or conceptual frameworks, formulate mathematical claims, provide critical ingredients for proving mathematical claims, assist in writing proofs, propose or refine hypotheses, implement methods. Translation and transcription were not applicable to this work. We have reviewed all AI-assisted work: reconstructed reference proposals were used only as evaluation references and were generated through the documented evidence-grounded reconstruction pipeline; experimental analysis suggestions and result interpretations were independently assessed by the authors against the experimental setup and outputs; qualitative analysis summaries (to improve readability) were checked against the underlying proposal content; scientific figures were reviewed for consistency with the reported results; and all AI-assisted edits to the manuscript were reviewed for factual accuracy, appropriate attribution, and consistency with the authors' intended claims. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

Supporting human-led research. AI NIGHT-SCIENTIST is designed for early-stage scientific ideation, when research directions are still speculative and not yet fully testable. Its role is to help researchers explore a broader set of possible directions, including connections or assumptions they may not otherwise consider. The system is intended as a creative collaborator rather than an autonomous researcher: generated proposals should be treated as candidate ideas that require critical evaluation, refinement, and validation by domain experts before they can support scientific conclusions.

Broadening research exploration. Human ideation is often shaped by the concepts and examples already available in the local environment (Lubart, 2001). By encouraging proposals that depart from existing work while remaining relevant and feasible, AI NIGHT-SCIENTIST aims to surface directions beyond those most immediately accessible to a researcher or model. This may be especially useful for interdisciplinary exploration, where relevant ideas can be distributed across distant literatures and research communities. Importantly, broader exploration does not imply that unfamiliar ideas are inherently better; their scientific value must still be established through expert judgment and empirical validation.

Risks and responsible use. As with other generative systems for scientific writing, AI NIGHT-SCIENTIST can produce plausible but incorrect, infeasible, or insufficiently grounded proposals, and could be misused to generate low-quality scientific content at scale. Its outputs should therefore be presented as AI-generated suggestions rather than validated research plans. Responsible use requires transparent disclosure, independent verification of claims and citations, and meaningful human oversight before ideas are pursued, disseminated, or incorporated into scientific work.

REPRODUCIBILITY STATEMENT

We provide anonymous source code for AI NIGHT-SCIENTIST, including the agentic framework and training pipeline, at `https://anonymous.4open.science/r/ai_night_scientist-DE2D/`, with the reconstructed reference-proposal pipeline released separately at `https://anonymous.4open.science/status/synthetic_proposals-ECF8`. The main paper specifies the action space, creativity levels, reward formulation, baselines, and evaluation metrics. The appendix further provides the complete action and reward prompts (Appendices F–G), training compute and hyperparameters (Appendix E), NSF dataset filtering and domain construction (Appendix B), reconstructed reference-proposal procedure and prompt (Appendix L), and full evaluation protocols for predicted citation impact, originality, and research-idea diversity (Appendices H–J). We additionally report confidence intervals and human–LLM agreement in Appendices A and K. Together, these materials document the data construction, model training, prompting, retrieval, and evaluation procedures needed to reproduce the reported experiments.

REFERENCES

Osama Mohammed Afzal, Preslav Nakov, Tom Hope, and Iryna Gurevych. Beyond" not novel enough": Enriching scholarly critique with llm-assisted feedback. *arXiv preprint arXiv:2508.10795*, 2025.

Dhruv Agarwal, Bodhisattwa Prasad Majumder, Reece Adamson, Megha Chakravorty, Satvika Reddy Gavireddy, Aditya Parashar, Harshit Surana, Bhavana Dalvi Mishra, Andrew McCallum, Ashish Sabharwal, et al. Autodiscovery: Open-ended scientific discovery via bayesian surprise. *Advances in Neural Information Processing Systems*, 38:25181–25219, 2026.

Shivam Agarwal, Zimin Zhang, Lifan Yuan, Jiawei Han, and Hao Peng. The unreasonable effectiveness of entropy minimization in llm reasoning. *arXiv preprint arXiv:2505.15134*, 2025.

Margaret A Boden. Creativity and artificial intelligence. *Artificial intelligence*, 103(1-2):347–356, 1998.

Ziyu Chen, Yilun Zhao, and Arman Cohan. Measuring the gap between human and llm research ideas. *arXiv preprint arXiv:2607.01233*, 2026.

Leonora M Cohen. A continuum of adaptive creative behaviors. *Creativity Research Journal*, 2(3): 169–183, 1989.

Pier-Luc de Chantal and Henry Markovits. Reasoning outside the box: Divergent thinking is related to logical reasoning. *Cognition*, 224:105064, 2022.

Christopher P Dwyer, Deaglán Campbell, and Niall Seery. An evaluation of the relationship between critical thinking and creative thinking: Complementary metacognitive processes or strange bedfellows? *Journal of Intelligence*, 13(2):23, 2025.

Juraj Gottweis, Wei-Hung Weng, Alexander Daryin, Tao Tu, Anil Palepu, Petar Sirkovic, Artiom Myaskovsky, Felix Weissenberger, Keran Rong, Ryutaro Tanno, et al. Towards an ai co-scientist. *arXiv preprint arXiv:2502.18864*, 2025.

Tianyang Gu, Jingjin Wang, Zhihao Zhang, and HaoHong Li. Llms can realize combinatorial creativity: generating creative ideas via llms for scientific research. *arXiv preprint arXiv:2412.14141*, 2024.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

Tarun Gupta and Danish Pruthi. All that glitters is not novel: Plagiarism in ai generated research. *arXiv preprint arXiv:2502.16487*, 2025.

Diane F Halpern. The nature and nurture of critical thinking. *Critical thinking in psychology*, (1): 1–14, 2007.

Sarah Harvey and James W Berry. Toward a meta-theory of creativity forms: How novelty and usefulness shape creativity. *Academy of Management Review*, 48(3):504–529, 2023.

Joy He-Yueya, Anikait Singh, Ge Gao, Michael Y Li, Sherry Yang, Chelsea Finn, Emma Brunskill, and Noah D Goodman. Giants: Generative insight anticipation from scientific literature. *arXiv preprint arXiv:2604.09793*, 2026.

Jules Hirsch. An anniversary for cancer chemotherapy. *Jama*, 296(12):1518–1520, 2006.

Priyanka Kargupta, Ishika Agarwal, Tal August, and Jiawei Han. Tree-of-debate: Multi-persona debate trees elicit critical thinking for scientific comparative analysis. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 29378–29403, 2025a.

Priyanka Kargupta, Shuyue Stella Li, Haocheng Wang, Jinu Lee, Shan Chen, Orevaoghene Ahia, Dean Light, Thomas L Griffiths, Max Kleiman-Weiner, Jiawei Han, et al. Cognitive foundations for reasoning and their manifestation in llms. *arXiv preprint arXiv:2511.16660*, 2025b.

Priyanka Kargupta, Runchu Tian, and Jiawei Han. Beyond true or false: Retrieval-augmented hierarchical analysis of nuanced claims. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 29664–29679, 2025c.

Priyanka Kargupta, Shuhaib Mehri, Dilek Hakkani-Tur, and Jiawei Han. Sparking scientific creativity via llm-driven interdisciplinary inspiration. *arXiv preprint arXiv:2603.12226*, 2026.

B Lee Ligon. Penicillin: its discovery and early development. In *Seminars in pediatric infectious diseases*, volume 15, pp. 52–57. Elsevier, 2004.

Fang Liu, Yang Liu, Lin Shi, Houkun Huang, Ruifeng Wang, Zhen Yang, Li Zhang, Zhongqi Li, and Yuchi Ma. Exploring and evaluating hallucinations in llm-powered code generation. *arXiv preprint arXiv:2404.00971*, 2024.

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. *arXiv preprint arXiv:2405.06373*, 2024.

Todd I Lubart. Models of the creative process: Past, present and future. *Creativity research journal*, 13(3-4):295–308, 2001.

R Thomas McCoy, Shunyu Yao, Dan Friedman, Mathew D Hardy, and Thomas L Griffiths. Embers of autoregression show how large language models are shaped by the problem they are trained to solve. *Proceedings of the National Academy of Sciences*, 121(41):e2322420121, 2024.

Elspeth McFadzean. The creativity continuum: Towards a classification of creative problem solving techniques. *Creativity and Innovation Management*, 7(3):131–139, 1998.

Sarnoff Mednick. The associative basis of the creative process. *Psychological review*, 69(3):220, 1962.

D Anthony Miles. A taxonomy of research gaps: Identifying and defining the seven research gaps. In *Doctoral student workshop: finding research gaps-research methods and strategies, Dallas, Texas*, volume 1, pp. 1–10, 2017.

Bernard A Nijstad, Carsten KW De Dreu, Eric F Rietzschel, and Matthijs Baas. The dual pathway to creativity model: Creative ideation as a function of flexibility and persistence. *European review of social psychology*, 21(1):34–77, 2010.

Charles O'Neill, Tirthankar Ghosal, Roberta Răileanu, Mike Walmsley, Thang Bui, Kevin Schawinski, and Ioana Ciucă. Sparks of science: Hypothesis generation using structured paper data. *arXiv preprint arXiv:2504.12976*, 2025.

Nirav Patel, Josiah Crossman, Eva Aggarwal, and Emily Wenger. Are llms becoming similarly creative? evidence from three years of models. *arXiv preprint arXiv:2608.19437*, 2026.

Mark A Runco and Garrett J Jaeger. The standard definition of creativity. *Creativity research journal*, 24(1):92–96, 2012.

Jürgen Schmidhuber. Formal theory of creativity, fun, and intrinsic motivation (1990–2010). *IEEE transactions on autonomous mental development*, 2(3):230–247, 2010.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

Guangming Sheng, Chi Zhang, Zilingfeng Ye, Xibin Wu, Wang Zhang, Ru Zhang, Yanghua Peng, Haibin Lin, and Chuan Wu. Hybridflow: A flexible and efficient rlhf framework. *arXiv preprint arXiv: 2409.19256*, 2024.

Chenglei Si, Diyi Yang, and Tatsunori Hashimoto. Can llms generate novel research ideas? a large-scale human study with 100+ nlp researchers. In Y. Yue, A. Garg, N. Peng, F. Sha, and R. Yu (eds.), *International Conference on Learning Representations*, volume 2025, pp. 94003–94092, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/file/ea94957d81b1c1caf87ef5319fa6b467-Paper-Conference.pdf.

Gunther S Stent. The statue within: An autobiography. *Science*, 239(4847):1545–1547, 1988.

Jingqi Tong, Mingzhe Li, Hangcheng Li, Yongzhuo Yang, Yurong Mou, Weijie Ma, Zhiheng Xi, Hongji Chen, Xiaoran Liu, Qinyuan Cheng, et al. Ai can learn scientific taste. *arXiv preprint arXiv:2603.14473*, 2026.

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.

Yiping Wang, Qing Yang, Zhiyuan Zeng, Liliang Ren, Liyuan Liu, Baolin Peng, Hao Cheng, Xuehai He, Kuan Wang, Jianfeng Gao, et al. Reinforcement learning for reasoning in large language models with one training example. *arXiv preprint arXiv:2504.20571*, 2025.

Thomas B Ward, Steven M Smith, and Ronald A Finke. Creative cognition. *Handbook of creativity*, 189:212, 1999.

Solange Muglia Wechsler, Carlos Saiz, Silvia F Rivas, Claudete Maria Medeiros Vendramini, Leandro S Almeida, Maria Celia Mundim, and Amanda Franco. Creative and critical thinking: Independent or overlapping components? *Thinking skills and creativity*, 27:114–122, 2018.

Emily Wenger and Yoed Kenett. We're different, we're the same: Creative homogeneity across llms. *arXiv preprint arXiv:2501.19361*, 2025.

Jacob O Wobbrock. Seven research contributions in hci. *Intelligence*, 174(12-13):910–950, 2012.

Itai Yanai and Martin Lercher. Night science. *Genome Biology*, 20(1):179, 2019.

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

Yunpu Zhao, Rui Zhang, Wenyi Li, and Ling Li. Assessing and understanding creativity in large language models. *Machine Intelligence Research*, 22(3):417–436, 2025.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36:46595–46623, 2023.

Li Zhong and Zilong Wang. Can llm replace stack overflow? a study on robustness and reliability of large language model code generation. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pp. 21841–21849, 2024.

A Confidence Intervals for Main Results

Table 5: Results with 95% confidence intervals. Citation and originality are pairwise win rates against matched reconstructed references; diversity metrics report normalized effective category coverage.

	Proposal Quality		Research-Idea Diversity	
Method	**Citation (%)**	**Originality (%)**	**Paradigm**	**Contribution**
Zero-shot				
Qwen3-8B	0.46 [0.13, 1.66]	2.53 [1.42, 4.47]	0.738 [0.682, 0.785]	0.370 [0.340, 0.399]
GPT-4.1	3.91 [2.45, 6.17]	12.18 [9.44, 15.59]	0.785 [0.733, 0.826]	**0.439 [0.401, 0.472]**
RL Baselines				
Qwen3-8B + Temp	11.52 [8.85, 14.87]	19.82 [16.34, 23.82]	0.792 [0.742, 0.831]	0.339 [0.313, 0.364]
Qwen3-8B + ReAct	24.94 [20.93, 29.42]	46.60 [41.75, 51.52]	0.888 [0.847, 0.916]	0.390 [0.366, 0.412]
AI Night-Scientist				
Night-8B (R_{po})	26.21 [22.30, 30.53]	**61.61 [56.96, 66.06]**	0.853 [0.807, 0.887]	0.362 [0.338, 0.384]
Night-8B (R_{out})	**29.89 [25.77, 34.35]**	56.32 [51.63, 60.91]	**0.943 [0.907, 0.962]**	0.425 [0.412, 0.435]

Table 5 reports 95% confidence intervals for the primary proposal-quality and research-idea diversity results shown in Table 3. Citation and originality are pairwise win rates against matched reconstructed reference proposals; their intervals are Wilson score intervals over the evaluated proposal pairs, using the same treatment of ties and undecided judgments as the reported win rates. Paradigm and contribution diversity are measured as the normalized effective number of categories represented, $\exp(H)/K$, where H is Shannon entropy and $K = 7$; their intervals are 95% percentile intervals from 100,000 proposal-level bootstrap resamples within each method. These intervals quantify evaluation-sample uncertainty only.

B Dataset Construction & Domain Coverage

B.1 Filtering and Splits

We select NSF CSE awards from 2018 onwards. Non-research grants (e.g., travel and conference grants) are excluded. This yields 4,905 awards in total, split 90:10 into 4,414 training and 491 test instances. We focus on CSE awards to maximize the availability of open-access literature on arXiv, which underpins both the search action and the retrieval steps in the reward pipeline.

B.2 Domain Distribution

To characterize the topical breadth of the training set, we applied an automated three-stage taxonomy pipeline to the 4,414 NSF award titles. In the first stage, GPT-4.1 assigned one to three fine-grained research subfields to each title in batches of 50, yielding 11,071 raw subfield labels (2.5 per title on average) spanning 4,300 unique terms. In the second stage, these labels were grouped into intermediate clusters via a second LLM pass. In the third stage, the intermediate clusters were consolidated into a final canonical taxonomy of 46 research domains (Table 6).

The distribution reflects the intended scope of the NSF CISE directorate: *Artificial Intelligence & Machine Learning* accounts for 23.1% of all subfield label occurrences, followed by *Computer Science Theory* (11.2%) and *Networks & Communications* (5.9%). The concentration in core CS and engineering is expected, especially as open-access literature on arXiv is most comprehensive for these fields, making retrieval-grounded proposal generation most reliable there. Importantly, the training set is not exclusively CS-focused: the remaining $\sim$20% of labels span applied and interdisciplinary domains, including *Biomedical Engineering & Health Informatics* (2.7%), *Cognitive & Behavioral Sciences* (2.2%), *Earth & Geosciences* (1.0%), *Quantum Science & Engineering* (0.9%), and policy-adjacent areas such as *Ethics, Equity & Societal Impacts* and *Economics, Policy & Law*, providing a diverse training signal across 46 domains in total.

C Performance Across Research Domains

We report per-domain win rates for AI Night-Scientist-8B and Qwen3-8B zero-shot against the reference proposals across 36 canonical research domains (Table 7). The zero-shot baseline achieves near-zero impact win rates in almost every domain (overall 0.5%) and near-zero originality win rates (overall 2.5%), confirming that an untuned 8B model cannot produce proposals competitive with actual NSF awards on either metric. AI Night-Scientist-8B consistently overcomes this gap, delivering positive impact gains in all 34 domains with sufficient decided pairs, ranging from $+8$ percentage points (pp) (Materials Science & Nanotechnology) to $+63$ pp (Environmental Informatics & Sensing). Originality gains are uniformly large across all 36 domains, ranging from $+37$ pp (Natural Language Processing & Linguistics, Energy, Power & Automotive Systems) to $+83$ pp (Organizational Computing & Workflow), with particularly strong improvements in interdisciplinary and application-oriented fields such as Quantum Science & Engineering ($+80.0 \pm 17.9$ pp), Safety, Risk & Resilience Engineering ($+71.4 \pm 12.1$ pp), Earth & Geosciences ($+71.4 \pm 17.1$ pp), and Computational Social Sciences & Digital Humanities ($+70.6 \pm 11.1$ pp).

Core CS domains (AI & ML, Computer Science Theory, Robotics) show more moderate but still substantial originality gains ($+49$–$+57$ pp), possibly reflecting that these are the most prevalent domains within the base model's pretraining distribution. The broad consistency of gains across all 36 domains (including fields with little CS overlap such as Biomedical Engineering, Healthcare,

Table 6: Distribution of research domains across training proposals, grouped into high-level domain families. Fine-grained labels are derived from a three-stage automated taxonomy pipeline applied to 4,414 NSF award titles; high-level families are used only to organize the labels for presentation.

Domain	Count	% of Labels
Computing & AI	**6,418**	**58.0%**
Artificial Intelligence & Machine Learning	2558	23.1
Networks & Communications	657	5.9
Computer Systems & Architecture	626	5.7
Data Science & Analytics	611	5.5
Cybersecurity & Privacy	547	4.9
Robotics & Autonomous Systems	455	4.1
Embedded & Hardware Systems	310	2.8
Cloud & Distributed Computing	277	2.5
Software Engineering & Programming Languages	164	1.5
Natural Language Processing & Linguistics	157	1.4
Extended Reality & Immersive Technologies	34	0.3
Sustainable Computing & Infrastructure	22	0.2
Mathematical & Computational Foundations	**1,942**	**17.5%**
Computer Science Theory	1239	11.2
Operations Research & Optimization	171	1.5
Modeling, Simulation & Visualization	152	1.4
Mathematics & Theoretical Foundations	123	1.1
Experimental Methods & Evaluation	100	0.9
Statistics & Data Science	80	0.7
Systems Science & Engineering	77	0.7
Human, Social & Educational Research	**1,344**	**12.1%**
Human-Computer Interaction	538	4.9
Cognitive & Behavioral Sciences	241	2.2
Education Research & Pedagogy	236	2.1
Computational Social Sciences & Digital Humanities	173	1.6
Career Development & Workforce	50	0.5
Ethics, Equity & Societal Impacts	47	0.4
Organizational Computing & Workflow	25	0.2
Economics, Policy & Law	22	0.2
Legal Informatics & Law	7	0.1
Media, Creativity & Inclusive Technologies	5	0.0
Health & Life Sciences	**543**	**4.9%**
Biomedical Engineering & Health Informatics	295	2.7
Biological & Biomedical Sciences	145	1.3
Healthcare & Medical Sciences	103	0.9
Engineering & Physical Sciences	**532**	**4.8%**
Imaging, Instrumentation & Sensors	122	1.1
Materials Science & Nanotechnology	103	0.9
Quantum Science & Engineering	97	0.9
Electrical & Electronic Engineering	63	0.6
Energy, Power & Automotive Systems	46	0.4
Safety, Risk & Resilience Engineering	45	0.4
Physical Sciences	42	0.4
Chemistry & Materials Science	14	0.1
Earth, Environment & Urban Systems	**193**	**1.7%**
Earth & Geosciences	116	1.0
Agricultural & Environmental Sciences	31	0.3
Urban Informatics & Smart Cities	26	0.2
Environmental Informatics & Sensing	20	0.2
Research Infrastructure & Other	**99**	**0.9%**
Open Science & Research Infrastructure	76	0.7
Miscellaneous	23	0.2
Total	**11,071**	**100.0%**

Materials Science, and Earth Sciences) suggests that the improvements are not confined to a narrow research subdomain.

Table 7: Domain-level gains of NIGHT-8B (R_{out}) over Qwen3-8B zero-shot, grouped into the same high-level domain families as Table 6. Gains are differences in pairwise win rate (percentage points); uncertainty is propagated as $\sqrt{\sigma^2_{\text{Night}} + \sigma^2_{\text{zs}}}$. Domains within each family are ordered by originality gain. "—" indicates insufficient samples (< 5 pairs) for a domain-level estimate.

Domain	Citation Gain (pp)	Originality Gain (pp)
Computing & AI		
Networks & Communications	+40.0 ± 7.3	+59.6 ± 7.2
Robotics & Autonomous Systems	+34.0 ± 6.9	+57.4 ± 7.7
Artificial Intelligence & Machine Learning	+29.6 ± 3.4	+57.0 ± 3.7
Cybersecurity & Privacy	+43.2 ± 7.5	+55.6 ± 7.7
Computer Systems & Architecture	+30.6 ± 7.0	+54.0 ± 7.5
Data Science & Analytics	+28.6 ± 7.6	+50.0 ± 9.1
Cloud & Distributed Computing	+42.9 ± 9.4	+50.0 ± 9.1
Embedded & Hardware Systems	+36.1 ± 11.5	+47.8 ± 11.3
Software Engineering & Programming Languages	+23.1 ± 11.7	+46.2 ± 13.8
Natural Language Processing & Linguistics	+22.2 ± 9.8	+36.8 ± 11.1
Mathematical & Computational Foundations		
Systems Science & Engineering	+60.0 ± 21.9	+60.0 ± 25.3
Statistics & Data Science	+14.3 ± 13.2	+57.1 ± 18.7
Mathematics & Theoretical Foundations	+23.5 ± 10.3	+52.9 ± 13.2
Computer Science Theory	+26.3 ± 4.7	+49.0 ± 5.6
Modeling, Simulation & Visualization	+31.2 ± 11.6	+47.1 ± 13.4
Experimental Methods & Evaluation	+36.4 ± 14.5	+45.5 ± 15.0
Operations Research & Optimization	+25.0 ± 12.5	+38.5 ± 13.5
Human, Social & Educational Research		
Organizational Computing & Workflow	+20.0 ± 17.9	+83.3 ± 15.2
Computational Social Sciences & Digital Humanities	+18.8 ± 9.8	+70.6 ± 11.1
Ethics, Equity & Societal Impacts	+0.0 ± 0.0	+66.7 ± 15.7
Education Research & Pedagogy	+21.4 ± 11.0	+64.3 ± 13.9
Cognitive & Behavioral Sciences	+47.1 ± 12.1	+57.9 ± 11.3
Human-Computer Interaction	+29.3 ± 7.1	+51.2 ± 8.2
Career Development & Workforce	+0.0 ± 0.0	+50.0 ± 20.4
Health & Life Sciences		
Biomedical Engineering & Health Informatics	+41.2 ± 11.9	+61.1 ± 11.5
Healthcare & Medical Sciences	+43.8 ± 12.4	+56.2 ± 13.5
Biological & Biomedical Sciences	+25.0 ± 12.5	+50.0 ± 14.4
Engineering & Physical Sciences		
Quantum Science & Engineering	—	+80.0 ± 17.9
Safety, Risk & Resilience Engineering	+35.7 ± 12.8	+71.4 ± 12.1
Materials Science & Nanotechnology	+7.7 ± 7.4	+64.3 ± 13.9
Imaging, Instrumentation & Sensors	+35.3 ± 11.6	+61.1 ± 11.5
Chemistry & Materials Science	—	+60.0 ± 21.9
Energy, Power & Automotive Systems	+50.0 ± 17.7	+37.5 ± 17.1
Earth, Environment & Urban Systems		
Earth & Geosciences	+14.3 ± 13.2	+71.4 ± 17.1
Environmental Informatics & Sensing	+62.5 ± 17.1	+50.0 ± 17.7
Research Infrastructure & Other		
Open Science & Research Infrastructure	+25.0 ± 12.5	+58.3 ± 15.8

D QUALITATIVE ANALYSIS

We qualitatively compare proposals on the NSF topic *"Using Artificial Intelligence to Transform Online Video Lectures into Effective and Inclusive Agent-Based Presentations."* GPT-4.1 produces

Table 8: Qualitative comparison for the NSF topic *"Using AI to Transform Online Video Lectures."*

Method	Key Proposal Excerpt	Assessment
GPT-4.1	"Develop AI algorithms to automatically segment video lectures into coherent topics and extract key instructional elements. Design and implement conversational agents that re-present segmented lecture content interactively, supporting adaptive pacing, clarifications, and multimodal delivery (text, visuals, sign language, audio descriptions). Empirically evaluate effectiveness and inclusivity against standard video lectures. Assess scalability and real-world deployment challenges within existing platforms."	**Pros:** Well-structured four-phase plan; covers accessibility and inclusivity. **Cons:** Restates existing segmentation and chatbot approaches without introducing a new mechanism; no novel technical contribution beyond integration.
ReAct	"Develop a multimodal AI framework for real-time content adaptation leveraging visual, auditory, and physiological engagement metrics to dynamically adjust lecture content. Introduces a 'multimodal engagement score' synthesizing data across modalities. Paired with a content adaptation engine using reinforcement learning to optimize lecture pacing and content granularity [no state, action, or reward space defined]. Validated via controlled A/B study with 120 participants across three engagement conditions."	**Pros:** Concrete novel artifact (multimodal engagement score); controlled experimental design. **Cons:** RL formulation is underspecified; experimental design is elaborate relative to the degree of technical novelty.
AI NIGHT-SCIENTIST (Outcome)	"'Agentify': a framework that transforms passive lectures into live, agent-mediated sessions where AI agents interweave structured questions, personalized hints, and interactive dialogues based on real-time participant behavior. Proposes co-evolution of agent templates and lecture cognitive tiers. Randomizes 200 learners across 40 lectures (STEM, Social Science, Humanities) into four groups comparing passive, over-moderated, fixed-tier, and adaptive-tier conditions. Includes a forced-speed-slider component with 8 granular speed settings (1x–8x) to calibrate attention span variability."	**Pros:** Novel reframing of lectures as co-creative sessions rather than content delivery artifacts. Central hypothesis is testable. Ambitious but sensible multi-group experimental design. **Cons:** Some design elements appear contrived (speed-slider rationale). Cognitive tier co-evolution mechanism is underdeveloped.
AI NIGHT-SCIENTIST (Process + Outcome)	"Progressive Content Enactors (PCEs): virtual agents that shift from automated fidelity to intentional pedagogical maladaptation—deliberately embedding controlled errors and ambiguity so that learners are forced to detect and resolve them, promoting active problem-solving over passive reception. Uses a Layered Semantic Transformative Stack (LISS) to anchor distortions to content structure. Reports cognitive engagement score (NCIR) of c=0.12, $+3.5\sigma$ retention improvement (TAR-film, 50:1 decay), and a 2.78-σ increase in credibility-based efficacy—none of which are independently verifiable or grounded in standard evaluation frameworks."	**Pros:** Most novel direction: purposeful errors as a pedagogical tool mirrors well-established ideas (e.g., error-based learning, productive failure). PCE is a fresh, actionable concept that is clearly differentiated from prior work. **Cons:** Background section relies on unverifiable metrics and internally inconsistent citations. Evaluation framework is not grounded in standard methodology.

a well-structured but generic proposal that restates existing approaches (e.g., lecture segmentation and conversational agents) without introducing a distinguishing mechanism. ReAct adds a concrete multimodal engagement score but leaves its reinforcement learning formulation underspecified. AI NIGHT-SCIENTIST-8B (outcome-only) introduces the novel *Agentify* concept—treating lectures as live, co-created sessions in which an AI agent interleaves questions and dialogues in real time, rather than optimizing pre-recorded content after the fact. The process-reward variant proposes *Progressive Content Enactors* (PCEs), which deliberately introduce controlled errors and ambiguities into lectures to force students to actively resolve them (analogous to how working through a flawed proof teaches more than reading a perfect one), though its background section overreaches with unverifiable metrics. Table 8 provides excerpts and error analysis; key novel or well-specified contributions are highlighted in green, while vague, incremental, or questionable claims are highlighted in red.

E TRAINING HYPERPARAMETERS

We report all key hyperparameters used to train AI NIGHT-SCIENTIST-8B and AI NIGHT-SCIENTIST-14B. Training is performed with Verl (Sheng et al., 2024) on 4×8 NVIDIA H100 (80 GB) GPUs (32 GPUs total). We fine-tune `Qwen3-8B-Base` and `Qwen3-14B-Base` end-to-end without LoRA adapters, as LoRA is not supported for SGLang-based rollouts in Verl.

F REWARD IMPLEMENTATION DETAILS

This appendix provides the complete prompts used within the outcome-level reward pipeline described in Section 3.3.1. The reward is computed in three stages: (1) decomposing the final proposal into

Table 9: Training and inference hyperparameters for AI NIGHT-SCIENTIST.

Hyperparameter	Value
Data	
Train batch size (prompts)	32
Max prompt length	16,384 tokens
Max response length	25,000 tokens
Dynamic batching	Enabled
Rollout (SGLang)	
Rollouts per prompt (n)	8
Max agent actions per trajectory	5 (11 max assistant turns)
Max tool-response length	2,048 tokens
Sampling temperature	0.7
Max model context length	32,768 tokens
Serendipitous swap probability	0.5 initially; decay $\gamma = 0.001$
Actor (FSDP2)	
PPO mini-batch size (prompts)	32
Max tokens per GPU (training)	32,000
Reference Model (FSDP2)	
Parameter offload	Enabled
Max tokens per GPU (log-prob)	40,000
Optimization	
Algorithm	GRPO
KL loss coefficient	0.001
KL in reward	Disabled
Total training steps	170
Retrieval and Reward Infrastructure	
arXiv index	Kaggle arXiv snapshot
Retrieval model	`sentence-transformers/all-MiniLM-L6-v2`
External reward judge	`GPT-4.1`
Judge temperature	0

atomic ideas, (2) scoring each idea for originality and relevance, and (3) scoring the execution plan for feasibility against retrieved literature.

F.1 ACTION SPACE AND CREATIVITY LEVELS

Table 10 summarizes the creativity level descriptions ($\mathbb{C}_a$) for each creativity-enabled action.

F.2 PRECEDENCE REWARD PROMPT

Each decomposed idea is scored for precedence against a paper retrieved from arXiv using the idea's keyword phrase.

Table 10: Creativity level descriptions for each action in the research proposal generation task. The *write* and *complete* actions are not creativity-enabled (fixed at level 1).

c	Description
search	
1	Extremely relevant to the proposal, focusing on background information and related works based on terms extracted from the proposal title.
2	Very relevant, focusing on background and related works closely related to the proposal idea.
3	Somewhat relevant, focusing on background that is tangentially related to the proposal idea.
4	Not very relevant; explores broader topics or concepts that may not be directly related but still provide useful context.
5	Very distantly relevant; explores specific alternate perspectives, domains, or philosophical questions of general interest.
debate	
1	Discussion with a colleague in the same specific area; structured single-turn debate focused tightly on proposal elements.
2	Discussion with a colleague on related topics; structured debate focused on the proposal with minor deviation permitted.
3	Debate with a peer from the same high-level field but a different topic; open-ended multi-turn format with some deviation.
4	Debate with a domain-expert from a different discipline or a mixture of 2–3 experts; open-ended with broader thematic scope.
5	Debate with an expert from a completely different discipline or a mixture of domain-experts; unstructured, focused on exploring diverse perspectives and high-level ideas.
spark	
1	A minor conventional idea to challenge, very specific to the proposal's current state.
2	A minor conventional idea to challenge; somewhat broad but still focused on the proposal.
3	A moderate conventional idea to challenge; meaningful change questioning existing assumptions.
4	A significant conventional idea to challenge; broad and exploratory, with potential field-level impact.
5	A major conventional idea to challenge; very broad, with the potential to revolutionize the field.

```
You are a reviewer for NSF proposals who has been a professor in your
field for over 20 years.  Evaluate the precedence of a proposal's core
idea relative to the problem statement and a related paper.

Problem statement:  {problem}
Related paper:  {related_paper}

Evaluate the following core idea for its precedence.  Consider how
it introduces new concepts, methodologies, or perspectives that
differentiate it from the related paper or existing work in the field.
Idea:  {idea}

Score from 1 to 5:
1 = Not novel.  Rehash of existing work.  2 = Slightly novel.  Largely
derivative.
3 = Moderately novel.  4 = Very novel.  Potential to advance the
field.
5 = Highly novel.  Groundbreaking; challenges existing paradigms.

Output JSON: {"score":  <int>, "explanation":  "<string>"}
```

F.3 Feasibility Reward Prompt

Each idea's execution plan is scored for feasibility against a retrieved paper grounding the assessment in existing methodological precedent.

```
You are a reviewer for NSF proposals who has been a professor in your
field for over 20 years. Evaluate the feasibility of a given core
idea from a proposal. A related paper is provided to contextualize
the assessment.
Problem statement: {problem}
Related paper: {related_paper}
Evaluate the following implementation idea for its feasibility:
practicality, resources required, complexity, ethical implications,
and specificity. A highly feasible proposal includes specific
technical details; vague ideas should be penalized.
Implementation idea: {idea}
Score from 1 to 5:
1 = Not feasible. 2 = Slightly feasible. 3 = Moderately feasible.
4 = Very feasible. 5 = Highly feasible and specific.
Output JSON: {"score": <int>, "explanation": "<string>"}
```

F.4 Relevance Reward Prompt

Each decomposed idea is scored for relevance to the input problem p.

```
You are a reviewer for NSF proposals who has been a professor in your
field for over 20 years. Evaluate the relevance of the following idea
to the given problem statement: how well it addresses the problem,
aligns with its objectives, and fits within the proposed methods.
Problem statement: {problem}
Idea: {idea}
Score from 1 to 5:
1 = Not relevant. 2 = Slightly relevant. 3 = Moderately relevant.
4 = Very relevant. 5 = Highly relevant; directly aligned and
comprehensive.
Output JSON: {"score": <int>, "explanation": "<string>"}
```

F.5 Action Selection Prompt

The following prompt is provided to the model at each trajectory step to select the next action and creativity level.

```
You are a researcher selecting the best next action for your proposal
writing process. The possible actions are:
search: Generate specific search queries to retrieve information
(level: 1-5).
debate: Set up a discussion with one or more participants on a topic
(level: 1-5).
spark: Generate a novel research idea that challenges conventional
thinking (level: 1-5).
write: Write or revise the proposal (level: 1 only).
complete: Indicate satisfaction and end the process (level: 1 only).
Level 1-2 = focused, proposal-specific. Level 4-5 = exploratory,
creativity-inducing.
Target problem: {problem}
Output JSON: {"action": "<string>", "level": <int>}
```

G ACTION PROMPTS

Each action in the **AI Night-Scientist** framework is realized through a structured LLM prompt that takes the current problem p, creativity level c, and the accumulated trajectory context as inputs, and returns a JSON output. Below we document the core prompt structure and output schema for each creativity-enabled action (*search*, *debate*, *spark*) and the *write* action.

G.1 SEARCH ACTION

The search action generates up to five arXiv search queries. The creativity level controls query relevance: level 1 produces tightly targeted queries derived from proposal keywords; level 5 produces broad, exploratory queries spanning alternate domains, philosophical perspectives, or unrelated fields. Retrieved papers are parsed and appended to the trajectory context for use by subsequent actions.

```
You are a researcher starting a literature review for a research
proposal. Generate specific search queries whose relevance depends
on the creativity level (1-5):
Level 1: Extremely relevant; queries based on terms from the proposal
title.
Level 2: Very relevant; background and related works closely tied to
the proposal.
Level 3: Somewhat relevant; tangentially related background.
Level 4: Loosely related; broader topics that still provide useful
context.
Level 5: Very distantly relevant; alternate domains, perspectives, or
philosophical questions.
Problem: {problem} Level: {level} ({level_description})
Output JSON: {"search_queries": ["<query 1>", ..., "<query 5>"]}
```

G.2 DEBATE ACTION

The debate action is executed in two sequential steps: (i) generating a debate setup (participants, topics, and structure) and (ii) simulating the debate conversation. The creativity level controls participant diversity and discussion structure, ranging from a focused single-turn discussion with a same-field colleague (level 1) to an open-ended, multi-turn panel with cross-disciplinary experts (level 5). The conversation is appended verbatim to the trajectory context.

```
Step 1 -- Debate Setup. You are a researcher designing a debate. The
creativity level controls who you debate with, what topics you cover,
and how the debate is structured:
Level 1: Single senior colleague in the same area; structured
single-turn debate on proposal specifics.
Level 2: Peer on related topics; structured, focused on the proposal
with minor deviation.
Level 3: Peer from the same field but a different topic; open-ended
multi-turn, some deviation allowed.
Level 4: Domain-expert from a different discipline or a 2-3 expert
panel; open-ended, broader scope.
Level 5: Expert(s) from a completely different discipline;
unstructured, high-level philosophical exploration.
Problem: {problem} Level: {level}
Output JSON: {"debate_participants": [{"name": "...", "job": "...",
"expertise": "..."}],
"debate_topics": ["..."], "debate_structure": "..."}
```

Step 2 -- Debate Conversation. You are a researcher conducting the debate as set up above. Generate a coherent, multi-turn conversation where each participant draws on their specific expertise with concrete details. Responses should not be surface level.

Debate setup: {debate_setup}

Output JSON: {"conversation_history": [{"speaker_name": "...", "speaker_response": "..."}]}

G.3 Spark Action

The spark action generates a "Bit-Flip" idea (O'Neill et al., 2025): it inverts a commonly held assumption to produce a novel research direction. The creativity level governs the scope of the challenged assumption, from a specific proposal-level constraint (level 1) to a field-reshaping paradigm shift (level 5). The Bit-Flip is appended to the trajectory context and can influence subsequent write and action-selection steps.

You are a researcher who has just had a sudden insight that challenges conventional thinking. Generate a Bit-Flip: identify a prevailing belief in the field (the **Bit**), invert or challenge it (the **Flip**), and distill the core conceptual leap into a single phrase (the **Spark**). Creativity level controls the scope:

Level 1: Minor, proposal-specific assumption to challenge.
Level 2: Minor but somewhat broader assumption; questions existing approaches in the proposal.
Level 3: Moderate assumption; meaningful challenge to existing norms within the proposal.
Level 4: Significant, field-level assumption; substantial departure from the status quo.
Level 5: Major, paradigm-level assumption; potential to revolutionize the field.

Problem: {problem} Level: {level}

Output JSON: {"bit": "<2-3 sentences on the status quo and its limitation>",
"flip": "<novel approach or perspective, $\geq$2 sentences>", "spark": "<core phrase>"}

G.4 Write Action

The write action produces or revises the complete structured proposal. It takes all accumulated trajectory context as input and outputs a structured JSON proposal. The creativity level is fixed at 1 for this action; creative expression is instead encoded in the trajectory that feeds into it.

```
You are a researcher writing or revising a research proposal based on
all information gathered throughout the writing process. Create a
coherent, comprehensive proposal. You may deviate from prior ideas if
it improves the proposal.
Problem: {problem} Current proposal: {proposal} Context history:
{context}
Output JSON:
{"proposal_title": "...",
"proposal_summary": "concise statement of what, why, and what
problems it resolves",
"background_and_significance": "historical review of the field, what
remains to be done, and how this proposal advances it",
"research_plan": [{"phase": "...", "idea": "detailed hypothesis and
contribution",
"experimental_plan": "concrete methods, evaluation, pitfalls,
fallback plans"}]}
```

H SciJudge Evaluation Details

We evaluate proposal quality using **SciJudge** (OpenMOSS-Team/SciJudge-30B) (Tong et al., 2026), a 30B-parameter model trained to predict scientific impact via citation forecasting. For each problem instance in the test set, we compare a model-generated proposal against a matched reconstructed reference NSF proposal in a pairwise setting.

```
Today is {date}. Based on the titles and abstracts of the following
two proposals A and B, determine which proposal has a higher citation
count across its eventual papers.
{timing_assumption} (e.g., "Assume Proposal B was published earlier
than Proposal A.")
Show your reasoning process in <reason> </reason> tags. Return the
final answer in <answer> </answer> tags. The final answer should
contain only 'A' or 'B'.
Proposal A:
Title: {title_a}
Abstract: {abstract_a}
Proposal B:
Title: {title_b}
Abstract: {abstract_b}
```

Randomized A/B order. For each pair, proposal order is randomized using a deterministic per-pair seed to eliminate position bias. The model does not know which proposal originated from the system under evaluation versus the reconstructed reference.

Temporal framing. When comparing against the synthetically reconstructed proposal (Appendix L), we instruct the judge to assume the reconstructed reference proposal was published *earlier*. This grounds the comparison in a realistic temporal context: the model-generated proposal is treated as the "newer" proposal, which must surpass the quality of the funded reference to be preferred.

Output parsing. The model responds with chain-of-thought reasoning inside `<reason>...</reason>` tags and a final answer (A or B) inside `<answer>...</answer>` tags. We extract the answer tag; if absent or malformed, we fall back to a regex match on the first standalone 'A' or 'B' character.

Reported metric. We report pairwise win rate: the fraction of evaluated pairs for which `SciJudge` predicts the model-generated proposal to yield greater downstream citation impact than the matched reconstructed reference.

I CONTRIBUTION-TYPE CLASSIFICATION DETAILS

We classify each proposal by its primary scholarly contribution using `GPT-5.4-mini` and a seven-way taxonomy adapted from Wobbrock (2012). The contribution types are: *empirical*, which produces new findings from systematically gathered or analyzed data; *artifact*, which creates a novel instantiated system, tool, process, intervention, or other constructed artifact; *methodological*, which introduces or refines a reusable method; *theoretical*, which develops reusable concepts, models, principles, hypotheses, or frameworks; *benchmark or dataset*, which contributes a reusable data or evaluation resource; *survey*, which synthesizes existing work into higher-level understanding; and *opinion*, which advances an evidence-grounded position intended to persuade or redirect discussion.

```
You are an expert annotator of scholarly contribution types.
Label the proposal using a domain-general taxonomy of research
contributions.  Do not classify by topic, domain, or technical
substrate.
A contribution type is the main form of knowledge or scholarly output
the proposed work will add.
Proposal to label:
Original research topic:  {problem}

Proposal title:
{proposal_title}

Proposal summary:
{proposal_summary}

Background and significance:
{background_and_significance}

Research plan:
{research_plan}

Identify the proposal's principal contribution from its proposed
deliverables and research plan.  Classify the central knowledge
claim or deliverable rather than the topic, incidental methods, or
technical vocabulary.  When several contribution types appear, select
the one that best captures the proposal's primary scholarly output.
Use a secondary label only when a distinct second contribution is
substantively central.

Return JSON with exactly this shape:
{
  "contribution_type": {
    "primary": "<one contribution type>",
    "secondary": "<one contribution type or none>"
  },
  "confidence": <0.0-1.0>
}
```

Distinction from research-idea paradigms. Contribution type captures *what form of scholarly output* the proposal ultimately contributes, whereas the research-idea paradigm captures *what high-level research move* is used to turn an opportunity into a proposed direction. Following the research-idea annotation setup of Chen et al. (2026), our paradigm categories distinguish moves such as assumption relaxation, failure mitigation, formal derivation, empirical mapping, artifact construction, and optimization. The two axes are therefore complementary rather than redundant. For example, an artifact contribution may arise from relaxing an assumption, mitigating a failure, or optimizing resource use; conversely, a measurement-oriented research idea may ultimately contribute either empirical findings or a reusable benchmark.

Primary-label selection. The classifier is instructed to distinguish the proposal's central contribution from methods or artifacts that merely support it. In particular, empirical contributions are defined by the new findings produced from data, while benchmark or dataset contributions are defined by the reusable resource itself. Similarly, artifact contributions center on a novel instantiated invention, whereas methodological contributions center on a reusable way of conducting research or practice. Theoretical contributions may be empirically evaluated, but the reusable concept, explanation, model, or framework must remain the principal contribution.

Reported diversity. For our diversity analysis, we use only the primary contribution label for each proposal. We measure how broadly and evenly a model distributes its proposals across the seven contribution types using the normalized effective number of categories described in Section **??**.

J LLM Originality Evaluation Details

System: You are an experienced NSF proposal reviewer. Return valid JSON only.

User: You are a reviewer for NSF proposals who has been a professor in your field for over 20 years.
You are comparing two proposals that address the same award problem.
Judge which proposal is more original.
Originality must be judged relative to: (1) the other proposal; (2) the closest retrieved paper for Proposal A; (3) the closest retrieved paper for Proposal B. If a proposal appears more novel at first glance but is very similar to its closest retrieved paper, penalize its novelty accordingly.

Award title / target problem: {problem_title}

Proposal A: {proposal_a}
Closest Retrieved Paper for Proposal A:
Title: {paper_a_title} Abstract: {paper_a_abstract}

Proposal B: {proposal_b}
Closest Retrieved Paper for Proposal B:
Title: {paper_b_title} Abstract: {paper_b_abstract}

Criteria: Originality of the central concepts; distance from the closest retrieved paper; whether the proposal goes beyond recombining familiar ideas; depth of novelty rather than superficial novelty; whether the proposal opens genuinely new technical directions.

Calibration examples:

- *Borderline positive:* Proposal A may look less flashy but introduces a genuinely different technical lever, representation, or problem decomposition than both Proposal B and A's closest paper.
- *Borderline negative:* Proposal A is polished but most apparent creativity comes from extending the closest paper to a new application or evaluation setting without a clearly new core idea.
- *Borderline negative:* Proposal A combines familiar components into a multi-phase agenda; breadth alone is not novelty if the contribution is still an expected recombination of known methods.
- *Borderline negative:* Do not reward writing sophistication, dense terminology, or detailed milestones if the underlying idea remains close to prior work.

Return JSON only: {"winner": "A or B", "explanation": "brief comparative rationale"}

We assess Originality using `GPT-5.1` as a judge in a pairwise setting. For each test instance, the model-generated proposal and the matched reconstructed reference proposal are each submitted to our arXiv retrieval system; the top-1 closest paper is retrieved for each using `sentence-transformers/all-MiniLM-L6-v2` embedding-based search. The retrieval query is constructed from the proposal title concatenated with the first three sentences of the proposal

summary. Proposal order (A vs. B) is randomized per-pair using a deterministic seed to eliminate position bias.

The judge is instructed to assess originality relative to both the competing proposal *and* the closest retrieved paper for each proposal. This prevents superficially creative proposals from scoring well if their core idea closely mirrors existing literature. The judgment emphasizes: (i) originality of the central concept, (ii) distance from the retrieved nearest neighbor, (iii) whether the proposal recombines familiar ideas or opens a genuinely new technical direction, and (iv) depth of novelty over breadth or terminological density.

We report the fraction of pairs in which the model-generated proposal is preferred over the matched reference (model win rate).

K Human Evaluation Study

To assess the reliability of our automated evaluation metrics, we conducted a small-scale human evaluation in which two expert annotators independently judged 33 pairwise proposal comparisons (covering all three model variants: Qwen3-8B-Base, GPT-4.1, and AI Night-Scientist-8B). Each item presented annotators with a model-generated proposal alongside a matched reconstructed reference proposal; annotators selected the preferred proposal on two axes (predicted citation impact and precedence) or indicated a tie. Both independent annotators have 5+ and 15+ years of research experience in related fields, respectively.

K.1 Inter-Annotator Agreement

Table 11 reports inter-annotator agreement. Agreement is high on citation impact (80.0%, κ=0.625, substantial) and originality (86.7%, κ=0.766, substantial), indicating that both dimensions can be reliably assessed by domain-expert annotators.

Table 11: Inter-annotator agreement across evaluation dimensions (n=15 shared items).

Metric	Raw Agreement	Cohen's κ	Interpretation
Impact	80.0% (12/15)	0.625	Substantial
Originality	86.7% (13/15)	0.766	Substantial

K.2 Human-LLM Agreement

We also measure how well the LLM judges (SciJudge for predicted citation impact, GPT-5.1 for originality) agree with human annotators. For each annotator response, we map the human A/B choice to a *model* or *ground_truth* winner using the answer key, and compare against the corresponding LLM judgment; ties and null LLM outputs are excluded. Table 12 reports the results pooled across both annotators.

Table 12: Human-LLM agreement on predicted citation impact and originality. Ties and unparseable LLM outputs are excluded.

Metric	LLM Judge	Human–LLM Agreement
Impact	SciJudge-30B	76.7% (46/60)
Originality	GPT-5.1	72.4% (43/60)

Agreement in the low-to-mid 70s is consistent with prior work reporting LLM-judge alignment with human raters (Zheng et al., 2023), and notably these human–LLM agreement rates provide additional evidence that the automated judges broadly track expert preferences at the proposal level. These results support the use of `SciJudge` and `GPT-5.1` as scalable proxies for human judgment at the proposal level.

L RECONSTRUCTED REFERENCE PROPOSAL CONSTRUCTION

Original NSF grant proposals are not publicly available. We therefore construct *reconstructed reference proposals* from evidence surrounding each funded project. These reconstructions are not intended to reproduce the original proposal text; rather, they provide standardized, evidence-grounded references for research directions that were actually funded and subsequently pursued. The proposal-generating model never receives the award abstracts, project outcomes, or associated publications used in this reconstruction.

1. **Collect award evidence.** For each award, we collect its title, abstract, project outcomes report (POR), PI names, award ID, and award period from the NSF Awards Database.

2. **Retrieve likely associated papers by PI.** We use PI-first retrieval over arXiv, issuing up to 12 author-based queries using PI full names, surnames, and combinations of multiple PIs. Results are deduplicated by arXiv ID. This constrains retrieval to papers plausibly authored by the funded investigators before using topical information to determine which papers are most closely associated with the award.

3. **Rank candidate papers using award evidence.** Candidate papers are first ranked using PI-author overlap, topical overlap between the award evidence (title, abstract, and POR) and the paper title and abstract, and publication timing relative to the award period. We additionally reward cases in which the award title appears directly in the paper abstract. PI overlap receives the strongest weight so that topical similarity alone cannot make an unrelated paper a strong candidate.

4. **Verify candidates using full-text evidence.** For the highest-ranked candidates, we retrieve the paper text and extract the abstract, introduction, methods or approach, and experiments or results, excluding references and appendices where possible. We then rescore papers using full-text evidence, including explicit mention of the NSF award ID, NSF funding acknowledgments, PI surnames, and topical overlap with the award. We retain up to three highly ranked papers per award. If no paper passes the grounding threshold, we retain the strongest PI-matched papers with usable text so that reconstruction remains grounded in likely investigator-authored work.

5. **Reconstruct a plausible pre-award research plan.** We prompt `GPT-5.1` as an expert NSF PI using the award metadata, POR, and selected papers. Crucially, the prompt treats publications as *downstream evidence* of what the funded project ultimately produced and asks the model to infer a plausible pre-award research plan that could have led to those outcomes, rather than summarize or copy the papers. The model is instructed not to mention the reconstruction process and to produce the same structured format used by our generated proposals: a proposal summary, background and significance, and a multi-phase research plan specifying the central idea and experimental plan for each phase.

The resulting references should therefore be interpreted as plausible reconstructions of the funded research direction, not as recovered NSF proposals. We use them only as matched references for pairwise evaluation, providing a consistent comparison point grounded in projects that were funded and subsequently pursued.

```
You are an expert NSF principal investigator reconstructing a highly
realistic original NSF proposal. Use the award title, abstract,
project outcomes report, PI names, and likely resulting papers to
infer a plausible original proposal. Do not copy text verbatim from
the papers and do not mention that the proposal is reconstructed or
generated.
Award title: {title} PI names: {pi_names} Award period: {dates}
NSF award abstract: {abstract}
NSF project outcomes report: {por}
Likely papers supported by this award ({N} retrieved):
{paper_blocks} (title, authors, arXiv ID, evidence excerpt, processed
content)
Write a detailed proposal concrete enough that an evaluator can see
the tasks, data, algorithms, evaluation metrics, expected results,
potential pitfalls, and fallback strategies.
Output JSON: {"proposal_title": "...", "proposal_summary": "2-4
dense paragraphs",
"background_and_significance": "detailed technical review of the
field",
"research_plan": [{"phase": "...", "idea": "...",
"experimental_plan": "..."}]}
Requirements: 3-5 research phases; every phase must include phase,
idea, and experimental_plan; content must remain consistent with the
award metadata and likely papers; no markdown, commentary, or code
fences.
```

M PROCESS-LEVEL REWARD DETAILS

The process-level reward evaluates the quality of the agent's reasoning *trajectory* rather than only its final output. Each intermediate action (i.e., all actions before the final *write*) is scored along two complementary dimensions: **exploration** and **contribution**. Both scores are produced by an LLM judge, mapped from the 1–5 integer scale to $[-1, 1]$ via the centering transform $s' = (s - 3)/2$, and averaged equally into a per-action score. The overall process reward is the mean of all per-action scores; actions that were available but unused (i.e., the trajectory is shorter than the maximum allowed) are penalized with a score of -1.

M.1 EXPLORATION SCORING

The exploration dimension rewards actions that cover genuinely new ground relative to everything already explored in the trajectory. A high exploration score indicates that the action introduces a novel direction, perspective, or information source that prior actions did not cover; a low score indicates redundancy. If no prior actions have been taken, the action is considered maximally exploratory by default.

M.2 CONTRIBUTION SCORING

The contribution dimension rewards actions that meaningfully shaped the final proposal's quality. Each intermediate action is scored for how much it contributed to the final proposal's precedence, feasibility, and overall quality, conditioned on prior actions (to avoid rewarding redundant actions that happen to echo an earlier high-value contribution).

You are a reviewer for NSF proposals who has been a professor in your field for over 20 years. Evaluate whether the current action is sufficiently different and novel compared to previous actions in the research process.

Problem statement: {problem}
Current action type: {action_type}
Current action output: {action_output}
Prior actions taken (ordered): {previous_actions}

Assess how meaningfully different the current action is from the previous actions. Consider:
• Does it explore a new direction or perspective not previously covered?
• Would it provide substantially different information from prior actions?
• For *search*: are the queries fundamentally different from prior queries or debate topics?
• For *debate*: are the topics/participants substantially different from prior debates?
• For *spark*: does it challenge different assumptions than prior spark actions?

Assign a low score if this action would likely generate outputs similar to prior actions; assign a high score if it explores genuinely new ground.

Score from 1 to 5:
1 = Not exploratory; highly redundant with prior actions.
2 = Slightly exploratory; mostly overlaps with prior actions.
3 = Moderately exploratory; some new perspectives with limited novelty.
4 = Very exploratory; covers substantially new ground.
5 = Highly exploratory; genuinely novel direction with minimal overlap.

Output JSON: {"score": <int>, "explanation": "<string>"}

You are a reviewer for NSF proposals who has been a professor in your field for over 20 years. Evaluate how much a single intermediate action contributed to the final proposal across three dimensions: precedence, feasibility, and overall quality.

Problem statement: {problem}
Final proposal: {final_proposal}
Prior actions before this one (ordered): {previous_actions}
Intermediate action type: {action_type}
Intermediate action output: {action_output}

When scoring, condition on prior actions. If this action mostly repeats prior actions with similar outputs, assign low contribution scores.

Score each dimension from 1 to 5:
Precedence contribution: 1 = No contribution or harmful; 5 = Essential contribution to novelty.
Feasibility contribution: 1 = No contribution or harmful; 5 = Essential contribution to feasibility.
Overall quality contribution: 1 = No contribution or harmful; 5 = Essential contribution to clarity, coherence, rigor, and alignment with the problem.

Output JSON: {"precedence_score": <int>, "precedence_explanation": "<string>",
"feasibility_score": <int>, "feasibility_explanation": "<string>",
"quality_score": <int>, "quality_explanation": "<string>"}

M.3 Score Aggregation

For each intermediate action i, the combined per-action process score is $r_i = (\text{explore}_i + \text{contribute}_i)/2$, where each component has been normalized to $[-1, 1]$. The contribution score is the mean of its three sub-components (novelty, feasibility, overall quality) after normalization. The final process reward for a trajectory of N intermediate actions out of a maximum of $N_{\max}$ is:

$$R_{\text{proc}} = \frac{1}{N_{\max}} \left(\sum_{i=1}^{N} r_i \; + \; (N_{\max} - N) \cdot (-1) \right)$$

Missing actions are assigned -1 to penalize overly short trajectories. When combined with outcome-level rewards (the PO setting), the final reward is $R_{\text{po}} = (R_{\text{proc}} + R_{\text{out}})/2$.

M.4 Process Rewards Shift the Learned Action Policy

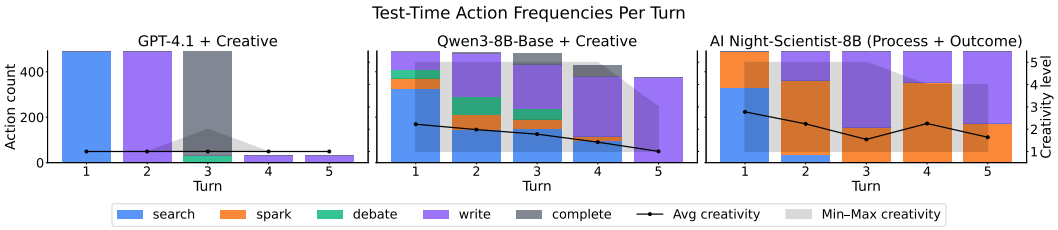

Figure 4: Test-time action frequencies across five reasoning turns for three model variants.

Figure 4 shows how the learned policies differ at test time. `GPT-4.1` typically follows a short *search* → *write* → *stop* pattern, while `Qwen3-8B-Base` uses a broader mix of actions across the full trajectory. AI Night-Scientist-8B trained with process+outcome rewards shows a different preference: it selects *spark* more frequently across turns and tends to use it at moderate-to-high creativity levels. This suggests that process supervision shifts the policy toward assumption-challenging actions that more directly open new research directions.

Debate is selected relatively rarely. One practical reason may be that it requires more of the trajectory budget, since the agent first identifies participants and relevant evidence before generating the grounded discussion. More generally, these results suggest that the learned value of an action depends not only on its creative potential, but also on how efficiently it contributes within a limited reasoning horizon.

N Action-Level Entropy and Similarity Metrics

During training, we track two diagnostic metrics—**action-level entropy** and **action-level similarity**—that measure whether the model's actual output behavior is consistent with its selected creativity level. Neither metric is used as a reward signal; they serve purely as interpretability probes.

The core premise of our creativity-aligned framework is that selecting a higher creativity level should *demonstrably change* how the model acts—producing outputs with higher token entropy (more lexical variety) and outputs that are more distinct from the evolving proposal (higher divergence). If a model selects level 5 but generates outputs indistinguishable from level 1, the creativity selector is decorative rather than functional. These metrics operationalize that sanity check.

Action-level entropy. For each action, we collect the per-token log-probabilities from the model's generation and compute the mean Shannon entropy across top-k vocabulary candidates. This is averaged over all tokens in the action output to yield a single per-action entropy score. We then compute the Pearson correlation between the action's selected creativity level (equivalently, its inverse noise level $1 - \text{noise}/\text{max_level}$) and this entropy value across all actions in a training batch. A positive correlation indicates that higher creativity levels lead to higher-entropy, more diverse token distributions.

Action-level similarity. For each action, we compute the cosine similarity between the action output's sentence embedding and the embedding of the most recently written proposal draft (or the problem statement if no draft exists yet). We again compute the Pearson correlation between the creativity level and this similarity score across the batch. A negative correlation (similarity decreases as creativity level increases) is desirable: higher creativity should yield actions that cover genuinely new territory rather than paraphrasing what has already been written.

O LIMITATIONS

Computational cost. Training creativity-aligned agents requires substantial GPU compute. We mitigate this by working with an efficient 8B-parameter architecture, but scaling to larger models would amplify the footprint. Additionally, our reward pipeline relies on LLM judges (GPT-4.1 for feasibility/relevance during training, SciJudge-30B and GPT-5.1 for evaluation). Unlike similarity-based reward signals that compare generated text against existing corpora, we deliberately use LLM judges to assess genuinely *new* ideas—minimizing data leakage and preventing the model from learning to simply paraphrase known work from our dataset, which spans proposals as far back as 2018. This design choice comes at higher inference cost but is central to the validity of the precedence and originality signal.

Training stability. Given the inherent complexity of our task, particularly the implicit objective of increasing the entropy of a model's outputs, we have observed that training can become unstable, especially over longer training runs. High-entropy generation is at odds with the stability assumptions underlying standard RL algorithms designed for low-variance, verifiable reward settings. We hope future work will explore more robust RL algorithms that can handle the entropy increases required for creative reasoning. We also note that our infrastructure relies on VERL as the training backend and SGLANG as the inference backend; their current joint support for structured, multi-step agentic frameworks is limited, and tighter integration would enable more scalable and reliable training pipelines.