Learning to Generate Research Idea with Dynamic Control

Anonymous submission

Abstract

The rapid advancements in large language models (LLMs) have demonstrated their potential to accelerate scientific discovery, particularly in automating the process of research ideation. LLM-based systems have shown promise in generating hypotheses and research ideas. However, current approaches predominantly rely on prompting-based pre-trained models, limiting their ability to optimize generated content effectively. Moreover, they also lack the capability to deal with the complex interdependence and inherent restrictions among novelty, feasibility, and effectiveness, which remains challenging due to the inherent trade-offs among these dimensions, such as the innovation-feasibility conflict. To address these limitations, we propose a novel framework that employs a two-stage approach combining Supervised Fine-Tuning (SFT) and controllable Reinforcement Learning (RL). In the SFT stage, the model learns foundational patterns from pairs of research papers and follow-up ideas. In the RL stage, multi-dimensional reward modeling, guided by fine-grained feedback, evaluates and optimizes the generated ideas across key metrics. Dimensional controllers enable dynamic adjustment of generation, while a sentence-level decoder ensures context-aware emphasis during inference. Our framework provides a balanced approach to research ideation, achieving high-quality outcomes by dynamically navigating the trade-offs among novelty, feasibility, and effectiveness.

Introduction

In recent years, advances in LLM have been made in their capacity to accelerate scientific discovery. Specifically, LLMs like GPT-4 and LLama have demonstrated their capability to produce coherent and contextually relevant text across diverse applications, including sentiment analysis (Zhang et al. 2023; Yang et al. 2024a; Zhu et al. 2024), question answering (spr 2024), and summarization (Authors 2023). Moreover, their remarkable performance in multi-step reasoning and complex decision-making (Zhou et al. 2023; Park et al. 2023) has underscored their potential in the field of research ideation. Typically, a well-developed research idea (or hypothesis)[1] is established with a *methodology* and an *experiment plan*, as shown in Figure 1.

By automating the research ideation process, these research idea generation systems can swiftly synthesize vast data and insights, uncovering novel connections that might elude human researchers. This capability is evidenced by the growing number of studies employing research agents to autonomously generate and validate new ideas (Wang and Zhou 2023; Du, Kim, and Park 2023; Bornstein and Singh 2024). However, despite notable progress in LLM-based research ideation as demonstrated in prior work, these efforts primarily rely on pre-trained models with no task-specific learning. Such reliance restricts the full exploitation of optimizing the generated content, underscoring the urgent need for further refinement and development in this area.

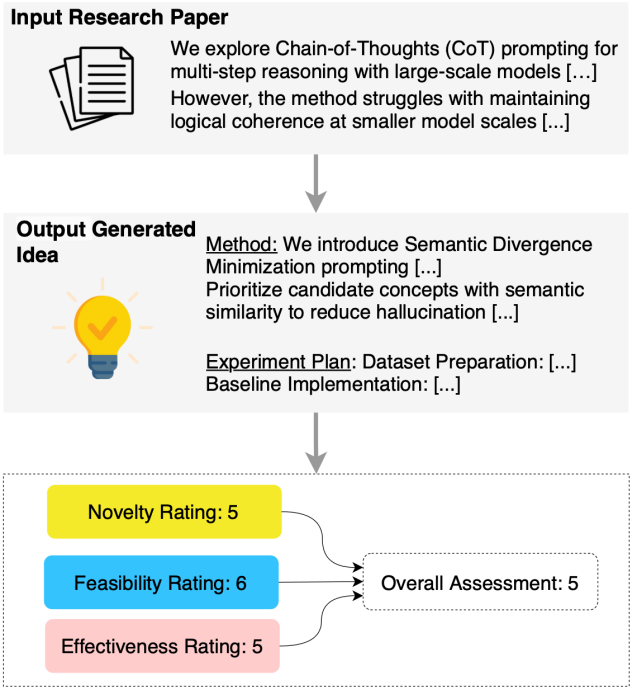

Figure 1: Research ideas generation from research papers. Each idea is measured across dimensions of novelty, feasibility, and effectiveness.

The quality assessment of a research idea involves multiple aspects, centered around key metrics in three dimensions: (1) Novelty, which assesses how unique or original

[1] In this paper, research idea and hypothesis are used interchangeably.

the ideas generated by the system are, distinguishing them from existing ideas; (2) Feasibility evaluates how practical or implementable the ideas are given current resources and constraints and (3) Effectiveness, which measures how likely the generated ideas will achieve their intended outcomes or solve identified problems. These metrics can serve as optimization objectives to guide the research ideation process, by leveraging techniques such as reinforcement learning (RL)(). Recent studies have explored Reinforcement Learning from Human Feedback (RLHF) to benefit LLM training (). However, existing techniques lack of capability to deal with the complex interdependence and inherent restrictions among the metrics used for assessing research idea quality. For instance, a recent study highlights these challenges by revealing the inevitable innovation-feasibility trade-off: highly novel hypotheses may lack feasibility, while overly feasible ideas often limit the scope for groundbreaking discoveries. (Chen et al. 2024b). How to optimize research ideation towards each of the key metrics while balancing them with satisfying trade-offs remains a critical, unresolved question.

To address this issue, we propose a novel research ideation framework designed to dynamically control the emphasis on key assessment metrics through a two-stage approach: SFT and controllable RL. In the SFT stage, the idea generator learns foundational patterns by training on pairs of research papers and corresponding follow-up ideas. In the RL stage, we employ multi-dimensional reward modeling as a real-world assessment approximation (Wu et al. 2023). Reward models, trained on fine-grained feedback from review data, score each metric—novelty, feasibility, and effectiveness—providing detailed guidance for model refinement. To enable precise and adaptive control, we introduce dimensional controllers, trained alongside the RL process, which adjusts the generation style to prioritize specific metric dimensions when necessary. This is complemented at inference time by a sentence-level decoder that dynamically adjusts the weights of controllers, ensuring context-aware emphasis - such as prioritizing novelty in the method section and feasibility in the experiment planning. Together, these mechanisms, guided by feedback signals from the reward models, result in more balanced and high-quality idea generation.

Our contributions are summarized as follows:

- We propose a research ideation framework to dynamically control the optimization of the generated idea towards novelty, feasibility, and effectiveness.
- We first introduce dynamic decoding into the RL-based supervised fine-tuning framework, achieving satisfying performance with a balanced trade-off among different assessment metrics of research ideation.
- We train our reward models using collected real-world datasets, enabling research idea scoring in a fine-grained manner.
- We conduct a comprehensive evaluation with a human study, demonstrating the effectiveness of our proposed method for optimized, controllable research ideation.

Related Work

NLP for scientific discovery NLP techniques have significantly advanced scientific discovery by enabling researchers to manage extensive literature, identify knowledge gaps, and analyze trends effectively (Raghu and Schmidt 2020; Hope et al. 2021). Models such as SciBERT (Beltagy, Lo, and Cohan 2019) and BioBERT (Lee et al. 2020) pre-trained on scientific materials have enhanced these abilities by improving performance on fundamental tasks. Recent developments in LLMs have extended their utility to creative and generative tasks in scientific research. For example, LLMs have been employed to formulate research questions, generate hypotheses, draft research proposals, and even outline experimental designs (Brown et al. 2020; Zhong et al. 2023; Qi et al. 2023; Yang et al. 2023; Wang et al. 2024a). Several prior works have specifically explored methods to enhance idea generation. Approaches such as iterative novelty boosting (Wang et al. 2024b), multi-agent collaboration (Baek et al. 2024), and multi-module retrieval and revision (Yang et al. 2024b) have been proposed to advance ideation capabilities beyond baseline prompting methods. Beyond ideation, another branch of research leverages LLMs for automating experimental workflows. Works like MLAgent (Huang et al. 2024) and SciCode (Tian et al. 2024) have used LLMs to generate code for executing research experiments, while systems such as AI Scientist (Lu et al. 2024) and MLR-Copilot (Li et al. 2024) combine idea generation with code implementation to directly test AI-generated concepts. However, these approaches are often limited to constrained problem spaces or rely on proxy metrics for evaluation, such as LLM-based scoring, which can be inconsistent and unreliable.

Fine-tuning LLM with RL RLHF has shown success in diverse NLP tasks (Christiano et al. 2017; Stiennon et al. 2020; Ouyang et al. 2022), including text summarization (Ziegler et al. 2019), instruction following (Ouyang et al. 2022), and question answering (Nakano et al. 2021). While most works focus on optimizing a single holistic reward combining multiple objectives, recent efforts have explored fine-grained rewards for specific attributes, such as reasoning or ethical considerations (Glaese et al. 2022; Uesato et al. 2022).

Additionally, non-RL methods have leveraged feedback to improve model outputs. For example, supervised fine-tuning has been used with high-scoring samples selected by reward models (Rafailov et al. 2023). Conversational models have incorporated binary user satisfaction signals to enhance response generation (Askell et al. 2021), while natural language feedback has been stored in memory banks and retrieved during task execution (Madaan et al. 2022). Some approaches refine outputs conditioned on human feedback and subsequently use reward models to select the best refinements (Scheurer et al. 2022; Menick et al. 2022).

Method

We introduce a scientific idea proposer with multi-dimension feedback, which consists of two stages: supervised fine-tuning stage, and reinforcement learning stage that has three components: reward modeling, multi-dimension reward augmented controllable reinforcement learning, and decoding.

Overview

Suppose we have a training set $\mathcal{D} = \{X_i, Y_i\}_{i=1}^{N}$, where X_i and Y_i are research paper and idea, respectively. Then we **fine-tune** the language model $\mathcal{M}$ with the training set. Thereafter, we collect a reward training set $\mathcal{D}_r = \{(X_i^r, Y_i^n, Y_i^f, Y_i^e)_{i=1}^{N}\}$, where X_i include the textual content of research paper and research idea, and Y_i^n, Y_i^f, Y_i^e are the labels which show the scores of novelty, feasibility, and effectiveness of research idea. We could utilize this training set to **train three reward models** as follows,

$$\begin{cases} F_n = \mathcal{R}_n(X_i^r, Y_i^n|\Theta_n), \\ F_f = \mathcal{R}_f(X_i^r, Y_i^f|\Theta_f), \\ F_e = \mathcal{R}_e(X_i^r, Y_i^e|\Theta_e) \end{cases} \tag{1}$$

where $\Theta_{n/f/e}$ is the parameters of the reward model $\mathcal{R}_{n/f/e}$. $\mathcal{R}^{n/f/e}$ denotes reward models that aim to score the novelty, feasibility, and effectiveness of the research idea. $F_{n/f/e}$ is reward values from reward models. Then, we use a set of N_f research papers $\{P_i\}_{i=1}^{N_f}$ as input to the language model to generate research ideas, which are assessed with reward models based on three criteria: novelty, feasibility, and effectiveness. Finally, we conduct reinforcement learning on the language model as,

$$H = \mathcal{M}(P|\Theta_m, \Theta_n, \Theta_f, \Theta_e), \tag{2}$$

where Θ_m is final optimized parameters of the language model $\mathcal{M}$. During which the dimensional controllers are jointly trained to improve its ability to generate high-quality research ideas with fine-grained control at inference time. During this process, three dimensional controllers are trained jointly with the language model to enable fine-grained control at inference time.

Supervised Fine-Tuning

To make the model training more stable in reinforcement learning (Chen et al. 2024a), we also introduce the supervised fin-tuning stage.

Data Collection. To conduct a Supervised Fine-Tuning stage, we first collect papers from the ICLR 2023 and 2024. We selected papers from ICLR as training data due to its prestigious standing as a top-tier conference in the field of machine learning, offering cutting-edge research and high-quality technical discussions. We sample 1,000 instances $P = \{p\}$ for training. We utilize the LLaMA with a prompt (detailed in appendix) to extract the research idea y from the sampled paper p as the golden output. To extract the one corresponding input paper x for each output, we select the one most significant supporting paper from all related works $\hat{x_1}, \hat{x_2}..., \hat{x_n}$ by prompting LLaMA of the abstract and introduction section of p, together with the citation counts of $\hat{x_1}, \hat{x_2}..., \hat{x_n}$ within the sampled paper p.

Fine-Tuning. Based on the collected training set, we fine-tune the language model $\mathcal{M}$ as follows,

$$\mathcal{L}_{sup} = CE(Y, \hat{Y}) \tag{3}$$

where $CE(\cdot)$ denotes the cross-entropy loss and $\hat{Y}$ is the predicted research idea from $\mathcal{M}$, formulated as $\hat{Y} = \mathcal{M}(X)$. X is the research paper and Y is the research idea.

Reward Modeling

Researchers mainly consider three aspects when they devise a research idea: novelty, feasibility, and effectiveness. Therefore, we train three distinct reward models to score the generated idea in reinforcement learning, each corresponding to one of the quality dimensions.

Multi-dimension Feedback Collection. To train reward models, we need to collect three kinds of feedback. Similar to the supervised fine-tuning stage, we also use the papers from ICLR[2]. Specifically, we collect the review data from OpenReview platform[3], and we also get the research idea by prompting the language model $\mathcal{M}$. For the Novelty score of the research idea in ICLR 2023, we could use the novelty score from the review directly. As for ICLR 2024, we prompt the LLM to get novelty scores since they don't provide direct ratings (see Appendix for prompt). Similarly, since there is no feasibility score or effectiveness score in the review, we prompt the LLM to get scores for every research idea. Feasibility score is mainly based on the experiment setup and method sections, taking into account factors such as dataset size, model complexity, and relevant review comments, while Effectiveness score is derived primarily from the experimental results and corresponding review comments. The detailed Scoring Criteria for Novelty, Feasibility, and Effectiveness are outlined in Appendix .

Notably, all the collected novelty, feasibility, and effectiveness are subsequently normalized to a 0-1 scale for training.

Reward Model Training. We select an LLM as the backbone of reward models. To make the model predict the score for each dimension, we add a Multi-Layer Perceptron as follows,

$$\begin{cases} \mathbf{F}_{n/f/e} = \mathcal{A}_{n/f/e}(X^r), \\ \hat{F}_{n/f/e} = \mathcal{C}_{n/f/e}(\mathbf{F}_{n/f/e}), \end{cases} \tag{4}$$

where $\mathcal{C}_{n/f/e}$ are MLPs which can output score for each dimension. $\mathcal{A}_{n/f/e}$ is the LLM backbone. Each reward model takes the generated idea as input and outputs a score $F_{n/f/e}$ between 0 and 1, representing its evaluation of novelty, feasibility, or effectiveness. To optimize the reward models, we utilize cross-entropy loss as follows,

$$\mathcal{L}_{n/f/e} = CE(\hat{F}_{n/f/e}, F_{n/f/e}), \tag{5}$$

where $F_{n/f/e}$ is the ground-truth label.

Multi-dimension Reward Augmented Controllable Reinforcement Learning

In this stage, we fine-tune the research idea proposer with controllable steering through reinforcement learning **??**, refining the model based on feedback across three dimensions: novelty, feasibility, and effectiveness.

Dimensional Controllers Inspired by the existing work (Han et al. 2024), we introduce the dimensional controllers of the novelty, feasibility, and effectiveness of the generated idea, as these dimensions often exhibit interdependency and

[2]https://iclr.cc/

[3]https://docs.openreview.net/reference/api-v2.

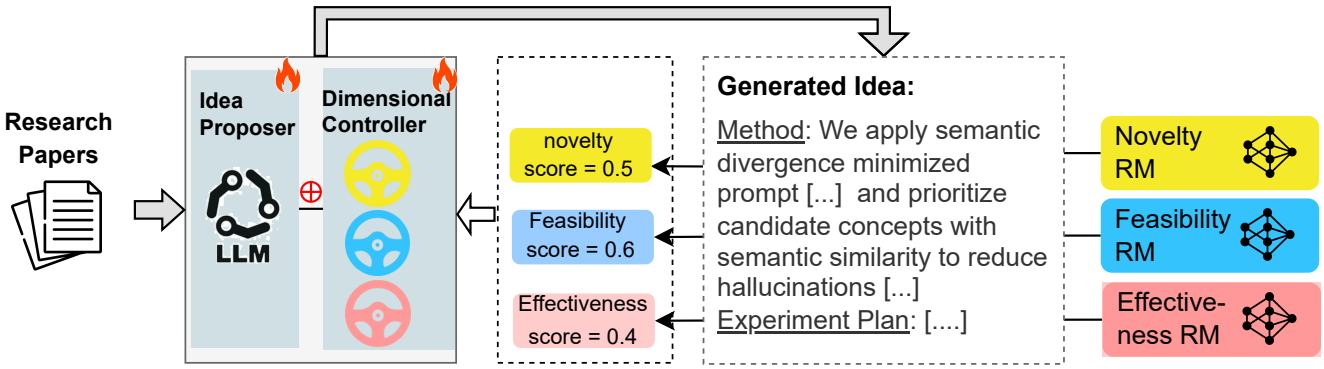

Figure 2: The learning framework with dynamic control across 3 dimensions. Generated research ideas are assessed by corresponding reward models, which provide scores for each dimension. These scores guide the fine-tuning process during reinforcement learning, optimizing both the idea proposer and the corresponding dimensional control parameters to enhance the quality of idea generation.

trade-offs. We achieve this by adding additional control parameters (i.e. the steers) as follows,

$$\begin{cases} \mathbf{M}_n^l = \mathbf{M}_l + \epsilon_n \mathbf{W}_n \mathbf{M}_l, \\ \mathbf{M}_f^l = \mathbf{M}_l + \epsilon_f \mathbf{W}_f \mathbf{M}_l, \\ \mathbf{M}_e^l = \mathbf{M}_l + \epsilon_e \mathbf{W}_e \mathbf{M}_l, \end{cases} \tag{6}$$

where $\mathbf{M}_l$ represents the output of l-th layer in the LLM. ϵ_n, ϵ_f, and ϵ_e are the hyper-parameters for controlling novelty, feasibility, and effectiveness. $\mathbf{W}_n$, $\mathbf{W}_f$, and $\mathbf{W}_e$ are learnable parameters. In the training stage, we set all ϵ_n, ϵ_f, and ϵ_e as 1. By this, we use $\mathbf{M}_{n/f/e}^l$ to replace the original output of the l-th layer. We denote the parameters for each resulting model as $\Theta_n = \{\Theta_{LLM}, \Theta_{\epsilon_n \mathbf{W}_n \mathbf{M}_l}\}$, $\Theta_f = \{\Theta_{LLM}, \Theta_{\epsilon_f \mathbf{W}_f \mathbf{M}_l}\}$ and $\Theta_e = \{\Theta_{LLM}, \Theta_{\epsilon_e \mathbf{W}_e \mathbf{M}_l}\}$.

Reward. Specifically, we get all three kinds of rewards for each research idea based on the well-trained reward model. We define r_n, r_f, and r_e as the novelty, feasibility, and effectiveness rewards for the research idea. Then we have a reward function for each dimension of the research idea at timestep t as follows,

$$\begin{cases} r_t^n = -\sum_{i=1}^{t} \mathbb{I}(i = K) w_l r_n, \\ r_t^f = -\sum_{i=1}^{t} \mathbb{I}(i = K) w_l r_f, \\ r_t^e = -\sum_{i=1}^{t} \mathbb{I}(i = K) w_l r_e, \end{cases} \tag{7}$$

where K is the token length of the research idea. t is the timestep. $\mathbb{I}(\cdot)$ is the indicator function. w_l is a weight assigned to rewards. Thereafter, we utilize the PPO algorithm (Schulman et al. 2017) to train the model following the existing work (Jing and Du 2024). More details of PPO algorithm can be found in Appendix.

Decoding

In this part, we devise two decoding methods for the inference stage.

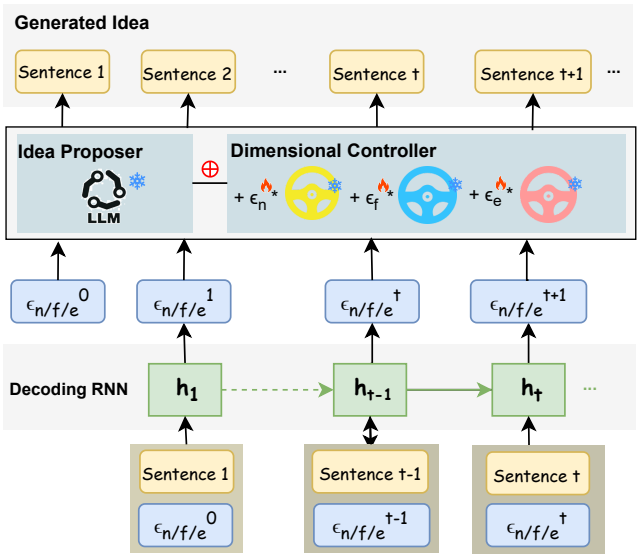

Figure 3: Decoding RNN dynamically steers the dimensions for a balanced and context-aware generation. It starts with $(\epsilon_n^0, \epsilon_f^0, \epsilon_e^0)$, and predicts the control parameter weights for the next sentence, based on the context generated by the combined proposer and controller.

Naive Static Decoding. In this decoding method, we set ϵ_n, ϵ_f, and ϵ_e as fixed values for the steers. To achieve a high score over novelty, feasibility, and effectiveness, we set all ϵ_n, ϵ_f, and ϵ_e as 1, because we set them as 1 in the training stage for maximum novelty, feasibility, and effectiveness.

Goal-driven Dynamic Decoding. The goal of achieving a good research idea is not only to blindly improve the result of a certain dimension but also to consider the overall quality. For example, too high a degree of novelty may result in a low effectiveness (Si, Yang, and Hashimoto 2024a), while different parts of a research idea, such as the method and experiment planning, may require varying levels of focus on novelty and feasibility. Therefore, how to balance novelty, feasibility, and effectiveness in the inference stage is important for generating a good idea. To achieve this, we utilize an

RNN (Sherstinsky 2020) to predict the steer value ϵ_n, ϵ_f, and ϵ_e, because RNN is good at sequence-level prediction (Figure 3).

To optimize the RNN for steer values prediction, we first collect 1,000 high-quality research ideas generated with Idea Proposer (above 8 in overall score). hereafter, we get the corresponding controller weights using our three reward models for each sentence of the high-quality research idea. Specifically, we feed each sentence in the research idea into our reward models to get the rewards as $\hat{r}_n, \hat{r}_f, \hat{r}_e$. Furthermore, we normalize the reward and get the corresponding steer values of each sentence as $\hat{\epsilon}_{n/f/e} = (\hat{r}_{n/f/e} - s_{n/f/e})/(a_{n/f/e} - s_{n/f/e}) \times w'$, where $s_{n/f/e}$ and $a_{n/f/e}$ are the minimum value and maximum value for all rewards and w' is the maximal controller weight, which is 5 in our case. This reflects the controller-weight ratios between 3 controllers, as well as the absolute scale of each controller weight from 0-5. After the data collection, we can use the pair $(S^t, s^{t+1}_{n/f/e})$ to train the model as follows,

$$\mathcal{L}_{rnn} = CE(RNN(S^{<t}), s^t_{n/f/e}), \tag{8}$$

where $S^{<t}$ is the previous $t-1$ sentences in the research idea and $s^t_{n/f/e}$ is steer values ϵ_n, ϵ_f, and ϵ_e of t-th sentence. Therefore, we can use the well-trained RNN to predict the controller weights of the next sentence based on the current generated sentence in the inference phrase.

Experiment

Dataset and Analysis

We collect a dataset of 6,765 usable research papers in total submitted to ICLR[4]and NeurIPS[5] in the years 2023 and 2024, including both accepted and rejected submissions and filtered 5,687 usable data. 4,271 of them are used for training, and 500 are sampled for evaluation. Each paper contains its abstract, methodology, and experiment sections. Additionally, review data from OpenReview[6] provides human ratings for overall quality as well as the review contents and key sub-dimensions - novelty, feasibility, and effectiveness. Paper content is scraped with title from the Semantic Scholar[7] and arXiv APIs[8] and then cleaned up with regular expression to extract corresponding sections. These papers and ratings are used to: 1. Derive ground-truth ideas for supervised fine-tuning. 2. Train reward models for the key dimensions. 3. Optimize idea generation using reinforcement learning with multi-dimensional steering.

The dataset is split into the following subsets:

1. ***Supervised Fine-Tuning* split.**: We use 1,000 papers from only ICLR to derive the golden generated idea, paired with the most supporting related work idea as input to fine-tune the model.
2. ***Reinforcement Learning* split.**: 3,271 research papers from both ICLR and NeurIPS with detailed reviews are used to train three distinct reward models for novelty, feasibility, and effectiveness, each capturing expert evaluations for further reinforcement learning.
3. ***Evaluation* split.**: 500 research papers from both ICLR and NeurIPS are sampled for evaluation, of which 30 are randomly selected for manual expert evaluation.

Figures 4 and 5 provide an overview of the dataset distribution and top keywords.

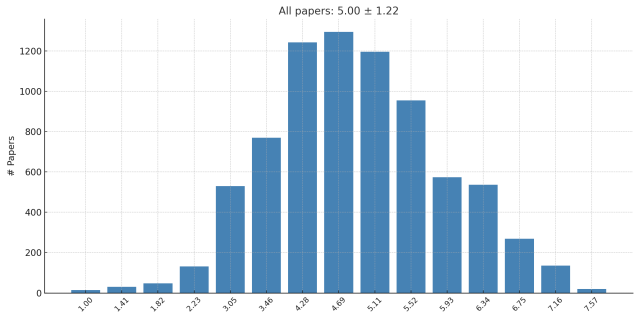

Figure 4: Rating distribution statistics of our dataset.

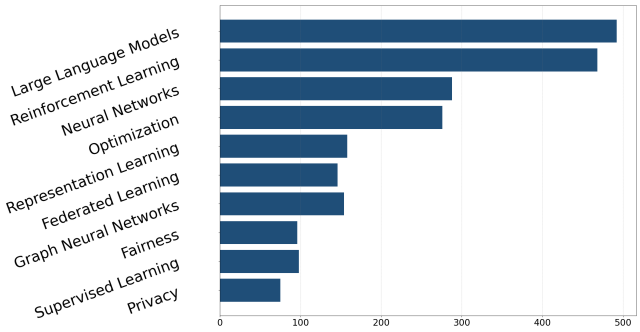

Figure 5: Top 10 topic distribution of our dataset.

Evaluation The evaluation is performed on two datasets: 500 papers of the evaluation split for automatic evaluation, and a subset of 30 papers are selected for manual expert evaluation. We measure performance across three core metrics (details in Appendix):

- **Novelty**: Evaluates how original and creative the generated ideas are, compared to existing works.
- **Feasibility**: Assesses the practical implementation and the likelihood that the idea can be executed within typical resource constraints.
- **Effectiveness**: Measures the potential improvement or impact of the generated idea when compared to baseline models.

We split our evaluation into two types:

1. **Automatic Evaluation**: For automatic evaluation, we evaluate novelty, feasibility, and effectiveness of the generated ideas with prompt-based method. We adopt GPT-4 as our reviewing agent.

[4]https://iclr.cc/
[5]https://neurips.cc/
[6]https://docs.openreview.net/reference/api-v2.
[7]https://www.semanticscholar.org/product/api.
[8]https://arxiv.org/help/api.

Model	Novelty	Feasibility	Effectiveness	Overall
T5-SFT	3.3	5.1	4.2	4.2
T5-RLHF	3.9	5.3	4.9	4.7
LLaMA2-SFT	4.8	5.9	5.2	5.3
LLaMA2-RLHF	5.5	6.1	5.6	5.8
LLaMA2-RLHF + Novelty Ctrl	6.4	5.9	5.5	6.0
LLaMA2-RLHF + Feasibility Ctrl	5.3	7.2	5.2	5.6
LLaMA2-RLHF + Effectiveness Ctrl	5.6	6.0	6.4	5.9
LLaMA2-RLHF + All Ctrls (Static)	5.8	6.0	5.5	5.9
LLaMA2-RLHF + All Ctrls (Dynamic)	6.0	6.1	5.8	6.2

Table 1: Experiment Results with Novelty (*N*), Feasibility (*F*), Effectiveness (*E*), and Overall Scores. *N/F/E Ctrl (abbrev. for Control)* represents that only 1 corresponding controller is enabled, whereas *All Ctrl* activate all the 3 controllers. Static and dynamic denote different decoding strategies.

2. **Manual Evaluation**: For manual evaluation, we select 30 papers and have domain experts assess the quality of the generated ideas of the selected model (SFT, RLHF and RLHF with Dynamic Controls), providing human scores for novelty, feasibility, and effectiveness. These scores are then compared with the scores generated by our automatic reviewing agent to measure the alignment between human judgment and the agent's reviews.

Main Experiments

Baselines and Setups We establish several baselines to evaluate the effectiveness of different control strategies applied to the LLaMA2-RLHF model. The baselines include T5-SFT, T5-RLHF, and LLaMA2-SFT, representing varying levels of model capacity and reinforcement learning application. These baselines are chosen to compare the impact of applying reinforcement learning fine-tuning (RLHF) and enabling targeted controls for Novelty, Feasibility, and Effective.

- ***T5-SFT***: A version of the T5 model trained using SFT on 1,000 examples, without reinforcement learning or control strategies, in which ideas are generated based on the prompt structure, serving as the simplest baseline.
- These baselines allow for a comprehensive comparison, highlighting the incremental improvements brought by RLHF, control strategies, and advanced decoding techniques.

  For the RL for idea proper and dimensional controllers training, we use The RL split to optimize our model using PPO and multi-dimensional reward augmentation. We incorporate the three distinct reward models for novelty, feasibility, and effectiveness, allowing for controllable generation combined with 3 control parameters, and experiment with different decoding stategies.

Main Results

Main Results and Statistical Analysis

Table 1 presents the experimental results for *Novelty (N)*, *Feasibility (F)*, *Effectiveness (E)*, and *Overall* metrics.

The baseline models establish foundational performance levels, with *T5-SFT* and *T5-RLHF* showing modest improvements in *Feasibility* and *Effectiveness* due to reinforcement learning, though their *Novelty* scores remain limited by the lack of mechanisms to encourage innovation. In contrast, *LLaMA2-SFT* achieves higher overall scores, benefiting from larger model capacity and superior pretraining, yet its reliance on supervised fine-tuning leaves room for enhancement through reinforcement learning and control strategies.

Adding targeted controls to *LLaMA2-RLHF* demonstrates the potential for metric-specific optimizations. For instance, introducing *Novelty Control* significantly boosts creativity while maintaining balanced practicality and performance, highlighting the feasibility of improving originality without major trade-offs. Similarly, *Feasibility Control* achieves the highest observed feasibility, albeit with minor reductions in novelty and effectiveness, showcasing its focus on practicality. The *Effectiveness Control*, on the other hand, enhances impact without compromising the balance across dimensions.

When all controls are combined, *Static Decoding* provides reliable, balanced performance, but its fixed nature limits its adaptability. In contrast, *Dynamic Decoding* emerges as the most effective approach, leveraging contextual dynamic strategy to balance creativity, practicality, and impact, ultimately producing higher-quality ideas.

These results show the importance of rl and dynamic control in tailoring model behavior to complex requirements, while also illustrating trade-offs inherent in single-metric optimizations.

To validate the observed improvements, we conducted paired t-tests to evaluate statistical significance. Results show that *LLaMA2-RLHF + Novelty Ctrl* achieved a statistically significant improvement in *Novelty* (p-value $<$ 0.01) compared to *LLaMA2-RLHF* without controls. Similarly, *Feasibility Ctrl* significantly enhanced *Feasibility* (p-value < 0.05), while *Effectiveness Ctrl* showed a notable gain in *Effectiveness* (p-value < 0.05). Furthermore, *Dynamic Decoding* demonstrated statistically significant improvements across all metrics (p-value < 0.01) com-

Model	Novelty	Feasibility	Effectiveness	Overall
LLaMA2-SFT	4.3	5.6	4.8	4.6
LLaMA2-RLHF	4.9	6.2	5.2	5.3
LLaMA2-RLHF + Dynamic	5.5	6.4	5.1	5.5

Table 2: Human evaluation results, *LLaMA2-RLHF + Dynamic* denotes the Dynamic decoding with all the 3 controllers enabled.

pared to the static approach, validating its superior adaptability and performance.

Human Evaluation

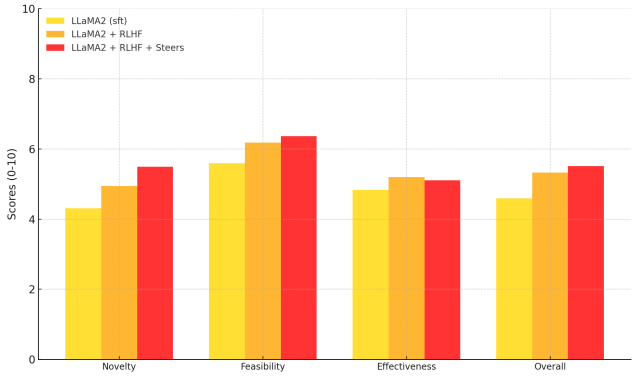

Figure 6: Human Evaluation Results

Metrics	Novelty	Feasibility	Effectiveness	Overall
Pearson (r)	0.995	0.972	0.839	0.970
Spearman (p)	1.000	0.866	0.500	1.000

Table 3: Correlation Coefficients (Pearson and Spearman) between human and reviewing agent scores.

Domain experts validated the effectiveness of our framework of the generated idea as shown in Table 2 and bar plot in Figure 6, with human scores showing a strong correlation with the automatic scores produced by our reward models. The Correlation Coefficients computed with both Pearson and Spearman between human and reviewing agent scores are shown in table 3.

Experts also highlighted the trade-off between novelty and feasibility, noting that the fine-tuned model with novelty steering produced more creative, though sometimes less practical, ideas compared to the equal-weighted model.

Analysis

Novelty and Feasibility Trade-off

We learn from (Si, Yang, and Hashimoto 2024b) that increasing novelty will likely reduce the feasibility of an idea. To test this idea, we controlled the weight of the novelty steer in the RLHF + Steer1 setup and observed its impact on both novelty and feasibility scores. The results are shown in Table 4.

Novelty Weight	Novelty Score	Feasibility Score
1.0	6.4	6.1
2.0	6.7	5.8
3.0	7.0	5.3
4.0	7.3	4.9

Table 4: Novelty and Feasibility trade-off by increasing the novelty controller weight.

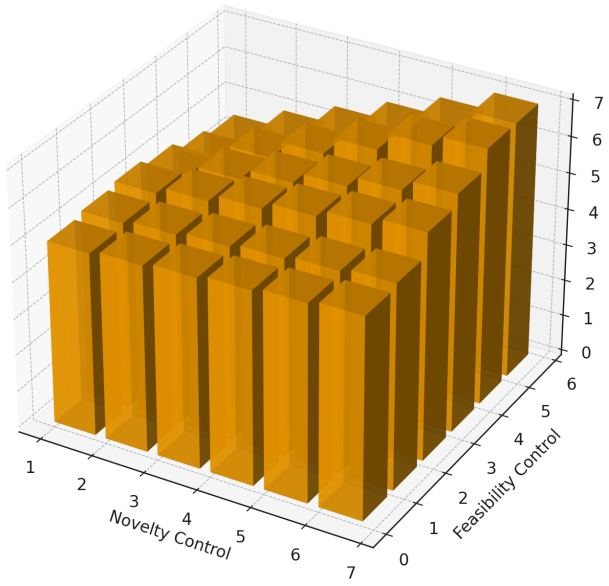

Figure 7: Novelty and Feasibility control analysis

As expected, increasing the novelty steer weight led to higher novelty scores but lower feasibility scores. This demonstrates the trade-off between generating highly creative ideas and ensuring their practical feasibility.

Decoding Strategy Motivation

Dynamic decoding adapts research ideation outputs to the varying demands of different parts of the idea, as shown in Figure 8. The observed novelty jump at the 6th sentence illustrates a shift in focus, aligning feasibility with experiment plan while reducing emphasis on novelty. By dynamically adjusting decoding weights, this strategy ensures that the generated ideas are coherent, contextually aligned, and balanced across key dimensions.

Model	Idea (Method part)	Novelty / Feasibility / Effectiveness	Overall
T5-SFT	Proposing a reinforcement learning algorithm with stochastic agent interactions, focusing on decentralized learning in dynamic environments. The method avoids shared policies and uses predefined heuristics for adaptability.	3.3 / 6.0 / 4.2	3.8
textitLLaMA2-SFT	Developing a reinforcement learning model that employs implicit environmental feedback for agent collaboration. The method eliminates the need for direct communication protocols and uses fixed reward functions for learning.	4.8 / 5.9 / 5.2	5.3
LLaMA2-RLHF	Introducing a reinforcement learning algorithm that combines stochastic interactions with an adaptive reward mechanism. This method enables efficient multi-agent collaboration in dynamic environments while ensuring scalability and practical feasibility.	5.5 / 6.2 / 5.6	5.8
LLaMA2-RLHF + Dynamic	Presenting a multi-agent reinforcement learning approach where agents utilize minimal communication protocols and enhanced environmental feedback. The method dynamically adjusts learning strategies to improve scalability and effectiveness in real-world applications.	6.3 / 6.4 / 6.8	6.6

Table 5: Comparison of method part of ideas and scores. Mdel and reviewing agent settings are the same as main experiment.

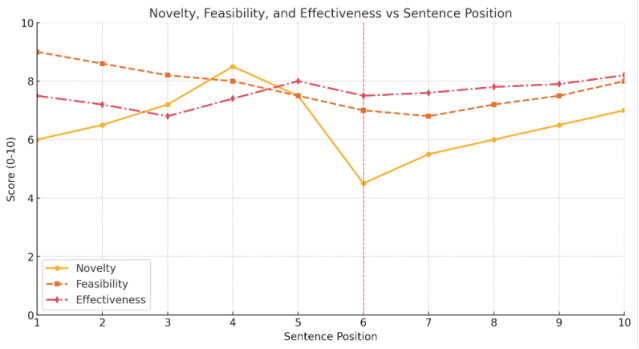

Figure 8: Dimensional variation w.r.t. normalized sentence position (1-10 according to idea length).

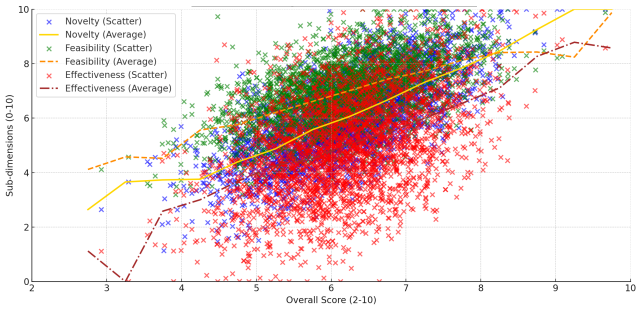

Figure 9: Scatters of different dimensions virus overall scores.

Case Study

Table 5 compares the evolution of ideas generated by models, progressing from SFT to advanced configurations with dynamic control. Baseline models with SFT exhibit moderate feasibility but struggle to achieve a balance between novelty and effectiveness, highlighting their limitations in fostering creative yet practical solutions. With RL fine-tuning, LLaMA2-RLHF demonstrates clear improvements across all metrics, leveraging reward mechanisms to enhance collaboration of fine-grained dimensions. The addition of dynamic control strategies further elevates performance, with LLaMA-RLHF + Dynamic achieving the highest overall score through dynamic adjustments that seamlessly balance creativity, feasibility, and impact. This progression underscores the potential of RL fine-tuning combined with context-aware dynamic control for innovative, practical, and highly effective idea generation.

Conclusion

In this work, we introduced a novel LLM-based framework for research idea generation that optimizes and dynamically balances key metrics—novelty, feasibility, and effectiveness—through a two-stage process combining supervised fine-tuning and controllable reinforcement learning. By leveraging multi-dimensional reward models and integrating the dimensional controller with sentence-level dynamic decoding, our approach effectively navigates the improvement and the inherent trade-offs among these metrics, ensuring context-aware and high-quality idea generation. Comprehensive evaluations, including human studies, highlight the robustness and effectiveness of our method, giving the path for more advanced and controllable systems in automated research ideation.

References

2024. An Analysis of Large Language Models: Their Impact and Potential Applications. *Knowledge and Information Systems*.

Askell, A.; Bai, Y.; Chen, A.; Drain, D.; Ganguli, D.; Henighan, T.; Joseph, N.; Kadavath, S.; Kernion, J.; and Conerly, J. H. 2021. A general language assistant as a laboratory for alignment. *arXiv preprint arXiv:2112.00861*.

Authors, V. 2023. Abstractive Summarization with Deep Reinforcement Learning Using Semantic Similarity Rewards. *Natural Language Engineering*.

Baek, J.; Jauhar, S. K.; Cucerzan, S.; and Hwang, S. J. 2024. ResearchAgent: Iterative Research Idea Generation over Scientific Literature with Large Language Models. *ArXiv*, abs/2404.07738.

Beltagy, I.; Lo, K.; and Cohan, A. 2019. SciBERT: A pretrained language model for scientific text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 3615–3620.

Bornstein, M.; and Singh, R. 2024. HypothesisCraft: Towards Automated Hypothesis Generation and Refinement Using LLMs. *Journal of AI Research*.

Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J. D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33: 1877–1901.

Chen, J.; Han, X.; Ma, Y.; Zhou, X.; and Xiang, L. 2024a. Unlock the Correlation between Supervised Fine-Tuning and Reinforcement Learning in Training Code Large Language Models. *CoRR*, abs/2406.10305.

Chen, X.; et al. 2024b. Can LLMs Generate Novel Research Ideas? A Large-Scale Human Study with 100+ NLP Researchers. *arXiv preprint arXiv:2409.04109*.

Christiano, P. F.; Leike, J.; Brown, T. B.; Martic, M.; Legg, S.; and Amodei, D. 2017. Deep reinforcement learning from human preferences. *Advances in Neural Information Processing Systems*, 30.

Du, X.; Kim, A.; and Park, M. 2023. ResearchAgent: Automating Scientific Discovery with LLMs and Feedback Loops. *Proceedings of NeurIPS*.

Glaese, A.; McAleese, N.; Trager, S.; Askell, A.; Bai, Y.; Chen, A.; Drain, D.; Elhage, N.; Ganguli, D.; Godwin, J.; et al. 2022. Improving alignment of dialogue agents via targeted human judgements. *arXiv preprint arXiv:2209.14375*.

Han, C.; Xu, J.; Li, M.; Fung, Y.; Sun, C.; Jiang, N.; Abdelzaher, T.; and Ji, H. 2024. Word Embeddings Are Steers for Language Models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 16410–16430. Association for Computational Linguistics.

Hope, T.; Portnoff, R.; Briscoe, E.; Krishnan, R.; Murphy, K.; Cohen, W. W.; Anandan, P.; Yates, A.; Benton, A.; Daumé III, H.; et al. 2021. SciSight: Combining faceted navigation and research group detection for COVID-19 exploratory scientific search. *arXiv preprint arXiv:2005.12683*.

Huang, Q.; Vora, J.; Liang, P.; and Leskovec, J. 2024. MLAgentBench: Evaluating Language Agents on Machine Learning Experimentation. In *ICML*.

Jing, L.; and Du, X. 2024. FGAIF: Aligning Large Vision-Language Models with Fine-grained AI Feedback. arXiv:2404.05046.

Lee, J.; Yoon, W.; Kim, S.; Kim, D.; Kim, S.; Kim, C.; and Kang, J. 2020. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4): 1234–1240.

Li, R.; Patel, T.; Wang, Q.; and Du, X. 2024. MLR-Copilot: Autonomous Machine Learning Research based on Large Language Models Agents. *ArXiv*, abs/2408.14033.

Lu, C.; Lu, C.; Lange, R. T.; Foerster, J.; Clune, J.; and Ha, D. 2024. The AI Scientist: Towards Fully Automated Open-Ended Scientific Discovery . *ArXiv*, abs/2408.06292.

Madaan, A.; Zhang, S.; Gupta, P.; Zhao, S.; and Rush, A. M. 2022. Memory-assisted prompt editing to improve language model performance. *arXiv preprint arXiv:2211.08859*.

Menick, J.; Uesato, J.; Reichert, D.; Tils, D.; Huang, P.-S.; Aravind, A. N.; Cai, T.; Mellor, J.; Anderson, K. C.; et al. 2022. Teaching language models to support answers with verified quotes. *arXiv preprint arXiv:2212.04891*.

Nakano, R.; Hilton, J.; Balaji, S.; Wu, J.; Ouyang, L.; Kim, C.; Jones, M.; Saunders, W.; Hesse, C.; Elhage, N.; et al. 2021. WebGPT: Browser-assisted question-answering with human feedback. *arXiv preprint arXiv:2112.09332*.

Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; Wainwright, C. L.; Mishkin, P.; Zhang, C.; Agarwal, S.; Slama, K.; Ray, A.; et al. 2022. Training language models to follow instructions with human feedback. *arXiv preprint arXiv:2203.02155*.

Park, J. S.; O'Brien, J.; Cai, C. J.; Morris, M. R.; Liang, P.; and Bernstein, M. S. 2023. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th annual acm symposium on user interface software and technology*, 1–22.

Qi, B.; Zhang, K.; Li, H.; Tian, K.; Zeng, S.; Chen, Z.-R.; and Zhou, B. 2023. Large language models are zero shot hypothesis proposers.

Rafailov, R.; Ie, E.; Liu, Y.; and Liang, P. 2023. Direct preference optimization: Your language model is secretly a reward model. *arXiv preprint arXiv:2305.18290*.

Raghu, M.; and Schmidt, E. 2020. A survey of deep learning for scientific discovery. *arXiv preprint arXiv:2003.11755*.

Scheurer, L.; Schleier-Smith, J.; Chaffin, M.; Salomon, D.; and Christiano, P. F. 2022. Reinforcement learning with human feedback: Leveraging preferences and demonstrations in training reward models. *arXiv preprint arXiv:2212.05185*.

Schulman, J.; Wolski, F.; Dhariwal, P.; Radford, A.; and Klimov, O. 2017. Proximal Policy Optimization Algorithms. *CoRR*, abs/1707.06347.

Sherstinsky, A. 2020. Fundamentals of recurrent neural network (RNN) and long short-term memory (LSTM) network. *Physica D: Nonlinear Phenomena*, 404: 132306.

Si, C.; Yang, D.; and Hashimoto, T. 2024a. Can LLMs Generate Novel Research Ideas? A Large-Scale Human Study with 100+ NLP Researchers. *CoRR*, abs/2409.04109.

Si, C.; Yang, D.; and Hashimoto, T. 2024b. Can LLMs Generate Novel Research Ideas? A Large-Scale Human Study with 100+ NLP Researchers. arXiv:2409.04109.

Stiennon, N.; Ouyang, L.; Wu, J.; Ziegler, D. M.; Lowe, R.; Voss, C.; Radford, A.; Amodei, D.; and Christiano, P. F. 2020. Learning to summarize with human feedback. *Advances in Neural Information Processing Systems*, 33: 3008–3021.

Tian, M.; Gao, L.; Zhang, S. D.; Chen, X.; Fan, C.; Guo, X.; Haas, R.; Ji, P.; Krongchon, K.; Li, Y.; Liu, S.; Luo, D.; Ma, Y.; Tong, H.; Trinh, K.; Tian, C.; Wang, Z.; Wu, B.; Xiong, Y.; Yin, S.; Zhu, M.; Lieret, K.; Lu, Y.; Liu, G.; Du, Y.; Tao, T.; Press, O.; Callan, J.; Huerta, E. A.; and Peng, H. 2024. SciCode: A Research Coding Benchmark Curated by Scientists. *ArXiv*, abs/2407.13168.

Uesato, J.; Reichert, D.; Mellor, J.; Tils, D.; Huang, P.-S.; Cai, T.; Aravind, A. N.; Desai, S.; Moats, I.; Glaese, A.; et al. 2022. Fine-grained reward models for helpful and harmless assistants. *arXiv preprint arXiv:2209.11895*.

Wang, J.; and Zhou, L. 2023. AI-Scientist: Iterative Planning and Search for Scientific Discovery with Large Language Models. *arXiv preprint arXiv:2308.12345*.

Wang, Q.; Downey, D.; Ji, H.; and Hope, T. 2024a. SciMON: Scientific Inspiration Machines Optimized for Novelty. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL2024)*.

Wang, Q.; Downey, D.; Ji, H.; and Hope, T. 2024b. SciMON: Scientific Inspiration Machines Optimized for Novelty. In *ACL*.

Wu, Z.; Hu, Y.; Shi, W.; Dziri, N.; Suhr, A.; Ammanabrolu, P.; Smith, N. A.; Ostendorf, M.; and Hajishirzi, H. 2023. Fine-Grained Human Feedback Gives Better Rewards for Language Model Training. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems*.

Yang, H.; Zhao, Y.; Wu, Y.; Wang, S.; Zheng, T.; Zhang, H.; Che, W.; and Qin, B. 2024a. Large Language Models Meet Text-Centric Multimodal Sentiment Analysis: A Survey. *arXiv preprint arXiv:2406.08068*.

Yang, Z.; Du, X.; Li, J.; Zheng, J.; Poria, S.; and Cambria, E. 2023. Large language models for automated open-domain scientific hypotheses discovery. *arXiv preprint arXiv:2309.02726*.

Yang, Z.; Du, X.; Li, J.; Zheng, J.; Poria, S.; and Cambria, E. 2024b. Large Language Models for Automated Open-domain Scientific Hypotheses Discovery. *ACL Findings*.

Zhang, W.; Deng, Y.; Liu, B.; Pan, S. J.; and Bing, L. 2023. Sentiment Analysis in the Era of Large Language Models: A Reality Check. *arXiv preprint arXiv:2305.15005*.

Zhong, R.; Zhang, P.; Li, S.; Ahn, J.; Klein, D.; and Steinhardt, J. 2023. Goal driven discovery of distributional differences via language descriptions. *Advances in Neural Information Processing Systems*, 36: 40204–40237.

Zhou, S.; Xu, F. F.; Zhu, H.; Zhou, X.; Lo, R.; Sridhar, A.; Cheng, X.; Ou, T.; Bisk, Y.; Fried, D.; et al. 2023. Webarena: A realistic web environment for building autonomous agents. *arXiv preprint arXiv:2307.13854*.

Zhu, X.; Gardiner, S.; Roldán, T.; and Rossouw, D. 2024. The Model Arena for Cross-lingual Sentiment Analysis: A Comparative Study in the Era of Large Language Models. *arXiv preprint arXiv:2406.19358*.

Ziegler, D. M.; Stiennon, N.; Wu, J.; Brown, T.; Radford, A.; Amodei, D.; and Christiano, P. F. 2019. Fine-tuning language models from human preferences. *arXiv preprint arXiv:1909.08593*.

Appendix

PPO

To optimize our idea proposer, we utilize Proximal Policy Optimization (PPO), an actor-critic RL algorithm widely used in previous RLHF works. PPO enables the proposer (i.e. the policy model) to be refined against multiple reward models that simulate human feedback, ensuring high-quality idea generation. In PPO, the value model $V_\psi(s_t)$ estimates the expected cumulative reward for a given state s_t, providing a baseline for the advantage function. The proposer is optimized with a PPO clipped surrogate training objective. The advantage A_t at timestep t is estimated by a generalized advantage estimation function (**?**): $A_t = \sum_{t'=t}^{T} (\gamma\lambda)^{t'-t}(r_{t'} + \gamma V_\psi(s_{t'+1}) - V_\psi(s_{t'}))$, with γ as a hyperparameter and λ as the discounting factor for rewards. r_t is the reward assigned to a_t, which in our case is acquired using multiple learned reward models. The value model $V_\psi(s_t)$ is optimized with an expected squared-error loss with the value target as $V_{\text{targ}}(s_t) = \sum_{t'=t}^{T-1} \gamma^{t'-t} r_{t'} + \gamma^{T-t} V_{\psi_{\text{old}}}(s_T)$, where $V_{\psi_{\text{old}}}$ is the lagging value model. Finally, PPO is trained to optimize both the proposer ($\mathcal{M}_\theta$) and value (V_ψ) models with their respective objectives. No reward model is being optimized during PPO training. See Algorithm 1 in the Appendix for more details.

Definition of Novelty, Feasibility, and Effectiveness

This appendix provides detailed definitions and scoring guidelines for **Novelty**, **Feasibility**, and **Effectiveness**—the three primary dimensions used to evaluate research ideas.

1. Novelty

Novelty evaluates how different a proposed research idea is compared to existing works. Following previous work (), the guidelines for scoring are as follows:

- **1**: *Not novel at all* — The idea is identical to many existing works.
- **3**: *Mostly not novel* — Very similar ideas already exist.
- **5**: *Somewhat novel* — There are differences, but not enough for a standalone paper.
- **6**: *Reasonably novel* — Notable differences, potentially sufficient for a new paper.
- **8**: *Clearly novel* — Major differences from all existing ideas.
- **10**: *Highly novel* — Highly different and creative in a clever, impactful way.

2. Feasibility

Feasibility measures how practical it is to execute the proposed idea within 1–2 months under the following assumptions:

- Ample access to OpenAI/Anthropic APIs.
- Limited GPU computing resources.

Scoring guidelines:

- **1**: *Impossible* — The idea or experiments are fundamentally flawed.
- **3**: *Very challenging* — Major flaws or significant resource limitations.
- **5**: *Moderately feasible* — Possible with careful planning and modifications.
- **6**: *Feasible* — Achievable with reasonable planning.
- **8**: *Highly feasible* — Straightforward to implement and run.
- **10**: *Easy* — Quick to implement without requiring advanced skills.

3. Effectiveness

Effectiveness assesses the likelihood of the research idea achieving meaningful experimental performance improvement. The scoring is defined as:

- **1**: *Extremely unlikely* — Significant flaws, almost certain to fail.
- **3**: *Low effectiveness* — Limited potential, might work in very specific scenarios.
- **5**: *Somewhat ineffective* — A slight chance of marginal or inconsistent improvement.
- **6**: *Somewhat effective* — A decent chance of moderate improvement on certain benchmarks.
- **8**: *Probably effective* — Likely to deliver significant improvement on benchmarks.
- **10**: *Definitely effective* — Highly likely to outperform existing benchmarks by a substantial margin.

To ensure reliability, we require the model to provide:

1. A brief justification for the score (minimum 2–3 sentences).
2. References to related works, especially if the score is low.

Detailed Algorithm for Multi-dimension reward augmented RL

Prompt for Research Idea Extraction

System Prompt: You are an AI assistant whose primary goal is to extract specific details from scientific literature to aid researchers in understanding and replicating the methodologies and experiment plans of the work.

User Message

You are tasked with extracting the **Method** and **Experiment Plan** from an academic paper. These should include:

- **Method**: A concise summary of the methodological approach employed in the study.

Algorithm 1: Multi-dimension reward augmented Reninformace Learning

Input: Initial policy model $\mathcal{M}_{\theta_{init}}$; initial value model $V_{\psi_{init}}$; 3 well-trained reward models $\mathcal{R}_{n/f/e}$; task prompts $\mathcal{D}$; hyperparameters γ, λ, ϵ

Output: Updated policy models $\mathcal{M}_{\theta_{n/f/e}}$.

Initialize policy model $\mathcal{M}_{\theta_{n/f/e}} \leftarrow \mathcal{M}_{\theta_{init}}$, value model $V_{\psi}^{n/f/e} \leftarrow V_{\psi_{init}}$

for step $= 1, \ldots, M$ **do**

Sample a batch $\mathcal{D}_b$ from $\mathcal{D}$

Sample output sequence $y_n^n \sim \mathcal{M}_{\theta_n}(\cdot \mid x^n)$, $y_f^n \sim \mathcal{M}_{\theta_f}(\cdot \mid x^n)$, $y_e^n \sim \mathcal{M}_{\theta_e}(\cdot \mid x^n)$ for each prompt $x^n \in \mathcal{D}_b$

Compute rewards $\{r_t^{n/f/e}\}_{t=1}^{|y^n|}$ for each sampled output y_n^n, y_f^n, y_e^n by running $\mathcal{R}^{o/a/r}$

Compute advantages $\{A_t^{o/a/r}\}_{t=1}^{|y^n|}$ and value targets $\{V_{\text{targ}}^{o/a/r}(s_t)\}_{t=1}^{|y^n|}$ for each y_n^n, y_f^n, y_e^n with V_ψ

for PPO iteration $= 1, \ldots, \mu$ **do**

Update the policy model by maximizing the PPO clipped surrogate objective for $\mathcal{M}_{\theta_{n/f/e}}$:

$$\theta \leftarrow \arg\max_{\theta} \frac{1}{|\mathcal{D}_b|} \sum_{n=1}^{|\mathcal{D}_b|} \frac{1}{|y^n|} \sum_{t=1}^{|y^n|} \min(\frac{\mathcal{M}_\theta(a_t \mid s_t)}{\mathcal{M}_{\theta_{\text{old}}}(a_t \mid s_t)} A_t, \text{clip}(v_t, 1-\epsilon, 1+\epsilon) A_t)$$

end for

Update the value model $\psi_{n/f/e}$ by minimizing a square-error objective:

$$\psi \leftarrow \arg\min_{\psi} \frac{1}{|\mathcal{D}_b|} \sum_{n=1}^{|\mathcal{D}_b|} \frac{1}{|y^n|} \sum_{t=1}^{|y^n|} (V_\psi(s_t) - V_{\text{targ}}(s_t))^2$$

end for

- **Experiment Plan**: Key details of the experiment, including dataset preparation, baseline implementation, and evaluation metrics or procedures.

Ensure that the output is clear, focused, and formatted to align with the given structure.

Input Details

I am going to provide the target paper, related papers, and entities as follows:

- **Target paper title:** {`paper['title']`}
- **Target paper abstract:** {`paper['abstract']`}
- **Entities:** {`Entities`}

Objective

With the provided target paper and entities, extract and summarize the **Method** and **Experiment Plan** in the following format:

- **Method:** [Provide a concise description of the methodology used in the study.]
- **Experiment Plan:** [Summarize the dataset preparation, baseline implementation, and evaluation procedures.]

Example Input

- **Target paper title:** "Transformer Models for Legal Text Analysis"
- **Target paper abstract:**

  "Deep learning has transformed the field of natural language processing, yet challenges remain in domain-specific applications. This paper explores the use of transformer models for legal text analysis, addressing the question: 'Can pre-trained language models be adapted effectively for legal case prediction?' The study employs fine-tuning techniques and evaluates performance on a benchmark dataset of legal cases. Results show a significant improvement in prediction accuracy compared to traditional methods."

Expected Output

- **Method:** We introduce fine-tuning techniques to adapt pre-trained transformer models for legal text analysis.
- **Experiment Plan:**
  - **Dataset Preparation:** A legal benchmark dataset of case documents is used.
  - **Baseline Implementation:** Models are compared against traditional NLP methods.
  - **Evaluation Procedure:** Performance is measured in terms of prediction accuracy on unseen legal cases.

Prompt for Novelty Score Extraction

System Prompt: You are a specialized assistant for scientific text evaluation. Your task is to evaluate the novelty of scientific papers.

User Prompt

Based on the following information about a scientific paper, please evaluate its novelty:

- **Title:** {`title`}
- **Abstract:** {`abstract`}

- **Related Works (top 3 from citations since 2023):** {recent_works}
- **Review Comments:** {reviews}

Novelty Evaluation Prompt

Evaluate how creative and different the idea is compared to existing works on the topic. Consider all papers that appeared online prior to July 2024 as existing work. Your evaluation should consider the degree to which the paper brings new insights and differentiates itself from prior research.

Scoring Criteria

Please assign a novelty score on a scale from 1 to 10 based on the following criteria:

Novelty Definition

Following the definition of the previous work (), we score the novelty of papers based on how different it is from existing works. The guidelines for scoring novelty are:

- **1:** Not novel at all — many existing ideas are the same.
- **3:** Mostly not novel — very similar ideas exist.
- **5:** Somewhat novel — differences exist but not enough for a new paper.
- **6:** Reasonably novel — notable differences, could lead to a new paper.
- **8:** Clearly novel — major differences from all existing ideas.
- **10:** Very novel — highly different and creative in a clever way.

Novelty Rationale

After assigning a score, provide a short justification for your rating. If the score is below 6, specify similar works that closely resemble this paper. The rationale should be at least 2-3 sentences.

Output Format

The result must be output in JSON format, as shown in the example below:

```
{"score": 8, "reason": "This paper introduces a novel machine learning approach for earthquake prediction using real-time seismic data, which represents a significant improvement over traditional statistical models. By incorporating both real-time data and deep learning techniques, this approach enables more accurate and timely earthquake forecasts. Although there are existing works using machine learning for seismic analysis, the integration of real-time data and advanced neural networks distinguishes this paper. The comprehensive validation of the method, including comparisons with conventional models, highlights its contribution to the field."}
```

The response should **only contain JSON content**.

Prompt for Research Idea Generation

System Prompt: You are an AI assistant specializing in extracting and generating structured research ideas from scientific papers. Your task is to assist researchers in developing concise, clear, and innovative research ideas based on the provided input.

User Instructions: You are tasked with generating a structured research idea that includes:

- **Method:** A concise summary of the methodological approach employed in the study.
- **Experiment Plan:** Key details of the experiment, including dataset preparation, baseline implementation, and evaluation procedures.
- **Problem:** A clear statement of the research problem or gap the study aims to address.
- **Related Works:** Identify and summarize the top 3 most relevant related works, emphasizing how the target paper builds upon or differs from them.

Ensure that the output adheres to the following requirements:

1. **Contextual Relevance:** The generated idea must align with the main theme of the provided paper and incorporate any specified entities or constraints.
2. **Clarity and Structure:** The output must be structured, clear, and concise, formatted as follows:

   Problem: [Description of the research problem or gap being addressed.]
   Method: [Concise description of the methodology used in the study.]
   Experiment Plan:
   - Dataset Preparation: [Details of the dataset used.]
   - Baseline Implementation: [Details of the baseline setup.]
   - Evaluation Procedure: [Evaluation metrics and procedures used.]

   Related Works:
   - **Work 1:** [Summary of the first related work.]

- **Work 2:** [Summary of the second related work.]
- **Work 3:** [Summary of the third related work.]

Example Input:

- **Target Paper Title:** "Transformer Models for Legal Text Analysis"
- **Abstract:** "This study explores fine-tuning transformer models for legal text analysis and evaluates their performance on a benchmark dataset, achieving significant accuracy improvements over traditional methods."
- **Problem:** Traditional NLP methods often fail to capture the complex linguistic structure and contextual dependencies in legal text, leading to suboptimal accuracy in legal text analysis tasks.
- **Entities:** Legal datasets, transformer models, benchmark evaluation.
- **Related Works:**
  - Work 1: "BERT for Legal Case Prediction" focuses on fine-tuning BERT models for legal document classification.
  - Work 2: "Legal NLP with Statistical Models" applies traditional NLP techniques for legal text analysis.
  - Work 3: "Adapting Transformers for Domain-Specific Tasks" investigates transformer models in specialized fields like healthcare and law.

Example Output:

Problem: Traditional NLP methods often fail to capture the complex linguistic structure and contextual dependencies in legal text, leading to suboptimal accuracy in legal text analysis tasks.

Method: We introduce fine-tuning techniques to adapt pre-trained transformer models for legal text analysis, focusing on improved generalization.

Experiment Plan:

- **Dataset Preparation:** A benchmark dataset of legal case documents is preprocessed and tokenized.
- **Baseline Implementation:** Traditional NLP methods are used as the baseline for comparison.
- **Evaluation Procedure:** Prediction accuracy is measured on unseen legal cases using cross-validation techniques.

Related Works:

- **Work 1:** "BERT for Legal Case Prediction" explores fine-tuning BERT for classification, but lacks transformer-level insights specific to domain challenges.
- **Work 2:** "Legal NLP with Statistical Models" applies rule-based methods but achieves lower accuracy and generalizability compared to transformer models.
- **Work 3:** "Adapting Transformers for Domain-Specific Tasks" provides foundational techniques but does not address challenges in legal text structure.

Prompt for Automatic Evaluation

System Prompt: You are an AI reviewer specializing in evaluating the quality of research ideas based on specific criteria: **Novelty**, **Feasibility**, and **Effectiveness**. Your task is to assess each criterion and provide structured feedback for automatic evaluation.

User Instructions: For a given research idea, evaluate the following dimensions:

1. **Novelty:** Assess how creative and unique the idea is compared to existing works.
2. **Feasibility:** Evaluate the practicality of executing the idea within typical resource constraints.
3. **Effectiveness:** Judge the potential of the idea to achieve its intended objectives or performance improvements.

Scoring Criteria: Provide a score between 1 and 10 for each dimension, adhering to these guidelines: {Add detailed definition of 3 Metrics HERE}

Evaluation Output Requirements: Provide a structured evaluation as follows:

- Score for each dimension (**Novelty**, **Feasibility**, **Effectiveness**).
- Brief justification (minimum 2–3 sentences) for each score.
- If the score is below 6, include references to related works or specific reasons for the low rating.

Example Input:

- **Title:** "Transformer Models for Legal Text Analysis"
- **Abstract:** "This paper explores fine-tuning transformer models for legal text analysis, demonstrating significant accuracy improvements over traditional methods."
- **Generated Idea:**

  Method: Fine-tune pre-trained transformer models for legal case prediction. Experiment Plan: Use a benchmark legal dataset, traditional NLP methods as baselines, and evaluate using prediction accuracy.

Example Output:

```
{ "novelty": 8,
"novelty_justification": "The idea introduces transformer-based approaches to legal text analysis, offering a clear improvement over rule-based and statistical methods.",
"feasibility": 6,
"feasibility_justification": "Implementation is feasible with access to pre-trained models and benchmark datasets, though computational cost may be a concern.",
"effectiveness": 7,
"effectiveness_justification": "The method has a high likelihood of outperforming traditional baselines based on prior research in similar domains."
}
```