# LDC: Learning to Generate Research Ideas with Dynamic Control

Anonymous ACL submission

## Abstract

Recent advancements in large language models (LLMs) have demonstrated their potential in automating the scientific research ideation. Existing approaches primarily focus on prompting techniques, often producing ideas misaligned with expert standards – novelty, feasibility, and effectiveness, which are widely recognized by the research community as the three key subdimensions of high-quality ideas. Also, balancing these dimensions remains challenging due to their inherent trade-offs. To address these limitations, we propose the first framework that employs a two-stage approach combining Supervised Fine-Tuning (SFT) and controllable Reinforcement Learning (RL) for the task. In the SFT stage, the model learns foundational patterns from pairs of research papers and their corresponding follow-up ideas. In the RL stage, multi-dimensional reward models guided by fine-grained feedback evaluate and optimize the model across key dimensions. During inference, dimensional controllers coordinated by a sentence-level decoder enable dynamic context-aware steering of the idea generation process. Our framework provides a balanced approach to research idea generation, achieving high-quality outcomes in the experiment by dynamically navigating the trade-offs among novelty, feasibility, and effectiveness.

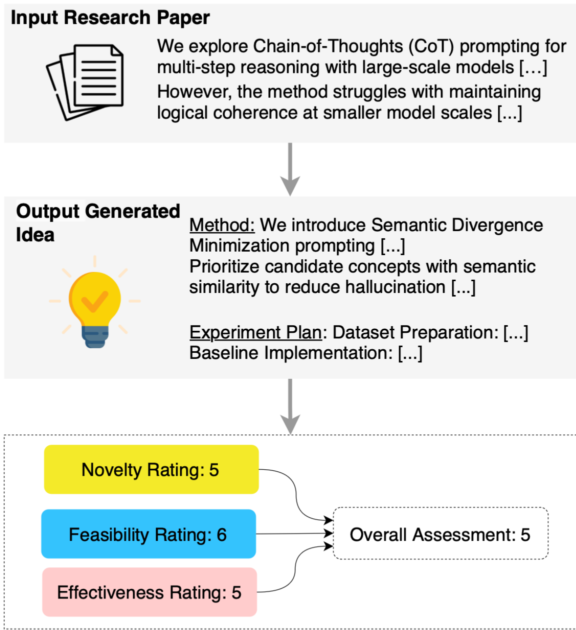



Figure 1: Research idea generation from research papers. Each idea is measured across the dimensions of novelty, feasibility, and effectiveness.

## 1 Introduction

Typically, a well-developed scientific research idea (or hypothesis[1]) consists of a *methodology* and an *experiment plan*, as illustrated in Figure 1. The *methodology* introduces the novel concept or approach, while the *experiment plan* provides a structured guide for its validation. Formulating such research ideas is fundamental to the research process. Traditional methods, which rely heavily on human intuition and experience, are often time-consuming and prone to biases. In contrast, automated research idea generation systems can swiftly synthesize vast data and insights, uncovering novel connections beyond human researchers. Recent work using LLM-based agents has demonstrated their potential for generating and validating innovative ideas (Baek et al., 2025; Bornstein and Singh, 2024). Despite the notable progress, these efforts primarily rely on pre-trained models without task-specific learning, which restricts the full exploitation of optimizing the generated content toward scientific expert standards.

Recent studies and expert interviews show that novelty, feasibility and effectiveness are widely recognized by the research community as the three key subdimensions of high-quality research ideas (Si et al., 2024; Baek et al., 2025). Specifically, novelty reflects the originality of the idea; feasibility assesses its practicality given current resources and constraints; and effectiveness measures the like-

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

lihood that the idea will achieve its intended outcomes. These fine-grained metrics, alongside the overall rating, can help evaluate ideas and guide generation through optimization techniques such as reinforcement learning (RL); more specifically, Reinforcement Learning from Human Feedback (RLHF) can be used to optimize LLM toward scientist standards (Ouyang et al., 2022). Despite these advancements, existing approaches cannot tackle the complex interdependence and inherent restrictions among these dimensions. One notable challenge identified is to reveal the inevitable *innovation-feasibility trade-off* (Yang et al., 2024b; Si et al., 2024): highly novel ideas often lack feasibility, while overly feasible ideas tend to limit the scope for groundbreaking discoveries. Optimizing idea generation towards each of the key dimensions while achieving a balanced trade-off remains a critical yet unresolved question.

To address this, we propose a framework to improve the intrinsic capabilties of LLMs on generating research ideas. It dynamically adjusts the emphasis on key dimensions of the research idea to achieve high overall quality through a two-stage training process: SFT and controllable RL. In the SFT stage, the idea proposer learns foundational patterns by training on pairs of research papers and corresponding follow-up ideas. In the RL stage, we employ multi-dimension reward modeling as a real-world assessment approximation (Wu et al., 2023). Reward models, trained on automatically obtained fine-grained feedback from review data, score each dimension–providing detailed guidance for model refinement. To enable precise and adaptive control, we introduce dimensional controllers, trained alongside the RL process, which adjust the generation to prioritize specific dimensions when necessary. This is done 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 part 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 introduce a two-stage fine-tuning framework for LLM-based research ideation, which dynamically optimizes idea generation towards novelty, feasibility, and effectiveness.
- We introduce a dynamic decoding to address interdependent subdimensions such as novelty and feasibility.
- We leverage automatically collected real-world data to train reward models that provide automated, fine-grained feedback aligned with expert evaluations.
- We conduct comprehensive evaluations, which demonstrate the effectiveness of our method for optimized and controllable research idea generation.

## 2 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 et al., 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., 2024b; 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., 2025), and multi-module retrieval and revision (Yang et al., 2024b) have been proposed to advance ideation capabilities beyond baseline prompting methods. Beyond ideation, other researchers leverage LLMs for automating experimental workflows. Works like MLAgent (Huang et al., 2024) and SciCode (Tian et al., 2024) use LLMs to generate code for executing research experiments, while 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),

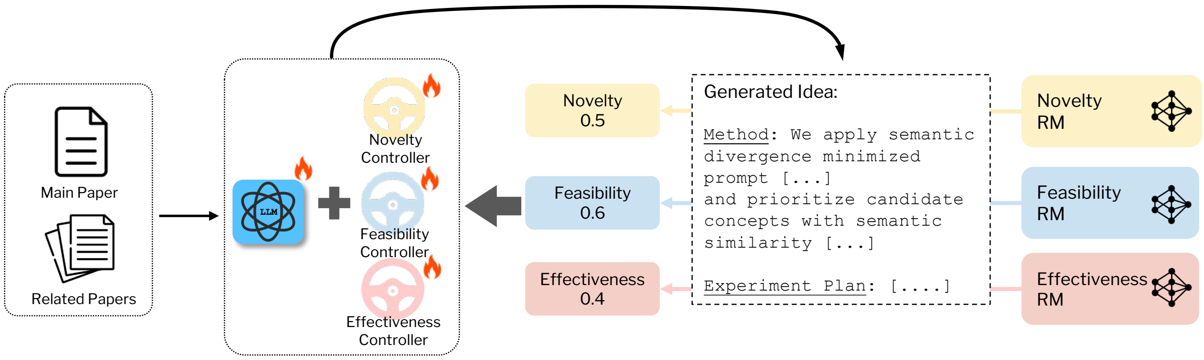



Figure 2: The learning framework with dynamic control across 3 dimensions. Generated research ideas are assessed by corresponding reward models, which provide scores for each dimension. These scores guide the fine-tuning process during reinforcement learning, optimizing both the idea proposer and the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

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 rewards modeling for multiple specific attributes, such as reasoning or ethical considerations (Glaese et al., 2022; Uesato et al., 2022). In this work we investigate fine-grained rewards for the more challenging problem of optimizing multiple dimensions.

## 3 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.

### 3.1 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 the reward models that aim to score the novelty, feasibility, and effectiveness of the research idea. $F_{n/f/e}$ are 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. 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 $\Theta_m$ 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.

### 3.2 Supervised Fine-Tuning

To improve model training stability in RL(Chen et al., 2024), we also introduce the supervised fine-tuning stage. The goal of this stage is to introduce the model with the general task format and stabilize the subsequent RL stage. Therefore, the training data at this stage does not need to achieve high scores in terms of the metrics, which will later be optimized through the fine-grained RL.

**Data Collection.** To conduct a supervised fine-tuning stage, we need to collect a set of research papers $\{X_i\}_{i=1}^{N}$, which we name as supporting papers, and a collection of research ideas $\{Y_i\}_{i=1}^{N}$, each inspired by a corresponding supporting paper. To collect high-quality research ideas, we first collect papers from ICLR 2023 and 2024. As a top-tier conference in the field of machine learning that covers diverse domains and topics, ICLR is renowned for its cutting-edge research and high-quality technical discussions, making it an ideal source for this purpose. We sample 1,000 instances of papers $\{p\}$, and then utilize the LLaMA with a prompt (detailed in Appendix J) to extract the research idea $y$ from the sampled paper $p$ as the golden output. To extract the one corresponding supporting paper $X_i$, i.e. the input of each extracted research idea $Y_i$, 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 $p$. For all extraction, we use LLaMA3 70B to ensure high-quality results

**Fine-Tuning.** Based on the collected training set $\mathcal{D} = \{X_i, Y_i\}_{i=1}^{N}$, 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 $\hat{Y} = \mathcal{M}(X)$.

### 3.3 Reward Modeling

Researchers mainly consider three aspects when they devise research ideas: novelty, feasibility, and effectiveness. These aspects are also used in the review process as fine-grained dimensions of research ideas besides an overall quality. 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 use the papers from ICLR[2] and NeurIPS[3] due to their availability and high quality. Specifically, we collect the review data from OpenReview, and we extract the research ideas also with prompting. For the Novelty score of the research ideas in the year 2023, we could use the novelty score from the review directly. As for those in the year 2024, we prompt Llama3 to get novelty scores since they don't provide direct ratings (see Appendix K for prompts). Similarly, since there is no feasibility score or effectiveness score in the review, we prompt Llama3 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. For all extraction with Llama3 we use the 70B API. The detailed Scoring Criteria for Novelty, Feasibility, and Effectiveness are outlined in Appendix F.

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.

### 3.4 Multi-dimension Reward Augmented Controllable Reinforcement Learning

In this stage, we fine-tune the research idea proposer with controllable steering through reinforcement learning (Figure 2), 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 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. $\epsilon_n$, $\epsilon_f$, and $\epsilon_e$ 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 $\epsilon_n$, $\epsilon_f$,

[2]https://iclr.cc/
[3]https://neurips.cc/

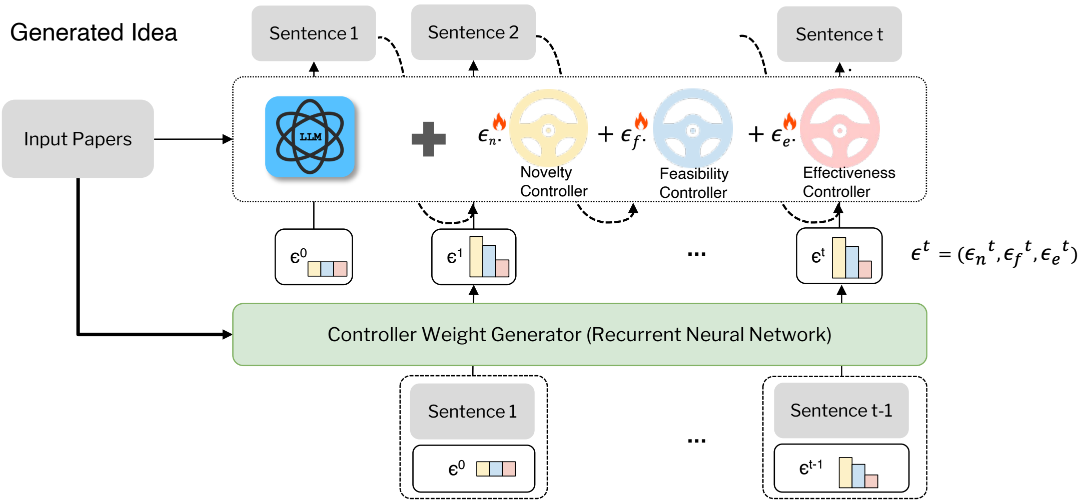



Figure 3: Decoding RNN dynamically steers the dimensions for a balanced and context-aware generation. The process starts with $\epsilon^0$ and predicts the control weights for the next sentence condition on the generated context.

and $\epsilon_e$ as 1. By this, we use $\mathbf{M}^l_{n/f/e}$ 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 are in Appendix A.

### 3.5 Decoding

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

**Naive Static Decoding.** In this decoding method, we set $\epsilon_n$, $\epsilon_f$, and $\epsilon_e$ as fixed values for the steers. To achieve a high score over novelty, feasibility, and effectiveness, we set all $\epsilon_n$, $\epsilon_f$, and $\epsilon_e$ 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 improve the result of a certain dimension but also to consider the overall quality. For example, very high degree of novelty may result in low effectiveness (Si et al., 2024; Yang et al., 2024a), while different parts of a research idea, such as method and experiment planning, may require varying levels of focus on novelty and feasibility. Therefore, how to balance novelty, feasibility, and effectiveness during inference is important. To achieve this, we utilize a recurrent neural network (RNN) (Sherstinsky, 2020) to predict the steer value $\epsilon_n$, $\epsilon_f$, and $\epsilon_e$ (Figure 3), as RNN is good at sequence-level prediction.

To optimize the RNN for steer values prediction, we first collect 1,000 high-quality research ideas generated with Idea Proposer (scoring above 8 overall). Thereafter, 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 to reflect the controller-weight ratios between three controllers, as well as the absolute scale of each controller weight from 0.0–5.0. The corresponding steer values of each sentence $s_t$ are computed as: $\hat{\epsilon}_{n/f/e} = (\hat{r}_{n/f/e} - r_{\min})/(r_{\max} - r_{\min}) \cdot \epsilon_{\max}$ where $r_{\min}$ and $r_{\max}$ denote the minimum and maximum value for all rewards, and $\epsilon_{\max}$ is the maximal controller weight. After the data collection, we can use the pair $(S^t, \hat{\epsilon}^t_{n/f/e})$ to train the model:

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

where $S^{<t}$ is the precedding $t$-1 sentences generated in the research idea. Afterward, we use the trained RNN to predict the controller weights of the next sentence $\epsilon^t = (\epsilon_n^t, \epsilon_e^t, \epsilon_f^t)$ based on $\epsilon^{t-1}$ and previous sentence.

Finally, during inference, we apply the controller weights by adding them on top of the LLM last layer embedding $\mathbf{M}_l$ to steer the generation:

$$\mathbf{M}_n^l = \mathbf{M}_l + \epsilon_n \mathbf{W}_n \mathbf{M}_l + \epsilon_f \mathbf{W}_f \mathbf{M}_l, + \epsilon_e \mathbf{W}_e \mathbf{M}_l \tag{9}$$

# 4 Experiment

## 4.1 Dataset

We collect 6,765 research papers from ICLR and NeurIPS (2023–2024), including both accepted and rejected submissions, and filtered 5,687 usable data. These papers cover diverse ML-related domains and topics. Each paper includes *Abstracts, Methodology, and Experiment* sections[4], supplemented with human reviews automatically obtained from OpenReview[5] for novelty, feasibility, effectiveness, and overall ratings. Statistics of topics and rating distributions are reported in Appendix E. The dataset is split into the following subsets: 1) *Supervised Fine-Tuning* split: 1,000 ICLR papers to derive the golden ideas and the most supporting paper for fine-tuning; 2) *Reinforcement Learning* split: 3,271 papers with detailed reviews to train reward models for novelty, feasibility, and effectiveness; and 3) *Evaluation* split: 500 sampled papers for evaluation, including 30 randomly selected for manual expert review.

To ensure data reliability, we implement multi-stage quality control across automated extraction, retrieval, and filtering. We conduct a manual audit of 100 examples on topical match, plausibility, and completeness. Full details of the data processing and quality checks are provided in Appendix C.

## 4.2 Evaluation Settings

The evaluation is conducted following the settings in recent works (Si et al., 2024; Baek et al., 2025). We evaluate three key dimensions—*novelty, feasibility, and effectiveness*—using both automatic and manual evaluation following the standard definition from OpenReview and human study (Si et al., 2024) as covered in Appendix F.

**Automatic Evaluation.** Following the recent trends in using LLMs to judge the quality of generated ideas (Yang et al., 2024a; Baek et al., 2025), we use a prompt-based method with *GPT-4* as the reviewing agent to score the generated ideas on all three metrics. Different from their reference-free evaluations, we employ retrieval-augmented evaluation by fetching the latest related work from Semantic Scholar to ensure more faithful evaluations, especially for novelty. We further validate the validity of this approach by measuring its correlation with human expert ratings.

**Manual Evaluation.** For manual evaluation, we randomly select a subset of 30 papers and have 15 domain experts across different institutes, recruited according to reviewer criteria adopted by leading conferences (e.g. NeurIPS, ACL, EMNLP) to independently assess ideas highly relevant to their field of expertise to assign a score for each criteria. Each idea is rated by three experts, and they are also required to provide written justifications for their ratings. We also report inter-annotator agreement and manual feedback examples as in Appendix B, along with further details on the recruiting criteria and annotation process.

## 4.3 Main Experiments

**Baselines and Setups.** We establish a comprehensive set of baselines to evaluate the effectiveness of different control strategies for the LLaMA2-RLHF model and to support ablation studies. Our baselines include *T5-SFT*, *T5-RLHF*, *LLaMA2-SFT*, and *ResearchAgent*, representing different model capacities and reinforcement learning configurations: ***T5-SFT*** is the simplest baseline–supervised fine-tuned T5 on 1,000 examples, without reinforcement learning or control mechanisms. ***T5-RLHF*** is fine-tuned with RLHF, but without dimensional controllers, to isolate the impact of RLHF. ***LLaMA2-SFT*** is LLaMA2-7B fine-tuned on 1,000 examples without reinforcement learning or control mechanisms. ***ResearchAgent*** is an agent-based baseline with GPT-4 (Baek et al., 2025) that iteratively generates ideas with retrieval from citation graph traversal and knowledge base. For RL and dimensional controllers training, we use The RL split to optimize the model with PPO and multi-dimension reward augmentation. The three reward models (novelty, feasibility, effectiveness) enable controllable generation via tunable control param-

[4] Paper content scraped from Semantic Scholar (https://www.semanticscholar.org/product/api) and ArXiv (https://arxiv.org/help/api) APIs, and then cleaned with regular expressions.

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

| Model | Novelty(N) | Feasibility(F) | Effectiveness(E) | Overall |
|---|---|---|---|---|
| *T5-SFT* | 3.3 | 5.1 | 4.2 | 4.0 |
| *T5-RLHF* | 3.8 | 5.3 | 4.8 | 4.5 |
| *LLaMA2-SFT* | 4.5 | 6.0 | 5.2 | 5.1 |
| *LLaMA2-RLHF* | 5.3 | 5.9 | 5.5 | 5.4 |
| *ResearchAgent* | 5.2 | 6.0 | 5.3 | 5.3 |
| *LLaMA2-RLHF + Novelty Ctrl* | 6.1* | 5.9 | 5.4 | 5.6 |
| *LLaMA2-RLHF + Feasibility Ctrl* | 5.1 | 6.8* | 5.0 | 5.5 |
| *LLaMA2-RLHF + Effectiveness Ctrl* | 5.2 | 5.9 | 6.2* | 5.6 |
| *LLaMA2-RLHF + All Ctrls (Static)* | 5.4 | 5.9 | 5.5 | 5.5 |
| *LLaMA2-RLHF + All Ctrls (Dynamic)* | 5.7* | 6.1* | 5.8* | 5.8* |

Table 1: Experiment results with Retrieval-Augmented evaluation. *N/F/E Control (Ctrl)* denotes only 1 corresponding controller enabled. *All Ctrl* activate all three controllers. * Significance checked with p-value $< 0.05$.

eters, and we experiment with various decoding strategies for ablation analysis.

**Main Results.** Table 1 summarizes the experimental results for novelty, feasibility, effectiveness, and overall scores. While baseline models (T5-SFT, T5-RLHF) show modest improvements in feasibility and effectiveness, their novelty remains limited. LLaMA2-SFT achieves higher overall scores due to its larger capacity and pretraining but benefits further from reinforcement learning and control strategy. Adding targeted control to LLaMA2-RLHF enables metric-specific optimization and enhances its respective target dimension: Novelty control boosts creativity, with feasibility setting enhances practicality, and effectiveness improves impact. Combining all controls, dynamic decoding outperforms the static approach across all metrics, balancing creativity, practicality, and impact effectively. Paired t-tests validate the significance. These results highlight the importance of RL and dynamic control in optimizing model performance across complex requirements. Notably, although ResearchAgent employs more advanced retrieval, our controllable models outperform it on all metrics, highlighting the effectiveness of controllable generation over complex retrieval alone.

### 4.4 Human Evaluation Results

| Model | N | F | E | Overall |
|---|---|---|---|---|
| *ResearchAgent* | 4.9 | 5.8 | 5.1 | 5.2 |
| *LLaMA2-SFT* | 4.2 | 5.6 | 4.6 | 4.4 |
| *LLaMA2-RLHF* | 4.9 | 6.0 | 5.1 | 5.3 |
| *LLaMA2-RLHF + Dynamic** | 5.3 | 6.2 | 5.2 | 5.6 |

Table 2: Human evaluation results. * This setting denotes dynamic decoding with all 3 controllers enabled.

The human evaluation is rigorously conducted according to the manual evaluation setting. Domain experts validated the effectiveness of our framework of generated ideas, as in Table 2, with human scores showing a strong correlation with the automatic scores produced by our reward models.

| Metrics | N | F | E | Overall |
|---|---|---|---|---|
| *Pearson (r)* | 0.997 | 0.772 | 0.744 | 0.884 |
| *Spearman (p)* | 0.949 | 0.949 | 0.949 | 1.000 |

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

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.

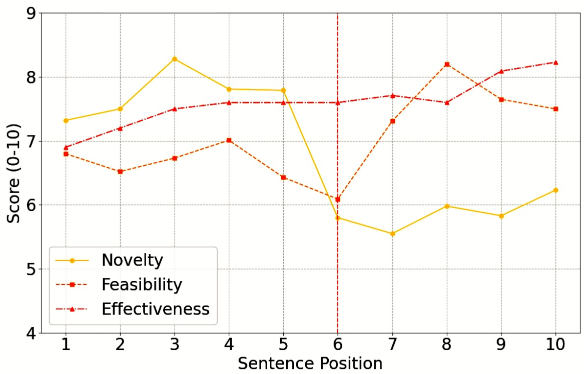



Figure 4: Dimensional variation w.r.t. normalized sentence position (1-10 according to idea length).

## 5 Analysis

### 5.1 Novelty and Feasibility Trade-off

Si et al. (2024) find that increasing novelty will likely reduce the feasibility of an idea. To test this

| 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 |
| *LLaMA2-SFT* | Developing a reinforcement learning model that employs implicit environmental feedback for agent collaboration. The method eliminates the need for direct communication 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 effectiveness in real-world applications. | 6.3 / 6.4 / 6.8 | 6.6 |

Table 4: Comparison of ideas (method part) and scores with all settings consistent with main experiments.

| Novelty Weight | N | F |
|---|---|---|
| 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 5: Novelty(N) and Feasibility(F) trade-off by increasing the novelty controller weight.

idea, we control the weight of the novelty steer in the RLHF + novelty ctrl setup and observed its impact on both novelty and feasibility scores. The results are shown in Table 5. As expected, increasing the novelty steer weight leads to higher novelty scores but lower feasibility scores. This demonstrates the trade-off between generating highly creative ideas and ensuring their practical feasibility.

### 5.2 Decoding Strategy Motivation

Dynamic decoding adapts research ideation outputs to the varying demands of different parts of the idea, as shown in Figure 4. Note that all the sentences are normalized to 1-10 and put in the nearest integer bracket for better averaging. The observed novelty jump in the 6th sentence illustrates a shift in focus, aligning feasibility with the experiment plan while reducing the emphasis on novelty. By dynamically adjusting decoding weights, this strategy ensures that the generated ideas are coherent, contextually aligned, and balanced across key dimensions.

### 5.3 Case Study and Other Analysis

Table 4 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, 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. We also include a novelty and feasibility control analysis and a scatter analysis in Appendix G and I.

## 6 Conclusion

We present a novel framework with LLM for research idea generation that optimizes and dynamically balances key dimensions—novelty, feasibility, and effectiveness—through a two-stage process combining supervised fine-tuning and controllable reinforcement learning. By leveraging multi-dimension 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 a path for more advanced and controllable systems in automated research idea generation.

## Limitations

Firstly, although the research ideas generated are of good quality, citation prediction could be explored as another way to judge their quality, which also makes it easier for human researchers to select promising ideas. Secondly, the interpretability of learned adjustments of dimension controllers is still a remaining open question for future exploration.

## References


Jinheon Baek, Sujay Kumar Jauhar, Silviu Cucerzan, and Sung Ju Hwang. 2025. Researchagent: Iterative research idea generation over scientific literature with large language models. *NAACL 2025*.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. Scibert: A pretrained language model for scientific text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3615–3620.

Michael Bornstein and Rahul Singh. 2024. Hypothesiscraft: Towards automated hypothesis generation and refinement using llms. *Journal of AI Research*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Jie Chen, Xintian Han, Yu Ma, Xun Zhou, and Liang Xiang. 2024. Unlock the correlation between supervised fine-tuning and reinforcement learning in training code large language models. *CoRR*, abs/2406.10305.

Paul F Christiano, Jan Leike, Tom B Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. Deep reinforcement learning from human preferences. *Advances in Neural Information Processing Systems*, 30.

Amal Glaese, Natasha McAleese, Samuel Trager, Amanda Askell, Yuntao Bai, Andy Chen, Dawn Drain, Nelson Elhage, Deep Ganguli, Jonathan Godwin, et al. 2022. Improving alignment of dialogue agents via targeted human judgements. *arXiv preprint arXiv:2209.14375*.

Chi Han, Jialiang Xu, Manling Li, Yi Fung, Chenkai Sun, Nan Jiang, Tarek Abdelzaher, and Heng Ji. 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)*, pages 16410–16430. Association for Computational Linguistics.

Tom Hope, Rebecca Portnoff, Erica Briscoe, Rahul Krishnan, Kevin Murphy, William W Cohen, Padmanabhan Anandan, Andrew Yates, Adrian Benton, Hal Daumé III, et al. 2021. Scisight: Combining faceted navigation and research group detection for covid-19 exploratory scientific search. *arXiv preprint arXiv:2005.12683*.

Qian Huang, Jian Vora, Percy Liang, and Jure Leskovec. 2024. Mlagentbench: Evaluating language agents on machine learning experimentation.

Liqiang Jing and Xinya Du. 2024. Fgaif: Aligning large vision-language models with fine-grained ai feedback.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Kim, and Jaewoo Kang. 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Ruochen Li, Teerth Patel, Qingyun Wang, and Xinya Du. 2024. MLR-Copilot: Autonomous Machine Learning Research based on Large Language Models Agents. *ArXiv*, abs/2408.14033.

Chris Lu, Cong Lu, Robert Tjarko Lange, Jakob Foerster, Jeff Clune, and David Ha. 2024. The AI Scientist: Towards Fully Automated Open-Ended Scientific Discovery . *ArXiv*, abs/2408.06292.

Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Matthew Jones, William Saunders, Chris Hesse, Nelson Elhage, et al. 2021. Webgpt: Browser-assisted question-answering with human feedback. *arXiv preprint arXiv:2112.09332*.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll L Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *arXiv preprint arXiv:2203.02155*.

Biqing Qi, Kaiyan Zhang, Haoxiang Li, Kai Tian, Sihang Zeng, Zhang-Ren Chen, and Bowen Zhou. 2023. Large language models are zero shot hypothesis proposers.

Maithra Raghu and Eric Schmidt. 2020. A survey of deep learning for scientific discovery. *arXiv preprint arXiv:2003.11755*.

John Schulman, Philipp Moritz, Sergey Levine, Michael I. Jordan, and Pieter Abbeel. 2016. High-dimensional continuous control using generalized advantage estimation. In *ICLR*.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *CoRR*, abs/1707.06347.

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

Chenglei Si, Diyi Yang, and Tatsunori Hashimoto. 2024. Can llms generate novel research ideas? a large-scale human study with 100+ nlp researchers.

Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel M Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. 2020. Learning to summarize with human feedback. *Advances in Neural Information Processing Systems*, 33:3008–3021.

Minyang Tian, Luyu Gao, Shizhuo Dylan Zhang, Xinan Chen, Cunwei Fan, Xuefei Guo, Roland Haas, Pan Ji, Kittithat Krongchon, Yao Li, Shengyan Liu, Di Luo, Yutao Ma, Hao Tong, Kha Trinh, Chenyu Tian, Zihan Wang, Bohao Wu, Yanyu Xiong, Shengzhu Yin, Min Zhu, Kilian Lieret, Yanxin Lu, Genglin Liu, Yufeng Du, Tianhua Tao, Ofir Press, Jamie Callan, E. A. Huerta, and Hao Peng. 2024. SciCode: A Research Coding Benchmark Curated by Scientists. *ArXiv*, abs/2407.13168.

Jonathan Uesato, David Reichert, Joe Mellor, David Tils, Po-Sen Huang, Trevor Cai, Arvind Neelakantan Aravind, Sankalp Desai, Isaac Moats, Amal Glaese, et al. 2022. Fine-grained reward models for helpful and harmless assistants. *arXiv preprint arXiv:2209.11895*.

Qingyun Wang, Doug Downey, Heng Ji, and Tom Hope. 2024a. Scimon: Scientific inspiration machines optimized for novelty.

Qingyun Wang, Doug Downey, Heng Ji, and Tom Hope. 2024b. SciMON: Scientific Inspiration Machines Optimized for Novelty. In *ACL*.

Zeqiu Wu, Yushi Hu, Weijia Shi, Nouha Dziri, Alane Suhr, Prithviraj Ammanabrolu, Noah A. Smith, Mari Ostendorf, and Hannaneh Hajishirzi. 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*.

Hao Yang, Yanyan Zhao, Yang Wu, Shilong Wang, Tian Zheng, Hongbo Zhang, Wanxiang Che, and Bing Qin. 2024a. Large language models meet text-centric multimodal sentiment analysis: A survey. *arXiv preprint arXiv:2406.08068*.

Zonglin Yang, Xinya Du, Junxian Li, Jie Zheng, Soujanya Poria, and E. Cambria. 2024b. Large Language Models for Automated Open-domain Scientific Hypotheses Discovery. *ACL Findings*.

Ruiqi Zhong, Peter Zhang, Steve Li, Jinwoo Ahn, Dan Klein, and Jacob Steinhardt. 2023. Goal driven discovery of distributional differences via language descriptions. *Advances in Neural Information Processing Systems*, 36:40204–40237.

Daniel M Ziegler, Nisan Stiennon, Jeffrey Wu, Tom Brown, Alec Radford, Dario Amodei, and Paul F Christiano. 2019. Fine-tuning language models from human preferences. *arXiv preprint arXiv:1909.08593*.

# A PPO and Detailed Algorithm for Multi-dimension reward augmented RL

**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 $\gamma$, $\lambda$, $\epsilon$

**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 $V_\psi$

**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**

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 (Schulman et al., 2016): $A_t = \sum_{t'=t}^{T} (\gamma\lambda)^{t'-t} (r_{t'} + \gamma V_\psi(s_{t'+1}) - V_\psi(s_{t'}))$, with $\gamma$ as a hyperparameter and $\lambda$ 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 ($V_\psi$) models with their respective objectives. No reward model is being optimized during PPO training. See Algorithm 1 for more details.

## B Manual Evaluation Details

For manual evaluation, we randomly select 30 papers and have 15 domain experts from different institutes (including several faculty members) assess the quality of the generated ideas for each model (SFT, RLHF, and RLHF with Dynamic Controls), with each idea independently annotated by three experts. To ensure the rigor and authority of human evaluation, all annotators meet widely accepted reviewer criteria used by leading conferences such as NeurIPS, ACL, and EMNLP. Specifically, our experts satisfy a combination of the following requirements:

- Hold a PhD or are authors of multiple peer-reviewed publications in relevant fields;
- Have at least two first-author publications in major conferences or journals (e.g., NeurIPS, ACL, EMNLP, ICML, ICLR, etc.) within the past five years;
- Have served as a reviewer in these conferences or journals, or have demonstrated substantial research expertise via citation record and research experience.

For each evaluation, annotators are required to provide a written justification for their ratings. On average, each evaluation took approximately three minutes to complete. Each expert provides human scores for novelty, feasibility, and effectiveness, which are then compared with those generated by our automatic reviewing agent to measure the alignment between human judgment and the agent's evaluations.

Additionally, we conduct inter-annotator agreement for all evaluation criteria to quantify the consistency among experts. The average Fleiss' kappa scores across all criteria are *novelty: 0.41*, *feasibility: 0.70*, and *effectiveness: 0.65*, reflecting good inter-annotator agreement. We also collect and analyze written feedback from annotators to better understand the qualitative aspects of novelty, feasibility, and effectiveness.
A representative example of human evaluation is below:

- **Idea:** *We tackle multimodal mental health assistance (text + voice tone + facial expressions). We introduce adaptive fine-tuning with emotion and sentiment feedback for state tracking; and incorporate trust and transparency feedback drawing insights from explainable AI. Experimental plan with setup, dataset, baselines, metrics, ablation, and expected results...*
- **Scores:** Novelty = 7, Feasibility = 6, Effectiveness = 8, Overall = 7
- **Feedback:** " This idea provides a novel multimodal setting to propose adaptive fine-tuning. It focuses on fine-grained aspects over single-score feedback. The experiment design is solid and appears feasible. Data collection may pose minor challenges."

## C Quality Control of SFT Data

To ensure the quality and relevance of the supervised fine-tuning (SFT) data, we implement a multi-stage quality control in terms of the following aspects:

- **Source and Extraction**: Ideas are not freely generated, but are extracted from ICLR/NeurIPS 2023–2024 papers. We prompt LLaMA3 to extract the central research idea from each paper's abstract and introduction.
- **Supporting Paper Selection**: The associated main supporting paper is identified through a retrieval process that integrates citation graph statistics with LLaMA3-based prompting.
- **Screening Filters**: We apply automated filters to remove incoherent, or incomplete off-topic samples.
- **RL-based Optimization**: We rely on RL-based fine-tuning for subsequent optimization, using data that has been evaluated by human experts to ensure high performance (see Section 3.4).
- **Human Feedback in Reward Modeling**: The OpenReview data used for reward modeling is manually scored to reflect human judgments.

**Manual Audit of SFT Samples:**

To further evaluate data quality, we (authors) conducted a manual audit of 100 randomly sampled SFT examples. Three criteria were evaluated on a scale from 0 to 10:

- **Paper-Idea Topical Match:** How well the extracted idea matches the main topic of the paper.
- **Plausibility of Idea:** Whether the idea is realistic and logically follows from the paper context.
- **Completeness:** Whether the extracted idea is sufficiently complete and self-contained.

These quality control procedures and manual audit results in Table 6 demonstrate that our SFT dataset is generally of high quality and well-suited for model training.

| Criterion | Mean Score (0–10) | Std. Dev. |
|---|---|---|
| Paper-Idea Topical Match | 8.1 | 1.2 |
| Plausibility of Idea | 7.8 | 1.4 |
| Completeness | 7.5 | 1.6 |

Table 6: Manual audit of 100 randomly selected SFT samples: mean scores and standard deviations.

## D Comparison with ChatGPT as a Generation Baseline

Our primary objective is to develop a controllable open-source framework to guide smaller models for better research idea generation, which motivates our focus on open models such as T5 and LLaMA2. ChatGPT as a generation baseline is not entirely fair, since it is a significantly larger, proprietary model that is not accessible for training or fine-grained control. Furthermore, in our setup, GPT-4 also serves as the evaluator, which would introduce bias if used as the baseline model.

Nevertheless, for reference, we report the result of ChatGPT generations in our evaluation using the same prompts and context inputs as our method with automatic evaluation.

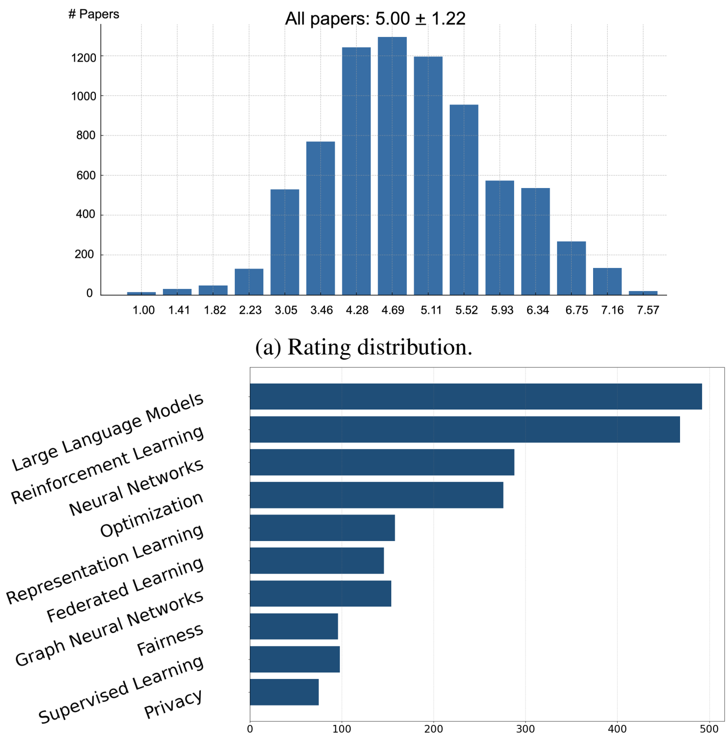
| **Model** | **Novelty** | **Feasiablity** | **Effectiveness** | **Overall** |
|---|---|---|---|---|
| ChatGPT (gpt-4-0314) | 6.2 | 5.3 | 5.4 | 5.6 |
| Ours (LLaMA2-RLHF+Dynamic) | 6.0 | 6.3 | 5.8 | 6.0 |

Table 7: Comparison between ChatGPT and our method with GPT-4 as reviewer.

As shown in Table 7, while ChatGPT achieves a higher novelty score, it tends to over-optimize for novelty at the expense of feasibility and grounding. In contrast, our method produces more balanced and controllable outputs, which we believe are better suited for real-world research ideation workflows.

# E Data Statistics

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




(a) Rating distribution.



(b) Top 10 topic distribution.

Figure 5: Rating and topic statistics of our dataset.

# F 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.

# G Novelty and Feasibility Control analysis

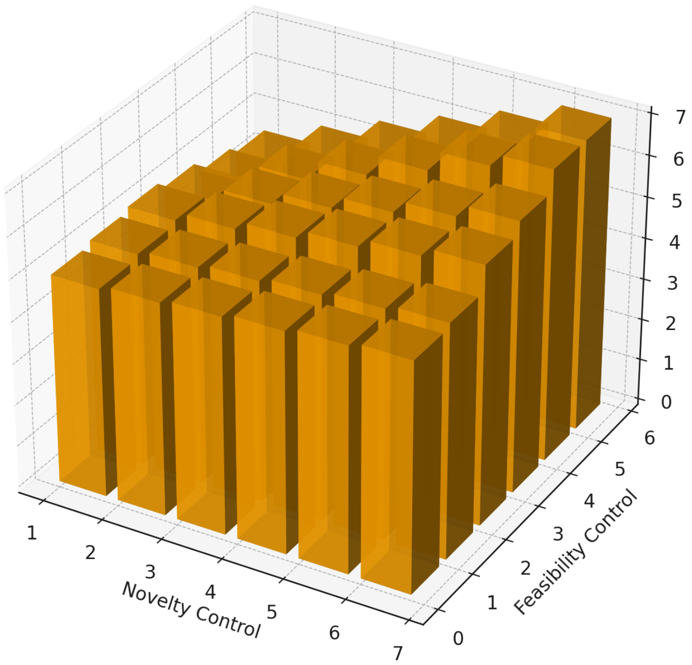



Figure 6: Novelty and Feasibility control analysis

We present the overall score analysis with the control of novelty and feasibility. We can clearly see that with the increase in the control of both dimensions, the overall score increases.

# H Human Evaluation Barplot

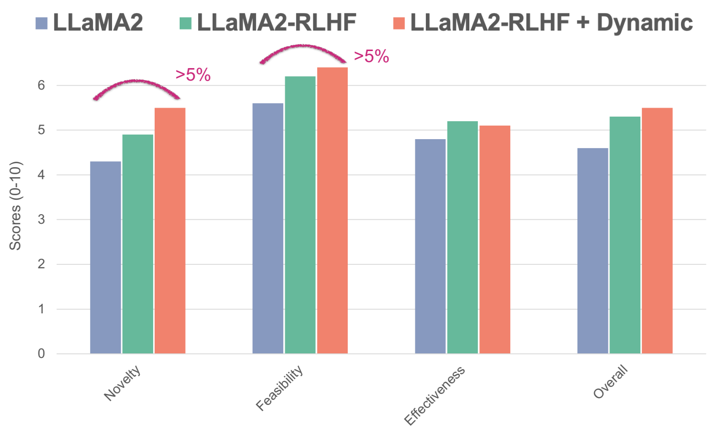



Figure 7: Human Evaluation Results

# I Scatter of Three Dimension v.s. Overall

# J Prompt for Research Idea Extraction

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

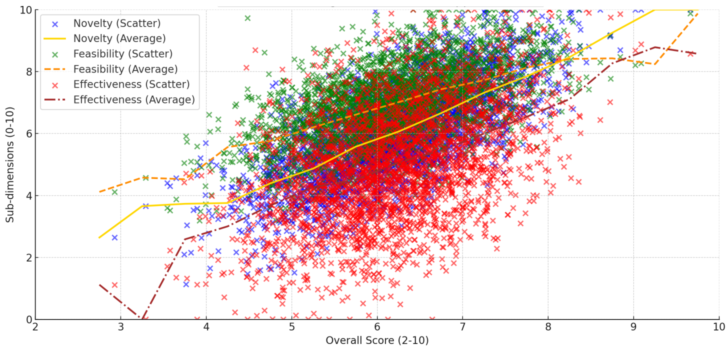



Figure 8: Scatters of different dimensions virus overall scores.

**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.
- **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.

## K 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:**
We score the novelty of papers based on how different they are 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**.

## L 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 pre-processed 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.

## M 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."
}
```