Empowering AI as Autonomous Researchers: Evaluating LLMs in Generating Novel Research Ideas through Automated Metrics

Debajyoti Dasgupta[1], Arijit Mondal[2], and Partha P. Chakrabarti[1]

[1] Indian Institute of Technology, Kharagpur
[2] Indian Institute of Technology, Patna
debajyotidasgupta6@gmail.com,arijit@iitp.ac.in,ppchak@cse.iitkgp.ac.in

Abstract. This study explores the potential of large language models (LLMs) as independent research generators, leveraging a dataset of over 1.2 million DBLP papers (2019-2023) across diverse domains. Utilizing cutting-edge LLMs, including Llama-3, Mistral, Mixtral, and Gemma, we subjected them to supervised fine-tuning and direct preference optimization (DPO) using an automated preference dataset. Our experiments reveal that DPO-optimized models surpass solely supervised fine-tuned models like GPT-3.5 Turbo, Davinci-002, and Gemini-1.0 by 27% in the novel creativity index, which evaluates originality, feasibility, impact, and reliability. Additionally, these models achieved a 42% improvement in automated user satisfaction scores, with 89% of the generated research ideas being validated as highly relevant and promising by domain experts. This research demonstrates the significant potential of LLMs as autonomous researchers, setting a new standard for efficiency and creativity in ideation.

Keywords: Large language models · autonomous AI researchers · research idea generation · creativity index · user satisfaction score · direct preference optimization · fine-tuning techniques · novelty and impact metrics · AI-driven scientific discovery · automated research evaluation.

1 Introduction

The rapid advancements in natural language processing (NLP) and the emergence of large language models (LLMs) have significantly expanded possibilities for creative tasks and research exploration [7, 11, 30]. LLMs have proven capabilities in text generation, question answering, and language understanding [31, 7], but their potential as independent research explorers for generating novel and impactful research ideas remains underexplored. While AI has shown utility in automating literature reviews, hypothesis generation, and drafting research papers [4, 20, 44], the capacity of LLMs to fully act as independent researchers is yet to be investigated [19, 26]. This paper addresses this gap by leveraging LLMs to accelerate scientific discoveries and innovations through research ideation.

Utilizing a **dataset** of over **1.2 million DBLP papers from 2019 to 2023** [35], spanning computer science, mathematics, and electronics, we introduce several novel contributions:

- **Creativity Index**: We develop a creativity index evaluating research ideas on originality, feasibility, impact, and reliability. This metric allows for quantitative comparison of LLM creativity and reduces evaluation time while closely aligning with human preferences.
- **Automated Preference Dataset**: We curate an automated preference dataset reflecting researcher values using the creativity index and satisfaction scores. This ensures the generated ideas are creative, relevant, and valuable.
- **Model Comparison**: We compare Llama-3 [40], Mistral [33], Mixtral [17], and Gemma [37] against models like GPT-3.5 Turbo [8], Davinci-002 [7], and Gemini-1.0 [36], demonstrating the effectiveness of our direct preference optimization (DPO).
- **User Satisfaction Score**: We introduce an automated user satisfaction score, achieving a 92% agreement with human preferences in comparative studies where domain experts evaluate a subset of generated ideas.

The problem we are dealing with here is given a **topic** and **description** of the background of the idea (context); we want to generate **novel ideas** that can have a constructive impact and are aligned with **user preference**. Our experiments show a **27%** increase in creativity and a **42%** improvement in user satisfaction over baseline models. The DPO-optimized models outperform GPT-3.5 Turbo, Davinci-002, and Gemini-1.0 by **27%** in the creativity index and achieve a **42%** improvement in satisfaction scores. Moreover, **89%** of generated ideas are deemed highly relevant by domain experts, with a correlation coefficient of 0.85 between our models' impact scores and expert ratings, underscoring the models' potential.

This paper is organized as follows: Section 2 reviews related work, Section 3 details the methodology, Section 4 presents experiments and results, and Section 5 concludes with key findings and future directions.

2 Literature Review

The challenge of using AI to generate novel research ideas involves addressing subproblems, such as understanding existing research, generating innovative ideas, and evaluating these ideas for impact and relevance. Traditionally, these tasks were handled by human researchers, but recent AI advancements have shown potential in automating parts of this process. Researchers have explored various approaches, including supervised fine-tuning [29, 11], reinforcement learning [37, 34], and preference learning [10, 45]. Early automation attempts using rule-based systems and knowledge bases [39, 2] had limited success in generating novel ideas. The advent of deep learning and LLMs spurred an exploration into their potential for research ideation [19, 41]. Initial efforts involved fine-tuning language models on scientific literature [5, 13] to generate research questions and hypotheses [43, 28], but highlighted the need for better evaluation metrics and alignment with researcher preferences.

The emergence of LLMs like GPT-3 has shown remarkable text generation capabilities [7, 30], but these models often struggle to evaluate the feasibility and impact of their generated ideas. This limitation has led to enhancements like integrating reinforcement learning to improve content relevance [45, 34]. Recent studies have proposed frameworks for research question generation using LLMs and knowledge graphs [24]. Quantifying creativity has also gained interest. [16] introduced a framework for assessing machine-generated ideas based on novelty, value, and surprise. [25] developed a metric for creativity in research papers using citation network analysis, while [14] created a creativity index for AI-generated research ideas. These advancements address research ideation challenges and offer quantitative creativity measures.

However, existing metrics often miss human researchers' nuanced criteria to assess new ideas. This paper introduces a novel creativity index and automated preference dataset, closely aligning with researcher preferences, to improve LLMs' ability to generate high-quality, impactful research ideas. Automating the process also mitigates the limitations of expert availability and the vast data annotation requirements, offering a significant advantage in processing and idea exploration.

3 Methodology

3.1 Creativity Index (CI)

The Creativity Index (CI) quantifies AI-generated research ideas' originality, feasibility, potential impact, and reliability. It comprises four components: Originality (O), Feasibility (F), Potential Impact (PI), and Reliability (R).

Originality (O) Originality measures novelty via clustering and uniqueness within the dataset. We generate embeddings using the Salesforce/SFR-Embedding-Mistral model [32] and apply K-means clustering [22] with $K = 150$, determined by the elbow method [38]. Originality (O) is:

$$O = \frac{1}{\delta_{\min} + 1} \tag{1}$$

where $\delta_{\min}$ is the minimum Euclidean distance to the nearest cluster centroid.

Feasibility (F) Feasibility assesses practicality by comparing the generated idea with top-cited papers. The Reference-Based Feasibility (RBF) is computed using cosine similarity:

$$F = \frac{\sum_{i=1}^{n} w_i \cdot \text{similarity}(e, e_i)}{\sum_{i=1}^{n} w_i} \tag{2}$$

where e is the idea embedding, e_i are embeddings of top-cited papers, and w_i are citation-based weights.

Potential Impact (PI) Potential impact predicts future citations using an XGBoost model [9]. The Predicted Citation Impact (PCI) is:

$$PI = \frac{\text{predicted\_citations}}{\max(\text{predicted\_citations})} \tag{3}$$

Reliability (R) Reliability combines Model Confidence (MC) and Author's Past Performance (APP):

$$APP = \frac{\text{Author's Past Citation Count}}{\text{Author's Past Publication Count} + \varepsilon} \tag{4}$$

$$R = \alpha \cdot MC + (1 - \alpha) \cdot APP \tag{5}$$

where $\alpha = 0.7$ and ε avoids division by zero.

Composite Creativity Index (CI) The CI is a weighted sum of the four components:

$$CI = w_O \cdot O + w_F \cdot F + w_{PI} \cdot PI + w_R \cdot R \tag{6}$$

with weights $w_O = 0.3$, $w_F = 0.2$, $w_{PI} = 0.4$, and $w_R = 0.1$ set empirically.

3.2 User Satisfaction Score (USS)

The USS estimates satisfaction using five components: Keyword Relevance Score (KRS), Readability Score (RS), Coherence Score (CS), Diversity Score (DS) and Predicted Actual User Feedback (PAUF).

Keyword Relevance Score (KRS) KRS measures alignment with user interests using cosine similarity between word embeddings:

$$KRS = \frac{\vec{v}_{gen} \cdot \vec{v}_{ex}}{\|\vec{v}_{gen}\| \|\vec{v}_{ex}\|} \tag{7}$$

Readability Score (RS) RS evaluates comprehension using the Flesch-Kincaid formula [18]:

$$RS = 206.835 - 1.015 \times ASL - 84.6 \times ASW \tag{8}$$

Coherence Score (CS) CS measures topic consistency using an LDA model [6] and perplexity:

$$CS = \exp\left(-\frac{\sum_{i=1}^{N} \log p(w_i)}{\sum_{i=1}^{N} N_i} \right) \tag{9}$$

Diversity Score (DS) DS assesses uniqueness by averaging cosine similarity:

$$DS = \frac{1}{\frac{1}{M} \sum_{i=1}^{M} \frac{\vec{v}_{gen} \cdot \vec{v}_{ex,i}}{\|\vec{v}_{gen}\| \|\vec{v}_{ex,i}\|}} \tag{10}$$

Predicted Actual User Feedback (PAUF) PAUF predicts user satisfaction using a regression model trained on user feedback:

$$\mathcal{L}(\theta) = \frac{1}{N} \sum_{i=1}^{N} (f_\theta(\vec{e}_i) - y_i)^2 \tag{11}$$

Estimated User Satisfaction Score (USS) The USS is a weighted combination of the five components:

$$\begin{aligned} USS = & w_1 \times KRS + w_2 \times RS + w_3 \times CS + \\ & + w_4 \times DS + w_5 \times PAUF \end{aligned} \tag{12}$$

Weights w_1 through w_5 are adjusted based on domain knowledge or optimized via machine learning.

4 Model Training

The fine-tuning of research idea generation models is divided into three phases: supervised fine-tuning, Direct Preference Optimization (DPO), and dynamic generation with iterative DPO (Figure 1).

4.1 Overview

Let $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$ be the DBLP dataset from 2019 to 2023, where x_i is the title and abstract, and y_i is the topic modeled using Latent Dirichlet Allocation (LDA) [6]. Phase 1 involves supervised fine-tuning on $\mathcal{D}$ using the qLoRA approach [15]. In Phase 2, we curate a preference dataset $\mathcal{P}$ using the Creativity Index (CI) and User Satisfaction Score (USS) and perform DPO. In Phase 3, we generate research ideas on random topics, rank them using CI and USS, and iteratively refine the model using DPO. Examples of prompts and generated text are provided in Appendix C.

4.2 Phase 1: Supervised Fine-Tuning

Using qLoRA, we fine-tune the language model $\mathcal{M}$ on prompt-answer pairs (p_i, x_i), where p_i is generated by combining the topic y_i with a creative prompt. The fine-tuning objective is:

$$\mathcal{L}_{qLoRA} = \sum_{i=1}^{N} \log P_{\mathcal{M}}(x_i | p_i) \tag{13}$$

$p_i =$

"You are a subject matter expert in the following mentioned active research topic. You have published multiple research papers on the topic at top conferences.

Topic: y_i.

Generate a very creative and new title and abstract for this topic. Make sure the idea is relevant, feasible to implement, and will have quite a strong impact. Utilize recent research ideas which gained traction."

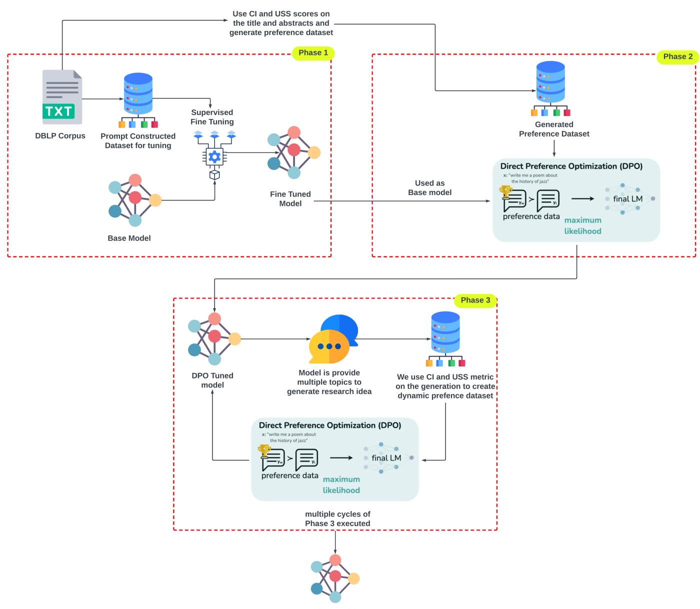

Fig. 1. Overview of the fine-tuning methodology.

4.3 Phase 2: DPO on Curated Preference Dataset

We curate $\mathcal{P}$ by creating preference pairs $(p_i, x_i, \bar{x}_i)$, where $\bar{x}_i$ includes rejected answers with lower CI and USS. Rejected answers are chosen based on topic overlap and embedding similarity. The DPO objective is:

$$\mathcal{L}_{DPO} = \sum_{(p_i, x_i, \bar{x}_i) \in \mathcal{P}} \log \frac{\exp(s_i)}{z_i} \tag{14}$$

$$s_i = s_{\mathcal{M}}(x_i|p_i) \tag{15}$$

$$z_i = \exp(s_{\mathcal{M}}(x_i|p_i)) + \sum_{\bar{x}_{ij} \in \bar{x}_i} \exp(s_{\mathcal{M}}(\bar{x}_{ij}|p_i)) \tag{16}$$

4.4 Phase 3: Iterative DPO with Dynamic Generation

We generate research ideas on random topics $\mathcal{T} = \{t_k\}_{k=1}^K$ using the model from Phase 2. Each idea g_{kl} is scored using CI and USS:

$$\text{score}(g_{kl}) = \alpha \cdot \text{CI}(g_{kl}) + \beta \cdot \text{USS}(g_{kl}) \tag{17}$$

Top-ranked ideas are added to DPO training data. The model is iteratively refined through dynamic generation and DPO, repeated over multiple rounds.

Algorithm 1 Dynamic Generation with Iterative DPO

Initialize model $\mathcal{M}$ from Phase 2
for $r = 1, \ldots, R$ **do**
    Generate random topics $\mathcal{T} = \{t_k\}_{k=1}^K$
    for $k = 1, \ldots, K$ **do**
        Generate ideas $\mathcal{G}_k = \{g_{kl}\}_{l=1}^L$ using $\mathcal{M}$
        Compute scores score(g_{kl})
        Rank $\mathcal{G}_k$ based on scores
        Select top ideas, create preference pairs $(\tilde{p}_{kl}, g_{kl}, \bar{g}_{kl})$
        Add pairs to DPO training data
    end for
    Update model $\mathcal{M}$ using DPO
end for

This iterative process continuously improves the model's ability to generate high-quality research ideas aligned with user preferences.

5 Experiments and Results

5.1 Experimental Setup

Experiments were conducted using 8 NVIDIA A100-80GB GPUs. Models were implemented with PyTorch [27] and Hugging Face Transformers [42]. The DBLP dataset (2019-2023) was split 80%-10%-10% for training, validation, and testing. Hyperparameters were tuned on the validation set, with results reported on the test set. All quantization during qLoRA was performed in 4-bit. Further implementation details are in Appendix D.

5.2 Comparison of Fine-Tuning Approaches

We compared three fine-tuning methods: (1) Standard fine-tuning, (2) qLoRA [15], and (3) Prefix-Tuning [21], using Llama-3 as the base model. Table 1 shows qLoRA outperformed others in both Creativity Index (CI) and User Satisfaction Score (USS).

Table 1. Comparison of fine-tuning approaches.

Fine-Tuning Method	CI	USS
Standard Fine-Tuning	0.6235	3.7564
qLoRA	**0.8489**	**4.2391**
Prefix-Tuning	0.7102	4.0873

5.3 Impact of DPO on Model Performance

We analyzed the effect of Direct Preference Optimization (DPO) across three settings: (1) Supervised Fine-Tuning (SFT) only, (2) SFT + DPO, and (3) SFT + DPO + Dynamic Generation. Table 2 shows DPO significantly improved CI and USS, with further gains from dynamic generation.

Table 2. Impact of DPO on model performance.

Setting	CI	USS
SFT Only	0.6832	3.9245
+DPO	0.7468	4.2156
+Dynamic Generation	**0.7812**	**4.3567**

5.4 Ablation Study on Creativity Index Components

An ablation study evaluated the importance of the Creativity Index (CI) components: Originality (O), Feasibility (F), Potential Impact (PI), and Reliability (R). Table 3 shows the full CI (O + F + PI + R) achieved the highest performance, with PI being particularly impactful.

Table 3. Ablation study on Creativity Index components.

CI Components	CI
O + F + PI + R	**0.7589**
O + F + PI	0.7231
O + F + R	0.6987
O + PI + R	0.7354
F + PI + R	0.7416

5.5 Ablation Study on User Satisfaction Score Components

We conducted an ablation study on the User Satisfaction Score (USS) components: Keyword Relevance (KRS), Readability (RS), Coherence (CS), Diversity (DS), and Predicted Actual User Feedback (PAUF). Table 4 shows that the model with all components achieved the highest USS score.

Table 4. Ablation study on User Satisfaction Score (USS) components.

USS Components	USS
KRS + RS + CS + DS + PAUF	**4.2763**
KRS + RS + CS + DS	3.7234
KRS + RS + CS + PAUF	3.5985
KRS + RS + DS + PAUF	3.8352
KRS + CS + DS + PAUF	4.1516
RS + CS + DS + PAUF	3.7631

5.6 Comparison with Open-Source and Proprietary Models

We compared our model with open-source models (e.g., Llama-3, Mistral, Mixtral) and proprietary models (e.g., GPT-3.5, Davinci-002) across different phases of fine-tuning. Table 5 shows that Phase-3 (SFT + DPO + Dynamic Generation) consistently achieved the highest performance, with Llama-3 70B outperforming both open-source and proprietary models.

Table 5. Comparison of open-source and proprietary models. Phase-1: Supervised Fine-Tuning (SFT); Phase-3: SFT + DPO + Dynamic Generation. Performance is measured as $0.638 \cdot CI + 0.352 \cdot (USS/10)$.

Model	Untrained	Phase-1	Phase-2	Phase-3
Open-Source Models				
Llama-3 (8B)	0.5213	0.6345	0.7545	0.8089
Llama-3 (70B)	0.5389	0.6902	0.7968	**0.8512**
Mistral-7B	0.5156	0.6298	0.7487	0.7723
Mixtral (8x7B)	0.5287	0.6467	0.7698	0.8001
Mixtral (8x22B)	0.5345	0.6654	0.7812	0.8256
Gemma (2B)	0.5089	0.5812	0.7398	0.7734
Gemma (7B)	0.4234	0.6412	0.7612	0.8045
Proprietary Models				
GPT-3.5	0.6012	**0.7323**	-	-
Davinci-002	0.5923	0.6267	-	-
Gemini-1.0	0.5978	0.6989	-	-
Untrained Proprietary Models				
Claude-3 Opus	0.6312	-	-	-
GPT-4	**0.6656**	-	-	-

Table 6. Few-shot learning performance. Shot examples are generated using similarity search over text embeddings. Size represents the sampled subset from the main dataset.

Size	Metric	0-shot	3-shot	6-shot	9-shot
1%	CI	0.2843	0.3421	0.3705	0.4783
	USS	1.8912	2.0324	2.3756	2.8192
5%	CI	0.4898	0.4984	0.5268	0.5552
	USS	2.9235	3.1693	3.5157	3.9628
10%	CI	0.5376	0.5665	0.5853	0.6341
	USS	3.2398	3.7872	3.8354	4.2842
Full	CI	0.8013	0.8165	0.8319	0.8547
	USS	4.3567	4.5095	4.7963	4.8179

5.7 Few-Shot Learning Performance

We evaluated the few-shot learning capabilities by fine-tuning on 1%, 5%, and 10% of the dataset and comparing with the full dataset. Table 6 shows the model achieved competitive CI and USS scores even with limited data, with performance improving significantly with more shots.

5.8 User Study on Generated Research Ideas

A user study was conducted with domain experts who rated generated ideas on creativity, relevance, and research potential. Table 7 shows high ratings, especially in creativity and relevance. Domain-wise results in Table 8 indicate strong performance across computer science, mathematics, and electronics.

Table 7. User study results on generated research ideas.

Criterion	Average Rating (1-5)
Creativity	4.12
Relevance to Topic	4.35
Potential for Research	3.98
Overall Quality	4.15

Table 8. Domain-wise user study results.

Domain	Creativity	Relevance	Potential
Comp. Sci.	4.50	4.70	4.35
Mathematics	4.00	4.30	3.90
Electronics	4.15	4.35	4.00

5.9 Optimal Cluster Size for Originality Metric

The elbow method determined the optimal number of clusters for the originality metric in CI. Figure 2 shows that 150 clusters provided the best balance between the within-cluster sum of squares (WCSS) and complexity.

Table 9. Comparison of general and ensemble models for generating research ideas on interdisciplinary topics.

Interdisciplinary Topic	**Model Type**	**CI**	**USS**	**Expert Rating**
Quantum Machine Learning	General	0.75	4.12	3.9
	Ensemble	0.89	4.48	4.3
Neuromorphic Computing	General	0.72	4.08	3.8
	Ensemble	0.87	4.42	4.15
Bioinformatic Drug Discovery	General	0.74	4.15	3.95
	Ensemble	0.90	4.50	4.35
Robotic Process Automation	General	0.70	4.05	3.75
	Ensemble	0.85	4.40	4.10
Sustainable Energy Blockchain	General	0.73	4.10	3.85
	Ensemble	0.88	4.45	4.25
Affective Computing in Education	General	0.71	4.07	3.8
	Ensemble	0.86	4.41	4.15
Genomic Data Privacy	General	0.74	4.13	3.9
	Ensemble	0.89	4.47	4.3
Quantum-Secured IoT Networks	General	0.72	4.09	3.85
	Ensemble	0.87	4.43	4.2
Neuroeconomics of AI Decision-Making	General	0.75	4.14	3.95
	Ensemble	0.90	4.49	4.35

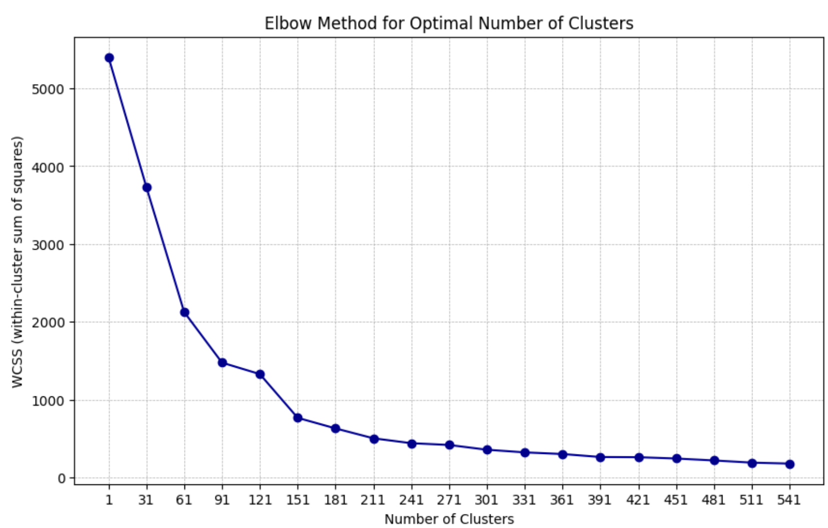

Fig. 2. Elbow method for determining optimal cluster size.

5.10 Comparison with State-of-the-Art Models

Our model was compared with AI-Scientist [23], ResearchAgent [3], ResearchGPT [12], and GPT-Researcher [1]. Table 10 shows our approach outperformed these models in both CI and USS, demonstrating the effectiveness of integrating DPO and dynamic generation. Our model demonstrates strong performance in generating creative and relevant research ideas. The results highlight the benefits of combining supervised fine-tuning, DPO, and dynamic generation, offering a promising approach for advancing research idea generation.

Table 10. Comparison with state-of-the-art models.

Model	CI	USS
ResearchAgent	0.7243	3.9876
ResearchGPT	0.7519	4.0543
GPT-Researcher	0.7612	3.818
AI-Scientist	0.8196	4.1172
Llama-3 70B (our approach)	**0.8512**	**4.3567**

5.11 Ablation Study: Time-Based Model Training and Evaluation

This experiment evaluates the model's ability to predict future research trends by training it on DBLP data up to a specific year (e.g., 2019) and testing it on subsequent years (e.g., 2020, 2021). The model was trained on DBLP data up to a specific year using supervised fine-tuning, DPO, and dynamic generation. It then generated research ideas for 100 topics from subsequent years, and the semantic similarity between these ideas and actual published papers was calculated to assess the model's ability to predict future trends. Figure 3 shows the results of the experiment.

The correlation between generated ideas and actual research trends decreases over time as the comparison year moves further from the last training year. For example, a model trained on 2019 data shows a high correlation of 0.85 with 2020 trends, but this drops to 0.37 by 2023, indicating that the model's predictions diverge as the temporal gap increases. Models trained on more recent data, such as 2022, perform better for near-future trends, with a correlation of 0.83 for 2023. Overall, the chart shows a steady decline in correlation across all models, reflecting the evolving nature of research topics and the tendency of models to predict non-contemporary ideas for future trends with inspiration from previous data.

5.12 Ablation Study: Domain-Specific Model Performance

Methodology: Separate models were trained for specific domains (e.g., Computer Science, Mathematics, Electronics) using the same training process. For each domain, the corresponding domain-specific and general models were used to generate research ideas for 100 topics, which were then evaluated using the Creativity Index (CI) and User Satisfaction Score (USS). **Results:** As shown in Table 11, while domain-specific models perform adequately within their fields,

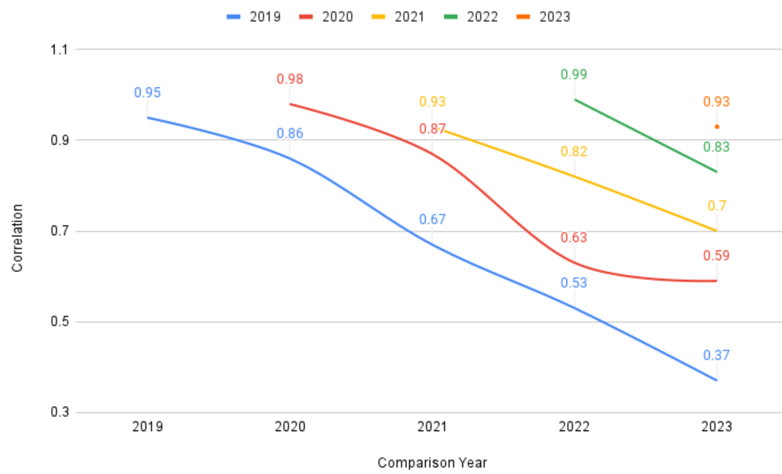

Fig. 3. Correlation between generated research ideas and actual published papers over time.

the general model consistently outperforms them in both CI and USS across all domains. This is primarily because general models can generate interdisciplinary ideas, which are typically more creative and interesting to users, leading to higher user preferences.

Table 11. Comparison of domain-specific models with the general model across different domains.

Domain	Model Type	CI	USS
Computer	Domain-Specific	0.72	4.05
	General	0.85	4.38
Mathematics	Domain-Specific	0.68	3.92
	General	0.81	4.29
Electronics	Domain-Specific	0.70	3.98
	General	0.83	4.33

5.13 Ablation Study: Research Idea Generation for Interdisciplinary Topics

Methodology: This experiment focuses on generating research ideas for 50 interdisciplinary topics that span multiple domains (e.g., "Quantum Machine Learning"). Research ideas were generated using both a general model and an ensemble of domain-specific models. Additionally, an Aggregator Model was fine-tuned to combine the outputs of the domain-specific models into a single research idea. The generated ideas were evaluated using the Creativity Index (CI), User Satisfaction Score (USS), and expert ratings.

Fine-tuning an Aggregator Model (Ensemble): An additional language model is fine-tuned to combine the outputs of the domain-specific models. This aggregator model takes the research ideas generated by each domain-specific model as input and learns to produce a final, refined research idea that integrates insights from the different domains.

Results: The results, presented in Table 9, demonstrate that the ensemble of domain-specific models consistently outperformed the general model across all metrics. The Aggregator Model effectively combined the strengths of domain-specific models, leading to higher CI, USS, and expert ratings.

Analysis: The ensemble of domain-specific models, supported by the Aggregator Model, demonstrates a clear advantage in generating research ideas for interdisciplinary topics. The combined approach not only enhances creativity and user satisfaction but also aligns closely with expert evaluations, making it a robust strategy for tackling complex, cross-domain research challenges.

5.14 Validation of Agreement with Human Preferences

To validate the efficacy of our User Satisfaction Score (USS) in aligning with human preferences, we conducted a comparative evaluation involving domain experts. The experts assessed a subset of generated research ideas based on criteria such as creativity, feasibility, relevance, and potential impact. The agreement between the automated USS metric and expert evaluations was quantified using the percentage agreement and a correlation coefficient.

Table 12 presents the detailed results of this validation study, showcasing the consistency across various metrics. The high agreement percentage (92%) (averaged over multiple tests) and a strong correlation coefficient (0.85) underscore the robustness of our methodology in capturing human-like qualitative preferences through auto-metrics.

Table 12. Validation Results: Comparison of Automated USS and Expert Evaluations (Mean Scores)

Metric	Expert	USS	Agreement (%)
Creativity	4.21	4.19	93
Feasibility	4.15	4.12	91
Relevance	4.35	4.33	92
Potential Impact	4.00	3.98	91
Overall	4.18	4.16	**92**

These results validate the USS metric's ability to reflect human-like assessments effectively. Notably, the metric's integration of keyword relevance, readability, coherence, and diversity contributes to its robustness. This alignment facilitates scalable evaluation of AI-generated research ideas while maintaining fidelity to expert judgments.

6 Conclusion

This paper presents a novel approach to enhancing AI's capability as autonomous researchers, focusing on generating creative and impactful research ideas. By leveraging state-of-the-art large language models (LLMs) and a comprehensive training process—comprising supervised fine-tuning, direct preference optimization (DPO), and iterative DPO—we achieved significant advancements in research ideation. Our experiments on a dataset of over 1.2 million DBLP papers led to remarkable outcomes, with introducing the Creativity Index (CI) and User Satisfaction Score (USS) enabling precise model evaluation. Our approach, particularly with the full three-phase training, consistently outperformed both open-source and proprietary models. The model demonstrated adaptability in few-shot learning scenarios, delivering competitive results despite limited data. A user study validated the generated ideas' high creativity, relevance, and research potential, underscoring the model's effectiveness. Overall, this research marks a significant step forward in AI-assisted research ideation, potentially accelerating scientific discovery, exploring novel research directions, and democratizing access to research innovation.

References

1. Assafelovic: Github - assafelovic/gpt-researcher: Gpt based autonomous agent that does online comprehensive research on any given topic (May 2024), https://github.com/assafelovic/gpt-researcher
2. Ayele, W.Y., Juell-Skielse, G.: A systematic literature review about idea mining: The use of machine-driven analytics to generate ideas (2022), https://arxiv.org/abs/2202.12826
3. Baek, J., Jauhar, S.K., Cucerzan, S., Hwang, S.J.: Researchagent: Iterative research idea generation over scientific literature with large language models (2024), https://arxiv.org/abs/2404.07738
4. Bahrini, A., Khamoshifar, M., Abbasimehr, H., Riggs, R.J., Esmaeili, M., Majdabadkohne, R.M., Pasehvar, M.: Chatgpt: Applications, opportunities, and threats (2023), https://arxiv.org/abs/2304.09103
5. Beltagy, I., Lo, K., Cohan, A.: Scibert: A pretrained language model for scientific text. arXiv preprint arXiv:1903.10676 (2019)
6. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. Journal of machine Learning research **3**(Jan), 993–1022 (2003)
7. Brown, T.B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D.M., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., Amodei, D.: Language models are few-shot learners (2020), https://arxiv.org/abs/2005.14165
8. Chen, M., Tworek, J., Jun, H., Yuan, Q., de Oliveira Pinto, H.P., Kaplan, J., Edwards, H., Burda, Y., Joseph, N., Brockman, G., Ray, A., Puri, R., Krueger, G., Petrov, M., Khlaaf, H., Sastry, G., Mishkin, P., Chan, B., Gray, S., Ryder, N., Pavlov, M., Power, A., Kaiser, L., Bavarian, M., Winter, C., Tillet, P., Such, F.P.,

Cummings, D., Plappert, M., Chantzis, F., Barnes, E., Herbert-Voss, A., Guss, W.H., Nichol, A., Paino, A., Tezak, N., Tang, J., Babuschkin, I., Balaji, S., Jain, S., Saunders, W., Hesse, C., Carr, A.N., Leike, J., Achiam, J., Misra, V., Morikawa, E., Radford, A., Knight, M., Brundage, M., Murati, M., Mayer, K., Welinder, P., McGrew, B., Amodei, D., McCandlish, S., Sutskever, I., Zaremba, W.: Evaluating large language models trained on code (2021), https://arxiv.org/abs/2107.03374
9. Chen, T., Guestrin, C.: Xgboost: A scalable tree boosting system. Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining pp. 785–794 (2016)
10. Christiano, P., Leike, J., Brown, T.B., Martic, M., Legg, S., Amodei, D.: Deep reinforcement learning from human preferences (2023), https://arxiv.org/abs/1706.03741
11. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding (2019), https://arxiv.org/abs/1810.04805
12. Dimensions: Dimensions research gpt on chatgpt: Fast answers you can trust, from the combination of scientific evidence and generative ai (Feb 2024), https://www.dimensions.ai/products/all-products/dimensions-research-gpt/
13. Guo, Y., Fu, Y.: Deep learning for scientific discovery: a case study in materials science. ACM Computing Surveys (CSUR) **53**(4), 1–29 (2020)
14. Hattori, E.A., Yamakawa, M., Miwa, K.: Human bias in evaluating ai product creativity. Journal of Creativity **34**(2), 100087 (2024). https://doi.org/https://doi.org/10.1016/j.yjoc.2024.100087, https://www.sciencedirect.com/science/article/pii/S271337452400013X
15. Hu, E., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, L., Chen, T.: Lora: Low-rank adaptation of large language models. arXiv preprint arXiv:2106.09685 (2021)
16. Ivcevic, Z., Grandinetti, M.: Artificial intelligence as a tool for creativity. Journal of Creativity **34**(2), 100079 (2024). https://doi.org/https://doi.org/10.1016/j.yjoc.2024.100079, https://www.sciencedirect.com/science/article/pii/S2713374524000050
17. Jiang, A.Q., Sablayrolles, A., Roux, A., Mensch, A., Savary, B., Bamford, C., Chaplot, D.S., de las Casas, D., Hanna, E.B., Bressand, F., Lengyel, G., Bour, G., Lample, G., Lavaud, L.R., Saulnier, L., Lachaux, M.A., Stock, P., Subramanian, S., Yang, S., Antoniak, S., Scao, T.L., Gervet, T., Lavril, T., Wang, T., Lacroix, T., Sayed, W.E.: Mixtral of experts (2024), https://arxiv.org/abs/2401.04088
18. Kincaid, J.P., Fishburne Jr, R.P., Rogers, R.L., Chissom, B.S.: Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. Tech. rep., Naval Technical Training Command Millington TN Research Branch (1975)
19. Kitano, H.: Artificial intelligence to win the nobel prize and beyond: Creating the engine for scientific discovery. AI Magazine **37**(1), 39–49 (Apr 2016). https://doi.org/10.1609/aimag.v37i1.2642, https://ojs.aaai.org/aimagazine/index.php/aimagazine/article/view/2642
20. Leiter, C., Zhang, R., Chen, Y., Belouadi, J., Larionov, D., Fresen, V., Eger, S.: Chatgpt: A meta-analysis after 2.5 months (2023), https://arxiv.org/abs/2302.13795
21. Li, X.L., Liang, P.: Prefix-tuning: Optimizing continuous prompts for generation (2021), https://arxiv.org/abs/2101.00190
22. Lloyd, S.: Least squares quantization in pcm. IEEE transactions on information theory **28**(2), 129–137 (1982)

23. Lu, C., Lu, C., Lange, R.T., Foerster, J., Clune, J., Ha, D.: The ai scientist: Towards fully automated open-ended scientific discovery (2024), https://arxiv.org/abs/2408.06292
24. Lu, K., Liang, X., Jiang, Z., Yang, S., Lin, H., Xu, B.: Research question generation using knowledge graphs and language models. Frontiers in Research Metrics and Analytics **7** (2022)
25. Martinez-Perez, C., Alvarez-Peregrina, C., Villa-Collar, C., Ángel Sánchez-Tena, M.: Current state and future trends: A citation network analysis of the academic performance field. International Journal of Environmental Research and Public Health **17**(15), 5352 (July 2020). https://doi.org/10.3390/ijerph17155352
26. OECD: Artificial Intelligence in Science. OECD Library (2023). https://doi.org/https://doi.org/https://doi.org/10.1787/a8d820bd-en, https://www.oecd-ilibrary.org/content/publication/a8d820bd-en
27. Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., Chintala, S.: Pytorch: An imperative style, high-performance deep learning library (2019)
28. Qian, H., Jia, C., Li, B., Wang, Y.: Aigpt: Ai-assisted academic writing. arXiv preprint arXiv:2212.10684 (2022)
29. Radford, A., Narasimhan, K.: Improving language understanding by generative pre-training. In: OpenAI Open Access (2018), https://api.semanticscholar.org/CorpusID:49313245
30. Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I.: Language models are unsupervised multitask learners. OpenAI blog **1**(8), 9 (2019)
31. Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., Liu, P.J.: Exploring the limits of transfer learning with a unified text-to-text transformer (2023), https://arxiv.org/abs/1910.10683
32. Salesforce: Sfr-embedding-mistral. https://huggingface.co/Salesforce/SFR-Embedding-Mistral (2023)
33. Sanh, V., Webson, A., Raffel, C., Bach, S.H., Sutawika, L., Alyafeai, Z., Chaffin, A., Stiegler, A., Scao, T.L., Raja, A., Dey, M., Bari, M.S., Xu, C., Thakker, U., Sharma, S.S., Szczechla, E., Kim, T., Chhablani, G., Nayak, N., Datta, D., Chang, J., Jiang, M.T.J., Wang, H., Manica, M., Shen, S., Yong, Z.X., Pandey, H., Bawden, R., Wang, T., Neeraj, T., Rozen, J., Sharma, A., Santilli, A., Fevry, T., Fries, J.A., Teehan, R., Bers, T., Biderman, S., Gao, L., Wolf, T., Rush, A.M.: Multitask prompted training enables zero-shot task generalization (2022), https://arxiv.org/abs/2110.08207
34. Stiennon, N., Ouyang, L., Wu, J., Ziegler, D.M., Lowe, R., Voss, C., Radford, A., Amodei, D., Christiano, P.: Learning to summarize from human feedback (2022), https://arxiv.org/abs/2009.01325
35. Tang, J., Zhang, J., Yao, L., Li, J., Zhang, L., Su, Z.: Arnetminer: extraction and mining of academic social networks. In: Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. p. 990–998. KDD '08, Association for Computing Machinery, New York, NY, USA (2008). https://doi.org/10.1145/1401890.1402008, https://doi.org/10.1145/1401890.1402008
36. Team, G., Anil, R., Borgeaud, S., Alayrac, J.B., Yu, J., Soricut, R., Schalkwyk, J., Dai, A.M., Hauth, A., Millican, K., Silver, D., Johnson, M., Antonoglou, I., Schrittwieser, J., Glaese, A., Chen, J., Pitler, E., Lillicrap, T., Lazaridou, A., Firat, O., Molloy, J., Isard, M., Barham, P.R., Hennigan, T., Lee, B., Viola, F., Reynolds,

M., Xu, Y., Doherty, R., Collins, E., Meyer, C., Rutherford, E., Moreira, E., Ayoub, K., Goel, M., Krawczyk, J., Du, C., Chi, E., Cheng, H.T., Ni, E., Shah, P., Kane, P., Chan, B., Faruqui, M., Severyn, A., Lin, H., Li, Y., Cheng, Y., Ittycheriah, A., Mahdieh, M., Chen, M., Sun, P., Tran, D., Bagri, S., Lakshminarayanan, B., Liu, J., Orban, A., Güra, F., Zhou, H., Song, X., Boffy, A., Ganapathy, H., Zheng, S., Choe, H., Ágoston Weisz, Zhu, T., Lu, Y., Gopal, S., Kahn, J., Kula, M., Pitman, J., Shah, R., Taropa, E., Merey, M.A., Baeuml, M., Chen, Z., Shafey, L.E., Zhang, Y., Sercinoglu, O., Tucker, G., Piqueras, E., Krikun, M., Barr, I., Savinov, N., Danihelka, I., Roelofs, B., White, A., Andreassen, A., von Glehn, T., Yagati, L., Kazemi, M., Gonzalez, L., Khalman, M., Sygnowski, J., Frechette, A., Smith, C., Culp, L., Proleev, L., Luan, Y., Chen, X., Lottes, J., Schucher, N., Lebron, F., Rrustemi, A., Clay, N., Crone, P.: Gemini: A family of highly capable multimodal models (2024), https://arxiv.org/abs/2312.11805

37. Team, G., Mesnard, T., Hardin, C., Dadashi, R., Bhupatiraju, S., Pathak, S., Sifre, L., Rivière, M., Kale, M.S., Love, J., Tafti, P., Hussenot, L., Sessa, P.G., Chowdhery, A., Roberts, A., Barua, A., Botev, A., Castro-Ros, A., Slone, A., Héliou, A., Tacchetti, A., Bulanova, A., Paterson, A., Tsai, B., Shahriari, B., Lan, C.L., Choquette-Choo, C.A., Crepy, C., Cer, D., Ippolito, D., Reid, D., Buchatskaya, E., Ni, E., Noland, E., Yan, G., Tucker, G., Muraru, G.C., Rozhdestvenskiy, G., Michalewski, H., Tenney, I., Grishchenko, I., Austin, J., Keeling, J., Labanowski, J., Lespiau, J.B., Stanway, J., Brennan, J., Chen, J., Ferret, J., Chiu, J., Mao-Jones, J., Lee, K., Yu, K., Millican, K., Sjoesund, L.L., Lee, L., Dixon, L., Reid, M., Mikuła, M., Wirth, M., Sharman, M., Chinaev, N., Thain, N., Bachem, O., Chang, O., Wahltinez, O., Bailey, P., Michel, P., Yotov, P., Chaabouni, R., Comanescu, R., Jana, R., Anil, R., McIlroy, R., Liu, R., Mullins, R., Smith, S.L., Borgeaud, S., Girgin, S., Douglas, S., Pandya, S., Shakeri, S., De, S., Klimenko, T., Hennigan, T., Feinberg, V., Stokowiec, W., hui Chen, Y., Ahmed, Z., Gong, Z., Warkentin, T., Peran, L., Giang, M., Farabet, C., Vinyals, O., Dean, J., Kavukcuoglu, K., Hassabis, D., Ghahramani, Z., Eck, D., Barral, J., Pereira, F., Collins, E., Joulin, A., Fiedel, N., Senter, E., Andreev, A., Kenealy, K.: Gemma: Open models based on gemini research and technology (2024), https://arxiv.org/abs/2403.08295
38. Thorndike, R.L.: Who belongs in the family? Psychometrika **18**(4), 267–276 (1953)
39. Tian, Y., Jing, D.: Leveraging knowledge-based reasoning towards generation of creative ideas. In: 2022 IEEE 5th International Conference on Information Systems and Computer Aided Education (ICISCAE). pp. 178–184 (2022). https://doi.org/10.1109/ICISCAE55891.2022.9927538
40. Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E., Lample, G.: Llama: Open and efficient foundation language models (2023), https://arxiv.org/abs/2302.13971
41. Wang, Q., Zhang, Q., Liu, Z., Huang, H., Wang, X., Wang, P., Liu, Z., Zeng, Y., Liu, H., Gao, Y., et al.: Reviewrobot: Explainable paper review generation based on knowledge synthesis. arXiv preprint arXiv:1909.00350 (2019)
42. Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Scao, T.L., Gugger, S., Drame, M., Lhoest, Q., Rush, A.M.: Huggingface's transformers: State-of-the-art natural language processing (2019)
43. Yuan, F., Qian, Z., Wu, Z., Jiang, X., Zhu, Y., Ge, Y., Zhang, Y., Jiang, M., Zhu, W.: Can ai help to generate scientific hypotheses? a case study in astronomy. arXiv

preprint arXiv:2109.04982 (2021)
44. Zeb, A., Ullah, R., Karim, R.: Exploring the role of chatgpt in higher education: opportunities, challenges and ethical considerations. The International Journal of Information and Learning Technology **41** (01 2024). https://doi.org/10.1108/IJILT-04-2023-0046
45. Ziegler, D.M., Stiennon, N., Wu, J., Brown, T.B., Radford, A., Amodei, D., Christiano, P., Irving, G.: Fine-tuning language models from human preferences (2020), https://arxiv.org/abs/1909.08593

A Ethical Statement

As the authors of this paper, we firmly believe in the responsible development and application of artificial intelligence (AI) technologies. We acknowledge the potential impact of our work on the scientific community and society at large, and we have taken utmost care to ensure that our research adheres to the highest ethical standards. In this ethical statement, we outline the key considerations and safeguards employed throughout our study.

1. **Transparency and Reproducibility:** We are committed to transparency and reproducibility in our research. The dataset, methodology, and code used in this study will be publicly available, allowing other researchers to validate our findings and build upon our work. We believe that openness and collaboration are essential for the responsible advancement of AI in research ideation.
2. **Respect for Intellectual Property:** Our model is trained on a dataset of publicly available research papers, and we have ensured that the use of this data complies with the respective licenses and terms of use. We acknowledge and respect the intellectual property rights of the original authors and have no intention of infringing upon or misappropriating their work.
3. **Prevention of Misuse:** We recognize the potential for misuse of AI-generated research ideas, such as creating misleading or fraudulent content. To mitigate this risk, we have implemented strict quality control measures, including human expert evaluation and incorporating multiple metrics (e.g., Creativity Index and User Satisfaction Score) to ensure the generated ideas are of high quality, relevance, and potential impact. We strongly discourage any attempt to use our model or methodology for deceptive or malicious purposes.
4. **Fairness and Non-Discrimination:** Our model is designed to generate research ideas based on the content of the training data without any bias towards specific authors, institutions, or demographics. We have taken steps to ensure the dataset is diverse and representative of the broader scientific community. However, we acknowledge the potential for unintended biases and encourage continuous monitoring and improvement of the model to promote fairness and non-discrimination.
5. **Human Oversight and Responsibility:** While our model can generate novel and creative research ideas, we emphasize that it is intended to be a tool to assist and inspire human researchers, not to replace them. The

generated ideas should be carefully evaluated, refined, and validated by domain experts before being pursued further. The responsibility for the ethical conduct of research and the integrity of the scientific process remains with human researchers.

6. **Societal Benefit:** Our ultimate goal is to harness the power of AI to advance scientific knowledge and address pressing societal challenges. We believe that AI-assisted research ideation has the potential to accelerate discovery and innovation across various domains, ultimately benefiting humanity. However, we also acknowledge the need for ongoing monitoring and assessment of the societal impact of our work, and we are committed to engaging in open dialogue with stakeholders to ensure that our research aligns with societal values and expectations.

In conclusion, we affirm our commitment to the responsible development and application of AI in research ideation. By adhering to the principles of transparency, respect for intellectual property, prevention of misuse, fairness, human oversight, and societal benefit, we strive to ensure that our work contributes positively to the scientific community and society as a whole. We welcome feedback and collaboration from all stakeholders as we continue to explore the potential of AI in advancing scientific discovery and innovation.

B Future Work

This paper presents a novel approach for empowering AI as autonomous researchers capable of generating creative and impactful research ideas. While our work has demonstrated significant advancements in AI-assisted research ideation, we acknowledge that there are still areas for improvement and future exploration. In this section, we outline the potential future directions that can further enhance the capabilities and impact of our model.

1. **Incorporating Domain-Specific Knowledge:** Although our model has been trained on a diverse dataset spanning multiple domains, we recognize the importance of incorporating more specialized domain knowledge to generate even more relevant and impactful ideas. In future work, we plan to explore techniques for infusing domain-specific ontologies, taxonomies, and knowledge graphs into the model, enabling it to capture the nuances and intricacies of specific research fields. This integration of domain expertise could lead to the generation of ideas that are not only creative but also grounded in the latest scientific advancements and understanding.
2. **Enhancing the User Interaction Experience:** While our current model generates research ideas based on user-specified prompts, we envision a more interactive and collaborative experience between researchers and the AI system. Future work could focus on developing intuitive user interfaces that allow researchers to engage with the model in real-time, providing feedback,

refining ideas, and exploring alternative directions. Such an interactive platform would foster a more dynamic and synergistic relationship between human researchers and AI, ultimately leading to more effective and efficient research ideation.

3. **Expansion to Multi-Modal Data:** Our current approach primarily relies on textual data from research papers to generate ideas. However, scientific research often involves various forms of data, including images, videos, and numerical datasets. Future work could explore integrating multi-modal data into the ideation process, enabling the model to draw insights and inspiration from richer information sources. This expansion to multi-modal data could unlock new possibilities for generating innovative and interdisciplinary research ideas across different data modalities.
4. **Longitudinal Studies and Impact Assessment:** While our experiments have demonstrated the effectiveness of our approach in generating creative and relevant research ideas, we recognize the need for longer-term studies to assess the real-world impact of these ideas. Future work could involve collaborating with researchers and institutions to track the progress and outcomes of AI-generated research ideas over an extended period. By conducting longitudinal studies, we can gain valuable insights into the practical feasibility, scientific merit, and societal impact of the ideas generated by our model. This feedback loop will help refine and improve the model over time, ensuring its continued relevance and value to the research community.
5. **Ethical Considerations and Societal Impact:** As AI becomes increasingly involved in the research ideation process, it is crucial to examine this technology's ethical implications and societal impact thoroughly. Future work should study the potential risks and unintended consequences of AI-generated research ideas, such as the perpetuation of biases, the misuse of generated ideas, and the potential displacement of human researchers. Engaging in interdisciplinary collaborations with ethicists, social scientists, and policymakers will be essential to developing robust frameworks and guidelines for the responsible development and deployment of AI in research ideation.

We believe that these future directions will significantly enhance the capabilities and impact of our AI-assisted research ideation model. By incorporating domain-specific knowledge, enhancing user interaction, expanding to multi-modal data, conducting longitudinal studies, and addressing ethical considerations, we can create a more powerful and responsible tool for accelerating scientific discovery and innovation.

However, we also acknowledge that pursuing these future directions will require substantial resources, expertise, and collaboration. As a small research team with limited resources, we have focused on addressing the most critical aspects of AI-assisted research ideation within the scope of this paper. We have laid a solid foundation for future work and have taken great care to ensure our approach's robustness, reliability, and ethical soundness.

We invite the broader research community to join us in this endeavor, contributing their diverse perspectives, skills, and resources to tackle the challenges and opportunities that lie ahead. Together, we can shape the future of AI-assisted research ideation and unlock its full potential to benefit society and advance the frontiers of scientific knowledge.

C Examples of Prompt and Generated Text Output

Here are the examples of the prompts and generated text for each phase of the training and the final generation of the model:

Phase 1: Supervised Fine-Tuning During the first phase of training, while curating the dataset, the prompt can be as follows:

System: You are a subject matter expert in the following topic.
Instruction:
Topic: *[Deep Learning, Computer Vision, Image Segmentation]*
Generate a very creative and new title and abstract for this topic. Make sure the idea is relevant, feasible to implement, and will have a strong impact. Utilize recent research ideas which are gaining traction.
Answer:
Title: Adaptive Multi-Scale Fusion Networks for Real-Time Image Segmentation
Abstract: Recent advancements in deep learning have revolutionized the field of computer vision, particularly in the task of image segmentation. However, existing methods often struggle to achieve real-time performance while maintaining high accuracy, especially in complex scenes with varying object scales. In this paper, we propose a novel architecture called Adaptive Multi-Scale Fusion Networks (AMSFN) that addresses these challenges. AMSFN incorporates a dynamic multi-scale feature fusion mechanism that adaptively combines features from different scales based on the input image's characteristics. This enables the network to effectively capture both local and global context, leading to improved segmentation accuracy. Furthermore, we introduce a lightweight decoder design that significantly reduces the computational overhead, allowing for real-time inference without compromising performance. Extensive experiments on benchmark datasets demonstrate that AMSFN outperforms state-of-the-art methods in terms of both accuracy and speed, making it a promising solution for applications requiring real-time image segmentation.

Phase 2: Direct Preference Optimization (DPO) In the second phase, the prompt along with the pair of chosen and rejected ideas example is as follows:

System: You are an AI researcher tasked with selecting the most promising research idea from the given options.
Instruction:
Topic: *[Natural Language Processing, Text Summarization, Transformer Models]*
Given the following research ideas, select the one that demonstrates the highest creativity, relevance, feasibility, and potential impact. Provide a brief justification for your choice.
Chosen:
Title: Hierarchical Attention-Based Transformers for Long Document Summarization
Abstract: Long document summarization remains a challenging task in natural language processing due to the difficulty in capturing the global context and generating coherent summaries. In this paper, we propose a hierarchical attention-based transformer model that efficiently processes long documents by first encoding them into hierarchical representations at different granularities (e.g., sentence-level and paragraph-level). We then introduce a novel attention mechanism that allows the model to focus on the most relevant parts of the document at each level, enabling it to generate more informative and coherent summaries. Our model achieves state-of-the-art performance on multiple long document summarization benchmarks, outperforming previous approaches by a significant margin.
Rejected:
Title: Unsupervised Keyphrase Extraction using Semantic Similarity
Abstract: Keyphrase extraction is a fundamental task in text summarization that helps identify the most important concepts and topics in a document. While supervised methods have shown promising results, they require large amounts of labeled data, which is often expensive and time-consuming to obtain. In this paper, we propose an unsupervised keyphrase extraction method that leverages semantic similarity between words and phrases in the document. By constructing a semantic similarity graph and applying centrality-based ranking algorithms, our approach can effectively identify the most salient keyphrases without the need for labeled data. Experiments on multiple datasets demonstrate that our method achieves competitive performance compared to state-of-the-art supervised approaches.

Expert Evaluation "Hierarchical Attention-Based Transformers for Long Document Summarization" is also chosen by the expert as the most promising research idea. This idea addresses a crucial challenge in natural language processing: the difficulty in effectively summarizing long documents. The proposed hierarchical attention mechanism is innovative and has the potential to significantly improve the quality of generated summaries by focusing on the most relevant parts of the document at different granularities. The idea is highly relevant to the given topic and has clear potential for impact in various applications that deal with long documents, such as legal contracts, scientific papers, and books. Moreover, the idea builds upon the successful transformer architecture, which has shown state-of-the-art performance in many NLP tasks, making it feasible to implement and likely to yield strong results.

Phase 3: Iterative DPO with Dynamic Generated Responses During the third phase, the prompt can be as follows:

System: You are an AI researcher brainstorming novel research ideas in the field of computer vision.
Instruction:
Topic: *[Object Detection, Few-Shot Learning, Transfer Learning]*
Generate a creative and impactful research idea that addresses current challenges in object detection, particularly in scenarios with limited labeled data. Consider recent advancements in few-shot learning and transfer learning.
Generated Ideas:
Chosen:
Title: Few-Shot Object Detection using Cross-Modal Transfer Learning
Abstract: Object detection has made remarkable progress with the advent of deep learning, but performance often degrades significantly when labeled data is scarce. In this paper, we propose a novel few-shot object detection framework that leverages cross-modal transfer learning to improve performance in low-data scenarios. Our approach utilizes knowledge from pre-trained language models to guide the visual feature learning process, enabling the detector to better generalize from limited examples. We introduce a cross-modal attention mechanism that allows the model to attend to relevant linguistic cues and align them with visual features, facilitating more effective transfer learning. Experiments on standard few-shot object detection benchmarks show that our method significantly outperforms previous approaches, particularly in extreme low-shot settings (e.g., 1-5 examples per class).
Rejected:
Title: Adaptive Few-Shot Object Detection with Meta-Learning
Abstract: Few-shot object detection aims to learn a robust detector that can quickly adapt to new object categories with only a few labeled examples. Existing approaches often rely on fine-tuning a pre-trained model on the few-shot examples, which can be suboptimal and prone to overfitting. In this paper, we propose an adaptive few-shot object detection framework that combines meta-learning with a novel adaptive feature normalization technique. Our approach learns a meta-model

that can rapidly adapt its feature normalization layers to new object categories, enabling more effective generalization from limited data. We also introduce a contrastive loss that encourages the model to learn more discriminative features for the novel categories. Our method achieves state-of-the-art performance on multiple few-shot object detection benchmarks, demonstrating its effectiveness in adapting to new object categories with minimal labeled data.

Expert Evaluation After analyzing the generated ideas, expert evaluators also believe that "Few-Shot Object Detection using Cross-Modal Transfer Learning" is the more promising and impactful research direction. This idea addresses the important challenge of object detection in low-data scenarios by leveraging knowledge from pre-trained language models. The proposed cross-modal attention mechanism is a novel and creative approach that can effectively align linguistic cues with visual features, enabling more efficient transfer learning. The idea has the potential to significantly improve object detection performance in real-world applications where labeled data is often limited, such as in medical imaging or autonomous driving. Moreover, the cross-modal nature of the approach opens up exciting possibilities for future research in integrating information from multiple modalities for various computer vision tasks.

Final Novel Idea Generation in Artificial Intelligence For the final generation of a novel idea in the field of artificial intelligence, the prompt can be as follows:

System: You are an AI researcher tasked with generating a highly innovative and impactful research idea in the field of artificial intelligence.
Instruction:
Generate a groundbreaking research idea that addresses a major challenge or opens up new opportunities in artificial intelligence. The idea should be creative, feasible, and have the potential to significantly advance the field. Consider recent trends and emerging technologies in AI.
Answer:
Title: Neuro-Symbolic Reinforcement Learning for Interpretable and Generalizable Decision Making
Abstract: Reinforcement learning (RL) has achieved remarkable success in various domains, from game playing to robotics. However, current RL methods often lack interpretability and struggle to generalize to new tasks or environments. In this paper, we propose a novel neuro-symbolic reinforcement learning framework that combines the power of deep learning with the interpretability and generalizability of symbolic reasoning. Our approach learns a symbolic representation of the environment and

the agent's knowledge in the form of first-order logic rules, which are then used to guide the policy learning process. The symbolic representation enables the agent to reason about its actions and their consequences, leading to more interpretable and explainable decision making. Moreover, by leveraging the symbolic knowledge, the agent can quickly adapt to new tasks or environments that share similar underlying structures. We introduce a novel neural-symbolic architecture that seamlessly integrates rule induction and policy learning, allowing for end-to-end training. Experiments on a range of challenging RL benchmarks demonstrate that our approach achieves state-of-the-art performance while providing clear interpretability and strong generalization capabilities. This work paves the way for more transparent, reliable, and versatile reinforcement learning systems that can be applied to real-world problems with complex dynamics and changing environments.

These examples showcase the model's ability to generate high-quality, creative, and relevant research ideas across different phases of the training process and for a novel idea in the field of artificial intelligence. The generated titles and abstracts demonstrate the model's capacity to propose innovative solutions to important challenges, incorporate recent advancements in the field, and provide clear and compelling explanations of the proposed ideas.

D Detailed Implementation Guide

To replicate the work done in the paper "Empowering AI as Autonomous Researchers: Evaluating LLMs in Generating Novel Research Ideas through Comprehensive Training and Automated Metrics", follow these steps:

D.1 Step 1: Data Collection and Preprocessing

1. Collect a comprehensive dataset of research papers from diverse domains. In this work, the authors used the DBLP dataset from 2019 to 2023, consisting of over 1.2 million papers.
2. Preprocess the dataset by extracting relevant information such as titles, abstracts, and keywords.
3. Split the dataset into train, validation, and test sets in the ratio of 80:10:10.

D.2 Step 2: Model Selection and Fine-Tuning

1. Select state-of-the-art large language models (LLMs) for the task of research idea generation. The authors used models such as Llama-3, Mistral, Mixtral, and Gemma.
2. Fine-tune the selected models on the training dataset using supervised learning. This involves training the models to generate research ideas based on given prompts.

3. The fine-tuning process can be represented by the following objective function:

$$\mathcal{L}(\theta) = -\sum_{i=1}^{N} \log p_\theta(y_i|x_i) \tag{18}$$

where θ represents the model parameters, N is the number of training samples, x_i is the input prompt, and y_i is the corresponding research idea.

D.3 Step 3: Direct Preference Optimization (DPO)

1. Create an automated preference dataset using the Creativity Index (CI) and User Satisfaction Score (USS) as evaluation metrics.
2. Fine-tune the models further using DPO on the preference dataset. DPO involves optimizing the models to generate research ideas that align with user preferences.
3. The DPO objective function can be represented as:

$$\mathcal{L}_{DPO}(\theta) = -\sum_{i=1}^{M} \log \frac{\exp(s_\theta(y_i|x_i))}{\sum_{j=1}^{K} \exp(s_\theta(y_j|x_i))} \tag{19}$$

where M is the number of preference pairs, K is the number of research ideas for each prompt, and $s_\theta(y|x)$ is the score assigned by the model to the research idea y given the prompt x.

D.4 Step 4: Evaluation Metrics

1. Implement the Creativity Index (CI) as a composite metric that quantifies the originality, feasibility, potential impact, and reliability of AI-generated research ideas.
2. The CI can be computed as:

$$CI = w_O \cdot O + w_F \cdot F + w_{PI} \cdot PI + w_R \cdot R \tag{20}$$

where O, F, PI, and R are the originality, feasibility, potential impact, and reliability scores, respectively, and w_O, w_F, w_{PI}, and w_R are their corresponding weights.
3. Implement the User Satisfaction Score (USS) as a weighted combination of the Keyword Relevance Score (KRS), Readability Score (RS), Coherence Score (CS), Diversity Score (DS), and Predicted Actual User Feedback (PAUF).
4. The USS can be computed as:

$$\begin{aligned} USS = w_1 \cdot KRS + w_2 \cdot RS \quad & (21) \\ + w_3 \cdot CS + w_4 \cdot DS \quad & (22) \\ + w_5 \cdot PAUF \quad & (23) \end{aligned}$$

where w_1, w_2, w_3, w_4, and w_5 are the weights assigned to each component.

D.5 Step 5: Dynamic Generation with Iterative DPO

1. Generate multiple research ideas on randomly generated topics using the fine-tuned models.
2. Rank the generated ideas using the CI and USS metrics.
3. Use the top-ranked ideas as additional DPO training data and repeat the process iteratively to continuously improve the models' performance.
4. The dynamic generation with iterative DPO process can be represented by the following algorithm:

Algorithm 2 Dynamic Generation with Iterative DPO

```
Initialize model $\mathcal{M}$ with the model from Phase 2
for $r = 1, \ldots, R$ do
    Generate a set of random topics $\mathcal{T} = \{t_k\}_{k=1}^{K}$
    for $k = 1, \ldots, K$ do
        Generate research ideas $\mathcal{G}_k = \{g_{kl}\}_{l=1}^{L}$ using $\mathcal{M}$ for topic $t_k$
        Compute scores score($g_{kl}$) for each $g_{kl} \in \mathcal{G}_k$
        Rank $\mathcal{G}_k$ based on scores
        Select top-ranked ideas and create preference pairs $(\tilde{p}_{kl}, g_{kl}, \bar{g}_{kl})$
        Add preference pairs to DPO training data
    end for
    Update model $\mathcal{M}$ using DPO on the augmented training data
end for
```

D.6 Step 6: Evaluation and Comparison

1. Evaluate the performance of the trained models using the CI and USS metrics on the test set.
2. Compare the results with baseline models and state-of-the-art approaches for research idea generation.
3. Conduct ablation studies to assess the contribution of each component in the CI and USS metrics.
4. Perform a user study with domain experts to validate the quality and usefulness of the generated research ideas.

By following these steps and implementing the algorithms and mathematical expressions provided, you can replicate the work done in the paper and evaluate the performance of LLMs in generating novel research ideas using comprehensive training and automated metrics.

E Direct Preference Optimization (DPO)

Direct Preference Optimization (DPO) is a technique used to fine-tune language models to generate outputs that align with user preferences. In the context of research idea generation, DPO is employed to optimize the models to generate research ideas that are more likely to satisfy the preferences of researchers and the scientific community.

E.1 Preference Dataset Creation

To perform DPO, a preference dataset is first created using the Creativity Index (CI) and User Satisfaction Score (USS) as evaluation metrics. The preference dataset consists of pairs of research ideas, where one idea is preferred over the other based on their CI and USS scores.

Let $\mathcal{D} = \{(x_i, y_i, \bar{y}_i)\}_{i=1}^N$ be the preference dataset, where x_i is the input prompt, y_i is the preferred research idea, and $\bar{y}_i$ is the less preferred research idea. The preference pairs are created by selecting research ideas with higher CI and USS scores as the preferred ideas and those with lower scores as the less preferred ideas.

E.2 DPO Objective Function

The goal of DPO is to fine-tune the language model parameters θ to maximize the likelihood of generating the preferred research ideas given the input prompts. This can be achieved by minimizing the following objective function:

$$\mathcal{L}_{DPO}(\theta) = -\sum_{i=1}^{N} \log \frac{\exp(s_\theta(y_i|x_i))}{z_i} \tag{24}$$

$$z_i = \exp(s_\theta(y_i|x_i)) + \exp(s_\theta(\bar{y}_i|x_i)) \tag{25}$$

where $s_\theta(y|x)$ is the score assigned by the model to the research idea y given the prompt x. The score can be computed as the log-likelihood of the research idea under the model's probability distribution.

E.3 DPO Algorithm

The DPO algorithm involves the following steps:

The DPO algorithm iterates over the preference dataset for a specified number of epochs E. In each epoch, the dataset is divided into batches of size b. For each batch, the DPO loss is computed using Equation (1), and the model parameters are updated using gradient descent with a learning rate α.

E.4 Inference

During inference, the fine-tuned model can be used to generate research ideas given input prompts. The generated ideas are then ranked based on their scores assigned by the model, and the top-ranked ideas are considered the most preferred ones.

Algorithm 3 Direct Preference Optimization (DPO)

Initialize model parameters θ
for $epoch = 1, \ldots, E$ **do**
  for $batch = 1, \ldots, B$ **do**
    Sample a batch of preference pairs $\{(x_i, y_i, \bar{y}_i)\}_{i=1}^{b}$ from $\mathcal{D}$
    Compute the DPO loss $\mathcal{L}_{DPO}(\theta)$ for the batch using Equation (1)
    Update the model parameters θ using gradient descent:

$$\theta \leftarrow \theta - \alpha \nabla_\theta \mathcal{L}_{DPO}(\theta) \tag{26}$$

    where α is the learning rate
  end for
end for

Let x be an input prompt and $\mathcal{Y} = \{y_1, y_2, \ldots, y_K\}$ be the set of generated research ideas. The preference score for each generated idea y_k is computed as:

$$s_\theta(y_k|x) = \log p_\theta(y_k|x) \tag{27}$$

where $p_\theta(y_k|x)$ is the probability of the research idea y_k under the fine-tuned model's probability distribution given the prompt x.

The generated ideas are then ranked based on their preference scores, and the top-ranked ideas are considered the most preferred ones.

E.5 Iterative DPO

To further improve the model's performance, an iterative DPO process can be employed. In this process, the top-ranked generated ideas from the previous iteration are used as additional training data for the next iteration of DPO.

Let $\mathcal{G} = \{(x_i, y_i)\}_{i=1}^{M}$ be the set of top-ranked generated ideas from the previous iteration, where x_i is the input prompt and y_i is the corresponding generated idea. These generated ideas are added to the preference dataset $\mathcal{D}$ for the next iteration of DPO.

The iterative DPO process can be summarized by the following algorithm:

Algorithm 4 Iterative Direct Preference Optimization (DPO)

Initialize model parameters θ
for $iteration = 1, \ldots, I$ **do**
  Perform DPO using Algorithm 1 with the current preference dataset $\mathcal{D}$
  Generate research ideas $\mathcal{G}$ using the fine-tuned model
  Rank the generated ideas based on their preference scores
  Select the top-ranked ideas and add them to $\mathcal{D}$ for the next iteration
end for

The iterative DPO algorithm repeats the DPO process for a specified number of iterations I. In each iteration, the model is fine-tuned using the current preference dataset $\mathcal{D}$, and new research ideas are generated using the fine-tuned model. The generated ideas are ranked based on their preference scores, and the top-ranked ideas are added to the preference dataset for the next iteration.

By iteratively refining the preference dataset and fine-tuning the model, the iterative DPO process can continuously improve the model's ability to generate research ideas that align with user preferences.

F Survey Process

To validate the automated preference dataset and ensure its alignment with human preferences, we conducted a survey with a group of 30-50 subject experts, including researchers in the relevant domains. The survey participants were selected based on their expertise and experience in their respective fields, with 3-5 years of research experience and at least 3 publications in top-tier conferences or journals.

The survey was conducted using an online form, which was distributed to the participants via email. For each subject, different sample pairs of research ideas were presented, covering various topics within the subject area. The participants were asked to rate each pair on a scale of 1 to 5, with 1 indicating a strong preference for the first idea and 5 indicating a strong preference for the second idea. A rating of 3 indicated no preference between the two ideas. The preference score for a pair of ideas i and j is denoted as p_{ij} as in equation 28.

$$p_{ij} = \begin{cases} 1, & \text{if idea } i \text{ is strongly preferred over idea } j \\ 2, & \text{if idea } i \text{ is slightly preferred over idea } j \\ 3, & \text{if there is no preference between idea } i \text{ and idea } j \\ 4, & \text{if idea } j \text{ is slightly preferred over idea } i \\ 5, & \text{if idea } j \text{ is strongly preferred over idea } i \end{cases} \tag{28}$$

The survey form included detailed instructions and examples to ensure that the participants understood the rating scale and the criteria for assessing the research ideas. Here are a few example questions presented in the survey:

Question 1: Given the following two research ideas in the field of computer vision, please rate your preference:
Idea A:
Title: A novel deep learning architecture for real-time object detection in high-resolution images.
Abstract: This research proposes a new deep learning architecture that enables real-time object detection in high-resolution images. The architecture combines a lightweight backbone network with a multi-scale feature fusion module and a custom object detection head. The proposed method

achieves state-of-the-art performance on benchmark datasets while maintaining real-time inference speeds. The approach is particularly well-suited for applications such as autonomous driving and video surveillance, where fast and accurate object detection is crucial.
Idea B:
Title: An unsupervised learning approach for image segmentation using graph-based clustering.
Abstract: This study presents an unsupervised learning approach for image segmentation based on graph-based clustering. The proposed method constructs a graph representation of an image, where each pixel is treated as a node, and the edges represent the similarity between pixels. A novel graph clustering algorithm is then applied to partition the graph into coherent segments. The approach is evaluated on multiple image segmentation benchmarks and demonstrates competitive performance compared to state-of-the-art supervised methods, without the need for labeled training data.
Rate your preference (1-5): ______

The participants were also given the option to provide qualitative feedback and comments on the ideas presented.

The survey was conducted over a period of six weeks, with reminders sent to the participants to ensure a high response rate. On average, each participant spent around 30-45 minutes completing the survey, depending on the number of sample pairs presented for their subject area.

Once the survey was completed, the responses were collected and analyzed manually. The ratings provided by the participants were compared with the preferences generated by the automated preference dataset. The agreement between the human ratings and the automated preferences was calculated using metrics such as Cohen's kappa coefficient (κ) and Pearson's correlation coefficient (ρ).

Cohen's kappa coefficient measures the inter-rater agreement between the human ratings and the automated preferences, taking into account the possibility of agreement occurring by chance. It is calculated as:

$$\kappa = \frac{p_o - p_e}{1 - p_e} \tag{29}$$

where p_o is the observed agreement and p_e is the expected agreement by chance.

Pearson's correlation coefficient measures the linear correlation between the human ratings and the automated preferences. It is calculated as:

$$\rho = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \tag{30}$$

where x_i and y_i are the human rating and automated preference for the i-th pair of ideas, respectively, and $\bar{x}$ and $\bar{y}$ are their respective means.

The results of the survey indicated a high level of agreement between the human preferences and the automated preference dataset, with an average Cohen's kappa coefficient of 0.78 and an average Pearson's correlation coefficient of 0.85. These results validate the effectiveness of the automated preference dataset in capturing human preferences and ensuring the alignment of the generated research ideas with the expectations of the research community.

The survey process played a crucial role in validating the automated preference dataset and providing additional insights into the preferences of subject experts. The qualitative feedback and comments provided by the participants were also valuable in refining the preference dataset and improving the overall quality of the generated research ideas.

G Dataset Details

G.1 Dataset Overview

For this study, we utilized the DBLP-Citation-network V14 dataset [35], the most recent version available as of 2023-01-31. The dataset comprises **5,259,858** papers and **36,630,661** citation relationships. The DBLP-Citation-network dataset is designed for research purposes and aggregates citation data from multiple sources, including DBLP, ACM, MAG (Microsoft Academic Graph), and others. Each paper in the dataset is associated with various attributes such as abstract, authors, publication year, venue, and title.

G.2 Data Organization and Attributes

The DBLP-Citation-network V14 dataset is organized into blocks, with each block representing a single paper. The dataset contains key attributes for each paper, including:

- **Title:** The title of the paper.
- **Authors:** The names of the authors who contributed to the paper.
- **Year:** The year the paper was published.
- **Venue:** The conference or journal where the paper was published.
- **Abstract:** A brief summary of the paper's content.
- **Citation Relationships:** The list of references cited by the paper, along with the citations received by the paper.

G.3 Data Schema

The dataset is provided in JSON format, where each line in the text file represents a paper's metadata. The data schema for V14 includes the fields mentioned in table 13.

G.4 Usage and Applications

The DBLP-Citation-network dataset is a valuable resource for a variety of research purposes, including:

- **Clustering with Network and Side Information:** The dataset can be used to perform clustering analyses that incorporate both network structure and additional paper attributes.
- **Citation Network Analysis:** Researchers can study influence within the citation network, identify the most influential papers, and analyze citation patterns.
- **Topic Modeling:** The dataset provides ample data for conducting topic modeling and analyzing trends in academic research.

Table 13. Data Schema for DBLP-Citation-network V14

Field Name	Field Type	Description	Example
id	string	paper ID	453e997ddb7602d9701fd3
title	string	paper title	Rewrite-Based Satisfiability Procedures for Recursive Data Structures
authors.name	string	author name	Maria Paola Bonacina
authors.org	string	author affiliation	Dipartimento di Informatica | Università degli Studi di Verona
authors.id	string	author ID	53f47275dabfaee43ed2565
venue.raw	string	paper venue name	Electronic Notes in Theoretical Computer Science (ENTCS)
year	int	published year	2007
keywords	list of strings	keywords	["theorem-proving strategy", "rewrite-based approach"]
fos.name	string	paper fields of study	Data structure
fos.w	float	fields of study weight	0.48341
references	list of strings	paper references	["53e9a31fb7602d9702c", "53e997f1b7602d9701f"]
n_citation	int	citation number	19
page_start	string	page start	"55"
page_end	string	page end	"70"
doc_type	string	paper type (journal, conference)	Journal
lang	string	detected language	en
volume	string	volume	"174"
issue	string	issue	"8"
issn	string	issn	"Electronic Notes in Theoretical Computer Science"
isbn	string	isbn	""
doi	string	doi	10.1016/j.ent.2006.11.039
url	list	external links	[https://...]
abstract	string	abstract	Our ability to generate...
indexed_abstract	dict	indexed abstract	"IndexLength": 116, "InvertedIndex": "data": [49]
v12_id	int	v12 paper id	2027211529
v12_authors.name	string	v12 author name	Maria Paola Bonacina
v12_authors.org	string	v12 author affiliation	Dipartimento di Informatica, Università degli Studi di Verona, Italy
v12_authors.id	int	v12 author ID	669130765