Personalising Idea Recommendation by Scientific Assistants

Mahmoud Elaref

Supervisor(s): Pradeep K. Murukannaiah , Shubhalaxmi Mukherjee

EEMCS, Delft University of Technology, The Netherlands

In Partial Fulfilment of the Requirements
For the Masters of Computer Science
at Delft University of Technology, EEMCS Faculty
To be defended publically on June 18, 2025 June 12, 2025

Name of the student: Mahmoud Elaref
Masters Computer Science - Artificial Intelligence
Thesis committee: Pradeep K. Murukannaiah, Shubhalaxmi Mukherjee, Jie Yang

Abstract

Large Language Models (LLMs) are increasingly transforming how scientists approach research, with emerging tools supporting ideation, experimentation, and publication in attempts to expedite the research process. This work focuses on the foundational first step: generating novel, high-quality research ideas. Building on an existing LLM idea generation method, we introduce a system for **personalised** research idea generation that generates ideas by analysing a researcher's publication history, providing tailored research recommendations. We evaluate the generated ideas via expert surveys across predefined metrics, offering insights into the system's effectiveness and LLMs' role in ideation in general. Findings show that personalised generation produces novel, interesting, and potentially impactful ideas comparable to baseline methods. Clarity remains a common limitation, while feasibility improves in the personalised variant, highlighting the potential of personalised LLM ideation.

1 Introduction

Recent advancements in Large Language Models (LLMs) have led to increasing interest in developing scientific assistants, systems that support and accelerate different stages of the research process (Lu et al., 2024; Li et al., 2024; Boyko et al., 2023). Tools have been proposed for tasks such as paper reviewing (Liang et al., 2023; Liu and Shah, 2023; Du et al., 2024), literature exploration (Whitfield and Hofmann, 2023), and experimental design.

Scientific research follows an iterative process, in which researchers build on previous work to formulate new hypotheses and explore novel directions. Traditionally, ideation remains a challenging and time-consuming process, requiring extensive knowledge of existing literature, an understanding of research gaps, and creative insight. Researchers spend significant effort brainstorming ideas, evaluating feasibility, and finding potential contributions to their field. LLMs can help in this process by identifying patterns in previous research, suggesting unexplored directions, and proposing innovative ideas (Li et al., 2024; Baek et al., 2025). Among the leading works in this domain, Si et al. (2024) systematically compare LLM-generated ideas with those written by human experts, finding that while LLMs can produce more novel ideas, they often struggle with feasibility. Their study highlights both the promise and current limitations of using LLMs for idea generation.

However, most existing approaches overlook the individual context of the researcher, treating idea generation as a one-size-fits-all task. In practice, a research idea is more valuable if it aligns with a researcher's prior knowledge, interests, and expertise. Personalisation has the potential to bridge a gap by tailoring suggestions to a researcher's academic trajectory, increasing both the excitement and practicality of the generated ideas. This motivates using researcher's previous publications alongside existing literature retrieval techniques. Moreover, there is little work in critically evaluating LLM generated ideas using human evaluators.

So our research question is: **Does personalised literature retrieval improve the quality of LLM-generated research ideas?**

To answer this question, we develop a system inspired by the pipeline from (Si et al., 2024), adapting it to generate personalized research ideas. We propose a system that tailors research recommendations to a researcher's own trajectory. The pipeline can be seen in Figure 1. To do this we generate and evaluate ideas under three generation variants, all explained further in Section 3. A key challenge then is evaluation. We first perform an analysis of the embeddings, and then conduct a user study with expert researchers, to compare the generation variants. Since recruiting domain experts at scale is difficult, our user study is exploratory, involving eight postdoctoral researchers.

We present the following contributions:

C1 We develop a personalised research ideation tool that generates research ideas grounded in both topic-specific literature reviews and the researcher's own prior publications.

C2 We conduct a user study with researchers to evaluate the generated ideas to assess the personalised and mixed generation methods against the non-personalised baseline proposed by Si et al. (2024).

2 Related Work

First, we inspect the growing role of LLMs as scientific assistants, outlining their various domains. Then, we focus on the use of LLMs for research idea generation specifically and how some of their findings influence our design choices. We present the baseline study on which this thesis builds on

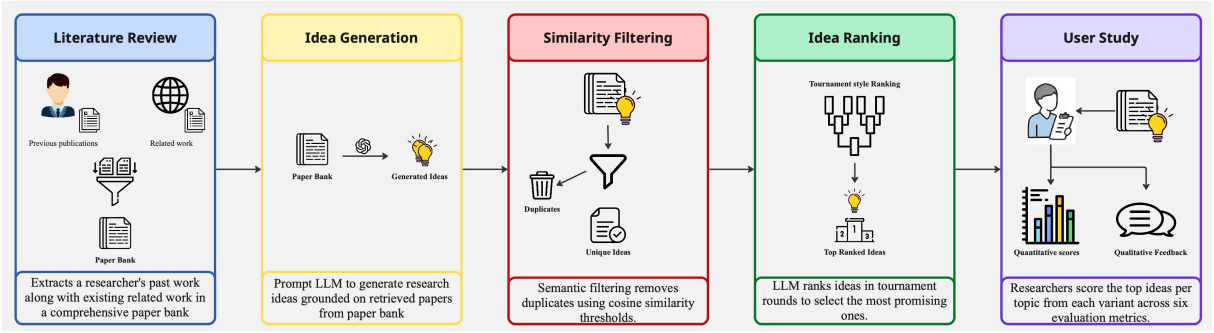


Figure 1: Overview of the methodology pipeline for personalized research idea generation and evaluation. The process begins with the collection of researcher publication histories, followed by idea generation. The generated ideas are then filtered and top-ranked ideas per method are evaluated by the corresponding researcher using a structured survey.

(Si et al., 2024) discussing their pipeline elements that were kept the same, as background knowledge.

2.1 LLMs in Scientific Research

A great deal of research is carried out on LLMs as Research Assistants. This can be seen in many domains: some works attempt hypothesis generation (Yang et al., 2024), others look at scientific paper reviewing and giving feedback, comparing that to human reviewers (Liang et al., 2023; Liu and Shah, 2023; Du et al., 2024). Whitfield and Hofmann (2023) develop a tool to carry out literature reviews for scientists, other tools are designed specifically to summarise research papers. Multiple papers utilise LLMs with Retrieval Augmented Generation (RAG) from specialized sources and large text archives to produce an effective research assistant, outperforming regular question answering LLMs (Garcia and Weilbach, 2023; Zheng et al., 2024). Some works attempt to automate the full research pipeline, ideation, experimentation, execution and writing (Lu et al., 2024; Li et al., 2024; Boyko et al., 2023).

Research shows us that, LLMs are helpful in assisting researchers, they can improve, expedite or boost the research process, however, they struggle on their own, and are most valuable when used in combination with human input.

2.2 LLMs for Idea Generation

The MLR-Copilot produces good results(Li et al., 2024); its IdeaAgent starts with a single research paper (specifically, its title, abstract, introduction, and related work), and uses LLMs to extract research tasks t, research gaps g, and keywords k. That is then used to retrieve relevant literature then generate new hypotheses based on identified trends and gaps in the existing research. We can see that retrieving related literature and using it as context for idea generation is the standard approach in this domain and is the most widely adopted technique. It also forms the basis of the literature-based prompting component in our pipeline Figure 1.

Further motivating the need for grounding idea generation on existing literature, Baek et al. (2025) attempt to mitigate hallucination of LLMs by augmenting with the target paper and knowledge base, to ground the generation process in more relevant and accurate information.

The AI Scientist (Lu et al., 2024) takes inspiration from evolutionary computation to iteratively grow an archive of ideas using LLMs. They note that the idea generation process often results in very similar ideas, even across different runs and models. Therefore, in our project we explicitly ask the LLM to avoid generating similar ideas to the previously generated ones we provided.

A common theme in idea generation with LLMs is the use of iterative refinement (Baek et al., 2025). LLMs may not always generate the best results on the first attempt. Techniques like chain of thought (Wei et al., 2023) and self reflection (Shinn et al., 2023) are often used to refine ideas. Self refinement is when an LLM is continuously prompted to improve on its own ideas (Shridhar et al., 2023; Madaan et al., 2023; Welleck et al., 2022). For example another method states that GPT-4 generates ideas with overall low technical depth and novelty and attempt to target novelty specifically: In SCIMON (Wang et al., 2024), the model takes the background problem contexts and provides suggested solutions that are novel. Our pipeline attempts a similar strategy to keep the LLM generating newer and better ideas by augmenting the previously gen-

erated ideas.

2.3 Background Paper

As mentioned in Section 1 this work builds on the pipeline developed in Si et al. (2024). In this section we will discuss the essential elements of their pipeline that were kept the same, as background knowledge for the methodology in this paper.

The pipeline NoviScl (2024) uses prompting based NLP, providing the LLM with the same instructions as the experts. This involves example ideas, instructions, and template to follow. The first is is a paper retrieval step for Retrieval Augmented Generation (RAG), by querying the Semantic Scholar API with relevant keywords to fetch the most relevant papers in a literature review. These papers are stored in a paper bank and given a score by the LLM based on criteria like their quality and relevance. The second step is generating the ideas grounded on the highest scoring papers and some given examples. Then, duplicates are removed after completing pairwise cosine similarity on embeddings of the generated ideas. The LLM is then prompted to turn the idea into a detailed project proposal, this is so that the evaluators can judge the quality of the generated ideas. The final step is automatic idea ranking using the LLM, adopting a Swiss tournament system where projects are paired with those of a similar score to produce an overall ranking.

The authors crucially commented that: in order to find the best ideas they generate as many candidate ideas as possible, to find the "diamond in the rough". We aim to adopt a similar strategy. Their findings reveal that LLM-generated ideas are statistically more novel than those of human experts, though less feasible.

For evaluation, four metrics were used to compare the ideas, Novelty, Feasibility, Excitement and Potential Effectiveness. The authors state that evaluation criteria can be highly subjective and that it is difficult for even the best experts to judge the quality of an idea. We take measures to mitigate the effects of this in our evaluation, while adding some more evaluation metrics, discussed in Section 3.4.

3 Methodology

The system developed in this work builds directly on the implementation from Si et al. (2024), using their publicly available GitHub repository as the foundation (NoviScl, 2024). While the core structure of the codebase remains largely the same explained in Section 2.3, several modifications were made to better align the system with the objectives of this thesis. This section will highlight these changes and justify the decisions we take to adapt for our research goals.

We adopt a prompting-based NLP framework. Prompting enables Large Language Models to perform complex tasks without additional training, it has become a widely used approach in recent NLP research, offering a balanced trade-off between feasibility and performance (Liu et al., 2021). We used OpenAI's gpt-4o model (OpenAI, 2024).

3.1 Design Choices

Defining a Research Idea

In this work, we define a research idea consisting of 5 key components: (1) the Problem, outlining the core research challenge, (2) the Motivation, explaining why the problem is important, (3) a brief overview of Existing Methods, summarising prior approaches, (4) the Proposed Method, detailing the novel approach, and (5) an Experiment Plan, describing how the idea can be executed in practice.

Based on this structured definition, we modified the original codebase from Si et al. (2024) by excluding the Project Proposal Generation module that expands each seed idea into a detailed project proposal. Instead, we keep the concise, yet informative research ideas in the discussed format. This allows us to streamline the idea generation process, produce a larger number of candidate ideas, to facilitate the user study given our available resources.

Personalisation

While the original implementation generates research ideas based solely on a general topic description and a standard literature review, our aim is to tailor the ideation process more closely to an individual researcher's prior work. To achieve this, we introduce a personalisation module that extracts a subset of the researcher's own publications. These papers are then used to inform idea generation, either exclusively or in combination with other literature, enabling a more context-aware and personalised output. To evaluate the impact of personalisation, we define three distinct experimental variants, shown in Figure 2:

1. **Baseline:** Research ideas are generated by retrieving the top-ranked papers from a general

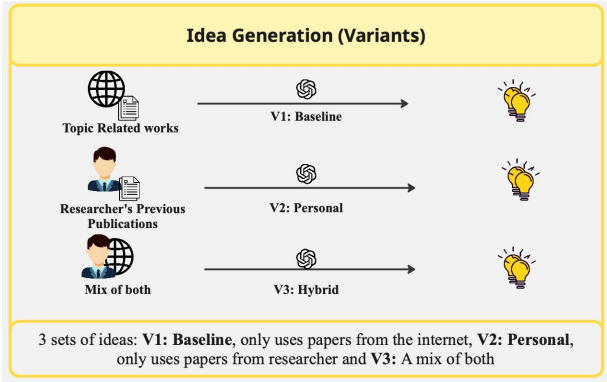


Figure 2: The three variants of idea generation, different in the retrieved literature. Variant 1: Baseline,

literature search.

2. Personal: The literature review is replaced entirely with the previous publications of a single researcher, personalizing the idea generation process.

3. Hybrid: A combination of both variants, using half of the papers from a general literature review and the other half from the researcher's past publications.

Paper selection

To improve consistency and quality in selecting a researcher's papers for the Personal and Hybrid variants, the papers were hand-picked. The intuition behind this choice is that, in practice, researchers themselves would curate and input their most representative or recent papers into the system to guide idea generation aligned with their expertise and interests. We applied the following selection criteria: (1) Topical Relevance: The paper must clearly fall within the defined topic area. (2) Recency: Papers among the most recent published by the author. (3) Reputation: Priority was given to papers published in well-regarded venues. (4) Contribution: The researcher should be the first or second author on the paper to indicate significant involvement.

3.2 Pipeline

Our approach follows a structured pipeline for research idea generation, adapted from the framework proposed by (Si et al., 2024) described in Section 2.3. The pipeline comprises four main stages: **Paper retrieval** (Literature Review), **Grounded Idea Generation**, **Filtering**, and **Ranking**. While maintaining the core structure of the original pipeline, we made several modifications to align with our research goals. Our project pipeline can be seen in Figure 1 and is explained in further detail.

Paper Retrieval:

The first step involves retrieving a relevant set of academic papers to inform the idea generation process. The LLM is prompted with the topic sentence and tasked with generating appropriate search keywords to query the Semantic Scholar API, retrieving metadata for each paper including title, authors and abstract. The papers are collected in a paper bank after which the LLM is prompted to score them based on quality and relevancy. For the Personal and Hybrid variants, the researcher's own selected publications are also added to the paper bank at this stage. The highest scoring papers are retrieved for the next step. This Retrieval Augmented Generation (RAG) (Lewis et al., 2021; Shi et al., 2023) step allows the LLM to ground the idea generation on the relevant literature.

Grounded Idea Generation:

The idea generation module, largely unchanged from the background paper (Si et al., 2024), prompts the LLM to produce research ideas based on the retrieved papers. The prompt follows a predefined template and includes example ideas. The module is repeated for each generation variant (Baseline, Personal, Hybrid), with only the input papers varying according to the specific variant being tested. Example prompt can be found in Appendix D.

Filtering:

To remove semantic duplicates, all generated ideas are encoded using the `all-MiniLM-L6-v2` model from Sentence-Transformers. Cosine similarity is computed pairwise, and any ideas exceeding a threshold of 0.8 are filtered out, for a final set of unique and varied ideas. This section is largely unchanged from the background paper.

Ranking:

We adopt the Swiss system tournament-style ranking module from the original pipeline (Si et al., 2024). In this stage, the LLM evaluates ideas in pairwise comparisons, deciding which of two ideas is more promising according to predefined evaluation criteria. Points are assigned based on wins across rounds, producing a final ranked list of top

ideas. This style of ranking is proven to be effective in LLM ranking (Qin et al., 2024).

3.3 Similarity Experiments

Throughout the development process, we conducted preliminary similarity based analyses to investigate whether there are statistical differences in the output of the different variants. We generated sets of 10 ideas from each variant, encoded them using `all-MiniLM-L6-v2` from Sentence-Transformers, and computed both average and maximum cosine similarities between ideas across the sets. These experiments serve as a sanity check, we wanted to deduce whether changing the literature retrieval method produced different ideas. Therefore we measured the distances between ideas and variants.

3.4 Evaluation Metrics

To evaluate the quality of generated research ideas, we adopt and adapt a set of six scoring metrics inspired by those used in (Si et al., 2024), refined to better suit the context of personalised idea generation. Table 1 outlines each metric, its origin, along with a brief explanation.

Metric	Description
Novelty	Measures how original or creative the idea is compared to existing work.
Feasibility	Evaluates how realistic or practical the idea is to implement as a research project.
Interestingness	Reflects how engaging the idea is perceived to be.
Impact	Assesses the potential academic and/or societal significance of the idea.
Clarity	Evaluates how clearly and understandably the idea is communicated.
Worth Pursuing	A judgment of whether the idea is worth exploring as a future research direction.

Table 1: Evaluation metrics used in the user study. Novelty and Feasibility are retained from Si et al. (2024), Interestingness and Impact are adapted from Excitement and Effectiveness respectively. Clarity and Worth Pursuing are new additions.

3.5 User Study

Si et al. (2024) emphasize the limitations of using LLMs as self-evaluators, advocating for human expert feedback instead. Therefore, to evaluate the impact of personalisation in idea generation, we conduct a user study involving scientific researchers, all subjects were Postdoctoral researchers, for strong research backgrounds and established publication records. Since the generated ideas are personalised, the most qualified evaluators are the researchers themselves. Due to project constraints, we opted for an exploratory study, where the goal is to discover early insights and use them to generate some hypotheses for further exploration.

For each participant, we generate research ideas tailored to their publication history. The ideas are in their field of expertise, with a matching topic description extracted from their personal webpage to closely represent the topic they work on. Along with their previous publications, ideas are generated using the three variants (Baseline, Personal, and Hybrid). The top two ranked ideas per variant are selected, resulting in six ideas in total per participant. These are then presented in a survey where participants rate each idea on the six different metrics, using a 1–10 Likert scale, and provide short justifications for their ratings. This setup enables a direct comparison of the perceived value of ideas using quantitative scores. Qualitative feedback provides insights into why certain ideas are preferred over others. Examples of the generated ideas are in Appendix E

This experimental design allows us to systematically assess the impact of personalization in LLM-based research ideation and explore its potential to enhance AI-assisted scientific discovery.

3.6 Analyses

Our evaluation involves both quantitative and qualitative analyses to comprehensively assess the effectiveness of each idea generation variant.

Quantitative Analysis. We employ the Kruskal-Wallis test to compare the distributions of more than two ordinal samples (5% significance level), said to be most appropriate for Likert-type data (Eiselen and Van Huyssteen, 2023).

Qualitative Analysis. In addition to numerical scores, participants provide short written justifications for each rating. These text responses form the basis of our qualitative analysis, through which we uncover deeper insights into how researchers perceive the generated ideas.

4 Results

4.1 Similarity Experiments

We conducted similarity experiments to explore whether statistical differences exist between ideas generated by the different variants.

In the experiments we looked at the cosine similarity of embeddings of the ideas across the different variants, the same method used by (Si et al., 2024) to find duplicate ideas. For the results in Figure 3 we computed the maximum cosine similarity for each idea to ideas from another variant, and within the same variant. While the experiments revealed some minor differences and high similarity numbers, the conclusions indicate that cosine similarity is not able to capture many differences between the variants. This is because ideas follow the same structure, use similar domain-specific terminology, and are generated on the same topics. Cosine similarity can signal textual diversity but is limited in assessing the *quality* or *effectiveness* of ideas, which is ultimately the intended difference between the variants. Hence, a user study is essential to evaluate whether one variant produces ideas that are actively more favourable to researchers. More results can be found in Appendix A.

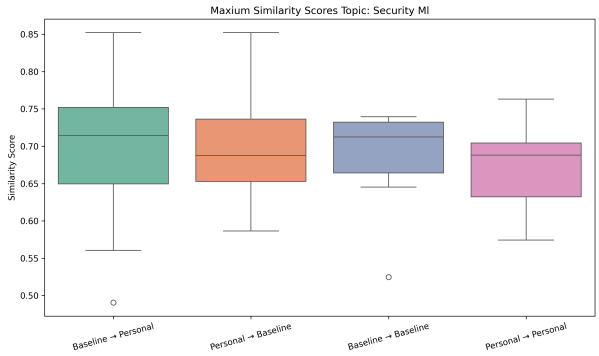


Figure 3: Graph of the distribution of each idea's highest similarity score with one in another set. Set A is the Baseline variant and Set B is the Personal variant. This graph is for ideas from the topic of Security in ML.

4.2 User Study

Eight researchers participated in the study in eight different areas of expertise with ideas generated for each. Summary statistics are shown in Table 2. We will first present some observations we derived from the graphs and Table, then we will analyse what some of the Qualitative feedback indicates while diving into a deeper discussion.

Condition	Mean	Highest	SD	p
Novelty				
Baseline	5.69	8.0	2.57	
Personal	**5.88**	9.0	2.13	0.65
Hybrid	5.12	8.0	2.36	
Feasibility				
Baseline	5.38	10.0	2.55	
Personal	**6.12**	10.0	2.73	0.62
Hybrid	5.31	9.0	2.3	
Interestingness				
Baseline	5.38	8.0	2.28	
Personal	**6.06**	10.0	2.35	0.29
Hybrid	4.81	9.0	2.37	
Impact				
Baseline	6.0	9.0	2.22	
Personal	**6.31**	10.0	2.36	0.52
Hybrid	5.44	8.0	2.34	
Clarity				
Baseline	4.31	9.0	2.65	
Personal	4.44	8.0	2.8	0.89
Hybrid	**4.75**	9.0	2.74	
Worth Pursuing				
Baseline	5.19	9.0	2.51	
Personal	**5.44**	10.0	3.08	0.47
Hybrid	4.38	9.0	2.66	

Table 2: Scores of the three experiment variants across all review metrics. Largest mean value for every metric is in bold. Highest score, Standard Deviation, Kruskal-Wallis p-values are presented for each metric. N is 16 ideas per variant, 48 ideas total.

Observation 1 *The personal variant demonstrates the highest averages across all metrics*

While the differences are not statistically significant, Table 2 shows that the personal variant consistently achieves higher average score across all metrics, except for Clarity. This trend suggests that grounding idea generation in a researcher's prior work can lead to more relevant and higher-quality research ideas.

Observation 2 *Clarity scores are very low for all variants.*

Scores for clarity are by far the lowest, across all variants, as seen in Table 2, this follows a limitation of LLMs, vagueness. This also aligns with feedback from (Si et al., 2024) where the AI generated ideas were said to be "too general and vague". This points to future possible improvements in idea clarity, like presenting the full project proposal with implementation steps instead of the idea alone. This could also present an opportunity of using chain of thought technique that helps the LLM produce better ideas due to having to write it all out with execution details.

Observation 3 *The personal variant had the highest rating among best rated ideas*

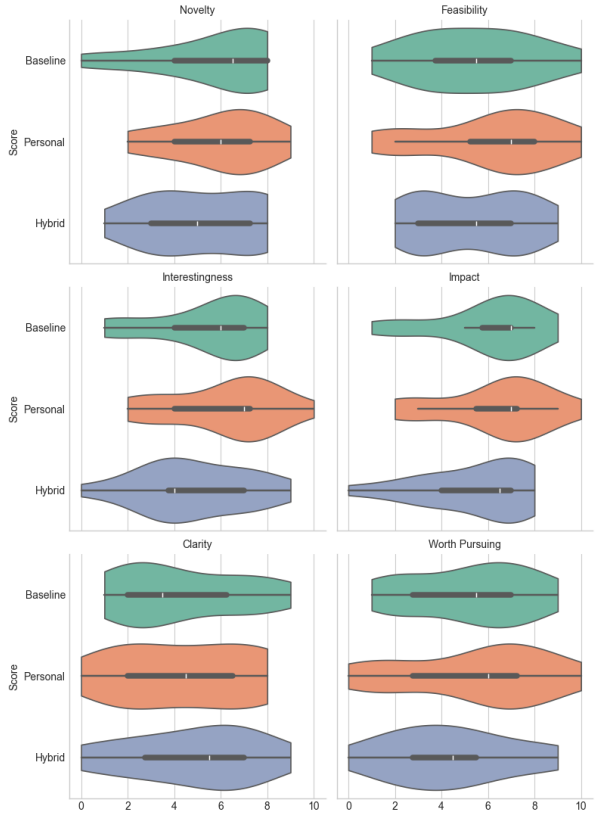


Figure 4: Violin graph showing the distribution of the three experiment variants across all 6 metrics.

The 'Highest' score in Table 2 and plots in Figure 4 consistently show that the personal variant (orange) is the superior compared to the other methods. The personal variant holds the highest score in every metric (except clarity). This suggests that while the numbers may show a lot of variability, the personalised approach is capable of producing standout, high-potential ideas.

Observation 4 *The hybrid variant struggles*

The Hybrid variant clearly underperformed across all metrics. While the baseline variant benefits from recent literature and the personal one from alignment with the researcher's history, the Hybrid variant appears to fall short, failing to fully leverage the strengths of either variant, resulting in lower overall idea quality.

4.3 Qualitative Analysis

Qualitative analysis is crucial, because the numbers are limited in interpretability, due to evaluator bias. E.g. a score of '6' for one evaluator might mean something different than another. Therefore in our qualitative analysis, we manually reviewed all the written justifications provided by participants. First, we categorised the responses into 3 buckets: positive, negative, and mixed sentiments. Looking for common reasons behind both high and low scores, grouping and counting the explanations by metric. From there, we were able to extract themes and patterns that appeared across responses. This helped identify what worked well, and the common limitations allowing us to gain deeper insight into why certain ideas were successful, providing actionable direction for improving the system. For each metric, we summarize the dominant sentiment category and highlight the most common reasons participants gave for high or low ratings.

Novelty and Interestingness.

- 50% of ideas were seen as novel or containing novel elements for Baseline and Personal.
- Baseline and Personal variants 70% and 85% of ideas were seen as interesting, respectively.
- Reasons for high interesting scores: Idea is novel or idea is in area of expertise.
- Leading reason for low scores: Ideas unclear.

Feasibility

- Around 50% of ideas were deemed feasible, for all variants.
- Reasons for unfeasible ideas were: Vagueness and unclarity, dataset difficulties.

Clarity struggles

- Lowest scoring across all metrics, only 20% of ideas were fully understandable.
- Even good ideas had some missing details and aspects.

Impact

- 85% of Personal variant ideas were deemed impactful.
- Baseline and Hybrid far behind at 50% and 35%.
- Low scores attributed to ideas being too niche or unclear.

Worth Pursuing. After analysis we found this metric to be some kind of summary for all other metrics. The Personal variant again performed best, with reviewers saying. "Yes it would be worth pursuing this idea" and even "I had this idea before." Low scores here were typically due to weaknesses in other metrics for example, low novelty, high effort (low feasibility), or vague (low clarity). Evaluators often said additional work was

needed before the idea could be considered viable.

5 Limitations

Next we discuss some possible limitations of this project and potential improvements.

L1 Survey fatigue and evaluation subjectivity. Evaluating research ideas is inherently subjective and cognitively demanding, particularly in surveys. Fatigue may affect scores, with more effort spent on early responses for example. To mitigate this, order of ideas was randomized. While judging unimplemented ideas is difficult, we argue that researchers are best suited to assess ideas tailored to their own work aligning with our goal of testing LLMs for personalised ideation.

L2 Extracting only abstracts from Research papers. Our system generates literature-grounded ideas using only titles and abstracts, following prior work, including the baseline. While some recent studies incorporate introductions and related work, it remains an unanswered question whether this improves idea quality.

L3 Baseline variant may over-rely on topic sentence The personalised variant of our system benefits from hand-picked prior publications provided by the researcher, ensuring a clearer topical focus and more grounded idea generation. In contrast, the baseline variant retrieves papers based solely on the our extracted topic sentence and relies on the LLM to assess papers' quality. Even though this may affect idea quality, it highlights the value of personalisation.

L4 Novelty filtering omitted Unlike the original background system, our pipeline omits the novelty-checking module that queries the Semantic Scholar database to discard duplicate ideas. Our intuition is that without the previous step of expanding the idea into a project proposal, some ideas might be wrongly discarded. We chose not to include this component, even though this might affect the Novelty score.

6 Future Work

Several avenues for future improvement emerged from the feedback and early analysis:

Feedback-Driven Self-Refinement. One promising direction is incorporating evaluator feedback into a self-refinement loop. Valuable reviewer comments could be used to re-prompt the LLM, allowing it to revise or elaborate on promising ideas, leading to more complete and tailored outputs, addressing the weak points.

From Ideas to Proposals. While some generated ideas showed strong potential, they lacked implementation-ready detail. A system to expand selected ideas into full project proposals, filling in missing elements would be useful.

Paper Ordering as a Factor. Another unexplored dimension is the order in which input papers are presented. Including a temporal element might help the LLM better understand the evolution of a field and improve idea quality.

7 Conclusion

This project explored the potential of large language models (LLMs) to generate personalised research ideas by comparing three generation strategies. In this exploratory research, we aimed to find if this avenue is worth further exploration. Results suggest that tailoring idea generation to a researcher's prior publications yields interesting and potentially impactful ideas comparable to baselines. While clarity remained a consistent challenge across all variants, feasibility scores improved for the personal variant. Qualitative feedback confirmed that, despite occasional vagueness or impracticality, researchers often saw value and inspiration in the generated ideas, especially in the personalised ones. The hybrid approach underperformed, indicating that combining personal papers and existing ones may require more sophisticated integration techniques. Overall, under specific circumstances, LLMs are able to generate good ideas, even if it takes some iterations to find the most valuable ones.

8 Ethical Considerations

This work explores the use of LLMs to generate research ideas. Our goal is supporting academic creativity and discovery. While the system is intended to assist researchers, there is a risk that the system may be misused. Mass production of low quality ideas and over-reliance on LLM-generated content may dilute academic standards. This system is designed as a support tool requiring active human engagement and critical evaluation. Researchers remain fully responsible for the ideas they choose to develop and submit. Ensuring accountability, maintaining academic integrity, and applying the same rigorous standards to AI-assisted outputs as to traditional research are essential to

the responsible use of this technology.

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.

James Boyko, Joseph Cohen, Nathan Fox, Maria Han Veiga, Jennifer I-Hsiu Li, Jing Liu, Bernardo Modenesi, Andreas H. Rauch, Kenneth N. Reid, Soumi Tribedi, Anastasia Visheratina, and Xin Xie. 2023. An interdisciplinary outlook on large language models for scientific research. *ArXiv*, abs/2311.04929.

Jiangshu Du, Yibo Wang, Wenting Zhao, Zhongfen Deng, Shuaiqi Liu, Renze Lou, Henry Peng Zou, Pranav Narayanan Venkit, Nan Zhang, Mukund Srinath, Haoran Ranran Zhang, Vipul Gupta, Yinghui Li, Tao Li, Fei Wang, Qin Liu, Tianlin Liu, Pengzhi Gao, Congying Xia, Chen Xing, Jiayang Cheng, Zhaowei Wang, Ying Su, Raj Sanjay Shah, Ruohao Guo, Jing Gu, Haoran Li, Kangda Wei, Zihao Wang, Lu Cheng, Surangika Ranathunga, Meng Fang, Jie Fu, Fei Liu, Ruihong Huang, Eduardo Blanco, Yixin Cao, Rui Zhang, Philip S. Yu, and Wenpeng Yin. 2024. Llms assist nlp researchers: Critique paper (meta-)reviewing.

Roald Eiselen and Gerhard Van Huyssteen. 2023. A comparison of statistical tests for likert-type data: The case of swearwords. *Journal of Open Humanities Data*, 9.

Giselle Gonzalez Garcia and Christian Weilbach. 2023. If the sources could talk: Evaluating large language models for research assistance in history. *ArXiv*, abs/2310.10808.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2021. Retrieval-augmented generation for knowledge-intensive nlp tasks.

Ruochen Li, Teerth Patel, Qingyun Wang, and Xinya Du. 2024. Mlr-copilot: Autonomous machine learning research based on large language models agents.

Weixin Liang, Yuhui Zhang, Hancheng Cao, Binglu Wang, Daisy Ding, Xinyu Yang, Kailas Vodrahalli, Siyu He, Daniel Scott Smith, Yian Yin, Daniel A. McFarland, and James Zou. 2023. Can large language models provide useful feedback on research papers? a large-scale empirical analysis. *ArXiv*, abs/2310.01783.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2021. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing.

Ryan Liu and Nihar B. Shah. 2023. Reviewergpt? an exploratory study on using large language models for paper reviewing. *ArXiv*, abs/2306.00622.

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.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, Shashank Gupta, Bodhisattwa Prasad Majumder, Katherine Hermann, Sean Welleck, Amir Yazdanbakhsh, and Peter Clark. 2023. Self-refine: Iterative refinement with self-feedback. In *Advances in Neural Information Processing Systems*, volume 36, pages 46534–46594. Curran Associates, Inc.

NoviScl. 2024. Ai-researcher. `https://github.com/NoviScl/AI-Researcher`.

OpenAI. 2024. Gpt-4o technical report. `https://openai.com/index/gpt-4o`.

Zhen Qin, Rolf Jagerman, Kai Hui, Honglei Zhuang, Junru Wu, Le Yan, Jiaming Shen, Tianqi Liu, Jialu Liu, Donald Metzler, Xuanhui Wang, and Michael Bendersky. 2024. Large language models are effective text rankers with pairwise ranking prompting.

Weijia Shi, Sewon Min, Michihiro Yasunaga, Minjoon Seo, Rich James, Mike Lewis, Luke Zettlemoyer, and Wen tau Yih. 2023. Replug: Retrieval-augmented black-box language models.

Noah Shinn, Federico Cassano, Edward Berman, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. 2023. Reflexion: Language agents with verbal reinforcement learning.

Kumar Shridhar, Koustuv Sinha, Andrew Cohen, Tianlu Wang, Ping Yu, Ram Pasunuru, Mrinmaya Sachan, Jason Weston, and Asli Celikyilmaz. 2023. The art of llm refinement: Ask, refine, and trust.

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

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

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-thought prompting elicits reasoning in large language models.

Sean Welleck, Ximing Lu, Peter West, Faeze Brahman, Tianxiao Shen, Daniel Khashabi, and Yejin Choi. 2022. Generating sequences by learning to self-correct.

Sharon Whitfield and Melissa A. Hofmann. 2023. Elicit: Ai literature review research assistant. *Public Services Quarterly*, 19(3):201–207.

Zonglin Yang, Xinya Du, Junxian Li, Jie Zheng, Soujanya Poria, and Erik Cambria. 2024. Large language models for automated open-domain scientific hypotheses discovery.

Yuxiang Zheng, Shichao Sun, Lin Qiu, Dongyu Ru, Cheng Jiayang, Xuefeng Li, Jifan Lin, Binjie Wang, Yun Luo, Renjie Pan, Yang Xu, Qingkai Min, Zizhao Zhang, Yiwen Wang, Wenjie Li, and Pengfei Liu. 2024. Openresearcher: Unleashing ai for accelerated scientific research.

A Full Idea Similarity Experiments

We conducted similarity experiments to explore whether noticeable differences exist between ideas generated by the different variants. Cosine similarity of the embeddings are used.

The first experiment Figure 5 shows the average cosine similarity of embeddings of the ideas across the different variants. Focusing on Baseline vs Personal variants we can see subtle differences when comparing ideas from different variants in comparison to ideas within the same variant, however the difference is not significant enough to draw any conclusions. The second experiment Figure 6 shows box plots when finding the idea with highest similarity to another, again across different generation variants and within the same variant. Once more, differences are too small to draw any conclusions. This experiment is also repeated on all the final generated idea topics in Figure 8. In the third and final experiment, Figure 7 we graph the maximum similarity for each idea but with the same set across different seed runs, the point was to find how much ideas vary from one generation method to the other across different runs.

While these experiments revealed some minor differences, the conclusions indicate that cosine similarity metrics do not capture statistically significant differences between the variants. This is expected, as ideas follow the same structure, use domain-specific terminology, and are based on the same topics. Cosine similarity can signal textual diversity but is limited in assessing the *quality* or *effectiveness* of ideas, which is ultimately the intended difference between the variants. Hence, a user study is essential to evaluate whether one variant produces ideas that are genuinely more favourable to researchers.

A.1 Final similarity results

After the preliminary runs shown above (Figure 5, 6, 7), we ran the experiments on the final ideas for each generated topic. (Figures: 8a, 8b, 8c, 8d, 8e, 8f, 8g, 8h)

B Correlation Analysis

To understand how our evaluation metrics relate to one another and whether they capture distinct dimensions of idea quality, we performed a correlation analysis across all scores, in Table 3. This was partly motivated by the observation that many reviewer comments under Interestingness overlapped with Novelty. The analysis indeed revealed a strong positive correlation between Interestingness and Novelty ($r = 0.72$). 'Worth Pursuing' showed moderate-to-strong correlations with several metrics, particularly Impact ($r = 0.71$), Interestingness ($r = 0.66$), and Novelty ($r = 0.56$), indicating it may reflect an overall value judgment as discussed. The results indicate that the metrics have some overlap.

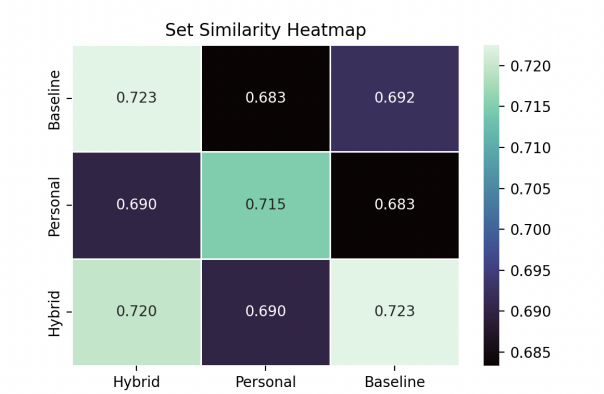


Figure 5: Preliminary experiments: Average cosine similarity of embeddings of the ideas across the different variants. This graph is for the top 3 ideas in the topic of Argument Mining.

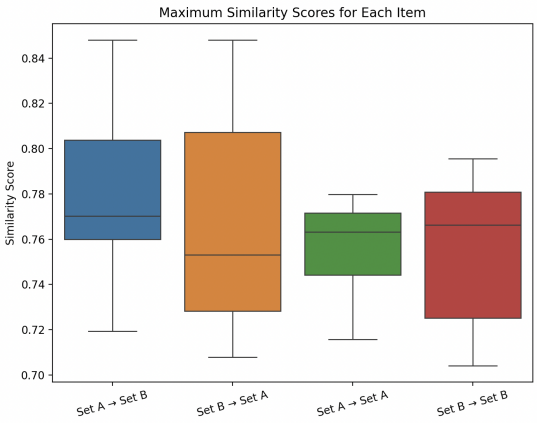


Figure 6: Preliminary experiments: Graph of the distribution of each idea's highest similarity score with one in another set. Set A is the Baseline variant and Set B is the Personal variant. This graph is for ideas from the topic of Argument Mining.

Table 3: Table showing pairwise Pearson correlation between all metrics. Symmetric matrix

	Novelty	Feasibility	Interestingness	Impact	Clarity	Worth Pursuing
Novelty	1.00	-0.23	0.72	0.48	0.03	0.56
Feasibility	-0.23	1.00	-0.22	0.03	0.48	0.15
Interestingness	0.72	-0.22	1.00	0.73	0.21	0.66
Impact	0.48	0.03	0.73	1.00	0.41	0.71
Clarity	0.03	0.48	0.21	0.41	1.00	0.41
Worth Pursuing	0.56	0.15	0.66	0.71	0.41	1.00

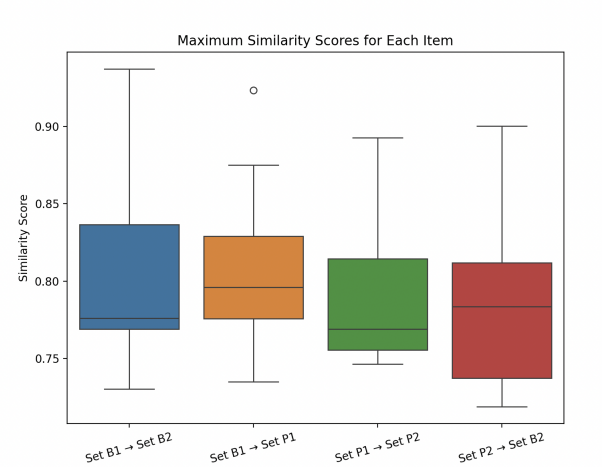


Figure 7: Preliminary experiments: Graph of maximum similarity like Figure 3 but with the same set across different seed runs. Set B is the Baseline variant and Set P is the Personal variant, 1 and 2 refers to the different seed runs. This graph is for ideas from the topic of Argument Mining.

C Survey Instructions

Welcome to the survey on Research Idea Generation using Large Language Models, Thank you for taking the time to help me with this evaluation on my Master Thesis.
You will be presented with 6 different, detailed Research Ideas related to your Research Area. You will be asked to evaluate each of the ideas by answering a few questions.
There are 5 Questions on each idea, revolving around 6 metrics: Novelty, Feasibility, Interestingness and Impact, Clarity and Worth Pursuing. You will be asked to grade each idea on each of the 6 metrics on a 1-10 scale. And then write a few sentences of rationale and explanation for that score.

C.1 Metrics Explanation

Following is a brief explanation of each of the metrics for clarity, and details for the scale numbering.

Novelty Score:
Whether the idea is creative and different from existing works on the topic, and brings fresh insights. You are encouraged to search for related works online.
1: Not novel at all - there are many existing ideas that are the same.
10: Very novel - very different from all existing ideas in a very interesting and clever way

Feasibility Score:
How feasible it is to implement and execute this idea as a research project? Specifically, how feasible the idea is for yourself to execute given resources you typically use.
1: Impossible: the idea doesn't make sense or the proposed experiments are flawed and cannot be implemented.
10: Easy: The whole proposed project can be quickly executed with my current resources and skills.

Interestingness Score:
How exciting and interesting you find this idea.
1: Poor: You cannot identify the contributions of this idea, or it's not interesting at all.
10: Transformative: would change the research field profoundly and worth a best paper award at major AI conferences.

Impact Score:
How likely the proposed idea is going to have both Societal and Research Impact.
1: Extremely Unlikely: The idea has major flaws and definitely won't work well
10: Definitely Effective: You are very confident that the proposed idea will be impactful if executed as a full project.

Clarity Score:
How clear, well-defined, and easy to understand the research idea is. This includes whether the Methods, and Experimental plan of the idea are easy to grasp and well presented.
1: Very unclear: The idea is confusing, vague, or poorly structured. It is hard to understand what the research is about or how it would be carried out.
10: Extremely clear: The idea is presented in a very clear, logical, and coherent manner. The problem, proposed methods, and experimental plan are all easy to understand.

D Idea Generation Prompt

The following prompt was used to instruct the language model to generate research ideas, from Background paper Si et al. (2024).

You are an expert researcher in AI. Now I want you to help me brainstorm some new research project ideas on the topic of: *[topic description]*.
If available, here are some relevant papers on this topic just for your background knowledge:
[Relevant papers inserted here]
You should generate N different ideas on this topic. Try to be creative and diverse in the idea generation, and do not repeat any similar ideas. The above papers are only for inspiration — you should not cite them or simply make incremental modifications. Instead, make sure your ideas are novel and distinct from prior literature.
You should aim for projects that could potentially win best paper awards at top AI conferences like ACL and NeurIPS.

Each idea should be described as follows:
Problem: State the problem statement, which should be closely related to the topic and address a limitation that large language models still face. Existing Methods: Mention some existing benchmarks and baseline methods, if applicable. Motivation: Explain the inspiration behind the proposed method and why it could be effective. Proposed Method: Propose your new method and describe it in detail. The method should be maximally different from existing work, more advanced and effective than current baselines. Be as creative as possible — unhinged ideas that sound crazy are welcome. Experiment Plan: Specify the experimental steps, baseline comparisons, and evaluation metrics. You can follow the formatting of the following example ideas (but do not copy the ideas themselves):
[Example ideas inserted here]

Make sure your ideas are your own and different from the following existing ideas:
[List of existing ideas to avoid]
Focus on proposing novel empirical methods, which can include prompting strategies, fine-tuning, inference-time interventions, etc. The Proposed Method section should include all relevant details, such as data collection, training objectives, and prompt construction.
Please write down your N ideas. Each idea should be described as a single paragraph and returned in JSON format as a dictionary. The dictionary should use a short idea title (e.g., "Non-Linear Story Understanding" or "Multi-Agent Negotiation") as the key, and the full idea description (following the above format) as the value.
Do not repeat idea names or content.

E Example Ideas

This section includes selected ideas, one from each variant. 3 different topics.

E.1 Example Idea: 'Interactive Human-AI Collaboration in Fact-Checking'

This is an example of generated idea from Variant: 'Baseline' in the topic of 'Information Retrieval for Fact Checking'

Interactive Human-AI Collaboration in Fact-Checking
Problem: Automated systems often lack nuanced human understanding and the ability to adaptively learn from experts during fact-checking.
Existing Methods: Current AI-assisted approaches are largely static without continuous human model refinement. **Motivation**: Incorporating a feedback loop where human experts interact directly with AI to refine its understanding provides adaptability and experiment-driven learning opportunities, blending human expertise with AI efficiency.
Proposed Method: Develop an interactive human-AI fact-checking platform where users contribute real-time reviews on LLM predictions. Utilize active learning techniques to refine model predictions based on human feedback, deploying a reinforcement learning pipeline where human-provided verdicts update model parameters. Include interactive visualization tools for experts to investigate reasoning paths within an LLM, fostering a collaborative environment where human insight improves AI decision-making. Construct interfaces that adapt with knowledge acquisition by prompting iterative refinements and recording human decisions to seed further model training.
Experiment Plan: Establish controlled environments with expert fact-checkers to engage with the system and collect high-quality feedback data. Evaluate system improvement with iterative A/B testing against baseline predictions without human intervention. Measure success with metrics like improvement over initial accuracy, user satisfaction ratings, percentage of improved claim verifications, and system adaptability as echoed by feedback loop effectiveness.

E.2 Example Idea: 'Adversarial Knowledge Injection for Secure Reasoning'

This is an example of generated idea from Variant: 'Personal' in the topic of 'Security aspects for ML'.

Adversarial Knowledge Injection for Secure Reasoning Problem: Existing models are prone to adversarial attacks that compromise the integrity and security of their decision-making processes. The challenge is to leverage neuro-symbolic AI to develop adversarial thinking that secures reasoning abilities.
Existing Methods: Current state-of-the-art models focus on enhancing prediction accuracy or reasoning capabilities, but very few are robust against adversarial conditions.
Motivation: The integration of symbolic reasoning with neural models theoretically offers a robust framework capable of cross-verifying decisions against established logical constructs, thereby increasing security against adversarial attacks.
Proposed Method: We propose a multi-tier neuro-symbolic architecture involving Symbolic Constraint Injection (SCI). SCI interacts with neural layers to inject logical constraints dynamically, which align model outputs with secure symbolic rules. We use a hybrid training paradigm with adversarial training loops that simulate attacks and enforce compliance with symbolic logic under threat conditions. SCI serves as a symbolic firewall, dynamically adapting rules based on perceived threats, thus reinforcing model integrity.
Experiment Plan: Experiments will explore dataset benchmarks like CIFAR-10 for image recognition and real-time cybersecurity datasets for intrusion detection. We will evaluate baseline models against SCI-enhanced models focusing on metrics such as attack success rates, accuracy retention post-attack, and computational overhead for logical checks. Baseline models will consist of standard neural networks without SCI integration.

E.3 Example Idea: 'Cultural Contextualization in Hate Speech Detection'

.

This is an example of generated idea from Variant: 'Hybrid' on the topic of 'Hate Speech in NLP'.

Cultural Contextualization in Hate Speech Detection
Problem: "Current hate speech detection models fail to incorporate cultural subtleties, leading to inaccuracies when applied across diverse communities.
Existing Methods: "Traditional hate speech detection systems rely on static linguistic features, primarily trained on Western datasets, often missing cultural nuances.
Motivation: "Cultural context deeply influences language interpretation. A culturally aware model can use specific features derived from cultural practices and linguistic variations, improving accuracy and reducing false positives in cross-cultural settings.
Proposed Method: "Develop a framework for cultural embedding, where we augment datasets with cultural context features and integrate them within transformer-based models. This involves collecting culturally diverse datasets, performing BERT-based cultural embeddings that capture socio-linguistic elements, and fine-tuning language models on this enriched dataset. We will also explore prompt engineering using cultural context clues, allowing the model to dynamically switch contexts or interpretations based on cultural hints present in the input data.
Experiment Plan: "Use multilingual benchmarks (e.g., HASOC) across culturally diverse languages. Evaluate models on precision, recall, and F1-score, with a particular focus on cross-cultural adaptability. Baselines include language models trained on homogeneous datasets without cultural integration.

F User Study Results

F.1 Quantitative Results

Figure 9 Shows a plot of the results for all metrics for each variant.

F.2 Qualitative Feedback

This section shows feedback given on the 3 example ideas presented in Appendix E. Each metric score and justification is given.

Feedback for idea E.1

Novelty 8.0: The idea is quite novel. While online evaluation has been used to benchmark fact checking systems leveraging human in the loop with online RL is an interesting and novel direction.
Feasibility 7.0: While the idea is novel, it is not clear whether the updates to the system based on feedback happens in online or offline setting. Online RL updates would be slow and impractical. Offline RL would be happening periodically and is also resource intensive. It is also not clear if feedback from users would be denoised and if so how. And is human feedback only on reasoning aspect or also on retrieval aspect where new relevance judgments would be collected from users ? The feasibility of this idea is quite challenging given the current state-of-the-art but not impossible.
Interestingness 8.0: Human AI collaboration in automated fact-checking is an interesting idea.
Impact 9.0: It has immense impact on fostering trust in such automated fact-checking systems.
Clarity 6.0: Some aspects are not clear. The questions have been posed in the feasibility rationale text box.
Worth pursuing 9.0: It is of high impact and hence immensely worth pursuing.

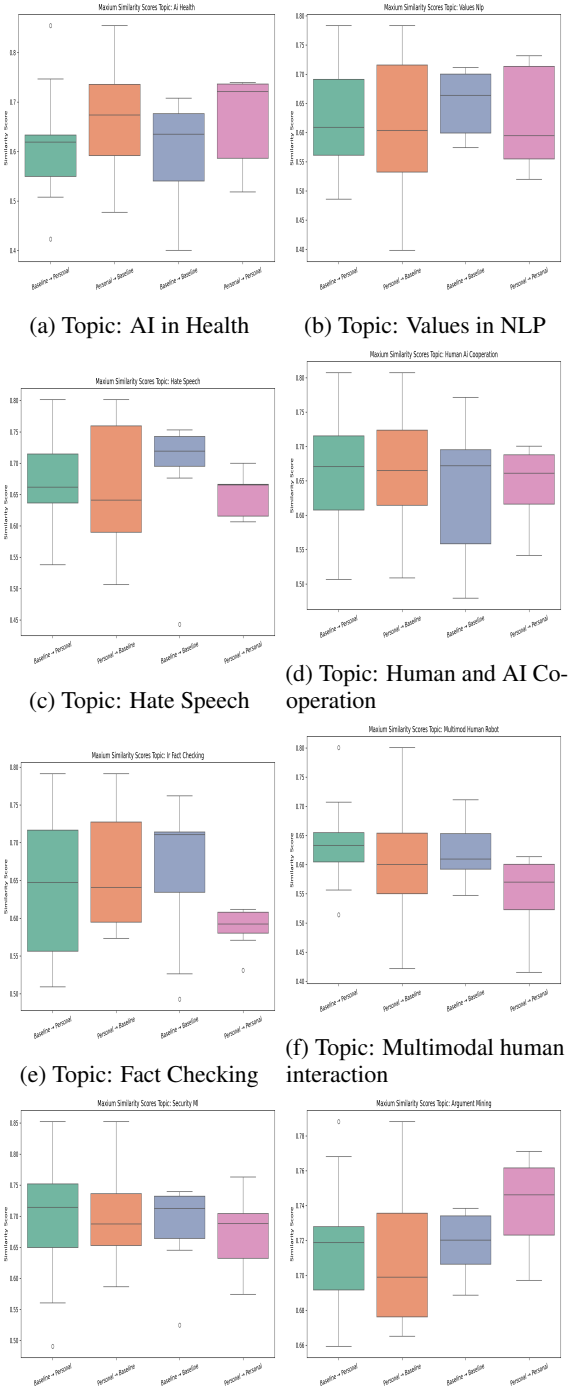
Feedback for idea E.2

Novelty 8.0: The idea of leveraging neuro-symbolic integration to protect models against adversarial threats is definitely interesting and novel.
Feasibility 6.0: The high level idea of this project seems feasible. However, as it lacks some implementation details, it is not easy to tell for sure.
Interestingness 8.0: The idea is definitely interesting since checking the capability of NeSy integration to define adversarially-robust models would provide valuable insights and would enable extending the application domain of NeSy.
Impact 7.0: The impact of the idea might be highly positive if the results will provide positive findings on the effect of NeSy for defending against adversarial attacks. Meanwhile, if the results will highlight a negative or neutral impact of NeSy integration on the adversarial robustness of models, the impact might be limited. Therefore, I would say that on average the impact of the idea might be moderately positive.
Clarity 4.0: While interesting, the idea of enforcing compliance of symbolic constraints against adversarial attacks is not made completely clear. How is the enforcing process enabled?
Worth Pursuing 8.0: The novelty and possible positive impact of the idea makes it worth pursuing even if it is not clearly defined in its details. Indeed, I am pursuing a closely related idea in my research agenda.

Feedback for idea E.3

Novelty 3.0: It is hard to tell how the idea wants to achieve its goal of a culturally-aware model. To me it is not clear that cultural embeddings are novel.
Feasibility 5.0: While prompt-engineering is definitely doable, I am unsure of what e.g., "performing BERT-based cultural embeddings" really means? There are parts of the idea that are just unclear or unfinished to me.
Interestingness 4.0: I think it is trying to do a lot at once (dataset, fine-tuning, prompting) that it is unclear what lies at the core of the research idea.
Impact 6.0: Incorporating culture is of high value, though in its current (unclear) setup, I am not sure how much impact can be made.
Clarity 2.0: The setup is really unclear to me, how they idea will be addressed.
Worth Pursuing 4.0: In its current state, there are just too much unclarities and ideas happening at once.

G Responsible NLP Checklist



(a) Topic: AI in Health (b) Topic: Values in NLP

(c) Topic: Hate Speech (d) Topic: Human and AI Co-operation

(e) Topic: Fact Checking (f) Topic: Multimodal human interaction

(g) Topic: Security in ML (h) Topic: Argument Mining

Figure 8: Graph of the distribution of each idea's highest similarity score with one in another set.

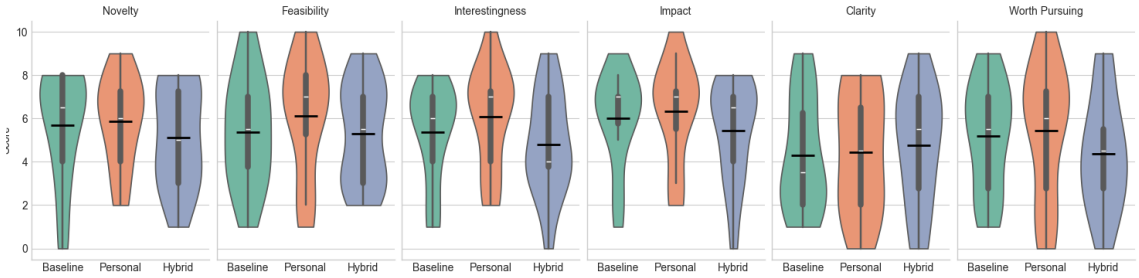


Figure 9: Violin plot showing scores of the three experiment variants across all review metrics Black line is the mean and white line is the median.

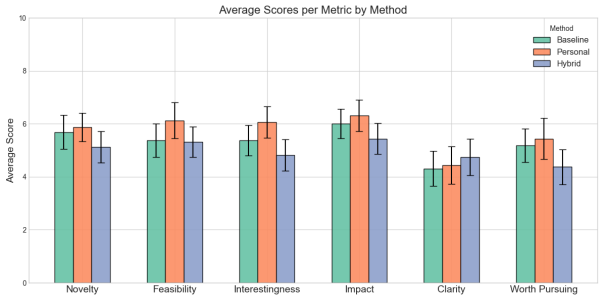


Figure 10: Bar chart showing average scores of the three experiment variants across all review metrics, SE plotted.

Responsible NLP Research Checklist

Members of the ACL are responsible for adhering to the ACL code of ethics. The ARR Responsible NLP Research checklist is designed to encourage best practices for responsible research, addressing issues of research ethics, societal impact and reproducibility.

Please read the Responsible NLP Research checklist guidelines for information on how to answer these questions. Note that not answering positively to a question is not grounds for rejection.

All supporting evidence can appear either in the main paper or the supplemental material. For each question, if you answer **Yes**, provide the section number; if you answer **No**, provide a justification.

Please do **not** modify, reorder, delete or add questions, question options or other wording of this document.

A For every submission

A1 Did you discuss the *limitations* of your work?

If you answer **Yes**, provide the section number; if you answer **No**, provide a justification.

Yes No N/A

Section or
Justification

A2 Did you discuss any potential *risks* of your work?

If you answer **Yes**, provide the section number; if you answer **No**, provide a justification.

Yes No N/A

Section or
Justification

A3 Do the abstract and introduction summarize the paper's main claims?

If you answer **Yes**, provide the section number; if you answer **No**, provide a justification.

Yes No N/A

Section or
Justification

B Did you use or create *scientific artifacts*?

If you answer **Yes**, answer the questions below; if you answer **No**, you can skip the rest of this section.

Yes No

If yes:

B1 Did you cite the creators of artifacts you used?

If you answer **Yes**, provide the section number; if you answer **No**, provide a justification.

Yes No N/A

Section or
Justification

B2 Did you discuss the *license or terms* for use and/or distribution of any artifacts?

If you answer **Yes**, provide the section number; if you answer **No**, provide a justification.

Yes No N/A

Section or
Justification

B3 Did you discuss if your use of existing artifact(s) was consistent with their *intended use*, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?

If you answer **Yes**, provide the section number; if you answer **No**, provide a justification.

Yes No N/A

Section or
Justification

B4 Did you discuss the steps taken to check whether the data that was collected/used contains any *information that names or uniquely identifies individual people* or *offensive content*, and the steps taken to protect / anonymize it?

If you answer **Yes**, provide the section number; if you answer **No**, provide a justification.

Yes No N/A

Section or
Justification

B5 Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?

If you answer **Yes**, provide the section number; if you answer **No**, provide a justification.

Yes No N/A

Section or
Justification

B6 Did you report relevant statistics like the number of examples, details of train/test/dev splits, etc. for the data that you used/created?

If you answer **Yes**, provide the section number; if you answer **No**, provide a justification.

Yes No N/A

Section or
Justification

C Did you run *computational experiments*?

If you answer **Yes**, answer the questions below; if you answer **No**, you can skip the rest of this section.

Yes No

If yes:

C1 Did you report the *number of parameters* in the models used, the *total computational budget* (e.g., GPU hours), and *computing infrastructure* used?

If you answer **Yes**, provide the section number; if you answer **No**, provide a justification.

Yes No N/A

Section or
Justification

C2 Did you discuss the experimental setup, including *hyperparameter search* and *best-found hyperparameter* values?

If you answer **Yes**, provide the section number; if you answer **No**, provide a justification.

Yes No N/A

Section or
Justification

C3 Did you report *descriptive statistics* about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?

If you answer **Yes**, provide the section number; if you answer **No**, provide a justification.

Yes No N/A

Section or
Justification

C4 If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?

If you answer **Yes**, provide the section number; if you answer **No**, provide a justification.

Yes No N/A

Section or
Justification

D Did you use *human annotators* (e.g., crowdworkers) or *research with human subjects*?

If you answer **Yes**, answer the questions below; if you answer **No**, you can skip the rest of this section.

Yes No

If yes:

D1 Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?

If you answer **Yes**, provide the section number; if you answer **No**, provide a justification.

Yes No N/A

Section or
Justification

D2 Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such *payment is adequate* given the participants' demographic (e.g., country of residence)?

If you answer **Yes**, provide the section number; if you answer **No**, provide a justification.

Yes No N/A

Section or
Justification

D3 Did you discuss whether and how *consent* was obtained from people whose data you're using/curating (e.g., did your instructions explain how the data would be used)?

If you answer **Yes**, provide the section number; if you answer **No**, provide a justification.

Yes No N/A

Section or
Justification

D4 Was the data collection protocol *approved (or determined exempt)* by an ethics review board?

If you answer **Yes**, provide the section number; if you answer **No**, provide a justification.

Yes No N/A

Section or
Justification

D5 Did you report the basic demographic and geographic characteristics of the *annotator* population that is the source of the data?

If you answer **Yes**, provide the section number; if you answer **No**, provide a justification.

Yes No N/A

Section or
Justification

E Did you use *AI assistants* (e.g., ChatGPT, Copilot) in your research, coding, or writing?

If you answer **Yes**, answer the question below; if you answer **No**, you can skip the rest of this section.

Yes No

E1 Did you include information about your use of AI assistants?

If you answer **Yes**, provide the section number; if you answer **No**, provide a justification.

Yes No N/A

Section or
Justification