An Interactive Co-Pilot for Accelerated Research Ideation

Harshit Nigam and **Manasi Patwardhan** and **Lovekesh Vig** and **Gautam Shroff**

TCS Research
{h.nigam, manasi.patwardhan, lovekesh.vig, gautam.shroff}@tcs.com

Abstract

In the realm of research support tools, there exists a notable void in resources tailored specifically for aiding researchers during the crucial ideation phase of the research life-cycle. We address this gap by introducing 'Acceleron', a 'Co-Pilot' for researchers, designed specifically to accelerate the ideation phase of the research life-cycle. Leveraging the reasoning and domain-specific skills of Large Language Models (LLMs) within an agent-based architecture with distinct personas, Acceleron aids researchers through the formulation of a comprehensive research proposals. It emulates the ideation process, engaging researchers in an interactive fashion to validate the novelty of the proposal and generate plausible set-of hypotheses. Notably, it addresses challenges inherent in LLMs, such as hallucinations, implements a two-stage aspect-based retrieval to manage precision-recall trade-offs, and tackles issues of unanswerability. Our observations and end-user evaluations illustrate the efficacy of Acceleron as an enhancer of researcher's productivity.

1 Introduction

With fast-paced research happening in every field, we are witnessing an exponential growth in the number of scientific articles and research papers on the web. It is difficult for an individual researcher or a small research team to keep abreast of the relevant advances amidst this information explosion. This has a downstream impact on the ability to be consistently appraised and ensure novelty of a proposed solution at various stages of the research life cycle. Thus there is an urgent need for a tools that can aid researchers to 1) understand, evaluate and incorporate the latest developments in the literature and 2) Formulate/Modify the current proposed solution accordingly to ensure novelty and impact.

Most of the existing tools focus on notifying and recommending researchers with relevant literature, facilitate exploration of existing literature and/or writing research manuscripts. Researchers have also proposed learning representations for retrieval of relevant scientific articles (Singh et al., 2022; Cohan et al., 2020; Ostendorff et al., 2022; Mysore et al., 2021), literature Review Generation (Hu and Wan, 2014; Kasanishi et al., 2023; Chen et al., 2021), Question Answering over scientific articles (Saikh et al., 2022; Dasigi et al., 2021; Lee et al., 2023), Scientific document summarization (Hayashi et al., 2020), citation recommendation (Ali et al., 2021, 2022; Medic and Snajder, 2023) citation intent detection (Cohan et al., 2019; Berrebbi et al., 2022; Roman et al., 2021; Lauscher et al., 2021), critical review and rebuttal generation (Ruggeri et al., 2022; D'Arcy et al., 2023; Kennard et al., 2021; Dycke et al., 2022; Wu et al., 2022), etc. However, to the best of our knowledge, no tool or no approach in the literature facilitates a researcher during the most arduous ideation stage of the research life-cycle. (Wang et al., 2024) attempts ideation in completely automated fashion. However, their results demonstrate ∼40% gap in the generation of ideas 'helpful' from the novelty

Figure 1: Acceleron Interface

Proceedings of the Third Workshop on Bridging Human–Computer Interaction and Natural Language Processing (3rd HCI+NLP Workshop), pages 60–73

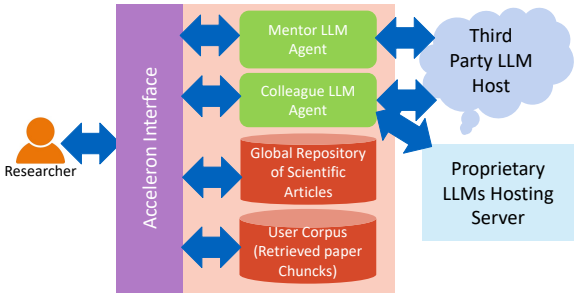

Figure 2: System Architecture

perspective.

Most of the tasks involved in research require domain expertise and complex reasoning skills. The recent advancement in Large Language Models (LLMs) has made it possible to partially automate some of these tasks (Liu and Shah, 2023; Liang et al., 2023; Zhang et al., 2023a; Lahiri et al., 2023; Kunnath et al., 2023). However complete automation of these tasks may not yield qualitative outcomes. In this work, we propose 'Acceleron' (Figure 1), a tool to accelerate the research life cycle. The tool exploits the reasoning and domain specific skills of LLM based agents not to replace researchers but to assist them for research ideation. With LLM powered *mentor* and *colleague* agents, Acceleron provides relevant inputs to researchers in an interactive fashion via a user-friendly interface. Thus, it aids the researcher to develop the research proposal consisting of a validated motivation, a well-defined research problem focusing of research gaps in the literature, a proposed approach selected from a set-of plausible synthesized methods and possible set-of experiments to be conducted to evaluate the approach for the research problem. To the best of our knowledge, we are the first ones to mimic the research ideation process using LLMs and execute it using human-machine interaction ensuring accelerated as well as qualitative outcomes, in terms of novel ideas.

2 System Architecture

Acceleron provides a web-based interface for researchers to interact. The system architecture is illustrated in Figure 2. We define an LLM Agent based architecture (Wang et al., 2023b), with agents of two distinct types of profiles or personas. A *Colleague* persona[1] performs less complex tasks including extraction of relevant information from user inputs, generation of relevant questions from extracted information or retrieval of relevant data from scientific documents. Whereas, *mentor* persona[2] performs more complex tasks requiring reasoning such as understanding the limitations or gaps of the existing work, identifying problems similar to the problem discussed in the proposal, identifying sub-tasks of the problem being solved in the proposal, solving similar problems and/or sub-tasks to synthesize a solution to the proposed problem and re-write the proposal given a plausible set-of approaches or possible limitations of related work. The architecture is flexible such that the LLM agents can interact with (i) LLMs like GPT-3.5-Turbo[3], Cohere[4] and Gemini[5] using API calls or (ii) open-source LLMs like Llama-2[6], Zephyr[7], Mixtral[8] which reside on an internal hosting server.

We expect to have a global repository which is a vector store of domain specific scientific articles[9] which are indexed by the Specter embeddings (Cohan et al., 2020) produced using the paper's title and abstract. We also have a User Specific corpus which has chunks of all the retrieved papers relevant to the current proposal the researcher is working on. The paper chunks are created with our in-house parser[10] treating paragraphs as semantic segments. If a paragraph does not fit into the the maximum token length of LLM agents, while chunking it is further split to fit into the maximum token length. The chunks are further converted to vector embeddings and indexed for efficient retrieval based on semantic similarity with a query. This user corpus acts as a shared 'memory' for the LLM agents.

3 Approach

The Acceleron Ideation simulation involves interaction between a researcher and the LLM agents, where the LLM agents perform actions based on the feedback received by the researcher or another agent. The process takes a proposal as an input from a researcher with a research problem description specified at a high level along with the motiva-

[1]OpenAPI's GPT-turbo-3.5 model

[2]OpenAPI's GPT4 model

[3]https://platform.openai.com/docs/models/gpt-3-5-turbo

[4]https://cohere.com/

[5]https://gemini.google.com/

[6]https://llama.meta.com/

[7]https://huggingface.co/HuggingFaceH4/zephyr-7b-beta

[8]https://mistral.ai/news/mixtral-of-experts/

[9]We use more than 2 million scientific articles in semantic scholar fetched using S2ORC dataset (Lo et al., 2020) as the global repository

[10]We built a PDF parser using PDFminer

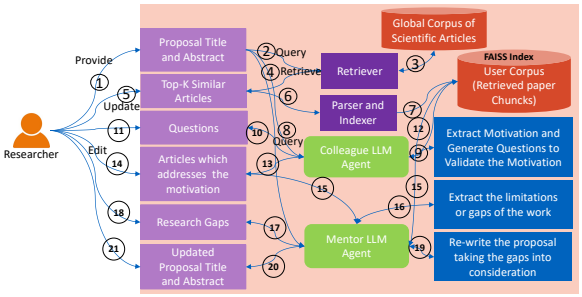

Figure 3: Motivation Validation Pipeline

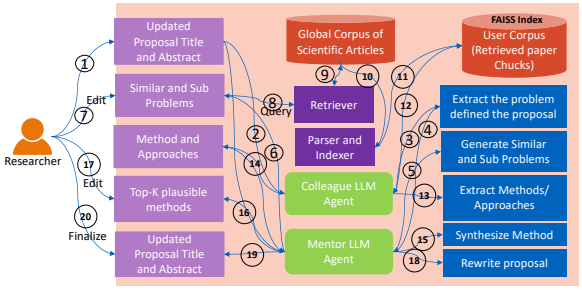

Figure 4: Method Synthesis Pipeline

tion behind the problem. The process involves: (i) Analyzing the existing literature to critically evaluate the motivation behind the research problem a researcher is trying to address to ensure that the mentioned research gap(s) still exist(s), (ii) Reformulating the proposed research problem and objectives based on the validation stage output and re-identification of research gaps, (iii) Identifying analogous research problems or sub-problems addressed in the literature and utilizing their solutions, available in the literature, to derive a set-of approaches or synthesizing a set-of plausible methods as a solution to the problem, (iv) Designing experimentation strategy for the given problem and selected methodology. The output of the ideation process is the updated proposal with a (i) A research problem with validated motivation (ii) Plausible methods to address the research problem. The overall ideation task is split into two pipelines: (i) Motivation Validation and (ii) Method Synthesis. The detailed prompts for the steps in each of the pipeline are illustrated in the Appendix Section A.2.

3.1 Motivation Validation Pipeline

As elaborated in Figure 3, the workflow begins with the researcher providing the title and abstract for their proposal. Acceleron identifies and extracts the motivation behind the proposal and retrieves relevant scientific articles relevant to the proposal and presents them to the researcher for review. The researcher can edit the selection of articles as needed. Subsequently, the system generates binary questions to validate the proposal's motivation against the retrieved articles. After review and potential edits by the researcher, the system retrieves relevant sections from the selected articles to answer these questions. If all articles fail to sufficiently address the proposal's motivation, the researcher is notified. Otherwise, identified gaps in the literature are presented to the researcher for consideration. The researcher can select relevant gaps or propose new ones, which are then used to refine the proposal's motivation and problem statement. The revised proposal is presented to the researcher for further editing or approval. This iterative process continues until the proposal's novelty is validated or until no relevant articles are found.

3.2 Method Synthesis Pipeline

The Method Synthesis workflow is illustrated in Figure 4. The method synthesis phase begins with the motivation validated proposal being accepted by the researcher. The system employs the *colleague* agent to extract and define the proposal's problem, followed by the *mentor* agent generating similar research problems and decomposing the main problem into sub-tasks. The researcher can refine these generated problems. Each refined problem is used to retrieve relevant scientific articles which is then parsed and stored in the user corpus. The *colleague* agent then consolidates similar problems and their solutions from these articles, presenting them to the researcher for further editing. This information, along with the original proposal, is provided to the *mentor* agent, which synthesizes a list of plausible methods to solve the problem. The researcher selects preferred methods, which are incorporated into the updated proposal by the *mentor* agent. The revised proposal is then reviewed and finalized by the researcher.

4 Novel Components

With Acceleron our aim is to bridge Human-Computer Interaction and Natural Language Processing using an interactive tool infused with the best of NLP and goodness of HCI. We created several novel components within Acceleron that fixes known shortcomings of NLP based systems using HCI inspired ideas.

4.1 LLM Agents for Research Ideation

To the best of our knowledge, ours is the first LLM agent based tool which assists in the complex task of ideation for research. We have devised with two novel portfolios for LLMS, viz., *colleague* and *mentor*, allocating less complex tasks to the *colleague* agent and more complex reasoning based tasks to the *mentor* agent. The user corpus acts as the shared memory for the agents, whereas the agents perform fixed set of actions at various stages of the workflow based on the provided inputs as discussed in the prior sections. Rather than using a costly LLM like GPT4 for all the tasks involved in the workflows; dividing the tasks as per the difficulty level and leveraging less costly LLM such as GPT-turbo-3.5 for colleague agent, performing less complex tasks, provides a cost-effective solution for workflows.

4.2 Mitigation of Hallucination

Hallucination is one of the major difficulties of using LLMs for knowledge based tasks (Zhang et al., 2023b; Wang et al., 2023a). We mitigate this problem using a two-fold solution: (i) There are retrieval augmented components of the workflows, viz. the motivation validation workflow poses questions generated to validate the motivation of the proposal on the retrieved articles stored in the user corpus or extract limitations of the articles which address the proposal motivation or the method synthesis workflow extracts approaches used to solve similar or sub problems from the retrieved articles. For these retrieval augmented tasks through proper prompt engineering, we ensure that the answers are provided by restricting the knowledge to the retrieved context only. We observe this helps to mitigate hallucinations. (ii) There are components of the workflows which rely on parametric knowledge of LLMs, for example the motivation validation involves re-writing the proposal and the method synthesis involves generating similar sub problems for the problem defined in the proposal and synthesizing methods. For these tasks the output can not be restricted to the provided input. In such cases, there is a higher chance of hallucinated outputs. For such scenarios, we ensure mitigation of hallucinated outputs, by keeping the system semi-automated and allowing user-interactions at every step to edit or delete hallucinated outputs. Moreover at every stage of the workflow, the LLM agents are asked to justify their outputs and the provided justification is exposed to the researcher through the interface. This forces the model to apply Chain-of-Thoughts (COT) (Wei et al., 2022) and allows the researcher to validate the output and check if it is in sync with the justification provided. This assists in alleviating the effect of hallucinations.

4.3 Two-Stage Aspect Based Retrieval

The global corpus contains a large number of scientific articles stored with the Specter embedding of the title and abstract of the papers. The title and abstract of the papers contains information about motivation and problem statement of the papers and a high level mention of the methodology and the results. For ideation we require more in-depth information from the papers across various aspects such as methodology, limitations, etc. To achieve this we perform retrieval in two stages. In motivation validation workflow, we first retrieve top-K papers from the global corpus with the proposal as the query and high value of K for good recall. This allows us to have a set-of papers with similar motivation and problem statement to that of the proposal. These papers are chunked and stored in the user corpus for further aspect based retrieval, such as papers with similar motivation to that of the proposal and paper paragraphs mentioning the research gaps of these papers. In method synthesis workflow, we first retrieve top-K papers from the global corpus with similar sub problem statements as the query and high value of K for good recall. This allows us to have a set-of papers with problems similar to the problem described in the proposal or similar to any of the sub-tasks of the problem described in the proposal. These papers are chunked and stored in the user corpus for further aspect based retrieval such as extracting the approaches of the papers. Note that keeping high-recall for the first stage of retrieval ensures coverage of papers, whereas for the second stage we favor more precise outcomes for aspect based retrieval.

4.4 Introduction of Unanswerability

The output of aspect based retrieval is always top-K paragraphs from the retrieved and chunked papers. We keep the value of K low to get more precise retrieval for the given aspect based query. However, there is a possibility that the retrieved paragraphs do not have the answer to the query (the query is unanswerable). For example, in the motivation validation workflow the retrieved paragraphs from the

papers do not answer the question of whether the paper addresses a specific motivation of the proposal and does not specify the limitations of the paper which would help to refine the problem defined in the proposal. Similarly, for the method synthesis workflow the retrieved paragraphs may not have an approach to solve a similar problem. In such cases, the LLM based agents check the relevancy of retrieved paragraphs for the given query and identifies the query as 'unanswerable' in case if all the retrieved paragraphs are irrelevant, avoiding irrelevant outputs. Allowing unasnwerability also assists in reduction of hallucinations.

4.5 Moderation of GenerativeAI

Output generated by API based closed-source Large Language Models like GPT-3.5 or Cohere, are always unmoderated relative to the domain they are being used in. Even though first party moderation in form of censorship and guardrails(Gehman et al., 2020)(Welbl et al., 2021) exist, these measures are focused on moderating offensive and inappropriate content being provided as input and generated as output by the LLM. Domain specific contextual moderation is necessary for a LLM to provide on-topic and context relevant outputs. An output generated as part of one domain may be irrelevant or inappropriate when taken out of context or when being provided as input to a LLM for a different task. To counter this issue we have specially designed our system using a novel expert-in-the-loop architecture where at each and every step where a LLM agent is called to generate an output, a context is created using our two-stage aspect based retrieval technique and task specific prompt provided by the human user themselves. This allows for the human to be in control of what the LLM is being fed as context for the output generation acting as a pseudo first layer of moderation. This in turn allows the LLM to generate domain relevant and topic appropriate output which is provided to the human for a review with option to edit if needed, so that the output can be used as context further down the pipeline, making a encapsulation of moderation on the LLM agents, negating the need for third party content moderation.

5 Qualitative Analysis of the Workflows

In the absence of an appropriate dataset for the tasks relevant to the ideation process, we evaluate our workflows by user-studies. We allow researchers working in distinct domains like computer science, material science and life science, to use Acceleron for ideation of their research problems. For computer science domain , we use Semantic Scholar data fetched using S2ORC dataset (Lo et al., 2020) as our global repository. Whereas, for material science and life science domain we use our repository of papers downloaded from 'Science Direct'[11] and 'PubMed'[12], respectively. We utilize the logging functionality of 'Acceleron' to keep track of the interactions between the researcher and the LLM Agents. For space constraints and data confidentiality preservation of unpublished work, here, we provide a qualitative analysis of the workflows with 2 proposals from distinct researchers, specifically in the domain of Artificial Intelligence (AI), Machine Learning (ML) and Natural Language Processing (NLP). The topics of these proposals are: (i) Topic-based citation retrieval for research proposal and (ii) Reference-free evaluation metric for retrieval augmented question answering.

We receive an input from a researcher with a proposal titled '*Topic-based citation retrieval for research proposal*' and the corresponding abstract '*Retrieval of research articles pertinent to a given query represents a thoroughly investigated research challenge. Typically, queries take the form of a title and abstract of a research article, or a specific sentence or paragraph from an existing research article requiring citation. However, existing approaches presuppose the availability of a well-constructed manuscript, an assumption that is inappropriate during the initial research proposal writing stage. At this initial phase, researchers seek pertinent literature for citing in their proposals, often focusing on specific topics or intents and further build the proposal. In this work, we aim to tackle the issue of topic-based citation retrieval for research proposals. We anticipate researchers providing the title and abstract of their research proposals, encompassing elements such as the research gap, problem statement, and a high-level overview of the proposed methodology and experiments. Additionally, researchers will furnish a list of topics for which relevant scientific articles need to be retrieved. Our proposed algorithm intends not only to fetch research articles pertinent to the given proposal from a corpus, but also to establish a crucial many-to-many mapping between these*

[11] https://www.sciencedirect.com/

[12] https://www.ncbi.nlm.nih.gov/pmc/tools/openftlist/

articles and the specified topics.' The colleague LLM agent generates the following questions for validation of the motivation: 1. "Is the research paper specifically addressing the retrieval of research articles relevant to a topic of a research proposal?" and 2. "Is the research paper developing a technique to map research articles to specified topics in research proposals?". Out of top-50 research articles used to validate the motivation of the proposal by posing the above mentioned questions, four (Appendix A.1.1) got retrieved to be answering as 'yes' to at the least one of the above questions, and thus invalidating the motivation behind the proposal. However, the justifications provided for these papers highlight that paper no. 1 and 3 introduce an approach for citation recommendations during the writing phase of the target manuscripts and not at the proposal writing stage. Also, scientific article 2 leverages contents of a target paper and citation graph to extract scientific information. The outcome of the scientific article 4 is a dataset which can be useful for the proposal, but does not address the task of 'topic-based citation retrieval for research proposal'. Thus, we observe that after evaluating the retrieved scientific articles claimed to be invalidating the proposal, the researcher disagrees with the justifications provided for each of the retrieved articles for addressing the motivation behind the proposal, hence validating the novelty of the proposal. This exemplifies the need as well as the effectiveness of this human computer interaction facility provided by the tool for the workflow. This example demonstrates acceleration of motivation validation stage of the research-life cycle ($\sim$8x for this proposal as stated by the researcher), by eliminating the need for the researcher to manually go through multiple relevant research articles retrieved by generic or academic search engines to ensure that the literature does not have a solution for the specific problem the researcher is trying to address, leading to a time consuming process.

We receive input from another researcher with the proposal titled '*Reference-Free evaluation metric for Retrieval augmented question answering task*' and the abstract '*We observe that questions with long answers on long documents do not have unique reference evidences (relevant paragraphs from the document) and answers. Rather, there is a distribution over reference answers, making expert based evaluation expensive and existing unique reference-based evaluation metrics inadequate. We also do not find any reference-free evaluation metric designed for evaluating retrieval augmented question answering task. Hence, this this work we propose to define this metric.*'. The colleague LLM agent generates the following question to validate the motivation of the proposal: "Is the research paper proposes a reference-free evaluation metric designed for evaluating retrieval augmented question answering tasks?". We observe that out of top-50 retrieved scientific articles relevant to the proposal, none of the articles provides answer as 'yes' to the question, leading to retrieval of zero relevant paper hence invalidating the motivation of the proposal. Manual analysis of the top-50 retrieved articles performed by the researcher (as well as other relevant articles manually visited by the researcher) to evaluate this outcome of the workflow, substantiates the results.

For the next workflow of method synthesis for the above proposal, the mentor LLM agent generates following set of research problems similar to the problem defined in the proposal: 1. "Evaluating complex tasks where there is no unique correct answer or reference", 2. "Designing evaluation metrics for tasks that involve retrieval and interpretation of large amounts of data", 3. "Creating reference-free evaluation metrics for tasks where reference-based metrics are inadequate or impractical", 4. "Assessing the quality of answers in tasks where the answers can be long and drawn from extensive documents". The mentor LLM agent also generates the following sub-tasks for the problem defined in the proposal: 1. "Defining a new metric that can effectively evaluate retrieval augmented question answering tasks" and 2. "Overcoming the inadequacy of existing unique reference-based evaluation metrics for questions with long answers on long documents". With these similar and sub-problems as queries, the colleague LLM agent first retrieves Top-10 similar scientific articles per statement (total 40 articles with some overlap as same article may be retrieved for multiple queries) and then poses the question that 'if the article provides a methodology or an approach to solve the above defined problem or sub-task'. The researcher receives total 17 scientific articles which answers 'yes' to the question along with a description of the methodology implemented for each of the above problems or task. Out of these papers, the researcher accepts 11 (Appendix A.1.2) scientific articles, finding them to be more relevant to the problem the researcher is trying to address.

Based on approaches discussed in these scientific

articles to solve the above enlisted similar problems and/or sub-tasks, the mentor LLM agent suggests the following 10 plausible methodologies for the proposal. Theses suggestions with the exact verbatim provided by the agent are as follows: 1. AVA (Automatic Evaluation Approach): This approach uses Transformer-based language models to encode the question, answer, and reference text. By measuring the similarity between the reference and an automatic answer, biased towards the question semantics, AVA can estimate the system accuracy. This approach can be used to evaluate retrieval augmented question answering tasks and overcome the inadequacy of existing unique reference-based evaluation metrics. 2. Informedness and Markedness Measures: These new probabilistic and information theoretic measures can be used to evaluate complex tasks where there is no unique correct answer or reference. They can also be used to design evaluation metrics for tasks that involve retrieval and interpretation of large amounts of data. 3. Large Language Models (LLMs): LLMs can be used as automated metrics for evaluating tasks that involve retrieval and interpretation of large amounts of data. They can also be used to assess the quality of answers in tasks where the answers can be long and drawn from extensive documents. 4. SacreROUGE: This open-source library dedicated to the development of summarization evaluation metrics can be used to develop reference-free evaluation metrics for tasks where reference-based metrics are inadequate or impractical. 5. Quantified Reproducibility Assessment (QRA): This method, based on concepts and definitions from metrology, can be used to develop reference-free evaluation metrics for tasks where reference-based metrics are inadequate or impractical. 6. RoSE Benchmark: This benchmark can be used to provide a more robust evaluation across different tasks. It can also be used to develop reference-free evaluation metrics for tasks where reference-based metrics are inadequate or impractical. 7. GENIE6: This human-in-the-loop leaderboard framework can be used to score answers in tasks where the answers can be long and drawn from extensive documents. 8. Sparse Attention and Contrastive Retriever Learning: This system conditions answer generation on identified documents and uses a dense retriever trained by scaling up a distantly supervised algorithm. It can be used to overcome the inadequacy of existing unique reference-based evaluation metrics for questions with long answers on long documents. 9. Unified Evaluation Benchmark for Long-form Answers: This approach involves conducting a thorough study of evaluation, including both human and automatic evaluation protocols. It can be used to overcome the inadequacy of existing unique reference-based evaluation metrics for questions with long answers on long documents. 10. Training an LFQA Evaluation Metric Directly on Human-Annotated Preference Judgments: This approach involves fine-tuning pre-trained Language Models based on human judgement scores for the task. This output showcases the quality of method recommendations provided by the tool for the given proposal. Though mentioned at high-level, the researcher agrees that most of these methods are well-suited as a plausible approach for the given proposal. Though there is a need for further work to finalize the most appropriate plausible method for proposal, the researcher finds this first cut of output provided by the tool to be relevant and the overall process to be significantly more efficient than the regular process followed by the researcher for constructing a plausible set-of approaches for the given problem, by searching through the relevant literature from scratch.

These examples illustrating the outcomes of the motivation validation and method synthesis phases of the ideation workflow of the tool, demonstrates the efficacy of the tool, in terms of providing relevant outputs at each stage of the workflow. The observations made in terms of time saved by the researchers with the tool usage for the respective tasks demonstrates the power of the tool with regards to time efficiency gains.

6 Conclusion

In this work, we have demonstrated a tool called 'Acceleron', developed to accelerate the ideation phase of the research life-cycle. To the best of our knowledge this is the first tool which addresses the tasks involved in the ideation stage. To emulate the ideation process, we use LLM agents with colleague and mentor personas to execute two workflows, viz. motivation validation and method synthesis, which engage researchers in an interactive fashion to develop the research proposal. Our workflow involves novel components to (i) alleviate the hallucinations of LLMs through user interaction, (ii) ensure relevant outcomes by two-stage aspect based retrieval, where first stage introduces higher recall reducing False Negatives and False Positives

are corrected by user interaction and second stage of more precise fine-grained aspect-based retrieval, (iii) introduction of unanswerability and (iv) Moderation of GenerativeAI via human interaction acting as a pseudo first layer of moderation increases user involvement in the final task specific outcome. The qualitative analysis performed with proposals from researchers in distinct domains, demonstrates qualitative outcomes for various stages in the workflow with ∼7.5x gains in the time efficiency for various stages of the ideation phase. Most importantly, expert-interaction avoids error propagation through the stages of workflows yielding qualitative outputs in terms of generation of novel and diverse ideas.

7 Future Works

This is an ongoing work. In future, we plan to emulate the domain specific aspects of the ideation process creating domain specific instances of the workflows. For example, there can be a specialized workflow for synthesis of alloys in material science domain or drug discovery or synthesis of clinical trials in life science domain. This would result into a meta-process for ideation, which is domain independent and instances of this meta-process customized for specific domains and / or tailor made for specific tasks.

The logging functionality of 'Acceleron' keeps track of every input provided to the researcher as well as LLM agents and every output from them along with the corresponding timestamps. We are saving these logs for each user interactions for all the sessions. We plan to use these logs with treating user validated inputs as ground truth annotations, to develop a datasets for the ideation process. The logs would be used for developing datasets for tasks such as: (i) retrieval of research papers with similar motivation (ii) proposal re-writing with addressing research-gaps (iii) retrieval of research papers with similar problems and/or (iv) method-synthesis from a set-of relevant papers. The datasets will be used to instruction-tune the Open-Source LMs, which can replace the existing LLMs yielding more cost-effective solutions.

We plan to extend the implementation of current phase to generate a list of experiments to be performed for the problem defined in the proposal and the methodology selected by the researcher. This would lead to generation of a (set-of) results table(s) in a semi-automated fashion, with baseline approaches, planned experiments (ablations) and appropriate metric(s) used for evaluation.

8 Limitations

The current version of ideation part of 'Acceleron' has certain limitations. The system generates descriptions for every generated question at every stage for the researcher to elaborate and explain of the outcomes of these stages. For example, if an existing paper is retrieved to be already addressing the motivation behind the proposal, the tool provides LLM generated description of the same to explain how the paper is already addressing the motivation. However, these descriptions sometimes are not sufficient for the researcher to evaluate if the retrieved outputs are correct, further hindering the process of updating the outputs. To counter this we are planning to extend this functionality by providing a facility to showcase the whole paper and highlight the chunk of context in the paper using which the description is generated. This would not only provide vital context to the researcher to understand the answer but also provide backtracking ability to check the context retrieved to generate the description for a particular question.

We typically observe that we do not get qualitative results for extracting limitations of user proposal as relevant retrieved papers do not specifically mention the limitations. In future we plan to enhance the reasoning capabilities of LLMs to extract limitations from a research paper. The open-source locally run LLMs like Llama-2 (Touvron et al., 2023) and Zephyr(Tunstall et al., 2023) are slow and produce less qualitative outcomes as compared to API based LLMs like GPT-3.5-Turbo and GPT4 driving up the cost of running the system. A single execution of the 2 workflows for a single proposal cost the researcher somewhere around $0.5 to $1 for GPT-3.5-Turbo depending on the inputs and context provided by the user and the number of papers retrieved for the proposal, whereas this cost is almost 10-fold for GPT4. To achieve better quality of explanations from the retrieved papers we plan to decontextualize the citations embedded in the retrieved papers by using an approach similar to (Newman et al., 2023). Moreover, we need a benchmark and metric to evaluate our idea generation pipeline. Right now, we are doing it by user-studies and expert feedback. However, we plan to use the newly released SciMon(Wang et al., 2024) Dataset to benchmark the ideation workflows and further enhance them.

References

Z. Ali, Guilin Qi, Pavlos Kefalas, Shah Khusro, Inayat Khan, and Khan Muhammad. 2022. Spr-smn: scientific paper recommendation employing specter with memory network. *Scientometrics*, 127:6763–6785.

Z. Ali, Guilin Qi, Khan Muhammad, Pavlos Kefalas, and Shah Khusro. 2021. Global citation recommendation employing generative adversarial network. *Expert Syst. Appl.*, 180:114888.

Dan Berrebbi, Nicolas Huynh, and Oana Balalau. 2022. Graphcite: Citation intent classification in scientific publications via graph embeddings. *Companion Proceedings of the Web Conference 2022*.

Xiuying Chen, Hind Alamro, Li Mingzhe, Shen Gao, Xiangliang Zhang, Dongyan Zhao, and Rui Yan. 2021. Capturing relations between scientific papers: An abstractive model for related work section generation. In *Annual Meeting of the Association for Computational Linguistics*.

Arman Cohan, Waleed Ammar, Madeleine van Zuylen, and Field Cady. 2019. Structural scaffolds for citation intent classification in scientific publications. *ArXiv*, abs/1904.01608.

Arman Cohan, Sergey Feldman, Iz Beltagy, Doug Downey, and Daniel S. Weld. 2020. Specter: Document-level representation learning using citation-informed transformers. *ArXiv*, abs/2004.07180.

Mike D'Arcy, Alexis Ross, Erin Bransom, Bailey Kuehl, Jonathan Bragg, Tom Hope, and Doug Downey. 2023. Aries: A corpus of scientific paper edits made in response to peer reviews. *arXiv preprint arXiv:2306.12587*.

Pradeep Dasigi, Kyle Lo, Iz Beltagy, Arman Cohan, Noah A Smith, and Matt Gardner. 2021. A dataset of information-seeking questions and answers anchored in research papers. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4599–4610.

Nils Dycke, Ilia Kuznetsov, and Iryna Gurevych. 2022. Nlpeer: A unified resource for the computational study of peer review. *ArXiv*, abs/2211.06651.

Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A. Smith. 2020. RealToxicityPrompts: Evaluating neural toxic degeneration in language models. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3356–3369, Online. Association for Computational Linguistics.

Hiroaki Hayashi, Wojciech Kryscinski, Bryan McCann, Nazneen Rajani, and Caiming Xiong. 2020. What's new? summarizing contributions in scientific literature. *ArXiv*, abs/2011.03161.

Yue Hu and Xiaojun Wan. 2014. Automatic generation of related work sections in scientific papers: An optimization approach. In *Conference on Empirical Methods in Natural Language Processing*.

Tetsu Kasanishi, Masaru Isonuma, Junichiro Mori, and Ichiro Sakata. 2023. Scireviewgen: A large-scale dataset for automatic literature review generation. In *Annual Meeting of the Association for Computational Linguistics*.

Neha Nayak Kennard, Timothy J. O'Gorman, Akshay Sharma, Chhandak Bagchi, Matthew Clinton, Pranay Kumar Yelugam, Rajarshi Das, Hamed Zamani, and Andrew McCallum. 2021. Disapere: A dataset for discourse structure in peer review discussions. *ArXiv*, abs/2110.08520.

Suchetha Nambanoor Kunnath, David Pride, and Petr Knoth. 2023. Prompting strategies for citation classification. *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*.

Avishek Lahiri, Debarshi Kumar Sanyal, and Imon Mukherjee. 2023. Citeprompt: Using prompts to identify citation intent in scientific papers. *2023 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*, pages 51–55.

Anne Lauscher, B. R. Ko, Bailey Kuehl, Sophie Johnson, David Jurgens, Arman Cohan, and Kyle Lo. 2021. Multicite: Modeling realistic citations requires moving beyond the single-sentence single-label setting. *ArXiv*, abs/2107.00414.

Yoonjoo Lee, Yoonjoo Lee, Kyungjae Lee, Sunghyun Park, Dasol Hwang, Jaehyeon Kim, Ho Hin Lee, and Moontae Lee. 2023. Qasa: Advanced question answering on scientific articles. In *International Conference on Machine Learning*.

Weixin Liang, Yuhui Zhang, Hancheng Cao, Binglu Wang, Daisy Ding, Xinyu Yang, Kailas Vodrahalli, Siyu He, Daniel 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.

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

Kyle Lo, Lucy Lu Wang, Mark Neumann, Rodney Kinney, and Daniel Weld. 2020. S2ORC: The semantic scholar open research corpus. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4969–4983, Online. Association for Computational Linguistics.

Zoran Medic and Jan Snajder. 2023. Paragraph-level citation recommendation based on topic sentences as queries. *ArXiv*, abs/2305.12190.

Sheshera Mysore, Arman Cohan, and Tom Hope. 2021. Multi-vector models with textual guidance for fine-grained scientific document similarity. *ArXiv*, abs/2111.08366.

Benjamin Newman, Luca Soldaini, Raymond Fok, Arman Cohan, and Kyle Lo. 2023. A question answering framework for decontextualizing user-facing snippets from scientific documents. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 3194–3212, Singapore. Association for Computational Linguistics.

Malte Ostendorff, Nils Rethmeier, Isabelle Augenstein, Bela Gipp, and Georg Rehm. 2022. Neighborhood contrastive learning for scientific document representations with citation embeddings. In *Conference on Empirical Methods in Natural Language Processing*.

Muhammad Roman, Abdul Shahid, Shafiullah Khan, Anis Koubâa, and Lisu Yu. 2021. Citation intent classification using word embedding. *IEEE Access*, 9:9982–9995.

Federico Ruggeri, Mohsen Mesgar, and Iryna Gurevych. 2022. Argscichat: A dataset for argumentative dialogues on scientific papers. *ArXiv*, abs/2202.06690.

Tanik Saikh, Tirthankar Ghosal, Amish Mittal, Asif Ekbal, and Pushpak Bhattacharyya. 2022. Scienceqa: a novel resource for question answering on scholarly articles. *International Journal on Digital Libraries*, 23:289 – 301.

Amanpreet Singh, Mike D'Arcy, Arman Cohan, Doug Downey, and Sergey Feldman. 2022. Scirepeval: A multi-format benchmark for scientific document representations. *ArXiv*, abs/2211.13308.

Hugo Touvron, Louis Martin, Kevin R. Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Daniel M. Bikel, Lukas Blecher, Cristian Cantón Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony S. Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel M. Kloumann, A. V. Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, R. Subramanian, Xia Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zhengxu Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open foundation and fine-tuned chat models. *ArXiv*, abs/2307.09288.

Lewis Tunstall, Edward Beeching, Nathan Lambert, Nazneen Rajani, Kashif Rasul, Younes Belkada, Shengyi Huang, Leandro von Werra, Clémentine Fourrier, Nathan Habib, Nathan Sarrazin, Omar Sanseviero, Alexander M. Rush, and Thomas Wolf. 2023. Zephyr: Direct distillation of lm alignment.

Cunxiang Wang, Xiaoze Liu, Yuanhao Yue, Xiangru Tang, Tianhang Zhang, Cheng Jiayang, Yunzhi Yao, Wenyang Gao, Xuming Hu, Zehan Qi, Yidong Wang, Linyi Yang, Jindong Wang, Xing Xie, Zheng Zhang, and Yue Zhang. 2023a. Survey on factuality in large language models: Knowledge, retrieval and domain-specificity. *ArXiv*, abs/2310.07521.

Lei Wang, Chengbang Ma, Xueyang Feng, Zeyu Zhang, Hao ran Yang, Jingsen Zhang, Zhi-Yang Chen, Jiakai Tang, Xu Chen, Yankai Lin, Wayne Xin Zhao, Zhewei Wei, and Ji rong Wen. 2023b. A survey on large language model based autonomous agents. *ArXiv*, abs/2308.11432.

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, Ed Huai hsin Chi, F. Xia, Quoc Le, and Denny Zhou. 2022. Chain of thought prompting elicits reasoning in large language models. *ArXiv*, abs/2201.11903.

Johannes Welbl, Amelia Glaese, Jonathan Uesato, Sumanth Dathathri, John Mellor, Lisa Anne Hendricks, Kirsty Anderson, Pushmeet Kohli, Ben Coppin, and Po-Sen Huang. 2021. Challenges in detoxifying language models. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2447–2469, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Po-Cheng Wu, An-Zi Yen, Hen-Hsen Huang, and Hsin-Hsi Chen. 2022. Incorporating peer reviews and rebuttal counter-arguments for meta-review generation. *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*.

Yang Zhang, Yufei Wang, Kai Wang, Quan Z. Sheng, Lina Yao, Adnan Mahmood, Wei Emma Zhang, and Rongying Zhao. 2023a. When large language models meet citation: A survey. *ArXiv*, abs/2309.09727.

Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, Longyue Wang, Anh Tuan Luu, Wei Bi, Freda Shi, and Shuming Shi. 2023b. Siren's song in the ai ocean: A survey on hallucination in large language models. *ArXiv*, abs/2309.01219.

A Appendix

A.1 Qualitative Analysis of the Workflow: Retrieved Papers

A.1.1 Papers Retrieved during Motivation Validation of Proposal 1

1. "Citation Recommendation: Approaches and Datasets"
2. "CitationIE: Leveraging the Citation Graph for Scientific Information Extraction"
3. "Content-Based Citation Recommendation"
4. "unarXive 2022: All arXiv Publications Pre-Processed for NLP, Including Structured Full-Text and Citation Network"

A.1.2 Papers retrieved during Method Synthesis Workflow of Proposal 2

1. "AVA: an Automatic eValuation Approach to Question Answering Systems"
2. "Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation"
3. "Re-visiting Automated Topic Model Evaluation with Large Language Models"
4. "SacreROUGE: An Open-Source Library for Using and Developing Summarization Evaluation Metrics"
5. "Quantified Reproductibility Assessment of NLP Results"
6. "Revisiting the Gold Standard: Grounding Summarization Evaluation with Robust Human Evaluation"
7. "A Critical Evaluation of Evaluations for Long-form Question Answering"
8. "Think you have Solved Direct-Answer Question Answering? Try ARC-DA, the Direct-Answer AI2 Reasoning Challenge"
9. "More Than Reading Comprehension: A Survey on Datasets and Metrics of Textual Question Answering"
10. "Hurdles to Progress in Long-form Question Answering"
11. "A Critical Evaluation of Evaluations for Long-form Question Answering"

A.2 Prompts for different stages of the Workflows

1. Motivation Extraction Prompt

System Message:
You are a researcher and trying to understand the following proposal written by another researcher:{proposal}

Human Message:
Describe in a bulleted list what is not addressed in the current literature which serves as the Motivation behind solving the above research problem proposed in the Proposal. Answer without a heading line and just the bullet points. Each bullet should mention one gap in the literature as a bullet point and not a sentence.

2. Motivation Question Generation Prompt

System Message:
You are a researcher and trying to understand the following proposal written by another researcher:{proposal}

Human Message:
Describe in a bulleted list what is not addressed in the current literature which serves as the Motivation behind solving the above research problem proposed in the Proposal. Answer without a heading line and just the bullet points. Each bullet should mention one gap in the literature as a bullet point and not a sentence.

AI Message:
{motivation}

Human Message: Convert each of the above bullets in to a binary question. The question should begin with 'Is the research paper'.

3. Ask Question for Motivation Validation Prompt

System Message:
You are a researcher. You have been given a context, which are paragraphs from a research paper. You have been given a question. Answer the given Question in 'Yes' OR 'No' OR 'Unanswerable'. Answer solely based on the provided context of the research paper. If the question can not be answered with the facts mentioned in the available context or there is any ambiguity in answering the question answer as 'Unanswerable'.
Answer as 'Yes' only when the question can be very clearly answered considering the facts in the research paper provided in the context. Do not repeat the question as the part of the answer.
Provide a concise explanation about how the answer to the question is 'Yes' mentioning the paragraphs used in the context to answer it as 'Yes'. If the answer is 'No' or 'Unanswerable' only output that with NO description or elaboration.

Human Message:
Question: {question}
Research Paper Context: {paper_chunks}

4. Extract Limitation Prompt

System Message:
You are a researcher. You have been given the following proposal: {proposal}

A different research paper provided in the context already addresses the gap mentioned as the motivation behind the proposal.
{descriptions}

Human Message:
Research Paper: {paper_chunks}

Identify the limitations or gaps of this research paper which can serve as the new motivation for the proposal. Provide a bulleted list of limitations, where each bullet is concise. Answer WITHOUT a heading line and just the bullet points.

5. Re-write Research Proposal Prompt

System Message:
You are a researcher and have written a proposal: {proposal}

Human Message:
Re-write the proposal by taking into consideration the mentioned gaps in the current literature as the new motivation behind of the problem defined in the proposal.
Answer in a Single detailed paragraph WITHOUT any bullet points or list.
Gaps in the current literature: {limitations}

6. Research Problem Extraction Prompt

System Message:
You are a researcher and trying to understand the following proposal written by another researcher: {proposal}

Human Message:
What is the problem solved in the proposal?

7. Similar Problem Generation Prompt

System Message:
You are a researcher and trying to understand the following proposal written by another researcher: {proposal}

Human Message:
What is the problem solved in the proposal?

AI Message:
{problem_statement}

Human Message:
Give me a bulleted list of a more generalised or similar problems to the problem defined in the proposal. Don't give a heading just the answer in a bulleted list.

8. Sub Problem Generation Prompt

System Message:
You are a researcher and trying to understand the following proposal written by another researcher: {proposal}

Human Message:
What is the problem solved in the proposal?

AI Message:
{problem_statement}

Human Message:
Provide a bulleted list of sub-problems or sub-tasks involved to solve the problem. Don't give a heading just the answer in a bulleted list.

9. Similar and Sub Problem Question Creation Prompt

Human Message:
{statement}
For the statement given above generate a question to be posed on a research paper to find out if the paper is proposing an approach or method to perform the task defined by the statement. Start the question with: 'Is the research paper proposing an approach or method to'.

10. Methodology Extraction Prompt

System Message:
You are a researcher and trying to answer the question posed on a research paper provided as the context.
Research Paper: {paper_chunks}

Human Message:
Answer the given Question in 'Yes' OR 'No' OR 'Unanswerable'. Answer solely based on the provided context of the research paper. If the question can not be answered with the facts mentioned in the available context or there is any ambiguity in answering the question, answer as 'Unanswerable'. Answer as 'Yes' only when the question can be very clearly answered considering the facts in the research paper provided in the context. Do not repeat the question as the part of the answer. If the answer to the question is 'Yes', provide detailed approach or methodology to perform the task. If the answer is 'No' or 'Unanswerable' only output that with NO description.

Question: {question}

11. Method Synthesis Prompt

System Message:
You are a researcher and have been given a proposal and the research problem the proposal is trying to solve. You have been given the approaches in the literature trying to solve, similar problems and sub problems or sub tasks of the problem defined in the proposal. Your task is to synthesize and propose a possible set of methods or approaches to solve the problem defined in the proposal.
Proposal: {proposal}
Research Problem in the Proposal: {problem}

Human Message:
{method_context}

Based on the above information suggest the top 3 possible methods or approaches to solve the problem defined in the proposal.