Published by :
http://www.ijert.org

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 14 Issue 05, May-2025

Evaluating the Effectiveness of Large Language Models in Conceptualizing Applied Research Proposals

A. Chakraborty
Engineering Officer, R&D Management
Central Power Research Institute

***Abstract*— This paper explores the effectiveness of Large Language Models (LLMs) in the conceptualization and composition of applied research proposals, particularly focusing on the domain of energy storage. By quantitatively assessing the capability of LLMs to generate critical sections of research proposals, such as background information, literature gaps, and research objectives, the study aims to understand their effectiveness and limitations in drafting research proposals. A total of 20 research proposals from Subject Matter Experts were selected, a section of the proposal was generated using GPT-4 and evaluated using metrics like BLEU, ROUGE, METEOR, and BERTScore to understand patterns in generative AI's performance. The results suggest that while LLMs offer significant advantages in generating semantically relevant content, their effectiveness varies across different proposal sections, with the Background and deliverable section showing relatively high semantic relevance despite lower exact word or phrase matches, and the Literature Gap section demonstrating varying levels of precision and quality. The study also highlights the potential of integrating LLMs with domain expertise to enhance the creativity and efficiency of research proposal development. Additionally, the paper outlines key considerations for R&D managers, emphasizing the importance of recognizing the acceptability of these proposals and suggesting critical areas for evaluation in content generated by LLMs. In its conclusion, this paper also identifies the potential areas for future research in this domain.**

***Keywords:* LLMs, Applied Research, Energy Storage, AI, Evaluation matrix**

I. INTRODUCTION

In the rapidly evolving world of artificial intelligence, large language models (LLMs) stand out as incredibly powerful and versatile tools, offering huge possibilities in many different areas. LLMs, such as GPT-4, InstructGPT and GPT 3.5, have demonstrated much proficiency in generating text that closely resembles human language, spanning a wide range of applications. [1,2,3,4].Their ability to streamline content creation has resulted in widespread utilization across domains like content writing, marketing, web page development, and crafting social media captions [2]. LLMs are also becoming more popular among individuals with vital information needs, such as students and patients, because of their exceptional ability to manage various applications, including general natural language tasks and tasks specific to certain domains [2].

II. LITERATURE REVIEW

Large language based applications like ChatGPT has already demonstrated a notable proficiency in coding [5]. However, in tasks related to common-sense planning, even in scenarios where humans excel, LLMs may not exhibit strong performance [6, 7]. As the adoption of these models continues to rise, it is anticipated that they may find application in the preparation of research proposals—a task traditionally perceived as time-consuming for researchers. While these models excel in various linguistic tasks, the unique requirement of research proposals, requiring in-depth literature survey, research gap identification, innovation and novelty, offer a distinct challenge to articulate technical details in an accurate and precise manner that necessitates careful examination. In few research works carried out to find the effectiveness of LLMs for academic research only qualitative approach has been taken to evaluate the performance [8]. To the best of our knowledge no work has been carried out to evaluate ability of LLMs for proposal conceptualization for applied research through a quantitative approach.

Through this investigation, the paper seeks to contribute insights into the strengths and limitations of LLMs in the process of applied research proposal conceptualization and composition. The findings aim to inform researchers, R&D managers, and decision-makers about the potential applications and considerations when integrating advanced language models into the critical task of drafting applied research proposals.

III. METHODOLOGY

To conduct a thorough assessment, a total of 20 applied research proposals were selected in the area of energy storage. These proposals were submitted by Subject Matter Experts in the area for consideration for funding support. One specific domain, Energy Storage, was selected for the study to keep the generation theme constant and to observe the repeatability across the content, their innovativeness and how effectively GPT-4 can distinguish the nuances involved. For the selection of the 20 applied research proposals in the field of energy storage, a stratified sampling technique was utilized to achieve broad coverage of topics in this area, encompassing technologies such as Sodium-ion batteries, Lithium-ion batteries, Metal Air Batteries, Flow batteries and Others. A representation of the mix of the technologies is indicated in Fig. 1:

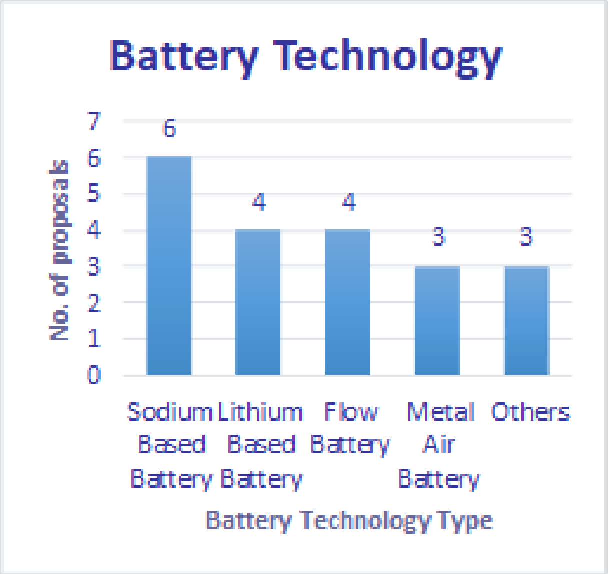

Fig 1: Distribution of Battery Technologies for the selected proposals

It should be noted that the selected proposals focus on emerging technological fields currently under research across various labs. Further the proposals span across TRL levels 2 to 5 with various levels as indicated below in Fig. 2:

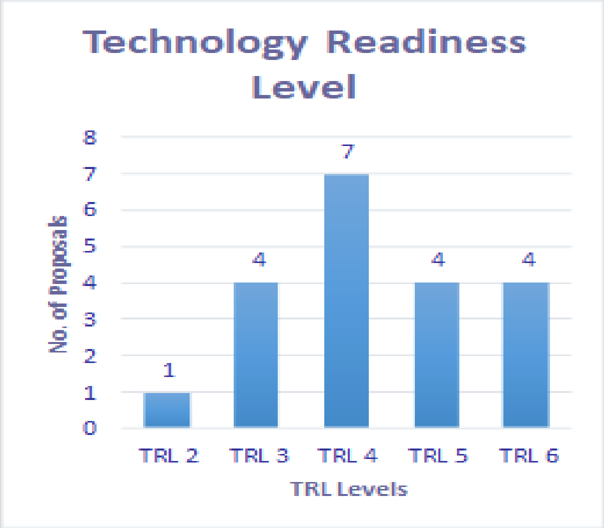

Fig. 2: TRL levels envisaged for the proposals under study

A. Brief Description of proposal sections

A typical proposal is structured into several key sections, including the Background, Literature Gap, Objectives,

Methodology, Project Cost, Justification of the Cost, Equipment and Facilities, Deliverables, References, and Sponsor Support. Among these, the Background, Literature Gap, and Objectives were specifically selected for closer examination. This selection was made because these sections are foundational to the proposal, providing a broad overview without delving deeply into the intricate details. They serve as the initial step in articulating the need, scope, and direction of the project. By focusing on these sections, it's possible to observe the shifts and developments in the proposal's thematic and conceptual framework, without delving deep in the more complex aspects of project execution, such as methodology specifics, budgeting, or technical requirements. Analysis of these sections allows examination of the proposal's core ideas and their coherence leading towards a more detailed analysis of the project's feasibility and strategic planning in subsequent sections.

B. Process of generation of proposal sections

The three key sections Background, Literature Gap and Objectives were extracted from the proposals. The said part of the proposal content was then generated using ChatGPT 4.0 model using prompt engineering. The prompts were crafted based on the title of the project. For Background generation, prompts were structured to include relevant context, key terms, and an overview of the field's current state, ensuring the AI produces a comprehensive backdrop of the subject area. For the Literature Gap, the prompts were designed to highlight specific areas where current research is lacking, prompting ChatGPT to identify and articulate these gaps based on the information fed into it. Finally, for the Objectives section, prompts focused on framing clear, achievable goals derived from the identified literature gaps, directing the AI to formulate objectives that are in tune with the current state of research. The parameters selected for the generation is provided in Table 1:

TABLE 1
PARAMETERS USED FOR GENERATING THE SECTIONS IN GPT-4

Parameter	Value
Temperature	0.7
Max Tokens	Equal to the length of the section in the original proposal.
Top P (Top-k sampling)	1.0
Frequency Penalty	0.5
Presence Penalty	0.5
Prompting	Based on the relevance and the topic with few filtration steps

C. Measuring the effectiveness of generated content

The effectiveness of the generative AI content was measured by comparing the output with the text from original proposals submitted by subject matter experts using metrics such as BLEU, ROUGE, METEOR, and BERTScore [9]. The rationalization of using the scores are briefed below:

1) BLEU Score (Bilingual Evaluation Understudy)

The BLEU score is a metric used to evaluate the quality of text which has been machine-translated from one language to another. However, its application has broadened to assess the quality of generated texts against a set of reference texts. BLEU evaluates the coherence and fluency of the generated text by measuring the overlap of n-grams (contiguous sequences of n items from a given sample of text) between the generated and reference texts. It scores from 0 to 1, where 1 indicates a perfect match with the reference. In the context of evaluating AI-generated research proposals, a high BLEU score would suggest that the generated text closely matches the expert-submitted proposals in terms of phrasing and specific terminology, indicating effective mimicry of expert writing styles and the proper use of domain-specific language.

2) ROUGE Score (Recall-Oriented Understudy for Gisting Evaluation)

The ROUGE score is primarily used to evaluate the quality of summaries by comparing an automatically generated summary to one or more reference summaries. It includes several measures such as ROUGE-N (evaluating the overlap of n-grams), ROUGE-L (evaluating the longest common subsequence), and ROUGE-S (evaluating the skip-bigram co-occurrence statistics). These measures focus on both precision (the portion of the generated text that is relevant) and recall (the portion of the reference that is captured by the generated text). In evaluating AI-generated sections of research proposals, the ROUGE score can assess how well the AI captures key points and concepts from the original expert proposals, indicating the model's ability to retain and reproduce critical information.

3) METEOR Score (Metric for Evaluation of Translation with Explicit ORdering)

The METEOR score is another metric for evaluating machine translation accuracy, which overcomes some of the limitations of BLEU by incorporating synonyms, stemming, and paraphrasing while aligning for meaning and fluency. It also introduces a penalty for overly long sentences to ensure conciseness. METEOR scores range from 0 to 1, with higher values indicating better quality translations. For assessing AI-generated research proposals, the METEOR score can provide insights into how well the AI understands and uses language nuances, including the use of synonyms and the generation of text that is both semantically and syntactically aligned with the expertise demonstrated in the original proposals.

4) BERTScore

BERTScore leverages the pre-trained contextual embeddings from models like BERT to compare the semantic similarity between the generated and reference texts. Unlike traditional metrics that rely on exact word matches, BERTScore evaluates the contextual relationship between words in the generated and reference texts, offering a more nuanced assessment of text quality. This makes BERTScore particularly relevant for evaluating research proposals generated by AI, as it can assess whether the AI-generated text captures the underlying concepts and ideas of the expert proposals, even if the exact wording differs. This metric is crucial for understanding the depth of conceptualization and idea representation in the generated text, which is essential for applied research proposals.

IV. RESULTS AND DISCUSSIONS

D. Analysis of the scores obtained

The average of the scores obtained for the evaluation metrics are presented below in Table 2:

TABLE 2
AVERAGE OF THE SCORES FOR COMPARING THE GEN-AI TEXTS

Section	BLEU score	ROUGE score	METEOR score	BERTscore
Background	0.01737	0.30619	0.21133	0.84437
Literature Gap	0.04171	0.29506	0.22918	0.83278
Deliverables	0.00374	0.289146	0.19571	0.84472

From the above it can be inferred that the Background section, despite recording the lowest BLEU score, achieved relatively high scores in other metrics, particularly the BERTscore. This suggests that, although there might be a lower incidence of exact word or phrase matches, the semantic relevance of the generated content to the reference materials remains high. Conversely, the Literature Gap section demonstrated an enhancement in both BLEU and METEOR scores compared to the Background, indicating a higher precision and quality in articulating the sections through GPT-4. Although the BERTscore of Literature Gap was marginally lower, it still indicates a good level of semantic similarity. The Deliverables section, like the Background, has the lowest BLEU scores, signifying minimal direct word or phrase correspondence with reference texts. However, its BERTscore was comparable to that of the Background, suggesting the relevance of its semantic content. This pattern across the sections underscores the nuanced capabilities of the model in maintaining semantic integrity, even when literal linguistic replication is low. This pattern shows that the GPT-4 model is good at maintaining semantic integrity across different sections, even if it doesn't always use the exact same words.

E. Generic inference from the generated text

This paper has also identified a few patterns in the generative AI's performance, identifying areas of strength and potential areas of improvement with the major takeaways for R&D mangers. The generic description provided in the background and initial literature gap identified are helpful as a starting point for the proposal. LLMs works well with proposals where more description is required and less quantitative targets are to be defined. As one generates similar proposals the content repeatability becomes obvious and hence the parameters are required to be changed to bring in more creativity. However, the downside may be generation of irrelevant content that might lead to hallucinations [10].

F. Observations for Researchers

Combining individual expertise with the generative capabilities of ChatGPT-4 presents a powerful synergy for writing research proposals. Experts can leverage their knowledge and insights in their specific fields to guide and refine the content generated by ChatGPT-4, ensuring that it aligns with the latest research trends, methodologies, and ethical standards. By initially setting detailed, informed prompts, researchers can direct ChatGPT-4 to produce drafts that captures necessary technical details, innovative ideas, and coherent narratives tailored to their project's goals. Following the initial generation, experts can then meticulously review and edit the output, integrating their nuanced understanding of the subject matter to enhance accuracy, add depth, and insert critical analysis where ChatGPT-4 might not fully grasp the intricacies or emerging developments in the field. This collaborative process allows for the creation of comprehensive, high-quality research proposals that benefit from the speed and breadth of AI-assisted writing without sacrificing the depth and precision that only human expertise can provide. Through this integration, researchers can streamline the proposal writing process, elevate the quality of their submissions, and better position their projects for approval and funding.

G. Takeaways for R&D mangers

In evaluating proposals generated with the assistance of ChatGPT-4, R&D managers must adopt a meticulous approach, emphasizing the importance of accuracy, innovation, and strategic alignment. They should start by scrutinizing the accuracy and time of release of the information, ensuring it is supported by quantitative data and up-to-date references. This involves a thorough review of the facts, figures, and statistics presented, comparing them against established databases and recent publications to confirm their validity. For proposals targeting high Technology Readiness Levels (TRLs), managers should specifically look for clearly defined, quantifiable objectives that demonstrate a measurable impact or outcome, aligning with the practical application and commercialization potential of the research.

Moreover, the proposer's previous work and its relevance to the current project must be carefully evaluated. R&D managers should assess how the proposer's past achievements and expertise are being leveraged to address the research questions at hand, ensuring a logical progression from past research to the proposed objectives. This includes examining the integration of previous findings into the new proposal, the continuity of research themes, and how past innovations are being built upon or expanded.

In addition to these qualitative reviews, managers must also be vigilant for any biases or ethical issues in the AI-generated content, due to the training data's limitations. By focusing on these comprehensive evaluation criteria, R&D managers can ensure that proposals not only benefit from the efficiency and broad capabilities of AI but also meet the highest standards of scientific rigor, innovation, and strategic relevance.

V. CONCLUSION & FUTURE RESEARCH DIRECTION

From this study it can be inferred that Large Language Models (LLMs) like ChatGPT-4 bring substantial benefits to the drafting of research proposals, notably in enhancing efficiency and creativity. By leveraging a wide array of perspectives and immediate access to a vast compendium of knowledge, LLMs stand out as transformative tools in the research proposal process. Furthermore, when researchers synergize the generative power of LLMs with their specialized domain knowledge, they can craft proposals that are not only innovative and novel but also thoroughly aligned with contemporary scientific benchmarks and quantifiably measured outcomes.

Looking ahead, future research directions present exciting prospects for further integrating LLMs into the proposal drafting process. A pivotal area for exploration could involve the standardization of content generation through ChatGPT-4. By establishing specific criteria and conditions within prompts, researchers can harness more tailored and relevant outputs. Such standardization might include predefined structures for proposals or checklists of essential items to be covered, which could streamline the drafting process and ensure consistency across different proposals.

Moreover, enriching LLMs with reference materials could offer another avenue for advancement. This would enable the models to not only access and summarize pertinent information but also infer new insights from the digested content, thereby elevating the depth and relevance of the generated proposals. Experimenting with various parameters of the LLMs—such as adjusting temperature and token levels—could recommend for optimal settings for different types of research proposals across various TRLs.

To aid R&D managers in their evaluation process, the development of specific LLM filters or criteria for assessing the relevance/ percentage of AI mix is required especially for Subject matter assisted generation. Also, authenticity of LLM-generated content is essential. These tools could help in distinguishing the most promising proposals and ensuring that the proposal meet the high standards expected in the scientific community.

In summary, while LLMs already represent a significant leap forward in drafting research proposals, the path forward includes refining these models through standardization, enhanced input material integration, parameter optimization, and the creation of evaluative tools. These advancements promise to further unlock the potential of LLMs in producing research proposals that are not only innovative and efficient but also rigorously aligned with the highest standards of scientific inquiry.

VI. ACKNOWLEDGMENT

The author would like to acknowledge the management of CPRI for their constant support and encouragement for coming up with this paper.

REFERENCES

[1] Mark Chen, et al. 2021. Evaluating large language models trained on code. arXiv preprint arXiv:2107.03374 (2021)

[2] Enkelejda Kasneci, et al. 2023. ChatGPT for good? On opportunities and challenges of large language models for education. Learning and Individual Differences 103 (2023), 102274.

[3] Meyer, Jesse G., et al. "ChatGPT and large language models in academia: opportunities and challenges." BioData Mining 16, no. 1 (2023): 20.

[4] Rahman, et al. "ChatGPT and academic research: a review and recommendations based on practical examples." Rahman, M., Terano, HJR, Rahman, N., Salamzadeh, A., Rahaman, S.(2023). ChatGPT and Academic Research: A Review and Recommendations Based on Practical Examples. Journal of Education, Management and Development Studies 3, no. 1 (2023): 1-12.

[5] OpenAI. 2023. GPT-4 Technical Report. arXiv:2303.08774 [cs.CL]

[6] Yan Zhuang, et al. 2023. Efficiently Measuring the Cognitive Ability of LLMs: An Adaptive Testing Perspective. arXiv preprint arXiv:2306.10512 (2023).

[7] Karthik Valmeekam et al. On the Planning Abilities of Large Language Models–A Critical Investigation. arXiv preprint arXiv:2305.15771 (2023).

[8] AlZaabi, et al. "ChatGPT applications in academic research: A review of benefits, concerns, and recommendations." bioRxiv (2023): 2023-08.

[9] Saadany, Hadeel, and Constantin Orasan. "BLEU, METEOR, BERTScore: evaluation of metrics performance in assessing critical translation errors in sentiment-oriented text." arXiv preprint arXiv:2109.14250 (2021).

[10] Alkaissi, Hussam, and Samy I. McFarlane. "Artificial hallucinations in ChatGPT: implications in scientific writing." Cureus 15, no. 2 (2023).