Leveraging Artificial Intelligence Large Language Models for Writing Clinical Trial Proposals in Dermatology

Megan Hauptman, Daniel Copley, Kelly Young, Tran Do, Joseph Durgin, Albert Yang, Jungsoo Chang, Allison Billi, Mio Nakamura, Trilokraj Tejasvi

Submitted to: JMIR Dermatology
on: August 28, 2025

Table of Contents

Leveraging Artificial Intelligence Large Language Models for Writing Clinical Trial Proposals in Dermatology

Megan Hauptman[1] MD; Daniel Copley[2] BSc; Kelly Young[1] MD, PhD; Tran Do[1] MD, PhD; Joseph Durgin[1]; Albert Yang[1] MD; Jungsoo Chang[1] MD; Allison Billi[1] MD, PhD; Mio Nakamura[1] MD; Trilokraj Tejasvi[1] MD

[1]Department of Dermatology University of Michigan Ann Arbor US
[2] Anduril Costa Mesa US

Corresponding Author:
Trilokraj Tejasvi MD
Department of Dermatology
University of Michigan
1500 E Medical Center Dr
Ann Arbor
US

Abstract

Background: Large language models (LLMs) are becoming increasingly popular in clinical trial design but have been underutilized in research proposal development.

Objective: This study compares the performance of commonly used open access LLMs versus human proposal composition and review.

Methods: Ten LLMs were prompted to write a research proposal. Six physicians and each of the LLMs assessed 11 blinded proposals for capabilities and limitations in accuracy and comprehensiveness.

Results: Chat GPT o1 was rated the most accurate and LLaMA 3.1 the least accurate by human scorers. LLM scorers rated Chat GPT o1 and Deepseek R1 the most accurate. Chat GPT o1 was the most comprehensive and LLaMA 3.1 the least comprehensive by human and LLM scorers. LLMs performed poorly on scoring proposals, and on average rated proposals 1.9 points higher than humans for both accuracy and comprehensiveness.

Conclusions: Paid versions of ChatGPT remain the highest quality and most versatile option of available LLMs. These tools cannot replace expert input but serve as powerful assistants, streamlining the development process and enhancing productivity. Clinical Trial: n/a

(JMIR Preprints 28/08/2025:76674)
DOI: https://doi.org/10.2196/preprints.76674

Original Manuscript

Article Type: Original Article

Title: Leveraging Artificial Intelligence Large Language Models for Writing Clinical Trial Proposals in Dermatology

Megan Hauptman, MD[1] (ORCID: 0009-0009-2598-293X), Daniel Copley, BS[2] (ORCID: 0009-0003-9149-0518), Kelly Young, MD, PhD[1] (ORCID: 0000-0002-6714-7563), Tran Do, MD, PhD[1] (ORCID: 0000-0001-6317-9202), Joseph S. Durgin, MD[1](ORCID: 0000-0002-1114-817X), Albert Yang, MD[1] (ORCID: 0000-0002-0582-897X.), Jungsoo Chang, MD[1] (ORCID: 0000-0001-5901-6929), Allison Billi, MD, PhD[1] (ORCID: 0000-0001-7115-9113), Mio Nakamura, MD, MS[1] (ORCID: 0000-0002-1367-742X), Trilokraj Tejasvi, MD[1] (ORCID: 0000-0002-8186-871)

[1]Department of Dermatology, University of Michigan, Ann Arbor, MI, USA

[2]Anduril Industries, Costa Mesa, CA, USA

Corresponding author:

Trilokraj Tejasvi, MD
1500 E Medical Center Dr
Ann Arbor, MI 48105
Email: ttejasvi@med.umich.edu

Abstract:

Background: Large language models (LLMs) are becoming increasingly popular in clinical trial design but have been underutilized in research proposal development.

Objective: This study compares the performance of commonly used open access LLMs versus human proposal composition and review.

Methods: Ten LLMs were prompted to write a research proposal. Six physicians and each of the LLMs assessed 11 blinded proposals for capabilities and limitations in accuracy and comprehensiveness.

Results: Chat GPT o1 was rated the most accurate and LLaMA 3.1 the least accurate by human scorers. LLM scorers rated Chat GPT o1 and Deepseek R1 the most accurate. Chat GPT o1 was the most comprehensive and LLaMA 3.1 the least comprehensive by human and LLM scorers. LLMs performed poorly on scoring proposals, and on average rated proposals 1.9 points higher than humans for both accuracy and comprehensiveness.

Conclusions: Paid versions of ChatGPT remain the highest quality and most versatile option of available LLMs. These tools cannot replace expert input but serve as powerful assistants, streamlining the development process and enhancing productivity.

Keywords: artificial intelligence; large language model; research proposal; clinical research; clinical trials; deep learning; machine learning; research design

Main Text

Introduction

Advancements in artificial intelligence (AI) have led to development of large language models (LLMs) utilizing algorithms that learn from data and recognize patterns to make decisions based on all available data within a training set (1). However, AI is limited by the data it is trained on and an inability to account for nuanced contexts of individual research studies (2). Researchers are increasingly employing LLMs in clinical trial design to improve patient selection, cohort composition, and recruitment (3). In contrast, use of LLMs in research proposal development is largely unexplored and thus perhaps underutilized. This study aims to address this gap by comparing the performance of LLMs versus the current gold standard of human proposal composition and review. Our goals were threefold: to rate LLM in composing clinical trial proposals, to assess LLMs in scoring clinical trial proposals, and to evaluate the ease of using LLMs (including usability and efficiency).

Methods

Commonly used open access AI platforms (DeepSeek R1, ChatGPT o3-mini, ChatGPT o1, ChatGPT 4o, Claude Sonnet, Claude Opus, OpenEvidence, Grok 2, Gemini Advanced, and LLaMA 3.1) were evaluated for use in research proposal drafting. We requested each of the models to “Write a research proposal for a study looking at the use of narrowband-ultraviolet B phototherapy for psoriasis treatment for psoriasis patients of varying skin pigmentation with 3 aims: 1. To understand the factors that affect the response of NB-UVB in psoriasis patients of varying skin pigmentation. 2. Evaluate adverse effects of NB-UVB and their impact on psoriasis patients of varying skin pigmentation. 3. Compare the acute immunologic response to NB-UVB in psoriasis patients of varying skin pigmentation using bulk and single-cell RNA Sequencing. Include the following sections: 1 page “Specific Aims” with details on each of the 3 aims, 1/2 page background and significance of the topic, 1 page of “preliminary data/studies” relevant to the study, 1 page “experimental design” (include summary of study, inclusion and exclusion criteria, study visits and procedures with an associated table describing specifics of study visits), 1/2 page of “statistical methods, power calculations and bioinformatic analyses” specific for each aim, 1/4 page of “potential problems and alternative strategies”. Please have approximately 30 references from reputable sources. Make the proposal a total of 7 pages long in paragraph form, in formal scientific language and at a graduate level.” Six human scorers (physicians with strong research backgrounds) and each of the LLMs assessed 11 blinded proposals (10 LLMs and 1 human proposal) on a scale from 1 to 5 (1- strongly disagree, 5- strongly agree) for capabilities and limitations in the LLM’s accuracy and comprehensiveness (Table 1). LLM usability and efficiency, including pros and cons, were assessed by two investigators without computer science backgrounds.

Results

LLMs composing proposals

The human-written proposal scored 5 in accuracy and comprehensiveness across all human scorers and remains the gold standard (Table 2). Human scorers rated Chat GPT o1 to be the most accurate and LLaMA 3.1 to be the least accurate. When assessed in scoring LLM-derived clinical trial proposals, LLM scorers rated Chat GPT o1 and Deepseek R1 the most accurate. Chat GPT o1 was found to be the most comprehensive and LLaMA 3.1 the least comprehensive by both human and LLM scorers.

LLMs scoring proposals

Overall, LLMs performed poorly on scoring proposals, and on average rated proposals 1.9 points higher than humans for both accuracy (range 1.3-2.8) and comprehensiveness (0.7-3). The Claude Sonnet proposal showed the largest discrepancy between human and LLM scoring, with an average difference of 2.8 points for accuracy and 3 points for comprehensiveness. Interestingly, ChatGPT o1 and Deepseek proposals both received top scores of 5 for both accuracy and comprehensiveness from all LLMs versus human averages of 4.3 (variance 2.2) and 3.3 (1.9), respectively. The absence of variance at the top of the range (and wide variance in the middle of the range) suggests that the discriminatory power of the LLMs plateaued out at the top LLM quality.

Ease of using LLMs

All open access LLMs are highly efficient and run in a matter of seconds to minutes (minimum 20 seconds (Llama 3.1), maximum 1 minute and 37 seconds (ChatGPT 4o)). When assessed for ease of using, all available LLMs offer an intuitive interface and are highly usable for researchers (DC an MH) without computer science backgrounds.

Discussion

LLMs offer powerful tools to assist humans in clinical trial proposal creation. Whereas LLMs take only minutes to generate proposals, prior investigations into time commitment for generation of proposals by humans have reported estimates of 116 principal investigator hours, 55 co-investigator hours, and 38 working days (4, 5). Judicious use of LLMs in proposal development therefore allows the researcher to save significant time in organizing sections, formatting, and ensuring coherence.

The pros and cons overall and for each LLM are summarized in Table 3. All open-access LLMs can aid in initial outlining and creation of research proposals. They can assist in initial brainstorming of a clear researchable question and generating hypotheses based on exiting literature. LLMs are useful in literature review and can summarize existing studies related to the proposal topic and identify gaps in current knowledge. Furthermore, all open-access LLMs can propose data collection methods, define eligibility criteria based on study objection, recommend appropriate statistical tests based on study design, and help draft proposal sections. They also allow iterative refinements, enabling tailored outputs to meet specific requirements or needs. While human verification is always required, LLMs can greatly improve time spent with initial proposal drafting, and aid in mundane tasks associated with proposal writing, including proofreading and revisions, writing administrative sections, and optimizing citations.

Limitations to consider

All LLMs operate similarly to traditional autocomplete and work by using available contextual clues and a statistical model to predict the most likely next “token” or word. Due to training data cut-off of AI models, researchers must manually incorporate the latest literature findings. AI researchers are working on incorporating more access to real-time data, for example GPT actions (5), but these solutions come with their own tradeoffs. Another limitation is that users must verify citations, as the model may “hallucinate” or fabricate realistic sounding but false information. Lastly, although AI models such as DALL·E (or others) can create images, they are less effective at producing accurate, clinical grade figures.

Additionally, current LLMs were largely unable to score proposals and should not replace human review for a quality control. The high scores by the LLM raters indicate that the LLMs were unable to detect entire missed protocol sections. Other than Gemini Advanced (self-scored Gemini-advanced written proposal 3s for accuracy and comprehensiveness), Claude sonnet, and Llama 3.1, all the LLMs self-scored their own proposal 5s for both accuracy and comprehensiveness, suggesting overlapping “blind spots” in LLM proposal generation and evaluation.

One limitation of the study is that the order the proposals were sent for respondents to review was not randomized. Additionally, the “gold standard” (human proposal) was last, and there likely plays a role of question order, with kinder grading of the LLM-derived proposals prior to reviewing the human-written proposal. Had the human proposal been first, it would have highlights missing components of LLM-derived proposals and likely led to harsher human grading of the LLM proposals.

Conclusion

The future of AI in clinical research is expected to be transformative and far reaching. As AI algorithms continue to evolve, they are likely to become more accurate, comprehensive, efficient and interpretable, enabling researchers to leverage AI-driven insights for personalized medicine, disease prevention, and improved patient outcomes. In the coming years, AI is anticipated to play a crucial role in optimizing clinical trial design and accelerating drug discovery (6). The integration of AI with other emerging technologies, such as blockchain and the Internet of Medical Things (IoMT), could further revolutionize clinical research by improving data security, privacy, and real-time patient monitoring (7). As these advancements continue to unfold, AI has the potential to democratize access to novel therapies, reduce healthcare costs, and ultimately usher in an era of precision medicine (8).

LLMs offer a transformative approach to drafting research proposals (9). Paid versions of ChatGPT (ChatGPT-o3 mini and ChatGPT-o1) currently remains the highest quality (as determined by the Artificial Analysis Quality Index) and most versatile option of available LLMs, balancing usability, speed, accuracy, and customization (10). While these tools cannot entirely replace expert input, they serve as powerful assistants, streamlining the development process and enhancing productivity. For optimal results, researchers should combine AI-generated content with their expertise, ensuring precision and adherence to the latest research standards.

Funding: The authors received no specific funding for this work.

Conflicts of Interest: The authors have no relevant conflicts of interest to disclose.

Data Availability: The data used in this manuscript can be made available upon request to the corresponding author.

Authors Contributions: Study conception and design were completed by Megan Hauptman and Trilokraj Tejasvi. Material preparation and data collection were performed by Megan Hauptman, Daniel Copley, Kelly Young, Tran Do, Joseph S. Durgin, Albert Yang, Jungsoo Chang, Allison Billi, and Mio Nakamura. Data analysis was performed by Megan Hauptman. The first draft of the manuscript was written by Megan Hauptman and Daniel Copley and all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

Abbreviations:

AI: Artificial Intelligence

LLMs: Large Language Models

References

1. Fliorent R, Fardman B, Podwojniak A, Javaid K, Tan IJ, Ghani H, Truong TM, Rao B, Heath C. Artificial intelligence in dermatology: advancements and challenges in skin of color. Int J Dermatol. 2024 Apr;63(4):455-461.
2. Askin S, Burkhalter D, Calado G, El Dakrouni S. Artificial Intelligence Applied to clinical trials: opportunities and challenges. Health Technol (Berl). 2023;13(2):203-213.
3. Harrer S, Shah P, Antony B, Hu J. Artificial Intelligence for Clinical Trial Design. Trends Pharmacol Sci. 2019 Aug;40(8):577-591.
4. von Hippel T, von Hippel C. To apply or not to apply: a survey analysis of grant writing costs and benefits. PLoS One. 2015 Mar 4;10(3):e0118494.
5. Herbert DL, Barnett AG, Clarke P, Graves N. On the time spent preparing grant proposals: an observational study of Australian researchers. BMJ Open. 2013 May 28;3(5):e002800.
6. OpenAI Platform. GPT Actions: Customize ChatGPT with GPT Actions and API integration. https://platform.openai.com/docs/actions/introduction [Accessed 07 Jan 2025].

7. Bajwa J, Munir U, Nori A, Williams B. Artificial intelligence in healthcare: transforming the practice of medicine. Future Healthc J. 2021 Jul;8(2):e188-e194.
8. ICON. AI: The Future of Clinical Research. AI: The Future of Clinical Research [Accessed 19 Mar 2025]
9. Markey N, El-Mansouri I, Rensonnet G, van Langen C, Meier C. From RAGs to riches: Utilizing large language models to write documents for clinical trials. *Clinical Trials*. 2025;0(0).
10. Artificial Analysis. LLM Leaderboard- Comparison of GPT-4o, Llama 3, Mistral, Gemini and over 30 models [Accessed 15 Jan 2025].

Table 1: Criteria of assessing accuracy, usability, comprehensiveness and efficiency of LLMs

Table 2. Human and LLM scoring LLMs on accuracy and comprehensiveness

Table 3: Pros and Cons of open access LLMs

Table 1: Criteria of assessing accuracy, usability, comprehensiveness and efficiency of LLMs

Accuracy	All output was fact checked by rater and is correct. All references cited are verifiable and from reputable sources.
Comprehensiveness	The study design is of high quality and includes comprehensive specific aims, background and significance, preliminary data/studies, experimental design with inclusion/exclusion criteria and study visits and procedures, statistical methods, power calculations and bioinformatic analyses, and potential problems/alternative strategies. The study is approximately 7 pages and has approximately 30 references from reputable sources.
Usability	Judged based on intuitiveness and ease of use for researchers (MH) without a computer science background.

Efficiency	Measured as the time taken from providing input to obtaining the final output, with minimal delays or inefficiencies.

Table 2. Human and LLM scoring LLMs on accuracy and comprehensiveness

	Human Scoring							AI Scoring										
	1	**2**	**3**	**4**	**5**	**6**	**Average (Var)**	**Deepseek R1**	**ChatGPT o3-mini**	**ChatGPT o1**	**ChatGPT 4o**	**Claude Sonnet**	**Claude Opus**	**Grok 2**	**Gemini Advanced**	**Llama 3.1**	**Average (Var)**	**Overall average**
Accuracy																		
Human Proposal (Control)	5	5	5	5	5	5	**5 (0)**	5	5	5	5	5	5	5	4	5	**4.9 (0.1)**	**4.9**
ChatGPT o1	1	4	2	5	5	4	**3.5 (2.7)**	5	5	5	5	5	5	5	5	5	**5 (0)**	**4.4**
Deepseek R1	2	4	1	4	5	3	**3.2 (2.2)**	5	5	5	5	5	5	5	5	5	**5 (0)**	**4.3**
Claude Opus	3	4	1	3	4	5	**3.3 (1.9)**	5	5	5	5	5	5	5	4	5	**4.9 (0.1)**	**4.3**
ChatGPT o3-mini	1	4	1	4	5	2	**2.8 (3)**	5	5	5	5	5	5	5	4	5	**4.9 (0.1)**	**4.1**
Grok 2	2	4	1	3	5	3	**3.2 (2)**	5	5	5	5	5	4	5	4	5	**4.8 (0.2)**	**4.1**
Claude Sonnet	1	1	1	4	3	2	**2 (1.6)**	5	5	5	5	4	4	5	5	5	**4.8 (0.2)**	**3.7**
OpenEvidence	1	1	5	2	3	2	**2.3 (2.3)**	5	5	4	5	4	4	5	5	5	**4.7 (0.3)**	**3.7**
Gemini Advanced	2	4	1	3	3	2	**2.5 (1.1)**	5	4	5	5	3	4	5	3	5	**4.3 (0.8)**	**3.6**
ChatGPT 4o	1	2	1	4	3	2	**2.2 (1.4)**	5	5	5	5	3	3	4	4	5	**4.3 (0.8)**	**3.5**
Llama 3.1	1	3	1	3	1	1	**1.7 (1.1)**	3	2	4	5	2	2	3	2	4	**3 (1.3)**	**2.5**
Comprehensiveness																		
Human Proposal (Control)	5	5	5	5	5	5	**5 (0)**	5	4	5	5	5	5	5	4	5	**4.8 (0.2)**	**4.9**
ChatGPT o1	4	4	4	5	5	4	**4.3 (0.3)**	5	5	5	5	5	5	5	5	5	**5 (0)**	**4.7**
ChatGPT o3-mini	4	4	3	4	5	4	**4 (0.4)**	5	5	5	5	5	5	5	4	5	**4.9 (0.1)**	**4.5**
Deepseek R1	4	4	1	3	5	3	**3.3 (1.9)**	5	5	5	5	5	5	5	5	5	**5 (0)**	**4.3**
Claude Opus	4	3	1	3	1	4	**2.7 (1.9)**	5	5	5	5	5	5	5	4	5	**4.9 (0.1)**	**4**
Grok 2	3	3	4	3	3	2	**3 (0.4)**	5	5	5	5	5	4	5	3	5	**4.7 (0.5)**	**4**
Claude Sonnet	2	1	2	3	1	2	**1.8 (0.6)**	5	5	5	5	5	3	5	5	5	**4.8 (0.4)**	**3.6**
ChatGPT 4o	2	2	1	2	2	2	**1.8 (0.2)**	5	5	5	5	3	3	4	4	5	**4.3 (0.8)**	**3.3**
OpenEvidence	2	1	2	1	1	1	**1.3 (0.3)**	4	5	2	5	3	2	3	3	5	**3.6 (1.5)**	**2.7**
Gemini Advanced	1	1	2	2	1	2	**1.5 (0.3)**	4	2	3	5	3	3	4	3	5	**3.56 (1)**	**2.7**
Llama 3.1	1	1	1	3	1	2	**1.5 (0.7)**	3	2	3	5	2	1	3	2	5	**2.9 (1.9)**	**2.3**

Table 3: Pros and Cons of open access LLMs

LLM (AI platform)	Pros	Cons
Overall	-Generally reliable, very user friendly, highly comprehensive and efficient	-Occasional factual inaccuracies and hallucinations (eg. fabricated references) -Lack access to the most recent studies due to their training data cut-off (11)
ChatGPT	-Most advanced and versatile option of available LLMs -GPT-4o is the lowest latency and cheapest model	-Offers more advanced paid “reasoning” models, Gpt-o1 and GPT-o3, but are computationally expensive and slower
Claude	-Designed with emphasis on alignment with human values -Tends to be more cautious about controversial or sensitive topics	-Models less tailored to clinical contexts compared to ChatGPT
DeepSeek	-Fully open-source, promoting transparency and community contributions -Does not have associated license fees	-Struggles with fine-tuning on dialogue -large models (e.g. DeepSeek-Coder 33B) require large amounts of GPU memory
Gemini	-Gemini 1.5 Pro boasts the largest context window as a part of Google’s ecosystem -Gemini 1.5 Flash is one of the fastest models	-Struggles to produce quality responses without significant prompt engineering -Concerns about data privacy and usage with integration into various Google services
Grok 2	-Integration into X’s ecosystem allows Grok to stay up to date with current events and trends -Offers conversational capabilities tailored for social interaction	-Remains suboptimal compared to Claude 3.5 or GPT-4o -As a result of being directly linked to X, a platform with frequent user-generated content, Grok struggles to moderate sensitive/controversial interactions
Llama 3.1	-Llama 3.2 is one of the fastest models (with Gemini 1.5) -Optimized for efficiency with lower computational requirements compared to other models	-Technical expertise required to run properly -Less user-friendly for researchers without technical support
OpenEvidence	-Offers access to most recently curated medical research -Most robust and relevant citations	-Weaker reasoning capabilities than leading frontier models

[1]LLM training data cut-off: ChatGPT October 2023, Claude April 2024 (Sonnet), July 2024

(Haiku), Llama 3.1 December 2023, Gemini May 2024, OpenEvidence and Grok Unknown
Definitions:

1. Context window: maximum number of combined input and output tokens
2. Latency: Time to first token of tokens received, in seconds, after API request sent