Empowering Clinical Trial Design through AI: A Randomized Evaluation of PowerGPT

Yiwen Lu, BS[1,2], Lu Li, BA[1,2], Dazheng Zhang, PhD[1,3], Xinyao Jian, MS[1,3], Tingyin Wang, MS[1,2], Siqi Chen, BS[1,2], Yuqing Lei, MS[1,3], Jiayi Tong, PhD[1,3,4], Zhaohan Xi, PhD[5], Haitao Chu, MD, PhD[6,7], Chongliang Luo, PhD[8,9], Alexis Ogdie, MD[10], Brian Athey, PhD[11], Alparslan Turan, MD[12], Michael Abramoff, MD, PhD[13,14], Joseph C Cappelleri, PhD[15], Hua Xu, PhD[16], Yun Lu, PhD[17], Jesse Berlin, ScD[18,19,*], Daniel I. Sessler, MD[12,*], David A. Asch, MD[20,21,*], Xiaoqian Jiang, PhD[22,*], Yong Chen, PhD[1-3,20,23,24,*]

1. Center for Health AI and Synthesis of Evidence (CHASE), Department of Biostatistics, Epidemiology and Informatics, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA
2. The Graduate Group in Applied Mathematics and Computational Science, School of Arts and Sciences, University of Pennsylvania, Philadelphia, PA, USA
3. Department of Biostatistics, Epidemiology, and Informatics, University of Pennsylvania Perelman School of Medicine
4. Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD, USA
5. School of Computing, Binghamton University, The State University of New York (SUNY), Binghamton, NY, USA
6. Division of Biostatistics and Health Data Science, University of Minnesota, Minneapolis, MN, USA
7. Statistical Research and Data Science Center, Pfizer Inc, New York, NY, USA
8. Division of Public Health Sciences, Department of Surgery, Washington University School of Medicine, St. Louis, MO USA
9. Siteman Cancer Center Biostatistics Shared Resource, Division of Public Health Sciences, Department of Surgery, Washington University School of Medicine, St. Louis, MO USA
10. Department of Medicine and Rheumatology, University of Pennsylvania Perelman School of Medicine, Philadelphia, PA, USA
11. Department of Computational Medicine and Bioinformatics, University of Michigan Medical School, Ann Arbor, MI, USA
12. Department of Anesthesiology, Critical Care and Pain Medicine, McGovern Medical School, Houston, TX, USA
13. Department of Ophthalmology and Visual Sciences, University of Iowa Hospital and Clinics, Iowa City, IA, USA
14. Electrical and Computer Engineering, University of Iowa, Iowa City, IA, USA
15. Statistical Research and Data Science Center, Pfizer Inc, Groton, CT, USA
16. Department of Biomedical Informatics and Data Science, School of Medicine, Yale University, New Haven, CT, USA
17. Center for Biologics Evaluation and Research, Food and Drug Administration, Silver Spring, MD, USA
18. Epidemiology, Rutgers the State University of New Jersey, New Brunswick, NJ, USA
19. Statistical Editor, JAMA Network Open
20. Leonard Davis Institute of Health Economics, University of Pennsylvania, Philadelphia, PA, USA
21. Division of General Internal Medicine, Department of Medicine, Perelman School of Medicine, University of Pennsylvania, Philadelphia, PA, USA
22. Department of Health Data Science and Artificial Intelligence, McWilliams School of Biomedical Informatics, The University of Texas Health Science Center at Houston, Houston, TX, USA
23. Penn Medicine Center for Evidence-based Practice (CEP), Philadelphia, PA, USA
24. Penn Institute for Biomedical Informatics (IBI), Philadelphia, PA, US

*: Senior authors

Background

Statistical power analysis is essential for clinical trial design but remains a technical barrier for many researchers due to the complexity of statistical test selection and sample size estimation[1,2]. Traditional tools, while widely used, often require substantial statistical knowledge and lack flexibility for non-standard designs[3,4]. Moreover, limited access to statistical expertise further hampers timely and accurate trial planning. Recent advances in artificial intelligence—particularly large language models (LLMs)—offer opportunities to simplify complex analytical workflows through natural language interaction[5–7]. However, general-purpose LLMs have shown limited reliability in specialized statistical applications, highlighting the need for domain-specific solutions.

Methods

We developed PowerGPT, an open-source, agent-based system that integrates large language models (LLMs) with statistical engines (R/Python), external APIs, and domain-specific data to automate test selection and sample size calculation through natural language prompts. PowerGPT comprises a dynamic architecture combining user interfaces, computational layers, short- and long-term storage, and external libraries. As shown in **Figure 1**, users interact with the system via conversational input, while PowerGPT selects appropriate statistical methods, performs computations, and returns results in plain language. To evaluate its effectiveness, we conducted a stratified randomized trial with 36 participants from the University of Pennsylvania and UTHealth, equally divided by statistical expertise. Participants used either PowerGPT or conventional tools to complete eight common clinical power analysis tasks. Primary outcomes included task completion rate, accuracy, and time spent. Comparisons across groups assessed PowerGPT's ability to improve efficiency and reduce performance gaps between statisticians and non-statisticians.

Results

PowerGPT significantly outperformed conventional methods across all evaluation metrics (**Figure 2**). Participants using PowerGPT achieved higher task completion rates for both test selection (99.3% vs. 88.9%) and sample size calculation (99.3% vs. 77.8%). Accuracy was also improved in the PowerGPT group for test selection (95.6% vs. 83.6%) and sample size estimation (94.1% vs. 55.4%, $p < 0.001$). Average task completion time was reduced by more than half (4.0 vs. 9.3 minutes, $p < 0.001$), with reduced variability across test types. Performance gains were consistent across all eight statistical scenarios and particularly notable in complex analyses such as Cox models and log-rank tests. Stratified analysis (**Figure 3**) demonstrated that PowerGPT effectively closed performance gaps between statisticians and non-statisticians, enabling comparable levels of accuracy, completion, and efficiency. Notably, non-statisticians using PowerGPT performed on par with statisticians, highlighting its potential to democratize access to high-quality statistical design.

Conclusion

PowerGPT substantially improves the accuracy, efficiency, and accessibility of statistical power analysis. By enabling natural language interaction and closing expertise gaps, it offers a scalable solution to support rigorous and timely clinical trial design across diverse research settings.

Figure 1: Work process of PowerGPT. The step-by-step workflow begins with understanding user objectives through natural language inputs and progresses to delivering actionable statistical results. The figure illustrates adaptive refinement of parameters, integration with external APIs, and presentation of outputs in a user-friendly format.

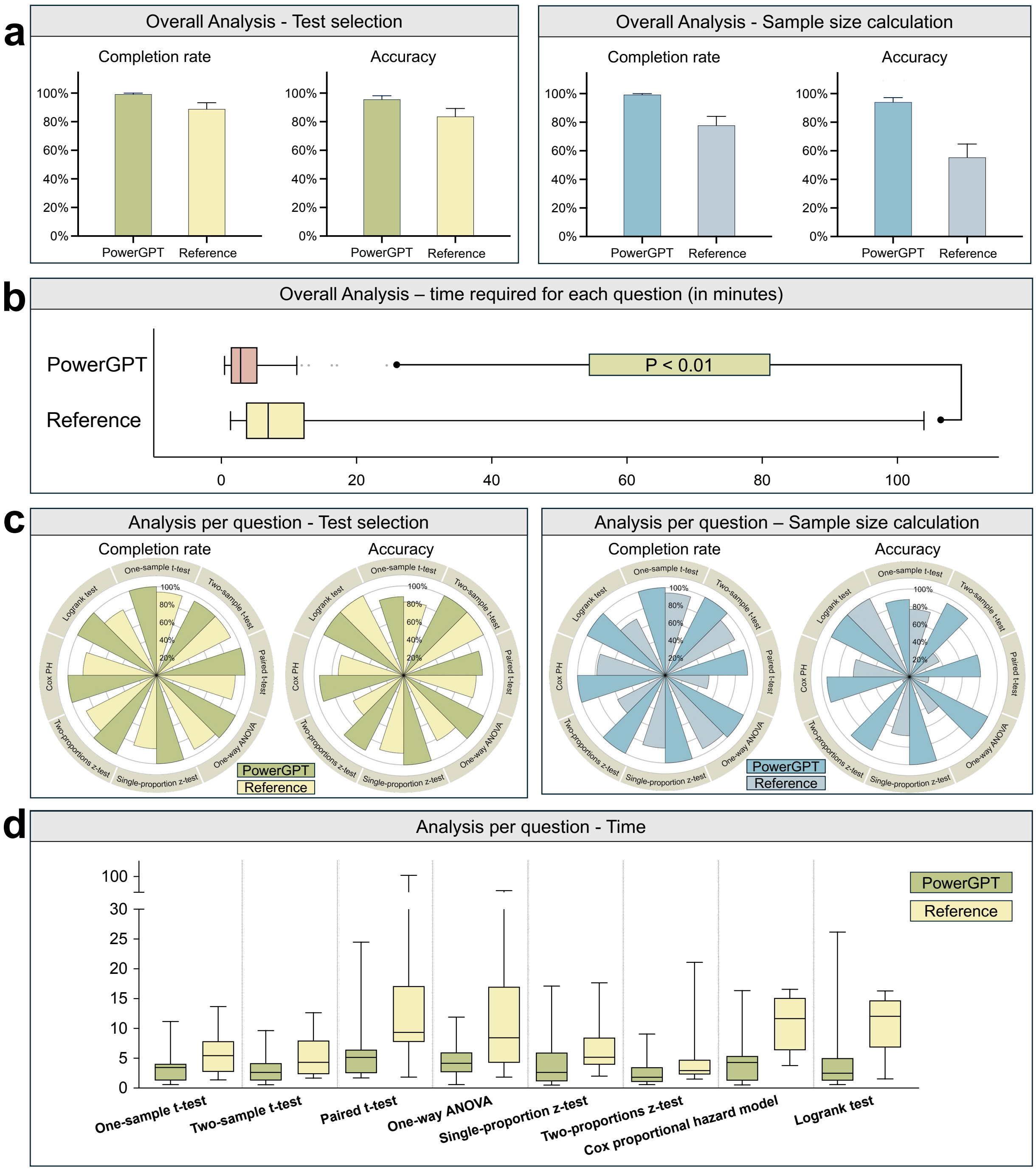

Figure 2. Evaluation results of PowerGPT. (a) Task completion rate and accuracy for test selection and sample size calculation in the PowerGPT and reference groups. PowerGPT users exhibited significantly higher completion rates and accuracy across both tasks. (b) Time required per question for both groups. PowerGPT substantially reduced the time required to complete statistical tasks compared to traditional methods. (c) Task completion rate and accuracy across different statistical tests. While PowerGPT maintained high performance across all tests, the reference group exhibited substantial variability, particularly in more complex tests. (d) Time required per question across different statistical tests. The reference group showed greater variability and longer completion times. PowerGPT provided more consistent and efficient performance across all scenarios.

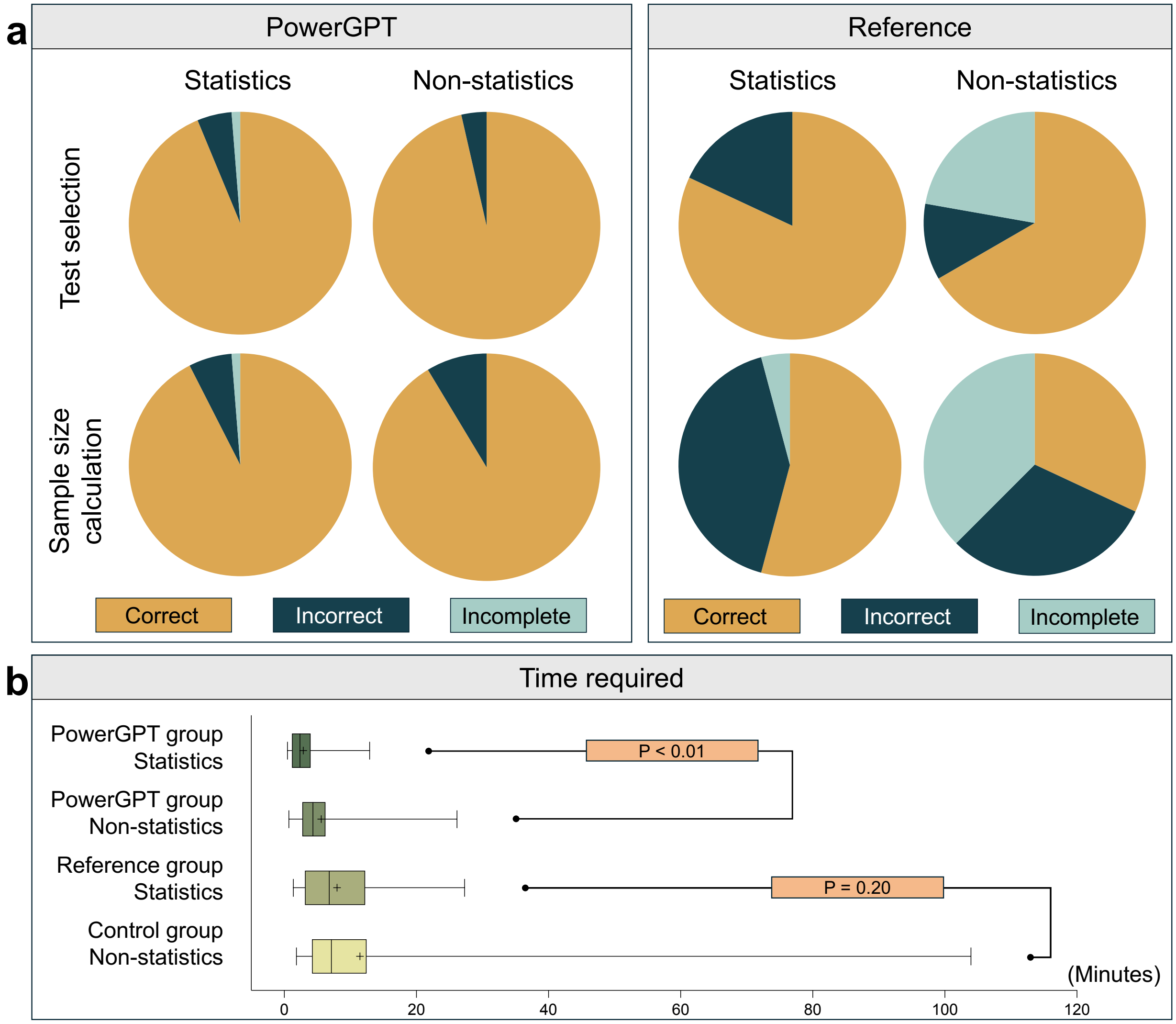

Figure 3. Stratified analysis of PowerGPT's impact by domain of expertise. (a) Task completion rate and accuracy stratified by expertise level. In the reference group, non-statisticians had significantly lower accuracy and higher incompletion rates compared to statisticians. PowerGPT improved performance across both groups and reduced the performance gap. (b) Time required per question stratified by expertise level. In the reference group, non-statisticians took significantly longer to complete tasks, with a long-tailed distribution indicating extreme delays for some participants.

References

1. Cohen J. Statistical Power Analysis. *Curr Dir Psychol Sci* 1992; 1: 98–101.
2. Cohen J. Statistical Power Analysis for the Behavioral Sciences. *Statistical Power Analysis for the Behavioral Sciences*. Epub ahead of print 13 May 2013. DOI: 10.4324/9780203771587.
3. Champely S, Ekstrom C, Dalgaard P, et al. pwr: Basic functions for power analysis, https://nyuscholars.nyu.edu/en/publications/pwr-basic-functions-for-power-analysis
4. Erdfelder E, Faul F, Buchner A. GPOWER: A general power analysis program. *Behavior Research Methods, Instruments, and Computers* 1996; 28: 1–11.
5. Naveed H, Khan AU, Qiu S, et al. A Comprehensive Overview of Large Language Models, https://arxiv.org/abs/2307.06435v10 (2023, accessed 12 February 2025).
6. Sun M, Han R, Jiang B, et al. A Survey on Large Language Model-based Agents for Statistics and Data Science, https://arxiv.org/abs/2412.14222v1 (2024, accessed 12 February 2025).
7. Thirunavukarasu AJ, Ting DSJ, Elangovan K, et al. Large language models in medicine. *Nat Med* 2023; 29: 1930–1940.