Prompted generative artificial intelligence versus interdisciplinary clinicians for academic abstract writing: the AbChat Collaborative long protocol.

Introduction

In the era of artificial intelligence (AI), large language models (LLMs) are increasingly being used within medicine and academia. Recently, chatGPT-4 has demonstrated its ability to generate patient discharge summaries using unstructured patient data in the MIMIC-3 dataset (unpublished, accessible at: https://openreview.net/forum?id=1kDJJPppRG). A study in *Nature* demonstrated that generative AI (GPT-4) can produce convincing abstracts when trained on a corpus of abstracts from leading medical journals (JAMA, the BMJ, the New England Journal of Medicine, the Lancet). In the study, 68% of the abstracts produced by generative AI were identified correctly, whilst 86% of abstracts written by humans were also identified correctly as human abstracts(1).

Even still, the ability of LLMs to synthesise a convincing, de-novo abstract is not scientifically useful nor should it be encouraged. Indeed, it is expected that scientists with a wide understanding of scientific literature would be able to tell which abstracts were created by generative pre-trained transformers versus humans. For example, in the study, the layout of the abstract was a significant factor which allowed scientists to tell which abstracts were produced by AI.

Whilst many have shown that generative AI can summarise entire research studies efficiently, research is often ready for presentation long before it is transcribed into a full manuscript for journal submission. It is known that the quality of prompt when fed to generative AI significantly affects the quality of the output. A significant gap in the current literature is whether, given similar information in the form of an abbreviated prompt, generative AI can produce a scientific abstract of equivalent quality to humans.

Aim

This study will determine the ability of a leading LLM (chatGPT-4-turbo) to generate academic conference abstracts using pre-specified prompts and compare these to the same abstracts as written by clinicians from a variety of medical fields, with the goal of validating LLMs as an academic tool.

Study Design

The study will be prospectively registered with Imperial College NHS Trust once the scientific committee has approved the protocol. No ethical approval is required for this study.

Abstracts submitted by clinicians in a particular field, previously accepted for conference presentation, will be summarised into approximately 100-word bullet point summary prompts. The prompts will be provided to chatGPT-4-turbo via a Python API to generate a 300-word abstract, to be directly compared against the original abstract. Abstracts will be assigned a random code using inbuilt python functionality. Four independent, blinded, senior academic adjudicators from each specialty will score a random selection of AI or clinician abstracts according to a previously validated proforma (2). The adjudicators will additionally be asked whether they think the abstract was written by human or generative AI. Four adjudicators will each score n=47 abstracts, a mixture of AI or clinician generated, such that each human and AI abstract receives a total score from two raters. Each individual project will use the same four adjudicators and this methodology will be followed for each medical discipline included within the study. The abstract scores from each adjudicator will be compared with intra-class correlation coefficients (ICCs) and the internal consistency of the survey will be measured throughout the study.

Outcomes

The primary outcome will be the abstract score for the generative AI abstracts versus the clinician abstracts. The secondary outcomes will be accuracy of the generated abstract as compared to the original prompt, the performance of the LLM as an abstract scorer versus field experts (including the intra-class correlation coefficients (ICC)) and the percentage of plagiarism (using an online free plagiarism checker) as well as originality (scored using an AI output detector).

Statistical Analyses

Average abstracts across all domains will be computed and compared with an unpaired Student's t test/ Wilcoxon Rank test (normality dependent). Reliability of the scorers, including the LLM, and internal consistency will be measured with ICC and Cronbach's alpha respectively. Overall accuracy of the abstracts from the given bullet points will be determined as an overall proportion of abstracts with a significant error that affects the interpretation.

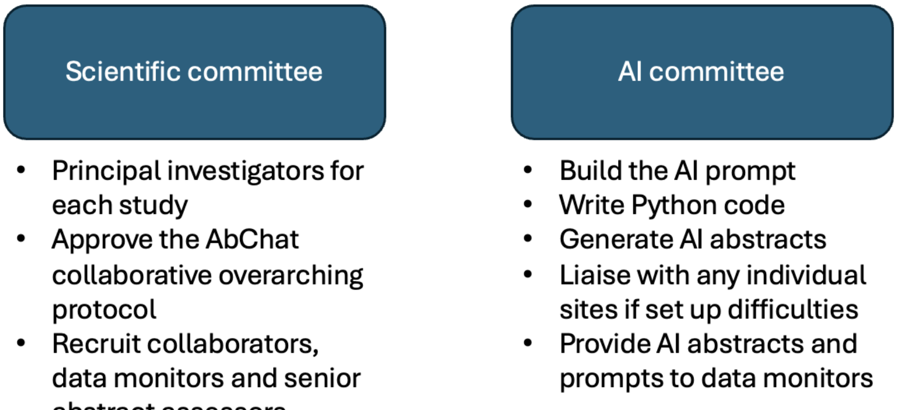
Roles & responsibilities

The AbChat Collaborative is headquartered in Charing Cross Hospital, London, UK. It is the coordinating centre, run by Dr Benedict Turner and the AI committee.

The **Chief Investigator** of the AbChat Collaborative is Dr Benedict Turner. The **Co-Chief Investigator** is Dr Henry Bergman. A complete Team Member list will be made available online and the study will be prospectively registered.

The **Scientific Committee** comprises all the principal investigators for each study. The responsibilities include, but are not limited to, overseeing, and controlling the scientific part of the project to give strategic direction and support to the collaborators they recruit (Figure 1). The Members of the Scientific Committee will approve the study design and protocol of the AbChat Collaborative.

The **AI Committee's** responsibilities include, but are not limited to, maintaining a homogeneous output from the LLM by coding a specific prompt that is piloted and validated prospectively, that can be applied to all abstracts (Figure 1). The AI team will also be responsible for generating the AI abstracts using the specified Python code and chatGPT-4-turbo API and providing these to the principal investigator at each centre.



Scientific committee

- Principal investigators for each study
- Approve the AbChat collaborative overarching protocol
- Recruit collaborators, data monitors and senior abstract assessors

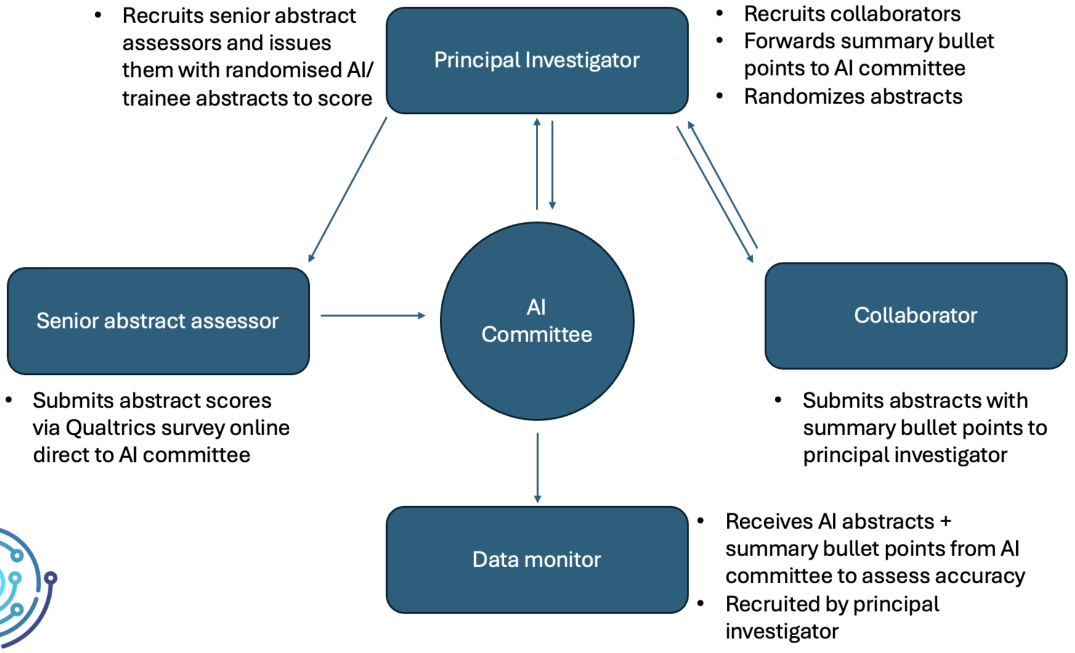
AI committee

- Build the AI prompt
- Write Python code
- Generate AI abstracts
- Liaise with any individual sites if set up difficulties
- Provide AI abstracts and prompts to data monitors

Figure 1 – the roles of the two committees

There are four roles to be fulfilled in each individual study (please see Figure 2).

- Principle investigator
    - Organiser of each individual study and primary author on the publication.
    - Responsible for recruiting the collaborators, senior abstract assessors and independent data monitors.
    - Responsible for ensuring adequacy of the summary bullet points for each clinician abstract and for writing the manuscript for their project.
- Collaborators
    - Submit their previously accepted conference abstracts of approximately 300 words along with the summary bullet points for each abstract.
    - Will be a collaborative author only, but on every publication. There is no limit to the number of collaborators that can be involved in the study.
- Senior abstract assessors
    - Responsible for assessing n=47 abstracts according to the proforma.
    - Must have a record of academia and at least five years of training within the specialty.
    - There will be four senior abstract assessors required for each study and they must not have previously reviewed any of the abstracts submitted.
- Independent data monitors
    - Responsible for reading the summary bullet points of each abstract and verifying that there are no discrepancies between the bullet points and the generative AI abstracts.
    - Two independent data monitors will review each bullet point summary and AI generated abstract to determine the accuracy. Any disagreements will be mediated by the principal investigator.



Principal Investigator

- Recruits senior abstract assessors and issues them with randomised AI/ trainee abstracts to score

- Recruits collaborators
- Forwards summary bullet points to AI committee
- Randomizes abstracts

Senior abstract assessor

- Submits abstract scores via Qualtrics survey online direct to AI committee

AI Committee

Collaborator

- Submits abstracts with summary bullet points to principal investigator

Data monitor

- Receives AI abstracts + summary bullet points from AI committee to assess accuracy
- Recruited by principal investigator

Figure 2 – site set up structure with roles, responsibilities and flow of information

Authorship

All who partake in the study will hold collaborative authorship status on every publication. Additionally, principal investigators will be primary authors on their specific publication, whilst data monitors and senior assessors will also hold named authorship on their given study. The chief and co-chief investigators will be senior authors on all published articles.

Medical disciplines

Principal authors that are current specialty clinicians will be recruited from all 26 disciplines of medicine and will all liaise directly with the single central AI committee (please see Figure 3).

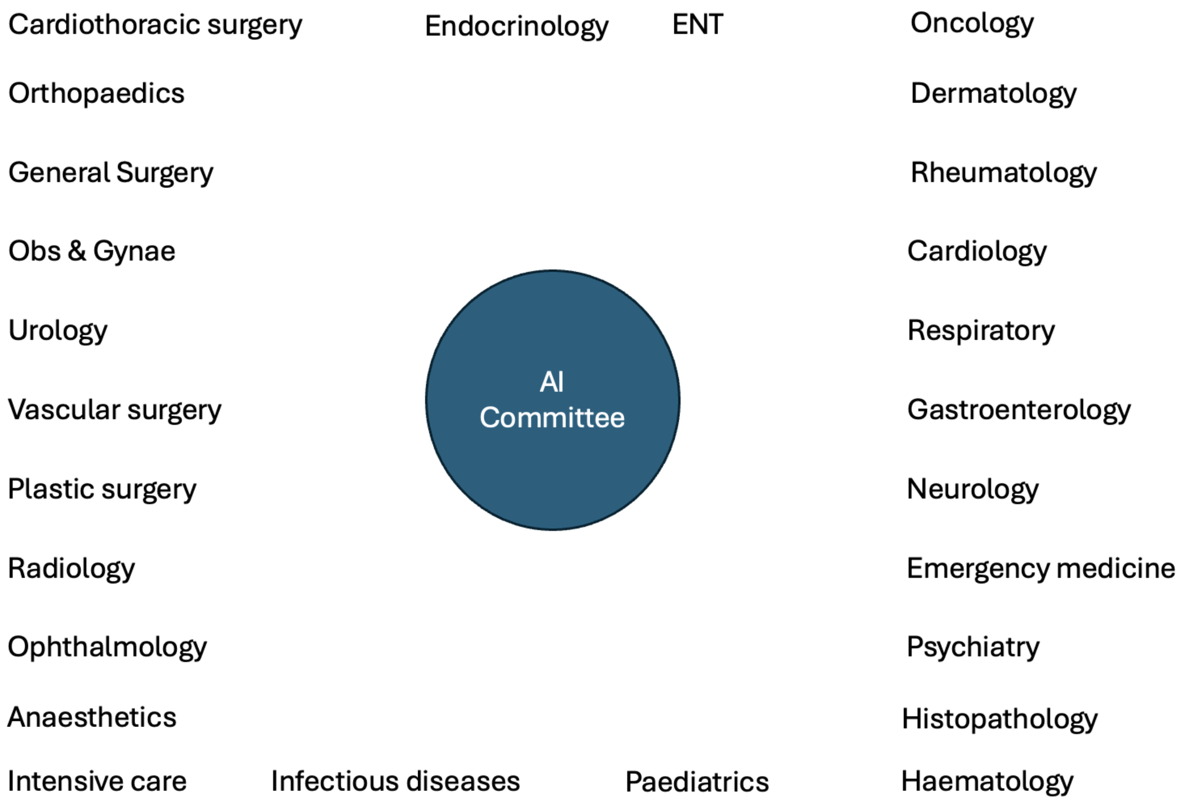
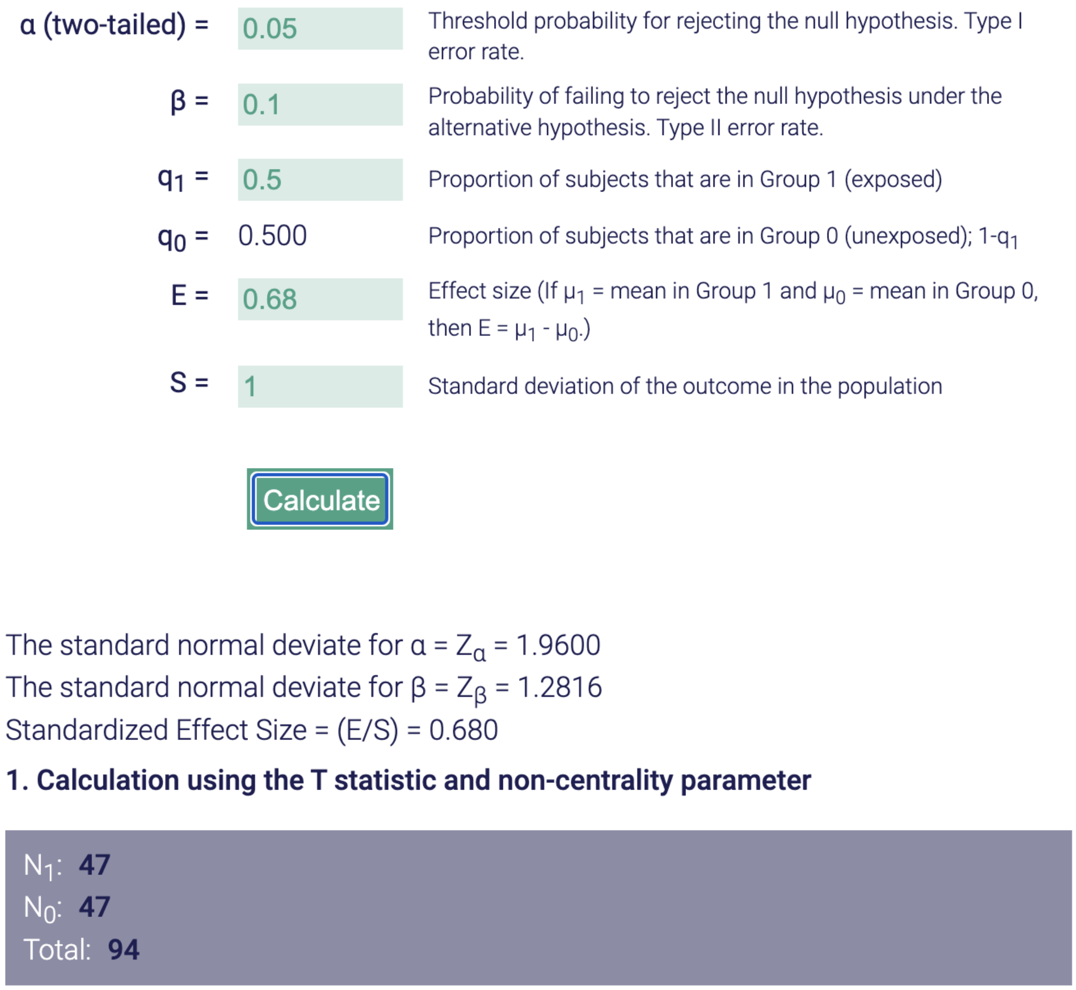


Figure 3 – the medical disciplines to be included within the study.

Power calculations

As 68% of abstracts produced by generative AI were identified as not being human, an effect size of 68% is the best estimate of how much better humans might perform than AI. Taking alpha of 0.05 and beta of 0.1 for a 90% power to detect an effect, as well as an assumed standard deviation of 1 point between the scores, a total of 94 abstracts (47 in each group) will be required for the present study (Figure 4).



α (two-tailed) = 0.05 Threshold probability for rejecting the null hypothesis. Type I error rate.

β = 0.1 Probability of failing to reject the null hypothesis under the alternative hypothesis. Type II error rate.

q_1 = 0.5 Proportion of subjects that are in Group 1 (exposed)

q_0 = 0.500 Proportion of subjects that are in Group 0 (unexposed); $1-q_1$

E = 0.68 Effect size (If μ_1 = mean in Group 1 and μ_0 = mean in Group 0, then $E = \mu_1 - \mu_0$.)

S = 1 Standard deviation of the outcome in the population

The standard normal deviate for $\alpha = Z_\alpha = 1.9600$
The standard normal deviate for $\beta = Z_\beta = 1.2816$
Standardized Effect Size = (E/S) = 0.680

1. Calculation using the T statistic and non-centrality parameter

N_1: **47**
N_0: **47**
Total: **94**

Figure 4 – an excerpt from an online power calculator demonstrating the sample size required.

Data ownership

The AbChat Collaborative will act as the custodian of the data. All participants will be able to access their own submitted data without the need for permission from the AbChat Collaborative. The Chief Investigators and Scientific committees together will decide about data sharing requests and will consider all such requests based on the quality and validity of the proposed project.

Data confidentiality

There will be no individual or centre-related information included within the abstracts, all data will be fully anonymized.

Timeline

From the go-live date for each individual project, the turnaround time for completion of data collection and abstract scoring is 8 weeks. This allows sufficient time for the recruitment of collaborators and their abstracts with summary bullet points (4 weeks) followed by abstract scoring and audit by the independent data monitors (4 weeks). Principal investigators will be contacted every fortnight by the chief investigators to check for any problems in establishing the project. A single extension of 2 weeks may be applied to any individual study at the discretion of the chief investigators, after which individual projects may be terminated and a new principal investigator sought.

Publication

A standardised template for the methods and the statistical analyses for the collected data will be provided to the principal investigator for each project. The principal investigator will be wholly responsible for the write up of the project they lead and must do so in a timely fashion (4 weeks) once provided with the results. A meta-analysis with all specialty data will be conducted once all data has been received from the principal investigators. The chief investigators are currently contacting journals to gauge appetite for a special issue on AI in academia, in which all of the AbChat Collaborative manuscripts would be published collectively. If this target is not feasible, principal investigators will be responsible for publishing their results in a specialty-specific or medical education journal.

No participant in this study will have to pay any DOI charges, nor open access nor other publication fees. These will be covered by the chief investigators through Imperial College London affiliation.

Presentation

After data analysis has been completed, authors may apply to present their findings at any conference but we kindly ask that you complete a presentation form so that a record of presentation can be kept.

References

1. Gao CA, Howard FM, Markov NS, Dyer EC, Ramesh S, Luo Y, et al. Comparing scientific abstracts generated by ChatGPT to real abstracts with detectors and blinded human reviewers. NPJ Digit Med. 2023;6(1):75.
2. Timmer A, Sutherland LR, Hilsden RJ. Development and evaluation of a quality score for abstracts. BMC Med Res Methodol. 2003 Feb 11;3:2. doi: 10.1186/1471-2288-3-2. Epub 2003 Feb 11. PMID: 12581457; PMCID: PMC149448.