A pilot study evaluating the performance of a custom-built large language model-based app that uses reporting guideline items to generate manuscript abstracts

Karen King,[1] Niall Harrison,[1] Lauren Cerruto,[2] Kristjan Kalm,[3] Henrique Brito,[3] Antoine Gallier,[3] Kate Unsworth,[1] Jennifer van Zwieten[1]

[1]OPEN Health, London, UK; [2]OPEN Health, Parsippany, NJ, USA; [3]Artefact, Paris, France

OPEN HEALTH

INTRODUCTION

- Researchers are evaluating the potential applications of large language models (LLMs) in medicine, including in medical writing[1]
- LLMs such as ChatGPT have shown promise in the generation of readable, accurate plain language summaries based on academic manuscripts[2-4]
- Studies using ChatGPT to generate abstracts from the body text of published manuscripts have generally found the outputs to be inferior to the original human-written abstracts[5,6]
  - A study in which abstracts for 62 articles were given an Overall Quality Score (OQS) based on adherence to CONSORT, readability, and hallucination rate found that original abstracts achieved a higher mean OQS (11.89 [95% CI 11.23-12.54) than those generated using either GPT-3.5 (7.89; 95% CI: 7.32-8.46) or GPT-4 (5.18; 95% CI: 4.64-5.71)[5]
  - A study of 30 abstracts generated using ChatGPT from the body text of published manuscripts found that experts evaluated the quality of LLM-generated abstracts as lower than that of the original abstracts (Likert scale [1-10]: mean 4.72, SD 2.09 vs mean 8.09, SD 1.03; P<0.001)[6]
- This study was conducted to assess the appropriateness and accuracy of abstracts generated by Conspectus, a custom-built LLM-based app that compiles a prompt from automated instructions based on reporting guidelines and user-defined specifications

METHODS: PILOT STUDY DESIGN

USER

Upload final submitted text of published manuscripts, without abstract

Specify:

- Output type (abstract)
- Word count
- Spelling (UK/US)
- Trial type

CONSPECTUS

Generates proposed prompt structure, based on user input and relevant reporting guideline

USER

Review proposed prompt structure and modify if required

CONSPECTUS

Generates draft text

USER

- Record time taken
- Evaluate Conspectus usability

REVIEWER

- Evaluate output quality

RESULTS: CONSPECTUS PERFORMANCE

Table 1. Users, studies, and reviewers

Users, n (%)	**22**
Medical Writer	8 (36)
Senior Medical Writer	4 (18)
Principal Medical Writer	4 (18)
Scientific Director	6 (27)
Abstracts generated, n (%)	**36**
Clinical trial	17 (47)
Retrospective cohort	9 (25)
Prospective cohort	3 (8)
Survey	3 (8)
Systematic literature review	2 (6)
Population pharmacokinetic modelling	1 (3)
Post hoc pooled data	1 (3)
Reviews conducted by 14 reviewers, n (%)	**36**
Scientific Director or higher	27 (75)
Principal Medical Writer	6 (17)
Senior Medical Writer	2 (6)
Medical Writer	1 (3)

Table 2. Example from development process showing improved parameters vs default performance of underlying LLM

Parameter	Underlying LLM performance	Conspectus performance
Numeric values included per abstract, n	5.7	19.7
Proportion of numeric values aligned with original abstract, %	36	77
Proportion of values aligned with original manuscript, %	40	85
Mean deviation of generated word count from specified word count, n	42.5	16.6

Based on 13 paired abstract generations reviewed by OPEN Health.

Table 3. Usability (n=22 users)

Parameter	Value
Mean (range) clarity of interface rating (minimum 1, maximum 5)	4.36 (3-5)
n (%) agree that Conspectus improves process	18/22 (82)
n (%) would recommend Conspectus to colleagues	21/22 (95)
Mean (range) overall rating (minimum 1, maximum 5)	4.18 (3-5)

Table 4. Output quality (n=36 reviews)

- All generated abstracts had an appropriate structure that matched the briefing form submission

Section	**Accuracy:** is the included information correct?*	**Appropriateness:** is the included information relevant?*
Introduction	30/36 83%	25/36 69%
Methods	33/36 92%	30/36 83%
Results†	765/782 98%	21/36 58%
Conclusion	28/36 78%	24/36 67%

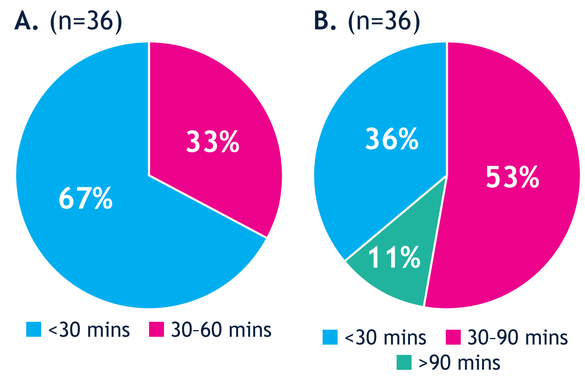
*Proportion of abstracts rated by reviewer as average, above average, or excellent (5-point scale from very poor to excellent) (except for results accuracy).

†Percentage of included datapoints that were accurate; 30/36 results sections were 100% accurate (range 56%-100%).

Workflow and time commitment

- At first use of Conspectus, 13/16 users (81%) spent less than 15 minutes familiarizing themselves with the tool before being ready to use it
- Users had the option to comment on their experience; 11/18 (61%) who commented said that they felt Conspectus had the potential to save them time

Figure 1. (A) Actual time taken to review Conspectus outputs (B) Reviewer-estimated time for human writer to revise Conspectus outputs



LIMITATIONS

- The pilot did not fully reflect real-world workflow, and in particular did not collect comparative data on the time required for a human writer to draft an abstract vs use of Conspectus (including refinement of prompt and checking of output)
  - Generated abstracts require senior scientific review in both standard and Conspectus workflow
- Lower scores for accuracy and appropriateness were associated with:
  - Manuscripts reporting study types not included in the current Conspectus configuration (e.g. pharmacokinetics/pharmacodynamics)
  - Manuscripts reporting analyses that do not have a dedicated reporting guideline (e.g. secondary and *post hoc* analyses of clinical trials)
  - User acceptance of default prompt – only 3/22 users customised the prompt

CONCLUSIONS

- **Based on reviews by experienced medical writers and scientific directors, abstracts generated using Conspectus were correctly structured and largely accurate**
  - It may be possible to increase appropriateness through refinement of the briefing form and additional user training on prompt customisation
- **Conspectus was rated positively by users**
- **Generated abstracts require human review to validate data and proposed framing**
  - Use of Conspectus in live projects requires approval of authors and sponsors, and should be disclosed to target journals and congresses
- **Future work could assess Conspectus in the real-world setting (live projects), and in other use cases where more robust time-saving benefits may be expected**

REFERENCES

1. Meng et al. The application of large language models in medicine: a scoping review. *iScience* 2024;27:109713. 2. Norman et al. Blinded assessment of potential of generative artificial intelligence to generate plain language summaries. *CMRO* 2024;40(S2):S45. 3. Shepherd et al. A multistakeholder assessment of generative artificial intelligence platforms for creating plain language summaries. *CMRO* 2024;40(S2):S41. 4. Satam et al. AI's might versus human prowess in crafting plain language summaries. *CMRO* 2024;40(S2):S7. 5. Hwang et al. Can ChatGPT assist authors with abstract writing in medical journals? Evaluating the quality of scientific abstracts generated by ChatGPT and original abstracts. *PLoS One* 2024;19:e297701. 6. Cheng et al. Comparisons of quality, correctness, and similarity between ChatGPT-generated and human-written abstracts for basic research: cross-sectional study. *J Med Internet Res* 2023;25:e51129.

ACKNOWLEDGEMENTS

The authors would like to acknowledge the following for their assistance in generating the data presented in this study: Amy Pashler, Arezou Seyed Hossein, Brian Jepson, Caroline McGown, Casey McKeown, Grant Kirkpatrick, Hannah Brazier, Harriet Burt, Hayley Dawson, Jon Viney, Leigh Church, Liam Gillies, Pallavi Patel, Rebecca Cunningham, Reece Bracewell, Sabah Farooq, Sara Black, Sarah Case, Tony Reardon, Viji Venugopalan.

The authors would like to thank Karen King for input throughout the pilot study, including interpretation of the results.

DISCLOSURES

Niall Harrison, Lauren Cerruto, Kate Unsworth, and Jennifer van Zwieten are employees of OPEN Health Communications.

Kristjan Kalm, Henrique Brito, and Antoine Gallier are employees of Artefact.