MSR68

Variability and Improvements of Answers Generated with Different Versions of Large Language Models

E Benbow,[1] T Reason,[1] B Malcolm,[2] S L Klijn,[3] N Hill,[3] S Teitsson[2]

[1]Estima Scientific, London, UK; [2]Bristol Myers Squibb, Uxbridge, UK ; [3]Bristol Myers Squibb, Princeton, NJ, USA

Key Messages

What's already known on this topic

As part of a systematic review or NMA, there is a need to extract data from relevant publications. This is a time-consuming and error-prone part of this process and usually requires input from a minimum of three people[1-8].

It has previously been demonstrated that there is potential to use LLMs, such as GPT-4, to automate data extraction for NMA. However, there are other LLMs that are now available, so it is unclear which might be optimum for this specific task, and whether prompts are generalizable across LLMs.

What this study adds

This study has demonstrated that, of the models considered, the best models for data extraction are currently OpenAI's GPT-4 Turbo and Anthropic's Claude-3 models. We have shown that the same prompts can be used with all models considered, to achieve an accurate extraction of required data items >95%, when using the most recently released models.

How this study might affect research, practice, or policy

The ability to use LLMs to automate data extraction could result in significant time savings and reduce human error in the systematic review and NMA process. There is a need to test the LLM-based processes developed across a greater number of disease areas and outcome types.

Generalizability of prompts across models and model version (as demonstrated in this study) can have a large impact on the use and acceptability of LLMs in HEOR.

Introduction

- The emergence of artificial intelligence (AI), capable of human-level performance on some tasks[9,10], presents an opportunity to revolutionize development of systematic reviews and network meta-analyses (NMAs).
- We have previously demonstrated that there is potential to use large language models (LLMs), such as GPT-4, to automate data extraction for NMA[11].
- Since OpenAI's release of the GPT-3.5 LLM in March 2022, subsequent updates have introduced new and enhanced models. In addition, other models have become available, such as Meta's Llama2 and Anthropic's Claude-3.
- The impact of response variations among these models on the accuracy of automated data extraction remains uncertain, along with whether the same prompts can be used with all models.

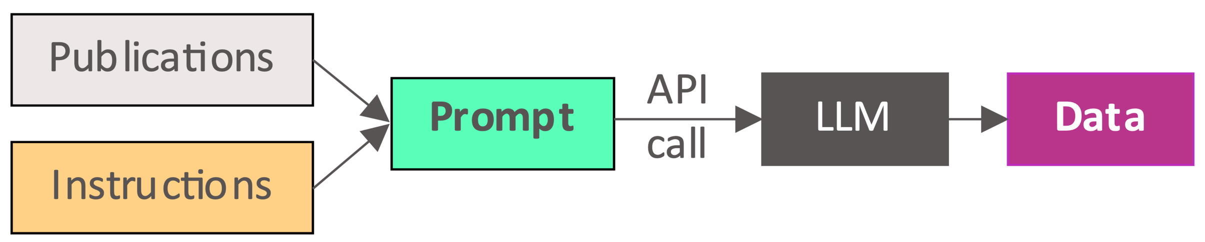
Aims

- To evaluate the variability and accuracy of answers generated by different LLMs during data extraction for an NMA using four case studies.

Methods

- In previous work, we tested GPT-4's ability to extract data using four case studies. These four studies required GPT-4 to review publications and extract the data required to conduct an NMA. The case studies covered three different outcomes (two time-to-event [TTE] and one binary) and two disease areas (metastatic non-small-cell lung cancer and moderate-to-severe hidradenitis suppurativa).
- For each case study, we wrote a Python script that sent a prompt via an API call to the LLM for each publication in the NMA. The prompt included text from the publication and a request to extract all relevant data from the supplied publication (Figure 1). Detail of the case studies and the process used is provided in Reason et al., 2024[11].
- We used the same four case studies in this current study and did not alter the prompts used to instruct the LLMs (developed and tuned for use with GPT-4). We replicated the approach taken in the previous research, to run the Python script for each case study, and for each LLM, 20 times.
- Six LLMs were included in this study:
  - GPT-3.5 Turbo
  - GPT-4
  - GPT-4 Turbo Beta (Nov '23 release) [GPT-4 Turbo '23]
  - GPT-4 Turbo Beta (Jan '24 release) [GPT-4 Turbo '24]
  - Claude-3 Opus
  - Claude-3 Sonnet.
- Thus, for each LLM, we obtained 20 sets of their attempts to extract data per trial publication (22 publications in total). These results were then compared with the results of the data extraction conducted (and checked) by systematic literature review and NMA experts.

Figure 1. LLM-based process for extracting data required for NMA



Results

- The data extraction accuracy achieved by each LLM for each case study is shown in Table 1.
- GPT-3.5 Turbo was consistently the worst performing LLM of those considered, with an average correct extraction rate between 65.1% and 86.3%, when considering all data items and runs. This is perhaps not surprising, since it is the oldest of the models considered.
- GPT-4 Turbo '23 was found to be the second worst performing LLM, which was initially surprising, as this was expected to be an improvement on GPT-4. However, OpenAI have recognized that this model was "lazy" and would sometimes fail to complete a task and have tried to address this in the latest version of GPT-4 Turbo[12]. They appear to have achieved this, as GPT-4 Turbo '24 was the best performing model that we considered, achieving an excellent average extraction rate of between 99.1% and 100% for the case studies.
- The two Claude models seemed to perform similarly to each other and achieved a very good average extraction rate of over 95%. For three of the four case studies, the Claude models achieved a higher average extraction rate than GPT-4.
- We found that the output from the Claude models was much more consistent between the 20 runs than the GPT models achieved. Where Claude did not achieve 100% data extraction, it was due to it consistently responding "not reported" for a particular data item (e.g. number at risk in one publication) but if it was able to correctly extract a data item once, then it correctly extracted that data item for the remaining 19 runs. Conversely, the GPT models would sometimes extract the correct data, but other times would respond "not reported". The Claude models also seemed to be better at following the instructions to provide data in a specific format, without additional narrative (Figure 2). In contrast, the GPT models seemed to choose how to report the data requested (making it much harder to automate data extraction from the LLM's response, without asking the LLM itself to do this). GPT-4 Turbo '24 seemed to be the best GPT model when following the format instructions (Figure 2).
- GPT-3.5 Turbo was the only LLM that was found to report incorrect data – where the other LLMs did not achieve 100% extraction rate success, it was due to failure to extract the data items (they stated "not reported" for these items), and not incorrect reporting.

Discussion

- This study has shown that, when choosing which model to use for HEOR-related tasks in the future, all relevant currently available models should be tested, in order to determine which is most suited/can achieve the highest level of accuracy.
- The cost of using an LLM will also need to be considered, when choosing a model (currently, Claude-3 Opus is more expensive than, but Claude-3 Sonnet is less than half the price of, GPT-4 Turbo).
- The same prompts were used for all models herein. However, there are other approaches that could be considered in future research:
  - Using AI agent workflows across different models.
  - Using auto-prompting across different models.
  - Tailoring the prompting for different models.

Table 1. Large Language Model (LLM) Extraction Accuracy

Case Study	LLM	Per run: Number of required data points	Minimum extracted correctly	Maximum extracted correctly	Average data extraction success rate over all runs
Binary outcome in hidradenitis suppurativa	GPT-3.5 Turbo	30	12	30	74.7%
	GPT-4	30	22	30	97.3%
	GPT-4 Turbo (Nov '23)	30	22	30	93.3%
	GPT-4 Turbo (Jan '24)	30	30	30	100%
	Claude-3 Opus	30	30	30	100%
	Claude-3 Sonnet	30	30	30	100%
Time to event (TTE) outcome in mNSCLC 1	GPT-3.5 Turbo	46	32	46	86.3%
	GPT-4	46	44	46	95.4%
	GPT-4 Turbo (Nov '23)	46	43	46	93.9%
	GPT-4 Turbo (Jan '24)	46	44	46	99.7%
	Claude-3 Opus	46	44	46	95.7%
	Claude-3 Sonnet	46	44	46	95.7%
Time to event (TTE) outcome in mNSCLC 2	GPT-3.5 Turbo	98	64	90	77.9%
	GPT-4	98	91	98	97.9%
	GPT-4 Turbo (Nov '23)	98	91	98	97.4%
	GPT-4 Turbo (Jan '24)	98	96	98	99.9%
	Claude-3 Opus	98	94	98	95.9%
	Claude-3 Sonnet	98	94	98	95.9%
Time to event (TTE) outcome in mNSCLC 3	GPT-3.5 Turbo	46	10	33	65.1%
	GPT-4	46	42	46	96.9%
	GPT-4 Turbo (Nov '23)	46	36	46	89.6%
	GPT-4 Turbo (Jan '24)	46	44	46	99.1%
	Claude-3 Opus	46	46	46	100%
	Claude-3 Sonnet	46	46	46	100%

Figure 2. Requested output format (top), example output from Claude-3 Opus (middle) and from GPT-4 Turbo '24 (bottom) for Case Study 1

Requested format:
Study name, treatment names, number or % of patients experiencing clinical response, number at risk in each treatment arm, total number of patients in trial, randomisation ratio.

Claude-3 Opus extraction:
Based on the information provided, here is the extracted data in the requested format:
PIONEER I:
Treatment names: Placebo, Adalimumab Weekly,
Clinical response: 40, 64
Number at risk: 154, 153
Number of patients in trial: 307
Randomisation ratio: 1:1

The number of clinical responses was determined from the percentages and number at risk provided in Figure 3 for the week 12 timepoint. The total trial size and randomization ratio were stated directly in the text. No other treatments were mentioned.

GPT-4 Turbo '24 extraction:
Based on the provided information from the study "Adalimumab for Hidradenitis Suppurativa" published in the New England Journal of Medicine on August 4, 2016, the clinical response outcomes for the treatments in the PIONEER I trial are as follows:
Treatment 1: Placebo, **Treatment 2:** ADA Weekly
Number of clinical responses 1: Not explicitly reported for PIONEER I Placebo at week 12 but inferred as 26.0% of participants based on the efficacy section.
Number of clinical responses 2: Not explicitly reported for PIONEER I ADA Weekly at week 12 but inferred as 41.8% of participants based on the efficacy section.
Number at risk 1: 154 (Placebo)
Number at risk 2: 153 (ADA Weekly)
Total trial size: 307
Randomisation ratio: Not explicitly reported, but based on the numbers provided, it appears to be approximately 1:1.
Please note, the exact number of clinical responses for each treatment group at week 12 is not directly provided in the text excerpt. The percentages given in the efficacy section were used to infer the number of clinical responses, but without the exact numbers, these are approximations based on the total number at risk and the reported percentages.

Conclusions

- The generalizability of prompts across LLMs and LLM versions has been demonstrated within this study.
- GPT models have exhibited notable enhancements in accurately extracting required NMA data. GPT-4 Turbo '24 demonstrated superior performance in this limited test, although it was not significantly better than GPT-4 or the two Claude-3 models.
- Whilst achieving a marginally lower data extraction success rate than GPT-4 Turbo '24, the Claude-3 models appear to be more consistent and concise in their responses than the GPT models and appear to follow instructions better.
- It may be possible to achieve even better performance with Claude-3 models if the prompts are tailored to these models.

References

1. Higgins J, Thomas J, Chandler J, et al. Cochrane Handbook for Systematic Reviews of Interventions Version 6.4. Cochrane; 2023. www.training.cochrane.org/handbook
2. Mathes T, Klaßen P, Pieper D. Frequency of data extraction errors and methods to increase data extraction quality: a methodological review. BMC Med Res Methodol. 2017;17(1):152. doi:10.1186/s12874-017-0431-4
3. Saldanha IJ, Schmid CH, Lau J, et al. Evaluating Data Abstraction Assistant, a novel software application for data abstraction during systematic reviews: protocol for a randomized controlled trial. Syst Rev. 2016;5(1):196. doi:10.1186/s13643-016-0373-7
4. Carroll C, Scope A, Kaltenthaler E. A case study of binary outcome data extraction across three systematic reviews of hip arthroplasty: errors and differences of selection. BMC Res Notes. 2013;6(1):539. doi:10.1186/1756-0500-6-539
5. Horton J, Vandermeer B, Hartling L, Tjosvold L, Klassen TP, Buscemi N. Systematic review data extraction: cross-sectional study showed that experience did not increase accuracy. J Clin Epidemiol. 2010;63(3):289-298. doi:10.1016/j.jclinepi.2009.04.007
6. Gøtzsche PC, Hróbjartsson A, Marić K, Tendal B. Data Extraction Errors in Meta-analyses That Use Standardized Mean Differences. JAMA. 2007;298(4). doi:10.1001/jama.298.4.430
7. Buscemi N, Hartling L, Vandermeer B, Tjosvold L, Klassen TP. Single data extraction generated more errors than double data extraction in systematic reviews. J Clin Epidemiol. 2006;59(7):697-703. doi:10.1016/j.jclinepi.2005.11.010
8. Jones AP, Remmington T, Williamson PR, Ashby D, Smyth RL. High prevalence but low impact of data extraction and reporting errors were found in Cochrane systematic reviews. J Clin Epidemiol. 2005;58(7):741-742. doi:10.1016/j.jclinepi.2004.11.024
9. Wang A, Pruksachatkun Y, Nangia N, et al. SuperGLUE: A Stickier Benchmark for General-Purpose Language Understanding Systems. Published online 2019. doi:10.48550/ARXIV.1905.00537
10. Adiwardana D, Luong MT, So DR, et al. Towards a Human-like Open-Domain Chatbot. Published online 2020. doi:10.48550/ARXIV.2001.09977
11. Reason T, Benbow E, Langham J, Gimblett A, Klijn SL, Malcolm B. Artificial Intelligence to Automate Network Meta-Analyses: Four Case Studies to Evaluate the Potential Application of Large Language Models. PharmacoEconomics - Open. 2024;8(2):205-220. doi:10.1007/s41669-024-00476-9
12. OpenAI. OpenAI Documentation - Models: GPT-4 and GPT-4 Turbo. Accessed March 21, 2024. https://platform.openai.com/docs/models/gpt-4-and-gpt-4-turbo

Presented at ISPOR 2024; May 5-8, 2024; Atlanta, GA

Email: emma.benbow@estima-sci.com