Automated Extraction of Cost-Effectiveness Models Data from Health Technology Assessment Submissions Using Large-Language Models (LLMs): Does the Prompting Approach Matter?

Poster No. MSR179

Szabo, Gabor[1]; Pinsent, Amy[2]; Slim, Mahmoud[3]; Sullivan, Shannon[4]; Benedict, Agnes[5]; Rivolo, Simone[6]

[1]Evidera Ltd., a business unit of PPD, part of Thermo Fisher Scientific, Budapest, Hungary; [2]Evidera Ltd., a business unit of PPD, part of Thermo Fisher Scientific, London, UK; [3]Evidera Inc., a business unit of PPD, part of Thermo Fisher Scientific, Montreal, Canada (at time of study); [4]Evidera Ltd., a business unit of PPD, part of Thermo Fisher Scientific, Paris, France; [5]Evidera Ltd., a business unit of PPD, part of Thermo Fisher Scientific, Vienna, Austria; [6]Evidera Ltd., a business unit of PPD, part of Thermo Fisher Scientific, Milan, Italy

Background

- Over the past few years, multiple studies have explored the capabilities of using large language models (LLMs) to automatically extract key data of interest from medical literature.[1-3]
- Recently, LLMs have been successfully applied[4] to automatically extract information from prior technology appraisals (TAs). The National Institute for Health and Care Excellence (NICE) recently published a position statement[5] providing guidance on the use of artificial intelligence (AI) in evidence generation.
- While the use of LLMs can expedite the review of existing TAs, making health economic model conceptualization faster, LLM output quality is highly dependent on adequate prompting strategies.[3,5] Therefore, emphasis remains on the user to devise prompts that maximize accuracy of automatic extraction in LLMs.

Objectives

- Evaluate the performance of alternative prompting strategies for automated cost-effectiveness model (CEM) data extraction from nine published TAs, using an LLM (Generative Pre-trained Transformer 4 [GPT-4]).

Methods

- An overview of the process followed for this study is displayed in **Figure 1**, with more details provided in the Supplemental Materials.
- A single AI-agent persona was developed, through a series of pre-prompts (i.e., text instructions) characterizing the AI-agent task objective, personality (e.g., formal, knowledgeable, evidence-based), rules and behaviors (e.g., detail-oriented, factual accuracy paramount), and step-by-step process to be followed. The same AI-agent was used throughout all the analyses.
- Four alternative prompting strategies were examined: 1) simple prompt; 2) a sequence (chain) of three prompts with increasing complexity; 3) complex chain-of-thought prompt; and 4) complex chain-of-thought prompt combined with a domain-specific example.[6] As an example, **Figure 2** shows the different prompting strategies for the model structure domain.

Results

- Success rates of the four prompting strategies across simple and advanced domains are displayed in **Figure 3**.

Simple domains

- For simple domains, the "chain of 3" prompts strategy (Strategy 2) outperformed (or performed as good as) the other prompting strategies, correctly extracting information most of the time (model structure: 100%, time horizon: 78%, cycle length and health states: 67% each, list of comparators: 44%).
- The list of comparators was not adequately extracted by any prompting strategy (<45% correct extractions).

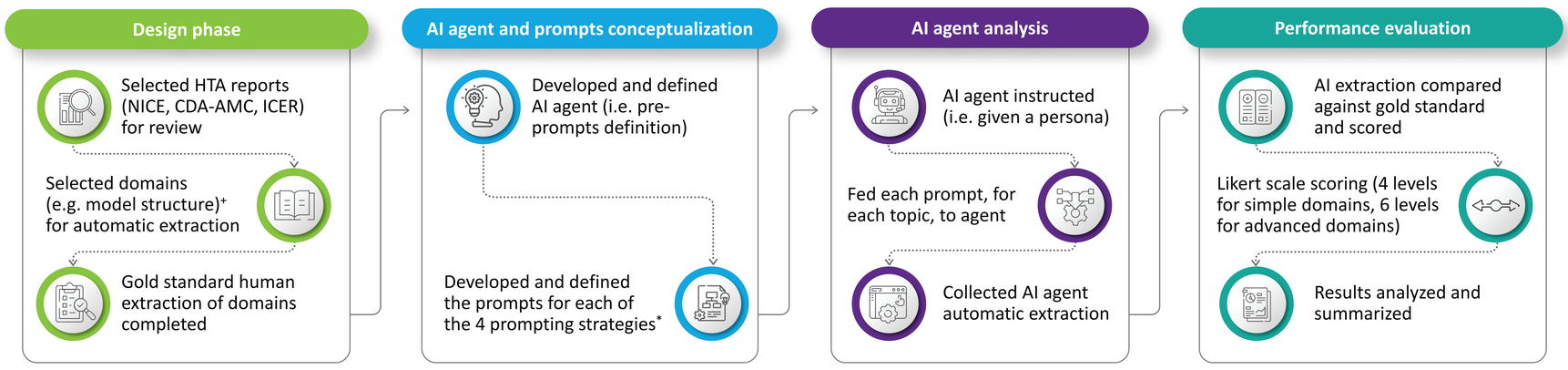
Advanced domains

- The advanced domains were more challenging to extract across all prompting strategies, with no single strategy consistently outperforming the others.
- Committee critiques and modeling assumptions were the most challenging domains to extract correctly (33%–56%) across all prompting strategies. On the other hand, the utility modeling approach was correctly extracted most of the time (67%–78%) by Strategies 2 through 4, with only the simplest prompting strategy (Strategy 1) struggling to achieve satisfactory performance (44% correct extractions).

Additional findings

- It was observed that more complex prompting strategies (e.g., Strategies 3 and 4, where a "chain-of-thought" prompt with or without domain-specific examples was provided) can fail on simple domains (e.g., model structure) even if simpler prompting strategies are successful, highlighting the need to align the prompts' complexity to the domain that needs to be extracted. In other words, it appears that prompts may be overengineered and LLM may "overthink" a problem.
- A strategy with a sequence (chain) of prompts (Strategy 2) has the advantage that it allows the LLM to "learn" after each prompt (e.g., in one TA extraction with Strategy 2, the LLM incorrectly reported that the information was not available for the first two prompts, then extracted the correct information for the third prompt). Note that the same third prompt will not necessarily give a satisfactory answer if used as a single prompt in a new "independent" automatic extraction.
- It was expected that a strategy with a "chain of 3" prompts (Strategy 2) might result more often in the LLM wrongly reporting information not available vs. single prompt strategies. This is because the sequence (chain) of prompt strategy "pushes" the LLM to improve the answer prompt after prompt. However, this was not observed.

Figure 1. Study overview



+ The 10 CEM domains extracted were categorized in five simple domains (model structure, time horizon, cycle length, comparators list, health states) and five advanced domains (key outcome modeling approach, cost categories, utility approach, committee critiques, modeling assumptions).

* Details of each prompt, for the four prompting strategies, for each of the 10 domains can be found in the Supplemental Materials.

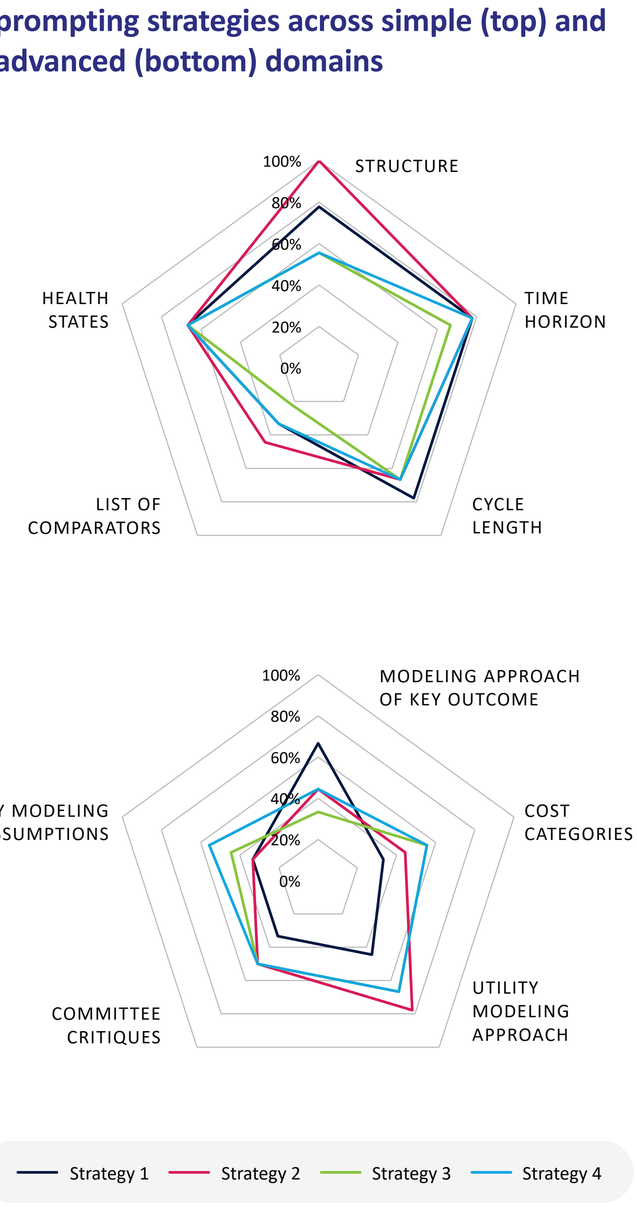
Abbreviations: AI = artificial intelligence; CDA-AMC = Canada's Drug Agency; HTA = health technology assessment; ICER = Institute for Clinical and Economic Review; NICE = National Institute for Health and Care Excellence

Figure 2. Example of prompts used for model structure extraction

- **Prompt 1**: What model structure was used in the cost-effectiveness model?
- **Prompt 2**: In the document's section discussing the model used for the analysis (often but not always titled "Model Structure"), the authors describe the type of model employed. Common health economic models include decision trees, Markov models, microsimulation models, discrete event simulation models, partitioned-survival models, and compartmental models. Can you identify the type of model used in this analysis?
- **Prompt 3**: In the document's section discussing the model used for the analysis (often but not always titled "Model Structure"), the authors describe the type of model employed. Common health economic models include decision trees, Markov models, microsimulation models, discrete event simulation models, partitioned-survival models, and compartmental models. The type of model employed is commonly found at the beginning of the Economic Analysis or Cost-effectiveness Analysis section. Can you identify the type of model used in this analysis?

Strategy 1	Strategy 2	Strategy 3	Strategy 4
Simple prompt	Sequence (chain) of three prompts with increasing complexity	Complex chain-of-thought prompt	Complex chain-of-thought prompt combined with a domain-specific example
Prompt: What model structure was used in the cost-effectiveness model?	**Same as Prompt 3 with the following text added:** Use this format: Q: <repeat_question> A: Let's think step by step. <give_reasoning> Therefore, the answer is <final_answer>.	**Same as Prompt 3 with the following text added:** An example of model structure extraction is: "The cost-effectiveness analysis adopted a Markov cohort model" Use this format: Q: <repeat_question> A: Let's think step by step. <give_reasoning> Therefore, the answer is <final_answer>.	

Figure 3. Comparison of the success of prompting strategies across simple (top) and advanced (bottom) domains



Conclusions

- The sequence (chain) of prompts strategy was the most promising prompting strategy for LLM-assisted CEM data extraction from TAs. However, all prompting strategies showed suboptimal performance in extracting advanced domains.
  - Further research is needed to optimize the health economics and outcomes research (HEOR)-specific prompting strategies and inform best practices for LLM-assisted data extraction, along with LLM continuous evaluation. A standardized database to test and evaluate HEOR-specific automatic extraction will be key, given the fast pace at which updated LLMs are released (e.g., between this study and the poster development an improved LLM, named GPT-4o, has been released).
- Overall, as of 2024, using LLM compared with extraction by an experienced HEOR researcher with average prompting knowledge may not be worthwhile for extracting a standard set of information out of nine TAs. An experienced researcher may complete the task faster and more precisely, considering the time requirement to set up proper prompts and the observed error rate, which suggests that complete checking of the extraction would be required.
  - Nevertheless, this is likely to change with updated LLM releases and HEOR-specific, optimized prompting strategies.

References

1. Lee K, et al. *Informatics in Medicine Unlocked*. 2024;50:101589.
2. Tam TYC, et al. *NPJ Digit Med*. 2024;7(1):258.
3. Tang L, et al. *NPJ Digit Med*. 2023;6(1):158.
4. Ayer T, et al. Generative AI to Identify Comparative Critiques in HTA Reports From Multiple Countries. presented at: ISPOR Europe 2024; November 17-20, 2024, Barcelona. Session HTA326.
5. National Institute For Health and Care Excellence. Use of AI in evidence generation: NICE position statement. 2024. Updated 15 August, 2024. Accessed 21 October, 2024.
6. Polat F, et al. *Semantic Web*. 2024;Preprint:1-34.

Presented at the ISPOR Europe Conference ● 17–20 November 2024 ● Barcelona, Spain

Funding provided by Evidera Ltd., a business unit of PPD, part of Thermo Fisher Scientific