Data Extraction in Literature Reviews Using an Artificial Intelligence Model: Prompt Development and Testing

MSR84

Lunn L,[1] Cross S,[2] Kumar S,[2] Boulton E,[1] Khan A,[3] Magri G,[2] Slater D,[4] Tiwari S,[3] Murton M[2]

[1]Costello Medical, Manchester, UK; [2]Costello Medical, Cambridge, UK; [3]Costello Medical, London, UK; [4]Costello Medical, Remote, UK

Objective

To develop prompts for the extraction of economic data from publications using Artificial Intelligence (AI) and compare the accuracy of AI-extracted data against human-extracted data.

Background

- Growing volumes of published literature make comprehensive literature reviews increasingly resource intensive. AI-based tools have the potential to enhance efficiency in the data extraction process, by using prompts to guide large language models (LLMs) in summarizing data from publications. However, rigorous testing is essential to avoid inaccuracies.

Methods

- A summary of the project approach is presented in **Figure 1**.

Model Selection

- Prompts for two LLMs provided by the OpenAI Application Programming Interface (API) were investigated. The models were GPT-4o and o3-mini; a temperature setting of 1 was used.

Development Phases 1 and 2

- Prompts were iteratively developed alongside a context prompt including the publication text.
  - In Phase 1, various development options (from individual prompts for each data point in the grid to a single prompt for all data) were tested on articles from a single disease area (NSCLC n=3) reporting both economic evaluations (EEs) and cost and/or resource use (CRU) data.
- In Phase 2, the best-performing prompts were then assessed through multiple iterations on articles from different disease areas (ALS EEs: n=7, CRU data: n=4; PsA utility data: n=4). After each iteration, F1 scores (the harmonic mean of precision and recall; score range 0–1) were calculated and the prompts were refined with the aim of achieving an F1 score of ≥0.70.

Testing Phase

- The performance of the prompts from the final iteration of Development Phase 2 was assessed in four articles in NSCLC which reported on EE, CRU and utility data.
- F1 scores for AI-extracted data were compared to the score for human extraction of the same data and performance of the GPT-4o model was compared to the GPT o3-mini. The accuracy of the AI extractions in the test set for each stream was also analyzed in more granular detail (**Figure 1**).

Results

Prompt Structure

- No single approach worked universally; different strategies were effective for different streams, though a single long prompt often yielded better results by returning more data points and capturing greater context.

F1 Scores Accuracy

- The F1 scores for the AI and corresponding human-extractions are summarized in **Figure 2**.
- AI extraction performance for EEs, CRU, and utility data improved across most iterations in Development Phase 2, and good performance was maintained in the Testing Phase, even equalling human standards for the EE set.
- When specifically looking at performance on individual extraction grid sections, more than 80% of the data were extracted accurately in most cases (**Figure 3**).
- Average F1 scores did not vary considerably between GPT-4o and o3-mini models (**Table 1**). The only difference was found for utility studies, with the o3-mini model extracting data that were missed by GPT-4o, resulting in an improved F1 score. However, the time taken to return results was longer with o3-mini.

Key Limitations of AI Extractions

- For EE, errors were commonly observed for complex data e.g., model input sources and rationale for model design/discounting. Extractions from HTA documents were particularly challenging with irrelevant data from the clinical/budget impact sections also being extracted.
- For CRU, distinguishing between costs and resource use was challenging.
- For utilities, issues included missing some utilities (e.g., later time points, placebo arm/subgroups) and calculating absolute utilities using baseline and change from baseline values.

Conclusion

The AI models showed promising results after prompt development and refinement on a small set of articles, especially for simple information. However, current performance is more limited for complex economic data compared to human extractions.

The o3-mini model takes longer to process data and shows comparable performance to GPT-4o; however, its ability to extract additional utility data, indicates potential value in specific contexts.

Further optimization and testing across more articles and disease areas are needed. Models should also be re-evaluated over time as performance improves, alongside testing other models.

FIGURE 1

Summary of project approach

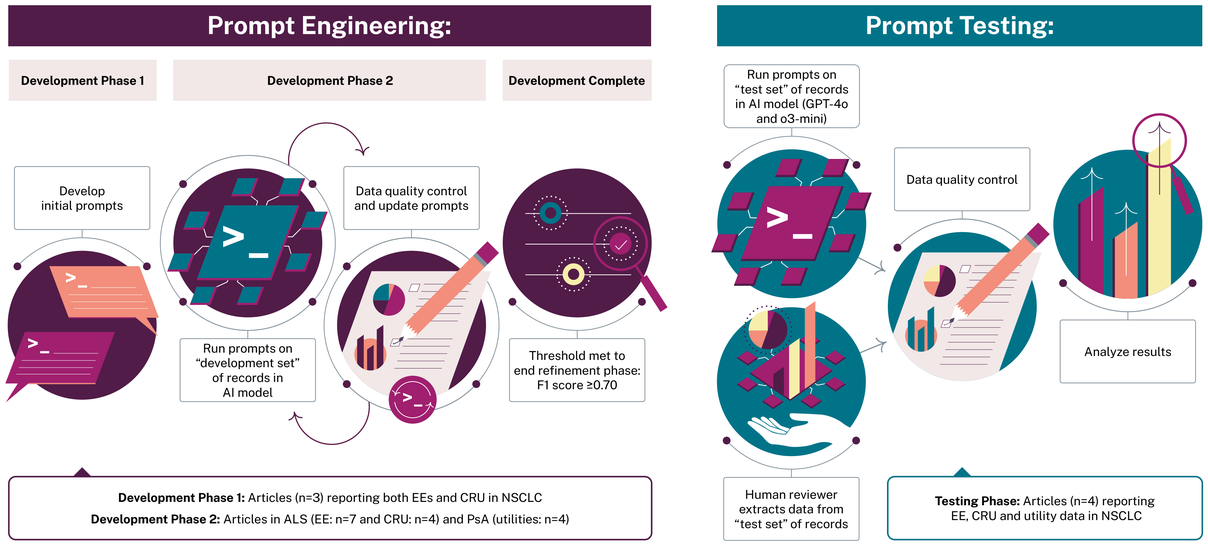
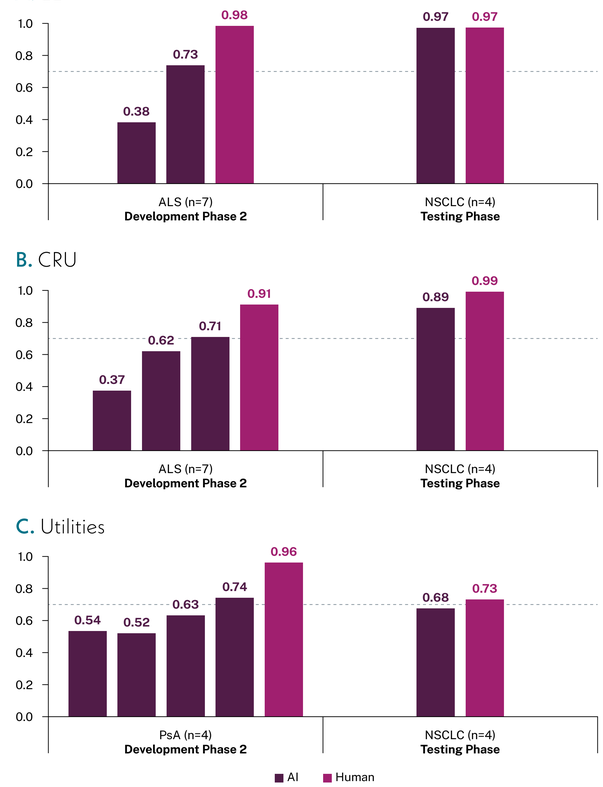


FIGURE 2

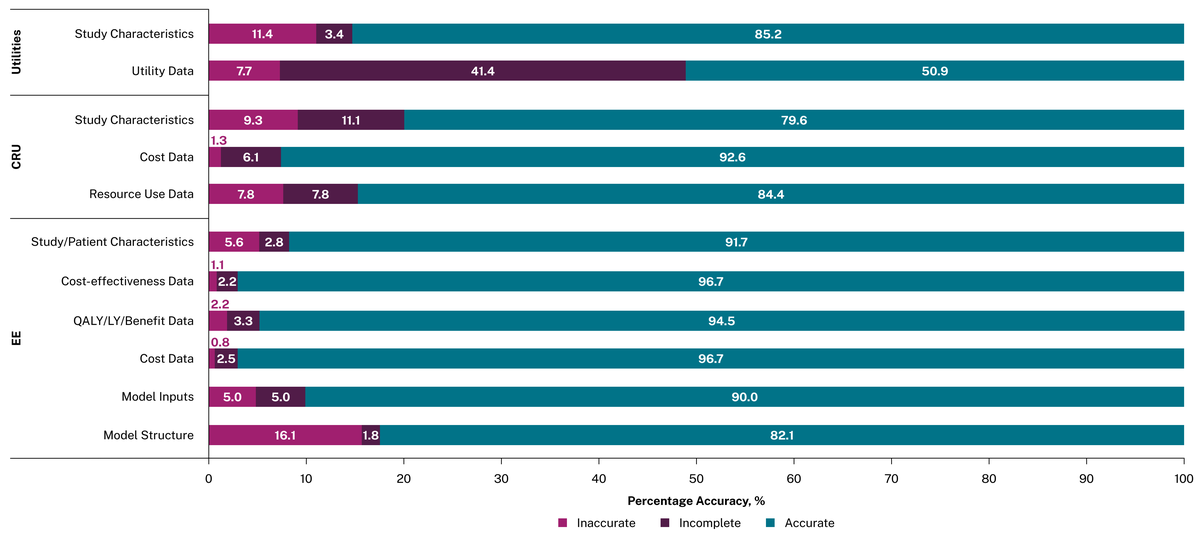
Comparison of F1 scores for AI-extracted and human-extracted data



The dashed line shows the target F1 score threshold of 0.7. Each iteration for the relevant stream is shown as a separate data point in the figure.

FIGURE 3

Granular accuracy of AI-extractions on test set papers, by outcome type (n=4)*



*Four articles from a single disease area (NSCLC). "Incomplete": AI model failed to identify data point that was present in the article; "Inaccurate": AI model extracted data point incorrectly; "Accurate": AI model correctly identified data point. Accuracy was assessed separately for different sections of the extraction grid. Note: Results may not add up to 100 due to value rounding.

TABLE 1

Comparison of GPT-4o vs GPT o3-mini

	GPT-4o		GPT o3-mini	
	F1 score	Time taken (secs)*	F1 score	Time taken (secs)*
EE	0.935	130.94	0.945	308.89
CRU	0.845	40.55	0.845	126.60
Utilities	0.812	18.44	0.917	66.25

A single iteration from the test set in NSCLC (n=2: EE, n=4: CRU/utilities) was used for these comparisons and average F1 scores are presented in the table. *Time taken was calculated based on the running of prompts for a single study for each of the three streams (EE, CRU and utilities).

Abbreviations: AI: artificial intelligence; **ALS:** amyotrophic lateral sclerosis; **API:** Application Programming Interface; **CRU:** cost and/or resource use data; **EE:** economic evaluation; **HTA:** Health Technology Assessment; **LLM:** large language model; **NSCLC:** non-small-cell lung cancer; **PsA:** psoriatic arthritis; **QALY:** quality-adjusted life year; **LY:** life-years.

Acknowledgements: The authors thank Becky Chesworth, Costello Medical, for graphic design assistance.

Presented at ISPOR International 2025 | Montreal, Quebec, Canada | 13–16 May