From GPT-4 to GPT-5.2: A Comparative Evaluation of Large Language Models for Extracting Clinical Real-world Evidence Data

Mariana Farraia[1], Anuja Pandey[2], Kassandra Schaible[3], Caroline von Wilamowitz-Moellendorff[2]

[1]Thermo Fischer Scientific, Ede, Netherlands, [2]Thermo Fischer Scientific, London, UK, [3]Thermo Fisher Scientific, Waltham, MA, USA

MSR33

Background

- Real-world evidence (RWE) is becoming an increasingly important source of information for clinical decision-making across therapeutic areas, but data extraction from published studies remains resource-intensive.[1] Large language models (LLMs) have emerged as promising tools to support data extraction from literature; however, prior evaluations of their accuracy have predominantly focused on randomized controlled trials.[2,3]
- We previously developed a structured data extraction framework for RWE in non-small cell lung cancer (NSCLC) to evaluate the performance of an LLM.[4] This framework was originally implemented using a proprietary Generative Pre-trained Transformer (GPT)-4–based model. As LLMs continue to evolve rapidly, ongoing evaluation across model versions and variants is necessary to assess performance improvements and ensure methodological robustness.

Objectives

- To compare the older GPT-4 model to the newer GPT-5.2 Pro version for structured data extraction of RWE within a standardized extraction framework.

Methods

- A previously developed GPT-4–based extraction framework was replicated using GPT-5.2 Pro.

Publication Set

- The analysis was originally intended to be conducted using the same set of primary studies as in the prior GPT-4 analysis.[4]
  - For each publication included in the original dataset, permissions for artificial intelligence (AI)-based use were systematically assessed.
  - Only two of the 10 primary publications initially used with the GPT-4 model had AI usage rights, and there was no way to obtain the necessary permissions for the other eight; therefore, GPT 5.2 Pro could not be tested on these publications. Both included publications were retrospective cohort studies.

Model Replication and Extraction Process

- To ensure comparability and isolate model performance, the same prompts (one-shot prompts), extraction instructions, and data extraction template (DET) structures were applied.
  - No prompt engineering modifications were introduced when transitioning between models.
- Each publication was processed using GPT-5.2 Pro, and outputs were captured in the same structured DET format used in the original GPT-4 analysis, enabling direct, cell-level comparison.
  - Extracted variables included study characteristics, population characteristics, efficacy and safety outcomes, and subgroup efficacy outcomes.[4]

Output Validation and Comparison

- Outputs generated by GPT-4 and GPT-5.2 Pro were systematically reviewed/validated by an experienced human reviewer. Each extracted data point was classified into one of four categories:

Correct (accurately extracted)

Incorrect (factually wrong)

Missing (reported in publication but not extracted)

Incomplete (partially extracted or affected by formatting/truncation relevant issues)

Performance Metrics

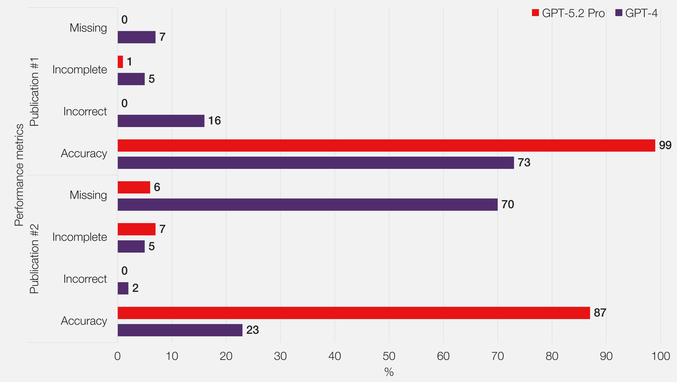
- Consistent with prior analyses,[3,5] quantitative performance was assessed at the section level, including study characteristics, patient characteristics, and outcomes. Overall accuracy was calculated as:

$$Accuracy\ (\%) = \frac{\text{(Number of correctly extracted variables)}}{\text{(Total number of expected variables)}} \times 100$$

Results

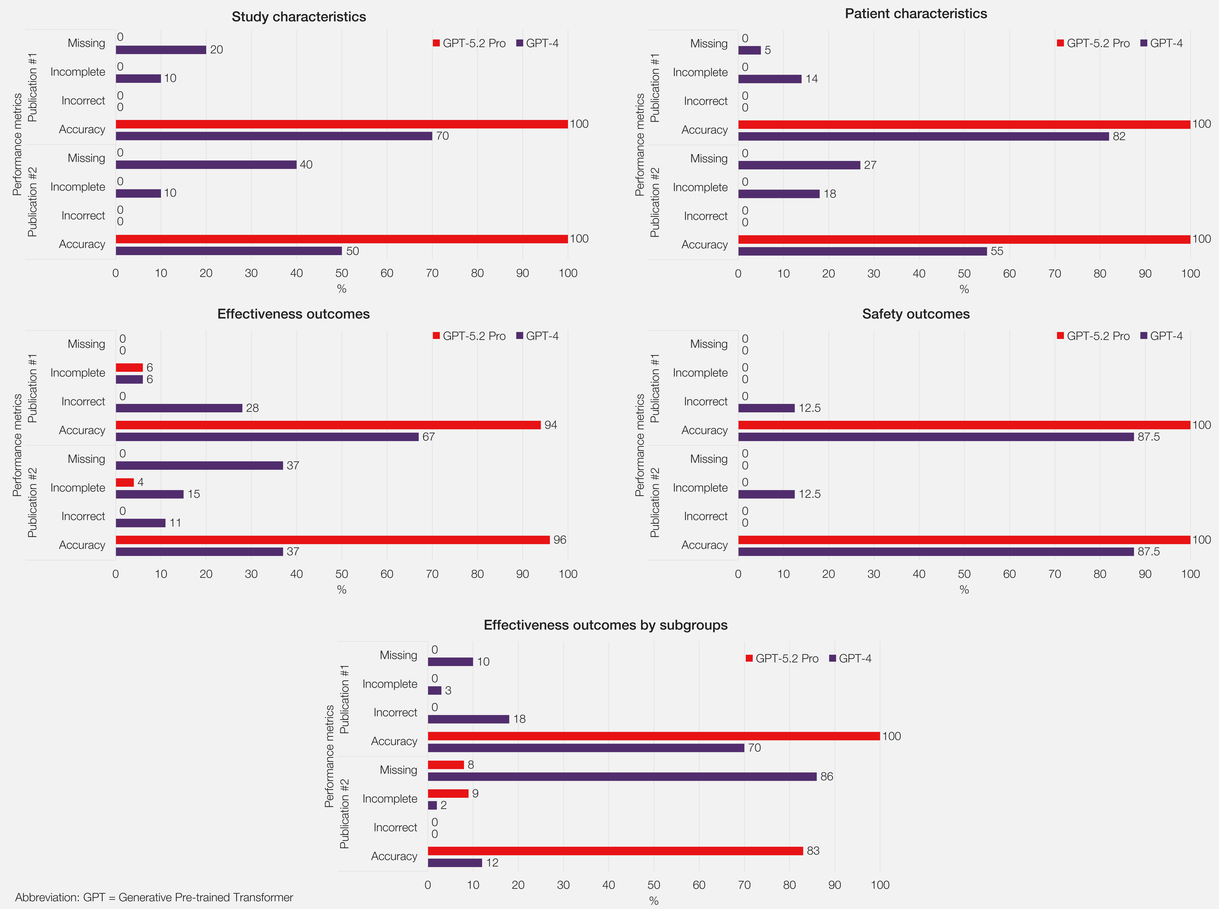
- Data extraction using the GPT 5.2 Pro model generally showed higher accuracy overall compared with the GPT-4 model (**Figure 1**). Overall accuracy for Publication #1 and Publication #2 was 99% and 87% with GPT 5.2 Pro compared with 73% and 23% with GPT-4, respectively.
  - Across all individual sections (**Figure 2**) in both publications (study characteristics, patient characteristics, effectiveness outcomes, safety outcomes, and effectiveness outcomes by subgroups), GPT 5.2 Pro performed better than GPT-4, with accuracy being consistently higher. In contrast, missing, incomplete, or incorrect extractions were lower with GPT 5.2 compared with GPT-4.
  - Incomplete data were observed for effectiveness outcomes when using GPT 5.2 Pro in both publications; however, this was the same or lower than incomplete data when using GPT-4.
  - When considering effectiveness outcomes by subgroups, GPT 5.2 Pro achieved 100% accuracy for Publication #1 compared with 70% using GPT-4. For Publication #2, accuracy was 83% with GPT 5.2 Pro vs. 12% with GPT-4. The considerable improvement was due to GPT 5.2 correctly identifying where specified outcomes were not reported in the publication for patient subgroups (these patient subgroups were available in the patient characteristics and other outcomes were reported for them within the publication, but they were not targeted for these specific extractions).

Figure 1. Overall Accuracy of LLM-based Data Extraction Using GPT-4 vs. GPT-5.2 Pro, Including the Percentage of Incorrect, Incomplete, and Missing Data



Abbreviation: GPT = Generative Pre-trained Transformer

Figure 2. Accuracy of LLM-based data Extraction Using GPT-4 vs. GPT-5.2 Pro, Including the Percentage of Incorrect, Incomplete, and Missing Data, for Each Section Level (Study Characteristics, Patient Characteristics, Effectiveness Outcomes, Safety Outcomes, and Effectiveness Outcomes Stratified by Subgroups)



Abbreviation: GPT = Generative Pre-trained Transformer

Discussion

Summary of Findings

- GPT-5.2 Pro outperformed GPT-4 in data extraction, achieving higher accuracy by reducing the percentage of incorrect, incomplete, and missing outputs. GPT-5.2 Pro was particularly accurate when extracting study and patient characteristics and safety outcomes.
- Nevertheless, human validation remains necessary, as GPT-5.2 Pro still failed to extract some relevant data, without any noticeable trend of why this data was missing. For example, when extracting subgroup data for Publication #2, GPT-5.2 Pro extracted data for male patients but not for the female subgroup. In Publication #1 this was not an issue.

Limitations

- Due to limited testing on only two of 10 NSCLC RWE publications, results should be interpreted with caution.
  - Both studies compared two treatment regimens (closer to randomized controlled trial-like designs) with a retrospective study design that may not reflect other real-world project settings.
  - In both cases, the study population fully matched the population, intervention/comparator, outcomes, study design (PICOS) criteria (NSCLC, programmed death-ligand 1 ≥50%), so findings do not address scenarios where the PICOS group is a subpopulation—a known challenge from prior work.[4]
- The objective of this work was a controlled replication of a previously used framework; therefore, prompts were not optimized for GPT-5.2 Pro, and results may not reflect maximum achievable performance under prompt engineering refinement.
- Outputs were compared against the publication by an experienced human reviewer and variable-level classification (e.g., incomplete vs. incorrect) may retain some degree of subjective interpretation. We did not perform a comparison of accuracy between a dataset fully extracted by a human vs. LLM-based data extraction.
- Performance differences were not statistically tested, and findings should be interpreted as exploratory.

Future Research

- This work focused on structured data extraction from published manuscripts; performance in other document types (e.g., regulatory submissions, registry datasets) was not assessed.
- Time performance was not formally evaluated; however, GPT-5.2 Pro appeared to take more time to process the prompts (i.e., slower) compared with GPT-4. Future work should quantitatively assess time metrics of using LLMs for data extraction and assess other ChatGPT model variants (e.g., Thinking) and alternative LLMs.

AI Legal Rights

- The main challenge in this work was related to obtaining the necessary permissions to use AI on full-text publications as recommended by a recent position statement from Cochrane.[6]
  - As a result, AI could only be applied to two of 10 papers, reflecting current legal and operational constraints.
  - Publisher policies (e.g., Elsevier) prohibit uploading full-text content into AI systems without explicit permissions, as this constitutes text and data mining and copying. Access to publications (including via purchase) does not grant rights for AI-based analysis.
  - The legal landscape around AI and copyright is still evolving, and clearer guidance is needed to enable broader and compliant use of LLMs in future evidence synthesis research, ensuring alignment with RAISE (Responsible AI in Evidence Synthesis) recommendations.[7]

References

1. ussbaumer-Streit B, et al. *J Clin Epidemiol*. 2021;139:287-296.
2. Gartlehner G, et al. *Res Synth Methods*. 2024;15(4):576-589.
3. Shree A, et al. *Value Health*. 2024;27(12):S475.
4. Farraia M, et al. *Value Health*. 2025;28(12):S666.
5. Shree A, et al. *Value Health*. 2025;28(6):S295.
6. Flemyng E, et al. *Cochrane Database of Systematic Reviews*. 2025;(10).
7. Thomas J, et al. Responsible use of AI in evidence SynthEsis (RAISE): recommendations and guidance. 2024. Updated March 14, 2026. Accessed April 21, 2026. https://osf.io/fwaud/overview

Disclosures

This study was supported by Thermo Fisher Scientific. **MF**, **AP**, **CvWM**, and **KS** are employees of PPD™ Evidera™ Health Economics & Market Access, Thermo Fisher Scientific.

Acknowledgments

Editorial and graphic design support were provided by Michael Grossi and Richard Leason of Thermo Fisher Scientific.