MSR92

Can artificial intelligence (AI) large language models (LLMs) such as generative pre-trained transformers (GPT) be used to automate literature reviews?

*Ines Guerra[1], Julia Gallinaro[1], **Ketevan Rtveladze[1]***, Alexandrina Lambova[2], Elia Asenova[2]*

*[1]IQVIA, London, UK, [2]IQVIA, Sofia, Bulgaria, *presenting author*

BACKGROUND AND OVERALL METHODOLOGY

- Systematic literature reviews (SLR) are commonly required to support market access activities for new products.
- The SLR process involves multiple time-consuming, and potentially error-prone, steps such as publication screening, data extraction and reporting.
- We implemented an algorithm that used AI LLMs, such as GPT[1,2], to generate the first version of the clinical data extraction file in Excel® for a SLR.
- We assessed performance by measuring the accuracy of the GPT-based extraction compared to a manual extraction performed by humans (benchmark) and with a human quality check performed on the GPT outputs (QC).
- The project was divided into two phases. In Phase 1, variable extraction was performed using a GPT-3 based algorithm. During Phase 2, the algorithm was adapted to use the most recent GPT version (GPT-4).

PHASE 1: GPT-3

Methods

- We devised a pipeline for variable extraction using GPT-3 (**Figure 1**).
- We extracted variables related to "Study details" from 23 papers and measured the accuracy of the GPT-based extraction compared to a manual extraction performed by humans (benchmark).
- For the variables that consisted of free-flow text, accuracy was estimated using BERTScore[3], an evaluation metric for text generation.
- We then iteratively engineered parts of the GPT-based extraction algorithm and re-evaluated performance for selected variables with poor performance.

Results

- For the measured variables, the accuracy of extraction with the pre-engineered version of the algorithm ranged from 17% to 100% (**Figure 2**).
- By iteratively engineering the GPT-based algorithm, the extraction accuracy was improved for variables where AI initially had low performance. For example, for patient inclusion criteria accuracy increased from 40% to 70%, and for patient exclusion criteria the accuracy increased from 35% to 80%, across the studies (**Figure 2**).
- Using the new version of the algorithm, we extracted the "Study details" variables from a new set of 26 papers and quality check (QC) was performed by human researchers (**Figure 3, left**).

Conclusion

- Iteratively engineering the extraction algorithm can lead to better performance of data extraction.

PHASE 2: GPT-4

Methods

- Using GPT-4, we extended the previous approach and generated the first draft of the clinical data extraction file in Excel® for 7 papers, including information on the study details, patient characteristics, interventions and the study outcomes (**Figure 4**).
- We evaluated performance with a human QC.

Results

- 497 variables were extracted from each paper on average at 22 minutes per paper.
- The performance was worst for variables that are normally found in tables (e.g., patient characteristics, **Figure 5**).
- Compared to the previous phase, we observed improvement in overall accuracy for "Study details" related variables (**Figure 3, right**).

Conclusion

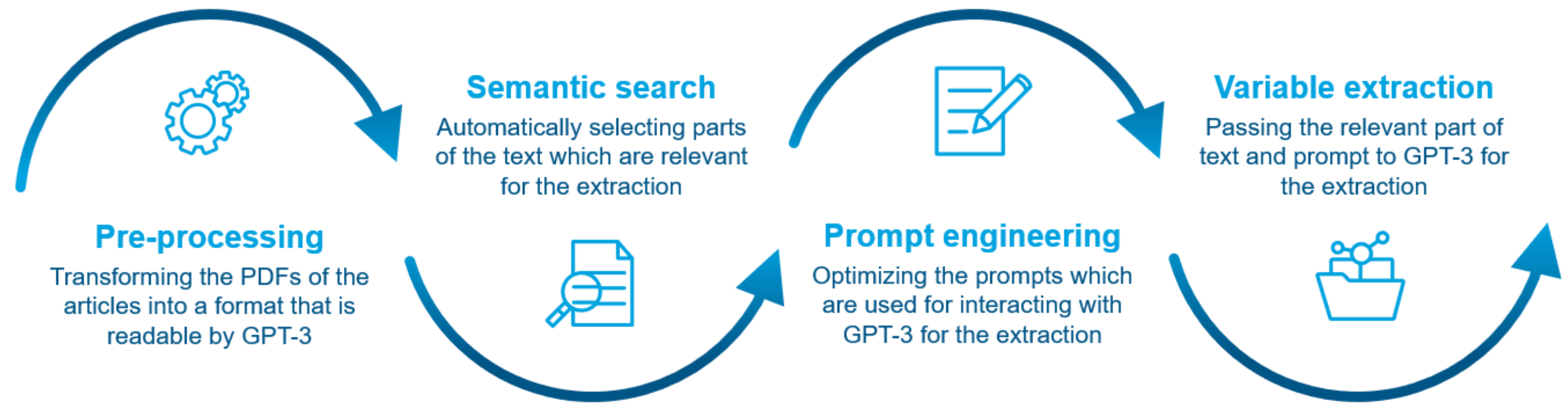
- Improvements were observed with the upgrade of GPT version, specially for variables found in free-flow text.
- Better performance for extracting information from free-flow text compared to tables.

SUMMARY AND CONCLUSION

These results suggest that AI LLMs such as GPT, in conjunction with iterative algorithm engineering, could be used for generating first version of extraction file with good accuracy.

The current advancements in technology are expected to improve accuracy further.

Provided a human subject matter expert undertakes a QC, these tools could provide more efficient data extraction versus manual, human-only extraction.



Pre-processing

Transforming the PDFs of the articles into a format that is readable by GPT-3

Semantic search

Automatically selecting parts of the text which are relevant for the extraction

Prompt engineering

Optimizing the prompts which are used for interacting with GPT-3 for the extraction

Variable extraction

Passing the relevant part of text and prompt to GPT-3 for the extraction

Figure 1: Overview of steps for extracting variables from scientific papers using GPT-3, including semantic search and prompt engineering.

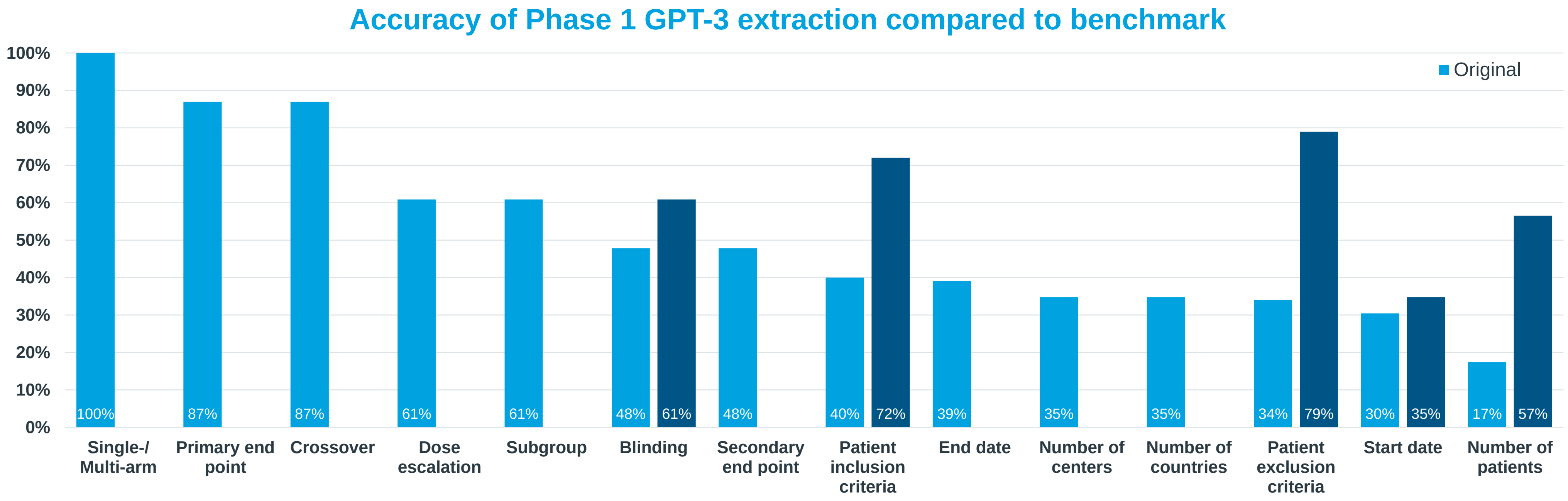


Figure 2: Accuracy of Phase 1 GPT-3 extraction compared to human benchmark. For yes/no or numeric variables, accuracy is given as percentage of correct variables extracted by GPT-3. For free-flow text variables, accuracy is given as BERTScore[3]. Light blue bars indicate accuracy. For some variables, prompt engineering was performed to try to improve accuracy. Dark blue bars indicate accuracy after prompt engineering.

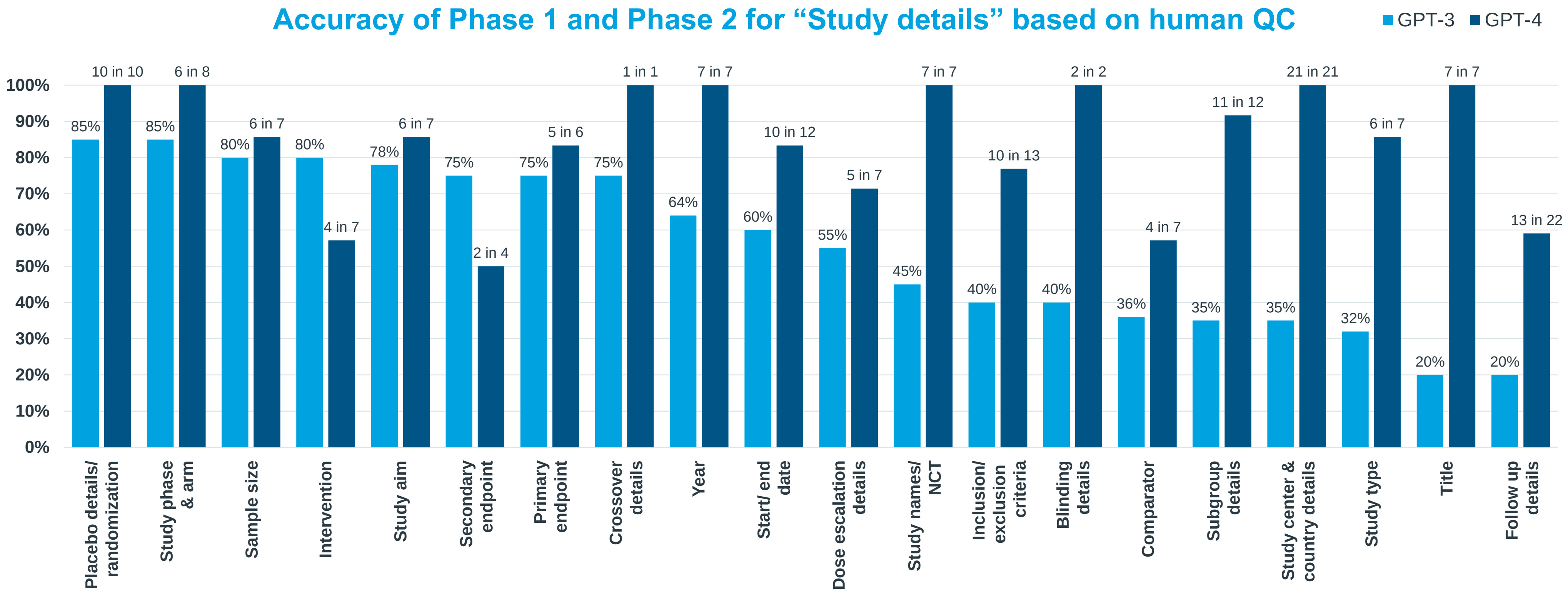
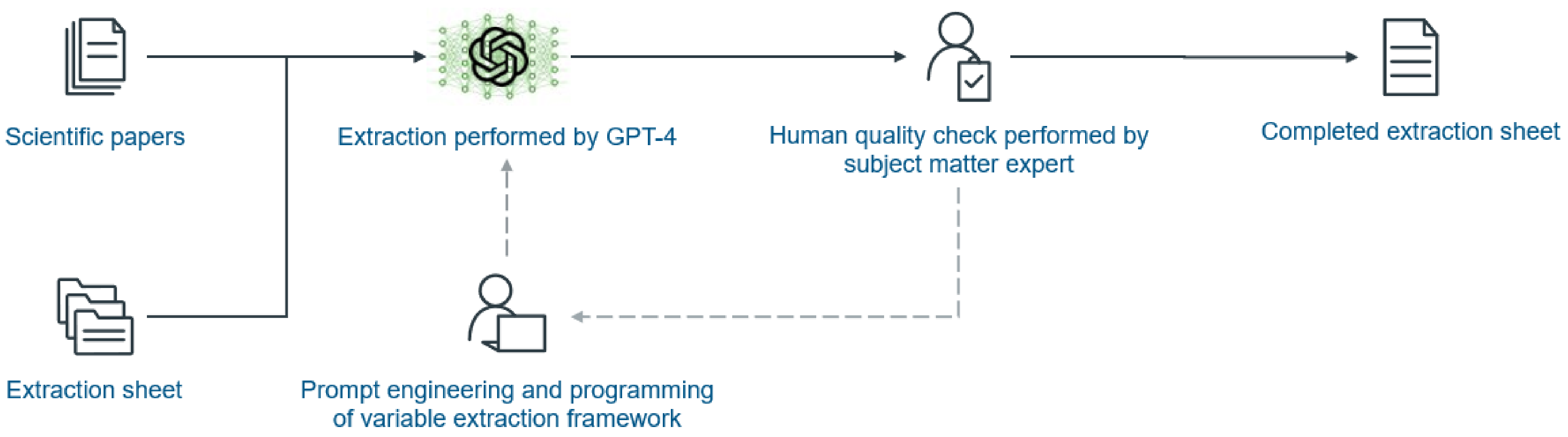


Figure 3: Accuracy of Phase 1 GPT-3 and Phase 2 GPT-4 extraction of "Study details" variables based on human QC. Accuracy is given as percentage of variables extracted by GPT-3 assigned as correct during the QC (*left, charts*) and number of variables extracted by GPT-4 assigned as correct during the QC *(right, tables*).



Scientific papers

Extraction performed by GPT-4

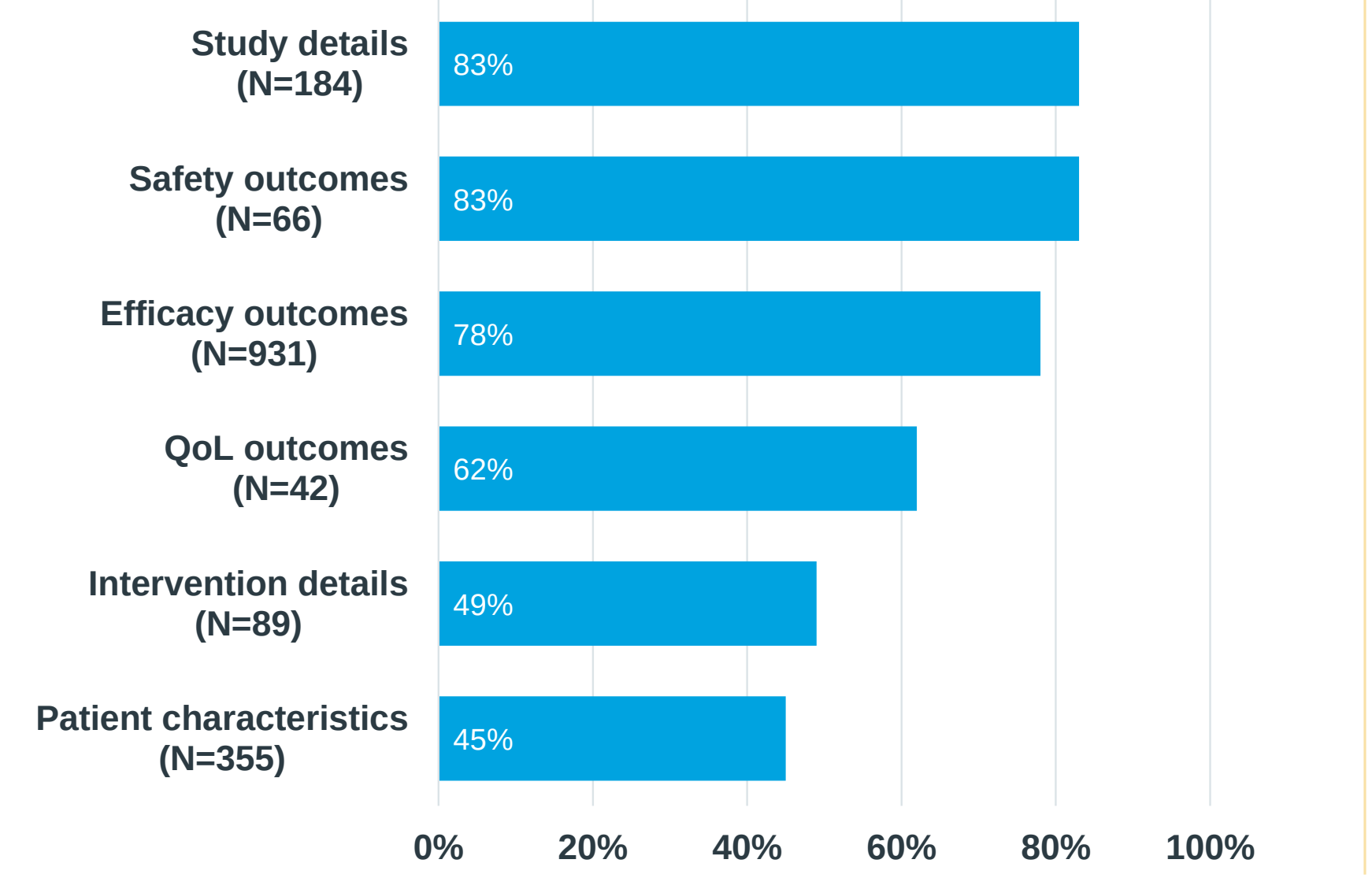
Human quality check performed by subject matter expert

Completed extraction sheet

Extraction sheet

Prompt engineering and programming of variable extraction framework

Figure 4: Overview of steps for extracting variables from scientific papers using GPT-4 in Phase 2. Dotted grey lines indicate steps to be performed only during optimisation of the algorithm. Solid dark lines indicate steps performed during optimisation and application of the optimised algorithm for extraction.



High performance	
Variable	**Accuracy**
Unit of follow up duration	30 in 30
Description of event free survival	7 in 7
Definition of FLT3-ITD positive	8 in 8
Data cut-off	5 in 5
Number of patients assigned to treatment	11 in 12

Low performance	
Variable	**Accuracy**
Intervention dosing frequency	5 in 9
Route of administration	3 in 6
Median of event free survival duration	1 in 4
Number of patients with complete response	2 in 8
Definition of transplantation rate	1 in 8

Figure 5: Accuracy of Phase 2 GPT-4 extraction based on human QC. Accuracy is given as percentage of variables that are not NA (not applicable) / NR (not reported) and that were assigned as correct during QC. *Left, chart:* Average accuracy across all papers and all variables by topic (Accuracy) and number of not NA/NR variables across all papers per topic (N). *Right, tables:* Example of variables with high and low accuracy.

References

1. Radford A, Narasimhan K, Salimans T, Sutskever I. "Improving Language Understanding by Generative Pre-Training." (2018).
2. Brown TB, Mann B, Ryder N, Subbiah M et al. "Language Models are Few-Shot Learners." *Advances in neural information processing systems* 33 (2020): 1877-1901.
3. Zhang T, Kishore V, Wu F, Weinberger KQ, Artzi Y. "BERTScore: Evaluating Text Generation with BERT." *arXiv preprint arXiv: 1904.09675* (2019).

ISPOR Europe 2023. 12-15 November 2023. Copenhagen, Denmark.