Heva
A subsidiary of DOCAPOSTE

Fanny RAGUIDEAU, Anthony PIEROTTI, Daniel PODEVIN, Arnaud PANES, Martin PRODEL, Ludovic LAMARSALLE

Heva, Lyon, Fance

40th ICPE
August 24-28, 2024 - Berlin, Germany
Poster n°ASPOT15

Automatisaton of study report abstract generation by Large Language Models

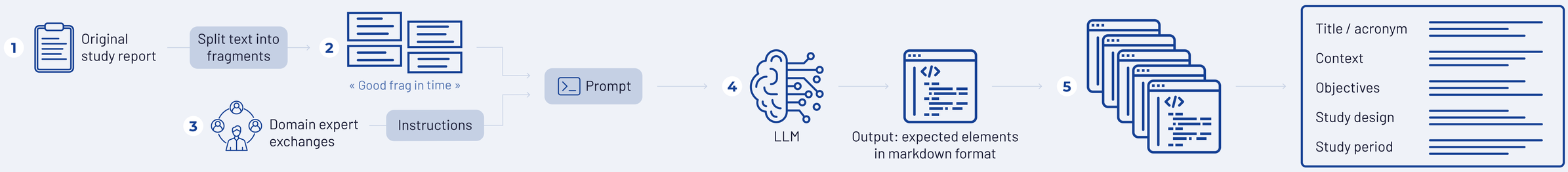
Methodology and application example

BACKGROUND

Since 2022, ChatGPT by OpenAI and Large Language Models (LLMs) have been used in scientific and academic fields. It seems a promising tool to facilitate thorough literature reviews and summarizing scientific contents (e.g. scientific articles, patient records, etc.).

Our objective was to assess the feasibility of automatically generating clinical study report abstracts testing several LLMs methods.

METHODOLOGY



Using a semi-autonomous approach

1 A dataset of eight clinical study reports (CSR) was used. It was a representative sample of pharmaco-epidemiological studies conducted on primary or secondary data, with various study designs and various objectives (description of patients, epidemiological assessment, survival studies, pharmacoeconomic evaluation).

2 We explored a full methodology to tackle the challenges of automatic summarizing. First, MapReduce was used to handle large documents (tens of pages).

3 Then, an abstract template containing the sections to be respected was supplied to the LLM. For each pre-defined section, instructions detailing the purpose and practical information (number of characters to be respected, whether it was necessary to transliterate certain parts without modifying the content, etc.) was specified.

4 The technical implementation was done in python 3.12 language on LangChain framework, using one of possible LLM OpenAI's ChatGPT (**3.5 Turbo**, **4 Turbo**, **4o**) and MistralAI (**Large**, **Open-Mixtral-8x22b**) in a private instance (Heva's Tenants OpenAI and MistralAI).

5 To obtain an abstract, we built a transformation chain using the study report as a starting point and a template of the expected parts of the abstract. Each study report was subdivided into labeled parts according to the sections that make it up and the information it contains. The generated abstract is built by filling in a template using the Heva theme.

How do you evaluate a text created by an LLM?

Several quality indicators for the generated abstract were calculated, by comparing it with a manually written abstract (at the time the report was drafted).

We used 3 ways

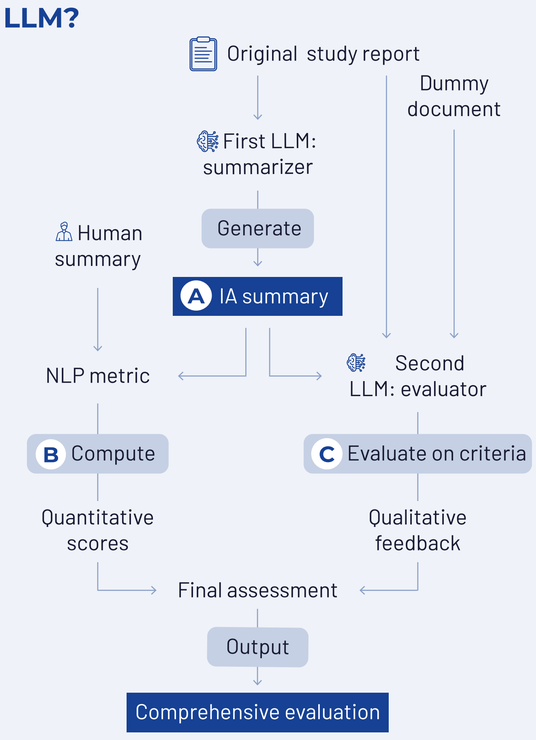
A By an expert who proofreads

B With a NLP metric: comparison with the "human-written" summary

C With a 2nd LLM based on specific instructions: evaluate 4 metrics (from 0 to 10):

- **Relevance**: redundancies and excess information are penalized
- **Coherence**
- **Consistency**: hallucinations are penalized
- **Fluency**

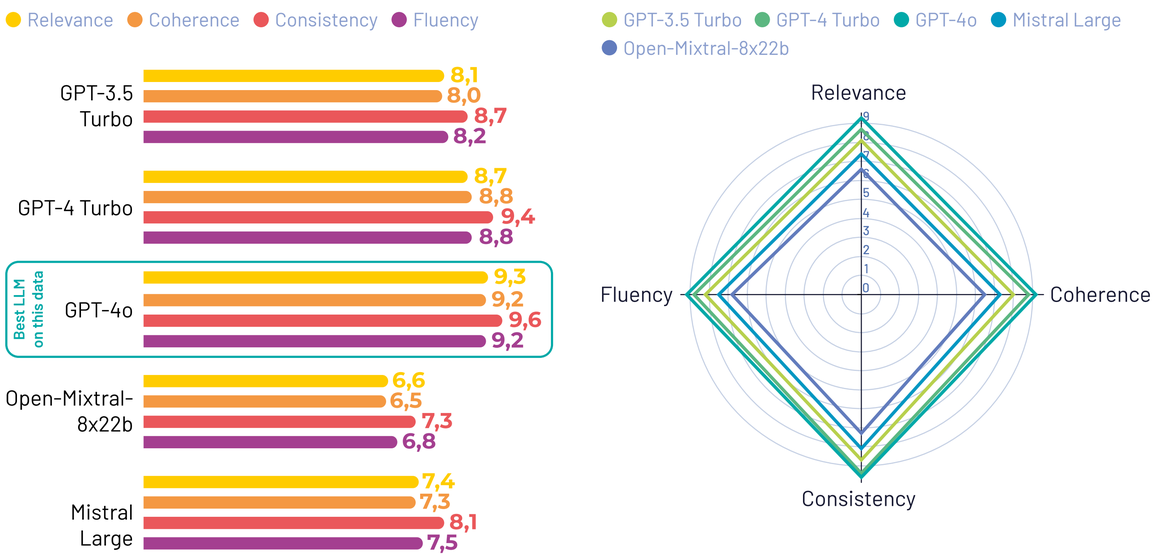
Metrics were validated with a dummy document.



RESULTS

On 8 reports

Summary evaluation by a 2nd LLM



→ High quality of summaries is achievable with LLM (≈9 /10)
→ The best performing LLM on this specific task is chatGPT-4o

Other metrics

	GPT-3.5 Turbo	GPT-4 Turbo	GPT-4o	Mistral Large	Open-Mixtral-8x22b
Duration	5 min	6 min	5 min	7 min	4 min
Cost per summary	0,5€	3€	2€	1€	1€
Number of input* tokens	70,000 (for all models)				
Number of output* tokens	114	104	185	175	147

* GPT tokenizer cl100k_base was used.

On a single report

	GPT-3.5 Turbo	GPT-4 Turbo	GPT-4o	Mistral Large	Open-Mixtral-8x22b	Dummy Document	Human summary
LLM summary vs. the « human-written » summary							
NLP: cosine similarity	94%	94%	94%	93%	94%	1%	
LLM summary vs. the original full report							
Relevance	8	8	9	7	6	2	10
Coherence	8	8	9	7	6	2	9
Consistency	9	9	10	8	7	1	10
Fluency	8	8	9	7	6	4	9
Mean	8.3	8.3	9.3	7.3	6.3	2.3	9.5

Explanation provided by the LLM when evaluating the summary of "chatGPT4o"

Relevance (Score: 9)
The summary excels in capturing the main points of the original report, ensuring that all critical elements are included. The summary covers the context and rationale, study objectives, design, data sources, outcomes, and results effectively. However, to achieve a perfect score, it could be slightly more concise by eliminating minor redundant details.

Coherence (Score: 9)
The summary demonstrates strong coherence, presenting information in a logical and well-organized manner. It follows a clear sequence, starting with the context and rationale, moving through the objectives, study design, data sources, outcomes, and statistical analyses, and concluding with the results and implications. Each section flows smoothly into the next, making it easy for readers to follow the narrative.

Consistency (Score: 10)
The GPT-4 summary is factually consistent with the original report, accurately reflecting all key points, data, and conclusions without any errors or discrepancies. It ensures that the findings and statistics mentioned in the summary are directly supported by the original document.

Fluency (Score: 9)
The summary is written in clear and grammatically correct language, making it easy to read and understand. It uses appropriate terminology and sentence structures that suit the academic and technical nature of the content. There are very few minor grammatical errors, and the language is precise and professional.

CONCLUSION

LLMs appear to be promising tools for producing summaries of clinical contents allowing to focus on tasks requiring significant human intellectual input, thereby saving valuable time in the research process. Human proofreading of LLM-generated summaries remains essential to ensure final consistency.