MSR110

Evaluating Large Language Models' (LLM) Performance in Content Generation for Global Value Dossiers (GVD)

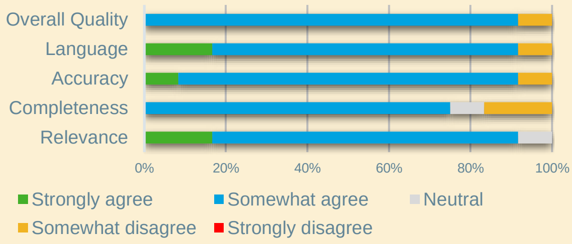
Jessica Walters[1], Ketevan Rtveladze[1], Weiwei Xu[2], Natalie Green[1], Joe Joseph[1], Kiril Matev[3], Julia Gallinaro[1], Inês Guerra[1]

[1]IQVIA, London, United Kingdom; [2]IQVIA, Amsterdam, Netherlands; [3]IQVIA, Sofia, Bulgaria

OBJECTIVES

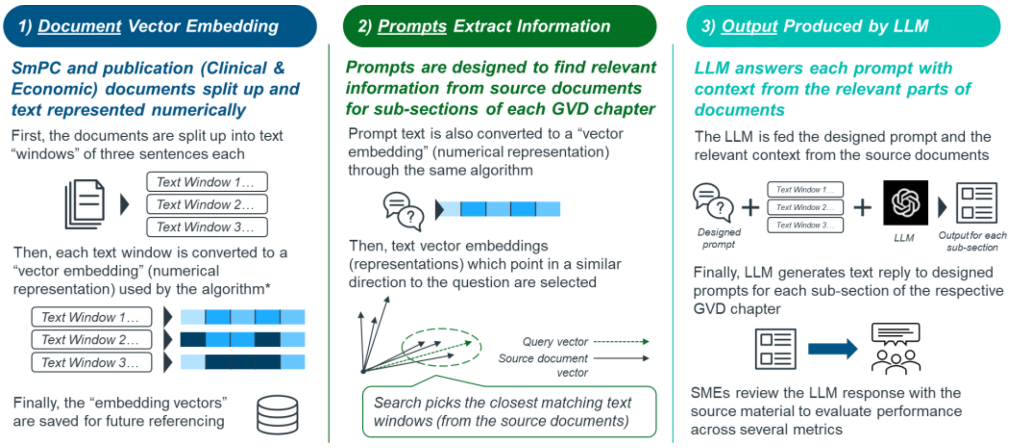
- Global Value Dossiers (GVDs) provide a summary of a medical product's value story, underpinned by up-to-date, comprehensive evidence.
- Multiple chapters need to be developed when creating a GVD, including Disease background, Burden of disease, Treatment landscape, Unmet need, Product profile (PP), Clinical value, and Economic value. Drafting the content for these chapters can be time consuming, leading to inefficiencies.
- Large Language Models (LLMs) have the capability to revolutionise content generation, potentially attaining a degree of expertise that could parallel human proficiency. This study investigated the practicality and effectiveness of LLMs in generating content for GVDs.

Figure 1. SME evaluation of LLM generated GVD content



METHODS

- Subject Matter Experts (SMEs) in GVDs used an IQVIA-built LLM Retrieval Augmented Generation (RAG) application to generate content for PP, Clinical, and Economic GVD chapters for four drugs.
- The drugs were chosen to cover diverse therapeutic areas.
- Publicly available sources were used for chapter development; the European Medicines Agency Summary of Product Characteristics (SmPC) and journal publications on clinical trials and cost-effectiveness models (CEMs) were used to generate PP, Clinical, and Economic content, respectively.
- LLM performance at generating the required output was evaluated by SMEs across several dimensions, including relevance, completeness, accuracy, language quality, and overall quality, using a five-point Likert Scale.



1) *Document* Vector Embedding

SmPC and publication (Clinical & Economic) documents split up and text represented numerically

First, the documents are split up into text "windows" of three sentences each

Then, each text window is converted to a "vector embedding" (numerical representation) used by the algorithm*

Finally, the "embedding vectors" are saved for future referencing

2) *Prompts* Extract Information

Prompts are designed to find relevant information from source documents for sub-sections of each GVD chapter

Prompt text is also converted to a "vector embedding" (numerical representation) through the same algorithm

Then, text vector embeddings (representations) which point in a similar direction to the question are selected

Search picks the closest matching text windows (from the source documents)

3) *Output* Produced by LLM

LLM answers each prompt with context from the relevant parts of documents

The LLM is fed the designed prompt and the relevant context from the source documents

Finally, LLM generates text reply to designed prompts for each sub-section of the respective GVD chapter

SMEs review the LLM response with the source material to evaluate performance across several metrics

***Text vector embeddings** *are generated by a neural network that is trained to create similar vectors for sentences with similar meaning*

RESULTS

- Overall, SMEs agreed that the generated content was relevant and accurate (92% strongly or somewhat agreed; **Figure 1**).
- The responses were largely complete (75% somewhat agreed; **Figure 1**) but could be further improved by LLM prompt editing and RAG improvements.
- In all chapters there were instances of language being overly simple, repetition throughout sub-sections and some hallucinations, however SMEs concurred that the responses were mostly well-written (92% strongly or somewhat agreed; **Figure 1**).
- SMEs agreed that the overall quality of the generated responses was good enough to incorporate into de novo GVD development (92% somewhat agreed; **Figure 1**), albeit with manual checking and editing required.
- Responses to all metrics were most positive on the Clinical chapter (100% strongly or somewhat agreed; **Figure 2**), followed by PP chapter (95% strongly or somewhat agreed; **Figure 3**) then the Economic chapter (70% somewhat agreed; **Figure 4**).
- Other notable findings included:
  - Restricted information within publications impacted the quality of responses but was improved by providing the LLM with supplementary material, when available.
  - Varying performance was observed for Economic responses due to variations of CEM design reported in each publication, meaning the differing content was not aligning with previously designed prompts.
  - Multiple indications within the drug label caused issues with incomplete or incorrect information within the PP section.

Figure 2. SME responses on Clinical

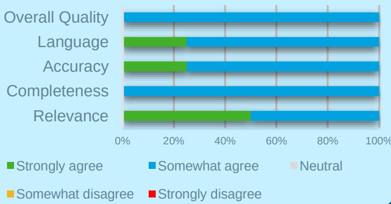


Figure 3. SME responses on PP

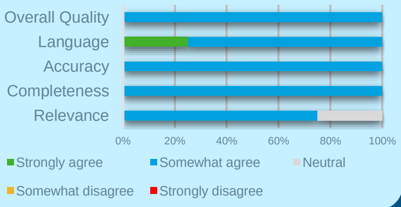
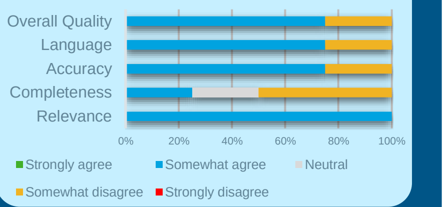


Figure 4. SME responses on Economic



CONCLUSIONS

- **This study demonstrated that LLMs can significantly aid in GVD content generation, when drafting PP, Clinical and Economic chapters.**
- **However, performance is contingent on the level of detail provided in the instructions to the LLM and within the source material.**
- **SME review is essential to ensure accuracy and completeness of the generated output.**
- **Further testing and refining of prompts, as well as evaluating LLM performance at generating other chapters of GVDs, is planned.**

> "The LLM was generally good at answering each sub-section and capturing the relevant details, however some sections have overlapping context thus the LLM is repeating information. Some prompts could be improved to include more of the necessary information and help avoid the repetition."

> "The LLM output is close to what is needed for the first draft; would still need a human to check to ensure no important detail is missing. In places the language appears a little lay, but this could be easily edited."

> "The information in the publication is really limited for the level of detail we need so quite a few sections were unanswered. There are a couple instances of inaccurate information and hallucinations in the LLM response, however generally fine considering the source document. Human is required to identify and input relevant tables and figures from the source, which is currently not done by the LLM."

Abbreviations: CEM: Cost-effectiveness model; GVD: Global Value Dossier; LLM: Large Language Model; PP: Product profile; RAG: Retrieval Augmented Generation; SME: Subject Matter Expert; SmPC: Summary of Product Characteristics.

Correspondence: Inês Guerra (Ines.Guerra@iqvia.com)

International Society for Pharmacoeconomics and Outcomes Research Europe 2024, November 17-20, 2024, Barcelona, Spain