EQ-5D-5L Value Set Retrieval Using Large Language Models

Ivan Au, The University of Hong Kong
Carlos Wong, The University of Hong Kong, Laboratory of Data Discovery for Health, London School of Hygiene & Tropical Medicine

Aim: To evaluate the performance of common large language models (LLMs) using the retrieval-augmented generation (RAG) technique in answering queries related to EQ-5D-5L utility values

Introduction:

- With the recent advancements in LLMs, there is a potential to develop an AI-powered domain-specific chatbot capable of answering EQ-5D related questions from researchers and end-users.
- These questions may include specific utility values, the choice of value set, questionnaire version selection, and findings from published studies.
- RAG is a technique that enhances LLMs by enabling access to external databases (e.g., publications, documents, and datasets).
- This approach is particularly suitable for accessing value sets for various EQ-5D instruments across different countries and regions.

Methods:

- The RAG approach was applied to five LLMs: Llama 3.2-3B, Llama 3.1-8B, Gemma 2-9B, Mistral-7B, and Qwen 2.5-7B, using the Hong Kong EQ-5D-5L value set as the reference database.
- Two types of queries were designed to assess the capabilities of LLMs with RAG :
  - One-state queries: "*What is the utility value for health state X using the Hong Kong value set?*" (where X represents an EQ-5D-5L health state). All 3,125 health states were assessed.
  - Two-state queries: "*What are the utility values for health states X1 and X2 using the Hong Kong value set?*". A total of 1,000 pairs of randomly selected health states were assessed.
- All queries were processed by the five LLMs locally using the Python package Ollama.
- Accuracy was measured as the proportion of responses with correct utility value(s) reported.

Figure. Study workflow

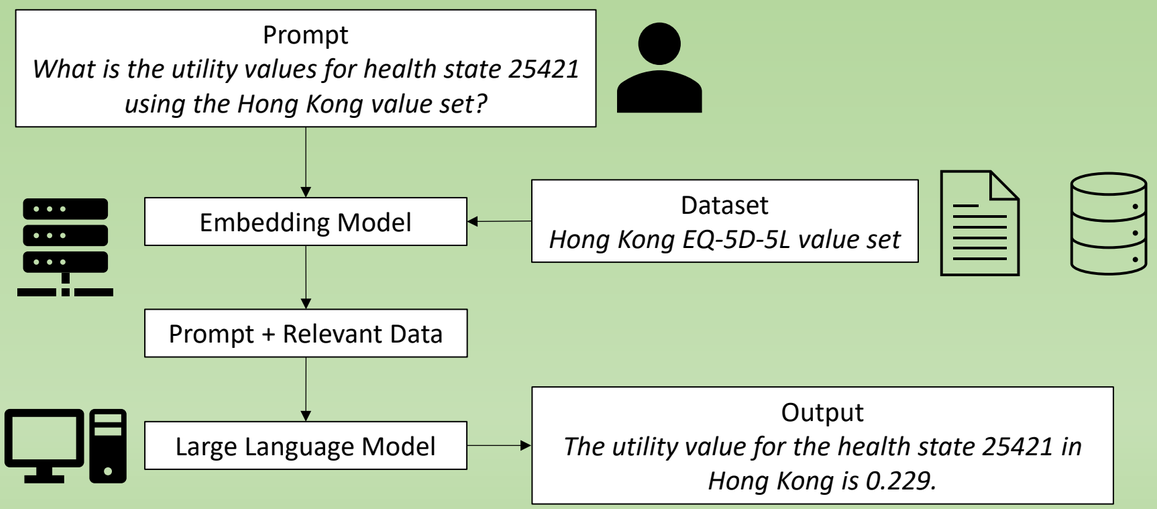

Table. Results of the large language models

One-state query						
Model	Parameters	Size	Queries	Correct responses	Accuracy	Time per task (sec)
Llama 3.2	3B	2.0GB	3,125	3,125	100.0%	1.6
Llama 3.1	8B	4.7GB	3,125	3,125	100.0%	1.5
Gemma 2	9B	5.5GB	3,125	3,125	100.0%	3.4
Mistral	7B	4.1GB	3,125	3,125	100.0%	1.1
Qwen 2.5	7B	4.7GB	3,125	3,125	100.0%	1.7

Two-state query						
Model	Parameters	Size	Queries	Correct responses	Accuracy	Time per task (sec)
Llama 3.2	3B	2.0GB	1,000	994	99.4%	1.5
Llama 3.1	8B	4.7GB	1,000	999	99.9%	2.5
Gemma 2	9B	5.5GB	1,000	1,000	100.0%	5.4
Mistral	7B	4.1GB	1,000	998	99.8%	2.8
Qwen 2.5	7B	4.7GB	1,000	1,000	100.0%	2.9

Conclusion:

- This study demonstrated the capabilities of LLMs equipped with RAG in accurately retrieving EQ-5D-5L utility values, highlighting their potential to assist researchers and end-users as an AI-powered EQ-5D chatbot.
- Future studies may evaluate the use of large online models which may process a batch of utility score estimation efficiently and accurately.
- Future developments could expand the scope of the EQ-5D database to include all published value sets for the EQ-5D instruments across different countries and regions, as well as EQ-5D related knowledge obtained from published studies.
- However, rigorous validation of responses before real-world deployment will be essential to ensure accuracy and reliability.