Exploring the Development of Briefing Books for Early Scientific Advice Using Large Language Models: A Proof-of-Concept Study

OP18

Ryan Thaliffdeen[1], Matthew Radford[1], Inês Guerra[2], João Leite[3], Shaktidhar Pullagurla[4], Vishalie Shah[2], Edel Falla[2], Raja Shankar[2], Yumi Asukai[5]

[1]Gilead Sciences, Inc., Foster City, CA, USA; [2]IQVIA, London, UK; [3]IQVIA, Lisbon, Portugal; [4]IQVIA, Bangalore, India; [5]Gilead Sciences, Ltd., Stockley Park, UK

QR code to be replaced with new one specific to a poster

Conclusions

- The LLM performed better in areas that relied on pre-trained knowledge (e.g., choice of comparator), compared to areas that required advanced reasoning (e.g., ITC, economic modeling)
- Overall, the LLM, although successful in retrieving information from the knowledgebase, could not generate an HTA-grade BB
- Improving the knowledgebase with relevant literature and clinical feedback, coupled with expert prompting guidance, could enhance the LLM's performance. However, this improvement would come at the cost of considerable human effort

Introduction

- Health Technology Assessment (HTA) bodies (e.g., National Institute for Health and Care Excellence (NICE)) are market-specific groups that evaluate the clinical, safety, and economic evidence surrounding a new medicinal product coming to a local market[1]
- Early Scientific Advice (ESA) is an opportunity to inform Clinical Development and Reimbursement Strategies.[2] Prior to any ESA engagement, a briefing book (BB) will be developed to seek advice on particular topics and summarize the company's position on those questions
- While Large Language Models (LLMs) have demonstrated efficiencies for various Health Economics and Outcomes Research deliverables, BBs pose unique challenges for LLMs, as BBs are generated earlier in a product's lifecycle when evidence is limited
- Additionally, BBs require strategic thinking to develop a company's position and justification on questions for HTA in order to optimize the BB to the specific necessities of each HTA process

Objective

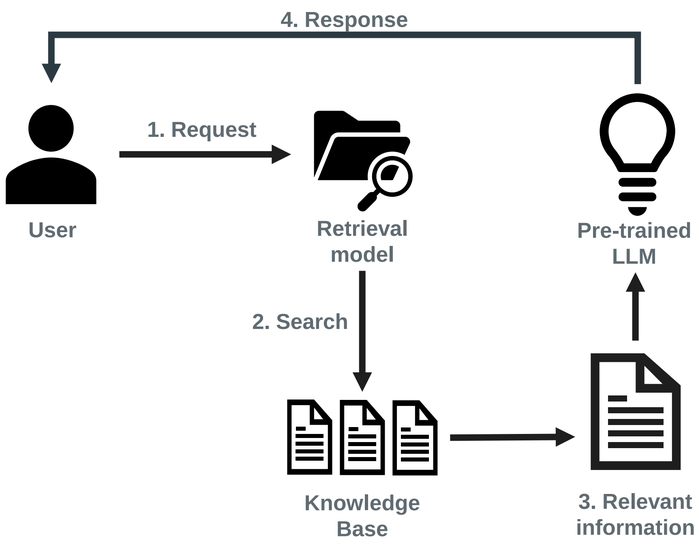
- This proof-of-concept study aimed to assess the feasibility of LLM-based generation of BBs for ESA

Methods

Sections and Knowledge Base

- Sections on approaches for trial comparator selection, indirect treatment comparison (ITC), and economic modeling were created using GPT-4 via python API
- To supplement the model's pre-trained knowledge, retrieval-augmented generation (RAG) was used for content generation and answer retrieval (**Figure 1**)
- The model's knowledgebase included the trial protocol, internal strategic documents, previous HTA appraisals, HTA BB guidance, and published trial results in similar indications
- Prompts were developed iteratively upon review of outputs
- Key Evaluation Metrics were output quality and human-led effort needed for revisions

Figure 1. Retrieval Augmented Generations (RAG) Framework



Results

Figure 2. The first question focused on the justification for the choice of comparator in the trial

Question	
Does the Agency agree that neoadjuvant nivolumab in combination with chemotherapy, followed by investigator's choice adjuvant treatment, is an appropriate comparator representing the SOC for patients with early-stage solid tumour?	
Prompt	**Results**
Prompt with **no sources** was tested as control	LLM retrieved the message "The information needed to answer the question is **not provided in the context**."
Initial prompt introduced a **clinical trial of nivolumab** as a source and requested the model to utilize **NCCN guidelines** for justification	Model provided **short summary** of CheckMate 816 trial and **hallucinated** using the NCCN guideline as justification
Request to use NCCN guideline was **removed** from the prompt	Model provided **more extensive summary** of CheckMate 816 trial and **limited top-level** information regarding standard of care.
Revised prompt included **two main updates**: (1) **additional knowledge sources** (HTA recommendation for Nivolumab, HTA recommendation for Atezolizumab, HTA briefing book template and briefing book guidance); and (2) a **clearer distinction** between the neoadjuvant and adjuvant setting within the context.	The response from the LLM was generally **similar to earlier responses**, although the LLM was able to identify that the investigator's choice of therapy is consistent with country-specific guidelines in the adjuvant setting. While the model was able to **recognize the relevant sources**, it did not always retrieve **all of the relevant information.**

Results

Figure 3. The answers by the LLM were often high level and did not always utilize the relevant information from the given sources

The LLM gave very general and top-level responses

- "The ESMO guidelines are also expected to include recommendations on PD-1 inhibitors in the (neo)adjuvant setting due to results from clinical trials"
- "Clinical studies have demonstrated the effectiveness of the comparator therapy"
- "The choice of Nivolumab + chemotherapy as a comparator may be supported by its demonstrated cost-effectiveness, its impact on quality of life, and its alignment with current therapeutic strategies in the […] healthcare setting"

The LLM often recognized the relevant source but not the relevant information

- "The choice of neoadjuvant Nivolumab in combination with chemotherapy as the comparator is justified by the clinical trial protocol"
- "The Journal of Clinical Oncology's recommendations section likely emphasizes the clinical benefit and the role of Nivolumab in combination with chemotherapy in the neoadjuvant setting"

Figure 4. The second question focused on whether the LLM could generate and justify an economic model based on the relevant context provided

Question	
Does the Agency agree that the proposed economic model structure captures the key health state relevant for this patient population, and do the available trial data/endpoints allow those health states to be accurately populated over time? Is there an alternative model structure/approach that should be considered?	
Prompt	**Results**
Initial prompts for the RAG model included providing a **summary of the economic model** and requesting for an **expansion on the model design** and any considerations from the LLM	The answer by the LLM **regurgitated** the summary of the economic model in different words, providing no new information
Another approach explored **using pre-trained knowledge** to provide context on the objective of the model and request the LLM to provide a **suggested model design** for the objective and clinical context.	The LLM suggested **slightly different health states in different attempts**, and it proved **inconsistent** in terms of model design. Some of the responses were more aligned with the health states provided than others. Additionally, the model was **not capable of conceptualizing specific design aspects** for the decision problem
A **tutorial on economic models** was provided to the LLM in order to provide further context and options for model designs	The tutorial **did not improve the response**, and introduced some **confusion to the model**, which took the model examples from the tutorial and utilized their approach for the model design

Figure 5. As the response by the LLM was not satisfactory, two different approaches were explored to further elaborate on the economic model approach and design suggested by the LLM

Selected approaches	
Additional knowledge sources (local guidelines for the economic evaluation of health technologies and HTA briefing book guidelines) and a **list of bullet points detailing the expected content** and format of the response were added to the prompt	The second approach updated the first prompt with a Q**&A style conversation** with the LLM, asking these follow-up questions: 1. What is your rationale behind choosing a Markov model? 2. Why have you assumed that the transition probabilities are constant over time, given the time-to-event nature of the proposed study? 3. Would you be able to revise your proposed economic model to a semi-Markov model and explain the model parameters, assumptions and limitations using this approach?
The response was similar to earlier responses in terms of having insufficient detail to be practicably useful.	The LLM was **able to provide some additional justification** on its choice of economic model following this Q&A, but again with insufficient detail and "textbook" style responses. It also **did not commit to a specific model design** and highlighted the possibility of selecting other designs.

Figure 6. The LLM could not conceptualize an economic model beyond the basic assumptions and standard model designs

The LLM gave very general and top-level responses, even in a Q&A setting

- "The model could be a Markov model with health states reflecting the progression of the disease and the treatment process. The health states could include pre-surgery, no surgery, surgery, event-free survival, loco-regional recurrence, metastasis, cure, and death. The model would allow for transitions between these states based on the probabilities derived from the […] trial data and other relevant literature."
- "Q: Why have you made the assumption that the transition probabilities are constant over time, given the time-to-event nature of the proposed study?; A: " The assumption that transition probabilities are constant over time is a simplifying assumption […] To address this limitation, time-dependent transition probabilities can be incorporated into the model […] Alternatively, semi-Markov or non-Markov models can be used, which allow for the transition probabilities to depend on the time already spent in a particular health state"

The LLM lacks the reasoning necessary to conceptualize an economic model

- "The transition between these states will be based on the efficacy endpoints from the […] study, specifically the pathological complete response rate (pCR), event-free survival (EFS), overall survival (OS), and major pathological response rate (MPR)"
- "It will be assumed that the efficacy and safety of the treatments are accurately represented by the […] study data. It will be assumed that all relevant costs are included in the model, although there may be other indirect costs that are difficult to quantify."

Figure 7. The third question focused on whether the LLM could assess the feasibility of conducting an NMA/ITC based on the pivotal trial design

Question	
Can the Agency comment on the feasibility and utility of conducting a network meta-analysis (NMA)/ITC to compare the EFS and pCR of product X with other relevant comparators?	
Prompt	**Results**
Initial prompts included only the Nivolumab HTA **report and trial protocol** as the sources	The answer by the LLM was **top level and non-specific**, and it did not provide any assessment on the possibility of an NMA
One solution tested was to **provide the LLM with a list of possible relevant comparator trials** and retrieve their relevant information for comparison with the pivotal trial via clinicaltrials.gov. Only Inclusion/Exclusion Criteria, Arms and Interventions and Endpoints were compared.	Response was more positive, and it provided a **general comparison between the trials**, their characteristics and a short assessment on the feasibility of conducting an ITC. However, the LLM **failed to recognize** most of the necessary steps and criteria for conducting an NMA
Two clinical trial publications were given to the LLM in order to assess a **possible network** with the trial, instead of the Clinical Trials website. This ideally allowed the LLM to assess the reported **baseline characteristics** that were not reported in the website.	Response was similar to previous ones. The LLM provided a **short analysis on the feasibility** of comparing study populations, possible subgroups of interest, differences between trials, and a possible anchor comparator for an ITC. The LLM lacked the reasoning capabilities to provide any justification for its assessment and **it did not compare** baseline characteristics.

Figure 8. The LLM recognized some criteria necessary for an NMA but lacked the depth to properly assess the feasibility of an ITC

The LLM recognized aspects of assessing the feasibility of an NMA

- "The eligibility criteria across the trials showed similarities […]. However, there are variations in the disease stages included, with some trials specifying different stages […]. Additionally, there are differences in the requirements for PD-L1 expression levels and the exclusion of patients with certain autoimmune diseases or other malignancies"
- "Event-Free Survival and Pathological Complete Response are common endpoints across the trials, which supports the feasibility of comparing these outcomes"
- "The interventions across the trials are similar in that they all involve immune checkpoint inhibitors combined with platinum-based chemotherapy in the neoadjuvant setting, followed by continuation of the immunotherapy in the adjuvant setting"

The LLM stopped the comparison at a high level and did not completely assess the feasibility of an NMA

- "While there is consistency in the use of EFS and pCR as primary endpoints, there may be differences in how these are defined and measured across the trials"
- "However, there are differences in […] Atezolizumab, Nivolumab, and Pembrolizumab […], which could affect the comparability of the trials"

References: 1. Trowman R et al. Health technology assessment 2025 and beyond: lifecycle approaches to promote engagement and efficiency in health technology assessment. International Journal of Technology Assessment in Health Care. 2023;39(1):e15. doi:10.1017/S0266462323000090. **2.** Wang T, et al. Building HTA insights into the drug development plan: Current approaches to seeking early scientific advice from HTA agencies. Drug Discov Today. 2022 Jan;27(1):347-353. doi: 10.1016/j.drudis.2021.09.014. Epub 2021 Sep 28. PMID: 34597755.

Correspondence: Ryan Thaliffdeen (Ryan.Thaliffdeen@gilead.com)

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