Automating Economic Modeling: The Potential of Generative AI for Updating Modeling Reports

William Rawlinson,[1] Siguroli Teitsson,[2] Tim Reason,[1] Bill Malcolm, [2] Andy Gimblett,[1] Sven L. Klijn[3]

[1]Estima Scientific, London, United Kingdom; [2]Bristol Myers Squibb, Uxbridge, UK; [3]Bristol Myers Squibb, Princeton, NJ, USA

Introduction

Large language models

- Large language models (LLMs), such as Generative Pre-trained Transformer 4 (GPT-4), are mathematical models that are trained on vast corpuses of textual data to learn underlying statistical relationships in language.[1]
- LLMs can generate human-like text. Consequently, LLMs have a wide range of applications in automating tasks that are currently performed manually in HEOR.[2]

Figure 1. Generation of human-like text

Adapting a cost-effectiveness model

- Model adaptation is the process of updating an existing health economic analysis to fit a new decision problem. Adaptations vary in complexity and scope, but can involve changing input values (commonly, more than 100 inputs required updating), updating methods, and adding or removing comparators.
- Adapting cost-effectiveness models is a core component of health technology assessment (HTA) workflows. Typically, the HTA cycle initiates through a manufacturer developing a global cost-effectiveness model and technical report. These materials evaluate the cost-effectiveness of a technology from the perspective of a single country. The global materials are then adapted to create country-specific models and reports, which are subsequently used in HTA submissions around the world.

Study aims

- The aim of this study was to assess the capabilities of GPT-4 in automatically adapting a global technical report (built in Microsoft Word) to a country-specific setting.
- The study assumed the corresponding country-specific cost-effectiveness model (built in Microsoft Excel) was already available.

Methods

Study design

- We developed an LLM-based solution in Python and Visual Basic for Applications (VBA) to enable automated adaptation of a technical report written in Microsoft Word, using a cost-effectiveness model built in Microsoft Excel.
- We tested the solution on a global technical report that compared treatments for muscle-invasive urothelial carcinoma (MIUC) from a National Health Service and Personal Social Services (UK) perspective. The report and the corresponding model were previously used in HTA submissions. The report was 137 pages in length.
- First, we manually adapted the Global MIUC Excel model to the Czech Republic perspective, by changing input values.
- The LLM-based solution then automatically updated the results and discussion section of the Global technical Word report to the Czech Republic perspective, using data from the Excel model.
- The results of the country-specific MIUC model were randomised to ensure confidentiality. A locally hosted and/or secure LLM instance should be used in a real-world scenario to ensure confidentiality.

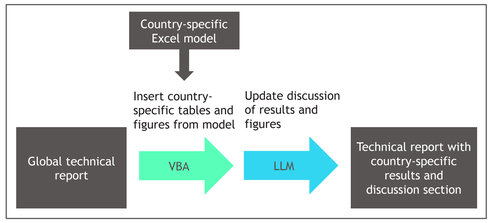
The LLM-based solution

- The LLM-based solution worked in two stages
  1. Inserting country-specific results tables and figures into the global report
  2. Updating text in the results and discussion section of the global report

Inserting country-specific result tables and figures into the global report

- A VBA script was developed to automatically copy country-specific results tables and figures from an Excel model into a global report.
- Some manual set up was required, but once performed, this would facilitate any number of adaptations.

Figure 2. High-level overview of the LLM-based adaptation solution

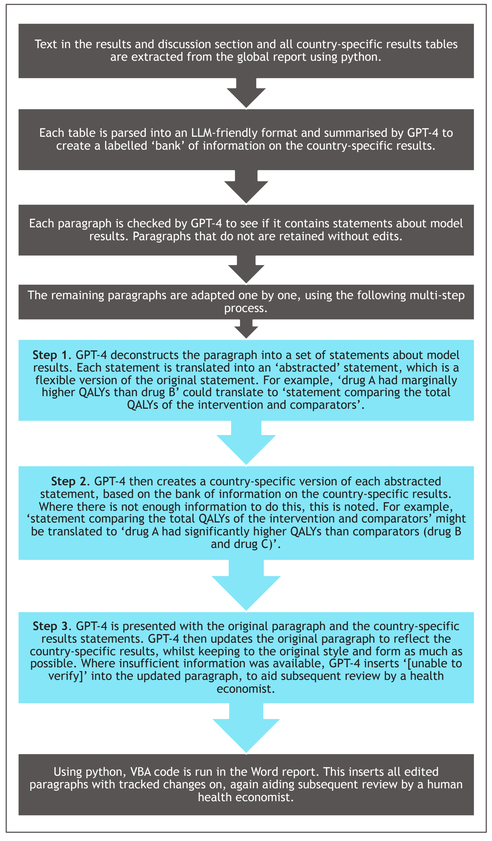


- The set up involved:
  - Creation of 'Live' results tables and figures in the Excel model. These are tables and figures built in the Excel model that use technical report formatting. 'Live' refers to the fact that the tables and figures are linked to the model results cells, meaning they will update to reflect the latest model results
  - Adding bookmarks (technical report) and named ranges (model) to link the 'live' results tables and figures to corresponding results tables and figures in the technical report
- Once set up was performed, the results tables and figures in the global report could be automatically updated to match the current model results with the click of a button. This process could also be set up to run automatically after sensitivity analyses.

Updating text in the results and discussion section of the global report

- A Python script was developed to automatically update text discussing model results in a global report.

Figure 3. Updating report text



- The script used application programming interface (API) calls to pass information to, and receive text from, a large language model.
- The script relies upon country-specific results tables and figures having been already inserted to the global report (step 1).
- At a high level, GPT-4 is provided with a representation of the country-specific results tables that have been inserted into the report. GPT-4 then modifies existing paragraphs in the report to ensure that they reflect the updated results.
- In practice, a multi-step process was required to satisfy the following key requirements:
  - Retainment of the original style and messaging of the report
  - Flexibility to handle significant changes to the model (country-specific vs global) such as inclusion of new comparators
  - Factual accuracy

Functionality

- As described in **Figure 3**, all edits made by GPT-4 were automatically inserted as tracked changes in the global report. This provides transparency and facilitates subsequent review by a human health economist.
- In addition, cross-references and references in sections of the report edited by GPT-4 were preserved. This was achieved through the following process:
  - Cross-references and references are stored as fields in Word, whereas an LLM generates unformatted text. Consequently, simply inserting LLM-generated text into a report would overwrite cross-references and references
  - Therefore, prior to sharing report text with GPT-4, VBA code was automatically triggered to store all fields in the report, and replace these by unique text keys (for example, '%field1%')
  - When editing text, GPT-4 was instructed to keep the text keys intact.
  - After GPT-4's edits had been inserted into the report, VBA code was automatically triggered to insert the fields back into the report, based on the locations of the text keys.

Assessment of performance

- Two experienced health economists blindly assessed the country-specific results and discussion section of the AI-generated report alongside a report that had been manually adapted by a third health economist.
- Performance was evaluated in two categories: accuracy and style.
- Style was evaluated qualitatively, and accuracy was evaluated based on two dimensions:
  - Were edits factually correct? (did GPT-4 make changes that correctly described the country-specific model results?)
  - Was original text retained when appropriate? (did GPT-4 avoid making changes when original report text was still applicable to the country-specific model results?)

Results

- The results of the study are displayed in **Table 1**.
- The reviewers agreed in their evaluation of the accuracy of the AI-generated and manually adapted reports.
- Accuracy was 94.3% for the AI-generated report and 97.1% for the manually adapted report.
- There were 2 incorrect edits in the AI-generated report. This included a rounding error, and an incorrect description of a scenario analysis.
- Qualitatively, the reviewers generally approved of the tone of edits made by GPT-4. However, there were a small number of factually correct edits where the reviewers preferred the language chosen by the human health economist. For example:
  - The use of decimal places by GPT-4 was sometimes inconsistent
  - The qualifiers used by GPT-4 (for example, higher vs twice as high) sometimes reduced the impact of statements

Table 1. Accuracy of LLM-based edits to technical report

Dimension	LLM-based edits	Manual edits by health economist
Correct change	22	20
Incorrect change	2	1
Correct retainment of original text	11	13
Incorrect retainment of original text	0	0

Discussion

- This study is promising evidence that LLMs can be leveraged as part of a pipeline for automatically updating cost-effectiveness model technical reports in Microsoft Word.
- The accuracy achieved in our study (94.3% vs 97.1% for manual adaptation) suggests that the process is currently suitable for first edits prior to human review.
  - However, it's likely that accuracy can be enhanced through techniques such as LLM self-evaluation[3] and fine tuning a model specifically for report adaptation[4]
  - Therefore, there is potential for automated pipelines to reduce the rate of errors in the reporting of cost-effectiveness model results
- The study demonstrated the technical feasibility of using tracked changes to highlight edits made by an LLM, thereby enhancing transparency and supporting subsequent review (as shown in **Figure 4**). Also, we were able to devise a method to preserve references and cross-references in text edited by an LLM.
- Although tested on only one model and report, the prompts used in our study were agnostic to disease area and therefore should prove generalisable.
- Some manual set up was required to enable insertion of country-specific results tables and figures into a global report.
  - Despite the initial manual set up, large efficiency gains can be made versus the existing manual workflow.
  - This is compounded given that manual set up is only required once per global report / global model and would support any number of subsequent automated adaptations.
  - In addition, the process can be used any number of times to update a technical report if results change in the cost-effectiveness model. For example, if an error is found or if input values are changed.

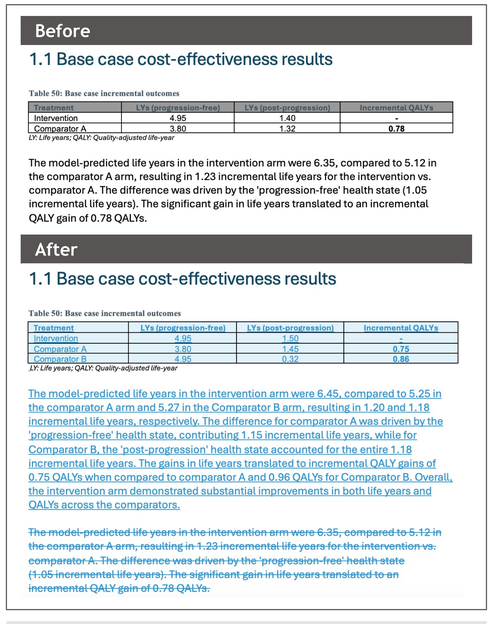
Limitations

- Interpretation of figures by GPT-4 was not included, meaning that result statements referencing figures (such as the cost-effectiveness plane) were not updated. Instead, these were classified as 'unable to verify'
  - Vision models are now widely available (such as gpt-4-vision-preview) that could be integrated into the process to interpret figures

Future research

- For this study, only the results and discussion section of the report were updated using an LLM. Technical reports usually also contain an introduction, methods, and inputs section.
- Further research is required to develop LLM-based solutions for updating the other sections of technical reports. There are two primary challenges to be overcome:
  - Writing these sections requires a broader knowledge of the cost-effectiveness model. For example, what the new country-specific input values are, and why these data sources were chosen. Potentially, this information could be communicated through providing a log of changes made to the cost-effectiveness model during adaptation (listing new input values, sources, and justification, etc.). This information would have to be collected regardless, if the report were manually adapted.
  - Secondly, these sections are likely to require creation and insertion of new references. Referencing information could be collected in the aforementioned log of changes, but a method would need to be developed for inserting formatted references automatically.
- Further research could also explore reducing the manual set up required
  - As methods for interpreting tabular data advance, it may be possible to skip step 1 in the process and use an LLM to update tables in the technical report by simply providing the results worksheet of a country-specific Excel file.
  - This approach was not used in this study due to accuracy limitations with currently available methods.

Figure 4. Example edits made by the automated process (using dummy results)



Key takeaways

- Technical reports are frequently adapted to reflect updated model results. For example, when adapting a global report to create a country-specific report for an HTA submission, when an error is discovered in a model, or when input values and methods are changed.
- This process can prove time consuming and error-prone when conducted manually.
- This study is promising evidence that LLMs can be leveraged as part of a pipeline for automatically updating Word technical reports for Excel cost-effectiveness models.
- Automated pipelines could save significant time over the course of a global model and report's lifespan, and (given further research) there is potential to reduce the rate of errors in reporting of model results.

References

[1] S. R. Bowman, 'Eight Things to Know about Large Language Models'. 2023.
[2] Reason T, Rawlinson W, Langham J, Gimblett A, Malcolm B, Klijn S. Artificial Intelligence to Automate Health Economic Modelling: A Case Study to Evaluate the Potential Application of Large Language Models. 2024 Mar. doi: 10.1007/s41669-024-00477-8.
[3] Madaan A et al, Self-Refine: Iterative Refinement with Self-Feedback. arXiv:2303.17651
[4] Jeong, C. Fine-tuning and Utilization Methods of Domain-specific LLMs.arXiv:2401.02981

Presented at ISPOR 2024; May 5-8; Atlanta, MA

Email: will.rawlinson@estima-sci.com