HTA225

LLM Engineering in HEOR – Approaches to Improving Accuracy in Clinical Data Extraction

Herman Eaves, MSc [1], Ahmad Hecham Alani, PharmD [1], Mackenzie Mills, PhD [1], Fiona Tolkmitt [1], Panos Kanavos, BSc, MSc, PhD [2]
[1] Hive Health Optimum Ltd. (HTA-Hive), London, United Kingdom
[2] The London School of Economics and Political Science (LSE), London, United Kingdom

Background

- Data extraction continues to be a difficult challenge for LLMs to overcome in the HEOR space.
- Often data comes from complex and large documents, such as HTA reports or SPC docs requiring context and understanding.
- HEOR represents a small portion of the overall training data for these LLMs, leading to limited out-of-the-box performance.

Objectives

- This study aims to assess which techniques lead to improved results for data extraction within the HEOR context, focusing on the use of emerging LLM tools.

Methods

- Variables related to ITCs assessed in HTA reports from HAS, G-BA, NICE, and PBAC (n=471) evaluating drugs for solid tumors were manually extracted to serve as an accuracy reference.
- The extracted variables were:
  1. Does the report contain an ITC;
  2. Does the ITC use any of the following: adjustment, anchoring, or matching;
  3. What was the agencies overall sentiment on the ITC.
- A series of structured extractions using LLMs were then run on text extracted from the same reports
- Several approaches were tested to improve accuracy, starting from a baseline generated with default API settings and a Gemini-2.0-flash prompt "Extract the defined information about indirect treatment comparisons from the following report: “ with the schema shown in **figure 1**.
- The additional methods were: 1. Taking the most common response when multiple outputs were produced; 2. “Improving” the prompt; 3. Adding a system prompt; 4. Using a larger model (Gemini-2.5-flash); 5. Using the biggest model (Gemini-2.5-Pro); 6. Using a lower temperature; 7. Using another LLM call to validate and correct the original extraction; 8. Adding detailed HEOR context from two published articles [1][2] related to ITCs.
- Results were compared across all methods to evaluate improvements in accuracy for the ITC-related extractions.

Figure 1: *Extraction schema definition*

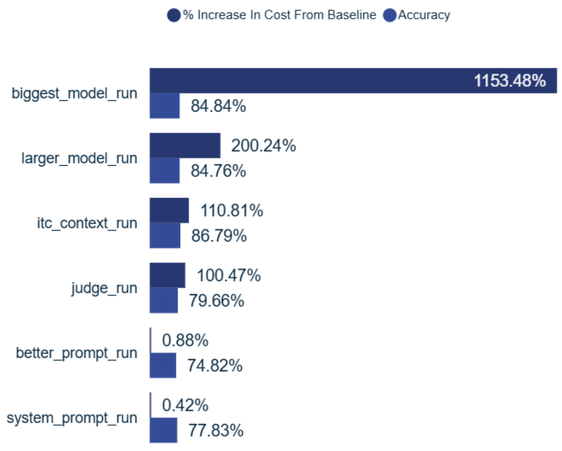
```python
class ITCAcceptance(Enum):
    POSITIVE = "positive"
    NEGATIVE = "negative"
    NEUTRAL = "neutral"
    UNKNOWN = "unknown"


class ITCInformation(BaseModel):
    has_itc: bool = Field(
        description=(
            "If the report describes an indirect treament comparison"
        )
    )
    has_adjustment: bool | None = Field(
        description=(
            "If the report describes an indirect treament comparison which is adjusted"
            )
    )
    has_anchoring: bool | None = Field(
        description=(
            "If the report describes an indirect treament comparison which has anchoring"
            )
    )
    has_matching: bool | None = Field(
        description=(
            "If the report describes an indirect treament comparison which has matching"
            )
    )
    itc_acceptance: ITCAcceptance | None
```

RESULTS

- The baseline approach (default API for Gemini-2.0 + base prompt) produced an average accuracy of 75.5% across all ITC variables.
- Adding ITC context produced the highest average accuracy 86.8%.
- The lowest accuracy came from attempts to improve the prompt, reducing accuracy to 74.8% (0.7% less than the baseline).
- When assessing cost, using a better model increased cost from baseline by 200.24%, using the additional context increased cost by 110.81%, and Gemini-2.5-Pro increased cost 1153.48% with an average accuracy of 84.8%. The results can be seen in **figure 2.**

Figure 2: *Cost of run as a percentage of the baseline with accuracy (low-cost values excluded) (Baseline accuracy = 75.5%)*



- **Figure 3** shows accuracy results by agency. The highest accuracy was achieved with the *ITC context* run on G-BA documents (91.7%), while the lowest was observed with the *improved prompt* run on NICE documents (68.9%).
- **Figure 4** presents accuracy by variable, showing that the inclusion of additional context significantly improved performance in the *has_itc* and *itc_acceptance* fields—by 28% and 31%, respectively, compared with the lowest-performing runs for those variables.

Figure 3: *Average accuracy results by agency separated by method*

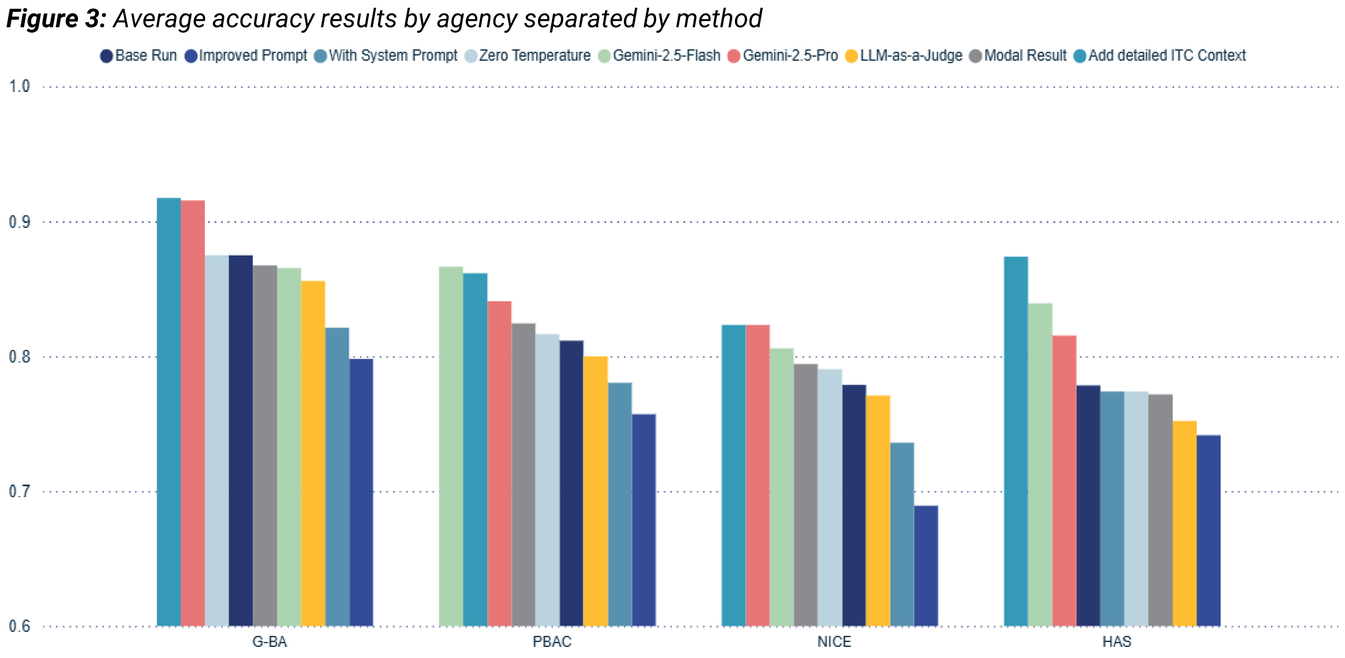
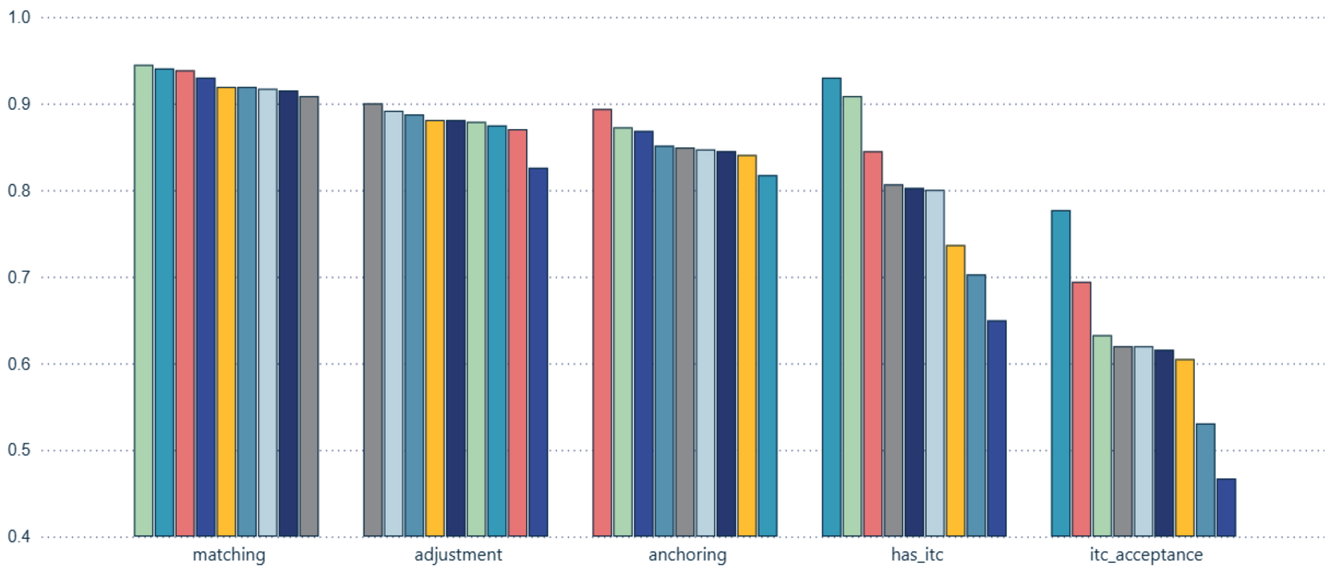


Figure 4: *Average accuracy results by variable separated by method (same colours as above)*



Conclusions

- For this task, the accuracy trade-off supports using smaller models such as Gemini-2.0-Flash, provided that sufficient contextual information related to the task and HEOR is included. Model size alone appeared to have minimal impact on performance.
- Agencies with clear guidance and reporting style, such as the G-BA, achieved stronger results, while HAS performed less well. HAS reports are often complex and difficult for LLMs to extract from, typically including large tables and extensive supplementary information.
- Initial analyses indicate that overall accuracy remains below the level generally required for routine data extraction in HEOR applications.
- Minor prompt adjustments or excessive task-specific details tended to reduce output quality and added limited value. Specific prompt updates were also difficult to evaluate due to the flexible nature of model interactions.
- Looking ahead, applying a RAG approach that focuses on selecting only the most “*relevant*” sections of an ITC may further enhance extraction performance.

References

[1] Igarashi, A., Tanaka, S., De Moor, R. *et al.* Indirect Treatment Comparisons in Healthcare Decision Making: A Targeted Review of Regulatory Approval, Reimbursement, and Pricing Recommendations Globally for Oncology Drugs in 2021–2023. *Adv Ther* **42**, 52–69 (2025). https://doi.org/10.1007/s12325-024-03013-6
[2] Macabeo, B., Rotrou, T., Millier, A. *et al.* The Acceptance of Indirect Treatment Comparison Methods in Oncology by Health Technology Assessment Agencies in England, France, Germany, Italy, and Spain. *PharmacoEconomics Open* **8**, 5–18 (2024). https://doi.org/10.1007/s41669-023-00455-6

Abbreviations

Large Language Models, LLMs; Indirect Treatment Comparisons, ITCs; Summary of Product Characteristics, SPC; Health Economics and Outcomes Research, HEOR; Health Technology Assessment, HTA; Haute Autorité de Santé (French National Authority for Health), HAS; Gemeinsamer Bundesausschuss (Federal Joint Committee), G-BA; National Institute for Health and Care Excellence, NICE; Pharmaceutical Benefits Advisory Committee, PBAC; Application Programming Interface, API; Retrieval-Augmented Generation, RAG