A Comparative Assessment of LLM Agreement for Clinical Data Extraction Tasks

Herman Eaves[1], Victoria Zaitceva[1], Ryan Lin[1], Mackenzie Mills, PhD[1,2], Panos Kanavos, 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

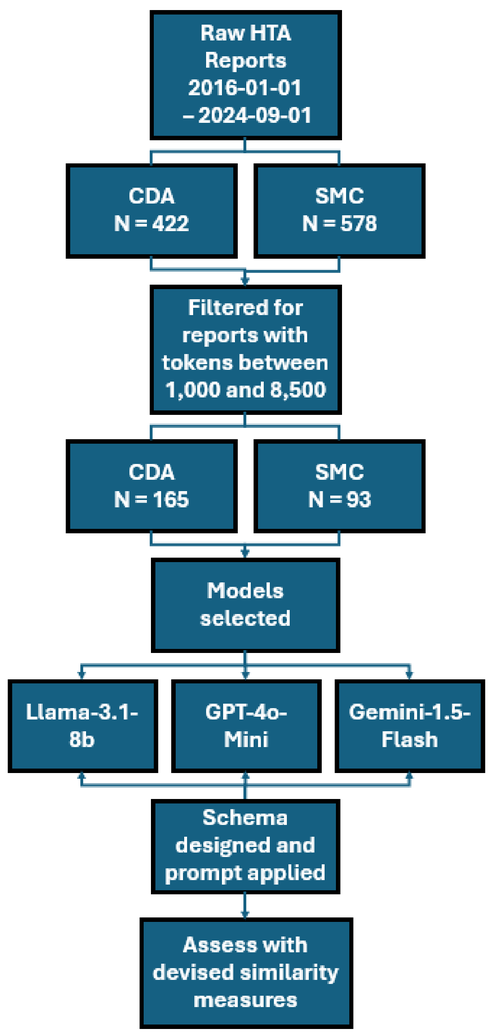
MSR17

Introduction

- Recent emergence in AI tools which are showing increasing promise in data extraction tasks for SLRs, meta-analyses, and text extraction from unstructured documents.
- AI models are prone to “hallucinations”, and each model has its own strengths and weaknesses.
- In this study we evaluate the comparative output of different LLM models against each other when extracting clinical assessment data from HTA reports. Evaluating ability to detect and describe distinct forms of clinical evidence.

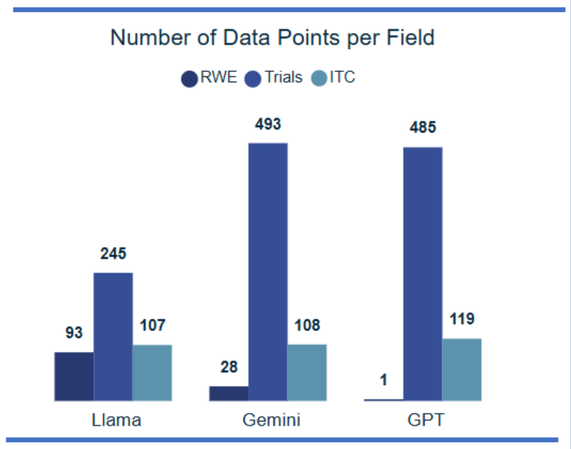
Methodology

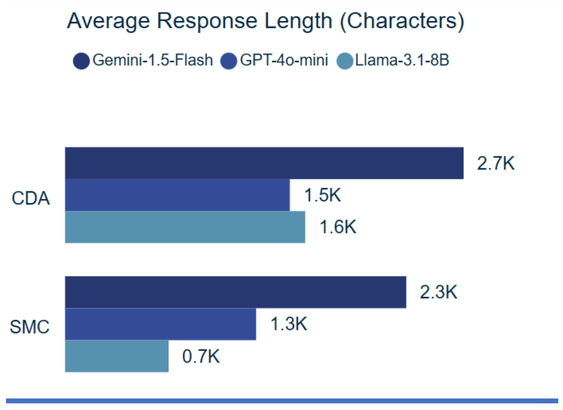
- We selected GPT-4o-Mini, Llama-3.1-8B, and Gemini-Flash-1.5-002 for the analysis, due to their popularity and low cost. A JSON schema was specified as an output format for these models, including a breakdown of clinical trials, real world evidence, and indirect treatment comparisons considered in the HTA report.
- The models were fed the same prompt and sampling parameters, including an example schema.
- Model outputs were assessed in terms of similarities and differences. A similarity scoring system was devised, directly comparing clearly defined outputs and comparing free text (such as a description) similarity using SentenceTransformers. A match was defined as an exact string match, or a cut-off at 65% similarity for larger text fields, each match contributed to a higher score, normalized against number of comparisons.
- Length of output string was measured to determine average amount of detail in a model response.
- When there were discrepancies in the number of each variable captured, we chose to only compare the ones available in both. Higher weighting was placed on clinical trial details.
- A subset of the 5 highest and 5 lowest similarities were selected for inspection.

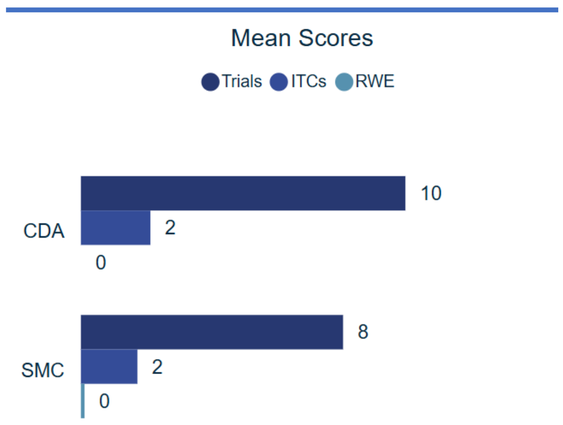


Results

- 20.5% of Llama and 4.7% of Gemini outputs (24.4% total) were not valid JSON and were omitted for analysis, every GPT response was in the required format.
- Gemini had by far the largest mean output length (2539 characters), while GPT and Llama had similar values (1430 and 1270 respectively). With maximum output length, Gemini, Llama and GPT had values of 11689, 7646 and all respectively.
- The mean total similarity score across all models was 11.9 (IQR 6.9 – 15.7), the maximum agreement was 41. The maximum agreement was derived from an SMC report assessing Nivolumab for melanoma, this report contained 6588 tokens, it had the highest agreement between GPT and Gemini (16.5).
- When extracting trial information, GPT and Gemini had the greatest similarity score (5.1), while Llama and Gemini had the least (3.8). For ITCs, the same pattern as before emerged (3.2, 2.4). GPT consistently failed to capture real world evidence and could not be compared.
- No model correctly separated the endpoints and reported all the results under one endpoint.







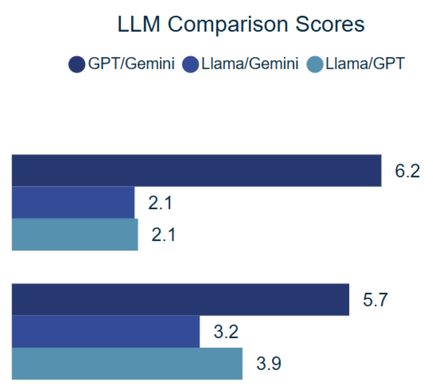


Table 1: Example extraction with the best score in the sample

Variable	Llama-3.1-8b	GPT-4o-mini	Gemini-1.5-flash
Clinical Trials			
Trial Name	CheckMate 238	CheckMate 238	CheckMate 238
NCT Code	Failed	Failed	Failed
Comparators	Ipilimumab	Ipilimumab	ipilimumab
Description	A Randomized, Double-blinded, Phase III Study of Recurrence-free Survival Comparing Nivolumab (Opdivo) Versus Ipilimumab in Patients With Completely Resected Stage IIIB, IIIC, or IV Melanoma	An ongoing double-blind, randomised, phase III study comparing nivolumab with ipilimumab in patients with completely resected stage IIIB, IIIC, or IV melanoma.	an ongoing double-blind, randomised, phase III study, CheckMate 238, that recruited patients at least 15 years of age with histologically confirmed melanoma (stage IIIB, IIIC, or IV according to American Joint Committee on Cancer [AJCC] 7[th] edition), metastases to regional lymph nodes or distant (including brain) metastases that had been surgically resected, complete regional lymphadenectomy or resection within 12 weeks before randomisation and Eastern Cooperative Oncology Group (ECOG) performance status score 0 or
Variable	**Llama-3.1-8b**	**GPT-4o-mini**	**Gemini-1.5-flash**
Indirect Treatment Comparisons (ITCs)			
Type of ITC	Bayesian	Patient level data meta-regression and parametric survival analysis	Patient level data meta-regression in a parametric survival analysis
			Bucher comparison
Comparators	Ipilimumab Placebo	Placebo Ipilimumab	placebo
ITC Description	The manufacturer submitted ITC used Bayesian methods to conduct the ITC.	The submitting company indirectly compared nivolumab with placebo, as a proxy for routine surveillance, using patient level data meta-regression in a parametric survival analysis.	The submitting company indirectly compared nivolumab with placebo, as a proxy for routine surveillance, using patient level data meta-regression in a parametric survival analysis (used in the economics base case) and also via a Bucher comparison (used in a scenario analysis in the economic case). Both indirect treatment comparisons (ITCs) included two studies: the pivotal CheckMate 238 study, comparing nivolumab with ipilimumab in patients with completely resected stage IIIB, IIIC, or IV melanoma and CA 184-029, an international, double-blind, randomised, phase III study comparing adjuvant ipilimumab therapy with placebo in 951 patients with completely resected stage III melanoma at high risk of recurrence.
Trial Endpoints			
Endpoint Name	**Recurrence free survival (RFS)**	**Recurrence Free Survival (RFS)**	**Recurrence free survival (RFS)**
Endpoint Results	**Measure**	Hazard Ratio (HR)	RFS rate at 12 months	RFS rate
	Value	0.65	70%	70% (66 to 74)
	CI	0.51 to 0.83	66 to 74	None
	P Value	0.001	0	None
	Measure	Hazard Ratio (HR)	RFS rate at 18 months	RFS rate
	Value	0.66	66%	66% (62 to 70)
	CI	None	62 to 70	None
	P Value	0.0001	0	None
	Measure		RFS rate at 24 months	RFS rate
	Value		63%	63% (58 to 67)
	CI		58 to 67	None
	P Value		0	None
	Measure		Recurrence or death at 18 months	Recurrence or death at 18 months
	Value		34%	34% (154/453)
	CI		154/453	None
	P Value		0.001	None
	Measure		Recurrence or death at 24 months	Hazard Ratio (HR)
	Value		38%	0.65
	CI		171/453	0.51 to 0.83
	P Value		0.0001	<0.001
	Measure		Hazard Ratio (HR) for recurrence at 18 months	Recurrence or death 24 months
	Value		0.65	38% (171/453)
	CI		0.51 to 0.83	None
	P Value		0.001	None
	Measure		Hazard Ratio (HR) for recurrence at 24 months	Hazard Ratio (HR)
	Value		0.65	0.66
	CI		0.51 to 0.83	None
	P Value		0.0001	<0.0001

Link to source report:

Conclusion

- We find that the quality of output varies greatly across models, in subsequent tests, we found that using a larger model (Chat-GPT-4o) still presented similar issues.
- Gemini had best overall performance in terms of level detail captured and accuracy. Llama generated concise but accurate responses, showing a good understanding of report context. GPT was effective at extracting specific values, such as trial endpoint results, but had inconsistencies with the value and the confidence intervals, reporting a percentage as the result, then hazard ratios in the confidence intervals in some examples.
- Further work in this area could include prompt engineering, fine tuning of a model for this specific purpose, additional pre-processing steps such as feeding in portions of the text which contain the clinical information and avoiding the noise of other discussion or modifications of the output schema.
- Stakeholders should remain aware of the underlying limitations in these tools and adjust accordingly.