Evaluating GenAI for HTA Analogue Analysis: Identifying Rare Disease Approvals with Surrogate Endpoints and Single-Arm Trials

CONCLUSIONS

GenAI models identified analogue products based on multiple identification criteria. However, no genAI model was 100% accurate nor was able to consistently reproduce results.

While none of the genAI models achieved human-level precision, ChatGPT-5 Pro Deep Search demonstrated the most consistent alignment with analogue identification criteria.

GenAI models have the potential to improve the efficiency of HTA analogue analyses and streamline the identification of relevant analogues to inform analyses on HTA drivers and barriers. However, human curation is currently essential to ensure accuracy and reliability.

As a next step, the capability of genAI models to augment data extraction, synthesis and visualization to support HTA analogue analyses will be explored. Further research on the relationship between prompt specificity and the accuracy of analogue identification by genAI models would be of interest.

AUTHORS

Rachel Beckerman[1], Charlie Davis[1], Sylwia Lach[1], Chloe Sheppard[1]

AFFILIATIONS

1. Maple Health Group, LLC. New York, USA

M🍁PLE HEALTH GROUP

INTRODUCTION

- HTA analogue analyses are crucial for understanding drivers and barriers to positive health technology assessments (HTA). They are particularly important in rare disease, where specific evidence constraints often necessitate reliance on surrogate endpoints and single-arm studies.
- Generating high-quality, fit-for-purpose analogue research is dependent on the accurate identification of relevant products. Currently, the analogue identification process is led by human consultants and takes several hours.
- Reducing the amount of human time spent on identifying analogues would allow these resources to be repurposed for more complex and strategic tasks.
- Generative artificial intelligence (genAI) models have become increasingly capable of producing novel, coherent outputs in response to a prompt. We hypothesise that genAI models could improve the efficiency of analogue analyses by assisting with product identification.

STUDY OBJECTIVE & METHODOLOGY

- The aim of this research was to evaluate the extent to which different genAI models can replicate human-led analogue analyses by accurately and efficiently identifying analogue therapies which meet specific criteria.
- Five analogue identification criteria were defined: (1) rare disease indication, (2) surrogate primary endpoint, (3) single-arm pivotal trial, (4) available HTA decision in at least one major jurisdiction (e.g. United Kingdom, Germany, France), (5) approval by the European Medicines Agency after January 1st, 2020.
- A detailed prompt was engineered to instruct three leading genAI models, ChatGPT-5, Gemini 2.5 and Grok 4, to identify analogues which meet the criteria. Different sub-models were tested. The prompt was run twice in each sub-model.
- The model outputs were analyzed for intra-model reliability, accuracy, and time taken to retrieve results compared to a reference analogue set (previously curated by experienced Maple consultants) which met the five criteria.

RESULTS

- The human-led reference analogue set was 100% accurate. The average accuracy of the genAI models ranged from 57% (Flash Regular, Gemini 2.5) to 78% (Pro Deep Research, ChatGPT-5) (Figure 1).
- Higher levels of accuracy of analogue identification were generally associated with a longer time taken to complete the analysis (Figure 2).
- All 9 sub-models performed poorly on intra-model reliability. Pro Regular Gemini 2.5 demonstrated the highest level of consistency between the two runs across all 9 sub-models, although the intra-model reliability was low (Table 1).
- The average number of analogues retrieved by the genAI models ranged from 0 (Auto Regular ChatGPT-5) to 11 (Pro Deep Research, ChatGPT-5). In comparison, human identification yielded 12 analogues as the accuracy benchmark.
- The most frequent rationale for genAI analogue misclassification was inclusion of "*not orphan*" analogues. ChatGPT-5 configurations demonstrated low to moderate error frequencies, with the most consistent overall performance and stability across runs. Grok 2.5 models displayed errors mostly similar to ChatGPT-5 but with higher "*non-surrogate*" error rates. Gemini 4 configurations showed the highest average error rates across most domains (Figure 3).

FIGURE 1. AVERAGE ACCURACY OF ANALOGUE IDENTIFICATION PER MODEL

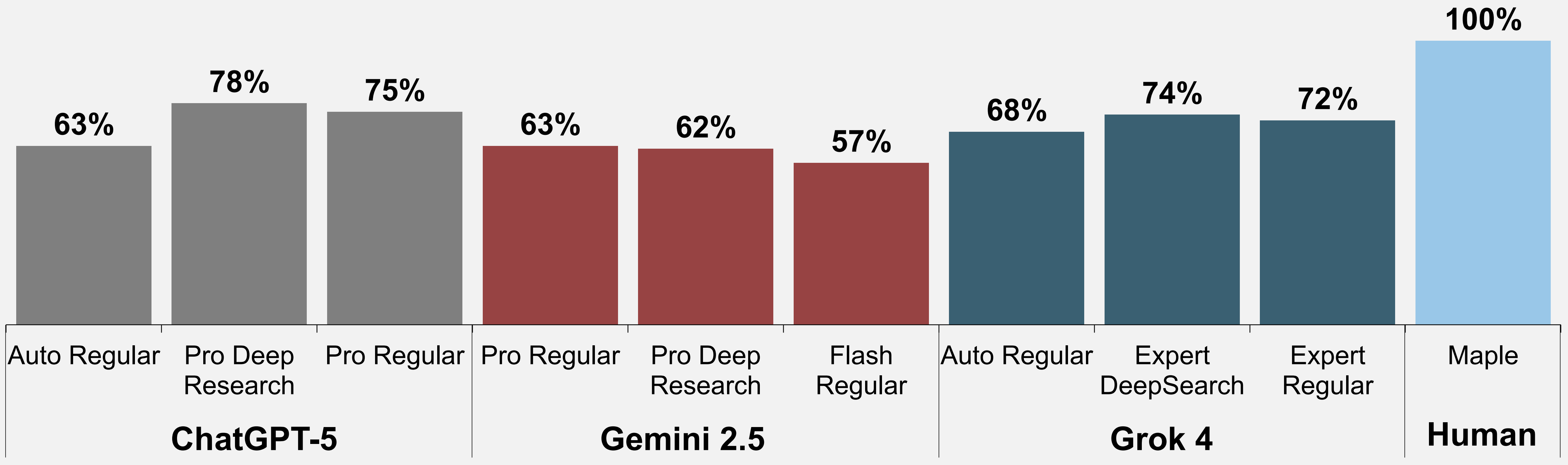


TABLE 1. INTRA-MODEL RELIABILITY

		COHEN'S KAPPA
ChatGPT-5	Auto Regular	-1.00
	Pro Deep Research	-0.29
	Pro Regular	0.00
Gemini 2.5	Pro Regular	0.22
	Pro Deep Research	-0.15
	Flash Regular	-0.92
Grok 4	Auto Regular	-0.29
	Expert DeepSearch	0.00
	Expert Regular	-0.62

FIGURE 2. ACCURACY OF ANALOGUE IDENTIFICATION AND TIME TO COMPLETE

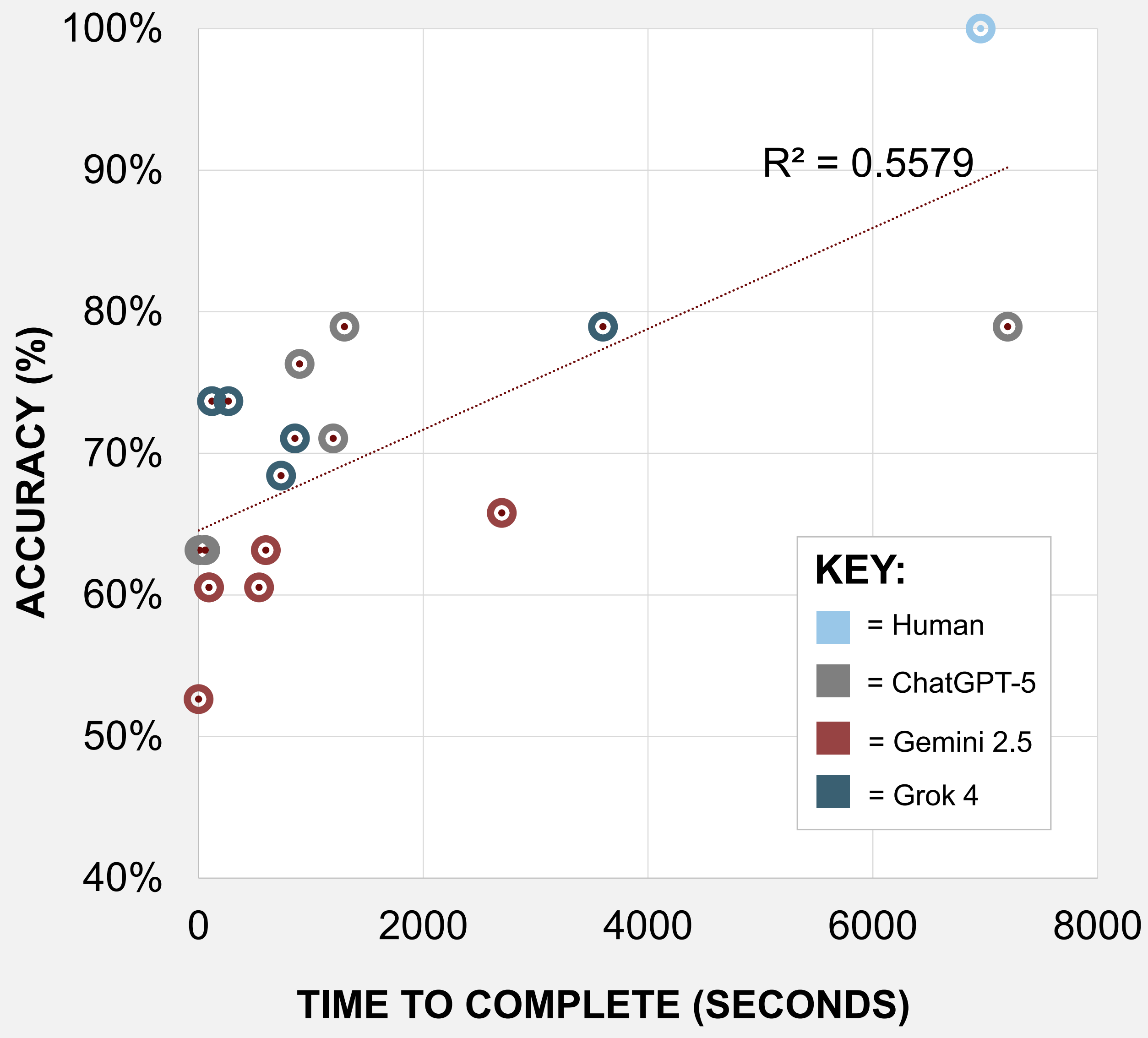
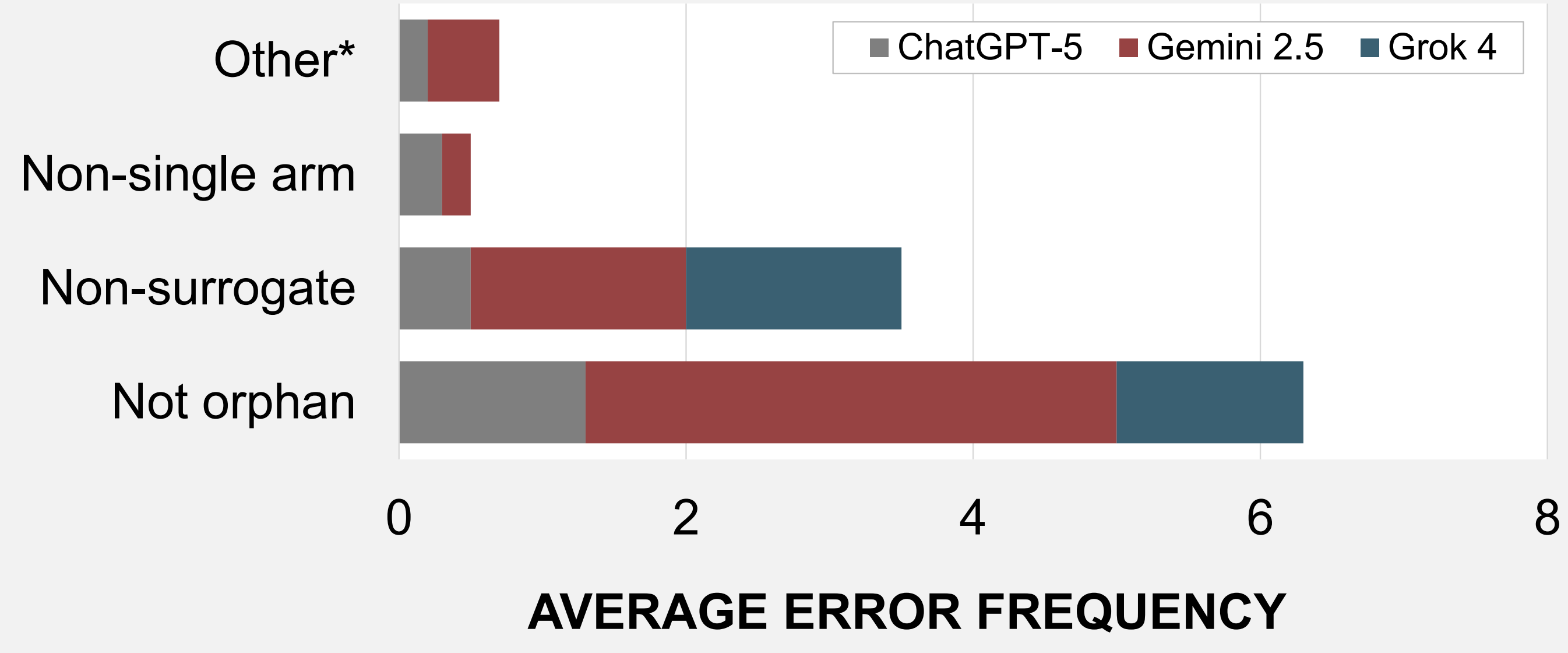


FIGURE 3. ERROR FREQUENCY BY MODEL FAMILY AND COMMON REASONS FOR EXCLUSION

	CORRECTLY IDENTIFIED ANALOGUES	AVERAGE ERROR FREQUENCY
ChatGPT-5	3.7 (31%)	2.3
Gemini 2.5	2.8 (24%)	5.8
Grok 4	4.0 (33%)	2.8

* Analogues outside the pre-specified approval window or not approved by EMA


ChatGPT-5
Gemini 2.5
Grok 4
Other*
Non-single arm
Non-surrogate
Not orphan
AVERAGE ERROR FREQUENCY