

case 11 courvoisier llm

Participant-visible method context. Results, discussion, captions, and the target figure are excluded.

2.1. Abstracts analyzed The included abstracts are primary reports of RCTs in rheumatology published between 2009 and 2022, used in a previous study.²⁶ They were used for 2 tasks: - Binary classification where the abstracts were classified based on interventions testing a drug or not (nondrug). - Extraction of the number of randomized patients in the RCTs, or reporting as “missing” if the information was not present. For the second task, humans also reported if the number was given as total or by arm, thus requiring calculations.

2.2. List of LLMs considered In all, 8 openly available LLMs of less than 10 billion parameters were considered: - Llama 3 8B from Meta AI²⁸ - Ministral 8B from Mistral²⁹ - Qwen 2.5 7B from Alibaba Qwen³⁰ - Yi 1.5 9B from 01.ai³¹ - Gemma 2 9B from Google³² - Deepseek 7B from Deepseek AI³³ - Phi 3 small (7B) from Microsoft³⁴ - Aya expanse 8B from Cohere For AI³⁵ We used the instruct version of these models, which were downloaded from Huggingface and used in python through the transformer library. Two state-of-the-art models were considered: - Llama 3.3 70B, from Meta AI³⁶ - Qwen 2.5 72B, from Alibaba Qwen³⁷

2.3. The gold standard Both tasks were performed independently by 2 human reviewers (health care professionals: statisticians and epidemiologists). In case of disagreement, adjudication was performed by a third reviewer, yielding the human gold standard.

2.4. The platinum standard When more than 4 LLMs provided a concordant answer that was different from the human gold standard, the human gold standard was revised by a fourth reviewer, who had access to the LLM justifications. The subsequent reference, named the platinum standard, was considered as the reference standard in all future analyses. <https://doi.org/10.1017/rsm.2025.10054> Published online by Cambridge University Press

Courvoisier et al.

Figure 1. Structure of the prompt for the 2 classification tasks. 2.5. LLM inference Let n be the number of LLMs included in the task and $thres$ the agreement threshold. Our framework considered that an answer was valid if provided by at least $thres$ LLMs. If this number of LLMs agreeing was not reached, then the task was considered as nonanswered for the given abstract (Figure 1). For the classification task, agreement was reached if $thres$ LLMs classified the abstract as being of the same category (drug or nondrug). For instance, if $thres$ is 5 and 7 LLMs considered, and 5 LLMs or more classify the abstract as nondrug, then agreement is considered reached. If only 4 LLMs classify the abstract as nondrug, then agreement is not reached, and the task is considered as nonanswered. For the sample size task, agreement was reached if $thres$ LLMs reported the exact same number. n was varied between 3 and 7, and all combinations among the 8 LLMs were tested, leading to 56, 70, 56, 28, and 8 LLM combinations, respectively. $thres$ varied between half n rounded to the lowest integer and n . For example, considering the sample size task, fixing n as 6 and $thres$ as 4, each abstract would be evaluated by a selection of 6 specific LLMs among 8, and the procedure would provide a classification only if 4 models provided the exact same sample size (perfect agreement). This procedure would then be done for each possible draw of 6 LLMs among 8. 2.6. Prompt and parameters All LLMs were given the same prompt and parameters. We set a random sampling with cumulative probability and the most likely next word sampling (top_p and top_k), considering $top_k = 40$ and $top_p = 0.95$ and a temperature of 0.4. A sensitivity analysis varied the top_p , top_k , and temperature parameters, using the Qwen 2.5 7B model. We performed one inference without sampling (greedy search), and, with sampling, the top_p parameter was varied from 0.95 to 0.80, top_k from 40 to 20, and temperature between 0.2, 0.4, and 0.8. The prompt was structured as follows (Figure 1): - Definition of LLM role and objective - Definition of the task and context (what is a drug and nondrug intervention, what is the number of randomized patients) - Definition of the output structure in JSON, with a first entry for an explanation, following the idea of chain of thought structure³⁸ - few shots prompting (5 examples) <https://doi.org/10.1017/rsm.2025.10054> Published online by Cambridge University Press

Research Synthesis Methods

Prompts, python code, inference results, computer configuration, and R code for the analysis are available at https://gitlab.unige.ch/trial_integrity/llm_majority_public. 2.7. Statistical analysis The main performance metric for both tasks is the accuracy, defined as the number of correct answers compared to the platinum reference divided

by the total number of queries. The accuracy between the combination of LLMs and the human gold standard was compared using t-test. For the classification task, recall (also called sensitivity, defined as the proportion of the label properly guessed) and F1 scores, defined as the harmonic mean of the precision and recall (or positive predictive value) are also provided for both categories (drug and nondrug). All analyses and figures were performed in R 4.4.2.39