Category: Clinical Trial Operations (Trial Start-up, Regulatory, Finance, Data Management, IITs) - Work in Progress

Evaluating Different Large Language Model/AI Agents for the Ability to Summarize SUSAR Reports

R. Lush, A. Barnett, A. Putur, A. Lucas, B. Sebsebie, L. Obara

GW Cancer Center

1. Background

The U.S. Food and Drug Administration (FDA) updated their guidance suggesting that Principal Investigators (PIs) "should review all Investigational New Drug (IND) safety reports received from sponsors as a part of the investigator's responsibility to protect the rights, safety, and welfare of trial participants (see § 312.60)." This will increase the burden of investigators at every research site. With the explosion of Artificial Intelligence (AI) tools available, we wanted to evaluate their use in summarizing these reports for investigators.

2. Goals

The goal of this project was to distill summary information from the reports for review by the PIs. We compared the different Large Language Model (LLM)/AI agents available within GW Cancer Center (GWCC) Box for their ability to accurately summarize the information. We also investigated the factors that might lead to higher/lower quality summaries using these tools.

3. Solutions and Methods

We evaluated nine different LLM/AI agents available on the internal Box system at George Washington University (GWU). The LLM/AI agents within GWU Box are institutionally approved for "regulated data" while others available outside of Box are not approved for "regulated data." We used a batch of Suspected Unexpected Serious Adverse Reaction (SUSAR) reports from a National Cancer Institute (NCI)-sponsored study that were emailed to the study team. The batch of reports (10) were submitted to each LLM/AI agent using the same prompt. The output summaries were graded by the team based upon a preset group of variables (accuracy, missing information). We recorded the time required to produce the summary as well as other factors that affected the outcome.

4. Outcomes

We determined that Claude 4.5 Haiku produced the most accurate summary of the batch of SUSAR reports and was also the fastest in task completion (<20 seconds). Overall, the time taken to produce the summary spreadsheet ranged from 19 seconds (Claude Haiku) to 2 minutes and thirty-one seconds (Claude Opus 4). GPT-4o did not include all ten reports in the summary and therefore was not deemed fit for purpose. PDF reports that were images of the actual reports could not be summarized by the LLM/AI agents. One LLM/AI agent (Gemini 2.5 Flash) did not produce an output summary after repeated prompts. Many of the LLM/AI agents produced comma separated values (CSV) files along with the summary spreadsheet results, although several CSV files had formatting issues which deemed the file unusable.

5. Lessons Learned and Future Directions

Different LLM/AI agents have properties that affect their ability to provide accurate summaries of a batch of SUSAR reports. We will continue to evaluate and optimize the prompt used during this process to optimize the output.