Contribution ID: **168** Type: **Presentation**

Evaluating the Effectiveness of an Open-Source Large Language Model in Drafting NIH Data Management Plans

Tuesday, October 14, 2025 11:41 AM (11 minutes)

As funding agencies increasingly emphasize responsible data stewardship in alignment with the FAIR (Findable, Accessible, Interoperable, and Reusable) principles, Data Management Plans (DMPs) have become a core requirement in research proposals. This emphasis reflects a growing recognition that data serves as the foundation for scientific discovery and progress. Since 2023, the National Institutes of Health (NIH) requires the inclusion of a DMP in all grant applications, encouraging investigators to proactively think about how scientific data will be managed, preserved, and shared throughout the course of the research. Creating a high-quality, policy-compliant DMP is essential but remains a complex and time-consuming endeavor, particularly because researchers are often not trained in data management and/or lack support. Tools like DMP Tool, DMPonline, Data Steward Wizard are available to help but are limited to providing guidance and samples. Recent advancements in large language models (LLMs) present promising opportunities to streamline and automate aspects of data management planning. We envision a workflow where a DMP is drafted with the help of an LLM, given basic information about a grant proposal and data to be collected. The draft could then be reviewed and revised by researchers, effectively reducing time and complexity in the process. In this work, we evaluated the performance of an open source LLM in creating drafts of DMPs that are compliant with the NIH guidelines. The goal was to investigate if an LLM could be used off-the-shelf for DMP drafting without undergoing fine-tuning or other domain-specific performance improvements.

We assessed the capabilities of Meta's Llama 3.3 70B, which is one of the most advanced open-source LLMs with demonstrated high-performance for writing tasks. We tested its performance in reproducing 26 DMP examples provided by the NIH from previously funded proposals that cover different study and data types. As per the NIH guidelines, each DMP follows a standard structure consisting of 12 sections, each requiring information about different aspects of data management and sharing, including what data types will be collected, the standards that will be followed for formatting data and metadata, where and how data will be shared, and more. Our prompting strategy consisted of a single prompt that includes the name of the NIH Institute/Center where the proposal was submitted, details about the data collection (from human or non-human participants, number of subjects, data types to be collected) and the full NIH-provided DMP template. Since the research strategy was not included in the NIH-provided DMP examples, we included Section 1A of each DMP as an input into the prompt since this section is expected to provide details about data collection. We then compared side-by-side the 12 sections of the NIH DMP generated by the LLM with the related NIH-provided example to assess content accuracy and completeness using SBERT-semantic Similarity score.

Usually, a score of 0.7 or higher is considered a strong indication of semantic similarity, while scores above 0.8 suggest a very high degree of similarity. Our results show that the SBERT-semantic similarity score is the highest on average for Section 1A (0.86). This is expected since content from Section 1A was included as part of the prompt. The similarity score is low for the other 11 sections of the DMP, ranging between 0.46 and 0.64 across all 26 DMPs. Section 4A, which requires information about the repository where data will be archived, had the lowest similarity score on average (0.46). The second lowest was Section 1B, which requires details about the data that will be preserved and shared along with the rationale for the decision. The highest score on average (excluding section 1A) was for sections 4B, 4C, and 6 (all around 0.63-0.64), which require details about how data will be made findable, when data will be shared, and who will manage compliance with the DMP, respectively. Looking at the DMP-specific score (average score across all 12 sections), we observed the highest scores for a DMP about clinical and genomic data (0.67) and a DMP about survey and interview data

(0.65). The lowest scores were observed for a DMP about non-human genomic data (0.55) and Hela cell whole genome sequence data (0.56).
These preliminary findings suggest that even a powerful open-source LLM like Llama 3.3 may not be ready off-the-shelf to use for DMP drafting. Performance improvements are likely needed in certain areas of data management such as best practices for sharing data, and on specific data domains such as non-human genomic data. Such improvements can be achieved through several approaches that will be tested in future studies, such as segmented prompting, retrieval-augmented generation (RAG), and domain-specific fine-tuning. We may also find that additional details from the researcher are needed in the prompt that may help the LLM draft the DMP correctly for that particular study we will be exploring. Future work will also assess the performance of additional LLMs, including commercial ones, to achieve a thorough benchmarking of LLMs performance off-the-shelf prior to any improvements. To ensure more robust evaluation, we also plan to incorporate expert human reviews alongside automated metrics, offering deeper insight into the quality, completeness, and usability of LLM-drafted DMPs, as it is possible that the generated DMPs are perfectly valid options for a study in that domain, even if it didn't happen to match the examples in this case. This work has practical implications for research data managers, librarians, grant administrators, tool developers, and funders interested in leveraging LLMs to support compliance with evolving data-sharing policies.

Primary authors: ZEINALI, Nahid (FAIR Data Innovations Hub, California Medical Innovations Institute, San Diego, California, United States of America, 9212); Dr PATEL, Bhavesh (FAIR Data Innovations Hub, California Medical Innovations Institute, San Diego, California, United States of America, 92121)

Co-authors: Dr HOFSTEIN GRADY, Becky (California Digital Library, University of California Office of the President, Oakland, CA, United States of America, 94607); Dr PRAETZELLIS, Maria (California Digital Library, University of California Office of the President, Oakland, CA, United States of America, 94607); Dr RILEY, Brian (California Digital Library, University of California Office of the President, Oakland, CA, United States of America, 94607)

Presenter: Dr HOFSTEIN GRADY, Becky (California Digital Library, University of California Office of the President, Oakland, CA, United States of America, 94607)

Session Classification: Presentations Session 3: Rigorous, responsible and reproducible science in the era of FAIR data and AI

Track Classification: SciDataCon2025 Specific Themes: Rigorous, responsible and reproducible science in the era of FAIR data and AI