PharmaSUG 2025 – Paper [AI-209]

Integrating Generative AI into Medical Writing: Building an Interactive Drafting and Search Framework

Tadashi Matsuno, Yuki Yoshida, Yoshitake Kitanishi, Shionogi & Co., Ltd.

Abstract

The increasingly lengthy timelines of clinical development have underscored the growing need to shorten the medical writing process, particularly for key documents such as Clinical Study Reports (CSRs) and protocols. At Shionogi, we have been pursuing a 15–25% improvement in drafting efficiency through the deployment of an in-house Generative AI solution on a secure cloud platform. This solution leverages organizational templates and historical trial documents along with a Large Language Model (LLM). Nevertheless, due to risks such as so-called “hallucination,” it is unrealistic to rely solely on LLMs for producing perfect, submission-ready documents. We have therefore developed a web-based front-end application that enables medical writers to interactively review and refine AI-generated text, focusing on using AI for initial drafts and leaving the finalization to human oversight.

In parallel, we are advancing a document search application that vectorizes CSRs, protocols, and regulatory meeting notes. This application enables natural language queries to be performed on large volumes of historical documents, which we believe will benefit various teams involved in the development process. By embedding Generative AI and enhanced search capabilities at earlier stages of clinical development, we aim to transform the efficiency of document creation and retrieval. This paper describes our technical framework, presents how we integrate human input into automated processes, and discusses the lessons learned and challenges encountered so far.

1. Introduction and Background

In the realm of clinical development, the creation of regulatory and clinical documentation—such as Clinical Study Reports (CSRs) and protocols—often becomes a significant bottleneck in drug development timelines. Drafting, reviewing, and finalizing these documents can take several months. In one study, the average time from database lock to completion of the final CSR was reported to be approximately 83 days (around 12 weeks) [1]

To address this challenge, Shionogi has initiated a project to enhance medical writing efficiency, aiming for a 15–25% reduction in overall drafting time by integrating Generative AI technology early in the clinical development process.

Historically, external outsourcing has been a common strategy to manage limited human resources in medical writing; however, this approach has not led to a dramatic reduction in document creation times. Furthermore, the need for rapid and precise revisions has underscored the demand for internal optimization. In discussions between our Data Science and Medical Writing departments regarding these challenges, we identified opportunities to reduce the time medical writers spend gathering relevant information and generating initial drafts. We concluded that rapidly advancing generative AI technology, combined with a secure cloud environment, could help us achieve this efficiency gain.

In this paper, we provide an overview of our approach, which combines an interactive drafting interface and an advanced search application. Although both systems are still under development, they are being designed to work with in-house data while meeting the security and compliance requirements necessary for eventual production use.

2. Project Overview

Our project is driven by two primary goals. First, we aim to automate the initial draft creation for essential clinical documents—such as Clinical Study Reports (CSRs) and protocols—by employing a Generative AI system that references existing organizational templates and historical materials. Second, we plan to develop an internal application that enables cross-document searches, allowing medical writers and other development personnel to efficiently locate previously created content.

To achieve these objectives, we have initially focused on three key applications, with the intention of continuously refining and expanding them over time. The first is a CSR generation application, which uses our Large Language Model to interactively produce a preliminary report structure and text. This application references our in-house CSR template as well as existing protocols and figures from past clinical trials. The second application—based on the same core framework—supports protocol generation by referring to our internal protocol template, the Investigator's Brochure, and a protocol synopsis. Finally, we are creating a document search application, which leverages vectorization and natural language processing to index and retrieve data from a comprehensive library of documents.

Rather than producing fully automated outputs, our approach leverages Generative AI primarily for the interactive creation of initial drafts, while final reviews and edits remain the responsibility of human medical writers. Crucially, any AI-generated text is strictly intended as a preliminary draft and is never used verbatim for regulatory submissions. Instead, each draft undergoes comprehensive human oversight to ensure accuracy, clarity, and compliance with regulatory standards. Taken together, these initiatives represent a significant step toward more efficient medical writing and information retrieval, all underpinned by robust security and adherence to regulatory requirements.

3. Technical Architecture

Our technical solution is hosted in a secure private cloud environment that has undergone rigorous security assessments, and it is built independently from our existing pharmaceutical development document management system. Within this environment, organizational templates, past study documents, and relevant medical literature are stored in an encrypted repository, and strict access controls are enforced to comply with data privacy regulations.

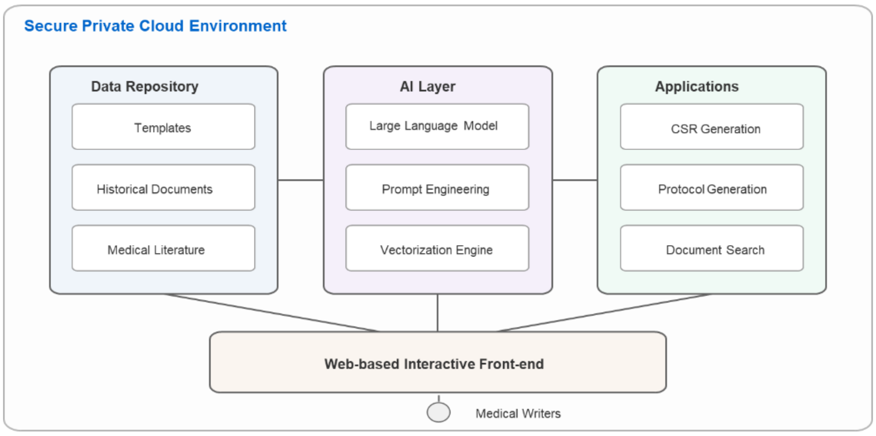

Figure 1 Technical Architecture of Generative AI System for Medical Writing

The Generative AI models interact with these data sources through a carefully designed prompt engineering process. Because Large Language Models can sometimes produce inaccurate or irrelevant content—often referred to as “hallucination”—we employ a multi-faceted approach informed by extensive trial runs:

Section-by-Section Generation

Rather than generating the entire document in a single pass, we create text for each section (or item) individually. This not only reduces ambiguous or overly broad outputs but also optimizes the token count supplied to the Large Language Model.

Template and Reference Mapping

For CSR creation, we pre-map our in-house CSR template to specific protocols and figures that might serve as references, storing this information in a database. By matching user requests against this mapping, we can efficiently extract the most accurate text and reduce the risk of unsuitable content. The same logic applies to protocol creation, where we also map the protocol template, the protocol synopsis, and the Investigator’s Brochure in order to retrieve essential information quickly and precisely.

Ideal Text Prompt Design

We integrate “ideal” text samples—drawn from high-quality past documents—into the prompts for each section. In close collaboration with medical writers, we iteratively tune these prompts to improve clarity and consistency in the AI-generated text.

Building on these steps, the generated text is refined through a “human-in-the-loop” process within the front-end application, where medical writers enhance its quality and verify accuracy and regulatory compliance before moving on to subsequent review stages. Additionally, the AI workflow includes logging and monitoring mechanisms that record each generation request and its output, thereby enabling detailed investigations of any issues that arise and ensuring transparency throughout the content creation process.

Taken together, this architecture meets stringent security and regulatory standards while enabling more efficient and higher-quality drafting of documents.

4. Front-End Application

The front-end application is a web-based system designed to provide medical writers with an accessible, user-friendly, and interactive environment. For each section of a document, the application displays AI-generated text that writers can either refine by adjusting the prompt and requesting a new generation, or edit directly. When editing text, they can seamlessly launch Microsoft Word within the interface, allowing them to retain their familiar day-to-day drafting workflow. Additionally, the application leverages asynchronous JavaScript to minimize page reloads, reducing friction and ensuring a more fluid experience for medical writers.

By capturing and storing each update, the platform supports an iterative cycle of text refinement without losing track of earlier versions and subsequent modifications. Once the drafting process is complete, a consolidated document is generated for review by experienced medical writers, who confirm its regulatory compliance and scientific accuracy. Crucially, our generative system is designed not to replace human expertise, but to augment and enhance it.

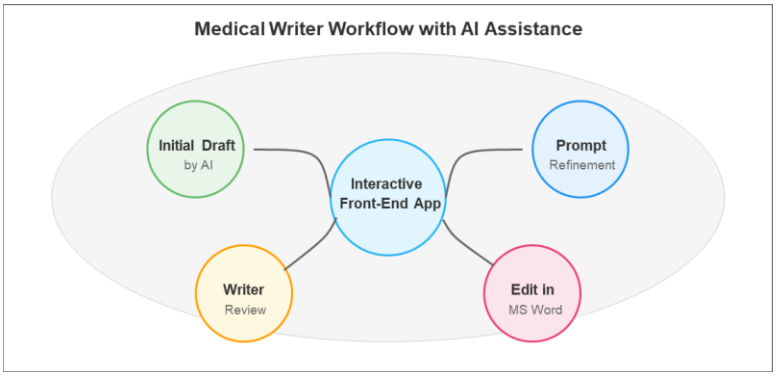

Figure 2 Interactive Drafting Workflow with Human-in-the-Loop Process

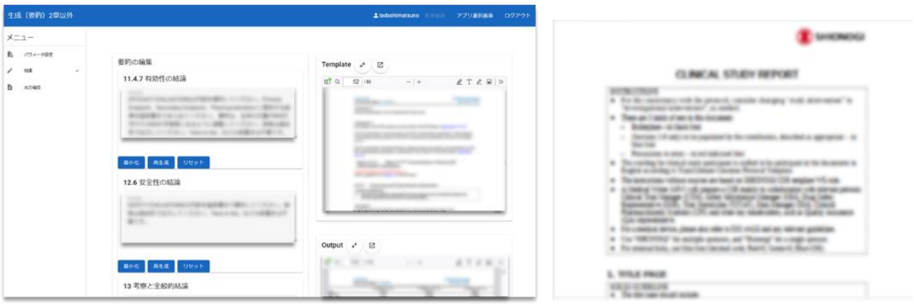

Figure 3 User Interface

5. Document Search Application

In parallel with our drafting initiatives, we have been developing a search application to help medical writers efficiently navigate a broad range of internal documents—including CSRs, protocols, and regulatory meeting notes—used in clinical development. In practice, medical writers frequently rely on a corporate document management system; however, finding specific references can be challenging if keywords must match exactly or if it is unclear which part of a lengthy file is relevant. Over time, these limitations cause significant delays. Medical writers therefore need a system that can handle more flexible, natural-language queries, even when the wording may be ambiguous.

To address this requirement, our solution employs a hybrid approach that combines semantic search with traditional keyword-based queries. Documents are split into sections or “chunks” based on a fixed character count, and each chunk is transformed into an embedded vector within a high-dimensional space. When a user enters a natural-language query, the query itself is likewise vectorized; the system then retrieves relevant chunks according to similarity scores while also applying keyword matching to ensure precision for specific terms. Preliminary evaluations indicate that this method substantially improves recall and precision compared to keyword-only searches.

We continue to refine our indexing strategies, optimize embedding algorithms, and carefully select the most suitable vectorization model by evaluating retrieval accuracy for commonly asked questions and analyzing user feedback. In addition, we have integrated a Large Language Model (LLM) that uses the retrieved chunks to generate concise, targeted answers for the user. Each cited chunk and its similarity score are displayed so that users can directly verify the source material. Crucially, if the system finds no relevant chunks, it declines to provide an answer, minimizing the risk of AI "hallucination."

This enhanced search capability is designed to serve the entire development organization. By rapidly connecting users with templates, data summaries, or regulatory communications, it addresses the day-to-day inefficiencies of keyword-only searches and fosters a more efficient research and writing process.

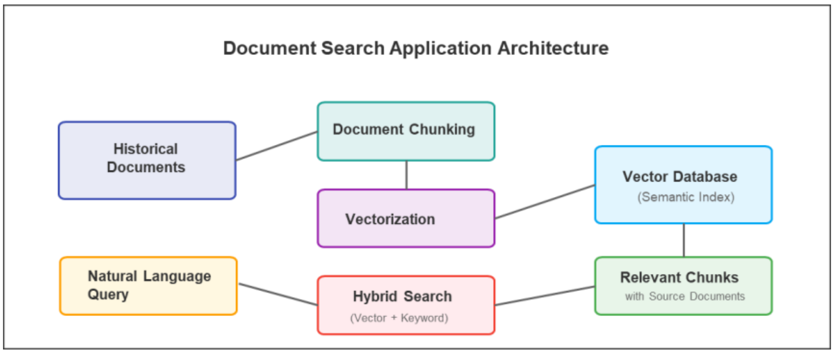

Figure 4 Vector-based Semantic Search System for Clinical Documents

6. Current Pilots and Observations

Our project is still in the development phase, and we have yet to conduct a definitive quantitative evaluation of its day-to-day impact on medical writing. Nevertheless, the early pilot outcomes—primarily focused on CSR (Clinical Study Report) generation and protocol generation—indicate promising potential for increasing efficiency in document preparation. In addition, we have begun assessing a document search application that addresses some of the broader needs of clinical development teams.

CSR Generation

In a pilot test involving a small group of medical writers, our CSR generation tool showed notable efficiency gains for sections that largely rely on existing protocol text (Sections 1–8). Early feedback suggests that these portions can be drafted within a few minutes and reach a quality level suitable for further editing. Some chapters that depend heavily on figures and tables initially revealed inaccuracies in data extraction, but improvements to the underlying Large Language Model (LLM) have enhanced its ability to interpret and transcribe figures with minimal error.

Despite this progress, certain chapters remain challenging. For instance, Section 11—which covers efficacy assessments—varies significantly from one study to another, making it difficult to craft highly precise prompts in advance. Consequently, medical writers still fine-tune prompts or compose text manually for these specialized sections. Section 9 (Study Design), while comparatively straightforward given its reliance on protocol excerpts, often exhibits minor yet time-consuming issues such as tense inconsistencies or overly verbose text. These shortcomings necessitate final reviews by medical writers to ensure clarity and correctness.

Protocol Generation

We have also developed a prototype protocol generation application, complete with a basic front-end interface. Given that protocol documents typically follow a well-defined structure, an extensive analysis of past protocols has resulted in generally acceptable drafts for each chapter. That said, the quality of the final output depends significantly on the synopsis prepared by the clinical trial team. Moving forward, we plan to study how variability in synopsis quality affects the generated text.

Furthermore, we are exploring the possibility of building a standalone solution to generate the synopsis itself—an advancement that would push automation to another level. However, achieving this goal is highly challenging: synthesizing a synopsis of sufficient quality would require the integration of diverse internal and external data sources, including global clinical trial analyses, drug–drug interaction details, and comparisons with existing medications. Successfully merging these complex data streams into a coherent synopsis demands careful design and robust data handling capabilities.

Common Challenges: Formatting and Beyond

Both CSR and protocol generation, as well as the new document search application, face certain shared obstacles. One recurring issue in generation tasks is the loss of formatting details, such as superscripts, subscripts, and table layouts, when text is fed into the LLM. Although each individual error may be small, the cumulative effect can become time-consuming to correct. To mitigate this, we have developed Python-based scripts that automatically resolve the most frequent formatting inconsistencies, leaving only the more complex adjustments to manual intervention by medical writers. While not a complete end-to-end automation, this strategy helps maintain high-quality deliverables without imposing excessive manual work.

Document Search Application

In parallel to document generation, we have been piloting a document search application for clinical development materials. This system is built around a simple Python-based front-end framework and is currently being evaluated by medical writers for search accuracy and overall usability. Unlike conventional document management systems that rely primarily on keyword-based retrieval, our approach supports natural language queries—an aspect that has been well-received thus far.

However, the application still faces obstacles specific to medical contexts. For example, in Japanese clinical documentation, “盲検” should directly map to “blind” in English, but the embedding model does not always interpret these pairs as closely related. One major reason is that we are currently using a general-purpose embedding model. Looking ahead, we plan to explore developing a model specifically tailored to Japanese clinical development documents, which we believe will offer more accurate semantic representation and ultimately improve retrieval performance.

From a user interface (UI) standpoint, requests from medical writers and other clinical development stakeholders have been diverse. Some teams prefer a lightweight interface that simply displays an index of relevant search results (similar to many existing document management systems), while others would like an LLM-driven interface that provides direct answers to user queries. Because our user base extends beyond medical writers to include a wide range of clinical development professionals, we plan to accommodate flexible UI configurations. Determining which features are highest in priority will require ongoing dialogue with end-users so that we can align system capabilities with real-world needs.

7. Summary and Future Directions

The pilot applications discussed thus far demonstrate how Generative AI can streamline both medical writing and document retrieval within the clinical development process. By focusing on a “human-in-the-loop” workflow, we leverage automation for initial drafts while preserving—and indeed emphasizing—the critical oversight and specialized knowledge of medical writers. It is important to clarify that the AI-generated content serves only as a preliminary draft; it is not intended for direct use in regulatory submissions. This distinction helps avoid misunderstandings and ensures that any final document is thoroughly reviewed and validated by qualified experts.

Looking ahead, we plan to pursue the following key initiatives:

Quantitative Evaluation

While initial results have been encouraging, a robust, data-driven assessment of time savings and productivity gains is still pending. We aim to collect systematic metrics (e.g., average drafting time per section, number of revisions needed) to quantitatively measure the overall impact. Additionally, building an automated or rule-based mechanism to evaluate certain aspects of draft accuracy will be essential. Since it would be impractical to have human reviewers inspect every line of AI-generated text at scale, we intend to incorporate AI-driven quality checks to supplement manual review.

Continuous Tuning of Large Language Models

Some sections with highly specialized or variable content (e.g., efficacy assessments, atypical study designs) have highlighted the need for more sophisticated prompting and ongoing model refinement. Our goal is not to achieve perfection at the outset, but rather to incrementally improve the LLM through real-world usage logs and iterative feedback. By monitoring how writers interact with the system on a daily basis, we can adjust prompts, retrain the model if necessary, and ensure that it adapts to evolving clinical documentation requirements.

Development of a Japanese-Specific Embedding Model

For the document search application, we aim to tackle the semantic gaps between Japanese and English medical terminology. Creating or fine-tuning an embedding model specifically trained on Japanese clinical documents could yield more accurate retrieval results—especially for domain-specific phrases like "盲検," which should map closely to "blind" in English but are not always recognized as such by general-purpose models.

Synopsis Generation and Complex Data Integration

Extending our protocol generation application to include automated synopsis creation is a potentially significant step toward further reducing workload for clinical teams. However, this requires integrating complex data sources—ranging from global trial results to drug–drug interaction databases and comparisons with existing therapies—into a coherent and scientifically valid summary. We intend to research and prototype data pipelines that can support such a comprehensive and reliable synopsis generation process.

Flexible Front-End for Document Search

Users across different functions—medical writers, regulatory affairs, clinical operations—have varying needs. Some prefer a simple index-based interface resembling existing document management systems, while others seek LLM-driven, real-time query responses. Designing a front-end that accommodates these diverse requirements remains a priority. Through ongoing dialogue with stakeholders, we will identify and prioritize features for progressive development and deployment.

Security and Compliance

As our tools progress toward production use, we will continue to enforce strict security measures and regulatory compliance—a necessity in pharmaceutical development. This includes maintaining rigorous audit trails, adhering to data governance policies, and deploying monitoring mechanisms that track and validate AI outputs to ensure responsible use.

By continuously refining these tools in close collaboration with medical writers and other stakeholders, we aim to strike an optimal balance between speed and quality for essential clinical documents. At the same time, by expanding the utility of vectorized search and natural language querying, we hope to cultivate a more agile, data-driven culture across our organization.

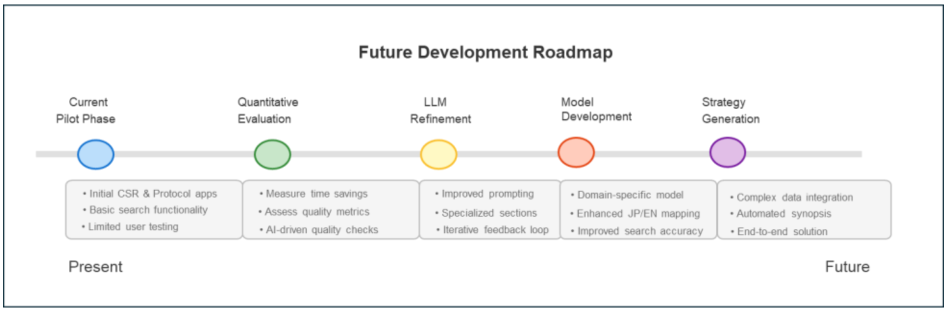

Figure 5 Roadmap for Future Development of AI-Enhanced Medical Writing System

8. Conclusion

Generative AI holds considerable promise for accelerating and improving the drafting of key documents such as CSRs and protocols—especially those adhering to standardized formats. Early pilots suggest meaningful gains in drafting speed and document retrieval, while underscoring the ongoing necessity of rigorous human review to ensure accuracy, regulatory compliance, and scientific integrity. Importantly, the text generated by AI is intended solely as a temporary draft, not as a submission-ready document; human expertise is paramount to finalize and validate content before it reaches any regulatory body.

Nevertheless, challenges remain before broad-scale adoption can be realized. Addressing highly variable study sections, preserving formatting details, and bridging linguistic gaps will require continuous innovation and careful strategy. Additionally, integrating these AI-driven solutions into day-to-day clinical workflows calls for thoughtful change management and systematic validation—especially in an industry that places a premium on data integrity and patient safety.

Despite these hurdles, the rapid advances in AI research and growing familiarity with AI tools among clinical professionals strongly indicate that Generative AI will play an increasingly central role in medical writing. By refining our methods, improving the technology, and drawing on feedback from those who use it most, we can work toward a future in which authors devote less time to mechanical or repetitive tasks and more time to critical scientific reasoning. Ultimately, this shift stands to benefit both the speed and quality of clinical documentation, and by extension, the patients who depend on timely and effective new therapies.

9. Acknowledgments

The authors wish to thank all colleagues at Shionogi, particularly the Medical Writing team and the Development organization, as well as our partners at Deloitte Tohmatsu Consulting and Hitachi, Ltd., for their collaboration and support throughout the course of this project. Their continuous input and willingness to adapt to new workflows have been essential to our ongoing initiatives.

10. Contact Information

Comments and questions are always welcome. Please contact the authors at:

- **Tadashi Matsuno**
  Shionogi & Co., Ltd. (Data Science Department)
  tadashi.matsuno@shionogi.co.jp
- **Yuki Yoshida**
  Shionogi & Co., Ltd. (Data Science Department)
  yuuki.yoshida@shionogi.co.jp
- **Yoshitake Kitanishi**
  Shionogi & Co., Ltd. (Data Science Department)
  yoshitake.kitanishi@shionogi.co.jp

References

[1] Clinical Researcher—September 2020, Volume 34, Issue 8.