innovate GenAI Use Case

enago Life Sciences

Leveraging GenAI for Medical Writing

BACKGROUND

The Context

Effective medical writing requires a combination of scientific expertise and effective writing skills. Medical writers face numerous challenges when drafting reports and documents. These challenges include dealing with complex terminology, maintaining clarity and precision, adhering to grammar and style rules, and ensuring that the document is appropriate for its intended audience while meeting regulatory and ethical standards. In addition, although individuals from various backgrounds can become medical writers, they often acquire additional training and knowledge in medical and scientific writing to excel in their roles.

Some common writing problems encountered in medical writing are as follows:

- Consistent and accurate use of complex medical terms and phrases.
- Writing for different target audiences, mandating writing at different levels of complexity and nuance. Target readers may be scientists, healthcare professionals (HCPs), regulatory affairs professionals, patients, or lay readers.
- Ensuring comprehension and readability for individuals with different backgrounds (e.g., HCPs and lay audiences).
- Ensuring grammatical precision, accurate spelling, and consistency, given the complexity of sentence construction and varied terminology in medical writing.
- Challenging turnaround times with iterations to finally deliver ready-to-publish manuscripts.

Generative AI models are being increasingly adopted for a variety of tasks in medical writing for stakeholders in pharmaceutical companies, scientific research, health care, and medical affairs. Some models can help writers by reducing the time spent on initial-draft creation, revision, and proofreading. These models can also effectively distill vast amounts of clinical data for insights generation, making it easier for medical professionals to derive meaningful conclusions. Generative AI can swiftly summarize and highlight key findings from a multitude of research papers, ensuring that biopharma and healthcare professionals stay updated with the most recent

advancements. This greatly benefits those conducting literature surveys and reviews. AI-driven tools can expedite the process of unveiling novel patterns or trends in medical data, which is invaluable when exploring new therapeutic avenues. These models can also bring to the fore intricate relationships and interactions in seemingly disconnected information, aiding the different stakeholders in understanding the underlying mechanisms of action and predicting suitable candidates or unforeseen interactions.

Given the transformative potential of these tools in the pharma and healthcare sectors, we developed a domain-specific generative AI model to directly tackle the prevalent writing challenges encountered by medical professionals.

The Challenge

Off-the-shelf large language models (LLMs) pose a problem to writers and readers of medical documents, both for academic and nonacademic purposes. These models are trained on nonspecialized data from the internet and hence are not suitable for specialized fields such as the biomedical, biotech, and pharma domains. Furthermore, the lack of security and tendency to produce falsehoods make these models unfit for high-stakes research and content generation, such as that required in the medical domain. Developing bespoke LLMs is the obvious solution, but this is difficult and requires significant investment because of the extensive training time, substantial resource consumption, and associated costs. Moreover, these bespoke LLMs are designed for very specific use cases, thus limiting their adaptability to new domains.

Our primary challenge was to train our language models to recognize and understand the specific terminology, phrasing, and syntax commonly encountered in medical literature and communications. This is crucial to ensure that the models effectively generate content that mirrors the precision and clarity required in the medical field. Additionally, we aimed for our models to provide pertinent revisions for improving clarity and readability, and our focus was on refining the structure, fluency, coherence, and overall readability. We treated this challenge as a generative AI problem to provide contextual grammar correction and to rephrase or restructure sentences to make the text fluent and coherent.

OUR SOLUTION - TRINKA AI MEDICAL WRITING ASSISTANT

As a solution, we created a state-of-the-art medical writing variant of our flagship AI academic writing assistant *Trinka AI*. We refer to it henceforth as *Trinka AI Medical Writing Assistant* (*Trinka MWA*). After fine-tuning and testing, *Trinka MWA* showed high levels of accuracy, confirming its applicability in optimizing medical writing tasks across various workflows.

Trinka MWA brings in benefits of speedy editing, consistency checks, and proofreading, which saves time and cost in quality checks. It will be useful for pharmaceutical companies, laboratories, regulatory bodies, research institutions, health education agencies, and STM publishers, among others, all of whom are involved in documentation, information-sharing, and communication in the domains of life sciences and healthcare.

METHOD

We used proprietary data amounting to millions of sentences from the biomedical, biotech, and biopharma domains for training, fine-tuning, and evaluating the language models for *Trinka MWA*. First, we trained multiple models across various transformer architectures, both encoder–decoder and decoder-only (including models such as BART[1], T5[2], GPT[3], OPT[4]). The model that yielded the best accuracy was then fine-tuned using the hyperparameter optimization technique. Once the *Trinka MWA* model training converged, we carried out evaluations using previously naïve medical and biopharma test data. The test data comprised 10,000 sentences previously edited by professional copy-editors from the abovementioned domains.

We used the ERRANT Scorer5, an automatic scorer used to evaluate grammar error correction systems. The ERRANT Scorer uses the following labels to compute the score.

- *True positive* (TP): correct edit
- *False positive* (FP): incorrect edit
- *False negative* (FN): missed edit

We used the $F_{0.5}$ score as the primary evaluation metric. The $F_{0.5}$ score is a weighted harmonic mean of the precision and recall, which gives more weight to precision than to recall.The $F_{0.5}$ score is ideal for decision-making when selecting a language model for high-stakes use cases where accuracy is paramount. We chose this metric because medical writing needs to be precise.

We then conducted a second round of evaluation to further validate the *Trinka MWA* model's effectiveness in real-world medicine/life-science scenarios. Three SMEs from different therapeutic areas evaluated the *Trinka MWA* revisions. In order to observe its contextual revisions, we chose a test set of 24 paragraphs, comprising 100 sentences. The SMEs used the same metrics used by the ERRANT Scorer (i.e., TP, FP, FN, $F_{0.5}$). Two additional labels were used by the SMEs: *enhancement* and *preferential*. *Enhancements* were changes that improved the text in terms of simplicity, conciseness, or better phrasing, whereas *preferential* revisions were subjective language choices that do not fix errors or enhance the text. This 5-class label system allowed us to understand finer nuances of *Trinka MWA*'s performance.

RESULTS

Table 1 shows the results of *Trinka MWA* after the two rounds of evaluation. The system detected and revised ungrammatical or awkward phrasing (*recall*), and most of its revisions were contextually correct (*precision*).

Test set (sentences)	Evaluation mode	Recall	Precision	$F_{0.5}$
10,000	ERRANT Scorer	66.9%	51.7%	54.1%
100	Human (SMEs)	90.2%	93.7%	93.0%

Table 1. Evaluation results of *Trinka MWA* on two test sets. See footnote† for further details.

†*In terms of the automated evaluation, the* Trinka MWA *model achieves an $F_{0.5}$ score of 54.1% on 10,000 sentences. By contrast, the human evaluation of 100 sentences results in a high $F_{0.5}$ score of 93%. This gap occurs because of the mode of evaluation. In the automated evaluation mode (ERRANT Scorer), the scorer labels the edits by referring to one version of the edited sentence. However, a sentence can be revised in different ways which are correct. Because of this limitation, the ERRANT Scorer marks revisions it has not seen as incorrect even though they are correct. This explains the higher scores in the human evaluation results and the lower ones in the automated results.*

CONCLUSION

Medical writing targeted at medical professionals of varying levels of expertise as well as lay readers requires specialized knowledge of the field as well as a strong grasp of the English language. Complex concepts, interactions, and outcomes need to be communicated in a clear, concise, fluent manner. Hence, documents undergo multiple rounds of revisions, which is time-consuming, resource and expertise intensive, and cost inefficient. We fine-tuned a *t*ransformer model (*Trinka MWA*) trained on text in papers from several academic disciplines in the medical domain.

Trinka MWA's execution was fast and its recall and precision were high, as seen in the $F_{0.5}$ score, making it highly suitable for generating, correcting, and refining medical documents. The high accuracy offered by *Trinka MWA* is a highly desirable characteristic for real-world applications in medicine and the life sciences. The system described here can significantly reduce the time taken for improving grammar accuracy in medical documents so that the writers can focus more on the science and the target audience.

Trinka MWA offers an efficient and cost-savings automation workflow alternative that ensures that high-stakes documents such as clinical study reports (CSRs), medical affairs communications, medico-legal reports, and marketing and educational material can be made human-readable, error-free, and clear.

References

1. Lewis M, Liu Y, Goyal N, et al. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. arXiv (Cornell University). Published online October 29, 2019. doi:https://doi.org/10.48550/arxiv.1910.13461

2. Raffel C, Shazeer N, Roberts A, et al. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. Journal of Machine Learning Research. 2020;21:1-67. Accessed October 15, 2023. https://jmlr.org/papers/volume21/20-074/20-074.pdf

3. OpenAI. GPT-4 Technical Report. arXiv (Cornell University). Published online March 15, 2023. doi:https://doi.org/10.48550/arxiv.2303.08774

4. Zhang S, Roller S, Goyal N, et al. OPT: Open Pre-trained Transformer Language Models. arXiv (Cornell University). Published online May 2, 2022. doi:https://doi.org/10.48550/arxiv.2205.01068

5. Bryant C, Felice M, Briscoe T. Automatic Annotation and Evaluation of Error Types for Grammatical Error Correction. *NA*. January 2017. doi:10.18653/v1/p17-1074

For more information please visit: https://lifesciences.enago.com/