Proceedings of Machine Learning Research 333:834–879, 2026 Conference on Health, Inference, and Learning (CHIL) 2026

Generation of Bilingual Synthetic Clinical Notes for Realistic Data Augmentation

David Seung U Lee* DLEE0880@SNU.AC.KR
Seoul National University, South Korea

Seeun Park* SEEUNP@SNU.AC.KR
Seoul National University, South Korea

Seoyoon Jang SEOYOONJ@SNU.AC.KR
Seoul National University, South Korea

Sunyoung Lee JKSUN5937@SNU.AC.KR
Seoul National University, South Korea

Chaeyoung Chang CHAEYOUNGC@SNU.AC.KR
Seoul National University, South Korea

Sungwook Choi KKUMIR1205@GMAIL.COM
Seoul National University, South Korea

Howard Lee HOWARDLEE@SNU.AC.KR
Seoul National University Hospital, South Korea

Abstract

Synthetic clinical notes offer a promising solution to data scarcity and privacy constraints in clinical natural language processing. However, existing generation approaches often prioritize semantic accuracy while not adequately reproducing the linguistic and structural (i.e., surface) characteristics of real-world clinical documentation, limiting their utility for downstream clinical tasks. In this study, we propose an expert-informed prompt with feedback-loop generation framework to improve the fidelity of synthetic clinical notes across both semantic and surface-level dimensions. Using individual case safety reports from FAERS, we formulated synthetic note generation as a controlled text generation task conditioned on adverse drug reaction descriptions and clinical narratives. We evaluated the performance of the proposed approach by comparing it with other generation strategies (in-context learning and multi-agent generation) and prompting methods (base and expert-informed) under a unified experimental condition. Generation quality was assessed using embedding-based semantic similarity, surface-level statistical and distributional metrics, and blinded human evaluation. The feedback-loop generation framework achieved superior performance across semantic (mean clinical BERTScore = 0.885) and surface-level distributional metrics (token-level Jensen-Shannon divergence = 0.344), producing synthetic clinical notes that more closely resembled real-world clinical notes than other approaches. Expert-informed prompting further improved semantic fidelity and lexical diversity.

Data and Code Availability

In this study, we used two publicly available databases. First, we used individual case safety reports (ICSRs) provided by FDA's Adverse Event Reporting System (FAERS) (Potter et al., 2025). The ICSRs are publicly available on the FAERS Quarterly Data repository[1] and contain information about patient demographics, medications used, adverse reactions, and

* These authors contributed equally

1. https://fis.fda.gov/extensions/FPD-QDE-FAERS/FPD-QDE-FAERS.html

seriousness of the reactions. Additionally, we used MIMIC-IV (v3.1) (Johnson et al., 2023). The dataset is available on the PhysioNet repository[2] and provide deidentified and anonymized electronic medical records of over 265,000 patients. Specifically, we used the "Present Illness" section of Discharge Summaries. Finally, for comparison, we used real-world clinical notes of the patients treated in the endocrinology department at Seoul National University Hospital (SNUH). Due to the sensitive patient health information, the clinical notes are not to be shared. Source codes used in this study are available at https://github.com/dlee0880/multi-lingual-Synthetic-Clinical-Notes.git.

Institutional Review Board (IRB)

Institutional review board approval was obtained from SNUH (IRB No. H-2308-065-1457).

1. Introduction

Leveraging clinical notes for downstream model training comes with several challenges. First, strict privacy constraints arise from the inclusion of sensitive protected health information (PHI) and the inherent difficulty of fully deidentifying these records (Norgeot et al., 2020). Second, effective preprocessing of clinical notes is nontrivial because they are highly heterogeneous and susceptible to data drift, driven by factors such as differences in documentation style, time of document entry, clinical department, and patient characteristics (Kim et al., 2026). In particular, clinical notes frequently contain incomplete or fragmented sentences, domain-specific abbreviations, and shorthand symbols like "↑" and "↓" to indicate clinical findings and record patient status. These stylistic quirks may further complicate generation models, as telegraphic phrasing, clinician-specific shorthand, and limited contextual information make it difficult to accurately reproduce realistic documentation patterns. Also, this issue becomes more serious when there is a frequent code-switching between different languages (e.g., Korean and English) because it adds another layer of complexity to the preprocessing pipeline (Kim et al., 2023). Finally, because clinical notes are written at the point of care to capture every clinical detail, it is difficult to obtain clinical note corpora that are both sufficiently large and of high quality (i.e., focused on essential information without extraneous noise) (Roberts et al., 2009).

To mitigate privacy risks and expand access to high-quality datasets (i.e., having high signal-to-noise ratio), the generation of synthetic clinical notes has gained traction as an effective data augmentation strategy (Biswas and Talukdar, 2024; Alshaikhdeeb et al., 2025; Li et al., 2021). For instance, Kweon et al. generated a dataset of synthetic notes (in the form of clinical narrative) based on the clinical vignettes in case reports (Kweon et al., 2024). Recent progress in large language models (LLMs) has accelerated this trend, making it possible to create large, contextually rich synthetic corpora that improve model performance across a wide range of clinical tasks. Through fine-tuning or prompt-based techniques, LLMs have shown strong capacity to adapt from general text generation to specialized medical applications. For example, Litake et al. employed LLMs to generate synthetic discharge summaries on rare disease cases in order to obtain more balanced cohorts (Litake et al., 2024). However, most previous works focused primarily on accurately representing clinical content (i.e., the semantics) while insufficiently reproducing the stylistic and structural properties of real-world clinical notes (Songsiritat, 2025; Wang et al., 2025; Litake et al., 2024). Moreover, understandably due to the scarcity of clinical notes in the training data during the pre-training of LLMs, synthetic notes may often deviate from genuine clinical writing practices. Such deviation may cause a distributional gap between real and synthetic data that diminishes the performance, robustness, and generalizability of downstream models when tested on actual clinical notes. This discrepancy is particularly detrimental for tasks like adverse drug reaction (ADR) signal information extraction, where model effectiveness critically depends on subtle surface-level characteristics (e.g., linguistic patterns) of clinical notes.

To bridge the gap between real and synthetic notes, we argue that it is crucial to accurately capture the linguistic, stylistic, and structural characteristics (i.e., surface-level characteristics),

2. https://physionet.org/content/mimiciv/3.1/

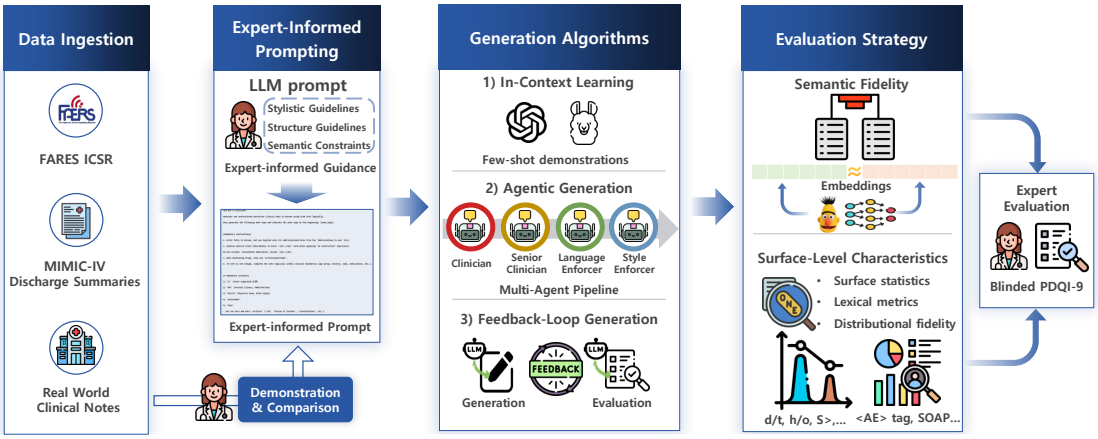


Figure 1: Overall workflow of the study. Three data sources (FAERS ICSR, MIMIC-IV discharge summaries, and real-world clinical notes) were used, from which the healthcare professional included specific guidances into the prompt. Three generation algorithms (ICL, multi-agent system, and feedback-loop) were used to generate synthetic clinical notes. Semantic fidelity and surface-level characteristics of the syntehtic and real-world notes were assessed. Finally, the healthcare professional manually evaluated and rated the quality and realism of the synthetic notes. Abbreviations: ICSRs, individual case safety reports; ICL, in-context learning.

along with semantics, inherent in real-world clinical notes and to incorporate these features into synthetic clinical notes. In this study, we defined semantic characteristics as the clinical validity of the underlying medical content (e.g., accurate documentation of ADR onset and resolution), whereas surface-level characteristics capture the linguistic and structural expression of that content (e.g., abbreviations, telegraphic phrasing, symbols such as "↑" and "↓" and code-switching). Based on this understanding, we propose an LLM-based framework which builds upon expert-informed prompting using iterative feedback-loop. We hypothesized that an expert-prompting strategy that integrates detailed expert-informed linguistic and structural guidelines directly into LLM prompts can enhance the generation of bilingual synthetic notes (i.e., Korean and English) that more faithfully reflect both semantics and surface-level realism. Through series of experiments, we demonstrated that expert-informed prompting using iterative feedback-loop generally outperforms other synthetic note generation algorithms in semantic fidelity and surface-level realism. Moreover, we also showed that using synthetic clinical notes with both semantic fidelity and surface-level realism offers a performance gain in a downstream task of extracting ADR signal information from real-world clinical notes. To the best of our knowledge, this study is one of the first attempts to explicitly incorporate domain-expert stylistic guidance into the prompt design process for the generation of synthetic clinical notes to ensure the fidelity across stylistic and semantic dimensions.

2. Methodology

The overall workflow of this study is summarized in Figure 1.

2.1. Data Sources

In this study, we used two publicly available data sources. First, we collected 116,000 individual case safety reports (ICSRs) from the FDA Adverse Event Reporting System (FAERS) (Potter et al., 2025), which were submitted between 2024 and the first quarter of 2025. Each ICSR represents a single adverse event case report and contains information about 1) patient demographics (i.e., age, sex, and body weight), 2) drug exposures, including primary suspect and concomitant medications, 3) reported adverse events, 4) reporter characteristics, and 5) an optional clinical narrative describing the case. We used ICSRs

as the primary sources of clinical information, from which synthetic clinical notes were generated. Specifically, we structured patient demographics, drug exposures, and adverse event data into JSON format and used them as seed inputs for the synthetic note generation (Appendix A.1).

Additionally, we collected 326,393 discharge summaries from the MIMIC-IV database (Johnson et al., 2023). Each discharge summary details a longitudinal record of the entire inpatient episode from hospital admission to discharge. Core components of discharge summaries include 1) admissions information (i.e., reason for admission, chief complaint, and present illness), 2) hospital course (i.e., confirmed diagnoses, interventions, progression), 3) medications, 4) test results, 5) discharge diagnoses and conditions, and 6) follow-up plans. In contrast to ICSRs, which are semi-structured safety report snippets, discharge summaries are written as fully unstructured clinical narratives. While discharge summaries themselves constitute clinical notes, we argue that they do not fully capture the fidelity of real-world point-of-care documentation, as the discharge summaries tend to lack key surface-level characteristics commonly observed in routine clinical notes, such as shorthand expressions, telegraphic phrasing, and incomplete or fragmented sentences. Accordingly, we used discharge summaries to assess the effectiveness of leveraging unstructured clinical narratives (rather than semi-structured safety report snippets) for generating quasi-realistic synthetic clinical notes, especially with regards to semantic fidelity and surface-level realism (Appendix A.2).

Lastly, we collected 310,067 real-world clinical notes (written in the mix of Korean and English) of the patients treated in the endocrinology department at Seoul National University Hospital (SNUH) between 2010 and 2020. This real-world clinical note corpus (the SNUH corpus, hereafter) was used for two primary purposes: 1) to construct demonstration pool (i.e., candidate few-shot examples) for in-context learning and 2) to serve as a comparator dataset in distributional analyses (see Section 2.5). Specifically for distributional analyses, we randomly selected clinical notes from the SNUH corpus, which met the following inclusion criteria: 1) number of tokens between 200 and 800, 2) number of sentences between 5 and 30, and 3) number of special characters between 0 and 30 per note. The criteria were determined to adjust for the potential confounders in distributional analyses. A total number of 118,971 notes satisfied the predefined conditions (mean token count = 417.63; mean number of sentences = 18.32; and mean number of special characters = 11.88). Moreover, a clinical nurse (S.Y.L) manually created five exemplar synthetic notes for each of the five clinical note categories available in the SNUH corpus (i.e., admission note, inpatient progress note, initial outpatient note, outpatient follow-up note, and discharge summaries). The exemplars were subsequently used as demonstrations for in-context learning. All exemplars were reviewed and validated for clinical plausibility and linguistic authenticity (Appendix A.3). The collection and use of the SNUH corpus were approved by the Institutional Review Board of SNUH (IRB No. H-2308-065-1457).

2.2. Task Formulation

We formulated synthetic clinical note generation as a controlled conditional text generation problem under a fixed pre-trained language model. In this study, the scope of the generation was limited to generating synthetic clinical notes written in English and Korean. Let $x \in \mathcal{X}$ denote the input clinical context, consisting of either an ICSR describing ADRs or a discharge summary describing present illness, and let $y \in \mathcal{Y}$ denote a synthetic clinical note written in Korean and English. Given a pre-trained LLM ($p_\theta(\cdot)$) with fixed parameters θ, the generation task is defined as conditional inference

$$y \sim p_\theta(y \mid x, c) \tag{1}$$

where c denotes a context in expert-informed prompts (see Section 2.3), role-specific agent instructions, or iterative feedback signals (see Section 2.4).

Unlike conventional learning which updates model parameters, our framework operates entirely at inference time and performs generation by manipulating the conditioning context c, without modifying the underlying model weights.

Accordingly, the objective of generation can be considered as generating a clinical note y that accurately and coherently captures the corresponding patient-level clinical information. In addition

to semantic fidelity, the generation of synthetic clinical notes required faithfully reproducing the surface-level characteristics of real-world clinical documentation, including stylistic and structural features, such that

$$y = \underset{y \sim p_\theta(\cdot|x,c)}{\operatorname{argmax}} \; \mathcal{S}_{\text{sem}}(y, x) + \mathcal{S}_{\text{surf}}(y) \qquad (2)$$

where $\mathcal{S}_{\text{sem}}(y, x)$ measures semantic fidelity between the generated note and the input clinical information, $\mathcal{S}_{\text{surf}}(y)$ quantifies agreement with the stylistic and structural surface characteristics required in the prompt (see Section 2.5).

2.3. Expert-informed Prompting

To capture both the semantic fidelity and surface-level characteristics of real-world clinical notes, we used expert-informed prompting strategy (Appendix A.4, A.5). In expert-informed prompting, a healthcare professional-derived context was explicitly integrated into the prompts to enforce stylistic and structural alignment with real-world clinical notes. Specifically, we included the following information as the context: 1) a list of common clinical abbreviations and 2) a list of common expression patterns for ADR (in case of using ICSRs for the generation). These lists included 74 abbreviations (Appendix A.6) and 49 ADR expression patterns (Appendix A.7) most commonly used in the SNUH corpus. All abbreviations and ADR expressions were reviewed and validated by the healthcare professional.

2.4. Generation Algorithms

In this study, we compared three generation algorithms: 1) single-model in-context learning (Appendix A.8), 2) multi-agentic generation (Appendix A.8), and 3) feedback-loop generation. First, the single-model in-context learning approach leveraged few-shot learning through prompt (or context) engineering (Brown et al., 2020), in which varying numbers of exemplars were provided to a single LLM to guide the generation process. This approach represents one of the most widely adopted methods for synthetic clinical note generation and serves as a baseline generation method in our study.

Specifically, under the feedback-loop generation setting, two distinct models (i.e., a *Generator* and an *Evaluator*) were configured in a collaborative and iterative framework (Appendix A.15). The *Generator* produced candidate synthetic clinical notes, while the *Evaluator* (Appendix A.16, A.17) assessed each generated note according to predefined scoring rubric targeting both semantic fidelity and surface-level realism. Specifically, the *Evaluator* assessed (1) appropriate Korean–English code-switching, (2) clinically consistent abbreviation usage, (3) correctness of ADR tag (i.e., <AE>) insertion (in case of using ICSRs), (4) adherence to established clinical writing conventions, and (5) richness and completeness of clinical information. The *Evaluator* then provided structured and detailed feedback, which was incorporated directly into the prompt for the *Generator* in the subsequent generation cycle. The number of iterative regeneration process was set at three rounds after an ablation study to determine the iteration number (Appendix A.18).

Formally, the control context is updated dynamically, such that given an initial control context $c^{(0)}$, generation proceeds as

$$y^{(t)} \sim p_\theta(y \mid x, c^{(t-1)}), \quad c^{(t)} = \Phi\Big(c^{(t-1)}, y^{(t)}\Big) \qquad (3)$$

where Φ denotes a feedback update operator from the *Evaluator*'s assessments of the generated note at step $t-1$.

In contrast to the multi-agent approach, which enabled implicit guidance and goal-setting through role-specific objectives distributed across agents, the feedback-loop algorithm imposes a more explicit and constrained task structure, facilitating controlled refinement of generation quality through targeted evaluation criteria.

2.5. Evaluation Strategy

We evaluated the quality of the synthetic clinical notes along two complementary dimensions: 1) semantic fidelity and 2) surface-level characteristics. First, semantic fidelity was defined as the extent to which the clinical information contained in the original source input (i.e., ICSRs or discharge summaries from MIMIC-IV) is preserved in the generated synthetic notes. To quantitatively assess semantic fidelity, we computed clinical BERTScore (Shor et al., 2023) between vector embeddings of the input and the corresponding synthetic note (i.e., the output), where em-

beddings were obtained using BioClinical ModernBERT (Sounack et al., 2025). The details are summarized in (Appendix A.19).

Additionally, surface-level characteristics were assessed using 1) surface statistics, 2) lexical diversity, and 3) distributional fidelity. Surface-level statistics were used to characterize low-level structural properties of the notes. The assessment of surface statistics included measurement of token length (using tiktoken[3] as the tokenizer), number of sentences (defined as either the use of period after an alphabet or a line-break), and special character frequency (Appendix A.20). Lexical diversity assessed vocabulary-level characteristics. We computed 1) type-token ratio (TTR) (Litvinova et al., 2017) to measure lexical variations, 2) self-BLEU (Alsajri et al., 2024) to measure the diversity of n-grams within the synthetic corpus (Appendix A.21), 3) the extent of bilinguality (i.e., the proportion of Korean characters per document), and 4) medical terminology density (i.e., the proportion of medical terminology per document) using scispaCy (Neumann et al., 2019) for English entities and KoELECTRA finetuned on Korean Bio-Medical Corpus (KBMC) for Korean entities (Byun et al., 2024). Finally, distributional fidelity was assessed by 1) goodness of fit (i.e., R^2) with the Zipf's Law (Saichev et al., 2009) on a log scale and 2) Jensen-Shannon divergence (JSD) (Lu et al., 2020) from the SNUH corpus (Appendix A.22).The former assessed whether word-frequency rank distributions followed natural language scaling behavior. The latter measured the corpus-level and token-level distributional deviation from the real-world clinical notes.

Finally, human evaluation was performed using blinded evaluation by adapting the Physician Documentation Quality Instrument (PDQI-9) (Stetson et al., 2012). The details of the human evaluation are summarized in Appendix A.23.

2.6. Extraction of Evidence of ADR

To evaluate the impact of different data augmentation methods (i.e., algorithms for generating synthetic clinical notes), we conducted a study on the extraction of ADR evidence from real-world clinical notes. Specifically, we assessed the ability of an LLM to perform in-context learning for ADR evidence extraction from real-world clinical notes (n=107) using the synthetic clinical notes as demonstrations. The real-world clinical notes had been annotated for ADR evidence in a previous study conducted by the authors (Kim et al., 2026). For each generation algorithm, three synthetic clinical notes containing ADR evidence were provided as contextual demonstrations. We used LLaMA-3.3-70B as the base model. The performance for this task was assessed using token-level F1, where BioClinical ModernBERT (Sounack et al., 2025) was used as the tokenizer and the token normalization was conducted by removing subword markers.

2.7. Experiment Setup

In the in-context learning setting, we evaluated three models: 1) GPT-3.5-turbo, 2) GPT-4o, and 3) LLaMa-3.1 8B. Each model was provided with varying numbers of demonstrations (k=0, 1, 3, 5) except for agentic generation which was consistently set up at k=0. The demonstrations were selected randomly from the 25 examplars constructed by the healthcare professional. We used GPT-4o as the base model for agentic generation and feedback-loop generation. In all settings, we used expert-informed prompt, while a base prompt (i.e., the prompt without expert guidance on common clinical abbreviations and expression patterns for ADR; Appendix A.25, A.26) served as comparator. Each combination of generation setup was prompted to generate 500 synthetic clinical notes. All hyperparameters for GPT-3.5-turbo and GPT-4o were set at their default values. The hyperparameters used for LLaMA-3.1 8B and LLaMA-3.3 70B are summarized in Appendix A.27. Inferences were made using a single A100 GPU.

3. Results

3.1. Semantic Fidelity

We found that using the feedback-loop generation retained the clinical content of the original document more faithfully than other generation algorithms (Figure 2). Notably, the semantic fidelity of the synthetic notes generated by the iterative feedback-loop algorithm reaches that of Wang et al. (2025) despite the latter being shorter in length and more prone to re-using

3. https://github.com/openai/tiktoken

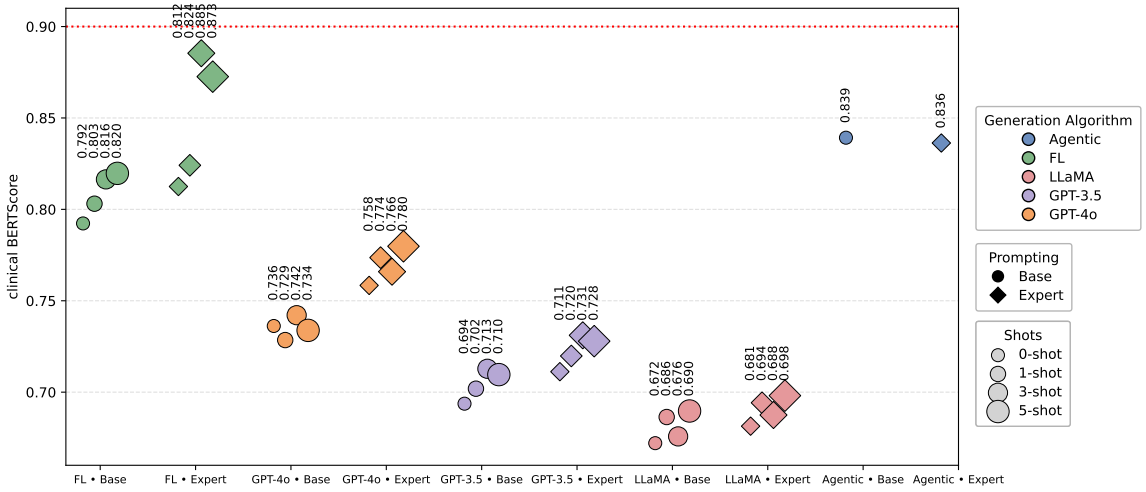


Figure 2: Mean clinical BERTScore (i.e., semantic fidelity) with ICSRs by prompting strategies, generation algorithms, and the number of demonstrations. The horizontal red line indicates the mean clinical BERTScore (0.900) of the synthetic notes generated by the dialogues provided by Wang et al. (2025). Abbreviations: FL, feedback-loop generation; GPT-3.5, GPT-3.5-turbo; LLaMA, LLaMA-3.1 8B.

the same medical terminologies (Figure 2, Section 3.2). Similarly, using expert-informed prompts tended to ensure semantic fidelity better than using base prompts (Figure 2). However, increasing the number of demonstrations did not consistently lead to better semantic fidelity. In some cases (e.g., using GPT-4o with base prompt for in-context learning), increasing the number of demonstrations led to model confusion and thus slightly reduced semantic fidelity (Figure 2). Similar results were observed when generating synthetic clinical notes using MIMIC-IV discharge summaries (Appendix B.9).

3.2. Surface-level Statistics

We found that the feedback-loop algorithm generated the largest number of sentences and showed the highest overall token count, whereas the agentic generation algorithm produced comparatively terse outputs (i.e., fewer tokens per document (Figure 3a and 3b)). Despite the differences in global terseness, the agentic and the feedback-loop algorithms (using expert-informed prompt) produced synthetic clinical notes that are concise at the sentence level (mean token count per sentence [i.e., mean token count normalized by the mean number of sentences] = 23.90 and 25.54, respectively) (Figure 3a and 3b). This finding contrasts with the verbosity of the synthetic clinical notes generated by previous works, such as Songsiritat (2025), whose mean token count per sentence is over 37 (Figure 3a and 3b).

Similarly, the feedback-loop algorithm tended to use most number of special characters compared with other generation algorithms (Figure 3c). After normalized by the mean number of sentences, the agentic and the feedback-loop algorithms (using base prompt) tended to include more number of special characters in synthetic notes (mean number of special characters per sentence [i.e., mean special character count normalized by the mean number of sentences] = 1.50 and 0.96, respectively) (Figure 3c).

Using expert-informed prompts or changing the number of demonstrations did not have consistent impact on surface-level statistics (Appendix B.3). Synthetic clinical notes generated based on MIMIC-IV discharge summaries demonstrated similar results except they were generally higher on token and sentence counts (Appendix B.10).

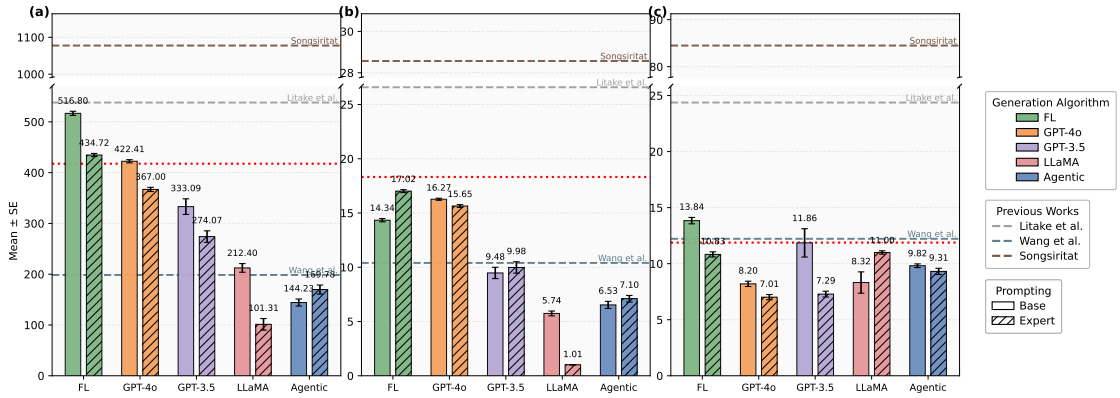


Figure 3: Surface-level statistics by generation algorithms using ICSRs (k=0). (a) mean token count per synthetic clinical note, (b) mean number of sentences per synthetic clinical note, (c) mean number of special characters used per synthetic clinical note. Here, previously reported synthetic clinical notes (Litake et al. (2024); Wang et al. (2025); Songsiritat (2025)) were used for comparison. The horizontal red line indicates the corresponding values of the real world clinical notes from SNUH corpus (mean token count = 417.63, mean number of sentences = 18.32, and mean number of special characters = 11.88) Abbreviations: FL, feedback-loop generation; GPT-3.5, GPT-3.5-turbo; LLaMA, LLaMa-3.1 8B.

3.3. Lexical Diversity

Generally, using expert-informed prompts improved the medical terminology density in synthetic clinical notes, although the effectiveness of the prompting strategy varied widely between the generation algorithms (Table 1). Across generation algorithms, the agentic algorithm generated synthetic clinical notes with the highest medical terminology density, with generally high type-token ratio (Table 1). Particularly, we observed that as type-token ratio increased, medical terminology density also tended to increase in synthetic notes generated by the agentic or the feedback-loop algorithms, while other algorithms showed negative association between the two metrics (Appendix B.2). On the other hand, the density of Korean characters was highest in the synthetic clinical notes generated by the feedback-loop algorithm (Table 1). However, the generation algorithms tended to have less n-gram diversity as measured by self-BLEU (Table 1), potentially due to the content similarity of the synthetic notes documenting ADRs. Changing the number of demonstrations did not have a significant impact on the lexical diversity (Appendix B.4). Synthetic notes generated by MIMIC-IV showed similar results (Appendix B.11).

3.4. Distributional Fidelity

Using expert-informed prompts tended to generate synthetic clinical notes with better behavioristic characteristics of human-written language, as measured by the goodness-of-fit with Zipf's Law (Appendix B.5). Moreover, using the agentic or feedback-loop generation algorithms further improved such tendency. All generation algorithms generally demonstrated lower divergence from the real-world clinical notes at token-level than at corpus-level (Appendix B.5). Particularly, using smaller models (e.g., LLaMA-3.1 8B) and base prompt tended to reduce the divergence (Appendix B.5). Increasing the number of demonstrations did not significantly affect the results (Appendix B.5). Using MIMIC-IV discharge summaries as the input showed similar results (Appendix B.12).

3.5. Human Evaluation

Overall, synthetic clinical notes generated using the feedback-loop algorithm (k=0) achieved the highest aggregate scores in human evalua-

Table 1: Lexical Diversity of Synthetic Clinical Notes per Generation Algorithm using ICSRs (k=0).

Algorithm	Self-BLEU (SD)	Medical Term Density (SD)	TTR (SD)	bilinguality (SD)
Agentic (base)	93.32 (12.28)	0.312 (0.061)	0.746 (0.093)	0.078 (0.148)
Agentic (expert)	93.14 (11.36)	**0.336** (0.082)	0.746 (0.068)	0.041 (0.082)
FL (base)	82.95 (5.65)	0.261 (0.047)	0.671 (0.078)	**0.395** (0.054)
FL (expert)	84.52 (5.54)	**0.314** (0.047)	0.645 (0.080)	0.346 (0.052)
LLaMA-3.1 8B (base)	**55.71** (26.55)	0.142 (0.089)	0.805 (0.269)	**0.367** (0.230)
LLaMA-3.1 8B (expert)	99.80 (4.47)	0.018 (0.021)	0.001 (0.031)	0.097 (0.000)
GPT-3.5-turbo (base)	69.98 (12.17)	0.061 (0.037)	**0.859** (0.107)	0.345 (0.089)
GPT-3.5-turbo (expert)	62.63 (10.46)	0.127 (0.066)	**0.817** (0.103)	0.258 (0.104)
GPT-4o (base)	83.82 (6.10)	0.156 (0.048)	0.653 (0.082)	0.335 (0.058)
GPT-4o (expert)	81.51 (6.83)	0.182 (0.057)	0.641 (0.089)	0.261 (0.063)
Litake et al. (2024)	87.79 (6.86)	0.211 (0.032)	0.503 (0.093)	–
Wang et al. (2025)	76.67 (8.25)	0.196 (0.045)	0.637 (0.091)	–
Songsiritat (2025)	**45.24** (7.24)	0.164 (0.028)	0.647 (0.063)	–

Note. Self-BLEU measures intra-corpus n-gram diversity (lower indicates greater diversity). Medical term density denotes the mean proportion of recognized clinical terms per document. Type–token ratio reflects lexical diversity over each set of synthetic notes generated by each algorithm. bilinguality measures the mean proportion of Korean characters per document. Abbreviations: FL, feedback-loop generation; Agentic, multi-agent generation; LLaMA-3.1 8B, generation by in-context learning of LLaMA-3.1 8B; GPT-3.5-turbo, generation by in-context learning of GPT-3.5-turbo; GPT-4o, generation by in-context learning of GPT-4o; SD, standard deviation.

tion compared with all other generation algorithms (Figure 4, Appendix B.6). Notably, among the 11 expert-assessed attributes, clinical realism (i.e., the degree to which synthetic notes conform to real-world clinical documentation practices) was rated highest for the notes generated by the feedback-loop algorithm (Figure 4). Similarly, the use of expert-informed prompts was associated with higher overall scores and improved clinical realism, although these differences did not reach statistical significance (Wilcoxon signed-rank test: $W = 19.5$ and 21.0; p-values = 0.069 and 0.283, respectively).

Additionally, the blinded evaluation demonstrated that the proportion of real-world clinical notes correctly identified from paired real-world and synthetic notes was lowest for synthetic notes generated by the iterative feedback-loop algorithm. The difficulty in distinguishing between real-world and synthetic clinical notes suggests a higher degree of realism in the notes generated by this algorithm.(Appendix B.7).

3.6. Performance of Extraction of ADR Evidence

Overall, the use of synthetic clinical notes generated by in-context learning with GPT-4o and the feedback-loop algorithm as demonstrations led to improved extraction performance of the base model (Appendix B.8). Furthermore, synthetic notes generated with a larger number of demonstrations or using an expert prompting strategy were associated with enhanced extraction performance (Appendix B.8).

4. Discussion

We showed that iterative generation strategy employing feedback loops between *Generator* and *Evaluator* (the feedback-loop algorithm) is a viable way to ensure both semantic fidelity (Figure 2) and realistic surface-level characteristics of synthetic clinical notes (Figure 3, Table 1). Importantly, the synthetic clinical notes generated by the feedback-loop algorithm demonstrated reduced distributional divergence from real-world clinical notes (Table **??**), with greater conformity to the statistical and structural properties of natural language (Table **??**). Moreover, they demonstrated greater tendency for terse or concise outputs (Table 1), with increased use of medical terminology and domain-specific surface features (e.g., shorthand, abbreviations, and telegraphic expressions, Table 1) that are characteristic of real-world clinical notes. We also showed

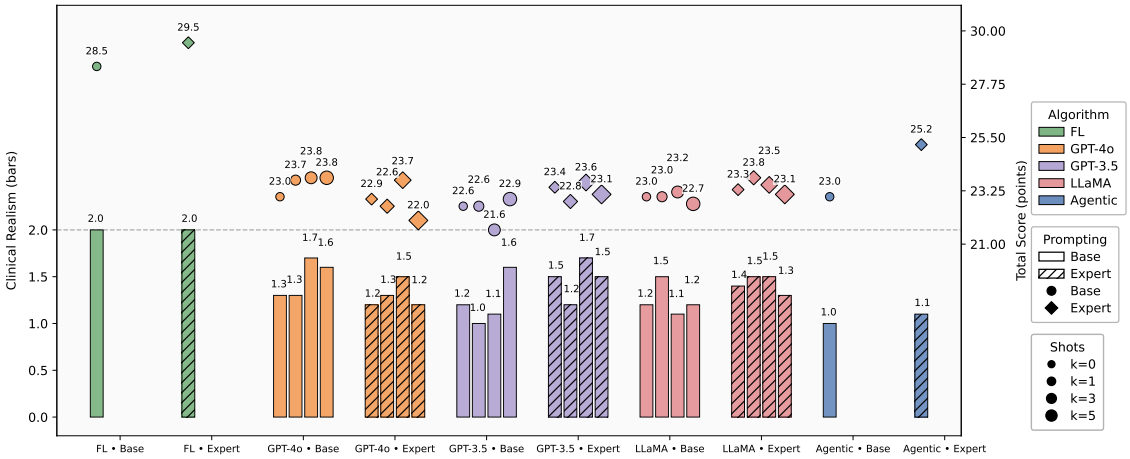


Figure 4: Human evaluation results of synthetic clinical notes based on ICSRs by generation algorithms and number of demonstrations. The maximum scores are 33 and 3 for total score and clinical realism, respectively, with a higher score indicating better quality on the evaluation metric(s). The vertical axis on the left indicates clinical realism, while the one on the right indicates total score. The scale of each vertical axis was adjusted for appropriate score ranges. Of note, we used the setup (k=0) only for the feedback-loop algorithm since there was no significant differences across the varying number of demonstrations for the algorithm. Abbreviations: FL, feedback-loop generation; GPT-3.5, GPT-3.5-turbo; LLaMA, LLaMa-3.1 8B.

that this strategy can be effectively employed in generating bilingual synthetic clinical notes (Table 1). These quantitative findings were corroborated by a manual human evaluation and a blinded evaluation by the healthcare professional (Figure 4). Our post-hoc qualitative error analysis also showed that the synthetic clinical notes generated by the iterative feedback-loop generation algorithm generally scored lowest on all error metrics (Appendix B.14). Furthermore, despite increased token usage (Appendix B.13), the iterative feedback-loop algorithm confers an advantage in generating realistic bilingual clinical notes for data augmentation, thereby improving downstream task performance (Appendix B.8).

These findings can be attributed to the advantage of iterative generation, which can be viewed as an extension of repeated sampling. In fact, repeated sampling by LLMs is a well-established strategy for improving output quality by effectively narrowing the gap between *top-1* and *top-k* predictions (Bateni et al., 2025). This method allows better and more effective sampling of outputs at the expense of increased computation (Brown et al., 2024). Iterative generation extends repetitive sampling by incorporating adaptive refinement. Rather than providing the same queries across iterations, the generation process is progressively guided by structured feedback from a performant *Evaluator* model, enabling targeted improvements in output quality. Because this feedback-driven refinement reflects the core mechanisms of reasoning in LLMs, the proposed approach may be viewed as an instantiation of LLM-based reasoning in the context of controlled text synthesis. Notably, our results indicate that such iterative feedback-loop–based generation is more advantageous than a generation strategy involving multiple agents, which lack an explicit iterative refinement cycle (Figures 2, 4). The differences underscore the importance of feedback-guided iteration over mere agentic multiplicity.

Furthermore, we showed that using expert-informed prompt (i.e., the direct incorporation of expressionistic and stylistic guidance curated by the healthcare professional) generally improved the similarity of synthetic notes to real-world clinical notes. The advantages of such prompting strategy were most pronounced in metrics reflecting medical terminology density and surface-level clinical realism (Table 1, Figure 4), both of which

are critical indicators of realism in synthetic clinical notes.

These findings can be attributed to the importance of context in LLM-based text generation. Context supplies task-specific information that enables effective in-context learning and guides the model toward desired behaviors (Mei et al., 2025). Thus, the deliberate construction of inference-time inputs as context (i.e., context engineering) has emerged as a well-established and powerful approach for inducing high-quality, task-aligned model outputs (Kim et al., 2026). In this sense, our proposed prompting strategy constitutes a guided form of context engineering for the task of controlled clinical text generation. In contrast, we observed that varying the number of demonstrations had no significant effect on model performance (Figure 1, Appendix B.3, B.4, B.5, B.6). This finding suggests that, for the task of generating semantically and expressionistically realistic synthetic clinical notes, the inclusion of domain-specific, expert-driven guidance plays a more crucial role on output quality than providing additional demonstrations.

The significance of this study lies in the introduction of a generation framework including a structured prompting strategy that jointly ensures semantic fidelity and surface-level realism in synthetic clinical notes. While previous studies have predominantly emphasized semantic correctness (Litake et al., 2024; Wang et al., 2025; Songsiritat, 2025), we argue that preserving surface-level characteristics is equally, if not more, critical as even subtle deviations in linguistic form can induce distributional shift and undermine downstream model generalizability. Thus, the greatest contribution of this study lies in addressing both dimensions of realism through a unified framework that enables practical and effective data augmentation in clinical settings. Particularly, our approach is well suited for automatic ADR detection using clinical natural language processing (NLP), where access to large volumes of high-quality, annotated text remains persistently challenging. Importantly, our approach is not a framework for learning but for more effectively accessing pre-trained knowledge. In the context of synthetic clinical note generation, we demonstrated that such a knowledge-accessing framework can be both more effective and computationally efficient than a learning framework, which often requires substantially greater resources for training or fine-tuning. Secondly, the systematic and quantitative evaluation scheme introduced in this study, including a structured protocol for human evaluation, offers a reusable and extensible methodology for future research, enabling rigorous evaluation of both semantic and surface-level properties of synthetic clinical notes. Finally, this study is one of the few attempts to generate bilingual synthetic clinical notes. By extending synthetic note generation beyond a single language (e.g., English), our framework can support cross-lingual and bilingual clinical ML applications, where stylistic conventions and linguistic structures can vary substantially across languages. In this sense, our proposed framework can improve the applicability of synthetic data augmentation to non-English clinical settings and facilitate the development of language-robust clinical NLP models.

This study had a few limitations. First, our framework relied on a specific set of LLMs, when varying model capacity may affect the fidelity and quality of the generated synthetic notes. Nevertheless, we partially mitigated this limitation by intentionally employing models of varying sizes for comparison. This setup allowed us to establish the baseline performance levels against which the superiority of the proposed framework could be more fairly evaluated. Second, our evaluation was conducted using real-world clinical notes from a single institution, which may limit the generalizability of the findings. Yet, we intended to improve the external validity of our study by incorporating expert-based blinded evaluations performed by the healthcare professional who was not accustomed to the institution-specific writing styles. Despite these limitations, this study provides a robust foundation for clinical ML and LLM applications through scalable, bilingual, and institution-agnostic synthetic clinical note generation.

All in all, despite the advantages of the iterative feedback-loop strategy, the choice of generation method should balance realism, downstream utility, computational cost, and task-specific requirements (Figure B.15). For applications requiring high fidelity and distributional alignment with real-world clinical notes, the iterative feedback-loop algorithm is preferred, as it yields more realistic outputs and improves downstream

performance despite increased token usage. In contrast, simpler strategies may be suitable in resource-constrained settings but at the cost of reduced realism and performance. Notably, expert-informed prompting consistently provides substantial gains in output quality, exceeding the impact of increasing the number of demonstrations. These findings highlight that feedback-guided iterative refinement, coupled with expert-driven context engineering, constitutes the most effective strategy for generating high-quality synthetic clinical notes.

Author Contributions

David Seung U Lee: Conceptualization, Project administration, Data curation, Formal analysis, Methodology, Visualization, Writing – Original Draft, Writing – Review and Editing. **Seeun Park**: Project administration, Formal analysis, Methodology, Writing – Original Draft, Writing - Review. **Seoyoon Jang**: Formal analysis, Writing – Original Draft, Writing – Review. **Sunyoung Lee**: Human evaluation, Writing – Review. **Chaeyoung Chang**: Formal analysis, Writing – Review. **Sungwook Choi**: Formal analysis, Writing – Review. **Howard Lee**: Conceptualization, Securing Funding, Writing – Review and Editing, Supervision, Project administration.

Acknowledgments

This study was supported by the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (No.2022R1A6A1A03063039), a grant (25202MFDS003) from Ministry of Food and Drug Safety in 2025, and Korea Institute for Advancement of Technology(KIAT) grant funded by the Korea Government(MOTIE) (RS-2025-02214034, HRD Program for Industrial Innovation).

References

Abdulazeez Alsajri, Hasan Ahmed Salman, and Amani Steiti. Generative models in natural language processing: a comparative study of chatgpt and gemini. *Babylonian Journal of Artificial Intelligence*, 2024:134–145, 2024.

Basel Alshaikhdeeb, Ahmed Abdelmonem Hemedan, Soumyabrata Ghosh, Irina Balaur, and Venkata Satagopam. Generation of synthetic clinical text: A systematic review. *arXiv preprint arXiv:2507.18451*, 2025.

MohammadHossein Bateni, Vincent Cohen-Addad, Yuzhou Gu, Silvio Lattanzi, Simon Meierhans, and Christopher Mohri. Algorithmic thinking theory. *arXiv preprint arXiv:2512.04923*, 2025.

Anjanava Biswas and Wrick Talukdar. Enhancing clinical documentation with synthetic data: Leveraging generative models for improved accuracy. *arXiv preprint arXiv:2406.06569*, 2024.

Bradley Brown, Jordan Juravsky, Ryan Ehrlich, Ronald Clark, Quoc V Le, Christopher Ré, and Azalia Mirhoseini. Large language monkeys: Scaling inference compute with repeated sampling. *arXiv preprint arXiv:2407.21787*, 2024.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

Sungjoo Byun, Jiseung Hong, Sumin Park, Dongjun Jang, Jean Seo, Minseok Kim, Chaeyoung Oh, and Hyopil Shin. Korean bio-medical corpus (kbmc) for medical named entity recognition. *arXiv preprint arXiv:2403.16158*, 2024.

Alistair EW Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom J Pollard, Sicheng Hao, Benjamin Moody, Brian Gow, et al. Mimic-iv, a freely accessible electronic health record dataset. *Scientific data*, 10(1):1, 2023.

Siun Kim, Taegwan Kang, Tae Kyu Chung, Yoona Choi, YeSol Hong, Kyomin Jung, and Howard Lee. Automatic extraction of comprehensive drug safety information from adverse drug event narratives in the korea adverse event reporting system using natural language

processing techniques. *Drug Safety*, 46(8):781–795, 2023.

Siun Kim, David Seung U Lee, Yujin Kim, Hyung-Jin Yoon, and Howard Lee. Beyond fine-tuning: Leveraging domain-aware in-context learning with large language models for clinical named entity recognition. *Journal of Biomedical Informatics*, page 104982, 2026.

Sunjun Kweon, Junu Kim, Jiyoun Kim, Sujeong Im, Eunbyeol Cho, Seongsu Bae, Jungwoo Oh, Gyubok Lee, Jong Hak Moon, Seng Chan You, et al. Publicly shareable clinical large language model built on synthetic clinical notes. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 5148–5168, 2024.

Jianfu Li, Yujia Zhou, Xiaoqian Jiang, Karthik Natarajan, Serguei Vs Pakhomov, Hongfang Liu, and Hua Xu. Are synthetic clinical notes useful for real natural language processing tasks: A case study on clinical entity recognition. *Journal of the American Medical Informatics Association*, 28(10):2193–2201, 2021.

Onkar Litake, Brian H Park, Jeffrey L Tully, and Rodney A Gabriel. Constructing synthetic datasets with generative artificial intelligence to train large language models to classify acute renal failure from clinical notes. *Journal of the American Medical Informatics Association*, 31 (6):1404–1410, 2024.

Tatiana Litvinova, Pavel Seredin, Olga Litvinova, and Olga Zagorovskaya. Differences in type-token ratio and part-of-speech frequencies in male and female russian written texts. In *Proceedings of the Workshop on Stylistic Variation*, pages 69–73, 2017.

Jinghui Lu, Maeve Henchion, and Brian Mac Namee. Diverging divergences: Examining variants of jensen shannon divergence for corpus comparison tasks. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 6740–6744, 2020.

Lingrui Mei, Jiayu Yao, Yuyao Ge, Yiwei Wang, Baolong Bi, Yujun Cai, Jiazhi Liu, Mingyu Li, Zhong-Zhi Li, Duzhen Zhang, et al. A survey of context engineering for large language models. *arXiv preprint arXiv:2507.13334*, 2025.

Mark Neumann, Daniel King, Iz Beltagy, and Waleed Ammar. Scispacy: fast and robust models for biomedical natural language processing. *arXiv preprint arXiv:1902.07669*, 2019.

Beau Norgeot, Kathleen Muenzen, Thomas A Peterson, Xuancheng Fan, Benjamin S Glicksberg, Gundolf Schenk, Eugenia Rutenberg, Boris Oskotsky, Marina Sirota, Jinoos Yazdany, et al. Protected health information filter (philter): accurately and securely de-identifying free-text clinical notes. *NPJ digital medicine*, 3(1):57, 2020.

Emeri Potter, Melissa Reyes, Jennifer Naples, and Gerald Dal Pan. Fda adverse event reporting system (faers) essentials: A guide to understanding, applying, and interpreting adverse event data reported to faers. *Clinical Pharmacology & Therapeutics*, 2025.

Angus Roberts, Robert Gaizauskas, Mark Hepple, George Demetriou, Yikun Guo, Ian Roberts, and Andrea Setzer. Building a semantically annotated corpus of clinical texts. *Journal of biomedical informatics*, 42(5):950–966, 2009.

Alexander I Saichev, Yannick Malevergne, and Didier Sornette. *Theory of Zipf's law and beyond*, volume 632. Springer Science & Business Media, 2009.

Joel Shor, Ruyue Agnes Bi, Subhashini Venugopalan, Steven Ibara, Roman Goldenberg, and Ehud Rivlin. Clinical bertscore: an improved measure of automatic speech recognition performance in clinical settings. In *Proceedings of the 5th Clinical Natural Language Processing Workshop*, pages 1–7, 2023.

Piyawoot Songsiritat. Syngp500: A clinically-grounded synthetic dataset of australian general practice medical notes. *arXiv preprint arXiv:2512.15259*, 2025.

Thomas Sounack, Joshua Davis, Brigitte Durieux, Antoine Chaffin, Tom J Pollard, Eric Lehman, Alistair EW Johnson, Matthew McDermott, Tristan Naumann, and Charlotta Lindvall. Bioclinical modernbert: A state-of-the-art long-context encoder for

biomedical and clinical nlp. *arXiv preprint arXiv:2506.10896*, 2025.

Peter D Stetson, Suzanne Bakken, Jesse O Wrenn, and Eugenia L Siegler. Assessing electronic note quality using the physician documentation quality instrument (pdqi-9). *Applied clinical informatics*, 3(02):164–174, 2012.

Hanyin Wang, Chufan Gao, Bolun Liu, Qiping Xu, Guleid Hussein, Mohamad El Labban, Kingsley Iheasirim, Hariprasad Reddy Korsapati, Chuck Outcalt, and Jimeng Sun. Towards adapting open-source large language models for expert-level clinical note generation. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 12084–12117, 2025.

Appendix A. Supplementary Methods

In this section, we further elaborate on the details employed in our methodology. Specifically, we focus on the formal mathematical definitions of metrics, examples of data, prompts, generation algorithms, evaluation criteria, hyperparameter setups as used in this study.

A.1. A Selected Example of ICSR formatted as used in the study

```
[Patient Information]
The patient was Female and 65 year-old
And the patient's body weight is 50.45 kg

[Patient Condition]
The patient needed medication(s) for the following health condition(s):
Malignant melanoma

[Medication(s) Taken by the Patient]
The patient was treated with the following medications:
<Drug 1>
{
  drugcharacterization: The drug is considered as the suspect for the adverse event,
  medicinalproduct: OPDIVO,
  drugstructuredosagenumb: 240,
  drugstructuredosageunit: mg,
  drugadministrationroute: intravenous drip,
  drugindication: Malignant melanoma,
  drugtreatmentduration: 43,
  drugtreatmentdurationunit: day(s),
  actiondrug: The drug was withdrawn after the adverse event occurred,
  drugdosagetext: UNK,
  drugstartdate: 20220106,
  drugenddate: 20220217,
  drugadditional: The adverse event abated after the drug use stopped or dose reduced,
  activesubstancename: NIVOLUMAB
}

[Adverse Event(s) or Reaction(s) Information]
The patient was reported to have experienced the following adverse event(s) or reaction(s):
<Reaction 1>
{
  reactionmeddrapt: Hypersensitivity pneumonitis,
  reactionoutcome: The patient is recovering from the reaction
}

<Reaction 2>
{
  reactionmeddrapt: Pneumonia,
  reactionoutcome: The outcome of the reaction is unknown
}
The patient experienced the event(s) on the following date (YYYYMMDD):
20220324

[The occurrence of serious adverse event(s)]
Serious: Serious adverse event has occurred
Death: Patient did not die
Life-threatening: No life threatening event has occurred
Hospitalization: The patient was hospitalized
Disabling: No disabling outcome has occurred
Congenital anomaly: No birth defect has occurred
```

A.2. A Selected Example of Discharge Summary in MIMIC-IV (the Present Illness Section)

```
___yo RHF with no significant PMHx who yesterday morning developed dysarthria as well as feeling tired
     and slow in all her daily activities. She also noted she couldn't read the newspaper but was
    able to see all the letters normally. She could write but mentioned that it wasn't her usual
    neatness and had to redo an envelope three times. She also noted some incoordination of her R
    hand though did not try to do same activities with L hand. She couldn't remember exactly where
    things went when she was drying off the dishes. At 3 am, she awoke and went to play solitaire on
    the computer when she noted she was unsteady and had to hold on to things when ambulating, as
    well she was unable to properly control the cursor with her R hand. This am she was unable to get
     out of the bathtub as both her legs felt weak. She also noted difficulty finding her mouth when
    trying to eat and that her coffee would spill out of the right side of her mouth. No trouble
    swallowing. The dysarthria is worse today.
```

A.3. A Selected Example of Real-World Clinical Note as Adapted by the Healthcare Professional

```
CC
S> itching

O> CRP 2.1, BT 36.8

A> r/o exfoliative dermatitis
exfoliative dermatitis d/t drug eruption
-> <AE>Clobetasol associated skin atrophy</AE>

Assessment & Plan
P> prednisolone lotion 100ml + olive oil KP mix 2/day
any kind of anti histamine (prn)
palm and sole - optiderm apply
```

A.4. Expert-informed Prompt for Synthetic Clinical Note Generation (ICSR)

```
You are a clinician.

Generate one unstructured narrative clinical note in Korean using ICSR info logically.
Only generate the following note type and indicate the note type at the beginning: {note_type}.

[Mandatory Instructions]
1) Write fully in Korean, but use English only for abbreviations/terms from the 'Abbreviations to use'
     list.
2) Mention adverse event individually in basic '<AE></AE>' form while applying 'AE Instruction'
    expression.
Do not include 'concomitant medication' inside '<AE></AE>'.
3) When mentioning drugs, only use 'activesuspectname'.
4) If info is not enough, complete the note logically within clinical boundaries (age group, history,
    labs, medications, etc.).

[5 Mandatory sections]
1) 'CC' (Chief Complaint)
2) 'HPI' (Present Illness, PMHx/FHx/SHx)
3) 'PE/V/S' (Physical Exam, Vital Signs)
4) 'Assessment'
5) 'Plan'

You can also add more 'sections' ('Lab', 'Review of Systems', 'Consultations', etc.).

[Telegraphic Style Guidelines]
```

```
Focus on keywords, symbols, and numbers.
Abbreviate dates/periods.

[Abbreviations to use]
{abbr_text}

[AE Instruction]
{ae_instruction}

[Examples]
{demonstration}

[ICSR]
{orig}

Only return the clinical note itself.
### Response:
[Clinical Note]
```

A.5. Expert-informed Prompt for Synthetic Clinical Note Generation (MIMIC)

```
You are a clinician.
Generate one unstructured narrative clinical note in Korean using MIMIC info logically.
Only generate the following note type and indicate the note type at the beginning: {note_type}.

[Mandatory Instructions]
1) Write fully in Korean, but use English only for abbreviations/terms from the 'Abbreviations to use'
     list.
2) If info is not enough, complete the note logically within clinical boundaries (age group, history,
    labs, medications, etc.).

[5 Mandatory sections]
1) 'CC' (Chief Complaint)
2) 'HPI' (Present Illness, PMHx/FHx/SHx)
3) 'PE/V/S' (Physical Exam, Vital Signs)
4) 'Assessment'
5) 'Plan'
- You can also add more 'sections' ('Lab', 'Review of Systems', 'Consultations', etc.).

[Telegraphic Style Guidelines]
Focus on keywords, symbols, and numbers.
Abbreviate dates/periods.

[Abbreviations to use]
{abbr_text}

[Examples]
{shot_block}

[MIMIC]
{orig}

### Response:
[Clinical Note]
```

A.6. List of Clinical Abbreviations Provided in Expert-informed Prompt

```
CC, Sx, dx, r/o, s/p, h/o, f/u, w/u, WNL, abn, V/S, opd, ABGA, LFT, Bx, Cx, ROM, CSF, BM, QD, BID, TID
     , QOD, prn, LMWH, op, PCI, PEG, CABG, KTPL, HTPL, DDKT, appe, CAG, CRRT, IV, IM, PO, DM, HTN, CKD
     , ESRD, CML, AML, COPD, NSCLC, HCC, HBV, TB, BPH, GB stone, CVA, PTE, afib, aflutter, vfib, PSVT,
      NSTEMI, STEMI, HF, PA, AP, RUL, RML, RLL, LUL, LLL, NP, I/O, MN, MD, HD, POD
```

A.7. List of Expressions Commonly Found in Real-World Clinical Notes Pertaining to Adverse Drug Reactions (Some Expressions were Translated into English for Comprehensibility)

```
In real clinical notes, information about adverse drug reaction (ADR) is written using the following
     common expression patterns.
[AE] = adverse drug reaction, [drug] = drug name, [date] = onset date, [action] = clinical action (e.g
     ., hold, stopped, discontinued, changed, dose adjusted).
When adding ADR information to a synthetic note, the expert-informed prompt required using one of
     these patterns and enclose the inserted AE mention with `<AE>` and `</AE>` tags:
1. [AE] d/t [drug]
2. [AE] -> [drug], [drug] hold
3. [drug] --> [AE] worsens [action]
4. [AE] occurred (culprit: [drug])
5. [date] [AE] --> [drug] needs checking
6. [drug] [action] d/t [AE]
7. while on [drug], [AE] worsens [action]
8. [AE] -> r/o d/t [drug]
9. [drug] is removed d/t [AE]
10. [AE] with [drug]
11. Previous [drug] use caused [AE]
12. [drug] use worsens [AE]
13. After [drug], [AE] occurred/worsened, Tx switching
14. [drug] : [action] d/t [AE]
15. [drug] induced [AE]
16. After [drug], hold d/t [AE]
17. After [drug], [AE]
18. [AE] d/t/ r/o [drug]
19. [AE] after [drug] ([date])
20. [AE] d/t [drug]
21. h/o [AE] d/t [drug]
22. A day after [drug], [AE] occurred
23. After [drug], hospitalized d/t [AE]
24. [AE] r/o [drug]-related
25. [drug] hold d/t [AE]
26. hold [drug] for [AE]
27. r/o [AE] (culprit drug: [drug])
28. [drug] associated [AE]
29. [AE], [drug] suspected
30. Possibility of [AE] (culprit: [drug](?))
31. Consider the possibility of [AE] d/t [drug]
32. [AE] (r/o d/t [drug])
33. [AE] [drug]
34. On [drug] complaints of [AE]
35. Complication: (+) [AE] > [drug]
36. [drug] => [action] d/t [AE]
37. [action] for [AE] d/t [drug]
38. No issues with other medications, but developed [AE] after starting [drug].
39. [AE] developed after [drug] administration yesterday, with ~ findings observed.
40. [drug]([date], hold d/t [AE])
41. S/E after [drug] ([date]) [AE] list
42. [AE] has been progressively worsening due to [drug].
43. [AE] improved after switching medications, suggesting possible side effect from [drug]; will
     continue monitoring
44. [AE] is highly likely to have been caused by [drug].
45. [drug] (s/e [AE])
46. [drug] S/E: [AE]
```

```
47. On [date], the patient visited the ER for [AE] after using [drug].
48. [AE] symptoms were present during [drug] administration.
49. he/she took [drug], and had [AE] afterwards
```

A.8. Generation Algorithms

The single-model in-context learning approach leveraged few-shot learning through prompt (or context) engineering (Brown et al., 2020), in which varying numbers of exemplars were provided to a single LLM to guide the generation process. This approach represents one of the most widely adopted methods for synthetic clinical note generation and serves as a baseline generation method in our study. Additionally, the multi-agentic generation approach was implemented as a serially structured multi-agent pipeline, in which four role-based agents were connected sequentially in the note-generation process under a zero-shot setting. Specifically, a *Clinician* agent first produced an initial draft of the synthetic clinical note. A *Senior Clinician* agent (Appendix A.9, A.12) then revised the draft by eliminating unnecessary phrasing, reinforcing a telegraphic style, and aligning the narrative more closely with expert-level clinical documentation conventions. Subsequently, a *Language Reviewer* agent (Appendix A.10, A.13) examined the draft to ensure appropriate use of Korean phrasing across sections to enable the generation of bilingual synthetic notes. Finally, a *Style Evaluator* agent (Appendix A.11, A.14) assessed abbreviation usage and if relevant ADR tagging (i.e., <AE>, in case of using ICSRs), as well as adherence to common ADR expression patterns, to produce the finalized synthetic clinical note. The multi-agent pipeline was implemented using the CrewAI framework[4].

A.9. Prompt for Agentic Generation - Senior Clinician (ICSR)

```
Refine the following draft into a realistic telegraphic clinical note style.

[Mandatory Rules]
- The overall ratio of Korean sentences = 50% (approximately half Korean, half English)
- Each paragraph must contain at least two Korean sentences
- Section headers must be in Korean
- Keep all medical abbreviations/English medical terms
- Do NOT delete or modify <AE> tags
- Reduce unnecessary full sentences and use telegraphic/memo-style expressions
- Do not simply translate; adjust the tone and structure to resemble real clinical documentation

[Output]
Output only the revised final clinical note.

### Response:
[Clinical Note]
```

A.10. Prompt for Agentic Generation - Language Enforcer (ICSR)

```
Check whether the document complies with the Korean and English mixing requirements.

[Review Items]
1) Is the proportion of Korean sentences (based on sentence count) approximately 50%?
2) Does each paragraph contain at least two Korean sentences?
3) Are all section headers written in Korean?
4) Are <AE>...</AE> tags fully preserved without modification?

If any requirement is not met, immediately rewrite the document to satisfy all conditions.
```

4. https://github.com/crewAIInc/crewAI

```
Output only the final revised document.

### Response:
[Clinical Note]
```

A.11. Prompt for Agentic Generation - Style Enforcer (ICSR)

```
Check whether the following document adheres to the expression formats and narrative styles of real
     clinical notes.

[Review Items]
1) Has all expression formatting from the provided ICSR been removed?
2) Does it follow the expression style of actual clinical notes?
3) Does it follow the narrative style of actual clinical notes?
4) Does each paragraph include at least one clinical abbreviation?
5) Are <AE> tags preserved based on the list in the 'AE_expressions'?

If any requirement is not met, immediately rewrite the document to satisfy all conditions.
Output only the final revised clinical note.

### Response:
[Clinical Note]
```

A.12. Prompt for Agentic Generation - Senior Clinician (MIMIC)

```
Refine the following draft into a realistic telegraphic clinical note style.

[Mandatory Rules]
- The overall ratio of Korean sentences = 50% (approximately half Korean, half English)
- Each paragraph must contain at least two Korean sentences
- Section headers must be in Korean
- Keep all medical abbreviations/English medical terms
- Reduce unnecessary full sentences and use telegraphic/memo-style expressions
- Do not simply translate; adjust the tone and structure to resemble real clinical documentation

[Output]
Output only the revised final clinical note.
```

A.13. Prompt for Agentic Generation - Language Enforcer (MIMIC)

```
Check whether the document complies with the Korean and English mixing requirements.

[Review Items]
1) Is the proportion of Korean sentences (based on sentence count) approximately 50%?
2) Does each paragraph contain at least two Korean sentences?
3) Are all section headers written in Korean?

If any requirement is not met, immediately rewrite the document to satisfy all conditions.
Output only the final revised document.
```

A.14. Prompt for Agentic Generation - Style Enforcer (MIMIC)

```
Check whether the following document adheres to the expression formats and narrative styles of real
    clinical notes.

[Review Items]
1) Has all expression formatting from the provided MIMIC been removed?
2) Does it follow the expression style of actual clinical notes?
3) Does it follow the narrative style of actual clinical notes?
4) Does each paragraph include at least one clinical abbreviation?

If any requirement is not met, immediately rewrite the document to satisfy all conditions.
Output only the final revised clinical note.
```

A.15. Algorithm for Feedback-loop framework-based Synthetic Note Generation

Algorithm 1 Feedback-loop Framework

1. **Input:** Generator G, Evaluator E, seed clinical input x, number of iterations T=3
2. Initialize prompt $p^{(0)} \leftarrow \textsc{ComposePrompt}(x)$
3. **For** $t = 1$ **to** T **do**
   (a) Generate candidate note:
   $$y^{(t)} \leftarrow G\Big(p^{(t-1)}\Big)$$
   (b) Evaluate $y^{(t)}$ using the following criteria:
      i. Korean–English code-switching consistency
      ii. Clinically appropriate abbreviation usage
      iii. Correctness of ADR tag insertion (i.e., `<AE>`)
      iv. Adherence to clinical writing conventions
      v. Richness and completeness of clinical information
   (c) Generate structured feedback $f^{(t)} \leftarrow E(x, y^{(t)})$
   (d) Update prompt:
   $$p^{(t)} \leftarrow \textsc{UpdatePrompt}\Big(p^{(t-1)}, f^{(t)}\Big)$$
4. **Output:** Final synthetic clinical note $y^{(T)}$

A.16. Prompt for Feedback-Loop-Based Generation - Evaluator (ICSR)

```
You are an experienced clinician and evaluator.
Your task is to evaluate the given synthetic clinical note that was generated based on the ICSR
    according to the evaluation criteria below.
For each criterion, assign a score and explain in detail the reason for your scoring.
At the end, calculate and present the total score.

[Evaluation Criteria]
1. Ratio of Korean-English Mixing
1 point: Korean accounts for 0 to 20% of the entire note
2 points: Korean accounts for 21 to 40% of the entire note
3 points: Korean accounts for 41 to 60% of the entire note
2 points: Korean accounts for 61 to 80% of the entire note
1 point: Korean accounts for 81 to 100% of the entire note
```

```
2. Use of Abbreviations
1 point: 0 to 2 abbreviations used
2 points: 3 to 5 abbreviations used
3 points: 6 or more abbreviations used

3. Use of <AE> and </AE> Tags
0 points: Tags not used at all
1 point: Tags used but applied to incorrect mentions
2 points: Tags used but not for all relevant mentions
3 points: Tags used appropriately and completely for all relevant mentions

4. Emulation of Real Clinical Note Style
0 points: Only complete sentences used
1 point: Only incomplete sentences or telegraphic notes used (but not both)
2 points: Both incomplete sentences and telegraphic notes used, but complete sentences still remain
3 points: Only incomplete sentences and/or telegraphic notes used (no complete sentences)

5. Addition of Missing Clinical Information
0 points: No new information beyond what is in the ICSR
1 point: New information added but clinically inappropriate
2 points: New information added and clinically appropriate

[ICSR]
{report_text}

[Clinical Note]
{note}

[Output Format]
- Score for each criterion (numeric):
- Explanation for each score (reasoning):
- Total Score:
```

A.17. Prompt for Feed-Back-Loop-Based Generation - Evaluator (MIMIC)

```
You are an experienced clinician and evaluator.
Your task is to evaluate the given synthetic clinical note that was generated based on the MIMIC
     according to the evaluation criteria below.
For each criterion, assign a score and explain in detail the reason for your scoring.
At the end, calculate and present the total score.

[Evaluation Criteria]
1. Ratio of Korean-English Mixing
1 point: Korean accounts for 0 to 20% of the entire note
2 points: Korean accounts for 21 to 40% of the entire note
3 points: Korean accounts for 41 to 60% of the entire note
2 points: Korean accounts for 61 to 80% of the entire note
1 point: Korean accounts for 81 to 100% of the entire note

2. Use of Abbreviations
1 point: 0 to 2 abbreviations used
2 points: 3 to 5 abbreviations used
3 points: 6 or more abbreviations used

3. Emulation of Real Clinical Note Style
0 points: Only complete sentences used
1 point: Only incomplete sentences or telegraphic notes used (but not both)
2 points: Both incomplete sentences and telegraphic notes used, but complete sentences still remain
3 points: Only incomplete sentences and/or telegraphic notes used (no complete sentences)

4. Addition of Missing Clinical Information
```

```
0 points: No new information beyond what is in the MIMIC
1 point: New information added but clinically inappropriate
2 points: New information added and clinically appropriate

[MIMIC]
{report_text}

[Clinical Note]
{note}

[Output Format]
Score for each criterion (numeric):
Explanation for each score (reasoning):
Total Score:
```

A.18. Ablation Study for the Impact of Iteration Number on the Key Evaluation Metrics

We conducted an ablation study to determine the optimal iteration number for the iterative feedback loop generation algorithm. Iteration numbers ranging from 1 to 5 were evaluated with respect to semantic fidelity (clinical BERTScore), diversity (self-BLEU), medical terminology density, lexical diversity (type-token ratio; TTR), and bilinguality (proportion of Korean characters). Based on the results, we selected three iterations as the optimal trade-off between semantic quality and diversity.

Table 2: Ablation study on the number of iterations (i) for the feedback loop generation algorithm. Values are reported as mean (SD). Lower self-BLEU indicates higher diversity. The optimal iteration ($i = 3$) achieves the best balance between semantic fidelity, lexical diversity, and output diversity.

Algorithm	Clinical BERTScore	Med. Term Density	TTR	Self-BLEU	bilinguality
FL (base), $k = 0$, $i = 1$	0.7748 (0.1712)	0.258 (0.052)	0.651 (0.081)	86.74 (5.91)	0.381 (0.058)
FL (base), $k = 0$, $i = 2$	0.7610 (0.1627)	0.254 (0.045)	0.674 (0.074)	84.11 (5.47)	0.389 (0.051)
FL (base), $k = 0$, $i = 3$	0.7923 (0.1649)	0.261 (0.047)	0.671 (0.078)	82.95 (5.65)	0.395 (0.054)
FL (base), $k = 0$, $i = 4$	0.7894 (0.1698)	0.257 (0.050)	0.663 (0.082)	83.88 (5.83)	0.395 (0.057)
FL (base), $k = 0$, $i = 5$	0.7837 (0.1661)	0.272 (0.048)	0.658 (0.080)	84.96 (5.59)	0.387 (0.055)

Abbreviations: FL, feedback-loop generation; Agentic, multi-agent generation; LLaMA-3.1 8B, generation by in-context learning of LLaMA-3.1 8B; GPT-3.5-turbo, generation by in-context learning of GPT-3.5-turbo; GPT-4o, generation by in-context learning of GPT-4o; SD, standard deviation.

A.19. Details of Clinical BERTScore Calculation

We calculated the clinical BERTScore between a reference document (i.e., an Individual Case Safety Report, ICSR) and a synthetic clinical note as follows:

$$R_{\text{BERT}} = \frac{1}{|x|} \sum_{x_i \in x} \max_{\hat{x}_j \in \hat{x}} x_i^T \hat{x}_j \tag{4}$$

$$P_{\text{BERT}} = \frac{1}{|\hat{x}|} \sum_{\hat{x}_j \in \hat{x}} \max_{x_i \in x} x_i^T \hat{x}_j \tag{5}$$

$$\text{BERTScore} = 2\frac{P_{\text{BERT}} \cdot R_{\text{BERT}}}{P_{\text{BERT}} + R_{\text{BERT}}} \tag{6}$$

Building on this formulation, we define the clinical BERTScore (CBERTScore) as

$$\text{CBERTScore}(x, \hat{x}) = k \times \text{BERTScore}_{\text{medical}}(x, \hat{x}) + (1 - k) \times \text{BERTScore}_{\text{all}}(x, \hat{x}), \tag{7}$$

where $0 \leq k \leq 1$ controls the relative contribution of medical terminology similarity and overall semantic similarity. $\text{BERTScore}_{\text{all}}$ is computed over all tokens in the documents, whereas $\text{BERTScore}_{\text{medical}}$ is calculated over the subset of tokens corresponding to medically relevant terms.

Based on the assumption that medically relevant terms are likely to be present in both the reference documents (i.e., ICSRs) and the synthetic clinical notes, we used BioClinical ModernBERT to identify those terms using the top-1 similarity search. This process enabled the mapping of conceptually equivalent medical temrs that are expressed in different languages (i.e., English and Korean). Moreover, the assumption can hold valid because the reference documents are not free text, but a semi-structured JSON format consisting of medical terms like medications and adverse drug reactions.

In this study, we set k=0.7.

A.20. List of Special Characters Analyzed in This Study. The counterpart of a special character, if exists, is also included and analyzed accordingly.

Category	Characters
Punctuation	" ? - { | _ ’ " | ~ ‘
Math / Labs	+ = % ^ ≤ ≥ ± ×
Bullets / Structure	• ○
Encoding / Markup	\ @ $ &
Unicode Clinical	° µ → ←
Others	、※ ◯ △

A.21. The formal mathematical definition of self-BLEU as used in this study.

Sentence-level BLEU. Given a hypothesis text d and a set of reference texts $R = \{r_1, \ldots, r_M\}$, we compute sentence-level BLEU (reported on a $[0, 100]$ scale) as

$$\text{BLEU}_n(d, R) = 100 \times \text{BP}(d, R)\ \exp\left(\sum_{k=1}^{n} w_k \log p_k(d, R)\right), \qquad w_k = \frac{1}{n}. \tag{8}$$

The modified k-gram precision is

$$p_k(d, R) = \frac{\sum_{g \in G_k(d)} \min\Big(\text{count}_d(g), \max_{r \in R} \text{count}_r(g)\Big)}{\sum_{g \in G_k(d)} \text{count}_d(g)}, \tag{9}$$

where $G_k(d)$ denotes the multiset of all k-grams appearing in d.

The brevity penalty is

$$\text{BP}(d, R) = \begin{cases} 1, & c_d > r, \\ \exp\left(1 - \dfrac{r}{c_d}\right), & c_d \leq r, \end{cases} \tag{10}$$

where c_d is the token length of the hypothesis d, and r is the effective reference length (as defined in the BLEU metric; typically the reference length closest to c_d with ties resolved by choosing the shorter length).

Tokenization and normalization. In our implementation, text is first normalized via a deterministic preprocessing function $\mathcal{N}(\cdot)$, and tokenization is disabled. Thus, BLEU is computed on the token sequence induced by the normalized string.

Self-BLEU (leave-one-out; $N = 500$). Let $\{d_1, \ldots, d_N\}$ be the set of generated synthetic clinical notes ($N = 500$). For each note d_i, we define its reference set as all other notes:

$$R_i = \{d_j\}_{j \neq i}. \tag{11}$$

We then compute self-BLEU as the average sentence-level BLEU across hypotheses:

$$\text{Self-BLEU}_n = \frac{1}{N} \sum_{i=1}^{N} \text{BLEU}_n(d_i, R_i), \qquad N = 500. \tag{12}$$

Symbol definitions.

- d : hypothesis text (one synthetic clinical note)
- $R = \{r_1, \ldots, r_M\}$: reference set (other synthetic notes)
- n : maximum n-gram order (BLEU-4 in this study)
- w_k : n-gram weights ($w_k = 1/n$)
- $G_k(d)$: multiset of k-grams in d
- $\text{count}_d(g)$: count of n-gram g in d
- $\text{count}_r(g)$: count of g in reference r
- c_d : hypothesis token length
- r : effective reference length
- BP : brevity penalty
- $\mathcal{N}(\cdot)$: normalization function used before BLEU computation
- N : number of synthetic notes ($N = 500$)

A.22. The formal mathematical definition of Jensen-Shannon divergence (JSD) as used in this study.

JSD was defined as

$$\text{JSD}(P \,\|\, Q) = \frac{1}{2}\text{KL}(P \,\|\, M) + \frac{1}{2}\text{KL}(Q \,\|\, M) \tag{13}$$

$$M = \frac{1}{2}(P + Q) \tag{14}$$

$$\text{KL}(P \,\|\, Q) = \sum_{i \in \mathcal{V}} P(i) \log \frac{P(i)}{Q(i)}. \tag{15}$$

where P and Q denote the empirical token frequency distributions (corpus- or document-level) of the synthetic clinical notes and the SNUH corpus, respectively; $\mathcal{V}$ denotes the vocabulary (i.e., the set of all tokens observed in either corpus); $P(i)$ and $Q(i)$ represent the normalized frequencies of token $i \in \mathcal{V}$ in the synthetic notes and the SNUH corpus, respectively, such that $\sum_{i \in \mathcal{V}} P(i) = \sum_{i \in \mathcal{V}} Q(i) = 1$; $M = \frac{1}{2}(P + Q)$ denotes the mixture distribution between P and Q; $\text{KL}(\cdot \| \cdot)$ denotes the Kullback–Leibler divergence. The logarithm is computed using base 2, such that $0 \leq \text{JSD}(P \| Q) \leq 1$. The n-gram range was set at 1 to 4.

A.23. Details of human evaluation

Randomly sampled 10 synthetic notes from each generation setting were provided to the healthcare professional, who (without knowing the identity of the generation schemes) rated them on 11 attributes (i.e., accurate, thorough, useful, organized, comprehensible, succinct, synthesized, internally consistent, clinical realism, terminology and abbreviation use, compliance with SOAP) using a three-point Likert scale, where higher scores indicate better documentation quality and realism (Appendix A.24). Moreover, an additional blinded evaluation was conducted (on random sample of 10 synthetic clinical notes per generation algorithm and setup) in which the healthcare professional was presented with pairs of real-world and synthetic clinical notes and asked to identify the real-world notes with rationales. The accuracy of the healthcare professional's ability to correctly identify the real-world note was subsequently evaluated.

A.24. Criteria and Scoring Strategy for Expert Evaluation

Attribute	1 (Not at all)	2 (Moderately)	3 (Extremely)	Description
Accurate	Contains clear errors or contradictions in clinical information	Generally accurate, but some information is unclear or imprecise	All clinical information is accurate with no errors	The note is true and free of incorrect information.
Thorough	Major clinical problems or key information are missing	Key information is included, but some important context is lacking	All important clinical issues are sufficiently described	The note is complete and documents all issues of importance to the patient.
Useful	Provides little to no assistance for clinical decision-making	Provides limited assistance for decision-making	Substantially contributes to clinical judgment and treatment decisions	The note is extremely relevant, providing valuable information and/or clinical analysis.
Organized	Structure is disorganized, making the clinical course difficult to follow	Basic structure exists, but logical flow and coherence are weak	Logically and systematically organized	The note is well-formed and structured to facilitate understanding of the patient's clinical course.
Comprehensible	Wording is vague or difficult to understand	Generally understandable, but partially ambiguous	Clear and straightforward to understand	The note is clear, without ambiguity or sections that are difficult to interpret.
Succinct	Excessive repetition or unnecessary verbosity	Somewhat verbose, but key points are maintained	Concise and effectively conveys only essential information	The note is brief, focused, and free of redundancy.
Synthesized	Merely lists information without integrated clinical judgment	Partial synthesis is present, but insight is limited	Integrates information to present clear assessment and plans	The note reflects the author's understanding of the patient's status and ability to develop an appropriate plan of care.
Internally Consistent	Internal contradictions are present within the note	Minor inconsistencies exist	Entire content is internally consistent	No part of the note contradicts or ignores any other part.
Clinical Realism	Written in a style not used in real EMRs	Generally similar to real clinical notes, but some expressions are artificial	Nearly indistinguishable from real EMR documentation	The note reflects real-world clinical documentation practices.
Terminology & Abbreviation Use	Inappropriate or unnatural terminology; inconsistent with EMR usage	Appropriate terminology, but partially spelled out terms	Natural use of terminology and abbreviations	The note uses professional medical terminology and abbreviations appropriately.
Compliance (SOAP)	Does not adhere to the SOAP structure	SOAP structure present with misplaced or extraneous content	Clearly adheres to the SOAP structure	The note faithfully follows the SOAP format.

A.25. Base Prompt for Synthetic Clinical Note Generation (ICSR)

```
You are a clinician.

Generate one unstructured narrative clinical note in Korean using ICSR info logically.

Only generate the following note type and indicate the note type at the beginning: {note_type}.

[Mandatory Instructions]
1. Write fully in Korean, but use English only for abbreviations/terms.
2. Mention adverse event individually in basic '<AE></AE>' form, do not include 'concomitant
     medication' inside '<AE></AE>'.
3. When mentioning drugs, only use 'activesuspectname'.
4. If info is not enough, complete the note logically within clinical boundaries (age group, history,
     labs, medications, etc.).

[5 Mandatory sections]
```

```
1) 'CC' (Chief Complaint)
2) 'HPI' (Present Illness, PMHx/FHx/SHx)
3) 'PE/V/S' (Physical Exam, Vital Signs)
4) 'Assessment'
5) 'Plan'

[Examples]
{demonstration}

[ICSR]
{orig}

### Response:
[Clinical Note]
```

A.26. Base Prompt for Synthetic Clinical Note Generation (MIMIC)

```
You are a clinician.
Generate one unstructured narrative clinical note in Korean using MIMIC info logically.
Only generate the following note type and indicate the note type at the beginning: {note_type}.

[Mandatory Instructions]
1. Write fully in Korean, but use English only for abbreviations/terms.
2. If info is not enough, complete the note logically within clinical boundaries (age group, history,
     labs, medications, etc.).

[5 Mandatory sections]
1) 'CC' (Chief Complaint)
2) 'HPI' (Present Illness, PMHx/FHx/SHx)
3) 'PE/V/S' (Physical Exam, Vital Signs)
4) 'Assessment'
5) 'Plan'

[Examples]
{shot_block}

[MIMIC]
{orig}

### Response:
[Clinical Note]
```

A.27. Hyperparameter Settings for the Inference by LLaMA-3.1 8B and LLaMA-3.3 70B

Hyperparameter	Value
max_new_token	700
do_sample	TRUE
top_k	50
top_p	0.9
temperature	0.01
repetition_penalty	1.2
num_return_sequences	1

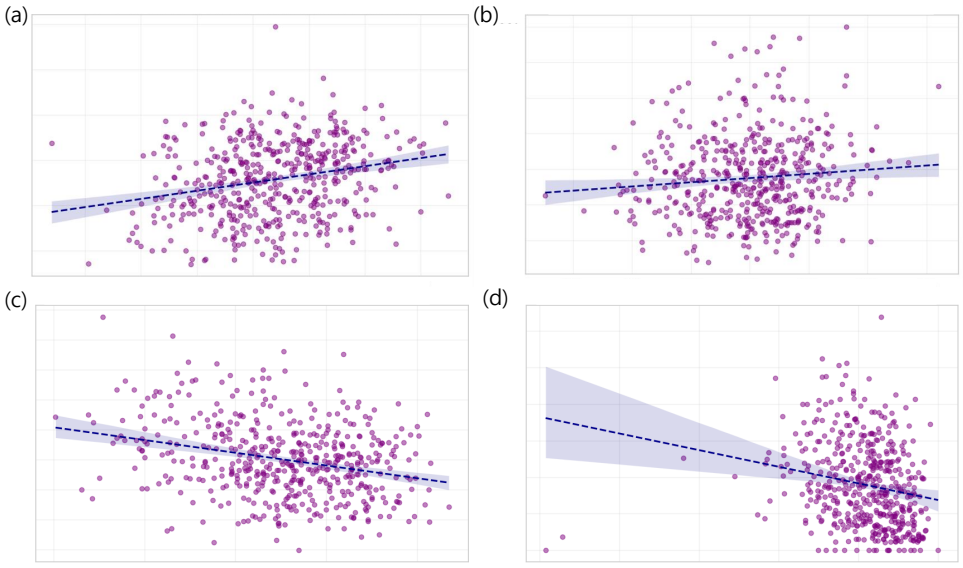
Appendix B. Supplementary Results

In this section, we report supplementary results of this study.

B.1. A Selected Example of Synthetic Clinical Notes.

The provided example of synthetic clinical note was generated by Feedback-Loop Generation Algorithm using Expert-informed Prompt at k=0. We showed that the syntax (surface-level characteristics) of the note highly resembles that of real-world clinical notes. Moreover, the clinical relevance and accuracy of the content in the note is also ensured.

```
1
2  ----- Generated Synthetic Clinical Note (Below) -----
3
4  CC: Palmar-plantar erythrodysaesthesia syndrome d/t cabozantinib
5
6  HPI: 60대 남성, h/o renal cell carcinoma로 cabozantinib 40 mg QD 복용. 2021년 12월 1일 Palmar-plantar
   ↪  erythrodysaesthesia syndrome 발생. cabozantinib 용량 감소 후 증상 호전.
7
8  PE/V/S: V/S WNL. 손바닥, 발바닥 홍반 및 통증 관찰.
9
10 Assessment: Palmar-plantar erythrodysaesthesia syndrome -- cabozantinib 용량 감소 후 회복 중.
11
12 Plan: cabozantinib 용량 감소 유지. 증상 지속 모니터링. 필요 시 피부과 상담 고려. 1주 후 f/u 예정.
```



B.2. Association between TTR and Medical Terminology Density.

The figure shows the association between type-token-ratio (TTR) and medical terminology density in synthetic clinical notes generated by (a) the feedback-loop generation, (b) the agentic generation, (c) a single GPT-4o using expert-informed prompt, and (d) a single GPT-3.5-turbo using expert-informed prompt (k=0). We found that there was a positive association between TTR and medical terminology density in synthetic clinical notes generated by feedback-loop and agnetic generation algorithms, while there was a negative association for other generation algorithms.

B.3. Surface-Level Statistics of Synthetic Clinical Notes across Generation Algorithms using ICSRs.

Algorithm	Mean token count (SD)	Mean sentence count (SD)	Mean special character count (SD)
Agentic (base), k=0	144.23 (154.12)	6.53 (7.34)	9.82 (3.65)
Agentic (expert), k=0	169.78 (199.32)	7.10 (6.76)	9.31 (6.20)
FL (base), k=0	516.80 (90.06)	14.34 (3.35)	13.84 (6.39)
FL (base), k=1	511.11 (85.72)	14.29 (3.62)	13.99 (6.66)
FL (base), k=3	514.47 (88.52)	14.33 (3.53)	13.68 (6.26)
FL (base), k=5	512.05 (88.42)	14.22 (3.59)	13.76 (6.66)
FL (expert), k=0	434.72 (70.03)	17.02 (2.99)	10.83 (5.13)
FL (expert), k=1	434.60 (70.49)	17.21 (3.16)	10.72 (4.91)
FL (expert), k=3	436.07 (71.03)	17.26 (3.17)	10.83 (4.86)
FL (expert), k=5	436.63 (73.87)	17.34 (3.41)	10.86 (4.94)
LLaMA-3.1 8B (base), k=0	212.40 (190.14)	5.74 (4.97)	8.32 (21.29)
LLaMA-3.1 8B (base), k=1	650.93 (88.36)	14.42 (2.40)	72.16 (14.13)
LLaMA-3.1 8B (base), k=3	333.73 (178.19)	10.69 (5.64)	16.68 (24.60)
LLaMA-3.1 8B (base), k=5	377.08 (160.27)	8.88 (5.42)	27.68 (31.13)
LLaMA-3.1 8B (expert), k=0	101.31 (255.21)	1.01 (0.18)	11.00 (3.32)
LLaMA-3.1 8B (expert), k=1	200.23 (6.93)	4.90 (5.97)	44.93 (39.46)
LLaMA-3.1 8B (expert), k=3	18.03 (279.73)	1.36 (3.08)	19.61 (38.42)
LLaMA-3.1 8B (expert), k=5	59.91 (86.66)	2.68 (1.01)	20.17 (19.11)
GPT-3.5-turbo (base), k=0	333.09 (344.72)	9.48 (11.64)	11.86 (28.26)
GPT-3.5-turbo (base), k=1	295.67 (99.57)	7.05 (1.53)	11.43 (9.76)
GPT-3.5-turbo (base), k=3	268.25 (260.48)	6.71 (5.72)	11.56 (10.43)
GPT-3.5-turbo (base), k=5	245.92 (193.15)	6.30 (1.55)	10.32 (8.85)
GPT-3.5-turbo (expert), k=0	274.07 (253.37)	9.98 (11.74)	7.29 (5.80)
GPT-3.5-turbo (expert), k=1	332.00 (258.94)	8.57 (2.64)	12.50 (9.85)
GPT-3.5-turbo (expert), k=3	337.08 (264.84)	7.66 (3.02)	14.72 (17.44)
GPT-3.5-turbo (expert), k=5	326.45 (312.13)	7.38 (2.39)	11.33 (10.09)
GPT-4o (base), k=0	422.41 (69.48)	16.27 (2.12)	8.20 (5.28)
GPT-4o (base), k=1	425.89 (66.04)	13.79 (2.48)	8.00 (4.08)
GPT-4o (base), k=3	436.13 (63.57)	14.33 (2.57)	6.99 (4.06)
GPT-4o (base), k=5	435.01 (62.98)	14.22 (2.57)	7.16 (4.03)
GPT-4o (expert), k=0	367.00 (92.64)	15.65 (2.81)	7.01 (5.26)
GPT-4o (expert), k=1	365.20 (74.66)	12.19 (2.48)	10.69 (6.08)
GPT-4o (expert), k=3	369.82 (73.98)	12.97 (2.33)	10.41 (6.42)
GPT-4o (expert), k=5	372.30 (77.70)	13.11 (2.50)	9.77 (6.48)
Litake et al. (2024)	538.04 (144.18)	26.59 (10.18)	24.37 (12.54)
Wang et al. (2025)	198.46 (83.96)	10.40 (4.74)	12.22 (7.07)
Songsiritat (2025)	1077.60 (368.41)	28.57 (10.96)	84.49 (33.19)

B.4. Lexical Diversity of Synthetic Clinical Notes across Generation Algorithms using ICSRs

Algorithm	Mean self-BLEU (SD)	Mean medical term prop. (SD)	Mean type-token ratio (SD)	Mean Korean char. prop. (SD)
Agentic (base), k=0	93.32 (12.28)	0.312 (0.061)	0.746 (0.093)	0.078 (0.148)
Agentic (expert), k=0	93.14 (11.36)	0.336 (0.082)	0.746 (0.068)	0.041 (0.082)
FL (base), k=0	82.95 (5.65)	0.261 (0.047)	0.671 (0.078)	0.395 (0.089)
FL (base), k=1	82.94 (5.75)	0.321 (0.044)	0.669 (0.077)	0.403 (0.079)
FL (base), k=3	83.07 (5.82)	0.254 (0.039)	0.668 (0.079)	0.333 (0.097)
FL (base), k=5	82.95 (6.07)	0.266 (0.038)	0.673 (0.077)	0.331 (0.092)
FL (expert), k=0	84.52 (5.54)	0.314 (0.047)	0.645 (0.080)	0.346 (0.104)
FL (expert), k=1	84.56 (5.47)	0.258 (0.045)	0.645 (0.080)	0.373 (0.094)
FL (expert), k=3	84.56 (5.44)	0.262 (0.048)	0.646 (0.079)	0.363 (0.106)
FL (expert), k=5	84.54 (5.35)	0.260 (0.046)	0.640 (0.081)	0.359 (0.102)
LLaMA-3.1 8B (base), k=0	55.71 (26.55)	0.142 (0.089)	0.805 (0.269)	0.367 (0.230)
LLaMA-3.1 8B (base), k=1	96.56 (13.13)	0.198 (0.052)	0.700 (0.044)	0.070 (0.029)
LLaMA-3.1 8B (base), k=3	72.48 (30.53)	0.214 (0.071)	0.878 (0.100)	0.254 (0.168)
LLaMA-3.1 8B (base), k=5	70.47 (24.15)	0.171 (0.063)	0.832 (0.132)	0.237 (0.173)
LLaMA-3.1 8B (expert), k=0	99.80 (4.47)	0.018 (0.021)	0.001 (0.031)	0.097 (0.000)
LLaMA-3.1 8B (expert), k=1	88.25 (23.07)	0.132 (0.118)	0.336 (0.393)	0.025 (0.028)
LLaMA-3.1 8B (expert), k=3	96.96 (13.42)	0.041 (0.084)	0.044 (0.183)	0.005 (0.016)
LLaMA-3.1 8B (expert), k=5	97.17 (15.17)	0.033 (0.079)	0.052 (0.120)	0.017 (0.020)
GPT-3.5-turbo (base), k=0	69.98 (12.17)	0.094 (0.051)	0.859 (0.107)	0.346 (0.054)
GPT-3.5-turbo (base), k=1	71.06 (10.34)	0.076 (0.044)	0.764 (0.091)	0.342 (0.055)
GPT-3.5-turbo (base), k=3	66.97 (10.10)	0.109 (0.057)	0.793 (0.101)	0.344 (0.059)
GPT-3.5-turbo (base), k=5	63.86 (10.13)	0.107 (0.054)	0.793 (0.096)	0.345 (0.054)
GPT-3.5-turbo (expert), k=0	62.63 (10.46)	0.127 (0.066)	0.817 (0.103)	0.258 (0.052)
GPT-3.5-turbo (expert), k=1	66.83 (11.93)	0.106 (0.061)	0.766 (0.103)	0.257 (0.054)
GPT-3.5-turbo (expert), k=3	69.91 (12.68)	0.095 (0.064)	0.762 (0.102)	0.258 (0.053)
GPT-3.5-turbo (expert), k=5	68.54 (13.52)	0.091 (0.060)	0.762 (0.101)	0.257 (0.052)
GPT-4o (base), k=0	83.82 (6.10)	0.156 (0.048)	0.653 (0.082)	0.335 (0.058)
GPT-4o (base), k=1	81.59 (6.59)	0.141 (0.045)	0.668 (0.071)	0.363 (0.059)
GPT-4o (base), k=3	79.32 (6.73)	0.143 (0.046)	0.666 (0.074)	0.364 (0.061)
GPT-4o (base), k=5	78.66 (7.25)	0.145 (0.047)	0.672 (0.073)	0.360 (0.066)
GPT-4o (expert), k=0	81.51 (6.83)	0.182 (0.057)	0.641 (0.089)	0.261 (0.063)
GPT-4o (expert), k=1	79.34 (6.90)	0.207 (0.060)	0.662 (0.085)	0.231 (0.058)
GPT-4o (expert), k=3	78.46 (7.18)	0.212 (0.061)	0.678 (0.083)	0.237 (0.060)
GPT-4o (expert), k=5	78.26 (7.09)	0.216 (0.062)	0.680 (0.081)	0.243 (0.060)
Litake et al. (2024)	87.79 (6.86)	0.211 (0.032)	0.503 (0.093)	–
Wang et al. (2025)	76.67 (8.25)	0.196 (0.045)	0.637 (0.091)	–
Songsiritat (2025)	45.24 (7.24)	0.164 (0.028)	0.647 (0.063)	–

B.5. Distributional Fidelity of Synthetic Clinical Notes across Generation Algorithms using ICSRs.

JSD-C and JSD-T denote corpus-level and token-level Jensen–Shannon divergence, respectively (the lower the higher distributional fidelity with the real-world clinical notes). GoF denotes goodness-of-fit (R^2) with Zipf's law on logarithmic scale (the higher, the better fit), while Slope represents the fitted slope after linear regression (close to -1 represents higher tendency for linearity and conformity with log-transformed Zipf's Law curve). Abbreviations: FL, feedback-loop generation; LLaMA, LLaMA-3.1 8B; GPT-3.5, GPT-3.5-turbo.

Algorithm	JSD-C	JSD-T	Slope	GoF
Agentic (base), k=0	0.8537	0.4622	-1.3114	0.9768
Agentic (expert), k=0	0.8278	0.3790	-1.2812	0.9798
FL (base), k=0	0.8055	0.3051	-1.3229	0.9765
FL (base), k=1	0.8081	0.3061	-1.3277	0.9756
FL (base), k=3	0.8060	0.3075	-1.3260	0.9753
FL (base), k=5	0.8067	0.3076	-1.3229	0.9763
FL (expert), k=0	0.8255	0.3439	-1.3249	0.9736
FL (expert), k=1	0.8255	0.3445	-1.3228	0.9730
FL (expert), k=3	0.8273	0.3455	-1.3304	0.9722
FL (expert), k=5	0.8251	0.3458	-1.3268	0.9727
LLaMA-3.1 8B (base), k=0	0.7352	0.2785	-0.8743	0.9439
LLaMA-3.1 8B (base), k=1	0.8757	0.4998	-1.6872	0.8473
LLaMA-3.1 8B (base), k=3	0.7688	0.2933	-1.2135	0.9382
LLaMA-3.1 8B (base), k=5	0.7562	0.2819	-1.1570	0.9566
LLaMA-3.1 8B (expert), k=0	0.9891	0.8035	-0.7438	0.5022
LLaMA-3.1 8B (expert), k=1	0.8137	0.4972	-1.0217	0.9341
LLaMA-3.1 8B (expert), k=3	0.8687	0.6488	-0.8142	0.9568
LLaMA-3.1 8B (expert), k=5	0.8714	0.6579	-0.9233	0.9671
GPT-3.5-turbo (base), k=0	0.8455	0.3142	-1.1786	0.9699
GPT-3.5-turbo (base), k=1	0.8347	0.3244	-1.1747	0.9752
GPT-3.5-turbo (base), k=3	0.8110	0.3118	-1.1451	0.9758
GPT-3.5-turbo (base), k=5	0.8142	0.3130	-1.1082	0.9758
GPT-3.5-turbo (expert), k=0	0.8196	0.2938	-1.0977	0.9733
GPT-3.5-turbo (expert), k=1	0.8163	0.2917	-1.1850	0.9773
GPT-3.5-turbo (expert), k=3	0.8181	0.3198	-1.2022	0.9751
GPT-3.5-turbo (expert), k=5	0.8312	0.3373	-1.1612	0.9750
GPT-4o (base), k=0	0.8372	0.3259	-1.3101	0.9757
GPT-4o (base), k=1	0.8112	0.3128	-1.2828	0.9790
GPT-4o (base), k=3	0.8020	0.3043	-1.2738	0.9780
GPT-4o (base), k=5	0.7987	0.3013	-1.2622	0.9790
GPT-4o (expert), k=0	0.8226	0.3554	-1.2999	0.9688
GPT-4o (expert), k=1	0.7966	0.3384	-1.2968	0.9723
GPT-4o (expert), k=3	0.7842	0.3274	-1.2968	0.9726
GPT-4o (expert), k=5	0.7855	0.3250	-1.2929	0.9742
Litake et al. (2024)	0.8515	0.7195	-1.8361	0.9476
Wang et al. (2025)	0.8086	0.6916	-1.5042	0.9637
Songsiritat (2025)	0.7862	0.6247	-1.5007	0.9487

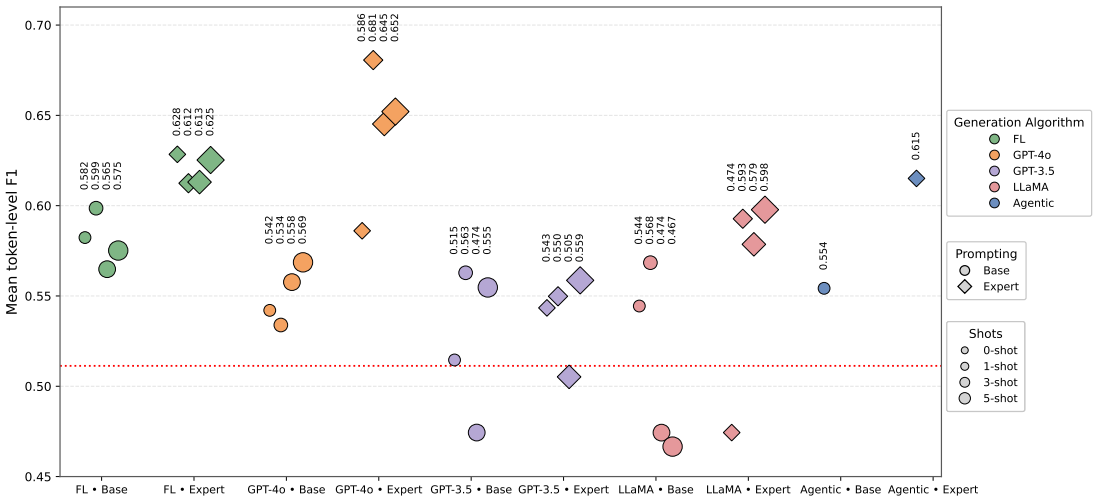
B.6. Human Evaluation Results of Synthetic Clinical Notes across Generation Algorithms using ICSRs.

Algorithm	Accurate	Thorough	Useful	Organized	Comprehensible	Succinct	Synthesized	Internally Consistent	Clinical Realism	Terminology & Abbreviation Use	Compliance	Total
Agentic (base), k=0	1.2	2.3	2.6	2.7	2.9	2.5	2.2	3	1	1	1.6	23
Agentic (expert), k=0	3	2.1	2.4	2.7	2.8	2.5	2.6	3	1.1	1	2	25.2
FL (base), k=0	3	2	2.9	3	2.8	2.8	2.9	3	2	1.3	2.8	28.5
FL (expert), k=0	3	2.1	2.9	3	2.7	3	3	2.9	2	2.3	2.6	29.5
LLaMA-3.1 8B (base), k=0	3	2	1.9	2.1	2.3	2.7	2	2.9	1.2	1	1.9	23
LLaMA-3.1 8B (base), k=1	3	2.1	2.3	2	1.9	2.3	2	2.9	1.5	1	2	23
LLaMA-3.1 8B (base), k=3	3	2	1.8	2.3	2.4	2.9	1.8	2.9	1.1	1	2	23.2
LLaMA-3.1 8B (base), k=5	2.9	1.8	2.4	1.9	2.4	2.5	1.9	2.7	1.2	1	2	22.7
LLaMA-3.1 8B (expert), k=0	3	2	2.3	2.1	2.4	2.4	1.9	2.8	1.4	1	2	23.3
LLaMA-3.1 8B (expert), k=1	3	1.9	2	2.1	2.7	2.7	2.1	2.8	1.5	1	2	23.8
LLaMA-3.1 8B (expert), k=3	3	2	1.7	2.1	2.6	2.9	1.9	2.9	1.5	1	1.9	23.5
LLaMA-3.1 8B (expert), k=5	2.9	2.9	2.9	2.9	2.9	2.9	2.9	2.9	1.3	1	2	27.5
GPT-3.5-turbo (base), k=0	3	1.9	1.7	2.3	2.2	2.9	1.8	2.8	1.2	1	1.8	22.6
GPT-3.5-turbo (base), k=1	3	1.9	2.1	2	2.4	2.5	1.7	3	1	1	2	22.6
GPT-3.5-turbo (base), k=3	3	1.9	1.9	2.1	1.8	2.1	1.7	3	1.1	1	2	21.6
GPT-3.5-turbo (base), k=5	3	1.9	2	2.1	2	2.5	1.8	3	1.6	1	2	22.9
GPT-3.5-turbo (expert), k=0	3	2	2.1	2.2	2.1	2.7	1.9	3	1.5	1	1.9	23.4
GPT-3.5-turbo (expert), k=1	3	2	2	1.9	2.4	2.5	1.9	2.9	1.2	1	2	22.8
GPT-3.5-turbo (expert), k=3	3	2.1	2.1	2.1	2.4	2.4	2.1	2.8	1.7	1	1.9	23.6
GPT-3.5-turbo (expert), k=5	3	2.1	2.2	2.1	1.8	2.4	2	3	1.5	1	2	23.1
GPT-4o (base), k=0	3	2	2.2	2	2	2.7	2.1	2.9	1.3	1	1.8	23
GPT-4o (base), k=1	3	1.9	2.1	2.1	2.6	2.8	2	2.9	1.3	1	2	23.7
GPT-4o (base), k=3	3	1.9	2.1	2.2	2.3	2.9	2	2.9	1.7	1	1.8	23.8
GPT-4o (base), k=5	2.9	1.8	2.4	2	2.4	2.6	2.1	3	1.6	1	2	23.8
GPT-4o (expert), k=0	2.9	2	2.3	2	2.2	2.5	2.2	2.6	1.2	1	2	22.9
GPT-4o (expert), k=1	3	2	2	2.1	1.9	2.6	1.9	2.8	1.3	1	2	22.6
GPT-4o (expert), k=3	2.6	2	2	2.1	2.6	2.9	2	3	1.5	1	2	23.7
GPT-4o (expert), k=5	2.9	1.8	2	2	2.3	2.5	2	2.4	1.2	1	1.9	22
Mean	2.903571429	2.014285714	2.189285714	2.221428571	2.364285714	2.628571429	2.085714286	2.882142857	1.382142857	1.057142857	1.996428571	23.725

B.7. Blinded Evaluation for the Identification of the Real-World Clinical Notes

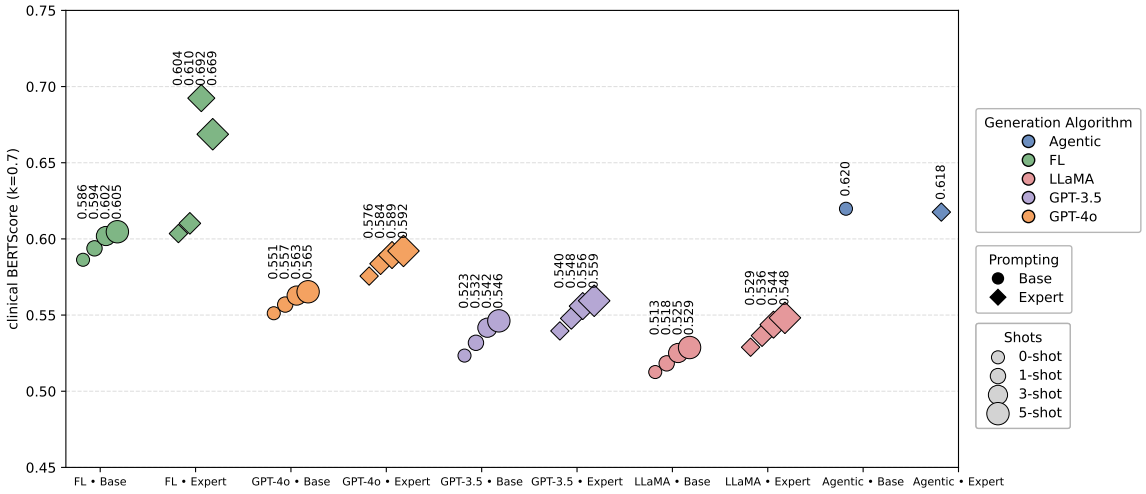
Accuracy was defined as the proportion of real-world clinical notes correctly identified within paired real-world and synthetic clinical notes ($n = 10$). Lower accuracy indicates greater difficulty for the healthcare professional in distinguishing real-world notes from synthetic ones. The metrics for Tone, Medical Expression (i.e., the use of appropriate medical terminology), Clinical Relevance (i.e., the correctness and relevance of clinical information), and Clinical Detail (i.e., the richness of clinical content) were defined as the proportion of instances in which each factor was cited as a reason for identifying a clinical note as real-world.

Algorithm	Accuracy	Tone	Medical Expression	Clinical Relevance	Clinical Detail
Agentic (base), $k = 0$	0.75	0.90	1.00	0.95	0.45
Agentic (expert), $k = 0$	0.35	0.65	1.00	0.90	0.70
FL (base), $k = 0$	0.30	0.75	0.90	1.00	0.70
FL (base), $k = 1$	0.15	0.60	0.85	1.00	0.95
FL (base), $k = 3$	0.25	0.50	0.90	1.00	0.75
FL (base), $k = 5$	0.15	0.55	0.90	1.00	0.85
FL (expert), $k = 0$	0.35	0.90	0.90	1.00	0.75
FL (expert), $k = 1$	0.20	0.85	0.95	1.00	0.85
FL (expert), $k = 3$	0.20	0.95	0.95	1.00	0.80
FL (expert), $k = 5$	0.35	0.95	0.85	1.00	0.85
GPT-3.5-turbo (base), $k = 0$	0.65	0.80	0.95	1.00	0.65
GPT-3.5-turbo (base), $k = 1$	0.90	0.95	0.90	1.00	0.60
GPT-3.5-turbo (base), $k = 3$	0.90	1.00	1.00	1.00	0.65
GPT-3.5-turbo (base), $k = 5$	0.90	1.00	0.95	1.00	0.85
GPT-3.5-turbo (expert), $k = 0$	0.80	0.95	1.00	1.00	0.75
GPT-3.5-turbo (expert), $k = 1$	0.90	0.95	0.95	1.00	0.65
GPT-3.5-turbo (expert), $k = 3$	0.85	0.90	1.00	1.00	0.65
GPT-3.5-turbo (expert), $k = 5$	0.95	0.95	1.00	1.00	0.65
GPT-4o (base), $k = 0$	0.20	0.75	0.75	1.00	0.80
GPT-4o (base), $k = 1$	0.20	0.60	0.95	1.00	0.80
GPT-4o (base), $k = 3$	0.25	0.75	0.90	1.00	0.85
GPT-4o (base), $k = 5$	0.15	0.75	0.80	1.00	0.85
GPT-4o (expert), $k = 0$	0.20	0.95	1.00	1.00	0.80
GPT-4o (expert), $k = 1$	0.15	0.85	0.85	1.00	0.90
GPT-4o (expert), $k = 3$	0.25	0.90	0.95	1.00	0.90
GPT-4o (expert), $k = 5$	0.20	0.90	0.95	1.00	0.75
LLaMA-3.1 8B (base), $k = 0$	0.90	0.95	0.95	1.00	0.80
LLaMA-3.1 8B (base), $k = 1$	0.80	1.00	1.00	1.00	0.85
LLaMA-3.1 8B (base), $k = 3$	0.90	0.95	1.00	0.95	0.65
LLaMA-3.1 8B (base), $k = 5$	0.95	1.00	1.00	1.00	0.85
LLaMA-3.1 8B (expert), $k = 0$	1.00	1.00	1.00	1.00	0.85
LLaMA-3.1 8B (expert), $k = 1$	0.90	1.00	1.00	1.00	1.00
LLaMA-3.1 8B (expert), $k = 3$	0.90	0.95	1.00	1.00	0.95
LLaMA-3.1 8B (expert), $k = 5$	1.00	1.00	1.00	1.00	0.85
Overall	0.554	0.865	0.943	0.994	0.781



B.8. Performance of ADR Evidence Extraction

The figure shows the token-level F1 score to assess the performance of extracting ADR evidence from the human-annotated real-world clinical notes (n=107). Three synthetic clinical notes generated by each algorithm and setup were provided as demonstrations for in-context learning by LLaMA-3.3 70B. The horizontal red dotted line represents the zero-shot performance of the base model (mean token-level F1 = 0.5113). Abbreviations: FL, feedback-loop generation; GPT-3.5, GPT-3.5-turbo; LLaMA, LLaMA-3.1 8B.



B.9. Semantic Fidelity with MIMIC-IV Discharge Summaries

The figure shows the mean clinical BERTScore (semantic fidelity) with MIMIC-IV Discharge Summaries per prompting strategies, generation algorithms, and the number of demonstrations. Abbreviations: FL, feedback-loop generation; GPT-3.5, GPT-3.5-turbo; LLaMA, LLaMA-3.1 8B.

B.10. Surface-Level Statistics of Synthetic Clinical Notes across Generation Algorithms using MIMIC IV Discharge Summaries.

Algorithm	Mean token count (SD)	Mean sentence count (SD)	Mean special character count (SD)
Agentic (base), k=0	491.95 (171.91)	19.64 (6.59)	22.57 (10.94)
Agentic (expert), k=0	502.52 (206.55)	20.52 (7.00)	23.45 (11.49)
FL (base), k=0	802.85 (180.91)	16.15 (4.62)	28.68 (10.64)
FL (base), k=1	803.82 (173.83)	16.21 (4.78)	29.01 (9.89)
FL (base), k=3	810.14 (192.44)	16.49 (5.25)	28.53 (10.20)
FL (base), k=5	802.70 (172.29)	16.42 (4.95)	28.49 (9.87)
FL (expert), k=0	616.66 (151.24)	20.16 (5.33)	27.65 (12.20)
FL (expert), k=1	619.65 (151.61)	19.94 (5.54)	27.85 (12.12)
FL (expert), k=3	612.16 (148.24)	19.86 (5.48)	27.19 (12.54)
FL (expert), k=5	616.41 (157.31)	19.85 (5.71)	27.80 (12.01)
LLaMA-3.1 8B (base), k=0	113.15 (269.83)	3.40 (5.86)	13.71 (13.96)
LLaMA-3.1 8B (base), k=1	531.54 (244.72)	11.01 (6.11)	47.79 (41.53)
LLaMA-3.1 8B (base), k=3	531.08 (392.30)	10.70 (9.43)	27.27 (39.30)
LLaMA-3.1 8B (base), k=5	497.77 (346.96)	11.56 (9.80)	31.02 (35.98)
LLaMA-3.1 8B (expert), k=0	41.03 (169.07)	1.79 (3.47)	14.44 (9.64)
LLaMA-3.1 8B (expert), k=1	616.58 (195.11)	13.22 (7.04)	56.80 (36.64)
LLaMA-3.1 8B (expert), k=3	514.82 (350.59)	11.23 (9.63)	34.42 (41.08)
LLaMA-3.1 8B (expert), k=5	619.38 (329.13)	13.61 (9.49)	34.67 (39.25)
GPT-3.5-turbo (base), k=0	762.66 (356.55)	17.22 (8.47)	28.05 (21.80)
GPT-3.5-turbo (base), k=1	801.55 (495.60)	18.51 (16.60)	36.16 (50.56)
GPT-3.5-turbo (base), k=3	840.00 (592.25)	18.78 (15.84)	39.47 (61.87)
GPT-3.5-turbo (base), k=5	785.98 (545.09)	17.85 (14.47)	32.31 (30.46)
GPT-3.5-turbo (expert), k=0	667.71 (413.24)	17.23 (6.89)	32.30 (93.05)
GPT-3.5-turbo (expert), k=1	771.91 (494.21)	18.28 (10.06)	37.29 (43.42)
GPT-3.5-turbo (expert), k=3	754.25 (526.28)	17.60 (8.49)	33.79 (47.94)
GPT-3.5-turbo (expert), k=5	766.82 (602.94)	16.93 (8.55)	36.26 (78.21)
GPT-4o (base), k=0	828.86 (153.87)	17.06 (3.83)	26.95 (7.20)
GPT-4o (base), k=1	786.00 (157.53)	15.35 (4.00)	16.55 (7.78)
GPT-4o (base), k=3	847.36 (152.25)	17.18 (3.91)	19.76 (8.78)
GPT-4o (base), k=5	859.86 (181.95)	17.71 (4.71)	20.69 (9.38)
GPT-4o (expert), k=0	601.11 (121.86)	19.95 (4.99)	30.60 (11.24)
GPT-4o (expert), k=1	640.50 (126.78)	20.35 (5.34)	27.79 (13.68)
GPT-4o (expert), k=3	663.63 (132.82)	20.07 (5.25)	26.08 (12.01)
GPT-4o (expert), k=5	673.46 (142.48)	19.67 (5.06)	28.34 (14.11)
Litake et al. (2024)	538.04 (144.18)	26.59 (10.18)	24.37 (12.54)
Wang et al. (2025)	198.46 (83.96)	10.40 (4.74)	12.22 (7.07)
Songsiritat (2025)	1077.60 (368.41)	28.57 (10.96)	84.49 (33.19)

B.11. Lexical Diversity of Synthetic Clinical Notes across Generation Algorithms using MIMIC-IV Discharge Summaries

Algorithm	Mean self-BLEU (SD)	Mean medical term prop. (SD)	Mean type-token ratio (SD)	Mean Korean char. prop. (SD)
Agentic (base), k=0	65.61 (10.59)	0.276 (0.072)	0.790 (0.083)	0.272 (0.214)
Agentic (expert), k=0	56.06 (11.05)	0.291 (0.075)	0.761 (0.084)	0.238 (0.225)
FL (base), k=0	60.39 (7.18)	0.212 (0.041)	0.802 (0.046)	0.506 (0.047)
FL (base), k=1	60.55 (6.81)	0.218 (0.039)	0.800 (0.048)	0.505 (0.047)
FL (base), k=3	60.37 (7.32)	0.209 (0.040)	0.800 (0.046)	0.502 (0.050)
FL (base), k=5	60.59 (6.79)	0.220 (0.042)	0.800 (0.047)	0.503 (0.048)
FL (expert), k=0	61.72 (8.03)	0.267 (0.061)	0.750 (0.060)	0.370 (0.067)
FL (expert), k=1	62.08 (8.01)	0.268 (0.063)	0.748 (0.058)	0.369 (0.069)
FL (expert), k=3	61.83 (8.00)	0.271 (0.065)	0.752 (0.059)	0.372 (0.065)
FL (expert), k=5	62.19 (8.21)	0.164 (0.060)	0.751 (0.058)	0.368 (0.067)
LLaMA-3.1 8B (base), k=0	83.63 (32.36)	0.041 (0.062)	0.199 (0.381)	0.523 (0.220)
LLaMA-3.1 8B (base), k=1	56.66 (32.00)	0.318 (0.118)	0.785 (0.182)	0.118 (0.213)
LLaMA-3.1 8B (base), k=3	45.63 (30.24)	0.239 (0.131)	0.684 (0.345)	0.292 (0.302)
LLaMA-3.1 8B (base), k=5	51.22 (32.64)	0.264 (0.123)	0.736 (0.286)	0.217 (0.268)
LLaMA-3.1 8B (expert), k=0	94.25 (22.06)	0.021 (0.038)	0.059 (0.228)	0.563 (0.200)
LLaMA-3.1 8B (expert), k=1	67.17 (36.14)	0.241 (0.102)	0.752 (0.124)	0.135 (0.211)
LLaMA-3.1 8B (expert), k=3	48.85 (30.42)	0.269 (0.144)	0.721 (0.319)	0.200 (0.276)
LLaMA-3.1 8B (expert), k=5	44.43 (28.53)	0.261 (0.129)	0.762 (0.224)	0.270 (0.289)
GPT-3.5-turbo (base), k=0	51.92 (11.14)	0.078 (0.034)	0.797 (0.081)	0.583 (0.055)
GPT-3.5-turbo (base), k=1	46.30 (12.45)	0.071 (0.037)	0.778 (0.124)	0.591 (0.079)
GPT-3.5-turbo (base), k=3	45.45 (12.19)	0.084 (0.041)	0.769 (0.144)	0.587 (0.072)
GPT-3.5-turbo (base), k=5	43.22 (11.66)	0.077 (0.039)	0.779 (0.136)	0.597 (0.072)
GPT-3.5-turbo (expert), k=0	51.63 (12.83)	0.132 (0.081)	0.800 (0.107)	0.535 (0.087)
GPT-3.5-turbo (expert), k=1	44.71 (12.15)	0.109 (0.056)	0.780 (0.125)	0.567 (0.089)
GPT-3.5-turbo (expert), k=3	46.18 (14.98)	0.126 (0.062)	0.789 (0.134)	0.553 (0.095)
GPT-3.5-turbo (expert), k=5	43.12 (12.21)	0.113 (0.071)	0.780 (0.142)	0.570 (0.092)
GPT-4o (base), k=0	62.30 (7.24)	0.091 (0.029)	0.813 (0.045)	0.545 (0.039)
GPT-4o (base), k=1	57.26 (8.00)	0.083 (0.031)	0.790 (0.052)	0.557 (0.041)
GPT-4o (base), k=3	54.57 (7.41)	0.092 (0.035)	0.771 (0.061)	0.575 (0.046)
GPT-4o (base), k=5	54.22 (7.72)	0.090 (0.034)	0.771 (0.064)	0.579 (0.046)
GPT-4o (expert), k=0	72.86 (12.97)	0.214 (0.066)	0.769 (0.057)	0.332 (0.062)
GPT-4o (expert), k=1	59.05 (8.88)	0.193 (0.071)	0.727 (0.070)	0.380 (0.080)
GPT-4o (expert), k=3	57.19 (8.19)	0.201 (0.073)	0.736 (0.068)	0.386 (0.092)
GPT-4o (expert), k=5	57.06 (8.58)	0.204 (0.077)	0.738 (0.068)	0.390 (0.102)
Litake et al. (2024)	87.79 (6.86)	0.211 (0.032)	0.503 (0.093)	–
Wang et al. (2025)	76.67 (8.25)	0.196 (0.045)	0.637 (0.091)	–
Songsiritat (2025)	45.24 (7.24)	0.164 (0.028)	0.647 (0.063)	–

B.12. Distributional Fidelity of Synthetic Clinical Notes across Generation Algorithms using MIMIC-IV Discharge Summaries.

JSD-C and JSD-T denote corpus-level and token-level Jensen–Shannon divergence, respectively (the lower the higher distributional fidelity with the real-world clinical notes). GoF denotes goodness-of-fit (R^2) with Zipf's law on logarithmic scale (the higher, the better fit), while Slope represents the fitted slope after linear regression (close to -1 represents higher tendency for linearity and conformity with log-transformed Zipf's Law curve). Abbreviations: FL, feedback-loop generation; LLaMA, LLaMA-3.1 8B; GPT-3.5, GPT-3.5-turbo.

Algorithm	JSD-C	JSD-T	Slope	GoF
Agentic (base), k=0	0.7292	0.2734	-1.3056	0.9646
Agentic (expert), k=0	0.6961	0.2485	-1.1974	0.9698
FL (base), k=0	0.7388	0.2492	-1.1072	0.9716
FL (base), k=1	0.7393	0.2505	-1.1061	0.9722
FL (base), k=3	0.7397	0.2497	-1.1067	0.9715
FL (base), k=5	0.7385	0.2494	-1.1102	0.9715
FL (expert), k=0	0.7367	0.2504	-1.1818	0.9717
FL (expert), k=1	0.7362	0.2519	-1.1832	0.9714
FL (expert), k=3	0.7370	0.2506	-1.1802	0.9720
FL (expert), k=5	0.7333	0.2505	-1.1783	0.9711
LLaMA-3.1 8B (base), k=0	0.7300	0.2497	-0.7123	0.9244
LLaMA-3.1 8B (base), k=1	0.7166	0.3007	-1.0577	0.9515
LLaMA-3.1 8B (base), k=3	0.7272	0.2695	-0.9789	0.9501
LLaMA-3.1 8B (base), k=5	0.7040	0.2488	-1.0359	0.9529
LLaMA-3.1 8B (expert), k=0	0.7751	0.2787	-0.4795	0.8453
LLaMA-3.1 8B (expert), k=1	0.7194	0.2878	-1.0452	0.9456
LLaMA-3.1 8B (expert), k=3	0.7293	0.2933	-1.0170	0.9531
LLaMA-3.1 8B (expert), k=5	0.6928	0.2245	-1.0502	0.9549
GPT-3.5-turbo (base), k=0	0.7387	0.2539	-1.0385	0.9646
GPT-3.5-turbo (base), k=1	0.7291	0.2423	-1.0341	0.9661
GPT-3.5-turbo (base), k=3	0.7273	0.2493	-1.0419	0.9664
GPT-3.5-turbo (base), k=5	0.7246	0.2511	-1.0187	0.9647
GPT-3.5-turbo (expert), k=0	0.7246	0.2439	-1.0448	0.9632
GPT-3.5-turbo (expert), k=1	0.7147	0.2371	-1.0245	0.9640
GPT-3.5-turbo (expert), k=3	0.7121	0.2356	-1.0369	0.9655
GPT-3.5-turbo (expert), k=5	0.7182	0.2441	-1.0305	0.9647
GPT-4o (base), k=0	0.7529	0.2598	-1.1299	0.9722
GPT-4o (base), k=1	0.7468	0.2538	-1.1156	0.9730
GPT-4o (base), k=3	0.7376	0.2490	-1.1136	0.9738
GPT-4o (base), k=5	0.7393	0.2497	-1.1106	0.9740
GPT-4o (expert), k=0	0.7283	0.2582	-1.3015	0.9722
GPT-4o (expert), k=1	0.6870	0.2214	-1.1730	0.9762
GPT-4o (expert), k=3	0.6713	0.2104	-1.1391	0.9754
GPT-4o (expert), k=5	0.6698	0.2117	-1.1348	0.9750
Litake et al. (2024)	0.8515	0.7195	-1.8361	0.9476
Wang et al. (2025)	0.8086	0.6916	-1.5042	0.9637
Songsiritat (2025)	0.7862	0.6247	-1.5007	0.9487

Model	Configuration	Mean Token Usage (SD)
Agentic		
Base	–	17008.4 (842.6)
Expert	–	23592.7 (1104.8)
Feedback Loop (FL)		
Base	$k = 0$	8400.3 (402.5)
Base	$k = 1$	11248.6 (498.3)
Base	$k = 3$	15893.9 (712.4)
Base	$k = 5$	20374.5 (934.7)
Expert	$k = 0$	12431.8 (556.9)
Expert	$k = 1$	16458.2 (711.3)
Expert	$k = 3$	21492.6 (958.4)
Expert	$k = 5$	25963.7 (1176.2)
GPT-4o		
Base	$k = 0$	1058.4 (72.5)
Base	$k = 1$	1645.2 (96.4)
Base	$k = 3$	2090.6 (118.7)
Base	$k = 5$	2473.8 (136.9)
Expert	$k = 0$	2135.5 (129.6)
Expert	$k = 1$	2734.4 (146.2)
Expert	$k = 3$	3241.7 (175.8)
Expert	$k = 5$	3561.9 (189.4)
GPT-3.5-turbo		
Base	$k = 0$	1131.6 (81.2)
Base	$k = 1$	1671.3 (94.5)
Base	$k = 3$	2112.7 (109.6)
Base	$k = 5$	2550.8 (121.7)
Expert	$k = 0$	1288.4 (86.3)
Expert	$k = 1$	1950.2 (101.5)
Expert	$k = 3$	2422.9 (118.6)
Expert	$k = 5$	2817.5 (133.4)
LLaMA-3.1 8B		
Base	$k = 0$	964.7 (66.1)
Base	$k = 1$	1421.3 (79.4)
Base	$k = 3$	1868.6 (92.7)
Base	$k = 5$	2214.9 (108.2)
Expert	$k = 0$	1743.8 (101.6)
Expert	$k = 1$	2195.4 (116.3)
Expert	$k = 3$	2657.2 (134.7)
Expert	$k = 5$	2983.5 (151.8)

B.13. Comparison of overall mean token usage (SD) across different algorithms and prompting strategies. Token usage was calculated as the sum of the tokens used for input and output.

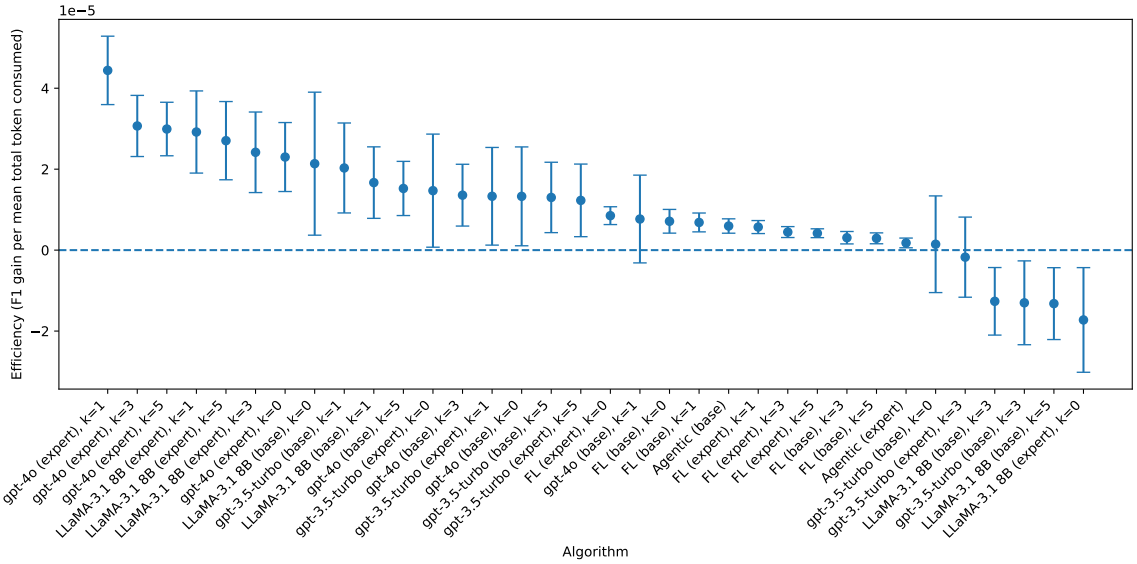
Abbreviations: FL, feedback-loop generation; LLaMA, LLaMA-3.1 8B; GPT-3.5, GPT-3.5-turbo.

B.14. Qualitative Error Analysis Results

Qualitative error analysis was conducted by manually reviewing random samples of 10 synthetic clinical notes per generation algorithms and configurations (i.e., number of demonstrations). The manual review was conducted by the authors. Error analysis was done using the following criteria: 1) hallucination, 2) omission, 3) time error, 4) absence of medical abbreviation, 5) use of complete sentences only, and 6) use of single language only. Hallucination was strictly defined as the inclusion of new clinical information that was not present in the reference document and not relevant or appropriate considering the medical context. Omission was defined as exclusion in the synthetic notes of any clinical information that was present in the reference document. Time error refers to inclusion of temporal information (e.g., dates) that is neither present in the reference document nor contextually relevant. The qualitative error analysis was conducted on the random samples of the synthetic notes generated based on ICSRs. Abbreviations: FL, feedback-loop generation; Agentic, multi-agent generation; LLaMA-3.1 8B, generation by in-context learning of LLaMA-3.1 8B; GPT-3.5-turbo, generation by in-context learning of GPT-3.5-turbo; GPT-4o, generation by in-context learning of GPT-4o; Halluc., Hallucination; Time Err., Time Error; No Abbrev., No Abbreviation; Complete Sent., Complete Sentences Only; Single Lang., Single Language Only.

Table 3: Evaluation metrics across algorithms and configurations.

Algorithm	Halluc.	Omission	Time Err.	No Abbrev.	Complete Sent.	Single Lang.
Agentic (base), k=0	0.85	0.90	0.30	0.40	0.75	0.55
Agentic (expert), k=0	0.20	0.40	0.05	0.05	0.75	0.30
FL (base), k=0	0.25	0.20	0.00	0.00	0.60	0.00
FL (base), k=1	0.05	0.20	0.00	0.00	0.65	0.00
FL (base), k=3	0.05	0.05	0.00	0.00	0.70	0.00
FL (base), k=5	0.00	0.15	0.00	0.00	0.80	0.00
FL (expert), k=0	0.05	0.00	0.00	0.00	0.00	0.00
FL (expert), k=1	0.00	0.05	0.00	0.00	0.05	0.00
FL (expert), k=3	0.00	0.05	0.00	0.00	0.00	0.00
FL (expert), k=5	0.00	0.05	0.00	0.00	0.00	0.00
GPT-3.5-turbo (base), k=0	0.15	0.25	0.10	0.55	0.90	0.45
GPT-3.5-turbo (base), k=1	0.25	0.15	0.10	0.50	0.50	0.50
GPT-3.5-turbo (base), k=3	0.10	0.00	0.10	0.40	0.65	0.75
GPT-3.5-turbo (base), k=5	0.00	0.05	0.05	0.60	0.80	0.70
GPT-3.5-turbo (expert), k=0	0.30	0.35	0.15	0.30	0.80	0.30
GPT-3.5-turbo (expert), k=1	0.25	0.25	0.00	0.30	0.85	0.45
GPT-3.5-turbo (expert), k=3	0.35	0.45	0.00	0.20	0.65	0.10
GPT-3.5-turbo (expert), k=5	0.20	0.15	0.00	0.10	0.70	0.15
GPT-4o (base), k=0	0.00	0.20	0.00	0.00	0.50	0.00
GPT-4o (base), k=1	0.05	0.05	0.00	0.00	0.50	0.00
GPT-4o (base), k=3	0.00	0.00	0.05	0.00	0.50	0.00
GPT-4o (base), k=5	0.00	0.20	0.00	0.00	0.50	0.00
GPT-4o (expert), k=0	0.15	0.00	0.00	0.00	0.00	0.00
GPT-4o (expert), k=1	0.05	0.05	0.00	0.00	0.05	0.00
GPT-4o (expert), k=3	0.00	0.05	0.00	0.00	0.00	0.00
GPT-4o (expert), k=5	0.00	0.00	0.00	0.00	0.05	0.00
LLaMA-3.1 8B (base), k=0	0.10	0.50	0.05	0.45	0.70	0.55
LLaMA-3.1 8B (base), k=1	0.05	0.50	0.00	0.25	0.90	0.35
LLaMA-3.1 8B (base), k=3	0.00	0.45	0.00	0.75	0.80	0.35
LLaMA-3.1 8B (base), k=5	0.00	0.00	0.45	0.50	1.00	0.30
LLaMA-3.1 8B (expert), k=0	0.35	0.45	0.05	0.45	0.40	0.45
LLaMA-3.1 8B (expert), k=1	0.80	0.80	0.35	0.10	0.30	0.15
LLaMA-3.1 8B (expert), k=3	0.70	0.65	0.10	0.75	0.65	0.65
LLaMA-3.1 8B (expert), k=5	0.50	0.55	0.25	0.65	0.70	0.95
Overall	0.171	0.240	0.063	0.215	0.521	0.235



B.15. Efficiency of Each Synthetic Note Generation Algorithm

The figure shows efficiency of each synthetic note generation algorithm calculated as the performance gain adjusted for by the token consumption. This analysis was performed in the downstream task where the LLaMA-3.3 70B model served as the base model for extracting adverse drug reaction evidence from real-world clinical notes. For each algorithm, three synthetic clinical notes were used as demonstrations. Our results show that the highest efficiency was achieved using in-context learning with GPT-4o (expert-informed prompting, k = 1), with an efficiency of 4.44×10^{-5} ($SD = 8.45 \times 10^{-6}$). In contrast, the feedback-loop (FL) approach with expert-informed prompting (k = 0) yielded an efficiency of 8.51×10^{-6} ($SD = 2.2 \times 10^{-6}$). Abbreviations: FL, feedback-loop generation; GPT-3.5, GPT-3.5-turbo; LLaMA, LLaMA-3.1 8B.