ConTextual: Improving Clinical Text Summarization in LLMs with Context-preserving Token Filtering and Knowledge Graphs

Fahmida Liza Piya LIZAPIYA@UDEL.EDU
Department of Computer & Information Sciences
University of Delaware
Newark, Delaware, USA

Rahmatollah Beheshti RBI@UDEL.EDU
Department of Computer & Information Sciences
University of Delaware
Newark, Delaware, USA

Abstract

Unstructured clinical data can serve as a unique and rich source of information that can meaningfully inform clinical practice. Extracting the most pertinent context from such data is critical for exploiting its true potential toward optimal and timely decision-making in patient care. While prior research has explored various methods for clinical text summarization, most prior studies either process all input tokens uniformly or rely on heuristic-based filters, which can overlook nuanced clinical cues and fail to prioritize information critical for decision-making. In this study, we propose `Contextual`, a novel framework that integrates a Context-Preserving Token Filtering method with a Domain-Specific Knowledge Graph (KG) for contextual augmentation. By preserving context-specific important tokens and enriching them with structured knowledge, `ConTextual` improves both linguistic coherence and clinical fidelity. Our extensive empirical evaluations on two public benchmark datasets demonstrate that `ConTextual` consistently outperforms other baselines. Our proposed approach highlights the complementary role of token-level filtering and structured retrieval in enhancing both linguistic and clinical integrity, as well as offering a scalable solution for improving precision in clinical text generation[1].

1. Introduction

Electronic health records (EHRs) are central to modernmedical informatics, providing a rich repository of structured and unstructured data that drives clinical decision-making and research (Piya et al., 2024; Wornow et al., 2023; Poulain et al., 2022). While structured data enables systematic analyses, unstructured components—such as discharge summaries and progress notes—contain nuanced clinical insights that are difficult to summarize due to their reliance on complex medical terminology, subtle contextual cues, and intricate interrelationships (Hossain et al., 2023). Efficiently extracting and summarizing these insights is critical for improving patient care (Piya and Beheshti, 2025), yet it remains an open challenge.

1. Our code repository is publicly available at: https://github.com/healthylaife/ConTextual.

© 2025 F.L. Piya & R. Beheshti.

Recent advancements in large language models (LLMs) have significantly accelerated progress in natural language processing (NLP), particularly in tasks such as clinical summarization and entity extraction (Chen et al., 2024; Aali et al., 2025; Ellershaw et al., 2024). In addition to the rapid development of general-purpose LLMs—such as `GPT-4` (Achiam et al., 2023), `LLaMA 3` (Grattafiori et al., 2024), and `Gemma` (Team et al., 2025)—domain-specific models like `BioInstruct` (Wang et al., 2023), `MediSwift` (Bhardwaj et al., 2024), and `BioMedLM` (Boag et al., 2024) have been introduced to better capture the structure and semantics of biomedical corpora, improving the understanding of medical narratives. These types of models have demonstrated considerable promise in enhancing the accuracy and efficiency of clinical documentation and information extraction (Chen et al., 2024; Hu et al., 2024a). However, they face critical limitations in real-world clinical settings. Specifically, unstructured clinical narratives are often verbose, redundant, and contain various types of nomenclature and jargon, increasing computational demands and obscuring essential information captured in the notes (Hu et al., 2024b). Reliance on fine-tuning and pretraining domain-specific LLMs further amplifies these challenges, as such processes require substantial computational resources and time (Christophe et al., 2024; Liu et al., 2024b). These constraints highlight the need for scalable, domain-aware methodologies that balance contextual fidelity with computational efficiency. Model compression and optimization techniques such as pruning (Ling et al., 2024), quantization (Lang et al., 2024), and distillation (Muralidharan et al., 2024) address some of these computational challenges by reducing the resource demands of deploying LLMs without significantly sacrificing their performance (Zhu et al., 2024).

Such approaches are particularly relevant in healthcare, where the efficient processing of large volumes of data is crucial for timely and accurate patient care. Optimizing LLMs for clinical use can mitigate their high resource requirements, making them more suitable for integration into existing clinical workflows and systems.

Recent studies evaluating LLMs on clinical note summarization highlight significant limitations in fidelity and coherence. Models often generate factually incorrect or fabricated content (hallucinations), posing a major obstacle for clinical use (Oeshy et al., 2024; Fayyaz et al., 2024; Poulain et al., 2024). For example, a physician review of GPT-4-generated emergency department summaries found hallucinated details in 42% of cases and omission of relevant clinical information in 47% (Williams et al., 2024). Such omissions of key medications, diagnoses, or events are common, sometimes due to oversimplification of complex cases (Lee et al., 2024). LLMs also struggle with temporal reasoning: they may misrepresent the chronology of care by focusing on outdated or irrelevant diagnoses (treating them as current) and not recognizing when earlier presumptive diagnoses were later ruled out (Xiong et al., 2024). Even when factual coverage is adequate, the narrative quality can be suboptimal – generated summaries are often less concise and lack the realistic clinical tone or structured flow of human-written notes (Van Veen et al., 2024). Moreover, current LLMs face context length constraints, often requiring truncation or segmentation of long patient records; as a result, many evaluations use isolated note segments instead of full longitudinal records, risking loss of important context (Ravaut et al., 2023). These limitations – hallucinations, omissions, poor handling of temporal context, and subpar coherence – underscore that while LLMs show promise in reducing documentation burden, they are not yet fully reliable for autonomous clinical summarization (Wang et al., 2024; Hager et al., 2024).

To address these challenges, we propose `ConTextual`, a novel framework that integrates a context-preserving token filtering (CPTF) approach with a domain-specific knowledge graph (KG) to enhance clinical text summarization. CPTF leverages attention mechanisms to dynamically identify and retain semantically significant tokens, minimizing computational costs while preserving critical clinical information. To mitigate information loss from token filtering, the domain-specific KG encodes structured relationships among clinical entities, such as diagnoses, treatments, and outcomes. This integration ensures that the retained tokens are enriched with domain-relevant context, enabling the framework to maintain contextual fidelity in complex clinical scenarios.

Generalizable Insights about Machine Learning in the Context of Healthcare

The proposed framework provides a scalable and efficient solution for processing verbose clinical narratives. By prioritizing clinically significant tokens and enriching them with structured knowledge, `ConTextual` achieves superior information retention, computational efficiency, and contextual depth. Our extensive evaluations demonstrate superior performance in summarization ability and reduction in latency and computational costs, making the framework particularly suited for complex and resource-constrained healthcare environments. Although tailored for clinical note summarization, the modular design of `ConTextual` supports broader applicability to other biomedical domains requiring efficient and domain-specific natural language understanding, such as medical literature review. In particular, our contributions are as follows:

- We propose a context-preserving token filtering (CPTF) method that dynamically compresses unstructured clinical text by removing redundancy while retaining essential information.
- We construct a domain-specific knowledge graph (KG) and integrate it with the CPTF to form a structured and interpretable framework that enhances contextual fidelity during token selection.
- We improve the contextual input for retrieval-augmented generation (RAG), enabling more effective LLM-based reasoning and demonstrating improved performance across two clinical datasets.
- We validate the scalability and efficiency of `ConTextual` through extensive evaluations using metrics, including lexical, semantic, and LLM-based measures.

2. Related Work

Clinical Text Summarization Existing summarization models rely heavily on standard attention mechanisms, which scale quadratically with sequence length. For instance, `BioGPT` (Luo et al., 2022b) and `PubMedBERT` (Gu et al., 2021) utilize biomedical corpora to refine performance, but their reliance on uncompressed token sequences results in inefficiencies for lengthy clinical notes. Models like `Flan-T5` (Lyu et al., 2024) have introduced instruction-tuned objectives to improve summarization; however, they may fail to address the redundancy of verbose clinical narratives, where attention mechanisms struggle to focus

on critical contextual cues (Hu et al., 2024b). To mitigate these issues, recent work (Han and Choi, 2024) has proposed models such as `Pointer-GPT`, which replace standard attention mechanisms with a pointer network to enhance content retention during summarization. However, such models may still suffer from factual inconsistency, highlighting the ongoing need for summarization approaches that balance precision, coherence, and domain specificity.

Model Compression and Optimization While model compression techniques such as pruning (Frantar and Alistarh, 2023), quantization (Dettmers et al., 2022), and distillation (Hinton, 2015) effectively reduce model size and latency, they often degrade performance on domain-specific tasks due to loss of fine-grained contextual information (Tinn et al., 2023). For instance, pruning reduces model complexity by zeroing out low-magnitude weights, but in clinical NLP tasks, even small weights can encode critical semantic relationships (Ma et al., 2023). Similarly, knowledge distillation transfers knowledge from large to smaller models (Ho et al., 2022), but these smaller models may lack the capacity to retain nuanced biomedical context (Magister et al., 2022).

Token Filtering and Attention Mechanisms Token filtering methods(e.g.,`PoWER-BERT`), rely on attention scores to progressively prune less relevant tokens (Goyal et al., 2020). However, these approaches operate in encoder-only architectures and are incompatible with the autoregressive decoding required by generative models typically employed in large language model applications. `Prunepert` introduced a differentiable perturbed top-k mechanism for token selection, but its reliance on stochastic perturbations increases variance in summarization outcomes. By contrast, `CPTF` operates natively within the multi-layered attention framework of modern open-source LLMs, dynamically weighting attention layers to compute token importance without introducing architectural modifications.

Moreover, long and heterogeneous texts significantly increase latency and computational overhead, which limits their scalability in resource-constrained clinical environments (He et al., 2025). Token filtering methods aim to mitigate this by dynamically retaining contextually important tokens while discarding less relevant ones (Lin et al., 2024; Lou et al., 2024). While effective in reducing computational demands, these methods often result in partial information loss, particularly in complex, domain-specific scenarios such as healthcare. Additionally, they are not readily integrable with structured knowledge, such as knowledge graphs, which can compensate for the loss of contextual information (He et al., 2025; Liu et al., 2024a).

Knowledge Graph Integration in NLP KGs are extensively utilized for encoding structured relationships, offering enhanced contextualization and interpretability in NLP applications (Peng et al., 2023; Sharma et al., 2022). In the biomedical domain, KGs like UMLS and SNOMED-CT have been employed for tasks such as entity linking and ontology-based query expansion (Lu et al., 2025; Arsenyan et al., 2024; Hu et al., 2023). By encoding explicit relationships (e.g., between diseases, symptoms, and treatments), KGs can enhance a model's contextual understanding and interpretability, helping align NLP outputs with established medical knowledge (He et al., 2025; Liu et al., 2024a; Piya and Beheshti, 2025). However, most prior approaches integrate KGs in a static fashion—treating the graph as a fixed resource—which means the knowledge base does not dynamically update or adapt to new data or task-specific needs, potentially limiting scalability and flexibility in fast-evolving clinical settings (Authors, 2024). Recent work has begun exploring more

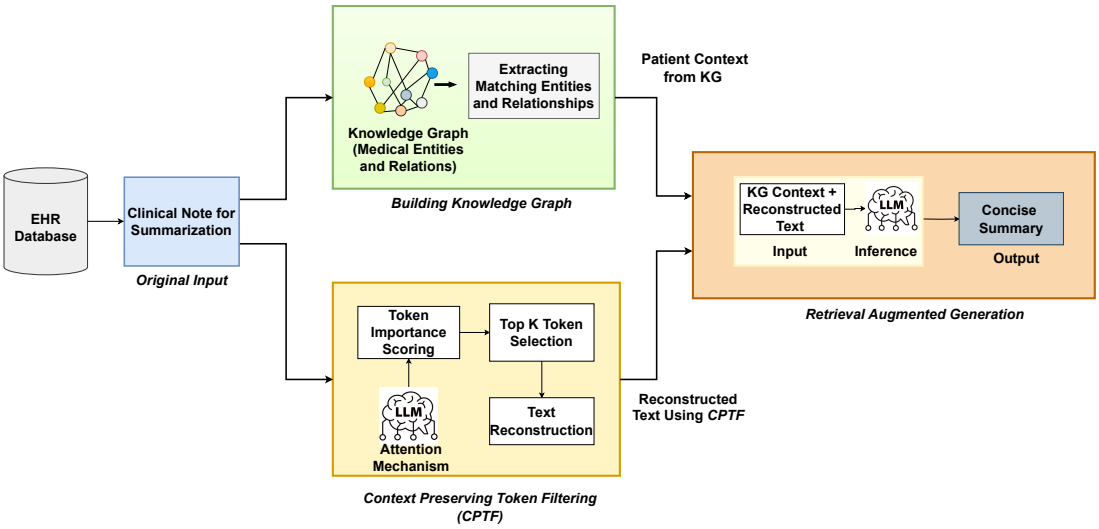

Figure 1: Overview of the `ConTextual` Framework for clinical text summarization.

.

dynamic KG integration strategies (e.g., continual graph updates or tailored subgraph retrieval) to improve adaptability (Arsenyan et al., 2024), but static KGs remain invaluable as authoritative repositories of biomedical knowledge. In this context, our work `ConTextual` contributes a novel method that effectively leverages a static domain-specific biomedical KG within a clinical summarization framework, demonstrating that a fixed, curated graph can be harnessed to provide relevant structured context and significantly improve the factuality and clinical fidelity of LLM-generated summaries (Lu et al., 2025).

3. Methods

We propose `ConTextual`, a framework for clinical text summarization that addresses the challenges of long and verbose narratives through three key components: (1) Context-Preserving Token Filtering (CPTF), (2) Domain-specific KG integration, and (3) an LLM inference with retrieval-augmented generation (RAG).

CPTF dynamically reduces redundancy by filtering out tokens with low contextual significance based on the attention mechanisms of LLMs, ensuring that the retained input is computationally efficient and semantically rich. The KG integration component enriches this reduced input by embedding structured relationships among clinical entities, such as diagnoses, treatments, and medications, to provide domain-specific context. Finally, the RAG component retrieves additional relevant context from the KG during inference, integrating it into the summarization process to maintain accuracy and adaptability for specific queries. Figure 1 shows the overall structure of the framework and its three components.

Problem Formulation

Prior to presenting the components of the proposed method in detail, we first present an overall description of the full framework. Let $\mathcal{D} = \{d_1, ..., d_N\}$ denote the dataset comprising

clinical notes. Each note $d \in \mathcal{D}$ consists of an input sequence $d = \{t_1, ..., t_n\}$, where $t_i \in \mathcal{V}$ and $\mathcal{V}$ denotes the vocabulary of tokens. The objective is to generate a contextually enhanced, reduced representation d_{reduced} that retains essential medical information while minimizing sequence length.

The text reduction task can be formalized as identifying a mapping function $f \in \mathcal{F}$, where $\mathcal{F}$ is a family of candidate reduction functions. The optimization objective is expressed as:

$$\max_{f \in \mathcal{F}} \sum_{d \in \mathcal{D}} \text{sim}(d, f(d)), \tag{1}$$

where $\text{sim}(d, f(d))$ denotes a general similarity function that captures the semantic preservation between the original and reduced representations. This similarity function serves as a conceptual abstraction to formalize the goal of retaining semantic fidelity. In practice, this objective is operationalized via deterministic ranking of tokens based on layer-weighted attention scores.

The reduced representation $f(d)$ is subject to a length constraint:

$$|f(d)| = \lfloor r \cdot |d| \rfloor, \quad r \in (0, 1], \tag{2}$$

where r denotes the *retention ratio*, specifying the proportion of tokens retained in d_{reduced}.

To ensure coherence and clinical accuracy in the reconstructed and reduced text, LLM leverages knowledge from a reference medical knowledge graph, denoted as $G = (V, E, \mathcal{R})$, where V is the set of nodes (e.g., medical entities), E is the set of edges (relationships), and $\mathcal{R}$ represents the types of relationships. We define a context retrieval function $\eta : d \to 2^V$, where 2^V denotes the power set of V, i.e., the set of all subsets of V. This function retrieves a subset of relevant nodes from the KG based on the input sequence d.

The final summarization objective combines both the reduced text (through CPFT) and the enhanced context (through KG) as:

$$s^* = \arg\max_{s \in \mathcal{S}} P(s \mid [f(d); \eta(d)]; \theta), \tag{3}$$

where $\mathcal{S}$ is the set of candidate summaries, $f(d)$ is the reduced text representation, $\eta(d)$ is the retrieved context, and θ denotes the model parameters. We now present the model and its three components in more detail.

3.1. Context Preserving Token Filtering (CPTF)

The CPTF framework is illustrated in Figure 2, and a detailed algorithm corresponding to this part is presented in Appendix A. This framework processes clinical text sequences by leveraging multi-head attention mechanisms from LLM to compute token-level importance, thereby preserving semantic fidelity while optimizing computational efficiency.

We consider multiple attention heads $(B_1, B_2, \ldots, B_n)$ that independently compute token interactions, capturing diverse semantic aspects as depicted in Figure 2. Each input token i is first projected into three representations: *query* (q), *key* (k), and *value* (v). The query (q)

interacts with keys (k) across the sequence to compute attention weights, which are then applied to the values (v) to generate weighted representations of the input tokens.

The attention outputs from multiple heads are concatenated into a single vector (B), which aggregates critical semantic features across tokens. The extraction and selection of tokens are guided by their computed importance scores, allowing for the retention of the most contextually significant tokens. These tokens are then reconstructed into a reduced clinical narrative, maintaining the focus on preserving essential information with high semantic relevance.

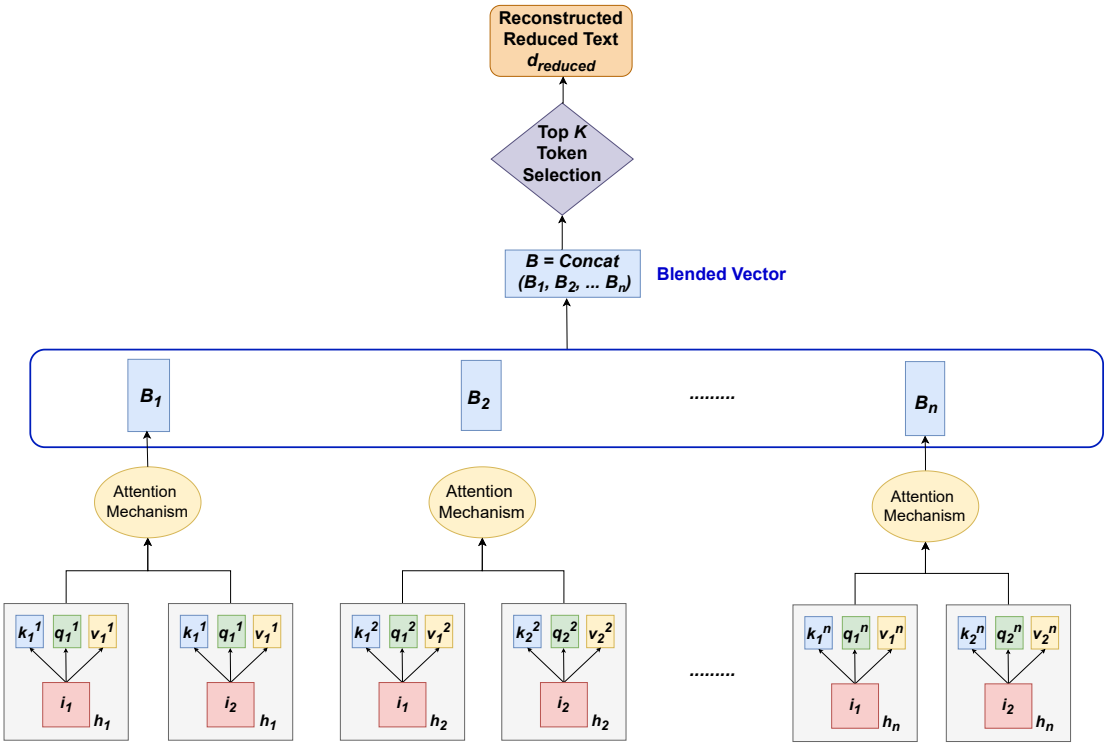

Figure 2: Overview of the Context Preserving Token Filtering algorithm.

Token Selection Mechanism Our methodology continues by extracting attention patterns from an LLM, wherein each input token is processed through transformer layers. Each layer $l \in \{1, ..., L\}$ computes attention matrices $A_l^h \in \mathbb{R}^{n \times n}$, indicating token-to-token attention weights. The attention matrices from multiple heads are blended into a single matrix for each layer, calculated as:

$$\bar{A}_l = \frac{1}{H} \sum_{h=1}^{H} A_l^h, \tag{4}$$

where H is the total number of attention heads. This consolidated attention information helps in determining the hierarchical semantic structure across layers, ultimately influencing the selection of tokens that contribute most significantly to the clinical narrative's context

and meaning. To capture the hierarchical semantic structure across layers, we compute layer weights as:

$$w_l = \alpha + (1 - \alpha)\frac{l}{L}, \tag{5}$$

where $\alpha \in [0, 1]$ is a tunable hyperparameter that balances the contribution of base-level features extracted from the lower layers with the more abstract features from higher layers. This balance aims to maintain a robust representation of both fundamental and complex features in the processed text, thereby preserving the integrity and richness of the clinical narrative.

Token Importance Scores and Selection The token importance score I_i is computed for each token at position i by aggregating the weighted attention patterns across all layers. This step captures both local syntactic relationships and global semantic dependencies:

$$I_i = \sum_{l=1}^{L} w_l \cdot \frac{1}{n} \sum_{j=1}^{n} \bar{A}_l[i, j], \tag{6}$$

where w_l is the layer-specific weight, n denotes sequence length, and $\bar{A}_l[i, j]$ represents the attention strength between tokens t_i and t_j. Using these scores, we select the top k tokens by solving the constrained optimization problem:

$$S^* = \arg\max_{S \in \mathcal{S}} \sum_{i \in S} I_i, \tag{7}$$

subject to:

$$|S| = k, \quad k = \lfloor r \cdot n \rfloor \tag{8}$$

$$\forall i, j \in S : i < j \implies \text{pos}(i) < \text{pos}(j), \tag{9}$$

where $r \in (0, 1]$ is the retention ratio, and $\text{pos}(i)$ maps token position i to its original sequence index.

Reconstruction of Reduced Narrative The final reduced sequence is reconstructed by mapping the selected token indices to their corresponding tokens and positional encodings:

$$d_{reduced} = \{(t_i, \text{pos}(i)) \mid i \in S^*\} \tag{10}$$

3.2. Domain-specific Knowledge Graph Construction

To enhance the understanding of the general medical context, we construct a reference KG $G = (V, E, \mathcal{R})$ using clinical records from the same cohort of patients for whom we analyze the clinical notes. The vertex set V comprises entity types representing the clinical domain hierarchy: $V = V_d \cup V_m \cup V_t$, where V_d, V_m, and V_t correspond to diagnoses, medications, and treatments, respectively. These entities are interconnected through a set of edges $E \subseteq V \times V$, which capture the general clinical relationships. The relationship types in the graph are defined as: $\mathcal{R} = \{r_{dm}, r_{dt}, r_{mt}\}$, where r_{dm}, r_{dt}, and r_{mt} represent the relationships between diagnoses and medications, diagnoses and treatments, and medications and treatments,

respectively. These relationships are used to map out the complex interactions within the clinical data.

For a given clinical note d, we utilize a retrieval function $\eta : d \to V_p$, where $V_p \subset V$ is the set of entities matching the unique patient identifier in the clinical note. Each entity $e \in V_p$ represents a clinical concept linked to the patient. The contextual information retrieved from the KG is formalized as:

$$C(d) = \{(e', r) \mid e \in V_p, (e, e', r) \in E\}, \tag{11}$$

where e' denotes an entity in the KG that shares a relationship r with e.

This arrangement retrieves entities directly linked to the same patient identifier as in the clinical notes, ensuring a match of patient-specific entities and their relationships. The extracted context enriches the input by concatenating the reduced clinical note with the retrieved context, forming the final enriched input for downstream processing:

$$\hat{d} = d_{\text{reduced}} \oplus C(d), \tag{12}$$

where, $\oplus$ denotes concatenation. This approach ensures that the enriched representation $\hat{d}$ retains key patient-specific details while incorporating domain knowledge from the knowledge graph. The constructed knowledge graph was evaluated using standard metrics, including MRR and Hit@10, as described in Appendix C.

3.3. Summarization with RAG

The summarization process leverages the patient-specific reduced text from CPTF d_{reduced} and dynamically retrieved KG context $C(d)$ using a RAG approach. The summarization objective aims to generate the most clinically relevant and coherent summary s^* from a set of possible summaries $\mathcal{S}$. This optimal summary is selected by the model as:

$$s^* = \arg\max_{s \in \mathcal{S}} P(s \mid \hat{d}; \theta), \tag{13}$$

where θ denotes the parameters of the model (i.e., the summary generator LLM). This approach conditions the generation of summaries on both the extracted entities and their associated knowledge from the KG, allowing the LLM to refine its understanding of patient-specific narratives and improve factual consistency and medical coherence. Since entity retrieval is performed via overlapping patient identifiers, the retrieved context remains directly relevant to the clinical note, ensuring adaptability across diverse cases. By dynamically incorporating structured medical knowledge from the KG, this approach enables LLMs to generate clinically coherent and factually consistent summaries while maintaining computational efficiency.

4. Experiments

4.1. Experimental Setup

We implement both the CPTF module and the primary summary generation component using the instruction-tuned LLaMA 3.2 1B model (AI, 2024). All experiments are conducted with a fixed generation budget of 200 tokens and a decoding temperature of 0.7 to ensure consistency and comparability across runs.

Data For the summarization task, we utilized the MIMIC-IV-Ext-BHC dataset (Aali et al., 2024a), derived from the MIMIC-IV-Note database, consisting of 270,033 clinical notes with corresponding brief hospital course (BHC) summaries. The preprocessing involved standardizing the structure of the note, cleaning the formatting, and normalizing the length of the token to an average of 2,267 tokens per note. The resulting curated dataset provides a structured resource for clinical text summarization research (Aali et al., 2024b).

The second dataset we incorporated comprises 1,473 patient-doctor conversations from the FigShare (Singh, 2011) and MTS-Dialog (Abacha et al., 2023) collections, specifically designed for generating clinical summaries. These conversations have been annotated to create structured SOAP (Subjective, Objective, Assessment, and Plan) summaries, available on Hugging Face datasets (Neupane, 2024).

Prompt Design We use a few-shot design to provide the language model with structured input-output examples that establish a consistent format for clinical summarization. Few-shot prompting outperformed zero-shot and one-shot strategies across key evaluation metrics, as shown in Table 7 (Appendix). These observed improvements indicate that providing the model with representative clinical examples helps constrain its output format, enhancing factual coherence and reducing hallucinations. Additionally, few-shot learning ensures that the model remains aligned with clinical terminology and structured reporting conventions, addressing concerns about variability in generated summaries.

The offered examples in the few-shot design align structured input tags—such as `<SEX>`, `<SERVICE>`, `<CHIEF COMPLAINT>`, and `<HISTORY OF PRESENT ILLNESS>`—with corresponding target summaries, capturing linguistic patterns and contextual nuances prevalent in clinical narratives. A description of the few-shot prompt design, including illustrative examples, is provided in Table 8.

Baselines We selected a range of models, chosen for their diverse approach to handling the challenges of clinical text summarization:

- `Longformer` by Beltagy et al. (2020) utilizes a sparse attention mechanism to handle long documents by extending attention spans up to 16K tokens, particularly suited for detailed clinical narratives.
- `BioBART` by Yuan et al. (2022) is specifically pre-trained on biomedical corpora and fine-tuned for medical summarization.
- `T5-Large` by Raffel et al. (2020) is a general-purpose sequence-to-sequence model that excels in diverse text-to-text tasks, testing the adaptability of transformers in specialized domains.
- `Flan-T5` by Chung et al. (2022) extends `T5` with instruction tuning, aiming to improve the model's ability to learn from descriptive tasks and generalizing across various NLP applications.
- `BioGPT` by Luo et al. (2022a) offers an adaptation of the GPT architecture tailored to understand and generate biomedical text, focusing on maintaining clinical accuracy and relevance.
- `Gemma3-Instruct(1B)` by Google (Team et al., 2025) is a open-source general-purpose LLM.
- `Mistral-7B-Instruct` by Mistral AI (Jiang et al., 2023) is the second open-source general-purpose LLM we use.

Evaluation Metrics We use the following statistical metrics to evaluate the accuracy and relevance of the generated summaries.

- **BLEU** (Papineni et al., 2002): Measures n-gram precision to quantify lexical overlap between generated and reference summaries.
- **ROUGE-L** (Lin, 2004): Captures the longest common subsequence between generated and reference summaries, emphasizing recall and precision while reducing redundancy.
- **BERT-Score** (Zhang et al., 2019): Computes semantic similarity using contextual embeddings, ensuring accurate representation of clinical meaning.

We evaluate using BLEU-1 (B-1), BLEU-2 (B-2), and ROUGE-L (R-L) for surface-level fidelity, and precision (P), recall (R), and F1-score derived from BERTScore to quantify semantic alignment with gold-standard references.

To complement statistical evaluation metrics and better assess the clinical fidelity and coherence of generated summaries, we also employ an 'LLM as a judge' framework. We use an instruction-tuned `Gemma 3.1B` model (Team et al., 2025) to evaluate the generated summaries relative to their corresponding reference along three critical axes: Main Idea Retention, Coherence, and Factual Consistency. Evaluators—instantiated through LLM prompting—were instructed to assign a score from 1 (poor) to 5 (excellent) for each criterion, guided by a structured evaluation prompt.

Additionally, we use two metrics to study the scalability and resource optimization of the proposed method.

- **Throughput** (Vaswani, 2017): Calculates summaries generated per second, showcasing scalability for large datasets.
- **Latency** (Narang et al., 2021): Evaluates the time taken to generate a single summary, reflecting the computational cost and efficiency of different prompting strategies.

4.2. Results

Ablation Analysis Table 1 reports the performance of three model configurations evaluated on the MIMIC-BHC and SOAP datasets. We compare the full model against two systematically reduced variants: (i) a baseline LLM without Context-Preserving Token Filtering (Vanilla LLM) and (ii) an intermediate variant that incorporates CPTF but excludes Knowledge Graph augmentation (Vanilla LLM + CPTF). The consistent performance degradation across both datasets—particularly in ROUGE-L and BERT-F1—underscores the complementary contributions of CPTF and KG integration to the overall effectiveness of the proposed framework.

Table 1: **Performance Comparison of Clinical Summarization Models on MIMIC-BHC and SOAP Datasets.** BERT-P, BERT-R, and BERT-F1 refer to precision, recall, and F1 score using BERT embeddings.

Dataset	Model	BLEU-1 (↑)	BLEU-2 (↑)	ROUGE-L (↑)	BERT-P (↑)	BERT-R (↑)	BERT-F1 (↑)
MIMIC-BHC	`LLaMA 3.2`	4.52 ± 6.1	1.79 ± 2.6	8.85 ± 3.5	$\mathbf{83.02} \pm 2.0$	78.64 ± 2.2	80.77 ± 1.7
	`LLaMA 3.2 +CPTF`	5.99 ± 5.5	1.82 ± 2.0	7.08 ± 2.9	81.82 ± 2.4	80.12 ± 2.1	80.97 ± 1.8
	`ConTextual`	$\mathbf{9.06} \pm 6.1$	$\mathbf{3.35} \pm 2.6$	$\mathbf{9.98} \pm 3.5$	82.72 ± 2.0	$\mathbf{80.32} \pm 2.2$	$\mathbf{81.48} \pm 1.7$
SOAP Summary	`LLaMA 3.2`	4.13 ± 5.5	2.22 ± 3.7	8.16 ± 6.5	82.87 ± 2.8	82.93 ± 3.0	82.90 ± 2.8
	`LLaMA 3.2 + CPTF`	9.29 ± 7.3	3.98 ± 4.1	8.75 ± 5.5	82.58 ± 3.2	82.38 ± 2.8	82.45 ± 2.6
	`ConTextual`	$\mathbf{11.55} \pm 5.5$	$\mathbf{6.09} \pm 4.2$	$\mathbf{10.70} \pm 4.5$	$\mathbf{83.51} \pm 1.6$	$\mathbf{83.70} \pm 3.0$	$\mathbf{83.60} \pm 2.2$

On MIMIC-BHC, ConTextual achieves a BERT-F1 of 81.48 ± 1.7, outperforming LLaMA 3.2 (80.77 ± 1.7) and CPTF-enhanced LLaMA 3.2 (80.97 ± 1.8). A similar pattern is observed on the SOAP dataset, where ConTextual attains the highest BERT-F1 of 83.60 ± 2.2 compared to 82.90 ± 2.8 and 82.45 ± 2.6 from the respective baselines. Improvements are also evident in lexical metrics: on SOAP, ConTextual increases BLEU-1 from 4.13 to **11.55** and BLEU-2 from 2.22 to **6.09**, indicating enhanced surface-level coherence and informativeness. We also explore the effects of temperature scaling and token limit adjustments and present the results in Appendix F.

We also use the LLM as a judge evaluator to score each output according to defined criteria. Table 2 presents the mean and standard deviation of these scores across datasets and models. Each column corresponds to one of the three evaluation criteria, with the final column (Avg.) representing the average of the three scores per instance. `ConTextual` stands

Table 2: **LLM-as-a-Judge Evaluation.** Scores have a 1-5 scale.

Dataset	Model	Main Ideas (↑)	Coherence (↑)	Factuality (↑)	Average Score (↑)
MIMIC-BHC	`LLaMA 3.2`	**4.56 ± 1.33**	3.78 ± 0.67	3.89 ± 1.17	4.07 ± 0.95
	`LLaMA 3.2 + CPTF`	3.93 ± 1.33	3.38 ± 0.77	3.77 ± 1.09	3.73 ± 0.87
	`ConTextual`	4.45 ± 1.21	**3.82 ± 0.75**	**4.55 ± 0.82**	**4.27 ± 0.55**
SOAP Summary	`LLaMA 3.2`	4.39 ± 0.85	3.71 ± 0.85	4.06 ± 1.03	4.09 ± 0.69
	`LLaMA 3.2 + CPTF`	4.70 ± 0.66	3.80 ± 0.52	3.84 ± 1.17	4.13 ± 0.60
	`ConTextual`	**5.00 ± 0.00**	**4.50 ± 0.55**	**4.67 ± 0.52**	**4.72 ± 0.25**

out as the winner model across the evaluation dimensions, consistently achieving the highest scores in most criteria. Its performance advantage is particularly evident in the SOAP dataset, where it attains near-ceiling ratings with minimal variance. These results position `ConTextual` as the most effective approach among those evaluated, highlighting the value of integrating context-preserving token filtering with structured knowledge representations in clinically grounded summarization tasks.

Table 3: Performance Comparison with Baseline Models Across Datasets.

Dataset	Model	BLEU-1 (↑)	BLEU-2 (↑)	ROUGE-L (↑)	BERT F1 (↑)
MIMIC-BHC	Longformer	2.76 ± 5.0	0.75 ± 2.0	3.10 ± 3.0	74.70 ± 1.5
	BioBART	6.88 ± 5.2	2.05 ± 2.1	8.00 ± 3.2	78.10 ± 1.6
	T5-Large	4.95 ± 5.3	1.41 ± 2.2	7.96 ± 3.3	79.84 ± 1.6
	Flan-T5	10.52 ± 5.8	2.54 ± 2.3	9.90 ± 3.4	77.91 ± 1.6
	BioGPT	6.15 ± 5.4	6.17 ± 2.5	7.47 ± 3.2	77.83 ± 1.6
	Gemma3-Instruct(1B)	7.89 ± 5.6	3.20 ± 2.4	9.85 ± 3.4	79.78 ± 1.6
	Mistral-7B-Instruct	4.43 ± 5.2	2.09 ± 2.2	9.87 ± 3.4	80.71 ± 1.7
	ConTextual (Ours)	**9.06** ± 6.1	**3.35** ± 2.6	**9.98** ± 3.5	**81.48** ± 1.7
SOAP Summary	Longformer	2.18 ± 3.6	1.19 ± 2.6	6.76 ± 5.8	75.00 ± 2.9
	BioBART	5.27 ± 4.2	1.83 ± 2.8	7.41 ± 4.9	77.32 ± 2.5
	T5-Large	3.86 ± 3.9	1.24 ± 2.4	7.52 ± 5.2	78.46 ± 2.7
	Flan-T5	8.35 ± 5.1	2.17 ± 2.9	8.91 ± 5.0	77.05 ± 2.8
	BioGPT	5.62 ± 4.8	5.11 ± 3.7	7.08 ± 4.7	76.93 ± 2.6
	Gemma3-Instruct(1B)	7.25 ± 5.3	2.84 ± 3.2	8.97 ± 5.3	79.25 ± 2.5
	Mistral-7B-Instruct	4.08 ± 4.4	1.85 ± 2.7	9.34 ± 5.6	81.18 ± 2.4
	ConTextual (Ours)	**11.55** ± 5.5	**6.09** ± 4.2	**10.70** ± 4.5	**83.60** ± 2.2

Comparison with Baselines As shown in Table 3, `ConTextual` consistently outperforms all baselines across both datasets and evaluation metrics. The improvements in BLEU-1 and BLEU-2 reflect superior lexical overlap with reference summaries, while higher ROUGE-L and BERT F1 scores suggest stronger structural alignment and semantic preservation. These gains are particularly pronounced on the MIMIC-BHC dataset, highlighting our model's effectiveness in handling longer and more complex clinical narratives. Notably, even when compared with instruction-tuned models such as Mistral-7B and Gemma3, `ConTextual` delivers higher performance, underscoring the value of structured retrieval and token filtering in medical summarization tasks.

Table 4: **Efficiency Metrics Comparison Across Models.**

Dataset	Model	Throughput (↑)	Latency (↓)
MIMIC-BHC	`LLaMA 3.2`	36.72 ± 2.44	$\mathbf{3.61 \pm 1.32}$
	`LLaMA 3.2 + CPTF`	139.10 ± 10.04	12.38 ± 1.25
	`ConTextual`	$\mathbf{142.87 \pm 16.94}$	14.29 ± 2.17
SOAP Summary	`LLaMA 3.2`	28.35 ± 36.26	16.15 ± 10.69
	`LLaMA 3.2 + CPTF`	108.43 ± 12.75	$\mathbf{14.22 \pm 2.73}$
	`ConTextual`	$\mathbf{116.82 \pm 19.64}$	15.78 ± 3.42

Computational Efficiency Table 4 reports the efficiency characteristics of the proposed models and the relationship between computational performance and knowledge integration. Notably, both CPTF and `ConTextual` improve throughput compared to the base `LLaMA` 3.2 model, suggesting that context-aware token filtering enhances generation efficiency by reducing irrelevant token processing. The latency increases in the case of the MIMIC-BHC dataset, possibly due to the additional steps introduced by structured retrieval and filtering. This may reflect the trade-off between performance and computational efficiency, where modest increases in processing time are offset by substantial gains in output quality and overall generation efficiency.

5. Discussion

This work introduces `ConTextual`, a structured framework for clinical text summarization that combines context-preserving token filtering (CPTF) with domain-specific knowledge graphs (KGs). CPTF dynamically reduces textual redundancy while preserving essential clinical information, and KG integration ensures that token selection aligns with structured domain knowledge. By improving the quality of inputs in a RAG structure, `ConTextual` enables more accurate and semantically grounded LLM-based reasoning. We validate the framework across two clinical datasets, demonstrating improvements in both generation quality and system efficiency, as measured by BLEU, ROUGE, BERTScore, LLM-based scores, latency, and throughput. Notably, it achieves up to a 1.5x improvement in BLEU-1 and a 31% increase in ROUGE-L on SOAP summaries while delivering the highest BERTScore-F1 across all settings. Beyond summarization, this framework has broader applicability in real-world clinical environments: alleviating documentation burden for providers, streamlining cohort identification for clinical trial recruitment, and enabling

small to mid-sized healthcare organizations to deploy high-quality language models under constrained computational budgets.

Limitations While `ConTextual` advances clinical summarization, it is constrained by its reliance on a static, domain-specific knowledge graph. This may restrict its generalizability to broader or evolving clinical domains, particularly in the context of rare conditions or emerging practices. Additionally, the framework assumes consistent quality and structure in clinical documentation, which may limit its robustness when applied to noisy, incomplete, or institution-specific records. Our preliminary analysis showed the value of developing a KG specific to the targeted domain (as determined by the input data). Future work can incorporate strategies for leveraging publicly available KGs to enhance adaptability. We also plan to support dynamic knowledge graph construction, expand entity coverage, and implement mechanisms for handling variability in input quality. These enhancements can further improve generalizability and resilience across diverse healthcare settings.

Acknowledgments

Our study was supported by the NIH award U54-GM104941 and P20-GM103446, as well as computing credits from Amazon Web Services (AWS).

References

A. Aali, D. Van Veen, Y. Arefeen, J. Hom, C. Bluethgen, E. P. Reis, S. Gatidis, N. Clifford, J. Daws, A. Tehrani, J. Kim, and A. Chaudhari. MIMIC-IV-Ext-BHC: Labeled Clinical Notes Dataset for Hospital Course Summarization. PhysioNet, 2024a. URL https://physionet.org/content/labelled-notes-hospital-course/1.1.0/.

A Aali, D Van Veen, YI Arefeen, et al. A dataset and benchmark for hospital course summarization with adapted llms. *Journal of the American Medical Informatics Association*, 32(3):470–479, 2025.

Asad Aali, Dave Van Veen, Yamin Ishraq Arefeen, Jason Hom, Christian Bluethgen, Eduardo Pontes Reis, Sergios Gatidis, Namuun Clifford, Joseph Daws, Arash S Tehrani, et al. A benchmark of domain-adapted large language models for generating brief hospital course summaries. *arXiv preprint arXiv:2403.05720*, 2024b.

Asma Ben Abacha, Wen-wai Yim, Yadan Fan, and Thomas Lin. An empirical study of clinical note generation from doctor-patient encounters. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2291–2302, 2023.

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Meta AI. Llama 3.2 1b instruct. https://huggingface.co/meta-llama/Llama-3.2-1B-Instruct, 2024. Accessed: April 2025.

Vahan Arsenyan et al. Large language models for biomedical knowledge graph construction: Information extraction from emr notes. In *Proceedings of the 23rd Workshop on Biomedical Natural Language Processing*, Bangkok, Thailand, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.bionlp-1.23. URL https://aclanthology.org/2024.bionlp-1.23/.

Authors. A review on knowledge graphs for healthcare: Resources, applications, and promises. *Journal Name*, Volume(Number):Page Range, 2024.

Iz Beltagy, Matthew E Peters, and Arman Cohan. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*, 2020.

Rishabh Bhardwaj, Kishan Patel, Simran Khanuja, Diptesh Kanojia Sharma, and Pushpak Bhattacharyya. Mediswift: Building fast and accurate biomedical language models with sparse pretraining. *arXiv preprint arXiv:2403.00952*, 2024.

William Boag, Felix Tan, Ritwik Das, Brett K Beaulieu-Jones, and Andrew L Beam. Biomedlm: A domain-specific foundation model for biomedical natural language processing. *arXiv preprint arXiv:2403.18421*, 2024.

F Chen, BY Miao, AJ Butte, et al. Evaluating llms for drafting ed discharge summaries. *medRxiv*, 2024. Preprint available: medRxiv.

Clément Christophe, Praveen K Kanithi, Prateek Munjal, Tathagata Raha, Nasir Hayat, Ronnie Rajan, Ahmed Al-Mahrooqi, Avani Gupta, Muhammad Umar Salman, Gurpreet Gosal, et al. Med42–evaluating fine-tuning strategies for medical llms: full-parameter vs. parameter-efficient approaches. *arXiv preprint arXiv:2404.14779*, 2024.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, et al. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*, 2022.

Tim Dettmers, Mike Lewis, Younes Belkada, and Luke Zettlemoyer. Gpt3. int8 (): 8-bit matrix multiplication for transformers at scale. *Advances in Neural Information Processing Systems*, 35:30318–30332, 2022.

Darren Edge, Ha Trinh, Newman Cheng, Joshua Bradley, Alex Chao, Apurva Mody, Steven Truitt, and Jonathan Larson. From local to global: A graph rag approach to query-focused summarization. *arXiv preprint arXiv:2404.16130*, 2024.

S Ellershaw, C Tomlinson, OE Burton, et al. Automated generation of hospital discharge summaries using clinical guidelines and llms. In *AAAI Spring Symposium: Clinical Foundation Models*, 2024.

Hamed Fayyaz, Raphael Poulain, and Rahmatollah Beheshti. Enabling scalable evaluation of bias patterns in medical llms. *arXiv preprint arXiv:2410.14763*, 2024.

Elias Frantar and Dan Alistarh. Sparsegpt: Massive language models can be accurately pruned in one-shot. In *International Conference on Machine Learning*, pages 10323–10337. PMLR, 2023.

Saurabh Goyal, Anamitra Roy Choudhury, Saurabh Raje, Venkatesan Chakaravarthy, Yogish Sabharwal, and Ashish Verma. Power-bert: Accelerating bert inference via progressive word-vector elimination. In *International Conference on Machine Learning*, pages 3690–3699. PMLR, 2020.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare (HEALTH)*, 3(1):1–23, 2021.

Paul Hager, Friederike Jungmann, Kunal Bhagat, Inga Hubrecht, Manuel Knauer, Jakob Vielhauer, Robbie Holland, Rickmer Braren, Marcus Makowski, Georgios Kaisis, et al. Evaluating and mitigating limitations of large language models in clinical decision making. *medRxiv*, pages 2024–01, 2024.

Hyunkyung Han and Jaesik Choi. Optimal path for biomedical text summarization using pointer gpt. *arXiv preprint arXiv:2404.08654*, 2024.

Kai He, Rui Mao, Qika Lin, Yucheng Ruan, Xiang Lan, Mengling Feng, and Erik Cambria. A survey of large language models for healthcare: from data, technology, and applications to accountability and ethics. *Information Fusion*, page 102963, 2025.

Geoffrey Hinton. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

Namgyu Ho, Laura Schmid, and Se-Young Yun. Large language models are reasoning teachers. *arXiv preprint arXiv:2212.10071*, 2022.

Elias Hossain, Rajib Rana, Niall Higgins, Jeffrey Soar, Prabal Datta Barua, Anthony R Pisani, and Kathryn Turner. Natural language processing in electronic health records in relation to healthcare decision-making: a systematic review. *Computers in biology and medicine*, 155:106649, 2023.

Linmei Hu, Zeyi Liu, Ziwang Zhao, Lei Hou, Liqiang Nie, and Juanzi Li. A survey of knowledge enhanced pre-trained language models. *IEEE Transactions on Knowledge and Data Engineering*, 2023.

Y Hu, Q Chen, J Du, et al. Improving llms for clinical ner via prompt engineering. *Journal of the American Medical Informatics Association*, 31(9):1812–1820, 2024a.

Zhiyuan Hu, Yuliang Liu, Jinman Zhao, Suyuchen Wang, Yan Wang, Wei Shen, Qing Gu, Anh Tuan Luu, See-Kiong Ng, Zhiwei Jiang, et al. Longrecipe: Recipe for efficient long context generalization in large language models. *arXiv preprint arXiv:2409.00509*, 2024b.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, et al. Mistral 7b, 2023.

Jiedong Lang, Zhehao Guo, and Shuyu Huang. A comprehensive study on quantization techniques for large language models. In *2024 4th International Conference on Artificial Intelligence, Robotics, and Communication (ICAIRC)*, pages 224–231. IEEE, 2024.

Chanseo Lee, Kimon A Vogt, and Sonu Kumar. Prospects for ai clinical summarization to reduce the burden of patient chart review. *Frontiers in Digital Health*, 6:1475092, 2024.

Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81, 2004.

Zhenghao Lin, Zhibin Gou, Yeyun Gong, Xiao Liu, Yelong Shen, Ruochen Xu, Chen Lin, Yujiu Yang, Jian Jiao, Nan Duan, et al. Rho-1: Not all tokens are what you need. *arXiv preprint arXiv:2404.07965*, 2024.

Gui Ling, Ziyang Wang, and Qingwen Liu. Slimgpt: Layer-wise structured pruning for large language models. *Advances in Neural Information Processing Systems*, 37:107112–107137, 2024.

Fenglin Liu, Hongjian Zhou, Yining Hua, Omid Rohanian, Lei Clifton, and David Clifton. Large language models in healthcare: A comprehensive benchmark. *medRxiv*, pages 2024–04, 2024a.

Qidong Liu, Xian Wu, Xiangyu Zhao, Yuanshao Zhu, Derong Xu, Feng Tian, and Yefeng Zheng. When moe meets llms: Parameter efficient fine-tuning for multi-task medical applications. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 1104–1114, 2024b.

Chao Lou, Zixia Jia, Zilong Zheng, and Kewei Tu. Sparser is faster and less is more: Efficient sparse attention for long-range transformers. *arXiv preprint arXiv:2406.16747*, 2024.

Yuxing Lu et al. Biomedical knowledge graph: A survey of domains, tasks, and real-world applications. *arXiv preprint arXiv:2501.11632*, 2025. URL https://arxiv.org/abs/2501.11632.

Renqian Luo, Liai Sun, Yingce Xia, Tao Qin, Sheng Zhang, Hoifung Poon, and Tie-Yan Liu. BioGPT: generative pre-trained transformer for biomedical text generation and mining. *Briefings in Bioinformatics*, 23(6), 09 2022a. ISSN 1477-4054. doi: 10.1093/bib/bbac409. URL https://doi.org/10.1093/bib/bbac409. bbac409.

Renqian Luo, Liai Sun, Yingce Xia, Tao Qin, Sheng Zhang, Hoifung Poon, and Tie-Yan Liu. Biogpt: generative pre-trained transformer for biomedical text generation and mining. *Briefings in bioinformatics*, 23(6):bbac409, 2022b.

Mengxian Lyu, Cheng Peng, Xiaohan Li, Patrick Balian, Jiang Bian, and Yonghui Wu. Automatic summarization of doctor-patient encounter dialogues using large language model through prompt tuning. *arXiv preprint arXiv:2403.13089*, 2024.

Xinyin Ma, Gongfan Fang, and Xinchao Wang. Llm-pruner: On the structural pruning of large language models. *Advances in neural information processing systems*, 36:21702–21720, 2023.

Lucie Charlotte Magister, Jonathan Mallinson, Jakub Adamek, Eric Malmi, and Aliaksei Severyn. Teaching small language models to reason. *arXiv preprint arXiv:2212.08410*, 2022.

Saurav Muralidharan, Sharath Turuvekere Sreenivas, Raviraj Joshi, Marcin Chochowski, Mostofa Patwary, Mohammad Shoeybi, Bryan Catanzaro, Jan Kautz, and Pavlo Molchanov. Compact language models via pruning and knowledge distillation. *Advances in Neural Information Processing Systems*, 37:41076–41102, 2024.

Sharan Narang, Hyung Won Chung, Yi Tay, William Fedus, Thibault Fevry, Michael Matena, Karishma Malkan, Noah Fiedel, Noam Shazeer, Zhenzhong Lan, et al. Do transformer modifications transfer across implementations and applications? *arXiv preprint arXiv:2102.11972*, 2021.

Subash Neupane. Soap summary dataset, 2024. URL https://huggingface.co/datasets/SubashNeupane/dataset_SOAP_summary.

Nafisa Tabassum Oeshy, Ajwad Abrar Mostofa, and Prianka Maheru. *Improving Faithfulness in Medical Text Summarization: An LLM-Based Approach*. PhD thesis, Department of Computer Science and Engineering (CSE), Islamic University of . . . , 2024.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318, 2002.

Ciyuan Peng, Feng Xia, Mehdi Naseriparsa, and Francesco Osborne. Knowledge graphs: Opportunities and challenges. *arXiv preprint arXiv:2303.13948*, 2023.

Fahmida Liza Piya and Rahmatollah Beheshti. Advancing feature extraction in healthcare through the integration of knowledge graphs and large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 29293–29294, 2025.

Fahmida Liza Piya, Mehak Gupta, and Rahmatollah Beheshti. Healthgat: Node classifications in electronic health records using graph attention networks. In *2024 IEEE/ACM Conference on Connected Health: Applications, Systems and Engineering Technologies (CHASE)*, pages 132–141. IEEE, 2024.

Raphael Poulain, Mehak Gupta, and Rahmatollah Beheshti. Few-shot learning with semi-supervised transformers for electronic health records. In *Machine Learning for Healthcare Conference*, pages 853–873. PMLR, 2022.

Raphael Poulain, Hamed Fayyaz, and Rahmatollah Beheshti. Aligning (medical) llms for (counterfactual) fairness. *arXiv preprint arXiv:2408.12055*, 2024.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140):1–67, 2020.

Mathieu Ravaut, Aixin Sun, Nancy F Chen, and Shafiq Joty. On context utilization in summarization with large language models. *arXiv preprint arXiv:2310.10570*, 2023.

Shikhar Sharma, Rahul Thotempudi, et al. Knowledge graphs in biomedical natural language processing: A survey. *Briefings in Bioinformatics*, 23(1), 2022.

Jatinder Singh. Figshare. *Journal of pharmacology & pharmacotherapeutics*, 2(2):138, 2011.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, et al. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*, 2025.

Robert Tinn, Hao Cheng, Yu Gu, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. Fine-tuning large neural language models for biomedical natural language processing. *Patterns*, 4(4), 2023.

Dave Van Veen, Cara Van Uden, Louis Blankemeier, Jean-Benoit Delbrouck, Asad Aali, Christian Bluethgen, Anuj Pareek, Malgorzata Polacin, Eduardo Pontes Reis, Anna Seehofnerová, et al. Adapted large language models can outperform medical experts in clinical text summarization. *Nature medicine*, 30(4):1134–1142, 2024.

A Vaswani. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017.

Chonghua Wang, Haodong Duan, Songyang Zhang, Dahua Lin, and Kai Chen. Ada-leval: Evaluating long-context llms with length-adaptable benchmarks. *arXiv preprint arXiv:2404.06480*, 2024.

Zhexin Wang, Vishakh Padmakumar Kumar, Tianyu Kang, Chang Xu, Tianle Cai, Xuanjing Ma, Yefeng Zheng, Zhengping Liu, and Meng Jiang. Bioinstruct: Bionlp tasks instruction tuning for llama model. *arXiv preprint arXiv:2310.19975*, 2023.

Christopher YK Williams, Jaskaran Bains, Tianyu Tang, Kishan Patel, Alexa N Lucas, Fiona Chen, Brenda Y Miao, Atul J Butte, and Aaron E Kornblith. Evaluating large language models for drafting emergency department discharge summaries. *medRxiv*, 2024.

Michael Wornow, Rahul Thapa, Ethan Steinberg, Jason Fries, and Nigam Shah. Ehrshot: An ehr benchmark for few-shot evaluation of foundation models. *Advances in Neural Information Processing Systems*, 36:67125–67137, 2023.

Siheng Xiong, Ali Payani, Ramana Kompella, and Faramarz Fekri. Large language models can learn temporal reasoning. *arXiv preprint arXiv:2401.06853*, 2024.

Hongyi Yuan, Zheng Yuan, Ruyi Gan, Jiaxing Zhang, Yutao Xie, and Sheng Yu. Biobart: Pretraining and evaluation of a biomedical generative language model. *arXiv preprint arXiv:2204.03905*, 2022.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*, 2019.

Xunyu Zhu, Jian Li, Yong Liu, Can Ma, and Weiping Wang. A survey on model compression for large language models. *Transactions of the Association for Computational Linguistics*, 12:1556–1577, 2024.

Appendix A. Context-Preserving Token Filtering Algorithm

We introduce an algorithm—*Context-Preserving Token Filtering (CPTF)*—designed to retain semantically important tokens from an input sequence while minimizing overall length. By leveraging internal attention dynamics from a multi-layer transformer model, CPTF computes layer-weighted token importance scores to identify and preserve tokens that are critical for maintaining contextual coherence and clinical accuracy. While our implementation uses the instruction-tuned `LLaMA 3.2 1B`, the method is model-agnostic and applicable to a range of transformer-based language models. Its computational efficiency and minimal architectural assumptions make it well-suited for deployment in low-resource or compute-limited settings.

Algorithm 1: Context-Preserving Token Filtering (CPTF)

Input: Text sequence x (string), language model M with L layers and H attention heads, tokenizer T, retention ratio $r \in (0, 1]$ (float), base weight $\alpha \in [0, 1]$ (float)

Output: Reduced sequence (string) retaining the most informative tokens

Step 1: Tokenization and Initialization $tokens \leftarrow T.\text{encode}(x)$ `// Convert text to tokens`

$n \leftarrow |tokens|$ `// Determine total number of tokens`

$k \leftarrow \lfloor n \cdot r \rfloor$ `// Determine number of tokens to retain`

$I \leftarrow [0] * n$ `// Initialize importance scores array`

Step 2: Calculate Token Importance **for** $l = 1$ **to** L **do**

- $w_l \leftarrow \alpha + (1 - \alpha) \cdot \frac{l}{L}$ `// Layer-specific weight`
- $A_l \leftarrow M.\text{attention}(tokens, l)$ `// Compute attention for layer` l
- $\bar{A}_l \leftarrow \frac{1}{H} \sum_{h=1}^{H} A_{l,h}$ `// Average over all heads`
- **for** $i = 1$ **to** n **do**
  - $I[i] \leftarrow I[i] + w_l \cdot \frac{1}{n} \sum_{j=1}^{n} \bar{A}_l[i, j]$ `// Update importance score`
- **end**

end

Step 3: Token Selection and Reconstruction $S \leftarrow \text{argsort}(-I)[:k]$ `// Select indices of top-`k` important tokens`

$tokens_{\text{reduced}} \leftarrow [tokens[i] for i in \text{sorted}(S)]$ `// Retrieve and sort tokens`

return $T.\text{decode}(tokens_{\text{reduced}})$ `// Reconstruct reduced sequence`

Appendix B. Knowledge Graph Construction

Traditional retrieval-augmented generation (RAG) models face limitations in synthesizing information from diverse sources, particularly when understanding requires identifying shared attributes or underlying semantic relationships Edge et al. (2024). To address these limitations, we constructed a domain-specific knowledge graph (KG) that encapsulates critical aspects of patient-level clinical data. The KG represents key entities, including *Problems*, *Treatments*, *Tests*, and *Patients*, along with their relationships, providing a structured and queryable representation of clinical interactions.

The constructed KG consists of 7,095 nodes distributed as follows: **Patients** (100 nodes), **Problems** (3,841 nodes), **Treatments** (1,686 nodes), and **Tests** (1,468 nodes). These

entities are interconnected through 11,443 relationships categorized into three primary types: `HAS_PROBLEM` (6,760 edges), `UNDERWENT_TEST` (5,469 edges), and `WAS_TREATED_WITH` (3,214 edges). These relationships encode clinically meaningful associations, such as diagnoses associated with patients, tests performed, and treatments administered.

Entity extraction was performed using a domain-specific named entity recognition (NER) pipeline built with the `samrawal/bert-base-uncased_clinical-ner` model. To process long clinical narratives, the pipeline segmented text into overlapping chunks, adhering to the model's token limits while preserving entity coherence. Extracted entities were grouped into the predefined categories and linked to patients, forming the basis for graph construction. Overlapping patient nodes denote cases where multiple diagnoses, tests, and treatments are associated with a single individual, effectively capturing the multi-relational structure inherent in clinical datasets.

The graph structure supports dynamic traversal to extract contextually relevant subgraphs for specific queries or tasks. For example, when responding to a patient-centric query, the KG enables the efficient extraction of related diagnoses, tests, and treatments, leveraging its multi-relational structure to ensure specificity and precision. This capability facilitates integration with downstream tasks, such as retrieval-augmented generation (RAG), contextual language modeling, and predictive analytics.

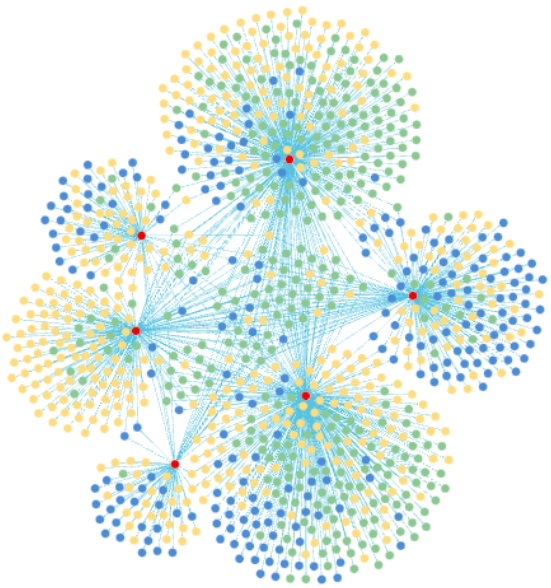

Figure 3: Visualization of the clinical knowledge graph constructed from unstructured EHR data. The graph comprises 1,000 nodes, categorized into **Patients** (red), **Problems** (yellow), **Tests** (green), and **Treatments** (blue). Relationships between nodes (**1,187 edges**) include *HAS_PROBLEM*, *UNDERWENT_TEST*, and *WAS_TREATED_WITH*, encoding critical clinical entity interactions. This structured representation facilitates interpretable and domain-aware contextualization for downstream tasks, such as summarization and retrieval.

Figure 3 illustrates a subset of the constructed KG, showing the relationships between patients, problems, treatments, and tests. By embedding clinical data into a graph structure, the KG provides a scalable and interpretable framework for contextualizing patient information and supporting advanced machine learning applications in the healthcare domain. In addition to the MIMIC-IV dataset, we applied the same knowledge graph construction pipeline to a structured SOAP-format clinical summary dataset. This yielded a domain-specific graph aligned with the Subjective, Objective, Assessment, and Plan sections. While the core methodology remained consistent, minor adaptations were introduced to account for the structural characteristics of patient–provider dialogue. The resulting graph was similarly integrated into our retrieval framework to evaluate the generalizability of our approach across diverse clinical documentation formats.

Appendix C. Knowledge Graph Evaluation

To assess the quality of the constructed knowledge graph (KG), we performed a quantitative evaluation using three widely adopted metrics: *Entity Matching Accuracy*, *Hit@10*, and *Mean Reciprocal Rank (MRR)*.

Entity Matching Accuracy was found to be 100%, indicating that all nodes—including patients, problems, treatments, and tests—were correctly and completely labeled with valid identifiers.

For retrieval-based evaluation, we considered a set of Q query triples $\{q_1, q_2, \ldots, q_Q\}$, where each query q_i involves retrieving the correct associated entity a_i^* from a ranked list of candidates $\{a_i^{(1)}, a_i^{(2)}, \ldots, a_i^{(k)}\}$. Let rank_i denote the rank position of a_i^* in the retrieved list for query q_i.

The *Mean Reciprocal Rank (MRR)* is defined as:

$$\text{MRR} = \frac{1}{Q} \sum_{i=1}^{Q} \frac{1}{\text{rank}_i} \tag{14}$$

This metric reflects how highly the correct entity appears in the ranked list, with higher MRR values indicating better ranking performance. In our evaluation, MRR was computed as 0.750, suggesting that correct entities were, on average, ranked among the top two results.

The *Hit@10* score is defined as:

$$\text{Hit@10} = \frac{1}{Q} \sum_{i=1}^{Q} \mathbb{I}(\text{rank}_i \leq 10) \tag{15}$$

where $\mathbb{I}(\cdot)$ is the indicator function, which equals 1 if the correct entity appears within the top-10 retrieved results, and 0 otherwise. A Hit@10 value of 0.95 indicates that 95% of the queries successfully returned the correct entity within the top 10 candidates.

These results confirm that the knowledge graph maintains high data integrity and strong retrieval performance. Notably, the Hit@10 metric exceeds the 90% threshold commonly expected for clinical knowledge graphs, supporting the KG's effectiveness as a reliable context source for downstream retrieval-augmented summarization and clinical reasoning tasks.

Appendix D. Layer Weighting Strategy

The choice of $\alpha = 0.5$ for our experiments was driven by its role in balancing the contributions of features from different transformer layers, which is pivotal for achieving a robust performance across various metrics. This balance ensures an effective integration of nuanced, deep contextual information from lower layers with more immediate, surface-level features from higher layers, thus preventing the model from overfitting to syntactic structures at the expense of semantic coherence. Such equilibrium is essential for the application in clinical environments where both types of information are crucial. The stability of ROUGE-L scores across different values of α, as shown in Fig 4, supports this choice by indicating a consistent capture of relevant content regardless of slight variations in expression or phrasing. Therefore, $\alpha = 0.5$ represents a strategic decision to optimize the overall efficacy and reliability of the model in real-world applications. For a detailed mathematical formulation of how α influences layer weighting, refer to Equation 5 where we define the weighting scheme across the layers.

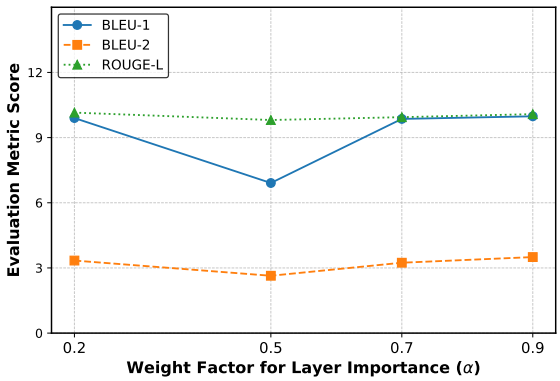

Figure 4: Comparing results with Weight factor for layer importance

Appendix E. Additional Experiments

To better understand how different generation settings and model improvements affect performance, we conduct additional experiments across a range of token limits and temperature settings. Table 5 presents performance metrics for all model variants. The results show that the inclusion of Context-Preserving Token Filtering (CPTF) consistently improves LLaMA 3.2 across all configurations, while the ConTextual model achieves the best performance on several metrics, particularly at higher token budgets. Notably, the highest BERT Precision (84.09) is achieved by ConTextual at Token=100, Temp=0.1, while the highest ROUGE-L (11.04) and B-2 (4.65) scores are observed at Token=300, Temp=0.1, indicating strong performance with extended generation. These findings suggest that both architectural modifications and generation hyperparameters play a critical role in optimizing clinical summarization quality.

The temperature parameter, T, was varied across $\{0.1, 0.7, 0.9\}$, influencing the stochasticity of token sampling. Lower values ($T = 0.1$) enforce deterministic outputs, ensuring

high precision in extracted entities, while higher values ($T = 0.9$) promote diversity in text generation, potentially capturing a broader semantic spectrum. Similarly, the maximum token constraint was adjusted across {100, 200, 300}, enabling a direct investigation of sequence length on linguistic coherence and computational feasibility.

Table 5: **Performance Metrics for Model Variants.** Higher values indicate better performance for all metrics (↑). The best result for each metric is highlighted in **bold**.

Model	Token	Temp	B-1 (↑)	B-2 (↑)	R-L (↑)	BERT-P (↑)	BERT-R (↑)	BERT-F1 (↑)
LLaMA 3.2	100	0.1	10.92	5.45	7.37	79.58	80.23	79.89
		0.7	10.90	5.40	7.33	79.58	80.23	79.89
		0.9	10.91	5.41	7.32	79.58	80.23	79.89
	200	0.1	10.76	5.37	7.33	79.58	80.23	79.89
		0.7	10.73	5.35	7.24	79.58	80.23	79.89
		0.9	10.73	5.34	7.25	79.58	80.23	79.89
	300	0.1	10.70	5.34	7.28	79.58	80.23	79.89
		0.7	10.68	5.33	7.26	79.58	80.23	79.89
		0.9	10.64	5.30	7.20	79.58	80.23	79.89
LLaMA 3.2 + CPTF	100	0.1	15.12	6.75	8.92	80.39	81.63	80.99
		0.7	15.25	6.81	8.99	80.39	81.63	80.99
		0.9	**15.28**	6.79	9.00	80.39	81.63	80.99
	200	0.1	14.63	6.52	8.74	80.39	81.63	80.99
		0.7	14.95	6.65	8.88	80.39	81.63	80.99
		0.9	15.10	6.72	8.93	80.39	81.63	80.99
	300	0.1	14.26	6.36	8.58	80.39	81.63	80.99
		0.7	14.75	6.58	8.80	80.39	**81.63**	80.99
		0.9	14.82	6.54	8.77	80.39	81.63	80.99
ConTextual	100	0.1	3.80	1.48	8.73	**84.09**	79.30	81.36
		0.7	3.60	1.30	8.30	83.76	79.29	81.44
		0.9	3.24	1.21	7.83	83.43	79.11	81.19
	200	0.1	9.49	3.60	10.68	83.05	80.35	**81.65**
		0.7	9.06	3.35	9.98	82.72	80.32	81.48
		0.9	8.64	2.90	9.33	82.23	80.18	81.18
	300	0.1	12.63	**4.65**	**11.04**	82.19	80.61	81.37
		0.7	12.17	4.26	10.37	82.11	80.66	81.36
		0.9	11.96	3.73	9.83	81.66	80.48	81.05

Appendix F. CPTF Workflow Example

To demonstrate the practical application and efficacy of the proposed Context-Preserving Token Filtering (CPTF) mechanism, we present an example workflow that illustrates the transformation of a verbose clinical note into a reduced, semantically significant representation and, finally, into a concise summary. This process highlights the capability of CPTF to optimize input text for efficient processing while retaining clinically relevant information.

Table 6: **Example** CPTF Workflow. From Original Clinical Note to Final Summary. This workflow demonstrates how CPTF reduces verbosity while retaining clinically relevant insights.

Stage	Content
Input	<SEX> M, <SERVICE> MEDICINE, <ALLERGIES> ibuprofen, <CHIEF COMPLAINT> Fever, altered mental status. This is a middle-aged male with a past medical history significant for ruptured AVM, status post craniotomy, and prior intracranial abscess, who is presenting today to the emergency department with fever and altered mental status. On admission, he was noted to be febrile and somewhat confused. A non-contrast CT scan of the head was performed which showed no acute intracranial abnormalities. Laboratory workup revealed elevated CRP and thrombocytosis. The patient was subsequently diagnosed with a urinary tract infection due to Klebsiella species and prostatitis. He was started on broad-spectrum antibiotics. Neurology and infectious disease teams were consulted for further management. A PET scan was obtained which was suggestive of prostatitis.
CPTF Output	<SERVICE> MEDICINE, <CHIEF COMPLAINT> Fever, altered mental status. Middle-aged male with history of ruptured AVM, craniotomy, and intracranial abscess presenting with fever and altered mental status. On admission, febrile and confused. CT head showed no acute intracranial abnormalities. Labs showed elevated CRP and thrombocytosis. Diagnosed with Klebsiella UTI and prostatitis. Started on antibiotics. Neurology and infectious disease consulted. PET scan suggestive of prostatitis.
Summary	Patient with ruptured AVM and intracranial abscess presented with fever and altered mental status. Imaging showed no acute intracranial changes. Labs revealed elevated CRP. Diagnosed with Klebsiella UTI and prostatitis, treated with antibiotics. Follow-up with neurology and infectious disease advised.

The workflow, illustrated in Table 6, demonstrates the transformation of unstructured clinical text through three distinct stages. It begins with the original clinical note, which comprises verbose and unstructured text. This input, while containing critical medical insights, is often interspersed with redundant and extraneous details that hinder computational efficiency. The Context-Preserving Token Filtering (CPTF) mechanism is then applied to process the input text, dynamically identifying and retaining semantically significant tokens essential for downstream tasks. By filtering out irrelevant and redundant information, CPTF reduces verbosity and computational overhead while preserving key clinical insights. Finally, the reduced text, enriched with contextual domain knowledge, is utilized to generate a concise and clinically actionable summary. This final output aligns with domain-specific requirements and effectively supports clinical decision-making by distilling complex narratives into precise and meaningful insights. This example highlights the role of CPTF in improving efficiency and preserving essential clinical details. By combining token filtering with the knowledge-enhanced summarization pipeline, the workflow ensures that the final output is both computationally optimized and clinically relevant.

Appendix G. Prompting Strategies

To guide the language model's generation process, we construct prompts in a structured manner for each input instance. Each prompt begins with a task-specific instruction that explicitly defines the summarization objective, optionally includes exemplar demonstrations (in one-shot or few-shot settings), and concludes with the clinical input to be summarized. This design grounds the model's output in domain-specific linguistic and clinical conventions while retaining the flexibility to accommodate the variability inherent in clinical narratives. Integrated into our framework, this prompting strategy significantly improves the model's ability to produce coherent, concise, and clinically actionable summaries that facilitate downstream tasks such as entity extraction and structured knowledge graph construction. We evaluate three prompting paradigms—zero-shot, one-shot, and few-shot—using standard summarization metrics, as detailed in Table 7. Empirically, few-shot prompting consistently outperforms both zero-shot and one-shot configurations across all evaluation metrics. In particular, higher ROUGE-L scores indicate improved lexical and structural alignment with reference summaries. Similarly, BERT-based metrics reveal superior F1 scores, reflecting a more effective balance between semantic precision and recall.

These results are consistent with prior findings in in-context learning, demonstrating that incorporating a small number of representative exemplars enhances the model's generalization capabilities. Our findings underscore the importance of carefully engineered prompts in optimizing language model performance on clinical summarization tasks.

Table 7: **Performance Comparison of Different Prompting Strategies.** Results show that few-shot prompting achieves superior performance on ROUGE-L and BERT metrics compared to zero-shot and one-shot approaches. Higher values indicate better performance across all metrics.

Prompting Strategy	**Lexical Alignment**			**Semantic Alignment**		
	BLEU-1 (%)	**BLEU-2 (%)**	**ROUGE-L (%)**	**BERT-P (%)**	**BERT-R (%)**	**BERT-F1 (%)**
Zero-shot	11.85	6.41	8.59	79.83	**81.95**	80.58
One-shot	12.99	6.17	8.12	79.85	81.69	80.74
Few-shot	**12.63**	**4.65**	**11.04**	**82.19**	80.61	**81.36**

Appendix H. Few-Shot Prompt Design

For each input instance, the model is guided by a dynamically constructed prompt. The prompt begins with a clear instruction, contextualizing the task as a clinical summarization problem. It incorporates curated examples as demonstrations of the desired output style, concluding with the specific input instance requiring summarization. This structured approach transitions seamlessly from exemplar summaries to the new input, providing the model with implicit guidelines for the task.

Table 8: Few-Shot Prompt Template for Clinical Summarization. Each example demonstrates the input structure, task-specific context, and desired output style. These examples guide the model in generating high-quality clinical summaries for unseen instances.

Component	**Example: Oncology**
Instruction	Summarize the provided clinical notes to produce a concise, domain-specific summary. Focus on clinically relevant information while omitting redundant details.
Input	45-year-old female with stage IV metastatic breast cancer. Chief Complaint: Severe thoracic back pain. **Imaging:** MRI spine reveals T5-T7 vertebral compression fractures; PET-CT shows multiple bone metastases. **Treatments:** Morphine IV PCA for pain management, radiation oncology consult, zoledronic acid 4mg IV for bone metastases, continued letrozole 2.5mg daily. **Labs:** CA 15-3: 68 U/mL (elevated), alkaline phosphatase: 220 U/L.
Target Summary	Patient with metastatic breast cancer underwent comprehensive pain management and palliative interventions, including morphine PCA, radiation consultation, and bone-targeted therapy.
Component	**Example: Cardiology**
Instruction	Summarize the provided clinical notes to generate a focused cardiology case summary.
Input	55-year-old male with history of hypertension and smoking. Chief Complaint: Acute chest pain radiating to left arm. **Diagnostics:** ECG shows ST-segment elevation in inferior leads; Troponin I: 12.4 ng/mL (significantly elevated); Cardiac ultrasound reveals anterior wall hypokinesis. **Interventions:** Immediate cardiac catheterization, primary PCI to right coronary artery, drug-eluting stent placement. **Medications:** Aspirin 325mg, atorvastatin 80mg, metoprolol 25mg. **Labs:** CK-MB: 22.5 ng/mL, LDL: 142 mg/dL.
Target Summary	Patient diagnosed with acute myocardial infarction underwent immediate primary percutaneous coronary intervention with right coronary artery stenting and initiated comprehensive cardiac medical management.
Component	**Example: Internal Medicine**
Instruction	Summarize the provided clinical notes to generate a concise and medically accurate case summary.

Continued on next page

Component	Example: Oncology
Input	65-year-old female with severe COPD and 40-pack-year smoking history. Chief Complaint: Acute respiratory distress. **Diagnostics:** Chest X-ray shows bilateral hyperinflation; ABG: pH 7.32, PaCO2 65 mmHg; Spirometry: FEV1 32% predicted. **Interventions:** Non-invasive ventilation, IV methylprednisolone 125mg, nebulized albuterol and ipratropium. **Medications:** Prednisone 40mg daily, azithromycin 500mg. **Labs:** WBC: 14,200/µL, sputum culture: Pseudomonas aeruginosa.
Target Summary	*(To be generated)*

The design emphasizes domain-specific contextualization by leveraging structured tags that align with clinical documentation conventions. The curated few-shot examples, as illustrated in Table 8, act as semantic anchors, enabling the model to adapt effectively to the intricacies of medical narratives. This approach dynamically accommodates the variability inherent in clinical notes while preserving consistency across varying inputs. By grounding the language model's generative capabilities in domain-specific examples, the prompt design facilitates the production of high-quality, semantically faithful summaries. When integrated into our Knowledge Graph-Enhanced Attention-Guided Summarization pipeline, this methodology enhances the extraction of actionable insights from complex medical narratives, advancing both the accuracy and utility of clinical summarization.

Appendix I. Model Selection

LLaMA 3.2 1B was chosen due to its popularity as a lightweight yet effective foundation model. Developed by Meta AI, LLaMA (Large Language Model Meta AI) is a family of transformer-based autoregressive models optimized for efficiency and scalability. The 1B variant is specifically designed for resource-constrained environments while maintaining strong performance across various NLP tasks. Unlike traditional large-scale models, LLaMA 3.2 1B prioritizes computational efficiency, making it a suitable choice for real-world applications where inference speed and accessibility are critical factors. The model's architecture incorporates improvements in tokenization, attention mechanisms, and optimization techniques, enabling it to handle text generation tasks effectively with a relatively small parameter count. Its strong performance in retrieval-augmented and few-shot learning scenarios further supports its applicability in domains requiring efficient text processing. To address our choice of a lightweight model over newer large-scale commercial LLMs, we conducted a comparative experiment with GPT-4o during the rebuttal phase. While GPT-4o is a state-of-the-art proprietary model with access to extensive pretraining resources, it achieved a BERTScore F1 of 81.20% on our evaluation benchmark. In contrast, our proposed method *ConTextual*, which uses the LLaMA 3.2 1B model, achieved a slightly higher BERTScore F1 of 81.37%. Though the improvement is marginal, ConTextual demonstrates more stable and consistent performance across clinical inputs, an important consideration for safety-critical

applications. Moreover, this performance is attained without the reliance on commercial LLM infrastructure and avoids the risk of potential data contamination—an issue often unaccounted for when benchmarking proprietary models on public datasets. Given these results, LLaMA 3.2 1B not only serves as a representative lightweight baseline but also enables transparent and reproducible experimentation in healthcare NLP applications.

Appendix J. Use Case: Clinical Trial Recruitment

Efficient participant identification remains a major bottleneck in clinical trial recruitment. Key eligibility criteria—such as age, comorbidities, prior treatments, and disease progression—are typically embedded in unstructured clinical narratives, including discharge summaries and progress notes. The heterogeneous and verbose nature of these records makes manual review time-consuming and error-prone. We address this challenge by introducing a framework that leverages summarization as a preprocessing step to distill unstructured texts into concise, semantically rich representations. Central to this approach is *Context-Preserving Token Filtering (CPTF)*, which selectively retains clinically salient information. The resulting summaries are then aligned with a domain-specific *Knowledge Graph (KG)* that encodes structured relationships among clinical entities (e.g., diagnoses, medications, outcomes). In the context of a multiple sclerosis (MS) clinical trial, our framework automates the identification of relevant patient characteristics—such as relapse history or immunotherapy exposure—by first summarizing clinical notes and then validating these attributes through KG-based reasoning. This process minimizes manual effort, improves consistency, and accelerates recruitment. Summarization thus serves as a critical abstraction layer, transforming unstructured narratives into actionable representations. Combined with KG integration, the framework enables scalable and accurate patient screening, even in resource-limited settings, and illustrates the broader applicability of summarization-driven workflows in clinical research.