

CLARUS: Hierarchical Hypergraph Modeling for Unsupervised Discharge Summarization

Gaetano Cimino^{1*}, Mir Rayat Imtiaz Hossain², Laura De Santis¹
Amirhossein Abaskohi², Raymond T. Ng², Giuseppe Carenini², Vincenzo Deufemia¹

¹Department of Computer Science, University of Salerno, Italy

²Department of Computer Science, The University of British Columbia, Canada

{gcimino, ldesantis, deufemia}@unisa.it

{rayat137, aabaskoh, rng, carenini}@cs.ubc.ca

Abstract

Discharge summary generation from clinical records is challenging due to long, heterogeneous inputs and the need to preserve clinically relevant information. Although Large Language Models (LLMs) have improved abstractive summarization, their direct application to full patient records is limited by context length and difficulties in aggregating evidence across notes. We present CLARUS, a fully unsupervised framework that models patient records as hypergraphs encoding multi-dimensional relations. A hierarchical hypergraph encoder with depth-aware concept attention strengthens sparse representations, while a variational information bottleneck with Gumbel-Softmax relaxation enables differentiable selection of compact, informative subsets for LLM generation. Experiments on MIMIC-III show that CLARUS outperforms strong unsupervised baselines across lexical, semantic, and clinical metrics, demonstrating the value of higher-order structure and concept hierarchy for faithful multi-document clinical summarization.¹

1 Introduction

Discharge summaries are a critical component of electronic health records (EHRs), providing a concise account of a patient’s hospitalization, diagnoses, treatment course, and follow-up care (Kripalani et al., 2007). In practice, however, writing them is labor-intensive: clinicians must consolidate information scattered across heterogeneous notes, test reports, and temporally distant events, often under substantial time pressure (Li et al., 2013). These challenges motivate automated discharge summarization, especially in unsupervised settings where reference summaries are expensive to obtain.

Recent LLMs have shown strong performance on discharge summarization (Van Veen et al., 2024),

but the task remains challenging. Patient records are long, multi-document, and clinically heterogeneous, so relevant evidence is often distributed across many notes. Consequently, summarization is constrained not only by context length, but also by the need to synthesize information across documents while preserving clinically salient details (Adams et al., 2021). Important evidence may therefore be overlooked despite appearing in the input, a limitation related to the *lost-in-the-middle* phenomenon (Liu et al., 2024).

These limitations motivate *extract-then-summarize* pipelines, where a selection stage first identifies salient evidence for generation. In unsupervised settings, selection is commonly implemented through prompting or retrieval (Kruse et al., 2025; Naemi and Sahafi, 2026). Prompting-based methods remain sensitive to context limits and may omit clinically important information on long records (Asgari et al., 2025), while retrieval-based approaches typically rely on pairwise similarity between queries and text segments, potentially missing evidence without direct lexical or semantic overlap. In both cases, structural dependencies across notes may only be partially and implicitly modeled, despite clinically relevant evidence often being distributed across documents and linked through report types, medical concepts, and temporal context.

This view suggests two central challenges, illustrated in Figure 1. First, clinical evidence is inherently structured. Relevant information is distributed across notes, organized by report categories, and connected through shared concepts (Kim et al., 2023; Huang et al., 2025); these dependencies are not naturally represented by sequence-based or pairwise retrieval mechanisms. Second, highly specific clinical concepts may appear only a few times in a record (Choi et al., 2017; Song et al., 2019), making them harder to represent robustly and easier to miss during content selection.

*Work done while the author was a postdoctoral researcher at the University of British Columbia.

¹Our code is available at [CLARUS](#).

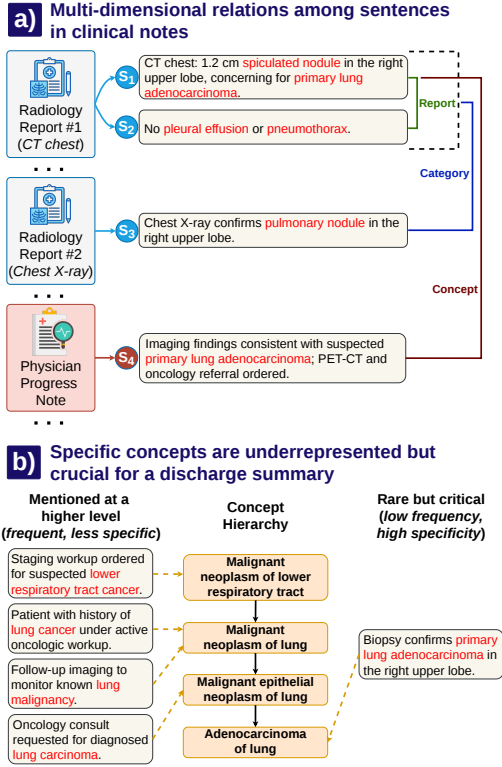


Figure 1: Two key challenges in clinical discharge summarization. **a)** Clinical evidence exhibits higher-order dependencies: sentences may belong to the same report, share a clinical category (e.g., *radiology*), or refer to the same medical concept across heterogeneous notes. **b)** Diagnostically critical concepts (e.g., *primary lung adenocarcinoma*) are often sparse, whereas their ontological ancestors (e.g., *lung carcinoma*, *lung cancer*) occur more frequently.

We address these challenges with CLARUS (CLinical Hypergraph-based Unsupervised Discharge Summarization), a fully unsupervised extract-then-summarize framework combining hierarchical hypergraph modeling and ontology-guided evidence selection for clinical summarization. Sentences are represented as nodes, while hyperedges encode note membership, clinical categories, and shared medical concepts. To better handle sparse clinical concepts, CLARUS leverages the hierarchical structure of expert-curated medical ontologies such as UMLS (Bodenreider, 2004). Rather than learning ontology embeddings, ontology depth modulates sentence aggregation, producing multi-granularity representations that reinforce specific concepts through frequently observed ancestors. These structures are encoded with a hierarchical hypergraph neural network (HGNN) (Feng et al., 2019), while a Variational Information Bottleneck

(VIB)-regularized (Alemi et al., 2019) Gumbel-Softmax (Jang et al., 2017) selector identifies compact, informative sentence subsets for downstream LLM generation. Because the framework is extractive, selected evidence can also be surfaced within the original notes, providing interpretable and source-grounded support for clinicians.

Experiments on MIMIC-III (Johnson et al., 2016) show that CLARUS consistently outperforms strong unsupervised baselines across lexical, semantic, and clinical metrics. Gains are particularly pronounced on semantic and clinical metrics, with improvements up to 3.1% on BERTScore and 9% on MEDCON over the best baseline. Gains further increase with input length, with BERTScore and MEDCON improvements rising from 2.2% and 8.5% on short records to 3.7% and 9.4% on long ones, indicating benefits of structural and ontology-aware modeling on heterogeneous long records. Human evaluation by board-certified clinicians provides additional support for these findings. CLARUS is preferred over the strongest baseline and receives higher ratings for comprehensiveness, correctness, and coherence. A targeted factuality audit shows substantially fewer clinically important omissions and no potentially harmful errors, providing additional evidence of the reliability of the proposed approach.

Our contributions are threefold: (i) we present an unsupervised framework for discharge summarization based on higher-order evidence selection over sentence-level hypergraphs; (ii) we introduce depth-aware concept attention, which uses ontology depth to modulate multi-granularity aggregation over sparse clinical concepts; and (iii) we show that this structured selection strategy consistently improves downstream summarization quality over strong unsupervised baselines.

2 Related Work

Unsupervised Discharge Summary Generation.

Recent unsupervised approaches mainly rely on prompt-based summarization with LLMs (Kruse et al., 2025; Naemi and Sahafi, 2026). However, performance degrades on long documents because of context truncation and the lost-in-the-middle effect, which may omit clinical information. To address this, extract-then-summarize pipelines have been proposed (Searle et al., 2023; Naemi and Sahafi, 2026). Existing extractors include zero-shot and chain-of-thought (CoT) prompting (Kruse

et al., 2025; Naemi and Sahafi, 2026), but they may still hallucinate or miss important details (Asgari et al., 2025). Alternatively, retrieval-augmented generation (RAG) methods (Lyu et al., 2024; Kruse et al., 2025; Myers et al., 2025) rely on pairwise similarity and may overlook clinically relevant evidence lacking lexical or semantic overlap (Gao et al., 2024; Edge et al., 2025). These limitations motivate structured extraction methods for modeling higher-order dependencies in clinical text.

Hypergraph Representation for Summarization. Hypergraph-based summarization methods model sentences as nodes and define hyperedges through latent topics (Xiong and Ji, 2016; Van Lierde and Chow, 2019) or keyword/topic-based strategies (Zhang et al., 2022). However, these approaches mainly rely on lexical or statistical signals, limiting their ability to capture semantically related clinical concepts expressed with different terminology. Other methods leverage document structure, such as section-level organization, to construct hyperedges (Onan and Alhumyani, 2024; Zhao et al., 2024; Huang et al., 2025), but overlook concept-level semantic information from the content itself, reducing their ability to model fine-grained domain relationships. In contrast, CLARUS combines structural cues with ontology-derived clinical concepts, linking sentences through shared entities to build semantically grounded hypergraphs.

Ontology-aware Representation Learning in Healthcare. Several healthcare models incorporate medical ontologies to address the sparsity of highly specific concepts. GRAM (Choi et al., 2017), KAME (Ma et al., 2018), and MMORE (Song et al., 2019) enrich representations through ontology ancestors, while HAP (Zhang et al., 2020a) and LINKO (Nayebi Kerdabadi et al., 2025) further improve hierarchical modeling through propagation and multi-ontology integration. In contrast, our framework directly leverages ontology structure to build sentence-level clinical hypergraphs, where ontology depth and shared concepts guide evidence aggregation for document-level summarization.

3 CLARUS

Figure 2 presents CLARUS, a hierarchical hypergraph-based framework for unsupervised discharge summarization. Clinical notes are represented as a sentence-level hypergraph encoding report, category, and concept relations. A hierarchical HGNN first encodes contextual sentence

representations, followed by a VIB-regularized Gumbel-Softmax selector that extracts a set of informative sentences for LLM generation.

3.1 Hypergraph Construction

Let $\mathcal{S} = \{s_1, \dots, s_N\}$ be the sentences obtained by segmenting all notes in a patient record. We build a hypergraph $\mathcal{G} = (\mathcal{S}, \mathcal{E})$, where nodes are sentences and hyperedges encode three relations: (i) report hyperedges $\mathcal{E}_{\text{rep}}$, linking sentences within the same note (Zhang et al., 2022); (ii) category hyperedges $\mathcal{E}_{\text{cat}}$, linking sentences from notes of the same clinical category (e.g., *radiology*) (Kim et al., 2023); and (iii) concept hyperedges $\mathcal{E}_{\text{con}}$, linking sentences mentioning the same clinical concept (e.g., *primary lung adenocarcinoma*). The hypergraph is represented by an incidence matrix $\mathbf{H} \in \{0, 1\}^{N \times |\mathcal{E}|}$ and a type map $\delta : \mathcal{E} \rightarrow \{\text{report}, \text{category}, \text{concept}\}$.

3.2 Hierarchical HGNN Encoder

Initial sentence embeddings. Each sentence s_i is encoded as $\mathbf{p}_i \in \mathbb{R}^{d_p}$ using a pre-trained text encoder. To model temporal dependencies among clinical events (Choi et al., 2016; Rajkomar et al., 2018), each sentence is associated with its note timestamp, normalized within each admission to capture relative order during the hospital stay. The normalized values are then mapped to a sinusoidal encoding $\mathbf{t}_i \in \mathbb{R}^{d_t}$, which is concatenated with the semantic embedding to obtain the initial node state $\mathbf{h}_i = [\mathbf{p}_i \parallel \mathbf{t}_i] \in \mathbb{R}^d$, with $d = d_p + d_t$.

HGNN layer. Message passing on hypergraphs follows a two-phase scheme: nodes aggregate into hyperedge representations, then hyperedges propagate information back to nodes. Resulting messages are integrated via a gated update mechanism.

Step 1: Hyperedge aggregation. Different hyperedge types capture distinct relations, and a shared transformation may conflate their contributions. Following prior work on heterogeneous graph modeling (Li et al., 2025), we use type-specific projections to compute hyperedge representations:

$$\mathbf{e}_j = \frac{1}{|e_j|} \sum_{i \in e_j} \mathbf{W}_{\delta(j)} \mathbf{h}_i \quad (1)$$

where $\mathbf{W}_{\delta(j)}$ depends on the hyperedge type, allowing each type to project sentence embeddings into a space tailored to its relational semantics.

Step 2: Hyperedge-to-sentence attention. report and category hyperedges are aggregated

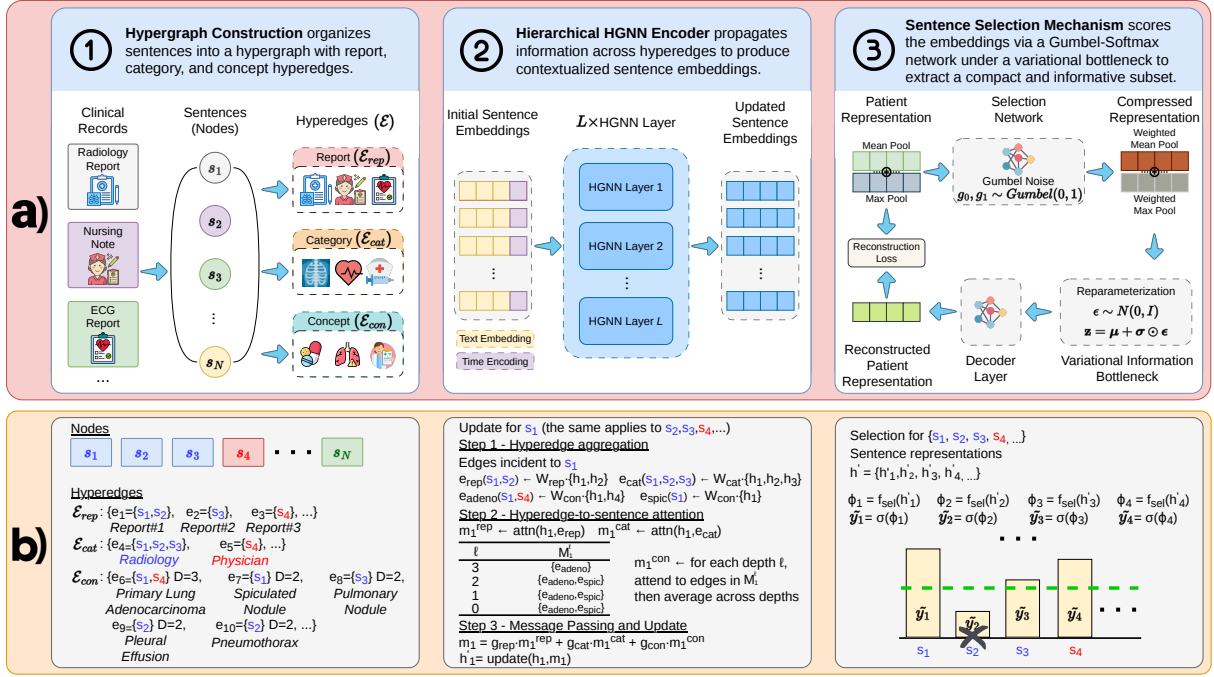


Figure 2: **a) Architecture overview.** CLARUS consists of: ① hypergraph construction, ② hierarchical HGNN encoding, and ③ VIB-based Gumbel-Softmax sentence selection. **b) Illustrative example.** Following the reports in Figure 1a, sentences $\{s_1, s_2, s_3, s_4, \dots\}$ are connected through report, category, and concept hyperedges, with concept edges annotated by ontology depth D . For sentence s_1 , the HGNN aggregates report, category, and depth-aware concept messages into the updated representation h'_1 ; the same process is applied to all sentences. The selector then assigns an importance score $\tilde{y}_i = \sigma(\phi_i)$ to each sentence and filters redundant ones (e.g., s_2).

using standard hyperedge-to-sentence attention (Ding et al., 2020), producing messages m^{rep} and m^{cat} . In contrast, concept hyperedges require specialized treatment because specific concepts are often sparse, whereas standard attention treats all concepts equally regardless of granularity. Following ontology-aware approaches that mitigate sparsity through ancestor information (Song et al., 2019), we use ontology depth to modulate aggregation over the concept hypergraph. Each concept hyperedge e_j is associated with ontology depth² D_j and projected into level-specific views $e_j^{(\ell)} = W_\ell^{\text{proj}} e_j$, where $\ell \in \{0, \dots, D_j\}$. Queries, keys, and values are computed as $Q_i^{(\ell)} = W_\ell^Q h_i$, $K_j^{(\ell)} = W_\ell^K e_j^{(\ell)}$, and $V_j^{(\ell)} = W_\ell^V e_j^{(\ell)}$, enabling multi-granularity representations for specific concepts.

For a sentence s_i , let $\mathcal{C}_i = \{e_j \in \mathcal{E}_{con} : \mathbf{H}_{ij} = 1\}$ denote its incident concept hyperedges. For attention level ℓ , we define $\mathcal{M}_i^{(\ell)} = \{e_j \in \mathcal{C}_i : \ell \leq D_j\}$ as the subset whose ontology depth is at least ℓ . Intuitively, larger ℓ correspond to finer-grained semantic levels, so only more specific concepts partic-

ipate at higher levels. Depth-constrained attention is computed by restricting the softmax to $\mathcal{M}_i^{(\ell)}$:

$$a_{ij}^{(\ell)} = \text{softmax}_{j \in \mathcal{M}_i^{(\ell)}} \left(Q_i^{(\ell)\top} K_j^{(\ell)} / \sqrt{d_k} \right). \quad (2)$$

The concept message is then

$$m_i^{con} = W^{\text{out}} \frac{1}{|\Lambda_i|} \sum_{\ell \in \Lambda_i} \sum_{j \in \mathcal{M}_i^{(\ell)}} a_{ij}^{(\ell)} V_j^{(\ell)}, \quad (3)$$

where W^{out} is a learnable output projection and $\Lambda_i = \{\ell : \mathcal{M}_i^{(\ell)} \neq \emptyset\}$ denotes the set of ontology levels for which sentence s_i has at least one valid concept hyperedge.

Step 3: Message Passing and Update. Different relations contribute unequally across sentences, making uniform aggregation of the three message types suboptimal. For example, findings repeatedly described within a note are better refined through report-level context, whereas clinically important concepts distributed across notes benefit more from concept-level aggregation. Prior graph neural network work has shown the effectiveness of adaptively weighting heterogeneous relations (Wang et al., 2019; Hu et al., 2020). Motivated by this, we

²Depth computation and depth-level analyses are provided in Appendices B and D, respectively.

introduce a learned gating mechanism that assigns sentence-specific weights to each message source:

$$\begin{aligned} \mathbf{u}_i &= [\mathbf{m}_i^{\text{rep}} \parallel \mathbf{m}_i^{\text{cat}} \parallel \mathbf{m}_i^{\text{con}}], \\ (g_{\text{rep}}, g_{\text{cat}}, g_{\text{con}}) &= \text{softmax}(f_{\text{gate}}(\mathbf{u}_i)), \quad (4) \\ \mathbf{m}_i &= g_{\text{rep}} \mathbf{m}_i^{\text{rep}} + g_{\text{cat}} \mathbf{m}_i^{\text{cat}} + g_{\text{con}} \mathbf{m}_i^{\text{con}}. \end{aligned}$$

The resulting message updates the sentence representation via residual connection and feed-forward network, each with layer normalization (Zhang et al., 2022). Stacking L such layers yields the final set of sentence embeddings $\mathbf{h}' = \{\mathbf{h}'_i\}_{i=1}^N$.

3.3 Sentence Selection Mechanism

Given the contextualized sentence embeddings produced by the hierarchical HGNN encoder, this stage selects a compact subset of clinically relevant sentences for downstream generation. In the fully unsupervised setting, no extractive labels or section-level pseudo-targets from reference summaries are available, as they would introduce supervision leakage. The selector therefore relies only on cues derived from the source notes. We formulate sentence selection as a compression problem characterized by three complementary properties: (i) *fidelity* to the global patient state, grounding selection in the full record rather than locally salient fragments; (ii) *distributed concept support*, encouraging preservation of evidence linked to clinical concepts across longitudinal notes; and (iii) *compactness*, reflecting the brevity of discharge summaries and the context limits of LLM-based generation. Architecturally, CLARUS adopts a modular extract-then-summarize design: the selector identifies a section-agnostic pool of clinically salient sentences, while section-specific synthesis is delegated to the downstream LLM via dedicated prompts.

Global patient representation. We summarize the full patient record using $\mathbf{r} = [\text{MeanPool}(\mathbf{h}') \parallel \text{MaxPool}(\mathbf{h}')] \in \mathbb{R}^{2d}$ (Chen et al., 2017; Kim et al., 2023), which serves as the reconstruction target. Mean pooling captures the overall clinical context by aggregating information across the entire patient record, while max pooling highlights the most salient latent features.

Differentiable sentence selection. To enable differentiable end-to-end training, a selection network produces a logit $\phi_i = f_{\text{sel}}(\mathbf{h}'_i)$ for each sentence, and a relaxed binary decision is sampled using Gumbel-Softmax (Jang et al., 2017):

$$y_i = \frac{\exp((\phi_i + g_1)/\xi)}{\exp(g_0/\xi) + \exp((\phi_i + g_1)/\xi)} \quad (5)$$

where $g_0, g_1 \sim \text{Gumbel}(0, 1)$ and ξ is an annealed temperature. This yields differentiable soft selection weights approaching discrete decisions as ξ decreases.

Variational bottleneck. From the soft selections, we construct a compressed record representation $\mathbf{c} = [\bar{\mathbf{h}} \parallel \tilde{\mathbf{h}}]$. Here, $\bar{\mathbf{h}}$ is a normalized weighted average of the transformed sentence representations, with the soft selection scores controlling the contribution of each element. The vector $\tilde{\mathbf{h}}$ is computed via feature-wise softmax pooling, which aggregates each feature dimension independently using weights determined by both the soft selection scores and the corresponding feature activations.

Mapping $\mathbf{c}$ directly to $\mathbf{r}$ would let the model retain sentence-level details irrelevant to a patient-level summary. We therefore apply an IB on $\mathbf{c}$, learning a latent $\mathbf{z}$ that keeps only what is needed to recover $\mathbf{r}$ and biases the selector toward sentences capturing the patient’s state across notes. Since the IB principle is intractable for high-dimensional continuous representations, we adopt its variational formulation (Alemi et al., 2019), optimized via the reparameterization trick (Kingma and Welling, 2022). To this end, we parameterize a diagonal Gaussian posterior by setting $\boldsymbol{\mu} = \mathbf{W}_\mu \mathbf{c}$ and $\log \boldsymbol{\sigma}^2 = \mathbf{W}_\sigma \mathbf{c}$, and sample $\mathbf{z} = \boldsymbol{\mu} + \boldsymbol{\sigma} \odot \boldsymbol{\epsilon}$ with $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. A linear decoder then reconstructs the record representation, $\hat{\mathbf{r}} = \mathbf{W}_{\text{dec}} \mathbf{z}$.

Training objective. Training is fully unsupervised and optimizes four complementary objectives.

Reconstruction Loss preserves the global patient information from the selected sentences:

$$\mathcal{L}_{\text{recon}} = \|\hat{\mathbf{r}} - \mathbf{r}\|^2 \quad (6)$$

KL Divergence regularizes the latent space through an IB, preventing the model from simply memorizing the input:

$$D_{KL} = D_{KL}(q(\mathbf{z}|\mathbf{c}) \parallel \mathcal{N}(\mathbf{0}, \mathbf{I})) \quad (7)$$

Concept-Incidence Loss encourages the preservation of concept mentions by rewarding the inverse-size-weighted fraction of selected sentences within each concept hyperedge:

$$\mathcal{L}_{\text{con}} = -\frac{1}{|\mathcal{E}_{\text{con}}|} \sum_{e_k \in \mathcal{E}_{\text{con}}} \frac{\sum_{i \in e_k} y_i}{|e_k|} \quad (8)$$

Compactness Loss encourages concise summaries:

$$\mathcal{L}_{\text{comp}} = \frac{\sum_i y_i}{N} \quad (9)$$

The final objective combines all terms:

$$\mathcal{L} = \mathcal{L}_{recon} + \beta D_{KL} + \lambda_{con} \mathcal{L}_{con} + \lambda_{comp} \mathcal{L}_{comp} \quad (10)$$

where β , λ_{con} , and λ_{comp} control the contribution of each component. $\mathcal{L}_{recon}$ biases the selector toward sentences that best capture the global patient state, as these enable accurate reconstruction, whereas the interaction between $\mathcal{L}_{con}$ and $\mathcal{L}_{comp}$ encourages the preservation of clinical evidence while preventing trivial selection of all sentences.

Inference. At test time, we compute $\tilde{y}_i = \sigma(\phi_i)$ using the selection network and keep sentences with probability greater than a threshold θ .

4 Experimental Setup

Datasets. Unlike prior work focusing on individual components of discharge summaries (Searle et al., 2023), our study considers the generation of all three sections: *Diagnosis*, *Brief Hospital Course*, and *Discharge Instructions*. For this setting, we use the MIMIC-III dataset³ (Johnson et al., 2016), following the experimental configuration of Kruse et al. (2025), which adopts the same three-section formulation. To further assess the generalizability of CLARUS, we conduct an additional evaluation on the MIMIC-IV dataset (Johnson et al., 2023), using the curated data and task formulation of Xu et al. (2024), which cover the generation of the *Brief Hospital Course* and *Discharge Instructions* sections. The results of this additional evaluation are reported in Appendix G.

Metrics. Following Xu et al. (2024), we report lexical quality using BLEU-4 (Papineni et al., 2002) and ROUGE-1/2/L (Lin, 2004), semantic similarity using METEOR (Banerjee and Lavie, 2005) and BERTScore (Zhang et al., 2020b), and clinical fidelity using AlignScore (Zha et al., 2023) for factual consistency and MEDCON (Yim et al., 2023) for concept-set agreement.

Backbone LLMs and baselines. For generation, we evaluate five open-source LLMs: Llama-2-13B (Touvron et al., 2023), Mistral-7B (Jiang et al., 2023), Llama-3-8B (Grattafiori et al., 2024), Qwen-2.5-7B (Yang et al., 2025), and DeepSeek-R1-Distill-Qwen-32B (Guo et al., 2025). These five models were also evaluated by Kruse et al. (2025), and we adopt the same model selection based on regulatory, operational, and empirical considerations. Specifically, the MIMIC Data Use Agreement prohibits the transmission of patient data to

closed-source APIs (Kruse et al., 2025). Moreover, clinical applications generally favor on-premise deployment to safeguard patient privacy (Wiest et al., 2024), particularly under constrained computational resources (Fime et al., 2025). We focus our analysis primarily on Llama-2-13B and Mistral-7B, as they provide the most favorable trade-off between computational resources and summarization performance. Results for the remaining models are reported in Appendix G. We compare CLARUS against four unsupervised baselines⁴: *Abs-ZS* (Kruse et al., 2025), which prompts the LLM on concatenated notes; *Ext-ZS* (Naemi and Sahafi, 2026), which performs zero-shot extraction before generation; *Ext-CoT* (Kruse et al., 2025), which adds CoT extraction; and *DisRAG* (Kruse et al., 2025), which retrieves note passages with section-specific queries before the final generation step.

5 Results and Analysis

Main results. Table 1 compares CLARUS with four unsupervised baselines using Llama-2 and Mistral as summary generators. Extract-then-summarize approaches outperform Abs-ZS on lexical metrics and often achieve competitive gains on semantic and clinical metrics, highlighting the benefits of an extraction stage before summarization. Notably, across both generators, CLARUS consistently attains the highest performance across lexical, semantic, and clinical evaluation metrics.

The strongest baseline varies across metrics and LLMs: DisRAG is generally most competitive with Llama-2, while no single baseline dominates with Mistral. Under Llama-2, CLARUS improves ROUGE-1/2 from 17.37/3.57 to 19.34/4.66, BERTScore from 71.98 to 74.22, and MEDCON from 19.24 to 20.98. Similar trends hold for Mistral, where CLARUS outperforms competing baselines across lexical, semantic, and clinical metrics. Consistent gains in BERTScore and MEDCON suggest improved preservation of clinically relevant information beyond lexical overlap.

Section-level analysis. Figure 3 compares CLARUS and DisRAG by discharge-summary section for Llama-2.⁵ Improvements vary across sections and metrics, suggesting that the proposed selection mechanism contributes differently depending on the target text. On *Diagnosis*, CLARUS

³Detailed dataset statistics are reported in Appendix C.

⁴Implementation details for both CLARUS and baselines are provided in Appendix A; the prompts are in Appendix E.

⁵The corresponding analysis for Mistral is provided in Appendix F.

Approach	Lexical				Semantic		Clinical	
	BLEU-4	ROUGE			BERT Score	METEOR	Align Score	MEDCON
		1	2	L				
Abs-ZS	0.56	15.14	3.13	14.36	67.66	8.59	39.21	18.20
Ext-ZS	1.34	15.94	2.99	15.25	69.01	12.18	39.89	17.30
Ext-CoT	1.44	15.28	2.62	14.64	67.50	12.28	40.12	17.44
DisRAG	1.21	17.37	3.57	16.64	71.98	11.32	41.08	19.24
CLARUS	1.76	19.34	4.66	18.51	74.22	12.88	41.53	20.98

Approach	Lexical				Semantic		Clinical	
	BLEU-4	ROUGE			BERT Score	METEOR	Align Score	MEDCON
		1	2	L				
Abs-ZS	0.66	14.86	3.78	14.19	66.40	8.89	29.92	17.84
Ext-ZS	0.77	16.31	3.22	15.63	67.14	9.47	41.28	16.08
Ext-CoT	0.78	15.66	2.98	14.98	69.22	10.02	38.23	17.75
DisRAG	0.64	16.83	3.47	16.13	65.97	8.58	39.55	17.92
CLARUS	1.39	19.04	5.13	18.26	70.01	10.62	41.54	18.95

Table 1: Comparison of CLARUS against unsupervised baselines using Llama-2-13B and Mistral-7B backbones. Scores are averaged over 5 random seeds; best results per backbone are in **bold**. Paired two-sided Wilcoxon signed-rank tests with Holm–Bonferroni correction show that CLARUS significantly outperforms competing methods ($p < 0.05$).

achieves the largest lexical- and concept-level gains, improving ROUGE-1 (+3.04), ROUGE-L (+2.77), and MEDCON (+1.77). Small drops in AlignScore (−0.60) likely reflect this section’s brevity, which limits the context available for entailment-based alignment. On *Brief Hospital Course*, improvements are positive across all metric families and most pronounced on BERTScore (+5.71), suggesting that the selected evidence supports generation of the narrative register typical of this section. On *Discharge Instructions*, lexical metrics remain largely flat, whereas semantic and clinical metrics improve substantially (BERTScore +4.37, AlignScore +1.14, MEDCON +0.67). This pattern reflects the templated nature of discharge instructions, where lexical overlap is often saturated while CLARUS produces semantically closer outputs, as captured by embedding-based metrics.

Robustness to input length. Figure 4 presents admissions grouped by the number of source notes for Llama-2.⁶ CLARUS achieves the strongest or near-strongest performance across nearly all bins and metric families. Competing methods show more heterogeneous trends. Prompting-based extractors (Ext-ZS, Ext-CoT) exhibit a clear decline as the number of source notes increases, particu-

⁶The corresponding analysis for Mistral is provided in Appendix F.

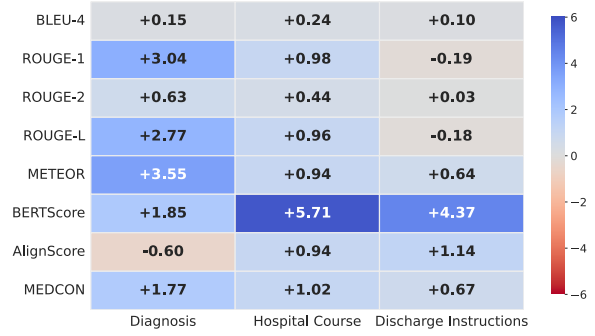


Figure 3: Section-level difference heatmap (CLARUS vs. DisRAG) for Llama-2-13B. Blue indicates gains for CLARUS, red favors DisRAG.

larly on lexical metrics. DisRAG follows a more nuanced pattern: while lexical scores deteriorate with increasing input length, semantic and clinical metrics show mixed or modest improvements, with both AlignScore and MEDCON exhibiting slight gains as the number of source notes increases. Importantly, CLARUS consistently leads on MEDCON across all bins and remains highly competitive on AlignScore, achieving the best performance in the two longer bins and comparable results in the shorter ones. This pattern aligns with the motivation of this work: prompting-based extractors are constrained by the LLM context window and the lost-in-the-middle phenomenon, both increasingly problematic as record length grows; DisRAG, although partially robust on concept-level matching, remains limited by pairwise query-passage similarity, which is less effective when relevant evidence is distributed across multiple notes and lacks strong lexical or semantic overlap with section-specific queries. Overall, these results suggest that structured aggregation through report-, category-, and concept-level relations may help integrate dispersed evidence in long clinical records.

Human evaluation. To complement automatic metrics, we conducted a human evaluation comparing CLARUS and DisRAG; full details of the protocol, clinician recruitment, and annotation procedure are provided in Appendix H. Table 2 summarizes results on 50 randomly sampled admissions evaluated by 10 clinicians, each reviewing 5 admissions (first value in each cell). Since each admission was rated by a single clinician in this initial study, inter-annotator agreement (IAA) could not be estimated. To address this limitation, we additionally evaluated a new subset of 20 admissions, each independently rated by 3 board-certified clin-

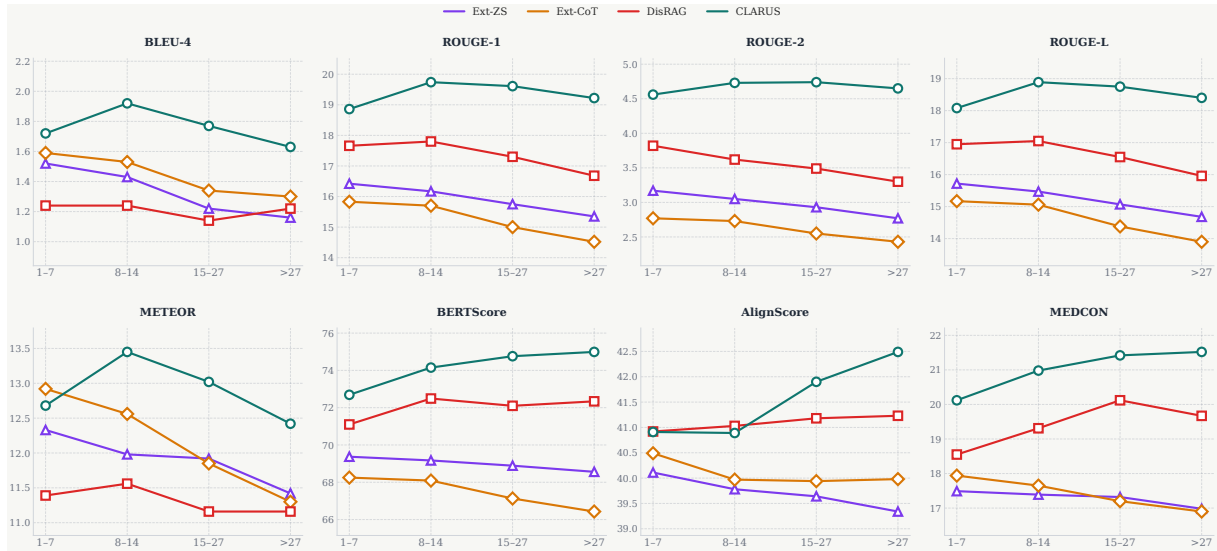


Figure 4: Performance of extract-then-summarize approaches with Llama-2-13B as a function of the number of clinical notes per admission.

Aspect	Criterion	CLARUS	DisRAG	IAA
Quality	Comprehensiveness	3.32*/3.23*	2.85/2.73	0.69
	Correctness	3.47*/3.57*	3.04/3.08	0.74
	Coherence	3.58*/3.67*	3.00/3.12	0.71
	Conciseness	3.11/3.32	3.09/3.21	0.56
Preference	Wins	33*/12	11/6	0.75
	Tie		6/2	–

Table 2: Human evaluation comparing CLARUS and DisRAG with Llama-2-13B. Each cell reports a/b , where a is from an initial study (50 admissions, 10 clinicians, 1 clinician per admission) and b is from a fully-crossed replication (20 admissions, 3 clinicians per admission), enabling estimation of inter-annotator agreement (IAA, Krippendorff’s α ; ordinal for quality dimensions, nominal for preference). Quality scores use a 1–5 Likert scale; best results per row are in **bold**. Statistical significance is assessed independently for each study using paired two-sided Wilcoxon signed-rank tests on Likert ratings (Holm–Bonferroni corrected) and exact two-sided sign tests on pairwise preferences; * denotes $p < 0.05$ in favor of CLARUS, reported separately for each value (a^*/b^*).

icians under the same protocol (second value in each cell), enabling us to estimate IAA via Krippendorff’s α , reported in the last column of Table 2.

CLARUS receives higher ratings than DisRAG on all four Likert-based dimensions and achieves more pairwise wins in both studies (33 vs. 11, with 6 ties, in the initial study; 12 vs. 6, with 2 ties, in the additional evaluation). The gains are significant ($p < 0.05$) in both studies on Comprehensiveness, Correctness, and Coherence, indicating that the improvements are also reflected in clinician

Error Type	CLARUS	DisRAG	IAA
Unsupported statement	35%	45%	0.68
Contradiction	0%	5%	0.59
Clinically important omission	20%	60%	0.66
Temporal error	15%	15%	0.62
Medication/discharge error	25%	45%	0.67
Potentially harmful error	0%	10%	0.64

Table 3: Clinical factuality audit comparing CLARUS and DisRAG with Llama-2-13B on the 20-admission fully-crossed replication sample. Values report the percentage of summaries flagged for each error type (majority vote across 3 clinicians per admission; lower is better); best results per row are in **bold**. IAA is Krippendorff’s α (nominal, computed on the binary per-clinician annotations before aggregation).

judgments. The additional evaluation yields Krippendorff’s α values ranging from 0.56 to 0.75, indicating moderate-to-substantial agreement among clinicians. On Conciseness, the two systems obtain nearly identical scores (3.11 vs. 3.09 in the initial study, 3.32 vs. 3.21 in the additional evaluation), with no significant difference in either study. This is expected, as both approaches rely on the same backbone generator, meaning summary length is driven mainly by the generation stage rather than the selection mechanism. Importantly, CLARUS improves the other evaluation dimensions without increasing summary length, a desirable property in clinical settings where concise discharge summaries are essential.

Clinical factuality audit. For the additional fully-crossed evaluation, we further assessed clinical fac-

Variant	LEXICAL				SEMANTIC		CLINICAL	
	BLEU-4	ROUGE-1	ROUGE-2	ROUGE-L	BERTScore	METEOR	AlignScore	MEDCON
CLARUS	1.76	19.34	4.66	18.51	74.22	12.88	41.53	20.98
<i>Hypergraph</i>								
w/o report hyperedges	1.61 (-0.15)	18.94 (-0.40)	4.50 (-0.16)	18.10 (-0.41)	73.45 (-0.77)	12.60 (-0.28)	41.34 (-0.19)	20.59 (-0.39)
w/o category hyperedges	1.59 (-0.17)	18.87 (-0.47)	4.47 (-0.19)	18.05 (-0.46)	73.48 (-0.74)	12.54 (-0.34)	41.48 (-0.05)	20.50 (-0.48)
w/o concept hyperedges	1.62 (-0.14)	18.60 (-0.74)	4.37 (-0.29)	17.76 (-0.75)	72.67 (-1.55)	12.56 (-0.32)	40.98 (-0.55)	20.17 (-0.81)
topic/keyword	1.43 (-0.33)	17.76 (-1.58)	3.92 (-0.74)	16.98 (-1.53)	71.95 (-2.27)	11.98 (-0.90)	40.76 (-0.77)	19.31 (-1.67)
graph-based	1.58 (-0.18)	18.21 (-1.13)	4.11 (-0.55)	17.44 (-1.07)	72.84 (-1.38)	12.31 (-0.57)	41.01 (-0.52)	19.88 (-1.10)
<i>Encoder</i>								
no type-proj	1.67 (-0.09)	19.10 (-0.24)	4.55 (-0.11)	18.25 (-0.26)	73.71 (-0.51)	12.71 (-0.17)	41.46 (-0.07)	20.72 (-0.26)
no gating	1.66 (-0.10)	19.01 (-0.33)	4.50 (-0.16)	18.18 (-0.33)	73.55 (-0.67)	12.68 (-0.20)	41.33 (-0.20)	20.58 (-0.40)
no depth	1.62 (-0.14)	18.44 (-0.90)	4.32 (-0.34)	17.59 (-0.92)	72.71 (-1.51)	12.52 (-0.36)	40.94 (-0.59)	19.96 (-1.02)
no time	1.71 (-0.05)	19.18 (-0.16)	4.59 (-0.07)	18.36 (-0.15)	73.95 (-0.27)	12.78 (-0.10)	41.42 (-0.11)	20.90 (-0.08)
<i>Selection</i>								
IB only (no variational)	1.70 (-0.06)	19.04 (-0.30)	4.53 (-0.13)	18.20 (-0.31)	73.68 (-0.54)	12.79 (-0.09)	41.42 (-0.11)	20.63 (-0.35)
w/o $\mathcal{L}_{con}$	1.49 (-0.27)	18.24 (-1.10)	4.10 (-0.56)	17.46 (-1.05)	72.93 (-1.29)	12.28 (-0.60)	41.13 (-0.40)	19.20 (-1.78)

Table 4: Ablation study of CLARUS with Llama-2-13B. Values report absolute scores and performance degradation (Δ) vs. the full model; cell shading indicates degradation magnitude (lighter = smaller, darker = larger).

tuality through a targeted audit covering six error categories⁷: unsupported statements, contradictions, clinically important omissions, temporal errors, medication/discharge errors, and potentially harmful errors. Clinicians were provided with the corresponding source notes and recorded binary judgments for each category. Table 3 reports the proportion of audited outputs exhibiting each error type, together with the corresponding Krippendorff’s α . CLARUS exhibits lower error rates than DisRAG across five of the six categories. Clinically important omissions decrease from 60% to 20%, medication/discharge errors from 45% to 25%, and potentially harmful errors from 10% to 0%. Contradictions are observed in none of the CLARUS summaries compared with 5% for DisRAG, while unsupported statements decrease from 45% to 35%. Temporal errors are the only category with identical rates for the two systems, at 15%. Across the six audited categories, inter-annotator agreement ranges from 0.59 to 0.68, indicating moderate-to-substantial agreement among clinicians.

Ablation study. Table 4 presents an ablation study assessing the contribution of each CLARUS component. All encoder modules positively affect performance: removing depth-aware concept attention causes the largest decline, while removing time encoding has only a minor effect, suggesting that hierarchical and heterogeneous structural information is more important than temporal cues.

The results also highlight the importance of concept-level modeling. Removing concept hyper-

edges reduces performance, suggesting that concept relations may provide complementary information beyond higher-level structure. However, concept hyperedges alone appear less effective without depth-aware modeling, indicating that hierarchical ontology information may help model clinical concepts. Replacing the proposed hypergraph with topic/keyword-based (Zhang et al., 2022) or graph-based (Wang et al., 2020) alternatives further degrades results, suggesting that hyperedge structure and ontology-derived concept relations jointly contribute to capturing cross-note dependencies.

Regarding the training objectives, replacing the variational bottleneck with a deterministic alternative slightly reduces performance, suggesting variational regularization improves generalization. Removing the concept-incidence loss causes a larger drop, especially on MEDCON, indicating its role in preserving clinical concepts in summaries.

6 Conclusion

We introduced CLARUS, an unsupervised extract-then-summarize framework that models patient records as hierarchical sentence-level hypergraphs enriched with report, category, and concept relations. Experiments with open-source LLMs show consistent improvements over strong unsupervised baselines across lexical, semantic, and clinical metrics, especially on longer and more heterogeneous records. The results, further supported by a human evaluation, provide evidence that modeling higher-order structure and ontology-based concept hierarchies can be beneficial for multi-document clinical summarization.

⁷Full protocol and category definitions are provided in Appendix H.

Limitations

Although CLARUS consistently improves over strong unsupervised baselines, some limitations should be acknowledged.

The concept hyperedges in CLARUS are constructed from clinical concepts extracted using a pre-trained *MedCAT* model (Kraljevic et al., 2021). Consequently, their quality is bounded by the underlying named entity recognition (NER) component. At the same time, MedCAT has been extensively validated on clinical corpora, with reported F_1 scores in the 0.80–0.95 range across concept recognition tasks (Kraljevic et al., 2021; Booker et al., 2024), suggesting that the resulting hyperedges are sufficiently reliable for practical use.

Our experiments are conducted on MIMIC-III and MIMIC-IV, both derived from the same U.S. tertiary academic hospital and comprising adult patient populations. Accordingly, the generalizability of our findings to outpatient settings, pediatric populations, and healthcare systems operating in non-English contexts remains to be empirically established. Nevertheless, CLARUS is not inherently dataset-specific, as it requires only a clinical NER system and a hierarchical ontology, both widely available across clinical domains.

The evaluation is limited to open-source LLMs, with no experiments using proprietary models. This follows from the MIMIC Data Use Agreement, which prohibits the transmission of patient data to closed-source APIs (Kruse et al., 2025), as well as from privacy constraints that favor on-premise deployment in clinical settings (Wiest et al., 2024).

Ethical Considerations

All experiments were conducted on MIMIC-III and MIMIC-IV, fully de-identified critical care databases made available for research use under applicable data use agreements. Access to the datasets was obtained through the standard PhysioNet credentialing process, including completion of the required human-subjects research training. The human evaluation described in Appendix H was conducted by 10 board-certified clinicians with documented experience in writing and reviewing discharge summaries. Participation was entirely voluntary and unpaid, and all clinicians provided informed consent for the use of their evaluations for research purposes prior to the start of the study. Reviewers were free to withdraw from the evaluation at any time without consequence. To minimize po-

tential bias, clinicians were blinded to the identity of the systems being compared, and the two candidate summaries were anonymized and presented in randomized order for each patient admission.

CLARUS is intended as a research prototype for studying unsupervised discharge summarization and should not be deployed in clinical practice without further validation, including prospective evaluation and integration with appropriate human oversight. As with any LLM-based system, generated summaries may contain factual errors, omissions, or hallucinations, and must not be used as a substitute for clinical judgment.

Acknowledgments

This work was supported by funding from the NSERC Alliance program and from the Provincial Health Services Authority. The authors thank the anonymous reviewers for their insightful comments and suggestions. We are also grateful to the clinicians who took part in the human evaluation, and in particular to Carmine Fabbriatore (Azienda Ospedaliera Universitaria Luigi Vanvitelli), Valeria Lonardo (Azienda Ospedaliera Antonio Cardarelli), and Paola Esposito (Azienda Ospedaliera Universitaria San Giovanni di Dio e Ruggi d’Aragona), for their time and clinical expertise. The remaining clinicians preferred not to be named and are gratefully acknowledged anonymously.

References

- Asad Aali, Dave Van Veen, Yamin Ishraq Arefeen, Jason Hom, Christian Bluethgen, Eduardo Pontes Reis, Sergios Gatidis, Namuun Clifford, Joseph Daws, Arash S Tehrani, Jangwon Kim, and Akshay S Chaudhari. 2025. [A dataset and benchmark for hospital course summarization with adapted large language models](#). *Journal of the American Medical Informatics Association*, 32(3):470–479.
- Griffin Adams, Emily Alsentzer, Mert Ketenci, Jason Zucker, and Noémie Elhadad. 2021. [What’s in a summary? laying the groundwork for advances in hospital-course summarization](#). In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4794–4811. Association for Computational Linguistics.
- Alexander A. Alemi, Ian Fischer, Joshua V. Dillon, and Kevin Murphy. 2019. [Deep variational information bottleneck](#). *Preprint*, arXiv:1612.00410.
- Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jindi, Tristan Naumann, and

- Matthew McDermott. 2019. [Publicly available clinical BERT embeddings](#). In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 72–78. Association for Computational Linguistics.
- Elham Asgari, Nina Montaña Brown, Magda Dubois, Saleh Khalil, Jasmine Balloch, Joshua Au Yeung, and Dominic Pimenta. 2025. [A framework to assess clinical safety and hallucination rates of LLMs for medical text summarisation](#). *NPJ Digital Medicine*, 8(1).
- Satanjeev Banerjee and Alon Lavie. 2005. [METEOR: An automatic metric for MT evaluation with improved correlation with human judgments](#). In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72. Association for Computational Linguistics.
- Olivier Bodenreider. 2004. [The unified medical language system \(UMLS\): integrating biomedical terminology](#). *Nucleic Acids Res.*, 32(suppl_1):D267–D270.
- James Booker, Jack Penn, Kawsar Noor, Richard JB Dobson, Jonathan P Funnell, Chan Hee Koh, Danyal Z Khan, Nicola Newall, David Rowland, Sidharth Sinha, Simon C. Williams, Parag Sayal, and Hani J. Marcus. 2024. [Early evaluation of a natural language processing tool to improve access to educational resources for surgical patients](#). *European Spine Journal*, 33(7):2545–2552.
- Harrison Chase. 2022. [LangChain](#). Accessed: 2026-05-23.
- Qian Chen, Xiaodan Zhu, Zhen-Hua Ling, Si Wei, Hui Jiang, and Diana Inkpen. 2017. [Enhanced LSTM for natural language inference](#). In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1657–1668. Association for Computational Linguistics.
- Edward Choi, Mohammad Taha Bahadori, Joshua A. Kulas, Andy Schuetz, Walter F. Stewart, and Jimeng Sun. 2016. [RETAIN: an interpretable predictive model for healthcare using reverse time attention mechanism](#). In *Proceedings of the 30th International Conference on Neural Information Processing Systems (NIPS 2016)*, page 3512–3520. Curran Associates Inc.
- Edward Choi, Mohammad Taha Bahadori, Le Song, Walter F. Stewart, and Jimeng Sun. 2017. [GRAM: Graph-based attention model for healthcare representation learning](#). In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD 2017)*, page 787–795. Association for Computing Machinery.
- Kaize Ding, Jianling Wang, Jundong Li, Dingcheng Li, and Huan Liu. 2020. [Be more with less: Hypergraph attention networks for inductive text classification](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP 2020)*, pages 4927–4936. Association for Computational Linguistics.
- Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvasy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. 2026. [The Faiss library](#). *IEEE Transactions on Big Data*, 12(2):346–361.
- Darren Edge, Ha Trinh, Newman Cheng, Joshua Bradley, Alex Chao, Apurva Mody, Steven Truitt, Dasha Metropolitan, Robert Osazuwa Ness, and Jonathan Larson. 2025. [From local to global: A graph RAG approach to query-focused summarization](#). *Preprint*, arXiv:2404.16130.
- Simon Ellershaw, Christopher Tomlinson, Oliver E Burton, Thomas Frost, John Gerrard Hanrahan, Danyal Zaman Khan, Hugo Layard Horsfall, Molie Little, Evaleen Malgapo, Joachim Starup-Hansen, Jack Ross, Martinique Vella-Baldacchino, Kawsar Noor, Anoop D. Shah, and Richard Dobson. 2024. [Automated generation of hospital discharge summaries using clinical guidelines and large language models](#). In *Proceedings of the AAAI 2024 Spring Symposium on Clinical Foundation Models*.
- Yifan Feng, Haoxuan You, Zizhao Zhang, Rongrong Ji, and Yue Gao. 2019. [Hypergraph neural networks](#). In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019*, volume 33, pages 3558–3565. AAAI Press.
- Awal Ahmed Fime, Md. Zarif Hossain, Saika Zaman, Abdur Rahman Bin Shahid, and Ahmed Imteaj. 2025. [Benchmarking large-language models for resource-efficient medical AI for edge deployment](#). In *Proceedings of the 2025 AAAI Spring Symposium Series*, volume 5, pages 137–142. AAAI Press.
- Markus Frohmann, Igor Sterner, Ivan Vulić, Benjamin Minixhofer, and Markus Schedl. 2024. [Segment any text: A universal approach for robust, efficient and adaptable sentence segmentation](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP 2024)*, pages 11908–11941. Association for Computational Linguistics.
- Matthias Ganzinger, Nicola Kunz, Pascal Fuchs, Cornelia K Lyu, Martin Loos, Martin Dugas, and Thomas M Pausch. 2025. [Automated generation of discharge summaries: Leveraging large language models with clinical data](#). *Scientific reports*, 15(1):16466.
- Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Meng Wang, and Haofen Wang. 2024. [Retrieval-augmented generation for large language models: A survey](#). *Preprint*, arXiv:2312.10997.

- Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. [The llama 3 herd of models](#). *Preprint*, arXiv:2407.21783.
- Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, and 175 others. 2025. [Deepseek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning](#). *Nature*, 645(8081):633–638.
- Amanda Hall and Graham Walton. 2004. [Information overload within the health care system: A literature review](#). *Health Information & Libraries Journal*, 21(2):102–108.
- Ziniu Hu, Yuxiao Dong, Kuansan Wang, and Yizhou Sun. 2020. [Heterogeneous graph transformer](#). In *Proceedings of The Web Conference 2020 (WWW 2020)*, page 2704–2710. Association for Computing Machinery.
- Zhenyu Huang, Xianlai Chen, Taixiang Li, Xing Zhao, Ying An, and Yunbo Wang. 2025. [Higraphsum: Hierarchical hypergraph-based dual-stage automatic summarization of discharge summaries](#). In *2025 IEEE International Conference on Bioinformatics and Biomedicine (BIBM 2025)*, pages 973–980.
- Eric Jang, Shixiang Gu, and Ben Poole. 2017. [Categorical reparameterization with Gumbel-Softmax](#). *Preprint*, arXiv:1611.01144.
- Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, L  lio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timoth  e Lacroix, and William El Sayed. 2023. [Mistral 7b](#). *Preprint*, arXiv:2310.06825.
- Alistair EW Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shammout, Steven Horng, Tom J Pollard, Sicheng Hao, Benjamin Moody, Brian Gow, Liwei H. Lehman, Leo A. Celi, and Roger G. Mark. 2023. [MIMIC-IV, a freely accessible electronic health record dataset](#). *Scientific data*, 10(1):1.
- Alistair EW Johnson, Tom J Pollard, Lu Shen, Liwei H. Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. 2016. [MIMIC-III, a freely accessible critical care database](#). *Scientific data*, 3(1):1–9.
- Nayeon Kim, Yinhua Piao, and Sun Kim. 2023. [Clinical note owns its hierarchy: Multi-level hypergraph neural networks for patient-level representation learning](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5559–5573. Association for Computational Linguistics.
- Diederik P Kingma and Max Welling. 2022. [Auto-encoding variational bayes](#). *Preprint*, arXiv:1312.6114.
- Zeljko Kraljevic, Thomas Searle, Anthony Shek, Lukasz Roguski, Kawsar Noor, Daniel Bean, Aurelie Mascio, Leilei Zhu, Amos A. Folarin, Angus Roberts, Rebecca Bendayan, Mark P. Richardson, Robert Stewart, Anoop D. Shah, Wai Keong Wong, Zina Ibrahim, James T. Teo, and Richard J.B. Dobson. 2021. [Multi-domain clinical natural language processing with MedCAT: The medical concept annotation toolkit](#). *Artificial Intelligence in Medicine*, 117:102083.
- Sunil Kripalani, Frank LeFevre, Christopher O. Phillips, Mark V. Williams, Preetha Basaviah, and David W. Baker. 2007. [Deficits in communication and information transfer between hospital-based and primary care physicians: Implications for patient safety and continuity of care](#). *JAMA*, 297(8):831–841.
- Maya Kruse, Shiyue Hu, Nicholas Derby, Yifu Wu, Samantha Stonbraker, Bingsheng Yao, Dakuo Wang, Elizabeth M. Goldberg, and Yanjun Gao. 2025. [Large language models with temporal reasoning for longitudinal clinical summarization and prediction](#). In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 20715–20735. Association for Computational Linguistics.
- Jordan Y. Z. Li, Tuck Y. Yong, Paul Hakendorf, David Ben-Tovim, and Campbell H. Thompson. 2013. [Timeliness in discharge summary dissemination is associated with patients’ clinical outcomes](#). *Journal of Evaluation in Clinical Practice*, 19(1):76–79.
- Yongkang Li, Zipei Fan, and Xuan Song. 2025. [Heterogeneous hyperbolic hypergraph neural network for friend recommendation in location-based social networks](#). *ACM Transactions on Knowledge Discovery from Data*, 19(3):1–29.
- Chin-Yew Lin. 2004. [ROUGE: A package for automatic evaluation of summaries](#). In *Text Summarization Branches Out*, pages 74–81. Association for Computational Linguistics.
- Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2024. [Lost in the middle: How language models use long contexts](#). *Transactions of the Association for Computational Linguistics*, 12:157–173.
- Mengxian Lyu, Cheng Peng, Daniel Paredes, Ziyi Chen, Aokun Chen, Jiang Bian, and Yonghui Wu. 2024. [UF-HOBI at “discharge me!”: A hybrid solution for discharge summary generation through prompt-based tuning of GatorTronGPT models](#). In *Proceedings of the 23rd Workshop on Biomedical Natural Language Processing*, pages 685–695. Association for Computational Linguistics.

- Fenglong Ma, Quanzeng You, Houping Xiao, Radha Chitta, Jing Zhou, and Jing Gao. 2018. [KAME: Knowledge-based attention model for diagnosis prediction in healthcare](#). In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management (CIKM 2018)*, page 743–752. Association for Computing Machinery.
- Skatje Myers, Timothy A Miller, Yanjun Gao, Matthew M Churpek, Anoop Mayampurath, Dmitriy Dligach, and Majid Afshar. 2025. [Lessons learned on information retrieval in electronic health records: a comparison of embedding models and pooling strategies](#). *Journal of the American Medical Informatics Association*, 32(2):357–364.
- Amin Naemi and Ali Sahafi. 2026. [Benchmarking large language models for MIMIC-IV clinical note summarization](#). *Journal of Healthcare Informatics Research*, 10(1):95–115.
- Mohsen Nayeibi Kerdabadi, Arya Hadizadeh Moghadam, Dongjie Wang, and Zijun Yao. 2025. [Multi-ontology integration with dual-axis propagation for medical concept representation](#). In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management (CIKM 2025)*, page 2201–2211. Association for Computing Machinery.
- Aytuğ Onan and Hesham Alhumyani. 2024. [Contextual hypergraph networks for enhanced extractive summarization: Introducing multi-element contextual hypergraph extractive summarizer \(MCHES\)](#). *Applied Sciences*, 14(11).
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. [BLEU: a method for automatic evaluation of machine translation](#). In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL 2002)*, pages 311–318. Association for Computational Linguistics.
- Alvin Rajkomar, Eyal Oren, Kai Chen, Andrew M. Dai, Nissan Hajaj, Michaela Hardt, Peter J. Liu, Xiaobing Liu, Jake Marcus, Mimi Sun, Patrik Sundberg, Hector Yee, Kun Zhang, Yi Zhang, Gerardo Flores, Gavin E. Duggan, Jamie Irvine, Quoc Le, Kurt Litsch, and 16 others. 2018. [Scalable and accurate deep learning with electronic health records](#). *NPJ Digital Medicine*, 1(1).
- Thomas Searle, Zina Ibrahim, James Teo, and Richard J.B. Dobson. 2023. [Discharge summary hospital course summarisation of in patient electronic health record text with clinical concept guided deep pre-trained transformer models](#). *Journal of Biomedical Informatics*, 141:104358.
- Lihong Song, Chin Wang Cheong, Kejing Yin, William K. Cheung, Benjamin C. M. Fung, and Jonathan Poon. 2019. [Medical concept embedding with multiple ontological representations](#). In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence (IJCAI 2019)*, pages 4613–4619. International Joint Conferences on Artificial Intelligence Organization.
- Michael Q. Stearns, Colin Price, Kent A. Spackman, and Amy Y. Wang. 2001. [SNOMED clinical terms: overview of the development process and project status](#). In *Proceedings of American Medical Informatics Association Annual Symposium (AMIA 2001)*, pages 662–666. AMIA.
- Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, and 49 others. 2023. [Llama 2: Open foundation and fine-tuned chat models](#). *Preprint*, arXiv:2307.09288.
- Hadrien Van Lierde and Tommy W.S. Chow. 2019. [Query-oriented text summarization based on hypergraph transversals](#). *Information Processing & Management*, 56(4):1317–1338.
- Dave Van Veen, Cara Van Uden, Louis Blanke-meier, Jean-Benoit Delbrouck, Asad Aali, Christian Bluethgen, Anuj Pareek, Malgorzata Polacin, Eduardo Pontes Reis, Anna Seehofnerová, Nidhi Rohatgi, Poonam Hosamani, William Collins, Neera Ahuja, Curtis P. Langlotz, Jason Hom, Sergios Gatidis, John Pauly, and Akshay S. Chaudhari. 2024. [Adapted large language models can outperform medical experts in clinical text summarization](#). *Nature medicine*, 30(4):1134–1142.
- Danqing Wang, Pengfei Liu, Yining Zheng, Xipeng Qiu, and Xuanjing Huang. 2020. [Heterogeneous graph neural networks for extractive document summarization](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL 2020)*, pages 6209–6219. Association for Computational Linguistics.
- Xiao Wang, Houye Ji, Chuan Shi, Bai Wang, Yanfang Ye, Peng Cui, and Philip S Yu. 2019. [Heterogeneous graph attention network](#). In *The World Wide Web Conference (WWW 2019)*, page 2022–2032. Association for Computing Machinery.
- Isabella Catharina Wiest, Dyke Ferber, Jiefu Zhu, Marko van Treeck, Sonja K. Meyer, Radhika Juglan, Zunamys I. Carrero, Daniel Paech, Jens Kleesiek, Matthias P. Ebert, Daniel Truhn, and Jakob Nikolas Kather. 2024. [Privacy-preserving large language models for structured medical information retrieval](#). *NPJ Digital Medicine*, 7(1).
- Christopher Y. K. Williams, Charumathi Raghu Subramanian, Syed Salman Ali, Michael Apolinario, Elisabeth Askin, Peter Barish, Monica Cheng, W. James Deardorff, Nisha Donthi, Smitha Ganeshan, Owen Huang, Molly A. Kantor, Andrew R. Lai, Ashley Manchanda, Kendra A. Moore, Anoop N. Muniyappa, Geethu Nair, Prashant P. Patel, Lekshmi Santhosh, and 5 others. 2025. [Physician- and large language model-generated hospital discharge summaries](#). *JAMA Internal Medicine*, 185(7):818–825.

- Shitao Xiao, Zheng Liu, Peitian Zhang, Niklas Muenighoff, Defu Lian, and Jian-Yun Nie. 2024. [C-Pack: Packed resources for general chinese embeddings](#). In *Proceedings of the 47th ACM International Conference on Research and Development in Information Retrieval (SIGIR 2024)*, page 641–649. Association for Computing Machinery.
- Shufeng Xiong and Donghong Ji. 2016. [Query-focused multi-document summarization using hypergraph-based ranking](#). *Information Processing & Management*, 52(4):670–681.
- Justin Xu, Zhihong Chen, Andrew Johnston, Louis Blankemeier, Maya Varma, Jason Hom, William J. Collins, Ankit Modi, Robert Lloyd, Benjamin Hopkins, Curtis Langlotz, and Jean-Benoit Delbrouck. 2024. [Overview of the first shared task on clinical text generation: RRG24 and “discharge me!”](#). In *Proceedings of the 23rd Workshop on Biomedical Natural Language Processing*, pages 85–98. Association for Computational Linguistics.
- An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, and 23 others. 2025. [Qwen2.5 technical report](#). *Preprint*, arXiv:2412.15115.
- Wen-wai Yim, Yujuan Fu, Asma Ben Abacha, Neal Snider, Thomas Lin, and Meliha Yetisgen. 2023. [Aci-bench: a novel ambient clinical intelligence dataset for benchmarking automatic visit note generation](#). *Scientific data*, 10(1):586.
- Yuheng Zha, Yichi Yang, Ruichen Li, and Zhiting Hu. 2023. [AlignScore: Evaluating factual consistency with a unified alignment function](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11328–11348. Association for Computational Linguistics.
- Haopeng Zhang, Xiao Liu, and Jiawei Zhang. 2022. [HEGEL: Hypergraph transformer for long document summarization](#). In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP 2022)*, pages 10167–10176. Association for Computational Linguistics.
- Muhan Zhang, Christopher R. King, Michael Avidan, and Yixin Chen. 2020a. [Hierarchical attention propagation for healthcare representation learning](#). In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD 2020)*, page 249–256. Association for Computing Machinery.
- Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020b. [BERTScore: Evaluating text generation with BERT](#). *Preprint*, arXiv:1904.09675.
- Chenlong Zhao, Xiwen Zhou, Xiaopeng Xie, and Yong Zhang. 2024. [Hierarchical attention graph for scientific document summarization in global and local level](#). In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 714–726. Association for Computational Linguistics.

A Implementation Details

Following Searle et al. (2023), we perform NER to extract *SNOMED CT* (Stearns et al., 2001) concepts using a pre-trained *MedCAT* model (Kraljevic et al., 2021), which was trained without supervision on MIMIC-III for SNOMED CT-related tasks. The extracted concepts are employed to construct the concept hyperedges of the clinical hypergraph. Clinical notes are first segmented into sentences using the *SaT* model (Frohmann et al., 2024), after which each sentence is embedded using *BioClinicalBERT* (Alsentzer et al., 2019), yielding representations of dimension 768. To incorporate temporal information, the timestamp associated with each note is min-max normalized within the corresponding admission and encoded through a 64-dimensional sinusoidal positional representation, following standard practice.

The hierarchical HGNN encoder is composed of 2 layers with a hidden dimension of 256. The maximum number of depth levels considered by the depth-aware concept attention mechanism is set to $D_{\max} = 3$. This configuration is motivated by the concept-depth analysis reported in Appendix C, which indicates that most extracted concepts are concentrated within the first four ontology levels. Additional support is provided by the ablation study in Appendix D, showing that increasing $D_{\max}$ beyond 3 introduces additional architectural complexity without producing measurable performance improvements. Consequently, concepts whose ontology depth exceeds $D_{\max}$ are mapped to level 3.

Selection logits are computed using a lightweight two-layer MLP applied to the contextualized sentence embeddings. To enable differentiable optimization, the discrete selection mechanism is relaxed through the Gumbel-Softmax technique (Jang et al., 2017). The temperature parameter is progressively annealed during training according to $\xi(t) = \max(\xi_{\min}, \xi_{\max} \cdot e^{-\gamma t})$, where $\xi_{\max} = 2$, $\xi_{\min} = 0.3$, and $\gamma = 0.1$. As the annealing process gradually pushes the selection probabilities toward near-binary values at inference time, the threshold θ primarily serves as a final binarization step rather than a highly sensitive hyperparameter. Empirical evaluation confirmed that any threshold within a broad interval (0.4–0.8) produces identical selections.

The compressed record representation, obtained by combining a weighted mean pooling operation with a softmax-weighted max-pooling oper-

ation, is encoded by the VIB module into a 128-dimensional Gaussian latent variable, from which a linear decoder reconstructs the global patient representation. The model is trained end-to-end with $\lambda_{\text{con}} = \lambda_{\text{comp}} = 1$. The KL-divergence coefficient β is linearly increased from 0 to $\beta_{\max} = 10^{-2}$ during the first three epochs and maintained constant thereafter, as this strategy proved important for preventing posterior collapse. Optimization is performed using Adam with a constant learning rate of 10^{-4} for 10 epochs, employing a per-step batch size corresponding to a single admission and gradient accumulation over 8 steps. Experiments were conducted on a Dell Precision 7960 Tower workstation (Intel Xeon w9-3595X, dual NVIDIA RTX 6000 Ada GPUs, 512 GB RAM).

Following Kruse et al. (2025), we employed 8-bit quantized versions of the language models to reduce GPU memory consumption and improve computational efficiency. This design choice further enhances the practicality of deploying LLMs in the medical domain, where computational resources are often constrained (Fime et al., 2025). Although the proposed approach is fully unsupervised, a validation set was used because the RAG framework requires validation for tuning retrieval-related hyperparameters, specifically top- K , chunk size, and chunk overlap. Accordingly, the validation set was used to identify the optimal parameter configuration for the LLMs, using the same parameter settings explored in Kruse et al. (2025). For Llama-2 and Mistral, the selected configuration was top- $K = 10$, a chunk size of 500, and a chunk overlap of 100, whereas Llama-3, Qwen, and DeepSeek used top- $K = 20$, a chunk size of 750, and a chunk overlap of 100. In contrast, the parameters of CLARUS were not optimized on the validation split, in line with the fully unsupervised nature of the proposed approach. With respect to the embedding model employed within the LangChain (Chase, 2022) and FAISS (Douze et al., 2026) retrieval framework, prior studies (Kruse et al., 2025; Myers et al., 2025) demonstrated the effectiveness of BGE (Xiao et al., 2024); therefore, BGE embeddings were adopted in all retrieval experiments. Consistent with Kruse et al. (2025), a maximum context length of 2000 tokens was imposed during the extraction phase of prompt-based approaches, while the generation phase was limited to 1000 tokens. A temperature of 0.2 was adopted based on recommendations from prior work (Ellershaw et al., 2024; Van Veen et al., 2024; Aali et al., 2025; Ganzinger et al., 2025),

showing that lower temperature values produce discharge summaries that are more fluent, coherent, and clinically accurate.

B Concept Depth Computation

This appendix details the procedure used to associate each concept hyperedge $e_j \in \mathcal{E}_{\text{con}}$ with an ontology depth D_j , which modulates the depth-aware concept attention described in Section 3.2.

Definition. Given the directed graph $\mathcal{G}_{\text{ont}} = (\mathcal{V}_{\text{ont}}, \mathcal{E}_{\text{ont}})$ induced by relations among UMLS concepts, we define the depth $D(v)$ of a concept $v \in \mathcal{V}_{\text{ont}}$ as the length of the shortest path from any root concept to v , where roots are concepts with no parents in $\mathcal{G}_{\text{ont}}$. Each edge $(p, c) \in \mathcal{E}_{\text{ont}}$ denotes that p is a parent of c . Concepts that are roots themselves are therefore assigned depth 0. For each concept hyperedge e_j , D_j is then obtained by mapping the underlying clinical concept to its corresponding UMLS Concept Unique Identifier (CUI) and computing the corresponding ontology depth $D(\text{CUI})$.

We adopted the minimum-path definition rather than the maximum-path definition because it provides a more conservative and stable lower bound on specificity that is less sensitive to local irregularities in the hierarchy, such as concepts positioned at different depths across parallel branches induced by multiple inheritance. This choice is also aligned with the role of D_j in the model, as it potentially determines the number of level-specific views through which a concept hyperedge is projected (Section 3.2). By relying on a conservative estimate of specificity, the model can avoid excessive fragmentation of the attention space for concepts whose greater depth is supported only by a subset of their ancestral paths.

Computation. In our experiments, concepts are extracted from UMLS and restricted to English and to SNOMED CT source vocabularies. We include both the hierarchical relations (PAR/CHD) and the broader/narrower relations (RB/RN), as the latter are needed to recover hierarchical links that, in some UMLS releases, are encoded only through RB/RN rather than PAR/CHD. Depths are computed via a standard breadth-first search (BFS) initialized with all root concepts at distance zero. The relation graph is first inverted into a parent-to-children adjacency map, after which the BFS proceeds systematically level by level, assigning to each newly visited concept a depth equal to one plus the depth

of its predecessor. Because the UMLS hierarchy forms a directed acyclic graph rather than a tree, the algorithm includes a conditional re-enqueuing step: when a concept is reached via a shorter path than previously recorded, its depth is updated and the improvement is propagated to its descendants. This guarantees that $D(v)$ corresponds to the minimum number of hierarchical steps separating v from the most general concept that subsumes it. The procedure terminates when no further updates are possible; termination is guaranteed because each concept can be re-enqueued only a finite number of times: the depth assigned to a node is monotonically non-increasing across updates, lower-bounded by zero, and integer-valued.

C Dataset Statistics

We evaluate CLARUS on MIMIC-III (Johnson et al., 2016), a freely accessible critical care database of de-identified EHRs. Rather than targeting a single component of the discharge summary as in prior work (Searle et al., 2023), we address the full summarization task across all three clinically relevant sections: *Diagnosis*, *Brief Hospital Course*, and *Discharge Instructions*. We follow the experimental setup of Kruse et al. (2025), who similarly frame the task as a three-section generation problem and define the regular expressions used to extract each target section from the discharge summaries. We adopt the train/validation/test partition of Searle et al. (2023), corresponding to an 80/10/10 split over 51,744 admissions (41,395 / 5,174 / 5,175).

The statistics reported in Table 5 directly reflect the challenges motivating CLARUS. Clinical records in MIMIC-III are long and highly heterogeneous: on average, each admission comprises 32.77 notes (median 15), spanning multiple clinical categories as shown in Figure 5. Input lengths computed with the Llama-2-13B tokenizer further underscore this heterogeneity, with a mean of 19,399 tokens but a median of only 6,170, indicating a heavily skewed distribution in which a substantial fraction of records far exceeds the effective context window of standard LLMs. This variability is precisely the regime in which prompting-based and retrieval-based approaches are most likely to fail. Discharge summaries, by contrast, are considerably more compact and uniform, averaging 3,750 tokens (median 3,455), highlighting the degree of compression and synthesis required. The mean hospital stay of 8.76 days (median 5) further

	All	Train	Val	Test
<i>Admission counts</i>				
Admissions	51,744	41,395	5,174	5,175
<i>Notes per admission</i>				
Mean	32.77	32.88	31.87	32.76
Median	15	15	15	15
<i>Length in tokens</i>				
Input mean	19,399	19,485	18,843	19,264
Input median	6,170	6,188	6,044	6,131
Discharge mean	3,750	3,750	3,737	3,760
Discharge median	3,455	3,455	3,443	3,462
<i>Stay duration (days)</i>				
Mean	8.76	8.78	8.63	8.72
Median	5	5	5	5
<i>Note-count bin distribution</i>				
1–7	12,816	10,180	1,306	1,330
8–14	12,636	10,168	1,254	1,214
15–27	11,363	9,078	1,136	1,149
>27	14,929	11,969	1,478	1,482

Table 5: Dataset statistics for MIMIC-III. Token lengths are computed with the Llama-2-13B tokenizer.

suggests temporally dispersed evidence that must be coherently integrated across notes.

To further analyze the hierarchical structure of the extracted clinical concepts, we computed the ontology depth distribution of concept hyperedges using the MedCAT-extracted SNOMED CT concepts and the depth computation procedure described in Appendix B. The resulting distribution is reported in Table 6. Most concepts are associated with intermediate ontology depths, with depth 2 accounting for 36.47% of all extracted concepts, followed by depth 1 (23.88%) and depth 3 (17.38%). Very shallow concepts (depth 0) represent only 8.77% of the dataset, while highly specific concepts at depths 4 and 5+ collectively account for less than 14%. This distribution supports the design rationale of CLARUS. Clinical records tend to contain a large proportion of moderately specific concepts connected through shared ancestors, while highly specific concepts appear more sparsely. The observed depth distribution therefore supports the design choice of incorporating depth-aware concept attention, which leverages ontology hierarchy to reinforce sparse fine-grained concepts through their more frequently observed ancestors.

D Depth Level Representation

In this appendix, we investigate the effect of the number of depth levels used in the depth-aware concept attention mechanism. The corresponding results are reported in Table 7. The param-

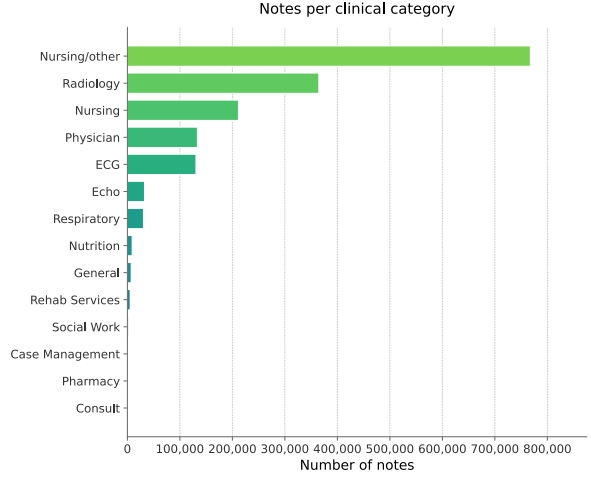


Figure 5: Distribution of clinical notes by category in the MIMIC-III dataset.

Depth	Percentage (%)
0	8.77
1	23.88
2	36.47
3	17.38
4	5.71
≥ 5	7.79

Table 6: Distribution of ontology depths for MedCAT-extracted clinical concepts in MIMIC-III.

ter $D_{\max}$ defines the maximum number of depth-specific representations associated with concept hyperedges. Concepts whose depth exceeds $D_{\max}$ are mapped to the highest available depth level. The *No depth* configuration disables the mechanism entirely, preserving the concept hypergraph structure while collapsing the attention computation into a single shared representation.

The results reveal a non-monotonic behavior. The introduction of a single additional depth level ($D_{\max} = 1$) already improves performance over the depth-agnostic baseline, with further gains observed up to $D_{\max} = 3$, which yields the best results across both semantic and clinical evaluation metrics. This outcome aligns with the depth distribution shown in Table 6, where concepts with depths between 0 and 3 account for approximately 86.5% of all extracted concepts, thereby encompassing the majority of clinically relevant entities at their natural granularity. In contrast, increasing the maximum depth to $D_{\max} = 4$ results in a marginal but consistent performance decline. A plausible explanation is that concepts at depth ≥ 4 account for less than 14% of extracted entities, limiting

Variant	BERTScore	METEOR	AlignScore	MEDCON
No depth	72.71	12.52	40.94	19.96
$D_{\max} = 1$	73.24	12.63	41.12	20.25
$D_{\max} = 2$	73.81	12.79	41.38	20.69
$D_{\max} = 3$	74.22	12.88	41.53	20.98
$D_{\max} = 4$	74.05	12.84	41.46	20.82

Table 7: Analysis of the depth-aware concept attention mechanism with Llama-2-13B on MIMIC-III. *No depth* omits the depth-aware attention mechanism. The remaining rows vary the maximum depth $D_{\max}$ used by the depth-aware attention.

the learning signal for additional depth-specific projections. As a consequence, the introduction of further level-specific representations increases model complexity without yielding proportional improvements in semantic expressiveness.

E Prompt Templates

The extraction phase of the prompting-based baselines is carried out by adopting the prompts defined in the corresponding reference studies. In particular, Ext-ZS relies on the prompt introduced in Naemi and Sahafi (2026), while Abs-ZS and Ext-CoT employ the prompts proposed in Kruse et al. (2025). For the generation phase, the format described in Kruse et al. (2025) is followed, where a task-specific query is provided to the LLM for each section of the discharge summary after the extraction step. More precisely, the queries “*What is the patient’s main diagnosis?*” for the Diagnosis section, “*Summarize the hospital course for this patient in a concise and accurate way.*” for the Brief Hospital Course, and “*Given the input EHR data, generate discharge instructions for this patient.*” for the Discharge Instructions are used.

F Further Analysis of Mistral’s Performance

Section-level analysis. Figure 6 presents a comparison between CLARUS and Ext-CoT, the strongest baseline for Mistral on semantic metrics, across discharge-summary sections. As with Llama-2, the pattern of improvements is markedly heterogeneous across sections and metric families, suggesting that the proposed selection mechanism contributes in different ways depending on the nature of the target text.

On *Diagnosis*, CLARUS achieves consistent and sizeable improvements across all metric families, with the largest deltas concentrated on METEOR (+2.73), AlignScore (+2.70) and MEDCON (+2.40), and uniformly positive gains

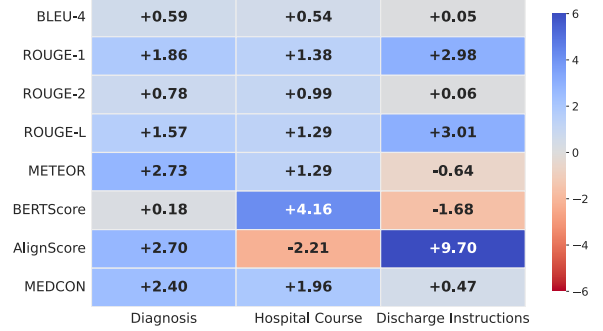


Figure 6: Section-level difference heatmap (CLARUS vs. Ext-CoT) for Mistral-7B. Blue indicates gains for CLARUS, red favors Ext-CoT.

on lexical metrics (BLEU-4 +0.59, ROUGE-1 +1.86, ROUGE-L +1.57). The relatively small BERTScore gain (+0.18) likely reflects the brevity of this section, where embedding-based similarity tends to saturate for both systems.

On *Brief Hospital Course*, the longest and most synthesis-heavy section, CLARUS shows uniformly positive gains on lexical metrics (BLEU-4 +0.54, ROUGE-1 +1.38, ROUGE-2 +0.99, ROUGE-L +1.29) and a particularly pronounced improvement on BERTScore (+4.16), which suggests that the structured aggregation performed over report-, category-, and concept-level hyperedges helps integrate evidence dispersed across many notes into a semantically coherent narrative. The drop observed on AlignScore (−2.21) appears localized to this section and may reflect the sensitivity of entailment-based metrics to phrasing variation in long-form narratives generated from multi-document inputs. Importantly, the consistent gain on MEDCON (+1.96) indicates that clinically relevant content is nevertheless preserved more faithfully than by the baseline.

On *Discharge Instructions*, the picture is more nuanced. Lexical metrics show a strong dissociation: ROUGE-1 (+2.98) and ROUGE-L (+3.01) improve substantially, while BLEU-4 (+0.05) and

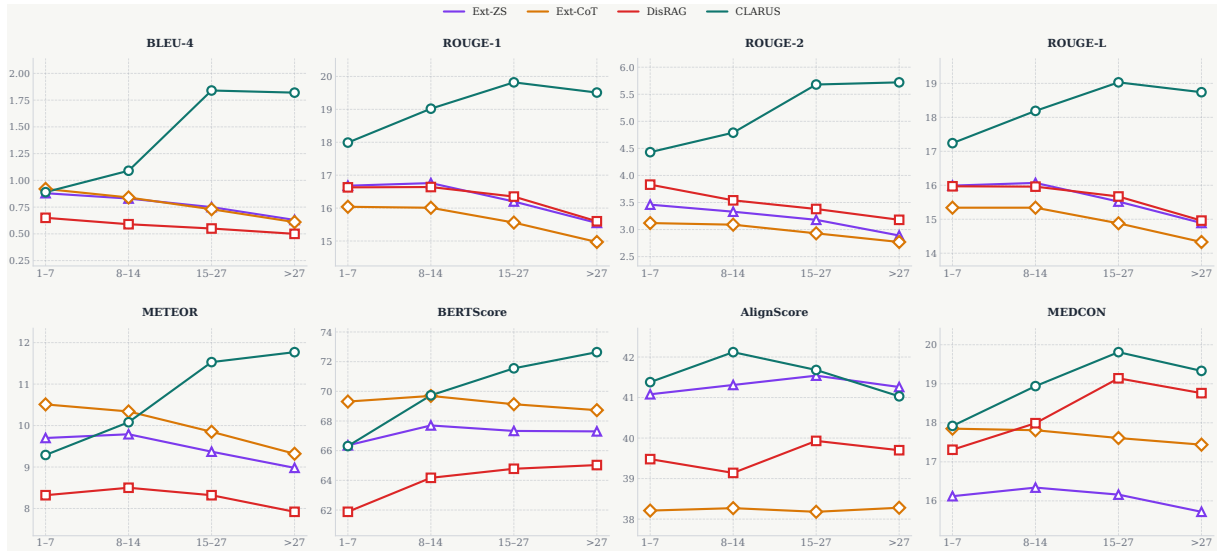


Figure 7: Performance of extract-then-summarize approaches with Mistral-7B as a function of the number of clinical notes per admission.

ROUGE-2 (+0.06) remain essentially flat, indicating that CLARUS recovers more salient content words and longer matching sub-sequences without necessarily aligning short n-grams. METEOR (−0.64) and BERTScore (−1.68) exhibit small negative deltas, which is consistent with the templated nature of discharge instructions, where surface-form and embedding-based similarity tend to saturate for both systems. In contrast, the clinical fidelity metrics improve markedly, with a particularly large gain on AlignScore (+9.70) accompanied by a positive shift on MEDCON (+0.47). This dissociation suggests that, even in a regime where retrieval- and prompting-based extractors are already competitive in terms of surface form, the structured aggregation performed by CLARUS produces outputs that are closer to the reference in terms of factual consistency and clinical content.

Robustness to input length. Figure 7 reports the performance of extract-then-summarize approaches with Mistral-7B as the number of clinical notes per admission increases. Overall, the trends are consistent with those observed for Llama-2 (Figure 4): CLARUS achieves the best results in most bin–metric combinations, particularly on longer and more heterogeneous records.

On lexical metrics, CLARUS improves as the number of source notes grows, with the largest gains between the 8–14 and 15–27 bins. In the > 27 bin, BLEU-4 and ROUGE-1 decline slightly, whereas ROUGE-2 continues to improve. In contrast, Ext-ZS, Ext-CoT, and DisRAG progressively

deteriorate with longer inputs. This trend reflects the limitations of prompting-based extractors, which are affected by context-window and lost-in-the-middle issues, as well as the dependence of DisRAG on pairwise query–passage similarity, which becomes less effective when relevant evidence is distributed across multiple notes.

Semantic metrics show a more nuanced behavior. On METEOR, CLARUS is the only method exhibiting a consistent improvement across bins, whereas prompting-based extractors decline and DisRAG remains relatively stable. On BERTScore, both CLARUS and DisRAG improve with longer inputs, but CLARUS maintains a substantial advantage that widens in the largest bin. In the 1–7 bin, prompting-based extractors, especially Ext-CoT, remain competitive, suggesting that when inputs are short, context limitations are less critical. From the 8–14 bin onward, however, CLARUS matches or surpasses Ext-CoT, with the gap increasing up to the > 27 bin.

Clinical metrics provide complementary evidence. On MEDCON, CLARUS achieves the highest score in every bin, consistently outperforming the strongest baseline, DisRAG. Although both methods achieve higher MEDCON scores on longer records overall, the gains of DisRAG are not reflected in other metrics, suggesting that higher concept overlap alone does not guarantee better summary integration or faithfulness. On AlignScore, differences across methods are smaller and less stable: CLARUS leads in the first three bins,

while Ext-ZS slightly outperforms it in the > 27 bin. Moreover, CLARUS shows a non-monotonic trend on this metric, indicating that improvements in semantic similarity and concept coverage are accompanied by stable rather than consistently increasing factual consistency.

Overall, the results obtained with Mistral broadly align with those observed for Llama-2: modeling higher-order structure often provides advantages on longer and more heterogeneous records, whereas baseline extractors remain competitive on shorter inputs where context limitations are less critical.

G Generalization Across Backbones and Datasets

This appendix reports two complementary evaluations that extend the experimental setting of the main paper: an assessment of CLARUS with additional LLM backbones on MIMIC-III and a cross-dataset evaluation on MIMIC-IV.

Additional LLM backbones. To assess whether the benefits of structured evidence selection extend beyond the two backbones used in the main paper, we extended the MIMIC-III evaluation to three additional open-source LLMs: Llama-3-8B, Qwen-2.5-7B, and DeepSeek-R1-Distill-Qwen-32B. All models use the same generation settings described in Appendix A, ensuring a controlled comparison across backbones. The results are reported in Figure 8. The relative performance patterns observed with Llama-2-13B and Mistral-7B are preserved across the additional backbones. CLARUS obtains the best score in the large majority of backbone–metric combinations, whereas the strongest baseline varies across models and metric families. Abs-ZS remains the weakest configuration across all five backbones, suggesting that an extraction stage is beneficial across different generators. The few cases in which a baseline is competitive with CLARUS are concentrated on metrics where the differences across methods are generally small.

Cross-dataset generalization. To assess the generalizability of CLARUS beyond the dataset used in the main experiments, we conducted an additional evaluation on MIMIC-IV, adopting the curated data and task formulation of Xu et al. (2024), which cover the generation of the *Brief Hospital Course* and *Discharge Instructions* sections. Results for all five backbones are reported in Figure 9. The performance trends observed on MIMIC-III are also observed on MIMIC-IV: CLARUS is the best-

performing approach in most backbone–metric combinations, whereas the strongest baseline varies across models and metric families.

H Human Evaluation Setup

To complement the automatic metrics, we conducted a human evaluation⁸ comparing CLARUS against DisRAG, the strongest unsupervised baseline under Llama-2-13B.

Reviewer recruitment. The human evaluation involved 10 board-certified clinicians recruited from 3 institutions to mitigate the risk that institution-specific documentation practices or templates could systematically bias the assessment. The clinicians included intensivists, emergency medicine specialists, and internists, as these specialties most closely match the clinical population represented in MIMIC-III, which consists of adult intensive care admissions. Clinicians in these specialties routinely write and review discharge summaries for medically complex patients. Participation was entirely voluntary and unpaid. Prior to the evaluation, all clinicians completed a calibration round consisting of 1–2 pilot cases (not included in the reported results). This calibration was used to align the interpretation of the evaluation criteria, familiarize clinicians with the annotation interface, and establish the expected time required per case. To enable IAA estimation, we conducted a fully-crossed evaluation involving 3 clinicians, one from each participating institution. These clinicians independently rated a new set of 20 admissions, distinct from those used in the initial study, with all three clinicians rating every admission.

Evaluation criteria. Following the protocol proposed by Williams et al. (2025), we adopted a dual evaluation framework consisting of:

- **Pairwise preference judgment**, in which clinicians were asked whether they preferred *Summary A*, *Summary B*, or considered them equivalent (*Tie*);
- **Four quality dimensions**, rated on a 5-point Likert scale ranging from 1 (worst) to 5 (best): *Comprehensiveness*, *Correctness*, *Coherence*, and *Conciseness*.

The definitions adopted in our annotation guidelines are reported in Table 8.

⁸The protocol was approved by the Ethics Committee of the authors’ institution.

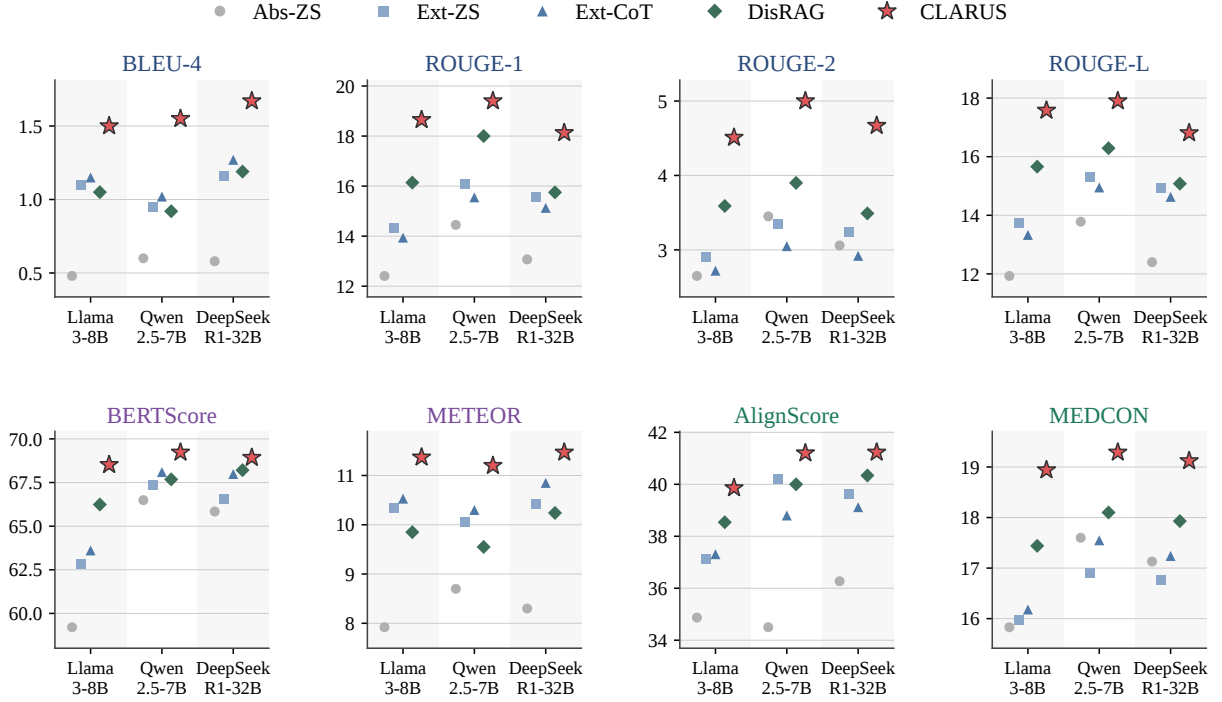


Figure 8: Comparison of CLARUS with the unsupervised baselines on MIMIC-III using three additional LLM backbones (Llama-3-8B, Qwen-2.5-7B, and DeepSeek-R1-Distill-Qwen-32B).

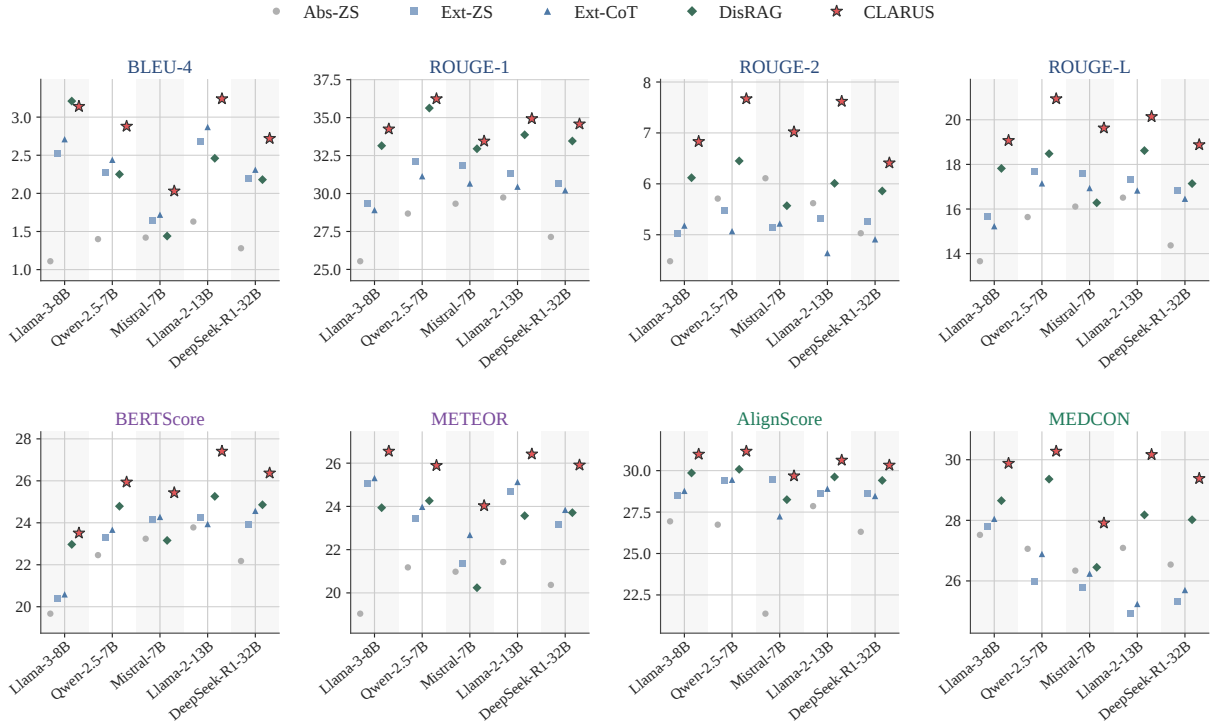


Figure 9: Cross-dataset evaluation on MIMIC-IV comparing CLARUS with the unsupervised baselines across five LLM backbones, for the generation of the *Brief Hospital Course* and *Discharge Instructions* sections.

Admission sampling. To obtain a robust assessment under realistic clinical conditions, admissions were randomly selected according to two complementary criteria. First, we targeted cases requiring the integration of multiple clinical notes, thereby

stressing the ability of LLMs to synthesize information across documents. Second, we constrained the number of notes per admission to limit the cognitive burden on clinicians and reduce potential evaluation bias associated with excessively long

Criterion	Definition
Preference	Overall preference between the two candidate summaries (Summary A, Summary B, or Tie).
Comprehensiveness	Whether the summary includes all clinically relevant information appropriate for a discharge summary.
Correctness	Whether the summary is factually accurate and faithful to the source notes, without errors, hallucinations, or missing critical information.
Coherence	Whether the text is logically structured, fluent, and cohesive across sentences.
Conciseness	Whether the summary is succinct, avoiding unnecessary verbosity while preserving all key information.

Table 8: Definitions of the evaluation criteria adopted in the human evaluation, following Williams et al. (2025).

records (Hall and Walton, 2004). Accordingly, in the initial study, we randomly sampled 50 admissions containing between 10 and 15 clinical notes, a range that captures the multi-document nature of the task while remaining feasible for human evaluation. For the additional fully-crossed evaluation, we sampled a separate set of 20 admissions, distinct from those in the initial study, from the same pool of admissions containing 10–15 clinical notes. On average, clinicians required approximately 25 minutes to evaluate a single admission.

Annotation protocol. In the initial study, each clinician evaluated 5 admissions, with no overlap across clinicians; thus, each admission was assessed by a single clinician. This allocation maximized coverage of the 50-admission sample while accounting for the time and cognitive demands of evaluating each case and the limited availability of board-certified clinicians. However, it precluded estimation of IAA for the initial sample. We therefore conducted a separate fully-crossed evaluation in which 3 clinicians independently assessed all 20 admissions in the additional sample, using the same annotation protocol and interface as in the initial study. In both evaluations, the two candidate summaries were anonymized and presented in randomized order for each admission to mitigate potential ordering and identity biases.

Clinical factuality audit protocol. Within the fully-crossed evaluation, the 3 clinicians additionally annotated each summary for six binary error categories, defined in Table 9. Clinicians had full access to the corresponding source notes during annotation and were instructed to flag an error only when it could be directly verified against the notes.

Error Category	Definition
Unsupported statement	Claims not attributable to the source notes.
Contradiction	Statements conflicting with the source notes or with other parts of the summary.
Clinically important omission	Missing information a clinician would consider necessary for safe handoff.
Temporal error	Incorrect ordering or timing of clinical events.
Medication/discharge error	Incorrect dosage, drug, or instruction.
Potentially harmful error	Any error judged capable of leading to patient harm if acted upon.

Table 9: Definitions of the error categories adopted in the clinical factuality audit.

The categories are not mutually exclusive, so a single summary could be flagged for more than one error type. For each admission, the final label of a category was then obtained by majority vote across the three clinicians.