Compressed Representations of Patient Records: A Comparative Study of Template-Based and LLM-Based Methods for Clinical Data Summarization and Visualization

Andreas Stöckl, Oliver Krauss, Sophie Bauernfeind
Digital Media Lab, Digital Media Lab, Research & Development
University of Applied Sciences Upper Austria
Hagenberg, Austria
{andreas.stoeckl, oliver.krauss, sophie.bauernfeind}@fh-hagenberg.at

Abstract

Electronic Health Records (EHRs) contain comprehensive patient information that is often voluminous and challenging to review efficiently. This paper presents a systematic evaluation of multiple methods for compressing patient records into standardized, comparable formats. Four compression approaches are implemented and compared: two template-based methods (structured extraction, extractive key-phrase) and two LLM-based methods (LLM, and hybrid LLM with 8 different models). Using a synthetic cohort of 75 patient records generated with realistic clinical patterns, each method is evaluated on information preservation (diagnosis, medication, allergy, lab value recall, and vital accuracy), compression efficiency, and output quality. Across methods, diagnosis recall ranged from 0.637 to 1.000, with medication and allergy recall consistently exceeding 0.880. In the test setup, the template-based approach yielded the highest compression ratio (7.6×), while the hybrid methods provided the most balanced trade-off between compression and clinical utility. These results suggest that combining structured extraction with LLM-generated summaries can be an effective strategy for scenarios requiring both compact representations and contextual clinical information.

Keywords: Electronic Health Records, Clinical Text Summarization, Large Language Models, Data Compression, Clinical NLP, Infographics

1. Introduction

The proliferation of Electronic Health Records (EHRs) has created both opportunities and challenges for healthcare delivery. While comprehensive digital records enable better continuity of care, the sheer volume of clinical documentation often exceeds what clinicians can efficiently review (Shickel et al., 2018). A typical patient encounter may require reviewing hundreds of pages of prior records, laboratory results, medication lists, and clinical notes (Pivovarov and Elhadad, 2015).

This information overload has motivated significant research into clinical text summarization and EHR compression. Recent advances in Large Language Models (LLMs) have shown promising results in generating coherent clinical summaries (Van Veen et al., 2024; Schilder et al., 2025; Schoonbeek et al., 2025). However, questions remain about the optimal approach for different clinical use cases, particularly when the goal is to enable rapid comparison across multiple patients.

This work investigates methods for generating compressed representations of patient records that retain essential clinical information while supporting efficient review. The contributions include:

1. A systematic comparison of four compression methods spanning template-based and LLM-based approaches with different models.
2. A synthetic patient dataset with realistic clinical patterns for reproducible evaluation.
3. Novel evaluation metrics focused on information preservation and comparability.
4. An open-source implementation with infographic visualization tools.

2. Related Work

2.1. Clinical Text Summarization

Clinical text summarization has evolved significantly with the advent of deep learning. Early approaches relied on extractive methods that selected key sentences from source documents (Pivovarov and Elhadad, 2015). More recent work has explored abstractive summarization using transformer architectures (Shickel et al., 2018).

Van Veen et al. (Van Veen et al., 2024) demonstrated that adapted LLMs can outperform medical experts in clinical text summarization tasks. Their study applied adaptation methods to eight LLMs across four clinical summarization tasks, finding that in 45% of cases LLM summaries were equivalent to expert summaries, and in 36% of cases they were superior.

A recent scoping review (Schilder et al., 2025) analyzed 30 studies on LLM-based clinical text

Proceedings of the Third Workshop on Patient-Oriented Language Processing (CL4Health), pages 96–106
12 May 2026.

summarization, identifying key limitations including narrow research focus (57% on radiology reports), heavy reliance on the MIMIC dataset (50% of studies), and insufficient clinical impact evaluation.

2.2. EHR Representation Learning

Deep representation learning for EHRs has been extensively studied (Si et al., 2021). The Graph Convolutional Transformer (Choi et al., 2020) demonstrated the value of learning graphical structures from EHR data. BioBERT (Lee et al., 2020) and Clinical BERT (Alsentzer et al., 2019) have shown strong performance on biomedical NLP tasks.

2.3. Large Language Models in Healthcare

GPT-4 (OpenAI, 2023) and similar LLMs have demonstrated remarkable capabilities in medical knowledge tasks (Singhal et al., 2023). Tang et al. (Tang et al., 2023) evaluated LLMs on medical evidence summarization, finding promising but inconsistent results. Concerns remain about hallucination and factual accuracy in clinical contexts (Wornow et al., 2023). MedGemma models were built upon Gemma 3, with optimization for medical domains, and has been shown to outperform other generative models of comparable size on clinically oriented tasks. (Sellergren et al., 2025)

2.4. Synthetic EHR Generation

Synthea (Walonoski et al., 2018) provides an open-source framework for generating realistic synthetic patient records. The synthetic data approach addresses privacy concerns while enabling reproducible research. Synthea generates complete patient histories conformant with FHIR standards, including diagnoses, medications, procedures, and laboratory values.

3. Methods

3.1. Dataset Generation

A synthetic cohort of 75 patient records was generated using a custom generator inspired by Synthea (Walonoski et al., 2018). Each record includes:

- Demographics (name, age, gender, MRN)
- Diagnoses with ICD-10 codes (1-7 per patient)
- Medications with dosing (1-12 per patient)
- Allergies with reaction severity
- Vital signs (BP, HR, RR, SpO2, weight, ...)
- Laboratory results with reference ranges and abnormal flags
- Social and family history
- Clinical notes (1,500-2,500 characters)

Patient complexity was varied across three levels (low: 29 patients, moderate: 33 patients, high: 13 patients), affecting the number of diagnoses, medications, and abnormal findings.

3.2. Example Patient Record

Listing 1 shows an excerpt from a synthetic patient record. The full clinical note spans approximately 1,874 characters and includes structured sections for chief complaint, history, medications, allergies, vital signs, physical examination, laboratory results, and assessment.

```
CHIEF COMPLAINT: COPD exacerbation

HISTORY OF PRESENT ILLNESS:
Patient presents COPD exacerbation.
Patient reports compliance with current
medications. Ongoing management of
Urinary tract infection.

PAST MEDICAL HISTORY:
- Urinary tract infection (N39.0)

MEDICATIONS:
- Aspirin 81mg daily
- Sertraline 50mg daily
- Lisinopril 10mg daily

ALLERGIES:
- No known allergies

VITAL SIGNS:
- Blood Pressure: 119/72 mmHg
- Heart Rate: 65 bpm
- Oxygen Saturation: 100%
- Weight: 111.9 kg, Height: 173 cm
- BMI: 37.4

LABORATORY RESULTS:
- WBC: 8.41 K/uL (4.5-11.0)
- Hemoglobin: 15.94 g/dL (12.0-17.5)
- Glucose: 92.76 mg/dL (70-100)
- eGFR: 49.0 mL/min (90-120) [L]
[...]
```

Listing 1: Excerpt from original patient record (Patient MRN000001, 56-year-old male). Full record: 1,874 characters.

3.3. Compression Methods

In total, four distinct compression methods were evaluated. The LLM-based approach and the hy-

brid LLM approach each incorporated eight different models, resulting in 18 individual compression tests across all four methods:

3.3.1. Template-Based Extraction

A rule-based approach that extracts key information using predefined templates without natural language generation. Key elements (diagnoses, medications, vitals, abnormal labs) are directly extracted and formatted into a standardized structure.

3.3.2. Extractive Key-Phrase

Uses TF-IDF-inspired scoring to identify and extract the most clinically significant phrases from the patient record. Medical terms are weighted by category (diagnoses: 3.0, allergies: 3.0, medications: 2.5, abnormal labs: 2.0).

3.3.3. LLM

Uses different LLM models to generate structured clinical summaries. The model receives formatted patient data and returns JSON-structured output including primary diagnoses, vitals summary, risk factors, active allergies, active medications, critical labs, clinical summary and care priorities. The evaluation included both cloud-based and self-hosted model deployments:

- GPT-4o-mini (cloud)
- GPT-4o (cloud)
- google_medgemma-4b-it-Q6_K_L (local)
- google_medgemma-27b-it-Q6_K_L (local)
- gemma-3-12b-it-Q4_K_M (local)
- lfm2.5-thinking (local)
- Meta-Llama-3.1-8B-Instruct-Q6_K_L (local)
- nemotron-3-nano (local)

Prompt: No prompt engineering was performed to avoid introducing potential bias in favor of specific models. Instead, a single concise and clear instruction set was developed that consistently produced valid outputs across all evaluated systems.

3.3.4. Hybrid-LLM

Combines template-based structured extraction, described in subsection 3.3.1 with LLM-generated clinical narratives. Structured data (vitals, labs, medications) are extracted algorithmically, while the LLM generates a comparative-friendly clinical summary. The models used in the hybrid LLM evaluation were identical to those utilized in the standard LLM approach.

```
You are analyzing a patient record for
↪ clinical summarization.
Create a comprehensive yet concise
↪ clinical summary optimized for:
1. Quick clinical review
2. Comparison with other patients
3. Identifying key risks and priorities

Patient Record:
[...]

Provide your analysis in this exact JSON
↪ structure:
{{
    "primary_diagnoses": ["List
    ↪ diagnoses in order of clinical
    ↪ significance"],
    "vitals": "List important vitals as
    ↪ short plain text",
    "active_allergies": ["List important
    ↪ known allergies"],
    "active_medications": ["List active
    ↪ medications as plain text
    ↪ including name, dose and
    ↪ frequency"],
    "risk_factors": ["Identified risk
    ↪ factors for adverse outcomes"],
    "critical_labs": ["List abnormal
    ↪ labs as plain text including
    ↪ values and clinical
    ↪ interpretation"],
    "care_priorities": ["Top 3 care
    ↪ priorities in order of
    ↪ urgency"],
    "clinical_summary": "A 2-3 sentence
    ↪ clinical overview covering: main
    ↪ conditions, current status, and
    ↪ key concerns"
}}

Respond ONLY with valid JSON.
```

Listing 2: Prompt structure for LLM compression, with [...] being the text of the patient shown in Listing 1.

Prompt: The prompt focuses on the overall summary, with the already extracted information to reduce the input token.

3.4. Output Format

All methods produce a standardized `CompressedPatientRecord` containing:

- Patient ID, name, age, gender
- Primary diagnoses (list)
- Active medications (list with dosing)
- Allergies

```
Based on this patient summary, generate
↪  a 2-3 sentence clinical overview
↪  optimized for comparing with other
↪  patients:

Patient: {template_result.name},
↪  {template_result.age}y
↪  {template_result.gender}
Diagnoses: {', '.join(template_result.p
↪  rimary_diagnoses)}
Vitals: {template_result.vital_summary}
Abnormal Labs: {', '.join(template_resu
↪  lt.critical_labs)}
Risk Factors: {', '.join(template_resul
↪  t.risk_factors)}

Provide your analysis in this exact JSON
↪  structure:
{{
    "clinical_summary": "2-3 sentence
    ↪  clinical overview optimized for
    ↪  comparing with other patients."
}}

Focus on: disease severity, control
↪  status, and key concerns. Be concise
↪  and just return the json.
```

Listing 3: Prompt to create summary for Hybrid Compression strategy using extracted template information.

- Vital signs summary (one-line format, including values of BP, HR, SpO2, weight)
- Critical/abnormal laboratory values
- Identified risk factors
- Clinical summary narrative (LLM methods only)

3.5. Evaluation Metrics

The methods were evaluated on multiple dimensions:

Information Preservation:

- *Diagnosis Recall*: Proportion of original diagnoses present in compressed output
- *Medication Recall*: Proportion of original medications preserved
- *Allergy Recall*: Proportion of allergies correctly captured
- *Lab Abnormal Recall*: Proportion of abnormal lab values identified
- *Vital Accuracy*: Correctness of vital sign representation

Compression Metrics:

- *Compression Ratio*: Original length / Compressed length
- *Field Completeness*: Proportion of expected fields populated (without clinical summary)

Quality Score: A composite metric combining preservation (40%), field completeness (25%), compression efficiency (15%), and presence of clinical summary (20%).

3.6. Visualization

An infographic visualization system was developed to render compressed patient records as interactive HTML dashboards. The visualization includes:

- Patient cards with color-coded risk indicators
- Comparison tables for cohort-level review
- Aggregate statistics (mean age, diagnoses, medications)
- Highlighted abnormal values and risk factors

Figure 1 shows a screenshot of the patient cohort dashboard. The header displays aggregate statistics (75 patients, average age 57, 2.9 diagnoses per patient, 3.7 medications per patient). Below, individual patient cards present compressed information with visual risk indicators (color-coded badges showing LOW/MEDIUM/HIGH risk). Each card displays vital signs with color-coded borders for abnormal values, diagnosis chips colored by severity, medication lists, and critical laboratory findings.

3.7. Compression Output Examples

Listing 4 shows the same patient as in Listing 1 with compressed record using three different methods, demonstrating the trade-off between compression ratio and narrative richness.

4. Results

4.1. Information Preservation

Table 1 presents the comparative results across all methods. The diagnosis recall rate varied between 0.637 (LLM-lfm2.5) and 1.000 (Extractive Key-Phrase & Template-Based), while the medication recall rate and allergy recall rate all had values above 0.880.

Lab abnormal recall varied between 0.346 (LLM-lfm2.5ing) and 1.000 (Extractive Key-Phrase). The slight reduction in LLM methods occurred when models reformulated lab findings rather than just list the exact values and some also misses flaged

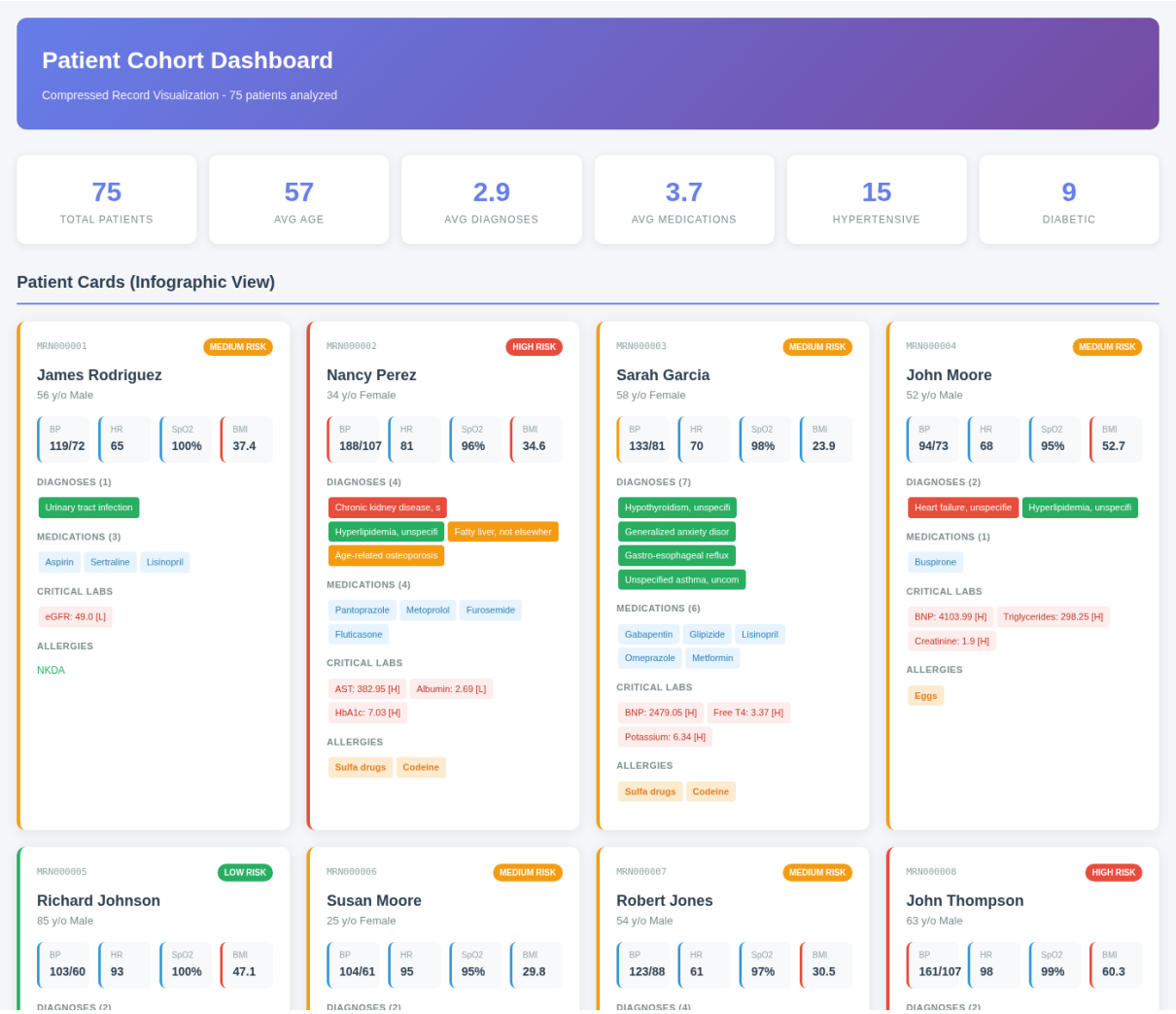

Figure 1: Patient Cohort Dashboard visualization. The header shows aggregate cohort statistics. Patient cards display compressed records with color-coded risk levels, base on the complexity (green=low, orange=medium, red=high), vital signs, diagnoses, medications, and critical lab values. This infographic format enables rapid visual comparison across patients.

lab values. As for the template-based compression is concerned, only the first 5 values were taken to achieve compression.

Vital sign accuracy varied between 0.600 (Extractive Key-Phrase) and 0.997 (LLM-gemma-3-12b).

As expected, there was hardly any difference in the quality score for the hybrid variant, as the difference between the methods is only apparent in one field, the summary, which is discussed in 4.4. The improved score of the lfm2.5 models can be explained by the higher compression rate.

4.2. Compression Efficiency

Template-Based extraction achieved the highest compression ratio (7.6x), reducing average clinical notes from approximately 1,874 characters to 247 characters. This aggressive compression comes at the cost of excluding clinical narratives. The Extractive Key-Phrase method has a lower compression ratio of 4.1.

In this setup, LLM methods produced the lowest compression ratios, ranging from 3.2 to 3.8, due to the inclusion of generated clinical summaries. The Hybrid-LLM method balanced compression (4.5x) with narrative quality.

4.3. Quality Comparison

The Hybrid-LLM method achieved the highest overall quality score (0.88), followed by the LLM approach using GPT-4o (0.878), the Key-Phrase method (0.868), and the remaining LLM models (0.783–0.868). In contrast, the Template-Based method yielded the lowest overall quality score (0.729).

The quality improvement from LLM methods primarily derives from the clinical summary narratives, which provide interpretive context not available in template-only approaches.

Template-Based Extraction (257 chars, 7.3x compression):

```
[MRN000001] James Rodriguez, 56y Male
Dx: Urinary tract infection
Meds: Aspirin, Sertraline, Lisinopril
Allergies: NKDA
VS: BP 119/72, HR 65, SpO2 100%
Labs: eGFR: 49.0 mL/min [L]
Risks: Obesity (BMI 37.4), Active smoker, Reduced kidney function (eGFR 49.0)
```

Hybrid-medgemma-27b (467 chars, 4.0x compression):

```
[MRN000001] James Rodriguez, 56y Male
Dx: Urinary tract infection
Meds: Aspirin, Sertraline, Lisinopril
Allergies: NKDA
VS: BP 119/72, HR 65, SpO2 100%
Labs: eGFR: 49.0 mL/min [L]
Risks: Obesity (BMI 37.4), Active smoker, Reduced kidney function (eGFR 49.0)
Summary: 56-year-old male presenting with a urinary tract infection. He has
significant risk factors including obesity (BMI 37.4), active smoking, and reduced
kidney function (eGFR 49.0 mL/min), which may impact treatment and recovery.
```

LLM-gpt-4o-mini (563 chars, 3.3x compression):

```
[MRN000001] James Rodriguez, 56y Male
Dx: Urinary tract infection (N39.0)
Meds: Aspirin, Sertraline, Lisinopril
Allergies: NKDA
VS: BP: 119/72 mmHg, HR: 65 bpm, SpO2: 100%, Weight: 111.9 kg
Labs: eGFR: 49.0 mL/min [L] - indicates potential renal impairment
Risks: Current smoker (0.5 pack/day), Family history of stroke and colon cancer,
Elevated weight
Summary: James Rodriguez is a 56-year-old male diagnosed with a urinary tract
infection. Current management includes medications for cardiovascular health and
mental health, but significant risk factors include his smoking status and family
history of chronic diseases.
```

Listing 4: Comparison of compression outputs for Patient MRN000001 using three methods. Template-based achieves highest compression; LLM methods provide clinical interpretation.

Method	Dx Recall	Med Recall	All. Recall	Lab Recall	Vital Acc	Comp. Ratio	Field Comp.	Preserv.	Quality
Template-Based	**1.000**	**1.000**	**1.000**	0.943	0.800	**7.6x**	0.940	0.949	0.729
Extractive Key-Phrase	**1.000**	**1.000**	**1.000**	**1.000**	0.600	4.1x	0.953	0.920	0.868
LLM-Meta-Llama	0.806	0.998	0.973	0.520	0.837	3.2x	0.967	0.827	0.820
LLM-gemma-3-12b	0.821	0.987	0.987	0.916	0.976	3.5x	0.953	0.937	0.865
LLM-medgemma-27b	0.835	0.972	**1.000**	0.939	0.997	3.2x	0.967	0.949	0.868
LLM-medgemma-4b	0.837	**1.000**	**1.000**	0.728	0.859	3.4x	0.967	0.885	0.847
LLM-lfm2.5	0.637	0.970	0.880	0.346	0.808	3.8x	0.940	0.728	0.783
LLM-nemotron-3	0.831	0.887	0.907	0.747	0.931	3.4x	0.962	0.860	0.836
LLM-gpt-4o-mini	0.967	0.995	**1.000**	0.917	**1.000**	3.0x	**0.971**	**0.976**	0.879
LLM-gpt-4o	0.906	**1.000**	0.987	0.938	0.992	3.0x	0.962	0.965	0.872
Hybrid-Meta-Llama	**1.000**	**1.000**	**1.000**	0.943	0.800	4.5x	0.940	0.949	0.882
Hybrid-gemma-3-12b	**1.000**	**1.000**	**1.000**	0.943	0.800	4.5x	0.940	0.949	0.882
Hybrid-medgemma-27b	**1.000**	**1.000**	**1.000**	0.943	0.800	4.5x	0.940	0.949	0.882
Hybrid-medgemma-4b	**1.000**	**1.000**	**1.000**	0.943	0.800	4.5x	0.940	0.949	0.882
Hybrid-lfm2.5	**1.000**	**1.000**	**1.000**	0.943	0.800	4.6x	0.940	0.949	**0.884**
Hybrid-nemotron-3	**1.000**	**1.000**	**1.000**	0.943	0.800	4.5x	0.940	0.949	0.882
Hybrid-gpt-4o-mini	**1.000**	**1.000**	**1.000**	0.943	0.800	3.4x	0.940	0.949	0.865
Hybrid-gpt-4o	**1.000**	**1.000**	**1.000**	0.943	0.800	3.2x	0.940	0.949	0.863

Table 1: Comparison of Patient Record Compression Methods (average values)

Figure 2 visualizes the key metrics across all methods, highlighting the trade-off between compression ratio and quality score.

4.4. Clinical Summary Analysis

The choice of model—whether applied in the LLM or hybrid approach—has only a marginal influence on the resulting summaries based on our manual assessment. Although the hybrid approach more consistently incorporates vital signs and their corresponding values, the overall structure and content of the outputs remain largely comparable to those produced by the full LLM approach. The following examples present summaries generated using the LLM approach:

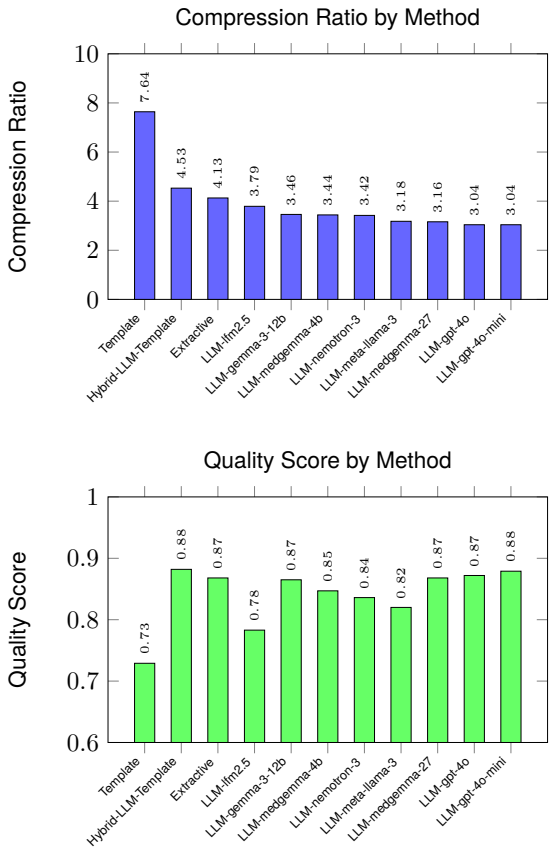

Figure 2: Infographic comparison of compression methods. Top: compression ratio; Bottom: quality score (higher is better for both). Template-Based leads in compression, Hybrid-LLM-Template and LLM-gpt-4o-mini in quality.

GPT-4o-mini produced concise 2-sentence summaries focusing on primary condition, current status and possible risks. Example: *"James Rodriguez is a 56-year-old male diagnosed with a urinary tract infection. Current management includes ... risk factors include his smoking status and family history of chronic diseases."*

GPT-4o produced a three-sentence summary that included medication status, diagnosis, and lifestyle factors, while the family history was not always present. The final sentence highlighted recommended next steps or focus areas. Example: *" ... is diagnosed with a urinary tract infection. His medical management includes ... controlled vital signs, but he is at risk due to obesity, smoking, and reduced renal function. ..."*

google_medgemma-4b-it-Q6_K_L produces a three-sentence overview: first a brief patient summary with the current diagnosis, followed by key lab values with possible interpretations, and finally treatment suggestions or important clinical considerations. Example: *"James Rodriguez, a 56-year-old male, presents ... elevated BUN/creatinine, suggesting possible early chronic kidney disease ... with a focus on encouraging smoking cessation."*

google_medgemma-27b-it-Q6_K_L produced a two- to three-sentence summary that included general patient information, diagnoses, and listed risk factors such as family history, while occasionally referring to or incorporating vital-sign values. Example: *"56M presents with ... current smoker with obesity and ... has Stage 3a CKD (eGFR 49). Key concerns include managing the UTI..."*

gemma-3-12b-it-Q4_K_M produced a three-sentence summary that included current diagnosis, incorporated key family history and relevant laboratory values without interpretation; and concluded with a brief treatment suggestion. Example: *"James Rodriguez is a 56-year-old male presenting ... He has a history of smoking and family history ... slightly low potassium. Management should focus on treating ..."*

lfm2.5-thinking produced highly inconsistent summaries, ranging from one to three sentences; in several cases, the output contained very little or no meaningful patient information and remained overly vague. Example: *"A 56-year-old male .. with normal vitals and no allergies, managed with standard antibiotics ... due to smoking-related risks."*

Meta-Llama-3.1-8B-Instruct-Q6_K_L produced one- to three-line summaries that included the patient's age, gender, and a variable-detail description of medical history and current diagnosis. However, it introduced additional descriptive terms or inferred diagnoses based on the listed medications. Example: *"A 56-year-old male patient ... and a history of hypertension. He is currently taking medications for heart health and depression. The primary concern is managing the infection ..."*

nemotron-3-nano produced one- to three-line summaries that included a patient's overview and a concise account of medical history and current diagnosis, highlighted relevant risk factors, and in some cases referenced or numerically reported vital-sign values with interpretation. Example: *"James Rodriguez is a ... his eGFR is reduced at 49 mL/min indicating early chronic kidney disease... family history of stroke ..."*

4.5. Processing Time

Non-LLM methods processed all 75 patients in under 0.1 seconds. Cloud LLM methods required 405–577 seconds (5.4–7.7 seconds per patient), while the Local hosted LLM's are slower with 2113–6682 seconds (35–111 minutes; 28–89 seconds per patient). The local Hybrid-LLM-Template methods are an improvement of the only LLM methods and took 589–1513 seconds (7.8–20 seconds per patient) and the cloud Hybrid-LLM-Template methods took 160-204 seconds (2.1–2.72 seconds).

4.6. Tabular Patient Comparison

Table 2 demonstrates that compressed records may facilitate quicker cross-patient comparison by presenting information in a standardized format. This structure could help clinicians more easily identify patients of interest and review key clinical parameters, though further study is needed to assess its impact in real-world clinical settings. The tabular representation appears to preserve essential information while potentially reducing cognitive load, but this requires additional validation.

5. Discussion

Barretto et al. (Barretto et al., 2026) compare GPT-4o and fine-tuned Llama3 models using several heuristic metrics as well as a novel LLM-based evaluation scheme. However, their analysis is limited to this small set of models, and domain-specialized models such as MedGemma, which are particularly relevant for medical summarization tasks were not included.

The comparative analysis of template-based, key-phrase, LLM-based, and hybrid compression methods highlights several important considerations for selecting an appropriate approach depending on clinical or computational requirements. While all methods successfully condense patient records into standardized formats, their performance varies substantially in terms of information preservation, compression efficiency, interpretability, and computational cost.

5.1. Method Selection Guidelines

Our results suggest different methods are optimal for different use cases:

High-throughput screening: Template-based extraction achieves the highest compression ratio (7.6×) while maintaining perfect preservation of diagnostic, medication, and allergy information with a fast processing time. However, this approach exhibits lower performance in lab abnormal recall and vital-sign accuracy, and the absence of a clinical summary further contributes to its comparatively low overall quality score. Consequently, this method is best suited for applications requiring minimal storage or high-throughput automated processing.

Clinical decision support: Hybrid methods provide the best balance of compression and clinical utility, with quality scores exceeding 0.88. The tested LLM hybrid methods were combined with the template-based approach. A combination with the extractive key phrase approach could lead to a better quality score, but compression could suffer as a result.

Patient comparison: All methods generate standardized outputs that enable structured, table-based comparison across patients, and the accompanying infographic visualization further facilitates rapid cohort-level assessment.

5.2. LLM Considerations

While LLM-based methods provide clear advantages in producing clinical narrative summaries, they also introduce notable considerations:

Interpretability and Trustworthiness: Certain LLM generated summaries provide clinical interpretation such as inferring chronic kidney disease (CKD) from reduced eGFR values, that template-based methods cannot deliver without substantial rule-based encoding. However, as outlined in the Clinical Summary Analysis 4.4, interpretive performance varies considerably across individual LLM models. Models trained specifically on medical data, such as MedGemma, are generally better suited for tasks requiring clinically grounded reasoning. Importantly, ensuring traceability and transparency remains essential, particularly in light of regulatory requirements under the European AI Act and the Medical Devices Directive (Bednarczyk et al., 2025).

Consistency: Template methods guarantee consistent output format while LLM outputs may vary in structure despite JSON formatting instructions. Notably, the lfm model deviated from the specified schema despite clear prompting, while the other LLMs consistently adhered to the intended structure under identical conditions. Nevertheless, the task of returning a consistently structured output in an LLM poses a challenge (Lu et al., 2025; Li et al., 2024; Geng et al., 2025).

Cost and Latency: API-based LLM methods incur per-token costs and network latency. Latency is particularly high with self-hosted models. Hybrid approaches can reduce this latency while still providing a clinical summary.

Hallucination Risk: While we observed no hallucinations in our evaluation, LLM summaries should be validated for safety-critical applications (Wornow et al., 2023).

6. Conclusion

A comprehensive evaluation of four distinct compression methods for compressing patient records into standardized, comparable formats was presented. The Template-Based and Extractive Key-Phrase achieved perfect preservation of diagnoses, medications and allergies, while the LLM-based methods provided enhanced clinical narratives with a quality improvement.

Patient	Age	Gender	Primary Dx	Key Labs	Risk Factors
James Rodriguez	56	Male	Urinary tract infection	eGFR: 49.0 mL/min [L]	Obesity (BMI 37.4) Active smoker, Reduced kidney function (eGFR 49.0)
Nancy Perez	34	Female	Chronic kidney disease, stage 4 Hyperlipidemia, unspecified Fatty liver, not elsewhere classified Age-related osteoporosis without fracture	Glucose: 249.25 mg/dL [H] Creatinine: 4.66 mg/dL [H] eGFR: 26.84 mL/min [L]	Hypertension Severe hypertension Obesity (BMI 34.6) Active smoker Reduced kidney function (eGFR 26.84)
John Moore	52	Male	Heart failure, unspecified Hyperlipidemia, unspecified	BNP: 4103.99 pg/mL [H] Potassium: 5.9 mEq/L [H] LDL: 111.29 mg/dL [H]	Obesity (BMI 52.7)
Susan Moore	25	Female	Fatty liver, not elsewhere classified Alcohol dependence, uncomplicated	Glucose: 290.45 mg/dL [H] Potassium: 6.97 mEq/L [H]	Active smoker
Richard Johnson	85	Male	Unspecified asthma, uncomplicated, Gastro-esophageal reflux disease with esophagitis	Glucose: 275.53 mg/dL [H]	Obesity (BMI 47.1), Heavy alcohol use

Table 2: Tabular Comparison of Compressed Patient Records (Template-Based Extraction)

The Hybrid-LLM-Template method and the LLM-gpt-4o-mini method achieved the best overall quality (0.88). For applications requiring interpretive summaries, the Hybrid-LLM approach with the a fitting model provides clinical context while maintaining structured data fidelity.

7. Outlook

Looking ahead, an important next step will be to evaluate the proposed approaches on real-world EHR datasets. Such an assessment would provide deeper insight into practical utility in authentic clinical environments. For this test only locally hosted LLM would be used to ensure compliance with data-protection and regulatory requirements. This would limit the range of deployable models.

Moreover, real-world EHRs typically contain longer, more heterogeneous, and less standardized documentation than the synthetic records used here, which may introduce additional challenges for both compression and summarization.

Another promising direction involves engaging clinicians directly in the evaluation process. Clinical expertise is essential for assessing the usefulness, safety, and interpretability of compressed records and summaries, and clinician-involved validation would offer a more grounded understanding of how these methods support real decision-making.

8. Ethical Considerations and Limitations

The synthetic patient data was generated for research purposes only and does not represent real individuals. The study also has several limitations:

1. **Synthetic data:** Although realistic patterns are produced by the generator, greater variability and complexity are present in real clinical records.
2. **English only:** All methods were evaluated on English-language records.
3. **No clinical validation:** Clinical utility was not assessed in cooperation with healthcare providers or on real-world patient data.
4. **Prompt engineering:** In this study, a short and clear prompt was used to generate valid JSON outputs. Because prompt engineering was not applied to the individual models, they may produce weaker results as a result.

9. Bibliographical References

Emily Alsentzer, John R Murphy, Willie Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, and Matthew BA McDermott. 2019. Publicly available clinical bert embeddings. *arXiv preprint arXiv:1904.03323*.

Daphne Barretto, Matthew Jin, and Bora Oztekin. 2026. Clinical text summarization with llm-based evaluation. Stanford CS224N Project.

Lydie Bednarczyk, Daniel Reichenpfader, et al. 2025. Scientific evidence for clinical text summarization using large language models: Scoping review. *Journal of Medical Internet Research*.

Edward Choi, Zhen Xu, Yujia Li, Michael W Dusenberry, Gerardo Flores, Yuan Xue, and Andrew M Dai. 2020. Learning the graphical structure of

electronic health records with graph convolutional transformer. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 606–613.

Saibo Geng, Hudson Cooper, Michal Moskal, Samuel Jenkins, Julian Berman, Nathan Ranchin, Robert West, Eric Horvitz, and Harsha Nori. 2025. JSONSchemabench: Evaluating constrained decoding with LLMs on efficiency, coverage and quality. In *ES-FoMo III: 3rd Workshop on Efficient Systems for Foundation Models*.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Diya Li, Yue Zhao, Zhifang Wang, Calvin Jung, and Zhe Zhang. 2024. Large language model-driven structured output: A comprehensive benchmark and spatial data generation framework. *ISPRS International Journal of Geo-Information*, 13(11):405.

Yaxi Lu, Haolun Li, Xin Cong, Zhong Zhang, Yesai Wu, Yankai Lin, Zhiyuan Liu, Fangming Liu, and Maosong Sun. 2025. Learning to generate structured output with schema reinforcement learning. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (ACL 2025)*, pages 4905–4918, Vienna, Austria.

OpenAI. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Rimma Pivovarov and Noemie Elhadad. 2015. Automated methods for the summarization of electronic health records. *Journal of the American Medical Informatics Association*, 22(5):938–947.

Frank Schilder, Daniel Reichenpfader, et al. 2025. Scientific evidence for clinical text summarization using large language models: Scoping review. *Journal of Medical Internet Research*, 27(1):e68998.

Rosanne C Schoonbeek, Jessica D Workum, Stephanie C E Schuit, Anne H Hoekman, Tarannom Mehri, Job N Doornberg, Tom P van der Laan, and Charlotte M H H T Bootsma-Robroeks. 2025. Quality and efficiency of integrating customised large language model-generated summaries versus physician-written summaries: a validation study. *BMJ Open*, 15(9).

Andrew Sellergren, Sahar Kazemzadeh, Tiam Jaroensri, Atilla Kiraly, Madeleine Traverse, Timo Kohlberger, Shawn Xu, Fayaz Jamil, Cían Hughes, Charles Lau, Justin Chen, Fereshteh Mahvar, Liron Yatziv, Tiffany Chen, Bram Sterling, Stefanie Anna Baby, Susanna Maria Baby, Jeremy Lai, Samuel Schmidgall, Lu Yang, Kejia Chen, Per Bjornsson, Shashir Reddy, Ryan Brush, Kenneth Philbrick, Mercy Asiedu, Ines Mezerreg, Howard Hu, Howard Yang, Richa Tiwari, Sunny Jansen, Preeti Singh, Yun Liu, Shekoofeh Azizi, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Riviere, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Elena Buchatskaya, Jean-Baptiste Alayrac, Dmitry Lepikhin, Vlad Feinberg, Sebastian Borgeaud, Alek Andreev, Cassidy Hardin, Robert Dadashi, Léonard Hussenot, Armand Joulin, Olivier Bachem, Yossi Matias, Katherine Chou, Avinatan Hassidim, Kavi Goel, Clement Farabet, Joelle Barral, Tris Warkentin, Jonathon Shlens, David Fleet, Victor Cotruta, Omar Sanseviero, Gus Martins, Phoebe Kirk, Anand Rao, Shravya Shetty, David F. Steiner, Can Kirmizibayrak, Rory Pilgrim, Daniel Golden, and Lin Yang. 2025. Medgemma technical report.

Benjamin Shickel, Patrick J Tighe, Azra Bihorac, and Parisa Rashidi. 2018. Deep ehr: A survey of recent advances in deep learning techniques for electronic health record (ehr) analysis. *IEEE Journal of Biomedical and Health Informatics*, 22(5):1589–1604.

Yuqi Si, Jingcheng Du, Zhao Li, Xiaoqian Jiang, Timothy Miller, Fei Wang, W Jim Jiménez, and Lucila Ohno-Machado. 2021. Deep representation learning of patient data from electronic health records (ehr): A systematic review. *Journal of Biomedical Informatics*, 115:103671.

Karan Singhal, Shekoofeh Azizi, Tao Tu, Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, et al. 2023. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180.

Liyan Tang, Zhaoyi Sun, Betina Idnay, Jordan G Nestor, Ali Soroush, Noémie Elhadad, Chunhua Weng, and Yifan Peng. 2023. Evaluating large language models on medical evidence summarization. *npj Digital Medicine*, 6(1):158.

Dave Van Veen, Cara Van Uden, Louis Berry, and Yiliang and";"; and"; others Chen. 2024. Adapted large language models can outperform medical experts in clinical text summarization. *Nature Medicine*, 30:1134–1142.

Jason Walonoski, Mark Kramer, Joseph Nichols, Andre Quina, Chris Moesel, Dylan Hall, Carlton Duffett, Kudakwashe Dube, Thomas Gallagher, and Scott McLachlan. 2018. Synthea: An approach, method, and software mechanism for generating synthetic patients and the synthetic electronic health care record. *Journal of the American Medical Informatics Association*, 25(3):230–238.

Michael Wornow, Yizhe Xu, Rahul Thapa, Birju Patel, Ethan Steinberg, Scott Fleming, Michael A Pfeffer, Jason Fries, and Nigam H Shah. 2023. The shaky foundations of large language models and foundation models for electronic health records. *npj Digital Medicine*, 6(1):135.