Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Local-First Clinical Text Structuring with Fine-Tuned MedGemma for Readmission Risk Assessment

Serhii Zabolotnii

zabolotnii.serhii@csbc.edu.ua

Cherkasy State Business College

Viktoriia Holinko

healthPrecision

Research Article

Keywords: clinical NLP, electronic health records, structured extraction, readmission prediction, MedGemma, local-first AI, fact-level tokenization

Posted Date: July 8th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10266000/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Local-First Clinical Text Structuring with Fine-Tuned MedGemma for Readmission Risk Assessment

Serhii Zabolotnii*[1] and Viktoriia Holinko†[2]

[1]Cherkasy State Business College, Cherkasy, Ukraine
[2]healthPrecision

Abstract

Background. Unstructured clinical notes remain a bottleneck for deployable healthcare AI; cloud-dependent pipelines raise privacy and infrastructure barriers.

Methods. We present **MedGemma StructCore**, a local-first two-stage extraction pipeline using compact MedGemma 4B models. Stage 1 applies Schema-Guided Reasoning to summarize notes into structured JSON across nine clinical clusters. Stage 2 projects summaries into canonical four-field KVT4 facts (**K**eyword–**V**alue–**T**imestamp tuples scoped by Cluster, i.e. `Cluster|Keyword|Value|Timestamp`) via a LoRA-adapted model. Deterministic normalization, a signal-integrity gate, and offline hybrid regeneration audit and reduce silent objective signal-loss between stages. Prompt KV-cache reuse yields a +10.6% speedup with bit-exact output [VERIFIED].

Results. On MIMIC-IV (N=50,000; patient-level split; N_{test}=9,857), a tabular baseline (A4) trained on the *full* cohort (N_{train}≈35,000) attains AUROC 0.685 (95% CI 0.670–0.699) [VERIFIED]—the task ceiling of tabular features under full supervision, not the comparator for our extraction contribution. That contribution is isolated only in a *matched-training* regime (N_{train}=1,500), where $A3_{factlevel}$ reaches AUROC 0.659 against 0.656 for the fairly refit baseline (XGBoost does not beat logistic regression on the same features). In that regime, $A3_{factlevel}$ yields verified AUPRC and Brier gains [VERIFIED], while the AUROC uplift is small (+0.0033) and not confirmed [PRELIMINARY]. A post-closure typed-fusion branch shows a single CI-backed gain—AUPRC +0.021 (95% CI [0.0018, 0.0395]) over the A4 arm—at near-parity with $A3_{factlevel}$; its larger deltas over the weak notes-only Stage 2 baseline reflect that baseline's weakness and are not headlined. KVT4 format validity is 99.74%; a signal-integrity audit (N=4,000) finds 15.55% doc-level objective loss, reduced to 8.48% by offline hybrid regeneration without extra LLM calls. Structured-reference validation covers a large LABS benchmark (full test split, micro-F1 ≈0.48) and a preliminary VITALS benchmark; beyond a 10-document curated set, the semantic clusters (PROBLEMS/SYMPTOMS/MEDICATIONS/PROCEDURES) have no quantitative extraction ground truth. A GPT-4.1-mini scaling pilot gives directional evidence—consistent with narrative-vs-table reference mismatch rather than model capacity—that LABS micro-F1 plateaus at ≈0.52 [PRELIMINARY, N=200].

Conclusion. The primary contribution is an auditable, format-stable local-first text-structuring pipeline on consumer hardware; "reliability" here means output-contract stability and signal-integrity auditing, not yet fully validated per-cluster extraction accuracy. Typed downstream fusion is a promising but not dominant continuation direction. Because predicted probabilities are over-confident in higher-risk bins (A3 calibration slope 0.741), pre-deployment recalibration is a prerequisite for any probability-based clinical use. Overall, notes-derived KVT4 facts add useful predictive signal, but stronger extraction-quality and fairness claims still require further validation.

*Corresponding author: `zabolotnii.serhii@csbc.edu.ua`; ORCID 0000-0003-0242-2234. Address: Chornovola 243, Cherkasy, Ukraine.

†ORCID 0009-0001-1882-7503.

Keywords: clinical NLP, electronic health records, structured extraction, readmission prediction, MedGemma, local-first AI, fact-level tokenization

1 Introduction

1.1 Clinical Context and Motivation

Discharge from the hospital is one of the most consequential moments in the clinical process. Clinical teams must assess readmission risks, plan follow-up care, and ensure reliable information transfer between levels of care. Approximately 20% of hospitalized patients return to the hospital within 30 days of discharge, imposing a substantial burden through increased mortality, reduced quality of life, and significant financial costs [1]. In the US Medicare system, the Hospital Readmissions Reduction Program (HRRP) penalizes hospitals with high readmission rates, further motivating the development of accurate predictive models [2].

Key risk signals—abnormal laboratory values, vital-sign instability, multiple comorbidities, polypharmacy—are often scattered across free-text clinical notes (discharge summaries, progress notes, operative reports, nursing notes) rather than in structured EHR rows and columns. These notes are written under time pressure, have variable structure, and are difficult to convert rapidly into quantitative features for predictive models.

1.2 The Problem of Unstructured Clinical Text

Despite decades of clinical NLP development, transforming free clinical text into structured, operationally useful data remains an open challenge. Clinical NLP has progressed from rule-based systems to transformer models and, most recently, to large language models [3].

Traditional IE systems such as cTAKES [4], MetaMap [5], and MedCAT [6] are optimized for *detecting individual medical concepts*, not for *comprehensive structured extraction*—converting an entire clinical document into a set of canonical facts with a predefined schema.

Large language models (LLMs) have substantially improved clinical text understanding. GPT-4 with an elaborate prompting framework demonstrated a 35% AUROC improvement over baseline prompts on EHR prediction tasks, including 30-day readmission on MIMIC-IV [7]. However, practical deployment faces critical barriers:

- **Privacy:** Cloud inference requires transmitting sensitive patient data. Even in research, some models (Gemini Pro) were not tested on MIMIC-IV "due to privacy issues" [7]. Federated approaches [8] mitigate but do not eliminate this barrier; on-device inference [9] remains the strongest privacy guarantee.
- **Format instability:** LLM outputs are non-deterministic—the same model can generate different formats for identical inputs, making downstream processing fragile.
- **Infrastructure access:** Not all clinical institutions have access to cloud APIs or GPU clusters.
- **Auditability:** Black-box proprietary models complicate clinical validation.

1.3 Our Thesis

We posit that *reliable healthcare AI begins with a robust text-to-structure layer.* Using **Schema-Guided Reasoning (SGR)** [10], we extend the clinical IE paradigm by converting the "black box" of generation into a controlled process with a stable, auditable output contract. We adopt SGR primarily as a *format-stability* mechanism for local models, where it compensates for reduced capacity through explicit schema guidance; we do *not* claim it as an independent driver of downstream accuracy—our own matched ablation shows a less constrained Stage 1 profile can match or exceed schema-guided Stage 1 on downstream AUROC (Sections 4.9 and 5.4).

Rather than building another "best predictive model," we focus on creating **reliable structured extraction infrastructure** that:

1. Runs locally on consumer hardware (Apple Silicon, standard GPUs)—critical for privacy and reducing dependency on external services.
2. Provides a stable, auditable output format (KVT4)—unlike the non-deterministic outputs of general-purpose LLMs.
3. Uses domain-adapted open-weight medical language models (fine-tuned MedGemma).
4. Can be reused for diverse clinical tasks, not limited to a single prediction endpoint.

30-day readmission prediction serves as a lighthouse demonstration—one concrete downstream task that validates the extraction layer and illustrates the utility of KVT4 facts for downstream ML models, rather than being the primary research goal.

1.4 Contributions

We present **MedGemma StructCore**—a system for local structured extraction from free-text EHR notes. Our main contributions are:

1. **Two-stage local extraction pipeline (StructCore).** Stage 1 applies Schema-Guided Reasoning to summarize notes into structured JSON across 9 clinical clusters. Stage 2 projects summaries into canonical **KVT4** facts—four-field **K**eyword–**V**alue–**T**imestamp tuples scoped by Cluster (`Cluster|Keyword|Value|Timestamp`)—via a LoRA-adapted model. A signal-integrity gate and deterministic offline hybrid regeneration audit and recover silent objective data loss between stages (15.55% → 8.48% doc-level on N=4,000). The deployed inference pipeline runs entirely on consumer hardware (GGUF Q5_K_M via llama.cpp, Apple Silicon / ≥8 GB VRAM) with no cloud dependency at inference time.
2. **KVT4 output contract with strict normalization.** We define an ontology of 9 clusters with canonical keywords, permitted values, and deduplication rules, ensuring reproducibility and auditability—unlike end-to-end embedding approaches [11] where structuring is implicit. Every metric is labeled VERIFIED, PRELIMINARY, or PLANNED, with traceable artifact links and a full audit trail.
3. **Fact-level KVT tokenization and downstream validation.** Instead of aggregating facts into fixed features, we propose sparse hashed tokenization at the individual fact level. On MIMIC-IV (N_{test}=9,857) in a matched-training regime, $A3_{factlevel}$ improves AUPRC and Brier over a fair tabular refit baseline with demographic covariates [VERIFIED], while the AUROC uplift is small [PRELIMINARY]. A typed downstream-fusion branch combining four semantic probabilities yields one CI-backed AUPRC gain over the canonical tabular baseline while remaining near-parity with $A3_{factlevel}$; we report it as a promising exploratory continuation direction, not a dominant arm.
4. **Structured-reference extraction validation and model scaling.** Beyond a curated 10-document set (N=10, PRELIMINARY), we validate LABS extraction against MIMIC structured tables at scale (N=9,857) and provide a preliminary VITALS benchmark with chartevents reference aligned to HL7 FHIR [12] observation codes. This is structured-reference (table-based) validation: the semantic clusters (PROBLEMS, SYMPTOMS, MEDICATIONS, PROCEDURES) are *not* yet quantitatively evaluated, and clean extraction ground truth beyond the 10 curated documents remains an open gap (Section 5.4). A model-scaling pilot (GPT-4.1-mini vs. MedGemma 4B) provides directional evidence that moderate per-keyword LABS F1 reflects narrative-vs-table reference mismatch rather than model capacity, and shows the pipeline supports pluggable Stage 1 replacement [PRELIMINARY, N=200].
5. **Comprehensive benchmark with conservative fairness reporting.** We compare against LACE [13], HOSPITAL [14], logistic regression, and XGBoost baselines. A first subgroup analysis across sex, age, and broad race categories is reported as PRELIMINARY supporting evidence.

1.5 Paper Organization

The remainder of this paper is organized as follows. Section 2 reviews related work in clinical NLP, LLMs in medicine, and readmission prediction. Section 3 describes the data, pipeline architecture, downstream task, and evaluation protocol. Section 4 presents extraction reliability and readmission benchmark results. Section 5 discusses clinical implications, limitations, and future work. Section 6 addresses ethical considerations and reproducibility.

2 Related Work

2.1 Evolution of Clinical NLP

2.1.1 Rule-Based and Dictionary Systems

The first generation of clinical NLP systems relied on medical terminologies, rules, and statistical models. **cTAKES** [4]—an open-source system based on Apache UIMA—performs NER, negation detection, and clinical concept classification with mapping to SNOMED CT and RxNorm. **MetaMap** [5] is an NLM tool that maps biomedical text to UMLS concepts through lexical variants and syntactic analysis. **MedCAT** [6] combines neural approaches with controlled vocabularies and demonstrated effectiveness on millions of NHS records. Their shared limitation is optimization for *individual concept detection* rather than *comprehensive document-level structured extraction*.

2.1.2 Deep Learning and Transformers

The limitations of rule-based systems motivated the transition to deep learning. Clinical NLP shared tasks—including the i2b2/VA challenges [15] and the n2c2 series [16]—catalyzed progress on concept extraction, relation classification, and medication information tasks. With the adoption of representation learning and transformers, neural models improved performance across clinical NLP tasks. **ClinicalBERT** [17]—a BERT adaptation for the clinical domain—demonstrated significant improvement in readmission prediction tasks compared to TF-IDF and traditional ML approaches. **Med-BERT** [18] pretrained on structured EHR codes from 28M patients showed gains in disease prediction, while **RETAIN** [19] introduced reverse-time attention for interpretable sequential EHR modeling.

2.1.3 Large Language Models in Clinical NLP

General-purpose LLMs have changed the landscape of clinical NLP while introducing new challenges. Zhu et al. [7] systematically studied LLM adaptation to structured longitudinal EHR in zero-shot mode: GPT-4 with an elaborate prompting framework showed ∼35% AUROC improvement on MIMIC-IV mortality prediction and outperformed traditional ML models in few-shot (10-example) scenarios.

Domain-adapted medical LLMs such as Med-PaLM 2 [20] reached expert-level performance on medical QA benchmarks. **MedGemma** [21]—a family of medical models from Google HAI-DEF based on Gemma [22]—combines domain adaptation with the advantages of open weights: local deployment, fine-tuning, and full control.

Pandey et al. [23] proposed a hybrid model using **ClinicalT5** combined with structured EHR data for 30-day readmission prediction on MIMIC-IV, confirming the importance of integrating both modalities.

Ferrazzi et al. [24] systematically benchmarked adaptation strategies for ∼1B-parameter LLMs (Gemma-3 1B, Llama-3.2-1B, Qwen3-1.7B) on five Italian medical NLP tasks. LoRA fine-tuning was the dominant strategy: a 1.7B model surpassed Qwen3-32B by +9.2 F1 on average ($p < 0.05$), while continual pretraining on 405M words of medical text rarely improved over

fine-tuning alone. These findings independently support our architectural choice of a compact LoRA-adapted MedGemma 4B over larger cloud-scale models.

2.1.4 Schema-Guided Reasoning for Local Models

Schema-Guided Reasoning (SGR), formalized by Abdullin [10] in a non-peer-reviewed technical report, proposes a technique for enforcing reasoning through predefined steps (schemas). Unlike simple constrained decoding that merely restricts output to JSON format, SGR uses *Cascade* (sequential reasoning), *Routing* (path selection), and *Cycle* (iterative refinement) patterns described via Pydantic schemas. This approach is intended for local models (4B–7B parameters), which have smaller "cognitive bandwidth" than cloud-scale models. That source reports SGR enabling 5–10% higher reliability than standard unstructured generation [10]; we cite this as a claim of the original source rather than as an independently reproduced result, and note that it has not undergone peer review. In our own matched Stage 1 ablation (Section 4.9), schema guidance does *not* improve downstream AUROC over a less constrained profile, so we adopt SGR as a format-stability and auditability mechanism rather than as a downstream-accuracy gain. Our work implements SGR in Stage 1 for structured clinical summarization.

Structured decoding tools. Outlines [25], LMQL [26], Guidance [27], and SGLang [28] provide constrained decoding at the grammar level, guaranteeing valid JSON or schema-conformant output. Unlike these tools, SGR operates at the *semantic* level—not merely constraining the output format but structuring the model's reasoning process through cascading schema steps. The two approaches are complementary: constrained decoding can serve as an additional format guarantee on top of SGR.

2.1.5 LLM-Based Pipelines for Readmission Prediction

Recent work explores diverse strategies for integrating LLMs into readmission prediction pipelines, which can be classified by output type and structuring level.

Zero-shot interpretable extraction. CHiLL [29] uses natural-language queries to extract clinician-meaningful features without labeled examples. Linear models trained on automatically extracted features performed comparably to models using reference features for 30-day readmission proxies, confirming that LLMs can generate interpretable predictors without supervised fine-tuning.

Fine-tuned transformers for domain-specific extraction. Shao et al. [30] fine-tuned Flan-T5 XL/XXL for SDoH extraction in heart-failure patients, achieving macro-F1 = 0.71 and identifying 93.8% of patients with adverse SDoH (versus 2.0% captured by ICD-10 codes alone). Yang et al. [31] applied domain-adapted LLMs for cardiovascular symptom extraction with clinician agreement $\kappa = 0.82$, noting the need to address hallucination and temporal ambiguity through prompt engineering and rule-based checks.

LLM as end-to-end predictor. CPLLM [32] applies prompt-tuning with quantization for direct readmission prediction, bypassing explicit extraction. The system exceeded baselines in PR-AUC and ROC-AUC but sacrifices interpretability and auditability.

Summarization as intermediate representation. Choudhuri et al. [33] prompt LLMs to generate clinical summaries whose vectorized representations feed ICU bounceback/LOS classifiers. On MIMIC-III: +7.17% AUC-ROC for bounceback, +14.16% AUPRC. This approach is closest to our Stage 1, but without strict schema enforcement or deterministic downstream processing.

Multi-agent pipelines. ClinNoteAgents [34] decomposes note understanding into extraction, interpretation, and abstraction agents that yield structured risk factors and clinician-style abstractions. On a heart-failure cohort (3,544 notes, readmission rate 35.16%) the system reports strong performance, though precise numeric metrics are not provided in the abstract.

Compared to these approaches, MedGemma StructCore uniquely combines: (1) a compact domain-adapted model (4B, not cloud APIs); (2) explicit schema-guided structuring (KVT4 contract, not implicit embeddings); (3) fully local deployment (GGUF + llama.cpp); and (4) a signal-integrity QA gate between pipeline stages, absent from all reviewed systems.

2.1.6 Programmatic LLM Frameworks

DSPy [35] provides a programmatic framework for compiling declarative LM calls into optimized pipelines. We explored DSPy for Stage 2 extraction but found that its optimization loop (requiring ground-truth metric feedback) introduced complexity without improving extraction quality over direct LoRA fine-tuning for our constrained KVT4 output format. DSPy remains a promising direction for tasks with richer output spaces.

2.2 30-Day Readmission Prediction

A systematic review by Kansagara et al. [1] analyzed 26 readmission prediction models and found that most achieve *moderate discrimination* (AUROC 0.55–0.65).

2.2.1 Traditional Risk Scores

Two widely cited discharge risk stratification tools:

- **LACE index** [13]—four components: Length of stay, Acuity of admission, Charlson comorbidity index, Emergency department visits. Original C-statistic 0.684 on external validation [13], though subsequent studies report substantially lower values; as summarised by Pandey et al. [23] (secondary citation), a Geisinger Health System study (>100,000 patients) reported AUC of only 0.60.
- **HOSPITAL score** [14]—seven components: Hemoglobin, Oncology service, Sodium, Procedure, index admission Type, number of Admissions, Length of stay. Also reported around AUC ∼0.60 on external cohorts (as summarised by Pandey et al. [23]; secondary citation).

2.2.2 ML Approaches on Structured Data

Lo et al. [36] built models for 14-day readmission on a cohort of 24,722 patients; CatBoost showed the best average performance (AUROC 0.99, AUPRC 0.77) with 5-fold cross-validation—though this was linked to a very low baseline readmission rate (1.22%) and single-institution data.

2.2.3 Multimodal Approaches

Almeida et al. [11] proposed combining clinical notes and structured EHR as nodes of a graph neural network (GraphSAGE) for 30-day readmission prediction on MIMIC-IV, achieving **AUROC 0.72** and balanced accuracy 66.7%. Golmaei and Luo [37] similarly modeled patient interactions as graphs with note-derived features for readmission. Pandey et al. [23] showed that hybrid models (ClinicalT5 + structured data) outperform text-only approaches.

2.3 Gap in the Literature

Table 1 summarizes the positioning of our work relative to existing approaches. Work combining *compact domain-adapted LLMs + fine-tuning + local deployment + schema-guided structured extraction + downstream risk prediction* is essentially absent.

3 Methods

We report this work following TRIPOD+AI guidance for clinical prediction models and the MI-CLAIM checklist for clinical AI modeling [38, 39].

Table 1: Positioning of MedGemma StructCore relative to existing approaches.

Approach	Extraction	Deploy	Structuring	Privacy
cTAKES / MetaMap / MedCAT	Concepts	Local	Rule-based	✓
ClinicalBERT / ClinicalT5	Embeddings	GPU	Implicit	✓
GPT-4 / Med-PaLM	Zero/few-shot	Cloud	Prompt-dep.	×
GNN + LLM emb. [11]	Graph	Training	Implicit	✓
CHiLL [29]	Zero-shot NLQ	Cloud	Interpretable	×
CPLLM [32]	End-to-end	Cloud	None	×
ClinNoteAgents [34]	Multi-agent	Cloud	Partial	×
LLM Summarization [33]	Summary	Cloud	Implicit	×
StructCore (ours)	**SGR KVT4**	**Local**	**Explicit**	✓

3.1 Data and Cohorts

3.1.1 Data Source

The study uses **MIMIC-IV** (Medical Information Mart for Intensive Care, version 3.1) [40]—a publicly available de-identified EHR database from Beth Israel Deaconess Medical Center (BIDMC), Boston, MA. MIMIC-IV contains de-identified clinical data (demographics, laboratory results, medications, procedures, discharge notes, and more) for over 500,000 hospitalizations. Data access is granted through PhysioNet Credentialed Access and requires completion of CITI training.

3.1.2 Cohort Construction

The analytical cohort was formed from the `hosp/admissions.csv.gz` table (MIMIC-IV v3.1):

- **Total cap:** 50,000 hospitalizations (chronologically first by `admittime`)
- **Exclusions:** Transfers (records with overlapping stays for the same `subject_id`)

Cohort characteristics. Table 2 summarizes available demographics from `admissions.csv` (N=50,000 cohort). Age (from `anchor_age`) and sex (from `patients.csv`) were incorporated into the fair-comparison canonical rerun (Section 4.5); both the tabular baseline (A4) and the fact-level arm ($A3_{\text{factlevel}}$) include these covariates for the canonical test split evaluation.

3.1.3 Outcome Definition

30-day unplanned readmission (binary label): a patient is considered readmitted if the next hospitalization of the same `subject_id` occurs within ≤30 days of the previous discharge date (`dischtime`). **Planned readmission exclusion:** records where the *next* admission has `admission_type` ∈ {`ELECTIVE`, `SURGICAL SAME DAY ADMISSION`} are not counted as positive readmissions (N_{excluded}≈2,300). `EW EMER.`, `URGENT`, `DIRECT EMER.`, `AMBULANCE`, `OBSERVATION ADMIT`, and `DIRECT OBSERVATION` are all treated as unplanned. Transfers (overlapping stays for the same `subject_id`) were already excluded at cohort construction. **Prevalence:** 18.8% overall (9,422/50,000 admissions); ∼18.8% in the test split.

Sensitivity analysis: We did not conduct a separate sensitivity analysis with alternative readmission definitions (e.g., all-cause including elective); this is noted as a limitation (Section 5.4).

3.1.4 Splitting Strategy

Train/validation/test splits are performed at the **patient level** (by `subject_id`), not at the hospitalization level, to avoid data leakage.

Table 2: Cohort demographics (N=50,000 hospitalizations).

Variable	**Category**	**N (%)**
Race/Ethnicity	White	30,889 (61.8%)
	Black/African American	6,948 (13.9%)
	Hispanic/Latino	1,629 (3.3%)
	Asian	1,416 (2.8%)
	Other/Unknown	9,118 (18.2%)
Insurance	Medicare	22,595 (45.2%)
	Private	15,754 (31.5%)
	Medicaid	9,499 (19.0%)
	Other/Missing	2,152 (4.3%)
Marital Status	Married	20,909 (41.8%)
	Single	18,648 (37.3%)
	Widowed	5,278 (10.6%)
	Divorced	3,915 (7.8%)
	Missing	1,250 (2.5%)
Language	English	44,767 (89.5%)
	Non-English	5,233 (10.5%)
30-day readmission prevalence (full cohort)		18.8% (9,422)
readmission prevalence (test split)		∼18.8%

Table 3: Data split summary.

Split	**%**	**N admissions**	**N patients**
Train	70%	35,036	—
Validation	10%	5,107	—
Test	20%	9,857	4,418

3.2 StructCore Extraction Pipeline

3.2.1 Overall Architecture

MedGemma StructCore implements a **two-stage pipeline** for converting free-text clinical documents into canonical structured facts (Figure 1):

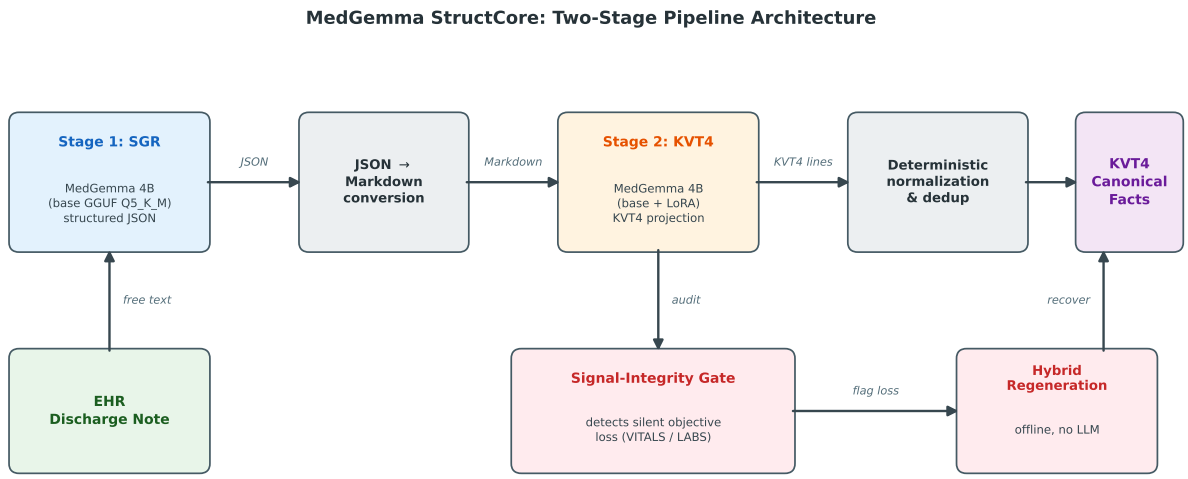


Figure 1: MedGemma StructCore two-stage pipeline architecture. Stage 1 applies Schema-Guided Reasoning (SGR) to produce structured JSON; Stage 2 projects the Markdown summary into canonical KVT4 facts via a LoRA-adapted model. A signal-integrity gate detects silent objective-data loss, and offline hybrid regeneration recovers facts without additional LLM calls. Stage 2 operates exclusively on Stage 1 output, not on raw EHR text.

Key architectural decisions:

1. Stage 2 operates *only on Stage 1 output* (not on raw EHR text), providing a controlled, filtered input.
2. Different model configurations for each stage enable independent optimization.
3. Deterministic normalization and quality gates between stages ensure format stability.

3.2.2 Stage 1: Schema-Guided Domain Summarization (SGR)

Model: MedGemma 4B (base GGUF Q5_K_M, no LoRA).

We apply **Schema-Guided Reasoning** [10], which structures the model's "thinking" process through a rigid Pydantic schema, compensating for the smaller parameter count (4B) with explicit encoding of the expert extraction algorithm.

Task: Convert an arbitrary free-text clinical document into structured JSON with a predefined schema (9 mandatory clusters) using the *Cascade* pattern (sequential field filling).

The schema (`readmission_domain_summary_sgr_v2.schema.json`) defines 9 mandatory top-level keys:

1. `DEMOGRAPHICS` — age, sex
2. `VITALS` — heart rate, BP, respiratory rate, temperature, SpO_2, weight
3. `LABS` — Hb, Hct, WBC, Na, K, Cr, BUN, Glucose, Bicarb, Plt
4. `PROBLEMS` — diagnoses and conditions
5. `SYMPTOMS` — symptoms and complaints
6. `MEDICATIONS` — integral medication flags
7. `PROCEDURES` — procedures and interventions

8. `UTILIZATION` — healthcare utilization history
9. `DISPOSITION` — discharge status and destination

Text processing strategy: Input condensation via head + tail + "keyword windows" around clinically significant terms (WBC, Na, Cr, BP, SpO_2, etc.); maximum input length: ∼4,000–6,000 characters; generation temperature: 0.0 (deterministic).

3.2.3 Stage 2: KVT4 Projection

Model: MedGemma 4B (base GGUF Q5_K_M + LoRA adapter trained on a hard200 sample).

Training set (hard200): 200 documents selected for maximum complexity in Stage 2 projection—those with the highest frequency of cluster-switching, non-standard lexicon, and greatest format drift in zero-shot mode. Ground truth was generated by Gemini 3 Flash Thinking (teacher model) with a deterministic sanitizer pipeline applied to every teacher output before inclusion in the training set.

Teacher verification protocol. Each teacher output passed through a multi-stage sanitizer: (1) KVT4 format validation (drop lines with $\neq 4$ pipe-delimited fields); (2) keyword canonicalization (drop non-canonical keywords: 136/3,488 raw lines = 3.9% on the hard100 pilot); (3) boolean medication normalization (195 corrections); (4) procedure flag inference (65 insertions); (5) deduplication (2 drops). No formal inter-annotator agreement was computed; we acknowledge this as a limitation (Section 5.4). The mean overlap between teacher output and the existing Stage 2 extraction was precision 0.528, recall 0.585, confirming that the teacher introduces genuinely new signal rather than merely restating current model output.

LoRA configuration. Stage 2 was fine-tuned to ensure high KVT4 format consistency (see table 4).

Table 4: LoRA fine-tuning hyperparameters for Stage 2.

Hyperparameter	**Value**
Rank / α / Dropout	16 / 32 / 0.05
Target Modules	`q,v,k,o,gate,up,down_proj`
Optimizer	AdamW ($\beta_{1,2}$: 0.9, 0.999; ϵ: 10^{-8})
Learning Rate	10^{-4} (Cosine decay to 10^{-6})
Batch Size	1 (grad_accum=8, effective=8)
Regularization	weight decay=0.01, grad_clip=1.0
Epochs	3 (Seed=42)
Hardware	Apple M3 Max (48 GB)
Training Time	≈2.5 hours

Task: Project Stage 1 Markdown into canonical KVT4 facts.

KVT4 format (`Cluster|Keyword|Value|Timestamp`):

```
VITALS|Heart Rate|91|Admission
LABS|Sodium|138|Discharge
DISPOSITION|Discharge Disposition|Home with Services|Discharge
PROBLEMS|Hypertension|chronic|Past
```

Output contract (STRICT):

- Exactly 4 fields separated by `|`
- `Cluster` ∈ {9 permitted clusters}
- `Timestamp` ∈ {Past, Admission, Discharge, Unknown}
- For VITALS/LABS/UTILIZATION: Value = numeric only (no units)
- Deduplication: at most 1 fact per (`Cluster`, `Keyword`) for objective clusters

3.2.4 Normalization and Quality Gates

After Stage 2, deterministic normalization is applied:

1. **Keyword canonicalization:** mapping to strict keyword sets (7 VITALS, 10 LABS, 4 UTILIZATION, 2 DISPOSITION).
2. **Numeric normalization:** unit stripping (mg/dL, %, bpm), BP splitting (120/80 → SBP=120, DBP=80).
3. **Timestamp normalization:** mapping to canonical values.
4. **Plausibility filter:** soft checks for physiological ranges (HR: 30–220, SpO_2: 50–100, Na: 100–180, etc.).
5. **Deduplication:** full duplicate removal; for objective clusters—one fact per keyword (prefer Discharge > Admission).

3.3 KVT4 Ontology

The ontology defines the *canonical contract* for structured extraction.

Objective clusters (canonical keywords, strict):

- **VITALS** (7 keywords, numeric): Heart Rate, Systolic BP, Diastolic BP, Respiratory Rate, Temperature, SpO_2, Weight
- **LABS** (10 keywords, numeric): Hemoglobin, Hematocrit, WBC, Platelet, Sodium, Potassium, Creatinine, BUN, Glucose, Bicarbonate
- **UTILIZATION** (4 keywords, numeric): Prior Admissions 12mo, ED Visits 6mo, Days Since Last Admission, Current Length of Stay
- **DISPOSITION** (2 keywords, allowlist): Discharge Disposition (Home | Home with Services | SNF | Rehab | LTAC | Hospice | AMA), Mental Status (alert | confused | oriented | lethargic)
- **DEMOGRAPHICS** (2 keywords): Age (numeric), Sex (male | female)

Semantic clusters (evidence-only policy):

- **PROBLEMS**: Keyword = diagnosis/condition name; Value ∈ {chronic, acute, exist, not exist}
- **SYMPTOMS**: Keyword = symptom name; Value ∈ {yes, no, severe}
- **MEDICATIONS**: Integral flags (Polypharmacy, Anticoagulation, Insulin Therapy, etc.)—`yes` only with explicit evidence
- **PROCEDURES**: Integral flags (Any Procedure, Surgery, Dialysis, Mechanical Ventilation)

3.4 Downstream Task: Readmission Risk Prediction

3.4.1 Experimental Arms

To enable comparative evaluation, we define several experimental arms differing in feature source (Table 5):

3.4.2 Predictors and Missing Data Handling

For the tabular baseline (A4), predictors include length of stay, 10 CCDE-style laboratory features (Hemoglobin, Hematocrit, WBC, Platelet, Sodium, Potassium, Creatinine, BUN, Glucose, Bicarbonate), simple utilization history features (prior admissions and gap time), and hashed categorical admission metadata (e.g., admission type, discharge location, insurance, language, marital status, race).

Logistic regression configuration (A4): SGDClassifier (sklearn) with `loss=log_loss`, $\alpha = 10^{-5}$, no explicit regularization penalty (`penalty=None`), `max_iter=1000`, `random_state=42`. Hyperparameters were selected via grid search on the validation split (N_{val}=5,107) with AUROC as the selection criterion. Final model is retrained on train+val before test evaluation.

Table 5: Experimental arms for readmission prediction.

Arm	Feature Source	Model	Description
A0	admissions.csv	Hash+LogReg	Metadata baseline
A1	admissions + labevents	Rule Engine	Labs-based (215-pt max)
A4	admissions + labs + util.	Hash+LogReg (tuned)	Tabular ML
A2	discharge notes	Two-stage + LogReg	Notes-only (15 aggregates) ($N_{notes} \leq 50$)*
A3	admissions + labs + notes	Hybrid Hash+LogReg	Tabular + 15 note aggr.
$A3_{fact}$	admissions + labs + notes (fact tokens)	Hybrid Hash+LogReg (AdaGrad)	Tabular + fact-level KVT sparse tokens
$A4_{XGB}$	admissions + labs + util.	XGBoost	Gradient-boosted tabular baseline
LACE	admissions.csv	Integer score	L + A (partial proxy)
HOSPITAL	admissions + labs	Integer score	H+S+I+A (partial proxy)

$\star$ A2 canonical-set evaluation limited to N=50 due to extraction throughput constraints (∼25 s/doc on Apple M2).

Gradient-boosted baseline ($A4_{XGB}$): To test whether a more powerful estimator could extract additional signal from the same feature set, we trained an XGBoost classifier [41] on the identical A4 features (labs, LOS, utilization, age, sex, label-encoded categoricals). Grid search over `max_depth` ∈ {3,5,7}, `min_child_weight` ∈ {1,5,10}, `subsample` ∈ {0.8,1.0}, `colsample_bytree` ∈ {0.8,1.0} (36 trials, early stopping on validation Brier, learning rate 0.05, `scale_pos_weight` for class imbalance). Final model refitted on train+val with best hyperparameters.

Missing data handling: Missing laboratory values are handled without imputation: for each lab we include a binary missingness indicator and set the corresponding numeric feature to zero. Other missing numeric predictors are encoded as zero. For notes-derived arms, absence of a fact results in the corresponding sparse token being absent (zero feature value).

3.4.3 ReadmissionRiskEngine

For arms A1 and A2, we use a rule-based scoring engine with 9 sub-clusters: vital instability (tachycardia, hypotension, tachypnea, hyperthermia, hypoxia), lab instability (hyponatremia, hyperkalemia, azotemia, hyperglycemia, anemia), discharge risk (non-Home disposition, altered mental status), and utilization (prior admissions, LOS). Maximum total score: 215 points. Score-to-probability calibration is performed via logistic calibration (α, β parameters), following the framework of Van Calster et al. [42]. Calibration assessment uses the methodology of Niculescu-Mizil and Caruana [43].

3.5 Signal-Integrity QA and Offline Hybrid Regeneration

3.5.1 Silent Signal-Loss in Two-Stage Pipelines

In a two-stage extraction pipeline, objective data (VITALS/LABS) can be silently lost between stages:

- **Generation loss:** Stage 2 fails to reproduce numeric facts present in Stage 1 Markdown (e.g., due to format drift—the model generates JSON or fenced code instead of KVT4 lines).
- **Postprocess drop:** The normalizer/deduplicator discards facts that do not conform to the strict KVT4 contract.

On a subset of N=4,000 documents: 3,196 contain numeric objective data in Stage 1; 497/3,196 (15.55%) exhibit any objective loss in final facts; key-level loss: 4,344/35,477 (12.24%).

Decomposition (not mutually exclusive): 299 documents with generation loss, 225 with postprocess drop; since $299 + 225 = 524 > 497$, 27 documents exhibit both loss types.

3.5.2 Signal-Integrity Gate

We implement an automated signal-integrity gate that: (1) parses Stage 1 output for numeric keys (VITALS/LABS), (2) compares against final `stage2_facts.txt`, (3) classifies each loss as generation or postprocess, and (4) produces a JSON report with per-document details and rerun candidates.

3.5.3 Offline Hybrid Regeneration

Deterministic hybrid recovery without additional LLM calls: (1) robust re-parse of `stage2_raw.txt` (JSON recovery, fenced-code extraction), (2) supplementation from `stage1.md` (MEDICATIONS/PROBLEMS aggregates + VITALS/LABS numeric keys when absent from Stage 2), (3) deduplication and normalization per the KVT4 contract.

Result on N=4,000: missing documents (hybrid): 271/3,196 (was 497); missing keys (hybrid): 1,862/35,477 (was 4,344).

3.6 Temporal Leakage Safeguards

The two-stage architecture provides a structural guarantee against temporal leakage:

- **Stage 1 input** consists exclusively of the discharge summary text—a document written at or before discharge. No post-discharge data (e.g., subsequent lab results, follow-up notes, or readmission records) enters Stage 1.
- **Stage 2 input** consists exclusively of the Stage 1 Markdown summary. Stage 2 never observes the raw EHR text or any data outside Stage 1 output.
- **Label construction** uses only `admittime` and `dischtime` from `admissions.csv`. The readmission label is computed from the *next* admission of the same `subject_id`; no information from the index admission's features can reverse-engineer this label.
- **Train/val/test splits** are at the patient level (`subject_id`), so no encounter from a training patient appears in the test set.

3.7 Fact-Level Sparse Tokenization

Instead of aggregating KVT facts via a rule-based scoring engine (9-cluster scoring → 15 numeric features), the $\text{A3}_{\text{factlevel}}$ arm uses **fact-level hashed tokenization**:

1. Each KVT fact (`Cluster|Keyword|Value`) is converted into one or more text tokens.
2. Tokens are hashed into a sparse feature vector via feature hashing (separate hash-block for notes-derived tokens vs. tabular features).
3. Filtering: `min_df=3` (minimum training frequency); DISPOSITION downweight (`scale=0.0`). The scale was selected by grid search on the validation split (`scale` $\in \{0.0, 0.1, 0.5, 1.0\}$; `scale=0.0` yielded the highest AUROC on val). Notably, the search drove the weight on DISPOSITION—clinically one of the strongest readmission predictors—to zero, indicating that the *extracted* discharge-disposition facts were too noisy in this regime to add downstream value. We candidly treat this as evidence of limited extraction reliability for this cluster rather than a tuning artifact, and discuss its implications in Section 5.4.
4. Combined with A4 tabular features (metadata + labs + utilization) in a single feature vector.
5. Optimization: AdaGrad (adaptive gradient) for better handling of sparse features compared to SGD.

This allows the downstream model to learn from *individual clinical facts* (e.g., `PROBLEMS|Hypertension|chronic`, `MEDICATIONS|Polypharmacy|yes`) rather than compressed aggregates.

3.8 Typed Downstream Fusion

As a post-closure continuation, we tested whether typed semantic signals could complement the fact-level features without modifying the free-form Stage 2 generator.

Typed labels. Four binary clinical phenotypes were defined from Stage 2 KVT4 facts via deterministic keyword-matching rules: *Respiratory Distress*, *Heart Failure*, *Acute Kidney Injury*, and *Sepsis/Severe Infection*. Each label is derived from combinations of PROBLEMS, SYMPTOMS, LABS, and VITALS keywords in the extracted facts.

Probability estimation. For each typed label, out-of-fold probability estimates were obtained via stratified K-fold TF-IDF + logistic regression on the structured PROBLEMS and SYMPTOMS fields of the training set. Validation and test probabilities were generated by the fold ensemble.

Fusion model. The four typed probabilities were concatenated with the existing A4 tabular features and the `ReadmissionRiskEngine` score, yielding a ∼40-dimensional dense vector. A Random Forest classifier (`max_depth`=5, `min_samples_leaf`=5, `max_features`=sqrt; selected by grid search on val Brier) produced the final readmission probability.

3.9 Evaluation Protocol

3.9.1 Prediction Metrics

- **AUROC** — discrimination (primary metric)
- **AUPRC** — discrimination accounting for class imbalance
- **Brier score** — overall probabilistic prediction quality

3.9.2 Bootstrap Confidence Intervals

Bootstrap resampling on the test set: K=2,000 resamples, random seed 42, 95% confidence level. Resampling is **grouped by `subject_id`** so that all encounters of the same patient appear in the same bootstrap draw (avoids within-patient information leakage).

3.9.3 Paired Subject-Bootstrap for Arm Comparison

To compare arms A and B on a shared test set, we compute $\Delta\text{AUROC} = \text{AUROC}_A - \text{AUROC}_B$ on each of K=2,000 subject-grouped bootstrap resamples and report the 95% percentile CI. ΔAUPRC and ΔBrier are computed analogously.

3.9.4 Primary vs. Exploratory Analyses and Multiplicity

To control the interpretation of multiple comparisons, we designate a small set of *primary* contrasts and treat all others as *exploratory (hypothesis-generating)*. The primary contrast is $\text{A3}_{\text{factlevel}}$ versus the matched-train tabular refit baseline on the full canonical test split (Table 15), for the three pre-specified metrics (AUROC, AUPRC, Brier). All remaining contrasts—subset-level paired deltas, typed-fusion deltas, signal-integrity downstream deltas, and every subgroup contrast (sex × age × broad race category)—are exploratory. We did *not* apply a formal multiplicity correction (e.g., Bonferroni or FDR) across the full family of bootstrap comparisons; given the number of exploratory tests, individual intervals that only marginally exclude zero should be read as hypothesis-generating rather than confirmatory. Accordingly, throughout the paper we reserve the VERIFIED label for primary, pre-specified contrasts whose CIs exclude zero, and flag borderline exploratory intervals explicitly where they appear.

3.9.5 Signal-Integrity Metrics

In addition to prediction metrics, we report:

- Missing-document rate: % of documents where Stage 1 had objective data but final facts contain none.
- Missing-key rate: % of individual VITALS/LABS keys lost between Stage 1 and final facts.
- Generation-loss vs. postprocess-drop decomposition.
- Post-hybrid rates (after offline regeneration): both document-level and key-level.

3.9.6 Extraction Quality Metrics

- **Format validity rate** — % of Stage 2 outputs conforming to the KVT4 contract
- **Parse success rate** — % of documents yielding ≥1 valid fact
- **Per-cluster Precision / Recall / F1** — extraction accuracy by cluster (against ground truth)

3.9.7 Evidence Discipline

Each metric is labeled with one of: VERIFIED (confirmed by committed artifacts and accepted protocol), PRELIMINARY (requires larger sample or replication), or PLANNED (not yet executed).

3.10 Deployment Architecture

- **Inference backend:** OpenAI-compatible local server (llama.cpp / llama-server, commit `a4ea7a18`, 2026-02-05) [44]
- **Model format:** GGUF (Q5_K_M quantization)
- **LoRA adapters:** dynamic loading/toggling via server API
- **GPU backend:** Apple Silicon MPS (primary), CUDA (secondary), CPU (fallback)

3.10.1 Stage 2 Prompt Caching (CAG)

To accelerate Stage 2 inference without affecting output quality, we employ **prompt KV cache reuse** in llama.cpp:

- **Mechanism:** `llama-server -cache-prompt -cache-reuse 256` — the KV cache for the stable prompt prefix (system prompt with ontology and extraction instructions) is reused across requests; only the per-document user prompt is freshly computed.
- **Cache-hit constraint:** the prompt prefix must be *byte-for-byte stable*; any dynamic fields (dates, run-IDs, randomized examples) would invalidate cache hits.
- **Validation (Verified):** A/B test on 200 new test-split documents (excluding the existing 3,200 subset): **+10.6% speedup** (6,297 s → 5,629 s) with **zero diffs** in `stage2_raw.txt`, `stage2_facts.txt`, `stage2_normalized.json`, and `stage2_metrics.json`.
- **Scale-out:** using the same CAG configuration, the rapid-iteration subset expanded from 3,200 to 4,200 documents (train/val/test = 1,000/200/3,000), and subsequent canonical scale-out completed extraction for 11,757 documents (train/val/test = 1,500/400/9,857), enabling evaluation on the full canonical test split (Section 4.5).

Table 6: Hardware requirements.

Component	Minimum	Recommended
RAM	8 GB	16+ GB
GPU VRAM	— (MPS shared)	8+ GB (CUDA)
Disk	∼5 GB (model)	∼10 GB
OS	macOS 13+ / Linux	macOS 14+ (Apple Silicon)

4 Results

4.1 Extraction Reliability

Table 7 summarizes the extraction quality metrics.

Table 7: Extraction reliability metrics.

Metric	Value	Scope	Status
Stage 2 KVT4 format validity	99.74%	—	Verified
Stage 1 JSON parse rate	98%+	test50	Verified
Single-document extraction F1	81.93%	1 doc	Preliminary
Curated10 VITALS+LABS micro-F1	0.984	10 docs (Adm.)	Preliminary
Curated10 VITALS+LABS Recall	1.000	10 docs	Preliminary
Curated10 VITALS+LABS Precision	0.968	10 docs	Preliminary
LABS keyword-only F1 (after canon.)	0.994	curated10	Preliminary

Format validity and parse rate are large-scale VERIFIED metrics. All single-/10-document extraction scores are PRELIMINARY (N≤10).

The two-stage pipeline achieves near-perfect format validity (99.74%), meaning that almost all Stage 2 outputs conform to the strict KVT4 4-field pipe-delimited contract. On the curated 10-document set, VITALS+LABS extraction reaches micro-F1 of 0.984 with perfect recall (1.000) and high precision (0.968); because this sample is only N=10, these figures are PRELIMINARY and are reported as an upper-bound illustration, not an extraction-quality guarantee. Beyond the small curated set, we validated extraction quality against MIMIC structured references at scale.

LABS (full canonical test, N=9,857): The structured-reference benchmark reaches coverage 97.92%, micro-F1 0.4809, macro-F1 0.4724 (PRELIMINARY). Table 8 shows per-keyword breakdown. Performance varies widely: Sodium achieves F1=0.837 (electrolyte values are well-represented in discharge text), while Platelet reaches only F1=0.201 (low recall suggests the model under-extracts platelet counts). The principal failure modes are: (1) value mismatch due to unit conversion or rounding; (2) false negatives where the lab is mentioned in text but not extracted; (3) reference sparsity where the structured table contains a value not present in the discharge summary.

Table 8: Per-keyword LABS extraction (structured reference, N=9,857; 10% numeric tolerance). All results PRELIMINARY.

Keyword	Evaluable	Precision	Recall	F1
Sodium	8,118	0.825	0.850	**0.837**
Hematocrit	8,256	0.561	0.586	0.573
Potassium	8,151	0.547	0.563	0.555
Bicarbonate	8,036	0.547	0.556	0.552
Hemoglobin	8,164	0.529	0.568	0.548
Creatinine	8,115	0.396	0.410	0.403
Glucose	8,061	0.361	0.361	0.361
WBC	8,175	0.338	0.360	0.349
BUN	8,084	0.342	0.349	0.345
Platelet	8,112	0.327	0.145	0.201
Micro-averaged		0.488	0.474	0.481

Figure 2 visualizes the per-keyword breakdown and overlays the GPT-4.1-mini scaling pilot results (see Section 4.10).

VITALS (coverage-aware, N=65 admissions with chartevents reference): Table 9 shows per-keyword performance. SpO_2 (F1=0.793) and Temperature (F1=0.777) are well-extracted; Heart Rate and Respiratory Rate are moderate; Weight remains weak (F1=0.065).

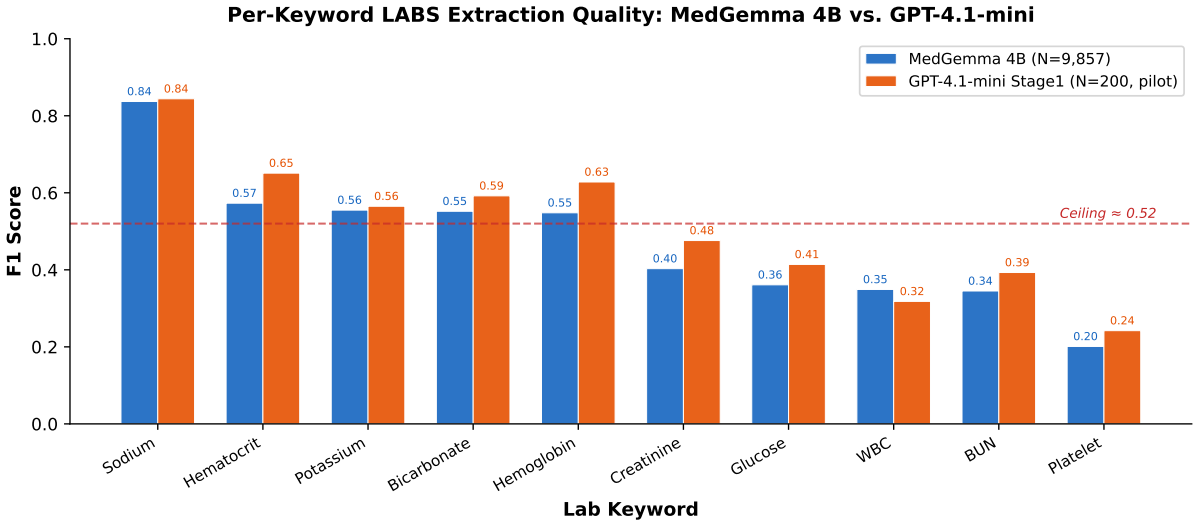


Figure 2: Per-keyword LABS extraction F1: MedGemma 4B (canonical N=9,857) vs. GPT-4.1-mini Stage 1 (pilot N=200). The dashed line marks the approximate plateau (≈0.52) at which the stronger Stage 1 model saturates, giving directional evidence (PRELIMINARY, N=200) consistent with narrative-vs-table reference mismatch rather than a model-capacity bottleneck.

These VITALS results remain supporting evidence rather than a headline extraction-quality claim (PRELIMINARY).

Table 9: Per-keyword VITALS extraction (coverage-aware chartevents reference, N=65 admissions). All results PRELIMINARY.

Keyword	Precision	Recall	F1
SpO_2	0.902	0.708	**0.793**
Temperature	0.839	0.723	0.777
Diastolic BP	0.429	0.415	0.422
Systolic BP	0.413	0.400	0.406
Heart Rate	0.290	0.277	0.283
Respiratory Rate	0.242	0.231	0.236
Weight	0.120	0.045	0.065
Micro-averaged	0.477	0.398	0.434

4.2 Track B Readmission Benchmark

Table 10 presents the main benchmark results on the MIMIC-IV test set (N=9,857 admissions, 4,418 unique patients).

The tabular ML arm (A4), which combines metadata, CCDE lab features, and utilization history, achieves the highest AUROC of 0.685—outperforming the labs-based rule engine (A1: 0.602), metadata-only baseline (A0: 0.584), and both traditional proxy scores (HOSPITAL: 0.581, LACE: 0.557). Importantly, the A4 arm is trained on the *full* training cohort (N_{train}≈35,000) and therefore represents the task ceiling of tabular features under full supervision; it is *not* the baseline against which the notes-derived contribution is measured. The fair, matched-training comparison (N_{train}=1,500) that isolates the effect of KVT4 facts is reported separately in Section 4.5 ($A4_{LR}$ 0.656 vs. $A3_{factlevel}$ 0.659); the 0.685-vs-0.656 gap is driven by training-set size, not by the choice of features or method. Figure 3 visualizes the comparison across all arms

Table 10: Track B readmission benchmark results (N_{test}=9,857). CI: 95% bootstrap (K=2,000, seed=42).

Arm	Description	AUROC	95% CI	AUPRC	Brier
A4	Tabular ML (hash-logreg)	**0.685**	[0.670, 0.699]	0.346	0.142
A1	Rule Engine (labs)	0.602	[0.588, 0.617]	0.248	0.158
A0	Metadata LogReg	0.584	[0.570, 0.598]	0.243	0.152
HOSPITAL	Proxy (partial)	0.581	[0.567, 0.595]	0.225	—
LACE	Proxy (partial)	0.557	[0.542, 0.571]	0.218	—
A2†	Notes-only (N=50)	0.462	[0.296, 0.625]	0.232	0.202

† Preliminary; small N, wide CI.

and metrics.

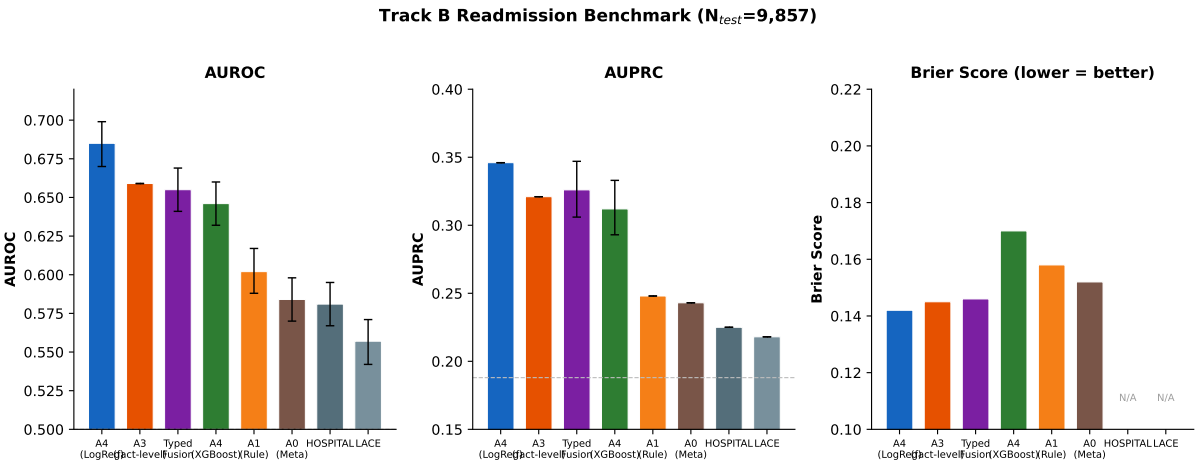


Figure 3: Track B readmission benchmark comparison across all arms (N_{test}=9,857). Error bars: 95% bootstrap CIs where available. HOSPITAL and LACE lack Brier scores (partial proxy implementations).

4.2.1 Calibration

Calibration was evaluated following [42]. The A4 baseline on the validation split (N=5,107) shows near-ideal calibration with slope 0.935 and intercept 0.035 (O/E ratio 1.094, ECE 0.021) [VERIFIED]. For the $A3_{factlevel}$ arm, we report calibration on the full canonical test split (N=9,857): O/E 0.980 (well-centered), ECE 0.018, but slope 0.741 (<1.0), indicating over-confidence in higher-risk bins [PRELIMINARY].

Table 11: Calibration metrics. A4: validation split; $A3_{factlevel}$: full canonical test split.

Arm	Split	N	Slope	Intercept	O/E	ECE
A4	Val	5,107	0.935	0.035	1.094	0.021
$A3_{fact}$	Test	9,857	0.741	−0.372	0.980	0.018

Figure 4 shows the reliability diagram.

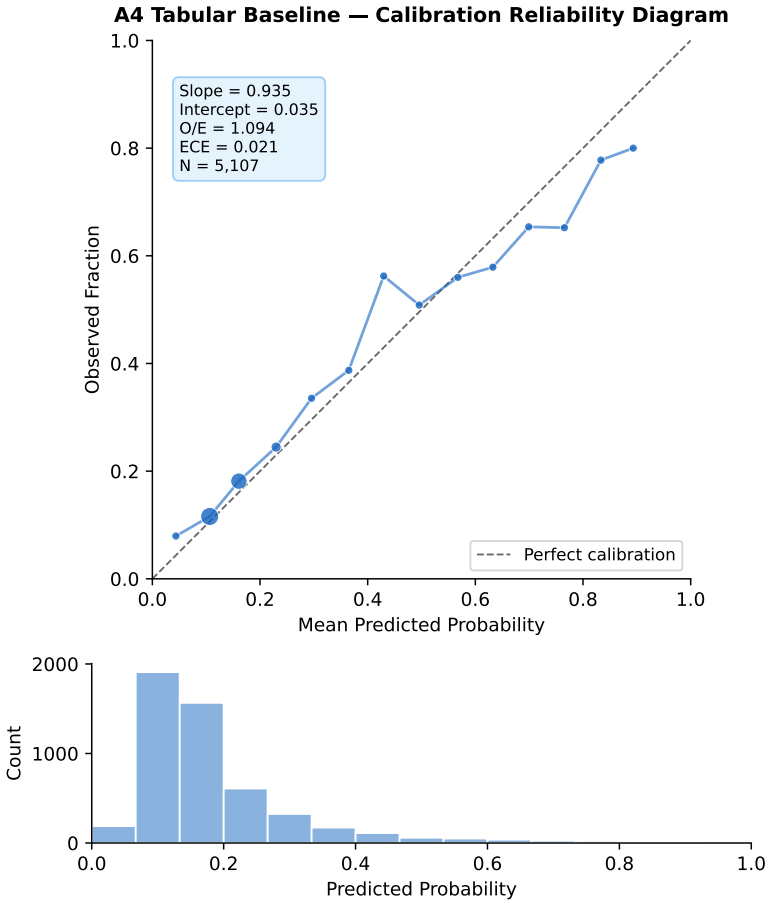


Figure 4: Calibration reliability diagram for the A4 tabular baseline (validation split, N=5,107). Dot size is proportional to bin count. The model shows near-ideal calibration in the clinically relevant range (predicted probability 0.05–0.40), with slight over-confidence in higher-risk bins.

4.3 Subset Benchmark (`subset4200`)

To enable rapid iteration on notes-derived features, we evaluate on a fixed subset (N_{train}=1,000, N_{val}=200, N_{test}=3,000) drawn from the canonical Track B cohort with the same label distribution.

Table 12: Subset benchmark results (N_{test}=3,000). CI: 95% subject-grouped bootstrap (K=2,000, seed=42). All results carry PRELIMINARY status.

Arm	Description	AUROC	AUPRC	Brier
$A3_{factlevel}$	Fact-level KVT tokens (AdaGrad)	**0.655**	0.306	0.146
A3	Aggregate scoring (train+val refit)	0.641	0.293	0.151
$A4_{subset}$	Tabular ML (train+val refit)	0.638	0.292	0.149

The fact-level KVT tokenization arm ($A3_{factlevel}$) achieves the highest subset AUROC (0.655), outperforming both the aggregate-based combination (A3: 0.641) and the tabular-only arm refitted on the subset ($A4_{subset}$: 0.638).

4.4 Paired Arm Comparisons

Table 13 presents paired subject-bootstrap deltas between the fact-level arm and baselines on the shared subset test set.

Table 13: Paired subject-bootstrap deltas (K=2,000, seed=42, N_{test}=3,000).

Comparison	ΔAUROC	95% CI	ΔAUPRC	ΔBrier
$A3_{fact}$ vs $A4_{subset}$	+0.017	[+0.007, +0.027]	+0.015	−0.002
$A3_{fact}$ vs $A3_{agg}$	+0.014	[+0.004, +0.024]	+0.013	−0.005

Both comparisons show statistically significant improvement: the 95% CI for ΔAUROC excludes zero in both cases. The fact-level arm gains +1.7 pp over tabular-only and +1.4 pp over aggregates, with consistent improvements in AUPRC and Brier score. In the bootstrap distribution, ΔAUROC > 0 in 99.9% and 99.7% of resamples for the tabular and aggregate comparisons, respectively.

4.5 Full Canonical Scale-Out (`test9857`)

As extraction throughput improved, we progressively scaled evaluation beyond the rapid-iteration subset. We report the final evaluation on the complete canonical Track B test split (N_{test}=9,857 admissions). To ensure a fair comparison with the notes-enabled $A3_{factlevel}$ arm, we refit the tabular baseline on the same reduced training splits (N_{train}=1,500, N_{val}=400), corresponding to 11,757 extracted documents in total (train+val+test).

Table 14: Consolidated canonical results on the full test split (N_{test}=9,857; N_{train}=1,500; N_{val}=400). All arms trained on the same reduced splits for fair comparison. 95% bootstrap CIs (K=2,000, seed=42).

Arm	Description	AUROC [95% CI]	AUPRC [95% CI]	Brier [95% CI]
$A3_{fact}$	Fact-level KVT (AdaGrad)	**0.659**	**0.321**	**0.145**
Typed fusion	Semantic prob. + RF	0.655 [0.641, 0.669]	0.326 [0.306, 0.347]	0.146 [0.142, 0.150]
$A4_{LR}$	Hash+LogReg (SGD)	0.656	0.304	0.146
$A4_{XGB}$	XGBoost (same features)	0.646 [0.632, 0.660]	0.312 [0.293, 0.333]	0.170 [0.166, 0.173]

Estimator comparison ($A4_{XGB}$ vs. $A4_{LR}$). To test whether the logistic regression baseline is too weak a comparator, we trained an XGBoost classifier on the identical feature set.

XGBoost does not outperform LogReg on these features: paired deltas show ΔAUROC = −0.010 [−0.023, +0.003] and ΔBrier = +0.023 [+0.020, +0.027] (higher = worse). The modest feature set (26 numeric + 7 categorical) favors linear models, confirming that the choice of logistic regression is not artificially weak and that downstream improvements from $A3_{factlevel}$ are attributable to the KVT4 features, not to estimator weakness.

Table 15: Paired subject-bootstrap deltas on the full canonical test split (K=2,000, seed=42). Brackets indicate 95% CIs. Positive Δ favors $A3_{factlevel}$.

Comparison	**ΔAUROC**	**ΔAUPRC**	**ΔBrier**
$A3_{fact}$ vs $A4_{refit}$	+0.0033 [−0.0048, 0.0107]	+0.0162 [0.0077, 0.0252]	−0.0012 [−0.0022, −0.0003]

On the full canonical test split, the fact-level arm yields VERIFIED improvements in AUPRC and Brier score (paired bootstrap CIs exclude zero), while the AUROC uplift is small and not statistically confirmed. Figure 5 summarizes the paired comparisons across all key arms.

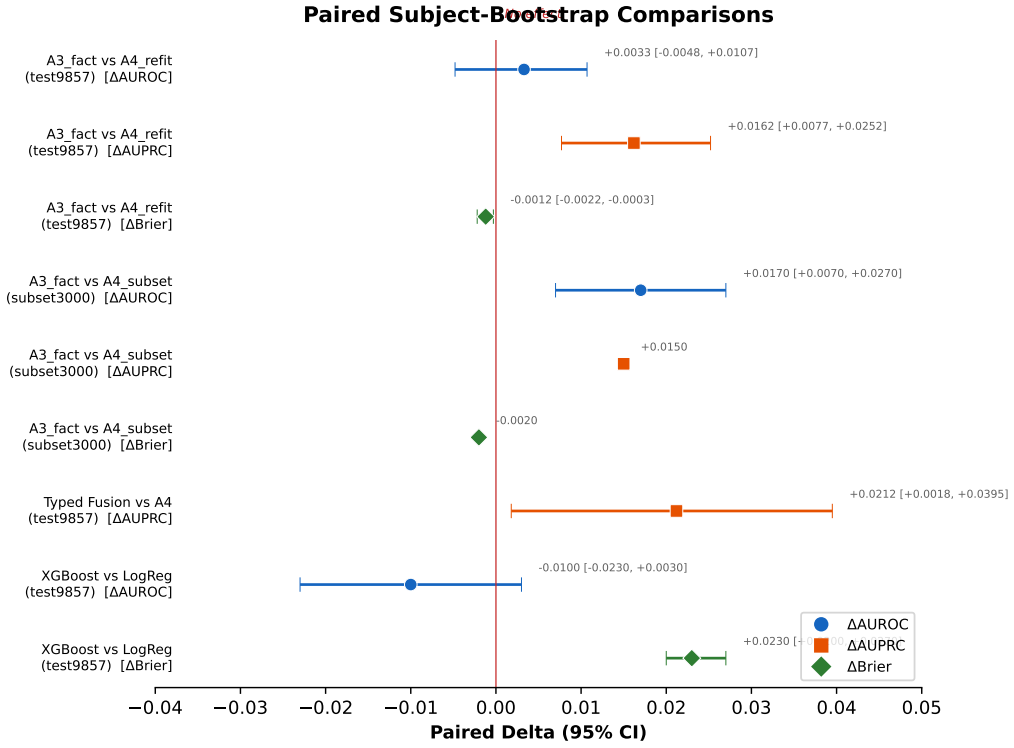


Figure 5: Forest plot of paired subject-bootstrap deltas (95% CIs). Comparisons whose CI excludes zero are statistically distinguishable from zero; primary pre-specified contrasts then carry the VERIFIED label, while exploratory ones remain hypothesis-generating (Section 3.9.4). Positive ΔAUROC/ΔAUPRC and negative ΔBrier favor the first-named arm.

4.5.1 Subgroup Fairness (Preliminary)

We performed a subgroup analysis with paired subject-bootstrap confidence intervals on the same fair-comparison canonical rerun, using age and sex covariates added to both A3 and A4. The strongest recurring pattern is not a universal AUROC gain, but improved AUPRC and Brier score for A3. In the female subgroup, ΔAUPRC is +0.0145 [0.0030, 0.0270], ΔBrier is -0.00144 [-0.00269, -0.00019], and ΔAUROC is +0.0102 [0.0004, 0.0206]. In the male subgroup, ΔAUPRC remains positive at +0.0167 [0.0026, 0.0296], while ΔAUROC and ΔBrier are mixed.

Across age strata, the <65 group shows CI-backed gains on both AUPRC (+0.0215 [0.0076, 0.0354]) and Brier (-0.00181 [-0.00326, -0.00025]), whereas the 65–79 and 80+ groups remain directionally compatible with the overall result but are not individually conclusive. Broad race-category analysis remains exploratory: the WHITE group shows stronger AUPRC and Brier, while non-WHITE categories are still too mixed for stronger claims. All subgroup contrasts here are *exploratory and hypothesis-generating*: they carry no multiplicity correction across the sex × age × race family (Section 3.9.4), and several intervals only marginally exclude zero (e.g., female ΔAUROC [0.0004, 0.0206], female ΔBrier [−0.00269, −0.00019]). We therefore do not attach a VERIFIED label to any individual subgroup effect.

4.6 Notes-Based Signal

On an independent, more balanced *proxy* cohort (not the canonical split), SGR v4.1 on N=50 yields AUROC 0.646 versus a 0.511 baseline (Δ +13.5 percentage points), indicating that notes-derived signal can be meaningful (PRELIMINARY). We stress that this proxy number is *not* comparable to the canonical notes-only result below: it comes from a favorable, class-enriched proxy cohort selected for rapid iteration, and must not be read as canonical notes-only performance.

On the canonical benchmark, the original notes-only arm (A2, N=50) achieves AUROC 0.462 with wide CI [0.296, 0.625]. Because the CI includes 0.5, this result is **not statistically distinguishable from random** classification; however, this is expected given the extremely small evaluation sample (N=50) due to extraction throughput constraints (see Section 3.4). On an earlier subset (N_{test}=2,000), A2 with expanded 15-aggregate features achieves AUROC 0.587 [0.552, 0.620], now above chance, confirming that MedGemma-extracted clinical information carries predictive value for readmission when evaluated on a sufficiently large sample.

4.7 Signal-Integrity Audit

Table 16 reports the signal-integrity audit on the subset (N=4,000 documents, 3,196 with objective data in Stage 1).

Table 16: Signal-integrity audit (N=4,000 documents).

Metric	Before hybrid	After hybrid
Documents with objective loss	497/3,196 (15.55%)	271/3,196 (8.48%)
Keys lost (doc+key level)	4,344/35,477 (12.24%)	1,862/35,477 (5.25%)
Generation loss (docs)	299	—
Postprocess drop (docs)	225	—

Before hybrid regeneration, 15.55% of documents with objective data suffered silent signal-loss. After deterministic offline hybrid recovery (no additional LLM calls), objective loss is reduced but not eliminated (8.48% doc-level, 5.25% key-level). Generation loss (299 docs) and postprocess drop (225 docs) are both observed and overlap: since $299 + 225 = 524 > 497$ affected documents, 27 documents exhibit both loss types.

Downstream impact of hybrid regeneration (Preliminary). To quantify the prediction effect, we ran the $A3_{factlevel}$ pipeline with and without hybrid regeneration on the canonical test split. Hybrid recovery nearly doubles the extracted fact count (352K vs. 184K total tokens), yet all three prediction deltas are statistically neutral: ΔAUROC = −0.005 [−0.022, +0.011], ΔAUPRC = +0.000 [−0.015, +0.016], ΔBrier = −0.001 [−0.004, +0.002]. This indicates that hybrid regeneration serves primarily as an extraction completeness and audit mechanism rather than a prediction performance driver: the downstream model already captures sufficient signal from the non-recovered facts.

4.8 Inference Speed (Stage 2 Prompt Caching)

Enabling prompt KV cache reuse (CAG) on the Stage 2 backend yields a **+10.6% wall-time speedup** (6,297 s → 5,629 s for 200 documents) with **zero output diffs** across all Stage 2 artifacts—confirming bit-exact equivalence at lower latency.

4.9 Matched Revision Ablations (`test200`)

To address reviewer concerns about component-level contributions, we ran two matched `test200` ablations under identical serving and evaluation code. All results in this subsection are PRELIMINARY because the slice is small compared with the canonical benchmark.

Table 17: Matched Stage 2 ablation on `test200` with Stage 1 fixed to `sgr_v2`. All results are PRELIMINARY.

Arm	Parse	Facts/doc	AUROC	AUPRC	Brier	Status
LoRA	0.955	19.235	0.573	0.308	0.174	Preliminary
Base-only	0.985	8.815	0.550	0.318	0.179	Preliminary

Table 18: Matched Stage 1 ablation on `test200` with Stage 2 fixed to LoRA. All results are PRELIMINARY.

Arm	Parse	Facts/doc	AUROC	AUPRC	Brier	Status
`sgr_v2` + LoRA	0.955	19.235	0.573	0.308	0.174	Preliminary
`strings_v1` + LoRA	1.000	20.780	0.614	0.303	0.172	Preliminary

The matched `test200` ablations indicate that the Stage 2 LoRA adapter materially improves extraction yield relative to base-only decoding, while the downstream proxy metrics are mixed on AUPRC but favor LoRA on AUROC and Brier. For Stage 1, the less constrained `strings_v1` profile produces slightly more facts and a higher AUROC on this slice, which means schema guidance should be presented as a stability choice rather than a universal downstream win. Using CAG, the rapid-iteration subset expanded from 3,200 to 4,200 documents for downstream evaluation, and subsequent canonical scale-out completed extraction for 11,757 documents (train/val/test = 1,500/400/9,857), enabling evaluation on the full canonical test split.

After the formal revision package was closed, we also tested a more radical downstream-only continuation branch that fused typed semantic probabilities directly into the downstream predictor rather than further modifying the free-form Stage 2 generator. We first observed a promising `test200` signal and then repeated the same formulation on the canonical `train1500/val400/test9857` split. On canonical `test9857`, the typed-fusion arm reached AUROC/AUPRC/Brier of 0.655/0.326/0.146.

The clinically meaningful comparison is against the *relevant* arms, not the weak internal one. Relative to the notes-only Stage 2 baseline (AUROC ≈0.545), paired subject-bootstrap gives large deltas (ΔAUROC = +0.110 [0.088, 0.132], ΔAUPRC = +0.114 [0.087, 0.142], ΔBrier = −0.0088 [−0.0128, −0.0051]); we report these only for completeness, because that baseline is itself far below the tabular and fact-level arms, so the gap mostly reflects the baseline's weakness rather than the strength of typed fusion. Against the relevant canonical `A4_tabular` arm, the *only* CI-backed gain is AUPRC +0.021 [0.0018, 0.0395] (lower bound near zero), with AUROC and Brier statistically compatible with parity; against the canonical `A3_factlevel` arm, all differences are mixed with confidence intervals crossing zero. We therefore treat typed fusion as a promising *exploratory* continuation direction with a single CI-backed AUPRC gain over the tabular arm—not as a headline result and not as a replacement for the fact-level arm. Because

this contrast is exploratory and its CI lower bound is essentially at zero, it does not carry the formal VERIFIED label reserved for primary contrasts (Section 3.9.4).

4.10 Model Scaling Pilot (GPT-4.1-mini vs. MedGemma 4B)

To probe whether moderate LABS extraction quality reflects a pipeline limitation or a reference-alignment plateau, we ran a three-arm ablation replacing Stage 1, Stage 2, or both with GPT-4.1-mini on N=200 documents. GPT-4.1-mini was accessed through a HIPAA-compliant Azure OpenAI environment provided by healthPrecision.

Table 19: Model scaling pilot (N=200). LABS metrics use the same structured-reference benchmark as Table 8. All results PRELIMINARY.

Arm	Stage1 / Stage2	Facts/doc	LABS Prec.	LABS Rec.	LABS F1
Arm 1	MedGemma / MedGemma+LoRA	27.0	0.476	0.492	0.484
Arm 2	GPT-4.1-mini / MedGemma+LoRA	15.5	0.505	0.539	**0.521**
Arm 3	GPT-4.1-mini / GPT-4.1-mini	15.7	0.478	0.487	0.482

Replacing Stage 1 with GPT-4.1-mini (Arm 2) yields the highest LABS F1 (0.521, +0.037 absolute over the MedGemma baseline), yet remains close to the ≈0.52 level. Because this rests on a single non-frontier alternative model at N=200, we read it as *directional evidence*—consistent with the inherent mismatch between discharge narrative text and structured laboratory tables (different lab draws, rounding, admission-vs-discharge values)—rather than as conclusive evidence of a hard ceiling. Cleanly separating residual extraction error from table mismatch would require a frontier-model arm or a clinician-adjudicated reference, which we leave to future work (Section 5.4). Arm 2 and Arm 3 produce zero semantic-cluster facts (PROBLEMS, SYMPTOMS, MEDICATIONS, PROCEDURES), demonstrating that the Stage 2 LoRA adapter is not merely a format enforcer but a trained output policy required for semantic extraction. The pipeline architecture supports pluggable Stage 1 replacement: if a stronger local model becomes available, the Stage 2 LoRA stage can be retained without retraining.

4.11 Ablation and Analysis

LABS alias canonicalization. Before canonicalization, LABS keyword matching yielded F1=0.467; after applying canonical alias mapping, F1 rose to 0.504—demonstrating the importance of deterministic post-processing.

Missingness. A non-trivial fraction of admissions lack laboratory events; the A4 baseline explicitly models this via per-lab missingness indicators.

5 Discussion

5.1 Clinical Implications

MedGemma StructCore demonstrates that compact, domain-adapted language models can produce *format-stable, auditable structured extraction* from clinical free text on consumer hardware. We are deliberately careful with the word "reliable": the pipeline reliably enforces the KVT4 output contract and audits objective signal-loss, but per-cluster extraction *accuracy*—especially for the semantic clusters—is only partially validated (Section 5.4). Rather than positioning this as a standalone readmission predictor, we emphasize structured extraction as **reusable infrastructure**: the same KVT4 extraction layer can feed discharge quality checks, care-gap flagging, cohort analytics, and other downstream tasks. For any probability-based clinical use, however, the over-confidence we observe at higher risk (A3 calibration slope 0.741) makes pre-deployment recalibration a prerequisite, not an optional refinement (Section 5.4).

The readmission use case serves as a *lighthouse demonstration*—one concrete endpoint that validates the extraction layer while illustrating the broader potential. The framework is inherently extensible, as the KVT4 ontology can be scaled by adding new clusters, keywords, or timestamp definitions, coupled with further task-specific model fine-tuning.

5.2 Local-First Deployment Value

Our pipeline processes clinical notes entirely on-device, with no data leaving the local machine. This addresses a fundamental barrier: even in research settings, some cloud-based LLMs cannot be applied to sensitive EHR data due to privacy constraints [7]. While federated learning [8] and edge deployment [9] are active research areas, fully local inference with quantized models remains the strongest privacy guarantee available today. Quantized models (GGUF Q5_K_M) running via llama.cpp make inference feasible on Apple Silicon laptops, Kaggle notebooks with 2x GPU (Tesla T4), or standard consumer GPUs with $\geq$8 GB VRAM. Prompt KV cache reuse (CAG) further improves throughput (+10.6%) without any output degradation, making large-scale local extraction practical even without GPU clusters.

5.3 Signal-Integrity and Hybrid Recovery

The signal-integrity audit (section 4.7) reveals that two-stage LLM pipelines can silently lose over 10% of objective data between extraction and feature engineering. This finding has broader implications: *any* multi-step NLP pipeline in clinical settings should include an automated signal-loss gate, not just an end-to-end metric.

Crucially, a substantial fraction of losses were recoverable through deterministic offline hybrid regeneration—without additional LLM calls or re-extraction. However, remaining loss persists even after regeneration (N=4,000: 8.48% doc-level, 5.25% key-level), indicating that both generation drift and postprocessing robustness are relevant bottlenecks.

5.4 Limitations

1. **Single institution.** MIMIC-IV contains data from one academic medical center (BIDMC). Documentation patterns, patient demographics, and clinical practices may differ substantially from other institutions.
2. **No temporal validation.** All results use a patient-level random split. An out-of-time split would better test real-world deployment scenarios.
3. **Constrained training size for $\mathbf{A3_{factlevel}}$.** While $A3_{\text{factlevel}}$ is evaluated on the full canonical test split (N_{test}=9,857), the notes-enabled models are trained on a reduced extracted training set (N_{train}=1,500, N_{val}=400). Scaling notes extraction to the full available training cohort may yield stronger discrimination gains. Matched `test200` revision ablations support this interpretation: Stage 2 LoRA improves extraction yield and downstream AUROC/Brier relative to base-only decoding, but the gains are still configuration-specific and should be labeled PRELIMINARY outside the canonical benchmark. On the canonical test split, AUROC uplift is small and not statistically verified, while AUPRC and Brier improvements are verified.
4. **PROBLEMS/SYMPTOMS recall.** Semantic cluster extraction remains a bottleneck; model performance on complex diagnostic reasoning is limited at 4B parameters.
5. **Most-predictive cluster downweighted to zero.** In the fact-level arm, validation grid search drove the DISPOSITION weight to `scale=0.0` (Section 3.7), effectively discarding discharge disposition—clinically one of the strongest readmission predictors. We candidly read this as a sign that the *extracted* DISPOSITION facts were too noisy in the current regime to add downstream value, consistent with the broader gap that DISPOSITION—like the semantic clusters—lacks a quantitative extraction-quality benchmark. Improving DISPOSITION extraction and re-testing its downstream contribution is future work.

6. **Notes-only branch (A2) is not viable at canonical scale.** Because Stage 1+Stage 2 extraction runs at ∼25 s/doc on consumer hardware, the pure notes-only arm (A2) could be evaluated on only N≤50 canonical admissions, with a CI that includes chance (Table 10). This is a structural throughput limitation rather than a footnote: at present the notes-only signal can be validated at canonical scale only when *fused* with tabular features (A3/A3$_{\text{factlevel}}$), not in isolation.
7. **Stage 2 format drift.** Despite near-perfect overall validity (99.74%), Stage 2 can emit 3-part lines or JSON fragments on certain notes, requiring recovery heuristics.
8. **Stage 1 schema guidance is not universally optimal.** Matched `test200` ablations show that the less constrained `strings_v1` profile can achieve higher downstream AUROC on some slices, even though `sgr_v2` remains the better structured default for auditability. This makes Stage 1 a stability/contract choice rather than a guaranteed performance win.
9. **Partial proxy scores.** Our LACE and HOSPITAL implementations are partial (missing Charlson comorbidity index, full ED visit counts) and thus represent lower bounds on the original scores' performance.
10. **Fairness analysis remains preliminary.** We now include a first subgroup analysis across sex, age, and broad race categories on the fair-comparison canonical rerun. The results are directionally reassuring, because A3 retains its strongest gains on AUPRC/Brier across sex and age slices, and subgroup paired-bootstrap confidence intervals are now available. However, the evidence remains preliminary because AUROC is still mixed across slices, formal disparity testing is absent, and no multiplicity correction was applied across the subgroup family, so individual subgroup intervals are hypothesis-generating (Section 3.9.4).
11. **No representation-learning baselines.** We do not compare against ClinicalBERT [17], ClinicalT5 [23], or graph-based approaches [11] that report AUROC up to 0.72 on MIMIC-IV readmission. Our focus is on the extraction infrastructure rather than achieving state-of-the-art discrimination. We do, however, include a gradient-boosted baseline (XGBoost) on the same feature set as A4: it does not outperform logistic regression (paired ΔAUROC = −0.010 [−0.023, +0.003], ΔBrier = +0.023 [+0.020, +0.027]), confirming that the modest feature set favors linear models and that our choice of estimator does not artificially handicap the tabular baseline. End-to-end representation-learning comparisons remain future work.
12. **Limited extraction ground truth.** The curated extraction evaluation covers only 10 documents (Curated10). While micro-F1 of 0.984 on VITALS+LABS is encouraging, this sample is insufficient to establish robust extraction quality bounds. A larger clinician-annotated evaluation ($N \geq 50$) across all 9 clusters is needed. We partially mitigate this gap with structured-reference benchmarks: a large LABS benchmark on full `test9857` and a preliminary VITALS benchmark on a recovered 500-admission canonical block. For VITALS, a coverage-aware chartevents view now shows that the raw precision collapse was driven in large part by source sparsity rather than by value mismatch alone; however, it still does not support a strong large-scale extraction-quality claim, especially because Weight remains weak. In particular, per-cluster F1 scores for PROBLEMS, SYMPTOMS, MEDICATIONS, and PROCEDURES have not been quantitatively evaluated; only VITALS and LABS micro-F1 is reported. All Curated10 metrics should be treated as PRELIMINARY.
13. **A3 calibration shows over-confidence at higher risk.** On the full canonical test split (N=9,857), A3$_{\text{factlevel}}$ has O/E 0.980 and ECE 0.018 (low overall miscalibration), but slope 0.741 (<1.0), indicating over-confidence in higher-risk bins. Recalibration (e.g., Platt scaling on a held-out set) would be needed before deployment-oriented interpretation of predicted probabilities.
14. **Clinical significance of ΔAUROC.** On the rapid-iteration subset, A3$_{\text{factlevel}}$ shows a +0.017 ΔAUROC over a refit baseline; however, on the full canonical test split the ΔAUROC is smaller (+0.0033) and not statistically confirmed. Established thresholds for minimally clinically important differences (MCID) in readmission AUROC are not well-defined. Whether small

discrimination shifts translate to improved clinical decision-making at actionable thresholds requires prospective evaluation. Given class imbalance, AUPRC and calibration may be more informative for deployment decisions.

15. **Scaling pilot scope.** The GPT-4.1-mini ablation (section 4.10) covers only N=200 documents and does not include downstream prediction. The ≈0.52 plateau finding is directional, not canonical, and rests on a single non-frontier alternative model; it does not by itself establish a hard extraction-quality ceiling.
16. **No definitive subgroup fairness validation.** Although subgroup confidence intervals and calibration-by-subgroup are now available, we do not yet report formal disparity tests or subgroup-aware recalibration. A first subgroup-aware recalibration audit was attempted, but did not yield a clean net benefit in the current validation regime. For that reason, the current fairness evidence should still be treated as preliminary rather than as a deployment-grade equity assessment.

5.5 Future Work

Near-term priorities include: (1) scaling notes extraction for the training cohort beyond N_{train}=1,500/N_{val}=400 and retraining $A3_{factlevel}$ at full scale; (2) further developing typed downstream fusion, which shows a single CI-backed (exploratory) AUPRC gain over `A4_tabular` and near-parity with `A3_factlevel` (its larger margin over the weak notes-only Stage 2 baseline is not, by itself, evidence of strength); the remaining gap to `A3_factlevel` appears to be a ranking problem rather than a generic extraction-capacity problem; (3) reducing remaining objective signal-loss identified by the signal-integrity audit; (4) strengthening the VITALS structured-reference benchmark, especially for the chartevents-backed BP/Weight path; (5) formal disparity testing and more robust subgroup-aware recalibration for a stronger fairness assessment; (6) temporal validation with an out-of-time split; (7) richer typed/fact-level features (e.g., value-binning, ontology embeddings, or typed downstream fusion features); (8) multi-institutional external validation; (9) evaluation on additional downstream clinical tasks beyond readmission; (10) testing stronger local models (Qwen2.5-7B, Phi-4, LLaMA-3.2-8B) as Stage 1 replacements, following the scaling pilot evidence that the pipeline architecture supports pluggable model substitution without Stage 2 retraining.

6 Ethical Considerations and Reproducibility

6.1 Ethical Considerations

6.1.1 Data and Privacy

This study uses **MIMIC-IV** [40]—a publicly available de-identified EHR database. Data were fully de-identified in accordance with the **HIPAA Safe Harbor** standard (removal of 18 identifier categories, random date shifts, masking of rare diagnoses/procedures that could be identifying).

- **Ethical approval:** MIMIC-IV is collected under BIDMC IRB approval. Secondary analysis of de-identified data does not require separate ethical approval under 45 CFR §46.104.
- **Data access:** PhysioNet Credentialed Access (CITI training + Data Use Agreement).
- **Local inference-time processing:** The deployed StructCore pipeline runs locally and does not transmit patient data to external servers or cloud APIs at inference time—a deliberate privacy-by-design architectural decision. A one-time, offline teacher-supervision step used Gemini 3 Flash Thinking through a HIPAA-compliant Google Cloud Vertex AI environment to construct the Stage 2 fine-tuning set. Separately, the GPT-4.1-mini scaling pilot used a HIPAA-compliant Azure OpenAI environment provided by healthPrecision. These cloud-assisted steps are disclosed below and are not part of the deployed inference pipeline.

- **Data hygiene:** The repository excludes raw note text and per-encounter folders from version control via `.gitignore`; contributions are reviewed to prevent accidental inclusion of patient-level text artifacts.

6.1.2 Clinical Disclaimer

MedGemma StructCore is developed **exclusively as a research tool**. The system is *not* a medical device, has no FDA/CE certification, is *not* intended to support clinical decisions in real-world practice without further clinical validation, and does *not* replace clinical judgment.

All results were obtained in a retrospective study of single-institution de-identified data and should not be extrapolated to other populations or clinical contexts without external validation.

6.1.3 Potential Biases

- **Institutional bias:** single academic medical center.
- **Language bias:** optimized for English clinical documents.
- **Model bias:** MedGemma trained predominantly on English medical texts.

6.2 AI Tool Declaration

In research: MedGemma 4B (Google HAI-DEF) was used as the primary extraction model; Gemini 3 Flash Thinking was used as a teacher model for generating ground truth datasets for LoRA adapter fine-tuning; GPT-4.1-mini was used in a limited model-scaling pilot. All models were used as data processing tools; final results were verified and interpreted by human authors.

In manuscript preparation: AI coding assistants (Claude, Gemini) were used for code development, document structuring, and bibliographic entry generation. All scientific claims, results, and interpretations were verified and approved by the human authors.

6.3 Reproducibility

Table 20: Availability of components and artifacts.

Component	Availability	Reference
Pipeline code	Open (GitHub)	https://github.com/SZabolotnii/MedGemma_StructCore-code-supplement
MedGemma 4B	Open weights (Google)	HuggingFace
LoRA adapters	Public	https://huggingface.co/DocUA/medgemma-1.5-4b-it-gguf-q5-k-m-two-stage
Interactive Demo	Open (Kaggle)	https://www.kaggle.com/code/zabolotnii/demo-medgemma-structcore
MIMIC-IV	Credentialed	PhysioNet
Benchmark protocol	Open	`STANDARD_BENCHMARK_PROTOCOL.md`
Run manifests + metrics	Generated (non-PHI)	`results/benchmark/<run_id>/`
Signal-integrity reports	Generated (non-PHI)	`results/benchmark/<run_id>/`
Paired bootstrap artifacts	Generated (non-PHI)	`results/benchmark/<run_id>/`
KVT ontology + allowlists	Open	`ONTOLOGY.md`
JSON schemas (Stage 1)	Open	`schemas/*.schema.json`
CAG A/B validation artifacts	Generated (non-PHI)	`results/benchmark/<run_id>/`

Fixed parameters: Random seed 42, bootstrap K=2,000, temperature 0.0, quantization Q5_K_M (GGUF), max tokens Stage 1: 256–384, Stage 2: 512.

Key reproduction commands:

1. **Canonical benchmark (A4):**
```
python3 Analysis_Readmission/run_trackb_tabular_baseline_hashlogreg.py \
-cohort-csv Analysis_Readmission/trackB_hosp_v3_20260207.cohort.csv \
-labels-csv Analysis_Readmission/trackB_hosp_v3_20260207.labels.csv \
-labs-features-csv Analysis_Readmission/trackB_hosp_v3_20260207.labs_ccde_features.
csv \
-split-manifest results/benchmark/20260207_trackB_hosp_v3_metadata50k_hash/split_m
anifest.csv \
-cohort-flow-json Analysis_Readmission/trackB_hosp_v3_20260207.cohort_flow.json \
-out-dir results/benchmark/<run_id> -final-train-on train+val
```
2. **Hybrid regeneration:**
```
python3 scripts/build_stage2_hybrid_facts_from_stage1_md.py \
-src-out-dir <RUN> -dst-out-dir <DST> -from-stage2-raw \
-recover-missing-timestamp -supplement-objective \
-supplement-medications -supplement-problems
```
3. **Signal-integrity audit:**
```
python3 Analysis_Readmission/report_stage_signal_loss.py \
-facts-root <RUN>/raw_stage_traces -out-json <OUT>.json -simulate-offline-regen
```
4. **Paired subject-bootstrap:**
```
python3 Analysis_Readmission/compute_paired_bootstrap_delta.py \
-predictions-a <A>/predictions_test.csv -predictions-b <B>/predictions_test.csv \
-group-manifest <SPLIT_MANIFEST>.csv -group-col subject_id -n-boot 2000 -seed 42
```
5. **Stage 2 with prompt caching (CAG):**
```
llama-server -model <GGUF> -lora <ADAPTER> -alias <ALIAS> \
-cache-prompt -cache-reuse 256
```

6.4 Author Contributions

Serhii Zabolotnii: Conceptualization, Methodology, Software, Formal analysis, Investigation, Writing — original draft. **Viktoriia Holinko:** Validation, Data curation, Resources, Writing — review & editing. Both authors read and approved the final manuscript.

6.5 Data Availability

This study used MIMIC-IV v3.1, available to credentialed users through PhysioNet (https://physionet.org/content/mimiciv/) under a Data Use Agreement; the data cannot be redistributed by the authors. All derived, non-PHI artifacts (run manifests, metrics, signal-integrity reports, paired-bootstrap outputs) and the benchmark protocol are described in the Reproducibility section. An author-deposited preprint is available on Zenodo (https://zenodo.org/records/20605433).

6.6 Code Availability

The pipeline code (data-free) is publicly available at https://github.com/SZabolotnii/MedGemma_StructCore-code-supplement. LoRA adapters and an interactive demo are linked in the Reproducibility section.

6.7 Conflict of Interest

Co-author Viktoriia Holinko is affiliated with healthPrecision, which provided access to the HIPAA-compliant Google Cloud Vertex AI environment used for the one-time teacher-supervision step and to a HIPAA-compliant Azure OpenAI environment used for the GPT-4.1-mini scaling pilot. The corresponding author, Serhii Zabolotnii, has no relationship with healthPrecision. The authors declare no other competing interests. This research was conducted as part of

The MedGemma Impact Challenge (Kaggle, Google Research)—a public competition; participation does not entail financial ties to, or publication obligations with, the organizer.

6.8 Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors. The only material support was in-kind: healthPrecision provided access to a HIPAA-compliant Google Cloud Vertex AI environment used solely for the one-time generation of teacher-model supervision for the Stage 2 fine-tuning set, and to a HIPAA-compliant Azure OpenAI environment used for the GPT-4.1-mini scaling pilot. The provider had no role in study design, analysis, interpretation, or the decision to submit for publication.

6.9 Patient and Public Involvement

No patients or members of the public were involved in the design, conduct, reporting, or dissemination plans of this research. The study is a secondary analysis of an existing, fully de-identified electronic health record database, and no participants could be approached.

6.10 Study Registration and Prior Dissemination

This study was not prospectively registered, and no a priori analysis protocol was filed; it is a retrospective secondary analysis of a publicly available, de-identified database. This work was developed for The MedGemma Impact Challenge (Kaggle, Google Research); a non-peer-reviewed competition write-up describing the system and preliminary results was publicly released under a Creative Commons Attribution 4.0 licence (https://www.kaggle.com/competitions/med-gemma-impact-challenge/writeups/new-writeup-1768562148448). An author-deposited preprint is available on Zenodo (https://zenodo.org/records/20605433). The present manuscript is the extended journal version of that prior, non-peer-reviewed work and contains a substantially broader set of analyses than the competition write-up: full-canonical-test benchmarking, paired subject-bootstrap significance testing, typed downstream fusion, subgroup-fairness and calibration analyses, structured-reference extraction validation at scale, a model-scaling pilot, and complete MI-CLAIM/TRIPOD+AI reporting.

Acknowledgements

We thank the Google Health AI team for creating and releasing the MedGemma (HAI-DEF) models; the PhysioNet team and MIMIC-IV developers for providing access to de-identified clinical data; the organizers of The MedGemma Impact Challenge (Kaggle, Google Research); and the llama.cpp community for quantization and local inference tools.

A TRIPOD+AI Adherence Checklist

Table 21 maps key TRIPOD+AI [38] items to the corresponding sections and tables in this manuscript.

References

[1] Devan Kansagara, Honora Englander, Amanda Salanitro, David Kagen, Cecelia Theobald, Michele Freeman, and Sunil Kripalani. Risk prediction models for hospital readmission: a systematic review. *JAMA*, 306(15):1688–1698, 2011.

Table 21: TRIPOD+AI adherence checklist. Section numbers refer to the manuscript.

TRIPOD+AI Item	**Manuscript Location**
Title identifies prediction model	Title
Abstract: objectives, methods, results	Abstract
Source of data	§3: MIMIC-IV, single institution
Participants: eligibility criteria	§3: inclusion/exclusion, Table 1
Outcome: definition, timing	§3: 30-day readmission, operationalized
Sample size rationale	§3: N=50,000 cohort; constrained training regime disclosed
Missing data handling	§3: per-lab missingness indicators
Feature selection / predictors	§3: KVT4 clusters, canonical keywords (ONTOLOGY)
Model specification	§3: Arms A0–A4, hashing, AdaGrad, LogReg, XGBoost
Model training / hyperparameters	§3: Table 2 (LoRA), Table 3 (arms)
Risk groups / thresholds	§3: risk engine scoring rules
Calibration assessment	§4.2.1: slope, intercept, ECE, O/E (Table 11)
Discrimination metrics	§4: AUROC, AUPRC, Brier (Tables 10 and 14)
Confidence intervals	Bootstrap CIs (K=2,000) throughout §4
Paired comparisons	§4: Tables 13 and 15, Figure 5
Subgroup analysis	§4.5.1: sex, age, race subgroups
Model performance figures	Figures 2 to 5
Data splitting strategy	§3: patient-level 70/10/20, no leakage
Temporal considerations	§5.4: random split; temporal validation = future work
Sensitivity analyses / ablations	§4.9: Stage 1/Stage 2 matched ablations
External validation	§5.4: single institution; multi-site = future work
Deployment considerations	§5: local-first, consumer hardware, privacy
AI-specific: training data provenance	§3: teacher verification protocol, hard200
AI-specific: model interpretability	§3: KVT4 explicit facts, not embeddings
Code and data availability	§6: GitHub, HuggingFace, reproducibility table

[2] Centers for Medicare & Medicaid Services. Hybrid hospital-wide (all-condition, all-procedure) risk-standardized readmission measure: Methodology report. Technical report, Centers for Medicare & Medicaid Services (CMS), 2023. Official CMS methodology combining claims-based and EHR-based predictors. https://qualitynet.cms.gov/inpatient/measures/hybrid-hwr/methodology.

[3] Arun James Thirunavukarasu, Daniel Shu Wei Ting, Karthik Elangovan, Laura Gutierrez, Ting Fang Tan, and Dani Sumantir Yee Ting. Large language models in medicine. *Nature Medicine*, 29:1930–1940, 2023.

[4] Guergana K. Savova, James J. Masanz, Philip V. Ogren, Jiaping Zheng, Sunghwan Sohn, Karin C. Kipper-Schuler, and Christopher G. Chute. cTAKES: A natural language processing system for clinical text extraction and coding. *Journal of the American Medical Informatics Association*, 17(5):507–513, 2010.

[5] Alan R. Aronson and François-Michel Lang. An overview of MetaMap: historical perspective and recent advances. *Journal of the American Medical Informatics Association*, 17(3):229–236, 2010.

[6] Zeljko Kraljević, Thomas Searle, Anthony Shek, Zhewei Ding, Raul Santos-Rodriguez, Richard Bendayan, James T. Teo, and Daniel Bean. Multi-domain clinical natural language processing with MedCAT: The Medical Concept Annotation Toolkit. *Artificial Intelligence in Medicine*, 117:102083, 2021.

[7] Yinghao Zhu, Zixiang Wang, Junyi Gao, Yuning Tong, Jingkun An, Weibin Liao, Ewen M. Harrison, Liantao Ma, and Chengwei Pan. Prompting large language models for zero-shot clinical prediction with structured longitudinal electronic health record data. *arXiv preprint*

arXiv:2402.01713, 2024. Zero-shot GPT-4 on MIMIC-IV and TJH datasets; 35% AUROC improvement with elaborated prompting. https://arxiv.org/abs/2402.01713.

[8] Dianbo Sui, Yubo Chen, Kang Liu, and Jun Zhao. Federated learning for clinical natural language processing: A systematic review. *ACM Computing Surveys*, 56(11):1–38, 2024.

[9] Zhongwei Xu, Zhenyu Liu, Yujie Li, et al. A survey on deploying large language models on edge devices. *arXiv preprint arXiv:2409.16245*, 2024. https://arxiv.org/abs/2409.16245.

[10] Rinat Abdullin. Schema-guided reasoning (SGR), July 2025. https://abdullin.com/schema-guided-reasoning/.

[11] Tiago Almeida, Plinio Moreno, and Catarina Barata. Prediction of 30-day hospital readmission with clinical notes and EHR information. In *arXiv preprint arXiv:2503.23050*, 2025. GraphSAGE + LLM embeddings on MIMIC-IV; AUROC 0.72, balanced accuracy 66.7%. https://arxiv.org/abs/2503.23050.

[12] Health Level Seven International. HL7 FHIR release 4 (R4). https://hl7.org/fhir/R4/, 2023. Fast Healthcare Interoperability Resources. Accessed: 2026-03-01.

[13] Carl van Walraven, Irfan A. Dhalla, Chaim Bell, Edward Etchells, Ian G. Stiell, Kelly Zarnke, Peter C. Austin, and Alan J. Forster. Derivation and validation of an index to predict early death or unplanned readmission after discharge from hospital to the community. *Canadian Medical Association Journal*, 182(6):551–557, 2010.

[14] Jacques Donzé, Drahomir Aujesky, David Williams, and Jeffrey L. Schnipper. Potentially avoidable 30-day hospital readmissions in medical patients: Derivation and validation of a prediction model. *JAMA Internal Medicine*, 173(8):632–638, 2013.

[15] Özlem Uzuner, Brett R. South, Shuying Shen, and Scott L. DuVall. 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text. *Journal of the American Medical Informatics Association*, 18(5):552–556, 2011.

[16] Sam Henry, Kevin Buchan, Michele Filannino, Amber Stubbs, and Özlem Uzuner. 2018 n2c2 shared task on adverse drug events and medication extraction in electronic health records. *Journal of the American Medical Informatics Association*, 27(1):3–12, 2020.

[17] Kexin Huang, Jaan Altosaar, and Rajesh Ranganath. ClinicalBERT: Modeling clinical notes and predicting hospital readmission. *arXiv preprint arXiv:1904.05342*, 2019. https://arxiv.org/abs/1904.05342.

[18] Laila Rasmy, Yang Xiang, Ziqian Xie, Cui Tao, and Degui Zhi. Med-BERT: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction. *npj Digital Medicine*, 4(1):86, 2021.

[19] Edward Choi, Mohammad Taha Bahadori, Jimeng Sun, Joshua Kulas, Andy Schuetz, and Walter Stewart. RETAIN: An interpretable predictive model for healthcare using reverse time attention mechanism. In *Advances in Neural Information Processing Systems*, volume 29, 2016.

[20] Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Mohamed Amin, et al. Towards expert-level medical question answering with large language models. *arXiv preprint arXiv:2305.09617*, 2023. https://arxiv.org/abs/2305.09617.

[21] Tao Yang et al. MedGemma: Medical foundation models from Google. *arXiv preprint arXiv:2501.03986*, 2025. Health AI Developer Foundations (HAI-DEF). https://arxiv.org/abs/2501.03986.

[22] Gemma Team et al. Gemma: Open models based on Gemini research and technology. *arXiv preprint arXiv:2403.08295*, 2024. https://arxiv.org/abs/2403.08295.

[23] Sanjib Raj Pandey, Joy Dooshima Tile, and Mahdi Maktab Dar Oghaz. Predicting 30-day hospital readmissions using ClinicalT5 with structured and unstructured electronic health records. *PLoS One*, 20(9):e0328848, 2025. Hybrid ClinicalT5 + structured data on MIMIC-IV. https://doi.org/10.1371/journal.pone.0328848.

[24] Pietro Ferrazzi, Tiziano Labruna, Silvia Casola, Alberto Lavelli, and Bernardo Magnini. Small LLMs for medical NLP: a systematic analysis of few-shot, constraint decoding, fine-tuning and continual pre-training in Italian. In *arXiv preprint arXiv:2602.17475*, 2026. Systematic benchmark: LoRA-fine-tuned 1.7B models outperform 32B baselines (+9.2 F1) on 5 medical NLP tasks; continual pretraining rarely helps. https://arxiv.org/abs/2602.17475.

[25] Brandon T. Willard and Rémi Louf. Efficient guided generation for large language models. *arXiv preprint arXiv:2307.09702*, 2023. Outlines: guided generation and structured output. https://arxiv.org/abs/2307.09702.

[26] Luca Beurer-Kellner, Marc Fischer, and Martin Vechev. LMQL: Programming large language models with constraints. *arXiv preprint arXiv:2212.06094*, 2024. Query language for constrained LLM generation. https://arxiv.org/abs/2212.06094.

[27] Scott Lundberg and Marco Tulio Ribeiro. Guidance: A language for controlling large language models. *GitHub*, 2024. https://github.com/guidance-ai/guidance.

[28] Lianmin Zheng, Liangsheng Yin, Zhiqiang Xie, Chuyue Sun, Jeff Huang, Cody Hao Yu, Siyuan Cao, Christos Kozyrakis, Ion Stoica, Joseph E. Gonzalez, Clark Barrett, and Ying Sheng. SGLang: Efficient execution of structured language model programs. *arXiv preprint arXiv:2312.07104*, 2024. https://arxiv.org/abs/2312.07104.

[29] Jingwei Huang. CHiLL: Zero-shot custom interpretable feature extraction from clinical notes with large language models. *arXiv preprint arXiv:2302.12343*, 2023. Linear interpretable models on LLM-extracted features comparable to reference features for 30-day readmission. https://arxiv.org/abs/2302.12343.

[30] Mingze Shao, Yan Kang, Xin Hu, Hyun Gon Kwak, Chuanren Yang, and Juan Lu. Mining social determinants of health for heart failure patient 30-day readmission via large language model. In *Studies in Health Technology and Informatics*, 2025. Flan-T5 XL/XXL for SDoH extraction; macro-F1=0.71, patient-level recall 93.8%.

[31] Hao Yang, Zheng Shen, Jie Shao, Lining Men, Xuguang Han, and Jiaqiang Dong. LLM-augmented symptom analysis for cardiovascular disease risk prediction: A clinical NLP. *arXiv preprint arXiv:2507.11052*, 2025. Domain-adapted LLM symptom extraction; kappa=0.82, addresses hallucination via prompt engineering. https://arxiv.org/abs/2507.11052.

[32] Ofir Ben Shoham and Nadav Rappoport. CPLLM: Clinical prediction with large language models. *PLOS Digital Health*, 3(12):e0000680, 2024. End-to-end LLM clinical predictor; exceeds baselines in PR-AUC and ROC-AUC for readmission. https://doi.org/10.1371/journal.pdig.0000680.

[33] Anindya Choudhuri, Philip M. Polgreen, Alberto M. Segre, and Bijaya Adhikari. Summarizing clinical notes using LLMs for ICU bounceback and length-of-stay prediction. *medRxiv preprint*, 2025. LLM-generated summaries for MIMIC-III; +7.17% AUC-ROC for ICU bounceback, +14.16% AUPRC. https://doi.org/10.1101/2025.01.19.25320797.

[34] Ruichen Zhou, Chang Li, Chuanren Yang, and Juan Lu. ClinNoteAgents: An LLM multi-agent system for predicting and interpreting heart failure 30-day readmission from clinical notes. *arXiv preprint arXiv:2512.07081*, 2025. Multi-agent extraction, interpretation, and abstraction for HF readmission (3,544 notes, rate 35.16%). https://arxiv.org/abs/2512.07081.

[35] Omar Khattab, Arnav Singhvi, Paridhi Maheshwari, Zhiyuan Zhang, Keshav Santhanam, Sri Vardhamanan, Saiful Haq, Ashutosh Sharma, Thomas T. Joshi, Hanna Mober-Sheikhi, et al. DSPy: Compiling declarative language model calls into self-improving pipelines. *arXiv preprint arXiv:2310.03714*, 2023. https://arxiv.org/abs/2310.03714.

[36] Yu-Tai Lo, Jay Chiehen Liao, Mei-Hua Chen, Chia-Ming Chang, and Cheng-Te Li. Predictive modeling for 14-day unplanned hospital readmission risk by using machine learning algorithms. *BMC Medical Informatics and Decision Making*, 21(1):288, 2021. CatBoost with 21 features; AUROC 0.99, AUPRC 0.77 (14-day, N=24,722).

[37] Sara Nouri Golmaei and Xiao Luo. DeepNote-GNN: predicting hospital readmission using clinical notes and patient network. *arXiv preprint arXiv:2108.01342*, 2021. https://arxiv.org/abs/2108.01342.

[38] Gary S. Collins, Karel G. M. Moons, Paula Dhiman, Richard D. Riley, Andrew L. Beam, Ben Van Acker, et al. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ*, 385:e078378, 2024.

[39] Beau Norgeot, Giorgio Quer, Brett K. Beaulieu-Jones, Ali Torkamani, Raquel Dias, Melissa Laber, Deidre Colmary Griffin, Elise Regan, Evan D. Muse, Eric J. Topol, et al. Minimum information about clinical artificial intelligence modeling: the MI-CLAIM checklist. *Nature Medicine*, 26:1320–1324, 2020.

[40] Alistair E. W. Johnson, Lucas Bulgarelli, Lu Shen, Alvin Gayles, Ayad Shamma, Simon Delbecq, Sharon Duber, Benjamin Moody, Brian Gow, Li-wei H. Wei, Peter Szolovits, Leo Anthony Celi, and Roger G. Mark. MIMIC-IV, a freely accessible electronic health record dataset. *Scientific Data*, 10(1):1, 2023.

[41] Tianqi Chen and Carlos Guestrin. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 785–794. ACM, 2016.

[42] Ben Van Calster, David J. McLernon, Maarten van Smeden, Ewout W. Steyerberg, et al. Calibration: the Achilles heel of predictive analytics. *BMC Medicine*, 17(1):230, 2019.

[43] Alexandru Niculescu-Mizil and Rich Caruana. Predicting good probabilities with supervised learning. In *Proceedings of the 22nd International Conference on Machine Learning*, pages 625–632, 2005.

[44] Georgi Gerganov. llama.cpp: Inference of LLaMA model in pure C/C++. https://github.com/ggerganov/llama.cpp, 2023. Accessed: 2026-02-15.