Closed-Loop Validation-Repair for Healthcare Interoperability: A Multi-Model Study of Schema Compliance in Clinical LLMs

Jianru Shen

arXiv:2607.24371v1 [cs.CL] 27 Jul 2026

***Abstract*— Healthcare interoperability requires AI systems to produce structured outputs conforming to standardized schemas including ICD-10 for diagnostic coding, CPT for procedure billing, and HL7 FHIR for data exchange. While large language models demonstrate clinical reasoning capabilities, their integration into electronic health record systems faces a critical barrier: schema noncompliance. We evaluate three open-source models, Qwen2.5 7B, Llama 3.1 8B, and Gemma2 9B, via local deployment across 320 clinical scenarios spanning ten medical specialties, yielding 960 model–scenario pairs assessed under paired baseline and validation-repair conditions. First, schema noncompliance is consistent across the three model families, with baseline compliance rates ranging from 85.9 to 91.6 percent despite varying architectures and training data, suggesting shared gaps in medical training corpora rather than model-specific limitations. Second, 96 percent of validator-detected failures are representation-level format violations such as alternative medical abbreviations and code prefixes, indicating models follow clinical writing conventions but lack awareness of healthcare IT standards. Third, the validation-repair framework achieves 99.0 percent overall compliance, ranging from 98.4 to 99.4 percent across models, with most errors resolving within one or two iterations. Exact McNemar p-values below 0.001 and absolute improvements of 7.8 to 12.5 percentage points across model sizes confirm statistical significance. These results support closed-loop validation-repair as an effective system-level safeguard for healthcare interoperability, improving schema-level readiness for downstream clinical system integration.**

I. INTRODUCTION

Modern healthcare delivery relies on seamless interoperability among diverse information systems. Electronic health record platforms, built on standardized schemas including HL7 FHIR for resource representation [4], ICD-10 for diagnostic coding [5], and CPT for procedure billing [6], enable data exchange across institutions and clinical workflows. However, integrating AI-powered clinical decision support into these ecosystems presents a critical challenge. While large language models demonstrate expert-level clinical reasoning on medical examinations and diagnostic tasks [1]–[3], their outputs frequently violate schema constraints required for system interoperability.

We identify a systematic deployment barrier that transcends individual model architectures: clinical competence does not guarantee schema compliance. Models produce clinically valid outputs that violate healthcare IT standards, for example using the international abbreviation BD rather than the US standard BID for twice daily dosing. Such violations occur at rates of 8 to 14 percent and prevent EHR integration, where single format violations cause billing rejections or workflow disruptions.

This paper presents a multi-model empirical study addressing healthcare interoperability gaps in clinical LLMs. We investigate three research questions. First, is schema noncompliance architecture-agnostic across different LLM vendors and model sizes? Second, what types of errors cause schema violations and do they reflect medical knowledge gaps or format compliance issues? Third, can iterative validation-repair achieve deployment-oriented schema compliance exceeding 98 percent across diverse architectures?

Our multi-model findings support schema validation-repair as an effective closed-loop system component, where validation provides feedback signals and repair executes corrective control, independent of advances in individual model architectures. Figure 1 provides an overview of the framework.

II. RELATED WORK

A. LLMs in Clinical Applications

LLMs demonstrate strong performance on medical question answering, diagnostic reasoning, and clinical text tasks [1]–[3], [16]. However, these evaluations focus on clinical accuracy metrics such as exam scores and diagnostic correctness, assuming generated outputs conform to required formats rather than addressing the structured output compliance required for healthcare system integration.

B. Healthcare Interoperability Standards

Healthcare information exchange relies on standardized schemas: HL7 FHIR for resource-based clinical data representation [4], ICD-10 for diagnostic coding with structured alphanumeric formats [5], and CPT for procedure billing with strict five-digit numeric codes [6]. Fragmented data standards have long been recognized as a barrier to health information exchange [19]. Validation tools exist for human-authored data, but traditional validators report conformance errors and do not themselves generate corrected content [28]. Recent work has applied LLMs to map clinical text to FHIR resources [22] and has coupled an LLM with iterative FHIR-validator feedback for syntactic validity [23]. Our study differs in evaluating three locally deployed open-weight models under a common study-defined schema, with paired compliance statistics and a cross-model error taxonomy.

C. Structured Output Generation and Validation

Constrained decoding and grammar-guided generation enforce syntactic validity by restricting outputs to permitted structures [8], [9], but do not address semantic compliance

Jianru Shen is with the University of Montana, Missoula, MT 59812, USA js258133@umconnect.umt.edu

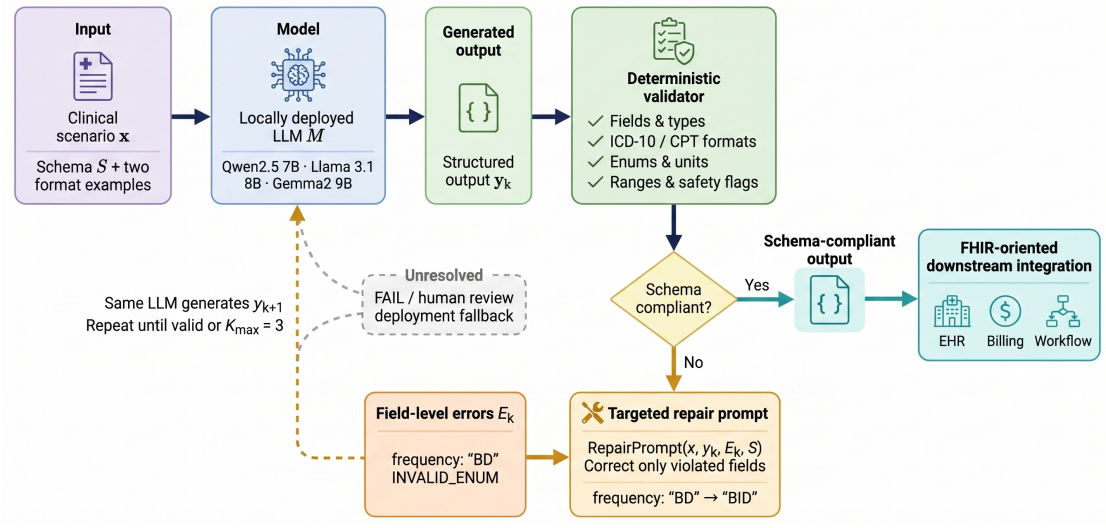

Fig. 1. Overview of the closed-loop validation-repair framework. A clinical scenario with schema specification and two format examples is processed by a locally deployed LLM; a deterministic validator checks fields, types, code formats, enumerations, ranges, and safety flags. Detected field-level errors are converted into a targeted repair prompt and the same model regenerates the output, repeating until compliance or $K_{\max} = 3$; unresolved scenarios are recorded as failures with human review as the proposed deployment fallback.

with domain-specific schemas. Grammar-constrained decoding extends such guarantees to a broad range of structured NLP tasks [29]; in contrast, our framework applies deterministic post-generation validation and targeted regeneration, keeping domain-specific checks external to the decoding engine. Self-refinement techniques including Reflexion [10] and Self-Refine [11] iteratively improve outputs through self-critique, but target task accuracy rather than format compliance. Constitutional AI aligns model behavior with specified principles rather than data format constraints [12].

D. Positioning and Contribution

Prior healthcare LLM studies typically evaluate single models, limiting conclusions about whether compliance issues are model-specific or systematic. Our multi-model evaluation shows that schema noncompliance is consistent across three vendors, and that validation alone is insufficient without repair. We contribute a paired multi-model compliance evaluation, an error taxonomy showing failures are predominantly representation-level format violations, and a convergence analysis of closed-loop repair.

III. VALIDATION-REPAIR FRAMEWORK

We present a closed-loop validation-repair framework that addresses schema noncompliance through cybernetic error correction [27]. The framework implements a feedback control architecture where validation acts as a compliance sensor detecting schema violations, and repair functions as a corrective controller generating schema-compliant outputs through targeted error feedback. This cybernetic approach enables vendor-independent compliance improvement while preserving the clinical reasoning capabilities of foundation models.

The framework consists of three components. First, we define healthcare interoperability requirements that LLM outputs must satisfy for downstream clinical system integration. Second, we describe the validation-repair algorithm that iteratively corrects schema violations. Third, we specify evaluation metrics that enable cross-model performance comparison and convergence analysis.

A. Healthcare Interoperability Requirements

Healthcare systems require LLM outputs to conform to standardized schemas that enable interoperability across electronic health record platforms, billing systems, and clinical workflows. Core requirements include diagnostic codes following ICD-10 format specifications, procedure codes adhering to CPT standards, medication frequencies using standardized clinical abbreviations, and dosage units from predefined medical vocabularies.

Specifically, outputs must satisfy the following constraints: diagnostic codes must match the ICD-10 regex pattern matching alphabetic prefix followed by numeric codes with optional decimal separators; procedure codes must be five-digit numeric strings or recognized standard letter codes; medication frequencies must belong to the enumeration set consisting of QD, BID, TID, QID, Q4H, Q6H, and PRN; dosage units must be selected from mg, ml, g, mcg, or units; and confidence scores must fall within the zero to one range. Additionally, safety flags indicating allergy checking and drug interaction screening must be set to true. These constraints constitute a structured interoperability schema incorporating FHIR-oriented fields together with ICD-10 and CPT format requirements. Validation targets code-format validity rather than full FHIR profile conformance or terminology-level verification that a given code exists in the current code set; the official FHIR $validate operation, by contrast, can assess resources against schemas, constraint rules, profiles, and terminology rules [28].

Algorithm 1 Closed-Loop Validation-Repair

Require: Clinical scenario x, LLM M, schema S, $K_{\max} = 3$
Ensure: Schema-compliant decision y, or FAIL

$y_0 \leftarrow M(x)$ {initial generation}
for $k = 0$ **to** $K_{\max}$ **do**
  $E_k \leftarrow \text{Validate}(y_k, S)$ {feedback sensor}
  if $E_k = \emptyset$ **then**
    return y_k {compliance achieved}
  end if
  if $k = K_{\max}$ **then**
    break
  end if
  $prompt \leftarrow \text{RepairPrompt}(x, y_k, E_k, S)$ {controller}
  $y_{k+1} \leftarrow M(prompt)$ {corrective action}
end for
return FAIL

Algorithm 2 Deterministic Field-Level Schema Validation

Require: Structured output y, schema S
Ensure: Field-level error set E

$E \leftarrow \emptyset$
for all required fields f in S **do**
  if f is absent from y **then**
    $E \leftarrow E \cup \{(f, \text{MISSING_FIELD})\}$; **continue**
  end if
  if $\text{Type}(y[f]) \neq S.\text{type}[f]$ **then**
    $E \leftarrow E \cup \{(f, \text{TYPE_MISMATCH})\}$; **continue**
  end if
  if $S.\text{enum}[f]$ is defined **and** $y[f] \notin S.\text{enum}[f]$ **then**
    $E \leftarrow E \cup \{(f, \text{INVALID_ENUM})\}$
  end if
  if $S.\text{pattern}[f]$ is defined **and** $y[f]$ does not match $S.\text{pattern}[f]$ **then**
    $E \leftarrow E \cup \{(f, \text{INVALID_FORMAT})\}$
  end if
  if $S.\text{range}[f]$ is defined **and** $y[f] \notin S.\text{range}[f]$ **then**
    $E \leftarrow E \cup \{(f, \text{OUT_OF_RANGE})\}$
  end if
end for
return E

Noncompliance with these requirements prevents integration with EHR systems, introduces billing errors that cause claim rejections, and disrupts clinical workflows. A single format violation can render an otherwise clinically accurate decision unusable for downstream system integration.

B. Validation-Repair Algorithm

Our framework employs iterative error-aware refinement to achieve schema compliance. Algorithm 1 presents the complete procedure.

The Validate function, formalized in Algorithm 2, performs deterministic field-level checking: for each required field it verifies presence, data type, enumeration membership, regex format including ICD-10 and CPT code structures, and numeric range, returning a detailed error set with field names and violated constraint types. Safety flags are enforced as single-value Boolean enumerations with $S.\text{enum}[f] = \{\text{true}\}$, so a value of false produces an INVALID_ENUM violation.

The RepairPrompt function constructs targeted error feedback for the model. Given the original scenario, current output, error list, and schema specification, it generates a prompt that explicitly identifies violated fields, specifies the violated constraints from the schema, and requests correction while preserving clinical accuracy. This error-aware approach enables the model to focus repair efforts on specific violations rather than regenerating the entire output.

Under this procedure, every generated output including the final repair attempt is validated before a scenario is recorded as a failure. In our evaluation runs, validation of the final-iteration outputs was applied to stored model outputs during analysis; because the validator is deterministic, this is equivalent to in-loop validation and yields identical compliance outcomes.

C. Evaluation Metrics

We employ four metrics to evaluate framework effectiveness and enable cross-model comparison.

Valid@K measures the percentage of scenarios achieving schema compliance by iteration k. This metric captures convergence efficiency: high Valid@1 indicates most errors resolve quickly, minimizing computational overhead. We report Valid@K for k from zero to three, where $k = 0$ represents baseline performance without repair.

Compliance Rate computes the binary success rate as the ratio of compliant scenarios to total scenarios. This serves as the primary metric for cross-model analysis, enabling direct comparison across models.

Error Distribution provides a taxonomy of violation types. We categorize errors into four classes: INVALID_ENUM for values outside allowed enumerations, INVALID_FORMAT for regex pattern mismatches, TYPE_MISMATCH for incorrect data types, and OUT_OF_RANGE for numeric values exceeding valid bounds; the validator additionally flags MISSING_FIELD for absent required fields, which did not occur in our evaluation. This taxonomy reveals whether failures stem from systematic patterns or random mistakes.

Statistical Significance testing employs McNemar's exact test for paired binary compliance outcomes, appropriate for comparing schema compliance before and after repair on the same clinical scenarios. We report exact two-sided p-values, absolute improvements in percentage points, 95 percent percentile confidence intervals for the improvement computed from 100,000 paired bootstrap resamples at the scenario level, and the relative reduction in noncompliance.

IV. EXPERIMENTAL SETUP

We conduct a multi-model evaluation to assess schema compliance across diverse architectures and identify whether noncompliance patterns are model-specific or systematic. Our experimental design consists of two components: a clinical scenario benchmark spanning ten medical specialties, and a multi-vendor model selection strategy that enables robust conclusions about architecture generalizability.

A. Clinical Scenario Benchmark

We constructed a benchmark of 320 clinical scenarios across ten medical specialties, each following the structured case-vignette format used in medical education and licensing examinations. Each scenario represents a typical clinical presentation documented in practice guidelines and major medical textbooks. Scenarios were designed based on cardiology guidelines from the American Heart Association and American College of Cardiology [20], infectious disease case definitions from the Centers for Disease Control and Prevention, and clinical presentations described in Harrison's Principles of Internal Medicine [13] and Kumar and Clark's Clinical Medicine [14].

Because the purpose of this benchmark is schema compliance evaluation rather than clinical diagnostic accuracy assessment, scenarios were designed as guideline-grounded structured prompts to test interoperability requirements rather than validated patient cases requiring clinical expert review.

Each scenario includes three components: patient demographics including age and relevant medical history, chief complaint describing the primary reason for clinical encounter, and key clinical findings sufficient for structured diagnostic and treatment decision-making. This structure mirrors real-world clinical documentation while providing sufficient information for schema-compliant output generation.

The dataset distribution was informed by encounter frequencies reported in national ambulatory care surveys [15]. Cardiology accounts for 48 scenarios at 15.0 percent, Respiratory Medicine and Infectious Disease each contain 40 scenarios at 12.5 percent, Gastroenterology and Neurology each include 32 scenarios at 10.0 percent, General Medicine comprises 36 scenarios at 11.3 percent, Endocrinology and Trauma/Orthopedics each contain 26 scenarios at 8.1 percent, and Dermatology and Psychiatry each include 20 scenarios at 6.3 percent. This distribution totals 320 scenarios providing sufficient statistical power for per-specialty analysis while maintaining practical dataset size.

B. Multi-Model Evaluation Design

We evaluate three open-source LLMs via local deployment using Ollama, selecting models from different companies to test whether schema noncompliance is architecture-specific or systematic across vendors. The selected models are Qwen2.5 7B [24] from Alibaba Cloud representing smaller models with non-Western training emphasis, Llama 3.1 8B [25] from Meta representing mid-size widely adopted architectures, and Gemma2 9B [26] from Google DeepMind representing larger instruction-tuned models.

This multi-vendor selection enables robust conclusions about noncompliance patterns. Consistent failure rates across the three companies would indicate shared characteristics of medical training corpora rather than model-specific architectural limitations. Divergent patterns would suggest vendor-specific issues addressable through training improvements. The parameter range from 7B to 9B tests whether model size influences schema compliance while remaining within practical deployment constraints for local healthcare environments.

TABLE I
MULTI-MODEL PERFORMANCE ACROSS 320 CLINICAL SCENARIOS

Model	Strong Baseline	Strong+Repair	Improvement (pp)
Qwen2.5 7B (Alibaba)	91.6% (293/320)	99.4% (318/320)	+7.8
Llama 3.1 8B (Meta)	85.9% (275/320)	98.4% (315/320)	+12.5
Gemma2 9B (Google)	89.4% (286/320)	99.1% (317/320)	+9.7
Overall	**89.0% (854/960)**	**99.0% (950/960)**	**+10.0**

Our evaluation protocol compares two conditions for each model. The strong baseline employs explicit schema specification in the prompt, representing current best practices for structured output generation. The strong plus repair condition applies our validation-repair framework to the same prompts. We use temperature zero to ensure deterministic reproducible outputs across all experiments. Each model processes all 320 scenarios once per condition, yielding 960 model–scenario pairs.

All experiments use identical prompt templates and schema specifications to ensure fair comparison. The prompt includes the clinical scenario, explicit field requirements, schema constraints, and two example outputs demonstrating proper format, i.e., few-shot prompting with explicit schema specification. This strong baseline establishes an upper bound on what current prompting techniques can achieve without post-hoc validation and repair.

V. RESULTS AND FINDINGS

Our multi-model evaluation across 960 model–scenario pairs reveals three key findings. First, schema noncompliance is consistent across the three evaluated vendors despite different training approaches. Second, error taxonomy shows 96 percent of failures are representation-level format violations. Third, validation-repair achieves consistent deployment-oriented schema compliance across all architectures with practical convergence within two iterations. We present detailed analyses addressing each research question.

A. RQ1: Cross-Vendor Consistency of Noncompliance

Table I presents compliance rates across the three models and two evaluation conditions.

Our multi-model evaluation addresses RQ1 by revealing that schema noncompliance is consistent across vendors. Despite different companies, training data sources, and parameter counts ranging from 7B to 9B, all three models exhibit remarkably similar baseline compliance rates spanning only 5.7 percentage points from 85.9 to 91.6 percent. This narrow variance suggests the problem stems from shared characteristics of medical training corpora, which emphasize

TABLE II
ERROR TAXONOMY BEFORE REPAIR

Error Type	Qwen	Llama	Gemma	Total
INVALID_ENUM	22	11	21	54 (51%)
INVALID_FORMAT	5	32	11	48 (45%)
TYPE_MISMATCH	0	2	0	2 (2%)
OUT_OF_RANGE	0	0	2	2 (2%)
Noncompliant Outputs	**27**	**45**	**34**	**106**

clinical accuracy over healthcare IT standards, rather than model-specific architectural limitations.

Even the best baseline model still produces 8.4 percent noncompliant outputs, while the weakest baseline produces 14.1 percent noncompliant outputs. This error rate remains problematic for healthcare system integration where single violations cause billing rejections or workflow disruptions. Notably, the smallest model at 7B parameters achieves the highest baseline compliance at 91.6 percent, demonstrating that parameter count alone does not predict interoperability readiness.

B. RQ2: Error Taxonomy and Root Causes

Table II presents error distributions across models and violation types. Each noncompliant output is counted once by its first detected violation; mandatory-field presence is likewise enforced, and no missing-field violations occurred in our evaluation, so the table reports the four observed error categories.

Cross-model error analysis addressing RQ2 reveals a critical insight: 96 percent of validator-detected failures are representation-level format violations. INVALID_ENUM and INVALID_FORMAT together account for 102 of the 106 noncompliant outputs. This pattern remains consistent across all three models with ENUM plus FORMAT comprising 100 percent for Qwen, 96 percent for Llama, and 94 percent for Gemma.

Representative error cases demonstrate models follow clinical writing conventions while violating healthcare IT standards. INVALID_ENUM errors include using BD instead of BID, where both terms mean twice daily but the former represents international Latin convention while the latter is the required US standard for EHR systems. Case sensitivity violations such as qd versus QD also appear. INVALID_FORMAT errors include CPT_74140 instead of 74140, where the prefix adds semantic clarity for human readers but violates billing system requirements expecting bare numeric codes. Period formatting inconsistencies such as I21.0 versus I210 and coding system mismatches such as G0438 representing HCPCS codes instead of CPT codes also occur. These errors reveal models trained on medical literature naturally learn international conventions and semantic annotations that improve human readability in research contexts but violate strict standardization required for healthcare system interoperability. Because clinical correctness was not independently evaluated in this study, we characterize these violations as representation-level rather than knowledge-level

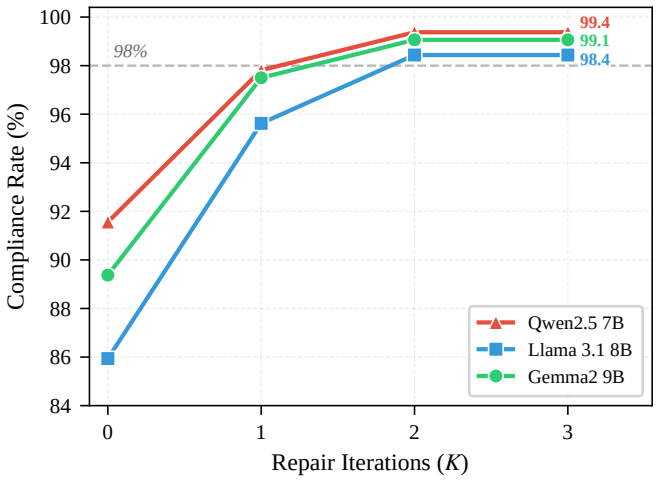

Fig. 2. Valid@K convergence across three LLMs, computed from the cumulative convergence counts in Table IV. All models exceed the 98 percent threshold by iteration two, and no additional scenarios converge at the third iteration.

TABLE III
STATISTICAL ANALYSIS OF THE REPAIR FRAMEWORK

Model	Impr./Regr.	Δ (pp)	95% CI (pp)	Exact p
Qwen2.5 7B	25 / 0	+7.8	[5.0, 10.9]	5.96×10^{-8}
Llama 3.1 8B	40 / 0	+12.5	[9.1, 16.2]	1.82×10^{-12}
Gemma2 9B	31 / 0	+9.7	[6.6, 13.1]	9.31×10^{-10}

errors. This finding nonetheless challenges assumptions that improved medical training corpora alone will solve deployment readiness: models can produce clinically conventional outputs while remaining systematically noncompliant with interoperability standards.

C. RQ3: Validation-Repair Effectiveness

Validation-repair achieves deployment-oriented schema compliance exceeding 98 percent across all three architectures. Table III confirms highly significant improvements: every scenario that changed compliance status moved from noncompliant to compliant, with 25, 40, and 31 improved scenarios and zero regressions, exact McNemar p-values below 10^{-7}, and absolute improvements ranging from 7.8 to 12.5 percentage points. These improvements correspond to relative reductions in noncompliance of 92.6, 88.9, and 91.2 percent for Qwen, Llama, and Gemma respectively, demonstrating robust evidence for framework effectiveness independent of model architecture.

Figure 2 and Table IV reveal practical convergence patterns. By iteration one, an average of 97.0 percent of scenarios achieve compliance, showing that most violations are corrected after a single feedback step. By iteration two, all models reach at least 98 percent compliance with Valid@2 ranging from 98.4 to 99.4 percent, supporting downstream interoperability requirements. No additional scenarios converge at the third iteration, indicating diminishing returns beyond two repair attempts; $K_{\max} = 3$ therefore serves as a conservative termination cap. The framework is also compu-

TABLE IV
CONVERGENCE DISTRIBUTION ACROSS ITERATIONS

Model	K=0 (Baseline)	K=1	K=2	Failed
Qwen2.5 7B	91.6% (293)	6.3% (20)	1.6% (5)	0.6% (2)
Llama 3.1 8B	85.9% (275)	9.7% (31)	2.8% (9)	1.6% (5)
Gemma2 9B	89.4% (286)	8.1% (26)	1.6% (5)	0.9% (3)

tationally light: including repair calls, it requires an average of 1.11, 1.20, and 1.14 LLM invocations per scenario for Qwen, Llama, and Gemma respectively.

Critically, convergence patterns remain consistent across the three models despite architectural differences. All models reach similar Valid@1 rates spanning 95.6 to 97.8 percent and Valid@2 rates spanning 98.4 to 99.4 percent. This cross-vendor convergence demonstrates the framework provides reliable interoperability compliance independent of model vendor or size, addressing RQ3 affirmatively with both statistical and practical validation.

a) Residual failures: Ten model–scenario pairs remained noncompliant after bounded repair, comprising two for Qwen2.5 7B, five for Llama 3.1 8B, and three for Gemma2 9B. A representative residual violation was the composite dosage unit `mg/kg/hour`, which falls outside the restricted unit enumeration {`mg`, `ml`, `g`, `mcg`, `units`} used in this study. Because compliance is defined conjunctively across all required fields, a single remaining field-level violation causes the complete output to remain noncompliant even when other violations have been corrected. These unresolved cases motivate the proposed human-review fallback after the bounded repair loop. This analysis concerns conformance to the study schema and does not assess the clinical correctness of the generated content.

D. Cross-Specialty Performance Analysis

Figure 3 shows compliance rates across ten medical specialties and three models.

Cross-specialty analysis reveals variation in baseline performance across medical domains. The most challenging specialties include Gastroenterology for Llama at 81.3 percent, Neurology for Llama at 84.4 percent, and Psychiatry for Qwen and Llama at 85.0 percent. These lower rates reflect domain-specific terminology variations and coding complexity rather than fundamental model limitations. Post-repair evaluation not shown in the figure demonstrates all model-specialty combinations achieve 95 percent or higher compliance, confirming framework robustness across diverse clinical contexts.

VI. DISCUSSION

Our findings support schema validation-repair as an effective system-level component for healthcare AI deployment. We discuss deployment implications, training data characteristics, and threats to validity.

A. Implications for Healthcare Interoperability

Our findings suggest that schema validation-repair should be integrated as a standard middleware layer in healthcare AI systems [17], [18], independent of model vendor selection: with baseline noncompliance consistent across three companies, the problem is unlikely to be solved by model choice alone. Validation-repair thus plays a role analogous to HL7 interface engines that ensure interoperability between heterogeneous clinical systems [7].

Figure 4 illustrates a cybernetic architecture with four functional layers: the LLM as generative component, validation as feedback sensor, repair as corrective controller with empirical convergence within two iterations, and human clinical oversight as a fallback layer handling the approximately one percent of cases that exceed automated correction. This layered architecture remains valid across model vendors, enabling LLM adoption without vendor lock-in.

Notably, the smallest model achieves the highest post-repair compliance: Qwen at 7B parameters reaches 99.4 percent, slightly exceeding Gemma2 9B at 99.1 percent. This enables lower-compute local deployment attractive for privacy-conscious healthcare settings, though regulatory compliance itself was not evaluated in this study.

B. Why Closed-Loop Repair Instead of Stronger Prompting?

A practical question is whether stronger prompting alone suffices. The strong baseline already constitutes few-shot prompting with explicit schema constraints and example outputs, yet all three models still produce noncompliant outputs at nontrivial rates, indicating a mismatch between clinical language conventions learned during training and strict interoperability requirements. Closed-loop repair addresses this by separating generation from compliance control: the LLM generates clinical content, the validator provides deterministic feedback, and the repair step performs targeted correction. Because compliance requirements are explicit and checkable, this division of responsibility is more reliable than expecting the model to internalize all schema constraints in a single pass.

C. Training Data Characteristics and Implications

The consistency of format violations across all three vendors suggests shared characteristics of medical training data: medical literature emphasizes clinical accuracy and international conventions, teaching models that BD and BID are interchangeable without distinguishing that US EHR systems require BID. Incorporating healthcare IT standards documentation such as HL7 specifications and coding guidelines into training corpora could help, but given the cross-vendor consistency of this gap, post-hoc validation-repair provides a more immediate solution that applies to any current or future model.

D. Threats to Validity and Future Directions

Schema compliance is evaluated using deterministic validators, so results depend on the selected schema constraints;

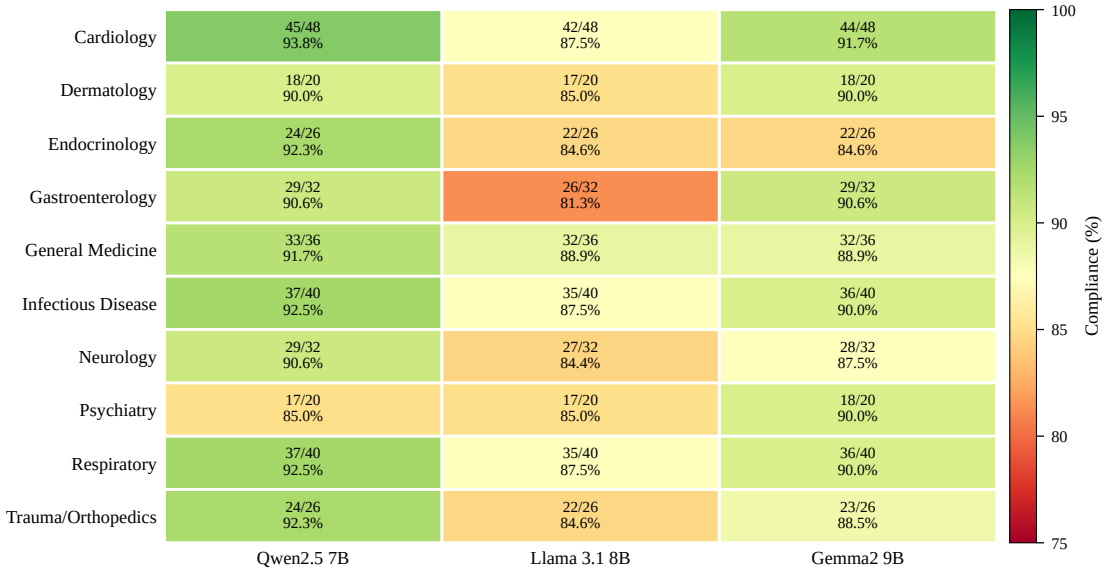

Fig. 3. Baseline compliance by model and medical specialty. Each cell reports compliant scenarios over specialty size together with the corresponding percentage; cell color encodes the compliance percentage. The most challenging combinations include Gastroenterology for Llama 3.1 8B at 81.3 percent and Psychiatry for Qwen2.5 7B and Llama 3.1 8B at 85.0 percent. Post-repair results not shown achieve 95 percent or higher across all combinations.

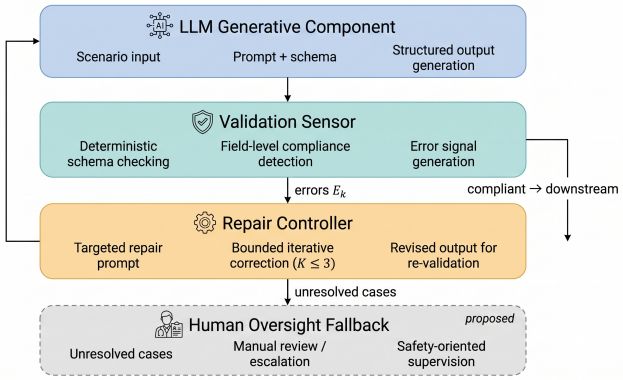

Fig. 4. Suggested four-layer deployment architecture for closed-loop compliance control. Compliant outputs exit after validation; the human oversight layer (dashed) is a proposed fallback for unresolved cases and was not evaluated in this study.

our validators check code formats rather than terminology-level validity, so a well-formed but nonexistent code would not be flagged. The benchmark uses guideline-grounded structured scenarios rather than live EHR records, so results characterize schema compliance behavior under controlled interoperability tasks rather than complete deployment readiness. Finally, schema compliance does not imply clinical correctness: our study intentionally isolates interoperability compliance from diagnostic accuracy. Content-level reliability, including hallucination detection for retrieval-augmented generation [21], is complementary to the schema-level compliance studied here.

Our evaluation is further limited to predefined schemas rather than free-text narratives, ten specialties, three open-source models of 7B to 9B parameters, and a static schema with single-turn prompts. Future work should evaluate additional and larger models, real-world and evolving institutional schemas, conversational clinical workflows, and hybrid rule-based and LLM-based repair, as well as prompting ablations and fine-tuning with compliance-based feedback signals.

VII. CONCLUSIONS

Through systematic multi-model evaluation across 960 model–scenario pairs, this work establishes that schema noncompliance presents a cross-vendor barrier to healthcare interoperability: three models from different vendors exhibit baseline compliance rates within 5.7 percentage points of one another, and 96 percent of failures are representation-level format violations such as international abbreviations and annotated code prefixes. This consistency suggests the problem stems from shared characteristics of medical training corpora and is unlikely to be solved by improved medical training alone.

Validation-repair achieves 99.0 percent overall compliance with practical convergence within two iterations, with statistically significant improvements of 7.8 to 12.5 percentage points across all models. Notably, the smallest model achieves the highest post-repair compliance, supporting lower-compute local deployment.

Our findings have direct implications for clinical AI deployment. Closed-loop validation-repair should be integrated as a standard system layer to support interoperability with electronic health records, independent of model vendor selection. This cybernetic approach enables healthcare organizations to adopt LLMs while maintaining schema-level compliance with FHIR-oriented, ICD-10, and CPT format requirements, transforming clinically capable but noncompliant models into interoperability-ready decision support systems suitable for downstream clinical system integration.

REFERENCES

[1] K. Singhal, S. Azizi, T. Tu, S. S. Mahdavi, J. Wei, H. W. Chung, et al., "Large language models encode clinical knowledge," *Nature*, vol. 620, pp. 172–180, 2023.

[2] K. Singhal, T. Tu, J. Gottweis, R. Sayres, E. Wulczyn, L. Hou, et al., "Towards expert-level medical question answering with large language models," *arXiv preprint arXiv:2305.09617*, 2023.

[3] H. Nori, N. King, S. M. McKinney, D. Carignan, and E. Horvitz, "Capabilities of GPT-4 on medical challenge problems," *arXiv preprint arXiv:2303.13375*, 2023.

[4] HL7 International, "HL7 FHIR Release 4," Health Level Seven International, 2019.

[5] World Health Organization, "International Statistical Classification of Diseases and Related Health Problems, 10th Revision," 2016.

[6] American Medical Association, "Current Procedural Terminology," Chicago, IL: American Medical Association, 2023.

[7] K. D. Mandl and I. S. Kohane, "No small change for the health information economy," *New England Journal of Medicine*, vol. 360, no. 13, pp. 1278–1281, Mar. 2009.

[8] C. Hokamp and Q. Liu, "Lexically constrained decoding for sequence generation using grid beam search," in *Proc. 55th Annual Meeting of the Association for Computational Linguistics*, Vancouver, Canada, 2017, pp. 1535–1546.

[9] T. Scholak, N. Schucher, and D. Bahdanau, "PICARD: Parsing incrementally for constrained auto-regressive decoding from language models," in *Proc. 2021 Conference on Empirical Methods in Natural Language Processing*, 2021.

[10] N. Shinn, F. Cassano, A. Gopinath, K. Narasimhan, and S. Yao, "Reflexion: Language agents with verbal reinforcement learning," in *Proc. NeurIPS*, 2023.

[11] A. Madaan, N. Tandon, P. Gupta, S. Hallinan, L. Gao, S. Wiegreffe, et al., "Self-refine: Iterative refinement with self-feedback," in *Proc. NeurIPS*, 2023.

[12] Y. Bai, S. Kadavath, S. Kundu, A. Askell, J. Kernion, A. Jones, et al., "Constitutional AI: Harmlessness from AI feedback," *arXiv preprint arXiv:2212.08073*, 2022.

[13] J. Loscalzo, A. S. Fauci, D. L. Kasper, S. Hauser, D. Longo, and J. L. Jameson, *Harrison's Principles of Internal Medicine*, 21st ed. McGraw-Hill Education, 2022.

[14] A. Feather, D. Randall, and M. Waterhouse, eds., *Kumar and Clark's Clinical Medicine*, 10th ed. Elsevier, 2020.

[15] L. Santo and K. Kang, "National Ambulatory Medical Care Survey: 2019 National Summary Tables," National Center for Health Statistics, 2023.

[16] K. Huang, J. Altosaar, and R. Ranganath, "ClinicalBERT: Modeling clinical notes and predicting hospital readmission," *arXiv preprint arXiv:1904.05342*, 2019.

[17] A. Wright, D. F. Sittig, J. S. Ash, J. Feblowitz, S. Meltzer, C. McMullen, K. Guappone, J. Carpenter, J. Richardson, L. Simonaitis, R. S. Evans, W. P. Nichol, and B. Middleton, "Development and evaluation of a comprehensive clinical decision support taxonomy: comparison of front-end tools in commercial and internally developed electronic health record systems," *Journal of the American Medical Informatics Association*, vol. 18, no. 3, pp. 232–242, May 2011.

[18] E. S. Berner, ed., *Clinical Decision Support Systems: Theory and Practice*, 3rd ed. Cham, Switzerland: Springer, 2016.

[19] R. D. Kush, E. Helton, F. W. Rockhold, and C. D. Hardison, "Electronic health records, medical research, and the Tower of Babel," *New England Journal of Medicine*, vol. 358, no. 16, pp. 1738–1740, Apr. 2008.

[20] P. T. O'Gara, F. G. Kushner, D. D. Ascheim, D. E. Casey, M. K. Chung, J. A. de Lemos, et al., "2013 ACCF/AHA guideline for the management of ST-elevation myocardial infarction: A report of the American College of Cardiology Foundation/American Heart Association Task Force on Practice Guidelines," *Circulation*, vol. 127, no. 4, 2013.

[21] J. Shen, "Evidence graph consistency in retrieval-augmented generation: A model-dependent analysis of hallucination detection," *arXiv preprint arXiv:2606.06748*, 2026.

[22] Y. Li, H. Wang, H. Z. Yerebakan, Y. Shinagawa, and Y. Luo, "FHIR-GPT enhances health interoperability with large language models," *NEJM AI*, vol. 1, no. 8, Art. no. AIcs2300301, Jul. 2024.

[23] P. Tabari, A. Piscitelli, G. Costagliola, and M. de Rosa, "Assessing the potential of an LLM-powered system for enhancing FHIR resource validation," *Studies in Health Technology and Informatics*, vol. 327, pp. 803–807, 2025.

[24] Qwen Team, "Qwen2.5 technical report," *arXiv preprint arXiv:2412.15115*, 2024.

[25] A. Grattafiori, et al., "The Llama 3 herd of models," *arXiv preprint arXiv:2407.21783*, 2024.

[26] Gemma Team, "Gemma 2: Improving open language models at a practical size," *arXiv preprint arXiv:2408.00118*, 2024.

[27] N. Wiener, *Cybernetics: Or Control and Communication in the Animal and the Machine*, 2nd ed. Cambridge, MA: MIT Press, 1961.

[28] HL7 International, "Resource $validate operation," HL7 FHIR Release 4 (v4.0.1), 2019.

[29] S. Geng, M. Josifoski, M. Peyrard, and R. West, "Grammar-constrained decoding for structured NLP tasks without finetuning," in *Proc. 2023 Conference on Empirical Methods in Natural Language Processing*, Singapore, 2023, pp. 10932–10952.