# MedGap: Detecting Clinical Inconsistencies in Electronic Health Records Using a Hybrid Rules + LLM Framework

David Lutala[a]

[a]Vietnam National University, Hanoi, Vietnam

***Abstract*—Electronic Health Records (EHRs) frequently contain inconsistencies such as prescriptions contradicting documented allergies, discordant clinical observations, or missing elements in the care plan. These issues increase clinicians' cognitive load and contribute to preventable adverse events. We introduce MedGap, a lightweight framework that combines formal clinical rules with a controlled Large Language Model (LLM) to automatically detect inconsistencies within FHIR structured EHR data. The approach integrates (1) clinical information extraction using NER and relation extraction, (2) a structured and auditable rule engine, and (3) a specialized LLM designed to identify narrative contradictions that are difficult to encode symbolically. The LLM module is wrapped within a strict control mechanism enforcing FHIR-grounded justification, a normalized response structure, and automatic validation through an NLI classifier, thereby reducing hallucinations and false positives. A preliminary evaluation on 120 de-identified patient records annotated by two clinicians (k = 0.82) shows that MedGap detects 92% of critical inconsistencies an improvement of +21% over rules alone while maintaining a controlled false positive rate.**

***Index Terms*—Electronic patient record, LLM, clinical decision support systems, FHIR, patient safety.**

## I. Introduction

Clinical notes, observations, and prescriptions within Electronic Health Records (EHRs) are often heterogeneous, redundant, or temporally misaligned. Manually identifying inconsistencies increases clinicians' cognitive load, slows decision-making, and exposes patients to preventable errors. Rule-based systems detect explicit inconsistencies effectively, yet they struggle with narrative contradictions for example, a note describing a patient as "stable" while vital signs are abnormal. Conversely, Large Language Models (LLMs) can capture such contextual signals but lack operational guarantees due to risks of hallucinations and insufficient justification.

MedGap adopts an intermediate strategy: audited clinical rules for critical inconsistencies, combined with a controlled LLM to identify narrative incoherence. For instance, MedGap flags a documented penicillin allergy when amoxicillin is prescribed, or detects a note reporting a "stable" patient despite an oxygen saturation of 82%.

## II. Method

### A. FHIR Ingestion and Clinical Extraction

FHIR resources including `MedicationRequest`, `Observation`, `AllergyIntolerance`, and clinical notes are retrieved from the EHR. A NER and relation-extraction pipeline (BERT-CRF) identifies allergies, medications and dosages, vital signs, diagnoses, and their structured relationships.

### B. Rule Engine

A set of explicit clinical rules covering thresholds, drug incompatibilities, and allergy prescription constitutes the first detection layer. These rules always prioritize critical alerts and ensure that all formalizable inconsistencies are captured in an auditable manner, without relying on the language model.

### C. LLM Module for Narrative Inconsistencies

A specialized LLM (13B model adapted on de-identified clinical notes) is invoked only when rules are insufficient. Prompting enforces strict constraints:

- justification grounded in FHIR snippets;
- a normalized response format (conclusion, evidence);
- refusal to generate an alert without a source.

LLM outputs are subsequently filtered by a lightweight Natural Language Inference (NLI) module. This LLM+NLI pairing forms a safety barrier, preventing unjustified alerts and strengthening clinical traceability.

### D. Hybrid Combination

MedGap merges signals into three alert levels: Critical (triggered by explicit rule violations), Moderate (LLM detected inconsistencies validated by NLI) and Informational (non-blocking suggestions)

This hierarchy facilitates operational integration into clinical workflows by preventing non-prioritized or noisy alerting.

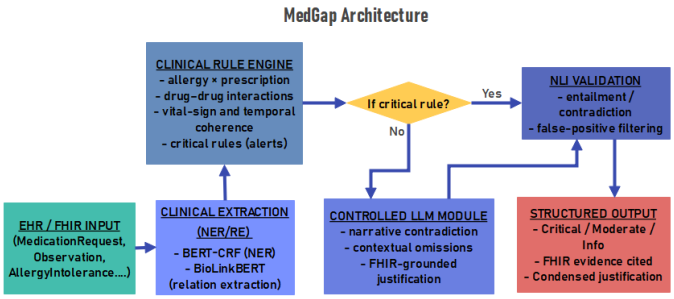


Fig. 1: Overall architecture of MedGap

## III. Evaluation

### A. Corpus annoté

The evaluation relies on a corpus of 120 de-identified patient records from four hospital departments (emergency medicine, infectious diseases, pulmonology, and internal medicine). Each record includes structured FHIR resources (MedicationRequest, Observation, Condition, ...) as well as narrative notes. Three categories of inconsistencies were annotated:

- critical (direct patient impact, e.g., allergy–medication conflicts),
- moderate (incoherent vital signs or dosages),
- narrative (contradictions between notes).

Inter-annotator agreement reached k = 0.82, confirming the reliability of the annotations despite task complexity.

### B. Baselines

We compared MedGap against two representative approaches:

1) Rules only: a deterministic engine without narrative analysis.
2) LLM only: the model detects inconsistencies from text + FHIR but without structural constraints.
3) MedGap (hybrid): controlled combination of `rules + LLM + NLI`.

These baselines isolate:

- the LLM's contribution to narrative coherence,
- the robustness provided by explicit rules,
- the true added value of the hybrid design.

### C. Results

The results indicate that each approach addresses a different facet of the problem, and that the hybrid combination is required for a stable system.

- Critical inconsistencies: MedGap achieves F1 = 0.92 (95% CI: [0.89–0.96]), clearly outperforming rules alone (0.76). This confirms that the LLM fills gaps where rules fall short (implicit formulations, allergies mentioned only in notes, etc.).
- Narrative inconsistencies: MedGap slightly surpasses the LLM alone (0.71 vs. 0.68) while substantially reducing false positives (7% vs. 19%). The NLI validation step is essential for controlling hallucinations.
- Moderate inconsistencies: improvements are smaller but meaningful: the hybrid model reaches F1 = 0.63 vs. 0.41 for rules alone.

Overall, the results validate the relevance of a constrained rules + LLM coupling and demonstrate that a relatively small model (13B) can perform effectively when structured by explicit logic.

### D. Error Analysis

Qualitative analysis highlights three main sources of limitations:

1) Insufficient temporal handling: Several errors arise from imprecise timestamps, creating apparent contradictions (e.g., outdated vital signs). This suggests the need for a more robust clinical timeline model.
2) Narrative variability within the EHR: Discrepancies between nursing and physician notes remain difficult to interpret automatically, especially when contradictions are implicit.
3) Overly conservative NLI filtering: While effective at reducing hallucinations, the NLI module sometimes rejects true inconsistencies when supporting evidence is fragmented across multiple notes.

These limitations highlight natural avenues for improvement without undermining the validity of the results obtained.

## IV. Discussion

MedGap demonstrates that a structured hybrid framework enables more stable inconsistency detection than purely symbolic or purely neural approaches:

- Auditability: critical inconsistencies remain explainable and traceable through explicit rules, which is essential in clinical settings.
- Hallucination control: MedGap maintains a low false-positive rate thanks to automatic NLI-based validation.
- Handling of narrative contradictions: a domain where deterministic rules systematically fail.
- Facilitated integration: the use of native FHIR structures makes the system deployable incrementally.

The identified limitations : modest corpus size, still basic temporal modeling, and the absence of real-world workflow evaluation do not diminish the relevance of the framework but highlight the need for broader validation.

In a real clinical environment, MedGap could appear as a notification at the moment a prescription is validated in the EHR or be integrated into a daily documentation coherence review dashboard.

The results remain preliminary but promising for a lightweight system aimed at improving documentation consistency in hospital settings.

## V. Conclusion

MedGap introduces a hybrid approach that leverages the complementarity between auditable clinical rules and a controlled LLM. Evaluation on 120 patient records demonstrates a substantial improvement in detecting critical inconsistencies (+21% compared to rules alone) and a notable reduction in false positives relative to an unconstrained LLM.

By relying on FHIR-native inputs, robust clinical information extraction, and systematic NLI-based validation, MedGap achieves a balanced compromise between performance, traceability, and safety. Future work includes:

Les prochaines étapes incluent :

- enhancing temporal modeling,
- expanding the annotated corpus, and
- conducting a pilot clinical deployment to assess real-world impact on cognitive load and documentation quality.