ARIS: A Fail-Closed LLM Framework for Uncertainty-Controlled Allergy Knowledge Graph Expansion

Anonymous Author(s)

Abstract

Large language models (LLMs) enable large-scale extraction of clinical signals from electronic health records, but directly inserting model-derived outputs into persistent knowledge graphs (KGs) risks propagating unsupported, weakly grounded or low-confidence information into downstream reasoning systems. We present the Allergy Risk Intelligence System (ARIS), an uncertainty-controlled framework for auditable KG expansion that replaces free-form generation with structured signal extraction and gated KG updates. Rather than treating provenance and uncertainty as post-hoc annotations, ARIS formulates KG expansion as an admission-control problem in which candidate updates are incorporated into persistent KGs when they satisfy deterministic audit constraints, Bayesian uncertainty thresholds, and provenance-completeness requirements. The framework combines encoder-based semantic representations, learned audit patterns, Bayesian uncertainty decomposition, and an uncertainty-weighted retraining buffer for conservative KG evolution. We instantiate ARIS for allergy-related KG expansion using MIMIC-IV electronic records and evaluate the framework on large-scale clinical extraction and routing tasks. Across 2000 evaluation records, ARIS achieves complete routing and provenance coverage, 99.27% audit specificity, 99.16% audit precision, and zero uncontrolled KG update events. The framework further demonstrates stable uncertainty decomposition, with epistemic uncertainty accounting for 97.0% of total uncertainty while remaining empirically uncorrelated from aleatoric noise ($r = 0.003$). Despite enforcing provenance-complete admission control, the Bayesian routing module contains only 1.12M parameters and incurs 0.042% computational overhead relative to the 4.30B-parameter base encoder. These results show that selective, provenance-aware admission control can support traceable and uncertainty-aware KG evolution under noisy clinical text, providing a general framework for reliable AI-mediated knowledge integration in regulated information domains.

CCS Concepts

• **Information systems** → **Information extraction**; • **Computing methodologies** → *Knowledge representation and reasoning*.

Keywords

knowledge graph expansion; information extraction; uncertainty-aware knowledge management

1 Introduction

Electronic health records (EHRs) contain large volumes of unstructured clinical narratives describing medications, laboratory findings, procedures, adverse reactions, and longitudinal patient histories [20, 40, 44]. Clinical knowledge graphs (KGs) derived from these records provide structured representations that support downstream retrieval, aggregation, and reasoning across heterogeneous healthcare data sources [15]. Recent advances in large language models (LLMs) and clinical natural language processing (NLP) have substantially expanded the scale and richness of information that can be extracted from clinical text [9, 25, 28, 32, 37], especially for normalizing heterogeneous mentions and supporting dynamic KG expansion [31].

However, expanding persistent KGs using model-derived signals introduces a fundamental system challenge. Clinical text is often noisy, incomplete, abbreviated, and internally inconsistent, frequently containing administrative or contextual information that does not directly reflect grounded clinical evidence [16]. In many existing pipelines, extracted entities and inferred relations are inserted into downstream knowledge structures through largely open-loop workflows, where outputs are accepted once generated or extracted [21, 26]. Under this paradigm, weakly supported or structurally ambiguous signals may silently propagate into persistent graph representations, potentially affecting downstream retrieval, reasoning, and analytics over time. These concerns are particularly important for LLM-assisted KG construction, where fluent semantic generalization can blur the boundary between source-grounded extraction and model-inferred knowledge [21, 26].

Prior work addresses related challenges through provenance tracking, explainability, and uncertainty estimation. Provenance mechanisms link predictions to source documents or text spans to improve interpretability and post hoc verification [15, 18, 34], while Bayesian and selective prediction models estimate uncertainty and support abstention or deferral under low confidence [10, 24]. However, these mechanisms are typically applied at the prediction layer rather than at the level of persistent knowledge mutation. Traceability can explain where an extraction originated, but it does not necessarily determine whether the extraction should modify a persistent KG [18]. Similarly, uncertainty estimates are commonly reported as auxiliary confidence measures without directly governing whether candidate entities or relations are admitted into evolving knowledge structures [1, 24]. As a result, existing clinical information-extraction and KG-construction pipelines largely emphasize extracting, normalizing, or predicting candidate knowledge, while comparatively less attention is given to admission control for persistent KG mutation [1, 15, 21, 40].

In this paper, we introduce the Allergy Risk Intelligence System (ARIS), an uncertainty-controlled framework for auditable clinical KG expansion. ARIS is instantiated for allergy-related clinical information, but its primary objective is not automated diagnosis or clinical decision support. Instead, ARIS studies how LLM-assisted extraction signals can be selectively admitted into persistent KGs under explicit structural, provenance, and uncertainty constraints. The framework formulates KG expansion as a controlled admission problem, in which candidate updates are incorporated into the KG only when they satisfy deterministic audit checks, Bayesian uncertainty thresholds, and provenance-completeness requirements.

ARIS is designed as a general KG update-control framework rather than a domain-specific prediction model. Although our implementation employs Bio_ClinicalBERT [4] for entity extraction and Gemma-3-4B-IT [39] for semantic encoding, the central contribution lies in the integration framework itself. ARIS combines learned deterministic audit patterns that identify structurally supported signals, a Bayesian neural network that estimates predictive uncertainty over candidate updates, and an uncertainty-weighted buffer that regulates how accepted signals contribute to KG evolution. This design shifts provenance and uncertainty from post hoc annotations to operational control signals for knowledge management.

We evaluate ARIS using allergy-related emergency and critical-care records derived from MIMIC-IV v2.2 [22], a sparse and heterogeneous EHR setting with incomplete documentation, abbreviated narratives, and variable clinical terminology. The evaluation focuses not on standalone diagnostic performance, but on whether the framework preserves traceability, exposes uncertainty, and selectively routes candidate updates under noisy clinical text. Specifically, we examine whether candidate updates can be structurally audited, made provenance-complete, routed by uncertainty, and conservatively integrated into persistent graph representations without uncontrolled mutation events.

Our main contributions are summarized as follows:

- We formulate LLM-assisted KG expansion as a controlled admission problem for persistent knowledge systems, where candidate updates are admitted only when structural validity, provenance completeness, and uncertainty thresholds are satisfied.
- We introduce an auditable KG expansion architecture combining encoder-based semantic representations, learned deterministic audit patterns, Bayesian uncertainty routing, and an uncertainty-weighted retraining buffer.
- We show how decomposed epistemic and aleatoric uncertainty can be used as operational signals for selective KG update, moving uncertainty from a post hoc confidence measure to a mechanism for regulating knowledge-base modification.
- Instantiated on allergy-related MIMIC-IV records, the framework demonstrates traceable, selective, and uncertainty-aware KG expansion under sparse and noisy clinical text.

Code and reproducibility documentation are available in an anonymized GitHub repository at https://github.com/Cro560/ARIS.

2 Related Work

2.1 NLP-driven KG Construction

Substantial research has studied how to transform unstructured clinical text into structured representations for downstream retrieval, analytics, and reasoning. Early efforts relied primarily on rule-based and ontology-driven clinical NLP systems, including dictionary matching, negation and temporality resolution, and concept normalization using established biomedical terminology and clinical NLP pipelines [14, 35]. These systems produce reliable outputs through strict selection rules but have limited capacity to represent uncertainty or temporal dynamics. Subsequent machine learning models improve scalability and recall by incorporating engineered textual and structured EHR features [42, 43]; however, their reliance on extensive feature engineering and site-specific preprocessing constrains generalizability across institutions and disease domains.

With the advent of deep learning, sequential and representation-learning models are introduced to capture longitudinal patient trajectories and complex interactions among clinical variables. Recurrent neural network models and unsupervised patient representations enable the prediction of future diagnoses and clinical events from EHR data [8, 29]. However, the traceability between predicted outcomes and underlying clinical evidence often remains limited and opaque [5, 36].

More recently, transformer-based clinical language models and LLMs have been applied to EHR text for entity extraction, relation inference, clinical question answering, and KG augmentation [4, 13, 17, 33, 42]. These models reduce manual feature engineering and enable cross-document synthesis, allowing richer semantic relations to be inferred from heterogeneous clinical narratives. However, their generative capacity introduces risks of hallucinated entities or relations, conflation of temporally distinct events, and ambiguity between extracted evidence and model-inferred knowledge [2].

Across these paradigms, NLP-driven KG construction and expansion pipelines have primarily been framed as extraction, normalization, or augmentation tasks, while the separate question of whether a candidate signal should be admitted into a persistent KG has received less explicit treatment [1, 15, 21, 40]. As a result, uncertainty, provenance, and long-term knowledge integrity are often treated as auxiliary concerns rather than operational constraints on KG mutation.

2.2 KG Expansion Traceability and Uncertainty

Given the safety-critical nature of healthcare applications, prior work has emphasized traceability, evidence linkage, and explainability in clinical NLP and EHR-based prediction systems [6, 36, 43]. Model traceability is typically defined as linking outputs to source documents, spans, or structured fields, enabling verification of the evidentiary basis underlying predictions [36, 43]. Evidence-linked extraction and attribution-based evaluation are proposed to reduce hallucinations and improve trustworthiness, particularly in clinical summarization and decision-support tasks [6]. However, most existing work treats explanations or attributions as post hoc evidence for model behavior rather than as operational constraints on knowledge-base mutation [18, 27, 34]. Traceability is predominantly treated as a post hoc explanatory mechanism to justify model outputs, rather than as a control signal that regulates the integrity of durable knowledge structures such as KGs [19, 27]. As a result, traceability mechanisms typically terminate at the prediction stage and do not directly influence knowledge persistence or mutation.

Similarly, uncertainty estimation has been extensively studied through Bayesian neural networks, approximate inference, selective prediction, and deep-learning uncertainty surveys [1, 7, 10, 11, 24]. Bayesian neural networks (BNNs), ensemble methods, and selective prediction frameworks enable models to quantify predictive uncertainty and to defer decisions under low confidence [36, 42]. However, uncertainty estimates are likewise rarely propagated beyond the prediction layer. Existing systems generally do not bind uncertainty signals to structural operations such as KG expansion or revision [1, 11]. Consequently, low-confidence or ambiguous

information extracted from clinical text may still modify downstream knowledge representations, even when models internally recognize high uncertainty. This disconnect between uncertainty awareness and knowledge mutation represents a critical gap in current NLP-driven KG pipelines.

3 ARIS: Controlled Admission for Auditable KG Expansion

3.1 Problem Formulation

Traditional clinical NLP and KG construction pipelines often emphasize extraction coverage, relation inference, or generation quality, treating errors primarily as prediction-level outcomes [15, 21, 26, 31, 40]. In persistent KGs, however, model-derived entities and relations become reusable knowledge objects for downstream retrieval, aggregation, and reasoning, so the relevant question is not only whether a model can produce a plausible signal, but whether that signal should be allowed to modify the knowledge structure [15, 18, 27]. We therefore formulate LLM-assisted KG expansion as a *controlled acceptance problem*: candidate signals are admitted into the KG only if they satisfy explicit structural, uncertainty, and provenance constraints.

Let x denote an input document and let $s(x)$ denote a candidate entity or relation signal derived from semantic extraction. The update risk of interest can be written as

$$P(\text{incorrect KG update} \mid s(x) \text{ is admitted}) \,. \tag{1}$$

ARIS does not directly estimate this probability as a calibrated formal bound. Instead, it reduces the admissible update space by requiring $s(x)$ to satisfy structural audit checks, uncertainty-based routing constraints, and provenance-completeness requirements before it can contribute to KG evolution.

The key design principle is to separate *semantic representation* from *knowledge admission*. LLMs and clinical encoders provide representations and candidate signals, but they do not directly mutate the KG. Admission is governed by three complementary controls: (1) deterministic audit checks for structural support, (2) Bayesian uncertainty routing for predictive uncertainty, and (3) controlled output formation that prevents free-form text from being treated as accepted knowledge. This design does not claim absolute correctness of every accepted update; rather, it makes the update process explicit, inspectable, and conservative. The overall pipeline architecture is illustrated in Figure 1.

3.2 Segment-Level Semantic Encoding

ARIS first converts each document into stable semantic representations and structured entity spans. We apply a frozen Bio_Clinical-BERT encoder [4] f_{BERT} to maintain representational consistency across KG evolution cycles. Given a document x, the encoder produces a 768-dimensional contextualized embedding $h = f_{\text{BERT}}(x) \in \mathbb{R}^{768}$, and segment extraction yields structured entity annotations

$$\{(\ell_i, s_i, e_i)\}_{i=1}^{K},$$

where K is the number of extracted entities, ℓ_i denotes the entity label, and $[s_i, e_i]$ denotes the character-level segment boundaries.

Transformer inference exhibits $O(NL^2d)$ complexity for sequence length L, embedding dimension d, and N layers. For typical clinical notes ($L = 512$, $N = 12$), this remains tractable on modern hardware. Although instantiated with Bio_ClinicalBERT [4], the framework is encoder-agnostic and can use alternative transformer architectures. For high-capacity semantic representation, we use Gemma-3-4B-IT [39] in encoder-only mode: hidden representations are extracted from fixed forward passes, while generative decoding, sampling, and free-form text emission are disabled in the controlled ARIS pipeline. This configuration separates semantic representation from persistent KG admission; the same base model is used as a generator only in the unconstrained LLM ablation described in Section 4.7.

3.3 Learned Deterministic Audit Gate

The deterministic audit gate serves as a reproducible admissibility filter for candidate KG updates. It is not intended to be a comprehensive clinical validator or a replacement for learned semantic models. Instead, it checks whether a candidate signal has sufficient structural support to proceed to uncertainty-based routing. This role is important for KG expansion because a low-quality or structurally incomplete extraction should not become a candidate for persistent graph update merely because it was produced by a fluent model.

We use a hybrid design. Audit patterns are *learned* from Bio_ClinicalBERT named entity recognition (NER) outputs on MIMIC-IV. They are then applied *deterministically* using Aho–Corasick pattern matching [3, 35]. The learned component provides domain adaptation, while deterministic execution ensures reproducible decisions and explicit failure logs. This design prioritizes conservative admission over aggressive coverage, making the update process inspectable without requiring manually handcrafted rules.

Rule learning from Bio_ClinicalBERT. Rather than handcrafting rules, we derive pattern-based audit rules directly from Bio_ClinicalBERT outputs on the MIMIC-IV training corpus (n = 4921 annotated allergy-related notes). The extracted audit lexicon contains 474 indexed patterns across 9 entity types, including 237 Drug patterns, 89 Dosage patterns, 52 Route patterns, 43 Frequency patterns, 31 Strength patterns, 14 Duration patterns, 6 Reaction patterns, 1 Severity pattern, and 1 Other pattern. The learning procedure proceeds in three stages.

(1) Pattern extraction. For each entity type ℓ (Drug, Dosage, Route, Frequency, etc.), we compute token frequency $f_\ell(t)$ and retain high-frequency patterns with minimum threshold $\tau_{\min} = 5$:

$$\mathcal{L}_\ell = \{t : f_\ell(t) \geq \tau_{\min}\}. \tag{2}$$

This threshold is not tuned for predictive optimality; it excludes idiosyncratic artifacts so that compiled rules reflect recurring operational patterns rather than isolated corpus noise. This yields domain-adapted lexicons while excluding rare one-off artifacts.

(2) Rule compilation. Extracted lexicons are compiled into three rule categories: *required-entity rules*, which require mandatory entity types such as Drug, Dosage, and Route; *forbidden-pattern rules*, which reject or flag exclusion indicators such as "no known allergy" or "NKDA"; and *structural-completeness rules*, which require the extracted segment count to satisfy a minimum threshold.

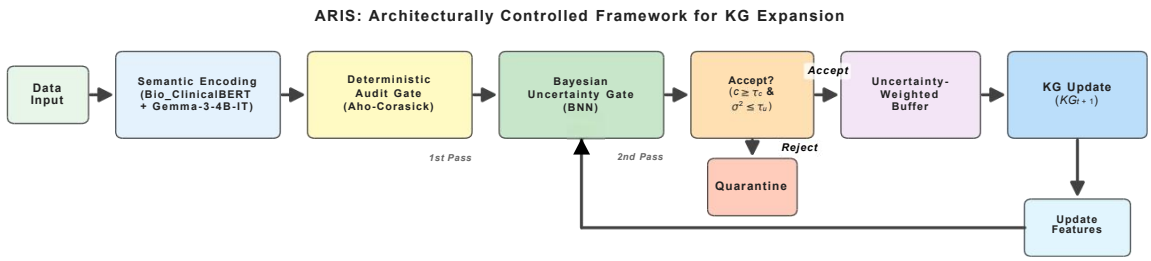

Figure 1: ARIS pipeline for controlled admission in auditable clinical KG expansion. Candidate signals are encoded, audited, uncertainty-scored, provenance-linked, and accumulated in an uncertainty-weighted buffer before contributing to KG evolution.

(3) Deterministic application via Aho–Corasick. Pattern lexicons are indexed using the `pyahocorasick` library [30], an optimized Python implementation of the Aho–Corasick algorithm [3], enabling exact string matching in $O(|x| + m)$, where m is the total pattern count. The audit decision is deterministic:

$$\text{audit_pass}(x) = r_{\text{req}}(x) \wedge r_{\text{forbid}}(x) \wedge r_{\text{complete}}(x), \tag{3}$$

requiring simultaneous satisfaction of required-entity, forbidden-pattern, and structural-completeness checks. Determinism here means that identical inputs yield identical audit decisions and failure reports, in contrast to stochastic or sampling-based generation. Although rule execution is deterministic, rule specification remains data-driven, enabling automatic domain adaptation while preserving interpretability.

The audit gate outputs structured logs containing triggered rules, hit counts by entity type, and rejection or warning reasons. These logs are retained as provenance metadata for downstream routing, allowing later inspection of why a candidate signal was admitted, rejected, or forwarded for uncertainty-based evaluation. In implementation, audit outcomes are recorded as structured *pass*, *warning*, or *rejection* states. Hard-failure cases are not admitted as reliable KG updates, while warning-level structural deficiencies may still be routed to the Bayesian uncertainty gate with explicit audit flags so that their uncertainty and provenance remain inspectable. This distinction allows ARIS to avoid silently discarding ambiguous records while still preventing unsupported signals from being treated as accepted KG knowledge.

3.4 Bayesian Uncertainty Gate

The deterministic audit gate checks structural admissibility, but it does not quantify whether a candidate signal is uncertain because of limited evidence, sparse context, or model ambiguity. ARIS thus introduces a Bayesian uncertainty gate, implemented as a Bayesian neural network (BNN) [7], to estimate uncertainty for audit-passed records before they can contribute to KG evolution. The role of this gate is not to certify correctness, but to expose uncertainty as an operational routing signal for selective KG admission.

Each audit-passed record is represented by concatenating the Bio_ClinicalBERT embedding $h = f_{\text{BERT}}(x)$ with KG-derived structural features $g_{\text{KG}}(x)$, forming

$$z = [h;\, g_{\text{KG}}(x)] \in \mathbb{R}^{d+d_{\text{KG}}}, \tag{4}$$

where $g_{\text{KG}}(x)$ encodes entity co-occurrence statistics, relation frequencies, and graph connectivity measures ($d_{\text{KG}} = 128$ dimensions). This joint representation conditions uncertainty estimates on both semantic content and the current state of the knowledge graph.

The supervised target y is an audit-derived structural admissibility score computed from the training records. It summarizes the structural support available for a candidate KG update and is used only to train the routing model, not as a clinical correctness label.

The BNN treats y as a noisy scalar routing target and models parameter uncertainty through a variational posterior $q_\phi(\omega)$ over weights $\omega = (\omega_{[k]})_{k=1}^{|\omega|}$. We use a Gaussian observation model:

$$y = f_{\text{BNN}}(z;\omega) + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \sigma^2_{\text{noise}}), \quad \omega \sim q_\phi(\omega). \tag{5}$$

We model $q_\phi(\omega)$ as a fully factorized Gaussian distribution:

$$q_\phi(\omega) = \prod_{k=1}^{|\omega|} \mathcal{N}(\omega_{[k]} \mid \mu_{[k]}, \sigma^2_{[k]}), \tag{6}$$

where $\phi = \{\mu_{[k]}, \sigma^2_{[k]}\}_{k=1}^{|\omega|}$ are learnable variational parameters. The predictive distribution is

$$p(y \mid z) = \int p(y \mid z, \omega) q_\phi(\omega)\, d\omega, \tag{7}$$

where $p(y \mid z, \omega) = \mathcal{N}\big(y \mid f_{\text{BNN}}(z;\omega), \sigma^2_{\text{noise}}\big)$ with the Gaussian noise variance σ^2_{noise} learned during BNN training.

The BNN is trained by minimizing the negative Evidence Lower Bound (ELBO) [7]:

$$\mathcal{L}_{\text{ELBO}} = \mathbb{E}_{q_\phi(\omega)}[-\log p(y \mid z, \omega)] + \text{KL}(q_\phi(\omega) \| p(\omega)), \tag{8}$$

where the first term is the expected negative log-likelihood and the second term regularizes the variational posterior toward the prior $p(\omega) = \mathcal{N}(0, I)$ [7]. The training configuration used in all experiments is summarized in Section 4.2.

At inference, Monte Carlo sampling draws $S = 10$ weight samples $\{\omega^{(s)}\}_{s=1}^{S}$. For each sample, we compute the noise-free BNN output

$$m^{(s)} = f_{\text{BNN}}(z;\omega^{(s)}),$$

and the predictive mean

$$\mu = \frac{1}{S}\sum_{s=1}^{S} m^{(s)}. \tag{9}$$

Following Bayesian approximate inference practice [10] and uncertainty decomposition in neural prediction [23], we separate epistemic uncertainty from aleatoric observation noise. Epistemic

uncertainty is estimated from Monte Carlo variation across sampled weights:

$$\sigma^2_{\text{epistemic}} = \frac{1}{S}\sum_{s=1}^{S}(m^{(s)} - \mu)^2. \tag{10}$$

Aleatoric uncertainty is represented by the learned observation-noise term:

$$\sigma^2_{\text{aleatoric}} = \sigma^2_{\text{noise}}. \tag{11}$$

The total uncertainty used for routing is therefore

$$\sigma^2_{\text{total}} = \sigma^2_{\text{epistemic}} + \sigma^2_{\text{aleatoric}}. \tag{12}$$

Here, epistemic uncertainty captures reducible model uncertainty induced by limited evidence or model ambiguity, while aleatoric uncertainty captures irreducible observation noise. The noise term is learned during training and held fixed at inference.

ARIS then computes an uncertainty-normalized confidence score,

$$c = \frac{|\mu|}{(\sigma^2_{\text{total}} + \varepsilon)^{1/2}}, \tag{13}$$

with $\varepsilon = 10^{-6}$. This score measures the signal-to-noise ratio of a candidate update and is used only as a routing statistic, not as a calibrated probability [7, 24].

The BNN gate admits a record only if it satisfies both a minimum confidence threshold and a maximum uncertainty threshold:

$$c \geq \tau_c \quad \wedge \quad \sigma^2_{\text{total}} \leq \tau_u, \tag{14}$$

where τ_c and τ_u are hyperparameters selected via grid search on the validation set before test-set evaluation. Records that fail these conditions are not treated as reliable KG updates; they are routed for rejection, quarantine, or lower-weighted use depending on downstream configuration. This design uses uncertainty as a control variable for knowledge admission rather than as a post hoc descriptive statistic.

BNN architecture specification. The Bayesian uncertainty gate consists of a 3-layer MLP with hidden dimensions 768, 512, and 256, respectively, followed by a scalar output layer. Each hidden layer uses a ReLU activation followed by a 0.1 dropout rate. The resulting BNN is lightweight relative to the frozen transformer encoder and is used only for routing candidate updates, not for generating clinical conclusions.

3.5 Controlled KG Signal Collection

ARIS restricts model-derived outputs to structured, provenance-complete candidate signals. This design is motivated by the distinction between transient generated text and persistent KG updates: a generated explanation can be ignored or regenerated, whereas a KG update may be reused by downstream retrieval, reasoning, or analytics modules. Therefore, ARIS does not allow free-form text to directly instantiate accepted knowledge.

Each admitted candidate signal is emitted as

$$\text{record}_i = \{E_i, T_i, P_i\}, \tag{15}$$

where E_i denotes accepted entities, including entity labels, character offsets, and normalized text; T_i denotes candidate subject-predicate-object triples derived from E_i; and P_i contains provenance metadata:

$$P_i = \{\text{audit_log}_i,\ \mu_i,\ \sigma^2_{\text{total},i},\ c_i,\ \text{hash}(x_i),\ \text{timestamp}\}. \tag{16}$$

This record structure makes each candidate KG update traceable to its source document and to the decision process that admitted it. The provenance field stores deterministic audit outcomes, Bayesian uncertainty statistics, confidence scores, source hashes, and timestamps. As a result, accepted triples can be inspected after construction, and uncertain or structurally weak updates can be traced back to the evidence and routing decisions that produced them.

The purpose of this design is not to eliminate all possible extraction errors, but to prevent unsupported model-derived signals from silently becoming persistent KG facts. In ARIS, LLMs and encoders provide semantic representations and candidate signals; KG admission is controlled by audit checks, uncertainty estimates, and provenance requirements.

3.6 Uncertainty-Weighted Retrain Buffer

Directly committing every accepted signal to a KG can make later correction difficult, especially when candidate updates are derived from sparse or incomplete text. ARIS therefore accumulates admitted signals in an uncertainty-weighted retraining buffer before they contribute to KG evolution. The buffer provides three functions: batch-level inspection, uncertainty-based reweighting, and reversible update management.

Each accepted record i receives a weight

$$w_i = \mu_i \cdot \exp(-\lambda \sigma^2_{\text{total},i}), \tag{17}$$

where w_i is the reweighting coefficient and λ controls the strength of the uncertainty penalty. The weighting scheme prioritizes candidate signals with higher predictive magnitude and lower uncertainty, while down-weighting uncertain records rather than treating all accepted records as equally reliable.

The buffer induces a weighted empirical distribution over candidate triples:

$$\hat{P}_{\text{buffer}}(T) = \frac{1}{Z}\sum_i w_i\, \delta(T = T_i), \tag{18}$$

where T_i denotes the triple set associated with record i, $\delta(\cdot)$ is the indicator function, and $Z = \sum_i w_i$ is a normalization constant. The KG model is trained to match $\hat{P}_{\text{buffer}}$ by minimizing the expected negative log-likelihood:

$$\begin{aligned}\mathcal{L}_{\text{KG}} &= \mathbb{E}_{\hat{P}_{\text{buffer}}(T)}[-\log p_{\text{KG}}(T \mid \text{KG}_t)] \\ &= -\frac{1}{Z}\sum_i w_i \log p_{\text{KG}}(T_i \mid \text{KG}_t),\end{aligned} \tag{19}$$

where $p_{\text{KG}}(T_i \mid \text{KG}_t)$ is the likelihood assigned by the current KG model to T_i.

This formulation turns uncertainty into a mechanism for controlling the influence of candidate updates on KG evolution. High-uncertainty signals are not necessarily discarded, but their contribution is reduced unless supported by stronger evidence in later update cycles. This conservative buffering strategy is particularly useful for clinical text, where missing documentation and abbreviated records can produce uncertain but potentially informative signals.

3.7 Closed-Loop Control

ARIS models KG expansion as an iterative knowledge-admission process. After accepted signals are accumulated in the uncertainty-weighted buffer, the KG is updated through an offline retraining procedure:

$$\mathrm{KG}_{t+1} \leftarrow \mathrm{Update}(\mathrm{KG}_t, \mathrm{Buffer}_t). \tag{20}$$

The updated KG is then used to recompute KG-derived structural features for subsequent records:

$$z_{t+1} = [h;\, g_{\mathrm{KG}_{t+1}}(x)]. \tag{21}$$

This feedback allows newly incorporated knowledge to influence future uncertainty estimates without altering the frozen semantic encoder.

To reduce self-reinforcement, ARIS restricts KG updates to signals that pass audit and uncertainty controls, and it encodes KG-derived features as aggregate structural statistics rather than entity-specific identifiers. This design limits the extent to which individual accepted records can directly reinforce themselves in later iterations. It also keeps the semantic encoder fixed, so changes in uncertainty estimates arise from graph evolution rather than from uncontrolled representation drift.

Under standard Bayesian intuition, incorporating additional validated evidence can reduce epistemic uncertainty as the posterior concentrates with increasing data [12]. ARIS uses this principle as a qualitative convergence signal rather than as a formal monotonic guarantee: if epistemic uncertainty decreases or plateaus over update cycles, the graph may be stabilizing relative to the available evidence; if uncertainty remains high, additional review, evidence, or domain-specific audit refinement may be needed.

We emphasize that ARIS is not intended as a fully automatic KG construction or clinical decision system. It is a controlled knowledge-admission framework for settings where persistent updates should remain traceable, auditable, and uncertainty-aware. Within this scope, ARIS deliberately trades aggressive coverage for selective admission: candidate signals are allowed to influence KG evolution only when they satisfy explicit structural, uncertainty, and provenance requirements.

4 Experiments

We evaluate ARIS as a controlled KG expansion framework on clinical EHR data. The goal of the evaluation is not to optimize a standalone clinical prediction model, but to examine whether LLM-assisted candidate signals can be admitted into a persistent KG under explicit structural, provenance, and uncertainty constraints. Accordingly, our experiments focus on five questions: (i) whether candidate updates remain fully routed and provenance-complete, (ii) whether learned deterministic audit patterns discriminate structurally reliable from unreliable inputs, (iii) whether Bayesian uncertainty provides usable routing signals for selective KG updates, (iv) whether uncertainty-weighted KG evolution preserves graph structure with low computational overhead, and (v) whether the full controlled pipeline provides benefits over unconstrained LLM generation and classical NER-to-KG construction.

This evaluation is designed as framework validation rather than deployment validation: the metrics measure admissibility and traceability of candidate KG updates, not autonomous clinical diagnosis. The final ablation isolates the central design choice of separating semantic representation from knowledge admission.

4.1 Data Preprocessing and KG Construction

All experiments are conducted on MIMIC-IV v2.2 [22], a public de-identified EHR database from Beth Israel Deaconess Medical Center. Records are filtered using allergy-related ICD-10 criteria [38, 41], including allergic reactions, anaphylaxis, allergy status, adverse effects, and drug-, food-, or environment-related hypersensitivity. The resulting cohort contains 1256 patients, 7030 admissions, and 7031 discharge summaries. The exact filtering scripts, configuration files, and data-access instructions are provided in the anonymized GitHub repository; raw MIMIC-IV records and derived patient-level files are not redistributed.

We use allergy-related emergency and critical care records as a challenging testbed for controlled KG expansion. These records are sparse and heterogeneous: they prioritize acute care over comprehensive allergy histories, contain frequent missing fields (mean 2.3 missing entity types per record), include abbreviated narratives (mean 89.3 tokens per emergency note versus 312.7 per elective admission note in the same MIMIC-IV corpus), and exhibit inconsistent terminology in 73.2% of records. This setting is appropriate for evaluating selective KG update because indiscriminate admission of weakly supported candidate signals may propagate errors into the persistent KG.

Named entity recognition is performed using a Bio_Clinical BERT-based token classification model fine-tuned on 4921 training notes, 1055 validation notes, and 1055 test notes, achieving an F1 score of 0.9778 on the test notes. The pipeline extracts 1,432,155 entity mentions across 9 entity types: Drug, Dosage, Route, Frequency, Strength, Duration, Reaction, Severity, and Other.

Following entity extraction and normalization, we construct the final clinical KG by integrating normalized entities, note–admission–patient links, entity co-occurrence relations, section-aware context, and structured annotations. The resulting graph provides 128-dimensional structural features for uncertainty-aware routing, including entity co-occurrence statistics, relation frequencies, and graph-connectivity measures. The final KG used in all experiments contains 63,532 nodes and 4,294,094 triples.

Data are split at the patient level into training, validation, and test sets (8:1:1) to prevent leakage. The training set is used for NER fine-tuning and BNN training, the validation set for threshold selection, and the test set for evaluation. The test partition ($n = 2000$) provides the evaluation set for routing, provenance, uncertainty, controlled-update analysis, and ablation against alternative KG construction strategies.

4.2 Model Configuration and Evaluation Protocol

The Bayesian uncertainty gate is trained on audit-passed records by minimizing the ELBO in Eq. (8) using Adam with learning rate 5×10^{-4}, batch size 32, and early stopping over 50 epochs with patience 5. Monte Carlo sampling uses $S = 10$ weight samples during training and inference [10]. Confidence and uncertainty thresholds are set to $\tau_c = 0.05$ and $\tau_u = 15.0$, selected via grid search on the validation set before test-set evaluation to keep conservative admission under

Table 1: Routing and provenance coverage over 2000 test records.

Field	Coverage
Routing decision coverage	2000/2000 (100%)
Audit metadata coverage	2000/2000 (100%)
Uncertainty estimate coverage	2000/2000 (100%)
Provenance hash coverage	2000/2000 (100%)
BNN bypass events	0/2000 (0%)
Uncontrolled generation events	0/2000 (0%)

the framework-level acceptance objective. Following [7], we set the prior distribution to $p(\omega) = \mathcal{N}(0, I)$. All experiments use FP16 precision on an NVIDIA V100 GPU.

For semantic encoding, ARIS uses Gemma-3-4B-IT [39] in encoder-only mode, with free-form generation disabled at the system level. The resulting fixed representations are passed to the deterministic audit gate and Bayesian uncertainty gate before any candidate signal can enter the uncertainty-weighted buffer.

For the system-level ablation, the unconstrained LLM baseline uses the same Gemma-3-4B-IT model with deterministic greedy decoding (temperature = 0, max tokens = 100) to generate free-form outputs without structural admission constraints, uncertainty estimates, or provenance-complete routing logs. For the classical KG construction ablation, the NER-to-KG baseline uses the same Bio_ClinicalBERT extraction backbone but directly maps extracted entities into KG records without post-extraction audit or Bayesian uncertainty routing. These baselines are used only to isolate the contribution of controlled knowledge admission; they are not intended as clinical deployment systems.

The Bayesian route selector is lightweight relative to the frozen Gemma-3-4B-IT encoder: it adds 1.12M parameters, corresponding to 0.026% of the approximately 4.30B-parameter backbone, and increases computation by 0.00225 GFLOPs/token, or 0.042% relative overhead. This cost is substantially smaller than full-model adaptation, repeated ensemble inference, or adapter-based alternatives, because ARIS uses the BNN only as a routing layer over fixed representations rather than as a replacement for the base encoder.

4.3 Experiment 1: Routing and Provenance Coverage

Persistent KG updates require more than producing candidate entities or relations. Each accepted signal should be recoverable, auditable, and associated with the decision process that admitted it. This experiment evaluates whether ARIS preserves complete routing and provenance records across the test set, and whether candidate updates remain inspectable after processing.

We analyze controlled-pipeline execution manifests generated by the routing and merge scripts for all 2000 test records. We measure coverage of routing decisions, audit metadata, uncertainty estimates, and cryptographic provenance hashes as update-control metadata rather than clinical correctness labels.

ARIS provides complete coverage for all provenance-related fields. Every record contains audit metadata, routing decisions, uncertainty estimates, and provenance hashes. No record bypasses the controlled routing path. Table 1 summarizes the coverage results.

A representative record (WD-0001) illustrates this behavior. The audit layer flags structural warnings (`min_total_hits: need=3, got=0`) but does not silently discard the record. Instead, the candidate signal is routed to the Bayesian uncertainty gate, which records high total uncertainty (18.77), epistemic uncertainty (18.42), aleatoric uncertainty (0.35), confidence score (0.051), and a cryptographic provenance hash. This example shows that ARIS does not merely block or accept records; it preserves the evidence needed to inspect why a candidate update was routed in a particular way.

Complete provenance coverage is important because KG updates are persistent structural changes. Unlike free-form generated text, a KG update may be reused by downstream retrieval or reasoning modules. ARIS therefore makes admission decisions auditable rather than treating generated signals as self-validating knowledge. This result supports the data-provenance aspect of controlled KG expansion: accepted signals remain linked to their source evidence, routing path, and uncertainty metadata.

4.4 Experiment 2: Synthetic Audit-Stress Validation for the Deterministic Audit Gate

Before a candidate signal is evaluated by the Bayesian uncertainty gate, ARIS first checks whether the input contains sufficient structural evidence for KG update. This experiment evaluates whether the learned deterministic audit gate can distinguish structurally reliable from unreliable records while preserving explicit rejection reasons. We frame this as synthetic audit-stress validation rather than adversarial robustness evaluation, because the goal is to test structural admissibility under controlled low-quality and high-quality input conditions.

The audit gate is evaluated over 100 independent runs, each containing 2000 synthetic clinical records: 1000 low-quality records and 1000 high-quality controls. This yields 200,000 synthetic audit-stress records in total, including 100,000 low-quality stress cases and 100,000 high-quality controls. Synthetic stress records are generated by varying structural completeness, entity-hit counts, missing required fields, and low-quality allergy-like templates, while high-quality controls preserve the required allergen, medication, route, dosage, reaction, or contextual evidence patterns. The generation and aggregation scripts are provided in the anonymized GitHub repository. The gate uses learned audit patterns extracted from Bio_ClinicalBERT outputs and applies them deterministically with a MIN_HITS threshold of 2. No manual tuning is performed on the synthetic evaluation runs.

Table 2 reports the audit gate discrimination results. Across 100 runs, the gate achieves an accuracy of 92.64% ± 0.58%, precision of 99.16% ± 0.33%, recall of 86.01% ± 1.06%, specificity of 99.27% ± 0.28%, and F1-score of 0.921 ± 0.007, where all values are reported as mean ± standard deviation (SD) across runs.

The audit gate is designed to support conservative KG update rather than maximize unconstrained extraction coverage. Its high precision and specificity indicate that accepted records are usually structurally reliable and that low-quality records are rarely admitted. The lower recall reflects the intended conservative behavior of the gate: ARIS prioritizes reliable admission over exhaustive coverage. This behavior is important for controlled KG expansion because the audit gate narrows the admissible update space before

Table 2: Synthetic audit-stress validation of the deterministic audit gate over 100 runs. Values with ± are reported as mean ± SD across runs.

Metric	Result
Accuracy	92.64% ± 0.58%
Precision	99.16% ± 0.33%
Recall	86.01% ± 1.06%
Specificity	99.27% ± 0.28%
F1-score	0.921 ± 0.007
MIN_HITS threshold	2
Runs	100
Records per run	2,000
Total synthetic records	200,000
Low-quality stress cases	100,000
High-quality controls	100,000

Bayesian uncertainty is applied. Unlike post hoc explanations, the audit decision directly participates in whether a candidate signal can proceed toward KG admission, be rejected, or be forwarded with explicit warning flags for uncertainty-based evaluation.

4.5 Experiment 3: Bayesian Uncertainty as a Selective Update Signal

Uncertainty estimates are often reported as prediction-level confidence scores, but they are not always connected to whether an extracted signal should modify a persistent KG. This experiment evaluates whether the Bayesian uncertainty gate provides usable record-level signals for selective KG update and whether these estimates remain stable under limited public-data regimes.

We evaluate the Bayesian uncertainty gate on the 2000-record test partition. For each routed record, the BNN produces a predictive mean, total uncertainty, epistemic uncertainty, aleatoric uncertainty, and confidence score using Monte Carlo sampling with $S = 10$ weight samples. Candidate updates are evaluated against the confidence and uncertainty thresholds $\tau_c = 0.05$ and $\tau_u = 15.0$, selected on validation data. To evaluate low-resource stability, we additionally subsample 50, 100, 200, 500, 1000, and 2000 records and compare total-uncertainty estimates.

Across the 2000 test records, the total uncertainty is 13.42 ± 3.29 (mean ± SD). Epistemic uncertainty (13.02 ± 3.27) dominates the decomposition, accounting for about 97.0% of total uncertainty, while aleatoric uncertainty (0.40 ± 0.30) accounts for about 3.0%. Epistemic and aleatoric uncertainties are effectively uncorrelated ($r = 0.003$, $p = 0.897$), suggesting that reducible model uncertainty and irreducible data noise are separated in the learned routing signal.

Total uncertainty varies substantially across records, with a range of 3.30–26.24 and 5th/50th/95th percentiles of 8.21/13.21/19.14. Low-resource subsampling shows that the total uncertainty stabilizes quickly: at 100 records, it is 13.88 ± 3.37, about 103.4% of the 2000-record baseline, and at 500 records it reaches 13.44 ± 3.28, approximately 100.1% of the baseline. Table 3 summarizes the uncertainty decomposition and low-resource stability results.

These results support the use of Bayesian uncertainty as an operational routing signal for KG update. The wide uncertainty range indicates that the BNN does not assign uniform uncertainty to all records, but instead distinguishes easier from harder candidate updates. The dominance of epistemic uncertainty is consistent with the evaluation setting: allergy-related emergency records are sparse, abbreviated, and often incomplete, so uncertainty is largely reducible through additional evidence rather than merely reflecting irreducible noise. The low-resource analysis suggests that the uncertainty signal remains stable under limited public-data regimes, but should not be interpreted as deployment-stage calibration against expert-labeled correctness.

Table 3: Bayesian uncertainty and low-resource stability. Values with ± are reported as mean ± SD across records.

Metric	Value
Uncertainty decomposition on 2000 records	
Total uncertainty	13.42 ± 3.29
Epistemic uncertainty	13.02 ± 3.27 (97.0%)
Aleatoric uncertainty	0.40 ± 0.30 (3.0%)
Epi.–alea. correlation	$r = 0.003$ ($p = 0.897$)
Confidence score	0.073 ± 0.019
Total unc. range	3.30–26.24
Total unc. 5th/50th/95th pct.	8.21 / 13.21 / 19.14
Low-resource total uncertainty	
50 records	14.16 ± 2.95 (105.5%)
100 records	13.88 ± 3.37 (103.4%)
200 records	13.82 ± 3.58 (103.0%)
500 records	13.44 ± 3.28 (100.1%)
1000 records	13.42 ± 3.32 (100.0%)
2000 records	13.42 ± 3.29 (100.0%)

4.6 Experiment 4: Graph Reweighting and Structure Preservation

A controlled KG expansion framework should not only route candidate updates, but also preserve the structure of the evolving graph. This experiment evaluates whether uncertainty-based reweighting changes global KG structure during offline graph evolution.

We evaluate uncertainty-based reweighting during offline KG retraining. Accepted candidate signals are weighted by uncertainty before contributing to KG evolution, following the weighting function in Eq. (17). We compare graph-level statistics before and after reweighting, including Gini coefficient, mutual information retention, entity retention, label overlap, and the total-uncertainty range of routed candidate signals.

Uncertainty-based reweighting preserves global graph structure. As shown in Table 4, the Gini coefficient changes from 0.9316 in the original graph to 0.9312 after reweighting, a difference of only $\Delta = 0.0004$. Mutual information retention remains 1.0, entity retention is 98.44%, and label overlap is preserved. These results indicate that total-uncertainty reweighting changes the influence of candidate signals without substantially distorting the global KG structure or disproportionately removing hub or long-tail entities.

The graph results suggest that uncertainty weighting changes the contribution of candidate signals without collapsing global graph structure. This is important for KG management because operators can adjust uncertainty penalties without rebuilding the entire graph

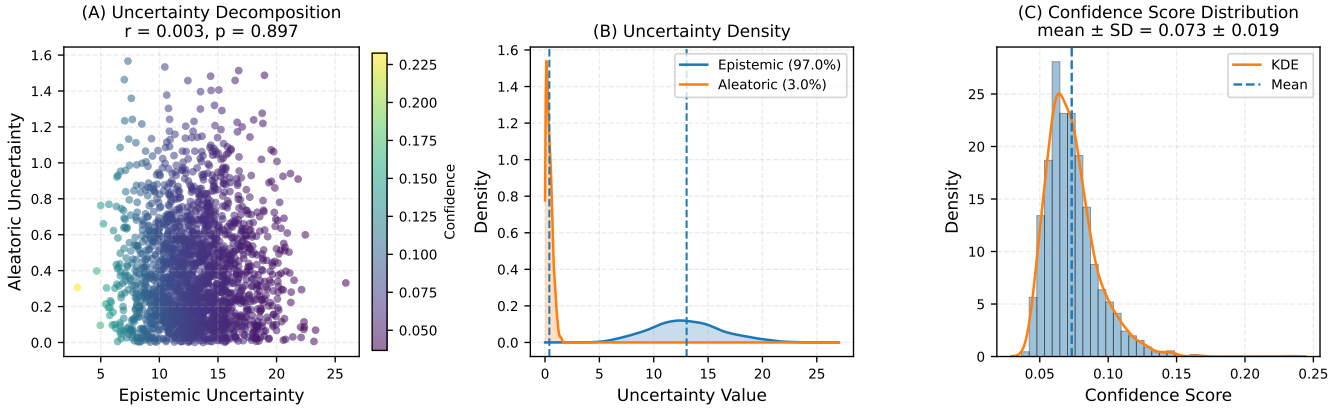

Figure 2: Bayesian uncertainty decomposition and confidence-score distribution over the 2000-record test set.

Table 4: Graph preservation under total-uncertainty reweighting.

Metric	Value
Original KG Gini	0.9316
Reweighted KG Gini	0.9312
Gini change	$\Delta = 0.0004$
MI retention	1.0
Entity retention	98.44%
Label overlap	Preserved
Total uncertainty range	3.30–26.24

or manually rebalancing hub and long-tail entities. Together with the lightweight computational footprint reported in Section 4.2, these findings support ARIS as a practical data-integration layer for controlled KG admission.

4.7 Experiment 5: Ablation Study

The preceding experiments evaluate ARIS routing, provenance, audit, uncertainty, and graph-evolution behavior. We now study whether these controls materially change the admissible update space. Because ARIS layers implement distinct admissibility conditions rather than additive predictors, removing one layer changes the update space and task semantics. We therefore conduct two system-level ablations: one against unconstrained LLM generation and one against classical NER-to-KG construction. The unconstrained LLM ablation uses the same Gemma-3-4B-IT baseline and 2000 test records described in Section 4.2.

To examine whether controlled admission depends on a single operating point, we also conduct an uncertainty-threshold sweep on the full 2000-record test set. We define three profiles by total-uncertainty percentiles: Strict accepts the lowest-uncertainty 10% of records, Balanced accepts the lowest-uncertainty 50%, and Permissive accepts the lowest-uncertainty 95%. This sweep changes only the total-uncertainty admission threshold, not the model architecture, routing mechanism, or underlying candidate records.

Table 5 reports the system-level comparisons and the full-test-set operating-profile sweep. The unconstrained LLM baseline produces free-form outputs for all records but provides no routing decisions, uncertainty estimates, audit metadata, or provenance hashes. In contrast, ARIS disables uncontrolled generation and provides complete coverage for routing and provenance fields across all 2000 records. Under the Strict profile, ARIS accepts 200/2000 records (10%) with threshold $\sigma^2_{\text{total}} \leq 9.19$ and mean total uncertainty 7.96. Under the Balanced profile, it accepts 1000/2000 records (50%) with threshold $\sigma^2_{\text{total}} \leq 13.22$ and mean total uncertainty 10.79. Under the Permissive profile, it accepts 1900/2000 records (95%) with threshold $\sigma^2_{\text{total}} \leq 19.14$ and mean total uncertainty 13.04. Across all profiles, ARIS maintains zero observed uncontrolled-generation bypass events.

We further compare ARIS with a classical NER-to-KG baseline using the same Bio_ClinicalBERT extraction backbone. The baseline directly maps extracted entities into KG records, while ARIS adds post-extraction audit and Bayesian uncertainty routing. We use "schema-complete structured records" to denote structural completeness of the allergen–reaction–temporal/contextual evidence needed for KG construction, not clinical diagnostic correctness.

As shown in Table 5, ARIS improves structurally coherent records from 95.2% to 100%, complete profiles from 72.7% to 76.0%, and schema-complete structured records from 28.8% to 76.0%.

The ablation results show why controlled admission is necessary. Compared with unconstrained LLM generation, ARIS provides routing, uncertainty, audit, and provenance fields that are absent from free-form outputs. Compared with a classical NER-to-KG pipeline, ARIS improves the structural completeness of candidate KG records by enforcing post-extraction admission requirements. Together, these results support the claim that the contribution is not merely stronger extraction or more fluent generation, but a controlled knowledge-admission layer that determines whether model-derived signals are eligible to influence persistent KG evolution.

4.8 Summary of Experimental Findings

The experimental results support ARIS as a data-integration framework for controlled KG expansion. Across five experiments, we evaluate whether candidate signals can be extracted, audited, uncertainty-scored, provenance-linked, selectively admitted, and incorporated

Table 5: Ablation study for controlled KG admission.

Ablation	Baseline	Full ARIS	Finding
Unconstrained LLM	Generator; no routing, audit, uncertainty, or provenance	Encoder-only; 100% routing, audit, uncertainty, and provenance	ARIS prevents free-form outputs from directly becoming KG updates.
Operating profile	N/A	10%/50%/95% accepted at thresholds 9.19/13.22/19.14; 0% bypass	ARIS tunes admission coverage while preserving controlled routing.
Classical NER-to-KG	95.2% coherent; 72.7% complete; 28.8% schema-complete	100% coherent; 76.0% complete; 76.0% schema-complete	ARIS improves structural completeness through post-extraction admission.

into a persistent KG without relying on unconstrained generation. Experiment 1 verifies complete routing and provenance coverage across all 2000 test records, making candidate updates inspectable through audit metadata, uncertainty estimates, and source hashes. Experiment 2 shows that the deterministic audit gate reliably narrows the admissible update space, achieving high precision and specificity over 200,000 synthetic audit-stress records. Experiment 3 shows that Bayesian uncertainty can serve as a record-level routing signal, with epistemic uncertainty dominating total uncertainty and low-resource subsampling producing stable estimates. Experiment 4 shows that uncertainty-weighted KG evolution preserves global graph structure, while the protocol-level computational analysis shows that the BNN route selector adds only 0.042% relative computational overhead, corresponding to 0.00225 GFLOPs/token. Finally, Experiment 5 shows that admission coverage is tunable over the full test set with 0% bypass, while the ablations show that ARIS supplies complete routing, uncertainty, audit, and provenance fields absent from unconstrained generation and improves schema-complete structured records from 28.8% to 76.0% over classical NER-to-KG construction. Overall, the experiments show that ARIS improves the traceability, structural completeness, and credibility of LLM-assisted KG updates without requiring unconstrained generation or expensive model adaptation.

5 Discussion and Conclusion

ARIS is motivated by a central challenge in LLM-assisted knowledge management: model-generated content should not automatically become persistent knowledge. A generated statement may be fluent, semantically plausible, and locally useful, but once it is written into a persistent KG, it may be reused in downstream retrieval, reasoning, analytics, or decision-support workflows. ARIS therefore reframes clinical KG expansion as a controlled knowledge-admission problem. Its goal is not to maximize free-form generation quality or to replace clinical judgment, but to determine when model-derived signals are sufficiently structured, traceable, and uncertainty-aware to become eligible for KG update.

The main contribution of ARIS is architectural. The framework separates semantic representation from knowledge admission: LLM-derived representations and clinical encoders provide candidate signals, while deterministic audit constraints, Bayesian uncertainty routing, and provenance requirements determine whether those signals may influence KG evolution. This distinction is important because stronger generative capabilities do not remove the need for governance. Even highly capable models may produce unsupported, weakly grounded, or structurally ambiguous outputs under noisy clinical text. Persistent knowledge systems therefore require explicit admission controls that determine how model-derived information becomes durable managed knowledge.

The experimental results support this admission-control perspective. ARIS achieves complete routing, audit, uncertainty, and provenance records across the evaluated test set while eliminating uncontrolled KG update events. The audit gate narrows the admissible update space with high precision and specificity, the Bayesian gate exposes record-level uncertainty for selective routing, and uncertainty-weighted graph evolution preserves global KG structure with minimal computational overhead. In comparison with unconstrained LLM generation and direct NER-to-KG construction, these findings show that the primary contribution of ARIS is not merely better extraction or more fluent generation. Rather, ARIS improves the governance of persistent KG mutation through controlled, provenance-aware integration.

Several scope boundaries should be acknowledged. ARIS is designed to make LLM-assisted KG updates inspectable, auditable, and selectively admissible under public-data constraints; it is not a prospective clinical deployment study and does not claim expert-validated diagnostic correctness. The uncertainty estimates are used operationally for routing and selective admission, but they are not presented as formal calibration guarantees against clinician adjudication. Similarly, the synthetic audit-stress evaluation tests structural admissibility under controlled low- and high-quality input conditions, rather than serving as a comprehensive adversarial robustness benchmark. Broader longitudinal, multi-institutional, and human-in-the-loop evaluations remain important future directions, particularly for deployment settings where high-uncertainty records may be escalated to clinical experts, data stewards, or domain curators.

Overall, ARIS argues for a broader principle in knowledge management: as LLMs become more capable of generating candidate knowledge signals, persistent information systems need provenance-aware, uncertainty-aware, and audit-controlled admission mechanisms. Although instantiated on clinical EHR data, the framework is domain-general and may support controlled knowledge integration in other regulated environments.

GenAI Usage Disclosure

We used a generative AI tool, OpenAI's ChatGPT, only for language polishing, grammar improvement, wording refinement, and clarity checks during manuscript preparation. The authors reviewed and approved all manuscript content and take full responsibility for the correctness, reproducibility, and integrity of the submitted work.

References

[1] Moloud Abdar, Farhad Pourpanah, Sadiq Hussain, Dana Rezazadegan, Li Liu, Mohammad Ghavamzadeh, Paul Fieguth, Xiaochun Cao, Abbas Khosravi, U. Rajendra Acharya, Vladimir Makarenkov, and Saeid Nahavandi. 2021. A review of uncertainty quantification in deep learning: Techniques, applications and challenges. *Information Fusion* 76 (2021), 243–297. doi:10.1016/j.inffus.2021.05.008

[2] Monica Agrawal, Irene Y. Chen, Freya Gulamali, and Shalmali Joshi. 2025. The evaluation illusion of large language models in medicine. *npj Digital Medicine* 8 (10 2025). doi:10.1038/s41746-025-01963-x

[3] Alfred V Aho and Margaret J Corasick. 1975. Efficient string matching: an aid to bibliographic search. *Commun. ACM* 18, 6 (1975), 333–340.

[4] Emily Alsentzer, John R Murphy, Willie Boag, et al. 2019. Publicly available clinical BERT embeddings. In *Clinical Natural Language Processing Workshop*. 72–78.

[5] Ali Amirahmadi, Mattias Ohlsson, and Kobra Etminani. 2023. Deep learning prediction models based on EHR trajectories: A systematic review. *Journal of Biomedical Informatics* 144 (2023), 104430. doi:10.1016/j.jbi.2023.104430

[6] Elham Asgari, Nina Montaña-Brown, Magda Dubois, Saleh Khalil, Jasmine Balloch, Joshua Au Yeung, and Dominic Pimenta. 2025. A framework to assess clinical safety and hallucination rates of LLMs for medical text summarisation. *npj Digital Medicine* 8 (05 2025), 274. doi:10.1038/s41746-025-01670-7

[7] Charles Blundell, Julien Cornebise, Koray Kavukcuoglu, and Daan Wierstra. 2015. Weight uncertainty in neural networks. In *International Conference on Machine Learning*. 1613–1622.

[8] Edward Choi, Mohammad Taha Bahadori, Andy Schuetz, Walter F. Stewart, and Jimeng Sun. 2016. Doctor AI: Predicting Clinical Events via Recurrent Neural Networks. In *Proceedings of the 1st Machine Learning for Healthcare Conference (Proceedings of Machine Learning Research, Vol. 56)*, Finale Doshi-Velez, Jim Fackler, David Kale, Byron Wallace, and Jenna Wiens (Eds.). PMLR, Northeastern University, Boston, MA, USA, 301–318. https://proceedings.mlr.press/v56/Choi16.html

[9] Cristóbal Esteban, Oliver Staeck, Stephan Baier, Yinchong Yang, and Volker Tresp. 2016. Predicting Clinical Events by Combining Static and Dynamic Information Using Recurrent Neural Networks. In *2016 IEEE International Conference on Healthcare Informatics (ICHI)*. 93–101. doi:10.1109/ICHI.2016.16

[10] Yarin Gal and Zoubin Ghahramani. 2016. Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning. In *Proceedings of The 33rd International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 48)*, Maria Florina Balcan and Kilian Q. Weinberger (Eds.). PMLR, New York, New York, USA, 1050–1059. https://proceedings.mlr.press/v48/gal16.html

[11] Jakob Gawlikowski, Cedrique Rovile Njieutcheu Tassi, Mohsin Ali, Jongseok Lee, Matthias Humt, Jianxiang Feng, Anna Kruspe, Rudolph Triebel, Peter Jung, Ribana Roscher, Muhammad Shahzad, Wen Yang, Richard Bamler, and Xiao Xiang Zhu. 2023. A survey of uncertainty in deep neural networks. *Artificial Intelligence Review* 56 (07 2023), 1513–1589. doi:10.1007/s10462-023-10562-9

[12] Andrew Gelman, John B Carlin, Hal S Stern, et al. 2013. *Bayesian data analysis* (3rd ed.). CRC press.

[13] Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2021. Domain-Specific Language Model Pretraining for Biomedical Natural Language Processing. *ACM Trans. Comput. Healthcare* 3, 1, Article 2 (Oct. 2021), 23 pages. doi:10.1145/3458754

[14] Henk Harkema, John N. Dowling, Tyler Thornblade, and Wendy W. Chapman. 2009. ConText: An algorithm for determining negation, experiencer, and temporal status from clinical reports. *Journal of Biomedical Informatics* 42, 5 (2009), 839–851. doi:10.1016/j.jbi.2009.05.002 Biomedical Natural Language Processing.

[15] Aidan Hogan, Eva Blomqvist, Michael Cochez, Claudia D'amato, Gerard De Melo, Claudio Gutierrez, Sabrina Kirrane, José Emilio Labra Gayo, Roberto Navigli, Sebastian Neumaier, Axel-Cyrille Ngonga Ngomo, Axel Polleres, Sabbir M. Rashid, Anisa Rula, Lukas Schmelzeisen, Juan Sequeda, Steffen Staab, and Antoine Zimmermann. 2021. Knowledge Graphs. *ACM Comput. Surv.* 54, 4, Article 71 (July 2021), 37 pages. doi:10.1145/3447772

[16] George Hripcsak and David J Albers. 2013. Next-generation phenotyping of electronic health records. *Journal of the American Medical Informatics Association* 20 (01 2013), 117–121. doi:10.1136/amiajnl-2012-001145

[17] Kexin Huang, Jaan Altosaar, and Rajesh Ranganath. 2020. ClinicalBERT: Modeling Clinical Notes and Predicting Hospital Readmission. arXiv:1904.05342 [cs.CL] https://arxiv.org/abs/1904.05342

[18] Alon Jacovi and Yoav Goldberg. 2020. Towards Faithfully Interpretable NLP Systems: How should we define and evaluate faithfulness? arXiv:2004.03685 [cs.CL] https://arxiv.org/abs/2004.03685

[19] Alon Jacovi, Oren Sar Shalom, and Yoav Goldberg. 2020. Understanding Convolutional Neural Networks for Text Classification. arXiv:1809.08037 [cs.CL] https://arxiv.org/abs/1809.08037

[20] Peter B. Jensen, Lars J. Jensen, and Søren Brunak. 2012. Mining electronic health records: towards better research applications and clinical care. *Nature Reviews Genetics* 13 (05 2012), 395–405. doi:10.1038/nrg3208

[21] Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of Hallucination in Natural Language Generation. *ACM Comput. Surv.* 55, 12, Article 248 (March 2023), 38 pages. doi:10.1145/3571730

[22] Alistair Johnson, Lucas Bulgarelli, Tom Pollard, Steven Horng, Leo Anthony Celi, and Roger Mark. 2023. MIMIC-IV. *PhysioNet* (Jan. 2023). doi:10.13026/6mm1-ek67 Version 2.2.

[23] Alex Kendall and Yarin Gal. 2017. What uncertainties do we need in Bayesian deep learning for computer vision?. In *Advances in Neural Information Processing Systems*. 5574–5584.

[24] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. 2017. Simple and Scalable Predictive Uncertainty Estimation using Deep Ensembles. In *Advances in Neural Information Processing Systems*, I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Eds.), Vol. 30. Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2017/file/9ef2ed4b7fd2c810847ffa5fa85bce38-Paper.pdf

[25] Katherine P. Liao, Ashwin N. Ananthakrishnan, Vishesh Kumar, Zongqi Xia, Andrew Cagan, Vivian S. Gainer, Sergey Goryachev, Pei Chen, Guergana K. Savova, Denis Agniel, Susanne Churchill, Jaeyoung Lee, Shawn N. Murphy, Robert M. Plenge, Peter Szolovits, Isaac Kohane, Stanley Y. Shaw, Elizabeth W. Karlson, and Tianxi Cai. 2015. Methods to Develop an Electronic Medical Record Phenotype Algorithm to Compare the Risk of Coronary Artery Disease across 3 Chronic Disease Cohorts. *PLOS ONE* 10, 8 (08 2015), 1–11. doi:10.1371/journal.pone.0136651

[26] Yang Liu, Yuanshun Yao, Jean-Francois Ton, Xiaoying Zhang, Ruocheng Guo, Hao Cheng, Yegor Klochkov, Muhammad Faaiz Taufiq, and Hang Li. 2024. Trustworthy LLMs: a Survey and Guideline for Evaluating Large Language Models' Alignment. arXiv:2308.05374 [cs.AI] https://arxiv.org/abs/2308.05374

[27] Andreas Madsen, Siva Reddy, and Sarath Chandar. 2022. Post-hoc Interpretability for Neural NLP: A Survey. *ACM Comput. Surv.* 55, 8, Article 155 (Dec. 2022), 42 pages. doi:10.1145/3546577

[28] Stéphane M Meystre, Guergana K Savova, Karin C Kipper-Schuler, and John F Hurdle. 2008. Extracting information from textual documents in the electronic health record: a review of recent research. *Yearbook of medical informatics* 17, 01 (2008), 128–144.

[29] Riccardo Miotto, Li Li, Brian A Kidd, and Joel T Dudley. 2016. Deep Patient: An Unsupervised Representation to Predict the Future of Patients from the Electronic Health Records. *Scientific Reports* 6 (05 2016), 26094–26094. doi:10.1038/srep26094

[30] Wojciech Muła et al. 2016. pyahocorasick: Python module implementing Aho-Corasick algorithm. https://github.com/WojciechMula/pyahocorasick. Version 2.0.0.

[31] Harsha Nori, Nicholas King, Scott Mayer McKinney, Dean Carignan, and Eric Horvitz. 2023. Capabilities of GPT-4 on Medical Challenge Problems. arXiv:2303.13375 [cs.CL] https://arxiv.org/abs/2303.13375

[32] Alvin Rajkomar, Eyal Oren, Kai Chen, Andrew M. Dai, Nissan Hajaj, Michaela Hardt, Peter J. Liu, Xiaobing Liu, Jake Marcus, Mimi Sun, Patrik Sundberg, Hector Yee, Kun Zhang, Yi Zhang, Gerardo Flores, Gavin E. Duggan, Jamie Irvine, Quoc Le, Kurt Litsch, Alexander Mossin, Justin Tansuwan, De Wang, James Wexler, Jimbo Wilson, Dana Ludwig, Samuel L. Volchenboum, Katherine Chou, Michael Pearson, Srinivasan Madabushi, Nigam H. Shah, Atul J. Butte, Michael D. Howell, Claire Cui, Greg S. Corrado, and Jeffrey Dean. 2018. Scalable and accurate deep learning with electronic health records. *npj Digital Medicine* 1 (05 2018). doi:10.1038/s41746-018-0029-1

[33] Laila Rasmy, Yang Xiang, Ziqian Xie, Cui Tao, and Degui Zhi. 2021. Med-BERT: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction. *npj Digital Medicine* 4 (05 2021), 86–86. doi:10.1038/s41746-021-00455-y

[34] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. "Why Should I Trust You?": Explaining the Predictions of Any Classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (San Francisco, California, USA) *(KDD '16)*. Association for Computing Machinery, New York, NY, USA, 1135–1144. doi:10.1145/2939672.2939778

[35] Guergana K Savova, James J Masanz, Philip V Ogren, Jiaping Zheng, Sunghwan Sohn, Karin C Kipper-Schuler, and Christopher G Chute. 2010. Mayo clinical Text Analysis and Knowledge Extraction System (cTAKES): architecture, component evaluation and applications. *Journal of the American Medical Informatics Association* 17, 5 (09 2010), 507–513. arXiv:https://academic.oup.com/jamia/article-pdf/17/5/507/5940551/17-5-507.pdf doi:10.1136/jamia.2009.001560

[36] Benjamin Shickel, Patrick James Tighe, Azra Bihorac, and Parisa Rashidi. 2018. Deep EHR: A Survey of Recent Advances in Deep Learning Techniques for Electronic Health Record (EHR) Analysis. *IEEE Journal of Biomedical and Health Informatics* 22, 5 (2018), 1589–1604. doi:10.1109/JBHI.2017.2767063

[37] Chaitanya Shivade, Preethi Raghavan, Eric Fosler-Lussier, Peter J Embi, Noemie Elhadad, Stephen B Johnson, and Albert M Lai. 2013. A review of approaches to identifying patient phenotype cohorts using electronic health records. *Journal of the American Medical Informatics Association* 21, 2 (11 2013), 221–230. arXiv:https://academic.oup.com/jamia/article-pdf/21/2/221/17375567/21-2-221.pdf doi:10.1136/amiajnl-2013-001935

[38] Luciana Kase Tanno, Moises Calderon, and Pascal Demoly. 2016. Supporting the validation of the new allergic and hypersensitivity conditions section of the World Health Organization International Classification of Diseases-11. *Asia Pacific Allergy* 6 (07 2016), 149–156. doi:10.5415/apallergy.2016.6.3.149

[39] Gemma Team et al. 2024. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295* (2024).

[40] Yanshan Wang, Liwei Wang, Majid Rastegar-Mojarad, Sungrim Moon, Feichen Shen, Naveed Afzal, Sijia Liu, Yuqun Zeng, Saeed Mehrabi, Sunghwan Sohn, and Hongfang Liu. 2018. Clinical information extraction applications: A literature review. *Journal of Biomedical Informatics* 77 (2018), 34–49. doi:10.1016/j.jbi.2017.11.011

[41] World Health Organization. 2004. *ICD-10: International Statistical Classification of Diseases and Related Health Problems* (2 ed.). World Health Organization, Geneva. https://iris.who.int/handle/10665/42980

[42] Siyue Yang, Paul Varghese, Ellen Stephenson, Karen Tu, and Jessica Gronsbell. 2022. Machine learning approaches for electronic health records phenotyping: a methodical review. *Journal of the American Medical Informatics Association* 30 (10 2022), 367–381. doi:10.1093/jamia/ocac216

[43] Xiangrong Zeng, Shizhu He, Kang Liu, and Jun Zhao. 2018. Large Scaled Relation Extraction With Reinforcement Learning. *Proceedings of the AAAI Conference on Artificial Intelligence* 32 (04 2018). doi:10.1609/aaai.v32i1.11950

[44] Zexian Zeng, Yu Deng, Xiaoyu Li, Tristan Naumann, and Yuan Luo. 2019. Natural Language Processing for EHR-Based Computational Phenotyping. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 16, 1 (2019), 139–153. doi:10.1109/TCBB.2018.2849968