

PRISM: Probabilistic Reasoning through Intelligent Situational Modeling - A Dynamic Context Engineering Framework for Uncertainty-Aware Agentic AI in Healthcare Administrative Workflows

Muthukumarapandian Chandrasekaran*

LTM Limited, Texas, USA

*Corresponding author: Muthukumarapandian Chandrasekaran, LTM Limited, Texas, USA & Email: muthuchand2013@gmail.com

ABSTRACT: Large language models (LLMs) used in healthcare administrative workflows often fail because the context they receive is incomplete, not because the model lacks capability. In prior authorization, a retrieved document set can be topically relevant yet still miss the evidence that a payer policy requires. Physicians report completing about 45 prior authorizations per week and 94% report associated care delays, so reliable automation matters. This paper presents PRISM (Probabilistic Reasoning through Intelligent Situational Modeling), a Dynamic Context Engineering (DCE) framework that runs as pre-inference middleware between healthcare data systems and an LLM. PRISM composes six mechanisms: Context Evidence Sufficiency scoring (CES), a Clinical Criteria Chain (CCC), Criteria-Seeded Gap retrieval (CSG), Context Mutation Tracking (CMT), Adversarial Context Hardening (ACH), and Uncertainty-Governed Escalation (UGE). Together they check a context bundle against policy requirements, retrieve missing evidence, record every transformation, and escalate when context quality is insufficient. On 1,200 synthetic prior authorization cases authored against InterQual 2024 criteria, PRISM reaches a determination F1 of 0.912 against 0.769 for dense retrieval with chain-of-thought prompting (18.6% relative improvement) and 0.710 for BM25 retrieval (28.5%), and lowers context brittleness from 0.158 to 0.049 (69.0%). Because the cases and the PRISM policy templates share one source, these results are an upper-bound estimate that requires confirmation on independently annotated and real-world data.

KEYWORDS: Context engineering, large language models, retrieval-augmented generation, prior authorization, healthcare administrative AI, uncertainty quantification, evidence graphs, adversarial robustness, regulatory compliance, agentic AI

1. Introduction

Natural language processing can, in principle, parse clinical notes, match documented evidence against payer coverage criteria, identify documentation gaps, and recommend coverage determinations faster and more consistently than manual review. In practice, deployed systems underperform laboratory benchmarks, and a common failure lies in the context supplied to the model rather than in its parametric knowledge or general reasoning ability.

The administrative burden is large. Physicians report completing about 45 prior authorizations per week, consuming nearly two business days of physician and staff time, and 94% report that prior authorization delays care [1]. A federal audit found that 13% of prior authorization requests denied by Medicare Advantage organizations met Medicare coverage rules [2], and administrative complexity accounts for an estimated \$266 billion of the \$760 to \$935 billion in annual U.S. health care waste [3].

Healthcare administrative decisions rest on three kinds of evidence: clinical evidence (diagnoses, procedure history, laboratory values, imaging), policy criteria (InterQual, MCG, or payer-specific coverage logic), and documentation standards that map clinical evidence to policy satisfaction.

Each is spread across separate systems, changes on its own schedule, and must be assembled into a coherent bundle before reasoning can begin. Retrieval-augmented generation (RAG) retrieves documents; it does not assemble policy-sufficient bundles. This distinction is the central problem addressed here.

1.1. The Context Assembly Problem

Let a prior authorization case X be defined by a procedure code p , a patient clinical profile $\mathcal{E}$, and a payer policy criterion set Π_p . A coverage determination $d \in \{\text{approve, deny, pend}\}$ can be made correctly only if

$$d \models C \text{ where } C \supseteq C_p^*$$

and C_p^* is the minimal sufficient context for policy Π_p . Standard RAG produces a bundle C ranked by topical similarity to X , without verifying that $C \supseteq C_p^*$. The result is systematic under-coverage: retrieved bundles are topically relevant but incomplete with respect to the policy.

1.2. Three Structural Failure Modes

Context fragmentation: clinical evidence is scattered across disconnected systems (electronic health record notes,

imaging, laboratory results, authorization history), so the bundle lacks the evidentiary chain that a human reviewer builds implicitly. *Missing evidentiary linkage*: policy criteria form dependency chains, yet RAG retrieves evidence items without the links between them. A confirmed L4-L5 herniation is necessary but not sufficient without the chain from diagnosis through conservative-care failure to surgical indication. *Overconfident decisions*: on ambiguous documentation, LLMs can emit high-confidence outputs that hide evidence gaps, and reviewers cannot audit them because the evidentiary basis is not exposed.

1.3. Research Gap

Existing work improves retrieval relevance, reasoning over a given context, or output-level uncertainty (Section 3). None of these treats the policy-defined minimal sufficient context C_p^* as an explicit target that is checked before inference. Corrective and self-reflective RAG judge retrieved passages against the query, graph RAG organizes a corpus by topic, and uncertainty methods measure the model's output. No existing method combines (i) a sufficiency check against a policy checklist, (ii) an explicit dependency structure over the evidence, (iii) retrieval targeted at the specific missing elements, (iv) a tamper-evident record of context changes, and (v) escalation driven by context deficits rather than model confidence. Table 1 summarizes this gap, and PRISM is designed to fill it.

1.4. Contributions

PRISM intercepts the raw context bundle before LLM inference, passes it through six stages, and delivers a structured, uncertainty-annotated context object to the reasoning layer. It builds on two prior streams of work. AURELIUS showed that uncertainty-aware hierarchical multi-agent reasoning with threshold-governed escalation improves safety and calibration in clinical AI [4]. U.S. Patent No. 12,475,154 B1 described an AI field-service assistance architecture that combines contextual retrieval, vectorization, adaptive optimization, and real-time context delivery at the point of dispatch [5]. PRISM applies these ideas to the context object itself. The contributions are:

- **Dynamic Context Engineering** as a pre-inference layer for healthcare administrative AI, complementary to prompt engineering and model fine-tuning.
- **Six specified mechanisms** (CES, CCC, CSG, CMT, ACH, UGE), each with pseudocode or equations and the operating parameters used.
- **A worked case study** tracing a lumbar spinal fusion authorization through the full pipeline.
- **An evaluation** on 1,200 synthetic cases with an 18.6% relative F1 improvement over dense retrieval with chain-of-thought prompting, a per-mechanism ablation, and an explicit statement of the dataset's limits.
- **A compliance mapping** relating each mechanism to CMS, ONC HTI-1, and HIPAA requirements.

2. Problem Formalization

Let C denote the context object for an administrative case, defined as the tuple

$$C \langle D, G, \sigma, \mathbf{q}, \kappa \rangle \quad (1)$$

where $D = \{d_1, \dots, d_n\}$ is the document set, $G = V, E, \mathbf{w}$ is the criteria-dependency graph over clinical entities, $\sigma : D \rightarrow [0, 1]$ is a provenance confidence function, $\mathbf{q} \in [0, 1]^5$ is the quality vector (completeness, consistency, linkage density, temporal coherence, source authority), and $\kappa \in [0, 1]$ is the aggregate context confidence.

A context object is *inference-ready* if and only if

$$\kappa C \geq \tau_v \quad \text{and} \quad \text{CESC} \geq \tau_{\text{CES}} \quad \text{and} \quad \Delta D = \emptyset \quad (2)$$

where τ_v is the vertical-specific escalation threshold, τ_{CES} is the minimum sufficiency score, and ΔD is the set of required but missing documentation elements. When (2) does not hold, PRISM routes the case through a remediation pathway before inference.

Standard RAG produces the approximation

$$C_{\text{RAG}} \langle D_k, \emptyset, \sigma, \mathbf{0}, \kappa \rangle$$

where D_k is the top- k retrieved documents, no dependency graph is built, σ is an uncalibrated retrieval score, no quality vector is computed, and κ reflects retrieval rank only. Transforming C_{RAG} into a context object that satisfies (2) is the problem PRISM addresses. A suitable framework should provide:

1. *Completeness verification*: check whether retrieved context satisfies policy-mandated documentation requirements.
2. *Explicit evidentiary structure*: make the chain from diagnosis to coverage criteria explicit and traversable.
3. *Uncertainty propagation*: map context quality deficits to reduced decision confidence.
4. *Auditability*: keep a tamper-evident record of all context transformations.
5. *Adversarial robustness*: keep determinations stable under realistic manipulation of the documentation bundle.

3. Related Work

Retrieval-augmented generation in medicine. RAG [6] is widely used in clinical NLP for decision support [7], medical coding [8], and summarization [9]. Domain-adapted models such as HuatuoGPT [10] and Med-PaLM 2 [11] improve factual accuracy. Recent work benchmarks medical RAG across corpora and retrievers [12] and grounds clinical answers in cited sources [13]. These systems are information-seeking: they retrieve passages relevant to a question and do not test whether the retrieved set satisfies a policy checklist.

Corrective, self-reflective, and graph RAG. Self-RAG trains the model to decide when to retrieve and to critique retrieved passages [14], and Corrective RAG (CRAG) adds a retrieval evaluator that triggers further retrieval when passages look unreliable [15]. GraphRAG builds an entity graph over a corpus to answer global questions [16],

and Medical Graph RAG applies graph retrieval to clinical evidence [17].

Agentic AI and context engineering. Agentic RAG lets an LLM plan and call retrieval tools iteratively [18], and multi-agent systems such as MedAgents [19] and MDAgents [20] coordinate several LLM roles for medical reasoning. MedAgentBench evaluates agents that act on a virtual record through FHIR (Fast Healthcare Interoperability Resources) interfaces [21]. A recent survey frames context engineering as the systematic design of what enters the context window [22], and long-context studies show that models use information unevenly depending on its position [23]. Prompting methods, including chain-of-thought (CoT) [24], self-consistency [25], least-to-most [26], Constitutional AI [27], AutoPrompt [28], DSPy [29], and ReAct [30], assume the bundle is adequate. PRISM governs the context object before any prompting strategy applies.

Uncertainty. Calibrated uncertainty is well studied in clinical prediction [31], and uncertainty-based referral of cases to radiologists has been shown to improve screening performance [32]. For LLMs, semantic entropy detects confabulation from output variability [33], and confidence elicitation studies show that verbalized confidence is often miscalibrated [34]. AURELIUS used threshold-governed escalation on model uncertainty and reached AUROC 0.952 on MIMIC-IV [4]. UGE moves this upstream: escalation depends on context quality, evidentiary linkage, and adversarial stability rather than model output alone.

Causal notation, robustness, and compliance. CCC borrows directed-graph notation from causal inference [35, 36] and relates to work on LLMs and causal reasoning [37], but it does not estimate causal effects (Section 6). Adversarial studies of NLP target generic perturbations such as misspellings, word substitutions, and universal triggers [38, 39, 40], while ACH targets the authorization decision with clinically plausible documentation manipulations. ONC HTI-1 [41] and CMS prior authorization rules [42, 43] set transparency and documentation expectations that CMT addresses through lineage recorded during processing rather than added afterwards.

Table 1 compares representative approaches on the five properties listed in Section 2 plus targeted gap retrieval.

4. PRISM Architecture

PRISM is a typed middleware layer between healthcare data systems and LLM inference, implementing the transformation

$$\mathcal{P} : C_{\text{raw}} \rightarrow C_{\text{structured}} \quad (3)$$

where $C_{\text{raw}} \langle D_{\text{retrieved}}, \theta, \sigma, \mathbf{0}, \kappa \rangle$ is the unprocessed RAG output and $C_{\text{structured}}$ satisfies (2).

Data ingestion covers three source categories aligned with FHIR R4 resource types: clinical documents (DocumentReference, DiagnosticReport, ClinicalImpression), administrative records (Claim, ClaimResponse, Coverage), and policy artifacts (payer policy documents, InterQual criteria tables, formulary data). Parsing of ASC X12 270/271 eligibility and 278 services-review transactions supplies electronic data interchange

(EDI) case metadata. The six mechanisms form a directed pipeline:

$$\begin{aligned} C_{\text{raw}} &\xrightarrow{\text{CES}} C_1 \xrightarrow{\text{CCC}} C_2 \xrightarrow{\text{CSG}} C_3 \\ &\xrightarrow{\text{CMT}} C_4 \xrightarrow{\text{ACH}} C_5 \xrightarrow{\text{UGE}} C_{\text{structured}} \end{aligned} \quad (4)$$

Each stage emits one of four directives: proceed (pass C_i on), augment (invoke CSG before proceeding), escalate (route to human review with an uncertainty report), or reject (stop and return a deficiency report). Figure 1 shows the arrangement.

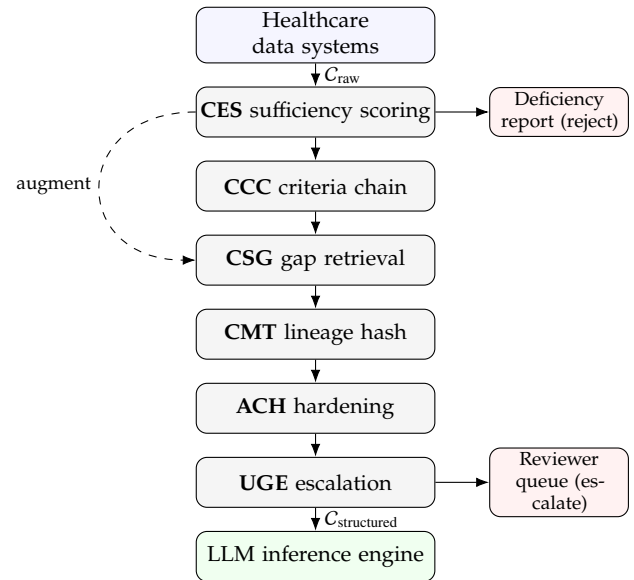


Figure 1: PRISM as a pre-inference middleware layer. A raw bundle C_{raw} from FHIR R4 and X12 sources passes through the six stages to the LLM as a structured object $C_{\text{structured}}$. A stage may invoke gap retrieval (augment), route to a reviewer (escalate), or return a deficiency report (reject).

PRISM runs on Azure Kubernetes Service (AKS) with namespace isolation per deployment vertical, each mechanism an independently scalable pod. The CMT hash chain is persisted to an append-only Cosmos DB collection under a write-once-read-many (WORM) policy. CSG issues FHIR R4 queries through an API Management gateway under OAuth 2.0 with SMART on FHIR, and UGE routes escalations to a reviewer queue as FHIR Task resources.

5. Context Evidence Sufficiency (CES)

An X12 278 request may carry as few as three elements (procedure code, diagnosis code, date of service), whereas a complete bundle adds clinical notes, imaging or laboratory evidence, a dated conservative-care history, a functional limitation assessment, and a necessity attestation. CES measures how far a bundle is from the minimum sufficient context for its procedure category. CES is a weighted sufficiency score; it is not an information-theoretic entropy.

CES is a weighted average of five quality dimensions, and Algorithm 1 gives the scoring and routing procedure:

$$\text{CESC} = \alpha C_{\text{comp}} + \beta C_{\text{cons}} + \gamma C_{\text{link}} + \delta C_{\text{temp}} + \epsilon C_{\text{source}} \quad (5)$$

The components are: completeness $C_{\text{comp}} = |D_{\text{available}} \cap D_{\text{required}}| / |D_{\text{required}}|$; consistency C_{cons}

Table 1: Comparison of PRISM with existing approaches

Approach	Main limitation for prior authorization	Co	Gr	Ga	Au	Ad	Es
Standard and dense RAG [6, 44]	Top- k similarity; no check against policy criteria	×	×	×	×	×	×
CoT, ReAct, DSPy [24, 30, 29]	Reasons over a given bundle; no sufficiency target	×	×	○	×	×	×
Self-RAG, CRAG [14, 15]	Re-retrieval judged against the query, not a policy checklist	○	×	○	×	×	○
GraphRAG, Medical Graph RAG [16, 17]	Topical entity graph, not a policy dependency structure	×	✓	×	×	×	×
Medical RAG [12, 13]	Cited answers, but information-seeking, not adjudicatory	×	×	×	○	×	×
Clinical LLM agents [19, 20, 21]	Multi-agent or FHIR tool use; no sufficiency gate or lineage	×	×	○	×	×	×
LLM uncertainty [33, 34]	Measures model output, not context deficits	×	×	×	×	×	○
AURELIUS [4]	Escalation driven by model uncertainty, not context	×	×	×	○	×	✓
PRISM (this work)	Evaluated on synthetic data only	✓	✓	✓	✓	✓	✓

✓ supported, ○ partial, × absent. Co: policy completeness check; Gr: evidence dependency graph; Ga: retrieval targeted at missing elements; Au: tamper-evident audit lineage; Ad: adversarial hardening; Es: context-driven escalation.

$\binom{n}{2}^{-1} \sum_{i < j} \text{agreed}_i, d_j$, the mean pairwise agreement under a contradiction-detection model; linkage density $C_{\text{link}} \min 1, |\text{linkage_spans}|/n_{\text{required}}$, the share of required evidence links supported by explicit connective phrases found by dependency parsing; temporal coherence $C_{\text{temp}} 1 - |\text{temporal_violations}|/|\text{temporal_pairs}|$; and source authority C_{source} , a four-tier provenance prior (attending at-testament 1.0, specialist 0.85, referring provider 0.65, patient-reported 0.40).

The default weights are α 0.30, β 0.20, γ 0.20, δ 0.15, ϵ 0.15. In deployment they are refit per authorization type as payer criteria change. The weights were calibrated on a labeled historical case cohort and were held fixed for all experiments in Section 12.

Algorithm 1: Context evidence sufficiency scoring

Input: Raw context C_{raw} , procedure type p , threshold τ_{CES}
Output: Annotated context C_1 , directive δ_1
 $D_{\text{req}} \leftarrow \text{LOADCRITERIA}(p)$
Compute $C_{\text{comp}}, C_{\text{cons}}, C_{\text{link}}, C_{\text{temp}}, C_{\text{source}}$ over C_{raw}
 $\text{CES} \leftarrow \alpha C_{\text{comp}} \beta C_{\text{cons}} \gamma C_{\text{link}} \delta C_{\text{temp}} \epsilon C_{\text{source}}$
Append CES to $\mathbf{q}$ in C_1
if $\text{CES} \geq \tau_{\text{CES}}$ **then**
| $\delta_1 \leftarrow \text{proceed}$
else if $C_{\text{comp}} < 0.50$ **then**
| $\delta_1 \leftarrow \text{augment}$ // activate CSG
else if $C_{\text{cons}} < 0.40$ **then**
| $\delta_1 \leftarrow \text{escalate}$ // contradictory evidence
else
| Generate a deficiency report listing low $\mathbf{q}$ entries
| $\delta_1 \leftarrow \text{reject}$
end
return C_1, δ_1

The threshold τ_{CES} 0.62 was selected on a 30-day opera-

tional calibration cohort that is separate from the synthetic evaluation, where CES routing agreed with human reviewer dispositions in 89.4% of cases.

6. Clinical Criteria Chain (CCC)

Coverage criteria for complex procedures form dependency chains. InterQual 2024 criteria for lumbar spinal fusion, for example, require a confirmed disc diagnosis at a level, attributable radiculopathy or myelopathy, failure of conservative treatment over a minimum duration, absence of contraindications, and a validated functional limitation score. CCC makes this chain explicit and inspectable.

Let V be a set of clinical entity nodes: diagnoses (V_{dx}), procedures (V_{px}), medications (V_{rx}), laboratory and imaging values (V_{lab}), policy clauses (V_{pol}), and outcomes (V_{out}). Edges $E \subseteq V \times V$ are identified in three ways: (i) dependency parsing of connective spans in clinical notes, (ii) traversal of the MeSH and SNOMED CT hierarchies for known clinical relationships, and (iii) lookup in a payer-specific policy graph that maps coverage criteria to required evidence nodes.

$$G_{\text{CCC}} \ V, E, \mathbf{w} \quad (6)$$

The weight $w_{u,v}$ is the geometric mean of the parsing confidence, ontology support score, and policy match score for edge u, v . These weights measure how confidently a required evidentiary link is documented. They are not causal effect estimates, and CCC makes no interventional claim. Edges with $w_{u,v} < \tau_{\text{edge}}$ 0.40 are kept but flagged as *speculative*, telling the LLM that the link is unconfirmed.

Given the policy-mandated dependency graph G_p^* for procedure p and the constructed graph G_{CCC} , the criteria gap is

$$\Delta G \quad EG_p^* \setminus EG_{CCC}. \quad (7)$$

$|\Delta G| = 0$ is a necessary but not sufficient condition for approval. When $|\Delta G| > 0$, the gap edges seed CSG retrieval. When $|\Delta G| > \tau_{\Delta G} - 3$, the case is escalated regardless of CES, because a gap of this size is unlikely to be closed by retrieval alone.

7. Criteria-Seeded Gap Retrieval (CSG)

Many avoidable denials stem from documentation that exists in the record but was left out of the submitted bundle. CSG asks which documentation, if retrieved, would make the bundle meet the policy. The question is counterfactual in form, but the mechanism is targeted retrieval: CSG issues queries for existing records and never generates clinical content. The missing set is

$$\Delta D \quad D_{\text{required}} \setminus D_{\text{available}}. \quad (8)$$

For each element $d \in \Delta D$, CSG issues a FHIR R4 query for the most probable resource type for that element class, parameterized by patient identifier, a date range derived from the surrounding criteria chain, and SNOMED CT or LOINC codes for the required evidence type.

A coverage assessor checks whether each returned resource satisfies the policy element at the semantic level, not only the resource-type level. A *DiagnosticReport* for a lumbar magnetic resonance imaging (MRI) study that does not show disc herniation is a *partial match*, and the element stays in ΔD with an *insufficient* annotation.

After retrieval, $C_3 \leftarrow C_2 \cup D_{\text{retrieved}}$ and CES is recomputed. If the new score meets τ_{CES} , processing proceeds. Elements still missing after retrieval are marked *confirmed-absent*, which instructs the LLM to acknowledge the gap rather than reason across it. Section 13.4 reports retrieval outcomes.

8. Context Mutation Tracking (CMT)

The ONC HTI-1 Final Rule sets transparency and source-attribute requirements for decision support interventions in certified health IT [41]. CMS rules require that prior authorization denials state specific reasons and that the process be documented and reviewable [42, 43]. The HIPAA Security Rule (45 CFR 164.312(b)) requires audit controls for systems that handle electronic protected health information (ePHI). CMT supports all three with one lineage mechanism.

Each transformation of the context bundle is recorded as

$$T_i \langle \text{type, stage, agent, delta, ces_before, ces_after, ts} \rangle$$

where *type* is one of *augment*, *prune*, *rerank*, *annotate*, or *retrieve*, *agent* is the PRISM mechanism that made the change, *delta* lists the document identifiers added or removed, and *ts* is a UTC timestamp from a trusted time source. The records form a hash chain

$$H_i \text{ SHA-256}(\text{canonical}T_i || H_{i-1}) \quad (9)$$

with H_0 seeded from the hash of the X12 278 transaction identifier, linking the initial submission to the final context object.

The chain is exposed through a FHIR R4 *AuditEvent* endpoint, so appeals reviewers and regulators can reconstruct the exact context object the LLM received and the sequence of transformations that produced it.

9. Adversarial Context Hardening (ACH)

ACH assumes an administrative rather than a technical adversary: a submitter under revenue pressure who knows payer coverage criteria and seeks a bundle that maximizes approval probability with weaker clinical evidence than the criteria require. The perturbation set $\mathcal{A}$ has four categories: *severity inflation*, replacing an ICD-10 code with a higher-acuity adjacent code (for example M51.16 to M51.06) without supporting imaging; *temporal manipulation*, reordering treatment events to suggest conservative-care failure; *policy keyword insertion*, adding approval-criteria phrases (such as “failed 6 weeks of conservative treatment”) that the clinical evidence does not support; and *source spoofing*, relabeling low-authority documents with high-authority authority.

For a batch of N perturbations, robustness is

$$R_{\text{ACH}} C \quad \frac{||\{i : aC_i^{\text{perturbed}} \neq aC\}||}{N} \quad (10)$$

where $a \cdot$ is the coverage determination. Contexts with $R_{\text{ACH}} < \tau_{\text{ACH}} = 0.75$ are routed for additional scrutiny and excluded from autonomous determination. In evaluation, PRISM reached $R_{\text{ACH}} = 0.847$ against 0.631 for BM25 RAG.

10. Uncertainty-Governed Escalation (UGE)

UGE follows an AURELIUS principle [4]: in safety-critical decisions, declining to decide autonomously has value, because a wrong autonomous determination costs more than routing to a human reviewer. PRISM decomposes context uncertainty into three sources. *Completeness uncertainty* is the normalized shortfall of CES below its threshold:

$$U_{\text{comp}} \quad \max\left(0, \frac{\tau_{\text{CES}} - \text{CES}}{\tau_{\text{CES}}}\right).$$

Linkage uncertainty is the share of speculative edges in G_{CCC} , where E counts confirmed edges:

$$U_{\text{link}} \quad \frac{|E_{\text{speculative}}|}{|E| + |E_{\text{speculative}}|}.$$

Adversarial uncertainty is the ACH robustness deficit:

$$U_{\text{adv}} \quad \max(0, \tau_{\text{ACH}} - R_{\text{ACH}}).$$

The aggregate uncertainty is

$$U_q \quad \lambda_1 U_{\text{comp}} + \lambda_2 U_{\text{link}} + \lambda_3 U_{\text{adv}} \quad (11)$$

with $\lambda_1 = 0.50$, $\lambda_2 = 0.30$, $\lambda_3 = 0.20$, held fixed for all reported experiments. The context confidence is $\kappa q = 1 - U_q$, and for a vertical-specific threshold τ_v the decision rule is

$$a^*q \begin{cases} \arg \max_a \text{Pra} \mid C_{\text{structured}}, & \kappa q \geq \tau_v \\ \text{escalate } q, Uq, \text{reason}, & \kappa q < \tau_v. \end{cases} \quad (12)$$

Thresholds rise with clinical consequence across the four deployment verticals: health plans and payers (HHP) τ 0.80, providers (HPR) τ 0.85, medical technology (HMT) τ 0.88, and life sciences (HLS) τ 0.90. Prior authorization falls under HHP. The payload **reason** lists the uncertainty sources that triggered escalation, so a reviewer can examine those items instead of re-adjudicating the whole case.

11. Clinical Case Study: Lumbar Spinal Fusion Authorization

We trace a lumbar spinal fusion case (CPT 22633) through the pipeline. The bundle contains an X12 278 transaction (ICD-10 M51.16, CPT 22633, date of service), an outpatient note from the referring provider documenting low back pain with left-leg radiculopathy, and an MRI report confirming L4-L5 disc herniation with moderate foraminal stenosis.

CES loads the InterQual lumbar fusion template. Of five required element classes, only the diagnosis and radiculopathy are present; the conservative-treatment record, the functional limitation score, and the surgical-risk clearance are absent. With C_{comp} 0.40, C_{cons} 0.91, C_{link} 0.30, C_{temp} 0.88, and C_{source} 0.65, (5) gives CES 0.59 $<$ τ_{CES} 0.62, and since $C_{\text{comp}} < 0.50$, CES emits *augment*.

CCC builds the policy graph for CPT 22633. The link herniation $\rightarrow$ radiculopathy is present (w 0.87), while radiculopathy $\rightarrow$ conservative failure and conservative failure $\rightarrow$ surgical indication are absent ($|\Delta G|$ 2). These gaps seed CSG with the conservative-treatment record and the functional limitation assessment. CSG queries FHIR R4 resources (SNOMED CT 229070002 for physical therapy, RxNorm 41493 for nonsteroidal anti-inflammatory drugs) over a 12-month window and returns six physical therapy visits across 8 weeks, a medication fill history, and a chiropractic referral. The radiculopathy $\rightarrow$ conservative failure link is now supported, but its weight falls below τ_{edge} , so it is flagged speculative. The Oswestry Disability Index (ODI) or visual analog scale (VAS) assessment is not found and is marked *confirmed-absent*, so $|\Delta G|$ 1. Recomputing with C_{comp} 0.60, C_{link} 0.67, and C_{source} 0.72 gives CES 0.74, which clears τ_{CES} . Figure 2 shows the resulting chain.

CMT records the template load, the CES annotation, and the CSG augmentation as H_3 SHA-256canonical $T_3||H_2$. ACH yields R_{ACH} 0.81 $>$ τ_{ACH} , so U_{adv} 0. CES is above threshold, so U_{comp} 0. One of the three chain links (MRI $\rightarrow$ herniation, herniation $\rightarrow$ radiculopathy, radiculopathy $\rightarrow$ conservative failure) is speculative, so U_{link} 13 and Uq 0.30 $\times$ 13 0.10. With κ 0.90 $>$ τ_{HHP} 0.80, UGE emits *proceed*. Given the retrieved documents, the annotated graph, and the quality vector, the LLM returns *pend*: conservative treatment is documented and surgical-indication criteria are partly met, but the functional limitation assessment is missing, so it requests an ODI or VAS score. The outcome is a targeted, auditable *pend* instead of a blanket denial or an overconfident approval.

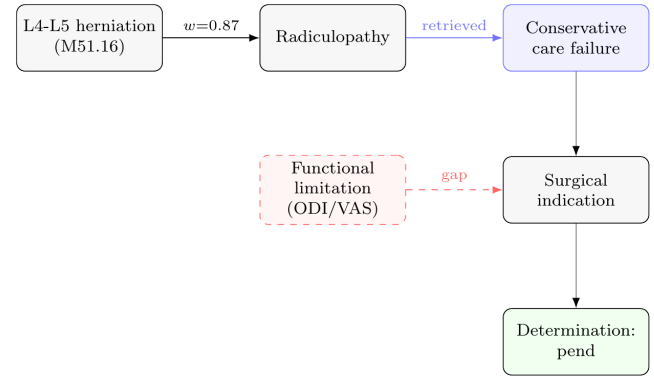


Figure 2: Clinical criteria chain for the lumbar spinal fusion example. Plain nodes are present in the submitted bundle, the blue node is recovered by gap retrieval, and the red dashed node is confirmed absent. The remaining gap, a missing functional limitation assessment, blocks a complete surgical-indication chain and yields a targeted *pend* rather than a denial.

12. Experimental Setup

12.1. Dataset Generation and Annotation

The corpus contains 1,200 synthetic prior authorization cases in four procedure categories with 300 cases each: lumbar spinal fusion, knee arthroplasty, advanced imaging, and specialty pharmaceuticals. Synthetic data was used because real authorization bundles contain protected health information and payer-proprietary decisions. Cases were produced in four steps.

1. *Criteria extraction.* For each category, a three-member panel of clinical informaticists encoded the InterQual 2024 criteria as a required-element list and a dependency graph G_p^* .
2. *Completeness assignment.* Each case was assigned a completeness stratum matching the distribution observed at a large U.S. commercial payer: 35% complete, 40% partially complete, and 25% substantially incomplete. Elements were removed at random within the stratum's range.
3. *Document authoring.* The panel prepared the X12 278 metadata and clinical documents (notes, imaging reports, medication and therapy histories) for the elements retained in each case.
4. *Labeling.* Each case received a ground-truth determination (approve, deny, or *pend*) and a gap annotation listing the missing required elements, assigned by the panel against the InterQual 2024 criteria.

A 720-case test set (180 per category) was held out, and all results below are reported on it.

Shared-schema risk. The PRISM policy templates and the case labels both derive from InterQual 2024, so PRISM may benefit from sharing the annotation schema with the cases. The baselines see the same policy text but not the structured templates. The size of this effect is not measured in this study and is discussed in Section 15.

12.2. Baselines

Every baseline uses a GPT-4-class LLM and draws on the same per-case document pool, which holds the case documents plus the applicable policy text. The baselines are:

1. *LLM-only*: the X12 metadata and policy text, without retrieval of clinical documents.
2. *BM25 RAG*: BM25 keyword retrieval of the top 5 documents.
3. *Dense RAG*: BioBERT [44] embedding retrieval of the top 5 documents.
4. *Dense RAG + CoT*: dense retrieval with chain-of-thought prompting [24].

PRISM was evaluated as a full pipeline and in six ablations that each remove one mechanism.

12.3. Metrics

Determination F1 compares each system's determination (approve, deny, or pend) with the ground-truth label over the test set; precision and recall are reported alongside it.

Relative improvement over baseline b is $F1_{PRISM} - F1_b$. For example, $0.912 - 0.710 = 0.202$ 28.5% against BM25 RAG and $0.912 - 0.769 = 0.143$ 18.6% against Dense RAG + CoT.

Context brittleness is the rate at which a system's determination changes when routine context engineering operations are applied to the bundle:

$$B = \frac{1}{|T|} \sum_{x \in T} \frac{1}{|P_x|} \sum_{x' \in P_x} \mathbf{1}[ax' \neq ax] \quad (13)$$

where T is the test set and P_x is the set of variants of case x produced by systemic context engineering operations: pruning, which removes documents from the bundle, and compaction, which condenses document content to fit the context budget. Pruning removes only documents outside $D_{required}$, and compaction preserves every required evidence element, so the correct determination is unchanged in every variant and any change counts as brittleness. These are operations that any production pipeline applies to context, not attacks.

Brittleness and R_{ACH} in (10) therefore measure different things and are not complements. Brittleness is measured on pruning and compaction variants and is reported for every system. R_{ACH} is measured on the adversarial set $\mathcal{A}$ (severity inflation, temporal manipulation, policy keyword insertion, and source spoofing), which is designed to push the determination toward approval, and PRISM uses it per case for routing. Adversarial manipulation flips determinations far more often than pruning or compaction, which is why PRISM has a brittleness of 0.049 but an adversarial flip rate of $1 - 0.847 = 0.153$. Brittleness reduction is $B_b - B_{PRISM}$, which gives $0.189 - 0.049 = 0.140$ 74.1% against BM25 RAG and $0.158 - 0.049 = 0.109$ 69.0% against Dense RAG + CoT.

Gap-detection F1 compares the predicted missing set ΔD with the annotated gap set, micro-averaged over cases, at two levels: document type (for example, "imaging report") and specific evidence element (for example, "MRI showing herniation at the treated level").

13. Results

13.1. Primary Results

Table 2 reports F1, precision, recall, and brittleness on the test set.

Table 2: Comparative evaluation on the test set ($n = 720$)

System	F1	Prec.	Rec.	Brittle.
LLM-only	0.611	0.584	0.641	0.218
BM25 RAG	0.710	0.698	0.723	0.189
Dense RAG	0.741	0.729	0.753	0.171
Dense RAG + CoT	0.769	0.755	0.784	0.158
PRISM (full)	0.912	0.901	0.924	0.049

Full PRISM reaches F1 0.912, a relative improvement of 18.6% over Dense RAG + CoT (0.769), 28.5% over BM25 RAG (0.710), and 49.3% over LLM-only (0.611). Brittleness falls by 69.0% against Dense RAG + CoT and 74.1% against BM25 RAG. Figure 3 plots the comparison.

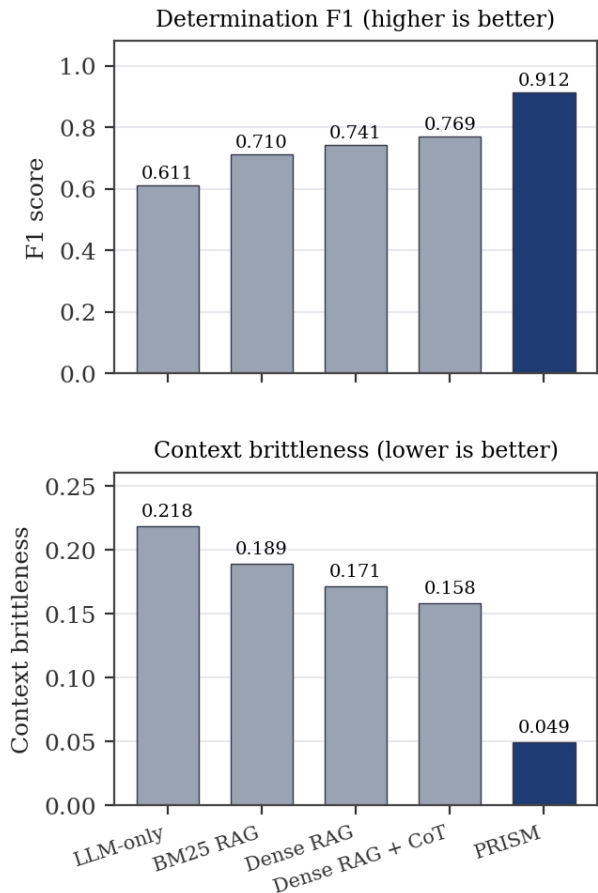


Figure 3: Determination F1 and context brittleness across baselines and full PRISM on the test set ($n = 720$).

13.2. Mechanism Ablation

Table 3 isolates each mechanism's contribution. Removing CES causes the largest loss in F1 (-0.081) and the largest rise in brittleness (0.033), so sufficiency gating is the most influential mechanism. UGE's contribution (-0.068 when removed) is concentrated in difficult cases: on the subset

that reviewers rated as non-obvious, removing UGE lowers F1 by 0.114. CMT has almost no effect on prediction, as expected, since its role is auditability.

Table 3: Ablation study: F1 and brittleness impact by module

Removed Module	F1	$\Delta F1$	Brittle.	$\Delta \text{Brittle.}$
None (full PRISM)	0.912	-	0.049	-
Remove CES	0.831	-0.081	0.082	0.033
Remove CCC	0.858	-0.054	0.061	0.012
Remove CSG	0.853	-0.059	0.068	0.019
Remove CMT	0.910	-0.002	0.050	0.001
Remove ACH	0.876	-0.036	0.072	0.023
Remove UGE	0.844	-0.068	0.063	0.014

13.3. Per-Category Performance

Full PRISM scores 0.908 on lumbar spinal fusion, 0.919 on knee arthroplasty, 0.924 on advanced imaging, and 0.897 on specialty pharmaceuticals, against 0.741, 0.782, 0.791, and 0.763 for Dense RAG + CoT. The largest gain is in lumbar spinal fusion (0.167), the most documentation-intensive category, where CSG and CCC recover multi-element chains that retrieval alone does not assemble.

13.4. Gap Retrieval and Escalation

CSG was activated in 274 of 720 test cases (38.1%). Of these, 185 (67.5%) were fully resolved, 39 (14.2%) were partially resolved, and 50 (18.2%) contained elements confirmed absent. Gap-detection F1 was 0.913 at the document-type level and 0.871 at the evidence-element level.

At τ_{HHP} 0.80, UGE escalated 106 test cases (14.7%). Reviewers judged 93 of these (87.7%) as warranting non-standard disposition, so 13 escalations (12.3%) were unnecessary. With the structured uncertainty report, review time per escalated case averaged 8.4 minutes, against 22.1 minutes for full re-adjudication, a 62% reduction.

14. Compliance and Governance Alignment

CES and CSG support the CMS requirements that adverse determinations rest on documented criteria and state specific reasons [42, 43]. The CES quality vector and the CCC graph supply source attribution for the decision rationale, which supports ONC HTI-1 transparency expectations [41]. CMT provides audit controls for ePHI workflows under HIPAA 45 CFR 164.312(b). ACH helps detect documentation patterns that are inconsistent with medical necessity, and UGE supplies the human-review triggers that keep clinicians responsible for adverse determinations.

15. Limitations and Future Work

Synthetic data. All 1,200 cases are synthetic. They do not reproduce the variability of real submissions, such as unstructured notes, scanned forms that require optical character recognition, and long attachments. The reported numbers should therefore be read as evidence that the

mechanisms work as designed, not as an estimate of production accuracy.

Shared schema. The case labels and the PRISM templates both derive from InterQual 2024, which may favor PRISM over baselines that do not see the criteria. This study does not report inter-annotator agreement or results on cases written without the templates. Independently annotated cases, labeled by reviewers blind to the PRISM templates and drawn from several payers' policies, are needed to confirm the gains.

Baselines. The baselines cover keyword, dense, and chain-of-thought retrieval. Comparisons with corrective RAG such as CRAG [15], graph retrieval such as GraphRAG [16], an agent that issues FHIR queries directly [21], and a baseline given the same policy checklist without the other PRISM mechanisms are needed to separate the value of the checklist from the rest of the pipeline.

Real-world validation. Evaluation on de-identified production cases, with determinations compared against final adjudicated outcomes including appeals, is ongoing and is the main next step.

Coverage and generality. CCC edge weighting depends on MeSH and SNOMED CT coverage, so novel procedures and rare diagnoses can yield low-confidence graphs. The policy graph is calibrated against a single payer's InterQual implementation, and per-payer policy graphs without refitting CES weights are planned. Claims adjudication, risk adjustment data validation (RADV) audits, utilization management, and appeals share the same evidentiary structure, and extending PRISM to them is future work.

16. Conclusion

This paper presented PRISM, a Dynamic Context Engineering framework that treats the context supplied to an LLM as a governed object in healthcare administrative AI. Its six mechanisms (CES, CCC, CSG, CMT, ACH, and UGE) check a context bundle against policy requirements, retrieve missing evidence, record every transformation, resist manipulation, and escalate when context is insufficient. On synthetic prior authorization cases, PRISM improved determination F1 by 18.6% over dense retrieval with chain-of-thought prompting and cut brittleness by 69.0%.

PRISM does not claim to solve healthcare administrative AI. It supports a narrower claim: that context sufficiency is a major source of failure, that it can be engineered systematically, and that doing so improves accuracy and stability without larger models. Confirming this claim on independently annotated and real-world data, across several payers' policies, is the priority for future work.

Conflict of Interest The author declares no conflict of interest.

References

- [1] American Medical Association, "2022 AMA Prior Authorization (PA) Physician Survey", Tech. rep., American Medical Association, 2023.
- [2] U.S. Department of Health and Human Services, Office of Inspector General, "Some medicare advantage organization denials of prior authorization requests raise concerns about beneficiary access to

- medically necessary care”, Tech. Rep. OEI-09-18-00260, U.S. Department of Health and Human Services, Office of Inspector General, 2022.
- [3] W. H. Shrank, T. L. Rogstad, N. Parekh, “Waste in the US Health Care System: Estimated Costs and Potential for Savings”, *JAMA*, vol. 322, no. 15, pp. 1501–1509, 2019, doi:[10.1001/jama.2019.13978](https://doi.org/10.1001/jama.2019.13978).
 - [4] M. Chandrasekaran, “AURELIUS: Agentic Uncertainty-Reasoning & Learning with Hierarchical Bayesian Multi-Agents Unified System for Handling Non-Determinism in HealthTech”, “2025 4th International Conference on Computing, Management and Telecommunications (ComManTel)”, pp. 52–57, IEEE, Madrid, Spain, 2025, doi:[10.1109/commantel68363.2025.11368546](https://doi.org/10.1109/commantel68363.2025.11368546).
 - [5] P. Anand, M. Chandrasekaran, V. Chandrasekaran, “System and method for artificial intelligence based field service assistance for telecommunications operations”, U.S. Patent 12,475,154 B1, 2025, issued November 18, 2025.
 - [6] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel, S. Riedel, D. Kiela, “Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks”, “Advances in Neural Information Processing Systems”, vol. 33, pp. 9459–9474, 2020.
 - [7] K. Singhal, S. Azizi, T. Tu, S. S. Mahdavi, J. Wei, H. W. Chung, N. Scales, A. Tanwani, H. Cole-Lewis, S. Pföhl, *et al.*, “Large language models encode clinical knowledge”, *Nature*, vol. 620, no. 7972, pp. 172–180, 2023, doi:[10.1038/s41586-023-06291-2](https://doi.org/10.1038/s41586-023-06291-2).
 - [8] J. Edin, A. Junge, J. D. Havtorn, L. Borgholt, M. Maistro, T. Ruotsalo, L. Maaløe, “Automated Medical Coding on MIMIC-III and MIMIC-IV: A Critical Review and Replicability Study”, “Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval”, pp. 2572–2582, ACM, 2023, doi:[10.1145/3539618.3591918](https://doi.org/10.1145/3539618.3591918).
 - [9] A. Ben Abacha, W.-w. Yim, G. Adams, N. Snider, M. Yetisgen, “Overview of the MEDIQA-chat 2023 shared tasks on the summarization & generation of doctor-patient conversations”, T. Naumann, A. Ben Abacha, S. Bethard, K. Roberts, A. Rumshisky, eds., “Proceedings of the 5th Clinical Natural Language Processing Workshop”, pp. 503–513, Association for Computational Linguistics, Toronto, Canada, 2023, doi:[10.18653/v1/2023.clinicalnlp-1.52](https://doi.org/10.18653/v1/2023.clinicalnlp-1.52).
 - [10] H. Zhang, J. Chen, F. Jiang, F. Yu, Z. Chen, J. Li, G. Chen, X. Wu, Z. Zhang, Q. Xiao, X. Wan, B. Wang, H. Li, “Hu-atuoGPT, towards Taming Language Model to Be a Doctor”, 2023, doi:[10.48550/arXiv.2305.15075](https://doi.org/10.48550/arXiv.2305.15075), preprint.
 - [11] K. Singhal, T. Tu, J. Gottweis, R. Sayres, E. Wulczyn, M. Amin, L. Hou, K. Clark, S. R. Pföhl, H. Cole-Lewis, *et al.*, “Toward expert-level medical question answering with large language models”, *Nature Medicine*, vol. 31, no. 3, pp. 943–950, 2025, doi:[10.1038/s41591-024-03423-7](https://doi.org/10.1038/s41591-024-03423-7).
 - [12] G. Xiong, Q. Jin, Z. Lu, A. Zhang, “Benchmarking retrieval-augmented generation for medicine”, L.-W. Ku, A. Martins, V. Srikumar, eds., “Findings of the Association for Computational Linguistics: ACL 2024”, pp. 6233–6251, Association for Computational Linguistics, Bangkok, Thailand, 2024, doi:[10.18653/v1/2024.findings-acl.372](https://doi.org/10.18653/v1/2024.findings-acl.372).
 - [13] C. Zakka, R. Shad, A. Chaurasia, A. R. Dalal, J. L. Kim, M. Moor, R. Fong, C. Phillips, K. Alexander, E. Ashley, J. Boyd, K. Boyd, K. Hirsch, C. Langlotz, R. Lee, J. Melia, J. Nelson, K. Sallam, S. Tullis, M. A. Vogel song, J. P. Cunningham, W. Hiesinger, “Almanac : Retrieval-Augmented Language Models for Clinical Medicine”, *NEJM AI*, vol. 1, no. 2, 2024, doi:[10.1056/aioa2300068](https://doi.org/10.1056/aioa2300068).
 - [14] A. Asai, Z. Wu, Y. Wang, A. Sil, H. Hajishirzi, “Self-RAG: Learning to Retrieve, Generate, and Critique through Self-Reflection”, “The Twelfth International Conference on Learning Representations”, 2024.
 - [15] S.-Q. Yan, J.-C. Gu, Y. Zhu, Z.-H. Ling, “Corrective Retrieval Augmented Generation”, 2024, doi:[10.48550/arXiv.2401.15884](https://doi.org/10.48550/arXiv.2401.15884), preprint.
 - [16] D. Edge, H. Trinh, N. Cheng, J. Bradley, A. Chao, A. Mody, S. Truitt, D. Metropolitan sky, R. O. Ness, J. Larson, “From Local to Global: A Graph RAG Approach to Query-Focused Summarization”, 2024, doi:[10.48550/arXiv.2404.16130](https://doi.org/10.48550/arXiv.2404.16130), preprint.
 - [17] J. Wu, J. Zhu, Y. Qi, J. Chen, M. Xu, F. Menolascina, V. Grau, “Medical Graph RAG: Towards Safe Medical Large Language Model via Graph Retrieval-Augmented Generation”, 2024, doi:[10.48550/arXiv.2408.04187](https://doi.org/10.48550/arXiv.2408.04187), preprint.
 - [18] A. Singh, A. Ehtesham, S. Kumar, T. T. Khoei, “Agentic Retrieval-Augmented Generation: A Survey on Agentic RAG”, 2025, doi:[10.48550/arXiv.2501.09136](https://doi.org/10.48550/arXiv.2501.09136), preprint; version 1.
 - [19] X. Tang, A. Zou, Z. Zhang, Z. Li, Y. Zhao, X. Zhang, A. Cohan, M. Gerstein, “MedAgents: Large language models as collaborators for zero-shot medical reasoning”, L.-W. Ku, A. Martins, V. Srikumar, eds., “Findings of the Association for Computational Linguistics: ACL 2024”, pp. 599–621, Association for Computational Linguistics, Bangkok, Thailand, 2024, doi:[10.18653/v1/2024.findings-acl.33](https://doi.org/10.18653/v1/2024.findings-acl.33).
 - [20] Y. Kim, C. Park, H. Jeong, Y. S. Chan, X. Xu, D. McDuff, H. Lee, M. Ghassemi, C. Breazeal, H. W. Park, “MDAgents: An Adaptive Collaboration of LLMs for Medical Decision-Making”, “Advances in Neural Information Processing Systems”, vol. 37, pp. 79410–79452, 2024, doi:[10.52202/079017-2522](https://doi.org/10.52202/079017-2522).
 - [21] Y. Jiang, K. C. Black, G. Geng, D. Park, J. Zou, A. Y. Ng, J. H. Chen, “MedAgentBench: A Realistic Virtual EHR Environment to Benchmark Medical LLM Agents”, 2025, doi:[10.48550/arXiv.2501.14654](https://doi.org/10.48550/arXiv.2501.14654), preprint.
 - [22] L. Mei, J. Yao, Y. Ge, Y. Wang, B. Bi, Y. Cai, J. Liu, M. Li, Z.-Z. Li, D. Zhang, C. Zhou, J. Mao, T. Xia, J. Guo, S. Liu, “A Survey of Context Engineering for Large Language Models”, 2025, doi:[10.48550/arXiv.2507.13334](https://doi.org/10.48550/arXiv.2507.13334), preprint.
 - [23] N. F. Liu, K. Lin, J. Hewitt, A. Paranjape, M. Bevilacqua, F. Petroni, P. Liang, “Lost in the Middle: How Language Models Use Long Contexts”, *Transactions of the Association for Computational Linguistics*, vol. 12, pp. 157–173, 2024, doi:[10.1162/tacl_a_00638](https://doi.org/10.1162/tacl_a_00638).
 - [24] J. Wei, X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, E. Chi, Q. V. Le, D. Zhou, “Chain-of-Thought Prompting Elicits Reasoning in Large Language Models”, “Advances in Neural Information Processing Systems”, vol. 35, pp. 24824–24837, 2022, doi:[10.52202/068431-1800](https://doi.org/10.52202/068431-1800).
 - [25] X. Wang, J. Wei, D. Schuurmans, Q. Le, E. Chi, S. Narang, A. Chowdhery, D. Zhou, “Self-Consistency Improves Chain of Thought Reasoning in Language Models”, “The Eleventh International Conference on Learning Representations”, 2023.
 - [26] D. Zhou, N. Schärli, L. Hou, J. Wei, N. Scales, X. Wang, D. Schuurmans, C. Cui, O. Bousquet, Q. Le, E. Chi, “Least-to-Most Prompting Enables Complex Reasoning in Large Language Models”, “The Eleventh International Conference on Learning Representations”, 2023.
 - [27] Y. Bai, S. Kadavath, S. Kundu, A. Askell, J. Kernion, A. Jones, A. Chen, A. Goldie, A. Mirhoseini, C. McKinnon, *et al.*, “Constitutional AI: Harmlessness from AI Feedback”, 2022, doi:[10.48550/arXiv.2212.08073](https://doi.org/10.48550/arXiv.2212.08073), preprint.
 - [28] T. Shin, Y. Razeghi, R. L. Logan IV, E. Wallace, S. Singh, “AutoPrompt: Eliciting Knowledge from Language Models with Automatically Generated Prompts”, B. Webber, T. Cohn, Y. He, Y. Liu, eds., “Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)”, pp. 4222–4235, Association for Computational Linguistics, Online, 2020, doi:[10.18653/v1/2020.emnlp-main.346](https://doi.org/10.18653/v1/2020.emnlp-main.346).
 - [29] O. Khattab, A. Singhvi, P. Maheshwari, Z. Zhang, K. Santhanam, S. Vardhamanan, S. Haq, A. Sharma, T. T. Joshi, H. Moazam, H. Miller, M. Zaharia, C. Potts, “DSPy: Compiling Declarative Language Model Calls into Self-Improving Pipelines”, 2023, doi:[10.48550/arXiv.2310.03714](https://doi.org/10.48550/arXiv.2310.03714), preprint.
 - [30] S. Yao, J. Zhao, D. Yu, N. Du, I. Shafran, K. Narasimhan, Y. Cao, “ReAct: Synergizing Reasoning and Acting in Language Models”, “The Eleventh International Conference on Learning Representations”, 2023.
 - [31] B. Kompa, J. Snoek, A. L. Beam, “Second opinion needed: communicating uncertainty in medical machine learning”, *npj Digital Medicine*, vol. 4, p. 4, 2021, doi:[10.1038/s41746-020-00367-3](https://doi.org/10.1038/s41746-020-00367-3), article 4.

- [32] C. Leibig, M. Brehmer, S. Bunk, D. Byng, K. Pinker, L. Umutlu, "Combining the strengths of radiologists and AI for breast cancer screening: a retrospective analysis", *The Lancet Digital Health*, vol. 4, no. 7, pp. e507–e519, 2022, doi:[10.1016/s2589-7500\(22\)00070-x](https://doi.org/10.1016/s2589-7500(22)00070-x).
- [33] S. Farquhar, J. Kossen, L. Kuhn, Y. Gal, "Detecting hallucinations in large language models using semantic entropy", *Nature*, vol. 630, no. 8017, pp. 625–630, 2024, doi:[10.1038/s41586-024-07421-0](https://doi.org/10.1038/s41586-024-07421-0).
- [34] M. Xiong, Z. Hu, X. Lu, Y. Li, J. Fu, J. He, B. Hooi, "Can LLMs Express Their Uncertainty? An Empirical Evaluation of Confidence Elicitation in LLMs", "The Twelfth International Conference on Learning Representations", 2024.
- [35] J. Pearl, *Causality: Models, Reasoning, and Inference*, Cambridge University Press, 2 ed., 2009, doi:[10.1017/CBO9780511803161](https://doi.org/10.1017/CBO9780511803161).
- [36] M. A. Hernán, J. M. Robins, *Causal Inference: What If*, Chapman & Hall/CRC, Boca Raton, FL, 2020.
- [37] E. Kıcıman, R. Ness, A. Sharma, C. Tan, "Causal Reasoning and Large Language Models: Opening a New Frontier for Causality", 2023, doi:[10.48550/arXiv.2305.00050](https://doi.org/10.48550/arXiv.2305.00050), preprint.
- [38] D. Pruthi, B. Dhingra, Z. C. Lipton, "Combating adversarial misspellings with robust word recognition", A. Korhonen, D. Traum, L. Márquez, eds., "Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics", pp. 5582–5591, Association for Computational Linguistics, Florence, Italy, 2019, doi:[10.18653/v1/P19-1561](https://doi.org/10.18653/v1/P19-1561).
- [39] M. Alzantot, Y. Sharma, A. Elgohary, B.-J. Ho, M. Srivastava, K.-W. Chang, "Generating natural language adversarial examples", E. Riloff, D. Chiang, J. Hockenmaier, J. Tsujii, eds., "Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing", pp. 2890–2896, Association for Computational Linguistics, Brussels, Belgium, 2018, doi:[10.18653/v1/D18-1316](https://doi.org/10.18653/v1/D18-1316).
- [40] E. Wallace, S. Feng, N. Kandpal, M. Gardner, S. Singh, "Universal adversarial triggers for attacking and analyzing NLP", K. Inui, J. Jiang, V. Ng, X. Wan, eds., "Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)", pp. 2153–2162, Association for Computational Linguistics, Hong Kong, China, 2019, doi:[10.18653/v1/D19-1221](https://doi.org/10.18653/v1/D19-1221).
- [41] Office of the National Coordinator for Health Information Technology, "Health Data, Technology, and Interoperability: Certification Program Updates, Algorithm Transparency, and Information Sharing", *Federal Register*, vol. 89, no. 6, pp. 1192–1438, 2024, final rule (HTI-1); published January 9, 2024; FR Doc. 2023-28857.
- [42] Centers for Medicare & Medicaid Services, "Medicare Program; Contract Year 2024 Policy and Technical Changes to the Medicare Advantage Program, Medicare Prescription Drug Benefit Program, Medicare Cost Plan Program, and Programs of All-Inclusive Care for the Elderly", *Federal Register*, vol. 88, no. 70, pp. 22120–22345, 2023, final rule (CMS-4201-F); published April 12, 2023; FR Doc. 2023-07115.
- [43] Centers for Medicare & Medicaid Services, "Medicare and Medicaid Programs; Patient Protection and Affordable Care Act; Advancing Interoperability and Improving Prior Authorization Processes for Medicare Advantage Organizations, Medicaid Managed Care Plans, State Medicaid Agencies, Children's Health Insurance Program (CHIP) Agencies and CHIP Managed Care Entities, Issuers of Qualified Health Plans on the Federally-Facilitated Exchanges, Merit-Based Incentive Payment System (MIPS) Eligible Clinicians, and Eligible Hospitals and Critical Access Hospitals in the Medicare Promoting Interoperability Program", *Federal Register*, vol. 89, no. 27, pp. 8758–8988, 2024, final rule (CMS-0057-F); published February 8, 2024; FR Doc. 2024-00895.
- [44] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, J. Kang, "BioBERT: a pre-trained biomedical language representation model for biomedical text mining", *Bioinformatics*, vol. 36, no. 4, pp. 1234–1240, 2020, doi:[10.1093/bioinformatics/btz682](https://doi.org/10.1093/bioinformatics/btz682).

Copyright: This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY-SA) license (<https://creativecommons.org/licenses/by-sa/4.0/>).



MUTHUKUMARAPANDIAN CHANDRASEKARAN heads a business AI practice at LTM Limited covering healthcare, life sciences, and travel and hospitality, where he developed the dynamic context packaging framework. He has more than 21 years of experience in AI and data architecture across healthcare, financial services, and enterprise technology.

He was previously Chief AI Architect and Vice President at CitiusTech, where he led Knewron, an agentic AI platform for healthcare, and created the AURELIUS framework. He has also held roles at Ascendion, Cognizant, and USAA. He holds U.S. Patent 12,475,154 B1, has served as a technical program committee reviewer for several IEEE conferences, and was named Professional of the Year 2024 in Data and Analytics/AI at OnConferences.