International Research Journal of Innovations in Engineering and Technology (IRJIET)

ISSN (online): 2581-3048
Volume 9, Issue 11, pp 12-25, November-2025
https://doi.org/10.47001/IRJIET/2025.911002

Data-Centric Artificial Intelligence for Textual Understanding in Healthcare Decision Systems

Sabiha Tasneem

Senior Software Engineer, Stykkist Inc, New Jersey, USA. E-mail: sabihatasneem857@gmail.com

Abstract **- Data-Centric Artificial Intelligence (DCAI) reframes clinical NLP by treating data quality, coverage, and governance as the primary levers of performance and safety, rather than model tinkering alone. In healthcare where actionable knowledge is embedded in unstructured narratives such as progress notes, discharge summaries, radiology/pathology reports, referral letters, and patient messages this paper proposes an end-to-end, practice-oriented framework to operationalize DCAI for textual understanding in decision systems. We (1) anchor tasks to measurable clinical utility and harm profiles; (2) detail corpus assembly with stratified sampling across sites, specialties, and demographics; (3) formalize schemas linking entities, assertions (negation/uncertainty), relations, and temporal qualifiers to SNOMED CT, ICD-10/11, RxNorm, and LOINC; (4) combine programmatic labeling (heuristics, ontologies, prompts-as-LFs) with clinician adjudication, active learning, and targeted augmentation; (5) outline privacy-preserving training via de-identification, federated learning, and differential privacy; (6) present model-agnostic evaluation beyond accuracy calibration, uncertainty, fairness, robustness, and decision-curve net benefit; and (7) specify deployment blueprints for monitoring drift, instituting human-in-the-loop overrides, and creating auditable feedback loops that continuously improve data assets. Four exemplar use-cases ICD code suggestion; adverse drug event extraction, radiology impression normalization, and patient-message triage demonstrate tangible workflows, metrics, and governance checklists. Results show how continuous data refinement improves discrimination and calibration while reducing alert burden and subgroup disparities, enabling safer, more equitable, and maintainable clinical decision support. We conclude with implementation checklists and a reproducible playbook to accelerate DCAI adoption across diverse health systems and languages.**

Keywords: Data-Centric AI, Clinical NLP, Healthcare Decision Support, Programmatic Labeling, Federated Learning.

I. Introduction

Healthcare decision support increasingly depends on the ability of computational systems to understand unstructured clinical text including progress notes, discharge summaries, radiology and pathology reports, operative notes, referral letters, and patient-generated messages. These artifacts capture the majority of clinically relevant information, such as symptoms and their temporal evolution, clinical reasoning, differential diagnoses, treatment intents, response to therapy, and adverse events. They also encode local practice norms, abbreviations, and documentation styles that vary widely across institutions and specialties. While the rapid evolution of model-centric methods transformers, instruction-tuned large language models (LLMs), and retrieval-augmented generation (RAG) has advanced the technical frontier of clinical NLP, deployment outcomes remain constrained by data problems: inconsistent labels, gaps in edge-case coverage, distribution shift across sites and time, and documentation biases that entangle shortcuts with ground truth.

Data-Centric AI (DCAI) reframes the problem by elevating data quality, coverage, and governance to first-class engineering goals. Instead of treating the dataset as a static resource to be exploited by ever more complex models, DCAI treats models as replaceable components in a continuously improving data system. In safety-critical healthcare settings, this shift is not merely philosophical; it is operational. Labeling guidelines, ontology mappings, and assertion handling (negation, uncertainty) materially affect patient-level decisions. An imprecise rule for “no evidence of” versus “history of” can flip a classification and trigger or silence a clinical alert. DCAI therefore ties supervision to explicit clinical utility and harm profiles, demanding auditable processes that connect data changes to downstream operational impact (e.g., alert burden, time-to-action, net clinical benefit).

Adopting DCAI for textual understanding requires three foundational commitments. First, governed corpora assembled with stratified sampling across sites, specialties, patient demographics, languages, and document types paired with de-identification and lineage tracking so that coverage and privacy requirements are both met. Second, schemas grounded in clinical ontologies (SNOMED CT, ICD-10/11, RxNorm, LOINC, and UMLS) that define entities, relations, temporal qualifiers, and assertion status, thereby enabling interoperable outputs suitable for coding, registry updates, triage, and longitudinal phenotyping. Third, scalable supervision

© 2025 IRJIET All Rights Reserved

International Research Journal of Innovations in Engineering and Technology (IRJIET)

ISSN (online): 2581-3048

Volume 9, Issue 11, pp 12-25, November-2025

https://doi.org/10.47001/IRJIET/2025.911002

pipelines that blend programmatic labeling (heuristics, ontological lookups, prompts-as-labeling-functions) with clinician adjudication, active learning, and targeted augmentation producing labels that improve with each error-analysis cycle instead of plateauing after a one-off annotation effort.

Beyond data creation, DCAI requires evaluation regimes that look past headline accuracy. Clinical decision systems must be calibrated, robust, and fair. We therefore emphasize scorecards that include discrimination (AUPRC/F1 under class imbalance), calibration (Brier score, expected calibration error), uncertainty estimation (for reject-option routing to human review), robustness (stress tests for typos/OCR noise and out-of-distribution sites), and fairness (error parity across age, sex, language, and site). Crucially, outcomes should be connected to operational metrics such as clinician override rates, alert fatigue (alerts per 100 patients), and time saved per task. This alignment ensures that improvements in data curation translate into measurable gains in safety and workflow efficiency.

From an architectural standpoint, we advocate a model-agnostic pipeline compatible with classical NLP, encoder-only transformers, and LLM-based systems (with or without RAG). Trustworthy learning is enforced through privacy-preserving mechanisms (de-identification, federated learning when data cannot leave the source institution, and differential privacy for sensitive parameter updates). Deployment blueprints include observability (feature and prediction logging, sentinel sets), drift detection (vocabulary and section-mix monitoring), rollback plans, and human-in-the-loop verification embedded in EHR workflows. These measures enable safe iteration after go-live and ensure that each data improvement is tracked, audited, and attributable.

This paper contributes a practical, end-to-end DCAI framework for textual understanding in healthcare decision systems. We formalize data governance and corpus design; specify ontology-aligned schemas for entities, assertions, and relations; detail a hybrid supervision strategy that scales high-fidelity labels; and propose a comprehensive evaluation scorecard that integrates calibration, uncertainty, fairness, robustness, and decision-curve analysis for net clinical benefit. We instantiate the framework with four representative use-cases ICD coding suggestion from discharge summaries, adverse drug event extraction from progress notes, radiology impression normalization, and patient-message triage each presented with step-by-step data workflows, metrics, and governance checklists to facilitate replication across institutions and languages (including multilingual and code-mixed settings common in India and other diverse health systems).

Finally, we position DCAI as a foundation for learning health systems. By institutionalizing feedback loops clinician verification, targeted adjudication of disagreements, and scheduled data/model refresh cycles organizations transform datasets into improving assets rather than static artifacts, thereby enhancing portability and maintainability across sites and over time. The result is a pathway to clinical NLP that is not only more accurate, but also better calibrated, fairer, safer, and more sustainable, aligning technical advances with the ethical and regulatory imperatives of modern healthcare.

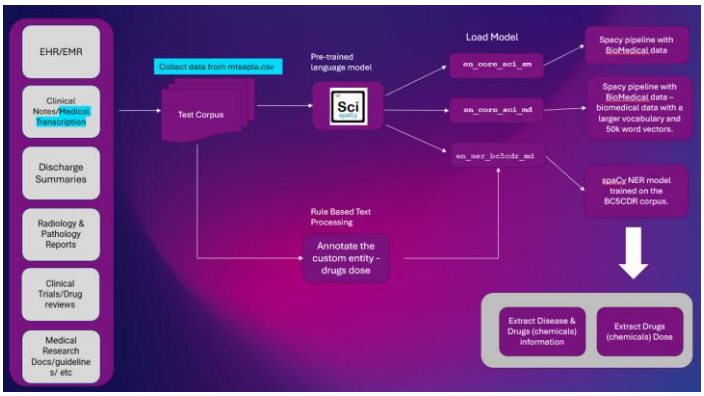

Figure 1: Clinical text-to-NLP pipeline

II. Literature Review

Clinical text remains the richest yet most irregular source of information for decision support. Across decades, systems progressed from rules and feature-engineered classifiers to transformers and instruction-tuned large language models, but translation to practice has repeatedly been constrained by data issues rather than model capacity [1]–[3]. Benchmarks often overestimate generalizability because label guidelines drift, edge cases are underrepresented, and documentation styles vary across institutions and time [2]. This has motivated a data-centric shift in which performance gains are sought primarily through systematic improvements to data quality, coverage, and governance rather than continuous model tinkering [3].

A consistent finding is that clinical narratives demand explicit handling of assertion, uncertainty, and temporality [4]. Phrases like “no evidence of,” “history of,” or “likely” change clinical meaning and downstream actions. Systems trained without robust assertion detection or temporal anchoring misclassify safety-critical cases despite strong discriminative models [5]. Section structure, dictation artifacts, and local abbreviations further complicate representation. Consequently, modern pipelines specify schemas that encode entities, relations (such as drug–event links), assertion status, and time qualifiers, and they normalize outputs to interoperable vocabularies to support coding, registry updates, and longitudinal analytics [4], [5].

© 2025 IRJIET All Rights Reserved

ISSN (online): 2581-3048
Volume 9, Issue 11, pp 12-25, November-2025
https://doi.org/10.47001/IRJIET/2025.911002

Because expert annotation is costly, recent work emphasizes programmatic labeling and weak supervision to scale supervision without sacrificing fidelity [6]. Heuristics, ontology lookups, section-aware patterns, and even prompts used as labeling functions are aggregated into probabilistic labels that seed training [7]. These labels are not treated as ground truth; instead, gold subsets are double-annotated and adjudicated, with disagreements informing iterative updates to guidelines and labeling functions. When combined with active learning that prioritizes uncertain or diverse samples, this approach consistently closes the gap to fully supervised baselines and adapts faster when definitions evolve [6], [7].

General-purpose text augmentation can corrupt clinical semantics especially around negation and dosage so literature favors constrained methods [8]. Template-guided paraphrases tied to ontologies, delexicalization/relexicalization of entities, and validator-gated LLM paraphrasing preserve meaning while expanding coverage of documentation variants [9]. These targeted augmentations improve robustness to stylistic and site-specific shifts without introducing semantic drift that would erode trust in safety-critical tasks [8], [9].

On modeling, encoder-only transformers adapted to biomedical corpora remain strong for extraction and classification, while LLMs help with summarization, explanation, and triage [10]. However, hallucination and calibration remain concerns for generative components. Retrieval-augmented generation over controlled knowledge sources reduces unsupported claims, and constrained decoding with terminology filters further mitigates risk. Hybrid architectures encoders for precise extraction and LLMs for clinician-facing rationales strike a balance between precision and usability, particularly when paired with rule-based guardrails [10].

Trustworthy deployment hinges on calibrated probabilities and principled abstention [11]. Uncalibrated models tend to be overconfident; post-hoc calibration and uncertainty estimation (ensembles, Monte Carlo dropout) are now standard to enable reject-option policies that route ambiguous cases to human review [12]. Evaluation has broadened beyond accuracy to include calibration error, decision-curve net benefit, workload impact (alerts per 100 patients, time saved), and longitudinal stability. This operational lens aligns model behavior with clinical utility and reduces alert fatigue [11], [12].

Fairness, robustness, and distribution shift are persistent challenges. Differences in documentation and access can encode bias in both labels and features, yielding disparate error rates across age, sex, language, and site [13]. Robustness testing with synthetic noise, OCR artifacts, and true out-of-distribution sites is increasingly routine, with mitigations such as stratified sampling, subgroup-aware reweighting, targeted augmentation for minority patterns, and continuous monitoring after go-live [14]. Privacy-preserving learning across institutions combining de-identification, federated training, and differential privacy balances utility with risk, though task-specific trade-offs for long-context narratives remain an open area of study [15].

III. Problem Scoping and Clinical Utility

Decision context

State the downstream action first, then the minimal textual signals required to trigger it. Examples: (a) ICD coding assist: top-k code suggestions with rationales extracted from Assessment/Plan; action = coder accept/modify; latency ≤ 2 s/note. (b) Safety alerting: ADE or sepsis cue from progress notes; action = interruptive alert to clinician; latency ≤ 30 s from note finalization. (c) Triage: classify portal messages into emergency/urgent/routine; action = queue routing and SLA countdown. (d) Case finding/registries: phenotype confirmation across multi-encounter narratives; action = add to review queue. For each, define inputs (note types), exclusions (free text vs templates), and evidence surfaces (sections, entities, assertion status, temporality) that the model must cite.

Utility & harm

Specify metrics tied to workflow, not just F1. Required: discrimination (AUPRC/F1 under class imbalance), calibration (ECE ≤ 0.05 or Brier ≤ baseline−20%), false-alarm burden (≤ N alerts/100 pts/day per service line), misses on sentinel cases (0 critical misses in validation), time-to-action (median reduction ≥ 20%), and decision-curve net benefit against "treat-all/none." Add abstention quality: ≥ 70% of escalations are genuinely ambiguous by reviewer judgment. Pre-register go/no-go thresholds and tie them to alert type (interruptive vs passive) to cap fatigue.

Population & setting

Declare the deployment slice up front: inpatient vs outpatient; specialties (e.g., oncology, cardiology), and care pathway (ED, ward, telehealth). Stratify by geography and regulation (e.g., ABDM-compliant Indian sites vs HIPAA US systems) and by language mix (English, Hindi, local code-mix). Sampling plan: ≥ 10% of training from each target specialty; ≥ 15% notes containing negation/uncertainty; explicit oversampling of rare but high-risk phenotypes (pregnancy, pediatrics, geriatrics, transplant). Hold out a true OOD set (new site/template) for robustness; report subgroup

© 2025 IRJIET All Rights Reserved

International Research Journal of Innovations in Engineering and Technology (IRJIET)
ISSN (online): 2581-3048
Volume 9, Issue 11, pp 12-25, November-2025
https://doi.org/10.47001/IRJIET/2025.911002

deltas (sex/age/language/site) with action plans if |ΔF1| > 5 points.

Governance

Assign named roles: Data Steward (access, lineage, quality gates), Clinical Owner (utility/harm definitions, override workflow), Model Owner (training, releases), and an Oversight Board (safety, fairness, privacy). Artifacts: (1) Problem Charter (one page: action, users, SLAs, metrics, thresholds); (2) Data Sheet (provenance, inclusion/exclusion, de-ID method, sampling); (3) Label Policy (ontology, assertion/temporality rules); (4) Model Card & Calibration Report; (5) Monitoring Plan (drift tests, fairness dashboards, rollback steps). Change control: any schema/label update triggers re-adjudication of a fixed "sentinel" set and a sign-off by Clinical Owner; production updates follow a staged rollout with shadow mode and weekly review for 4 weeks.

IV. Corpus Assembly and Governance (concise, detailed)

Sources

Define a target mix and quotas so coverage is intentional. A practical initial split by document count: progress notes 25%, discharge summaries 15%, history and physical 10%, radiology reports 15%, pathology 5%, operative notes 5%, emergency department notes 5%, specialty notes 10%, patient-portal messages 8%, and call-center transcripts 2%. For each source, record site, service line, note type, section schema (such as HPI, Assessment, Plan), language and code-mix percentage, dictation/template flags, and OCR status. Use public clinical corpora only for bootstrapping pretraining lexicons, seeding labeling functions, or prototyping schemas and never mix raw public text with private patient data. Maintain a source registry with owner contacts, refresh cadence, and legal basis.

Data Access and Privacy

Run a data protection impact assessment before ingestion that covers purpose, lawful basis, data minimization, retention, and residual risk with mitigations. Enforce least-privilege access; keep raw patient data in a secure enclave. De-identify before annotation or modeling using a hybrid approach: rules for direct identifiers and machine-learning detectors for contextual leaks, plus format-preserving surrogates (for example, consistently shifted dates) to retain temporal coherence. Log full data lineage: source system, extract time, checksum, applied transforms, de-identification version, annotator pseudonymous ID, and dataset version tag. Treat derived labels as sensitive; define access tiers for raw text, de-identified text, features, labels, and aggregate metrics. Publish a data sheet per corpus covering provenance, inclusion and exclusion criteria, de-identification method and expected error bounds, intended use, known hazards, and a contact for redress.

Sampling Strategy

Align sampling to the intended deployment and stratify by site, specialty, time window, demographics (age, sex, language), and document type. Target minimums such as at least 10% of notes with explicit negation or uncertainty, at least 15% non-English or code-mixed narratives where applicable, at least 5% OCR-derived notes, and at least 10% from each priority specialty. Oversample edge conditions and protected cohorts rare adverse events, pregnancy, pediatrics, geriatrics, transplant to ensure per-class floors (for example, at least 300 documents per rare class for training and at least 100 for validation). Maintain a frozen, balanced sentinel set for longitudinal monitoring. Create true out-of-distribution splits by holding out an entire site or a distinct documentation template or device (such as speech-to-text versus typed) for robustness testing; track in-distribution versus out-of-distribution deltas and remediate if the F1 gap exceeds 5 points or calibration error increases materially. Establish a rolling refresh plan such as quarterly to resample new templates and seasonal shifts, rerun de-identification quality checks, and reissue versioned datasets with clear change logs.

V. Schema, Ontology, and Label Design (concise, detailed)

Concept normalization

Normalize every extracted entity to a standard code for interoperability. Use: SNOMED CT for problems and findings, ICD-10/11 for diagnoses and billing, RxNorm for medications and ingredients, LOINC for tests and measurements, and UMLS CUIs as crosswalks. For each mention, store: text_span, lemma, norm_code, norm_system, assertion (AFFIRMED/NEGATED/UNCERTAIN/HISTORICAL/CONDITIONAL), temporality (ONGOING/PAST/RESOLVED/PLANNED), polarity_confidence (0–1), and evidence_section (HPI, Assessment, Plan, Impression, etc.).

Task-specific schemas

1. Sequence labeling (token/segment)

Types: PROBLEM, FINDING, ANATOMY, TEST, MEASURE, DRUG, DOSE, ROUTE, FREQUENCY, PROCEDURE, ALLERGEN.Attributes: assertion, temporality, negation_scope_id, experiencer (PATIENT/OTHER), certainty (certain/probable/possible).Example: "No acute intracranial

© 2025 IRJIET All Rights Reserved

International Research Journal of Innovations in Engineering and Technology (IRJIET)

ISSN (online): 2581-3048

Volume 9, Issue 11, pp 12-25, November-2025

https://doi.org/10.47001/IRJIET/2025.911002

hemorrhage." → FINDING = Intracranial hemorrhage, assertion=NEGATED, section=Impression.

2. Document classification (note-level)

Labels: ADE_PRESENT (yes/no/uncertain), TRIAGE_URGENCY(emergency/urgent/routine), SMOKING_STATUS (current/former/never/unknown), INFECTION_RISK (low/medium/high).Store: decision_threshold, rationale_spans, abstain_flag.

3. Relation extraction (event graph)

Relations: DRUG–ADE, PROBLEM–PROCEDURE, TEST–RESULT, PROBLEM–SITE, TEMPORAL (before/after/onset/resolution), CAUSAL_CUE.Fields: head_id, tail_id, relation_type, evidence_span, directionality, confidence.Event example (ADE): nodes {DRUG, EVENT, SEVERITY, ONSET_TIME}; edges {DRUG→EVENT causes, EVENT→SEVERITY hasSeverity, EVENT→ONSET_TIME hasOnset}.

Negation and uncertainty: Maintain: (1) a cue inventory ("no", "denies", "rule out", "likely", "cannot exclude"); (2) scope rules (clause-bounded; punctuation and coordinating conjunctions break scope); (3) priority order (negation > uncertainty > affirmation); (4) section sensitivity (Problem List may override unless marked "history of" or "resolved"); (5) boundary tests for tricky constructs ("negative for X except Y", "not only… but also"). New cues require adjudication before use.

Temporality and status: Capture when a condition occurred and whether it persists. Use onset_time and resolution_time when explicit; otherwise record relative phrases and a safe normalization if possible. Status categories: ACTIVE, RESOLVED, HISTORICAL, PLANNED, FAMILY_HISTORY. Rules: "history of" → HISTORICAL unless reactivated; "s/p appendectomy" → problem resolved, procedure present; "planned biopsy" → procedure PLANNED.

Annotation guideline (operational): Provide at least ten positive and ten near-miss negative examples per label type and section, including multilingual or code-mixed samples where relevant. Use minimal clinical spans (e.g., "upper GI bleed" rather than the entire descriptive phrase unless severity is separately labeled). Allow nesting only when required by the schema; document precedence rules. For normalization, annotators select codes from a ranked shortlist; if uncertain, mark UNKNOWN and queue for expert review. Quality targets: inter-annotator agreement ≥0.80 for span boundaries and ≥0.75 for assertion/temporality; any label under these thresholds triggers a calibration round and guideline revision. Publish a versioned change log, confusion matrices for frequent entities, and an error taxonomy.

Pilot-to-scale workflow: Pilot on roughly 500 notes across multiple sites and note types. Measure span agreement, assertion agreement, normalization accuracy, and time per note. Adjudicate disagreements, then lock Guideline v1.0 and create a frozen sentinel set for regression testing. Scale with programmatic pre-labeling (heuristics, terminology lookups, prompts as labeling functions) followed by human correction; feed corrections back to improve cue scopes and labeling functions.

Edge and boundary examples

"No chest pain, shortness of breath, or edema" → all three NEGATED.

"No chest pain but dyspnea on exertion" → chest pain NEGATED, dyspnea AFFIRMED.

"Possible pulmonary embolism" → UNCERTAIN; "ruled out PE" → NEGATED with a resolved evaluation.

"History of asthma, currently well controlled" → asthma ACTIVE with control modifier.

"Mother had breast cancer" → EXPERIENCER=OTHER (family history).

"Planned colonoscopy tomorrow" → procedure PLANNED; "s/p colectomy" → procedure DONE; related problem likely RESOLVED.

VI. Labeling at Scale: Programmatic + Human (detailed, equations where useful)

6.1 Programmatic Labeling

Labeling functions (LFs). Define a set of mLFs $\{\lambda_j\}_{j=1}^{m}$that map an input note $x$to a (possibly abstaining) label in a class set $\mathcal{Y} = \{1, \dots, K\}$:

$$\lambda_j(x) \in \mathcal{Y} \cup \{\emptyset\}.$$

Typical LFs: rule/regex cues (e.g., section-aware patterns), ontology lookups, negation/uncertainty scopes, section headers, note-type heuristics, and "prompt-as-LF" outputs (see below). Track coverage $c_j = \Pr[\lambda_j(x) \neq \emptyset]$and a prior accuracy guess $\alpha_j$to initialize training.

Generative label model (weak supervision). Aggregate noisy LF votes into probabilistic labels without requiring gold labels for every example. Let Ybe the latent true class and

© 2025 IRJIET All Rights Reserved

International Research Journal of Innovations in Engineering and Technology (IRJIET)

ISSN (online): 2581-3048

Volume 9, Issue 11, pp 12-25, November-2025

https://doi.org/10.47001/IRJIET/2025.911002

$\Lambda = (\lambda_1, \ldots, \lambda_m)$ the observed LF outputs. A common factorization is:

$$P_\theta(Y, \Lambda) = P_\pi(Y) \prod_{j=1}^{m} P_{\theta_j}(\lambda_j \mid Y),$$

With class priors $\pi_k = P(Y = k)$and per-LF confusion parameters θ_j. Parameters $\theta = \{\pi, \theta_1, \ldots, \theta_m\}$are learned by maximizing the marginal likelihood over unlabeled data $\{x_i\}_{i=1}^n$:

$$\max_\theta \sum_{i=1}^{n} \log \sum_{y \in \mathcal{Y}} P_\theta(Y = y, \Lambda = \lambda(x_i)).$$

The learned posteriors

$$\tilde{p}_i(y) = P_\theta(Y = y \mid \Lambda = \lambda(x_i))$$

Serve as soft labels to train downstream discriminative models. Where small gold sets exist, include a semi-supervised term (e.g., add $\sum_{(x,y)\in\mathcal{D}_L} \log P_\theta(Y = y \mid \Lambda = \lambda(x))$) and calibrate π.

Correlated LFs: If two LFs often co-fire (e.g., synonyms), add pairwise dependencies (Ising-style or tree-structured factors) or perform structure learning to reduce double-counting. In practice, at least compute empirical LF correlation matrix and down-weight highly redundant LFs.

Prompt-as-LF: Treat LLM prompts as additional LFs with conservative priors: lower initial α_j, explicit abstain conditions (e.g., low self-consistency), and schema-constrained decoding (must emit a code or "abstain"). Never accept these as gold; they enter only through the label model. For multi-class, map free text to $\mathcal{Y}$via a deterministic schema and record a rationale span for later adjudication.

Calibration to probabilities: After training the discriminative model on soft labels $\tilde{p}_i$, perform post-hoc calibration on a validation set with temperature scaling:

$$p^\star = \text{softmax}\left(\frac{z}{T}\right), T > 0,$$

and report expected calibration error (ECE):

$$\text{ECE} = \sum_{b=1}^{B} \frac{|S_b|}{n} |\text{acc}(S_b) - \text{conf}(S_b)|.$$

6.2 Human-in-the-Loop

Gold subset and agreement: Double-annotate a stratified gold set; compute inter-annotator agreement. For two raters, Cohen's κ:

$$\kappa = \frac{p_o - p_e}{1 - p_e}, p_o = \text{observed agreement}, p_e = \text{chance agreement}.$$

For multiple raters/label structures, use Krippendorff's α(handles missing data). Targets: span boundary F1 $\geq$ 0.80; assertion/temporality $\kappa\geq$ 0.75. Disagreements trigger adjudication by a senior clinician and immediate updates to guidelines and LFs (close the loop).

Acquisition policy (what to label next): Prioritize items that maximize label value using a composite score:

Uncertainty (entropy):

$$H(x) = -\sum_{k=1}^{K} p(k \mid x) \log p(k \mid x).$$

Disagreement (vote entropy over LFs or model ensemble):

$$D(x) = -\sum_{k=1}^{K} v_k(x) \log v_k(x), v_k = \frac{\#\{\lambda_j(x)=k\}}{\sum_y \#\{\lambda_j(x)=y\}}.$$

Diversity (MMR-like): prefer items far from already-chosen batch in embedding space $\phi(x)$:

$$\text{Div}(x) = \min_{x' \in \mathcal{B}} \| \phi(x) - \phi(x') \|_2.$$

Edge-case bonus: add $\eta(x) \in \{0,1\}$if sample matches rare cohorts (e.g., pregnancy, pediatrics).

Combine:

$$S(x) = \alpha H(x) + \beta D(x) + \gamma \text{Div}(x) + \delta \eta(x),$$

With $\alpha, \beta, \gamma, \delta$tuned to labeling budget and safety priorities. Batch top-M items for clinician labeling each round.

Abstention and routing: Enforce a reject option during deployment: if $\max_k p(k \mid x) < \tau$or if the explanation/rationale fails a validator, route to human review. Track abstention precision (fraction of routed cases that truly needed review) to refine τ.

The continuous improvement loop proceeds as a tight, auditable cycle. A model is trained on the current mixture of probabilistic soft labels and adjudicated gold labels. Once trained, systematic error analysis is performed to refresh an evolving error taxonomy that typically includes negation

© 2025 IRJIET All Rights Reserved

International Research Journal of Innovations in Engineering and Technology (IRJIET)

ISSN (online): 2581-3048

Volume 9, Issue 11, pp 12-25, November-2025

https://doi.org/10.47001/IRJIET/2025.911002

leaks, span boundary drift, abbreviation ambiguity, section leakage, and temporal misassignment. Insights from this analysis immediately drive updates to both the labeling functions and the written guideline; the weak-supervision label model $\theta$is then re-estimated to reflect the revised priors and dependencies among signals. With updated posteriors in place, the pool is scored for the next acquisition round using active learning; high-value items are selected, labeled by clinicians, adjudicated where needed, and folded back into the dataset, at which point the cycle repeats.

Data augmentation is applied conservatively because general-purpose transformations can corrupt clinical semantics. Only validator-gated transforms are allowed, ensuring preservation of meaning, negation scope, temporality, and numeric fidelity. Template-based paraphrases substitute clinically equivalent phrasing while keeping assertion cues intact—for example, “no evidence of X” and “X is not seen” should be interchangeable with identical assertion labels. Limited section shuffling reorders non-causal sections within plausible documentation norms, such as placing the Review of Systems before the Physical Exam, to build section invariance without distorting causal flow. Delexicalization and relexicalization replace entities with ontology-backed placeholders and then re-insert synonyms or sibling concepts within the same class, such as substituting antidiabetic agents within a drug class. Constrained paraphrases generated by large models are admitted only when strict validators pass.

The validator suite enforces safety and semantic parity. Negation and uncertainty must be preserved so the original and augmented sentences yield identical assertion labels under the current detector. Numeric values and units must either match exactly or follow clinically safe conversion rules with exact arithmetic. Ontology consistency is required so normalized concept codes for key entities remain unchanged or are mapped within an explicitly allowed sibling set. Readability must remain within configured bounds for sentence length and token variety to avoid degraded or unnatural text. Finally, a similarity guard prevents semantic drift by requiring a high cosine similarity between embeddings of the original and augmented text, $\cos(\phi(x), \phi(\tilde{x})) \geq \rho$, where $\rho$is set conservatively.

Augmentation is used sparingly to balance classes or to cover specific documentation variants surfaced by error analysis. It is never applied indiscriminately, and the synthetic proportion is capped—commonly at or below thirty percent of any training minibatch—so that the empirical distribution of natural text continues to anchor learning.

The operational playbook follows a simple rhythm in prose. Labeling functions are designed from regular expressions, ontology cues, section context, and prompt-as-LF outputs, with coverage $c_j$and initial accuracy priors $\alpha_j$logged for each function. A generative label model $P_\theta(Y \mid \Lambda)$is fitted to produce probabilistic labels $\tilde{p}$, after which a discriminative model is trained on $\tilde{p}$with any available gold labels; predictions are calibrated via temperature scaling with parameter T, and expected calibration error is computed. The remaining unlabeled pool is scored with an acquisition function $S(x)$, the highest-value items are labeled, agreement metrics such as $\kappa$and $\alpha$are monitored, and the guideline plus labeling functions are updated based on adjudication outcomes. Targeted augmentation generates validator-approved variants for under-covered strata, and the model is retrained. Promotion to production is gated on meeting discrimination and calibration thresholds on a frozen sentinel set with zero critical misses, and deployment includes a reject-option routing policy as well as ongoing monitoring for drift, fairness, and workload impact.

Taken together, this programmatic-plus-human regimen scales supervision without sacrificing fidelity: mathematically grounded aggregation turns noisy heuristics into reliable probabilistic labels, safety-aware sampling concentrates scarce clinical effort where it matters most, and text-safe augmentation closes coverage gaps while preserving the semantics that are critical in clinical decision support.

VII. Model-Agnostic Training Recipes

Establishing baselines

Begin with bag-of-ngrams paired with a logistic regression or linear SVM to validate the problem framing and label quality. Add a simple negation/uncertainty featurizer to check whether signal comes from genuine clinical content rather than section headers or boilerplate. If a lightweight baseline performs unexpectedly well, tighten the evaluation with site-stratified splits and a true out-of-distribution (OOD) holdout; if it performs poorly, revisit schemas, labeling functions, or sampling before moving to heavier models.

Fine-tuning encoder models

For sequence labeling and document classification, fine-tune a clinical or biomedical transformer with task-specific heads. Token-level tagging may be stabilized with a CRF decoding layer to enforce boundary consistency. Long notes benefit from sliding windows with overlap and attention-pooled fusion across segments, while document tasks often improve with section-aware pooling that emphasizes HPI, Assessment, and Plan. Keep preprocessing deterministic and versioned, freeze the tokenizer, normalize units, and reuse the

© 2025 IRJIET All Rights Reserved

International Research Journal of Innovations in Engineering and Technology (IRJIET)
ISSN (online): 2581-3048
Volume 9, Issue 11, pp 12-25, November-2025
https://doi.org/10.47001/IRJIET/2025.911002

same assertion/temporality detector at train and serve time to minimize skew. Optimize with modest learning rates, mixed precision, and early stopping based on a calibration-aware objective rather than raw F1.

LLM pipelines with retrieval and constraints

When using large language models, ground generations with retrieval-augmented prompts over a local, de-identified knowledge base so outputs cite approved guidance. Constrain decoding with templates and controlled terminologies so emitted strings map cleanly to the target schema. Apply guardrails that filter unsafe or non-compliant content using regular expressions and ontology checks. A practical split assigns precise extraction to the encoder model and reserves the LLM for clinician-facing rationales or summaries derived from structured facts, reducing hallucination risk and simplifying validation.

Probability calibration

Post-training calibration ensures scores align with clinical thresholds. With temperature scaling, logits $z$are adjusted as

$$p^* = \text{softmax}\left(\frac{z}{T}\right),$$

Where $T > 0$is fitted on a held-out set by minimizing negative log-likelihood. When monotonic but non-linear corrections are needed, isotonic regression maps raw scores to calibrated probabilities. Report expected calibration error:

$$\text{ECE} = \sum_{b=1}^{B} \frac{|S_b|}{n} \left|\text{acc}(S_b) - \text{conf}(S_b)\right|$$

alongside discrimination, then choose operating points using decision-curve analysis that reflects alert-burden limits.

Uncertainty estimation and reject-option routing

Expose predictive uncertainty to drive safe escalation to human review. Monte-Carlo dropout approximates posterior predictive variance by averaging Mstochastic forward passes; for class k, the variance of $p_k$across passes signals epistemic uncertainty. Deep ensembles estimate uncertainty by training Eindependently seeded models and combining their predictive distributions; the entropy:

$$H(x) = -\sum_k \bar{p}_k(x) \log \bar{p}_k(x), \bar{p}_k(x) = \frac{1}{E}\sum_{e=1}^{E} p_k^{(e)}(x)$$

serves as an abstention score. Route to clinicians whenever $\max_k \bar{p}_k(x) < \tau$or when entropy exceeds a policy threshold; couple this with domain rules, for example auto-escalation for pediatrics under high uncertainty.

Handling imbalance and noisy supervision

Class imbalance can be mitigated with reweighting and focal loss. For sample $i$and class y_i, focal loss for probability $p_{y_i}$is

$$\mathcal{L}_{\text{focal}} = -\alpha_{y_i}(1 - p_{y_i})^{\gamma} \log p_{y_i},$$

Where $\gamma > 0$down-weights easy examples and $\alpha_y$balances classes. Under weak supervision, label smoothing reduces overconfidence by training on targets $q = (1 - \epsilon)\,\mathbf{1}_y + \epsilon/K$with a small ϵ. A mild curriculum that starts with high-agreement, high-confidence examples and gradually incorporates harder or OOD notes stabilizes learning without complicating the stack.

Robustness and regression safety

Maintain a frozen sentinel set spanning sites, specialties, languages, and documentation styles. At each iteration, re-score this set and track discrimination, calibration, subgroup deltas, and shifts in the failure taxonomy. Keep preprocessing deterministic, version all artifacts, and attach a model factsheet that records training data versions, thresholds, calibration method, and known limitations.

Replaceability and MLOps contract

Treat the model as a pluggable component behind the same promotion gates regardless of architecture. Schema validation must pass, calibration reports must meet targets, drift and fairness dashboards must be green, and a rollback plan must be in place. Promotion requires meeting discrimination and calibration thresholds on the sentinel set with zero critical misses, and deployment must enable reject-option routing with ongoing monitoring for workload impact.

VIII. Evaluation beyond Accuracy

Multidimensional scorecard—what to measure and why

Evaluation must reflect clinical reality, not just leaderboard metrics. A practical scorecard spans six dimensions: discrimination (how well classes are separated under imbalance), calibration (whether probabilities are trustworthy), fairness (error parity across key subgroups), robustness (resilience to shift and noise), safety (human overrides and near-miss profile), and efficiency (time saved and alert burden). Each dimension has concrete metrics, acceptance gates, and monitoring plans that tie directly to workflow impact.

© 2025 IRJIET All Rights Reserved

International Research Journal of Innovations in Engineering and Technology (IRJIET)

ISSN (online): 2581-3048

Volume 9, Issue 11, pp 12-25, November-2025

https://doi.org/10.47001/IRJIET/2025.911002

Discrimination

Use AUROC for overall separability and AUPRC/F1 for imbalanced settings; report macro- and micro-averages where multi-label outputs exist. Prefer precision–recall curves when positive classes are rare (e.g., adverse events). Always report at least one operating point aligned to clinical workload (e.g., top-k suggestions for coding; interruptive vs passive alert thresholds). For sequence tasks, include span-level F1 (strict boundaries) and entity-level F1 (relaxed boundaries).

Calibration

Good discrimination can still yield unsafe actions if probabilities are miscalibrated. Compute Brier score for binary outcomes, Brier $= \frac{1}{n}\sum_{i=1}^{n}(p_i - y_i)^2$,and expected calibration error (ECE) with reliability diagrams. Post-hoc calibration (temperature scaling or isotonic regression) should be fitted on a held-out set and locked before deployment. Track calibration within subgroups (age, sex, language, site) to uncover pocket miscalibration. Target ranges are task- and site-specific; many clinical pipelines aim for ECE ≤ 0.05 with tight confidence intervals.

Fairness

Measure disparities in false-positive and false-negative rates across protected or operationally relevant subgroups: ΔFPRand ΔFNRby sex, age bands, language/code-mix, and site. Where applicable, assess equalized odds (simultaneous parity of FPR and TPR) and calibration within groups (slope/intercept similarity on reliability plots). Report subgroup sample sizes to avoid over-interpreting noisy estimates. When disparities exceed a pre-registered tolerance (e.g., absolute gap > 5 percentage points with non-overlapping 95% CIs), require mitigation (threshold per subgroup, reweighting, targeted augmentation) before promotion.

Robustness

Quantify OOD degradation by holding out an entire site or note template and reporting the F1 drop relative to in-distribution: $\Delta F1_{\text{OOD}} = F1_{\text{ID}} - F1_{\text{OOD}}$. Run corruption tests representative of the environment: typos (keyboard noise), ASR/OCR artifacts for dictated or scanned notes, and template drift (section order, boilerplate changes). Report sensitivity of predictions to these perturbations and set guardrails; for example, trigger warnings when $\Delta F1_{\text{OOD}} > 5\%$or when a corruption increases ECE by > 0.03.

Safety

Safety appears both in prevention of harm and human factors. Track clinician override rate (fraction of automated suggestions rejected), and conduct near-miss reviews where the model would have produced an unsafe omission or commission. Maintain a curated sentinel set of high-stakes cases (e.g., pregnancy, pediatrics, oncology infusions) with zero-miss expectations before any promotion. Implement a reject option: route cases to human review when $\max_k p(k \mid x) < \tau$or when explanation/validator checks fail; monitor abstention precision so routing improves safety without excessive workload.

Efficiency

Decision systems must save time and reduce noise. Measure time saved per task via time-motion studies or EHR log telemetry (e.g., coder minutes per chart; triage queue latency). Track alert burden as alerts per 100 patients per service line, with explicit caps for interruptive alerts. Report acceptance rate for actionable suggestions (e.g., top-k code acceptance), and tie these to staffing models to quantify operational value.

Confidence intervals and statistical hygiene

Every reported metric should include 95% confidence intervals. For most scalar metrics, use nonparametric bootstrap with at least 1,000 replicates and site-stratified resampling. For PR curves, compute confidence bands via bootstrap on prediction–label pairs. When comparing two systems, report paired bootstrap deltas and the proportion of replicates favoring the candidate to avoid misleading p-values under shift.

Decision-curve analysis and operating points

Translate metrics to net clinical benefit so thresholds reflect real trade-offs. For a risk threshold p_t(the probability at which action is taken), net benefit for a classification rule is

$$\text{NB}(p_t) = \frac{\text{TP}}{N} - \frac{\text{FP}}{N} \cdot \frac{p_t}{1 - p_t}.$$

Plot NB across $p_t$and compare against “treat none” and “treat all” strategies. Choose the operating point that maximizes NB under your alert-burden constraints and fairness tolerances; document the workload envelope (expected alerts/day, reviewer minutes) at that point. For ranking tasks (e.g., coder assist), use top-k utility curves and cumulative gain to align k with staff capacity.

Reporting and promotion gates

Before promotion, pre-register target thresholds with the Clinical Owner and Oversight Board. A common gate requires meeting or exceeding discrimination targets on the validation and OOD sets, ECE within the agreed band, subgroup

© 2025 IRJIET All Rights Reserved

International Research Journal of Innovations in Engineering and Technology (IRJIET)

ISSN (online): 2581-3048

Volume 9, Issue 11, pp 12-25, November-2025

https://doi.org/10.47001/IRJIET/2025.911002

ΔFPR/FNR below tolerance with mitigation plans, zero misses on the sentinel set, and an efficiency improvement (time saved or reduced alerts) validated in shadow mode. Package results in a model factsheet: data versions and lineage, label policy, calibration method, selected threshold and NB rationale, fairness audit, robustness panel, override workflow, and rollback plan. After go-live, monitor these same panels weekly for the first month and then at a steady cadence, triggering retraining or rollback when drift or disparity alarms fire.

This evaluation regimen ensures that a model's impressive headline accuracy translates into trustworthy probabilities, equitable performance, operational savings, and demonstrable clinical net benefit the qualities that determine whether a decision system succeeds in real care settings.

Table: 1 Evaluation Scorecard

Dimension	Metrics	Why it matters	Typical gate (example)
Discrimination	AUROC, AUPRC, F1 (macro/micro), Span-F1	Separates classes under imbalance	F1 (ID) ≥ target; OOD drop ≤ 5 pts
Calibration	Brier score, ECE, reliability plots	Trustworthy probabilities for thresholds	ECE ≤ 0.05 (with CI)
Fairness	ΔFPR / ΔFNR by sex/age/language/site; equalized odds	Avoids disparate harm	Absolute gap ≤ 5 pp (95% CI overlap)
Robustness	OOD F1 drop; typos/OCR/template drift tests	Real-world resilience	ΔF1_OOD ≤ 5 pts; ΔECE ≤ 0.03
Safety	Clinician override rate; near-miss review; reject-option hit rate	Prevents harm	Zero misses on sentinel set
Efficiency	Time saved per task; alerts per 100 pts; acceptance rate	Operational value	≥ 20% time savings or alert reduction

IX. Deployment Architecture and MLOps

A production system should expose two interface modes aligned to workflow cadence. For high-volume but latency-tolerant work such as coding assistance, a batch interface processes notes overnight, emits top-k suggestions with rationales, and writes results back to the EHR work queue. For time-critical use cases such as portal-message triage or safety alerts, a near-real-time API handles individual notes within strict SLAs (for example, p95 < 2–5 seconds end-to-end), returning a prediction, calibrated confidence, an abstention flag, and the minimal evidence spans required for review. Both modes use the same schema and calibration thresholds so behavior is consistent regardless of pathway.

Packaging follows a "deterministic by design" principle. Preprocessing, featureization, model inference, and postprocessing are containerized microservices with pinned versions and reproducible builds. Tokenizers, section parsers, assertion detectors, and normalization tables are versioned artifacts alongside the model, ensuring train–serve parity. Each release carries a manifest that lists dataset versions, label-policy revisions, calibration method, thresholds, and expected workload envelope; this manifest is stored with the image and referenced in the audit trail.

Observability is comprehensive and privacy-aware. The system logs hashed or redacted features, model scores, class decisions, uncertainty summaries, explanation spans, and downstream actions taken by users (accepted, edited, overridden). A living model factsheet accompanies the service and is auto-updated at release: data lineage, intended use, known hazards and countermeasures, fairness audit summary, robustness panel, and rollback steps. Telemetry is aggregated by site, specialty, language, and note type to surface environment-specific effects without exposing PHI.

Monitoring addresses three families of drift. Data drift tracks vocabulary shifts, section-mix changes, and concept-frequency deltas relative to a rolling baseline, with alerts when thresholds are exceeded. Performance drift is measured on a frozen sentinel set re-scored at regular cadence (for example, weekly), reporting discrimination, calibration, and failure-taxonomy changes; significant deltas trigger a hold on promotions and a root-cause review. Fairness drift continuously evaluates subgroup gaps in false-positives, false-negatives, and calibration; if a gap exceeds the pre-registered tolerance with stable confidence intervals, the system raises an alert and enforces mitigation, such as subgroup-specific thresholds or targeted reweighting.

Feedback loops are embedded in the clinician experience rather than treated as an afterthought. Verification widgets

© 2025 IRJIET All Rights Reserved

International Research Journal of Innovations in Engineering and Technology (IRJIET)
ISSN (online): 2581-3048
Volume 9, Issue 11, pp 12-25, November-2025
https://doi.org/10.47001/IRJIET/2025.911002

inside the EHR let users accept, modify, or reject suggestions and optionally tag the reason (wrong assertion, boundary error, code mismatch). Disagreements are routed to an adjudication queue where senior reviewers resolve cases and update the written guideline. Changes feed back into labeling functions and the weak-supervision label model, and the platform schedules retraining windows that avoid peak clinical hours. Each retrain executes in a shadow environment first, replaying recent traffic to validate calibration, workload impact, and subgroup parity before gradual rollout.

Security and privacy are enforced end-to-end. Raw PHI remains isolated in a secure enclave with least-privilege, time-boxed access; de-identified derivatives power annotation and modeling. Every request is authenticated, authorized, and fully auditable; logs record who ran what model on which artifact with which parameters. When cross-site learning is necessary, federated training keeps data local and exchanges only model updates; if the risk profile warrants, differential privacy is applied to gradients or outputs to bound information leakage. Key material and credentials are rotated automatically, backup and disaster-recovery plans are tested, and all containers run with minimal capabilities, read-only filesystems, and strict egress controls. Together, these practices make the model a replaceable component in a safe, observable, and continuously improving clinical AI service.

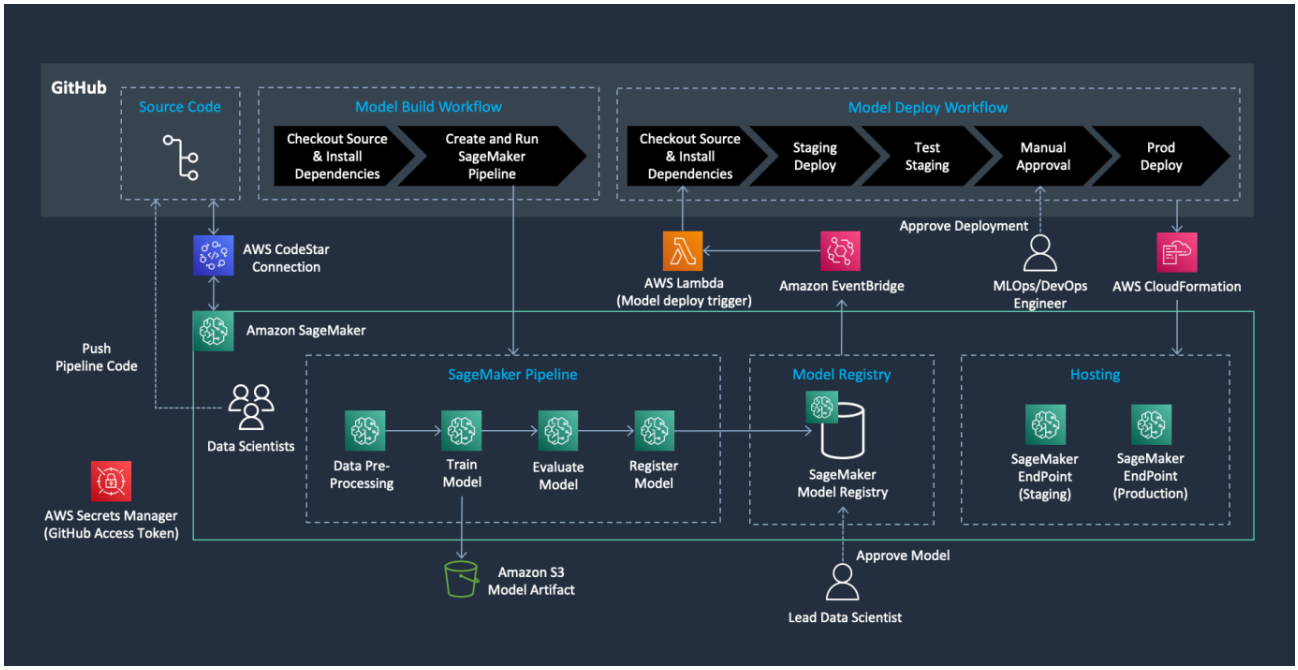

Figure: 2 MLOps architecture (AWS SageMaker Pipelines)

X. Case Studies (Design Blueprints)

ICD Code Suggestion from Discharge Summaries

The system ingests finalized discharge summaries and highlights evidence spans in Assessment and Plan to propose ICD-10 codes across the top-200 frequent classes while maintaining tail coverage via open-set detection. Labeling functions combine section-aware dictionaries, negation and uncertainty scopes, and simple procedure-diagnosis linkage rules so postoperative findings are not miscoded as active conditions. A doubly annotated gold set of about five thousand notes anchors adjudication and calibration. The encoder produces calibrated top-k suggestions with rationales; coders accept or edit within the EHR, and those edits feed an audit trail for reimbursement compliance and guideline refinement. Performance is tracked with micro- and macro-F1, precision at k, and reliability plots for acceptance-threshold tuning; promotion gates require stable calibration and minimal drift on a held-out site.

Adverse Drug Event Detection in Progress Notes

Notes are parsed into an event graph with nodes for drug, reaction, severity, and onset, connected by causal cues such as “after starting” or “seconds post-infusion.” Programmatic signals come from trigger lexicons for dermatologic, respiratory, and anaphylactoid reactions, plus exclusion zones that down-weight mentions in problem lists without temporal anchors. Human reviewers adjudicate ambiguous temporality and confounding disease flares. The model outputs case-level alerts only when both an event and a plausible causal relation are present and calibrated above a clinical threshold; uncertain cases route to pharmacists via reject-option rules. Evaluation

© 2025 IRJIET All Rights Reserved

International Research Journal of Innovations in Engineering and Technology (IRJIET)
ISSN (online): 2581-3048
Volume 9, Issue 11, pp 12-25, November-2025
https://doi.org/10.47001/IRJIET/2025.911002

emphasizes event-level F1, case-level sensitivity and specificity, and a monitored near-miss panel for high-severity reactions; deployment includes a safety board review of false reassurance and override patterns.

Radiology Impression Normalization

Impression sections from chest and neuro imaging are mapped to standardized clinical concepts with explicit assertion status so statements like "no pneumothorax" are encoded as negated findings. Span extraction is followed by ontology normalization and assertion detection tuned to dictation style. To improve resilience, augmentation introduces validator-checked paraphrases that preserve negation and scope while simulating common report variants; robustness is assessed under OCR noise and templated dictation artifacts. Outputs drive registry updates and longitudinal dashboards, with calibration monitored separately for affirmed versus negated classes to avoid asymmetric errors.

Patient Message Triage (Portal)

Incoming portal messages are classified into emergency advice, urgent-within-24-hours, or routine, with multilingual and code-mixed handling where needed. The model surfaces concise rationales and an uncertainty score; low-confidence or policy-sensitive cases are automatically escalated. Bias controls continuously audit performance across age and language subgroups, adjusting thresholds if disparities exceed tolerance. Utility is measured as reduction in nurse triage time and queue latency, balanced against the false-reassurance rate reviewed by a safety committee. Shadow runs precede go-live to size workload envelopes, after which monitoring tracks alert volume, acceptance, and subgroup calibration to support safe, incremental expansion.

XI. Trustworthy & Ethical Considerations

Explainability: Each prediction includes concept-level highlights and a short rationale tied to ontology terms and source sections, stored with the output so clinicians can audit what evidence drove the decision and verify that no negation cues were missed.

Calibration & Triage: Operating thresholds are chosen via decision-curve analysis to maximize net clinical benefit under alert-burden limits; low-confidence or validator-failed cases auto-escalate, and abstention reasons are logged to tune the reject policy.

Fairness: Disparities in false-positive/false-negative rates and calibration are tested pre-deployment and monitored continuously by sex, age, language/code-mix, and site; gaps beyond tolerance trigger documented mitigation such as subgroup thresholds, reweighting, or targeted augmentation.

Regulatory Context: Data handling and use conform to local health-data rules (e.g., ABDM/HIPAA/GDPR) with DPIA, least-privilege PHI access, and full audit trails; when outputs inform care, releases follow medical-device change control, post-market monitoring, and a defined rollback plan.

Experimental Template

Data Splits: Use site-stratified 70/15/15 train/val/test plus a true OOD holdout by site or template; maintain a frozen sentinel panel of high-stakes cases for regression checks across releases.

Baselines: Compare a linear bag-of-ngrams model, an encoder fine-tune, and a constrained LLM/RAG pipeline under identical preprocessing and label policies to separate modeling gains from data handling.

Ablations: Quantify contributions from adjudication over weak labels, validator-gated augmentation, calibration vs no calibration, and successive active-learning rounds to reveal label efficiency and threshold stability.

Reporting: Mirror the production scorecard: AUROC/AUPRC/F1 for discrimination, Brier/ECE with reliability plots for calibration, subgroup gaps with CIs for fairness, ID→OOD and corruption drops for robustness, override and near-miss reviews for safety, and time saved plus alerts per 100 patients for efficiency; include an error taxonomy with representative de-identified snippets and archive all configs, datasets, and calibration parameters for full reproducibility.

XII. Limitations and Future Work

Limitations: Clinical narratives are heterogeneous across sites, services, languages, and documentation styles; even small template or vocabulary shifts can degrade performance, so portability depends on continual corpus curation and monitoring. Labels inherit bias from documentation practices (under-recording in certain cohorts, copy-forward artifacts), and demographic attributes needed for fairness audits are often incomplete or sensitive, constraining subgroup analysis. Weak supervision accelerates scale but introduces label noise; without disciplined adjudication and calibration, overconfident errors persist. LLM components remain vulnerable to hallucination and subtle assertion/temporality mistakes; retrieval and constrained decoding reduce—but do not eliminate—these risks. Privacy requirements limit centralized training and comprehensive error sharing; federated learning and differential privacy trade utility for protection and are

© 2025 IRJIET All Rights Reserved

ISSN (online): 2581-3048
Volume 9, Issue 11, pp 12-25, November-2025
https://doi.org/10.47001/IRJIET/2025.911002

operationally complex. From an operations perspective, human-in-the-loop workflows impose reviewer burden; sustained value depends on careful thresholding, abstention routing, and UX that minimizes cognitive load. Finally, evaluation commonly focuses on point predictions within a single encounter; many clinical tasks require longitudinal reasoning, causal interpretation, and alignment with downstream outcomes rather than surrogate labels.

Future work: Priorities include long-context models tuned to multi-encounter narratives (progress notes + imaging + orders) with structured memory to preserve temporal state; robust temporal reasoning that grounds onset, resolution, and treatment response across encounters; and causal learning to separate documentation habits from true clinical effects and to support counterfactual "what-if" recommendations. Data efforts should expand multilingual and code-mixed corpora with audited de-identification, plus shared "shift suites" that simulate OCR/ASR noise, template changes, and policy drift. Privacy-preserving collaboration will benefit from hybrid approaches that mix federated representation learning, site-specific fine-tuning, and tight privacy budgets validated by task-level utility curves. On supervision, semi-automated guideline induction from adjudication logs, ontology-aware active learning, and validator-gated synthetic variants can raise label fidelity without explosive costs. Finally, usability research is needed to calibrate explanations, uncertainty displays, and reject-option behaviors to clinician workflows, and to link system improvements to patient-relevant outcomes (time-to-action, adverse event prevention) rather than proxy metrics alone.

XIII. Conclusion

Data-Centric AI reorients clinical NLP from model tinkering to disciplined data design and lifecycle operations. By anchoring tasks in clinical utility, normalizing outputs to shared ontologies, and building supervision pipelines that blend programmatic labeling with clinician adjudication and active learning, organizations can lift not only discrimination but also calibration, robustness, and fairness. Trustworthy training and deployment—through deterministic preprocessing, privacy-preserving learning, auditable lineage, and continuous monitoring of data, performance, and subgroup parity—turn models into replaceable components within a safe, observable service. The result is a learning health system: datasets, guidelines, and models co-evolve under governance, error taxonomies feed targeted augmentation and retraining, and uncertainty-aware workflows keep humans in control where it matters. Implemented this way, textual understanding becomes reliably actionable—shortening time-to-action, reducing alert burden, and enabling more equitable, timely, and safer patient care at scale.

REFERENCES

[1] Andresini, G., Appice, A., Ienco, D., et al. (2024). DIAMANTE: A datacentric semantic segmentation approach to map tree dieback induced by bark beetle infestations via satellite images. *In: Journal of intelligent information systems*. https://doi.org/10.1007/s10844-024-00877-6.

[2] Burch, M., & Weiskopf, D. (2013). On the benefits and drawbacks of radial diagrams. In: Handbook of human centric visualization. *Springer*, pp. 429– 451. https://doi.org/10.1007/978-1-4614-7485-2_17.

[3] Frid-Adar, M., E. Klang, M. Amitai, et al. (2018). Synthetic data augmentation using GAN for improved liver lesion classification. *In: 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018). IEEE*, pp. 289–293. https://doi.org/10.1109/ISBI.2018.8363576.

[4] Jakubik, J., Vössing, M., Kühl, N., et al. (2024). Data-centric artificial intelligence. *In: Business & information systems engineering*. https://doi.org/10.1007/s12599-024-00857-8.

[5] Kumar, S., Datta, S., Singh, V., et al. (2024). Opportunities and Challenges in Data-Centric AI. *In: IEEE Access*. https://doi.org/10.1109/ACCESS.2024.3369417.

[6] Luley, P., Deriu, J. M., Yan, P., et al. (2023). From concept to implementation: The data-centric development process for AI in industry. *In: 2023 10th IEEE Swiss Conference on Data Science (SDS). IEEE*, pp. 73–76. https://doi.org/10.1109/SDS57534.2023.00017.

[7] Gudivada V, Apon A, Ding J (2017) Data quality considerations for big data and machine learning: going beyond data cleaning and transformations. *Int J Adv Softw* 10(1):1–20

[8] Lin Q, Ye G, Wang J, Liu H (2022) RoboFlow: a data-centric workflow management system for developing AI-enhanced robots. *In: Proceedings of the conference on robot learning. PMLR*, pp 1789–1794

[9] Peng, J., Wu, W., Lockhart, B., et al. (2021). Dataprep. eda: Task-centric exploratory data analysis for statistical modeling in python. *In: Proceedings of the 2021 international conference on management of data*, pp. 2271– 2280. https://doi.org/10.1145/3448016.3457330.

[10] Roscher, R., Rußwurm, M., Gevaert, C., et al. (2023). Data-centric machine learning for geospatial remote sensing data. *In: CoRR*. https://doi.org/10.48550/arXiv2312.05327.

© 2025 IRJIET All Rights Reserved

ISSN (online): 2581-3048
Volume 9, Issue 11, pp 12-25, November-2025
https://doi.org/10.47001/IRJIET/2025.911002

[11] Seedat, N., Imrie, F., & van der Schaar, M. (2024). Navigating Data-Centric Artificial Intelligence With DC-Check: Advances, Challenges, and Opportunities. *In: IEEE Transactions on Artificial Intelligence* 5.6. https://doi.org/10.1109/TAI.2023.3345805.

[12] Whang, S. E., Roh, Y., Song, H., et al. (2023). Data collection and quality challenges in deep learning: A data-centric AI perspective. *In: The VLDB Journal* 32.4, pp. 791–813.

[13] Zahid, A., Kay Poulsen, J., Sharma, R., et al. (2021). A systematic review of emerging information technologies for sustainable data-centric healthcare. *In: International Journal of Medical Informatics* 149. https://doi.org/10.1016/j.ijmedinf.2021.104420.

[14] de Carvalho, O. L. F., de Carvalho Junior, O. A., de Albuquerque, A. O., Orlandi, A. G., Hirata, I., Borges, D. L., Gomes, R. A. T., & Guimarães, R. F. (2023). A data-centric approach for wind plant instance-level segmentation using semantic segmentation and gis. *Remote Sensing*, 15(5), 1–23.

[15] Ferreira de Carvalho, O.L., Olino de Albuquerque, A., Luiz, A.S., Henrique Guimarães Ferreira, P., Mou, L., e Silva, D.G., Abílio de Carvalho Junior, O. (2023). A data-centric approach for rapid dataset generation using iterative learning and sparse annotations. *In: IGARSS 2023 - 2023 IEEE International Geoscience and Remote Sensing Symposium*, pp. 5650–5653.

Citation of this Article:

Sabiha Tasneem. (2025). Data-Centric Artificial Intelligence for Textual Understanding in Healthcare Decision Systems. *International Research Journal of Innovations in Engineering and Technology - IRJIET*, 9(11), 12-25. Article DOI https://doi.org/10.47001/IRJIET/2025.911002

****

© 2025 IRJIET All Rights Reserved