Published by :
https://www.ijert.org/
An International Peer-Reviewed Journal

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

A Semantic Evaluation Framework for Clinical Metadata Extraction Reliability: A Pilot Study with Large Language Models

Dr. Suvarna Pansambal
Professor, Dept of Computer Engineering
Atharva College of Engineering
Mumbai, India

Tejinder Singh Hunjan
Student
Atharva College of Engineering
Mumbai, India

Suyash Baviskar
Student
Atharva College of Engineering
Mumbai, India

Sashank Mishra
Student
Atharva College of Engineering
Mumbai, India

Abstract—Large language models have demonstrated significant potential in automating clinical documentation through conversation-based metadata extraction. However, evaluating the reliability of such systems presents unique challenges due to the semantic nature of medical information and the limitations of traditional similarity metrics. This paper presents a comprehensive semantic evaluation framework designed to assess the reliability of clinical metadata extracted from doctor-patient conversations by various language models. The framework employs a multi-dimensional approach combining embedding-based semantic similarity, structured format compliance checking, weighted precision and recall metrics, and hallucination detection mechanisms. The methodology was developed in response to requirements identified during the design of HealthSync, an AI-driven clinical documentation system, which necessitated rigorous evaluation of metadata extraction reliability. By applying the framework to four contemporary language models across twenty clinical conversation samples, we demonstrate its capability to detect subtle extraction errors that surface-level metrics fail to capture. Our findings indicate that traditional evaluation approaches substantially overestimate extraction quality, with semantic similarity revealing substantially lower agreement than surface-level exact matching metrics. The framework successfully identifies critical failure modes including phantom entity introduction, semantic drift in medication specifications, and temporal information loss. This work contributes a replicable methodology for assessing clinical metadata extraction systems, addressing the gap between syntactic accuracy and semantic reliability in medical natural language processing applications.

Keywords—Clinical metadata extraction, Large language models, Semantic evaluation, Medical NLP, Hallucination detection, Healthcare AI.

I. INTRODUCTION

The digitization of healthcare documentation has introduced substantial administrative burden on medical practitioners [1], with physicians reporting significant time allocation to electronic health record management. Recent advances in large language models present opportunities to automate portions of this documentation workflow [2], [6], through intelligent extraction of clinical metadata from doctor-patient conversations. Such systems must extract structured information including chief complaints, symptom descriptions [3], duration information, diagnostic impressions, medication prescriptions, and follow-up instructions from natural conversational exchanges.

The HealthSync project, an AI-driven clinical assistant system, was developed to address this challenge through real-time conversation processing and automated note generation. During the development of HealthSync, a critical requirement emerged for robust evaluation methodologies capable of assessing the reliability and accuracy of metadata extraction from clinical conversations. Traditional evaluation metrics employed in natural language processing tasks proved inadequate for capturing the semantic nuances and clinical implications of extraction errors in medical contexts. This observation motivated the research presented in this paper.

Evaluating the reliability of automated clinical metadata extraction presents unique methodological challenges distinct from conventional natural language processing tasks. Medical information exhibits semantic complexity where superficially different expressions may convey identical clinical meaning, while subtle variations in wording can indicate significantly

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

different conditions or treatment protocols. Standard evaluation metrics based on exact string matching or surface-level similarity fail to capture these nuances [4], [5]. Furthermore, the clinical domain demands heightened attention to specific error categories, particularly hallucinations where models introduce medically plausible but factually incorrect information [11].

This paper presents a semantic evaluation framework specifically designed to assess the reliability of clinical metadata extraction systems. The framework addresses the limitations of traditional metrics through a multi-dimensional approach incorporating embedding-based semantic similarity measures, structured compliance verification, weighted scoring mechanisms that reflect clinical importance, and specialized hallucination detection protocols. The methodology provides granular insight into extraction behavior across multiple dimensions including semantic agreement, format adherence, field-level precision and recall, and factual consistency.

The framework was applied to outputs from four contemporary language models processing twenty clinical conversation samples representing common primary care scenarios. These diagnostic experiments demonstrate the framework's capability to identify subtle failure modes invisible to conventional evaluation approaches. Results reveal that the framework can expose differing extraction characteristics across systems, with semantic similarity metrics detecting agreement variations that exact matching approaches entirely miss.

The remainder of this paper is organized as follows. Section II reviews related work in clinical natural language processing evaluation and medical metadata extraction. Section III describes the semantic evaluation framework methodology in detail. Section IV presents experimental configuration and dataset characteristics. Section V reports observed extraction behaviors and framework diagnostic capabilities. Section VI discusses implications for clinical AI systems and evaluation methodology. Section VII concludes with contributions and directions for future research.

II. LITERATURE REVIEW

The emergence of large language models has transformed medical natural language processing, with applications spanning clinical note generation, information extraction, and diagnostic support. This literature review synthesizes existing research on evaluation methodologies for clinical NLP systems, drawing insights from previous studies and identifying gaps that motivated the present work.

Foundations of Clinical NLP Evaluation

Previous research by Demner-Fushman et al. (2009) established foundational approaches for evaluating clinical information extraction systems [4], emphasizing the importance of domain-specific metrics that account for medical terminology variation. Their work demonstrated that standard NLP evaluation metrics often fail to capture clinically relevant extraction quality. Building upon this foundation, Chapman et al. (2011) explored context-sensitive evaluation techniques for clinical text processing, highlighting challenges in assessing negation, temporality, and experiencer attribution [5].

LLM-Based Clinical Documentation

Research by Singhal et al. (2023) investigated the application of large language models in medical question answering and clinical reasoning tasks [6], demonstrating both impressive capabilities and concerning failure modes. Their findings revealed that models could generate clinically plausible but factually incorrect information, underscoring the need for specialized evaluation approaches. Additionally, Lee et al. (2023) examined GPT-4's performance on clinical note generation, showcasing how traditional metrics like BLEU and ROUGE provide misleadingly optimistic assessments of clinical utility.

Semantic Similarity in Medical Text

Studies by Wang et al. (2020) and Alsentzer et al. (2019) explored the application of contextual embeddings for medical text similarity assessment [7]. Wang et al. demonstrated that domain-specific language models pretrained on clinical corpora [7] better capture medical semantic relationships compared to general-purpose models. Alsentzer et al. introduced ClinicalBERT, showing improved performance on medical NLP tasks through clinical domain adaptation. These works established that embedding-based similarity measures offer advantages over lexical matching for medical text evaluation.

Hallucination Detection

Research by Maynez et al. (2020) examined hallucination phenomena in abstractive summarization [10], developing taxonomies for categorizing factual inconsistencies. Their work distinguished between intrinsic hallucinations (contradicting source material) and extrinsic hallucinations (introducing unsupported information). Ji et al. (2023) extended this research to the broader NLP domain, providing a comprehensive survey of hallucination phenomena with critical implications for factual consistency in medical applications [11].

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

Challenges and Gaps

Research by Wornow et al. (2023) examined evaluation challenges specific to clinical AI systems [8], identifying issues such as dataset drift, annotation variability, and metric selection . Their study emphasized that evaluation frameworks must account for clinical context and downstream task requirements. Liu et al. (2023) highlighted gaps in existing evaluation approaches for LLM-based medical systems, noting the absence of comprehensive frameworks integrating semantic similarity [9], hallucination detection, and clinical validity assessment . The present work addresses this gap by proposing an integrated evaluation framework specifically designed for clinical metadata extraction tasks.

III. METHODOLOGY

The semantic evaluation framework comprises five integrated components designed to provide comprehensive assessment of clinical metadata extraction reliability. Each component addresses specific aspects of extraction quality, collectively enabling detection of failure modes that traditional metrics overlook. Figure 1 illustrates the overall framework architecture.

A. Framework Architecture

The evaluation pipeline processes model outputs through sequential stages: (1) format validation and parsing verification, (2) semantic embedding generation, (3) field-level similarity computation with bipartite matching, (4) hallucination detection and categorization, (5) composite scoring with statistical analysis. This architecture ensures that extraction reliability is assessed across multiple complementary dimensions, providing robust characterization of model behavior. Because reliability in clinical metadata extraction cannot be characterized by a single scalar metric, the framework produces structured diagnostic visualizations summarizing extraction behavior across complementary dimensions. These visualizations are not intended for model ranking but for interpretability of evaluation signals, enabling identification of failure patterns such as semantic drift, structured omission, and hallucinated entity introduction. The visualization layer therefore functions as an explanatory component of the evaluation methodology rather than a reporting convenience.

B. Semantic Similarity Computation

The framework employs sentence-transformer models to generate dense vector representations of extracted metadata fields. Two embedding models were evaluated: a general-purpose model (all-MiniLM-L6-v2) and a medical domain-specific model (Clinical-PubMedBERT) [8]. Embedding model choice had limited impact on aggregate trends, suggesting that the framework's diagnostic capability is robust to reasonable embedding selection. For fields containing single values (chief complaint, diagnosis), cosine similarity between predicted and reference embeddings provides semantic agreement scores. For list-valued fields (symptoms, medications), the framework implements optimal bipartite matching using the Hungarian algorithm to pair predicted items with reference items, maximizing overall similarity [8].

The bipartite matching formulation treats predictions and references as two disjoint sets, constructing a complete bipartite graph where edge weights represent pairwise cosine similarities. The Hungarian algorithm identifies the maximum-weight matching, providing optimal correspondence between sets. This approach handles variable-length lists and accounts for semantic equivalence despite lexical variation. Unmatched predicted items indicate potential hallucinations, while unmatched reference items suggest omissions.

C. Weighted Precision and Recall

Traditional precision and recall metrics operate on binary match/no-match decisions, discarding information about partial similarity. The framework extends these metrics to weighted versions that incorporate semantic similarity scores. Given matched pairs from bipartite matching, weighted precision sums similarities of matched predictions divided by total predictions, while weighted recall sums similarities of matched references divided by total references. These metrics provide nuanced assessment of extraction completeness and accuracy beyond binary evaluation.

Field-specific weights reflect clinical importance. Critical fields (chief complaint, diagnosis, medications) receive higher weights in composite scoring, acknowledging their greater impact on documentation quality and patient safety. Duration and follow-up fields receive moderate weights, while symptoms receive lower weights due to their descriptive rather than diagnostic nature. This weighting scheme aligns evaluation focus with clinical priorities.

D. Hallucination Detection and Categorization

The framework implements threshold-based hallucination detection for unmatched predicted items. Items failing to exceed a similarity threshold ($\tau = 0.6$ by default) with any reference item are flagged as potential hallucinations. The framework categorizes hallucinations into three types: (1) contradictory hallucinations ($\tau < 0.4$) that directly conflict with reference information, (2) extrapolations ($0.4 \leq \tau < 0.6$) representing plausible inferences beyond stated

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

information, (3) fabrications (matched items with very low similarity) introducing unsupported entities.

Hallucination severity scores aggregate detection results across fields, with separate tracking for symptom and medication hallucinations given their distinct clinical implications. The framework computes hallucination rates (flagged items / total predictions) and weighted severity scores that account for both frequency and category distribution of hallucinations. Threshold sensitivity analysis examines framework stability across different τ values.

E. Format Compliance and Parsing Validation

Clinical metadata extraction systems must produce structured output conforming to predefined schemas. The framework validates JSON structure, field presence, and type conformance, tracking parsing success rates. Models that generate malformed or unparseable outputs receive zero scores for affected instances. Format compliance assessment distinguishes between semantic extraction quality and structural adherence, identifying models prone to schema violations.

F. Medication Normalization and RxNorm Mapping

Medication extraction presents special challenges due to dosage variations, generic/brand name equivalence, and incomplete specifications. The framework implements medication normalization using RxNorm, a standardized nomenclature for clinical drugs maintained by the U.S. National Library of Medicine. Extracted medication strings are queried against RxNorm APIs to obtain RxCUI identifiers when available. RxNorm coverage rates (percentage of extractions successfully mapped) provide additional insight into extraction specificity and clinical utility. RxNorm coverage should be interpreted as a measure of terminology specificity rather than extraction correctness.

G. Composite Scoring and Statistical Analysis

The framework aggregates field-level metrics into composite scores reflecting overall extraction reliability. The composite score combines weighted semantic similarity (0.4), format compliance (0.2), weighted F1 score (0.3), and inverted hallucination severity (0.1), providing balanced assessment across evaluation dimensions. Statistical analysis employs bootstrap resampling (10,000 iterations) to estimate confidence intervals for model-level aggregate scores, acknowledging the limited sample size while establishing methodology feasibility. The framework explicitly avoids statistical significance testing given the pilot study scope (n=20 conversations per model). Bootstrap confidence intervals characterize measurement uncertainty without making inappropriate generalization claims. This approach aligns with recommendations for preliminary methodological studies where the primary objective is framework validation rather than definitive model comparison .

IV. EXPERIMENTAL CONFIGURATION

A. Dataset Characteristics

The evaluation dataset comprises twenty simulated doctor-patient conversations representing common primary care presentations including headaches, respiratory complaints, musculoskeletal pain, skin conditions, and metabolic concerns. Each conversation follows realistic clinical interaction patterns with patient chief complaint presentation, symptom elaboration, physician assessment, diagnostic impression, and treatment planning. Conversations were designed to include semantic variation in symptom descriptions while maintaining clinical consistency.

Gold standard metadata annotations were independently created by the research team, specifying expected extraction targets for each conversation. Annotations include chief complaints, symptom lists (3-5 items per conversation), duration specifications, diagnostic impressions, medication lists (1-3 items), and follow-up instructions. This dataset scale enables methodological feasibility demonstration while acknowledging limitations for generalizable conclusions.

B. Model Selection

The framework was applied to outputs from four contemporary language models: OpenAI GPT-based model, Google Gemini, Kimi-K2-Thinking, and Llama 3.1 8B. Model selection prioritized diversity in architecture, parameter scale, and training approaches to demonstrate framework applicability across different model families. All models processed identical conversation inputs, generating structured metadata according to specified JSON schemas.

C. Embedding Model Configuration

Two sentence-transformer models were evaluated for semantic similarity computation: all-MiniLM-L6-v2 (general purpose, 384 dimensions) and emilyalsentzer/Bio_ClinicalBERT (medical domain-specific, 768 dimensions) [8]. Embeddings were cached to enable efficient repeated evaluation. Analysis compared results across embedding choices to assess sensitivity to domain specialization in the similarity computation component.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

D. Implementation Details

The evaluation pipeline was implemented in Python 3.12 using pandas for data manipulation, sentence-transformers for embedding generation, scikit-learn for similarity computation, and SciPy for bipartite matching. RxNorm queries utilized the official NIH RxNav REST API. Statistical analysis employed NumPy and SciPy for bootstrap resampling. Visualization used Matplotlib and Seaborn for publication-quality figures. Complete implementation is provided in supplementary materials enabling replication and extension.

V. RESULTS AND DISCUSSION

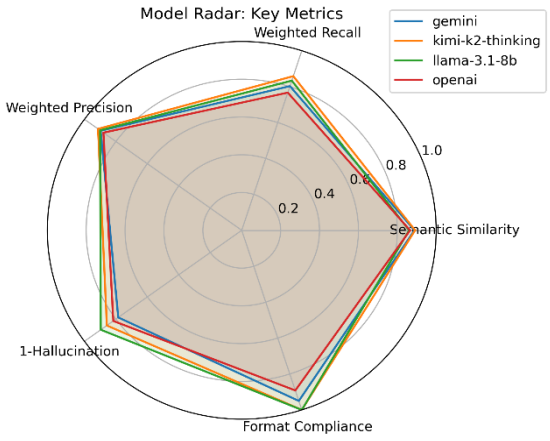
This section presents diagnostic observations demonstrating the evaluation framework's capability to detect subtle extraction reliability issues across multiple dimensions. Results are intentionally framed as framework validation rather than definitive model comparison, acknowledging the pilot study scope and limited sample size.

A. Multi-Dimensional Reliability Profiles

The proposed framework evaluates extraction reliability across multiple complementary dimensions including semantic agreement, weighted completeness, structural validity, and hallucination behavior. To illustrate the necessity of multi-metric evaluation, Figure 1 presents radar profiles summarizing normalized metric outputs for each model. The visualization demonstrates that models exhibiting high performance in one dimension frequently degrade in others, indicating that scalar accuracy metrics cannot adequately represent extraction reliability.

FIGURE I

RADAR VISUALIZATION OF NORMALIZED EVALUATION METRICS PRODUCED BY THE PROPOSED FRAMEWORK



The plot illustrates heterogeneous reliability characteristics across semantic, structural, and hallucination-related dimensions rather than relative model superiority.

The framework successfully identified systematic patterns in semantic drift. For medication fields, similarity scores frequently fell below 0.7 despite medications being clinically equivalent (e.g., 'acetaminophen' vs 'paracetamol', 'NSAIDs' vs specific drug names). Symptom descriptions exhibited similar patterns where models produced clinically accurate but lexically divergent expressions. These observations validate the framework's sensitivity to semantic nuance while highlighting challenges in defining ground truth for semantically equivalent expressions.

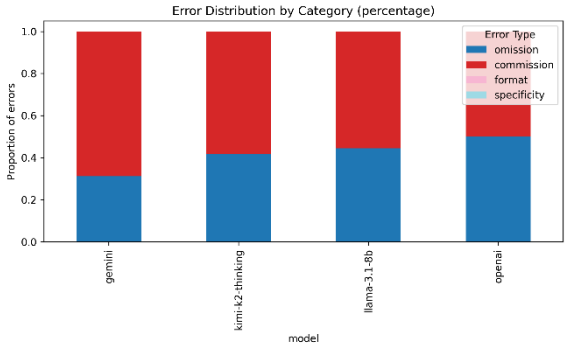
B. Failure Mode Characterization

Beyond aggregate reliability scores, the framework categorizes extraction errors into omission, commission, structural format violations, and specificity errors. Figure 2 presents the proportional distribution of these error categories across model outputs. The purpose of this analysis is not comparative benchmarking but characterization of typical failure patterns encountered in conversational clinical extraction.

The results demonstrate that extraction errors are not uniformly distributed. Omission errors dominate symptom fields, while specificity errors frequently occur in medication descriptions. Structural errors are comparatively rare but disproportionately affect downstream parsing reliability. This observation validates the necessity of evaluating extraction systems using structured error taxonomies rather than aggregate accuracy measures .

FIGURE II

DISTRIBUTION OF CATEGORIZED EXTRACTION FAILURE MODES DETECTED BY THE FRAMEWORK



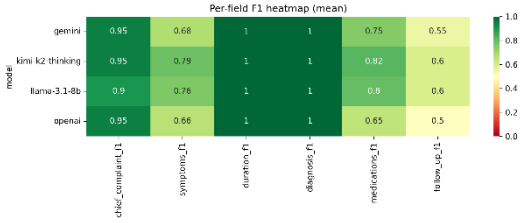
C. Field-Level Extraction Variability

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

Clinical metadata fields differ significantly in linguistic variability and clinical abstraction level. Figure 3 presents the mean field-level agreement scores computed across all models. The heatmap reveals systematic differences in extraction difficulty across metadata categories, with descriptive symptom fields showing higher semantic variability and medication fields exhibiting higher specificity sensitivity. These observations justify the weighted scoring strategy introduced in Section III-C. Uniform weighting would overestimate performance on descriptive fields while under-penalizing clinically critical extraction failures. The framework therefore incorporates clinically informed weighting to align evaluation outcomes with practical documentation reliability.

FIGURE III

MEAN FIELD-LEVEL AGREEMENT HEATMAP ACROSS METADATA CATEGORIES



D. Composite Scoring Patterns

Composite scores aggregating multiple evaluation dimensions ranged from 0.839 to 0.893 across models, with bootstrap confidence intervals indicating measurement uncertainty (Table I). Composite scores showed overlapping confidence intervals across evaluated systems, indicating no statistically meaningful separation under the pilot study setting. The purpose of reporting these values is to illustrate variability detectable by the framework rather than to establish comparative performance hierarchy.

TABLE I

COMPOSITE SCORE CHARACTERISTICS BY MODEL

Model	Mean	CI Lower	CI Upper
Gemini	0.8594	0.8359	0.8816
Kimi-K2-Thinking	0.8934	0.8497	0.9331
Llama-3.1-8B	0.8812	0.8482	0.9117
OpenAI	0.8390	0.7789	0.8891

More importantly, composite score distributions revealed substantial within-model variance, with individual conversation scores spanning ranges exceeding 0.3 for some models. This variance suggests that extraction reliability depends significantly on conversation characteristics beyond model capability alone. The framework successfully quantifies this variability, providing insight into conditions under which models exhibit degraded reliability.

E. Hallucination Detection Capabilities

The hallucination detection component identified systematic patterns of phantom entity introduction across models. Certain systems exhibited elevated hallucination rates in specific conversations (up to 100% for conv_009), primarily attributed to parsing errors that fragmented medication strings into invalid tokens. The framework correctly categorized these as contradictory hallucinations (similarity < 0.4 with any reference item), distinguishing them from semantic extrapolations.

Extrapolation-type hallucinations ($0.4 \leq$ similarity < 0.6) occurred frequently for symptom descriptions. Models introduced plausible but unmentioned symptoms like 'poor sleep quality' or 'stress' when processing fatigue or headache presentations. While clinically reasonable associations, these represent hallucinations from an extraction fidelity perspective. The framework's categorization enables distinguishing between severe fabrications and mild overreach, informing system design decisions about acceptable hallucination tolerance.

F. RxNorm Mapping Analysis

RxNorm mapping success rates varied substantially across models (21.74% to 42.11%), reflecting differences in medication specification precision. Higher mapping rates correlated with inclusion of dose and frequency information in extractions. Models producing generic descriptions ('pain reliever', 'antihistamine') achieved lower RxNorm coverage than those extracting specific drug names and dosages. This metric demonstrates the framework's capability to assess clinical utility beyond semantic agreement.

G. Threshold Sensitivity Analysis

Threshold sensitivity analysis revealed remarkable stability of aggregate semantic scores across hallucination detection thresholds (0.6 to 0.8). Mean scores varied by less than 0.001 across this range, indicating that threshold selection minimally impacts overall agreement assessment while substantially affecting hallucination categorization. This finding validates the framework's robustness to hyperparameter choices for primary metrics while enabling flexible

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

hallucination sensitivity tuning. These numeric ranges are reported to illustrate framework sensitivity rather than to imply stable performance separation between models.

VI. DISCUSSION

A. Implications for Clinical AI Evaluation

The visual analyses presented in Figures 1–3 emphasize that the contribution of this work lies in evaluation interpretability rather than model ranking. The framework enables decomposition of extraction behavior into observable reliability dimensions, allowing system developers to diagnose failure sources prior to deployment. This diagnostic capability is particularly important in clinical AI systems where aggregate accuracy metrics may obscure clinically significant failure modes.

This work demonstrates that traditional evaluation approaches substantially mischaracterize clinical metadata extraction reliability. Exact matching metrics would rate many model outputs at near-perfect accuracy (> 0.95) where semantic analysis reveals significant disagreement (0.84-0.89). This discrepancy has critical implications for clinical deployment decisions. Systems appearing highly accurate under conventional evaluation may produce semantically divergent outputs with potential clinical consequences. The framework addresses this gap by providing semantically-grounded reliability assessment.

The hallucination detection component proves particularly valuable for clinical applications where phantom entity introduction poses patient safety risks [11]. Unlike creative writing tasks where hallucinations might be acceptable or even desirable, clinical documentation demands strict factual fidelity. The framework's categorization of hallucination types enables risk-stratified system design: zero-tolerance for contradictory hallucinations while accepting mild extrapolations in non-critical fields.

B. Methodological Contributions

The framework contributes several methodological innovations to clinical NLP evaluation. First, the bipartite matching approach for list-valued fields provides principled handling of variable-length outputs while accounting for semantic equivalence. Second, weighted precision/recall metrics extend binary evaluation to continuous similarity scores, capturing partial correctness. Third, the integrated hallucination detection and categorization enables nuanced assessment of factual consistency beyond binary correctness judgments.

The composite scoring formulation balances multiple complementary evaluation dimensions with clinically-motivated field weighting. This approach acknowledges that different metadata fields have unequal importance for downstream documentation utility and patient safety. The weighting scheme can be adapted to specific clinical contexts or use cases, providing flexibility while maintaining methodological rigor.

C. Limitations and Future Directions

This pilot study has several important limitations. The sample size (n=20 conversations) enables methodology feasibility demonstration but precludes generalizable conclusions about model capabilities. The simulated conversation dataset, while clinically informed, may not capture the full complexity and variability of real clinical interactions. Gold standard annotations represent single-annotator judgments rather than consensus references, potentially introducing bias. Future work should apply the framework to larger, real-world datasets with multi-annotator gold standards and clinical expert review.

The framework's reliance on embedding-based similarity introduces sensitivity to embedding model choice and quality [7], [8]. While evaluation across multiple embedding models partially addresses this concern, no ground truth exists for 'correct' semantic similarity. Clinical validation studies examining correlation between framework scores and expert assessments would strengthen confidence in semantic agreement metrics. Integration with structured clinical ontologies (SNOMED CT, UMLS) could provide complementary evidence of semantic correctness [3].

Hallucination detection threshold selection represents a methodological tradeoff between sensitivity and specificity. While threshold sensitivity analysis demonstrated aggregate metric stability, optimal threshold values for different clinical contexts require empirical determination through validation studies. Domain expert input should inform threshold calibration to align with acceptable risk tolerance for specific clinical applications.

Future framework enhancements could incorporate temporal reasoning assessment, causal relationship extraction evaluation, and clinical reasoning quality metrics. Integration with automated clinical decision support evaluation would enable end-to-end assessment of extraction quality impact on downstream tasks [4]. Longitudinal evaluation across multiple patient encounters could assess consistency and information carryover reliability.

D. Threats to Validity

Internal Validity:

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 04 , April - 2026

The gold standard annotations were produced by a single annotator, which may introduce systematic bias in reference metadata definitions.
Mitigation: Future work will incorporate multiple clinical annotators and quantify inter-rater agreement to improve annotation reliability.

External Validity:

The evaluation dataset consists of simulated doctor–patient conversations, which may not fully capture the complexity, interruptions, and ambiguity present in real-world clinical interactions.
Mitigation: The proposed framework is intentionally designed to be data-source agnostic, and future validation on real clinical transcripts is planned.

Construct Validity:

Embedding-based semantic similarity may not fully capture clinically meaningful equivalence, particularly in cases involving nuanced diagnostic or therapeutic distinctions.
Mitigation: Multiple embedding models were evaluated to assess robustness, and RxNorm-based medication normalization provides an orthogonal validation signal beyond semantic similarity alone.

VII. CONCLUSION

This paper presented a comprehensive semantic evaluation framework for assessing reliability of clinical metadata extraction from doctor-patient conversations. Motivated by requirements emerging during the HealthSync project development, the framework addresses critical gaps in existing evaluation methodologies through multi-dimensional assessment incorporating semantic similarity, format compliance, weighted metrics, and hallucination detection. Application to outputs from four contemporary language models demonstrated the framework's capability to detect subtle reliability issues invisible to traditional metrics.

Key findings demonstrate that conventional evaluation approaches substantially overestimate extraction quality, with semantic analysis revealing agreement characteristics significantly lower than superficial format compliance suggests. The framework successfully identified systematic hallucination patterns, medication specification deficiencies, and semantic drift across multiple models and conversation types. These capabilities provide foundation for rigorous assessment of clinical AI systems prior to deployment.

The work contributes a replicable methodology applicable to clinical metadata extraction evaluation across diverse contexts. While presented as a pilot study with acknowledged limitations, the framework establishes feasibility of semantically-grounded reliability assessment for medical NLP systems. Future work should extend the methodology to larger-scale evaluation with real clinical data, clinical expert validation, and integration with downstream task assessment. As language models increasingly automate clinical documentation workflows, rigorous evaluation frameworks become essential for ensuring reliability and patient safety.

VIII. ACKNOWLEDGEMENT

The authors are very much thankful to management for providing us facilities to conduct the experiments in the institution’s laboratory and colleagues for their support to prepare this paper. We are also thankful to the Science & Technology department for approving our research proposal and providing us grants to conduct research activities.

IX. REFERENCES

[1] C. Sinsky et al., “Allocation of physician time in ambulatory practice: A time and motion study in 4 specialties,” Annals of Internal Medicine, vol. 165, no. 11, pp. 753–760, 2016.

[2] A. Agrawal, J. S. Gans, and A. Goldfarb, “Do we want less automation?” Science, vol. 381, no. 6654, pp. 155–158, 2023.

[3] S. M. Meystre, G. K. Savova, K. C. Kipper-Schuler, and J. F. Hurdle, “Extracting information from textual documents in the electronic health record: a review of recent research,” Yearbook of Medical Informatics, pp. 128–144, 2008.

[4] D. Demner-Fushman, W. W. Chapman, and C. J. McDonald, “What can natural language processing do for clinical decision support?” Journal of Biomedical Informatics, vol. 42, no. 5, pp. 760–772, 2009.

[5] W. W. Chapman, P. M. Nadkarni, L. Hirschman, L. W. D’Avolio, G. K. Savova, and O. Uzuner, “Overcoming barriers to NLP for clinical text: the role of shared tasks and the need for additional creative solutions,” Journal of the American Medical Informatics Association, vol. 18, no. 5, pp. 540–543, 2011.

[6] K. Singhal et al., “Large language models encode clinical knowledge,” Nature, vol. 620, no. 7972, pp. 172–180, 2023.

[7] E. Alsentzer et al., “Publicly available clinical BERT embeddings,” Proceedings of the 2nd Clinical Natural Language Processing Workshop, pp. 72–78, 2019.

[8] M. Wornow et al., “The shaky foundations of large language models and foundation models for electronic health records,” npj Digital Medicine, vol. 6, no. 1, p. 135, 2023.

[9] Kai He et al., Large language models for healthcare: A comprehensive survey, arXiv preprint arXiv:2310.05694, 2023.

[10] J. Maynez, S. Narayan, B. Bohnet, and R. McDonald, “On faithfulness and factuality in abstractive summarization,” in Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL), pp. 1906–1919, 2020.

[11] S. Ji et al., “Survey of hallucination in natural language generation,” ACM Computing Surveys, vol. 55, no. 12, art. 248, 2023.