Conformal Prediction and Verification of Large Language Model Extractions in EHR Data

Edward Kim[1,2], **Richard Foty**[1], **Manil Shrestha**[1,2], **Vicki Seyfert-Margolis**[1]

[1]RespondHealth, Washington, DC, USA
[2]Department of Computer Science, Drexel University, Philadelphia, PA, USA

Abstract

While Electronic Health Records (EHRs) promise comprehensive documentation of patient care, in reality there are significant challenges in data reliability and utilization. EHRs contain vast amounts of unstructured clinical narratives that, despite containing critical and relevant medical information, remain difficult to systematically extract and verify. Recent advances in large language models (LLMs) offer increasingly improving capabilities for extracting structured information from clinical notes, yet these approaches raise fundamental questions about output reliability, over-confident token predictions, and provide no guarantees (statistical or otherwise) for downstream clinical applications.

In this work, we present a conformal verification framework for unstructured EHR data extraction using generative AI. While LLMs have increasingly impressive capabilities, they are notoriously miscalibrated and overconfident in their predictions, necessitating rigorous verification methods to eliminate the need to trust AI models. Our approach (i) employs LLMs to extract medical entities and concepts from clinical narratives with LLM-as-a-judge verification, (ii) implements probabilistic calibration to quantify extraction confidence, and (iii) applies conformal prediction to provide finite-sample guarantees on error rates for accepted extractions. We evaluate our framework on 10k clinical visits across 898 clinical practices utilizing three different EHR systems. Our conformal verification approach can provide assurances that the future expected proportion of accepted but incorrect extractions remains below a pre-specified risk level with rigorous statistical verification. It also maintains formal guarantees over clinical data quality, and illuminates the miscalibrations present in state-of-the-art LLM models, requiring additional validation for safe deployment of automated extraction systems.

Introduction

Machine learning (ML) and natural language processing (NLP) have increasingly become a necessity for analyzing large-scale healthcare data such as electronic health records (EHRs). Yet a persistent barrier to safe clinical adoption is trustworthy and reliable ML, especially in the era of Large Language Models (LLMs) and the issues around hallucinations. In our study, we leverage recent work in reliable uncertainty quantification to aid clinicians in understanding how confident an algorithm is in each prediction. We present a verification system around an emerging area called *Conformal prediction* (CP), which provides clinicians and healthcare analysts with an interpretable, explainable, and model agnostic solution. Instead of a single point prediction (like one might see in typical statistical estimations), a conformal predictor outputs a prediction set (for classification) or an interval (for regression) calibrated to contain the true value with a user specified probability (e.g., 95 %) (Vazquez and Facelli 2022). There is no assumption of normal distributions, rather our framework is distribution free, further providing guarantees that hold under the relaxed i.i.d assumption of exchangeability, regardless of model complexity or data distribution.

In our implementation, we utilize CP to compute a nonconformity score on a held-out calibration set that can be computed with both open source LLMs and several proprietary cloud-based models that output predictive probabilities. We define a binary acceptance problem around entity extraction where the nonconformity scores measured as residual errors, establish a threshold whereby new cases are accepted if their scores fall within a critical distance determined by the calibration data. Cases with high confidence (low nonconformity) are automatically accepted with a user-specified confidence level (e.g., 95%), while uncertain cases exceeding the threshold are flagged for human review.

In summary, **we present a conformal verification framework that establishes rigorous statistical guarantees when utilizing an LLM to read and extract unstructured clinical notes**. We can provide the clinician calibrated confidences around medical concept extraction as well as an interpretable verification score that provides the rationale and justification of the aforementioned extraction. Our framework reduces the need for humans to *trust* the output of any machine learning model (including generative models) as the outputs are formally verified and interpretable. We note, this is particularly valuable in medicine, where erroneous automation can be critical to patient health.

Background and Related Work

There has been a significant body of working looking at confidence scoring and LLM calibration (Geng et al. 2023; Tonolini et al. 2024); however, this prior work on LLM "confidence" largely provides correlates of correctness, rather

Domain	Category	Count	Proportion (%)
Basic statistics	Total visits	10,000	–
	Unique patients	3,012	–
	Unique practices	898	–
	Unique providers	1,824	–
	Average visits processed per patient	3.32	–
Age (yrs)	20–39	7	0.07
	40–49	73	0.73
	50–59	326	3.26
	60–69	1,194	11.94
	70–79	2,452	24.52
	80–89	5,948	59.48
Gender	Female	4,478	44.78
	Male	5,522	55.22
Race	Caucasian	4,260	42.60
	Other	1,003	10.03
	Unknown / Refused	4,737	47.37

Table 1: Cohort demographics, geographic distribution, and clinical content completeness of our dataset.

than viewing confidence in the lens of coverage/risk control on the error rate of the produced outputs. Our work utilizes conformal prediction for generating predictive sets or intervals with guaranteed coverage probabilities (Angelopoulos and Bates 2021). It uses past observations to determine how unusual new observations appear, using a metric called a nonconformity measure. For any new prediction, conformal prediction computes a p-value indicating the proportion of previous examples with nonconformity scores equal to or more extreme than the new example. By choosing a desired confidence level, one can construct prediction sets that provably cover the true outcome with the specified probability.

This framework contains minimal assumptions, robust statistical validity, and wide applicability across tasks. Conformal prediction inherently adapts to data distributions and model uncertainties, providing reliable quantification of uncertainty. This approach is particularly useful when rigorous confidence guarantees are necessary, such as in our specific domain of health related safety-critical applications. Conformal methods have been applied to a variety of medical AI tasks including dermatology imaging (Fayyad, Alijani, and Najjaran 2024), intensive care sepsis prognostication (Yang et al. 2024), and genomic decision support (Papangelou et al. 2025).

In the EHR analysis space, CP for clinical text was used in gaining confidence for code prediction and analysis. Snyder et al (Snyder and Brodsky 2024) fine-tuned a LLaMA 3B model on pathology reports, and achieved 95% accuracy on five common CPT codes. By calibrating a softmax-based nonconformity score on a validation set, they allowed the model to abstain whenever its p-value fell below the desired 95% threshold. This "graded automation" saw the system handle 70 % of reports autonomously at 99.5% precision, while the remaining 30% were routed to human coders.

Genari et al. (Genari and Goedert 2025) used label-conditional inductive CP in an active-learning loop for epidemiological surveillance. At each round, the classifier predicted on unlabeled EHR notes, automatically accepts those whose conformal p-values exceed the confidence level, and queried experts only for the most nonconforming (i.e. uncertain) examples. This strategy achieved competitive performance with 200 labeled notes and provided an order-of-magnitude saving in annotation effort while still delivering per-case error guarantees.

However, even given these encouraging results, open questions remain around scaling to thousands or millions of clinical notes (and codes) (Kim et al. 2024), automating the validation of AI methods, extracting structure with statistical guarantees, and educating clinicians and patients on uncertainty quantification and set-valued predictions. In our work, our goal is all of the above. We first summarize how we build CP into our medical extraction pipeline, show how to automatically validate hundreds of thousands of records, demonstrate conformal prediction on medical extraction, and include a discussion of prevailing challenges and promising research directions.

Patient Data and Cohort Selection

The data used in this study was selected from a database containing more than 34 million unique patients across thousands of clinical practices. For this particular analysis, we were performing specific discovery and analysis focused on a cohort of 3,012 patients with Parkinson's disease (PD) and related movement disorders, identified using ICD-10 codes G20 (Parkinson's disease), G21 (secondary parkinsonism), and G22 (parkinsonism due to other conditions). We note that this cohort is not representative of the general population, and introduces biases toward older adults with more chronic conditions. This cohort also includes increased physician notes and a larger proportion of clinical visits. Table 1 summarises key characteristics of the extracted cohort (N=3,012 patients; and 10,000 encounters).

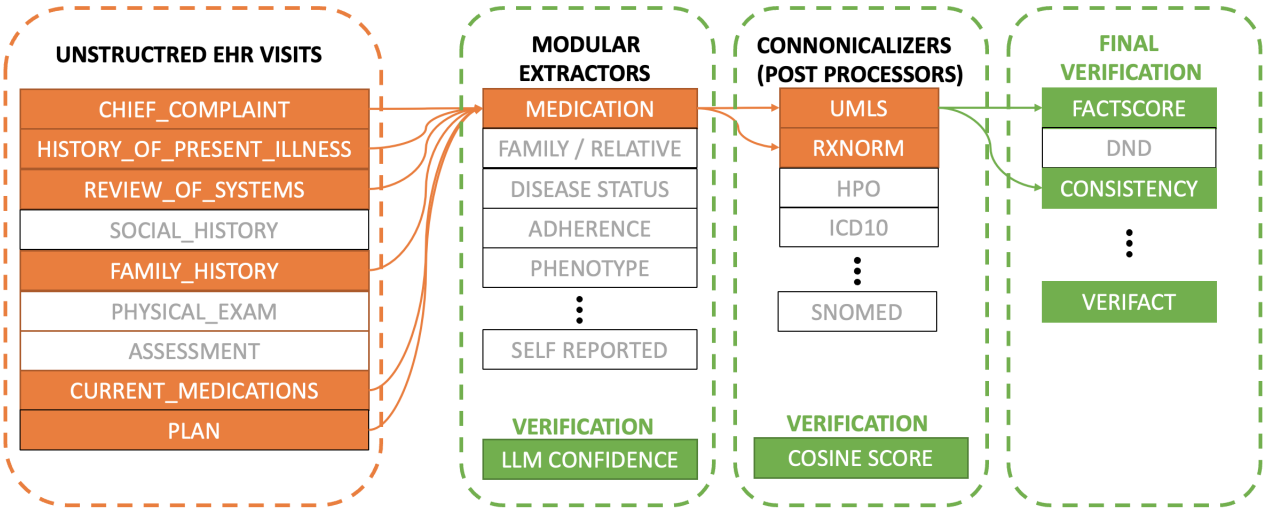

Figure 1: Illustration of the modular extraction, canonicalization, and verification process, highlighting how unstructured clinical notes are processed through modular extraction components, standardized via canonicalization, and subsequently verified against structured EHR data.

Methods

Our large-scale EHR data extraction pipeline leverages generative AI to extract, link, and verify clinical entities from unstructured clinical notes. The following sections outline the key components of the pipeline, including extraction modules, verification processes, and evaluation metrics. See the illustration of our pipeline in Figure 1. In the first step, the user selects which major headers they want to include in the extractor pipeline. In our illustration, we are extracting data around medication;therefore, we want to select unstructured text from the relevant sections of the EHR (Chief complaint, history of present illness, review of systems, family history, current medications and plan). We then extract data using the LLM based upon a key set of entities that the user would specify. In our medication case, we are interested in what medicaitons are mentioned, if the have any adverse effects, if they are starting new medications, or if they are stopping or switching to other medications. These are post processed to standardized mappings (existing ontologies like RxNorm, SNOMED, etc.). The postprocessing and verification steps are similarly modular, with verifiers (e.g., factscore, consistency, etc.) and postprocessors specified in configuration and loaded at runtime. The system provides robust tools for human-in-the-loop verification and result comparison. We have robust verification at every step shown illustrated in green. The verification gives the user visibility and control into the extraction quality.

Scalable Ground Truth Verification

While the overall task is a medical named entity recognition (NER) task, the majority of our focus is this work is towards verification of the underlying ML technology. A well known challenge in EHR data extraction is the lack of reliable ground truth for clinical entities. To address this, we implemented a scalable verification system that combines several automated processes and and human-in-the-loop intervention. Here, we provide an example from our pipeline,

Medical Note: *+upper back pain - felt like a hot pain shooting through back. neck feels fine. + anxiety*

Atomic Statement: *Patient reports upper back pain felt like a hot pain shooting through the back.*
Extracted Entity Name: *pain*
LLM confidence: *0.99998832706*
UMLS Mapping: *C0030193* **Cosine Similarity:** *1.0*
Verification FactScore: 3 *Explanation: The source states "+upper back pain - felt like a hot pain shooting through back," directly supporting this.*

Atomic Statement: *Patient reports anxiety.*
Extracted Entity Name: *anxiety*
LLM confidence: *0.78649887889*
UMLS Mapping: *C0030193* **Cosine Similarity:** *1.0*
Verification FactScore: 3 *Explanation: The current review of systems includes "+ anxiety".*

This example illustrates three of the verification processes put into place, (1) Atomic statement generation with verification FactScores (Min et al. 2023), (2) LLM Confidence and Conformal Prediciton, and (3) Standardization of terms with Cosine Similarity. Our approach decomposes clinical narratives into atomic statements, e.g. discrete, fact-like assertions such as "Patient diagnosed with Parkinson's disease," or "Levodopa prescribed." For each clinical visit, we systematically extract terms and ask the LLM to generate an atomic statement from the extraction, which we then verify from the narrative text. Our fact verification ranges from 0-3 where 0 is not verified, 1 is inferred but not explicitly stated, 2 is partially verified, and 3 is fully verified. The verification score is based on the LLM's confidence in the extraction

Model	Entity Extraction Performance				Human Verification		
	Precision	Recall	F1	Entities	Entities	Correct (%)	Acc. Rating (%)
GPT-4.1	0.951	0.511	0.665	2,796	215	93.5	99.1
LLaMA-4	0.966	0.476	0.638	2,564	228	93.4	96.0
LLaMA-3	0.953	0.469	0.629	2,563	228	94.3	97.8
o4-Mini	0.935	0.365	0.525	2,030	207	97.6	99.0
o3-Mini	0.945	0.223	0.361	1,229	160	95.0	98.7

Table 2: Human verification and entity extraction performance across different language models. Status verification measures factual correctness; Rating verification measures confidence score appropriateness. All models achieve exceptionally high precision with varying recall performance.

and its alignment with the original clinical narrative. Notable patterns in the verification of extractions show that categories such as family history and self-reported symptoms demonstrate higher average verification scores, suggesting more consistent documentation practices, while medication events exhibit lower scores, potentially reflecting gaps in structured data capture or differences in clinical documentation workflows.

Next, we needed to evaluate the reliability of our verification method, both in terms of the LLM used for the pipeline and in terms of human-level performance and agreement. For the supporting experiment, we utilized a 5% subset of the original data (500 clinical visits) from which we evaluated precision and recall across five different language models, as well as human evaluation and annotation on a set of random entities, see Table 2.

These results demonstrate the reliability of our extraction pipeline. Most remarkably, all models achieve outstanding precision scores ranging from 0.935 to 0.966, indicating that when these models extract clinical entities, they are correct over 93% of the time. This high precision is critical for clinical applications where false positives can be costly and potentially dangerous. The precision consistency across diverse model architectures validates the robustness of our extraction methodology. Looking at the correct set of extractions, we can compute the recall for each of the methods. While there is significant overlap between most LLMs, there do not always extract the same entities as reflected by the recall rate.

Human verification is needed to assess the quality of these methods. Our findings confirm the LLM is performing with nearly perfect accuracy. The verification study encompassed 743 total extracted items across all models, revealing status accuracy ranging from 93.3% (LLaMA-4) to 98.0% (GPT-4.1), with an overall average of 95.7%. Even more impressive is the rating accuracy, which measures whether confidence scores appropriately reflect extraction quality. This metric achieved near-perfect performance ranging from 96.0% to 99.1% across all models. This high performance is consistent with recent literature that shows demonstrated 99.8% accuracy in extracting data from radical prostatectomy pathology reports (Azar et al. 2025). And these results provide strong empirical evidence that our LLM-based extraction system identifies clinical entities with exceptional accuracy, and that the verification process aligns closely with human expert judgment. This alignment is crucial for ensuring that automated systems can be trusted to support clinical decision-making.

The last datapoint needed to confirm the validity of our approach is alignment with clinical subject matter experts. We use the VeriFact-BHC dataset benchmark (Chung et al. 2025) for evaluating proposition verification in clinical narratives. It consists of 13,290 proposition statements derived from both human-written and LLM-generated Brief Hospital Course (BHC) summaries for 100 patients from the MIMIC-III Clinical Database (Johnson et al. 2016). Each patient's BHC is paired with their longitudinal EHR, which includes all clinical notes prior to the discharge summary for the admission. Each BHC narrative is decomposed into two types of propositions: sentence-level and atomic claims. Propositions are assessed for validity using formal logic criteria, then annotated by three physicians as Supported, Not Supported, or Not Addressed by the patient's EHR. A majority vote and adjudication process establishes the ground truth for each proposition.

Method	LLM Atomic	LLM Sentences	Human Atomic	Human Sentences
Human (Verifact)	88.5%	84.7%	73.3%	66.6%
Ours (FactScore)	87.3%	87.5%	73.0%	67.6%

Table 3: Inter-clinician agreement on whether propositions are Supported, Not Supported, or Not Addressed by the patient's EHR for all N=13,290. Agreement is shown for LLM-written atomic claims, LLM-written sentences, human-written atomic claims, and human-written sentences.

The results demonstrate that our FactScore verification methodology achieves performance nearly identical to the majority vote adjudication of three expert subject matter experts. For atomic claims generated by LLMs, our approach achieves 87.3% agreement compared to 88.5% for human experts—a difference of only 1.2 percentage points. Similarly, for human-generated atomic claims, our method attains 73.0% agreement versus 73.3% for expert adjudication, representing an even smaller 0.3 percentage point difference. This alignment validates the reliability of our automated verification framework and demonstrates that algorithmic fact-checking can achieve expert-level performance in clinical proposition verification.

Conformal Prediction and Verification

Leveraging the results around the GPT-4.1 FactScore verification and the consistency (nearly identical) nature of the score to human expert performance, we establish that a FactScore of 3 (fully supported) is a very strong indicator of correctness. We can now utilize this verification score to implement a conformal prediction framework that can provide future finite-sample guarantees on the correctness of our extractions. The following sections outline the conformal prediction methodology, including token-level confidence aggregation, calibration confidence scores, and category-specific conformal thresholds.

Token-level confidence - Let an extracted span e consist of the ordered token sequence $\boldsymbol{w}_e = (w_{e,1}, \ldots, w_{e,m_e})$ with decoder logits $\boldsymbol{\ell}_{e,t} \in \mathbb{R}^{|\mathcal{V}|}$ for token $w_{e,t}$. The model's softmax probability for the realized token is

$$p_{e,t} \;=\; \frac{\exp\{\ell_{e,t}(w_{e,t})\}}{\sum_{v\in\mathcal{V}} \exp\{\ell_{e,t}(v)\}} \;\in [0,1]. \tag{1}$$

We aggregate the m_e token confidences into a span-level score

$$\hat{p}_e \;=\; \Big(\prod_{t=1}^{m_e} p_{e,t}\Big)^{1/m_e}. \tag{2}$$

Each realized span is labeled $y_e \in \{0,1\}$ (1 = correct using FactScore >= 3). Define the *non-conformity score*

$$s_e \;=\; -\log\left(\frac{\hat{p}_e}{1-\hat{p}_e}\right) = -\text{logit}(\hat{p}_e). \tag{3}$$

Split conformal calibration - For every extraction category c we partition the verified dataset using a 70/30 split, with 70% designated as the training set and 30% as the *calibration* set $\mathcal{D}_c^{\text{cal}} = \{(s_e, y_e)\}_{e=1}^{N_c^{\text{cal}}}$. This split conformal approach ensures proper separation between model training and threshold calibration.

Category-specific conformal threshold - Let $n_c = |\mathcal{D}_c^{\text{cal}}|$ and order the calibration non-conformity scores $\{s_{(1)}, \ldots, s_{(n_c)}\}$ from smallest to largest. For a desired risk level $\alpha \in (0,1)$ (e.g. $\alpha = 0.05$ for 95 % confidence) define

$$\tau_c \;=\; s_{\left(\lceil (1-\alpha)(n_c+1)\rceil\right)}. \tag{4}$$

Decision rule at deployment - For a new extraction in category c with score s:

$$\text{accept} \iff s \le \tau_c, \qquad \text{otherwise manually review.} \tag{5}$$

Finite-sample Guarantees Using the split conformal prediction for calibration, our framework ensures

$$\Pr\big(y = 0 \,\wedge\, s \le \tau_c\big) \le \alpha, \tag{6}$$

$$\text{equivalently} \quad \mathbb{E}\Big[\frac{1}{|A_c|}\sum_{e\in A_c} \mathbf{1}(y_e = 0)\Big] \le \alpha, \tag{7}$$

where $A_c = \{e : s_e \le \tau_c\}$ is the *accepted* set. Equation (6) holds *simultaneously* for every category c under the exchangeability assumption between calibration and future deployment data. In our experiments α is fixed globally (0.05), but τ_c varies markedly (Table 4), reflecting heterogeneity in extractor calibration. Categories with excellent calibration (e.g. family history) naturally yield $\tau_c \approx 0$, admitting almost all extractions, whereas challenging categories (e.g. phenotypes) impose a more conservative threshold.

Calibration of Model and Reported Confidence

Hypothetically all transformer-based LLMs with modern day generative capabilities accompany every output with a next-token probability. Although seemingly straightforward to use as a measure of trustworthiness, these raw soft–max scores are rarely in perfect alignment with real-world, semantic level confidence. To diagnose and correct this mismatch we produced calibration curves for each extraction module (Figure 2). The plot shows the x–axis records the model's predicted probability, while the y–axis records the empirical likelihood of the extraction being correct; a perfectly calibrated system would trace the 45° diagonal. Inspection of GPT-4.1's curves immediately reveals heterogeneity in over/under confident modeling. Family-history extractions closely match the diagonal, signaling near-ideal calibration. Self-reported symptoms, by contrast, fall far below the line, revealing a tendency toward over-confidence. *A notable observation is that we see that calibration is not a single global property of "the model"; but rather, depends on* what *the model is being asked to extract.*

Split Conformal Prediction Analysis

To provide statistical reliability guarantees for our extraction pipeline, we applied standard split conformal prediction with $\alpha = 0.05$ (targeting 95% coverage) to the 62,981-entity test corpus (via 70/30 split). The analysis reveals the typical overconfident LLM miscalibration across most categories, with conformal thresholds requiring high confidence levels to ensure the high coverage requirement is met. The results are summarized in Table 4.

The conformal analysis exposes the miscalibration of generative AI alignment; all categories converge to the maximum threshold of 6.9068 log-odds (equivalent to 99.9% probability) under standard conformal prediction, except phenotypes which requires a slightly lower but still extreme threshold of 3.2131 log-odds (96.1% probability). This convergence indicates reinforces known knowledge that the models are poorly calibrated (Achiam et al. 2023), sychophantic (Laban et al. 2023), and overconfident (Daniali and Kim 2023). Due to this fact, only predictions with near-certainty can satisfy the 95% coverage requirement. The flexible conformal approach provides some relief for the most miscalibrated categories: self-reported symptoms drops to a threshold of 4.325 log-odds, and disease status to 5.455 log-odds, though these remain extremely high confidence requirements.

The calibration metrics reveal distinct patterns: family history demonstrates the best overall calibration with low Brier score (0.091) and ECE (0.057), while self-reported symptoms exhibits the worst miscalibration (Brier = 0.404, ECE = 0.353). The FactScore metric further confirms this hierarchy: family history (2.87) and disease status (2.85) achieve the highest scores, while self-reported symptoms

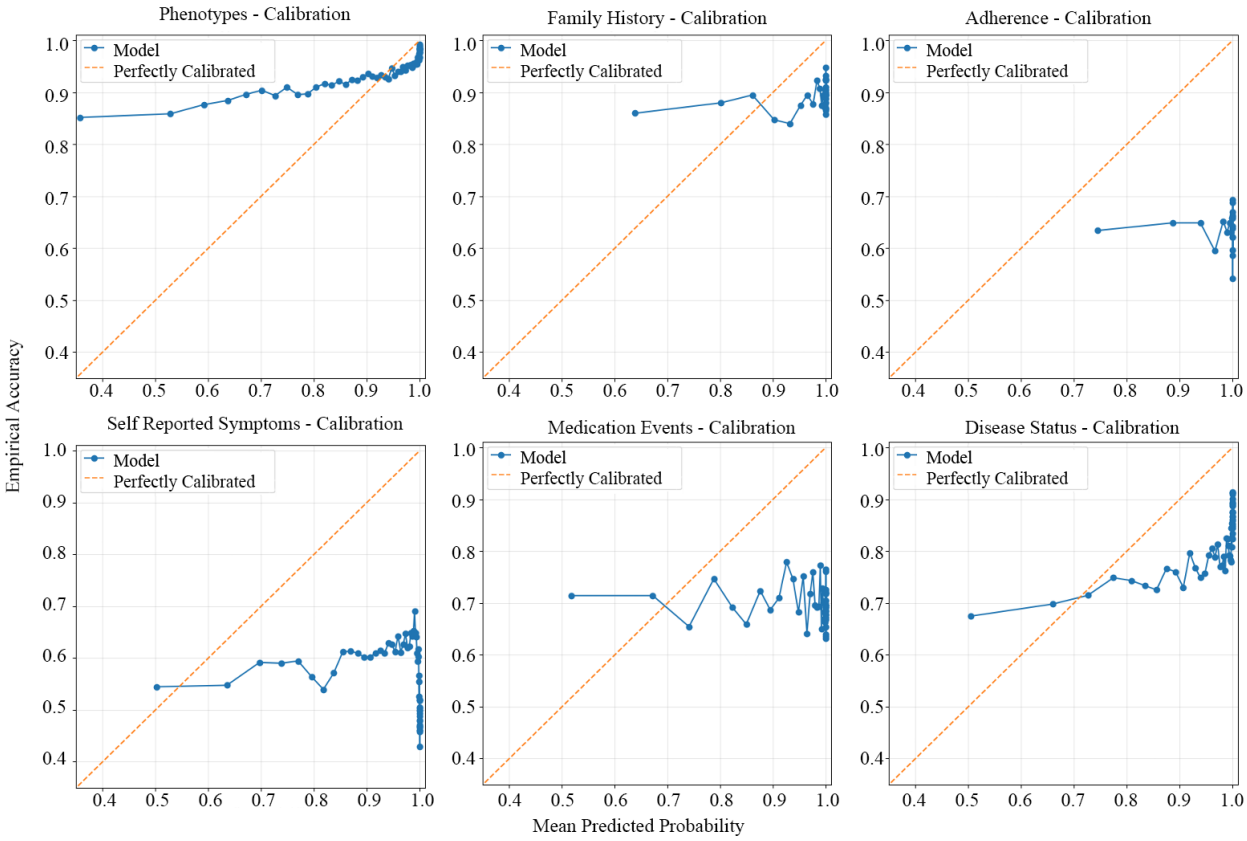

Figure 2: Calibration curves by extraction category using OpenAI's GPT-4.1 model. Each curve shows the relationship between predicted confidence scores (x-axis) and observed accuracy (y-axis) for different clinical information types. The diagonal line represents perfect calibration. We can see that the calibration is highly dependent on what kind of extraction the model is asked to perform.

(1.98) shows the poorest performance. Standard conformal prediction with $\alpha = 0.05$ yields rejection rates that render the system very selective: phenotypes reject 73.3% of entities, self-reported symptoms reject 63.1%, and disease status rejects 54.5%. This means to hit the 95% coverage requirement, the system needs to be highly conservative and only the most confident token extractions can meet the threshold.

A more flexible conformal approach provides modest improvements for the most severely miscalibrated categories. Self-reported symptoms sees its rejection rate drop from 63.1% to 45.7% with a reduced threshold (4.325 vs 6.907 log-odds), while disease status improves from 54.5% to 46.0% rejection. However, these improvements come at the cost of slightly reduced coverage guarantees. Even with flexible thresholds, rejection rates remain high necessitating complimentary verification steps to ensure clinical safety (such as the FactScore verification described earlier, human review, or other additional verification methods).

Coverage Analysis and Strategic Quality Assurance

The coverage analysis demonstrates the value of conformal prediction in providing transparent reliability assessments for LLM-based extraction systems. While phenotypes achieves strong 98.3% coverage with its 96.1% probability threshold, this comes with the trade-off of requiring human review for nearly three-quarters of predictions. This is a clear indication that blind trust in LLM outputs would be inappropriate.

The varying coverage rates across categories provide crucial insights into where these models excel and where additional verification mechanisms are essential. Conformal analysis successfully identifies the boundaries of reliable automation, enabling principled decisions about where to deploy automated processing versus human oversight This analysis reveals the strategic value of our multi-layered approach. Standard conformal prediction with 95% coverage requirements identifies tens of thousands of "highly confident" entities (63.7% of the corpus) requiring additional verification—precisely where our FactScore methodology, structured data cross-validation, and human-in-the-loop processes become essential. The framework transforms an "all-or-nothing" trust paradigm into a nuanced quality assurance pipeline where statistical guarantees protect against silent failures while complementary verification methods handle uncertain cases.

Category	# Ent.	Brier	ECE	FactS	Std. τ	Flex. τ	Rej. %	Cov.
Phenotypes	36,900	0.065	0.064	2.92	3.213	3.213	73.3 / 44.5	98.3 / 96.8
Family history	3,156	0.091	0.057	2.87	6.907	6.907	26.0 / 26.0	91.8 / 91.8
Adherence	2,001	0.355	0.124	2.54	6.907	6.907	16.0 / 16.0	63.5 / 63.5
Medication	3,341	0.295	0.228	2.47	6.907	6.907	47.4 / 47.4	68.1 / 68.1
Self-reported	12,570	0.404	0.353	1.98	6.907	4.325	63.1 / 45.7	48.6 / 52.7
Disease status	5,013	0.163	0.130	2.85	6.907	5.455	54.5 / 46.0	87.5 / 87.0
All categories	62,981	0.156	0.133	2.69	–	–	63.7 / 42.6	– / –

Table 4: Calibration metrics and conformal prediction analysis comparing standard ($\alpha = 0.05$) and flexible approaches for the 62,981-entity test corpus across the 70/30 calibration split.

For real-world clinical deployment, this represents a substantial advancement over current practice. Rather than requiring clinicians to trust opaque AI systems or manually review all extractions, our approach provides: (1) automated processing with statistical guarantees for high-confidence extractions, (2) transparent identification of cases requiring verification, and (3) structured workflows for handling uncertain predictions. This enables scalable, safe deployment while maintaining clinical oversight where it matters most.

Conclusion

EHRs offer rich but complex data. Reducing documentation burden, improving integration of structured and unstructured data, and leveraging NLP are key to maximizing EHR value for care and research. We present a conformal prediction framework that provides statistical guarantees on the reliability of LLM-based extractions from unstructured clinical narratives. Our approach combines probabilistic calibration, nonconformity scoring, and finite-sample guarantees to ensure that the expected proportion of incorrect extractions remains below a pre-specified risk level. This enables safe deployment of automated extraction systems while maintaining rigorous guarantees over clinical data quality. We demonstrate that scalable LLM extraction can be both performant and certifiably safe. Calibration, when treated as a domain-specific property and reinforced with conformal filtering, turns raw probability scores into actionable guarantees. In practical terms, the method protects downstream users from silent model errors without sacrificing scale: almost everything the system produces is delivered automatically, and what remains is accompanied by an explicit warning that its correctness cannot yet be vouched for. This ability to couple breadth of coverage with a formal error bound brings clinical-note extraction a decisive step closer to routine, trustworthy deployment.

References

Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F. L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Angelopoulos, A. N.; and Bates, S. 2021. Gentle Introduction to Conformal Prediction and Distribution-Free Uncertainty Quantification. *arXiv preprint arXiv:2107.07511*.

Azar, W. S.; Junkin, D. M.; Hesswani, C.; Koller, C. R.; Parikh, S. H.; Schuppe, K. C.; Williams, N.; Nethala, D.; Mendhiratta, N.; Kenigsberg, A. P.; et al. 2025. LLM-mediated data extraction from patient records after radical prostatectomy. *NEJM AI*, 2(6): AIcs2400943.

Chung, P.; Swaminathan, A.; Goodell, A. J.; Kim, Y.; Reincke, S. M.; Han, L.; Deverett, B.; Sadeghi, M. A.; Ariss, A.-B.; Ghanem, M.; Seong, D.; Lee, A. A.; Coombes, C. E.; Bradshaw, B.; Sufian, M. A.; Hong, H. J.; Nguyen, T. P.; Rasouli, M. R.; Kamra, K.; Burbridge, M. A.; McAvoy, J. C.; Saffary, R.; Ma, S. P.; Dash, D.; Xie, J.; Wang, E. Y.; Schmiesing, C. A.; Shah, N.; and Aghaeepour, N. 2025. VeriFact: Verifying Facts in LLM-Generated Clinical Text with Electronic Health Records.

Daniali, M.; and Kim, E. 2023. Perception over time: Temporal dynamics for robust image understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 5656–5665.

Fayyad, J.; Alijani, S.; and Najjaran, H. 2024. Empirical validation of conformal prediction for trustworthy skin lesions classification. *Computer Methods and Programs in Biomedicine*, 253: 108231.

Genari, J.; and Goedert, G. T. 2025. Mining Unstructured Medical Texts With Conformal Active Learning. *arXiv preprint arXiv:2502.04372*.

Geng, J.; Cai, F.; Wang, Y.; Koeppl, H.; Nakov, P.; and Gurevych, I. 2023. A survey of confidence estimation and calibration in large language models. *arXiv preprint arXiv:2311.08298*.

Johnson, A. E.; Pollard, T. J.; Shen, L.; Lehman, L.-w. H.; Feng, M.; Ghassemi, M.; Moody, B.; Szolovits, P.; Anthony Celi, L.; and Mark, R. G. 2016. MIMIC-III, a freely accessible critical care database. *Scientific data*, 3(1): 1–9.

Kim, E.; Shrestha, M.; Foty, R.; DeLay, T.; and Seyfert-Margolis, V. 2024. Structured Extraction of Real World Medical Knowledge using LLMs for Summarization and Search. In *2024 IEEE International Conference on Big Data (BigData)*, 3421–3430. IEEE.

Laban, P.; Murakhovs' ka, L.; Xiong, C.; and Wu, C.-S. 2023. Are you sure? challenging llms leads to performance drops in the flipflop experiment. *arXiv preprint arXiv:2311.08596*.

Min, S.; Krishna, K.; Lyu, X.; Lewis, M.; tau Yih, W.; Koh, P. W.; Iyyer, M.; Zettlemoyer, L.; and Hajishirzi, H. 2023.

FActScore: Fine-grained Atomic Evaluation of Factual Precision in Long Form Text Generation. arXiv:2305.14251.

Papangelou, C.; Kyriakidis, K.; Natsiavas, P.; Chouvarda, I.; and Malousi, A. 2025. Reliable machine learning models in genomic medicine using conformal prediction. *Frontiers in Bioinformatics*, 5: 1507448.

Snyder, C.; and Brodsky, V. 2024. Conformal Prediction and Large Language Models for Medical Coding. In *American Journal of Clinical Pathology*, volume 162. Oxford Univ Press Inc Journals Dept, NC 27513.

Tonolini, F.; Aletras, N.; Massiah, J.; and Kazai, G. 2024. Bayesian prompt ensembles: Model uncertainty estimation for black-box large language models. In *Findings of the Association for Computational Linguistics ACL 2024*, 12229–12272.

Vazquez, J.; and Facelli, J. C. 2022. Conformal Prediction in Clinical Medical Sciences. *Journal of Healthcare Informatics Research*, 6: 241–252.

Yang, M.; Chen, H.; Hu, W.; Mischi, M.; Shan, C.; Li, J.; Long, X.; and Liu, C. 2024. Development and Validation of an Interpretable Conformal Predictor to Predict Sepsis Mortality Risk: Retrospective Cohort Study. *Journal of Medical Internet Research*, 26: e50369.