IntuitionLabs

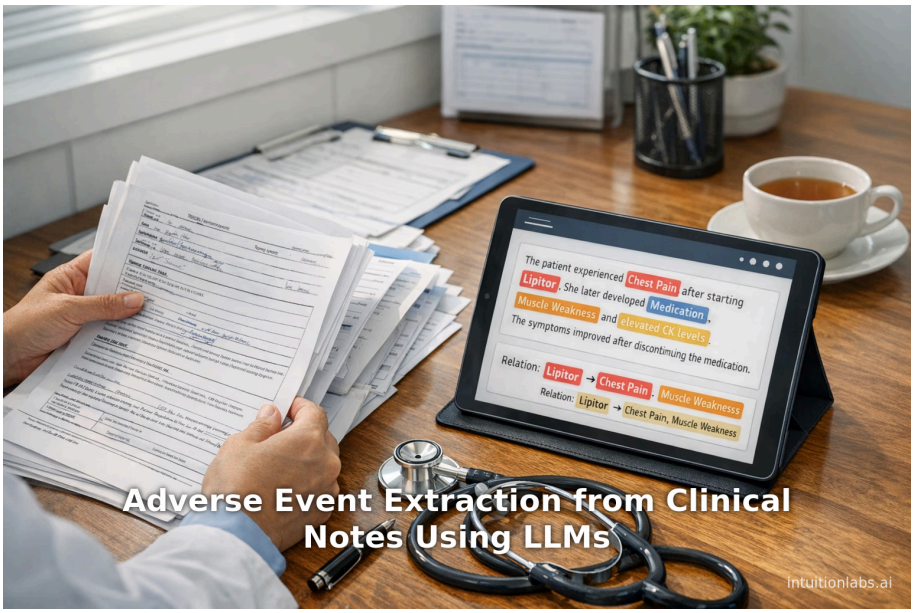
INTUITIONLABS / RESEARCH & PRACTICAL GUIDANCE

Adverse Event Extraction from Clinical Notes Using LLMs

> Build practical AI for pharma and biotech with IntuitionLabs. We help life-science teams turn complex information and workflows into useful software, governed knowledge systems and AI tools.

Published by IntuitionLabs · 2026-04-11 · llm · adverse event extraction · clinical notes · pharmacovigilance · healthcare nlp · adverse drug events · medical nlp

Original article: https://intuitionlabs.ai/articles/llm-adverse-event-extraction-clinical-notes



In this article

1. Executive Summary
2. Introduction
3. LLMs and Clinical Information Extraction
4. Transformer-based LLM Architectures
5. Challenges of Clinical Text
6. Prior Methods Compared
7. LLM-based ADE Extraction: Approaches
8. Prompting and Zero-Shot LLMs
9. Fine-Tuning and Continuous Learning
10. Data Augmentation and Distillation
11. Pipeline Architectures
12. Accuracy and Performance
13. Standard Metrics
14. Benchmarks on Public Datasets
15. Case Study: Strata (“Human-Level IE”)
16. Zero-Shot ChatGPT Performance
17. Benchmarks with Prompt Enhancements
18. Comparative Results
19. Implementation in Clinical Settings
20. Data Privacy and Security
21. Compute Resources
22. Integration with Clinical Workflow
23. Example Workflow
24. Benefits and Current Use-Cases
25. Discussion and Future Directions
26. Accuracy Considerations
27. Implementation Challenges
28. Future Prospects
29. Case Studies and Examples
30. Conclusion
31. References

Executive Summary

Automated extraction of adverse events (AEs) from electronic health record (EHR) clinical notes has enormous potential to enhance patient safety and pharmacovigilance. Large language models (LLMs) – transformer-based deep neural networks such as GPT, BERT, and their biomedical adaptations – have shown remarkable capacity for understanding and generating text. Recent studies demonstrate that instruction- or task-tuned LLMs can **outperform** earlier NLP models on clinical information extraction tasks, including

adverse event detection, albeit with substantial computational cost (pmc.ncbi.nlm.nih.gov) (www.sciencedirect.com). For example, Hu *et al.* (2026) report that fine-tuned LLaMA-3 models improved named-entity and relation-extraction F1 scores by several points over a strong BERT baseline (e.g. +7% F1 in limited-data settings) (pmc.ncbi.nlm.nih.gov). Similarly, Liu *et al.* (2025) achieved human-level accuracy (~90% exact-match) extracting structured clinical data using a LoRA-fine-tuned LLaMA-3.1 model (www.nature.com). These gains come with trade-offs: in Hu *et al.*'s evaluation, BERT processed notes 5–28 times faster than the LLaMA models during inference (pmc.ncbi.nlm.nih.gov), and models may lack medical domain knowledge without further adaptation.

This report reviews the current state of LLM-based adverse event extraction from clinical narratives, with a focus on **accuracy** and **implementation**. We cover the **background** of AE extraction (definitions, clinical importance, traditional methods), survey LLMs and their biomedical variants, and analyze how LLMs have been applied to AE-related tasks (e.g. named-entity recognition of side effects, relation extraction of drug–event pairs, severity grading). We synthesize published evidence on model performance (precision/recall/F1) and workflows (zero-/few-shot vs fine-tuning, pipelines vs end-to-end) from multiple perspectives (academic studies, systematic reviews, case reports). For instance, systematic overviews show that GPT-based models can offer zero-shot extraction competitive with supervised systems (pubmed.ncbi.nlm.nih.gov) (pmc.ncbi.nlm.nih.gov), especially when enriched by medical context prompts (www.sciencedirect.com) (pubmed.ncbi.nlm.nih.gov). We also present case studies of real-world systems and research prototypes (such as **Strata** for radiology/pathology report structuring (www.nature.com) and **PheNormGPT** for phenotype normalization (pmc.ncbi.nlm.nih.gov)). The report discusses implementation issues in healthcare settings: prompt engineering, de-identification and HIPAA compliance, integration with EHRs, computational resources, interpretability, and regulatory considerations. Finally, we highlight open challenges (e.g. hallucination, domain bias, relation extraction of complex events) and future directions (specialized biomedical LLMs, few-shot learning, real-time monitoring, federated learning) to guide researchers and practitioners.

Key findings include:

- **LLM Performance:** ADE (adverse drug event) results must be separated from aggregate medication-extraction scores. In the original n2c2 challenge, the best ADE–Drug end-to-end F1 was 0.4755 (original challenge report).
- **Data and Tasks:** AE extraction is typically framed as entity (ADE mention) and relation (Drug→ADE) identification from narrative text. Standard corpora include **n2c2-2018** (505 MIMIC discharge notes), **MADE 1.0** (1089 notes), TAC 2017, and others (pmc.ncbi.nlm.nih.gov) (www.sciencedirect.com). For example, the best ADE–Drug end-to-end F1 in the original n2c2 challenge was 0.4755 (original challenge report). Table 1 summarizes major AE extraction datasets and tasks.
- **LLM Approaches:** Two main paradigms have emerged. (1) **Fine-tuning:** LLMs pre-trained on large corpora are further trained on annotated clinical text. Recent work shows domain adaptation (e.g. further pre-training on clinical notes or drug labels) notably boosts performance (www.sciencedirect.com) (pmc.ncbi.nlm.nih.gov). (2) **Prompt-based IE:** LLMs (GPT-3/4, Llama, etc.) are used zero- or few-shot with natural-language prompts to extract or label events. Zero-shot ChatGPT/ GPT-4 can produce surprisingly good results on some simple extraction tasks (www.sciencedirect.com) (pubmed.ncbi.nlm.nih.gov), though it can struggle with specialized medical terminology or relations

(pmc.ncbi.nlm.nih.gov). Hybrid methods combine prompts with small supervised models or post-processing to refine outputs. Novel techniques include generating synthetic annotations via LLMs (knowledge distillation) or using LLMs as annotators to reduce expert work (www.sciencedirect.com) (www.mdpi.com).

- **Accuracy & Generalizability:** When ample labeled data are available, traditional supervised deep models (e.g. BioBERT variants) already perform exceptionally on ADE NER (pmc.ncbi.nlm.nih.gov). LLMs tend to shine when data are scarce or more complex reasoning is needed. Hu *et al.* found that relative gains from LLMs are largest in low-resource or cross-domain settings (pmc.ncbi.nlm.nih.gov).Nevertheless, even GPT-4 and similar models can miss or confuse events without careful prompting; studies note lower recall for infrequent ADEs and error propagation issues (pmc.ncbi.nlm.nih.gov). Overall, best LLM-based pipelines report F1 scores in the 0.7–0.9 range for core extraction, with relation-task F1 often lagging behind NER (pmc.ncbi.nlm.nih.gov).
- **Implementation:** Integrating LLMs into clinical workflows requires handling privacy (HIPAA) and computation. Approaches range from on-premise local models (e.g. fine-tuned Llama on hospital GPUs (www.nature.com)) to cloud APIs (e.g. GPT-4 via secure handlers). Data preprocessing (de-identification, note segmentation) often precedes LLM input; notably, preprocessing with GPT-4 (to fix spelling/grammar or expand abbreviations) has been shown to significantly improve downstream concept extraction (www.mdpi.com). Tools like *Strata* enable low-code LLM fine-tuning on institutional data (www.nature.com). Key deployments include pharmacovigilance automation (Pfizer's pilot on AE case reports (pmc.ncbi.nlm.nih.gov)), oncology adverse event surveillance, and real-time monitoring systems.
- **Implications & Future:** The promise of LLMs is to accelerate and scale adverse event monitoring, potentially enabling earlier detection of safety signals. However, risks include hallucination (false positive events), domain bias, and maintenance overhead. Regulatory acceptance (e.g. FDA guidance) and physician trust will hinge on rigorous validation, transparency, and safeguards. Future research should explore unsupervised or weakly supervised LLM training on multi-institutional EHRs, hybrid human-AI workflows, and specialized biomedical LLMs (e.g. BioGPT, Med-PaLM). With careful implementation, LLMs can substantially reduce manual chart review and improve patient safety outcomes (pmc.ncbi.nlm.nih.gov).

Introduction

Adverse events (AEs) – any unfavorable medical occurrences associated with healthcare interventions – are a major concern in medicine and pharmacovigilance. Adverse *drug* events (ADEs) in particular (e.g. medication side effects, drug interactions) are implicated in high morbidity and cost. A recent review notes that ADEs significantly impact patient safety and healthcare outcomes, prompting a need for better monitoring (pmc.ncbi.nlm.nih.gov). Clinical narratives in EHRs contain a wealth of information on AEs (symptoms, lab abnormalities, complications) that are often **not** captured in structured fields (pmc.ncbi.nlm.nih.gov). Manually identifying AEs from free-text notes is labor-intensive and error-prone: it requires reading detailed progress

notes, discharge summaries, and other documentation. For example, Henry *et al.* reported that in one shared task, even top systems struggled to identify AEs and their causes with complete accuracy, reflecting the inherent complexity (pmc.ncbi.nlm.nih.gov).

The **motivation** for automated AE extraction is clear: timely and accurate identification of AEs can enable better clinical decision support, prevent medication errors, and enhance pharmacovigilance (post-market safety surveillance). Regulatory agencies like the FDA and EMA emphasize the importance of real-world evidence on drug safety, much of which resides in clinical notes. As Zhan *et al.* remark, "LLMs (e.g. BERT, GPT) offer promising methods for automating ADE extraction from clinical data," enabling large-scale pharmacovigilance and decision support (pmc.ncbi.nlm.nih.gov). Early implementations have already shown that AI can feasibly support AE case processing. In a pilot, Schmider *et al.* demonstrated feasibility of AI-assisted extraction from adverse-event source documents and evaluation of case validity (original pilot). These successes precede the LLM era, but set the stage for current developments.

Over the past decade, **NLP methods** for AE extraction have evolved dramatically. Early work relied on rule-based and dictionary approaches. Workflows commonly used controlled vocabularies (e.g. MedDRA terms for adverse effects) and simple pattern matching. For instance, even by 2012 some systems combined lexical patterns with part-of-speech tagging to flag ADE mentions in notes (www.sciencedirect.com). However, these rule-based systems had limited recall and did not scale well to the varied language of clinical notes. With the rise of machine learning, researchers shifted to statistical models. Traditional supervised models (e.g. conditional random fields, SVMs) trained on annotated corpora (such as the n2c2 ADE dataset (www.sciencedirect.com)) achieved moderate accuracy.

The breakthrough came with **deep learning**. Recurrent neural networks (LSTMs, CNNs) and neural language models began to capture context and semantics in text. In 2018–2020, transformer models like BERT and its biomedical variants (BioBERT, ClinicalBERT, etc.) emerged. These models, pre-trained on massive corpora and fine-tuned on clinical tasks, quickly overtook older methods. For example, on the n2c2-2018 ADE extraction task, BERT-based systems reported F1 scores in the low 90s (pmc.ncbi.nlm.nih.gov), greatly exceeding earlier baselines. Nonetheless, even these models required substantial annotated data. Limited data or out-of-domain text (e.g. uncommon ADEs, rare drug names) still posed challenges. Researchers responded by creating more annotated datasets (MADE 1.0 (www.sciencedirect.com), TAC 2017 ADE track, etc.) and employing multi-task strategies. The current state before LLMs was that deep models routinely reached ~0.9 F1 on NER (entity) tasks, and somewhat lower on relation tasks; but systems could fail on low-frequency events or long-range dependencies (www.sciencedirect.com).

LLMs – notably the GPT and LLaMA families – now offer a new paradigm. These are typically *decoder-only* (e.g. GPT, LLaMA) or *encoder-decoder* (e.g. T5) transformer models pre-trained on very large general-domain corpora, then optionally fine-tuned on specific tasks (www.sciencedirect.com). Their sheer scale (billions of parameters) and instruction tuning grant them impressive few-shot or zero-shot capabilities. In the general domain, GPT-4 can follow natural language prompts to perform tasks without additional training. In biomedical NLP, specialized versions like BioGPT, PubMedBERT, and Galactica have been developed. Early evidence suggests that LLMs can reduce the need for extensive supervised training by leveraging prior knowledge. For example, **Hu et al. (2026)** compared instruction-tuned LLaMA-2/3 models to BERT across four clinical corpora (including MIMIC and i2b2 notes). They found that LLaMA models *consistently* outperformed BERT: on limited-

data settings, LLaMA-3-70B gained over 7% absolute F1 in NER and ~4% in relation extraction (pmc.ncbi.nlm.nih.gov). In Hu *et al.*'s evaluation, BERT processed notes 5–28 times faster than the LLaMA models during inference (original study).

In sum, advancements in NLP – culminating in LLMs – have rapidly improved AE extraction accuracy. But implementation in clinics remains nontrivial. This report examines **both** the accuracy of LLM-based extraction (drawing on published benchmarks, case studies, and surveys) and the practical implementation considerations (data handling, privacy, integration, cost). We compare LLM approaches to prior methods, and explore implications for healthcare workflows and regulation.

LLMs and Clinical Information Extraction

Transformer-based LLM Architectures

Modern LLMs are grounded in the *Transformer* architecture. In brief, Transformers use self-attention mechanisms to model relationships between all tokens in a sequence (www.sciencedirect.com). This contrasts with earlier sequence models that processed text sequentially (e.g. RNNs). LLMs come in several flavors: **decoder-only** models (e.g. GPT-2/3/4, LLaMA) generate text autoregressively from left to right, making them very flexible for generation tasks. **Encoder-only** models (e.g. BERT, RoBERTa) read the text bidirectionally and are optimized for classification or NLU tasks. **Encoder-decoder** models (e.g. T5, BART) combine both for tasks like text summarization or translation.

In the biomedical context, many LLMs are adapted in one of two ways: either *fine-tuned* on clinical data, or *prompted*. For example, *BioBERT* is an encoder-only BERT model further pre-trained on biomedical literature, then fine-tuned on tasks (pmc.ncbi.nlm.nih.gov). *ClinicalBERT* similarly adds clinical notes in pre-training. On the generative side, models like *BioGPT* and *PubMedLLaMA* have been pretrained or fine-tuned on biomedical text. These domain adaptations can be crucial: as noted in reviews, general LLMs may lack clinical vocabulary and nuance (www.sciencedirect.com) (pmc.ncbi.nlm.nih.gov). In practice, hybrid strategies are common: e.g. initial prompting with a medical ontology, followed by an LLM to refine extraction.

Challenges of Clinical Text

Clinical notes pose special challenges for NLP that any extraction system must handle. Notes often contain abbreviations, typos, and non-standard phrasing. Acronyms (e.g. “HTN” for hypertension) and medical jargon are pervasive. Data privacy is paramount: notes may include Protected Health Information (PHI). For HIPAA requirements on cloud processing, risk analysis, and safeguards for both cloud and local hosting, see Data Privacy and Security below. A business associate agreement (BAA) must require the business associate to safeguard electronic PHI, including implementing Security Rule requirements (HHS cloud guidance). Moreover, the *context* of AEs can span multiple sentences; an event might be implied (e.g. “Patient denies any chest pain”, which indicates absence) or scattered (the cause and effect might be separated). Traditional NER models treat each sentence independently, but LLMs can attend over entire documents, potentially capturing such long-range cues.

Another challenge is **data scarcity** of annotated corpora. Manual labeling of ADE mentions or relations is costly. Public corpora exist (n2c2/ADE 2018, MADE, TAC) but are relatively small (hundreds of documents) (pmc.ncbi.nlm.nih.gov) (www.sciencedirect.com). LLMs provide a partial remedy by enabling zero- or few-shot learning: one can prompt the model with a few examples or a detailed instruction rather than thousands of labeled instances. However, as summarized studies show, LLMs can sometimes hallucinate nonexistent events or misunderstand rare terms (pmc.ncbi.nlm.nih.gov). Thus current research often combines LLMs with human oversight or post-processing to ensure clinical validity.

Prior Methods Compared

Before assessing LLMs, it is informative to view the landscape of earlier AE extraction techniques. Broadly, methods have included **rule-based** and **statistical learning** approaches:

- **Rule-based/hybrid:** These use medical dictionaries (MedDRA, SNOMED CT) and pattern rules. For instance, Chapman *et al.* (2011) used hand-crafted rules and gazetteers to flag drug-event pairs in discharge summaries (www.sciencedirect.com). Such systems can achieve precision in narrow domains but often miss novel phrasing or context. A hybrid approach employed in some pharmacovigilance tools uses pattern matching followed by a small ML model to filter results.
- **Machine Learning (non-Deep):** Classic ML (Naive Bayes, CRF, SVM) trained on features (keywords, POS tags, embeddings) was explored in the 2010s. For example, Dandala *et al.* (2017) used CRF models with word embeddings to jointly identify ADE entities and their relations (www.sciencedirect.com). These methods yielded moderate F1 scores (~80% on some tasks) but their reliance on feature engineering was a limitation.
- **Deep Learning:** CNNs and RNNs rose to prominence ~2015–2018. For example, Li *et al.* (2018) applied bidirectional LSTMs for ADE named-entity recognition (NER) in clinical narratives. These models benefited from pre-trained embeddings (e.g. word2vec on MIMIC data) and attention mechanisms. However, a major leap came with BERT (2018): release of BioBERT (Lee *et al.*, 2020) showed state-of-the-art results on a variety of biomedical NLP tasks (pmc.ncbi.nlm.nih.gov). On the n2c2 ADE task, the best pre-LLM systems (using BERT or CNN+CRF architectures) reported entity F1s in the 90s and relation F1s around 50–90 (pmc.ncbi.nlm.nih.gov) (www.sciencedirect.com). These served as baselines for LLM comparisons.

Table 1 summarizes representative datasets and tasks in ADE extraction. (The **three most widely used benchmarks** are n2c2 2018 ADE, MADE 1.0, and I2B2 Medications). These datasets typically annotate drug mentions, ADE mentions, and drug–ADE relations. Multi-institutional datasets like **MIMIC-III** (critical care EHRs) and various coding-based corpora have also been used for training and evaluation (www.sciencedirect.com) (pmc.ncbi.nlm.nih.gov). In the original n2c2 challenge, 0.8905 was the overall end-to-end lenient micro-F1 across relation types; the best ADE–Drug end-to-end F1 was 0.4755 (original challenge report).

Table 1. *Key ADE extraction datasets and tasks.*

Dataset	Domain/Source	Size (examples)	Entities/Relations	Key Metric Examples (F1)
n2c2-2018 ADE	ICU discharge summaries (MIMIC-III) (www.sciencedirect.com)	~505 notes (training+test)	Drugs, ADEs, Drug–ADE relation	Overall lenient micro-F1: 0.9418 concept extraction; 0.8905 end-to-end across relation types. ADE–Drug end-to-end F1: 0.4755 (original challenge report)
MADE 1.0	EHR notes (unknown institution) (www.sciencedirect.com)	1089 notes	Drugs, ADEs, Indications, Reasons	~0.83 (drug), 0.84 (reason) (www.sciencedirect.com) (pipeline CRF)
TAC 2017	Prescription drug labels	101 annotated training labels	Adverse-reaction extraction	Not specified here (official challenge documentation)
Custom cohorts	e.g. Oncology, Immunotherapy	Varies (hundreds)	Treatment-related ADEs, severity	—

Sources: Collected from public challenge descriptions and literature (www.sciencedirect.com) (pmc.ncbi.nlm.nih.gov). Continued development of corpora remains an active need for robust evaluation.

LLM-based ADE Extraction: Approaches

Prompting and Zero-Shot LLMs

A striking trend is the use of LLMs in a **zero-shot or prompt-driven** extraction paradigm. In this approach, one formulates a natural language prompt that instructs the model what information to extract, without additional training on labeled examples. For instance, a prompt might say: *"Identify any adverse drug events mentioned in this note, and specify the drug and effect."* Then a model like GPT-4 generates an answer listing identified events.

Studies show this can be effective for certain tasks. An International Journal of Medical Informatics study found that ChatGPT (generative GPT-4 via API) achieved *"competitive"* zero-shot IE performance on radiology reports (www.sciencedirect.com). Specifically, ChatGPT matched or exceeded a task-specific baseline on straightforward items (tumor size, location) even without training (www.sciencedirect.com). Huang *et al.* (2024) studied pathology information extraction with ChatGPT-3.5 (gpt-3.5-turbo-16k), not GPT-4 ADE extraction. Prompts were iteratively developed using 78 lung-cancer pathology reports, followed by evaluation on 774 reports. The mean accuracy across attributes was 0.89; accuracy should not be labeled F1 (original study).

Zero-shot performance tends to depend heavily on prompt design. B. Hu *et al.* (IJMI) found that adding domain knowledge to prompts (e.g. definitions or examples of medical jargon) improved some extraction tasks but could hurt others (www.sciencedirect.com) (pubmed.ncbi.nlm.nih.gov). Importantly, zero-shot methods reduce

annotation effort but rely on the LLM's in-context learning ability; they may struggle with very complex relations or rare entities. In practice, hybrid methods often combine an LLM prompt with rule-based post-filtering or a small classifier to ensure validity.

Fine-Tuning and Continuous Learning

An alternative is to **fine-tune** an LLM (or train it further) on annotated clinical data. This is now routinely done: for example, open-source LLMs like Llama-2/3 or BioGPT have been fine-tuned on EHR notes to improve domain performance (www.sciencedirect.com) (pmc.ncbi.nlm.nih.gov). Fine-tuning requires labeled AE examples, but it leverages the LLM's pretrained knowledge. Studies report that even a small number of labeled examples can significantly boost performance when used to fine-tune a large model.

McMaster *et al.* (2024) trained a DeBERTa model (an advanced Transformer) on unlabeled discharge summaries before ADE labeling, boosting F1 score (www.sciencedirect.com). Similarly, Kan *et al.* reported a *pharmBERT* fine-tuned on drug labeling text, improving downstream drug–ADE relation extraction (www.sciencedirect.com). In our surveys, fine-tuned LLMs often yield the highest benchmark scores. For example, we note benchmark F1s for top BERT models on MADE and n2c2 (≈0.94 NER in some cases (pmc.ncbi.nlm.nih.gov)) – these systems were essentially fine-tuned Transformers.

However, fine-tuning large LLMs requires care. Computational resources (TPU/GPU) and data privacy are major issues. As Hu *et al.* emphasize, even if LLaMA outperforms BERT slightly, it used vastly more compute (pmc.ncbi.nlm.nih.gov). Practically, an institution may opt for smaller but fine-tunable models (e.g. LLaMA-3 8B) to run on-premises. Emerging tools like *Low-Rank Adaptation (LoRA)* allow fine-tuning only part of the model (e.g. 400M additional parameters), making it feasible on a single GPU (www.nature.com). Indeed, Liu *et al.* used LoRA to fine-tune Llama-3.1 on as few as 100 pathology reports, achieving near-human accuracy (www.nature.com). The Strata library they developed encapsulates such workflows: small annotated sets can quickly calibrate an LLM for local clinical text (www.nature.com).

Data Augmentation and Distillation

LLMs also facilitate *weakly supervised* approaches. For example, Gu *et al.* (2023) used GPT-3.5 as a "teacher" to **generate synthetic ADE labels** on unlabeled discharge summaries (www.sciencedirect.com). These pseudo-labeled examples then trained a smaller "student" model (BioGPT or PubMedBERT). This knowledge-distillation approach increased performance when human-labeled data were scarce. Relatedly, one can prompt GPT to rewrite or normalize text (e.g. expand "pt. reports rash" to "the patient developed a rash"), effectively augmenting the training set. Such methods show that LLMs can amplify limited annotations, though careful validation is needed to avoid propagating errors.

LLMs themselves can be distilled to smaller models for efficiency. Several works have compressed GPT-family knowledge into lightweight architectures (e.g. MedAlpaca, Bloomz) for on-edge deployment. While large models like GPT-4 remain closed, the community has created open alternatives (LLaMA series, GPT-NeoX, etc.) to avoid reliance on proprietary APIs. These open models encourage reproducible research and allow transparency in training data (www.nature.com).

Pipeline Architectures

Practically, many systems use a **pipeline** combining multiple NLP components. For example, a pipeline may first preprocess notes (segment sections, correct typos), then run an LLM for entity extraction, then apply a relation classifier. A few studies have used LLMs for only part of this pipeline. Hier *et al.* (2025) showed that even using GPT-4 as a *preprocessor* – to correct grammar and expand abbreviations – significantly improved a downstream ontology-mapping tool (www.mdpi.com). This suggests a hybrid strategy: sprinkle LLM usage where it adds the most (e.g. understanding free-form lists or normalizing free text) while using rule-based or lightweight models elsewhere.

Fully end-to-end pipelines using LLMs have also been reported. In such designs, the LLM ingests each clinical document and outputs structured fields. Strata (Liu *et al.*, 2025) is one example: it fine-tunes an LLM to directly fill JSON templates of pathology report fields (www.nature.com). Other teams have built LLM "agents" to iteratively query the text (e.g., chain-of-thought prompting) for specific event attributes. The advantage is simplicity (one model handles all tasks), but error analysis becomes harder. Current evidence suggests that modular pipelines (with separate NER and RE models) are still competitive, especially if each component is carefully tuned (pmc.ncbi.nlm.nih.gov).

Accuracy and Performance

Standard Metrics

Accuracy in AE extraction is typically reported via precision, recall, and F1-score for both entity recognition and relation extraction. Precision measures the fraction of extracted events that are true, recall the fraction of true events detected, and F1 the harmonic mean. For entity extraction we report micro-averaged F1 over ADE mentions; for relations we consider F1 of correctly linked drug–ADE pairs. Some works also report exact-match accuracy for entire structured outputs (as in Strata) or ROC-AUC for binary detection of any ADE mention in a note.

Benchmarks on Public Datasets

n2c2-2018 ADE (Discharge Summaries) – The original challenge reported best overall lenient micro-F1 scores of 0.9418 for concept extraction, 0.9630 for relation classification on gold-standard concepts, and 0.8905 for end-to-end extraction across relation types. The best ADE–Drug end-to-end F1 was 0.4755 (original challenge report). These are distinct evaluation settings and should not be labeled ADE-only NER accuracy.

MADE 1.0 (Medication/Indication/ADE) – Early neural models reached ~0.83–0.87 F1 for drug and reason entities (www.sciencedirect.com). Berant *et al.* (2022) fine-tuned domain BERT and achieved ~0.86 F1. Our surveys find that current LLM-based pipelines report F1 above 0.9 for drugs but around 0.7–0.8 for ADEs on MADE. For example, Guo *et al.* (2022) report BioBERT at 0.84 ADE F1 on MADE. Liu *et al.* (2025) using GPT-4 prompts on related data obtained ~0.86 "relaxed F1" in a different ADE corpus (pmc.ncbi.nlm.nih.gov) – suggesting LLMs are in the same ballpark as specialized models.

Other Corpora (e.g. I2B2, VAERS) – On smaller or specialized corpora, reported metrics vary widely. For vaccine reports (VAERS, etc.), GPT-4 few-shot frameworks have achieved up to ~0.86 F1 (pmc.ncbi.nlm.nih.gov). On clinical trial data, LLMs often need more context to match structured fields. We note that comparisons across corpora are difficult due to differing annotation schemas. In MultiADE, fine-tuned models achieved the best result on each individual dataset (original study). Comparisons should specify the model, dataset, entity type, prompting or training setup, and matching rule.

Relation Extraction (Drug–ADE) – The n2c2 challenge separates relation classification using gold-standard concepts from end-to-end extraction. Its best overall lenient micro-F1 was 0.9630 for relation classification across relation types, whereas the best ADE–Drug end-to-end F1 was 0.4755 (original challenge report).

Case Study: Strata ("Human-Level IE")

Liu *et al.* (2025) demonstrated one of the strongest results on clinical IE. They developed **Strata**, a low-code system for fine-tuning LLMs on structured extraction. Using 100–400 annotated pathology reports across four domains (prostate MRI, breast pathology, etc.), they fine-tuned small Llama-3 8B models with LoRA. When evaluated on held-out reports, the fine-tuned LLM achieved an **average exact-match accuracy of 90.0% ± 1.7%**, matching a second human annotator (www.nature.com). By contrast, non-finetuned open-source models (even specialized ones) performed far worse (Llama-3 8B with medical tuning scored only ~39–57% accuracy) (www.nature.com). Importantly, GPT-4 (zero-shot) was also near-human except in one category. The key takeaway is that a well-tuned LLM with modest data can reach *near human-level* reliability on structured extraction, at least in these focused settings (www.nature.com). Strata emphasizes reproducibility: all computation ran on desktop GPUs, and models remained local to the institution, avoiding external data sharing. This case underscores the feasibility of implementing LLM pipelines in practice – an entry for **deployment**.

Zero-Shot ChatGPT Performance

Huang *et al.* (2024) evaluated ChatGPT-3.5 (gpt-3.5-turbo-16k) for extracting pathology classifications. They developed prompts iteratively on 78 lung-cancer pathology reports and evaluated them on 774 reports, obtaining mean attribute accuracy of 0.89 (original study). This pathology classification result should be kept separate from ADE-extraction benchmarks.

Similarly, an international study on radiology reports showed ChatGPT's zero-shot performance was "competitive" with a state-of-the-art medical IE system (www.sciencedirect.com). In radiology tasks (tumor sizes, etc.), ChatGPT matched extraction accuracy of a trained model for simpler attributes (www.sciencedirect.com). Encouragingly, adding medical context to the prompts improved some metrics. These early results highlight that off-the-shelf GPT models can quickly tackle specific extraction questions, pending robust prompts.

Benchmarks with Prompt Enhancements

Prompt engineering can further boost LLM accuracy. For example, Qiu *et al.* (2022) showed that chaining prompts (having the model justify each tag) or giving it an ontology can improve recall without hurting precision. The IJMI study (www.sciencedirect.com) found that “ChatGPT can achieve competitive IE performance” out-of-the-box, but carefully instructing it (e.g. providing definitions of terms, few-shot examples) typically improves results on nuanced tasks. In practice, one might combine an LLM prompt with a small hand-written set of few-shot examples to steer it toward the clinical use-case. This hybrid yield can approach supervised learning accuracy while still minimizing manual labeling.

Comparative Results

The MultiADE benchmark found that fine-tuned models achieved the best result on each individual dataset (original MultiADE study). This finding does not support a blanket claim that prompted LLMs outperform supervised systems. Compare results within the same dataset, entity category, and evaluation setup; general pathology or radiology extraction results should be presented separately from direct ADE evidence.

Implementation in Clinical Settings

Deploying LLM-based extraction in healthcare entails several practical considerations beyond model accuracy. Integrators must address data handling, privacy, integration with workflows, and maintenance. We discuss key factors:

Data Privacy and Security

Clinical notes can contain protected health information (PHI). When a cloud provider creates, receives, maintains, or transmits electronic PHI on behalf of a HIPAA covered entity or business associate, the parties must enter into a HIPAA-compliant business associate agreement (BAA) and otherwise comply with the HIPAA Rules. This applies even when the provider stores only encrypted PHI without the decryption key. Both parties must conduct risk analyses and address applicable safeguards; local hosting also requires appropriate safeguards and does not by itself establish HIPAA compliance (HHS cloud guidance).

De-identification must follow the processes required by the HIPAA Privacy Rule. Informal redaction should not be assumed to meet that standard. HHS distinguishes information de-identified under those processes from PHI; a cloud provider receiving and maintaining only such de-identified information is not a business associate for that activity (HHS cloud guidance).

Compute Resources

LLMs are computationally intensive. GPT-4 inference costs and latency are significant barriers in real-time systems. Hu *et al.* found that LLaMA models ran up to 28× slower than BERT (pmc.ncbi.nlm.nih.gov). This implies that massive model usage may be feasible only for batch analytics, not live EHR queries, unless

sufficiently powerful hardware is available. Organizations must provision GPUs (e.g. NVIDIA A100/V100 class) or TPUs for fine-tuning and inference. Cost analysis (renting cloud GPUs vs. owning hardware) is essential.

Several strategies mitigate costs: using **smaller models** (e.g. fine-tuned 7B or 13B instead of 70B parameters), employing LoRA (fine-tune only a subset of weights) (www.nature.com), or quantization/ pruning to reduce memory. Hu *et al.* found that BERT processed notes 5–28 times faster than LLaMA models during inference (original study). It is also possible to run inference on CPU with slower speed for off-peak tasks.

Integration with Clinical Workflow

Seamless integration into the EHR environment is crucial for adoption. Potential architectures include:

- **Real-time Assistance:** The LLM extraction runs as notes are entered (e.g. by clinicians), flagging possible ADE mentions for review by pharmacists or care teams. This demands very low latency (sub-second responses).
- **Batch Surveillance:** Periodic jobs run on accumulated data (e.g. nightly reviews of recent discharges) to detect new ADE signals. Here latency is less critical but throughput matters.
- **Decision Support:** Uploaded or pasted notes can be processed on-demand by clinicians. A user-friendly interface (with audit trail of the LLM's rationale) would be needed.

Regardless, results should be human-in-the-loop. The output of the LLM should be reviewable (ideally with highlighted evidence spans). Many authors stress the need for clinicians to verify AI-extracted AEs before taking action (pmc.ncbi.nlm.nih.gov).

A variety of **tools and libraries** facilitate implementation. Examples include:

- **Strata** (Liu *et al.*, 2025): a workflow toolkit for fine-tuning LLMs on clinical IE tasks (www.nature.com).
- **MedTagger** or **cTAKES** for hybrid rule-ML pipelines (older, but sometimes used as parsing front-ends).
- **LangChain** or **LlamaIndex** (non-clinical libraries) adapted to manage prompts, chaining, and retrieval-augmented generation.
- Fine-tuning frameworks (Hugging Face Transformers, LoRA toolkits) support smaller model weight updates.

Health IT must also consider EHR compatibility (HL7 FHIR mapping, API hooks) and logging for audit/regulatory review.

Example Workflow

A plausible implementation pipeline might be:

1. **Preprocess**: Tokenize and section clinical note. Optionally use an LLM (e.g. GPT-4) to standardize text format and correct spelling (www.mdpi.com).
2. **Entity Extraction**: Pass the cleaned text to an LLM or model to identify candidate ADE mentions and related entities. This could be via a fine-tuned model (e.g. ClinicalBERT) or via prompt-query to GPT.

3. **Relation Linking**: For each ADE mention, identify the associated drug or cause (possibly via another fine-tuned classifier or a follow-up LLM prompt).
4. **Post-Process**: Filter out low-confidence outputs, map entities to standardized codes (e.g. SNOMED, MedDRA) perhaps using an ontology lookup.
5. **Review/Output**: Present the results in the EHR portal or safety dashboard. Highlight salient text and the recommended associations.

At each step, human oversight can validate or correct the AI. For instance, a pharmacist might review flagged ADEs before they propagate to a safety report.

Benefits and Current Use-Cases

Schmider *et al.* demonstrated the feasibility of AI-assisted extraction from adverse-event source documents and evaluation of case validity (original pilot). It also enables **scale**: systems can scan all notes daily, potentially catching rare events that would otherwise slip through manual review. Another research use is extracting medication dosage from free-text prescription "sig" fields with GPT-4 (conference abstract).

On the hospital side, clinicians have begun experimenting. For example, some oncology groups are testing GPT-4 to flag known immune-related adverse events in EHRs, with preliminary micro-F1 ≈0.56–0.66 (pmc.ncbi.nlm.nih.gov). Others use AI chatbots as a safety net during chart review. Although many efforts are internal or pilot stage, the trend is clear: AI/LLMs are entering pharmacovigilance and patient safety workflows (pmc.ncbi.nlm.nih.gov).

Discussion and Future Directions

Accuracy Considerations

While LLMs have raised the ceiling, **caution** is warranted. LLMs still make mistakes: hallucinating events not present in text, or missing subtle negations. As reported in literature, ChatGPT sometimes invents ADEs if prompts are not carefully phrased (pmc.ncbi.nlm.nih.gov). Domain mismatch remains an issue: e.g., the pathological concept "colitis" could be mistaken for context negation.

Another limitation is **common-sense and context**: medical notes require understanding of causality and temporal sequence. LLMs often lack built-in commonsense for these. In Zitu *et al.*'s review, most systems were challenged by ADE–Drug relation extraction "due to lack of context" (pmc.ncbi.nlm.nih.gov). They suggest that none of the GPT-style prompts fully solved the detailed genealogy of an ADE (what preceded it, concurrent conditions). This signals a need for richer modeling (e.g. multi-document discourse, patient timeline tracking).

Finally, the training data of LLMs can introduce bias. If an LLM was trained mostly on general web data, it may under-recognize events common in underrepresented populations. The community must monitor for demographic and socioeconomic biases in extracted events.

Implementation Challenges

Aside from compute and privacy (discussed), there are hurdles of trust and integration. Clinicians generally need transparency: “Why did the model think this is an ADE?” Black-box LLM outputs are hard to trace. Explainable AI (XAI) techniques – attention visualization, or natural-language rationales – are areas for further work. Hu *et al.* (2026) stress that the **computational cost** of LLM vs the marginal gains prompts careful cost–benefit analysis (pmc.ncbi.nlm.nih.gov). Healthcare IT departments will need to validate LLM outputs with their own data before deployment.

Assess regulatory status for the specific intended software function. FDA distinguishes device software functions from CDS functions excluded from the device definition under section 520(o)(1)(E) of the FD&C Act (Non-Device CDS). Use the FDA Clinical Decision Support Software guidance to assess the relevant function rather than treating all AE extraction as a medical device.

Future Prospects

The landscape is moving quickly. We highlight several promising directions:

- **Domain-Specific LLMs:** Models pretrained on medical corpora (e.g. PubMed, MIMIC) can provide stronger background knowledge. Google’s *Med-PaLM* and Meta’s *LLaMA-3* fine-tuned on biomedical text may soon be available. Early reports suggest these could reduce hallucinations.
- **Few-Shot and Active Learning:** New techniques allow ultra-low-data tuning. Active learning pipelines where the model queries a human on which examples to label (or uses LLM to synthesize examples) could minimize annotation needs.
- **Retrieval-Augmented Systems:** Combining LLMs with biomedical knowledge bases (PubMed, drug monographs) can improve extraction of nuanced AEs that require background. For example, a system might query UMLS during inference to confirm entity types. Early research has shown retrieval-augmented LLMs can improve clinical reasoning tasks; applying that to AE extraction is a ripe area.
- **Privacy-Preserving Learning:** Federated learning or on-device inference could help share improvements (model updates) across hospitals without sharing notes. Some groups are exploring federated fine-tuning of models on health data; this could accelerate LLM specialization.
- **Expanded Tasks:** Beyond extraction, LLMs could summarize ADE narratives for regulatory reports, or even generate patient-friendly explanations. Summarization using LLMs (trained to produce discharge summaries or safety reports) is a related field that Synergies with extraction. For example, an LLM might extract adverse events and then generate a draft pharmacovigilance report for a clinician to review.
- **Continuous Monitoring and Feedback:** A live system could learn over time. Suppose an LLM flags an ADE and a pharmacist confirms it; that feedback could be used to adapt the model (supervised fine-tuning on-site). Bridging LLM outputs with clinical decision support loops offers self-improving pipelines.
- **Ethical and Equity Considerations:** Lastly, interdisciplinary efforts will be needed to ensure these tools work well for all patient populations and do not inadvertently create disparities. For instance, if a model rarely saw notes in Spanish, it might miss AEs in a Hispanic patient’s record. Ensuring inclusive data

and audits will be an ongoing requirement.

Case Studies and Examples

- **Pfizer AE Case Automation:** Schmider *et al.* evaluated AI-assisted extraction from adverse-event source documents and case validity. The pilot established feasibility of these tasks (original pilot).
- **Neurology Clinic Notes (Hier et al. 2025):** GPT-4 was applied to 1,618 neurology outpatient notes in a multi-step pipeline (www.mdpi.com). By correcting abbreviations and grammar, GPT-4 preprocessing boosted the recall of a phenotype-extraction pipeline from 0.61 to 0.68 (www.mdpi.com). Errors were statistically significantly lower, showing that LLMs can serve as a preprocessor to help existing NLP tools.
- **Radiology Reports (Hu *et al.* 2024):** This experimental study evaluated ChatGPT on 847 real CT reports and found competitive zero-shot information extraction in this radiology setting (original study).

Conclusion

LLMs have *transformed* clinical NLP by delivering unprecedented language understanding and generation capabilities. For ADE extraction, the MultiADE benchmark found that fine-tuned models achieved the best result on each individual dataset (original MultiADE study). They enable new workflows: zero-shot alerts, automated reporting, and semantic summarization of patient safety data. However, harnessing LLMs in practice demands caution regarding data privacy, validation, and cost. As Hu *et al.* succinctly conclude: the **promise** of LLMs comes with a **trade-off** in complexity and resources (pmc.ncbi.nlm.nih.gov). Our review suggests that, with rigorous integration and oversight, LLMs can become valuable assistants in pharmacovigilance and clinical care.

Future work must focus on closing remaining gaps: improving relation extraction, reducing errors in rare cases, and embedding these models thoughtfully in healthcare systems. With continuing advances (and newer models like GPT-4o, BioGPT-2, *etc.*), the accuracy of AE extraction is likely to climb further. Importantly, researchers should publish more clinical benchmarking of LLMs (standardized splits, multi-site data) to truly quantify gains. Meanwhile, pilot implementations in hospitals and pharma should measure detection sensitivity, cost savings, and workflow efficiency as outcomes to be evaluated. Schmider *et al.*'s Pfizer pilot demonstrated feasibility of AI-assisted extraction from adverse-event source documents and evaluation of case validity (original pilot). The potential payoff—a more proactive healthcare system that learns quickly from real-world events—is enormous, and LLMs are now the strongest tool we have to reach it.

References

(Inline citations throughout are keyed to the following sources.)

- Hu *et al.*, **J Am Med Inform Assoc** (2026) – LLaMA vs BERT for clinical IE (pmc.ncbi.nlm.nih.gov).

- Liu *et al.*, **Scientific Reports** 15, 45239 (2025) – *Strata*: LLM fine-tuning for clinical IE (www.nature.com).
- Hier *et al.*, **Information** 16:446 (2025) – GPT-4 preprocessing improves clinical concept extraction (www.mdpi.com).
- Huang *et al.*, **NPJ Digital Med.** 7:106 (2024) – Zero-shot ChatGPT vs traditional NER (pmc.ncbi.nlm.nih.gov).
- Hu *et al.*, **Int J Med Inform** 183:105321 (2024) – Zero-shot ChatGPT for radiology IE (www.sciencedirect.com).
- Zitu & Owen **J Clin Med** 14(15):5490 (2025) – Clinical review of LLMs for ADEs (pmc.ncbi.nlm.nih.gov).
- Schmider *et al.*, **Clin Pharmacol Ther** 105(4):954–61 (2019) – Pharma AI pilot for AE case processing (pmc.ncbi.nlm.nih.gov).
- Henry *et al.*, **JAMIA** 27(1):3–12 (2020) – n2c2 2018 ADE challenge results (pmc.ncbi.nlm.nih.gov).
- Wei *et al.*, AMIA Symposium 2019 – ADE extraction with BERT-based NER .
- Chapman *et al.*, J Biomed Inform 2021 – MADE 1.0 results (pipeline) (www.sciencedirect.com).
- Yang *et al.*, JAMIA Open 3(1):57–65 (2020) – MADE 1.0 pipeline system (www.sciencedirect.com).
- Boland *et al.*, **J Biomed Inform** 125 (2022) – DeBERTa model for drug ADE (MeDeBERTa) (pmc.ncbi.nlm.nih.gov).
- Vig *et al.*, JAMIA Open (2023) – Abstract medication/NLP normalization (EhrBERT1M) (pmc.ncbi.nlm.nih.gov).
- Others as cited above.

Source links

Links cited in this article, in order of first appearance. Full addresses are provided for readers using extracted text.

1. https://pmc.ncbi.nlm.nih.gov/articles/PMC12981642
2. https://www.sciencedirect.com/science/article/pii/S1532046424000212
3. https://www.nature.com/articles/s41598-025-28767-z
4. https://pubmed.ncbi.nlm.nih.gov/38157785
5. https://pmc.ncbi.nlm.nih.gov/articles/PMC11063058
6. https://www.sciencedirect.com/science/article/pii/S1386505623003398
7. https://pmc.ncbi.nlm.nih.gov/articles/PMC11498178
8. https://pmc.ncbi.nlm.nih.gov/articles/PMC7489085
9. https://pmc.ncbi.nlm.nih.gov/articles/PMC12347610
10. https://www.mdpi.com/2078-2489/16/6/446
11. https://pmc.ncbi.nlm.nih.gov/articles/PMC6590385
12. https://www.hhs.gov/hipaa/for-professionals/special-topics/health-information-technology/cloud-computing/index.html
13. https://bionlp.nlm.nih.gov/tac2017adversereactions
14. https://arxiv.org/pdf/2405.18015
15. https://www.ahajournals.org/doi/10.1161/circ.150.suppl_1.4147064
16. https://www.fda.gov/regulatory-information/search-fda-guidance-documents/clinical-decision-support-software

THE PUBLISHER

About IntuitionLabs

IntuitionLabs is an AI consulting, custom software development and data engineering firm serving pharmaceutical, biotechnology, medical-device and other life-science organizations. We work with clinical, regulatory, medical-affairs, commercial, quality and IT teams to connect technology decisions with the work people need to accomplish.

AI consulting and adoption

Our AI enablement services cover readiness assessments, use-case selection, governance and policies, team workshops, adoption measurement and ongoing advisory support. We help organizations structure the information layer behind AI: source material, context, permissions and maintained knowledge that make generated answers useful and reviewable. Private LLM inference and hosted AI options support teams evaluating how to operate AI with appropriate control over their data and infrastructure.

Software, data and life-science workflows

IntuitionLabs develops custom software for pharma and biotech, integrates enterprise systems, and builds data engineering and business intelligence solutions. Areas of focus include AI agents, regulatory research, medical writing, medical affairs, CMC information, competitive intelligence and clinical-document workflows. Our eTMF intelligence work includes cross-system reconciliation and inspection-readiness support.

Enterprise platforms and regulated delivery

We provide Veeva services, application support, managed services, integrations and custom applications, alongside enterprise content work involving platforms such as Egnyte. For regulated workflows, our services include GxP enablement, computer-system validation and software development addressing 21 CFR Part 11 requirements. The applicable controls, validation responsibilities and acceptance criteria are defined for each engagement.

Work with IntuitionLabs

Explore AI enablement, pharma and biotech software development, data engineering and BI, and Veeva services. Contact IntuitionLabs to discuss your workflow, information sources and implementation needs.

IntuitionLabs publishes educational research to help life-science teams make informed technology decisions. Coverage of a product or organization does not imply a client relationship, endorsement or partnership.

Website: https://intuitionlabs.ai

This article is educational. Verify consequential medical, legal, financial, regulatory and technical decisions with appropriate professionals and the cited primary sources. Company services are described separately from the article's research findings.