Confidence-Calibrated LLM Pipeline for Adverse Drug Reaction Detection from Clinical Instant Messaging: Development and Temporal Validation

Dongxu Wang[1,†], Zihong Lu[1,†], Wenbo Yuan[1], Kaiqiang Yuan[2], Di Yin[1], Ying Yao[1,*], Sunmin Jiang[1,*]

[1]Department of Pharmacy, Affiliated Women's Hospital of Jiangnan University, Wuxi, China

[2]Guangzhou Pinyi Information Technology Co., Ltd., Guangzhou, China

[†]These authors contributed equally to this work.

[*]Corresponding authors: Ying Yao, Sunmin Jiang

Abstract

Objective: To develop and temporally validate a confidence-calibrated large language model (LLM) pipeline for adverse drug reaction (ADR) detection, entity extraction, and causality assessment from clinical instant messaging (IM).

Materials and Methods: Development proceeded in two phases. In the pilot phase, we evaluated a hybrid rule-LLM pipeline on 450 simulated messages (Fleiss' κ=0.719) and 1,792 real messages from a hospital pharmacy WeChat group. In the validation phase, we constructed a 2,023-message gold standard (2021–2025) annotated by five pharmacists via blind review, with temporal split into development (<2024, n=1,277) and locked test ($\geq$2024, n=746) sets. The pipeline uses Qwen 3.5 Plus with 1–10 confidence scoring. Evaluations included calibration analysis, multi-turn context ablation, cross-model robustness (four LLMs), a fine-tuned BERT baseline, specificity on 1,000 synthetic medical-but-non-ADR messages, and Naranjo/IMCT causality assessment on 200 cases.

Results: The pipeline improved from pilot F1=0.906 to temporally validated F1=0.970 (P=0.944, R=0.997), matching the BERT baseline. Confidence scores were well-calibrated (expected calibration error=0.030). Single-message classification outperformed multi-turn approaches (p <0.001). Entity extraction achieved lenient F1 of 0.750 (drug) and 0.738 (symptom). All causality instruments showed no agreement with pharmacist consensus (intraclass correlation coefficient≤ -0.236).

Discussion: The pipeline achieves near-expert ADR detection with calibrated uncertainty, but causality assessment is limited by the information content of brief IM messages.

Conclusion: Confidence-calibrated LLM classification is effective for IM-based ADR screening (projected positive predictive value=71% at 10% prevalence, based on synthetic negative controls). Single-center findings require multi-site validation. Causality assessment should be deferred to formal review.

Keywords: adverse drug reaction, pharmacovigilance, large language model, clinical instant messaging, confidence calibration, temporal validation

Lay Summary: Clinical pharmacists discuss suspected drug side effects in hospital messaging groups, but these valuable safety signals are not systematically captured. We developed an AI system that automatically identifies adverse drug reaction reports from these conversations with an F1-score of 0.970, matching the performance of a purpose-trained model without requiring task-specific training data. The system provides calibrated confidence scores that enable priority-based review. While effective for detection, automated causality assessment from brief messages remains infeasible, indicating that these systems should serve as screening tools with human expert follow-up.

1 Introduction

Adverse drug reactions (ADRs) account for 5–8% of hospital admissions and are a leading cause of preventable patient harm [1–3]. Although spontaneous reporting systems remain central to post-marketing pharmacovigilance [4], up to 94% of ADRs go unreported, primarily because voluntary reporting requires clinicians to interrupt workflow and complete structured forms [5–7].

Clinical instant messaging (IM) platforms—particularly WeChat and WeCom in China—have become integral to hospital communication [8]. Pharmacists routinely discuss suspected adverse reactions through these channels, generating pharmacovigilance signals as a byproduct of clinical practice. Yet scoping reviews of natural language processing (NLP)-based adverse event detection from electronic health records [9, 10] and social media [11, 12] have identified no system utilizing IM-derived clinical text.

Large language models (LLMs) have demonstrated strong performance in adverse event extraction from clinical notes [13, 14], surveillance reports [15], and social media [16], though challenges persist in domain variability and hallucination [17]. No prior work has applied LLMs to clinical IM conversations or attempted automated causality assessment from informal conversational text.

In a pilot study, we developed a hybrid rule-LLM pipeline that achieved F1=0.906 on 450 simulated messages and F1=0.905 on 1,792 real clinical messages, with zero false positives. However, the pilot evaluation relied on a single annotator's simulated benchmark, lacked temporal separation, and did not validate entity extraction or causality assessment. The present study extends this work through a comprehensive two-phase development and validation design with five contributions:

1. **Rigorous temporal validation** with a 2,023-message gold standard annotated by five pharmacists, strict development/test separation, and a fine-tuned BERT baseline.

2. **Confidence-calibrated classification with architectural ablation**, evaluating calibration quality and testing whether two-pass re-examination improves over single-pass classification.

3. **Multi-turn context analysis and entity extraction validation** against pharmacist annotations across four independent Chinese LLMs.

4. **Specificity and deployment feasibility**, including prevalence-adjusted positive predictive value for lower-prevalence settings.

5. **Systematic causality assessment** using three instrument variants, characterizing the limits of IM-based causality scoring.

2 Materials and Methods

2.1 Study Design Overview

This study followed a two-phase design. Phase 1 (pilot) developed the pipeline architecture and validated it on simulated and real-world messages. Phase 2 (temporal validation) constructed a multi-annotator gold standard and conducted comprehensive evaluation with confidence calibration, multi-turn analysis, cross-model robustness, and causality assessment.

2.2 Phase 1: Pilot Validation

2.2.1 Simulated Benchmark

A clinical pharmacist with 5 years of experience authored 450 messages based on authentic drug safety scenarios, stratified by difficulty: easy (n=110; explicit drug–symptom co-mentions), medium (n=165; brand names, abbreviations, colloquial phrasing), and hard (n=175; implicit causality, ambiguous referents, lab values). Five independent pharmacists validated clinical realism (Fleiss' κ=0.719, mean realism rating 3.81/5) [18].

2.2.2 Real-World Pilot Data

A non-overlapping set of 1,792 authentic messages from the hospital's pharmacovigilance WeChat group (March 2024–February 2025; 1,372 ADR+, 420 ADR−), annotated by a single pharmacist. These messages are temporally and numerically distinct from the

Phase 2 dataset; none were included in Phase 2 development or test sets. A negative control corpus of 3,897 messages from a separate pharmacy quality-control group (containing drug names in non-ADR contexts) served as the specificity test set.

2.3 Phase 2: Temporal Validation Datasets

2.3.1 Data Source

The Phase 2 dataset comprises 2,023 messages from a pharmacovigilance-dedicated WeChat group at Wuxi Maternity and Child Health Care Hospital, a tertiary obstetrics and gynecology hospital, collected between September 2021 and March 2025. Messages represent the complete unfiltered archive from this group.

2.3.2 Gold Standard Construction

Five clinical pharmacists (each with ≥3 years of experience) independently annotated each message as ADR-positive or ADR-negative using blind review. An ADR-positive message was defined as containing: (1) an identifiable drug name; (2) a described adverse symptom or clinical sign; and (3) an explicit or implied temporal association between drug administration and symptom onset. A calibration session using 50 pilot messages preceded independent annotation. Inter-annotator agreement was Fleiss' κ=0.697 (95% CI: 0.677–0.717), indicating substantial agreement [19], with full 5/5 agreement on 69.1% of messages (n=1,398), 4/5 agreement on 17.4% (n=351), and 3/5 agreement on 13.5% (n=274).

We determined the final label through a two-stage process: initial majority vote (≥3/5), followed by systematic adjudication by two senior pharmacists (not involved in initial annotation) who reviewed all cases where the majority vote conflicted with the objective annotation criteria. Adjudication was unidirectional (ADR− to ADR+ only), correcting cases where messages contained all three required elements (drug name, adverse symptom, and temporal association) yet were labeled negative by majority vote, primarily involving dechallenge patterns, abbreviated drug names, and chemotherapy laboratory abnormalities. In total, 734 labels were changed (439 in the development set, 295 in the test set): 497 had 0/5 positive votes, 118 had 1/5, and 119 had 2/5. The high correction rate among 0/5 cases reflects messages where all annotators overlooked implicit ADR patterns (e.g., dechallenge narratives without explicit drug names) that met the pre-specified criteria upon expert review. Adjudication criteria were defined independently of any pipeline output; all corrections with original votes and rationale are documented in the code repository. The final dataset contains 1,692 ADR-positive (83.6%) and 331 ADR-negative (16.4%) messages, compared with 958 (47.4%) positive by initial majority vote alone.

2.3.3 Temporal Split

We split messages by timestamp at January 2024 to prevent data leakage. We chose this cutoff a priori based on two criteria: (1) achieving approximately 60/40 development/test proportions, and (2) ensuring the test set captured the most recent clinical patterns, including any temporal drift in reporting conventions or drug formulary changes:

- **Development set** (< January 2024): 1,277 messages (1,054 ADR+, 223 ADR−)
- **Test set** (≥ January 2024): 746 messages (638 ADR+, 108 ADR−)

All optimization experiments used only the development set; the test set was accessed once for final validation.

2.3.4 Missing Data

No messages were excluded for missing or incomplete text content. All 2,023 messages in the pharmacovigilance group archive contained extractable text and were included in the annotation process.

2.3.5 Conversation Structure

We grouped messages into 401 multi-message conversations (≥2 messages): 245 (791 messages) in the development set and 156 (393 messages) in the test set, used for multi-turn context experiments.

2.3.6 Medical Negative Controls

To evaluate specificity, we generated 1,000 synthetic messages across 11 categories of medical-but-non-ADR content (e.g., confounding patterns where drug–symptom co-occurrence reflects expected pharmacological effects, n=120; disease symptoms without drug involvement, n=90; dosage consultations, n=90). Templates incorporated realistic drug names and clinical parameters, achieving a 63.5% drug keyword rate and 4.1% drug–symptom co-occurrence rate.

2.4 Pipeline Architecture

The pipeline uses a two-layer design developed in Phase 1: a deterministic rule-based prefilter and an LLM classifier. The sole input to all configurations is the raw text content of each individual IM message; no structured metadata (sender identity, timestamp, patient demographics) or external knowledge bases are used as predictors.

2.4.1 Rule Layer

The rule engine matches messages against vocabularies of 140+ drug names (brand, generic, and abbreviated) and 60+ symptom patterns. Messages matching both are classified as ADR-positive; those matching neither are classified as ADR-negative; remaining messages proceed to the LLM.

2.4.2 LLM Layer with Confidence Scoring

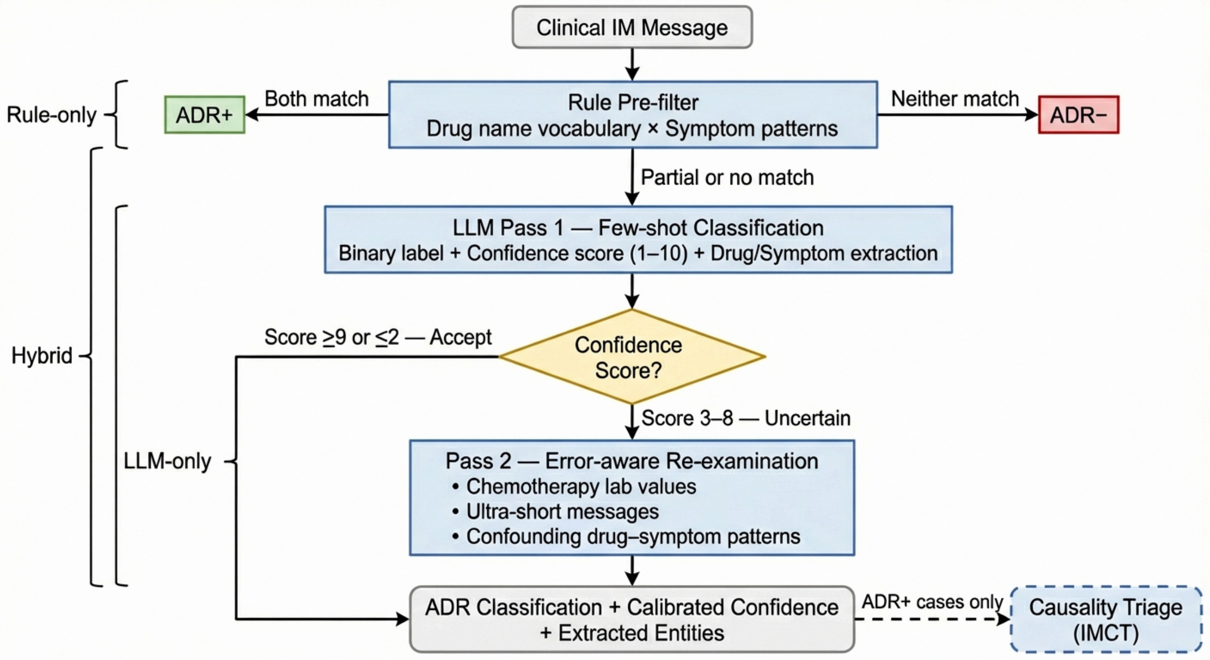
The LLM layer uses Qwen 3.5 Plus [20] via the Aliyun DashScope API. A few-shot prompt provides classification criteria with domain-specific guidance (chemotherapy laboratory abnormalities, abbreviated drug names), four annotated examples, and instructions to output JSON containing: binary classification, confidence score (1–10 integer), drug name, symptoms, patient identifier, and reasoning. The prompt was optimized on the development set and locked before test evaluation.

2.4.3 Two-Pass Architecture

We evaluated a confidence-gated two-pass design where intermediate-confidence cases (between thresholds θ_L and θ_H) undergo error-pattern-specific re-examination. Pass 2 incorporates targeted guidance for chemotherapy lab values, short texts (<20 characters), and confounding patterns. Thresholds (θ_H=9, θ_L=2) were optimized on the development set.

2.4.4 Pipeline Configurations

We compared three variants: **rule-only** (no LLM), **LLM-only** (no rule pre-filter), and **hybrid** (rule pre-filter with LLM fallback). Figure 1 illustrates the pipeline architecture.



Clinical IM Message
Rule-only
ADR+
Both match
Rule Pre-filter
Drug name vocabulary × Symptom patterns
Neither match
ADR−
Partial or no match
LLM Pass 1 — Few-shot Classification
Binary label + Confidence score (1–10) + Drug/Symptom extraction
Hybrid
Score ≥9 or ≤2 — Accept
Confidence Score?
Score 3–8 — Uncertain
LLM-only
Pass 2 — Error-aware Re-examination
• Chemotherapy lab values
• Ultra-short messages
• Confounding drug–symptom patterns
ADR Classification + Calibrated Confidence + Extracted Entities
ADR+ cases only
Causality Triage (IMCT)

Figure 1: Pipeline architecture for ADR detection from clinical IM messages. Three configurations were evaluated: rule-only (deterministic pre-filter), LLM-only (few-shot classification with confidence scoring), and hybrid (rule pre-filter with LLM fallback). Uncertain cases from Pass 1 are routed through error-pattern-specific re-examination in Pass 2.

2.5 Causality Assessment

2.5.1 Automated Naranjo Scale

For ADR-positive cases, we implemented LLM-based Naranjo assessment [21] using the 10-item scale (Definite ≥9, Probable 5–8, Possible 1–4, Doubtful ≤0). The LLM receives conversation context and outputs per-question answers with evidence and information source. Three pharmacists independently scored 200 cases as gold standard, with consensus by majority vote.

2.5.2 IM Causality Triage

Given the Naranjo scale's structural mismatch with IM data (Section 3.9), we designed a 3-question IM Causality Triage (IMCT) focusing on IM-answerable dimensions:

1. **Q1: Known drug–ADR association?** Whether the drug–symptom pair is a recognized adverse reaction.
2. **Q2: Temporal evidence?** Whether temporal markers indicate drug preceded symptom.
3. **Q3: Dechallenge evidence?** Whether symptom improvement after discontinuation is described.

Triage categories: HIGH (Q1=Yes AND [Q2 or Q3]=Yes), MEDIUM (partial evidence), LOW (all Unknown). Gold standard mappings used existing Naranjo annotations (Q2→Naranjo Q2, Q3→Naranjo Q3) with Q1 assessed against pharmacological knowledge.

2.6 Entity Extraction Evaluation

For ADR-positive messages with pharmacist annotations (n=958), extraction accuracy was evaluated using strict matching (normalized exact match after alias resolution) and lenient matching (token-overlap F1 with substring matching), stratified by split and annotator agreement level.

2.7 Experimental Design

Experiments were organized into four groups:

Group A: Core validation. Architecture comparison (Exp1), prompt optimization (Exp2), and locked test set validation (Exp3).

Group B: Multi-turn context. Context window ablation on the development subset (Exp4; windows 0, 1, 3, 5, full) and temporal validation on the test subset (Exp5).

Group C: Causality. Full Naranjo (Exp6), simplified 5-question Naranjo (Exp7), and IMCT (Exp10) on 200 cases.

Group D: Robustness and extraction. Inference stability (Exp8; five runs, temperature=0.1), error taxonomy (Exp9), medical negative control specificity, cross-model comparison (Qwen 3.5 Plus, DeepSeek V3.2 [22], Kimi K2.5, GLM-5 [23]), supervised Bidirectional Encoder Representations from Transformers (BERT)-base-Chinese baseline [24] (five-fold stratified CV on the development set; max sequence length=128; batch size=16; AdamW optimizer, lr=2×10^{-5}, weight decay=0.01; early stopping with patience=3 on validation loss; maximum 10 epochs), confidence calibration analysis, two-pass ablation, and entity extraction evaluation.

2.8 Statistical Analysis

We assessed classification performance using precision, recall, F1, and specificity with 95% bias-corrected and accelerated (BCa) bootstrap confidence intervals (10,000 resamples, seed=42) [25]. McNemar's test with continuity correction was used for paired comparisons; exact p-values are reported alongside χ^2 statistics. Where multiple pairwise comparisons were conducted within an experiment (e.g., context window ablation), Bonferroni correction was applied and adjusted significance thresholds are noted. Calibration was evaluated using expected calibration error (ECE, 10 bins) and Brier score. Causality agreement used intraclass correlation coefficient [ICC(2,1)], Cohen's κ, and mean absolute error (MAE). Entity extraction used strict and lenient matching stratified by split and agreement level. Inference stability used F1 coefficient of variation (CV) and unanimous agreement rate. All analyses used Python 3.13.

2.9 Ethics

This study was a retrospective analysis of de-identified clinical instant messages with no patient interaction or intervention. All messages were de-identified prior to analysis: patient identifiers were replaced with hashed anonymous codes, patient names were substituted with uniform placeholder tokens, and sender identities were anonymised with sequential codes. The institutional review board of Wuxi Maternity and Child Health Care Hospital determined that formal ethics approval was not required. Clinical pharmacists participating in the annotation study provided verbal informed consent. For the retrospective IM data, individual consent was waived given the de-identified, retrospective nature of the analysis. LLM inference used the Aliyun DashScope API under data processing agreements compliant with China's Personal Information Protection Law [26].

3 Results

3.1 Phase 1: Pilot Validation

On the 450-message simulated benchmark, the hybrid rule-LLM pipeline achieved F1=0.906 (P=0.994, R=0.833), substantially outperforming the rule-only baseline (F1=0.567, R=0.397). A fine-tuned BERT-base-Chinese established a supervised ceiling at F1=0.965 (5-fold CV). Performance varied by difficulty: easy 1.000, medium 1.000, hard 0.654—all 35 false negatives involved implicit causality or colloquial expressions. Four Chinese LLMs showed model-independent robustness (F1: 0.892–0.950). On 1,792 real clinical messages, the pipeline achieved F1=0.905 ($\Delta = -0.001$ vs. simulated) with zero false positives, and maintained 100% specificity on 3,897 negative control messages from a pharmacy quality-control group. These pilot results established feasibility but were limited by single-annotator benchmarks and the absence of temporal validation, motivating Phase 2.

3.2 Architecture Comparison

Table 1 presents the three pipeline architectures on both datasets.

Table 1: Pipeline architecture comparison on development (n=1,277) and locked test (n=746) sets. 95% bootstrap CI for F1 in brackets.

Dataset	Architecture	P	R	F1	95% CI
Dev	Rule-only	1.000	0.998	0.999	[0.998, 1.000]
	LLM-only	0.971	0.951	0.961	[0.952, 0.969]
	Hybrid	0.972	1.000	0.986	[0.981, 0.991]
Test	Rule-only	1.000	0.998	0.999	[0.998, 1.000]
	LLM-only	0.943	0.992	0.967	[0.957, 0.977]
	Hybrid	0.944	1.000	0.971	[0.961, 0.981]

The rule-only approach achieved F1=0.999 on both sets. **Importantly, this near-perfect performance is setting-specific** to this pharmacovigilance-dedicated group where standardized reporting conventions ensure high vocabulary coverage. On messages with diverse expression patterns—including brand names, abbreviations (e.g., "MTX"), and colloquial symptom descriptions—rule-only F1 dropped to 0.567 in Phase 1, because vocabulary gaps cause missed detections. This 0.432 F1 gap between settings underscores that rule-based approaches are not generalizable without extensive vocabulary engineering. The LLM-only configuration achieved F1=0.961 (95% CI: 0.952–0.969; Dev) and 0.967 (95% CI: 0.957–0.977; Test) without vocabulary engineering. The LLM-only vs. hybrid difference was not significant on the test set (McNemar $\chi^2(1)=3.20$, $p=0.074$). Given the LLM's vocabulary independence, it was used as the primary configuration for subsequent experiments.

3.3 Prompt Optimization and Temporal Validation

Three prompt strategies were compared on the development set (Table S3). The few-shot balanced strategy achieved F1=0.961, far exceeding the strict negative (0.614) and intermediate (0.830) strategies. Enhancement with confidence scoring and chemotherapy guidance raised F1 to 0.977 (P=0.962, R=0.992; FN reduced from 52 to 9; 95% CI in Table S3). On the locked test set, the enhanced prompt achieved F1=0.970 (95% CI: 0.960–0.978; P=0.944, R=0.997), confirming stable generalization ($\Delta=-0.007$).

3.4 Confidence Calibration and Two-Pass Ablation

Confidence scores were well-calibrated (ECE=0.030, Brier=0.039; Figure 2a). The distribution was strongly bimodal (Figure 2b): 85.6% of messages scored 8–10, 14.3% scored 1–3, and only 0.1% fell in the uncertain range (4–7). Accuracy increased monotonically with confidence: 100% at score 1 (n=133) to 99.3% at score 10 (n=454).

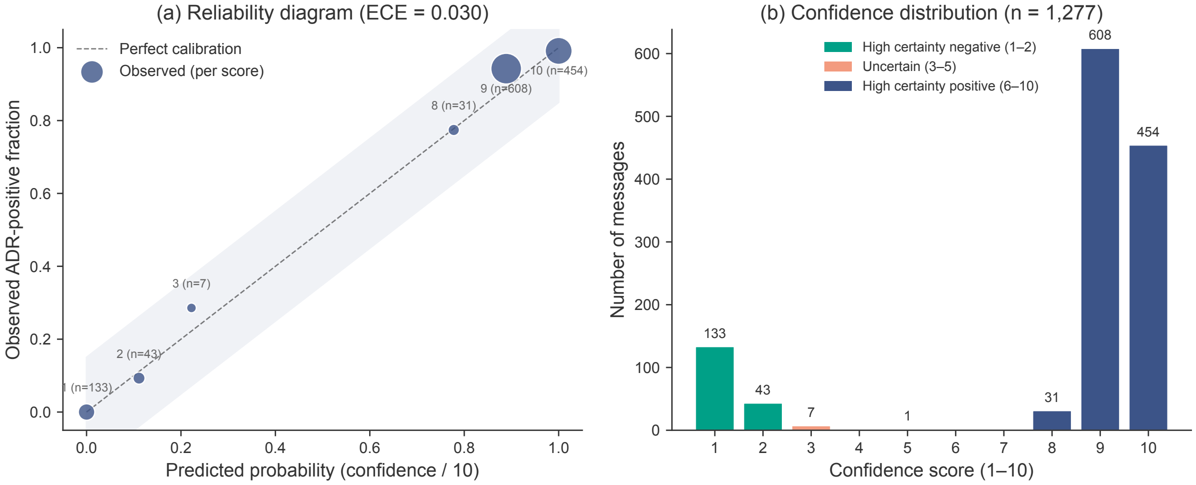


Figure 2: Confidence calibration on the development set (n=1,277). (a) Reliability diagram: each point represents one confidence score (1–10, mapped to predicted probability via $(s-1)/9$; scores with $n < 5$ omitted), sized by sample count; the dashed line indicates perfect calibration and the shaded band marks ±0.15 tolerance. (b) Confidence score distribution showing a bimodal pattern.

With optimized thresholds (θ_H=9, θ_L=2), 39 cases (3.1%) were routed to Pass 2. Two-pass F1=0.973, lower than single-pass F1=0.977 (McNemar $\chi^2(1)$=1.57, p=0.211), as re-examination introduced more errors than it resolved.

3.5 Multi-Turn Context Analysis

Single-turn classification achieved F1=0.967 on the development multi-turn subset (245 conversations, 791 messages), significantly outperforming all context windows ($\chi^2(1)$=12.90–19.53, all McNemar p <0.001 after Bonferroni correction for four comparisons; Table S5). Context increased false positives (32 to 46–50) without recovering any of 9 single-turn false negatives. On the test subset, single-turn F1=0.968 (95% CI: 0.954–0.982).

3.6 Cross-Model Comparison

Four Chinese LLMs showed consistent performance on the development set (Table S11): F1 ranged from 0.958 (Kimi K2.5) to 0.968 (GLM-5) with overlapping confidence intervals, confirming that the pipeline's effectiveness is not model-dependent. Five independent runs at temperature=0.1 yielded CV=0.0005 with 99.2% unanimous agreement (Table S6).

3.7 Specificity on Medical Negative Controls

Overall specificity on 1,000 synthetic messages was 95.5% (Figure S6). Confounding patterns were most challenging (72.5% specificity, 33/120 FP), reflecting the sensitivity–specificity trade-off of the enhanced prompt. Seven of 11 categories achieved ≥98% specificity.

Error analysis of 71 false positives on the real clinical data (development and test sets combined) revealed three dominant patterns: chemotherapy-related expected effects (36.6%), drug–symptom co-occurrences without clear causal framing (31.0%), and other borderline clinical narratives (29.6%). These patterns partially overlap with the confounding-pattern category in synthetic controls, suggesting that drug–symptom co-occurrence without causal framing is the primary challenge for the pipeline.

3.8 Supervised Baseline

A fine-tuned BERT-base-Chinese achieved F1=0.970 on the test set (Table S7), matching the few-shot LLM pipeline (F1=0.970). The bootstrapped F1 difference was −0.001 (95% CI: [−0.014, 0.013]), confirming non-inferiority within a ±0.015 margin.

3.9 Causality Assessment

Automated Naranjo assessment showed no agreement with pharmacist consensus (Table 2): ICC(2,1)=−0.286, κ=−0.010, MAE=3.68. Per-question analysis (Table S8) revealed that 5 of 10 questions were answered "Unknown" in ≥99.5% of cases. Restricting to five IM-answerable questions did not improve agreement. The IMCT showed similarly poor results (κ=−0.135, ICC=−0.236; Figure S4).

Table 2: Causality assessment: automated vs. pharmacist consensus (n=200).

Metric	Naranjo (10Q)	Naranjo (5Q)	IMCT (3Q)
ICC(2,1)	−0.286	−0.286	−0.236
Category κ	−0.010	−0.011	−0.135
MAE	3.68	3.67	N/A[a]

[a]Not applicable; IMCT uses categorical triage (High/Medium/Low).

3.10 Entity Extraction

Entity extraction achieved lenient drug F1=0.780 (Dev) and 0.750 (Test), with lenient symptom F1=0.764 and 0.738 respectively; 95% bootstrap CIs are reported in Table S12. Evaluation was limited to ADR-positive messages with pharmacist entity annotations (n=958); end-to-end performance on the full message stream—including true negatives where no entity should be extracted—was not assessed. Strict match failures primarily reflected systematic naming differences: the LLM translated abbreviations to full generic names (e.g., “MTX”→“methotrexate”). Cross-model extraction was consistent (lenient drug F1: 0.745–0.780; symptom F1: 0.735–0.764; Table S10).

Taken together, these results converge on three findings: (1) the pipeline achieves near-expert ADR detection with well-calibrated confidence scores and model-independent robustness; (2) ADR reports are self-contained within individual messages, making multi-turn modeling unnecessary; and (3) automated causality assessment is structurally infeasible from IM data alone, regardless of instrument design. The Discussion examines the mechanisms and implications of each finding.

4 Discussion

4.1 Principal Findings

This study validated a confidence-calibrated LLM pipeline for ADR detection from clinical IM conversations. Three principal findings emerged.

First, the pipeline achieved near-expert classification with well-calibrated confidence. On the locked test set, the LLM-only configuration reached F1=0.970 (P=0.944, R=0.997), matching the supervised BERT baseline without task-specific training. Confidence scores were well-calibrated (ECE=0.030, Brier=0.039) with a strongly bimodal distribution, providing a principled mechanism for deployment-time review prioritization: high-confidence predictions (scores 8–10, comprising 85.6% of messages) can be accepted with minimal oversight, while the rare uncertain cases can be routed for pharmacist review. Notably, two-pass re-examination of uncertain cases did not improve classification accuracy (McNemar $\chi^2(1)=1.57$, $p=0.211$), suggesting that the primary value of confidence scores lies in triage rather than iterative refinement. Four independent Chinese LLMs all achieved F1$\geq$0.958 with overlapping confidence intervals.

Second, ADR reports in this IM setting are self-contained within individual messages. Single-message classification significantly outperformed all multi-turn configurations (F1=0.967 vs. 0.946–0.950; $p < 0.001$), and the pipeline maintained 95.5% specificity on non-ADR medical content.

Third, automated causality assessment showed no agreement with pharmacist consensus across three instrument variants (ICC≤ -0.236), confirming that the barrier is information asymmetry rather than instrument design.

4.2 Why Single-Turn Classification Outperforms Multi-Turn

The superiority of single-message classification contradicts the general expectation that conversational context improves NLP tasks [8]. Three factors explain this. First, pharmacists reporting suspected ADRs include drug name, symptom, and temporal framing within a single message, making reports operationally self-contained. Second, preceding messages often describe different patients, introducing irrelevant drug–symptom co-occurrences that the LLM misinterprets as ADR signals—explaining the increased false positives (32 to 46–50) with context. Third, the 9 single-turn false negatives involved chemotherapy lab values and ultra-short messages requiring domain interpretation, not contextual information.

This finding has broader design implications: IM-based pharmacovigilance systems on dedicated reporting channels can adopt per-message classification, reducing complexity and latency.

4.3 Causality Assessment: A Structural Mismatch

The Naranjo scale's failure (ICC=-0.286) reflects a mismatch between its design—structured case evaluation with complete medical records [21]—and the limited information in IM messages. Five of 10 questions require data categorically absent from IM text, as confirmed by both LLM and pharmacists answering "Unknown" in $\geq$99.5% of cases.

For the five remaining questions, poor agreement (κ <0.1) reveals that pharmacists draw on tacit clinical knowledge when interpreting brief messages, while the LLM applies this knowledge inconsistently. The IMCT, designed specifically for IM-answerable dimensions, showed comparably poor results (ICC=-0.236). The consistency across all three instruments demonstrates that the barrier is not instrument complexity but a fundamental information asymmetry between what pharmacists infer from clinical context and what can be extracted from individual messages. IM-based pharmacovigilance should therefore be scoped as detection and triage, with causality assessment deferred to formal review.

4.4 Prevalence and Deployment Considerations

The high ADR prevalence in this pharmacovigilance-dedicated group (83.6%) warrants discussion of lower-prevalence settings. Using the observed sensitivity (0.997) and specificity from synthetic medical negative controls (0.955), Bayesian projections (Figure 3) yield: at 10% ADR prevalence, positive predictive value (PPV)=71% and negative predictive value (NPV)>99.9%; at 5% prevalence, PPV=54% and NPV>99.9%. These projections are based on specificity estimated from synthetic medical negative controls and assume that this specificity transfers to authentic non-ADR clinical communication. Real-world specificity may differ if authentic non-ADR messages exhibit patterns not represented in the synthetic corpus; validation on authentic non-ADR IM data is a priority for future work. At 10% prevalence, approximately 7 of 10 flagged messages would be true ADR reports while fewer than 1 in 3,000 unflagged messages would be missed. Confidence scores provide an additional filtering mechanism—restricting to high-confidence alerts (≥9) would increase PPV at the cost of modest recall reduction.

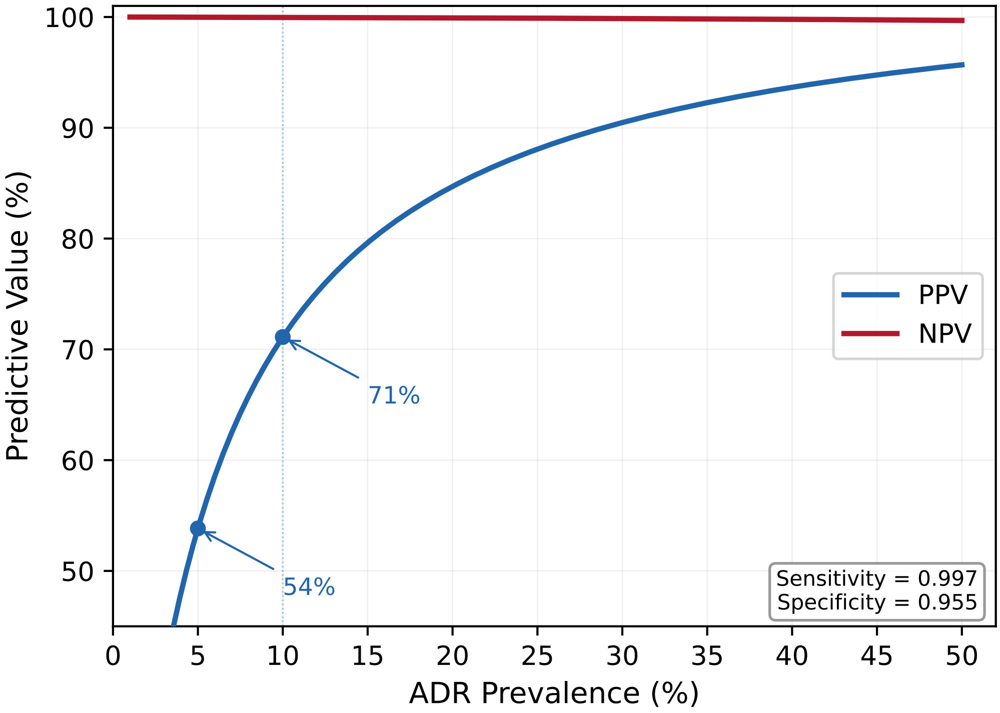


Figure 3: Prevalence-adjusted predictive values using observed sensitivity (0.997) and specificity (0.955). PPV increases steeply with prevalence, reaching 71% at 10% prevalence. NPV remains >99.9% across the displayed range.

In a practical deployment scenario, the pipeline would operate as a continuous monitor on clinical IM groups, flagging suspected ADR messages for pharmacist review in a dashboard ranked by confidence score. In a group generating approximately 50 messages per day at 10% ADR prevalence, the system would flag an estimated 7 messages daily (5 true ADRs, 2 false positives), representing a manageable review workload that could augment existing spontaneous reporting workflows.

4.5 Comparison with Prior Work

Compared to the Phase 1 pilot, Phase 2 demonstrates four improvements: F1 increased from 0.906 to 0.970 on temporally validated data; inference stability improved 10-fold

(CV: 0.0047→0.0005); the evaluation incorporated 5-annotator blind review with temporal splitting; and entity extraction was validated against pharmacist annotations for the first time. The few-shot LLM pipeline matched the supervised BERT ceiling (both F1=0.970; bootstrapped ΔF1=$-$0.001, 95% CI: [$-$0.014, 0.013]), demonstrating that domain-specific prompt engineering with in-context examples eliminates the need for task-specific fine-tuning.

In the broader NLP-based ADR detection literature, our results are consistent with evidence that LLMs achieve near-expert performance on clinical NLP tasks [13, 14, 17]. However, this is the first demonstration on clinical IM text—a source distinct from electronic health records (EHRs), clinical notes, and social media in its brevity and real-time conversational structure. While recent work has applied deep learning NLP to EHR discharge summaries for automated ADR reporting [27], IM-based pharmacovigilance involves unique challenges including ultra-short messages, absence of structured fields, and informal language that differentiate it from EHR-based text mining.

4.6 Limitations

Gold standard and adjudication. The two-stage adjudication process changed 734 of 2,023 labels (36.3%) from ADR$-$ to ADR+, substantially increasing prevalence from 47.4% (majority vote) to 83.6%. While all adjudicated cases met the pre-specified objective criteria (drug name, adverse symptom, and temporal association), this rate of override—particularly 497 cases where 0/5 annotators initially voted positive—warrants scrutiny. The high override rate suggests either that annotators systematically under-recognized certain ADR patterns (e.g., dechallenge narratives, chemotherapy lab values) or that adjudication criteria were overly inclusive. Future studies should consider adjudicator blinding, formal inter-adjudicator agreement measurement, and bidirectional adjudication. The Naranjo gold standard used only three annotators for 200 cases; inter-annotator ICC(2,1)=0.476 indicates moderate reliability for the gold standard itself, compounding uncertainty in the LLM-vs-consensus ICC=$-$0.286 reported in Table 2.

Generalizability and negative controls. This single-center study at a tertiary obstetrics and gynecology hospital limits external validity: the ADR spectrum is specialty-specific, all evaluated models are Chinese-language LLMs, and the pharmacovigilance-dedicated WeChat group has established reporting conventions that may not exist in general clinical communication channels. Furthermore, the synthetic negative controls may not fully capture authentic non-ADR clinical communication; validation on real non-ADR IM data is a priority for future work. Multi-center, multi-specialty, and multi-language validation is planned.

Technical scope. The near-perfect rule-only F1 (0.999) reflects this group's standardized conventions (F1=0.567 on diverse expressions in Phase 1) and should not be generalized. Entity extraction evaluation was limited to ADR-positive messages with annotations (n=958), and enhanced recall on chemotherapy ADRs reduced confounding-pattern specificity to 72.5%. Retrieval-augmented generation with drug-specific knowledge bases may improve discrimination.

Model and temporal constraints. All four evaluated models are Chinese-developed LLMs accessed via cloud APIs; international models (e.g., GPT-4, Claude) were not evaluated due to data residency requirements. API updates or model deprecation could affect reproducibility, and the 3.5-year temporal split may not capture longer-term drift in drug formularies or clinical practices.

Interpretability. The LLM provides natural language reasoning for each classification, but this reasoning is generated alongside the decision rather than derived from a transparent decision process. Feature-level attributions or attention-based explanations that would allow clinicians to verify which message elements drove each classification are not currently available. Future work should explore interpretability methods to increase clinical trust in automated ADR detection.

5 Conclusion

A confidence-calibrated LLM pipeline achieves near-expert ADR detection from clinical IM conversations (test F1=0.970), matching a supervised BERT baseline without task-specific training, with well-calibrated confidence scores (ECE=0.030) and consistent performance across four Chinese LLMs (F1=0.958–0.968). ADR reports are self-contained within individual messages, eliminating the need for multi-turn modeling. Prevalence-adjusted analysis supports deployment as a high-sensitivity screening tool (projected PPV=71% at 10% prevalence, based on synthetic negative controls); these projections and all findings require validation on authentic non-ADR IM data and across multiple centers and specialties. Causality assessment remains fundamentally limited by the information content of IM data regardless of instrument design (ICC≤ -0.236), indicating that these systems should serve as detection and triage tools with causality assessment deferred to formal review. Future work should prioritize multi-center validation and integration with electronic health records for post-detection causality assessment.

Acknowledgments

The authors thank the five clinical pharmacists who contributed to the blind-review annotation and the three pharmacists who provided Naranjo causality consensus scoring.

Author Contributions

D.W. and Z.L. conceived the study, developed the system, conducted the experiments, and wrote the manuscript. W.Y. contributed to data curation and the annotation study. K.Y. contributed to software development. D.Y. contributed to data analysis. Y.Y. and S.J. supervised the project and revised the manuscript. All authors reviewed and approved the final manuscript.

Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Conflicts of Interest

The authors declare no conflicts of interest.

Data Availability

The clinical messaging data contain protected health information and cannot be shared publicly under China's Personal Information Protection Law [26]. The pipeline source code, evaluation scripts, rule vocabulary lists, prompt templates, synthetic negative control generation code, and all statistical analysis code are available at https://github.com/Patrick647/adr-detection-confidence-calibrated. A synthetic demonstration dataset preserving the distributional characteristics of the original data is included to enable independent verification. Researchers seeking access to the de-identified clinical dataset may contact the corresponding author; access will be granted subject to institutional data sharing agreements and ethics committee approval.

References

[1] World Health Organization. Safety of medicines: A guide to detecting and reporting adverse drug reactions. WHO/EDM/QSM/2002.2, 2002.

[2] Ashish K. Jha, Itziar Larizgoitia, Carmen Audera-Lopez, Nittita Prasopa-Plaizier, Hugh Waters, and David W. Bates. The global burden of unsafe medical care: analytic modelling of observational studies. *BMJ Quality & Safety*, 22(10):809–815, 2013. doi: 10.1136/bmjqs-2012-001748.

[3] Ania Syrowatka, Wenyu Song, Mary G. Amato, Dinesh Chakraborty, Katharine Harris, Thomas Hartvigsen, Gretchen P. Jackson, Clemens Scott Kruse, Kathleen M. Mazor, Bill Mclean, et al. Key use cases for artificial intelligence to reduce the frequency of adverse drug events: a scoping review. *The Lancet Digital Health*, 4(2): e137–e148, 2022. doi: 10.1016/S2589-7500(21)00229-6.

[4] Rave Harpaz, Alison Perez, Herbert S. Chase, Raul Rabadan, George Hripcsak, and Carol Friedman. Biclustering of adverse drug events in the FDA's spontaneous reporting system. *Clinical Pharmacology & Therapeutics*, 89(2):243–250, 2011. doi: 10.1038/clpt.2010.285.

[5] Lorna Hazell and Saad A. W. Shakir. Under-reporting of adverse drug reactions: a systematic review. *Drug Safety*, 29(5):385–396, 2006. doi: 10.2165/00002018-200629050-00003.

[6] Elena López-González, Maria Teresa Herdeiro, and Adolfo Figueiras. Determinants of under-reporting of adverse drug reactions: a systematic review. *Drug Safety*, 32 (1):19–31, 2009. doi: 10.2165/00002018-200932010-00002.

[7] Philip Alexander Routledge. Improving the spontaneous reporting of suspected adverse drug reactions: an overview of systematic reviews. *British Journal of Clinical Pharmacology*, 89(8):2377–2385, 2023. doi: 10.1111/bcp.15791.

[8] Jiancheng Ye. Transforming and facilitating health care delivery through social networking platforms: evidences and implications from WeChat. *JAMIA Open*, 7(2): ooae047, 2024. doi: 10.1093/jamiaopen/ooae047.

[9] Rachel M. Murphy, Joanna E. Klopotowska, Nicolette F. de Keizer, Kitty J. Jager, Jan Hendrik Leopold, Dave A. Dongelmans, Ameen Abu-Hanna, and Martijn C. Schut. Adverse drug event detection using natural language processing: a scoping review of supervised learning methods. *PLOS ONE*, 18(1):e0279842, 2023. doi: 10.1371/journal.pone.0279842.

[10] Su Golder, Dongfang Xu, Karen O'Connor, Yunwen Wang, Mahak Batra, and Graciela Gonzalez Hernandez. Leveraging natural language processing and machine learning methods for adverse drug event detection in electronic health/medical records: a scoping review. *Drug Safety*, 48(4):321–337, 2025. doi: 10.1007/s40264-024-01505-6.

[11] Su Golder, Karen O'Connor, Yunwen Wang, Ari Klein, and Graciela Gonzalez Hernandez. The value of social media analysis for adverse events detection and pharmacovigilance: scoping review. *JMIR Public Health and Surveillance*, 10:e59167, 2024. doi: 10.2196/59167.

[12] Oladapo Oyebode and Rita Orji. Identifying adverse drug reactions from patient reviews on social media using natural language processing. *Health Informatics Journal*, 29(1):14604582221136712, 2023. doi: 10.1177/14604582221136712.

[13] Cheng Peng, Xi Yang, Aokun Chen, Kaleb E. Smith, Nima PourNejatian, Anthony B. Costa, Cheryl Martin, Mona G. Flores, Ying Zhang, Tanja Magoc, Gloria Lipori, Duane A. Mitchell, Naykky S. Ospina, Mustafa M. Ahmed, William R. Hogan, Elizabeth A. Shenkman, Jiang Bian, Yonghui Wu, and Yi He. A study of generative large language model for medical research and healthcare. *npj Digital Medicine*, 6: 210, 2023. doi: 10.1038/s41746-023-00958-w.

[14] Zhaoyue Sun, Gabriele Pergola, Byron C. Wallace, and Yulan He. Leveraging ChatGPT in pharmacovigilance event extraction: an empirical study. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, volume 2, pages 344–357, St. Julian's, Malta, 2024. doi: 10.18653/v1/2024.eacl-short.30.

[15] Yiming Li, Jianfu Li, Jingcheng He, and Cui Tao. AE-GPT: using large language models to extract adverse events from surveillance reports—a use case with influenza vaccine adverse events. *PLOS ONE*, 19(3):e0300919, 2024. doi: 10.1371/journal.pone.0300919.

[16] Fan Dong, Wenjing Guo, Jie Liu, Tucker A. Patterson, and Huixiao Hong. BERT-based language model for accurate drug adverse event extraction from social media: implementation, evaluation, and contributions to pharmacovigilance practices. *Frontiers in Public Health*, 12:1392180, 2024. doi: 10.3389/fpubh.2024.1392180.

[17] Mazin Mamun Zitu, Daniel H. Owen, Max E. Hartranft, Jakob D. Schmit, Nicolas

Villegas, James C. Byrd, et al. Large language models for adverse drug events: A clinical perspective. *Journal of Clinical Medicine*, 14(15):5490, 2025. doi: 10.3390/jcm14155490.

[18] Joseph L. Fleiss. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378–382, 1971. doi: 10.1037/h0031619.

[19] J. Richard Landis and Gary G. Koch. The measurement of observer agreement for categorical data. *Biometrics*, 33(1):159–174, 1977. doi: 10.2307/2529310.

[20] An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, et al. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

[21] Claudio A. Naranjo, Usoa Busto, Edward M. Sellers, Paul Sandor, Ivan Ruiz, E. A. Roberts, E. Janecek, Carlos Domecq, and David J. Greenblatt. A method for estimating the probability of adverse drug reactions. *Clinical Pharmacology & Therapeutics*, 30(2):239–245, 1981. doi: 10.1038/clpt.1981.154.

[22] DeepSeek-AI. DeepSeek-V3 technical report. *arXiv preprint arXiv:2412.19437*, 2024.

[23] GLM Team, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Dan Roessler, Jie Xie, Jinghan Zhao, Kai Yu, et al. ChatGLM: a family of large language models from GLM-130B to GLM-4 All Tools. *arXiv preprint arXiv:2406.12793*, 2024.

[24] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 4171–4186, 2019. doi: 10.18653/v1/N19-1423.

[25] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[26] National People's Congress of China. Personal information protection law of the people's republic of china, 2021. Effective November 1, 2021.

[27] Christopher McMaster, Julia Chan, David F. L. Liew, Elizabeth Su, Albert G. Frauman, Wendy W. Chapman, and Douglas E. V. Pires. Developing a deep learning natural language processing algorithm for automated reporting of adverse drug reactions. *Journal of Biomedical Informatics*, 137:104265, 2023. doi: 10.1016/j.jbi.2022.104265.

Supplementary Material

S1. Prompt Template

The enhanced V3 prompt (few-shot balanced with confidence scoring) is summarized below. The complete prompt with all examples is available in the code repository.

> **System instruction (abbreviated):** You are a clinical pharmacist reviewing instant messages from a hospital pharmacy group. Determine whether each message reports an adverse drug reaction (ADR). An ADR-positive message must contain: (1) an identifiable drug name (including brand names, abbreviations, or generic names); (2) a described adverse symptom or clinical sign; and (3) an explicit or implied temporal association. Chemotherapy-related laboratory abnormalities (e.g., bone marrow suppression after chemotherapy) with temporal framing should be classified as ADR-positive.
>
> **Output format:** JSON with fields: `is_adr` (boolean), `confidence` (integer 1–10), `drug_name` (string), `symptoms` (string), `patient_id` (string), `reasoning` (string).
>
> **Examples:** Four annotated examples are provided—two ADR-positive (explicit drug–symptom–temporal pattern; chemotherapy lab abnormality) and two ADR-negative (dosage consultation; disease symptom without drug involvement).

S2. Prompt Optimization

Table 3 presents the comparison of three prompt strategies on the development set (n=1,277). The few-shot balanced strategy was selected for all subsequent experiments. Enhancement with confidence scoring (1–10 integer scale) and domain-specific chemotherapy guidance further improved performance. Prompt versions V1–V2 were discarded during early development; V3–V5 represent the final candidate set.

S3. Medical Negative Control Categories

Table 4 provides descriptions of the 11 medical negative control categories with example messages.

Table 3: Prompt strategy comparison on development set (LLM-only, n=1,277).

Strategy	Description	P	R	F1	95% CI
Few-shot balanced	4 examples (2+, 2−)	0.971	0.951	0.961	[0.952, 0.969]
Intermediate	Relaxed criteria	—	—	0.830	—
Strict negative	High-specificity focus	—	—	0.614	—
Enhanced	+ confidence + chemo guidance	0.962	0.992	0.977	[0.966, 0.980]

Table 4: Medical negative control categories (n=1,000).

Category	n	Description
Confounding pattern	120	Drug + symptom, but expected pharmacological effect
Positive followup	100	Treatment success reports
Treatment planning	100	Future treatment discussions
Disease symptom	90	Symptoms without drug involvement
Dosage consultation	90	Drug dosing inquiries
Laboratory result	90	Lab values without drug–reaction framing
Medication order	90	Prescription and dispensing records
Drug information	80	General pharmacological queries
Drug interaction	80	Drug–drug interaction discussions
Medication inventory	80	Stock and supply management
Patient education	80	Patient counseling content

S4. Multi-Turn Context Analysis

Table 5 presents the multi-turn context window ablation on the development multi-turn subset (245 conversations, 791 messages). Single-turn classification significantly outperformed all multi-turn configurations (McNemar p <0.001 for all comparisons).

Table 5: Multi-turn context window ablation on development multi-turn subset (n=791). McNemar p-values compare each window to single-turn.

Window	P	R	F1	95% CI	FP	FN
0 (single)	0.949	0.985	**0.967**	[0.956, 0.976]	32	9
1	0.921	0.972	0.946	[0.932, 0.958]	50	17
3	0.926	0.975	0.950	[0.938, 0.962]	47	15
5	0.927	0.972	0.949	[0.935, 0.960]	46	17
Full	0.923	0.970	0.946	[0.933, 0.959]	49	18

Of 9 false negatives in the single-turn configuration, none were recovered by any multi-turn context window (window sizes 1, 3, 5, and full conversation). The 9 missed messages comprised: 5 chemotherapy-related lab values, 2 ultra-short messages, and 2 ambiguous causality descriptions. In all cases, the classification error stemmed from the message content itself (domain-specific interpretation) rather than missing contextual information from preceding messages.

S5. Inference Stability

Table 6 presents the inference stability analysis across five independent runs on the development set (n=1,277) at temperature = 0.1.

Table 6: Inference stability across five independent runs (Qwen 3.5 Plus, temperature=0.1, n=1,277).

Run	TP	FP	FN	P	F1
1	1,001	29	53	0.972	0.961
2	1,003	30	51	0.971	0.961
3	1,001	30	53	0.971	0.960
4	1,002	29	52	0.972	0.961
5	1,004	30	50	0.971	0.962
F1 CV					0.0005
Unanimous agreement				99.2%	(1,267/1,277)

Of 1,277 development set messages evaluated across five independent runs, 10 messages (0.8%) showed any classification disagreement. All 10 were borderline cases with short text length (median 18 characters) and ambiguous drug–symptom relationships. The maximum disagreement was 3/5 runs classifying as ADR-positive and 2/5 as ADR-negative, indicating that even unstable messages exhibited narrow decision boundaries.

S6. Supervised Baseline Comparison

Table 7 presents the fine-tuned BERT-base-Chinese supervised baseline alongside the few-shot LLM-only pipeline.

Table 7: Supervised BERT baseline vs. few-shot LLM pipeline on locked test set (n=746).

Model	Approach	P	R	F1	95% CI
BERT-base-Chinese	Fine-tuned (5-fold)	0.949	0.992	0.970	[0.961, 0.980]
Qwen 3.5 Plus	Few-shot LLM-only	0.944	0.997	0.970	[0.960, 0.978]

The fine-tuned BERT model achieved 5-fold cross-validation F1 = 0.983 (SD = 0.007) on the development set, indicating strong in-distribution performance. On the locked test set, both models achieved identical F1 = 0.970 with overlapping confidence intervals, confirming that task-specific fine-tuning provides no additional benefit when domain-specific prompt engineering is applied.

S7. Naranjo Per-Question Analysis

Table 8 presents the per-question agreement between automated LLM assessment and pharmacist consensus for the 10 Naranjo questions.

Table 8: Naranjo per-question agreement between LLM and pharmacist consensus (n=200). IM answerability reflects whether the question can typically be answered from IM text alone.

Q	Question (abbreviated)	Accuracy	κ	IM answerable
1	Prior ADR reports?	0.530	0.100	Medium
2	ADR after drug use?	0.270	−0.113	High
3	Improved after stopping?	0.555	0.006	High
4	Recurred on rechallenge?	0.995	0.000	Very low
5	Alternative causes?	0.080	−0.001	Low
6	Placebo response?	1.000	—	Very low
7	Toxic drug levels?	1.000	—	Very low
8	Dose–response?	1.000	—	Low
9	Prior similar reaction?	0.990	−0.005	Low
10	Objective evidence?	0.520	0.041	Medium

Questions 4, 6, 7, 8, and 9 were answered as “Unknown” by the LLM in $\geq$99.5% of cases, confirming their fundamental unanswerability from IM data. For Q6–Q8, both LLM and pharmacists agreed on “Unknown,” yielding perfect agreement but no discriminative value. The five IM-answerable questions (Q1, Q2, Q3, Q5, Q10) showed poor agreement (κ <0.1), reflecting the information asymmetry between what pharmacists infer from clinical context and what can be extracted from IM messages.

S8. Naranjo Score Distributions

The LLM systematically underscored cases compared to pharmacist consensus, with the modal category being “Possible” (63.5%) for the LLM vs. “Probable” (62.0%) for phar-

Table 9: Naranjo category distribution: LLM vs. pharmacist consensus (n=200).

Category	LLM (%)	Pharmacist (%)
Definite (≥9)	0.0	1.0
Probable (5–8)	6.5	62.0
Possible (1–4)	63.5	31.0
Doubtful (≤0)	30.0	6.0

macists. This reflects the LLM's conservative interpretation of limited IM evidence, defaulting to "Unknown" for questions where pharmacists applied tacit clinical knowledge to infer positive answers.

S9. IMCT Category Distribution

Figure 4 presents the distribution of IM Causality Triage (IMCT) categories comparing automated LLM assessment with pharmacist gold standard.

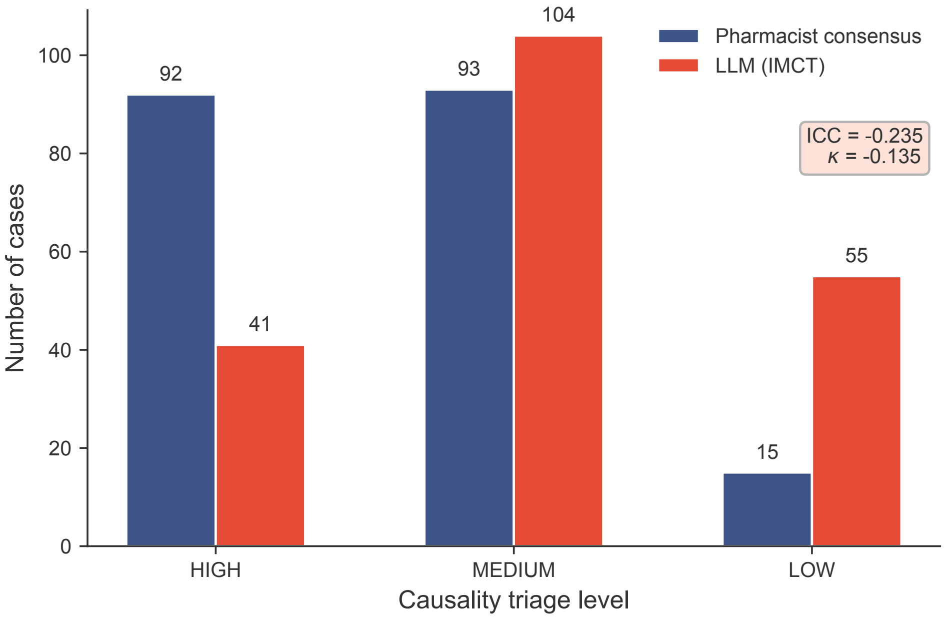


Figure 4: IMCT triage category distribution: automated LLM vs. pharmacist gold standard (n=200). The LLM systematically over-assigns Low (27.5% vs. 7.5%) and under-assigns High (20.5% vs. 46.0%), reflecting conservative interpretation of limited IM evidence.

S10. Cross-Model Entity Extraction

Table 10 and Figure 5 present entity extraction performance across four Chinese LLMs on the development set (n=615 ADR-positive messages).

Table 10: Cross-model entity extraction on development set (n=615 ADR-positive messages). Strict: normalized exact match; Lenient: token-overlap F1.

Model	Drug		Symptom	
	Strict	Lenient	Strict	Lenient
Qwen 3.5 Plus	0.761	0.780	0.803	0.764
GLM-5	0.699	0.776	0.836	0.761
Kimi K2.5	0.725	0.762	0.787	0.751
DeepSeek V3.2	0.698	0.745	0.772	0.735

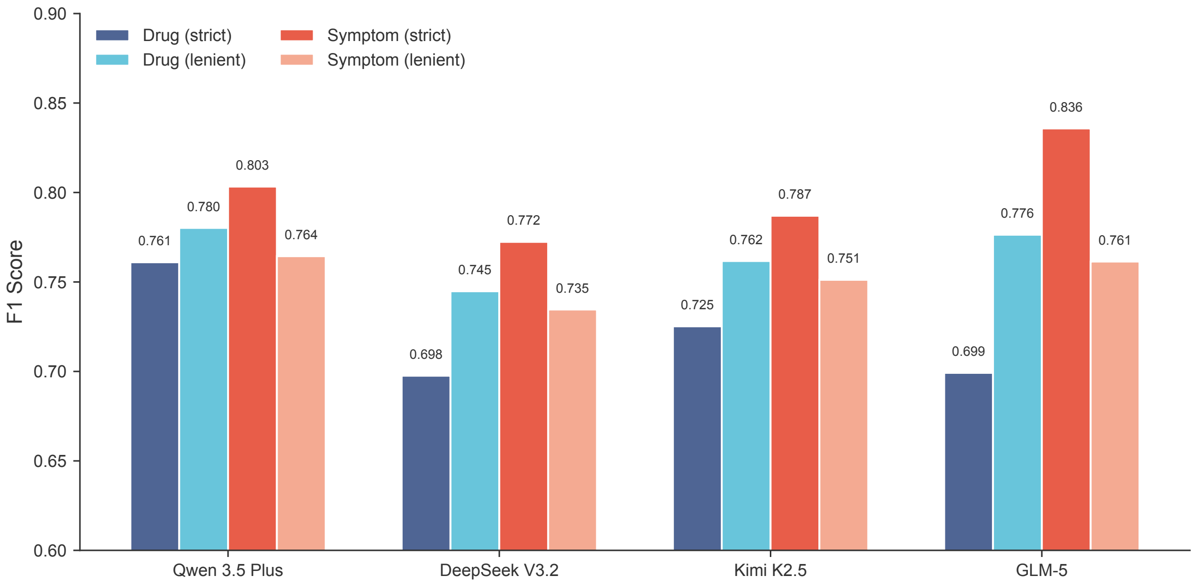


Figure 5: Cross-model entity extraction comparison on development set (n=615). All four models show consistent performance with lenient drug F1 in the range 0.745–0.780 and lenient symptom F1 in the range 0.735–0.764.

All four models achieved comparable extraction performance, with lenient drug F1 ranging from 0.745 to 0.780 and lenient symptom F1 from 0.735 to 0.764. No single model dominated across both entity types: Qwen 3.5 Plus achieved the highest drug extraction while GLM-5 achieved the highest strict symptom match.

S11. Cross-Model ADR Classification

Table 11 presents the cross-model comparison for ADR classification on the development set.

Table 11: Cross-model comparison on the development set (LLM-only, n=1,277).

Model	P	R	F1	95% CI	Spec
GLM-5	0.972	0.963	**0.968**	[0.960, 0.975]	0.870
DeepSeek V3.2	0.967	0.967	0.967	[0.959, 0.974]	0.843
Qwen 3.5 Plus	0.972	0.952	0.962	[0.952, 0.969]	0.870
Kimi K2.5	0.975	0.941	0.958	[0.948, 0.966]	0.883

S12. Entity Extraction

Table 12 presents entity extraction performance against pharmacist annotations.

Table 12: Entity extraction on ADR-positive messages (Dev n=615, Test n=343).

Entity	Split	n	Strict	Lenient	Empty
Drug	Dev	615	0.761	0.780	0.0%
	Test	343	0.706	0.750	0.3%
Symptom	Dev	615	0.803	0.764	0.2%
	Test	343	0.726	0.738	0.3%

Strict match failures primarily reflected systematic naming differences: the LLM translated abbreviations to full generic names (e.g., “MTX”→“methotrexate”) and provided more specific symptom descriptions than the gold standard. For symptoms, strict matching occasionally exceeded lenient matching (e.g., Dev: 0.803 vs. 0.764) because lenient token-overlap scoring penalizes cases where the LLM provides a detailed multi-token description against a terse gold standard label, while strict matching after normalization can still achieve an exact match.

S13. Specificity by Medical Negative Control Category

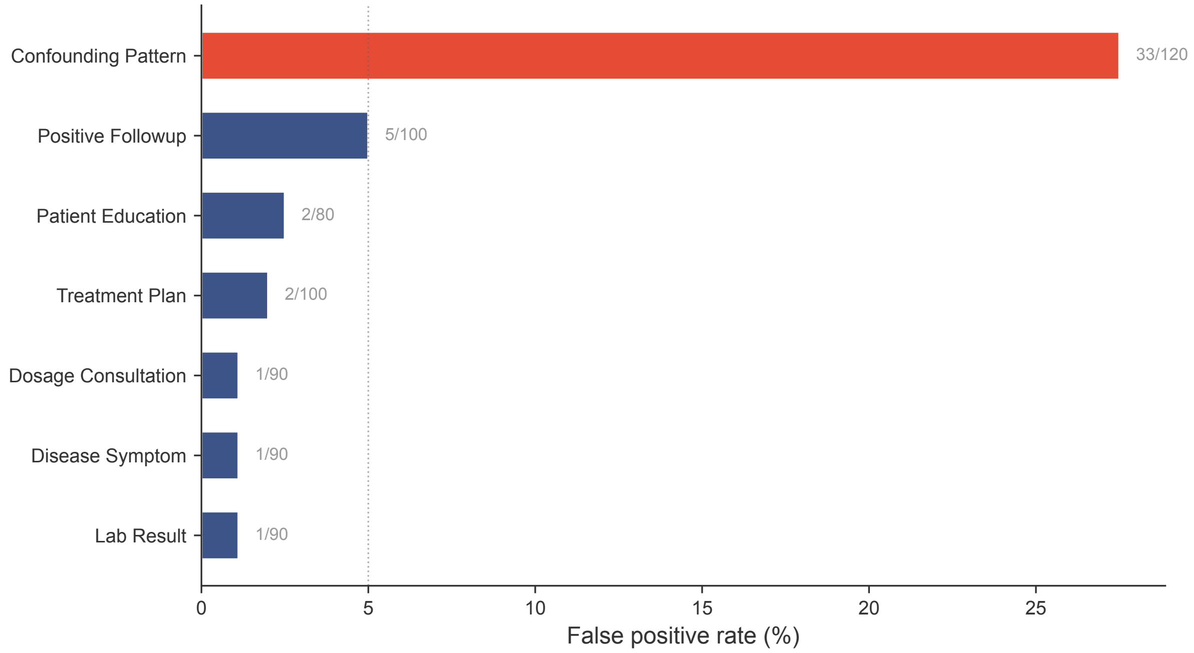


Figure 6: False positive rate by medical negative control category (n=1,000). Four categories with zero false positives (medication inventory, drug information, medication order, drug interaction) are omitted. Confounding patterns account for 73.3% of all false positives.