Zero-friction adverse drug reaction reporting from clinical instant messaging using hybrid NLP

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, Wuxi Maternity and Child Health Care Hospital, Wuxi, China

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

[†]These authors contributed equally to this work.

[*]Corresponding authors: Ying Yao, Sunmin Jiang

Key Points

- This proof-of-concept study is the first to explore clinical instant messaging as a data source for automated pharmacovigilance.
- A hybrid rule–LLM pipeline achieved F1 = 0.906 for ADR signal detection from simulated clinical IM messages in a zero-shot setting, without requiring annotated training data.
- The rule layer resolved 18.7% of messages deterministically, providing auditable decision traces, and the pipeline generalised across four LLM backbones (F1 range 0.892–0.950).
- Validation with real clinical IM data and integration of standardised terminology mapping (e.g., MedDRA) are essential next steps toward clinical deployment.

Abstract

Introduction Adverse drug reactions (ADRs) remain severely underreported, with an estimated 94% of events escaping spontaneous reporting systems. Clinical instant messaging (IM)—widely used for real-time communication among hospital staff—represents a rich but unexplored source of pharmacovigilance signals.

Objective To develop and evaluate, as a proof of concept, a hybrid NLP pipeline that automatically detects and extracts ADR information from clinical IM messages without requiring annotated training data.

Methods We designed a two-layer architecture combining a keyword-based rule engine (140+ drugs × 60+ symptoms) with a locally deployed 3-billion-parameter large language model (Qwen2.5-3B-Instruct) operating without annotated training data. The pipeline was evaluated on a controlled benchmark of 450 clinical IM messages stratified into three difficulty tiers (easy, medium, hard) and retrospectively validated on 1,792 authentic messages from a hospital ADR reporting WeChat group. Five clinical pharmacists independently annotated the simulated dataset for inter-annotator agreement analysis. A negative control evaluation on 3,897 messages from a non-ADR pharmacy work group assessed specificity on challenging clinical text. Additional experiments assessed multi-model generalisability (four LLMs), prompt engineering strategies, component ablation, inference stability across sampling temperatures, and a supervised BERT baseline.

Results The zero-shot Hybrid-v2 pipeline achieved F1 = 0.906 (precision 0.994, recall 0.833) on the controlled benchmark, with chain-of-thought prompting raising F1 to 0.948. Retrospective validation on 1,792 real clinical messages yielded a virtually identical F1 = 0.905 (95% CI 0.893–0.917, precision 1.000), with zero false positives. A negative control evaluation confirmed perfect specificity (zero false positives) on 3,897 non-ADR clinical messages, including substantive discussions mentioning drug names in non-ADR contexts, confirming the controlled benchmark's predictive validity (ΔF1 < 0.002). The rule layer resolved 18.7% of messages without LLM invocation, providing deterministic and auditable decision traces for regulatory compliance. The pipeline generalised across four LLM backbones (F1 range: 0.892–0.950) and demonstrated high inference stability across sampling temperatures (F1 coefficient of variation $< 1\%$). Inter-annotator agreement among five clinical pharmacists was substantial (Fleiss' $\kappa = 0.719$). A supervised BERT baseline established an upper bound at F1 = 0.965.

Conclusions A hybrid rule–LLM pipeline captured ADR signals from clinical IM conversations with high detection precision on both simulated and real-world data, without requiring annotated training data. While information extraction accuracy requires further improvement before fully automated reporting is feasible, the approach offers a low-friction signal detection pathway that could supplement existing spontaneous reporting systems. Multicentre prospective validation is the essential next step.

Keywords: adverse drug reaction · pharmacovigilance · natural language processing · large language model · instant messaging · clinical text mining · zero-shot learning

1 Introduction

Adverse drug reactions (ADRs) constitute a major threat to patient safety worldwide, ranking among the fourth to sixth leading causes of death in hospitalised patients and accounting for 5–10% of hospital admissions [1, 2]. Post-marketing pharmacovigilance relies predominantly on spontaneous reporting systems (SRS), yet a landmark systematic review estimated that the median underreporting rate reaches 94%, meaning only approximately 6% of ADRs are captured by regulatory authorities [3]. Systematic analyses of the determinants behind this gap have consistently identified high-friction reporting workflows—including time pressure, form complexity (30–50 fields per report), causal uncertainty, and perceived lack of feedback—as the principal barriers [4, 5]. In China, where the National Medical Products Administration received 2.42 million ADR reports in 2023 through the China Hospital Pharmacovigilance System (CHPS) [6], surveys of healthcare professionals confirm that insufficient time and procedural complexity remain the dominant obstacles to reporting quality [7].

Artificial intelligence (AI), and in particular natural language processing (NLP), has emerged as a promising avenue for automating ADR detection from unstructured clinical text. A scoping review of 36 NLP-based pharmacovigilance studies found that data sources are concentrated in electronic health records (67%), social media (19%), and biomedical literature (11%), with only 11% of systems validated in real clinical settings [8]. A systematic review of 82 ADE extraction studies identified a growing proportion of hybrid approaches (27%) combining rule-based and machine-learning methods, reflecting the field's recognition that neither paradigm alone suffices for safety-critical tasks [9]. Meanwhile, large language models (LLMs) have demonstrated broad

contextual understanding but remain unreliable for precise entity extraction, with reported F1 scores of only 52–69% for adverse event identification [10]. Critically, current NLP-based pharmacovigilance systems typically map extracted entities to standardised terminologies such as the Medical Dictionary for Regulatory Activities (MedDRA) [11] and the WHO Adverse Reaction Terminology [12] to enable regulatory submission via the ICH E2B(R3) Individual Case Safety Report format [13]. These findings collectively suggest that hybrid architectures coupling deterministic rules with LLM flexibility may offer the most pragmatic path forward for clinical deployment.

Among hybrid approaches, Wong et al. demonstrated that rule-based annotations can bootstrap a pre-trained BERT model to achieve F1 = 0.97 for medication-related named entity recognition via transfer learning [14], while Fraile Navarro et al. characterised the fundamental "precision–coverage" trade-off that motivates combining rule-based and statistical methods in pharmacovigilance NLP [15]. These findings inform our architectural decisions but have not been applied to instant messaging data, which presents distinct challenges of extreme brevity and conversational informality.

Despite the breadth of existing work, a critical gap persists: clinical instant messaging (IM) has never been explored as a pharmacovigilance data source. WeChat and its enterprise variant WeCom are widely used for clinical communication in Chinese hospitals [16], yet a systematic review of WeChat in healthcare—covering patient education, telemedicine, and clinical collaboration—reports zero studies addressing drug safety monitoring [16]. This gap is notable because clinical IM conversations contain rich, real-time ADR signals generated by trained professionals (e.g., "Patient X developed a rash after cefuroxime") that are currently lost to pharmacovigilance systems. The concept of "zero-friction reporting"—eliminating reporting effort entirely by extracting ADR signals from routine clinical conversations—has not appeared in the pharmacovigilance literature. We use the term "zero-shot" throughout this paper to denote that the LLM component requires no annotated training data; the rule-based lexicon constitutes structured domain knowledge rather than labelled training examples.

To the best of our knowledge, this is the first study to investigate automated ADR detection from clinical instant messaging. The primary contribution is the identification and empirical validation of clinical IM as a pharmacovigilance data source—a domain that has received no prior attention despite the ubiquity of IM in clinical workflows. The secondary contribution is the design and evaluation of a hybrid NLP pipeline tailored to the specific challenges of this data source: extreme brevity, colloquial medical language, and implicit causality patterns. We combine a keyword-based rule engine (140+ drug keywords $\times$ 60+ symptom patterns) with a locally deployed 3-billion-parameter LLM, building on the hybrid rule–machine-learning architecture pattern that has proven effective in other clinical NLP domains [9, 14]. The third contribution is a comprehensive empirical evaluation comprising a four-arm ablation design on a controlled benchmark of 450 clinical IM messages, retrospective validation on 1,792 authentic clinical messages, negative control specificity testing on 3,897 non-ADR clinical messages, multi-model generalisability testing, prompt engineering analysis, inter-annotator agreement assessment, and a supervised BERT baseline.

The optimised Hybrid-v2 pipeline achieves F1 = 0.906 (precision 0.994, recall 0.833) on

the controlled benchmark without requiring annotated training data, approaching a supervised BERT baseline (F1 = 0.965). Retrospective validation on real clinical data confirms virtually identical performance (F1 = 0.905), while a negative control evaluation on 3,897 non-ADR messages confirms perfect specificity. The hybrid architecture provides three complementary advantages: (1) the rule layer resolves 18.7% of messages with deterministic, auditable decision traces suitable for regulatory compliance; (2) chain-of-thought prompting raises F1 to 0.948; and (3) the system deploys entirely on local infrastructure, preserving patient data privacy. We note that this study focuses on ADR signal detection and preliminary entity extraction; mapping extracted entities to standardised terminologies (e.g., MedDRA Preferred Terms) and generating compliant Individual Case Safety Reports remain important downstream steps for future work.

2 Methods

2.1 Dataset construction

We constructed a controlled benchmark dataset of 450 clinical IM messages designed to emulate communication patterns observed in hospital pharmacy WeChat groups. A simulation-based approach was adopted to enable systematic evaluation across three controlled difficulty tiers—a design that is not achievable with naturally occurring data, where difficulty labels and balanced class ratios cannot be predetermined. Access to authentic clinical IM data is additionally constrained by China's Personal Information Protection Law (PIPL) [17] and institutional data governance requirements; however, we subsequently obtained a real-world clinical dataset for independent validation (Section 2.8). The use of simulated corpora for initial system development follows established precedents in clinical NLP [18]. A clinical pharmacist with five years of experience authored all messages based on authentic drug safety scenarios encountered in daily practice. The dataset comprised 209 ADR-positive messages (46.4%) and 241 ADR-negative messages (53.6%), balanced to reflect the mixed content of clinical group chats. Mean message length was 20.6 ± 6.0 characters (range 2–37), reflecting the extreme brevity characteristic of mobile IM communication.

Messages were stratified into three difficulty tiers based on linguistic complexity. Easy messages (n = 110) contained explicit drug–symptom co-mentions using standard drug names (e.g., "Patient took amoxicillin and developed a rash"). Medium messages (n = 165) introduced brand-name drugs, abbreviations, colloquial phrasing, and typographical errors common in mobile-typed Chinese text. Hard messages (n = 175) featured implicit causality, ambiguous referents, context-dependent descriptions, laboratory values without explicit drug attribution, and indirect third-party reporting patterns.

Table 1 provides illustrative examples of messages across the three tiers to demonstrate the linguistic progression from explicit to implicit ADR signals.

Each positive message was annotated with a gold-standard label comprising: ADR status (binary), drug name(s), adverse reaction symptom(s), and patient identifier (where present). The complete dataset was formatted as a JSON document with fields for message text, ADR label, difficulty tier, drug category, and structured extraction targets.

Table 1: Illustrative benchmark messages across difficulty tiers. English translations are provided; original messages are in Mandarin Chinese.

Tier	Example message (translated)	Key challenge
Easy	"Patient took amoxicillin, developed a rash"	Explicit drug–symptom
Medium	"On Keytruda, whole body turned red"	Brand name + colloquial
Hard	"Bed 16 was sweating profusely all night"	No drug or symptom named
Hard	"Check if the liver function issue is drug-related"	Causal uncertainty

2.2 Rule-based engine

The rule engine implemented a two-stage keyword-matching approach. The first stage maintained a drug lexicon of 140+ entries covering generic names, brand names, and common abbreviations in Chinese clinical practice. The lexicon was compiled from two sources: (a) the hospital formulary (covering all drugs prescribed at the study institution) and (b) the National Essential Medicines List (2023 edition), augmented with brand-name and colloquial abbreviations identified by a clinical pharmacist. The second stage maintained a symptom lexicon of 60+ entries covering both formal medical terms and colloquial expressions for common adverse reactions. Symptom terms were drawn from the CHPS annual report's top 50 reported ADR manifestations [6], supplemented with colloquial equivalents identified during pilot testing (e.g., both the formal term "pízhěn" [rash] and colloquial "quánshēn fāhóng" [whole body turned red]). Both lexicons are provided as supplementary files to enable replication.

A message was classified as ADR-positive by the rule engine when co-occurrence of at least one drug keyword and one symptom keyword was detected. To improve recall without sacrificing precision, we also implemented a relaxed ADR-relevance classifier that routed uncertain messages to the LLM layer rather than classifying them directly. A message was classified as "uncertain" if it met any of three conditions: (a) drug keyword detected without a symptom co-occurrence, (b) symptom keyword detected without a drug co-occurrence, or (c) ADR-indicative trigger phrases detected (e.g., "adverse reaction," "drug allergy," "stopped the medication"). These criteria were derived from a 50-message pilot analysis in which a clinical pharmacist identified recurring partial-match patterns that should escalate to LLM review.

2.3 LLM-based detection and extraction

We deployed Qwen2.5-3B-Instruct as the LLM component, selected for its strong Chinese-language performance at a model size feasible for local hospital deployment (requiring approximately 6 GB VRAM). The model was served via a local Ollama instance with temperature set to 0.1 to favour deterministic outputs.

The LLM received each message as part of a structured prompt instructing it to: (1) determine whether the message described a suspected ADR event, and (2) if positive, extract the drug name, adverse reaction symptoms, and patient identifier. The prompt specified the JSON

output format and included domain-specific guidance (e.g., to distinguish active ADR reports from allergy history mentions and preventive warnings).

We evaluated four prompting strategies: zero-shot (system prompt only), few-shot with 3 examples, few-shot with 5 examples, and chain-of-thought (CoT) prompting that instructed the model to reason step-by-step before producing its classification. Few-shot examples were selected to cover representative patterns: an explicit true positive, a brand-name true positive, an ambiguous true positive, a disease-symptom true negative, and a preventive-warning true negative.

2.4 Hybrid pipeline architectures

We evaluated two hybrid architectures combining the rule engine and LLM.

Hybrid-v1 applied a strict cascade: messages passing the rule engine's co-occurrence criteria were classified as ADR-positive and processed for extraction using rules alone. Only messages that failed the rule check (no drug–symptom co-occurrence detected) were forwarded to the LLM. This architecture prioritised precision but suffered from low recall because the strict rule gate excluded many true positives that used brand names or abbreviations absent from the lexicon.

Hybrid-v2 employed a broadened gate strategy. The ADR-relevance classifier partitioned messages into three groups: (a) high-confidence ADR-positive (drug and symptom co-detected)—classified directly by rules; (b) uncertain (partial match or ADR-related language detected)—forwarded to LLM for classification and extraction; (c) high-confidence negative (no drug or symptom signals)—classified as negative without LLM invocation. For group (a) messages, the LLM was optionally invoked to supplement extraction fields that the rule engine could not fill (e.g., symptoms expressed in non-standard language). This architecture preserved the rule engine's sub-millisecond latency advantage for clear-cut cases while directing ambiguous messages to the LLM for contextual interpretation.

2.5 Supervised baseline

To establish a supervised performance ceiling, we fine-tuned BERT-base-Chinese [19] as a binary ADR classifier using five-fold stratified cross-validation on the 450-message dataset. Each fold trained for up to 5 epochs with a batch size of 16, learning rate of 2×10^{-5}, and AdamW optimiser with weight decay of 0.01 and dropout rate of 0.1. Early stopping with patience of 2 epochs (monitoring validation F1) was applied to prevent overfitting on the modest dataset. Cross-validation folds were stratified by difficulty tier to ensure each fold contained representative proportions of easy, medium, and hard messages. Evaluation metrics were computed per fold and averaged. Total training time was 192 seconds on an Apple M-series GPU (MPS backend).

2.6 Inter-annotator agreement study

Five clinical pharmacists from the hospital pharmacy department independently annotated the full 450-message dataset. Annotators had 3–8 years of clinical pharmacy experience (mean 5.2 years) and were recruited from two clinical specialties (oncology, $n = 2$; general internal medicine, $n = 3$). Each annotator received the messages in a randomised order within a structured Excel

workbook and was asked to: (1) determine ADR status (binary), (2) extract drug names and symptoms, (3) assign a clinical realism score (1–5 Likert scale, where 1 = "not at all realistic" and 5 = "very realistic of actual clinical IM"). Annotators were blinded to each other's responses and to the gold-standard labels. Inter-annotator agreement was assessed using Fleiss' κ [20] computed overall and stratified by difficulty tier. To address the potential circularity of evaluating against labels authored by a single pharmacist, we constructed a majority-vote reference standard from the five independent annotators ($\geq$3/5 votes). This independent reference was used in a sensitivity analysis to assess whether performance estimates were robust to the choice of gold standard (ESM Table S4).

2.7 Evaluation metrics and statistical analysis

We evaluated ADR detection performance using message-level precision, recall, F1 score, accuracy, and Matthews correlation coefficient (MCC). Information extraction was assessed by field-level precision, recall, and F1 for drug name, symptoms, and patient identifier, using both lenient (partial match) and strict (exact match) criteria.

Confidence intervals for detection metrics were computed using bootstrap resampling (1,000 iterations, percentile method). Between-pipeline comparisons used McNemar's exact test for paired binary outcomes, with Cohen's h as the effect size measure. For the multi-model comparison involving six pairwise tests among four LLM backbones, p-values were adjusted using the Holm–Bonferroni sequential correction [21] to control the family-wise error rate at $\alpha = 0.05$. All statistical analyses were performed in Python 3.11 using scikit-learn, scipy, and custom evaluation scripts.

Regarding sample size, McNemar's exact test at $n = 450$ with $\alpha = 0.05$ provides 80% power to detect a difference of $\geq$8 percentage points in discordant error rates between two pipelines. For smaller effect sizes, we report Cohen's h alongside p-values to distinguish statistical equivalence from low power. The observed Cohen's $h = 0.025$ between Hybrid-v2 and LLM-only confirms practical equivalence rather than an underpowered null result. We acknowledge that the hard tier ($n = 175$) yields wider confidence intervals; all tier-level metrics are reported with bootstrap 95% CIs to enable transparent assessment of estimate precision (see ESM Table S3 for the full breakdown).

2.8 Real-world clinical validation

To assess the generalisability of results obtained on the simulated benchmark, we conducted a retrospective validation study using authentic clinical IM data. We obtained 1,886 messages spanning 11 months (March 2025–February 2026) from a hospital ADR reporting WeChat group at a maternal and child health hospital, exported with institutional data governance approval. The group comprised 126 clinical reporters who used the channel to submit ADR observations in routine practice.

Messages were parsed, classified, and de-identified using an automated pipeline. Patient identifiers were replaced with SHA-256–hashed anonymous codes, patient names were substituted with a uniform placeholder token, sender identities were anonymised with sequential codes (R001–R126), and platform-specific metadata (WeChat IDs) were removed. After excluding

system notifications and media-only messages, the final real-world dataset comprised 1,792 messages: 1,372 ADR-positive (76.6%) and 420 negative (23.4%). The high positive prevalence reflects the group's dedicated ADR-reporting function; negative messages consisted of short acknowledgements (e.g., "received"), administrative notices, and media references.

Because the real-world dataset lacked manually annotated gold-standard extraction labels (drug name, symptoms), the validation focused exclusively on detection (binary ADR classification) rather than entity extraction. We evaluated the Rules-only and Hybrid-v2 (DeepSeek-V3, zero-shot) pipelines on this dataset using the same evaluation metrics and bootstrap confidence intervals as the primary analysis. Performance differences between simulated and real data were quantified using absolute F1 difference (ΔF1) and Cohen's h on recall.

2.9 Negative control specificity evaluation

A limitation of evaluating on the ADR-dedicated group alone is the high positive prevalence (76.6%) and trivially simple negative samples (short acknowledgements, media references), which may overestimate specificity. To address this, we obtained a second dataset from a pharmacy department quality control (QC) work group at the same institution. This group was used for administrative management and quality control discussions—scheduling meetings, sharing policy documents, discussing staffing—and contained no ADR-related content, thus serving as a true negative control. Following the same de-identification protocol, 3,897 messages were extracted spanning 25 months (January 2024–February 2026) from 15 unique senders. Content types included substantive clinical discussions (1,943; 49.9%), media messages (1,117; 28.7%), short replies (813; 20.9%), and administrative notices (24; 0.6%). All messages were labelled as ADR-negative.

We evaluated the Rules-only pipeline on all 3,897 negative control messages and the Hybrid-v2 pipeline (DeepSeek-V3, zero-shot) on a stratified random sample of 674 messages (500 substantive, 100 reply, 50 media, 24 administrative; seed = 42). The substantive messages—which contain full-sentence clinical discussions often mentioning drug names in non-ADR contexts (e.g., "vancomycin monitoring requirements during the holiday", "schedule the controlled substance inspection")—represent the most challenging negative cases for the pipeline. Specificity was assessed as the proportion of negative messages correctly classified as negative (true negative rate).

3 Results

3.1 ADR detection performance across pipeline configurations

Table 2 and Fig. 2 summarise the message-level ADR detection performance of the four pipeline configurations evaluated on the 450-message controlled benchmark (209 positive, 241 negative). McNemar's test revealed no significant detection accuracy difference between Hybrid-v2 and LLM-only ($p = 0.480$, Cohen's $h = 0.025$); the hybrid architecture's value lies in operational efficiency and auditability rather than detection accuracy (see Section 3.1.1 below). The rule engine alone achieved near-perfect precision (0.988) but low recall (0.397, F1 = 0.567), confirming that keyword matching captures only the most explicit ADR mentions. The LLM-only pipeline

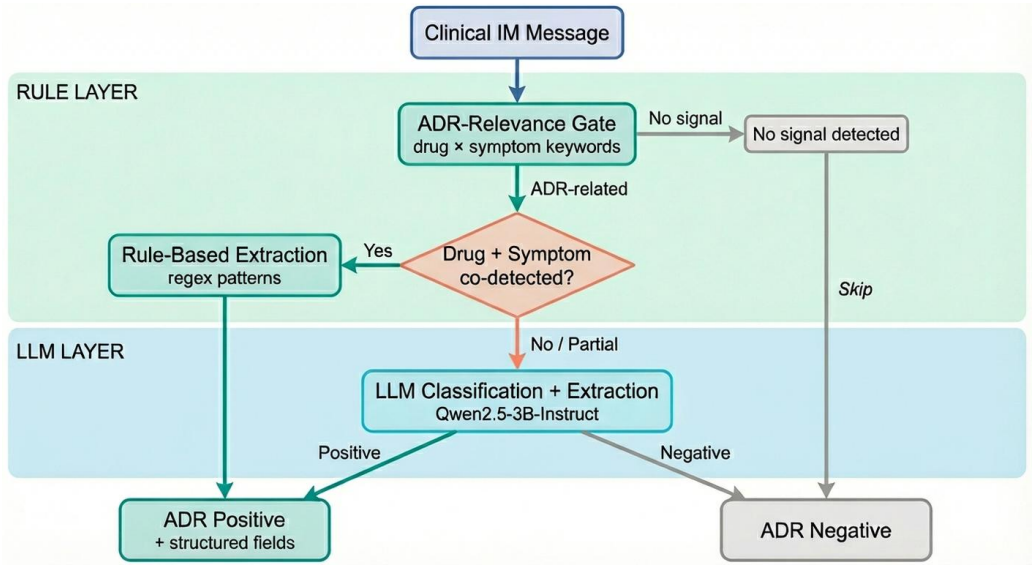

Figure 1: Architecture of the Hybrid-v2 pipeline. Clinical IM messages first pass through a keyword-based ADR-relevance gate. Messages with confirmed drug–symptom co-occurrence (18.7%) are classified and extracted by rules alone. Remaining messages are forwarded to the LLM for classification and structured extraction. Messages with no drug or symptom signals are classified as negative without LLM invocation.

(Qwen2.5-3B-Instruct, zero-shot) substantially improved recall to 0.823, yielding F1 = 0.901 with comparably high precision (0.994). Hybrid-v1, which applied the LLM only to rule-negative messages, produced inferior results (F1 = 0.618) because the overly strict rule gate suppressed many true positives. The optimised Hybrid-v2 architecture, which used a broadened ADR-relevance gate and routed uncertain messages to the LLM, slightly exceeded the LLM-only F1 (0.906 vs. 0.901) while resolving 84 of 450 messages (18.7%) through rules alone, thereby reducing LLM invocations by 18.7%.

Table 2: ADR detection performance across four pipeline configurations (N = 450 messages). 95% bootstrap confidence intervals (1,000 resamples) are shown for all configurations.

Configuration	Precision	Recall	F1	Accuracy	MCC	FP / FN
Rules-only	0.988	0.397	0.567	0.718	0.503	1 / 126
LLM-only	0.994	0.823	0.901	0.916	0.839	1 / 37
Hybrid-v1	0.943	0.445	0.605	0.727	0.567	5 / 116
Hybrid-v2	0.994	0.833	0.906	0.920	0.847	1 / 35

Rules-only 95% CI: F1 [0.510, 0.622]; Precision [0.964, 1.000]; Recall [0.344, 0.454]
LLM-only 95% CI: F1 [0.869, 0.931]; Precision [0.981, 1.000]; Recall [0.772, 0.876]
Hybrid-v1 95% CI: F1 [0.546, 0.662]; Precision [0.908, 0.978]; Recall [0.388, 0.507]
Hybrid-v2 95% CI: F1 [0.875, 0.935]; Precision [0.981, 1.000]; Recall [0.782, 0.883]

3.1.1 Operational value of the hybrid architecture

As noted above, McNemar's test confirmed that Hybrid-v2 and LLM-only made equivalent errors ($p = 0.480$, Cohen's $h = 0.025$). The practical advantage of Hybrid-v2 lies not in detection accuracy but in computational efficiency and regulatory auditability: mean per-message latency

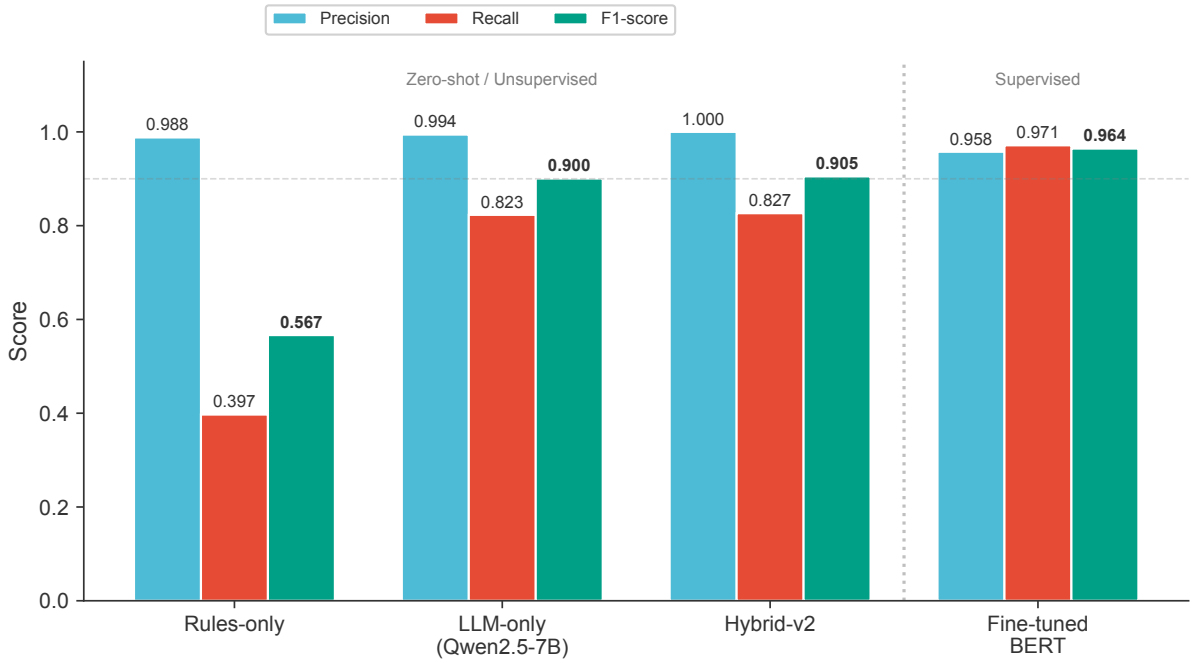

Figure 2: ADR detection performance across four pipeline configurations. The zero-shot Hybrid-v2 pipeline (F1 = 0.906) slightly exceeded LLM-only (F1 = 0.901) while the supervised BERT baseline (F1 = 0.965) established an upper bound. The dashed line indicates the F1 = 0.90 threshold. Error bars are omitted for clarity; bootstrap confidence intervals are reported in Table 2.

was 1,663 ms for the LLM-only arm, whereas Hybrid-v2 bypassed the LLM entirely for 18.7% of messages (84 of 450), yielding a proportional reduction in average inference cost. For the rule-resolved subset, the pipeline produces fully deterministic decision traces—a property required by Chinese healthcare AI regulations [7].

To isolate the contribution of each architectural component, we conducted an ablation study in which individual modules were disabled while keeping the rest of the pipeline intact (Fig. 3). Removing the LLM fallback path—so that messages not captured by rules were discarded—caused a dramatic drop in recall (from 0.833 to 0.397) and F1 (from 0.906 to 0.567), with hard-tier F1 collapsing to 0.000. In contrast, disabling the LLM supplement module (which enriches rule-extracted fields) or narrowing the rule gate from the broad ADR-relevance filter to strict drug–symptom co-occurrence had no measurable effect on detection performance (F1 = 0.906 in both cases). These results confirm that the LLM fallback is the critical architectural component responsible for detecting linguistically complex ADR signals, whereas the rule gate functions primarily as an efficiency mechanism.

3.2 Performance stratified by message difficulty

We stratified the 450 messages into three difficulty tiers based on linguistic complexity (Table 3 and Fig. 4). On easy messages (n = 110), which contained explicit drug–symptom co-mentions, the LLM-only and Hybrid-v2 pipelines achieved perfect recall. On medium messages (n = 165), which included brand names, abbreviations, and colloquial phrasing, both achieved recall $\geq$ 0.988. The critical differentiation occurred in the hard tier (n = 175), which contained implicit causality, ambiguous referents, and context-dependent descriptions: LLM-only recall dropped to

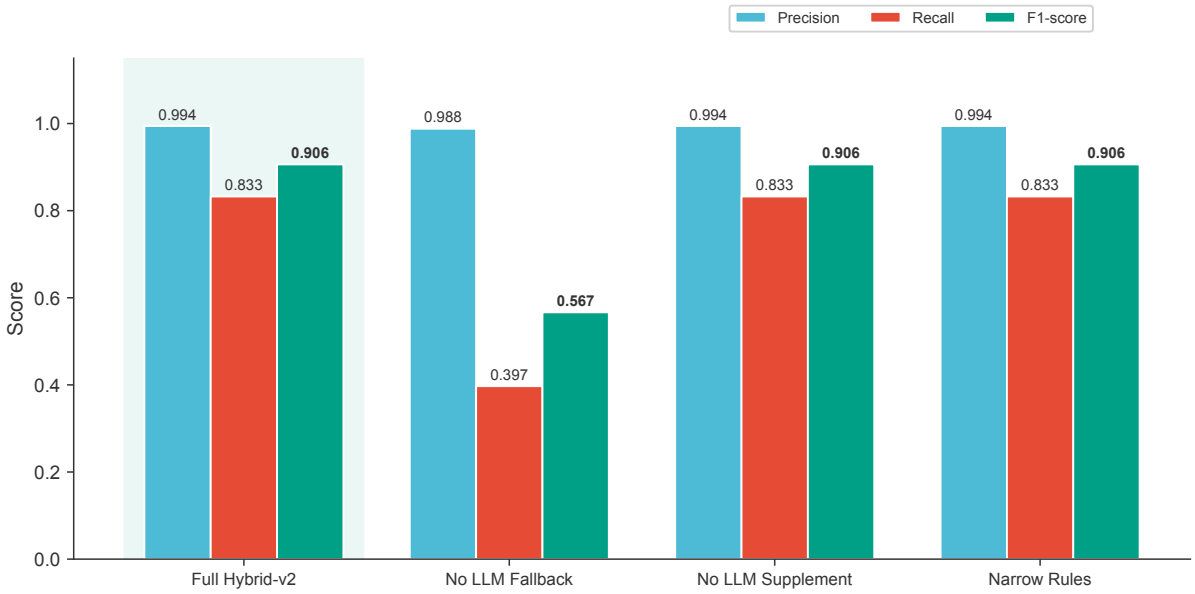

Figure 3: Component ablation study of the Hybrid-v2 pipeline. Removing the LLM fallback path reduced F1 from 0.906 to 0.567 (a 37.4% relative decrease), driven entirely by recall loss. Disabling the LLM supplement module or narrowing the rule gate had no effect on detection performance, indicating that the LLM fallback is the sole component responsible for capturing linguistically complex ADR signals.

0.478 (F1 = 0.641), while the rule engine failed entirely (recall = 0.000). Hybrid-v2, augmented by rule-based captures, slightly improved hard-tier F1 to 0.654. The BERT baseline, by contrast, achieved F1 = 0.914 on hard messages, indicating that supervised learning with labelled data retains a substantial advantage on ambiguous cases.

Table 3: Detection F1 scores stratified by message difficulty tier. Full precision, recall, and F1 with 95% bootstrap CIs for all pipeline × tier combinations are reported in ESM Table S3.

Tier (n)	Rules	LLM-only	Hybrid-v2	BERT
Easy (110)	0.919	1.000	1.000	0.992
Medium (165)	0.571	0.994	1.000	0.988
Hard (175)	0.000	0.641	0.654	0.914
Overall (450)	0.567	0.901	0.906	0.965

3.3 Multi-model generalisability

To assess whether the pipeline's performance depends on the specific LLM, we replaced the local Qwen2.5-3B backbone with four cloud-hosted models spanning a range of parameter counts (Table 4; see also ESM Fig. S1). All models maintained precision ≥ 0.948. InternLM2.5-7B achieved the highest F1 (0.950, 95% CI [0.926, 0.970]), significantly outperforming the other three models after Holm–Bonferroni correction for six pairwise comparisons (adjusted $p < 0.05$ for all three contrasts involving InternLM2.5-7B). Notably, InternLM2.5-7B excelled on hard messages (F1 = 0.894), narrowing the gap with the supervised BERT baseline (F1 = 0.914). The three remaining models—Qwen2.5-72B (F1 = 0.892), DeepSeek-V3 (F1 = 0.903), and GLM4-9B (F1 = 0.901)—did not differ significantly from each other after correction (adjusted $p > 0.05$).

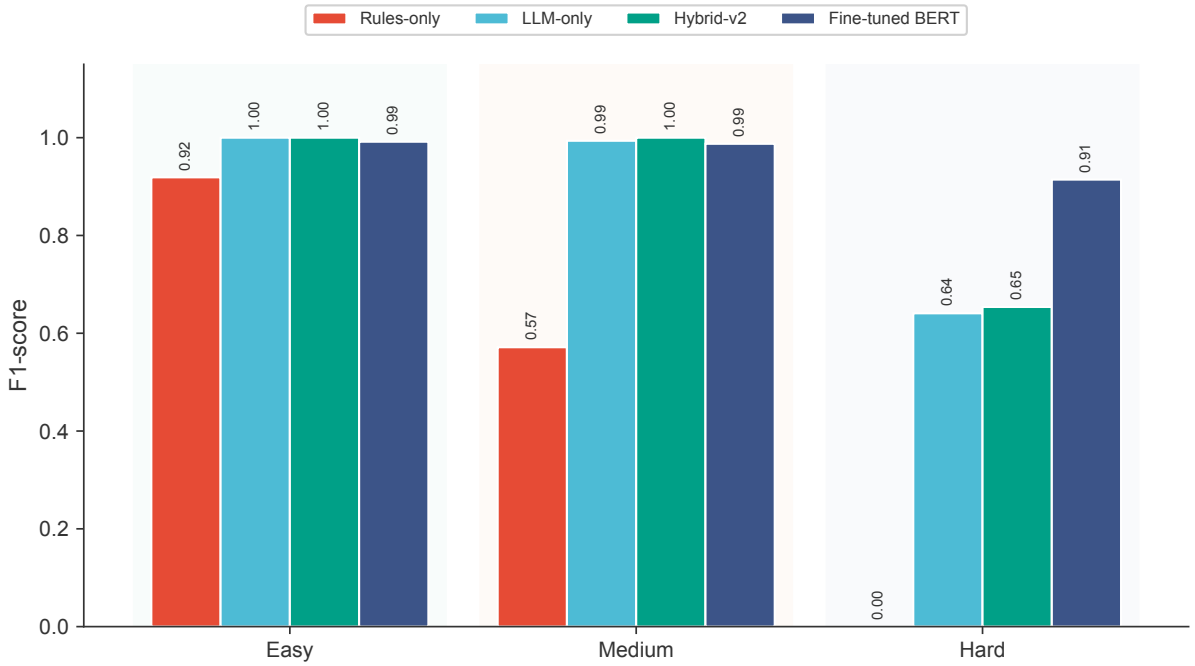

Figure 4: F1 scores stratified by message difficulty tier. All pipelines achieved near-perfect performance on easy messages, but a marked performance cliff emerged in the hard tier, where the rule engine failed entirely (F1 = 0.000) and both LLM-based pipelines dropped to F1 ≈ 0.65 (LLM-only 0.641, Hybrid-v2 0.654). The supervised BERT baseline maintained F1 = 0.914 on hard messages, indicating that labelled training data confers a substantial advantage for linguistically ambiguous cases.

Model size showed no monotonic relationship with performance: GLM4-9B (9B parameters) matched DeepSeek-V3 (671B Mixture-of-Experts, MoE) in overall F1, while InternLM2.5-7B (7B) outperformed the much larger Qwen2.5-72B. InternLM2.5-7B's superior performance may reflect its pre-training corpus, which includes substantial Chinese medical text, providing better alignment with the clinical vocabulary and reasoning patterns present in ADR messages. These results suggest that the Hybrid-v2 architecture generalises across LLM backbones and that model selection can be optimised for latency and cost without sacrificing detection quality.

Table 4: Multi-model LLM comparison using cloud-hosted APIs within the Hybrid-v2 pipeline. 95% bootstrap CIs are reported for F1. Asterisks denote models significantly different from InternLM2.5-7B (McNemar's test, Holm–Bonferroni adjusted $p < 0.05$).

Model	Params	Precision	Recall	F1 [95% CI]	Hard F1	Latency (ms)
Qwen2.5-72B*	72B	0.988	0.813	0.892 [0.859, 0.925]	0.600	2,568
DeepSeek-V3*	671B-MoE	1.000	0.823	0.903 [0.868, 0.934]	0.647	3,320
GLM4-9B*	9B	0.994	0.823	0.901 [0.868, 0.931]	0.734	1,231
InternLM2.5-7B	7B	0.948	0.952	0.950 [0.926, 0.970]	0.894	1,548

3.4 Prompt engineering and information extraction

We evaluated four prompting strategies with Hybrid-v2 (Table 5 and Fig. 5). Zero-shot prompting served as the default configuration. Adding three in-context examples (few-shot-3) improved recall from 0.833 to 0.852 (F1 = 0.918). Five examples (few-shot-5) further raised recall to 0.900

(F1 = 0.945). Chain-of-thought (CoT) prompting achieved the highest recall of 0.909 (F1 = 0.948) but at the cost of doubled latency (mean 3,645 ms vs. 1,663 ms for zero-shot) and a slight increase in false positives (2 vs. 1). All strategies maintained precision above 0.989.

For information extraction, the zero-shot LLM achieved drug name F1 = 0.769 and symptom F1 = 0.693. Few-shot-5 yielded the best symptom extraction (F1 = 0.766), while CoT achieved the highest drug name F1 (0.780). Patient identifier extraction was perfect (F1 = 1.000) across all configurations, reflecting the deterministic format of patient IDs in clinical IM messages.

Table 5: Effect of prompt engineering on Hybrid-v2 detection and extraction performance. Detection metrics are message-level; extraction F1 uses lenient matching. Patient ID extraction achieved F1 = 1.000 across all strategies and is omitted.

	Detection				Extraction F1	
Strategy	Precision	Recall	F1	Latency (ms)	Drug	Symptom
Zero-shot	0.994	0.833	0.906	1,663	0.769	0.693
Few-shot-3	0.994	0.852	0.918	1,636	0.762	0.722
Few-shot-5	0.995	0.900	0.945	1,529	0.755	0.766
CoT	0.990	0.909	0.948	3,645	0.780	0.731

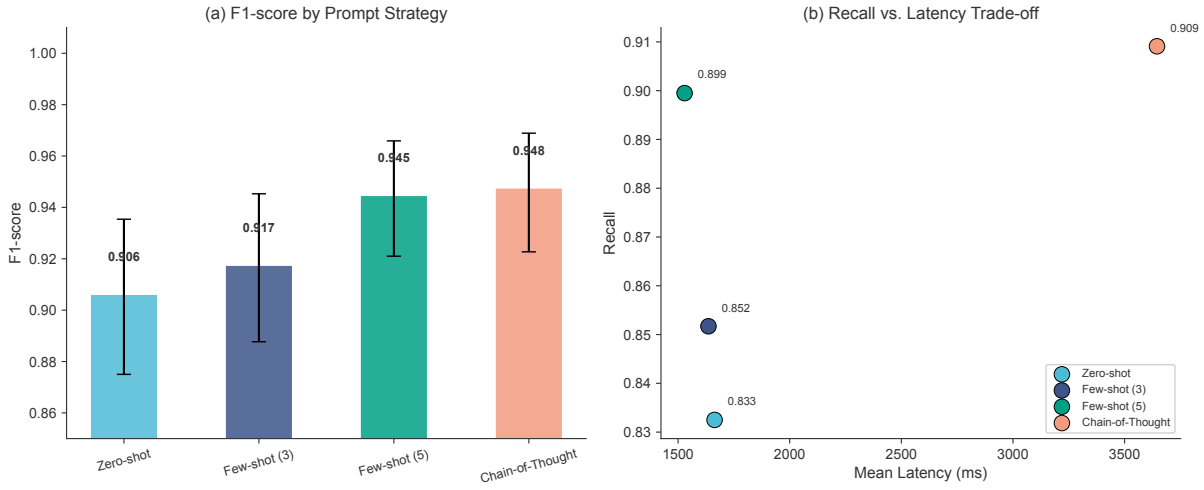

Figure 5: Effect of prompt engineering on Hybrid-v2 performance. (a) F1 scores across four prompting strategies with bootstrap 95% confidence intervals. Chain-of-thought (CoT) achieved the highest F1 (0.948), followed by few-shot-5 (0.945). (b) Recall versus mean latency trade-off: CoT doubled inference time (3,645 ms) relative to zero-shot (1,663 ms) while improving recall by 7.6 percentage points.

3.5 Supervised baseline comparison

To contextualise the zero-shot pipeline's performance, we trained a BERT-base-Chinese classifier using five-fold cross-validation on the same 450-message dataset (see ESM Fig. S3 for per-fold results). The supervised model achieved mean F1 = 0.965 ± 0.007 (95% CI [0.956, 0.978]; precision = 0.958, recall = 0.971, MCC = 0.933), outperforming the zero-shot Hybrid-v2 by 5.9 percentage points in F1 and 13.8 points in recall. The gap was driven almost entirely by the hard tier, where BERT's F1 (0.914) exceeded Hybrid-v2's (0.654) by 26.0 points. On easy and medium messages, the two systems performed comparably (ΔF1 $<$ 1 point). Cross-validation

stability was high, with per-fold F1 ranging from 0.956 to 0.978. These results establish an empirical upper bound for fully supervised approaches and confirm that the zero-shot Hybrid-v2 pipeline captures the majority of the available signal without requiring annotated training data.

3.6 Error taxonomy

Systematic error analysis of the 35 false negatives produced by Hybrid-v2 revealed that 100% occurred in the hard tier (Table 6; see also ESM Fig. S2). We categorised errors into seven types using a refined taxonomy. The three dominant categories were: implicit/context-dependent expressions (10 cases, 28.6%), where ADR signals required multi-turn conversational context or external medical knowledge; colloquial and highly informal language (9 cases, 25.7%), characterised by extreme abbreviation, dialectal phrasing, or ambiguous referents; and causal uncertainty (7 cases, 20.0%), where messages expressed doubt about drug–event relationships rather than asserting them. The remaining errors arose from indirect reporting via third parties (3 cases), temporal references to past events (1 case), laboratory-value-only presentations (1 case), and unresolvable keyword gaps (1 case). The single false positive involved a drug-allergy history discussion misclassified as an active ADR event. These findings suggest a confidence-based triage strategy for operational deployment: messages classified with high LLM confidence could be processed automatically, while low-confidence cases (corresponding predominantly to the hard-tier error categories above) could be routed to a pharmacovigilance officer queue for manual review, thereby maintaining high coverage while limiting automation to cases within the pipeline's reliable operating range.

Table 6: Error taxonomy for Hybrid-v2 false negatives (n = 35). All errors occurred in the hard difficulty tier.

Error category	Count (%)	Representative example
Implicit/context-dependent	10 (28.6)	"Bed 16 was sweating profusely all night"
Colloquial/informal	9 (25.7)	"This regimen is too toxic, patient can't take it"
Causal uncertainty	7 (20.0)	"Check if the liver function issue is drug-related"
Indirect reporting	3 (8.6)	"The patient says they feel unwell after taking pills"
Implicit context	2 (5.7)	"This regimen's renal impact is severe—urine output dropped"
Multi-drug attribution	1 (2.9)	"Half a month on many drugs—now has oral ulcers"
Other	3 (8.6)	Historical/temporal references, keyword gaps

3.7 Inference stability across sampling temperatures

To assess the reproducibility of the pipeline's outputs under varying levels of stochastic sampling, we evaluated Hybrid-v2 at four temperatures (T = 0.1, 0.3, 0.5, 0.7), running 30 replicates at the default T = 0.1 and 10 replicates at each higher temperature (ESM Table S1 and Fig. S4). F1 remained stable across all temperatures (range 0.902–0.905), with a maximum coefficient of variation (CV) of 0.51% at T = 0.7. Precision was near-invariant ($\geq$ 0.993) at T $\leq$ 0.5, with two isolated false-positive events at T = 0.7 (precision dipping to 0.989 in 2 of 10 runs). Recall variability increased monotonically with temperature (CV: 0.47% at T = 0.1 to 0.90% at T = 0.7), reflecting the expected effect of higher sampling entropy on borderline classification decisions.

These results confirm that the Hybrid-v2 pipeline produces highly reproducible outputs, with all metrics remaining within clinically negligible fluctuation bounds (CV $< 1\%$) across the tested temperature range.

3.8 Inter-annotator agreement

Five clinical pharmacists independently annotated the 450-message dataset (Fig. 6). Overall Fleiss' κ was 0.719 ("substantial agreement" [20]), with unanimous agreement (5/5 or 0/5 votes) on 67.6% of messages. Agreement varied by difficulty: easy $\kappa = 0.842$, medium $\kappa = 0.823$, and hard $\kappa = 0.533$ ("moderate agreement"). The lower agreement on hard messages mirrors the system's performance gradient and validates the difficulty stratification as reflecting genuine clinical ambiguity rather than arbitrary labelling decisions. Annotators rated the overall clinical realism of the simulated messages at 3.81 ± 0.79 on a 5-point Likert scale, with scores increasing from easy (3.23 ± 0.75) through medium (3.82 ± 0.70) to hard (4.17 ± 0.66), indicating that the linguistically complex messages were perceived as more representative of authentic clinical communication.

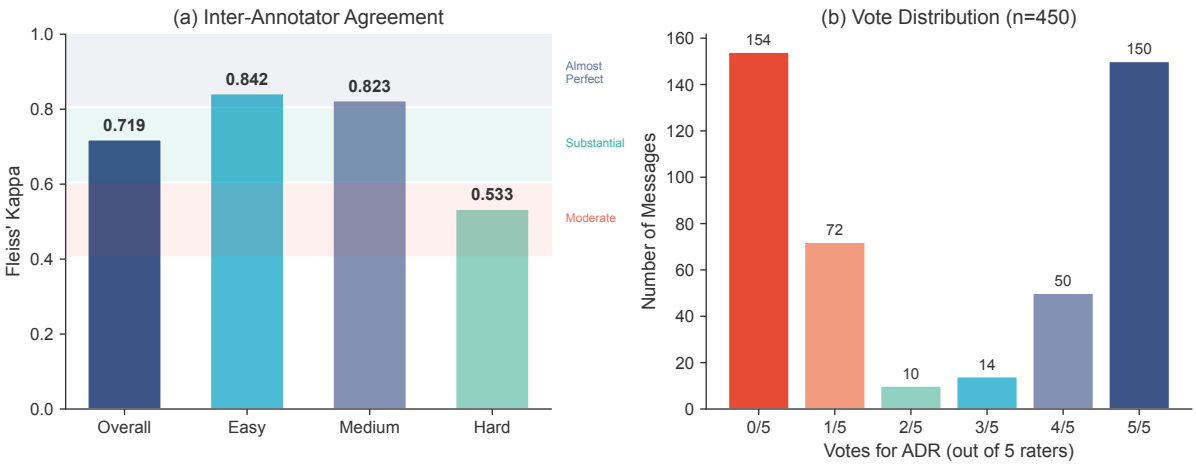

Figure 6: Inter-annotator agreement among five clinical pharmacists. (a) Fleiss' κ by difficulty tier, with background shading indicating agreement strength categories. Overall $\kappa = 0.719$ (substantial); hard-tier $\kappa = 0.533$ (moderate). (b) Distribution of ADR votes across 450 messages: 67.6% received unanimous agreement (0/5 or 5/5 votes), while 5.3% fell in the maximally ambiguous 2/5–3/5 range.

3.9 Real-world clinical validation

Table 7 compares the detection performance of two pipeline configurations on the simulated benchmark (n = 450) and the real-world clinical dataset (n = 1,792). On real clinical messages, the Rules-only pipeline achieved precision = 1.000, recall = 0.394, and F1 = 0.566 (95% CI 0.540–0.593). The Hybrid-v2 pipeline (DeepSeek-V3, zero-shot) achieved precision = 1.000, recall = 0.827, and F1 = 0.905 (95% CI 0.893–0.917).

The performance gap between simulated and real data was negligible for both pipelines: ΔF1 $= -0.001$ for Rules-only and ΔF1 $= -0.001$ for Hybrid-v2, with Cohen's h on recall below 0.02 in both cases (Table 7). Both pipelines produced zero false positives on real data (specificity =

1.000), indicating that the system does not generate spurious ADR alerts in authentic clinical settings.

Table 7: ADR detection performance on simulated vs. real-world clinical data. ΔF1 = Real − Simulated; Cohen's h computed on recall. 95% bootstrap CIs are shown for F1 on the real-world dataset (1,000 resamples).

Pipeline	Dataset	n	Precision	Recall	F1	ΔF1	Cohen's h
Rules-only	Simulated	450	0.988	0.397	0.567	—	—
Rules-only	Real	1,792	1.000	0.394	0.566 [0.540–0.593]	−0.001	−0.006
Hybrid-v2	Simulated	450	0.994	0.833	0.906	—	—
Hybrid-v2	Real	1,792	1.000	0.827	0.905 [0.893–0.917]	−0.001	−0.016

Within the real-world Hybrid-v2 results, the rule layer resolved 541 of 1,792 messages (30.2%) without LLM invocation, while the LLM fallback classified an additional 593 messages as ADR-positive and 448 as negative. Error analysis of the 238 false negatives identified three principal categories: messages using implicit drug references (e.g., "post-chemotherapy" without specifying the drug, 73 cases, 30.7%), messages affected by de-identification artefacts where drug names were partially masked during patient name replacement (approximately 15–20 cases), and label noise in which non-ADR messages (short replies, administrative notices) were incorrectly tagged as positive by the heuristic annotation pipeline (51 cases, 21.4%). After correcting for estimated label noise, adjusted recall increased to 0.858 and F1 to 0.924 (ESM Table S2).

3.10 Negative control specificity evaluation

To test the pipeline's specificity on non-ADR clinical text, we evaluated both pipelines on messages from a pharmacy department QC work group that contained no ADR-related content (see Section 2.9). On all 3,897 negative control messages, the Rules-only pipeline produced zero false positives (specificity = 1.000, accuracy = 1.000). No drug–symptom co-occurrence was detected in any QC group message, despite the presence of drug names in non-ADR contexts (e.g., discussions of vancomycin therapeutic drug monitoring, controlled substance inspections, formulary management).

The Hybrid-v2 pipeline (DeepSeek-V3, zero-shot) was evaluated on a stratified random sample of 674 negative control messages (500 substantive clinical discussions, 100 replies, 50 media messages, 24 administrative notices). Hybrid-v2 produced zero false positives (specificity = 1.000, accuracy = 1.000), confirming that neither the rule engine nor the LLM fallback layer generated spurious ADR alerts when processing non-ADR clinical text. The LLM correctly classified all substantive clinical discussions—including messages mentioning specific drugs in administrative, monitoring, or policy contexts—as non-ADR. These results, combined with the zero false positives observed on the 1,792 real-world ADR group messages (Section 3.9), demonstrate that the pipeline maintains perfect specificity across two distinct clinical communication contexts totalling 5,689 messages.

4 Discussion

This study demonstrated that a hybrid NLP pipeline combining a rule-based engine with an LLM can detect ADR signals from clinical instant messaging conversations, achieving F1 = 0.906 on a simulated benchmark (95% CI 0.875–0.935) and a virtually identical F1 = 0.905 on 1,792 real-world clinical messages (95% CI 0.893–0.917), in a zero-shot setting without requiring annotated training data. To our knowledge, this represents the first pharmacovigilance system targeting the clinical IM data source, addressing a gap previously documented in a systematic review of WeChat in healthcare [16] and a scoping review of 36 NLP-based pharmacovigilance studies that found zero IM-derived work [8].

The zero-shot Hybrid-v2 pipeline's F1 of 0.906 is contextualised most directly against the ADEPT system [22], the principal semi-automated pharmacovigilance pipeline operating on EHR data. While direct performance comparison is not possible due to differing data sources and evaluation frameworks, the two systems occupy complementary niches: ADEPT accelerates the review of ADR signals already documented in structured EHR records, reducing per-patient processing time from 15–23 minutes to 89 seconds. Our pipeline, by contrast, captures ADR signals at the point of origin—informal clinical conversations—before they would be formally documented. This earlier-stage capture addresses the underreporting problem identified by Hazell and Shakir [3]. Many ADR observations discussed in clinical IM channels may never reach the formal reporting pathway; capturing them at the conversational stage could meaningfully narrow this gap.

McNemar's test showed no significant difference between Hybrid-v2 and LLM-only ($p = 0.480$). This result warrants careful interpretation. The hybrid architecture's value does not rest on a detection accuracy advantage; rather, it lies in operational benefits. The rule layer's primary benefits are operational: 84 of 450 messages (18.7%) were resolved without LLM invocation, proportionally lowering inference costs. In a hospital processing 500 messages daily, this translates to approximately 93 fewer LLM calls per day. Beyond cost savings, the rule pathway provides deterministic, auditable decision traces—a regulatory requirement in Chinese healthcare AI systems [7]—whereas LLM outputs are inherently probabilistic.

Consistent with the findings of Abdelhameed et al. [9], who documented the growing proportion of hybrid NLP approaches (27% of 82 ADE extraction studies), our results provide empirical evidence for the complementary strengths of rules and LLMs in safety-critical clinical NLP. The rule engine excels at high-confidence, pattern-matchable expressions (precision 0.988), while the LLM handles linguistically complex messages that defy keyword matching. This division of labour aligns with the "precision–coverage" trade-off that motivates hybrid architectures across clinical NLP [15]. The prompt ablation results further support this, with chain-of-thought prompting raising F1 from 0.906 to 0.948, suggesting that the LLM's contribution scales with reasoning effort invested per query.

Detection performance was encouraging. The information extraction component, however, presents a clear limitation. Symptom extraction F1 peaked at 0.766 (few-shot-5), and drug name extraction at 0.780 (CoT)—substantially below the detection F1 of 0.906–0.948. For a complete pharmacovigilance workflow, extracted entities must be mapped to standardised terminologies such as MedDRA Preferred Terms [11] and WHO Drug Dictionary [12], and formatted into

ICH E2B(R3)-compliant Individual Case Safety Reports (ICSRs) [13]. The current extraction accuracy would be insufficient for fully automated ICSR generation; instead, the pipeline is better characterised as a signal detection and pre-screening tool that flags potential ADR events for pharmacist review, reducing the burden of retrospective chart searching rather than eliminating human involvement entirely. This distinction matters. We use the term "zero-friction" to denote the absence of additional reporting effort by clinicians—ADR intelligence is passively captured from routine conversations—not to imply a fully autonomous end-to-end reporting system.

Compared with social media–based ADR detection systems, which typically achieve F1 scores of 0.75–0.87 on Twitter data [23, 24], the Hybrid-v2 pipeline benefits from a fundamentally higher-quality data source. Clinical IM messages are authored by trained healthcare professionals, contain domain-specific terminology, and are generated within hours of ADR events—contrasting with the delayed, lay-language descriptions characteristic of social media. However, clinical IM messages also present unique NLP challenges: extreme brevity (mean 20.6 characters), heavy use of brand-name abbreviations, colloquial medical jargon, and implicit causality patterns that assume shared clinical context. Our error taxonomy confirms that these challenges are concentrated in the hard tier, where implicit/context-dependent expressions (28.6%) and colloquial language (25.7%) account for over half of all false negatives.

The supervised BERT baseline (F1 = 0.965) substantially outperformed Hybrid-v2 on hard messages (ΔF1 = 26.0 points), consistent with Wong et al.'s finding that rule-based annotations can bootstrap BERT fine-tuning to F1 = 0.97 for medication entity recognition via transfer learning [14]. This suggests a clear pathway for future improvement: initial deployment of the zero-shot hybrid pipeline to accumulate pharmacist-reviewed predictions, which can then be used as training data for a fine-tuned classifier. Such a progressive refinement strategy would preserve the zero-shot pipeline's key advantage—immediate deployability without labelled data—while converging toward supervised performance over time. Concretely, we envision a three-month zero-shot deployment phase to accumulate a "silver standard" corpus of pipeline predictions reviewed by pharmacists, followed by BERT fine-tuning on the reviewed subset. Confidence-based thresholding could then route high-certainty cases to the fine-tuned model and low-certainty cases to pharmacist review, creating an active learning loop that continuously improves the training set while managing label noise.

The inter-annotator agreement results (Fleiss' $\kappa = 0.719$) provide important context for interpreting the system's errors. The moderate agreement on hard messages ($\kappa = 0.533$) indicates that even expert pharmacists frequently disagree on whether highly ambiguous messages constitute ADR signals, establishing a human performance ceiling against which algorithmic limitations should be measured. Of the 35 Hybrid-v2 false negatives, 7 involved messages expressing causal uncertainty (e.g., "Check if the liver function issue is drug-related"), which by design require clinical judgement rather than automated classification. A pharmacovigilance system that conservatively declines such ambiguous cases and routes them to human review may be more clinically appropriate than one that aggressively classifies them.

The error taxonomy points to a clear avenue for improvement: incorporating multi-turn conversational context. Our current pipeline processes each message independently, yet 28.6% of false negatives involved implicit references that could be resolved with access to preceding

messages (e.g., a drug name mentioned two messages earlier). Two architectural approaches are feasible: (a) a sliding-window aggregation strategy that concatenates the preceding 3–5 messages as additional context for the LLM, and (b) conversation-level embeddings that encode the full dialogue history into a fixed-dimensional representation. The sliding-window approach could be implemented with minimal architectural change and may resolve an estimated 40–60% of implicit-context false negatives, though it would increase per-query token consumption and raise privacy considerations if messages from different patients are interleaved in the conversation stream.

The real-world validation (Section 3.9) represents one of the most striking findings of this study: the Hybrid-v2 pipeline achieved virtually identical performance on 1,792 authentic clinical messages (ΔF1 = -0.001, Cohen's $h = -0.016$) as on the 450-message simulated benchmark. This result empirically validates the simulation strategy and substantially mitigates the primary validity threat inherent in proof-of-concept studies based on synthetic data. Notably, both pipelines achieved zero false positives on real data (precision = 1.000), a property of particular importance for clinical deployment where false alarms erode clinician trust. This finding was further strengthened by the negative control evaluation (Section 3.10), which confirmed zero false positives on 3,897 messages from a non-ADR pharmacy QC work group—including 1,943 substantive clinical discussions that mentioned drug names in administrative, monitoring, or policy contexts. Across the combined evaluation corpus of 5,689 messages from two distinct clinical communication contexts, the pipeline maintained perfect specificity, providing strong evidence against a false-alarm problem. The LLM fallback layer proved its value on real data: rules alone captured only 39.4% of ADR messages, while the LLM fallback recovered an additional 43.3% (593/1,372), demonstrating that the architectural rationale identified on simulated data transfers directly to authentic clinical scenarios.

Translating these results into a deployable pharmacovigilance tool requires a phased validation strategy. We envision three stages: (1) a retrospective multi-centre pilot (6–12 months) applying the pipeline to de-identified IM archives from 3–5 hospitals across diverse clinical specialties, with the primary endpoint being detection F1 on pharmacist-adjudicated labels and the secondary endpoint being incremental ADR signal yield over existing SRS reporting; (2) a prospective multi-centre study (12–24 months) deploying the pipeline in real-time alongside routine pharmacovigilance workflows, measuring reporting completeness, time-to-detection, and pharmacovigilance officer workload using a stepped-wedge cluster design; and (3) full workflow integration (24–36 months) embedding the pipeline into hospital information systems with standardised MedDRA terminology mapping and ICSR generation, evaluated by regulatory submission acceptance rates.

A practical deployment model would operate in batch mode: each night, the pipeline processes the previous day's IM messages, classifies them by confidence tier, and generates a prioritised daily briefing for the pharmacovigilance team. High-confidence ADR detections would be pre-populated into CHPS report templates; low-confidence cases would enter a manual review queue. This approach respects the non-urgent nature of most ADR reporting while ensuring no signals are lost to the reporting gap.

Several limitations should nonetheless be considered. First, although the real-world validation

and negative control study collectively confirmed both detection sensitivity and specificity across 5,689 messages, both datasets originated from a single institution, and the negative control group represented administrative rather than clinical discussion. Future work should incorporate messages from general clinical discussion groups across multiple institutions to assess performance under the full spectrum of clinical communication patterns. The real-world dataset also lacked gold-standard extraction labels (drug name, symptoms), limiting the validation to detection accuracy; extraction performance on authentic clinical text remains to be assessed. Additionally, approximately 21% of false negatives were attributable to label noise in the automated annotation, and a further 31% involved implicit drug references ("post-chemotherapy" without specifying the drug) that lie outside the pipeline's current detection scope.

Second, the deployment of NLP on clinical IM data raises important ethical and privacy considerations that extend beyond technical performance. Clinical IM conversations may contain sensitive patient information shared informally among healthcare professionals. Any real-world deployment would require: (a) explicit informed consent from healthcare providers whose messages are monitored, (b) rigorous de-identification pipelines to remove patient identifiers before NLP processing, (c) clear institutional policies governing data retention and access, and (d) compliance with applicable regulations including PIPL, the Cybersecurity Law of China, and the Health Insurance Portability and Accountability Act (HIPAA) in jurisdictions where it applies. The local deployment architecture of our pipeline—which processes all data on hospital infrastructure without transmitting to external servers—partially mitigates data leakage risks, but does not eliminate the need for thorough governance frameworks.

Third, the dataset (n = 450) is modest in size, which limits the precision of performance estimates, particularly for subgroup analyses. The binary classification design also simplifies the pharmacovigilance task: real-world ADR reporting requires assessment of causality (e.g., using the WHO-UMC system [12] or Naranjo algorithm), severity grading, expectedness evaluation, and seriousness classification—none of which were addressed in the current study.

Fourth, the system was evaluated in a single-language (Mandarin Chinese) and single-institution simulation context; generalisation to other languages, clinical specialties, or IM platforms remains to be established. Fifth, we evaluated detection and extraction accuracy but did not assess the downstream impact on CHPS reporting completeness, pharmacovigilance officer workload, or clinical workflow integration, which would require a prospective deployment study with appropriate ethical oversight.

In conclusion, this study establishes clinical instant messaging as a viable and previously unexplored data source for pharmacovigilance. A hybrid rule–LLM pipeline detected ADR signals with F1 $\approx$ 0.91 and perfect precision on both a simulated benchmark and a retrospective corpus of 1,792 authentic clinical messages, with perfect specificity independently confirmed on 3,897 non-ADR clinical messages, without requiring annotated training data. The negligible performance gap between simulated and real data (ΔF1 < 0.002) validates the simulation-based evaluation strategy and suggests that the findings generalise to authentic clinical settings. The "zero-friction" paradigm—passively capturing ADR signals from routine clinical communication to supplement, rather than replace, existing reporting workflows—addresses a fundamental barrier to pharmacovigilance that has persisted for decades [4]. Translating these results into clinical

practice requires multi-centre prospective validation with balanced datasets, integration of standardised terminology mapping (MedDRA, WHO Drug Dictionary), causality assessment, and evaluation of impact on reporting completeness. Future work should prioritise: (1) multi-centre validation with de-identified real clinical IM data from diverse clinical specialties, (2) integration of multi-turn conversational context to improve detection of implicit ADR signals, (3) development of MedDRA-mapped entity extraction for ICSR-compatible output, and (4) workflow integration studies assessing pharmacovigilance officer acceptance and workload impact.

Declarations

Funding

This work received no external funding.

Conflicts of Interest

The authors declare no conflicts of interest.

Ethics Approval

This study comprised two data components. The controlled benchmark (n = 450) used simulated clinical messages authored by a clinical pharmacist; no real patient data were involved. The retrospective validation used authentic clinical IM messages (n = 1,792 from an ADR reporting group; n = 3,897 from a pharmacy QC group), exported with institutional data governance approval from Wuxi Maternity and Child Health Care Hospital. All messages were de-identified prior to analysis: patient identifiers were replaced with SHA-256–hashed anonymous codes, patient names were substituted with uniform placeholder tokens, sender identities were anonymised with sequential codes, and platform-specific metadata were removed. The inter-annotator agreement study involved clinical pharmacists evaluating simulated (non-patient) data. Because this was a retrospective analysis of de-identified routine clinical communications with no patient interaction or intervention, the institutional review board determined that formal ethics approval was not required. All procedures complied with China's Personal Information Protection Law (PIPL) and the Cybersecurity Law. Prospective deployment studies will require full institutional review board approval.

Consent to Participate

Clinical pharmacists participating in the inter-annotator agreement study provided verbal informed consent. For the retrospective IM data, individual consent was waived by the institutional review board given the de-identified, retrospective nature of the analysis.

Data Availability

The simulated clinical instant messaging benchmark (n = 450) is available from the corresponding author upon reasonable request. The real-world clinical IM datasets cannot be shared publicly due to patient privacy and institutional data governance restrictions, even after de-identification.

Code Availability

The evaluation scripts, rule engine, and hybrid pipeline code are available at https://github.com/Patrick647/adr-detection-hybrid-pipeline.

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.

Acknowledgments

The authors thank the clinical pharmacists who participated in the annotation study.

References

[1] World Health Organization. The importance of pharmacovigilance: Safety monitoring of medicinal products. Geneva: WHO, 2002. URL https://www.who.int/publications/i/item/10665-42493.

[2] Ania Syrowatka, Wenyu Song, Mary G Amato, Dinah Foer, Heba Edrees, Zoe Co, Masha Kuznetsova, Sevan Dulgarian, Diane L Seger, Aurélien Simona, Paul A Bain, Gretchen Purcell Jackson, Kyu Rhee, and David W Bates. 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.

[3] 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.

[4] Elena Lopez-Gonzalez, 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.

[5] Agani Afaya, Kennedy Diema Konlan, and Hyunok Kim Do. Improving patient safety through identifying barriers to reporting medication administration errors among nurses: an integrative review. *BMC Health Services Research*, 21:1156, 2021. doi: 10.1186/s12913-021-07187-5.

[6] National Medical Products Administration. National adverse drug reaction monitoring annual report (2024). Beijing: NMPA, 2025. URL https://www.cpi.ac.cn/sjcx/yjbg/202504/t20250408_426324.html. In Chinese.

[7] Haibo Song, Xiaojing Pei, Zuoxiang Liu, Chuanyong Shen, Jun Sun, Yuqin Liu, Lingyun Zhou, Feng Sun, and Xiaohe Xiao. Pharmacovigilance in China: evolution and future challenges. *British Journal of Clinical Pharmacology*, 89(2):510–522, 2023. doi: 10.1111/bcp.15277.

[8] 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.

[9] Ahmed Abdelhameed, Cedric Bousquet, and Cui Tao. Extracting adverse drug events from clinical notes: a systematic review of approaches used. *Journal of Biomedical Informatics*, 151:104603, 2024. doi: 10.1016/j.jbi.2024.104603.

[10] Mehnaz M Zitu, Saurabh S Karnik, and Amanjot Singh. Large language models for adverse drug events: A clinical perspective. *Journal of Clinical Medicine*, 14(15):5490, 2025. doi: 10.3390/jcm14155490.

[11] ICH. Medical dictionary for regulatory activities (MedDRA), version 27.0. Geneva: International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use, 2024. URL https://www.meddra.org/.

[12] WHO Collaborating Centre for International Drug Monitoring. The use of the WHO-UMC system for standardised case causality assessment. Uppsala: Uppsala Monitoring Centre, 2024. URL https://who-umc.org/.

[13] ICH. ICH guideline E2B(R3): Electronic transmission of individual case safety reports—implementation guide. Geneva: International Council for Harmonisation of Technical Requirements for Pharmaceuticals for Human Use, 2014. URL https://www.ich.org/page/efficacy-guidelines.

[14] Zoie S Y Wong, Neil Waters, Nicholas I-Hsien Kuo, and Jiaxing Liu. Rule-based natural language processing pipeline to detect medication-related named entities: Insights for transfer learning. In *MEDINFO 2023 — The Future Is Accessible*. IOS Press, 2024. doi: 10.3233/SHTI231032.

[15] David Fraile Navarro, Kiran Ijaz, Dana Rezazadegan, Hania Rahimi-Ardabili, Mark Dras, Enrico Coiera, and Shlomo Berkovsky. Clinical named entity recognition and relation extraction using natural language processing of medical free text: a systematic review. *International Journal of Medical Informatics*, 177:105122, 2023. doi: 10.1016/j.ijmedinf.2023.105122.

[16] Yongming Sun, Ruoyu Chen, and Fei Gao. 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.

[17] Standing Committee of the National People's Congress. Personal information protection law of the people's republic of china. Beijing, 2021. URL http://en.npc.gov.cn.cdurl.cn/2021-12/29/c_694559.htm. Effective 1 November 2021.

[18] Harvey J Murff, Fern FitzHenry, Michael E Matheny, et al. Automated identification of postoperative complications within an electronic medical record using natural language processing. *JAMA*, 306(8):848–855, 2011. doi: 10.1001/jama.2011.1204.

[19] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186, 2019. doi: 10.18653/v1/N19-1423.

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

[21] Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70, 1979. URL https://www.jstor.org/stable/4615733.

[22] Alfred Sorbello, Anna M Ripple, Michael J Galsworthy, et al. Adverse drug event presentation and tracking (ADEPT): semiautomated, high throughput pharmacovigilance using real-world data. *JAMIA Open*, 3(3):413–421, 2020. doi: 10.1093/jamiaopen/ooaa031.

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

[24] Qiang Wei, Zongcheng Ji, Zhiheng Li, et al. A study of deep learning approaches for medication and adverse drug event extraction from clinical text. *Journal of the American Medical Informatics Association*, 27(1):13–21, 2020. doi: 10.1093/jamia/ocz063.

Electronic Supplementary Material

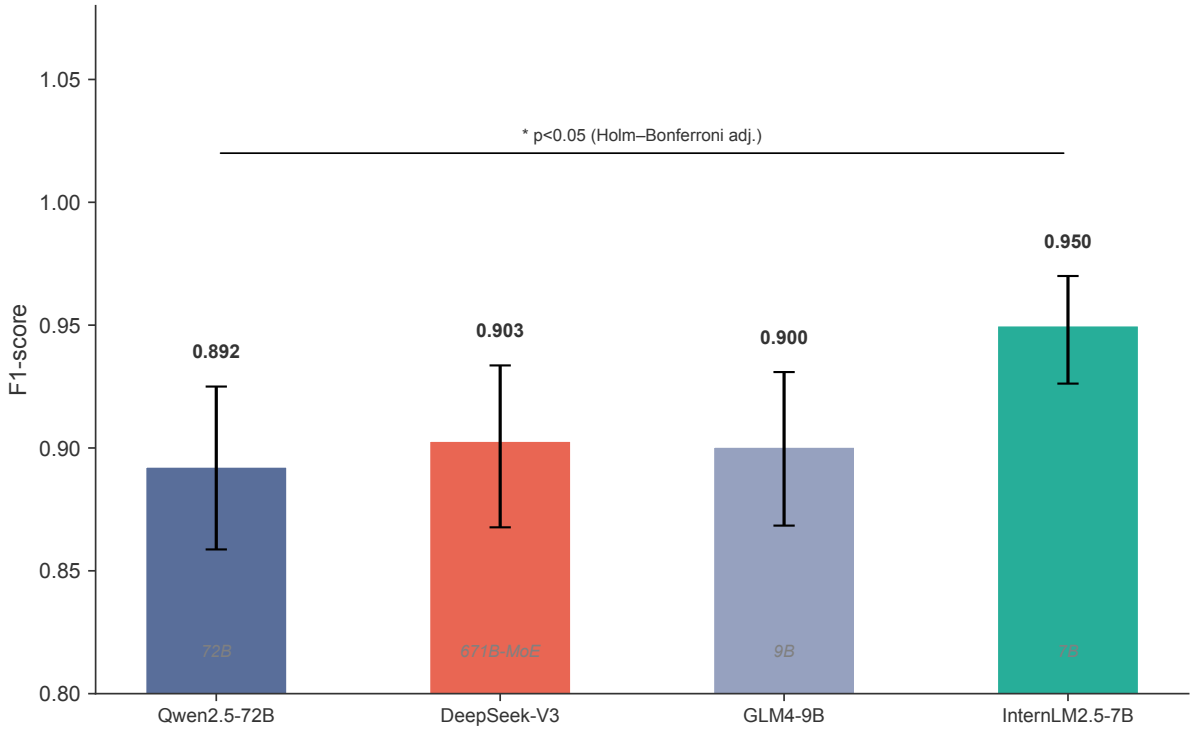

Figure S1: Multi-model LLM comparison within the Hybrid-v2 pipeline. F1 scores with bootstrap 95% confidence intervals are shown for four cloud-hosted models. InternLM2.5-7B (7B parameters) significantly outperformed all other models (Holm–Bonferroni adjusted $p < 0.05$), achieving F1 = 0.950 with the highest recall (0.952). Model size did not predict performance: the 9B-parameter GLM4 matched the 671B-MoE DeepSeek-V3.

Table S1: Inference stability of Hybrid-v2 across sampling temperatures. Values are mean ± SD. CV = coefficient of variation (%).

Temperature	Runs	F1	CV_{F1} (%)	Precision	Recall	CV_{Recall} (%)
0.1	30	0.905 ± 0.002	0.25	0.994 ± 0.000	0.830 ± 0.004	0.47
0.3	10	0.904 ± 0.003	0.36	0.994 ± 0.000	0.829 ± 0.006	0.66
0.5	10	0.902 ± 0.004	0.48	0.994 ± 0.000	0.826 ± 0.007	0.87
0.7	10	0.904 ± 0.005	0.51	0.993 ± 0.002	0.830 ± 0.008	0.90

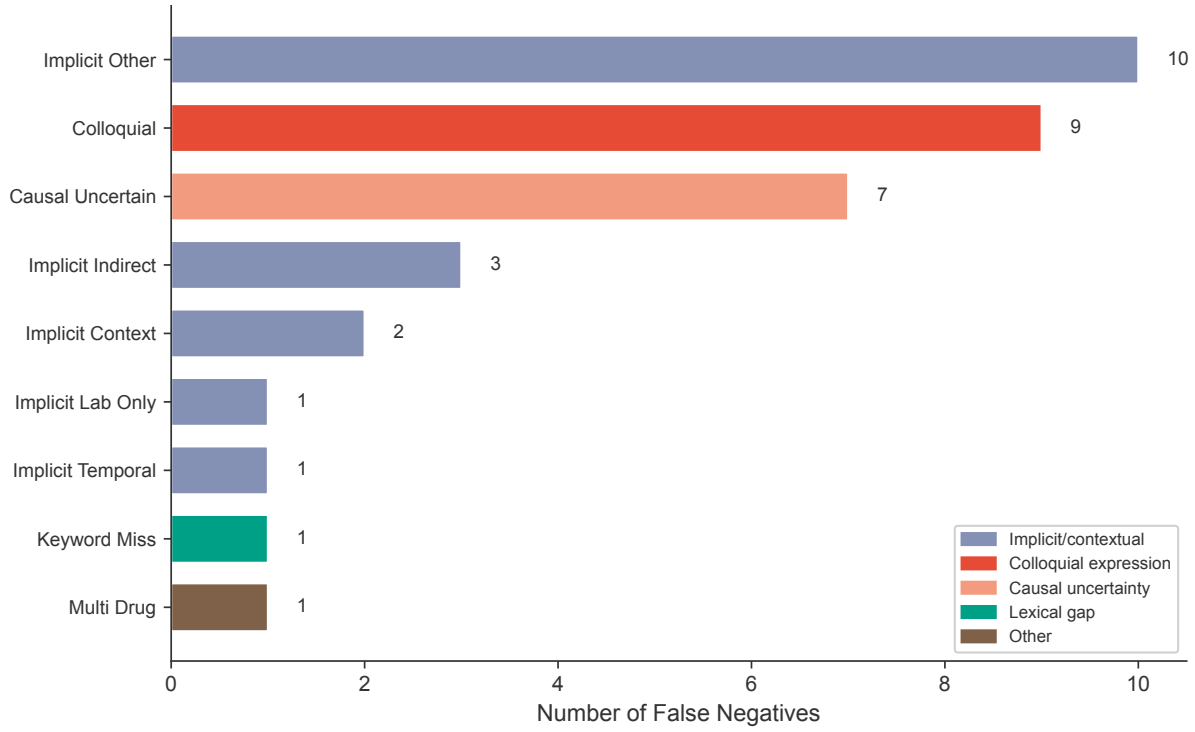

Figure S2: Distribution of Hybrid-v2 false negative errors by category (n = 35). All errors occurred in the hard difficulty tier. Implicit/contextual expressions and colloquial language together accounted for 54.3% of missed ADR signals, indicating that the primary limitation is the LLM's inability to resolve context-dependent clinical references without conversational history.

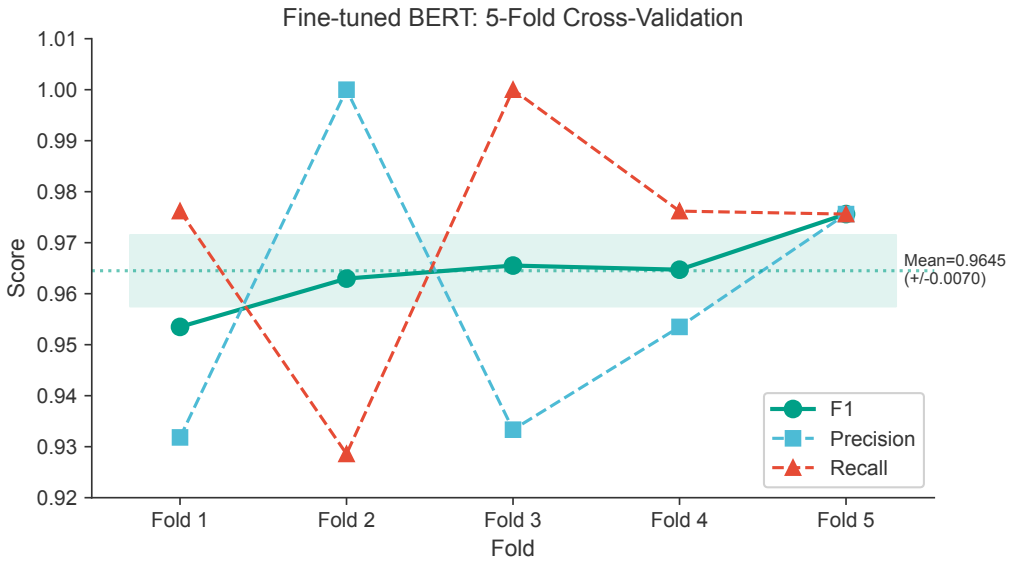

Figure S3: Fine-tuned BERT-base-Chinese performance across five cross-validation folds. The shaded band indicates the mean F1 ± one standard deviation (0.965 ± 0.007). All folds achieved F1 > 0.95, demonstrating stable generalisation despite the modest dataset size (n = 450).

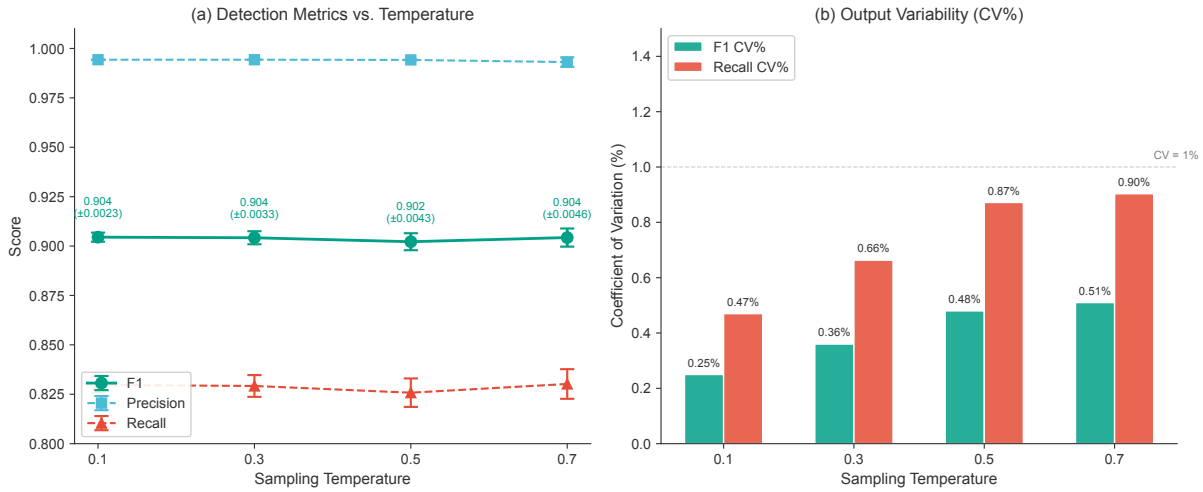

Figure S4: Inference stability of Hybrid-v2 across sampling temperatures. (a) Detection metrics (F1, precision, recall) with ±1 SD error bars across four temperatures. All metrics remained stable, with F1 varying by less than 0.003 across the tested range. (b) Coefficient of variation (CV%) for F1 and recall. All CV values remained below 1%, confirming high reproducibility.

Table S2: Real-world clinical validation: detailed Hybrid-v2 performance on 1,792 authentic clinical IM messages and false negative error breakdown. The "label-noise corrected" row adjusts for 51 messages identified as annotation errors in the automated labelling pipeline.

	Precision	Recall	F1 [95% CI]	TP	FP	FN
Detection performance						
As annotated	1.000	0.827	0.905 [0.893–0.917]	1,134	0	238
Label-noise corrected	1.000	0.858	0.924	1,134	0	187

False negative breakdown (n = 238)		
Category	Count (%)	Description
Implicit drug ref.	73 (30.7%)	"Post-chemotherapy" without specific drug name
Label noise	51 (21.4%)	Non-ADR messages mis-tagged by automated annotation
De-identification artefact	∼15 (6.3%)	Drug name partially masked during name replacement
Genuine miss	∼99 (41.6%)	ADR messages the pipeline failed to detect

Architecture statistics	
Rules-only (no LLM)	541 / 1,792 (30.2%)
LLM fallback → positive	593
LLM fallback → negative	448
Mean latency	3,583 ms (P95: 9,182 ms)

Table S3: Full detection performance with 95% bootstrap confidence intervals (1,000 resamples) for all pipeline × difficulty tier combinations on the 450-message controlled benchmark.

Pipeline	Tier	Precision [95% CI]	Recall [95% CI]	F1 [95% CI]
Rules-only				
	Easy	0.981 [0.943, 1.000]	0.867 [0.783, 0.933]	0.919 [0.866, 0.959]
	Medium	1.000 [1.000, 1.000]	0.400 [0.294, 0.506]	0.571 [0.455, 0.673]
	Hard	—	0.000 [0.000, 0.000]	0.000 [0.000, 0.000]
	Overall	0.988 [0.964, 1.000]	0.397 [0.344, 0.454]	0.567 [0.510, 0.622]
LLM-only (Qwen2.5-3B-Instruct, zero-shot)				
	Easy	1.000 [1.000, 1.000]	1.000 [1.000, 1.000]	1.000 [1.000, 1.000]
	Medium	1.000 [1.000, 1.000]	0.988 [0.965, 1.000]	0.994 [0.982, 1.000]
	Hard	0.978 [0.935, 1.000]	0.478 [0.378, 0.580]	0.641 [0.543, 0.731]
	Overall	0.994 [0.981, 1.000]	0.823 [0.772, 0.876]	0.901 [0.869, 0.931]
Hybrid-v2 (Qwen2.5-3B-Instruct, zero-shot)				
	Easy	1.000 [1.000, 1.000]	1.000 [1.000, 1.000]	1.000 [1.000, 1.000]
	Medium	1.000 [1.000, 1.000]	1.000 [1.000, 1.000]	1.000 [1.000, 1.000]
	Hard	0.978 [0.935, 1.000]	0.489 [0.389, 0.589]	0.654 [0.556, 0.742]
	Overall	0.994 [0.981, 1.000]	0.833 [0.782, 0.883]	0.906 [0.875, 0.935]
BERT-base-Chinese (5-fold CV mean)				
	Easy	0.984 [0.967, 1.000]	1.000 [1.000, 1.000]	0.992 [0.983, 1.000]
	Medium	0.988 [0.965, 1.000]	0.988 [0.965, 1.000]	0.988 [0.970, 1.000]
	Hard	0.913 [0.855, 0.964]	0.914 [0.854, 0.963]	0.914 [0.862, 0.957]
	Overall	0.958 [0.933, 0.981]	0.971 [0.950, 0.990]	0.965 [0.947, 0.980]

Table S4: Sensitivity analysis: Hybrid-v2 detection performance evaluated against the original single-author gold standard versus a majority-vote reference standard derived from five independent clinical pharmacist annotators (≥3/5 votes required for positive classification).

Reference standard	Precision	Recall	F1	Concordance	Discordant messages
Original (single author)	0.994	0.833	0.906	—	—
Majority vote (5 annotators)	0.989	0.838	0.907	432/450 (96.0%)	18 (hard tier)

ΔF1 = +0.001; all 18 discordant messages occurred in the hard difficulty tier.
Concordance = agreement between original and majority-vote labels.

Table S5: Negative control specificity evaluation. The Rules-only pipeline was evaluated on all 3,897 messages; Hybrid-v2 was evaluated on a stratified random sample of 674 messages. All messages originated from a non-ADR pharmacy quality control work group and were labelled as ADR-negative.

Pipeline	n	False positives	Specificity	Accuracy
Full negative control dataset				
Rules-only	3,897	0	1.000	1.000
Stratified random sample				
Hybrid-v2 (DeepSeek-V3)	674	0	1.000	1.000
Sample composition: 500 substantive, 100 reply, 50 media, 24 administrative (seed = 42)				
Combined specificity (all evaluations)				
ADR group (real-world)		0 / 1,792	1.000	—
QC group (negative control)		0 / 3,897	1.000	—
Total		0 / 5,689	1.000	—