CROSSDDI: Cross-Source Evidence-Grounded Drug–Drug Interaction Verification

Hui Wang **Bohao Chu** **Norbert Fuhr**
University Duisburg-Essen
{hui.wang.0111@stud, bohao.chu, norbert.fuhr}@uni-due.de

Abstract

LLM-based drug–drug interaction (DDI) assessment remains difficult to audit when predictions are not explicitly tied to evidence. While retrieval-augmented generation (RAG) improves grounding, predictions are not guaranteed to be entailed by retrieved items. We present CROSSDDI, a verification-first framework that separates LLM-based evidence extraction from deterministic arbitration over DrugBank and PubMed, requiring positive predictions to be linked to explicit supporting evidence. Evaluated on 1,000 DDInter 2.0 pairs under a positive–unlabeled setting, CROSSDDI achieves recall of 0.576–0.593 over confirmed positives with interaction prediction rates comparable to RAG, while reducing cross-backbone variation (0.018 vs. 0.066). Analysis identifies literature evidence acquisition and attribution as the primary bottleneck: PubMed retrieval covers only 40.5% of confirmed positives, and Path B-only evidence is substantially less reliable than structured evidence. These results suggest that verification-first architectures can improve traceability and backbone consistency, while broader and more reliable literature evidence is needed to extend coverage beyond structured sources.

1 Introduction

Users increasingly turn to conversational AI for medication-related questions, including safety-critical queries such as "Can I take warfarin with aspirin?" (White et al., 2025). Yet large language models (LLMs) can produce hallucinated medical claims (Singhal et al., 2023), and in drug–drug interaction (DDI) assessment, missed or incorrect identification of interactions may lead to adverse events (Huang et al., 2025). Reliable DDI assessment therefore requires every prediction to be traceable to explicit evidence.

Retrieval-augmented generation (RAG) augments language models with retrieved text passages from external sources (Asai et al., 2024) but does not guarantee that generated predictions are faithfully grounded in the retrieved evidence (Wu et al., 2024). Neural DDI models (He et al., 2022; Wang et al., 2025a) predict interactions from molecular features, but operate on structured inputs and do not directly support verification of free-form claims against source-level evidence.

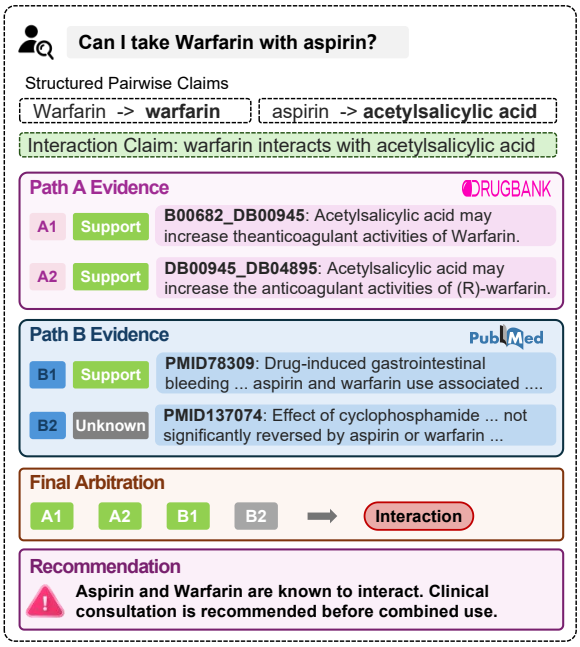

Figure 1: End-to-end example of CROSSDDI for the query "Can I take Warfarin with aspirin?" DrugBank provides structured evidence and PubMed complementary evidence; deterministic arbitration yields an *Interaction* verdict linked to explicit evidence.

We introduce DDI assessment as **evidence-grounded claim verification** and propose CROSSDDI, which separates LLM-based evidence extraction from deterministic decision-making over two complementary sources: DrugBank (Knox et al., 2024) for broad structured coverage and PubMed for long-tail or recent cases. LLMs assign a stance from each source; a rule-based arbitration layer combines these stances to produce a final ver-

Proceedings of the 25th Workshop on Biomedical Language Processing (BioNLP 2026), pages 911–919

dict, *guaranteeing* that every positive prediction is backed by explicit evidence (Figure 1).

Curated DDI resources such as DDInter (Tian et al., 2025) primarily provide annotated interaction records, while non-interactions are typically unlabeled. In such settings, drug–drug interaction modeling is commonly formulated as a positive–unlabeled (PU) learning problem (Bi et al., 2018). Across three LLM backbones (Qwen2.5-7B, GPT-4o, and Qwen3-235B-A22B (Qwen et al., 2025; Hurst et al., 2024; Yang et al., 2025)), CROSSDDI achieves recall (0.576–0.593) comparable to generation-based RAG baselines while ensuring full evidence traceability, and reduces cross-backbone variation (recall spread 0.018 vs. 0.066). Our analysis traces the remaining recall gap to sparse literature-side evidence acquisition.

Contributions. (i) We formulate DDI assessment as evidence-grounded claim verification over drug pairs under the positive–unlabeled (PU) setting, requiring explicit evidence support. (ii) We introduce CROSSDDI, a verification-first framework that separates LLM-based evidence extraction from deterministic decision-making, enabling explicit grounding, traceability, and consistent predictions across LLM backbones. (iii) Through staged error analysis, we identify literature-side evidence acquisition as a primary bottleneck for recall.

2 Related Work

DDI prediction and extraction. Prior DDI work includes deep learning methods (e.g., SSF-DDI (Zhu et al., 2024), CNN-DDI (Zhang et al., 2022)) and knowledge-graph-based approaches (e.g., LLM-DDI (Li et al., 2026), KGDB-DDI (Zhao et al., 2025)), which learn drug representations from molecular, structural, or relational data for interaction prediction. These methods focus on prediction rather than verification and do not provide evidence-grounded justification for individual claims. Recent LLM-based approaches have been applied to medical QA settings (Wang et al., 2025b), but remain prone to hallucinated or unsupported claims (Wang et al., 2025b; Sicard et al., 2025).

Biomedical verification. Prior work on scientific and medical fact verification, including SciFact, MultiVerS, HealthFC, and BioFactCheck (Wadden et al., 2022a,b; Vladika et al., 2024; Lamichhane et al., 2023), assesses whether claims are supported or refuted by retrieved textual evidence. These approaches typically operate over abstract- or document-level evidence and treat claims independently, without modeling structured pairwise interactions or integrating evidence across curated knowledge and clinical literature. CROSSDDI instead targets structured pairwise drug interaction claims and combines evidence from curated databases and clinical literature through verification with deterministic arbitration.

Biomedical RAG. Biomedical RAG systems (e.g., Almanac (Zakka et al., 2024)) improve grounding by incorporating retrieved evidence, but typically rely on the generator to reconcile evidence during generation. As a result, predictions are not guaranteed to be entailed by specific retrieved items (Wu et al., 2024). In contrast, CROSSDDI separates evidence extraction from deterministic arbitration, enabling traceable decisions under heterogeneous and incomplete evidence.

3 CrossDDIPipeline

CROSSDDI proceeds in four stages: (1) entity resolution and claim construction, (2) DrugBank verification (Path A), (3) PubMed verification (Path B), and (4) deterministic arbitration (see Appendix A). LLMs handle only evidence extraction and attribution; final verdicts follow a fixed rule.

3.1 Entity Resolution and Claim Construction

Given a query q, drug mentions are normalized via case folding, suffix stripping, and alias matching against DrugBank synonyms, yielding a canonical pair $\phi(q) = (d_1, d_2),\ d_1 \leq_{\text{lex}} d_2$. We then construct the binary claim $c = \text{Interacts}(d_1, d_2)$.

This normalization reduces surface variation but may fail for complex multi-token drug names; such cases are treated as part of the overall pipeline error profile analyzed later.

3.2 Path A: Structured DrugBank Verification

Path A queries a local DrugBank (Knox et al., 2024) index with a ranked retrieval procedure. The primary channel is an exact symmetric pair-key match over normalized DrugBank identifiers. Auxiliary lexical channels retrieve lower-specificity records whose descriptions mention both drug names or whose counterpart field matches one queried drug. Records are ranked by specificity, and the top K ($K = 3$) are retained.

Let $R_A^{(K)}$ denote the ranked top-K DrugBank output, and let $\text{Exact}(r, d_1, d_2)$ indicate that record

r is an exact pair-key match for the normalized pair (d_1, d_2). The path-level stance is defined as

$$y_A = \begin{cases} \textit{Support} & \exists r \in R_A^{(K)} : \mathrm{Exact}(r, d_1, d_2), \\ \textit{Unknown} & \text{otherwise.} \end{cases}$$

Exact pair matches are treated as structured support evidence, whereas lexical matches provide auxiliary context only. The absence of an exact match is interpreted as unresolved rather than negative evidence. Path A does not emit *Contradict*, consistent with the PU setting.

3.3 Path B: Literature-grounded Verification

Path B performs sentence-level verification over PubMed. Synonym-expanded queries retrieve up to three articles, from which an LLM extracts at most six claim-relevant sentences and assigns stance labels in $\{\textit{Support}, \textit{Contradict}, \textit{Unknown}\}$. Post-extraction filtering reduces spurious signals: sentences that merely co-mention the drugs are downgraded to *Unknown*, and *Contradict* is kept only when both explicit negation and interaction cues are present. These heuristics prioritize precision over recall. Let E_B^+ and E_B^- denote the filtered *Support* and *Contradict* sets. The path-level stance is defined as

$$y_B = \begin{cases} \textit{Support} & E_B^+ \neq \emptyset \wedge E_B^- = \emptyset, \\ \textit{Contradict} & E_B^- \neq \emptyset \wedge E_B^+ = \emptyset, \\ \textit{Unknown} & \text{otherwise.} \end{cases}$$

In practice, retrieval coverage is limited and many queries default to *Unknown*, making Path B a sparse but complementary signal. Despite this, it provides a literature channel that links decisions to explicit textual statements, supporting interpretability and enabling inspection of model outputs. Full prompt details are provided in Appendix F.

3.4 Deterministic Arbitration

Path-level outputs (y_A, y_B) are combined via a fixed rule:

$$y = \begin{cases} \text{Interaction} & y_A = \mathrm{Sup} \vee y_B = \mathrm{Sup}, \\ \text{Non-Interaction} & y_A = \mathrm{Unk} \wedge y_B = \mathrm{Con}, \\ \text{Unknown} & \text{otherwise.} \end{cases}$$

The rule is asymmetric: any supporting signal yields *Interaction*, reflecting the higher cost of missing harmful interactions, while *Non-Interaction* requires explicit contradiction without support.

A fixed rule is preferred over learned fusion given the absence of negative supervision in the PU setting. Positive predictions are linked by design to retrieved supporting evidence, though their correctness depends on extraction quality.

4 Experiments

4.1 Setup

Dataset and PU setting. We evaluate on 1,000 DDInter 2.0 pairs (Tian et al., 2025) (seed 42): 797 confirmed interactions and 203 unlabeled pairs. Because DDInter provides no verified *Non-Interaction* labels, evaluation is positive–unlabeled (PU) (Bi et al., 2018); precision and false-positive rate are not identifiable.

Metrics. We use **recall over the 797 confirmed positives** as the primary metric and report interaction prediction rate (P.Int) over all pairs, with P.Unk and P.NI describing abstention behavior. Backbone variation is summarized by recall spread. Bootstrap 95% confidence intervals (2,000 resamples) are reported for P.Int, not recall. Accuracy is reported in Appendix D only for completeness.

Pseudo-negative analysis. We form pseudo-negative (PN) subsets from unlabeled pairs with both drugs resolved and no Path A DrugBank match ($n = 91$–93 per suite). Under a closed-world assumption, *Interaction* predictions on these pairs approximate false positives, so PN-FPR is used only for within-suite comparison. Wilson 95% CIs reflect uncertainty within the PN subset, not clinical false-positive uncertainty.

Baselines. We compare CROSSDDI with *LLM-only*, *RAG*, *Path A*, and *Path B* across Qwen2.5-7B, GPT-4o, and Qwen3-235B. Main comparisons use normalized pair inputs after claim construction, so all retrieval-based conditions use the same DrugBank and PubMed evidence retrieved for CROSSDDI. RAG receives this evidence and generates a final verdict using a fixed prompt (Appendix G). These comparisons are matched at the evidence level, isolating differences in reasoning and arbitration; full parsing and claim-construction failures are analyzed separately. Because Path A is deterministic, its outputs are identical across backbones.

4.2 Results

Overall performance. CROSSDDI achieves the highest recall on two of three backbones (0.580, 0.593) and is comparable on the third (0.576 vs.

Suite	PN n	LLM	RAG	Path A	Path B	CrossDDI
Qwen2.5	91	27.5 [19.4, 37.4]	3.3 [1.1, 9.2]	0.0 [0.0, 4.1]	11.0 [6.1, 19.1]	12.1 [6.9, 20.4]
GPT-4o	93	6.5 [3.0, 13.4]	6.5 [3.0, 13.4]	0.0 [0.0, 4.0]	14.0 [8.4, 22.5]	19.4 [12.6, 28.5]
Qwen3	92	5.4 [2.3, 12.1]	13.0 [7.6, 21.4]	0.0 [0.0, 4.0]	14.1 [8.4, 22.7]	17.4 [11.0, 26.4]

Table 1: Pseudo-negative FPR (heuristic). Percentages with Wilson 95% CIs in brackets.

Model	Setting	P.Int	P.Int 95% CI	Recall	P.Unk	P.NI
Qwen2.5-7B	LLM-only	0.385	[.355,.413]	0.386	0.510	0.104
	RAG	0.546	[.513,.578]	0.566	0.434	0.019
	Path A	0.541	[.508,.573]	0.565	0.459	0.000
	Path B	0.051	[.037,.067]	0.041	0.937	0.004
	CROSSDDI	**0.556**	[.524,.587]	**0.580**	0.435	0.000
GPT-4o	LLM-only	0.351	[.320,.381]	0.407	0.311	0.338
	RAG	0.478	[.445,.511]	0.512	0.517	0.003
	Path A	0.541	[.508,.573]	0.565	0.459	0.000
	Path B	0.089	[.070,.110]	0.077	0.868	0.001
	CROSSDDI	**0.575**	[.543,.606]	**0.593**	0.424	0.000
Qwen3-235B	LLM-only	0.320	[.291,.350]	0.384	0.580	0.100
	RAG	0.564	[.531,.596]	**0.578**	0.432	0.004
	Path A	0.541	[.508,.573]	0.565	0.459	0.000
	Path B	0.083	[.064,.104]	0.074	0.908	0.001
	CROSSDDI	0.557	[.525,.588]	0.576	0.427	0.001

Table 2: Interaction prediction rate (P.Int) over all pairs, bootstrap confidence intervals for P.Int, and recall over confirmed positives. P.Unk and P.NI denote prediction rates for *Unknown* and *Non-Interaction*.

0.578 for RAG) under matched retrieval inputs. Interaction prediction rates are similar (0.556–0.575 vs. 0.478–0.564), indicating that gains are not driven by substantially more frequent positive predictions. Path A alone already provides strong coverage (recall = 0.565, Table 2). Among the 1,000 evaluation pairs, 558 (55.80%) have direct DrugBank pair coverage under the Path A resolver. Recall on confirmed positive pairs with direct DrugBank pair coverage reaches 85.81%, compared with 18.98% on confirmed positive pairs without such coverage, indicating that structured database coverage strongly influences end-to-end performance (Appendix E).

Backbone consistency. CROSSDDI exhibits lower cross-backbone variation (recall spread 0.018) than RAG (0.066), indicating more consistent behavior across models. This difference is partly attributable to the dominance of structured evidence in Path A, which reduces the influence of LLM-dependent variation from Path B during decision making.

Literature evidence. Path B alone yields low recall (0.041–0.077) and predicts *Unknown* for most pairs. PubMed retrieval applies to only 40.5% of confirmed positives, making literature-side evidence acquisition the primary bottleneck. Additional loss arises from entity resolution, query construction, and extraction reliability. Despite these limitations, Path B provides complementary evidence in cases where structured lookup abstains, linking decisions to explicit textual statements beyond curated sources. At the system level, the support-favoring arbitration rule leads to higher PN-FPR than RAG across all pseudo-negative suites (Table 1), reflecting an inherent recall–FPR tradeoff in the current design.

Evidence quality. Manual evaluation of 100 stratified *Interaction* predictions shows overall stance correctness of 84.0%, with 85.0% strict evidence relevance and 95.0% lenient evidence relevance. Evidence quality varies substantially by trigger source: Path A-only and both-path cases show high stance correctness (96.0% and 100.0%, respectively), while Path B-only cases are weaker (44.0% correctness). The main failure mode for Path B-only cases is selection of topically related but non-interaction statements during retrieval and extraction. This evaluation uses a single annotator and should be treated as preliminary; future work should include multi-annotator evaluation and agreement analysis to assess annotation reliability.

5 Conclusion

CROSSDDI shows that separating evidence extraction from deterministic decision-making yields more consistent behavior across LLM backbones (recall spread 0.018 vs. 0.066 for RAG). Under matched retrieval, CROSSDDI achieves higher recall on two of three backbones and comparable recall on the third, with interaction prediction rates similar to RAG. Coverage analysis shows that end-to-end recall is strongly influenced by direct DrugBank coverage: recall over confirmed positives reaches 85.81% on covered pairs but drops to 18.98% on uncovered pairs. The literature path remains limited by retrieval coverage (40.5% of confirmed positives) and extraction reliability. These results suggest that verification-first architectures can improve traceability and backbone consistency, while better literature retrieval and evidence quality are needed for coverage beyond structured sources.

Limitations

Positive–unlabeled evaluation. DDInter provides confirmed positive pairs but no verified *Non-Interaction* labels, resulting in a PU setting in which precision and false-positive rate are not identifiable. Our pseudo-negative FPR is therefore heuristic and intended only for relative comparison within each backbone suite.

Curated-resource overlap. Because DDInter and DrugBank are both curated DDI resources, their interaction coverage may partially overlap. Quantitative analysis suggests a strong dependence on direct DrugBank coverage: recall drops from 85.81% to 18.98% when moving from pairs with direct DrugBank coverage to those without, on the same 1,000-pair evaluation set used in the reported experiments (Appendix E, Table 4). The strong Path A results should therefore be interpreted primarily as evidence of structured database coverage, and the extent to which current results reflect independent cross-source verification should be interpreted cautiously.

Incomplete and imbalanced evidence. Literature retrieval covers only 40.5% of confirmed positives, and a much smaller fraction yields *Support* evidence, reflecting limited coverage and attribution challenges. In contrast, structured knowledge (DrugBank) provides the dominant signal for end-to-end performance. As a result, *Unknown* predictions may reflect missing evidence rather than true uncertainty, and current behavior is largely driven by structured sources.

LLM dependence and task scope. Although final arbitration is deterministic, Path B relies on LLMs for sentence selection and stance attribution, introducing model-dependent variability at the extraction stage. Furthermore, CROSSDDI formulates DDI assessment as pairwise binary verification and does not account for dosage, route, or patient-specific factors.

Future Work

Evaluation and benchmarks. A key limitation of current DDI resources is the absence of verified non-interaction labels. Constructing benchmarks with both confirmed interactions and validated negatives would enable reliable evaluation of precision and false-positive control in evidence-grounded settings.

Retrieval and evidence coverage. Improving literature-side coverage remains critical. Future work includes expanding retrieval to broader sources (e.g., full-text articles and regulatory data) and improving drug-name normalization to reduce loss from entity resolution and retrieval failures.

Stronger baselines and comparisons. Further evaluation against stronger RAG variants (e.g., few-shot or chain-of-thought prompting) and neural DDI models under matched retrieval conditions is needed to better characterize the strengths and limitations of the verification-first design.

Ethics Statement

CROSSDDI is a research prototype for evidence-grounded DDI assessment and is not intended to replace professional clinical judgment. DrugBank data were accessed under an academic research license. No patient-level or personally identifiable data were used in this study.

References

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2024. Self-rag: Learning to retrieve, generate, and critique through self-reflection. In *The Twelfth International Conference on Learning Representations (ICLR)*.

Xin Bi, He Ma, Jianhua Li, Yuliang Ma, and Deyang Chen. 2018. A positive and unlabeled learning framework based on extreme learning machine for drug-drug interactions discovery. *Journal of Ambient Intelligence and Humanized Computing*.

Haohuai He, Guanxing Chen, and Calvin Yu-Chian Chen. 2022. 3dgt-ddi: 3d graph and text based neural network for drug–drug interaction prediction. *Briefings in Bioinformatics*, 23(3):bbac134.

Wenzhun Huang, Xiao Wang, Yunhao Chen, Changqing Yu, and Shanwen Zhang. 2025. Advancing drug-drug interactions research: integrating ai-powered prediction, vulnerable populations, and regulatory insights. *Frontiers in Pharmacology*, 16:1618701.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Craig Knox, Michael Wilson, Christian M. Klinger, et al. 2024. Drugbank 6.0: the drugbank knowledgebase for 2024. *Nucleic Acids Research*, 52(D1):D1265–D1275.

Prajwol Lamichhane, Indika Kahanda, Xudong Liu, Karthikeyan Umapathy, Sandeep Reddivari, Catherine Christie, Andrea Arikawa, and Jenifer Ross. 2023. Poster: Biofactcheck: Exploring the feasibility of explainable automated inconsistency detection in biomedical and health literature. In *Proceedings of the 8th ACM/IEEE International Conference on Connected Health: Applications, Systems and Engineering Technologies*, pages 196–197.

Dongxu Li, Yue Yang, Ziwen Cui, Hengchuang Yin, Pengwei Hu, and Lun Hu. 2026. Llm-ddi: Leveraging large language models for drug-drug interaction prediction on biomedical knowledge graph. *IEEE Journal of Biomedical and Health Informatics*, 30(1):773–781.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. 2025. Qwen2.5 technical report.

Justine Sicard, François Montastruc, Coline Achalme, Annie Pierre Jonville-Bera, Paul Songue, Marina Babin, Thomas Soeiro, Pauline Schiro, Claire de Canecaude, and Romain Barus. 2025. Can large language models detect drug–drug interactions leading to adverse drug reactions? *Therapeutic Advances in Drug Safety*, 16:20420986251339358.

Karan Singhal, Shekoofeh Azizi, Tao Tu, S Sara Mahdavi, Jason Wei, Hyung Won Chung, Nathan Scales, Ajay Tanwani, Heather Cole-Lewis, Stephen Pfohl, et al. 2023. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180.

Yao Tian, Jiacai Yi, Ningning Wang, Chengkun Wu, Jinfu Peng, Shao Liu, Guoping Yang, and Dongsheng Cao. 2025. Ddinter 2.0: an enhanced drug interaction resource with expanded data coverage, new interaction types, and improved user interface. *Nucleic Acids Research*, 53(D1):D1356–D1362.

Juraj Vladika, Phillip Schneider, and Florian Matthes. 2024. Healthfc: Verifying health claims with evidence-based medical fact-checking. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 8095–8107.

David Wadden, Kyle Lo, Bailey Kuehl, Arman Cohan, Iz Beltagy, Lucy Lu Wang, and Hannaneh Hajishirzi. 2022a. Scifact-open: Towards open-domain scientific claim verification. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 4719–4734.

David Wadden, Kyle Lo, Lucy Lu Wang, Arman Cohan, Iz Beltagy, and Hannaneh Hajishirzi. 2022b. MultiVerS: Improving scientific claim verification with weak supervision and full-document context. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 61–76, Seattle, United States. Association for Computational Linguistics.

GuiShen Wang, Hui Feng, and Chen Cao. 2025a. Birnn-ddi: A drug-drug interaction event type prediction model based on bidirectional recurrent neural network and graph2seq representation. *Journal of Computational Biology*, 32(2):198–211.

Sheng Wang, Fangyuan Zhao, Dechao Bu, Yunwei Lu, Ming Gong, Hongjie Liu, Zhaohui Yang, Xiaoxi Zeng, Zhiyuan Yuan, Baoping Wan, et al. 2025b. Lins: A general medical q&a framework for enhancing the quality and credibility of llm-generated responses. *Nature communications*, 16(1):9076.

Christopher A White, Yehuda A Masturov, Eric Haunschild, Evan Michaelson, Dave R Shukla, and Paul J Cagle. 2025. Can chatgpt reliably answer the most common patient questions regarding total shoulder arthroplasty? *Journal of Shoulder and Elbow Surgery*, 34(5):e254–e264.

Di Wu, Jia-Chen Gu, Fan Yin, Nanyun Peng, and Kai-Wei Chang. 2024. Synchronous faithfulness monitoring for trustworthy retrieval-augmented generation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 9390–9406, Miami, Florida, USA. Association for Computational Linguistics.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. 2025. Qwen3 technical report.

Cyril Zakka, Rohan Shad, Akash Chaurasia, Alex R Dalal, Jennifer L Kim, Michael Moor, Robyn Fong, Curran Phillips, Kevin Alexander, Euan Ashley, et al. 2024. Almanac—retrieval-augmented language models for clinical medicine. *Nejm ai*, 1(2):AIoa2300068.

Chengcheng Zhang, Yao Lu, and Tianyi Zang. 2022. Cnn-ddi: a learning-based method for predicting drug–drug interactions using convolutional neural networks. *BMC Bioinformatics*, 23(Suppl 1):88.

Changpeng Zhao, Dongfang Han, Zicheng Zuo, and Turdi Tohti. 2025. Kgdb-ddi: Knowledge graph-based drug background data fusion model for drug–drug interaction prediction. *Artificial Intelligence in Medicine*, 168:103225.

Jing Zhu, Chao Che, Hao Jiang, Jian Xu, Jiajun Yin, and Zhaoqian Zhong. 2024. Ssf-ddi: a deep learning method utilizing drug sequence and substructure features for drug–drug interaction prediction. *BMC bioinformatics*, 25(1):39.

A Pipeline Overview

B Implementation Details for Path A

Retrieval parameter. The parameter K bounds the number of DrugBank records returned per query. We set $K = 3$ by default, yielding a compact evidence set. Records are ranked by specificity: exact pair-key matches are ranked before lexical description matches and lower-specificity counterpart matches. Let $R_A^{(K)}$ denote this retained ranked output. Path A emits *Support* iff an exact pair-key match appears in $R_A^{(K)}$; otherwise it emits *Unknown*. Thus K controls both the retained evidence set and the retrieval output used for Path A stance assignment.

Interpretive reporting module. The retained DrugBank records are additionally processed by an LLM-based summarization module that produces structured summaries and natural-language descriptions. These outputs are used for interpretation only and do not affect the deterministic Path A stance defined in §3.2.

C Implementation Details for Path B

Alias expansion. Path B expands normalized drug names using a small fixed alias map covering common generic and brand-name variants. When no aliases are available, retrieval uses the normalized name.

Retrieval scope. Path B issues synonym-expanded queries, aggregates PubMed results, deduplicates by PMID, and retains the first three unique articles D_B.

Joint evidence extraction. The retrieved documents are processed by an LLM-based module (Prompt B.1) that extracts up to six candidate sentences, each labeled as *Support*, *Contradict*, or *Unknown*, together with a short rationale.

Post-extraction guardrails. Two deterministic rules are applied after extraction. (i) *Pairwise relevance*: sentences that do not match both queried drugs are downgraded to *Unknown*. (ii) *Contradiction filter*: *Contradict* is retained only when explicit negation and interaction cues are both present; otherwise it is downgraded to *Unknown*. These rules favor precision over recall.

Guardrail cue lists. The following cue lists are fixed across all experiments. Matching is case-insensitive and implemented as substring matching over the extracted sentence.

Explicit pairwise negation cues (used to retain *Contradict* labels): `no interaction`, `no significant interaction`, `no clinically meaningful interaction`, `did not identify an interaction`, `does not interact`, `no effect on`, `not associated with an interaction`.

Interaction-related cues (required alongside a negation cue to retain *Contradict*): `interaction`, `interact`, `coadministration`, `co-administration`, `concomitant`, `combined`, `bleeding risk`, `toxicity`, `exposure`, `serum level`, `metabolism`, `excretion`, `anticoagulant`, `inr`, `monitor`, `contraind`.

Interpretive reporting module. Retrieved documents are also summarized by an LLM for interpretive purposes. These summaries do not affect path-level stance or final decisions.

D Accuracy Under PU Evaluation

Table 3 reports accuracy for completeness. Accuracy is not a reliable measure under PU evaluation, as unlabeled pairs may contain true interactions and are treated as negatives.

E Structured Coverage Analysis

DDInter and DrugBank are both curated DDI resources, and their interaction coverage may partially overlap, potentially contributing to strong performance for systems relying on structured DrugBank evidence. We therefore analyze direct DrugBank coverage in the randomly sampled 1,000-pair DDInter evaluation set.

Under the deterministic resolver used by Path A, 558 pairs (55.80%) have a direct DrugBank pair match and 442 pairs (44.20%) do not. We then evaluate recall over confirmed positives separately

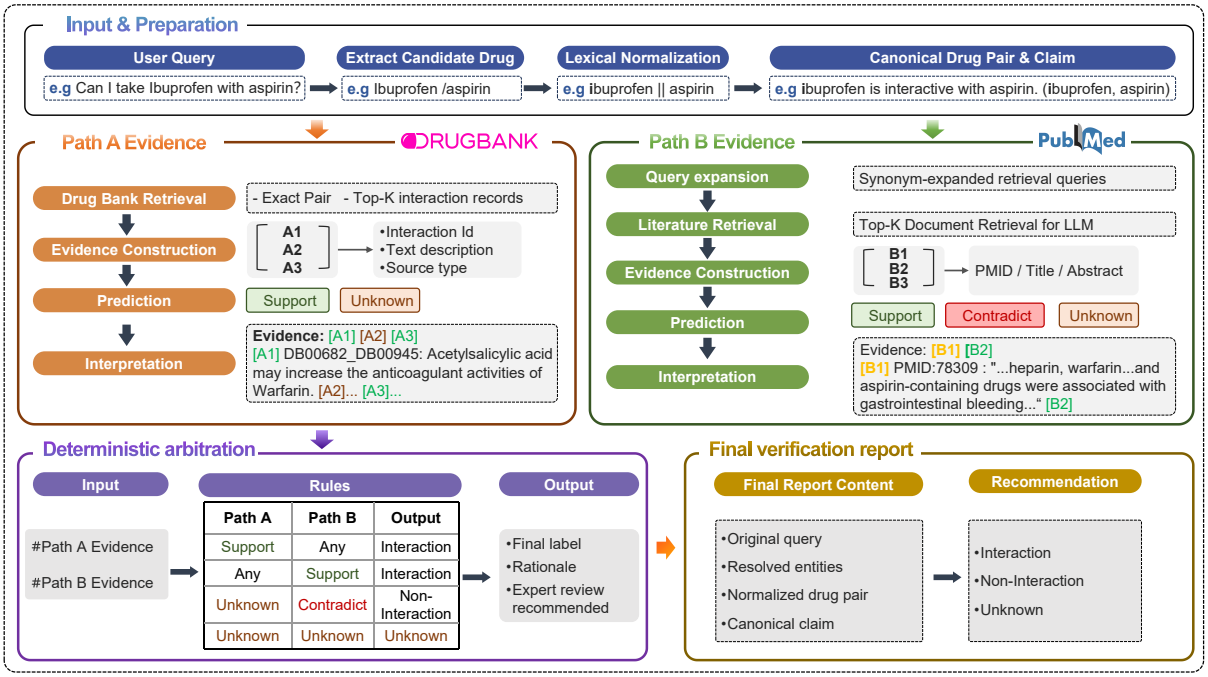

Figure 2: CROSSDDI pipeline. A drug-pair query is normalized into a canonical interaction claim and verified along two evidence paths: DrugBank (Path A) and PubMed (Path B), which provide structured and literature-based evidence, respectively. A deterministic arbitration rule combines path-level outputs into a final verdict linked to explicit supporting evidence.

for the two subsets. With direct DrugBank coverage, CROSSDDI recalls 399 of 465 positives (85.81%); without such coverage, it recalls 63 of 332 positives (18.98%). This gap suggests that structured database coverage strongly influences observed end-to-end recall.

F Prompt Template for Path B Evidence Extraction

Prompt B.1 is the only LLM component whose outputs directly influence the final decision, as it determines the evidence sets used for stance assignment.

Prompt B.1: Literature Joint Extraction.

```
You are an expert clinical pharmacologist
    performing joint
evidence extraction and stance labeling for
    PubMed drug-drug
interaction verification.
From the provided Top-K PubMed documents,
    extract the most
claim-relevant evidence sentences and assign
    each one a label:
Support, Contradict, or Unknown.

Rules:
- Prefer sentences that explicitly describe
    interaction,
  coadministration risk, bleeding risk, toxicity,
      exposure
  change, monitoring concern, or lack of
      interaction.
- Do not select sentences that merely co-mention
    both drugs
  unless the sentence discusses their pairwise
      relationship.
- Use Contradict only when the sentence
    explicitly denies a
  pairwise interaction or pairwise effect.
- Use Unknown when the sentence is indirect,
    weak,
  not pair-specific, or insufficient to resolve
      the claim.
- Return at most the requested number of
    sentences.
- Return valid JSON only.
- Use exactly this schema:
  {"selected_evidence":[{"pmid":"...","title
      ":"...",
  "sentence":"...","score":0.0,
  "source_type":"abstract_sentence",
  "label":"Support|Contradict|Unknown",
  "rationale":"one concise evidence-grounded
      sentence",
  "selection_rationale":"why this sentence was
      selected"}]}.
- Do not wrap the JSON in markdown fences.

Canonical claim: {claim}
Maximum sentences: {max_sentences}

Top-K PubMed documents:
{context}
```

Model	Setting	P.Int	P.Int 95% CI	Recall	P.Unk	P.NI	Acc.
Qwen2.5-7B	LLM-only	0.385	[.355,.413]	0.386	0.510	0.104	0.404
	RAG	0.546	[.513,.578]	0.566	0.434	0.019	0.553
	Path A	0.541	[.508,.573]	0.565	0.459	0.000	0.562
	Path B	0.051	[.037,.067]	0.041	0.937	0.004	0.215
	CROSSDDI	**0.556**	[.524,.587]	**0.580**	0.435	0.000	0.568
GPT-4o	LLM-only	0.351	[.320,.381]	0.407	0.311	0.338	0.372
	RAG	0.478	[.445,.511]	0.512	0.517	0.003	0.537
	Path A	0.541	[.508,.573]	0.565	0.459	0.000	0.562
	Path B	0.089	[.070,.110]	0.077	0.868	0.001	0.225
	CROSSDDI	**0.575**	[.543,.606]	**0.593**	0.424	0.000	0.574
Qwen3-235B	LLM-only	0.320	[.291,.350]	0.384	0.580	0.100	0.465
	RAG	0.564	[.531,.596]	**0.578**	0.432	0.004	0.557
	Path A	0.541	[.508,.573]	0.565	0.459	0.000	0.562
	Path B	0.083	[.064,.104]	0.074	0.908	0.001	0.236
	CROSSDDI	0.557	[.525,.588]	0.576	0.427	0.001	0.561

Table 3: Full results including accuracy. Confidence intervals correspond to P.Int. Accuracy is shown for completeness but should not be interpreted as a reliable measure under PU evaluation.

Subset	Pairs	Share (%)	Positives	Recall (%)
Direct DrugBank match	558	55.80	465	85.81
No direct match	442	44.20	332	18.98

Table 4: Direct DrugBank coverage and end-to-end recall over confirmed positives in the randomly sampled 1,000-pair DDInter evaluation set.

G RAG Baseline Prompt

For transparency, we report the full prompt template used by the RAG baseline. The baseline receives the same retrieved DrugBank and PubMed evidence used by the verification paths. DrugBank evidence is formatted as up to three structured records, each containing an evidence ID, DrugBank interaction ID, counterpart name, retrieval rank bucket, severity, evidence level, and description. PubMed evidence is formatted as up to three retrieved articles, each containing an evidence ID, PMID, title, journal, publication date, and abstract. If no evidence is retrieved from a source, the corresponding section explicitly states that no records or documents were retrieved.

Prompt RAG: Interaction Verdict Generation.

```
You are an expert clinical pharmacologist
    evaluating a possible
drug-drug interaction.
Your task is to determine whether there is an
    interaction
between {drug_a} and {drug_b} based only on the
    retrieved
evidence below. Do not use outside knowledge.

DrugBank Evidence:
{drugbank_context}

PubMed Evidence:
{pubmed_context}

Instructions:
1. Use only the evidence provided above.
2. Classify the pair as "Interaction" if the
    evidence
   explicitly supports an interaction, mechanism,
       or
   clinical risk involving both drugs.
3. Classify the pair as "Non-Interaction" if the
     evidence
   explicitly states there is no interaction
       between the drugs.
4. Classify the pair as "Unknown" if the
    evidence is
   insufficient, indirect, conflicting, or does
       not explicitly
   establish either of the above.
5. If the evidence is mixed or unresolved,
    choose "Unknown".
6. Return valid JSON only.

Output schema:
{"rationale": "One or two sentences grounded in
    the retrieved
evidence", "final_label": "Interaction | Non-
    Interaction |
Unknown"}
```