DIMAS-OMOP: A Deliberative Intelligence-Based Multi-Agent System for Chinese Medical Text Standardization toward OMOP

Hanlin Lv, MD, Xiao Wang, PhD, Lei Wang, MS*
RYTECH, Wuhan, 430074, China

Kesong Wu, MS
Ian Frazer Centre for Children's Immunotherapy Research
The University of Queensland, Brisbane, QLD, 4101, Australia

Lei Li, PhD*
School of Life Science and Technology
Southeast University, Nanjing, 210096, China
*

Abstract

Standardizing Chinese clinical imaging reports within the Observational Medical Outcomes Partnership (OMOP) framework is hindered by linguistic complexity and output inconsistency in existing methods. We propose DIMAS-OMOP, a Deliberative Intelligence-based Multi-Agent System designed for high-fidelity medical concept mapping toward OMOP standardization. Moving beyond single-model architectures, DIMAS-OMOP employs a hybrid three-stage workflow that integrates traditional natural language processing modules with selective Large Language Model reasoning and Retrieval-Augmented Generation. The core innovation lies in a hierarchical six-agent proposer-skeptic deliberation mechanism, complemented by a dynamic concept resolution approach and a four-dimensional quality control framework. Experimental results on 1,250 imaging reports demonstrate that DIMAS-OMOP achieves 95.2% mapping accuracy, significantly outperforming rule-based methods (+21.8 percentage points) and single-AI baselines (+8.1 percentage points). The system maintains a throughput of 1,200 reports/hour, with the multi-agent deliberation stage alone contributing an 8.9% relative accuracy gain. Furthermore, pilot deployment shows a 160.6% return on investment and a 31.5% increase in workflow efficiency. This study provides a novel, robust methodology for integrating unstructured non-English clinical data into the global Observational Health Data Sciences and Informatics (OHDSI) ecosystem through deliberative intelligence.

*Corresponding authors:
Lei Wang, raywong2121@gmail.com.
Lei Li, lei.li@seu.edu.cn

1 Introduction

1.1 Observational Research and Real-World Evidence Data Requirements

Modern medical research is experiencing a paradigm shift from traditional randomized controlled trials (RCT) to large-scale observational studies and Real-World Evidence (RWE). The Observational Medical Outcomes Partnership (OMOP) Common Data Model (CDM) launched by the Observational Health Data Sciences and Informatics (OHDSI) collaborative provides a unified framework for global medical data standardization, allowing large-scale observational studies across institutions and regions (Hripcsak et al., 2015; Suchard et al., 2019). The core of such research lies in extracting valuable information and features from massive unstructured medical data to meet complex analytical needs such as high-dimensional causal inference, drug safety monitoring, and treatment effectiveness evaluation. Clinical imaging reports, as an important component of medical records, contain rich diagnostic information, anatomical structure descriptions, pathological findings, and measurement data. This information is crucial for defining high-quality observational research cohorts, disease phenotypes, and evaluating treatment outcomes. However, imaging reports are unstructured free-text, containing dense professional terminology, abbreviations, and complex anatomical-pathological relationships. This necessitates precise information extraction to map concepts to the SNOMED CT vocabulary and populate standardized records within the OBSERVATION, MEASUREMENT, and FACT_RELATIONSHIP tables of the OMOP CDM v5.4 (Reps et al., 2018; Voss et al., 2015).

Proceedings of the 1st Workshop on Multilinguality in the Era of Large Language Models (MeLLM 2026), pages 108–118

1.2 Technical Evolution and Methodological Challenges

Technical Limitations Before LLMs Before Large Language Models (LLMs), medical text information extraction primarily relied on supervised or weakly supervised learning with inherent performance ceilings. BERT family pre-trained models, while performing well in medical text understanding, depended highly annotated corpora and pre-structured template designs (Lee et al., 2020; Alsentzer et al., 2019). This dependency constrained generalization across new domains and non-standard lexical variations, thereby limiting adaptability to the rapid evolution of medical terminology and the diversity of clinical language. Another key limitation of traditional methods lies in their dependence on large amounts of high-quality annotated data. Professional annotation in the medical domain requires experts with medical backgrounds, resulting in high costs and long cycles. Meanwhile, rule-based methods, while achieving high accuracy in specific scenarios, suffer from serious hard-coding problems and struggle to handle linguistic variations, abbreviations, and context-dependent semantic understanding in Chinese medical texts (Wang et al., 2018; Uzuner et al., 2011).

New Challenges After LLMs With LLMs, numerous prompt engineering and in-context Learning paradigms have emerged, bringing new possibilities for medical text processing. However, these methods have introduced new technical challenges. Low-parameter LLMs have extremely limited capabilities in complex medical reasoning tasks, struggling to accurately understand subtle differences in medical concepts and complex anatomical-pathological relationships (Singhal et al., 2023; Thirunavukarasu et al., 2023). High-parameter models have stronger reasoning capabilities, but their processing speed and alignment exhibit random fluctuations, making consistent engineering standards difficult to achieve.

Privacy Sensitivity and Local Deployment Requirements Clinical texts contain extensive protected health information (PHI). Processing them through closed-source public LLMs risks data leakage and patient privacy violations (Lee et al., 2023; Nori et al., 2023). Medical institutions require local deployment to ensure patient data never leaves institutional boundaries. Therefore, extraction of local deployment-based, reliable and engineering-grade imaging text terminology becomes our core focus. Our goal is to develop a system that can efficiently and accurately convert Chinese clinical imaging reports to OMOP CDM v5.4 standard format while protecting patient privacy. The system needs to utilize the SNOMED CT terminology system for precise concept alignment and, following OHDSI community best practices, appropriately distribute extracted information into the target OMOP standardized tables (Bodenreider, 2004; Overhage et al., 2012). Prompt engineering also suffers from sensitivity and output inconsistency when applied to medical domain texts. The non-deterministic nature of different prompt designs can yield disparate outputs for identical inputs, introducing an impermissible level of stochasticity within clinical decision support contexts. Meanwhile, most high-performance LLMs rely on API calls, raising cost and latency concerns while fundamentally precluding use due to stringent medical data privacy mandates (Meskó and Topol, 2023; Chen et al., 2023).

1.3 Research Objectives and Contributions

Leveraging the insights above, this research aims to design and evaluate a locally deployable system for the secure, robust, and computationally efficient conversion of Chinese clinical imaging reports into the OMOP CDM v5.4 standard. We therefore propose a **D**eliberative **I**ntelligence-based **M**ulti-**A**gent **S**ystem for Chinese Medical Text Standardization toward OMOP (DIMAS-OMOP), a hybrid technical architecture that integrates traditional natural language processing (NLP), selective LLM reasoning, and multi-agent collaboration.

Hybrid Technical Approach DIMAS-OMOP operates through a structured three-stage workflow: the system first establishes a reliable foundation by capturing explicit clinical entities, then selectively invokes reasoning capabilities to resolve semantic ambiguities, and finally employs a collaborative deliberation mechanism to gate-keep the quality of standardized outputs.

Key Innovative Contributions Our contributions are fourfold: (1) Architectural Innovation: a hybrid multi-agent collaborative framework integrating the stability of traditional NLP, the reasoning capabilities of LLMs, and the decision reliability of multi-agent systems; (2) Methodological Innovation: a dynamic concept resolution algorithm using multi-strategy search and adaptive threshold adjustment to minimize hard-coding; (3) Application Innovation: specialized OMOP CDM

v5.4 tools for Chinese clinical imaging reports with SNOMED CT alignment and adherence to OHDSI best practices; (4) Quality Assurance Innovation: a four-dimensional quality control system based on deliberative intelligence to ensure medical accuracy and engineering reliability.

2 Methods

2.1 Study Design

Design Objectives for Observational Research The core design objective is to efficiently and accurately extract structured information from Chinese clinical imaging reports to support downstream OMOP-based applications, including drug safety monitoring and comparative effectiveness research.

Hybrid Technical Architecture Design Principles Considering the technical evolution of clinical NLP and LLMs, our approach is implemented through a three-stage process of "extraction, reasoning, and deliberation" (Figure 1):

Stage 1 **- Baseline Extraction**: Employs stable NLP modules to ensure efficient coverage and reliable capture of foundational information. The Linguist and Structurer Agents perform text normalization and decompose clinical entities into atomic components, such as Chinese medical Named Entity Recognition (NER), terminology recognition, and numerical extraction;

Stage 2 **- Intelligent Reasoning**: An on-demand process that selectively invokes locally deployed LLMs and Retrieval-Augmented Generation (RAG) to generate evidence and provide auxiliary judgment when addressing complex semantic inference, concept disambiguation, or relationship determination characterized by high uncertainty;

Stage 3 **- Collaborative Deliberation**: Finalizes decisions through an adversarial proposer-skeptic loop. The Mapping Proposer Agent generates candidates via agentic reasoning, while the Skeptical Critic Agent challenges the results to ensure clinical accuracy. Subsequently, the Synthesis Audit Agent performs final validation based on the complete deliberation history. Verified outputs must pass a four-dimensional quality gate before the OMOP Builder populates the records into the corresponding CDM v5.4 clinical tables (Wang et al., 2020; Rudin, 2019).

2.2 Overall System Architecture Design

Hierarchical Multi-Agent Collaborative Architecture The DIMAS-OMOP system operates

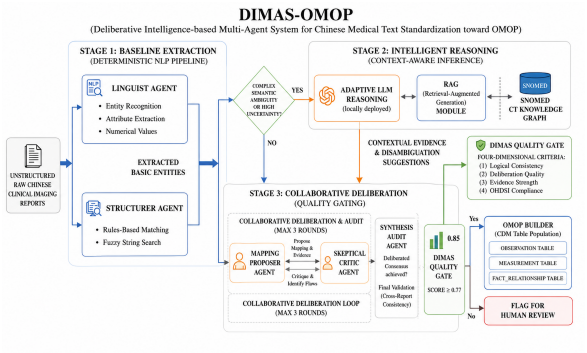

Figure 1: Overall Hybrid Architecture of DIMAS-OMOP: A three-stage pipeline integrating deterministic NLP with multi-agent adversarial deliberation.

through a hierarchical six-agent collaborative framework, organized into three specialized functional teams under the governance of a task orchestrator. This division of labor is designed to decouple foundational text analysis from high-level concept mapping and final OMOP construction, thereby enabling a standardization process that balances clinical depth with engineering rigor.

Task Orchestrator: Responsible for global state management, tool provisioning, and deliberation process control. The orchestrator implements asynchronous task processing, supports concurrent processing of multiple medical texts, and maintains complete deliberation history records.

Team 1 - Text Deep Analysis Team: Linguist Agent performs text normalization, grammatical analysis, ambiguity resolution, and semantic segmentation. Structurer Agent strictly follows OHDSI separation of concerns principles to decompose concepts into independent standard components.

Team 2 - Mapping and Deliberation Committee: Mapping Proposer Agent uses Reasoning-Acting (ReAct) mode to rapidly generate concept mapping proposals. Skeptical Critic Agent specifically challenges and validates proposer's mapping results to ensure quality control.

Team 3 - Final Audit and Construction Team: Synthesis Audit Agent performs final validation based on complete deliberation history. OMOP Builder, following OHDSI best practices, converts validated concept mappings to OMOP CDM v5.4 standard format and appropriately distributes them to corresponding CDM tables.

System Communication Protocol and Data Flow The system employs an asynchronous communication mechanism based on message passing,

with all agents interacting through standardized data structures. The data flow follows a regulated, phased processing pattern:

Raw Chinese Clinical Imaging Reports → [Linguist Agent] → Normalized Text & Semantic Segments → [Structurer Agent] → Atomic Entity Components → [Mapping Proposer Agent ↔ Skeptic Deliberation Agent] → Validated Concept Mapping → [Synthesis Audit Agent] → Four-Dimensional Quality Verification → [OMOP Builder] → Standardized CDM v5.4 Records → Local Database Storage

2.3 Core Deliberative Intelligence Algorithm Design

Proposer-Skeptic Adversarial Mechanism The core of the deliberative intelligence algorithm (Algorithm 1) is the proposer-skeptic adversarial collaboration mechanism. This mechanism ensures high-quality and reliable concept mapping through multi-round deliberation:

Algorithm 1: Deliberative Mapping Process

Input: Structured Entity E, Concept Database DB
Output: Validated Mapping M or NULL

```
PROPOSE_PHASE:
    mapping_proposal = MappingProposer.analyze(E, DB)

CRITIQUE_PHASE:
    challenges = SkepticalCritic.evaluate(mapping_proposal, E, DB)

DELIBERATION_PHASE:
    IF challenges.isEmpty():
        RETURN mapping_proposal
    ELSE:
        refined_proposal = MappingProposer.refine(mapping_proposal, challenges)
        GOTO CRITIQUE_PHASE

CONSENSUS_EVALUATION:
    IF consensus_achieved(deliberation_history):
        RETURN final_mapping
    ELSE:
        RETURN NULL
```

Algorithm 1: Deliberative Mapping Process

Consensus Achievement Determination Mechanism The system employs a multi-dimensional consensus evaluation mechanism that comprehensively considers the number of challenges, proposal confidence, and deliberation rounds. Specifically, the consensus is reached when either no challenges exist, or confidence ≥ 0.8 with ≤ 1 challenge, or max rounds reached with confidence ≥ 0.7.

The mathematical model is:

$$\text{Consensus}(h, c, r) = \begin{cases} \text{True} & \text{if } |\text{challenges}| = 0 \\ \text{True} & \text{if confidence} \geq 0.8 \\ & \text{and } |\text{challenges}| \leq 1 \\ \text{True} & \text{if confidence} \geq 0.7 \\ & \text{at max rounds} \\ \text{False} & \text{otherwise,} \end{cases} \tag{1}$$

where h is the deliberation history, c is the current proposal, and r is the current round.

2.4 Zero-Hardcoded Dynamic Concept Resolution

Design Philosophy to Overcome Traditional Method Limitations Addressing the hard-coding limitations of rule-based and supervised learning methods before LLMs, as well as the prompt sensitivity and consistency issues after LLMs, we design a completely dynamic concept resolution mechanism. This mechanism avoids the heavy dependence of traditional BERT family models on pre-structured templates while circumventing the random fluctuation problems of high-parameter LLMs in medical reasoning. The system achieves dynamic adaptation to the rapid evolution of medical terminology through multi-strategy search and adaptive threshold adjustment.

Multi-Strategy Concept Search Algorithm (Algorithm 2) The system implements a four-level matching strategy for SNOMED CT terminology alignment, ensuring high-precision concept mapping in local environments that protect patient privacy: exact string matching (1.0), synonym matching using OMOP synonym tables (0.95), fuzzy matching using edit distance (0.8), and semantic matching using SapBERT similarity (0.7).

Algorithm 2: Multi-Strategy Concept Search

Input: Medical Term T, Concept Database DB
Output: Ranked Candidate List C

```
Initialize candidates C = [ ]
FOR EACH strategy S IN [Exact, Synonym, Fuzzy, Semantic]:
    matches ← S.search(T, DB)
    FOR EACH match IN matches:
        score ← S.weight × match.similarity
        C.add((match.concept, score))
SORT C BY score DESCENDING
APPLY dynamic threshold filtering
RETURN top-k candidates
```

Algorithm 2: Multi-Strategy Concept Search Algorithm

Dynamic Threshold Adjustment The system implements adaptive threshold adjustment based on term complexity and context:

$$\text{Threshold}_{dynamic} = \text{Threshold}_{base} \times (1 + \alpha\text{Complexity}_{term} + \beta\text{Context}_{uncertainty}), \tag{2}$$

where α and β are learned parameters, complexity factors include term length, medical domain specificity, and ambiguity level.

2.5 Quality Control and Validation Mechanism

Quality Assurance System for Observational Research To meet the rigorous data quality standards

required for RWE generation, the system implements a four-dimensional assessment framework tailored for OMOP CDM v5.4 and SNOMED CT alignment. This mechanism ensures that extracted information adheres to OHDSI community best practices across all target clinical domains. By prioritizing structural integrity and medical precision, the system provides a high-fidelity data foundation suitable for complex analytical tasks, such as high-dimensional causal inference and drug safety monitoring.

Four-Dimensional Audit Standards The system establishes a comprehensive quality assessment system based on four dimensions: logical consistency (30%) to evaluate reasoning chain completeness, deliberation quality (25%) to evaluate discussion adequacy and consensus level, evidence strength (25%) to evaluate supporting evidence for mapping decisions, and OHDSI compliance (20%) to evaluate standard adherence level.

Overall Score Calculation:

$$\text{Score}_{\text{overall}} = \sum_{i=1}^{4} \omega_i \cdot \text{Score}_i, \tag{3}$$

where ω_i is the weight of the i-th dimension and $Score_i$ is the score of the i-th dimension.

Quality Gate Mechanism The system implements multi-level quality gates:

- Level 1: Automatic pass for high-confidence unanimous decisions (confidence ≥ 0.9, no challenges)
- Level 2: Secondary review for medium-confidence decisions ($0.7 \leq$ confidence < 0.9)
- Level 3: Human review required for low-confidence or conflicting decisions (confidence < 0.7 or unresolved challenges)

2.6 Experimental Design

Dataset Construction We constructed a comprehensive evaluation dataset consisting of 1250 Chinese clinical imaging reports retrospectively collected from three tertiary hospitals in China. For confidentiality, the institutions are not named in the manuscript. Two participating hospitals also served as pilot deployment sites. The dataset includes 412 CT reports (33.0%), 318 MRI reports (25.4%), 287 X-ray reports (23.0%), and 233 ultrasound reports (18.6%). Each report was manually annotated by medical informatics experts following OMOP CDM v5.4 standards and SNOMED CT terminology guidelines.

Evaluation Metrics Primary evaluation metrics include concept mapping accuracy (percentage of correctly mapped concepts), processing speed (reports processed per hour), and F_1-score. Quality metrics comprise average deliberation rounds per report, consensus achievement rate (percentage of cases reaching consensus), and explainability score (based on deliberation history completeness and reasoning chain clarity).

Baseline Method Comparison

Rule-Based Method: A deterministic implementation based on regular expressions and curated medical dictionaries, including manually crafted mapping rules and edit distance-based fuzzy matching algorithms.

Single AI Method: A standalone transformer-based model performing direct concept recognition followed by similarity-based concept matching, without retrieval augmentation, multi-agent review, or deliberation mechanism.

Only reproducible baselines with transparent implementation details were retained in the final comparative analysis.

3 Results

On the test dataset of 1250 Chinese clinical imaging reports, the DIMAS-OMOP system demonstrated superior performance compared to the selected baselines, achieving statistically significant improvement ($p < 0.05$) in concept mapping accuracy, processing efficiency, and quality control reliability. Details are in Figure 2 (A).

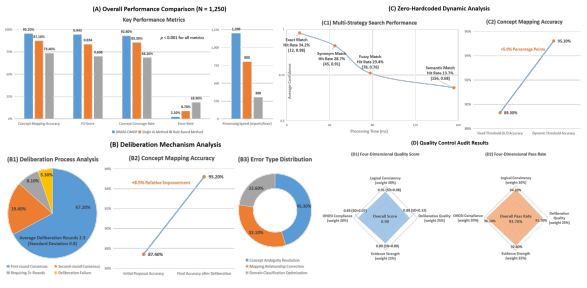

Figure 2: Highlights of DIMAS-OMOP Performance.

Paired t-tests confirmed that the performance gains of DIMAS-OMOP over both baseline methods reached high statistical significance ($p < 0.001$). Furthermore, effect size analysis revealed a substantial practical impact under current experimental setting:

- DIMAS-OMOP vs Rule-based method: Cohen's d = 2.34 (Large effect)
- DIMAS-OMOP vs Single AI method: Cohen's d = 1.67 (Large effect)

3.1 Deliberative Intelligence Mechanism Analysis

Figure 2 (B1) presents the impact of the deliberation mechanism on concept mapping accuracy under the defined experimental settings. Quality Improvement Effect is shown in Figure 2 (B2), indicating a 8.9% relative improvement. Error Type Distribution is presented in Figure 2 (B3).

3.2 Zero-Hardcoded Dynamic Resolution Effects

Analysis of multi-strategy concept search algorithm performance at different levels is shown in Figure 2 (C1). Dynamic Threshold Adjustment Effects are shown in Figure 2 (C2).

3.3 Quality Control System Validation

Quality assessment results based on four-dimensional audit standards showed that the DIMAS-OMOP system performed excellently in all dimensions. Details are in Figure 2 (D). The quality gate mechanism effectively prevented the generation of low-quality mappings:

- Mappings Blocked by Quality Gates: 6.3%
- Reprocessing Success Rate After Blocking: 78.4%
- Final Accuracy of Mappings Passing Quality Gates: 98.7%

3.4 Real Deployment Effectiveness Validation

Pilot deployment (Table 1) at two anonymized hospitals that were part of the same three-hospital participating network used for data collection showed 99.0% average availability and 31.5% workflow efficiency improvement.

Metric	Hospital A	Hospital B	Average
Reports Processed	15420	12680	14050
System Availability	99.2%	98.8%	99.0%
User Satisfaction	4.1/5.0	3.9/5.0	4.0/5.0
Data Quality Improvement	23%	19%	21%
Work Efficiency Improvement	35%	28%	31.5%

Table 1: Pilot Deployment Results.

Item	Cost/Benefit (10K RMB)
System Development Cost	-120
Deployment & Maintenance Cost	-45
Labor Cost Savings	+280
Data Quality Improvement Benefits	+150
Net Benefit	**+265**

Table 2: Cost-Benefit Analysis (Annual).

Annual cost-benefit analysis (Table 2) indicates a Return on Investment (ROI) of 160.6%.

4 Discussion

4.1 Comparison with Existing Research

Our work aligns with the broader movement toward OMOP-standardized observational data infrastructures (Hripcsak et al., 2015; Voss et al., 2015). While clinical NLP research has historically focused on English-language datasets (Bazoge et al., 2023), DIMAS-OMOP provides a specialized framework for Chinese clinical imaging reports. Unlike existing toolkits such as the Open Health NLP Toolkit that primarily serve English phenotyping (Wen et al., 2024), our system is engineered for the semantic complexities of Chinese clinical narratives.

Existing clinical NLP systems including MetaMap, cTAKES, CLAMP, and medspaCy offer mature component-based pipelines for concept extraction (Aronson and Lang, 2010; Savova et al., 2010; Soysal et al., 2018; Eyre et al., 2022), while tools like Usagi focus on terminology pairing (OHDSI, 2021). DIMAS-OMOP is complementary to these tools rather than a replacement: NLP pipelines can serve upstream extraction, and Usagi-style review can support downstream validation. For Chinese biomedical NER, domain-adapted models such as MC-BERT and imConvNet have demonstrated strong performance (Zhang et al., 2020; Zheng et al., 2022), but these are best positioned as extraction modules. Our focus is on cross-representation concept determination, ambiguity resolution, and OMOP-compliant table population. Regarding LLM and RAG techniques, while they improve knowledge coverage (Lewis et al., 2020; Gao et al., 2023; Brown et al., 2020), they face challenges in prompt sensitivity and hallucination (Shah, 2024). DIMAS-OMOP uses LLM/RAG selectively under explicit deliberation and quality

constraints to balance reasoning power with controllability.

Before LLMs, clinical text processing relied on deterministic pipelines and discriminative models using corpora such as MIMIC and PhysioNet (Johnson et al., 2016; Goldberger et al., 2000; Chapman et al., 2001). Rule-based methods like NegEx remained influential for their transparency (Chapman et al., 2001). The post-LLM era introduced in-context learning and retrieval-augmented workflows (Singhal et al., 2023; Thirunavukarasu et al., 2023; Lee et al., 2023), but also raised concerns about prompt dependence, reproducibility, and clinical safety (Meskó and Topol, 2023; Shah, 2024). DIMAS-OMOP synthesizes both paradigms: it retains mature NLP for first-pass extraction, invokes LLM/RAG only when semantic ambiguity remains, and applies proposer-skeptic deliberation with quality gates to reduce unverified output (Wang et al., 2020; Rudin, 2019).

4.2 Multi-Agent Systems in Medical Informatics

Multi-agent systems have a long history in distributed AI and coordinated decision-making (Stone and Veloso, 2000; Wooldridge, 2009; Jennings et al., 1998). Subsequent research in cooperative and adversarial multi-agent learning refined algorithms for coordination, debate, and consensus formation (Tampuu et al., 2017; Foerster et al., 2018; Rashid et al., 2020). Recent medical studies have applied multi-agent LLM frameworks to mitigate diagnostic cognitive bias and improve diagnostic capability (Ke et al., 2024; Chen et al., 2025). Relative to those clinically oriented systems, our contribution lies in applying adversarial multi-agent reasoning to terminology standardization and data engineering rather than direct bedside decision support.

4.3 Challenges and Breakthroughs in Chinese Medical NLP

Chinese clinical NLP remains challenging due to lexical variability, compact syntax, contextual dependence, and scarce non-English benchmarks (Wang et al., 2018; Uzuner et al., 2011; Bazoge et al., 2023). English clinical NLP has benefited from public data ecosystems and mature rule-based components (Johnson et al., 2016; Chapman et al., 2001), whereas Chinese imaging reports often compress anatomy, pathology, and measurements into highly compact expressions. DIMAS-OMOP addresses this through deterministic normalization, dynamic concept resolution, and structured deliberation.

4.4 Theoretical Contributions and Methodological Innovation

This study makes three theoretical contributions. First, it operationalizes a deliberative architecture for high-stakes informatics tasks, extending the argument that medical AI systems should favor transparent reasoning and controllable decision mechanisms (Wang et al., 2020; Rudin, 2019). Second, the dynamic concept-resolution module reduces dependence on hard-coded rules and improves maintainability as local terminology evolves. Third, the four-dimensional quality audit separates proposal generation, critique, evidence review, and OHDSI-compliance checking, rather than treating concept mapping as a single-step prediction problem.

4.5 Clinical Application Value and Practical Significance

Higher-quality standardization supports the broader OHDSI evidence-generation workflow, including network studies, patient-level prediction, comparative effectiveness research, and safety surveillance (Schuemie et al., 2020; Ryan et al., 2013; Hernán and Robins, 2016). In our study, DIMAS-OMOP improved mapping quality and workflow efficiency, offering practical value for institutions building OMOP-based infrastructure. For Chinese institutions specifically, this provides more reliable conversion of free text to OMOP-compatible representations and improved readiness for cross-site collaboration within the global OHDSI ecosystem.

4.6 Long-term Impact on Medical Informatics

Beyond immediate performance gains, this work enables the conversion of unstructured Chinese clinical data into reusable observational research pipelines, which can strengthen causal inference and pharmacovigilance analyses (Schuemie et al., 2020; Ryan et al., 2013). Our research also provides a preliminary example of privacy-conscious local deployment. A central challenge in the LLM era is benefiting from advanced models without transferring protected health information outside institutional boundaries (Lee et al., 2023; Meskó and Topol, 2023). The local-first architecture of DIMAS-OMOP is a practical design choice addressing this concern. Finally, this work contributes to the limited literature on non-English

OMOP-oriented clinical NLP, offering a potentially transferable framework for other Asian languages (Bazoge et al., 2023).

4.7 Study Limitations

Our study has several limitations. Although we used 1,250 real imaging reports, the dataset remains modest in scale and derived from only three tertiary hospitals, which may limit generalization to other institutions and reporting styles. Future research should include broader multi-center validation with held-out test sets. The evaluation was primarily based on aggregate metrics such as accuracy and processing speed. Future work would benefit from reporting inter-annotator agreement, exact data partitioning, error analysis for edge cases, and external validation. Additionally, user satisfaction data came mainly from medical informatics personnel; broader assessment from clinicians and data engineers would provide a more complete picture of usability.

4.8 Future Research Directions

Future work will focus on several directions. First, we plan to optimize the system architecture while reducing deployment complexity through model compression, knowledge distillation, and modular orchestration, enabling deployment on a wider range of local hardware. We also plan to incorporate newer medical-domain language models and explore federated learning strategies that avoid transferring raw clinical text across institutions. Second, we will expand the application scope to additional document types including pathology reports, laboratory reports, and operative notes, as well as multilingual extensions for Japanese and Korean. Third, we will conduct longitudinal studies in real clinical environments to assess downstream effects on cohort definition, causal inference, and pharmacovigilance, with large-scale multi-center validation to test transferability across institutions and specialties. Finally, we aim to align future development with evolving OHDSI community standards and implementation guidance.

5 Conclusion

This study presents DIMAS-OMOP, a novel deliberative intelligence-based multi-agent system designed to address the linguistic and semantic complexities of standardizing Chinese medical texts toward the OMOP framework. By integrating traditional NLP stability with the reasoning depth of LLM/RAG, DIMAS-OMOP introduces a proposer-skeptic deliberation mechanism that moves beyond static extraction toward an auditable, adversarial reasoning process. Our findings demonstrate that this hybrid architecture significantly enhances concept-mapping accuracy and operational efficiency while maintaining high controllability and interpretability through a four-dimensional quality control framework. To our knowledge, this work represents a pioneering systematic application of multi-agent deliberative intelligence in the field of medical informatics.

Beyond technical performance, DIMAS-OMOP provides a scalable bridge between unstructured non-English clinical narratives and globally interoperable observational data models. The pilot deployment's substantial ROI and workflow improvements underscore the practical viability of deploying such systems within institutional research infrastructures. While these preliminary results are promising, future research will focus on multilingual adaptation, lightweight deployment strategies, and large-scale external validation across diverse clinical domains. Ultimately, this research offers a robust methodological reference for the reliable integration of localized clinical data into the global OHDSI ecosystem, facilitating broader multi-center observational research.

6 Declarations

6.1 Acknowledgments

Some schematic elements in the figures were refined with the assistance from the GEMINI models; all figure concepts, annotations, and final edits were designed and verified by the authors.

6.2 Conflicts of Interest

The authors report no conflict of interest.

6.3 IRB Statement and Consent

This study was conducted in accordance with the Declaration of Helsinki. The use of clinical notes was approved by the IRB of participating institutions. All data were de-identified according to HIPAA standards prior to analysis.

6.4 Funding

This study was supported by Guangxi Key Research and Development Program (No. AB25069049).

6.5 Code Availability

The code is available at https://github.com/raywong2121/Multi-Agent.

6.6 Data Availability

Data may be available from the corresponding authors upon reasonable request and subject to approval by the relevant institutions.

6.7 Author Contribution

Hanlin Lv and Xiao Wang contributed equally to this work. Conceptualization: Lei Wang, Lei Li. Methodology and software: Hanlin Lv, Xiao Wang, Kesong Wu. Data curation: Hanlin Lv, Xiao Wang, Lei Wang. Formal analysis: Hanlin Lv, Xiao Wang. Supervision: Lei Wang, Lei Li. Writing-original draft: Hanlin Lv, Xiao Wang. Writing-review and editing: all authors. All authors read and approved the final manuscript.

References

Emily Alsentzer, John Murphy, William Boag, Wei-Hung Weng, Di Jindi, Tristan Naumann, and Matthew McDermott. 2019. Publicly available clinical bert embeddings. In *Proceedings of the 2nd clinical natural language processing workshop*, pages 72–78.

Alan R Aronson and François-Michel Lang. 2010. An overview of metamap: historical perspective and recent advances. *Journal of the American medical informatics association*, 17(3):229–236.

Adrien Bazoge, Emmanuel Morin, Béatrice Daille, and Pierre-Antoine Gourraud. 2023. Applying natural language processing to textual data from clinical data warehouses: systematic review. *JMIR medical informatics*, 11:e42477.

Olivier Bodenreider. 2004. The unified medical language system (umls): integrating biomedical terminology. *Nucleic acids research*, 32(suppl_1):D267–D270.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, and 1 others. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Wendy W Chapman, Will Bridewell, Paul Hanbury, Gregory F Cooper, and Bruce G Buchanan. 2001. A simple algorithm for identifying negated findings and diseases in discharge summaries. *Journal of biomedical informatics*, 34(5):301–310.

Shan Chen, Benjamin H Kann, Michael B Foote, Hugo JWL Aerts, Guergana K Savova, Raymond H Mak, and Danielle S Bitterman. 2023. Use of artificial intelligence chatbots for cancer treatment information. *JAMA oncology*, 9(10):1459–1462.

Xi Chen, Huahui Yi, Mingke You, WeiZhi Liu, Li Wang, Hairui Li, Xue Zhang, Yingman Guo, Lei Fan, Gang Chen, and 1 others. 2025. Enhancing diagnostic capability with multi-agents conversational large language models. *NPJ digital medicine*, 8(1):159.

Hannah Eyre, Alec B Chapman, Kelly S Peterson, Jianlin Shi, Patrick R Alba, Makoto M Jones, Tamara L Box, Scott L DuVall, and Olga V Patterson. 2022. Launching into clinical space with medspacy: a new clinical text processing toolkit in python. In *AMIA Annual Symposium Proceedings*, volume 2021, page 438.

Jakob Foerster, Gregory Farquhar, Triantafyllos Afouras, Nantas Nardelli, and Shimon Whiteson. 2018. Counterfactual multi-agent policy gradients. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yixin Dai, Jiawei Sun, Haofen Wang, Haofen Wang, and 1 others. 2023. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*.

Ary L Goldberger, Luis AN Amaral, Leon Glass, Jeffrey M Hausdorff, Plamen Ch Ivanov, Roger G Mark, Joseph E Mietus, George B Moody, Chung-Kang Peng, and H Eugene Stanley. 2000. Physiobank, physiotoolkit, and physionet: components of a new research resource for complex physiologic signals. *circulation*, 101(23):e215–e220.

Miguel A Hernán and James M Robins. 2016. Using big data to emulate a target trial when a randomized trial is not available. *American journal of epidemiology*, 183(8):758–764.

George Hripcsak, Jon D Duke, Nigam H Shah, Christian G Reich, Vojtech Huser, Martijn J Schuemie, Marc A Suchard, Rae Woong Park, Ian Chi Kei Wong, Peter R Rijnbeek, and 1 others. 2015. Observational health data sciences and informatics (ohdsi): opportunities for observational researchers. *Studies in health technology and informatics*, 216:574.

Nicholas R Jennings, Katia Sycara, and Michael Wooldridge. 1998. A roadmap of agent research and development. *Autonomous agents and multi-agent systems*, 1(1):7–38.

Alistair EW Johnson, Tom J Pollard, Lu Shen, Li-wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. 2016. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3(1):1–9.

Yuhe Ke, Rui Yang, Sui An Lie, Taylor Xin Yi Lim, Yilin Ning, Irene Li, Hairil Rizal Abdullah, Daniel

Shu Wei Ting, and Nan Liu. 2024. Mitigating cognitive biases in clinical decision-making through multi-agent conversations using large language models: simulation study. *Journal of Medical Internet Research*, 26:e59439.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Peter Lee, Sebastien Bubeck, and Joseph Petro. 2023. Benefits, limits, and risks of gpt-4 as an ai chatbot for medicine. *New England Journal of Medicine*, 388(13):1233–1239.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, and 1 others. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474.

Bertalan Meskó and Eric J Topol. 2023. The imperative for regulatory oversight of large language models (or generative ai) in healthcare. *NPJ digital medicine*, 6(1):120.

Harsha Nori, Nicholas King, Scott Mayer McKinney, Dean Carignan, and Eric Horvitz. 2023. Capabilities of gpt-4 on medical challenge problems. *arXiv preprint arXiv:2303.13375*.

OHDSI. 2021. Usagi documentation. Observational Health Data Sciences and Informatics. Available from: https://www.ohdsi.org/web/wiki/doku.php?id=documentation:software:usagi.

J Marc Overhage, Patrick B Ryan, Christian G Reich, Abraham G Hartzema, and Paul E Stang. 2012. Validation of a common data model for active safety surveillance research. *Journal of the American Medical Informatics Association*, 19(1):54–60.

Tabish Rashid, Mikayel Samvelyan, Christian Schroeder De Witt, Gregory Farquhar, Jakob Foerster, and Shimon Whiteson. 2020. Monotonic value function factorisation for deep multi-agent reinforcement learning. *Journal of Machine Learning Research*, 21(178):1–51.

Jenna M Reps, Martijn J Schuemie, Marc A Suchard, Patrick B Ryan, and Peter R Rijnbeek. 2018. Design and implementation of a standardized framework to generate and evaluate patient-level prediction models using observational healthcare data. *Journal of the American Medical Informatics Association*, 25(8):969–975.

Cynthia Rudin. 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature machine intelligence*, 1(5):206–215.

Patrick B Ryan, Martijn J Schuemie, Susan Gruber, Ivan Zorych, and David Madigan. 2013. Empirical performance of a new user cohort method: lessons for developing a risk identification and analysis system. *Drug safety*, 36(Suppl 1):59–72.

Guergana K Savova, James J Masanz, Philip V Ogren, Jiaping Zheng, Sunghwan Sohn, Karin C Kipper-Schuler, and Christopher G Chute. 2010. Mayo clinical text analysis and knowledge extraction system (ctakes): architecture, component evaluation and applications. *Journal of the American Medical Informatics Association*, 17(5):507–513.

Martijn J Schuemie, Patrick B Ryan, Nicole Pratt, Rui-Jun Chen, Seng Chan You, Harlan M Krumholz, David Madigan, George Hripcsak, and Marc A Suchard. 2020. Principles of large-scale evidence generation and evaluation across a network of databases (legend). *Journal of the American Medical Informatics Association*, 27(8):1331–1337.

Savyasachi V Shah. 2024. Accuracy, consistency, and hallucination of large language models when analyzing unstructured clinical notes in electronic medical records. *JAMA Network Open*, 7(8):e2425953.

K. Singhal, S. Azizi, T. Tu, S. S. Mahdavi, J. Wei, H. W. Chung, and 1 others. 2023. Large language models encode clinical knowledge. *Nature*, 620(7972):172–180.

Ergin Soysal, Jingqi Wang, Min Jiang, Yonghui Wu, Serguei Pakhomov, Hongfang Liu, and Hua Xu. 2018. Clamp–a toolkit for efficiently building customized clinical natural language processing pipelines. *Journal of the American Medical Informatics Association*, 25(3):331–336.

Peter Stone and Manuela Veloso. 2000. Multiagent systems: A survey from a machine learning perspective. *Autonomous Robots*, 8(3):345–383.

Marc A Suchard, Martijn J Schuemie, Harlan M Krumholz, Seng Chan You, RuiJun Chen, Nicole Pratt, Christian G Reich, Jon Duke, David Madigan, George Hripcsak, and 1 others. 2019. Comprehensive comparative effectiveness and safety of first-line antihypertensive drug classes: a systematic, multinational, large-scale analysis. *The Lancet*, 394(10211):1816–1826.

Ardi Tampuu, Tambet Matiisen, Dorian Kodelja, Ilya Kuzovkin, Kristjan Korjus, Juhan Aru, Jaan Aru, and Raul Vicente. 2017. Multiagent cooperation and competition with deep reinforcement learning. *PloS one*, 12(4):e0172395.

Arun James Thirunavukarasu, Darren Shu Jeng Ting, Kabilan Elangovan, Laura Gutierrez, Ting Fang Tan, and Daniel Shu Wei Ting. 2023. Large language models in medicine. *Nature medicine*, 29(8):1930–1940.

Özlem Uzuner, Brett R South, Shuying Shen, and Scott L DuVall. 2011. 2010 i2b2/va challenge on

concepts, assertions, and relations in clinical text. *Journal of the American Medical Informatics Association*, 18(5):552–556.

Erica A Voss, Rupa Makadia, Amy Matcho, Qianli Ma, Chris Knoll, Martijn Schuemie, Frank J DeFalco, Ajit Londhe, Vivienne Zhu, and Patrick B Ryan. 2015. Feasibility and utility of applications of the common data model to multiple, disparate observational health databases. *Journal of the American Medical Informatics Association*, 22(3):553–564.

Fei Wang, Rainu Kaushal, and Dhruv Khullar. 2020. Should health care demand interpretable artificial intelligence or accept "black box" medicine?

Yanshan Wang, Liwei Wang, Majid Rastegar-Mojarad, Sungrim Moon, Feichen Shen, Naveed Afzal, Sijia Liu, Yuqun Zeng, Saeed Mehrabi, Sunghwan Sohn, and 1 others. 2018. Clinical information extraction applications: a literature review. *Journal of biomedical informatics*, 77:34–49.

Andrew Wen, Liwei Wang, Huan He, Sunyang Fu, Sijia Liu, David A Hanauer, Daniel R Harris, Ramakanth Kavuluru, Rui Zhang, Karthik Natarajan, and 1 others. 2024. A case demonstration of the open health natural language processing toolkit from the national covid-19 cohort collaborative and the researching covid to enhance recovery programs for a natural language processing system for covid-19 or postacute sequelae of sars cov-2 infection: algorithm development and validation. *JMIR medical informatics*, 12(1):e49997.

Michael Wooldridge. 2009. *An introduction to multiagent systems*. John wiley & sons.

Ningyu Zhang, Qianghuai Jia, Kangping Yin, Liang Dong, Feng Gao, and Nengwei Hua. 2020. Conceptualized representation learning for chinese biomedical text mining. *arXiv preprint arXiv:2008.10813*.

Yuchen Zheng, Zhenggong Han, Yimin Cai, Xubo Duan, Jiangling Sun, Wei Yang, and Haisong Huang. 2022. An imconvnet-based deep learning model for chinese medical named entity recognition. *BMC Medical Informatics and Decision Making*, 22(1):303.