Neuro-Symbolic Resolution of Recommendation Conflicts in Multimorbidity Clinical Guidelines

Shiyao Xie[1,2], **Jian Du**[1,2*]

[1]Peking University, National Institute of Health Data Science
Beijing, CN
[2]Peking University Health Science Center, Institute of Medical Technology
Beijing, CN

Abstract

Clinical guidelines, typically developed by independent specialty societies, inherently exhibit substantial fragmentation, redundancy, and logical contradiction. These inconsistencies, particularly when applied to patients with multimorbidity, not only cause cognitive dissonance for clinicians but also introduce catastrophic noise into AI systems, rendering the standard Retrieval-Augmented Generation (RAG) system fragile and prone to hallucination. To address this fundamental reliability crisis, we introduce a Neuro-Symbolic framework that automates the detection of recommendation redundancies and conflicts. Our pipeline employs a multi-agent system to translate unstructured clinical natural language into rigorous symbolic logic language, which is then verified by a Satisfiability (SAT) solver. By formulating a hierarchical taxonomy of logical rule interactions, we identify a critical category termed Local Conflict—a decision conflict arising from the intersection of comorbidities. Evaluating our system on a curated benchmark of 12 authoritative SGLT2 inhibitor guidelines, we reveal that 90.6% of conflicts are Local, a structural complexity that single-disease guidelines fail to address. While state-of-the-art LLMs fail in detecting these conflicts, our neuro-symbolic approach achieves an F1 score of 0.861. This work demonstrates that logical verification must precede retrieval, establishing a new technical standard for automated knowledge coordination in medical AI.

Code and datasets are available at —
https://github.com/Shiyaoa/GuidelineCoordination

Introduction

Retrieval-Augmented Generation (RAG) has emerged as the prevailing paradigm to mitigate the intrinsic limitations of Large Language Models (LLMs), such as hallucinations and knowledge obsolescence (Vladika, Dhaini, and Matthes 2025; Wu et al. 2025), thereby enabling their widespread deployment in clinical decision-making (Liu, McCoy, and Wright 2025; Amugongo et al. 2025; Yang et al. 2025).By grounding model outputs in authoritative external sources like clinical guidelines, this approach aims to ensure safety and factual correctness (Zakka et al. 2024; Ong et al. 2024; Kresevic et al. 2024).However, the efficacy of this paradigm is fundamentally limited by the logical integrity of the underlying knowledge base. In reality, contemporary guidelines exhibit substantial redundancy, inconsistency, and conflict across organizations, disease areas, and updates, particularly for patients with multimorbidity (Blozik et al. 2013; Hoffmann, Jansen, and Glasziou 2018; Yaacoub et al. 2023; Tseng et al. 2025). When such conflicting documents are injected into the RAG context, even state-of-the-art LLMs can be misled, degrading performance and potentially amplifying clinical risk (Javadi et al. 2025). While some training-free (Jin et al. 2024) or training-based (Li et al. 2024) methods attempt to mitigate noise, in the medical domain, conflicts among guidelines are not merely noise, they are the cause of logical inconsistency in AI systems and potential clinical errors. Consequently, merely patching the retrieval component is insufficient; there is an urgent need for Knowledge Governance targeting the guidelines themselves.

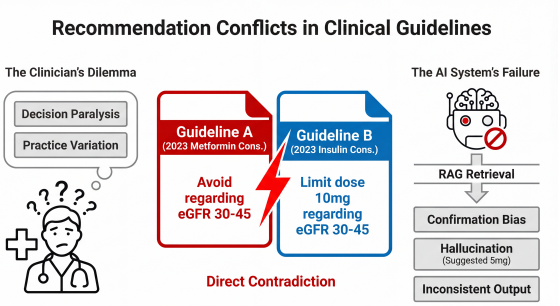


Figure 1: Decision-making crisis caused by clinical guideline conflicts. Illustrated using two guidelines published in 2023 on Empagliflozin use in patients with eGFR 30-45, demonstrating direct contradictions.

Conflicts among clinical guidelines primarily stem from organizational fragmentation (Tseng et al. 2025), variances in evidence interpretation (Nagavci, Gáspár, and Lakatos 2025), and a lack of standardized terminology (Ghai et al. 2021). Specifically, these inconsistencies manifest as direct contradictions (Raber et al. 2019), temporal discrepancies, and most critically, multimorbidity conflicts arising from adverse interactions when single-disease guidelines are applied to multi-disease patients (Dumbreck et al. 2015). Anal-

*Corresponding author.

ysis of primary cardiovascular disease prevention guidelines found that out of 124 recommendation clusters, only 44 clusters (35%) included consistent recommendations, and merely 4 clusters (3%) included highly consistent recommendations (Bredehorst et al. 2025). Such conflicts significantly increase polypharmacy risks, reduce patient adherence, and lead to medical resource wastage. Although the medical community has initiated harmonization efforts like Canada's C-CHANGE (Jain et al. 2022), existing approaches rely heavily on manual expert consensus. Given the exponential growth of medical literature, manual governance is no longer sustainable, creating an imperative for automated technical means to assist in this complex reasoning task.

From a computational perspective, guideline harmonization focuses on identifying and resolving redundancies and conflicts, which can be modeled as a Boolean Satisfiability (SAT) problem and solved using the Z3 theorem prover (de Moura and Bjørner 2008). To address the probabilistic nature of LLMs in rigorous logical reasoning, Neuro-Symbolic AI has emerged as a promising solution. By leveraging LLMs as semantic parsers alongside external logical solvers, this paradigm has achieved remarkable progress in mathematical proving and formal logic tasks (Pan et al. 2023). Despite this promise, neuro-symbolic methods have not yet been systematically explored for clinical guideline coordination. This is mainly due to two challenges: first, the extreme complexity of clinical natural language—replete with nested conditional clauses and ambiguity—makes high-fidelity Natural Language to Symbolic Language (NL2SL) translation difficult; second, the challenge of accurately modeling complex logical exclusions, particularly those involving multimorbidity.

To bridge these gaps, this paper presents an automated framework for clinical guideline conflict detection and harmonization. First, we designed a computable data model (schema) for clinical guidelines, formalizing unstructured recommendations into binary condition, action tuples. Second, drawing on Multi-Agent collaboration architectures, we developed a system to automate NL2SL translation and constructed a pipeline integrating a SAT solver to systematically verify guideline consistency. Finally, based on this pipeline, we evaluated the logical reasoning robustness of state-of-the-art LLMs under realistic RAG noise conditions. The main contributions of this paper are:

- We design a guideline-specific NL2SL schema that captures the nested conditional structure and action semantics of clinical recommendations.
- We develop an automated neuro-symbolic pipeline that combines multi-agent LLM parsing with SAT-based verification to detect and categorize redundancies and conflicts, including multimorbidity-specific local conflicts.
- We construct and open-source the first deep logical reasoning dataset dedicated to Guideline Harmonization, providing a new benchmark for knowledge governance methods for medical RAG systems.

Related Work

Computable Clinical Guidelines

The transition from narrative guidelines to machine-executable formats has evolved from early rule-based systems to modern interoperable standards. Early formalisms like Arden Syntax pioneered knowledge representation but lacked support for shared data models (Soares et al. 2021). Currently, Clinical Quality Language (CQL) serves as the standard for clinical decision support (CDS) and electronic clinical quality measures (eCQMs) (Odigie et al. 2019; McClure et al. 2020). As an HL7 standard, CQL distinguishes itself through a data-model-agnostic architecture—compatible with FHIR, OMOP, and QDM—and a dual-representation system that offers both human-readable logic and a machine-executable Expression Logical Model (ELM) (Brandt et al. 2020). This formalization enables the precise definition of cohort criteria and care gaps, leading to its widespread adoption by CMS and HEDIS for national quality programs (Soares et al. 2021). However, despite CQL's expressive power, translating complex clinical prose into valid CQL libraries remains a labor-intensive bottleneck. It requires interdisciplinary experts to manually map ambiguous narrative exclusion criteria to strict logic and standardized terminologies (Sittig et al. 2023), highlighting the urgent need for automated, neuro-symbolic translation mechanisms.

Neuro-Symbolic Reasoning and Consistency Verification

Neuro-symbolic frameworks employ LLMs as semantic parsers to translate natural language into formal representations such as First-Order Logic or ASP for verification by deterministic solvers. Systems such as Logic-LM (Pan et al. 2023) and LINC (Olausson et al. 2023) demonstrate that offloading inference to external theorem provers significantly boosts performance on logical benchmarks. Frameworks like NL2FOL (Lalwani et al. 2024) and LELMA (Mensfelt, Stathis, and Trencsenyi 2024) leverage SMT solvers to explicitly detect logical fallacies and validity errors in LLM outputs.Recent research has critically extended this paradigm to normative conflict detection in high-stakes domains. In the legal field, (Yadamsuren, Platt, and Diaz 2025) demonstrated that while pure LLMs struggle to identify statutory inconsistencies in tax codes, anchoring formalization in symbolic logic (Prolog) ensures deterministic detection. (Mantravadi et al. 2025) introduced LegalWiz, a multi-agent framework designed to stress-test legal RAG pipelines. Their work explicitly validates that unresolved contradictions in retrieved evidence lead to hallucinations, necessitating rigorous benchmarks with structured conflict types. While these legal frameworks offer valuable methodological parallels, clinical guidelines present a distinct challenge: unlike statutory contradictions, conflicts in multimorbidity often manifest as Local Conflicts—subtle logical intersections of conditional exclusions rather than direct negations—requiring the specialized hierarchical modeling and SAT-based verification proposed in this study.

Methodology

Our methodology transforms unstructured clinical guidelines into verifiable symbolic representations through a three-phase pipeline: Context Atomization, Neuro-Symbolic Formalization, and Logic-Based Verification.

Corpus Processing

Prior to formalization, a *Recommendation Extractor* functions as a pre-processor to stratify guideline content into four semantic categories: *Risk Assessment*, *Pharmacological Intervention*, *Non-pharmacological Intervention*, and *Other Opinions*. We specifically isolate **Pharmacological Interventions** for downstream processing using rigid syntactic inclusion criteria. To be retained, a recommendation must: (1) explicitly define both a *target population* and a *recommended action*; and (2) contain specific directive verbs (e.g., recommend, consider) and deontic modals (e.g., should, may). Following this extraction, we employ Locality-Sensitive Hashing (LSH) to group these recommendations into semantic clusters based on textual affinity

Multi-Agent Formalization

Within each cluster, a chained multi-agent system transforms natural language into executable symbolic logic (Terms $\rightarrow$ Predicates $\rightarrow$ Rules). This process addresses the semantic gap through three specialized agents:

1. Entity Agent To ground ambiguous clinical text, this agent maps mentions to canonical codes in SNOMED CT, LOINC, and RxNorm via external retrieval tools. Crucially, this mapping adheres to **HL7 FHIR** standards, guaranteeing interoperability with real-world Hospital Information Systems (HIS).

2. Predicate Agent A fundamental Semantic Gap exists between static *Ontological Entities* (e.g., T2DM, insulin) and verifiable *Logical Predicates* (e.g., Has T2DM, On insulin). Direct mapping leads to semantic ambiguity and restricts numerical reasoning. This agent bridges the gap between static entities and verifiable logic using a computable mapping $F : E \times O \rightarrow P$. It applies typed operators (O) to entities (E) to generate **Atomic Predicates** (P) compiled as SMT constraints (see Table 4 in Appendix):

- **Existential Logic**: $O_{\text{exist}}(E) \rightarrow P_{\text{bool}}$.
- **Arithmetic Logic**: $O_{\text{arith}}(E) \rightarrow P_{\text{lra}}$.
- **Categorical Logic**: $O_{\text{cat}}(E) \rightarrow P_{\text{enum}}$.

3. Rule Agent The final agent interprets semantic intent to assemble atomic predicates into **Compound Logical Structures** and maps directives to a standardized *action vocabulary* (Appendix). This produces a set of formal rules ready for solver-based verification.

Rule Relationship Reasoning

To ensure the clinical safety and consistency of the generated rules, we formulate the reasoning task as a **Satisfiability (SAT) problem**. By encoding the rules into the Z3 Solver (mathematical formulation detailed in **Appendix**), we classify rule interactions into a hierarchical taxonomy comprising two coarse-grained categories and five fine-grained types:

Category 1: Redundancy This category identifies rules that provide identical advice for overlapping populations, enabling knowledge base deduplication.

- **Complete Redundancy**: Occurs when logical conditions are strictly equivalent ($C_1 \iff C_2$) and actions are identical.
- **Contained Redundancy**: Occurs when one rule's condition is strictly subsumed by another ($C_{\text{specific}} \subset C_{\text{general}}$), yet they dictate the same action. This implies that the specific rule is logically redundant within the scope of the general rule.

Category 2: Conflict & Disagreement This category captures logical incompatibilities where adherence to one rule violates another.

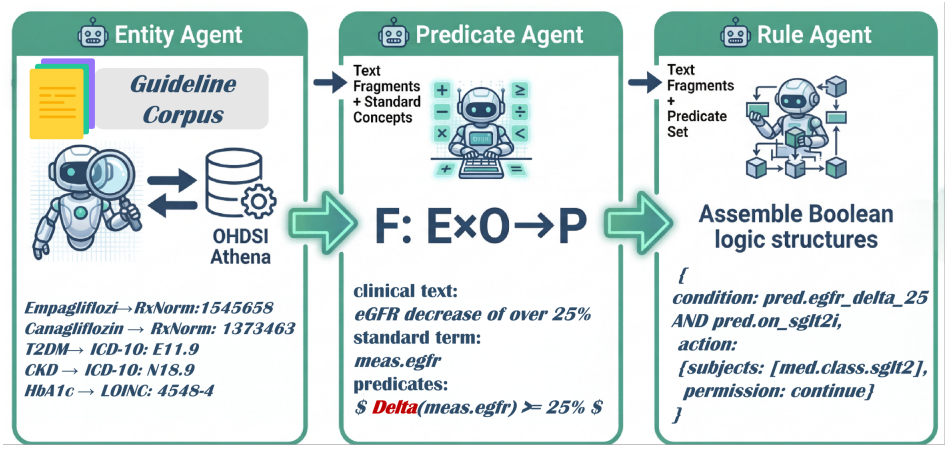
- **Intrinsic Conflict**: The most direct inconsistency, where rules addressing the **exact same population** ($C_1 \iff C_2$) dictate opposing actions (*Recommend* vs. *Contraindicate*).
- **Implication Conflict**: Arises from a **General vs. Specific** tension ($C_{\text{specific}} \subset C_{\text{general}}$), where a subgroup guideline contradicts the general recommendation. Detecting this prevents the accidental application of broad guidelines to high-risk subgroups.
- **Local Conflict**: The critical source of multimorbidity issues. It occurs when rule conditions merely **intersect** ($C_1 \cap C_2 \neq \emptyset$) rather than encompass one another. When a patient simultaneously satisfies the conditions of multiple such rules, these rules yield conflicting decision recommendations. The conflict is local because it manifests exclusively in the patient subset suffering from both conditions simultaneously.

Experiments

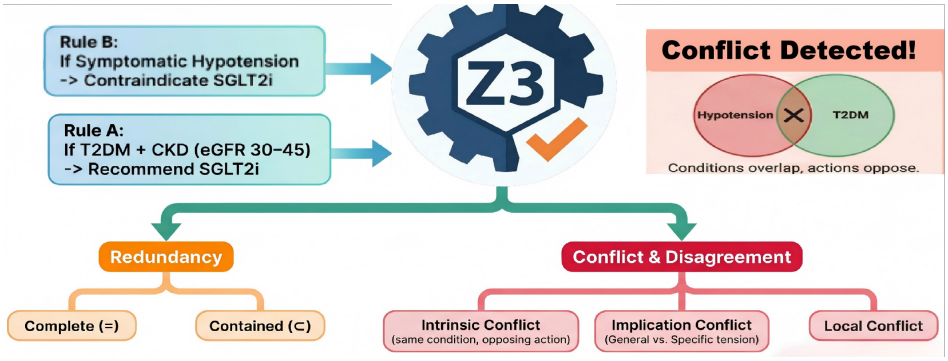
Our experimental evaluation is designed to address two pivotal research questions: RQ1: How accurately can our Multi-Agent system translate ambiguous, unstructured clinical natural language into the strict symbolic representations required by the solver? RQ2: Is the proposed Neuro-Symbolic pipeline truly necessary? specifically, can standalone State-of-the-Art (SOTA) LLMs operating in a standard RAG setting achieve comparable reasoning performance without symbolic formalization?

Experimental Setup

Dataset Construction. We curated a benchmark corpus comprising **12 authoritative clinical guidelines** related to SGLT2 inhibitor (SGLT2i) therapy, sourced from diverse specialty societies within the Chinese Medical Association. We selected this domain as a representative testbed because the independent development of these guidelines naturally engenders a **full type of logical inconsistencies**, ranging from simple redundancy to complex intrinsic and implication conflicts. This heterogeneity provides a rigorous environment to evaluate the system's capability in logic formalization and relationship detection. Comprehensive details regarding the corpus composition are provided in **Appendix** . Our pipeline utilizes DeepSeek-V3.1 (DeepSeek-



(a) Multi-Agent Formalization



(b) Rule Relationship Analysis

Figure 2: The Neuro-Symbolic Pipeline for Clinical Guideline Formalization and Verification

AI 2024) as the core semantic reasoning engine for all agentic tasks, while the Z3 SMT Solver serves as the symbolic verification backend. To answer RQ2, we benchmark our neuro-symbolic approach against top-tier proprietary and open-source LLMs: GPT-5 (OpenAI 2025), Gemini-2.5-Pro (Comanici et al. 2025), Qwen-MAX (Team 2025), and DeepSeek-V3.1 (standalone). We evaluate their ability to detect redundancy and conflict relationships under varying noise levels, simulating real-world retrieval scenarios.

Dataset Construction

Gold Standard Dataset To establish a rigorous benchmark, we constructed a high-quality dataset derived from the pipeline's output, subject to comprehensive human verification. Clinical experts first reviewed the candidate rule pairs to correct any misclassifications in the extraction or relation phases. To ensure dataset balance, we retained all instances of rare, critical relationships while down-sampling common ones. Additionally, to evaluate the models' capability to discern non-relationships, we incorporated negative samples consisting of rule pairs explicitly verified to have no logical interaction. **RAG Noise Dataset** To evaluate model robustness under realistic retrieval setting, we constructed a Noisy Test Set using a Graph-Based Strategy. We first modeled the entire rule base as a relationship graph, where nodes represent rules and edges represent verified logical relations. Within this graph, we identified Isolated Rules—nodes with a degree of zero, indicating no known logical interaction with any other rule in the corpus. We then generated test samples by injecting k ($k \in [1, 8]$) of these isolated rules into each verified base pair, followed by random shuffling to eliminate positional bias. This methodology guarantees that the injected noise functions purely as distractors without altering the fundamental logical relationship of the target pair.

Results

Evaluation of Formalization Accuracy

We first evaluated the foundational component of our pipeline—the NL-to-Symbolic translation—through rigorous manual audit. We posit that the validity of the downstream logical verification hinges entirely on the semantic fidelity of the LLM agents' formalization, as the Z3 solver's output is formally correct relative to its input. An expert author with dual expertise in clinical medicine and data science manually audited 565 Rule objects generated by the pipeline. A rule was deemed Correct only under a Strict Exact-Match criterion: the formalized Predicate IDs and action (Permission, Subject) must be semantically identical to

Table 1: Statistics of the Logical Benchmark Dataset

Label Category	#Pairs	Ratio
Conflict	97	0.429
local_conflict	20	0.088
implication_conflict_or_disagreement	40	0.177
intrinsic_conflict_or_disagreement	37	0.164
Redundancy	69	0.305
complete_redundancy	15	0.066
contained_redundancy	54	0.239
None	60	0.265
Total	**226**	**1**

the original text provenance. Partial matches were classified as failures. The system achieved a Formalization Accuracy of 80.1%. While this demonstrates strong zero-shot reasoning capabilities, 19.9% of cases required correction. A granular breakdown of the 111 error cases reveals four primary failure modes:

- **Contextual Loss** (56.7%, 63 cases): The most prevalent error involved missing implicit diagnostic context. For example, a recommendation extracted from a subsection titled Treatment for T2DM might fail to explicitly include `pred.has_t2dm` in its predicate list, rendering the rule overly broad.
- **Entity Grounding Failure** (19.6%, 23 cases): This includes hallucinations or misclassifications of drug entities, such as confusing specific gliflozins or failing to map brand names to standard RxNorm terminologies.
- **Predicate Distortion** (12.6%, 14 cases): Logical errors such as hallucinating non-existent conditions, omitting critical exclusions, or inverting numerical operators.
- **Action Alignment Error** (8.1%, 9 cases): Discrepancies in the strength or direction of the recommendation, such as mapping should be considered to Recommend instead of Consider, or conflating Avoid with Contraindicate.

Consistency Analysis of SGLT2i-Usage Guidelines

Following the manual correction of formalization errors, we deployed the Z3 SMT solver on the curated dataset to map the global logical landscape of SGLT2 inhibitor guidelines. This analysis revealed a stark contrast between general consensus and specific conflict.

The distribution of interaction types in Table 2 reveals that contradiction is rarely intrinsic, but frequently contextual. The overwhelming prevalence of Local Conflicts 2,442 pairs confirms that while guidelines are consistent for single diseases, they are structurally fragile for multimorbidity. The CKD vs. Hypotension example illustrates this perfectly: it is not a factual error, but a deadlock between competing clinical objectives long-term renal protection vs. immediate hemodynamic safety. This proves that RAG systems cannot simply retrieve documents; they require a Pathology Hierarchy to adjudicate such trade-offs.

The presence of 115 Implication Conflicts highlights the risk of semantic ambiguity, validating our need for a controlled action vocabulary. Conversely, true Intrinsic Conflicts are rare (37 pairs), suggesting that direct disputes over evidence are the exception rather than the norm.

Baseline Comparison

To answer RQ2 (Comparative Necessity), we benchmarked our fully automated pipeline against three state-of-the-art LLMs (DeepSeek-v3.1, Gemini-2.5-Pro, and GPT-5) operating in a standard RAG setting. It is important to note that the Ours metric reported here reflects the end-to-end performance of the automated agents without human correction. Despite the 20% extraction error rate above noted, our symbolic reasoning engine significantly outperforms pure neural baselines.

The results reveal a fundamental dichotomy in LLM capabilities: Baselines achieve moderate performance on Redundancy tasks (F1: 0.50–0.62). This is expected, as redundancy often manifests as semantic similarity, which matches the pre-training objective of embedding models. However, ours still leads (F1: 0.729) because the Z3 solver eliminates pseudo-redundancies where text is similar but logical conditions differ slightly. All baseline models failed catastrophically on Conflict detection, with Recall scores collapsing to 0.08–0.18. Even GPT-5 achieved an F1 of only 0.298. This indicates that LLMs struggle to distinguish topically related but contradictory information from topically related and consistent information. In contrast, our Neuro-Symbolic approach achieved an F1 of 0.861, demonstrating that mapping text to Boolean logic is the only reliable path for conflict detection.

Figure 3 decomposes the F1-scores across the five fine-grained sub-types. Baselines perform worst on Local Conflict and Implication Conflict. These categories require understanding intersection and subset logic. Pure LLMs tend to classify these pairs as Related or Neutral because they lack the symbolic machinery to verify if the conditions overlap. Our method maintains a consistent hexagonal shape on the radar chart, indicating balanced performance across all logical types. The solver successfully bridges the gap, specifically excelling in Complete Redundancy and Intrinsic Conflict where logic is binary and absolute.

Figure 4 illustrates the degradation of F1-scores as the number of noise rules increases from 0 to 8. All baseline models exhibit a sharp downward trend. As noise increases, the Signal-to-Noise Ratio in the context window drops, causing models to hallucinate relationships between unrelated rules or miss true conflicts due to attention dilution. These results confirm that pure RAG is insufficient for Guideline Harmonization. While LLMs are adequate for semantic retrieval, the Logic-Based Governance provided by our Neuro-Symbolic pipeline is indispensable for ensuring the safety and consistency of medical AI systems.

Conclusion

In summary, we highlighted a foundational yet often overlooked assumption in medical RAG: that clinical guidelines are logically consistent ground truth. We introduced

Table 2: Representative Examples and Prevalence of Logical Interactions in SGLT2i Usage Guidelines

Relation Type	**Count Pairs**	**Rule A**	**Rule B**	**Clinical explanations**
Local Conflict	2442	If T2DM + CKD (eGFR 30–45) → Recommend SGLT2i.	If Symptomatic Hypotension → Contraindicate SGLT2i.	A patient with both conditions faces a clash between organ protection and immediate hemodynamic safety. The system flags this intersection, prompting a clinical priority setting.
Implication Conflict	115	If eGFR $<$ 30 → Contraindicate Metformin.	If eGFR $<$ 45 → Avoid Metformin.	When a more specific condition triggers a stronger action than its broader, inclusive condition, creating contradictory guidance for the overlapping population.
Intrinsic Conflict	37	If eGFR $<$ 15 or Dialysis → Continue SGLT2i.	If eGFR $<$ 15 or Dialysis → Stop SGLT2i.	Evidence Contradiction: Direct contradiction on the exact same population. Rare but critical, likely reflecting differences in guideline versions or evidence interpretation.
Contained Redundancy	57	If T2DM + High CV Risk → Prioritize SGLT2i.	If T2DM + CKD (on ACEi/ARB) → Combine with SGLT2i.	Rule B is a subset of Rule A. The system validates that specific scenarios align with broader therapeutic principles.
Complete Redundancy	16	If Severe Liver Impairment → Avoid SGLT2i.	If Severe Liver Impairment → Avoid SGLT2i.	Both rules enforce the exact same constraint. Identifying this allows for safe deduplication of the knowledge base.

Table 3: Performance Comparison on Logical Relation Detection

Relation	**Model**	**Prec.**	**Rec.**	**F1**
Redundancy	DeepSeek-v3.1	0.515	0.768	0.616
	Qwen-Max	0.526	0.594	0.558
	Gemini-2.5-pro	0.543	0.638	0.587
	GPT-5	0.507	0.493	0.5
	Ours	**0.689**	**0.775**	**0.729**
Conflict	DeepSeek-v3.1	0.727	0.082	0.148
	Qwen-Max	0.75	0.093	0.165
	Gemini-2.5-pro	0.5	0.144	0.224
	GPT-5	0.75	0.186	0.298
	Ours	**0.826**	**0.898**	**0.861**

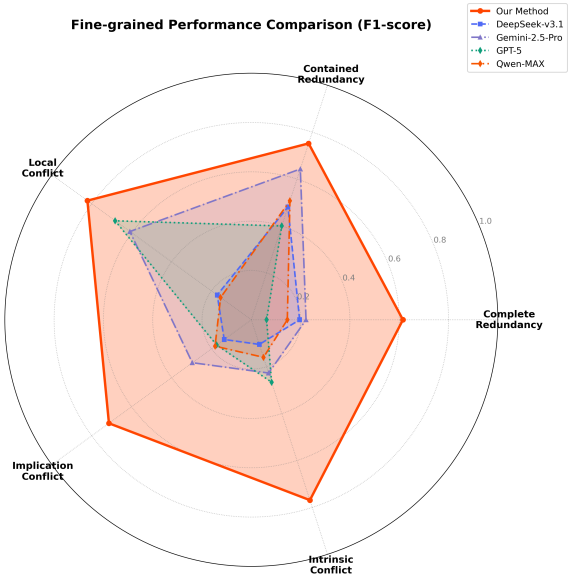


Figure 3: Performance Comparison across Logical Subcategories.

a novel Neuro-Symbolic framework that synergizes the semantic flexibility of LLMs with the rigorous deducibility of SAT solvers to perform automated Knowledge Governance on clinical guidelines. Our contributions are threefold. First, our methodology successfully formalizes ambiguous natural language into verifiable symbolic logic, achieving an 80% accuracy in zero-shot translation. Second, our large-scale consistency analysis of SGLT2i usage guidelines uncovered a critical blind spot in modern medicine: 90.6% of conflicts are Local, arising solely from multimorbidity intersections. This finding empirically demonstrates that single-disease guidelines are structurally inadequate for multi-disease patients, a complexity that standard vector-based retrieval cannot navigate. Third, our baseline comparisons reveal that even state-of-the-art LLMs suffer from a Reasoning Gap, failing catastrophically in conflict detection (Recall ¡ 18%) where our neuro-symbolic approach remains robust. We conclude that logical verification must become a prerequisite for Medical RAG deployment. Merely retrieving information is insufficient; AI systems must possess the symbolic

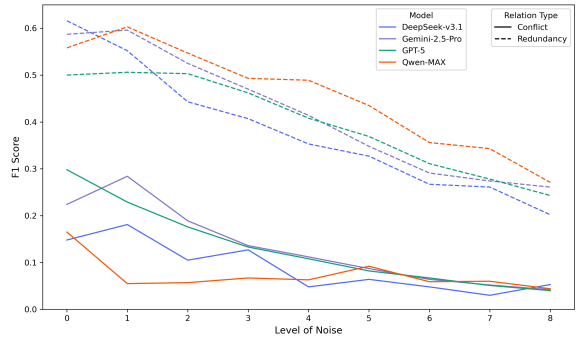


Figure 4: Impact of RAG Retrieval Noise on Reasoning Performance.

machinery to adjudicate the conflicting constraints of real-world pathology. Future work will extend this framework from conflict detection to automated harmonization, exploring how solver-generated proofs can guide LLMs to rewrite inconsistent guidelines into coherent, patient-centered decision pathways.

References

Amugongo, L. M.; Mascheroni, P.; Brooks, S.; Doering, S.; and Seidel, J. 2025. Retrieval augmented generation for large language models in healthcare: A systematic review. *PLOS Digital Health*, 4(6): e0000877.

Blozik, E.; van den Bussche, H.; Gurtner, F.; Schäfer, I.; and Scherer, M. 2013. Epidemiological strategies for adapting clinical practice guidelines to the needs of multimorbid patients. *BMC Health Services Research*, 13: 352 – 352.

Brandt, P. S.; Kiefer, R. C.; Pacheco, J. A.; Adekkanattu, P.; Sholle, E. T.; Ahmad, F. S.; Xu, J.; Xu, Z.; Ancker, J. S.; Wang, F.; Luo, Y.; Jiang, G.; Pathak, J.; and Rasmussen, L. V. 2020. Toward cross-platform electronic health record-driven phenotyping using Clinical Quality Language. *Learning Health Systems*, 4.

Bredehorst, M.; González-González, A. I.; Schürmann, L.; Firmansyah, D.; Muth, C.; Haasenritter, J.; van der Wardt, V.; and Puzhko, S. 2025. Recommendations for the primary prevention of atherosclerotic cardiovascular disease in primary care: a systematic guideline review. *Frontiers in Medicine*, 11.

Comanici, G.; Bieber, E.; Schaekermann, M.; Pasupat, I.; Sachdeva, N.; Dhillon, I.; Blistein, M.; Ram, O.; Zhang, D.; Rosen, E.; et al. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.

de Moura, L. M.; and Bjørner, N. S. 2008. Z3: An Efficient SMT Solver. In *International Conference on Tools and Algorithms for Construction and Analysis of Systems*.

DeepSeek-AI. 2024. DeepSeek-V3 Technical Report. arXiv:2412.19437.

Dumbreck, S.; Flynn, A.; Nairn, M.; Wilson, M.; Treweek, S. P.; Mercer, S. W.; Alderson, P.; Thompson, A.; Payne, K.; and Guthrie, B. 2015. Drug-disease and drug-drug interactions: systematic examination of recommendations in 12 UK national clinical guidelines. *The BMJ*, 350.

Ghai, V.; Subramanian, V.; Jan, H.; Loganathan, J.; and Doumouchtsis, S. K. 2021. Evaluation of clinical practice guidelines (CPG) on the management of female chronic pelvic pain (CPP) using the AGREE II instrument. *International Urogynecology Journal*, 32: 2899 – 2912.

Hoffmann, T. C.; Jansen, J.; and Glasziou, P. P. 2018. The importance and challenges of shared decision making in older people with multimorbidity. *PLoS Medicine*, 15.

Jain, R.; Stone, J. A.; Agarwal, G.; Andrade, J. G.; Bacon, S. L.; Bajaj, H. S.; Baker, B.; Cheng, G.; Dannenbaum, D.; Gelfer, M.; Habert, J.; Hickey, J.; Keshavjee, K.; Kitty, D.; Lindsay, P.; L'Abbé, M. R.; Lau, D. C. W.; Macle, L.; McDonald, M.; Nerenberg, K. A.; Pearson, G. J.; Pham, T.-D. T.; Poppe, A. Y.; Rabi, D. M.; Sherifali, D.; Selby, P.; Smith, E. E.; Stern, S.; Thanassoulis, G.; Terenzi, K. A.; Tu, K.; Udell, J. A.; Virani, S. A.; Ward, R.; Warburton, D. E. R.; Wharton, S.; Zymantas, J.; Hua-Stewart, D.; Liu, P.; and Tobe, S. W. 2022. Canadian Cardiovascular Harmonized National Guideline Endeavour (C-CHANGE) guideline for the prevention and management of cardiovascular disease in primary care: 2022 update. *CMAJ : Canadian Medical Association Journal*, 194: E1460 – E1480.

Javadi, S.; Mirabi, S.; Gangar, M.; and Ofoghi, B. 2025. When Evidence Contradicts: Toward Safer Retrieval-Augmented Generation in Healthcare. *arXiv preprint arXiv:2511.06668*.

Jin, B.; Yoon, J.; Han, J.; and Arik, S. Ö. 2024. Long-Context LLMs Meet RAG: Overcoming Challenges for Long Inputs in RAG. *ArXiv*, abs/2410.05983.

Kresevic, S.; Giuffrè, M.; Ajcevic, M.; Accardo, A.; Crocè, L. S.; and Shung, D. L. 2024. Optimization of hepatological clinical guidelines interpretation by large language models: a retrieval augmented generation-based framework. *NPJ digital medicine*, 7(1): 102.

Lalwani, A.; Kim, T.; Chopra, L.; Hahn, C.; Jin, Z.; and Sachan, M. 2024. Autoformalizing Natural Language to First-Order Logic: A Case Study in Logical Fallacy Detection.

Li, X.; Mei, S.; Liu, Z.; Yan, Y.; Wang, S.; Yu, S.; Zeng, Z.; Chen, H.; Yu, G.; Liu, Z.; Sun, M.; and Xiong, C. 2024. RAG-DDR: Optimizing Retrieval-Augmented Generation Using Differentiable Data Rewards. *ArXiv*, abs/2410.13509.

Liu, S.; McCoy, A. B.; and Wright, A. 2025. Improving large language model applications in biomedicine with retrieval-augmented generation: a systematic review, meta-analysis, and clinical development guidelines. *Journal of the American Medical Informatics Association*, 32(4): 605–615.

Mantravadi, A.; Dalmia, S.; Pospelova, O.; Mukherji, A.; Dave, N.; and Mittal, A. 2025. LegalWiz: A Multi-Agent Generation Framework for Contradiction Detection in Legal Documents. *ArXiv*, abs/2510.03418.

McClure, R. C.; Macumber, C. L.; Skapik, J. L.; and Smith, A. M. 2020. Igniting Harmonized Digital Clinical Quality

Measurement through Terminology, CQL, and FHIR. *Applied Clinical Informatics*, 11: 23 – 33.

Mensfelt, A.; Stathis, K.; and Trencsenyi, V. 2024. Towards Logically Sound Natural Language Reasoning with Logic-Enhanced Language Model Agents.

Nagavci, B.; Gáspár, Z.; and Lakatos, B. 2025. Defining expert opinion in clinical guidelines: insights from 98 scientific societies – a methodological study. *BMC Medical Research Methodology*, 25.

Odigie, E.; Lacson, R. C.; Raja, A. S.; Osterbur, D.; Ip, I. K.; Schneider, L. I.; and Khorasani, R. 2019. Fast Healthcare Interoperability Resources, Clinical Quality Language, and Systematized Nomenclature of Medicine—Clinical Terms in Representing Clinical Evidence Logic Statements for the Use of Imaging Procedures: Descriptive Study. *JMIR Medical Informatics*, 7.

Olausson, T. X.; Gu, A.; Lipkin, B.; Zhang, C. E.; Solar-Lezama, A.; Tenenbaum, J.; and Levy, R. 2023. LINC: A Neurosymbolic Approach for Logical Reasoning by Combining Language Models with First-Order Logic Provers. In *Conference on Empirical Methods in Natural Language Processing*.

Ong, C. S.; Obey, N. T.; Zheng, Y.; Cohan, A.; and Schneider, E. B. 2024. SurgeryLLM: a retrieval-augmented generation large language model framework for surgical decision support and workflow enhancement. *npj Digital Medicine*, 7(1): 364.

OpenAI. 2025. GPT-5 System Card. cdn.openai.com. System card, accessed on 2025-12-06.

Pan, L.; Albalak, A.; Wang, X.; and Wang, W. Y. 2023. Logic-LM: Empowering Large Language Models with Symbolic Solvers for Faithful Logical Reasoning. *ArXiv*, abs/2305.12295.

Raber, I.; McCarthy, C. P.; Vaduganathan, M.; Bhatt, D. L.; and McEvoy, J. W. 2019. The rise and fall of aspirin in the primary prevention of cardiovascular disease. *The Lancet*, 393: 2155–2167.

Sittig, D. F.; Boxwala, A. A.; Wright, A.; Zott, C.; Desai, P. J.; Dhopeshwarkar, R. V.; Swiger, J.; Lomotan, E. A.; Dobes, A.; and Dullabh, P. 2023. A lifecycle framework illustrates eight stages necessary for realizing the benefits of patient-centered clinical decision support. *Journal of the American Medical Informatics Association : JAMIA*, 30: 1583 – 1589.

Soares, A.; Jenders, R. A.; Harrison, R.; and Schilling, L. M. 2021. A Comparison of Arden Syntax and Clinical Quality Language as Knowledge Representation Formalisms for Clinical Decision Support. *Applied Clinical Informatics*, 12: 495 – 506.

Team, Q. 2025. Qwen3-Max: Just Scale it.

Tseng, O. L.; Brar, S.; Dawes, M.; Ranchod, H.; Lacaille, D.; Su, V. C.; and Mitton, C. 2025. Are Canadian Clinical Practice Guidelines Accounting for Adults With Multiple Chronic Diseases? A Systematic Review. *Journal of Evaluation in Clinical Practice*, 31(4): e70143.

Vladika, J.; Dhaini, M.; and Matthes, F. 2025. Facts Fade Fast: Evaluating Memorization of Outdated Medical Knowledge in Large Language Models. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, 9161–9174.

Wu, W.; Xu, X.; Gao, C.; Diao, X.; Li, S.; Salas, L. A.; and Gui, J. 2025. Assessing and Mitigating Medical Knowledge Drift and Conflicts in Large Language Models. In Christodoulopoulos, C.; Chakraborty, T.; Rose, C.; and Peng, V., eds., *Findings of the Association for Computational Linguistics: EMNLP 2025*, 707–730. Suzhou, China: Association for Computational Linguistics. ISBN 979-8-89176-335-7.

Yaacoub, S.; Chamseddine, F.; Jaber, F.; Blazic, I.; Frija, G.; and Akl, E. A. 2023. Exploring the concordance of recommendations across guidelines on chest imaging for the diagnosis and management of COVID-19: A proposed methodological approach based on a case study. *PLoS One*, 18(7): e0288359.

Yadamsuren, B.; Platt, S. K.; and Diaz, M. 2025. LLM-Assisted Formalization Enables Deterministic Detection of Statutory Inconsistency in the Internal Revenue Code.

Yang, R.; Wong, M. Y. H.; Li, H.; Li, X.; Zhu, W.; Liao, J.; Yu, K.; Liew, J. C. K.; Xuan, W.; Chen, Y.; et al. 2025. Retrieval-Augmented Generation in Medicine: A Scoping Review of Technical Implementations, Clinical Applications, and Ethical Considerations. *arXiv preprint arXiv:2511.05901*.

Zakka, C.; Shad, R.; Chaurasia, A.; Dalal, A. R.; Kim, J. L.; Moor, M.; Fong, R.; Phillips, C.; Alexander, K.; Ashley, E.; et al. 2024. Almanac—retrieval-augmented language models for clinical medicine. *Nejm ai*, 1(2): AIoa2300068.

SAT Modeling of Rule Relationships Using Z3-Solver

Problem Formulation

We model the problem of determining relationships between clinical rules as a **Satisfiability (SAT)** problem using the Z3 SMT solver. This approach provides a unified framework for reasoning about logical relationships between predicates, actions, and rules.

Notation and Preliminaries

Let $\mathcal{R} = \{R_1, R_2, \ldots, R_n\}$ be a set of clinical rules, where each rule R_i is represented as:

$$R_i : \quad C_i \Rightarrow A_i$$

where:

- C_i: a condition formula (Boolean combination of predicates over patient attributes)
- A_i: an action constraint formula (permissions, contraindications, dose restrictions, etc. on drugs/strategies)

Let $\mathcal{P} = \{P_1, P_2, \ldots, P_m\}$ be a set of predicates, where each predicate P_j has a formal definition ϕ_j over a set of variables $\mathcal{V}$ representing the attributes of the patient like measurements, conditions, procedures.

Predicate Relation Modeling

Variable Encoding For each predicate P_i with formal definition ϕ_i, we create a Boolean variable p_i in Z3:

$$p_i = \text{Z3Expr}(\phi_i)$$

where $\text{Z3Expr}(\phi_i)$ converts the formal definition into a Z3 Boolean expression over variables in $\mathcal{V}$.

Predicate Relation Definitions via SAT Given two predicates P and Q with Z3 expressions p and q, we define their relationship using SAT queries:

Equivalence ($P \Leftrightarrow Q$)

$$\mathsf{Equiv}(P, Q) \Leftrightarrow \text{SAT}(p \neq q) = \text{UNSAT}$$

Implication ($P \Rightarrow Q$)

$$\mathsf{Imply}_{P \to Q}(P, Q) \Leftrightarrow \text{SAT}(p \wedge \neg q) = \text{UNSAT}$$

Reverse Implication ($Q \Rightarrow P$)

$$\mathsf{Imply}_{Q \to P}(P, Q) \Leftrightarrow \text{SAT}(q \wedge \neg p) = \text{UNSAT}$$

Mutual Exclusion ($P \wedge Q$ is unsatisfiable)

$$\mathsf{Mutex}(P, Q) \Leftrightarrow \text{SAT}(p \wedge q) = \text{UNSAT}$$

Intersection (Satisfiable but neither implies the other)

$$\begin{aligned}\mathsf{Intersect}(P, Q) \Leftrightarrow\ & \text{SAT}(p \wedge q) = \text{SAT} \\ & \wedge \neg \mathsf{Imply}_{P \to Q}(P, Q) \\ & \wedge \neg \mathsf{Imply}_{Q \to P}(P, Q)\end{aligned}$$

Action Relation Modeling

Action Representation An action A is represented as:

$$A = \{(s_1, \pi_1), (s_2, \pi_2), \ldots, (s_k, \pi_k)\}$$

where:

- $s_i \in \mathcal{S}$: a subject (drug/strategy) identifier
- $\pi_i \in \Pi$: a permission type like `allow`, `contraindicate`, `reduce_dose`

Action Variable Encoding For each action A and each subject $s \in \mathcal{S}$, we define Boolean variables:

$$\text{allow}_A(s) : \text{subject } s \text{ is allowed by action } A \quad (1)$$

$$\text{prohibit}_A(s) : \text{subject } s \text{ is prohibited by action } A \quad (2)$$

Constraints:

$$\forall s \in \mathcal{S} : \neg(\text{allow}_A(s) \wedge \text{prohibit}_A(s))$$

Action Relation Definitions via SAT Given two actions A and B:

Equivalence

$$\begin{aligned}\mathsf{Equiv}(A, B) \Leftrightarrow \forall s \in \mathcal{S} :\ & \\ & \text{allow}_A(s) = \text{allow}_B(s) \\ & \wedge\ \text{prohibit}_A(s) = \text{prohibit}_B(s)\end{aligned}$$

This is checked by comparing the sets of subjects and permission values directly (no SAT query needed).

Conflict

$$\begin{aligned}\mathsf{Conflict}(A, B) \Leftrightarrow \exists s \in \text{overlap}(A, B) :\ & \\ & \text{allow}_A(s) \wedge \text{prohibit}_B(s) \\ & \vee\ \text{prohibit}_A(s) \wedge \text{allow}_B(s)\end{aligned}$$

where $\text{overlap}(A, B) = \{s : s \in \text{subjects}(A) \cap \text{subjects}(B)\}$.

Disagreement

$$\begin{aligned}\mathsf{Disagree}(A, B) \Leftrightarrow \exists s \in \text{overlap}(A, B) :\ & \\ & \text{permission}_A(s) \neq \text{permission}_B(s) \\ & \wedge \neg \mathsf{Conflict}(A, B)\end{aligned}$$

This occurs when actions have overlapping subjects but different permission within the same category.

Independent

$$\mathsf{Indep}(A, B) \Leftrightarrow \text{overlap}(A, B) = \emptyset$$

Rule Relation Modeling

Rule Pair Analysis For a pair of rules $(R_a : C_a \Rightarrow A_a, R_b : C_b \Rightarrow A_b)$, we determine their relationship by combining predicate and action relations.

Rule Relation Classification via SAT

Complete Redundancy

$$\begin{aligned}\mathsf{CompRed}(R_a, R_b) \Leftrightarrow\ & \mathsf{Equiv}(C_a, C_b) \\ & \wedge\ \mathsf{Equiv}(A_a, A_b)\end{aligned}$$

Contained Redundancy

$$\begin{aligned}\mathsf{ContRed}(R_a, R_b) \Leftrightarrow\ & \mathsf{Imply}_{P \to Q}(C_a, C_b) \\ & \vee\ \mathsf{Imply}_{Q \to P}(C_a, C_b) \\ & \wedge\ \mathsf{Equiv}(A_a, A_b)\end{aligned}$$

Intrinsic Conflict

$$\begin{aligned}\mathsf{IntrConf}(R_a, R_b) \Leftrightarrow\ & \mathsf{Equiv}(C_a, C_b) \\ & \wedge\ \mathsf{Conflict}(A_a, A_b)\end{aligned}$$

Intrinsic Disagreement

$$\begin{aligned}\mathsf{IntrDis}(R_a, R_b) \Leftrightarrow\ & \mathsf{Equiv}(C_a, C_b) \\ & \wedge\ \mathsf{Disagree}(A_a, A_b)\end{aligned}$$

Implication Conflict

$$\begin{aligned}\mathsf{ImpConf}(R_a, R_b) \Leftrightarrow\ & \mathsf{Imply}_{P \to Q}(C_a, C_b) \\ & \vee\ \mathsf{Imply}_{Q \to P}(C_a, C_b) \\ & \wedge\ \mathsf{Conflict}(A_a, A_b) \\ & \wedge\ \neg \mathsf{SpecPrior}\end{aligned}$$

Implication Disagreement

$$\begin{aligned}\mathsf{ImpDis}(R_a, R_b) \Leftrightarrow\ & \mathsf{Imply}_{P \to Q}(C_a, C_b) \\ & \vee\ \mathsf{Imply}_{Q \to P}(C_a, C_b) \\ & \wedge\ \mathsf{Disagree}(A_a, A_b)\end{aligned}$$

Local Conflict

$$\mathsf{LocalConf}(R_a, R_b) \Leftrightarrow \mathsf{Intersect}(C_a, C_b) \wedge \mathsf{Conflict}(A_a, A_b) \wedge \neg\mathsf{SpecPrior}$$

Study Corpus

The Chinese SGLT2i Multidisciplinary Guideline Dataset

We curated a specialized corpus comprising **12 authoritative clinical guidelines** published by major branches of the **Chinese Medical Association (CMA)** and related professional societies, specifically the Diabetes Society, Society of Cardiology, Society of Nephrology, and the Chinese Geriatrics Society. The dataset focuses exclusively on **Sodium-Glucose Cotransporter-2 inhibitors (SGLT2i)**.

Rationale for Corpus Selection: The SGLT2i Multimorbidity Nexus

We selected SGLT2i therapy as the primary testbed because it perfectly exemplifies the **Organizational Fragmentation** and **Multimorbidity Conflicts** highlighted in our Introduction. The independent development of guidelines by distinct specialty societies creates a complex logical landscape that rigorously stress-tests our **Neuro-Symbolic Verification Pipeline**.

While both Cardiology and Nephrology guidelines recognize SGLT2i as a foundational therapy, they advocate for distinct drug combination paradigms based on their specialty focus. **Cardiology Guidelines**: Advocate for the Quadruple Therapy (SGLT2i + ARNI/ACEI + Beta-blocker + MRA) as the standard for Heart Failure. **Nephrology Guidelines**: Emphasize the Three Pillars of cardiorenal protection (SGLT2i + RAASi + MRA). For a patient with simultaneous Heart Failure and CKD (Cardiorenal Syndrome), these guidelines present **non-identical actionable sets**. The Nephrology does not explicitly mandate Beta-blockers, creating a discrepancy in the recommended set compared to the Cardiology. This tests the **Rule Agent's** ability to parse *Conjunctions* and the **SMT Solver's** capacity to detect Inconsistent Detail when merging guidelines for multimorbid patients.

A primary source of inconsistency arises from how different specialties prioritize adverse risks. This creates contradictory stopping rules or monitoring logic, challenging our system's ability to handle complex exclusions.**Endocrinology**: Focuses heavily on *Diabetic Ketoacidosis (DKA)*. Guidelines typically mandate discontinuation upon ketosis detection. **Cardiology**: Prioritizes hemodynamic stability. The primary constraint for suspension is symptomatic *hypotension*, often tolerating metabolic fluctuations that might concern other specialists. **Nephrology**: Focuses on *eGFR fluctuations* and *Hyperkalemia*. Unlike simple binary triggers, renal guidelines contain complex conditional logic: an initial reversible dip in eGFR is expected and does not warrant discontinuation, whereas hyperkalemia requires managing the concomitant RAASi/MRA dosage rather than simply stopping SGLT2i. Detecting these nuanced constraints validates the **Entity Agent's** ontology mapping and the **Rule Agent's** synthesis of *Compound Logical Structures*.

SGLT2i guidelines published at different times present conflicting constraints due to rapidly updating evidence, particularly regarding eGFR thresholds. Older guidelines often set conservative exclusion criteria as eGFR < 30 mL/min/1.73m^2, while newer consensus documents lower this threshold to 20 or 25 or remove it for specific indications. This temporal discrepancy creates Implication Conflicts. It serves as a critical benchmark for our **Predicate Agent's** *Arithmetic Operators* (O_{arith}), verifying whether the system can precisely model continuous numerical boundaries to identify outdated advice.

Predicate Operator

The Predicate Agent employs a set of typed operators to transform ontological entities into verifiable logical predicates. Table 4 provides a comprehensive reference of all supported operators, their semantics, return types, Z3 mappings, and typical usage examples.

Action Vocabulary

The Rule Agent maps clinical directives to a standardized action vocabulary with explicit partial ordering. This vocabulary enables precise conflict detection by establishing semantic relationships between different recommendation strengths. Table 5 provides a comprehensive reference of all supported action types, organized by category.

Table 4: Predicate Operator

Operator	Semantics	Return Type	Z3 Mapping	Typical Example
HAS	Disease diagnosis / existence	Bool	Bool	`HAS(cond.t2dm)`
ON	Medication status	Bool	Bool	`ON(med.insulin)`
HISTORY	Past history / historical events	Bool	Bool	`HISTORY(cond.stroke)`
ASSESS	Subjective assessment / clinical status	Bool	Bool	`ASSESS(Intolerance, med.statins) = True`
RISK	Risk rating	String	String	`RISK(cond.ascvd) = High`
STAGE	Staging / grading	Int/String	Int/String	`STAGE(cond.ckd) >= 3 or STAGE(cond.ckd) = B`
VALUE	Lab value / measurement	Real	Real	`VALUE(meas.egfr) < 45`
DURATION	Duration / length of time	Real	Real	`DURATION(med.insulin) >= 30`
DELTA	Change / delta	Real	Real	`DELTA(meas.creatinine) >= 50%`

Table 5: Action Vocabulary

Action Type	Semantics	Category
Usage Control (with strength ordering)		
ALLOW	Allow use (optional alternative)	Usage Control
RECOMMEND	Recommend use (preferred choice)	Usage Control
REQUIRE	Require use (mandatory)	Usage Control
CONSIDER	Consider use	Usage Control
CAUTION	Use with caution (requires monitoring)	Usage Control
AVOID	Avoid use (better alternatives exist)	Usage Control
CONTRAINDICATE	Contraindicate (absolute prohibition)	Usage Control
Continuation Control		
CONTINUE	Continue current use	Continuation Control
STOP	Stop current use	Continuation Control
Dose Adjustment		
REDUCE_DOSE	Reduce dose	Dose Adjustment
INCREASE_DOSE	Increase dose	Dose Adjustment
START_LOW_DOSE	Start with low dose	Dose Adjustment
MAX_DOSE_LIMIT	Limit maximum dose	Dose Adjustment
TITRATE	Titrate adjustment based on efficacy/tolerance	Dose Adjustment