

InsLogicBench: An Argumentation Logic Grounded Benchmark for Complex Insurance Claims Adjudication

Jin Liu¹, Yunpeng Liu¹, Keyi Wang¹, Jie Shi^{1*}
Xiao Xu², Wenkang Huang², Xingzhong Xu², Xin Liang², Yanghua Xiao^{1*}

¹Shanghai Key Laboratory of Data Science,
College of Computer Science and Artificial Intelligence, Fudan University, ²Ant Group
liujin25@m.fudan.edu.cn; {jieshi, shawyh}@fudan.edu.cn

Abstract

Insurance claims adjudication demands not only accurate decisions but also interpretable reasoning grounded in policy clauses. However, existing benchmarks are limited to information retrieval or simple multiple-choice setups, which fail to require step-by-step inferences from facts to conclusions. To address this gap, we introduce **InsLogicBench**, a benchmark providing complete reasoning traces that link factual inputs, relevant policy clauses, and final verdicts. We construct the dataset using a controllable synthesis framework based on the **Nested Toulmin Model**. By capturing the defeasible logic of insurance policies through hierarchical truth assignment and enforcing validity via consistency verification, we ensure interpretability and logical rigor across generated examples. We evaluate eight Large Language Models (LLMs) on InsLogicBench. Results show significant difficulties in handling exception clauses and verifying missing conditions. Notably, models often produce correct final decisions but fail to provide precise justifications, highlighting a critical discrepancy between their decision accuracy and logical reasoning capabilities.

1 Introduction

Insurance claims adjudication is a critical component of the insurance business. It requires verifying policy terms to confirm claim validity with logical rigor and interpretability (Kant et al., 2025; Owens et al., 2022) (see Figure 1). In recent years, Large Language Models (LLMs) have demonstrated significant potential in the insurance domain, offering new approaches for claims adjudication (Balona, 2024; Cao et al., 2024; Li et al., 2025; Troxler and Schellendorfer, 2022; Winder et al., 2025). However, hallucinations and a lack of interpretability remain substantial barriers to the practical deployment of LLMs in this field (Fu, 2025; Dahl et al., 2024).

*Corresponding authors.

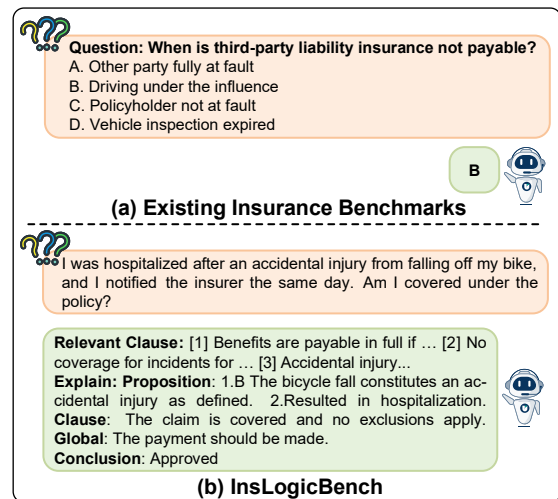


Figure 1: (a) Existing Insurance Benchmarks: Focus on multiple-choice questions that evaluate answer matching without reasoning. (b) InsLogicBench: Targets the claims adjudication task. This task requires retrieving relevant clauses and generating explanations to derive the final verdict.

To ensure reliability, it is imperative to establish a benchmark that evaluates both the accuracy of adjudication outcomes and the logical validity of the reasoning process.

While domain-specific evaluations have progressed, they remain insufficient for the demands of claims adjudication. Early retrieval-based datasets (Feng et al., 2015; Ding et al., 2025) focus primarily on simple answer matching. CUFEInse (Zhou et al., 2025) relies on multiple-choice formats, which oversimplify the complex reasoning steps required for valid judgments. INSEva (Chen et al., 2025) uses unstructured dialogues that fail to enforce explicit links between policy citations and conclusions. This leaves a critical gap for a benchmark that **demand both evidence-based adjudication and rigorous logical verification**. Compounding these limitations, the construction of such a benchmark faces significant challenges. Real-world claims data is re-

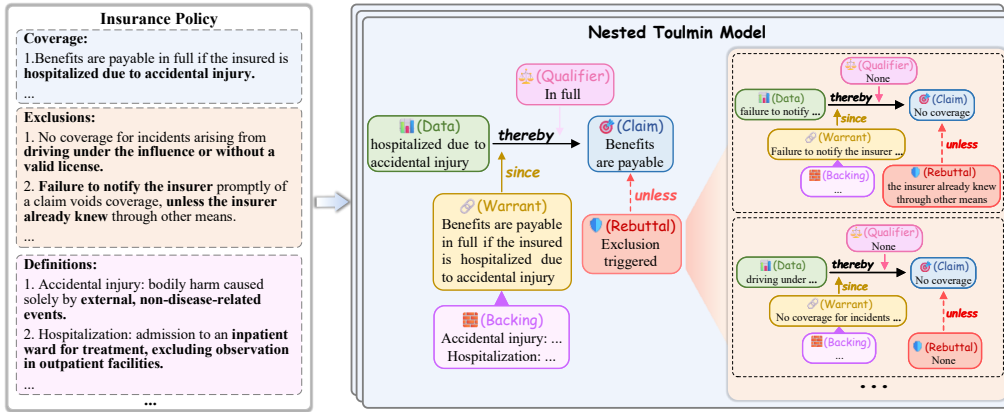


Figure 2: Nested Toulmin Model for Insurance Policy. Coverage clauses construct the argument for an Approved decision, while exclusion clauses form the argument for Denied. Arguments for *Denied* serve as the *Rebuttal* component to those for Approved, creating a nested hierarchy where valid exclusions defeat the Approved claim.

stricted by privacy regulations (Cao et al., 2024) and exhibits a long-tail distribution (Baran and Rola, 2022). Conversely, simple prompt engineering lacks constraints, often yielding unreliable data with logical inconsistencies (Long et al., 2024).

To address these challenges, we propose a controllable synthesis framework grounded in formal argumentation. We employ the Nested Toulmin Model (Toulmin, 2003) (NTM) to align with the defeasible nature of insurance contracts where general coverage rules apply only when specific exceptions are absent (Prakken and Sartor, 2015). By enforcing a rigid mapping from factual data and policy warrants to the final claim, our framework ensures interpretability and serves as the foundation for a consistency verification mechanism that guarantees the logical rigor of the generated dataset. Based on this framework, we construct **InsLogicBench**, a benchmark designed for claims adjudication. It features: (1) **Diverse Scenarios**: The dataset covers a wide range of scenarios, from standard liability triggers to complex interactions involving exclusions and their exceptions; (2) **Interpretability**: InsLogicBench includes complete reasoning traces that explicitly map factual propositions and policy clauses to the conclusion; (3) **Logical Consistency**: The ground truth is derived from formal symbolic rules and confirmed by our verification mechanism, ensuring the logical consistency of the data. In summary, our contributions are as follows:

- To the best of our knowledge, our work is the first to introduce the Nested Toulmin Model into the insurance domain. By integrating hierarchical truth assignment and verification, we generate a logically consistent dataset at scale.

- We introduce InsLogicBench, a benchmark tailored for claims adjudication. Unlike label-only datasets, it provides structured step-by-step reasoning that maps factual propositions to the final conclusion, thereby enabling interpretability of the adjudication process.
- We evaluate several LLMs on InsLogicBench. Our analysis reveals their strengths and critical limitations in claims adjudication, shedding light on the practical challenges of deploying current models in this high-stakes domain.

2 Related Work

2.1 Insurance Benchmarks

Many benchmarks have been developed for the insurance domain. InsuranceQA (Feng et al., 2015) pioneered insurance question answering, while InsQABench (Ding et al., 2025) incorporates commonsense and document-based knowledge. More recently, INSEva (Chen et al., 2025) uses multi-turn dialogues to assess professional service capabilities. CUFEInse (Zhou et al., 2025) employs multiple-choice questions to benchmark theoretical expertise. INS-MMBench (Lin et al., 2024) evaluates Large Vision-Language Models on visual perception. RAIAB (Dong, 2025) targets the reinsurance sector, emphasizing governance and compliance. While these benchmarks advance domain evaluation, they lack the logical rigor and interpretability required for claims adjudication.

2.2 Toulmin Model

The Toulmin Model (Toulmin, 2003) structures practical arguments into functional components: *claim*, *data*, *warrant*, *backing*, *qualifier*, and *rebut-*

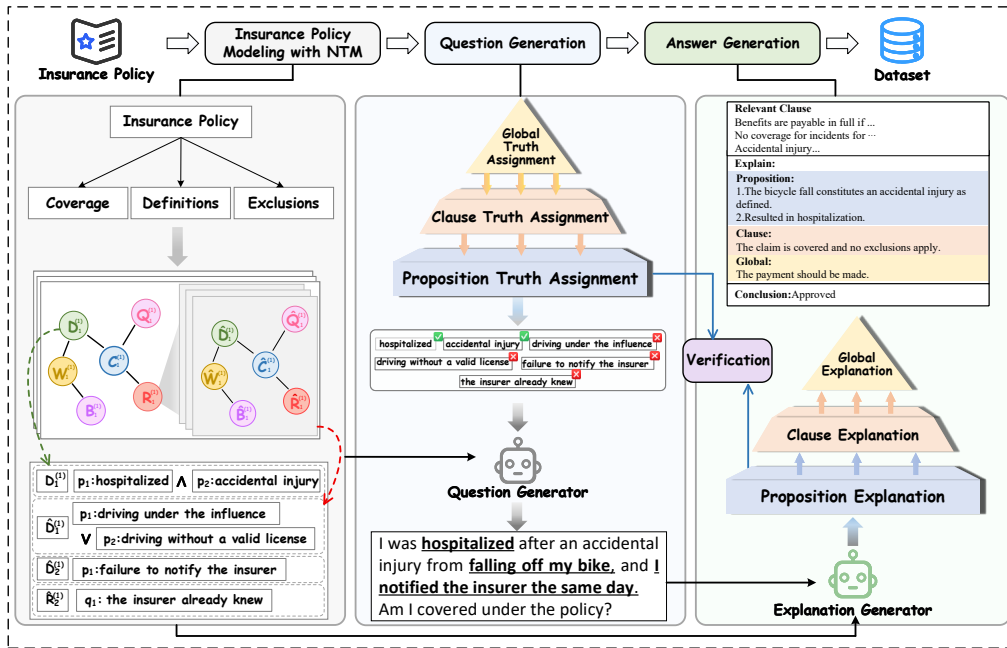


Figure 3: Overview of the data synthesis pipeline. The process consists of three stages: Insurance Policy Modeling with NTM, Question Generation, and Answer Generation.

tal. This model has long served as a foundational schema for analyzing legal texts. Pioneering works have utilized it to extract argumentation structures from legal judgments (Palau and Moens, 2011). Recent work has applied this framework to LLMs for argument evaluation (Dalzell et al., 2025), interactive reasoning steering via critical questioning (Castagna et al., 2024), and zero-shot extraction of structured argument components (Gupta et al., 2024). Extensions to multi-modal legal reasoning have also been explored (Li, 2024). These approaches predominantly treat the model as a descriptive schema for analyzing existing text. In contrast, we adapt the Toulmin model as a generative framework, utilizing its structured logic to synthesize logically consistent data.

3 InsLogicBench

We introduce **InsLogicBench**, a dataset constructed via the pipeline illustrated in Figure 3. This process consists of three stages: Insurance Policy Modeling with NTM, Question Generation, and Answer Generation.

3.1 Task Formalization

The task of claims adjudication is to predict a claim decision $\mathcal{Y} \in \{\text{Approved, Denied}\}$ given a claim query Q and the full policy text \mathcal{P} . This requires the system to retrieve relevant clauses $\mathcal{S} \subseteq \mathcal{P}$, reason over the interaction between Q and \mathcal{S} , and generate

an explanation \mathcal{E} that justifies \mathcal{Y} .

3.2 Insurance Policy Modeling with NTM

In this section, we convert unstructured policy documents into a computable argumentation framework based on the Nested Toulmin Model.

Clause Extraction Since coverage (cov), exclusion (exc), and definition (def) clauses serve as the primary evidence for assessing liability and interpreting terms, we extract them from the policy text \mathcal{P} using an extractor function Φ combining rules and an LLM. This is formalized as $\mathcal{S}_{\text{type}} = \Phi(\mathcal{P}, \text{type})$, where $\text{type} \in \{\text{cov}, \text{exc}, \text{def}\}$.

Toulmin Model Instantiation We adapt the classical Toulmin Model to the insurance domain. We define the domain-specific roles for the components as follows: **Claim** (C): The conclusion regarding liability status. **Data** (D): The factual conditions triggering the *Claim*. **Warrant** (W): The clause text authorizing the derivation of *Claim* from *Data*. **Qualifier** (Q): Constraints on the Claim, such as payout ratios. **Backing** (B): Definitions grounding the terms in *Warrant* and *Data*. **Rebuttal** (R): Conditions that defeat the *Claim*. Building on recent findings that prompting with Toulmin’s terminology aids structural identification (Gupta et al., 2024), we employ an LLM to parse policy clauses into the structural components defined above. Our prompts explicitly incorporate these

domain-specific definitions to ensure a consistent and guided extraction process (see Appendix B.2). Using this approach, we transform the extracted clause sets \mathcal{S}_{cov} and \mathcal{S}_{exc} into corresponding sets of Toulmin models. For the i -th coverage, we construct $\mathcal{M}^{(i)} = \{M_1^{(i)}, \dots, M_j^{(i)}\}$ from \mathcal{S}_{cov} , where each $M_j^{(i)}$ independently argues for liability. Conversely, we derive $\hat{\mathcal{M}}^{(i)} = \{\hat{M}_1^{(i)}, \dots, \hat{M}_k^{(i)}\}$ from \mathcal{S}_{exc} , where each $\hat{M}_k^{(i)}$ argues for claim denial.

Toulmin Model Formalization We adopt Verheij’s dialectical logic (Verheij, 2006) to construct a computational argumentation framework, utilizing \rightsquigarrow for support and \times for defeat. For the i -th coverage, we define the global liability conclusion $C^{(i)}$, coverage state $D^{(i)}$, and exclusion state $R^{(i)}$. The adjudication follows the defeasible structure $R^{(i)} \rightsquigarrow \times(D^{(i)} \rightsquigarrow C^{(i)})$, where an active $R^{(i)}$ defeats the $C^{(i)}$ derived from $D^{(i)}$. These global states result from disjunctive aggregation, where any individual clause supports the global state (i.e., $\forall j, D_j^{(i)} \rightsquigarrow D^{(i)}$; $\forall k, \hat{C}_k^{(i)} \rightsquigarrow R^{(i)}$). We represent exceptions via the nested structure $\hat{R}_k^{(i)} \rightsquigarrow \times(\hat{D}_k^{(i)} \rightsquigarrow \hat{C}_k^{(i)})$, where $\hat{R}_k^{(i)}$ defeats the specific exclusion claim.

Propositional Decomposition Following LogicLM (Pan et al., 2023), we decompose textual components into propositions via an LLM (prompts in Appendix B.3): 1) Conjunctive Decomposition. Coverage conditions $D_j^{(i)}$ comprise cumulative requirements, parsed into propositions $\{p_1, \dots, p_m\}$ such that $D_j^{(i)} \iff \bigwedge_{k=1}^m p_k$. 2) Disjunctive Decomposition. Exclusion conditions $\hat{D}_k^{(i)}$ and Exception conditions $\hat{R}_k^{(i)}$ share a disjunctive structure despite opposing roles. We parse such a condition X into $\{q_1, \dots, q_n\}$, formalized as $X \iff \bigvee_{k=1}^n q_k$, where satisfying any single q is sufficient to establish X .

3.3 Data Synthesis Pipeline

Utilizing the instantiated Nested Toulmin Model, we generate QA pairs with reasoning traces.

3.3.1 Question Generation

Scenarios Definition We formulate the generation process based on Logical Scenarios as detailed in Table 1.

Hierarchical Truth Assignment We implement the top-down hierarchical truth assignment described in Algorithm 1 to instantiate the scenar-

Scenario	v_D	v_R
Scenario I	T	F
Scenario II	F	F
Scenario III	T	T
Scenario IV	F	T

Table 1: Definition of scenarios. v_D represents the global coverage satisfaction, and v_R represents the global exclusion validity.

ios. 1) **Global Truth Assignment:** We set the global liability status $\langle v_D, v_R \rangle$ according to the selected scenario. 2) **Clause Truth Assignment:** The global state is valid if at least one corresponding clause is satisfied. Specifically, for an exclusion to be valid, the exclusion condition must be met while its exception remains inactive. 3) **Proposition Truth Assignment:** We assign truth values according to the conjunctive or disjunctive nature of the logical structure.

Algorithm 1 Hierarchical Truth Assignment

Require: Structures $\mathcal{M}^{(i)}, \hat{\mathcal{M}}^{(i)}$; Global states $\langle v_D, v_R \rangle$
Ensure: Truth assignment \mathcal{V}

- 1: $\mathcal{V} \leftarrow \emptyset$
- 2: $j^* \leftarrow \text{None}, k^* \leftarrow \text{None}$
- 3: **if** $v_D = T$ **then**
- 4: $j^* \leftarrow \text{Sample}(\text{Indices}(\mathcal{M}^{(i)}))$
- 5: **end if**
- 6: **for** $j \in \text{Indices}(\mathcal{M}^{(i)})$ **do**
- 7: Assign $\mathcal{V}[D_j^{(i)}]$ s.t. $\bigwedge p = (j = j^*)$
- 8: **end for**
- 9: **if** $v_R = T$ **then**
- 10: $k^* \leftarrow \text{Sample}(\text{Indices}(\hat{\mathcal{M}}^{(i)}))$
- 11: **end if**
- 12: **for** $k \in \text{Indices}(\hat{\mathcal{M}}^{(i)})$ **do**
- 13: **if** $k = k^*$ **then**
- 14: $\langle v_{\hat{D}}, v_{\hat{R}} \rangle \leftarrow \langle T, F \rangle$
- 15: **else**
- 16: $\langle v_{\hat{D}}, v_{\hat{R}} \rangle \leftarrow \mathcal{U}(\{\langle F, F \rangle, \langle T, T \rangle\})$
- 17: **end if**
- 18: Assign $\mathcal{V}[\hat{D}_k^{(i)}], \mathcal{V}[\hat{R}_k^{(i)}]$ s.t. $\bigvee q = v_{\hat{D}}$ and $\bigvee r = v_{\hat{R}}$
- 19: **end for**
- 20: **return** \mathcal{V}

Text Realization Finally, we transform the assignments \mathcal{V} into a natural language narrative using an LLM. This process involves two phases: 1) Semantic Validation. We discard assignments where truth values of two propositions conflict. For example, an *unlicensed driving* exclusion cannot be triggered if the claimant was a *passenger*. 2) Definition-Grounded Generation. We prompt the LLM with the complete set of assigned propositions and corresponding definitions to synthesize a coherent claim narrative. This ensures the model follows logical constraints within a unified context. For instance, to reflect a false status for *accidental injury* defined to exclude disease, the generated text describes a disqualifying event like *fell due to high blood pressure* directly within the story flow.

3.3.2 Answer Generation

Reasoning Trace Construction The inference process proceeds bottom-up through three logical layers: 1) **Proposition Explanation:** We verify the generated narrative against the proposition set. For each proposition, the LLM determines the truth value and generates a textual justification, referencing B for term interpretation. 2) **Clause Explanation:** We determine clause states $D_j^{(i)}$, $\hat{D}_k^{(i)}$, and $\hat{R}_k^{(i)}$ through conjunctive or disjunctive aggregation of proposition values. We then derive $\hat{C}_k^{(i)}$ following the structure $\hat{R}_k^{(i)} \rightsquigarrow \times (\hat{D}_k^{(i)} \rightsquigarrow \hat{C}_k^{(i)})$. These computed results constrain the LLM to generate explanations citing W as the basis. 3) **Global Explanation:** We derive global states $D^{(i)}$ and $R^{(i)}$ via disjunctive aggregation of coverage conditions and exclusion claims. The final liability decision follows $R^{(i)} \rightsquigarrow \times (D^{(i)} \rightsquigarrow C^{(i)})$. We constrain the LLM to generate the conclusion incorporating Q .

Consistency Verification We compare the *assigned truth values* \mathcal{V} from top-down assignment with the *inferred truth values* from bottom-up inference. We retain only samples where the inferred logic matches the assigned ground truth. This ensures a strict correspondence between the generated narrative and the logical labels.

4 Dataset Analysis

This section introduces the dataset and evaluates its reliability through human evaluation.

4.1 Dataset Statistics

We collected 110 Chinese-language insurance policy documents from publicly available sources, with an average length of 10,103 tokens. These documents span diverse insurance product types as detailed in Appendix C. The dataset contains 2,621 question–answer (QA) pairs, each consisting of a question and a detailed answer that includes clause citations, explanations, and a conclusion.

Distribution by Scenario Table 2 summarizes the basic statistics. The dataset maintains a reasonably balanced distribution across the four adjudication scenarios to reduce the risk of model bias toward specific outcomes. In contrast, instances involving nested exceptions are much less frequent than standard cases. We retain this natural imbalance to reflect the rarity of exception clauses in real-world policies, which supports a realistic evaluation of long-tail reasoning capabilities.

Category	Count	Len.	# Props.	# Cites.
<i>Logical Scenarios</i>				
Scenario I	881	321.74	3.80	2.86
Scenario II	682	324.75	3.88	2.89
Scenario III	556	328.86	4.97	4.46
Scenario IV	502	330.20	4.96	4.70
<i>Nested Structures</i>				
w/o Exception	2048	319.63	4.01	3.44
w/ Exception	573	347.19	5.31	3.98
Overall / Avg.	2621	325.66	4.29	3.56

Table 2: Statistics of the INSLOGICBENCH. **Len.** denotes the average number of tokens per question. **# Props.** and **# Cites.** represent the average number of propositions and clause citations. *Nested Structures w/o Exception* denotes scenarios with no internal exceptions within exclusion clauses, while *w/ Exception* refers to those that include such exceptions.

Complexity Analysis The claim inquiries have an average length of 326 tokens. Responses contain on average 4.29 propositional explanations and 3.56 clause citations. Figure 4 shows the distribution of these two measures per answer. Both exhibit a unimodal distribution. Most answers involve a moderate number of explanations and citations, with fewer at the low and high ends.

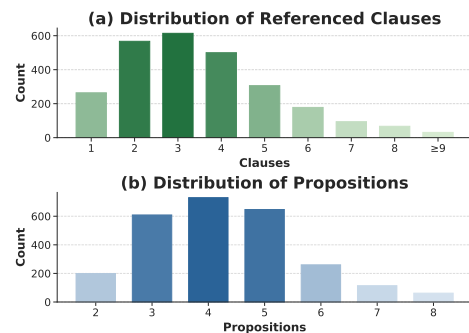


Figure 4: Distribution of the Clauses and Propositions.

Semantic Diversity Analysis The semantic distribution of claim queries is visualized in Figure 5. The t-SNE plot shows that queries from all four scenarios are distributed across a continuous space with significant overlap. This indicates that the linguistic features of the queries are not strictly tied to their logical categories. Consequently, the scenarios are not easily distinguishable based solely on surface-level text patterns, which suggests that correct adjudication depends more on underlying logical reasoning.

4.2 Quality Assurance

We engaged three domain experts with insurance underwriting experience to assess the quality of

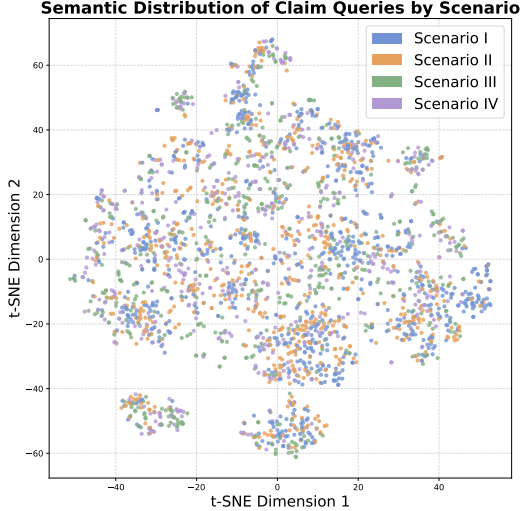


Figure 5: Semantic distribution of claim queries categorized by scenario.

200 instances selected via scenario-based stratified sampling. The evaluation covers three dimensions: (1) **Label Validity (LV)**: A categorical assessment (Supported / Insufficient / Contradicted) to verify whether the verdict is factually accurate and strictly substantiated by policy clauses; (2) **Reasoning Faithfulness (RF)**: A 5-point Likert scale measuring the precision of clause citations and the completeness of the explanation; (3) **Linguistic Naturalness (LN)**: A 5-point Likert scale evaluating professional fluency and readability. Detailed scoring rubrics are provided in Appendix D. We calculated inter-annotator agreement using Fleiss’ Kappa (κ) for categorical labels and Krippendorff’s Alpha (α) for ordinal scales. As shown in Table 3, the results suggest high dataset quality and strong expert consensus.

Metric	Score	Agreement
LV (Supported %)	94.5%	$\kappa = 0.85$
RF (1–5 scale)	4.67	$\alpha = 0.81$
LN (1–5 scale)	4.64	$\alpha = 0.83$

Table 3: Human evaluation results.

5 Experiments

5.1 Experimental Settings

Models and Setup We select 8 LLMs for evaluation, which are as follows: **DeepSeek-R1** (Guo et al., 2025), **GPT-5.1**, **Qwen3-235B-A22B-Instruct** (Team, 2025b), **Kimi-K2-Instruct** (Team, 2025a), **GLM-4.6**, **GPT-OSS-120B** (OpenAI, 2025), **Qwen2.5-72B-Instruct** (Yang et al., 2024), and **Qwen3-32B** (Team, 2025b). All selected models possess sufficient context windows to process

the complete policy text and claim description without truncation. Detailed prompt templates are provided in Appendix B.5.

Evaluation Metrics. We adopt the LLM-as-a-Judge approach (Zheng et al., 2023) to evaluate the model’s performance based on semantic consistency. We report three sets of metrics: (1) **Citation Precision (Prec.)**, **Recall (Rec.)**, and **F1**, where the evaluator determines whether a retrieved clause is semantically equivalent to the ground truth; (2) **Explanation Coverage (Cov.)**, which measures the proportion of ground-truth explanatory propositions covered by the generated rationale; (3) **Adjudication Accuracy (Acc.)**, derived by extracting the final verdict from the model’s response. To ensure evaluation reliability, we utilized an ensemble judge aggregating predictions from DeepSeek-R1, GPT-5.1, and Qwen3-235B via majority voting. We validated this protocol on 200 instances selected through stratified sampling and labeled by three domain experts (Fleiss’ $\kappa = 0.81$). The results show a strong correlation with human judgments ($r = 0.86$). Detailed prompts are provided in Appendix B.5.

Model	Prec.	Rec.	F1	Cov.	Acc.
DeepSeek-R1	0.62	0.70	0.66	0.71	0.82
Qwen3-235B	0.47	<u>0.79</u>	0.59	<u>0.75</u>	<u>0.79</u>
GPT-5.1	0.40	0.84	0.54	0.84	0.78
GLM-4.6	0.54	0.67	0.60	0.65	0.78
GPT-OSS-120B	0.56	0.67	0.61	0.60	0.78
Kimi-K2	<u>0.60</u>	0.69	<u>0.64</u>	0.62	0.77
Qwen3-32B	0.59	0.64	0.61	0.64	0.74
Qwen2.5-72B	0.59	0.64	0.61	0.64	0.73

Table 4: Main results on InsLogicBench. **Bold** indicates best, underline indicates second best.

5.2 Main Results

Table 4 presents the performance of various LLMs on InsLogicBench. DeepSeek-R1 achieves the highest Adjudication Accuracy of 0.82, indicating that current models remain below the reliability standards for claim adjudication. We observe two patterns regarding the trade-off between retrieval and reasoning: 1) Models like GPT-5.1 and Qwen3-235B prioritize high recall. GPT-5.1 records the highest Citation Recall of 0.84 and Explanation Coverage of 0.84, but its lower Citation Precision of 0.40 implies the retrieval of irrelevant policy clauses. This pattern may help explain why high retrieval rates do not translate into the highest Adjudication Accuracy. 2) Models like Qwen3-32B and Qwen2.5-72B place greater emphasis on Cita-

	Prec.	Rec.	F1	Cov.	Acc.	Prec.	Rec.	F1	Cov.	Acc.
Scenario	DeepSeek-R1					Qwen3-235B				
Scenario I	0.55	0.80	0.65	0.81	0.88	0.41	0.87	0.56	0.83	0.91
Scenario II	0.57	0.72	0.64	0.73	0.71	0.41	0.81	0.54	0.77	0.67
Scenario III	0.75	0.64	0.69	0.66	0.82	0.57	0.73	0.64	0.70	0.74
Scenario IV	0.68	0.58	0.63	0.57	0.85	0.55	0.69	0.61	0.64	0.81
	GPT-5.1					GLM-4.6				
Scenario I	0.33	0.90	0.48	0.92	0.93	0.47	0.75	0.58	0.73	0.83
Scenario II	0.35	0.87	0.50	0.85	0.64	0.48	0.71	0.57	0.66	0.66
Scenario III	0.51	0.78	0.62	0.81	0.71	0.66	0.61	0.63	0.61	0.79
Scenario IV	0.48	0.76	0.59	0.71	0.78	0.59	0.56	0.57	0.52	0.86
	GPT-OSS-120B					Kimi-K2				
Scenario I	0.49	0.77	0.60	0.68	0.85	0.54	0.78	0.64	0.71	0.93
Scenario II	0.50	0.67	0.57	0.61	0.66	0.55	0.72	0.62	0.64	0.62
Scenario III	0.69	0.60	0.64	0.54	0.76	0.70	0.62	0.66	0.56	0.70
Scenario IV	0.63	0.55	0.59	0.48	0.82	0.67	0.58	0.62	0.50	0.79
	Qwen3-32B					Qwen2.5-72B				
Scenario I	0.53	0.74	0.62	0.72	0.83	0.53	0.73	0.61	0.73	0.83
Scenario II	0.54	0.67	0.60	0.68	0.61	0.55	0.67	0.60	0.68	0.56
Scenario III	0.68	0.56	0.61	0.55	0.73	0.69	0.57	0.62	0.56	0.72
Scenario IV	0.65	0.52	0.58	0.54	0.80	0.66	0.52	0.58	0.54	0.78

Table 5: Model performance comparison across scenarios. **Bold** values indicate the best performance per metric for each model across all scenarios.

tion Precision, with comparatively higher Precision but lower Citation Recall. This imbalance suggests these models frequently overlook necessary policy clauses, resulting in incomplete reasoning chains. Finally, the results demonstrate that balancing Citation Precision and Recall is essential. DeepSeek-R1 achieves the highest F1 score of 0.66 and Adjudication Accuracy of 0.82, suggesting that correctly identifying applicable policy clauses while excluding irrelevant ones is critical for accurate adjudication.

5.3 Detailed Analysis

Scenario-based Analysis Table 5 presents the performance across different scenarios. A comparison between Scenario I and Scenario II reveals a performance gap in identifying unmet conditions. Across all models, performance in Scenario I is stronger than in Scenario II, with higher Explanation Coverage and higher Adjudication Accuracy. However, accuracy drops in Scenario II where the claim is denied because a coverage condition is missing. For instance, DeepSeek-R1’s Adjudication Accuracy decreases from 0.88 to 0.71. This result indicates that verifying the absence of a required condition is more difficult for models than verifying its presence.

Comparing Scenario I and Scenario III highlights the impact of exclusion clauses. Although coverage conditions are satisfied in both cases, the presence of an active exclusion in Scenario III causes a decline in Adjudication Accuracy (e.g.,

GPT-5.1 drops from 0.93 to 0.71). Furthermore, the Explanation Coverage also decreases. This suggests that when models identify an exclusion that justifies denial, they often stop processing the remaining coverage evidence, resulting in an incomplete reasoning chain.

In Scenario IV, claims are invalid due to both unmet coverage conditions and triggered exclusions. Although Adjudication Accuracy is higher here than in Scenario II (e.g., GPT-5.1 scores 0.78 in IV versus 0.64 in II), the Explanation Coverage is the lowest across all scenarios (e.g., DeepSeek-R1 records 0.57). This indicates that the high accuracy does not reflect a complete reasoning process. The models likely identify only one sufficient reason for denial and ignore the other factor. Therefore, the correct verdicts result from partial reasoning rather than a comprehensive evaluation of all evidence.

Overall, Scenarios I and II exhibit higher Citation Recall but lower Citation Precision. In these contexts, standard adjudication protocols do not require the citation of untriggered exclusion clauses. Despite this, the models frequently retrieve these irrelevant exclusions to augment their conclusions. Conversely, Scenarios III and IV exhibit lower Citation Recall compared to the first two scenarios. This reduction suggests that the models perform poorly in identifying the specific exclusion clauses when those clauses are actively triggered.

Impact of Exception Clauses Figure 6 compares model performance on samples with and without

triggered exception clauses. We observe a consistent decline in both Explanation Coverage and Adjudication Accuracy when exceptions are introduced. For example, DeepSeek-R1 drops by 6.65% in coverage and 15.62% in accuracy. This concurrent decline in both metrics reveals a fundamental challenge for current LLMs in handling complex exception scenarios.

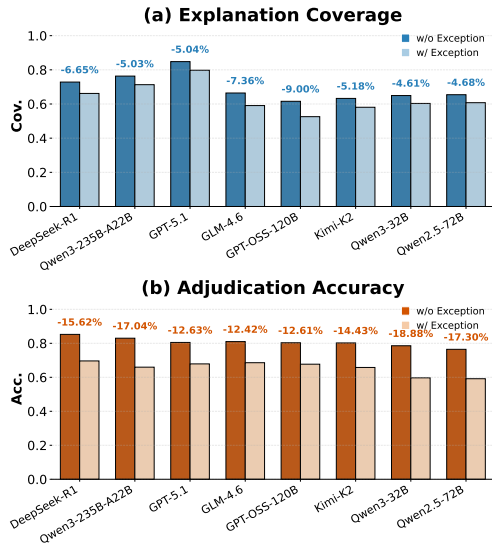


Figure 6: Impact of exception clauses. (a) Comparison of Explanation Coverage with and without Exception Clauses. (b) Comparison of Adjudication Accuracy with and without Exception Clauses.

Impact of Reasoning Chain Length We examine how the model performs as the reasoning chain lengthens, defined by the increasing number of ground-truth propositions required for a complete explanation. Figure 7 shows an overall negative trend, with performance generally decreasing as the proposition count rises from 2 to 8. For instance, DeepSeek-R1 sees its Explanation Coverage score drop significantly from approximately 0.87 in short chains to 0.54 in long chains. This decline suggests that as the logical process becomes more extensive, the model is more prone to overlooking specific necessary details. The failure to exhaustively capture these key points directly results in lower Explanation Coverage and consequently degrades the final Adjudication Accuracy.

Question-only Adversarial Check We further evaluate the top three models using only the question, without the policy text, and compare them with a majority-label baseline. The majority baseline always predicts the dominant label in the dataset and provides a simple reference for label-

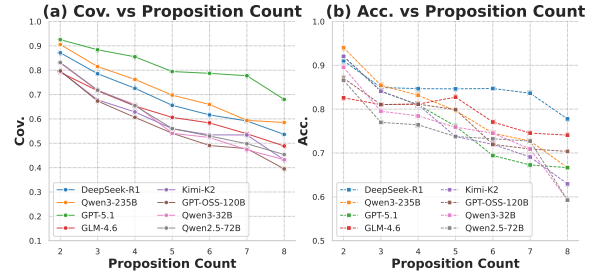


Figure 7: Performance vs. Proposition Count. (a) Explanation Coverage and (b) Adjudication Accuracy across varying numbers of ground-truth propositions.

distribution bias. As shown in Table 6, question-only performance remains much closer to this baseline than to the standard policy-provided setting, suggesting that surface-level phrasing alone is insufficient for reliable adjudication.

Model / Baseline	Question-only	Policy provided
Majority baseline	0.66	—
DeepSeek-R1	0.62	0.82
Qwen3-235B	0.62	0.79
GPT-5.1	0.63	0.78

Table 6: Question-only adversarial check on InsLogicBench. The majority baseline always predicts the majority label.

Case Study Figure 8 illustrates a typical failure in Scenario II. The Ground Truth (left) correctly denies the claim because the claimant’s visual metrics exceed the specific policy thresholds. However, the model (right) overlooks these critical numerical constraints, erroneously assuming that the payout conditions are satisfied and resulting in an incorrect Approved conclusion (more details in Appendix E).

Question
I would like to inquire about ...the best corrected visual acuity in both eyes was only 0.03, and the visual field radius was only 10 degrees ... Can we apply for a claim in this situation?

Ground Truth	Model Generation
<p>Relevant clause: Permanent and Total Blindness in Both Eyes: Refers to best corrected visual acuity below 0.02 on the international standard visual acuity chart, or a visual field radius of less than 5 degrees...</p> <p>Explain: Proposition: ...Corrected visual acuity is 0.03 (>0.02 standard) and visual field radius is 10° (>5° standard) for "Permanent and Total Blindness in Both Eyes" per the contract... Clause: The user... Total Disability are not satisfied. Global: ...not meet the definition ... thus failing the conditions for Total Disability.</p> <p>Conclusion: Denied</p>	<p>Citations: Total Disability: Refers to the Insured, during the ... from the date of accidental injury or disease diagnosis...</p> <p>Analysis: ...user's life depends entirely on the care of others, and they cannot continue to work. Insured's situation meets the definition of Total Disability ... In summary, the facts meet the conditions for insurance</p> <p>Conclusion: Approved.</p>

Figure 8: A representative case study from Scenario II illustrating model failure in numerical verification.

6 Conclusion

In this paper, we introduce a controllable synthesis framework based on the Nested Toulmin Model to construct InsLogicBench. We formalize policy documents into computable argumentation structures and employ a consistency verification mechanism to guarantee strict logical consistency. InsLogicBench addresses the critical need for interpretability in insurance claims adjudication by providing reasoning traces alongside final decisions. Our evaluation demonstrates that current LLMs still face significant challenges in this domain. Models struggle with defeasible reasoning involving exception clauses and fail to consistently identify unmet coverage conditions. These results indicate a gap between current retrieval capabilities and the requirements for rigorous logical verification.

Limitations

First, the dataset relies on synthesized scenarios where all necessary facts are explicitly stated. In contrast, real-world claims frequently contain ambiguous descriptions or incomplete information that complicates the decision process. Second, this study focuses solely on textual reasoning for liability determination. Practical adjudication often involves multimodal evidence and requires quantitative calculations for loss assessment, which are not covered in this work. Third, the quality of the synthesized data depends on both the construction quality of the Nested Toulmin Model and the capabilities of the LLMs used in the generation pipeline. Any errors or omissions in policy modeling, or limitations in the models' extraction and generation abilities, may affect the logical validity and overall quality of the resulting dataset.

Ethics Statement

The policy clauses used in this study were collected from publicly available insurance policy documents. All claim queries and scenarios in the dataset are synthetically generated and do not involve any real-world insurance claims or private information of individuals. The dataset is intended for research purposes only and does not constitute professional insurance or legal advice.

Acknowledgements

This work was supported by Ant Group.

References

- Caesar Balona. 2024. Actuarygpt: Applications of large language models to insurance and actuarial work. *British Actuarial Journal*, 29:e15.
- Sebastian Baran and Przemyslaw Rola. 2022. Prediction of motor insurance claims occurrence as an imbalanced machine learning problem. *CoRR*, abs/2204.06109.
- Chunling Cao, Xiuling Yao, Wenying Gong, Yutong Li, Yang Pan, Yuechen Yang, and Chaoran Zhao. 2024. Llms for insurance: Opportunities, challenges and concerns. *Preprints*.
- Federico Castagna, Isabel Sassoon, and Simon Parsons. 2024. Critical-questions-of-thought: Steering LLM reasoning with argumentative querying. *CoRR*, abs/2412.15177.
- Shisong Chen, Qian Zhu, Wenyan Yang, Chengyi Yang, Zhong Wang, Ping Wang, Xuan Lin, Bo Xu, Daqian Li, Chao Yuan, Licai Qi, Wanqing Xu, Zhenxing Sun, Xin Lu, Shiqiang Xiong, Chao Chen, Haixiang Hu, and Yanghua Xiao. 2025. Inseva: A comprehensive chinese benchmark for large language models in insurance. *CoRR*, abs/2509.04455.
- Matthew Dahl, Varun Magesh, Mirac Suzgun, and Daniel E Ho. 2024. Hallucinating law: Legal mistakes with large language models are pervasive. *Law, regulation, and policy*.
- Geordie Dalzell, Chen Liu, Wei Peng, and Xinghuo Yu. 2025. Analysis of large language model prompting and generation using toulmin's model. In *AI 2025: Advances in Artificial Intelligence - 38th Australasian Joint Conference on Artificial Intelligence, AI 2025, Canberra, ACT, Australia, December 1-5, 2025, Proceedings, Part I*, Lecture Notes in Computer Science, pages 3–14. Springer.
- Jing Ding, Kai Feng, Binbin Lin, Jiarui Cai, Qiushi Wang, Yu Xie, Xiaojin Zhang, Zhongyu Wei, and Wei Chen. 2025. Insqabench: Benchmarking chinese insurance domain question answering with large language models. *CoRR*, abs/2501.10943.
- Stella C. Dong. 2025. Prudential reliability of large language models in reinsurance: Governance, assurance, and capital efficiency. *CoRR*, abs/2511.08082.
- Minwei Feng, Bing Xiang, Michael R. Glass, Lidan Wang, and Bowen Zhou. 2015. Applying deep learning to answer selection: A study and an open task. In *2015 IEEE Workshop on Automatic Speech Recognition and Understanding, ASRU 2015, Scottsdale, AZ, USA, December 13-17, 2015*, pages 813–820. IEEE.
- Weilong Fu. 2025. The new quant: A survey of large language models in financial prediction and trading. *Preprint*, arXiv:2510.05533.
- Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang,

- Shirong Ma, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, Aixin Liu, Bing Xue, Bingxuan Wang, Bochao Wu, Bei Feng, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chong Ruan, Damai Dai, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, Fuli Luo, Guangbo Hao, Guanting Chen, Guowei Li, Hao Zhang, Hanwei Xu, Honghui Ding, Huazuo Gao, Hui Qu, Hui Li, Jianzhong Guo, Jiaoshi Li, Jingchang Chen, Jingyang Yuan, Jinhao Tu, Junjie Qiu, Junlong Li, J. L. Cai, Jiaqi Ni, Jian Liang, Jin Chen, Kai Dong, Kai Hu, Kaichao You, Kaige Gao, Kang Guan, Kexin Huang, Kuai Yu, Lean Wang, Lecong Zhang, Liang Zhao, Litong Wang, Liyue Zhang, Lei Xu, Leyi Xia, Mingchuan Zhang, Minghua Zhang, Minghui Tang, Mingxu Zhou, Meng Li, Miaojun Wang, Mingming Li, Ning Tian, Panpan Huang, Peng Zhang, Qiancheng Wang, Qinyu Chen, Qiushi Du, Ruiqi Ge, Ruisong Zhang, Ruizhe Pan, Runji Wang, R. J. Chen, R. L. Jin, Ruyi Chen, Shanghao Lu, Shangyan Zhou, Shanhuang Chen, Shengfeng Ye, Shiyu Wang, Shuiping Yu, Shunfeng Zhou, Shuting Pan, S. S. Li, Shuang Zhou, Shaoqing Wu, Tao Yun, Tian Pei, Tianyu Sun, Tao Wang, Wangding Zeng, Wen Liu, Wenfeng Liang, Wenjun Gao, Wenqin Yu, Wentao Zhang, W. L. Xiao, Wei An, Xiaodong Liu, Xiaohan Wang, Xiaokang Chen, Xiaotao Nie, Xin Cheng, Xin Liu, Xin Xie, Xingchao Liu, Xinyu Yang, Xinyuan Li, Xuecheng Su, Xuheng Lin, X. Q. Li, Xiangyue Jin, Xiaojin Shen, Xiaosha Chen, Xiaowen Sun, Xiaoxiang Wang, Xinnan Song, Xinyi Zhou, Xianzu Wang, Xinxia Shan, Y. K. Li, Y. Q. Wang, Y. X. Wei, Yang Zhang, Yanhong Xu, Yao Li, Yao Zhao, Yaofeng Sun, Yaohui Wang, Yi Yu, Yichao Zhang, Yifan Shi, Yiliang Xiong, Ying He, Yishi Piao, Yisong Wang, Yixuan Tan, Yiyang Ma, Yiyuan Liu, Yongqiang Guo, Yuan Ou, Yuduan Wang, Yue Gong, Yuheng Zou, Yujia He, Yunfan Xiong, Yuxiang Luo, Yuxiang You, Yuxuan Liu, Yuyang Zhou, Y. X. Zhu, Yanping Huang, Yaohui Li, Yi Zheng, Yuchen Zhu, Yunxian Ma, Ying Tang, Yukun Zha, Yuting Yan, Z. Z. Ren, Zehui Ren, Zhangli Sha, Zhe Fu, Zhean Xu, Zhenda Xie, Zhengyan Zhang, Zhewen Hao, Zhicheng Ma, Zhigang Yan, Zhiyu Wu, Zihui Gu, Zijia Zhu, Zijun Liu, Zilin Li, Ziwei Xie, Ziyang Song, Zizheng Pan, Zhen Huang, Zhipeng Xu, Zhongyu Zhang, and Zhen Zhang. 2025. [Deepseek-r1 incentivizes reasoning in llms through reinforcement learning](#). *Nat.*, 645(8081):633–638.
- Ankita Gupta, Ethan Zuckerman, and Brendan T. O’Connor. 2024. [Harnessing toulmin’s theory for zero-shot argument explication](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 10259–10276. Association for Computational Linguistics.
- Manuj Kant, Sareh Nabi, Manav Kant, Roland Scharer, Megan Ma, and Marzieh Nabi. 2025. [Towards robust legal reasoning: Harnessing logical llms in law](#). *CoRR*, abs/2502.17638.
- Dongchen Li, Zhuo Jin, Linyi Qian, and Hailiang Yang. 2025. [Textual analysis of insurance claims with large language models](#). *Journal of Risk and Insurance*, 92(2):505–535.
- Jiaxing Li. 2024. [Moving beyond text: Multi-modal expansion of the toulmin model for enhanced ai legal reasoning](#). In *Intelligent Computers, Algorithms, and Applications*, pages 299–308, Singapore. Springer Nature Singapore.
- Chenwei Lin, Hanjia Lyu, Xian Xu, and Jiebo Luo. 2024. [Ins-mmbench: A comprehensive benchmark for evaluating lvlms’ performance in insurance](#). *CoRR*, abs/2406.09105.
- Lin Long, Rui Wang, Ruixuan Xiao, Junbo Zhao, Xiao Ding, Gang Chen, and Haobo Wang. 2024. [On llms-driven synthetic data generation, curation, and evaluation: A survey](#). In *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, Findings of ACL, pages 11065–11082. Association for Computational Linguistics.
- OpenAI. 2025. [gpt-oss-120b & gpt-oss-20b model card](#). *CoRR*, abs/2508.10925.
- Emer Owens, Barry Sheehan, Martin Mullins, Martin Cunneen, Juliane Ressel, and German Castignani. 2022. [Explainable artificial intelligence \(xai\) in insurance](#). *Risks*, 10(12).
- Raquel Mochales Palau and Marie-Francine Moens. 2011. [Argumentation mining](#). *Artif. Intell. Law*, 19(1):1–22.
- Liangming Pan, Alon Albalak, Xinyi Wang, and William Yang Wang. 2023. [Logic-llm: Empowering large language models with symbolic solvers for faithful logical reasoning](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, Findings of ACL, pages 3806–3824. Association for Computational Linguistics.
- Henry Prakken and Giovanni Sartor. 2015. [Law and logic: A review from an argumentation perspective](#). *Artif. Intell.*, 227:214–245.
- Kimi Team. 2025a. [Kimi K2: open agentic intelligence](#). *CoRR*, abs/2507.20534.
- Qwen Team. 2025b. [Qwen3 technical report](#). *CoRR*, abs/2505.09388.
- Stephen E Toulmin. 2003. *The uses of argument*. Cambridge university press.
- Andreas Troxler and Jürg Schelldorfer. 2022. [Actuarial applications of natural language processing using transformers: Case studies for using text features in an actuarial context](#). *CoRR*, abs/2206.02014.
- Bart Verheij. 2006. [Evaluating arguments based on toulmin’s scheme](#). In David Hitchcock and Bart Verheij, editors, *Arguing on the Toulmin Model*, Argumentation Library, pages 181–202. Springer.

Philipp Winder, Laurenz Hommel, and Christian Alexander Hildebrand. 2025. Ai-based claims handling: A systematic performance and bias assessment of large language models for automated insurance claims handling.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiayi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. 2024. [Qwen2.5 technical report](#). *CoRR*, abs/2412.15115.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. [Judging llm-as-a-judge with mt-bench and chatbot arena](#). In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Hua Zhou, Bing Ma, Yufei Zhang, and Yi Zhao. 2025. [Design, results and industry implications of the world’s first insurance large language model evaluation benchmark](#). *CoRR*, abs/2511.07794.

A Expert Recruitment and Compensation

We recruited three domain experts specifically for the Quality Assurance phase. These participants possess over five years of experience in insurance underwriting and claims adjudication. Each expert was required to demonstrate proficiency in interpreting complex policy clauses and identifying logical fallacies. Participants were compensated at a rate of approximately \$30 per hour. All experts provided informed consent regarding the usage of their evaluation data for research purposes.

B Prompt Template

This section provides the specific prompt templates used in this paper.

B.1 Prompt for Clause Extraction

Figure 9 illustrates the prompt used to extract exclusions from the policy described in Section 3.2.

B.2 Prompts for Toulmin Model Instantiation

Following clause extraction, we employ specific prompts to parse the text into the Nested Toulmin Model components. Figure 10 displays the prompt

for processing **Coverage** clauses, while Figure 11 shows the prompt for **Exclusion** clauses.

B.3 Prompts for Propositional Decomposition

As detailed in Section 3.2, clause text is decomposed into propositions. Figure 12 illustrates the prompt for **Conjunctive Decomposition** (Logical AND). Figure 13 displays the prompt for **Disjunctive Decomposition** (Logical OR).

B.4 Prompts for Text Realization and Answer Generation

Figure 14 presents the prompt for **Text Realization**, where the model synthesizes a claim narrative based on the assigned truth values of propositions (Section 3.3.1). Figure 15 displays the prompt for **Answer Generation**.

B.5 Prompt for Claim Adjudication (Model Inference)

Figure 16 presents the standardized prompt used in our experiments (Section 5.1) to evaluate the baseline models.

B.6 Prompts for Automated Evaluation (LLM-as-a-Judge)

We employ an LLM-as-a-Judge approach to quantify model performance (see Section 5.1). Figure 17 shows the prompt used to assess Citation Precision, Recall, and F1. Figure 18 presents the prompt for Explanation Coverage, measuring the recall of ground-truth propositions. Finally, Figure 19 displays the prompt for extracting and verifying the final Adjudication Accuracy.

C Dataset Statistics and Analysis

This section provides a detailed breakdown of the insurance policy documents used to construct Ins-LogicBench.

C.1 Classification of Insurance Products

To ensure the benchmark covers a wide range of reasoning logic and domain terminology, we collected 110 real-world insurance policy documents. Table 7 presents the detailed classification, count, and representative examples for each category. As shown, the dataset includes diverse product lines ranging from standard life insurance to niche products like nursing care and income loss insurance.

C.2 Full Data Examples

In this section, we present a comprehensive visualization of a data sample from InsLogicBench. Table 8 demonstrates the end-to-end transformation of a raw policy clause into a logically consistent QA pair.

The visualization captures four key layers: **Raw Policy Input:** The original unstructured text from the insurance contract. **Toulmin Model:** The parsed Toulmin Model (Claim, Data, Warrant). **Truth Assignment:** The hierarchical assignment of boolean values to propositions. **QA Generation:** The final synthesized Question, Standard Answer, and the structured Reasoning Trace (Verdict + Citations + Explanation).

D Human Evaluation Rubrics

To ensure rigorous quality control, we provided the domain experts with detailed scoring guidelines. The evaluation is conducted across three dimensions: **Label Validity (LV):** Assesses the factual accuracy and supportability of the adjudication conclusion (Table 9). **Reasoning Faithfulness (RF):** Evaluates the precision of clause citations and the logical validity of the derivation chain using a 5-point Likert scale (Table 10). **Linguistic Naturalness (LN):** Measures the grammatical correctness and professional tone of the generated narrative using a 5-point Likert scale (Table 11).

E Case Studies

In this section, we provide a qualitative analysis of model performance on InsLogicBench. We selected five representative cases to illustrate both the capabilities of current LLMs and their specific failure modes in complex adjudication scenarios.

Figure 20 demonstrates a successful reasoning chain. Figures 21 through 24 highlight critical challenges, including overlooked conditions, incomplete evidence retrieval, and failures in processing nested exclusion logic.

Role: You are a professional Legal Analyst specializing in insurance contracts, expert at accurately extracting information from policy terms and conditions.

Task: Extract all circumstances from the provided insurance policy that explicitly state the insurer “does not assume insurance liability” (exclusions).

Rules

1. Strict Exclusions (What NOT to extract): Non-conclusive Clauses: Terms that only affect the payout amount (e.g., “deducting outstanding premiums/loans”). **Application Procedures:** All circumstances related to claim documentation, filing processes, or administrative formalities. **Non-Exclusionary Liability Clauses:** Clauses describing how liability is fulfilled under different scenarios (e.g., “returning premiums if an event occurs during the waiting period”). These are liability definitions, not exclusions.

2. Extraction Scope: Only extract scenarios that directly lead to a “no payout” conclusion. Criteria: The clause explicitly contains phrases such as “we shall not assume the liability to pay insurance benefits” or “no insurance liability shall be assumed”; OR the clause describes circumstances leading to contract termination where no further liability is assumed.

3. ID Generation Rules: No Duplicates. Base ID on the parent heading number. Format: Parent Heading_Sub-item Number (e.g., 2.4.1). Sub-items increment from 1. Segments separated by two line breaks are distinct clauses.

4. Content Field Rules: Maintain Integrity: Ensure content is semantically complete (retain exceptions like “unless Y occurs”). **Strict Adherence:** Direct copy only. Do not summarize. **Retain All References:** Keep parenthetical references like “(see...)”.

Output Format: Output strictly in the following JSON format without any explanatory text.

```
[ { "id": "x.x.x", "content": "Full sentence or clause copied directly from the original text." }, ... ]
```

Figure 9: Prompt template for Clause Extraction.

Role Positioning

You are a professional Insurance Policy Analysis Expert, specializing in using the **Toulmin Model of Argumentation** to analyze insurance liability clauses.

Core Task

Extract and analyze payout scenarios from insurance liability clauses. For **each independent payout trigger**, construct **one and only one** argumentation following the Toulmin Model. Different scenarios must be strictly split or merged based on logical connectors (e.g., “or”, “and”) found in the clauses.

Toulmin Model Element Specifications

1. Claim – Content: The legal conclusion of the payout. **Format:** Must use the phrase “Shall pay/be liable for [Specific Name of Insurance Benefit]”. **Requirement:** **Must not contain numerical values** such as specific amounts or percentages. **Example:** “Shall pay Critical Illness Benefit”

2. Data – Content: Specific conditions that trigger the Claim. **Requirement:** Extract objective facts directly from the clause. Maintain semantic integrity **without adding interpretations, summaries, or inferences**. **Example:** “The insured, due to accidental injury, suffers a disability categorized in the ‘Standards and Codes for Disability Assessment of Personal Insurance’ within 180 days from the date of the accident”

3. Warrant – Content: The **full original text** of the relevant clause in the insurance policy. **Requirement:** Must be the original text or a lossless transcription. **Splicing, rewriting, or deletions are strictly prohibited.**

4. Backing – Content: Contractual definitions of key terms appearing in the Data or Warrant. **Requirement:** Fill this only if the clause **explicitly cites a definition location** (e.g., “see 8.3”); otherwise, use []. **Format:** A list of strings, each being a definition index. **Example:** [“8.3”]

5. Qualifier – Content: All specific limiting conditions regarding “how to pay” or “how much to pay.” **Inclusions:** Payout ratios, maximum caps, deductible rules, etc. **Requirement:** Must come from the original text; if none, set to null. **Example:** “The Accidental Disability Benefit is limited to the Basic Sum Insured; the contract terminates when the cumulative payout reaches this limit.”

6. Rebuttal – Content: Explicit exceptions or exclusions that would invalidate the Claim. **Requirement:** Fill this only if the clause **explicitly lists exclusions**; otherwise, set to null. **Example:** “The disability is caused by intentional self-harm of the insured.” (If the clause does not mention it, set to null)

Element Existence Rules

- **Mandatory:** claim, data, warrant (must not be null).
- **Optional:** backing (set to an empty list [] instead of null), qualifier, rebuttal. If not mentioned in the clause, these must be set to null.
- **No Inference:** All content must be strictly based on the original text. Do not introduce external knowledge, common sense, or assumptions.

Argumentation Splitting Rules

When to MERGE (Single Argumentation): Multiple conditions connected by “and,” “as well as,” “simultaneously satisfy,” “both,” etc. All conditions together form an indivisible prerequisite.

Example: “In case A and case B, the insurance benefit shall be paid” → One argumentation. / “Simultaneously satisfy: (1) A; (2) B” → One argumentation.

When to SPLIT (Multiple Argumentations): Use of selective connectors like “or,” “any of the following,” “one of the circumstances.” Different paths are logically independent in terms of **Time, Cause, Result, or Payout Standards**. **Even if the benefit name is the same, if the triggers are connected by “or,” they must be split.**

Example: “Non-accident diagnosis within 180 days → Refund premium; Non-accident or accident diagnosis after 180 days → Pay sum insured”

Should be split into three argumentations: A: Non-accident within 180 days; B: Due to accident (regardless of time); C: Non-accident after 180 days.

Argumentation Identifier: Format: [Original Clause ID]_[Uppercase Letter]. Letters start from A and increment logically (A, B, C). Example: “2.3.1_A”, “2.3.1_B”.

Output Format: Perform a splitting analysis first inside <think></think> tags. Output the result as a **JSON list**. Each element is an argumentation object. The entire JSON list must be wrapped in <json></json> tags.

Below is the clause you need to process. Strictly follow the above requirements.

{policy}

Figure 10: Prompt template for instantiating the Toulmin Model from **Coverage** clauses.

Role Positioning

You are an Insurance Policy Analysis Expert, specialized in constructing the Toulmin Model of Argumentation using insurance policy clauses.

Core Task

Extract non-payment scenarios from insurance exclusion clauses. For each trigger that leads to a non-payment conclusion, construct an argumentation based on the Toulmin Model.

Toulmin Model Element Specifications

1. Claim – Content: The non-payment conclusion. **Format:** Must use phrases such as “Shall not pay/shall not assume/shall not give [Specific Name of Insurance Benefit]” or “Contract canceled/Contract rescinded.” **Requirement: Must not contain numerical values** such as amounts or percentages; such information should be placed in the Qualifier. **Example:** “Shall not assume the liability to pay insurance benefits”

2. Data – Content: The specific conditions that trigger the Claim. **Requirement:** Extract objective facts directly from the clause. Maintain the semantic integrity of the original text. **Do not add interpretations, summaries, or inferences.** **Example:** “The policyholder intentionally fails to fulfill the obligation of truthful disclosure regarding an insurance accident occurring before the rescission of the contract”

3. Warrant – Content: The **complete original text of the relevant clause** from the insurance policy. **Requirement:** Must be the original text of the clause or a lossless transcription. **Splicing, rewriting, or deletions are strictly prohibited.**

4. Backing – Content: Contractual definitions of key terms appearing in the Data or Warrant. **Requirement:** Fill this only if the clause **explicitly cites a definition location** (e.g., “see 8.3”); otherwise, set to null. **Format:** A list of strings, each being a definition index. **Example:** [“8.3”]

5. Qualifier – Content: All specific limiting conditions regarding “how it is not paid.” **Requirement:** Must come from the original text; if there is no relevant content, set to null. **Example:** null

6. Rebuttal – Content: Explicit exception scenarios that would invalidate the Claim. **Requirement:** Fill this only if the clause **explicitly lists exception cases**; otherwise, set to null. **Example:** “If the insurer knew about the policyholder’s failure to disclose truthfully at the time the contract was concluded, the insurer may not rescind the contract”

Element Existence Rules

- **Mandatory:** claim, data, warrant (must not be null).
- **Optional:** backing, qualifier, rebuttal. If not mentioned in the clause, these must be set to null.
- **No Inference:** All content must be strictly based on the original text. Do not introduce external knowledge, common sense, or assumptions.

Argumentation Identifier: Must be identical to the original clause ID.

Output Format: The result must be output as a **JSON object** containing the Toulmin Model elements. The JSON must be wrapped between <json></json> tags.

Figure 11: Prompt template for instantiating the Toulmin Model from **Exclusion** clauses.

I. Core Task

Decompose natural language text that expresses a **logical AND relationship**—i.e., where all conditions must be satisfied simultaneously—into a set of independent, atomic propositions. The conjunction of these propositions via logical AND must be **logically equivalent** to the original clause.

II. Core Decomposition Rules

Parallel Structure: For “a and b and c”, decompose into three propositions: a, b, c. **Compound Structure:** For “(a or b) and c”, treat “(a or b)” as one proposition and c as another. **Integrity:** If a segment cannot be split without altering its meaning, keep it as a single proposition. **Logical Equivalence:** The reconstructed meaning using only AND must exactly match the original—no loss, no addition.

III. Application Scenarios & Constraint Types

Primarily applied to **Insurance Liability Clauses** to extract all payment conditions that must hold *concurrently*. Common constraint categories include:

Subject Eligibility (e.g., “The Insured”); **Time Window** (e.g., “Within 180 days”); **Causality** (e.g., “Caused by accidental injury”); **Activity/Status Scope** (e.g., “While participating in Sports”); **Event Nature** (e.g., “Suffered accidental injury”); **Outcome Standard** (e.g., “Reached disability assessment standards”); **Exclusions** (Convert to positive form only if explicitly stated); **Procedural Obligations** (e.g., “Promptly notify”); **Quantitative Standards** (e.g., “Expenses exceed deductible”); **Location Restrictions** (e.g., “At a designated hospital”).

IV. Terminology Processing & Expression Standards

Terminology Handling: If the input provides a list of defined terms (e.g., [“Sports”, “Accidental Injury”]), and a term appears in the clause, represent it as an **independent proposition using the exact term title**.

Expression Standards: **Term-based:** Use term title directly. **Non-term:** Refine into concise “Subject + Condition” phrases (e.g., “within 180 days” → “Disability occurred within 180 days”). **Explicit Logic:** Expand pronouns (e.g., “said injury”).

Core Principle: Strictly adhere to the original text.

V. Input & Output Format

Input Format (JSON): { "Policy": "Clause text", "Definitions": ["Term 1", ...] }

Output Process: First, briefly outline reasoning inside <think>...</think>. Then, output result as JSON wrapped in <json>...</json>: { "propositions": ["Proposition 1", "Proposition 2", ...] }

Below is the input you need to process:

```
{ "Policy": {policy}, "Definitions": {definition} }
```

Figure 12: Prompt template for **Conjunctive Decomposition** (Logical AND).

Logic Condition Analysis Expert Task Description (Logical OR)

I. Core Task

Identify and extract all **logical OR relationships** from natural language text—i.e., scenarios where satisfying *any one* of several conditions is sufficient. Decompose each alternative into an independent, self-contained proposition. The disjunction (OR) of these propositions must be **logically equivalent** to the relevant portion of the original clause.

II. Core Decomposition Rules

Explicit Logical OR: For phrases like “a or b”, “either a or b”, decompose into two propositions: a and b. **Implicit Parallel Enumeration:** When items are listed in parallel (e.g., separated by commas) and the context implies mutual exclusivity or sufficiency, treat each item as an independent proposition. **Mixed Structure:** For compound forms such as “(a and b) or c”, preserve internal conjunctions; decompose into “(a and b)” and “c”. **Integrity:** If no clear logical OR relationship exists, retain as a single atomic proposition.

III. Application Scenarios & Constraint Types

Primarily applied to insurance clauses involving **alternative or selectable conditions**. Common enumeration patterns include:

Activity/Behavior (e.g., “Diving, skydiving”); **Cause/Event** (e.g., “War, riots”); **Status/Situation** (e.g., “Drunkenness, drug use”); **Location/Place** (e.g., “Overseas, war zones”); **Result/Consequence** (e.g., “Death, disability”); **Timepoint/Period** (e.g., “Before inception, during waiting period”).

IV. Terminology Processing & Expression Standards

Terminology Handling: The input provides defined terms [“Term 1”, ...]. If a defined term appears and *itself contains an internal logical OR* (per definition), decompose it. If it functions as a single alternative, represent it as one proposition using its exact title.

Expression Standards: Each proposition must be a complete, coherent statement. Prefer “Subject + Condition” structure (e.g., “The Insured engages in skydiving”). For bare enumerated items (e.g., “rock climbing”), reconstruct full propositions by inheriting context. **Core Principle:** Strictly adhere to source text; accurately identify OR relationships without adding, omitting, or inferring.

V. Input & Output Format

Input Format (JSON): { "Policy": "Clause text", "Definitions": ["Term 1", "Term 2"] }

Output Process: First, briefly describe decomposition logic inside <think>...</think>. Then, output result as JSON wrapped in <json>...</json>: { "propositions": ["Proposition 1", "Proposition 2", ...] }

Below is the input you need to process:

```
{ "Policy": {policy}, "Definitions": {definition} }
```

Figure 13: Prompt template for **Disjunctive Decomposition** (Logical OR).

Task Objective

You are an **Insurance Claims Logic Architect**. Given a Claim and a list of Propositions (conditions with truth values), generate a **natural, fluent user inquiry** simulating a policyholder asking about eligibility. The inquiry must reflect factual scenarios implied by truth values, avoid policy jargon, and sound authentic.

Input Structure

Claim: The benefit sought. **Propositions:** A list of objects containing: { "proposition": "...", "truth": true|false, "definition": "Optional definition" }.

Generation Rules**1. Generate Factual Narrative Fragments**

Case A (Definition provided): If true: Describe a scenario that **implicitly satisfies** the definition (e.g., "twisted ankle during cool-down"). If false: Describe a scenario that **implicitly fails** (e.g., "sprained wrist walking dog before match"). Never quote the definition.

Case B (Definition is null): If true: Provide concrete details fulfilling the proposition (dates, actions). If false: Provide details violating it.

Key Principle: Never state *why*—only present facts as a layperson.

2. Integrate into User Inquiry

Perspective: 1st or 3rd person. **Tone:** Natural, informal. **Prohibited:** No clause mentions, logical connectors ("therefore", "since"), or jargon. **Closing:** Must end with "Can I/they receive the [Claim]?"

Output Format: Optionally include reasoning in <think>. Final output must be JSON wrapped in <json>:
<json> { "question": "Natural narrative... + Closing question?" } </json>

Reference Example

Input: { "Claim": "Accidental Injury", "Propositions": [{ "prop": "Sports period", "truth": false, "def": "..."}, { "prop": ">180 days", "truth": true, "def": null }] }

Output: <json> { "question": "My brother bought insurance... Can he receive the Accidental Injury Benefit?" } </json>

Now, please generate the answer based on the following input:

Claim: {Claim}; Propositions: {Propositions}

Figure 14: Prompt template for **Text Realization** (Question Generation).

Task Objective

You will receive a **real-world insurance claim inquiry** and a **proposition** (with optional definition). Your task is to **exclusively evaluate this proposition**. Based strictly on the inquiry's facts and the provided definition, determine if the proposition holds true. Output a truth value (true/false) with a concise explanation. **Note:** Do not judge the overall claim; focus solely on this single proposition.

Input Format

question: Natural language inquiry (e.g., "My friend fell...").

proposition_json: { "proposition": "Text...", "definition": "Explicit definition or null" }

Judgment Rules

1. With Definition: Interpret strictly within defined boundaries. Facts must clearly fall inside/outside. **Prohibit speculation.** If vague or insufficient → false.

2. Without Definition (null): Judge based on **literal meaning** and common sense. Inquiry must contain clear facts. If insufficient → false.

Prohibited: No external knowledge (laws/practices). No assuming unmentioned details.

Output Format: Optional reasoning in <think>. Final JSON in <json>:

```
<json> { "explain": "Reasoning...", "truth": true|false } </json>
```

Reference Examples

Ex 1 (Matches Def):

In: Q: "Sprained ankle playing badminton after yoga." Prop: "Injured during sports period" (Def: "Includes warm-ups...")

Out: <json> { "explain": "Stretching is warm-up, injury occurred during process.", "truth": true } </json>

Ex 2 (Fails Def):

In: Q: "Fell on way to basketball court." Prop: "Injured during sports period" (Def: "...excluding travel to venue.")

Out: <json> { "explain": "Travel to venue is explicitly excluded.", "truth": false } </json>

Below is the input you need to process:

question: {question}

proposition_json: {proposition}

Figure 15: Prompt template for **Answer Generation**.

Task Setting and Roleplay

You are a **Professional Insurance Claims Audit Expert**. Your task is to conduct a **fair, rigorous, and clause-bound** analysis based solely on the provided [Original Policy Clauses] and [Claims Inquiry].

Core Rules: Base judgment **exclusively** on explicit clause content; **Do not introduce** external knowledge or assumptions; Demonstrate a **complete logical chain** from text to conclusion.

Input Data

[Original Policy Clauses]: Read carefully; this is the **sole basis** for judgment.

{policy}

[Claims Inquiry]: Judge strictly according to clauses.

{question}

Response Requirements (Strictly follow this 3-part structure)

1. Relevant Clauses: Precisely quote the **original text** of all clauses **directly relevant** (payout conditions, exclusions, definitions). **Do not paraphrase**; use verbatim excerpts only.

2. Explanation and Analysis: Analyze how factual elements in the inquiry **do or do not satisfy** the cited clauses. Articulate the **reasoning process** (clause → fact → inference). Maintain objective, contractual reasoning.

3. Conclusion: State final judgment explicitly as Payable or Not Payable.

Output Format: Response must be a JSON object wrapped in <json>...</json>:

```
<json> { "relevant_policy": ["policy1 details,..."], "analysis": "Logical linking...", "conclusion": "Payable/Not Payable + details" } </json>
```

Model Response Commencement: Output strictly in accordance with the above requirements.

Figure 16: Prompt template for **Model Inference**. This prompt is applied to all evaluated LLMs to generate adjudication decisions and reasoning traces.

Role: You are a rigorous **Insurance Clause Matching Evaluator**. Please determine which items in the **gold_list** and **pred_list** correspond to the same clause content.

Input Data
[gold_list]: Standard-answer clause list: {gold_list}
[pred_list]: Model-output clause list: {pred_list}

Matching Requirements

Step 1: Compare each item in gold_list with items in pred_list and judge whether they refer to the **same clause content**.

Step 2: A match does **not** require identical wording or the same granularity. Full clauses and fragments may match; fragments may also match each other, as long as the **core meaning** is the same.

Step 3: If a gold_item matches any pred_item, record its index in covered_gold_indices. If a pred_item matches any gold_item, record its index in supported_pred_indices.

Step 4: Focus on the **core content**: insurance liability clauses = payout condition + benefit content; exclusion clauses = exclusion condition; definition clauses = term + definition.

Step 5: Do **not** count as a match if: only a title/label is given without substantive text; the topic is similar but the core content does not align; or it is unclear whether the two items express the same clause meaning.

Output Format: Output **JSON only**, wrapped in <json>:
<json> { "covered_gold_indices": [], "supported_pred_indices": [] } </json>

Requirements: Indices must start from 0, be sorted in ascending order, contain no duplicates, and no extra text should be output.

Below is the input you need to process:
gold_list: {gold_list}
pred_list: {pred_list}

Figure 17: Prompt template for evaluating Citation Metrics.

Role: Insurance Claims Evaluation Expert (Semantic Coverage Comparison)

Task: Strictly evaluate how many explanation points in the **Reference Answer** are **covered** by the **Model Answer** via semantic comparison.

Input Data
Model Answer: {HA_explain}
Reference Answer: A list of independent explanation points. {SA_explain}

Evaluation Rules

- Point-by-Point:** Semantically compare Model Answer against each Reference point.
- Criteria:** A point is **covered** if mentioned, explained, or clearly implied (paraphrasing is acceptable; verbatim matching not required).
- Counting:** Count the total number of covered explanation points.

Output Format: Output a pure JSON object (no extra text):
{ "coverage": (integer count) }

Example
Model Answer: "...effective date Jan 1, diagnosis Feb 15 (46 days < 90)... invasive pulmonary adenocarcinoma meets Cancer definition... 'first time' meets Initial Diagnosis..."
Reference Answer: ["Contract Effective Date & Diagnosis Date (<90 days)", "Grade III Class A hospital requirement", "First time having illness (Initial Diagnosis)", "Pathological examination confirms Cancer"]
Output: { "coverage": 3 }
Explanation: 1. Time interval (46 days < 90) → Covered. 2. Hospital Grade info → Not mentioned → Not Covered. 3. First time diagnosis → Covered. 4. Cancer definition → Covered.

Figure 18: Prompt template for evaluating Explanation Coverage.

Role: You are an **extremely rigorous Insurance Claims Evaluation Expert**.

Task: Strictly evaluate consistency of the **Core Conclusion** (Payable vs. Not Payable) between the Answer and Standard Answer via **semantic comparison**.

Evaluation Rules

Consistent (1): The core conclusion in the Answer is **completely consistent** with the Standard Answer.

Inconsistent/Undetermined (0): Conclusions are contradictory, **or** the Answer **fails to explicitly state** a clear conclusion.

Input Data

Answer: {ha_answer}

Standard Answer: {sa_answer}

Output: Output **only** the following JSON object (no extra text):
 { "answer": 0 or 1 }

Figure 19: Prompt template for evaluating Adjudication Accuracy.

Category	Examples	Count
Disease Insurance (Critical Illness & Cancer)	1. PICC Life Renrenbao 2.0 Critical Illness Insurance (Type B) 2. Taikang Additional Child Leukemia Disease Insurance Clause	24
Life Insurance (Term, Whole, Endowment)	1. PICC Life Yisheng Whole Life Insurance Clause 2. Taikang Jixiang Xiangban (2018) Term Life Insurance	24
Medical Insurance	1. Individual Tax-Preferential Health Insurance 2. Taikang Group Critical Illness Medical Insurance for Residents	21
Annuity & Pension Insurance	1. Life Individual Tax-Deferred Pension Annuity Insurance 2. United Jubaopen Annuity Insurance	16
Accident Injury Insurance	1. Airline Passenger Accident Injury Insurance 2. United Borrower Accident Injury Insurance Clause	11
Nursing Care Insurance	1. Changqing Group Long-term Care Insurance Clause 2. Taikang Additional Zunxiang Hehu Long-term Care Insurance	8
Disability & Income Loss Insurance	1. PICC Life Pilot Group Disability Income Loss Insurance 2. Life Anxiang Pilot Disability Income Loss Insurance	6
Total		110

Table 7: Detailed classification statistics and examples of the 110 insurance product documents used in the dataset.

Category	Content
Insurance policy	<p>(1) Coverage Clause 2.3.1: If the Insured sustains an accidental injury and dies as a result of such accidental injury within 180 days from the date of the accident, or dies due to causes other than accidental injury within the waiting period stipulated in this Supplementary Contract, we will pay a Death Benefit equal to 1.6 times the premium paid for this Supplementary Contract (without interest), and this Supplementary Contract shall terminate.</p> <p>(2) Exclusion Clause 2.4(5): We shall not be liable to pay insurance benefits if the death or total disability of the Insured is caused by: (5) The Insured drives under the influence of alcohol, drives without a legal and valid driving license, or drives a motor vehicle without a valid vehicle license.</p> <p>(3) Definition — Accidental Injury: Refers to an objective event that is external, sudden, unintentional, and non-disease related, causing physical injury to the Insured.</p>
Toulmin model	<p>(1) Claim ID: 2.3.1_B</p> <ul style="list-style-type: none"> • Claim: Payment of Death Benefit • Data: The Insured sustained an accidental injury and died as a result within 180 days. • Warrant: If the Insured sustains an accidental injury and dies as a... (Coverage Clause 2.3.1). • Backing: Accidental Injury: Refers to an ... (Definition — Accidental Injury) • Qualifier: Pay 1.6 times premiums (no interest) • Rebuttal: Exclusion triggered <p>(2) Claim ID: 2.4.5</p> <ul style="list-style-type: none"> • Claim: Not liable to pay insurance benefits • Data: The Insured drives under the influence of alcohol, drives without a legal and valid driving license, or drives a motor vehicle without a valid vehicle license • Warrant: We shall not be liable to pay ... (Exclusion Clause 2.4(5)) • Backing: — • Qualifier: — • Rebuttal: —
Proposition Truth	<p>(1) The Insured sustained an accidental injury resulting in death → True</p> <p>(2) The death occurred within 180 days from the date of the accident → True</p> <p>(3) Driving without a legal and valid driving license → True</p>
QA-Pair	<p>Question: I would like to inquire about my father's insurance claim. While driving to work, he swerved sharply to avoid a pedestrian who suddenly ran across the road, causing the car to lose control and crash into a guardrail. My father suffered severe head injuries and, after 45 days of emergency treatment in the hospital, he unfortunately passed away. The traffic police investigation concluded that my father was not at fault, and the vehicle's safety inspection was up to standard. However, there is one issue: the vehicle license had expired one month prior to the accident. I would like to ask if this will affect the claim settlement?</p> <p>Citations:</p> <p>[1] If the Insured sustains an accidental injury and dies as a result of such ... (Coverage Clause 2.3.1).</p> <p>[2] We shall not be liable to pay insurance benefits ... (Exclusion Clause 2.4(5)).</p> <p>[3] Accidental Injury: Refers to an objective ... (Definition) to the Insured.</p> <p>Explanation:</p> <p>Proposition:</p> <p>[1] The user describes swerving sharply to avoid a pedestrian while driving to work, causing the vehicle to lose control and hit a guardrail, which satisfies the condition of sustaining an 'accidental injury'.</p> <p>[2] The user states that the father passed away after 45 days of emergency treatment due to the accidental injury, which is within 180 days of the accident, satisfying the condition of 'death caused by such accidental injury within 180 days from the date of the accident'.</p> <p>[3] The user explicitly mentions that the vehicle license had expired one month before the accident, meaning the license was invalid. This satisfies the trigger condition for the exclusion clause: 'driving a motor vehicle without a valid vehicle license'.</p> <p>Clause: Implying the license was invalid, which satisfies the trigger condition of the exclusion clause: 'driving a motor vehicle without a valid vehicle license.'</p> <p>Global: Satisfying the trigger condition for the Exclusion Clause.</p> <p>Conclusion: Claim Denied</p>

Table 8: A complete data sample from InsLogicBench. The table illustrates the mapping from the **Policy** to the **Toulmin Model**, followed by the **Proposition Truth Assignment**, and finally resulting in the generated **Question-Answer Pair** with a complete reasoning trace.

Category	Label	Description & Rubric
Correct	Supported	The conclusion (Approved/Denied) is factual and fully substantiated by explicit, unambiguous policy clauses. The logical reasoning is closed-loop and indisputable.
Insufficient	Ambiguous	The provided policy text is insufficient to derive the current conclusion (e.g., missing key definitions or exclusion clauses), or the policy itself contains ambiguity that prevents a unique verdict. <i>Note: These samples are typically treated as noise.</i>
Incorrect	Contradicted	The conclusion directly conflicts with the policy text (e.g., the policy explicitly states "not covered," but the model concludes "Approved"), or the decision relies on completely irrelevant clauses.

Table 9: Evaluation rubric for **Label Validity & Supportability**. This dimension assesses whether the verdict is factually consistent with the provided evidence.

Score	Rating	Detailed Rubric
5	Flawless	Precise citation, tight logic. The model cites all necessary and <i>only</i> necessary clauses (no omission, no redundancy). The explanation accurately maps the user's claim to the policy conditions. The conclusion is a logical necessity derived from the premises.
4	Good	Accurate citation, generally smooth logic. Core clauses are cited correctly, though minor auxiliary clauses might be missing (without affecting the verdict). The explanation is logically clear but may be slightly verbose or insufficiently concise. The overall deduction chain holds.
3	Acceptable	Relevant citation, minor logical flaws. Citations are relevant but may include irrelevant clauses (noise). The explanation may contain slight logical leaps (relying on common sense rather than strict derivation). The conclusion is correct, but the derivation process is not fully rigorous.
2	Weak	Biased citation, messy logic. The model cites incorrect clauses or misses critical exclusions. The explanation fails to establish a clear link between the claim and the policy terms. There are signs of "forced reasoning" to justify the output.
1	Hallucinated	Fabricated facts, complete failure. The model invents non-existent policy content. The explanation is unrelated to the cited clauses, or the logical chain is completely broken.

Table 10: Evaluation rubric for **Reasoning Faithfulness**. This dimension evaluates the precision of evidence retrieval and the logical validity of the Chain of Thought.

Score	Rating	Detailed Rubric
5	Professional	Expert-level fluency. The text flows smoothly with precise terminology. The style perfectly matches professional insurance adjudication practices (e.g., using terms like "falls within the scope of liability," "subject to the exclusions"). The tone is objective, calm, and authoritative.
4	Fluent	Natural but slightly colloquial. Grammar is correct and logic is coherent. However, some vocabulary choices are insufficiently formal or lack "insurance flavor" (e.g., using "won't pay money" instead of "liability denied"). The overall reading experience is good.
3	Average	Correct but stiff. No grammatical errors, but the sentence structure is monotonous or robotic. There may be slight repetition or verbosity. While understandable, it is clearly identifiable as machine-generated text.
2	Poor	Grammatical errors or obscure phrasing. The text contains obvious grammatical mistakes or unclear references. Sentences may be overly long and convoluted, making them difficult to parse. Terminology usage is incorrect.
1	Unreadable	Broken logic, gibberish. Severe grammatical errors, incoherent sentences, or excessive repetition of fragments. The text cannot be read effectively.

Table 11: Evaluation rubric for **Linguistic Naturalness**. This dimension assesses the grammatical correctness and the professional tone of the generated text.

<p>question: Hello, I would like to inquire about an insurance claim for my father. On the afternoon of March 15, 2024, while on his way home, he was knocked down by an electric bike. He suffered a head injury and a fracture in his left leg. He was conscious immediately after the accident and had not consumed alcohol. A kind passerby called an ambulance, which took him to a nearby Community Health Service Center for emergency bleeding control and initial fixation of the leg. However, because his injuries were severe and his condition unstable, the center transferred him to the City No. 2 People's Hospital that same night. After an examination in the Emergency Department, the doctor diagnosed him with a traumatic brain injury and hemorrhagic shock. He was admitted to a formal ward in Neurosurgery but was subsequently transferred immediately to the ICU, as they stated he required 24-hour monitoring and treatment. He stayed in the ICU for a total of 5 days, utilizing a ventilator and ECG monitoring, with doctors and nurses constantly attending to him for resuscitation and treatment. Additionally, I would like to confirm something: at the time of the accident, our supplementary insurance contract had already been in effect for more than 180 days, and the City No. 2 People's Hospital is a public Class II hospital. These factors should meet the claim conditions, correct?</p>
<p>relevant clause: 1. If the Insured receives inpatient treatment in a Hospital stipulated in this Supplementary Insurance Contract and is admitted to an Intensive Care 2. Accidental Injury: Refers to bodily injury directly caused by an objective event that is external, sudden, unintentional, and non-disease-related. 3. Hospital stipulated in this Supplementary Insurance Contract: Refers to the general department of a public hospital classified as Class II or higher by ... 4. Inpatient Treatment: Refers to the Insured being admitted to a formal ward of a hospital to receive treatment due to accidental injury or disease, 5. Intensive Care Unit: Refers to a formal intensive care unit established within a hospital as approved by the health administrative department.</p>
<p>Explain Proposition: 1.The user described being 'knocked down by an electric bike,' and this event meets the definition of Accidental Injury. 2. The user received treatment at 'City No. 2 People's Hospital,' confirmed as a 'public Class II hospital,' which meets the hospital tier and 3.The user was 'admitted to a formal ward in Neurosurgery,' satisfying the requirement for Inpatient Treatment in the contract. 4.The user was 'transferred to the ICU,' 'stayed for 5 days,' and 'used a ventilator and ECG monitoring equipment,' satisfying the requirements for Clause: User description... all essential elements for claim conditions are clearly and completely met. Global: Payout, calculated as the actual number of days of hospitalization in the Intensive Care Unit..., Conclusion: Approved</p>
<p>DeepSeek Answer</p>
<p>Citations 2.3 Insurance Liability: If the Insured receives inpatient treatment in a Hospital stipulated in this Supplementary Insurance Contract and is ... 9.5 Accidental Injury: Refers to bodily injury directly caused by an objective event that is external, sudden, unintentional, and non-disease-related. 9.6 Hospital stipulated in this Supplementary Insurance Contract: Refers to the general department of a public hospital classified as Class ... 9.7 Inpatient Treatment: Refers to the Insured being admitted to a formal ward of a hospital to receive treatment ... 9.11 Intensive Care Unit: Refers to a formal intensive care unit established within a hospital as approved by the health administrative department ... 8.3 Emergency Treatment for Accidental Injury: Emergency treatment for accidental injury is not subject to the restriction of ...</p>
<p>Analysis Based on the claim description and cited clauses, logical reasoning is performed as follows:1. **Nature of the incident meets the definition of Accidental Injury (9.5)**: The Insured was knocked down by an electric bike, causing a . . . - within the 90-day limit. The annual cumulative days (assuming not exceeded) and the sum insured (300,000 RMB per policy) are within limits. In summary, the facts of the claim completely meet the conditions for the payment of the Intensive Care Insurance Benefit, with no discrepancies.</p>
<p>Conclusion: Payout. Specific payout details:- **Payment Type**: - The cumulative annual payment days must ensure not to exceed 180 days (assuming no other claims in this policy year).- **Aggregate Limit**: Cumulative insurance benefits paid (including this one) shall not exceed the Sum Insured; the contract terminates upon reaching this limit.</p>

Figure 20: **Successful Adjudication.** The model correctly retrieves all relevant clauses, validates the conditions, and derives the correct approval verdict with a complete reasoning trace.

<p>question: I would like to inquire about a claim for my father. On May 10, 2023, while at work, he accidentally got chemicals splashed into his eyes, causing severe injuries. The hospital diagnosed him with severe corneal burns in both eyes. We handled it immediately and notified the insurance company via phone and in writing within 3 days of the accident. His insurance became effective on January 1, 2023. We truthfully disclosed his health and work situation when applying, and his declared age was 35, which should fit your required age range of 18 to 60. After 185 days of systematic treatment, his vision still has not recovered. He now relies entirely on others for daily life and cannot continue working. In the most recent examination, the best corrected visual acuity in both eyes was only 0.03, and the visual field radius was only 10 degrees. The ophthalmologist has already issued a certificate of permanent visual impairment. Can we apply for a claim in this situation?</p>
<p>relevant clause: 1.If the Insured suffers Total Disability, we shall pay the Total Disability Insurance Benefit according to the Basic Sum Insured ... 2. Permanent and Total Blindness in Both Eyes: Refers to best corrected visual acuity below 0.02 on the international standard visual acuity chart, or a visual field radius of less than 5 degrees." 3. Permanent and Total: Refers to the function remaining completely lost after one hundred and eighty days of treatment starting ... 4. Total Disability: Refers to the Insured, during the validity period of this Contract, resulting in any of the following circumstances due to accidental injury or disease: 2... best corrected visual acuity below 0.02 on the international standard visual acuity chart, or visual field radius less than 5 degrees, ... where recovery is obviously impossible, such as eyeball removal."</p>
<p>Explain Proposition: 1. The user's father suffered severe corneal burns in both eyes due to chemical splashes, clearly establishing the cause of injury." 2. After 185 days of treatment, the requirement for the 180-day treatment period in the clauses is met." 3. The corrected visual acuity of both eyes is 0.03, which is higher than the standard of 'below 0.02' defined in the contract for 'Permanent and Total Blindness in Both Eyes'." 4. The visual field radius is 10 degrees, which is greater than the standard of 'less than 5 degrees' defined in the contract for 'Permanent and Total Blindness in Both Eyes'." 5. Although there is severe visual impairment and a doctor's certificate, the ... thus failing the conditions for Total Disability payout." Clause: The user describes that ... eyes is 0.03 (higher than the 0.02 standard), and the visual field radius is 10 degrees .. Total Disability are not satisfied. Global: the specific numerical values do not meet the definition standards for 'Permanent and Total Blindness in Both Eyes' in the contract, thus failing the conditions for Total Disability payout., Conclusion: Denied</p>
<p>GLM Answer</p>
<p>Citations 2.3.1 Waiting Period: If the Insured dies or suffers Total Disability ... Insured dies or suffers Total Disability due to Accidental Injury. 2.3.3 Total Disability Insurance Benefit: If the Insured ... this Contract shall terminate. 10.4 Total Disability: Refers to the Insured, during the ... from the date of accidental injury or disease diagnosis... 10.5 Accidental Injury: Refers to an objective ... bodily injury that is external, sudden, unintentional, and non-disease-related. 3.3.1 Claim Application: When applying for ... the nature and cause of the insurance accident.</p>
<p>Analysis Based on the claim facts and clause analysis:1. ... the Insured's situation meets the definition of Total Disability under Item 8 of 10.4.... Compliant proof must be supplemented to complete the application. In summary, the facts meet the conditions for insurance liability, but documentation requirements must be satisfied.</p>
<p>Conclusion: Payout. The Insured suffered Total Disability due to Accidental Injury, which complies with the Total Disability Insurance Benefit liability in 2.3.3. ...</p>

Figure 21: **Failure due to Overlooked Condition.** Although the general coverage seems applicable, the model fails to verify a specific required condition (e.g., waiting period or specific definition), leading to an incorrect approval.

<p>question: I previously purchased accident and health insurance. After 180 days, I went to a Grade 3A hospital in our city because of persistent stomach pain and bleeding. The doctor diagnosed an ectopic pregnancy and suggested outpatient conservative treatment first, prescribing some medication. However, after taking the medicine for 15 days, the results were not good, and the pain actually worsened, so I was formally hospitalized for surgery on March 10. I was discharged on March 17, but on the 10th day after discharge, the surgical wound got slightly infected, so I went back to the outpatient department of the same hospital for a dressing change. This time, I incurred some special outpatient fees that I had to pay out of pocket. I want to ask, since these treatments were necessary requirements by the doctor and the costs are in line with local charging standards, can I apply for a claim?</p>
<p>relevant clause:</p> <ol style="list-style-type: none"> 1. General Medical Insurance Benefit: During the insurance period,... in this Contract. 2. Exclusions: We shall not assume liability ... caused thereby. 3. Outpatient/Emergency Treatment Before and After Hospitalization: Refers to... due to said outpatient/emergency treatment. 4. Hospital: The term Hospital in this Contract refers to hospitals designated by us... wards of the aforementioned hospitals. 5. Accidental Injury: Refers to an objective event causing bodily injury that is external, sudden, unintentional, and nondisease-related. 6. Onset of Disease: Refers to the appearance of precursors or abnormal physical conditions of... diagnosis, treatment, or nursing. 7. Hospitalization: Refers to the process where the Insured is admitted to a formal ward... diagnosis/treatment fees, bed fees, etc. 8. Reasonable and Necessary: Reasonable and Necessary refers to meeting the following... expert recognized by both parties.
<p>Explain</p> <p>Proposition:</p> <ol style="list-style-type: none"> 1. The Insured's onset of disease occurred after the waiting period, specifically more than 180... waiting period. 2. Treatment was conducted in a Grade 3A hospital, complying with the definition of a hospital in the insurance contract. 3. The user incurred special outpatient fees requiring out-of-pocket payment. 4. The outpatient treatment occurred on the 10th day after discharge, exceeding the time limit of within 7 days before and after hospitalization. 5. The user explicitly described seeking medical attention for an ectopic pregnancy, including outpatient conservative treatment, drug treatment, inpatient surgery, and postoperative infection management; these all fall under the category of 'pregnancy (including ectopic pregnancy)' <p>Clause:</p> <ol style="list-style-type: none"> 1. The user description indicates: ...However, the critical issue is: the user described outpatient treatment on the 10th day after discharge, which exceeds the time limit of 'within 7 days before and after hospitalization' defined in the policy terms, thus failing to meet the element requirements for 'Outpatient/Emergency Treatment Before and After Hospitalization'. 2. The user explicitly described seeking medical attention for an ectopic pregnancy, including outpatient conservative treatment, drug treatment, inpatient surgery, and postoperative infection management. These all fall under the category of 'pregnancy (including ectopic pregnancy) and complications caused thereby <p>Global: Failing to meet the element requirements for 'Outpatient/Emergency Treatment Before and After Hospitalization and exclusion remains in effect.</p> <p>Conclusion: Denied</p>
<p>DeepSeek Answer</p>
<p>Citations</p> <p>2.4 Exclusions: We shall not assume liability for paying insurance benefits for medical expense expenditures incurred by the Insured or for the occurrence of Specific Critical Illnesses of Children caused by any of the following circumstances: ...11. The Insured's pregnancy (including ectopic pregnancy), prenatal and postnatal checkups, miscarriage, abortion, childbirth (including dystocia), contraception, sterilization (including tubal ligation/vasectomy), infertility treatment, artificial insemination, and complications caused thereby;</p>
<p>Analysis</p> <p>Based on the claim description, the Insured received outpatient conservative treatment, inpatient surgery, and outpatient dressing changes for postoperative wound infection due to an ectopic pregnancy (explicitly included in 'pregnancy (including ectopic pregnancy)' ... Reasoning Chain: Ectopic pregnancy falls under Exclusions → All resulting medical expenses (including outpatient, inpatient, and complication treatment) do not meet claim conditions.</p>
<p>Conclusion:</p> <p>Not payable. All medical expenses, including conservative outpatient treatment, inpatient surgery, and postoperative outpatient dressing changes, were denied reimbursement because they fall within the scope of exclusions.</p>

Figure 22: **Incomplete Reasoning Trace.** While the final verdict is correct, the model fails to cite all necessary supporting clauses or definitions, resulting in low Explanation Coverage scores.

<p>question: I would like to inquire about a critical illness insurance claim for my mother. She bought this policy on January 1 2023, and truthfully disclosed all conditions at that time. Later, during a physical examination on May 15, 2023, an abnormality was found. After a pathological examination at the hospital, she was diagnosed with cervical carcinoma in situ and subsequently underwent conization surgery. Can we apply for a claim in this situation? Additionally, during the admission examination, she tested positive for HIV antibodies, but my mother currently has no related symptoms. Also, since she has been insured, she has never used drugs or injected drugs, nor has she committed suicide or crimes, not even drunk driving or driving without a license, let alone war or nuclear radiation. Our insurance contract has remained valid, we haven't received any termination notice, and there is no case of anyone intentionally harming her.</p>
<p>relevant clause:</p> <ol style="list-style-type: none"> 1. If the Insured is diagnosed for the first time with a Non-Life-Threatening Malignant Lesion stipulated in this Contract after 90 days from the effective date (or the date of the last reinstatement, whichever is later) of this Contract, the Company shall pay the Non-Life-Threatening Malignant Lesion Extra Benefit amounting to 30% of the Basic Sum Insured corresponding to the Insured. The payment of the Non-Life-Threatening Malignant Lesion Extra Benefit under this Contract is limited to one time; after payment, the Company's liability for this specific insurance benefit for the Insured terminates." 2. If the Insured suffers from a Malignant Tumor or Non-Life-Threatening Malignant Lesion stipulated in this Contract due to any of the following circumstances, the Company shall not assume liability for paying the Malignant Tumor Insurance Benefit or the Non-Life-Threatening Malignant Lesion Extra Benefit: (1) The Insured is infected with HIV or suffers from AIDS; 3. Non-Life-Threatening Malignant Lesion: Refers to the following malignant ... TNM staging of T1N0M0 or lighter staging. 4. Infection with HIV or Suffering from AIDS: HIV refers to ... it is considered Infection with HIV; ... Suffering from AIDS.
<p>Explain</p> <p>Proposition:</p> <ol style="list-style-type: none"> 1. The contract effective date is January 1, 2023. 2. An abnormality was found during a physical examination on May 15, 2023; this date is more than 90 days ..., satisfying the waiting period. 3. The diagnosed disease is cervical carcinoma in situ, which complies with 'Carcinoma in Situ' under the definition of 'Non-Life-Threatening Malignant Lesion' in the clauses. 4. The diagnosis was based on a pathological examination, satisfying the requirement for histopathological examination in the clauses. 5. The treatment method was conization surgery, indicating that corresponding treatment was received, which complies with the clause requirements. 6. The user description explicitly mentions that HIV antibodies were positive during the admission exam with no related symptoms. According to the terminology explanation, this fits the definition of 'Infection with HIV', satisfying the trigger condition for the Exclusion Clause. <p>Clause:</p> <ol style="list-style-type: none"> 1. The user description states 'bought this policy on ... This fully meets all elements for the basic payout conditions. 2. The user description explicitly mentions that HIV antibodies According to the terminology explanation, this fits the definition of 'Infection with HIV', ... <p>Global: Satisfying the trigger condition for the Exclusion Clause.</p> <p>Conclusion: Denied</p>
<p>DeepSeek Answer</p>
<p>Citations</p> <p>Insurance Liability (2.3): If the Insured is diagnosed for the first time with a ... benefit for the Insured terminates.</p> <p>Definition of Non-Life-Threatening Malignant Lesion (10.8): ... (5) Prostate cancer with TNM staging of T1N0M0 or lighter staging.</p> <p>Exclusions (2.5): If the Insured suffers from a Malignant Tumor or Non-Life-Threatening Malignant Lesion ... Insurance Benefit or the Non-Life-Threatening Malignant Lesion Extra Benefit: (1) The Insured is infected with HIV or suffers from AIDS; ... If the Insured is diagnosed with a Malignant Tumor or Non-Life-Threatening Malignant Lesion stipulated in this Contract caused by the aforementioned circumstances, the Company's insuranceliability for said Insured terminates, and the Company shall refund the corresponding cash value to the Applicant.</p> <p>Contract Effectiveness (1.2): This Contract becomes effective ... The effective date of the contract is specified in the insurance policy.</p> <p>Truthful Disclosure (8.1): The Applicant shall disclose truthfully... liability for paying the insurance benefit.</p>
<p>Analysis</p> <ol style="list-style-type: none"> 1. Insurance Liability (2.3): If the Insured is ...for the Insured terminates. 2. Definition of Non-Life-Threatening Malignant Lesion (10.8): ... Prostate cancer with TNM staging of T1N0M0 or lighter staging. 3. Exclusions (2.5): If the Insured suffers from a Malignant Tumor or Non-Life-Threatening Malignant ... (1) The Insured is infected with HIV or suffers from AIDS; Exclusions Not Applicable: Exclusion Clause 2.5(1) stipulates that the Company shall not assume liability for payment when a Non-Life-Threatening Malignant Lesion is caused by infection with HIV or suffering from AIDS. However, in this claim context: The status of being HIV antibody positive but asymptomatic classifies as 'Infection with HIV' (Definition 10.10), ... 4. Contract Effectiveness (1.2): This Contract becomes ... The effective date of the contract is specified in the insurance policy. 5. Truthful Disclosure (8.1): The Applicant shall disclose truthfully... the Company assumes the liability for paying the insurance benefit.
<p>Conclusion:</p> <p>Payout. The payout amount is 30% of the Insured's Basic Sum Insured (according to Clause 2.3). Simultaneously, the Non-Life-Threatening Malignant Lesion Extra Benefit is limited to one time (Clause 2.3); after this payment, this specific insurance liability terminates. The Basic Sum Insured is agreed upon at the time of application and specified in the insurance contract (Clause 2.1); the specific figure requires checking the policy</p>

Figure 23: **Failure in Exclusion Handling.** The model identifies the coverage clauses but overlooks an active exclusion clause that should have triggered a denial. This illustrates the "recall-bias" toward approval discussed in Section 5.3.

<p>question: I would like to inquire about a claim. My father accidentally slipped at home and hit his head on the corner of a table. He immediately felt his vision was blurry and had a severe headache. We rushed him to the Emergency Department of the City People's Hospital, which is a Grade 3A hospital. The doctor performed a CT scan, diagnosed a mild concussion, and suggested hospitalization for observation. During the hospitalization, an ophthalmologist came for a consultation and said the vision problem caused by the fall required laser surgery to correct. He underwent this vision correction surgery on the third day of hospitalization and was discharged after a total stay of 5 days. This accidental fall occurred after the insurance waiting period. Also, I checked the contract, and 'falling' is not in the exclusions. The City People's Hospital is also the grade of hospital stipulated in our contract. I want to ask, can we apply for a claim in this situation?</p>
<p>relevant clause:</p> <ol style="list-style-type: none"> 1.If the Insured receives inpatient treatment in a hospital classified as Class II ... Subsidy Insurance Amount'. 2.We shall not assume liability for paying insurance benefits ... (5) Surgery resulting from vision correction; 3.We shall not assume liability for paying ... (6) Plastic surgery, cosmetic surgery, beauty treatment, beauty surgery,... Accidental Injury; 4.Accidental Injury: Refers to bodily injury directly caused by ... and non-disease-related. 5.Other medical institutions recognized by us: Refers to designated ... institutions stipulated in the policy. 6.Hospitalization: Refers to the ... incurring only nursing fees, bed fees, etc. 7.Orthopedics, Orthopedic Surgery: Refers to restoring the human body to ... transplantation, or reconstruction and repair of body parts. 8. Vaccine: Refers to preventive biological products used for human vaccination to... of infectious diseases.
<p>Explain Proposition:</p> <ol style="list-style-type: none"> 1.The user accidentally slipped at home and hit his head on a table corner,... meets the definition of accidental injury. 2.The user was sent to the Emergency Department of the City People's Hospital,... meeting the requirement for a Class II or higher hospital. 3.The user was hospitalized for observation and stayed for 5 days, meeting the requirement for Inpatient Treatment. 4. The user description explicitly mentioned that This directly satisfies the trigger condition for Exclusion Clause (5) 'Surgery resulting from vision correction'. 5.The vision correction surgery the user's father needed due to the ... [Trigger Condition]. 6.According to the [Exception Circumstances] 'except for orthopedics, ... so this specific exclusion clause (Clause 6) does not take effect. <p>Clause: This directly satisfies the trigger condition for Exclusion Clause (5) Surgery resulting from vision correction</p> <p>Global: Satisfying the trigger condition for the Exclusion Clause.</p> <p>Conclusion: Denied</p>
<p>DeepSeek Answer</p>
<p>Citations</p> <p>2.5.1 Accidental Hospitalization Subsidy Insurance Benefit: ... Accidental Hospitalization Subsidy Amount.</p> <p>8.1 Accidental Injury: Refers to bodily injury directly caused by an objective event that is external, sudden, unintentional, and non-disease-related.</p> <p>8.3 Hospitalization: Refers to the process where the ... including cases with no treatment for several consecutive days during hospitalization, incurring only nursing fees, bed fees, etc.</p> <p>2.6 Exclusions: (5) Surgery resulting from vision correction; (6) Plastic surgery, cosmetic surgery, beauty treatment, beauty surgery, orthopedics, orthopedic surgery, plastic surgery (surgical), ... caused by Accidental Injury.</p>
<p>Analysis</p> <p>According to Clause 8.1, ..In summary, although vision correction surgery is an excluded item, this surgery was an auxiliary treatment performed during hospitalization caused by an Accidental Injury, and the hospitalization itself meets the payout conditions of Clause 2.5.1. Therefore, this hospitalization qualifies for the Accidental Hospitalization Subsidy Insurance Benefit.</p>
<p>Conclusion:</p> <p>Approved (Can apply for claim). The insurance company shall pay the Accidental Hospitalization Subsidy Insurance Benefit calculated as the actual number of hospitalization days (5 days) multiplied by the Daily Accidental Hospitalization Subsidy Amount stipulated in the contract. The payout amount is limited to the 'Accidental Hospitalization Subsidy Sum Insured', and the number of days for a single hospitalization shall not exceed the maximum days for single hospitalization subsidy stipulated in the contract.</p>

Figure 24: **Failure in Nested Logic.** The case involves a complex interaction between an exclusion and its exception. The model fails to correctly parse this nested structure, ignoring the specific conditions that validate (or invalidate) the exclusion.