Mitigating Misinterpretation in Policy Documents Through Automated Language Understanding

Momojit Biswas, Anka Chandrahas Tummepalli, Preethu Rose Anish
TCS Research, India
{momojit.biswas, ankachandrahas.t, preethu.rose}@tcs.com

Abstract

Policy documents often employ intricate and technical language, posing comprehension challenges for policyholders and increasing the risk of misinterpretation, financial losses, and legal disputes. To address these issues, we propose an automated framework leveraging Retrieval-Augmented Generation to identify and clarify potentially mis-interpretable paragraphs within policy documents. The framework consists of two key modules: the Annotation module and the Rectification module. The Annotation module employs both paragraph-level and document-level contextual reasoning to classify paragraphs into categories indicative of potential misinterpretation. The Rectification module resolves these ambiguities by generating targeted interpretation queries, retrieving relevant document-level context, and incorporating external knowledge sources. Applied to a corpus of 240 real-world policy documents, the Annotation module produced a benchmark dataset comprising 11,000 annotated paragraphs, enabling systematic evaluation of interpretability issues. We assessed the dataset's quality through expert-driven manual reviews and large-scale automated evaluations using fine-tuned Pretrained Language Model. For the Rectification module, we evaluated five open-source Large Language Models: Mistral-2-7B, Mistral-3-7B, LLaMA-2-7B, LLaMA-3-8B, and Saul-7B. Among these, Mistral-2-7B achieved the highest human evaluation scores: 0.912 for *Clarity*, 0.914 for *Fidelity*, and 0.934 for *Usefulness*. This work demonstrates the practical feasibility of utilizing automated frameworks to enhance the clarity and comprehensibility of complex policy documents, thereby mitigating risks associated with misinterpretation and its adverse consequences.

Keywords: RAG, Interpretation Queries (IQs), Policy Clarification, Misinterpretation Detection, Policy Rectification

1. Introduction

Policy documents across sectors such as insurance, employment, and finance often employ complex legal language, rendering them challenging for policy readers to understand (Senninger, 2023; Derguech et al., 2018). This complexity, while ensuring legal precision, frequently leads to misinterpretations that can result in financial loss, legal disputes, and diminished trust between policyholders and providers (Sumit Arora, 2024; Derguech et al., 2018; Han et al., 2024; French, 2013).

Although previous studies (Bhatia et al., 2016; Liu et al., 2016; Kotal et al., 2020; Bannihatti Kumar et al., 2020) have explored issues such as vagueness, ambiguity, and accessibility in policy documents, the focus has largely remained on identifying these problems rather than resolving them (further details regarding the previous studies are provided in Sec 2). As a result, there remains a critical gap in approaches that not only identify mis-interpretable content but also improve its clarity through rewriting. To address this gap, we propose a Retrieval-Augmented Generation (RAG)-based framework that automatically identifies potentially mis-interpretable paragraphs in policy documents and rectifies them to enhance comprehension for policy readers.

Our framework comprises two core modules: (1) Annotation and (2) Rectification. The Annotation module leverages open-source Large Language Models (LLMs) to classify each paragraph within a policy document into predefined misinterpretability categories. Paragraphs that do not fall into any of these categories are deemed interpretable and excluded from further processing. To enhance classification accuracy, the Annotation module incorporates both paragraph-level and document-level contextual reasoning. The output is a curated set of potentially mis-interpretable paragraphs, each accompanied by explanatory reasoning to promote transparency and interpretability.

In the Rectification module, targeted Interpretation Queries (IQs) are generated for each potentially mis-interpretable paragraph. These IQs are designed to elucidate the underlying causes of misinterpretation and to provide detailed justifications for the mis-interpretability categories assigned during annotation. Responses to the IQs are obtained using a RAG mechanism, which integrates contextual information from both internal policy document segments and external, reliable web sources. This dual-context strategy facilitates a comprehensive understanding of the ambiguities present in the paragraph prior to rectification. The final output includes the rectified paragraph, accompanied by the corresponding IQs and their responses, thereby enhancing both clarity and in-

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 8217–8234
11–16 May 2026.

terpretability.

We applied the framework to a diverse corpus of 240 real-world policy documents sourced from leading multinational corporations, resulting in a benchmark dataset comprising 11,000 annotated paragraphs. Each paragraph is labeled as either interpretable or belonging to a specific mis-interpretability category. This dataset provides a valuable resource for evaluating both the detection and rectification of potentially mis-interpretable content in policy documents.

To evaluate the Annotation module i.e. to ensure the reliability and quality of the annotations, we conduct a two-step evaluation process: (1) manual assessment by expert annotators, and (2) automated validation using Pretrained Language Models (PLM),including Legal-BERT (Chalkidis et al., 2020), BERT (Devlin et al., 2019), CaseLaw-BERT (Zheng et al., 2021), RoBERTa (Liu et al., 2019) and DeBERTa (He et al., 2020).

To evaluate the Rectification module, we utilized five open-source LLMs: Mistral-2-7B (Jiang, 2024), Mistral-3-7B (Jiang, 2024), LLaMA-2-7B (Touvron et al., 2023), LLaMA-3-8B (Grattafiori et al., 2024), and Saul-7B (Colombo et al., 2024). The quality of the rectified content was assessed through human evaluations based on three key metrics: *Clarity*, *Fidelity*, and *Usefulness*.

The contributions of this work are as follows:

1. We propose an automated framework leveraging RAG for the identification and rectification of potentially mis-interpretable paragraphs in policy documents. The framework comprises:
   (a) An Annotation module that systematically identifies and classifies potentially mis-interpretable paragraphs in policy documents using both paragraph-level and document-level contextual reasoning.
   (b) A Rectification module that leverages LLMs to rewrite potentially mis-interpretable content, thereby enhancing clarity and interpretability for policy readers.
2. We introduce a benchmark dataset developed using the Annotation module on a selected set of 240 real-world policy documents. This resulted in over 11,000 annotated paragraphs, each classified either potentially mis-interpretable or interpretable. The dataset is a key resource for improving research in automated misinterpretation detection and rectification.
3. Through extensive experimentation with five open-source LLMs, we demonstrate that the rectified output significantly improves comprehension of policy content, as validated by human evaluators across multiple dimensions.

2. Related Work

Recent research on policy documents, particularly privacy policies, has predominantly concentrated on identifying ambiguous language. While these efforts have enhanced our understanding of ambiguities present in policy documents, they often fall short of addressing the misinterpretations that such ambiguities can cause.

Bhatia et al. (2016) proposed a theory of vagueness in privacy policies through content analysis of 15 documents, categorizing vague expressions into Conditionality, Generalization, Modality, and Numeric Quantifiers. Liu et al. (2016) employed deep learning techniques, specifically LSTM-based models, to classify vague versus non-vague language in privacy policies. Wilson et al. (2016) leveraged crowdsourcing to annotate ambiguous elements in privacy policies, emphasizing the value of human judgment in interpreting complex legal language. Lebanoff and Liu (2018) constructed a large-scale corpus of vague words and sentences and employed BiLSTM and feedforward neural networks for vagueness classification. Their annotated dataset serves as a valuable resource for training and evaluation. Kotal et al. (2020) introduced ViCLOUD, a model that quantifies ambiguity in cloud-related legal documents using linguistic markers and the Dale-Chall readability formula. Ahmad et al. (2020) developed the PolicyQA dataset to support reading comprehension and improve information retrieval from lengthy privacy policies, facilitating the development of question-answering systems. Hosseini et al. (2021) proposed an automated method for detecting ambiguity in privacy policies by inferring semantic relationships, such as hypernymy and synonymy. Safaei and Longo (2024) explored the integration of natural language processing with human analysis to generate public policy briefing notes, enhancing analytical efficiency and reducing misinterpretation.

Despite significant advancements in detecting vagueness and ambiguity within policy documents, a critical gap remains: current methodologies predominantly focus on identification, often neglecting the equally vital task of rectifying linguistic imprecision and mitigating its potential for misinterpretation. To address this limitation, we propose a novel methodology that not only identifies sentences prone to misinterpretation but also generates rectified versions to improve clarity and accessibility.

3. Data Collection and Identification of Mis-Interpretable Categories

We compiled a corpus of 240 policy documents in PDF format from leading multinational corpo-

rations across the globe, encompassing a wide range of policy types, including privacy policies, employee rights, insurance, fraud prevention, anti-corruption measures, and codes of conduct. All contracts are in English, ensuring consistency across the dataset. To further enhance its diversity and representation, documents were sourced from multiple reputable repositories: the Mendeley Data Repository[1](which has privacy policies from the top 100 MNCs), the CKAN Repository[2] for Misinformation Policies, the CKAN Repository for AI Regulation Policies[3], and the PingAn Group's policy archive[4].

Text extraction was performed using the PyPDF2[5] library. For paragraph segmentation, each paragraph was further divided into individual sentences using SpaCy's[6] sentence boundary detection. To maintain semantic coherence and manage chunk sizes, we implemented a recursive sentence-grouping strategy based on predefined character length thresholds. This approach ensured that the resulting text blocks were contextually intact and suitable for annotation.

The final dataset comprised 11,000 samples. To develop a labeling scheme, the first two authors manually analysed a pilot subset of 300 samples. Each sample was assessed for potential misinterpretation, with documented reasoning for those deemed potentially mis-interpretable. Importantly, the review was conducted in the context of the full source document to mitigate risks of out-of-context misreadings. Based on this analysis, the authors reached consensus on four key linguistic features, referred to as mis-interpretable categories, that commonly contribute to misunderstanding in policy documents:

1. Conditional Sentences: These refer to sentences where an outcome, benefit, or action depends on the fulfillment of specific conditions. When multiple conditions are implied or not explicitly defined, the sentence can become complex, unclear, or open to interpretation, increasing the risk of misinterpretation.

- Example: *Currently, the Special Surrender Value (SSV) is the same as Guaranteed Surrender Value. The SSV may be revised from time to time with prior approval of the Authority.*
- Explanation: *This sentence introduces of the possibility of changes to the SSV but does not clearly define the specify conditions or criteria under which such changes might occur. The absence of clear conditions increases interpretive ambiguity and could lead to misunderstanding among policyholders.*

2. Cross Dependent Sentences: These sentences reference other sections of the same document or external sources, such as regulatory documents. Comprehension requires readers to navigate across multiple sections or documents. When cross-references are poorly organized or inadequately signposted, the risk of misinterpretation increases.

- Example: *Coverage for losses arising from natural disasters as defined under Section 2.1 shall only apply if the policyholder has fulfilled the obligations specified in Section 3.4, including the payment of additional premiums detailed in Appendix A."*
- Explanation: *To fully understand this sentence, readers must consult Section 2.1, Section 3.4, and Appendix A. The lack of an integrated summary or cohesive referencing may hinder understanding and increase the cognitive burden on readers, thereby increasing the risk of misinterpretation.*

3. Legal Terminology: Sentences featuring formal legal or technical vocabulary can obscure meaning, particularly for lay readers. These terms often lack immediate clarity and require contextual or domain-specific knowledge to interpret correctly.

- Example: *In accordance with the provisions outlined in this agreement, any claims for reimbursement submitted by the policyholder must be accompanied by all requisite documentation, including, but not limited to, certified copies of original receipts, detailed proof of loss forms, and corroborating evidence from third-party service providers, failing which the insurer reserves the right, at its sole discretion, to deny said claims without further obligation to notify the policyholder of deficiencies.*
- Explanation: *Terms such as "requisite documentation," "corroborating evidence," and "sole discretion" introduce legal complexity that can blur the meaning. Without clear definitions or supporting context, readers may struggle to understand their implications for claims processing and the extent of the insurer's authority.*

4. Ambiguity in Expression: This category includes sentences containing vague or non-committal language that can be interpreted in multiple ways, typically due to imprecise word choices.

- Example: *The company may offer coverage for damage caused by certain events, subject*

[1]https://data.mendeley.com/datasets/pcgvm6zh43/1

[2]https://ckan.uos.staging.datopian.com/dataset/policy-documents-misinformation/resource/85d7c125-5c69-4ed5-a41a-3e7e70c09866

[3]https://ckan.uos.staging.datopian.com/dataset/policy-documents-ai-regulation/resource/280a3400-4a6d-4e07-9a65-e9f03e008f2d

[4]https://group.pingan.com/

[5]https://pypi.org/project/PyPDF2/

[6]https://spacy.io/

to approval.

- Explanation: *Phrases such as 'may offer' and 'certain events' are vague and lack specificity, making the scope and certainty of coverage unclear. Such ambiguity can result in inconsistent interpretations and confusion regarding policy applicability.*

Our evaluation highlights the importance of contextual granularity in interpreting policy language, specifically: (a) Global Context (i.e., the broader document or associated materials) is essential for accurately interpreting sentences in the *Cross-Dependent Sentences* and *Legal Terminology* category. For example, a statement such as "*Coverage for losses due to natural disasters shall apply as per the conditions outlined in Section 4.2 and Appendix A*" cannot be fully understood in isolation. Similarly, legal phrases like 'reasonable effort','due diligence' or 'at the sole discretion' derive their precise meaning from definitions or explanations provided elsewhere in the document. (b) Local Context (i.e., confined to a single paragraph) may suffice for identifying sentences in *Ambiguity in Expression* category, but is generally inadequate for resolving cross-references or interpreting specialized legal terms. Therefore, effective analysis and clarification of mis-interpretable content necessitate access to the full document structure and supporting references to ensure accurate and context-aware interpretation.

4. Proposed Framework

Our framework comprises two core modules: (1) Annotation module, which identifies potentially mis-interpretable paragraphs through local and global context analysis and (2) Rectification module, which enhances the clarity by rewriting the potentially mis-interpretable paragraphs. Figure 1 illustrates the complete pipeline.

4.1. Annotation Module

We developed a two-stage Annotation module aimed at categorizing paragraphs extracted from policy documents, with an emphasis on interpretability and susceptibility to misinterpretation. The module leverages LLMs to perform both classification and explanation generation, thereby enhancing the transparency and reliability of the annotation process. Next, we explain the two stages of the Annotation module.

4.1.1. Local Context Annotation

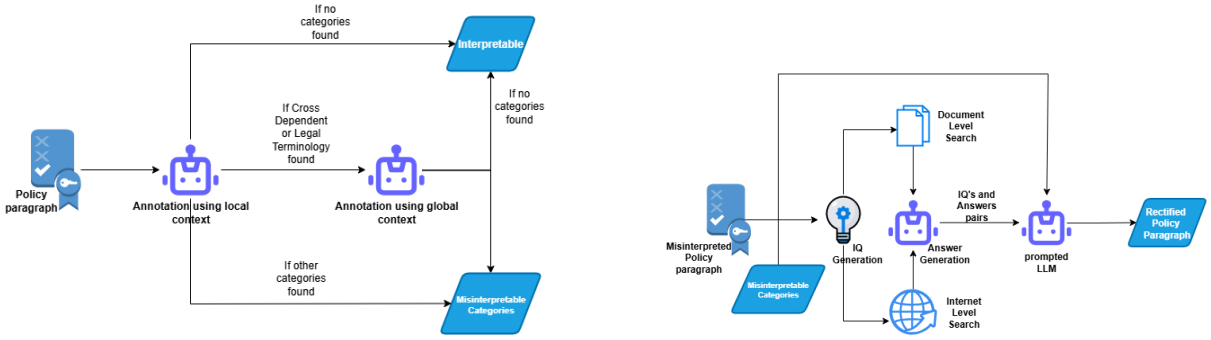
In the local context annotation stage, each paragraph p_i is independently processed using a structured classification prompt. The LLM assigns one or more labels l_i from the mis-interpretability categories: *Conditional Sentences*, *Cross-Dependent Sentences*, *Legal Terminology* and *Ambiguity in Expression*. Paragraphs that do not fall into any of these categories are labeled as *Interpretable*, indicating that the sentence is clear without any ambiguities. Each classification is accompanied by a model-generated explanation r_i, which provides insight into the rationale behind the assigned mis-interpretability category labels. These model-generated explanations are later used in the Global Context Annotation.

4.1.2. Global Context Annotation

Following the local context annotation stage, paragraphs initially classified as *Cross-Dependent Sentences* or *Legal Terminology* undergo a second stage of analysis, termed as Global Context Annotation, due to their reliance on broader document context for accurate interpretation. To retrieve relevant global context, the source policy document is segmented into a set of textual chunks $\{c_i\}$. Cosine similarity is computed between the chunks and model-generated explanation for the respective categories (i.e. cos(c_i, $r_{cross-dependent}$) and cos(c_i, $r_{legal\ terminology}$)). The top-k most relevant chunks are selected and combined with the original paragraph to form an enriched input for the LLM. This augmented input is processed using a global context prompt, yielding a refined classification l_i' and an updated explanation r_i', which incorporates the additional contextual information. Importantly, this stage allows for re-evaluation of the initial mis-interpretability category, meaning a paragraph originally tagged as mis-interpretable may be re-classified as *Interpretable* if the added context resolves the misinterpretation. For both Local Context and Global Context Annotation, we utilized Gemma-2-9B (Team et al., 2024) due to its advanced capabilities in understanding nuanced contexts. The prompts used in both stages of annotation are mentioned in the Appendix 11.3.

4.1.3. Annotated Dataset Description

All 11,000 samples were processed through the Annotation module for labeling, which classifies each sample as either *Interpretable* or assigns it to one or more of the four mis-interpretable categories: *Conditional Sentences*, *Cross-Dependent Sentences*, and *Legal Terminology*. The labeled dataset consists of three components: paragraph text, annotated categories, and global context. The global context includes surrounding text from the original policy documents, which is essential for addressing interpretability challenges beyond the local paragraph level. The dataset is distributed across various mis-interpretability categories. Among the single category assigned in-



(a) Annotation module

(b) Rectification module

Figure 1: Overview of the Annotation and Rectification modules

stances, *Interpretable* (6848) has the maximum number of instances, while *Ambiguity in Expression* (103) has the minimum. The predominance of the *Interpretable* category suggests that most paragraphs are clear and do not present significant interpretability challenges. Among the instances, the maximum combination occurs with 2,156 instances labeled as both *Conditional Sentences* and *Cross-Dependent Sentences*, while the minimum combination includes 30 instances annotated with all four mis-interpretability labels. The remaining instances are widely distributed across several other combinations, reflecting the complex linguistic nature of legal documents. Figure 2 provides a detailed category-wise distribution of the labeled dataset, highlighting both single and multi-class combinations. Further details regarding the category-wise distribution are provided in Appendix 11.1.

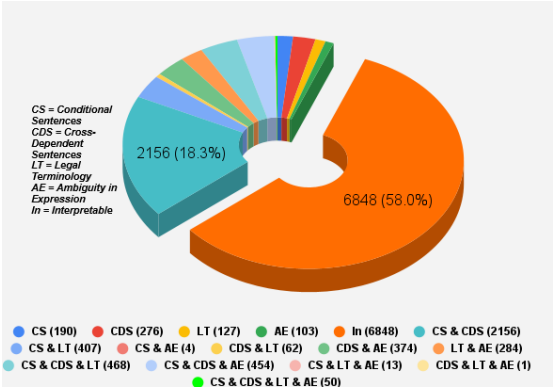


Figure 2: Category Wise Distribution of the Dataset

4.2. Rectification Module

The Rectification module enhances the clarity of potentially mis-interpretable paragraphs in policy documents. Let D denote the input document and P represent the set of paragraphs identified as mis-interpretable by the Annotation module in section 4.1, along with their mis-interpretability categories. The rectification process begins by passing each paragraph P, along with its assigned mis-interpretability categories, into an LLM to generate Interpretation Queries (IQs) = $\{q_1, q_2, \ldots, q_n\}$, which probe the reasons for the paragraph's mis-interpretability based on its category. IQ's are specific, targeted questions designed to retrieve precise information needed to resolve the ambiguity identified by the Annotation module. For each query q_i, relevant contextual information is retrieved using a RAG mechanism. The document D is segmented into chunks $C = \{c_1, c_2, \ldots, c_t\}$, and cosine similarity scores cos(q_i, c_j) are computed to assess the relevance of each chunk to the IQ. The top-k chunks with the highest scores are selected for further analysis. Additionally, the module retrieves the top-n internet search results, which includes URLs, titles, and content, based on a specified set of criteria that specifically prioritizing '.gov', '.org', and '.edu' websites over commercial '.com' domains. Next, each query q_i, along with its top-k document chunks and top-n internet results, is passed back into the LLM, which generates answers $A = \{a_1, a_2, \ldots, a_n\}$. If an answer is not found in the retrieved context, the LLM indicates that the answer is unavailable. Finally, the original paragraph P, its associated queries IQ, and answers A are reintroduced to the LLM to produce a rectified version of the paragraph, denoted as P', which incorporates insights from both document-level and external context. The module's final output, intended for human consumption to enhance clarity and interpretability for policy readers, includes the rectified paragraph P', the set of IQs, and their corresponding answers A, which together provide transparency into the rectification process. Appendix 11.2 presents sample outputs generated by the Rectification module and detailed prompt templates used for query generation and rectification are provided in Appendix 11.3.

5. Experimental Results

In this section, we outline the evaluation strategy and metrics employed to assess the Annotation and Rectification modules. For evaluating the annotation quality of the Annotation module, we utilized both manual evaluation and PLM-based assessment. In assessing the Rectification module, we applied manual evaluation to determine how effectively it enhances the clarity of potentially misinterpretable paragraphs.

5.1. Evaluation of the Annotation Module

Given that our dataset is generated using a LLM, rigorous evaluation of its quality and accuracy is imperative. We employ two complementary evaluation methodologies: (1) Manual evaluation, (2) Automated evaluation using Pre-Trained Language Model (PLM) validators.

5.1.1. Manual Evaluation

To evaluate the annotation quality of the Annotation module, we conducted a manual validation study. Eight researchers , each with over two years of experience in the legal domain participated in the manual validation exercise (henceforth referred to as participants). As manual cross-checking of all 11,000 samples was infeasible, we randomly selected 800 samples for review, distributing 100 unique samples to each participant. Participants were provided with detailed labeling guidelines and access to the full source documents to ensure contextual fidelity, and they independently labeled their assigned subsets. Once labeling was completed, we compared the annotation by the participants against the outputs of the Annotation module. We then compared human-generated labels with those from the Annotation module, calculating F_1 scores for each category. The *Legal Terminology* category achieved the highest F_1 score of 0.962, while the *Interpretable* category had the lowest at 0.891, as summarized in Table 1. To assess the statistical reliability of these F_1 scores across the full dataset, we computed 95% Confidence Intervals (CI) for each category. CI indicates the range within which the true F_1 score is expected to lie and is calculated using the formula CI $= F_1 \pm Z \times$ SE, where $Z = 1.96$ for a 95% confidence level, and SE is the standard error given by SE $= \sqrt{\frac{F_1 \times (1-F_1)}{n}}$; in this case, $n = 800$, which represents the number of annotated samples. The standard error reflects the variability of the F_1 score and influences the width of the confidence interval. The standard error reflects the variability of the F_1 score and influences the CI's width. As shown in Table 1, the CI scores for the F_1 scores are relatively narrow across all categories, suggesting consistent performance of the framework. For instance, the F_1 score for *Ambiguity in Expression* was 0.964 with a 95% CI of [0.957, 0.971], indicating high precision and low variability. To further assess inter-annotator agreement, an additional subset of 100 samples was annotated by all eight participants. Krippendorff's alpha, computed on this subset, yielded a value of 0.767, indicating substantial agreement. Overall, these results provide robust statistical support for the framework's reliability across distinct legal annotation categories.

Category	Mean F_1	95% Confidence Interval
Conditional Sentences	0.911	[0.904, 0.918]
Cross-Dependent Sentences	0.91	[0.903, 0.917]
Legal Terminology	0.962	[0.955, 0.969]
Ambiguity in Expression	0.964	[0.957, 0.971]
Interpretable	0.891	[0.884, 0.898]

Table 1: Manual Evaluation Results with 95% Confidence Intervals.

5.1.2. PLM-Based Evaluator

Manual evaluation (Section 5.1.1) of 800 samples provided a quality benchmark but was not scalable to the full dataset of 11,000 samples. To address this, we fine-tuned five transformer-based models: Legal-BERT, BERT, CaseLaw-BERT, RoBERTa, and DeBERTa on the human annotated data. To improve the training set, we employed Back Translation (Ciolino et al., 2021) for text augmentation by first translating the paragraphs into four languages: German, Italian, French, and Spanish, and then rendering it back into English, increasing sample size from 800 to 4,000. Each model was trained for multi-label classification across five misinterpretable categories using a sigmoid-based dense output layer. As shown in Table 2, all models performed strongly (F_1 > 0.919), with DeBERTa leading in *Ambiguity in Expression* (0.980) and *Conditional Sentences* (0.965), and Legal-BERT excelling in the *Interpretable* category (0.933). The fine-tuned models were then applied to the remaining 10,200 unlabeled samples to assess annotation consistency. As summarized in Table 3, F_1 scores remained high (> 0.85) across all categories. RoBERTa achieved top scores in *Ambiguity in Expression* (0.905) and *Legal Terminology* (0.915), while CaseLaw-BERT and DeBERTa led in *Cross-Dependent Sentences* (0.916). The *Interpretable* category had relatively lower scores (0.844–0.868), highlighting the subjectivity of legal interpretation. These results validate the Annotation module and demonstrate an effective strategy for scalable annotation quality assessment.

Model	Ambiguity in Expression			Conditional Sentences			Cross-Dependent			Legal Terminology			Interpretable		
	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1
BERT	0.966	0.987	0.976	0.975	0.949	0.961	0.983	0.962	0.972	0.988	0.978	0.983	0.898	0.958	0.926
CaseLaw-BERT	0.966	0.989	0.977	0.977	0.941	0.959	0.969	0.961	0.965	0.989	0.978	0.984	0.898	0.943	0.919
DeBERTa	0.969	0.991	0.98	0.975	0.955	0.965	0.985	0.956	0.97	0.987	0.98	0.983	0.898	0.958	0.926
Legal-BERT	0.975	0.977	0.976	0.976	0.954	0.965	0.987	0.963	0.975	0.99	0.969	0.979	0.901	0.968	0.933
RoBERTa	0.975	0.979	0.977	0.98	0.95	0.965	0.983	0.956	0.969	0.986	0.975	0.981	0.882	0.96	0.919

Table 2: Performance of PLMs using 5-Fold Cross-Validation on the Augmented Annotated Dataset

Model	Ambiguity in Expression			Conditional Sentences			Cross-Dependent			Legal Terminology			Interpretable		
	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1
BERT	0.882	0.907	0.894	0.895	0.898	0.896	0.91	0.925	0.913	0.905	0.907	0.907	0.861	0.869	0.858
CaseLaw-BERT	0.894	0.907	0.9	0.898	0.901	0.899	0.909	0.929	0.913	0.912	0.911	0.911	0.876	0.876	0.868
DeBERTa	0.896	0.908	0.902	0.897	0.906	0.9	0.912	0.936	0.916	0.912	0.911	0.912	0.863	0.867	0.86
Legal-BERT	0.873	0.907	0.888	0.886	0.89	0.887	0.91	0.922	0.913	0.893	0.902	0.897	0.849	0.86	0.844
RoBERTa	0.901	0.909	0.905	0.886	0.885	0.885	0.9	0.905	0.901	0.917	0.914	0.915	0.854	0.863	0.85

Table 3: Model Evaluation on Unlabeled Dataset

5.2. Evaluation of the Rectification module

5.2.1. Evaluation Metrics

To assess the Rectification module, we employed five open-source LLMs: Mistral-2-7B (Jiang, 2024), Mistral-3-7B (Jiang, 2024), LLaMA-2-7B (Touvron et al., 2023), LLaMA-3-8B (Grattafiori et al., 2024), and Saul-7B (Colombo et al., 2024). The manual evaluation was conducted by the same eight researchers. Each participant was provided with the 1000 samples generated by the Rectification module (200 for each LLM) to manually assess the quality of the rectified content. The average scores assigned by the participants for each LLM are presented in Table 4. Participants were given clear definitions and guidelines for four evaluation metrics: *Clarity*, *Fidelity*, *Usefulness*, and *Clarification Effectiveness*. *Clarity* measured how much the rectified paragraph improved readability, while *Fidelity* assessed whether the original meaning was preserved. *Usefulness* evaluated how well the rectified text resolved ambiguities, and *Clarification Effectiveness* assessed the relevance and helpfulness of the IQ Answer pair in addressing confusion. Each metric was scored on a 1–5 scale, ranging from poor to excellent performance. These structured guidelines ensured consistency and reliability in human judgment across evaluations. Scores were normalized for consistency, and we also calculated the aggregated mean, representing the mean of all four metrics.

Model	Clarity	Fidelity	Usefulness	Clarification Effectiveness	Aggregated Mean
Mistral-2-7B	0.912	0.914	0.934	0.947	0.927
Mistral-3-7B	0.887	0.931	0.923	0.962	0.926
LLaMA-2-7B	0.626	0.738	0.727	0.856	0.737
LLaMA-3-8B	0.758	0.812	0.815	0.950	0.834
Saul-7B	0.820	0.952	0.786	0.874	0.858

Table 4: Normalized human evaluation results for rectified content and QA effectiveness. Results are normalized to [0, 1].

5.2.2. Evaluation Results

As shown in Table 4, Mistral-2-7B and Mistral-3-7B excelled in human-rated criteria, with mean scores of 0.927 and 0.926, respectively. Mistral-3-7B had the highest ratings in *Fidelity* (0.931) and *Clarification Effectiveness* (0.962), showcasing its ability to preserve meaning and generate quality IQs. Mistral-2-7B excelled in *Usefulness* (0.934) and *Clarity* (0.912). Saul-7B performed well in *Fidelity* (0.952) but had a lower *Usefulness* score (0.786), leading to an overall mean of 0.858. LLaMA-3-8B showed moderate performance with an aggregated score of 0.834, while LLaMA-2-7B had the lowest mean score of 0.737. These results highlight the effectiveness of Mistral model in producing high-quality rectifications and IQs for applications requiring semantic precision.

5.2.3. Ablation Study on the Rectification Module

To evaluate the contribution of each component in our Rectification module, we conducted two ablation studies: (1) Ablation-1 involved providing the LLM with the potentially mis-interpretable paragraph and IQs but without any retrieved answers. This setup tested the model's ability to rectify based solely on internal knowledge and the guiding questions. (2) Ablation-2 presented only the potentially mis-interpretable paragraph, omitting both IQs and retrieved answers, to assess the impact of external knowledge. We evaluated the output of these ablated systems against the full Rectification module using the same evaluation setup mentioned in Section 5.2.1. The results are summarized in Table 5. The results indicated that the full Rectification module consistently outperformed both ablated versions, demonstrating that neither internal knowledge alone (Ablation-2) nor guidance from unanswered IQs (Ablation-1) is sufficient for optimal rectification. Notably, the Mistral family performed well compared to other models. However, they still achieve their peak per-

Model	Rectification module			Ablation-1			Ablation-2		
	Clarity	Fidelity	Usefulness	Clarity	Fidelity	Usefulness	Clarity	Fidelity	Usefulness
Mistral-2-7B	0.912	0.914	0.934	0.878	0.796	0.886	0.772	0.750	0.742
Mistral-3-7B	0.887	0.931	0.923	0.858	0.806	0.870	0.836	0.796	0.802
LLaMA-2-7B	0.626	0.738	0.727	0.675	0.665	0.633	0.560	0.540	0.540
LLaMA-3-8B	0.758	0.812	0.815	0.622	0.596	0.588	0.647	0.627	0.647
Saul-7B	0.820	0.952	0.786	0.612	0.760	0.536	0.705	0.731	0.696

Table 5: Results for Ablation Study. Results are normalized to [0, 1].

formance only with the full Rectification module. LLaMA models faced significant performance drop in the ablation settings, whereas Saul-7B significantly lost *Fidelity* in Ablation-1, suggesting its reliance on retrieved context for maintaining legal meaning and *Clarity*.

6. Runtime and Scalability Analysis

To evaluate the feasibility of our framework for real-world applications, we conducted a runtime and scalability analysis on an NVIDIA V100 GPU (32 GB memory, 60 GB RAM) by processing 100 policy documents with a batch size of 1 across five LLMs. The Annotation module averaged 102.95 seconds per document, 14.85 seconds for classification and 88.10 seconds for retrieval. The Rectification module's runtime varied by model, from 4.52 seconds for Saul-7B to 31.31 seconds for LLaMA-3-8B, as shown in Table 6. We also computed total processing time (annotation + rectification) and document-level throughput estimates for each model, for example using Mistral series in the Rectification module, the overall throughput of the entire framework is approximately 30 contracts per hour.

Model	Rectification Time (s)	Annotation Time (s)	Total Time per Document (s)	Throughput
Mistral-2-7B	16.98	102.95	119.93	30.02
Mistral-3-7B	12.89	102.95	115.84	31.08
LLaMA-2-7B	9.4	102.95	112.35	32.04
LLaMA-3-8B	31.31	102.95	134.26	26.81
Saul-7B	4.52	102.95	107.47	33.5

Table 6: Combined Runtime and Throughput Analysis

7. Data and Code Availability

The code and dataset are available in this public repository[7]

8. Limitations

Despite strong performance, our framework has practical limitations: (1) Scope of Interpretability vs. Genuine Ambiguity: Our framework effectively resolves misinterpretability caused by missing context using RAG, but is less suited for fundamental semantic ambiguities with multiple valid interpretations. In such cases, the RAG mechanism may retrieve conflicting information from different sources, making consistent disambiguation challenging. Future work will explore dedicated ambiguity detection and disambiguation modules to address this limitation. (2) Knowledge Data Limitations: Our approach depends on external knowledge retrieval, which risks outdated internet content, and our dataset may not fully capture global policy variations. Future work will integrate curated knowledge bases and expand dataset diversity. (3) Annotation Output Quality: The framework is subject to annotation subjectivity in legal interpretation and occasionally generates vague IQs. We plan to improve prompt engineering, diversify annotator pools, and implement an IQ refinement pipeline to enhance consistency.

9. Conclusion

In this work, we presented a comprehensive two-stage framework designed to mitigate misinterpretation in policy documents. The framework first identifies policy paragraphs that may be prone to misinterpretation through sentence-level and document-level reasoning, and subsequently rectifies it using targeted Interpretation Queries and Retrieval-Augmented Generation. Applied to a diverse corpus of 240 real-world policy documents, our approach produced a benchmark dataset comprising 11,000 annotated samples, systematically categorized into distinct mis-interpretability categories. The reliability of the Annotation module was validated through both expert manual review and large-scale evaluations using PLMs. Furthermore, the Rectification module demonstrated robust performance across key human-evaluated dimensions: *Clarity*, *Fidelity*, and *Usefulness*, when tested with multiple open-source LLMs.

In future work, we aim to enhance the framework by integrating curated, domain-specific knowledge bases to provide a more controlled and contextually relevant support during rectification. Additionally, we plan to implement a refinement loop to improve the specificity and effectiveness of IQs, thereby further elevating the quality of the rectified

[7] Repository

outputs.

10. Bibliographical References

Wasi Uddin Ahmad, Jianfeng Chi, Yuan Tian, and Kai-Wei Chang. 2020. Policyqa: A reading comprehension dataset for privacy policies. *arXiv preprint arXiv:2010.02557*.

Vinayshekhar Bannihatti Kumar, Roger Iyengar, Namita Nisal, Yuanyuan Feng, Hana Habib, Peter Story, Sushain Cherivirala, Margaret Hagan, Lorrie Cranor, Shomir Wilson, et al. 2020. Finding a choice in a haystack: Automatic extraction of opt-out statements from privacy policy text. In *Proceedings of The Web Conference 2020*, pages 1943–1954.

Jaspreet Bhatia, Travis D Breaux, Joel R Reidenberg, and Thomas B Norton. 2016. A theory of vagueness and privacy risk perception. In *2016 IEEE 24th International Requirements Engineering Conference (RE)*, pages 26–35. IEEE.

Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. 2020. Legal-bert: The muppets straight out of law school. *arXiv preprint arXiv:2010.02559*.

Matthew Ciolino, David Noever, and Josh Kalin. 2021. Back translation survey for improving text augmentation. *arXiv preprint arXiv:2102.09708*.

Pierre Colombo, Telmo Pessoa Pires, Malik Boudiaf, Dominic Culver, Rui Melo, Caio Corro, Andre FT Martins, Fabrizio Esposito, Vera Lúcia Raposo, Sofia Morgado, et al. 2024. Saullm-7b: A pioneering large language model for law. *arXiv preprint arXiv:2403.03883*.

Wassim Derguech, Syeda Sana e Zainab, and Mathieu d'Aquin. 2018. Assessing the readability of policy documents: The case of terms of use of online services. In *Proceedings of the 11th International Conference on Theory and Practice of Electronic Governance*, pages 247–256.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186.

Christopher C French. 2013. The aftermath of catastrophes: Valuing business interruption insurance losses. *Ga. St. UL Rev.*, 30:461.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Yu Han, Aaron Ceross, and Jeroen HM Bergmann. 2024. The use of readability metrics in legal text: A systematic literature review. *arXiv preprint arXiv:2411.09497*.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020. Deberta: Decoding-enhanced bert with disentangled attention. *arXiv preprint arXiv:2006.03654*.

Mitra Bokaei Hosseini, John Heaps, Rocky Slavin, Jianwei Niu, and Travis Breaux. 2021. Ambiguity and generality in natural language privacy policies. In *2021 IEEE 29th International Requirements Engineering Conference (RE)*, pages 70–81. IEEE.

Fengqing Jiang. 2024. Identifying and mitigating vulnerabilities in llm-integrated applications. Master's thesis, University of Washington.

Anantaa Kotal, Karuna Pande Joshi, and Anupam Joshi. 2020. Vicloud: Measuring vagueness in cloud service privacy policies and terms of services. In *2020 IEEE 13th International Conference on Cloud Computing (CLOUD)*, pages 71–79. IEEE.

Logan Lebanoff and Fei Liu. 2018. Automatic detection of vague words and sentences in privacy policies. *arXiv preprint arXiv:1808.06219*.

Fei Liu, Nicole Lee Fella, and Kexin Liao. 2016. Modeling language vagueness in privacy policies using deep neural networks. In *AAAI Fall Symposia*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Mehrdad Safaei and Justin Longo. 2024. The end of the policy analyst? testing the capability of artificial intelligence to generate plausible, persuasive, and useful policy analysis. *Digital Government: Research and Practice*, 5(1):1–35.

Roman Senninger. 2023. What makes policy complex? *Political Science Research and Methods*, 11(4):913–920.

Anil K. Dixit Sumit Arora. 2024. Demystifying insurance dispute settlement in india: A comprehensive analysis. *International Journal of Research Publication and Reviews*, 5:7443–7447.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, et al. 2024. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Shomir Wilson, Florian Schaub, Rohan Ramanath, Norman Sadeh, Fei Liu, Noah A Smith, and Frederick Liu. 2016. Crowdsourcing annotations for websites' privacy policies: Can it really work? In *Proceedings of the 25th International Conference on World Wide Web*, pages 133–143.

Lucia Zheng, Neel Guha, Brandon R Anderson, Peter Henderson, and Daniel E Ho. 2021. When does pretraining help? assessing self-supervised learning for law and the casehold dataset of 53,000+ legal holdings. In *Proceedings of the eighteenth international conference on artificial intelligence and law*, pages 159–168.

11. Appendix

11.1. Dataset Category Distribution Overview

Dataset category distribution is mentioned in Table 7. The first five rows correspond to instances exclusively assigned to a single category, while the subsequent rows capture various combinations of overlapping labels.

11.2. End-to-End Examples with Interpretation Queries and Rectified Paragraphs

Table 8 and Table 9 present examples containing the original input paragraph, the rectified paragraph generated by the Rectification module, and the corresponding Interpretation Queries for each input.

11.3. Prompts Used

All prompts employed in the proposed framework are listed in this section to ensure transparency and reproducibility.

Category	Number of Samples	Percentage Distribution (%)
Conditional Sentences	190	1.61
Cross-Dependent Sentences	276	2.34
Legal Terminology	127	1.08
Ambiguity in Expression	103	0.87
Interpretable	6848	58.05
Conditional Sentences & Cross-Dependent Sentences	2156	18.28
Conditional Sentences & Legal Terminology	407	3.45
Conditional Sentences & Ambiguity in Expression	4	0.03
Cross-Dependent Sentences & Legal Terminology	62	0.53
Cross-Dependent Sentences & Ambiguity in Expression	374	3.17
Legal Terminology & Ambiguity in Expression	284	2.41
Conditional Sentences & Cross-Dependent Sentences & Legal Terminology	468	3.97
Conditional Sentences & Cross-Dependent Sentences & Ambiguity in Expression	454	3.85
Conditional Sentences & Legal Terminology & Ambiguity in Expression	13	0.11
Cross-Dependent Sentences & Legal Terminology & Ambiguity in Expression	1	0.01
Conditional Sentences & Cross-Dependent Sentences & Legal Terminology & Ambiguity in Expression	30	0.25

Table 7: Category Wise Distribution of The Dataset

Original Paragraph	The transmission of your personal data is secured by Standard Contractual Clauses and Binding Corporate Rules in accordance with Art. 46 GDPR. Learn more about Zendesk and their commitment to the protection of personal data at https://www.zendesk.com/company/privacy-and-data-protection. For security reasons, registration in the FREE NOW Passenger App is verified with an SMS code to your mobile device ("Multi–Factor Authentication"). To send SMS, we use the service provider Twilio Inc., 375 Beale St, Suite 300, San Francisco, CA 94105, USA ("Twilio"). To protect your personal data from misuse, Twilio's Call Service is used to mask your phone number from the driver. The driver can still call you (via Twilio), but will not see your actual phone number. The legal basis for using Twilio's services is Art. 6(1)(f) GDPR, as these measures improve the security of our services.
Rectified Paragraph	The transmission of your personal data is protected through the European Commission's Standard Contractual Clauses (SCCs) and Twilio's Binding Corporate Rules, in accordance with Art. 46 GDPR. To enhance account security, the FREE NOW Passenger App uses Multi–Factor Authentication (MFA), requiring verification via an SMS code sent to your registered device. SMS delivery and phone-number masking are handled by Twilio Inc. During number masking, Twilio assigns a proxy number so drivers can contact you without seeing your real number. These security measures rely on the legitimate-interest basis of Art. 6(1)(f) GDPR to ensure safe and privacy-preserving communication.
Interpretation Queries 1	**What specific Standard Contractual Clauses and Binding Corporate Rules are used under Art. 46 GDPR?**
Answer	The SCCs referenced are the European Commission–approved clauses designed to safeguard personal data transfers within the EEA and to third countries. They impose EU-level data-protection obligations on the data recipient. Binding Corporate Rules (BCRs) are internal compliance frameworks adopted by multinational firms like Twilio to ensure that personal data transferred within their corporate group receives GDPR-equivalent protection.
Interpretation Queries 2	**How does Multi–Factor Authentication improve the security of the FREE NOW registration process?**
Answer	MFA strengthens the registration process by requiring a second verification factor beyond a password. In the FREE NOW app, this involves entering an SMS code sent to the user's device, ensuring that only individuals with access to that device can complete registration and preventing unauthorized account creation.
Interpretation Queries 3	**How does Twilio mask the passenger's phone number from the driver?**
Answer	Twilio's Call Service assigns a temporary proxy number, which routes the call to the passenger while concealing their real phone number. This enables necessary communication while protecting passenger privacy.
Interpretation Queries 4	**What is the legal basis for using Twilio's SMS and masking services under Art. 6(1)(f) GDPR?**
Answer	The processing relies on FREE NOW's legitimate interest in ensuring secure user authentication and preventing misuse of personal data. Twilio's SMS verification and number-masking functions are necessary to maintain secure, privacy-preserving communication, aligning with the legitimate-interest basis of Art. 6(1)(f) GDPR.

Table 8: Example 1: Original Paragraph, Rectified Version and Interpretation Queries

Original Paragraph	The Life Assured may be the Policyholder. 9) Maturity Date - means the date stated in the Policy Schedule, on which the Policy Term expires and this Policy terminates; 10) Nominee - means the person named by You and registered with Us in accordance with the Nomination Schedule, who is authorized to receive the Sum Assured under this Policy, on the death of the Life Assured; 11) Policy Anniversary - means the annual anniversary of the Date of Risk Commencement; 12) Policyholder, You, you, your – means or refers to the Policyholder stated in the Policy Schedule. 13) Policy Term - means the term of the Policy as stated in the Policy Schedule; 14) Premium(s) - means an amount stated in the Policy Schedule, payable by You to Us for every Policy Year by the due dates, and in the manner stated in the Policy Schedule, to secure th e benefits under this Policy, excluding service tax or any other taxes, cesses or levies;
Rectified Paragraph	The Life Assured, who may or may not be the Policyholder, is the individual whose life is insured under the policy. The Policyholder, You, you, or your refers to the person who purchases the policy and has control over it during their lifetime. They can assign the policy, surrender it, or change the beneficiary. The Nominee, registered with Us in accordance with the Nomination Schedule, is the person authorized to receive the Sum Assured under this Policy upon the death of the Life Assured. The Maturity Date is the date stated in the Policy Schedule on which the Policy Term expires, and this Policy terminates. The Policy Term refers to the term of the Policy as stated in the Policy Schedule. Premiums are amounts stated in the Policy Schedule, payable by the Policyholder to secure the benefits under this Policy, excluding service tax or any other taxes, cesses, or levies.
Interpretation Queries 1	**What is the exact meaning of "The Life Assured may be the Policyholder"?**
Answer	The term ""The Life Assured"" refers to the individual whose life is insured under the policy. It is the person whose death triggers the payment of the Sum Assured as per the terms and conditions of the policy. In some cases, the Policyholder and the Life Assured may be the same person.
Interpretation Queries 2	**Is it mandatory for the Life Assured and Policyholder to be the same person?**
Answer	The question asks whether it is mandatory for the Life Assured and Policyholder to be the same person. Based on the context provided, there is no explicit requirement stating that the Life Assured and Policyholder must be the same person. However, in some instances, they may be the same individual, particularly when the breadwinner of a family purchases the insurance policy for income replacement for their dependents in case of their death. The context focuses on the role of the Policyholder, the insurer, and the Nominee/Beneficiary in a life insurance policy. The Policyholder is the person who buys the insurance policy and has control over it during their lifetime. They can assign the policy, surrender it, or change the beneficiary. The insured is the person whose life is insured under the policy. The Nominee or Beneficiary is the person who will receive the death benefits under the policy upon the death of the insured.
Interpretation Queries 3	**Could you please specify the exact process for registering a Nominee with the company?**
Answer	To register a Nominee with a company in accordance with the Insurance Act, 1938, the Policyholder can follow these steps: (1) Identify the Nominee and ensure they are willing to act as such. (2) The Policyholder must send a written notice to the Company, intimating the change in nomination and registering the same with the Company. (3) If the Nominee is a minor, the Policyholder is required to appoint an Appointee to receive the benefits under the Policy during the minority of the Nominee. The Policyholder's identity and residence proof, along with the Nominee's details, may be required as per the company's norms. However, the exact documents required for registration may vary depending on the company's policies and the jurisdiction in which the company is incorporated.
Interpretation Queries 4	**What information is required for registration?**
Answer	To register for a policy under the given context, the following information is required: (1) Claimant's identity and residence proof. (2) In case of Accidental Death, original or certified copies of Police records and Post Mortem Report. (3) Nomination details, including the nominee's name and, if applicable, the appointment of an appointee during the nominee's minority.

Table 9: Example 2: Original Paragraph, Rectified Version and Interpretation Queries

Classification Prompt for Local context Annotation Task

```
You are a legal expert analyzing policy documents.  Classify the
following text into the most appropriate categories listed below.
Return the output as mentioned in Output Format below.
### Categories and Examples:  1.  **Conditional Sentences**:
Conditional Sentences state that an outcome, benefit, or action
depends on the fulfillment of specific conditions.  When multiple
conditions are implied or not explicitly defined, the sentence
can become complex, unclear, or open to interpretation, increasing
the risk of misinterpretation.  - Example:  Currently, the Special
Surrender Value (SSV) is the same as the Guaranteed Surrender Value.
The SSV may be revised from time to time with prior approval of the
Authority.
2.  **Cross-Dependent Sentences**:  These refer to sentences that
rely on references to other sections of the document for full
understanding.  Cross-references often require readers to navigate
multiple sections, making the content harder to follow and increasing
```

```
the risk of misinterpretation.  - Example:  Coverage for losses
arising from natural disasters as defined under Section 2.1 shall only
apply if the policyholder has fulfilled the obligations specified in
Section 3.4, including the payment of additional premiums detailed in
Appendix A.
3.  **Legal Terminology**:  These sentences are challenging to
understand due to intricate phrasing or technical legal vocabulary.
Such complexity can blur the intended meaning, increasing the
likelihood of misinterpretation.  - Example:  In accordance with the
provisions outlined in this agreement, any claims for reimbursement
submitted by the policyholder must be accompanied by all requisite
documentation, including, but not limited to, certified copies of
original receipts, detailed proof of loss forms, and corroborating
evidence from third-party service providers, failing which the insurer
reserves the right, at its sole discretion, to deny said claims
without further obligation to notify the policyholder of deficiencies.
4.  **Ambiguity in Expression**:  This class refers to sentences or
clauses in policy documents that are unclear or open to multiple
interpretations due to vague or imprecise language.  - Vague Terms
like some, many, or a few, which lack precise meaning and can be
interpreted differently by different readers.  - Modal Verbs like may,
could, might, or should that introduce uncertainty and flexibility in
the interpretation of the statement.  - Example:  The refund will be
processed soon, but it may take a while.
### Instructions:  - First, identify key linguistic indicators
that belong to each category in the given text for classification
(such as conditionals like may, could, shall; cross-references like
Section 2.1, Appendix A; or complex legal terms like provisions
and corroborating evidence).  - Assess if the sentence introduces
uncertainty, uses vague terms, or provides unclear definitions--any
of these would signal potential ambiguity or conditionality.  - Think
about whether parts of the text reference other sections, creating a
need to follow up for understanding.  If such references are crucial
for full comprehension, the Cross-Dependent Sentences category is
likely applicable.  - Do not merely repeat the examples and key
linguistic indicators provided in the categories.  Instead, **focus
on explaining why the text you've been given fits into a specific
category**, based on its own language and structure.  - Consider only
the given text for classification, Do not imply or consider any text
that's not in the given text for classification.
### Text for Classification:  {input_text}
### Output Format:  - Return two lists, One is applicable categories
from the above mentioned categories only and the other is explanation
for the same in the same order of first list in the exact format
below.  Do not add additional text.
labels list - ["Category A", "Category B"] reasons list - ["Category
A" :  Detailed reason for why Category A applies to the given
text","Category B" :  Detailed reason for why Category B applies to
the given text"]
- If no categories apply to the given text for classification, return
labels list - [] reasons list - []
- Strictly follow this format in all cases.  The **number of elements
in both lists must be the same**.  If there are 2 categories, then
there must be 2 reasons.  The reason for each category should be in
the following format:  "Category Name:  Reason for why the category
applies".  - Ensure that the punctuation in the **reasons list**
is correct so that it will not disrupt the Python list format when
```

```
loaded.  Each reason should not have stray commas or unclosed quotes.
Use **double quotes** " only for values in both lists, not single
quotes '.  - The maximum number of categories in the labels list
should not exceed **4**, and the reasons list should have the **same
number** of elements.  - labels list should only **Categories that are
mentioned above**
```

Classification Prompt for Global context Annotation Task

```
You are a legal expert analyzing policy documents.  Classify the
following text into the most appropriate categories listed below,
using the given global context as a reference.  Return the output as
mentioned in Output Format below.
### Categories and Examples:
1.  **Conditional Sentences**:  Conditional Sentences state that
an outcome, benefit, or action depends on the fulfillment of
specific conditions.  When multiple conditions are implied or not
explicitly defined, the sentence can become complex, unclear, or
open to interpretation, increasing the risk of misinterpretation.  -
Potential linguistic indicators for reference:  depending, necessary,
appropriate, inappropriate, as needed, as applicable, otherwise
reasonably, sometimes, from time to time.  - Example:  'Currently, the
Special Surrender Value (SSV) is the same as the Guaranteed Surrender
Value.  The SSV may be revised from time to time with prior approval
of the Authority.'
2.  **Cross-Dependent Sentences**:  These refer to sentences that
rely on references to other sections of the document or external
entities for full understanding.  Cross-references often require
readers to navigate multiple sections, making the content harder to
follow and increasing the risk of misinterpretation.  - Potential
linguistic indicators for reference:  According to section 3, As
per IRDAI guidelines.  - Example:  'Coverage for losses arising from
natural disasters as defined under Section 2.1 shall only apply if the
policyholder has fulfilled the obligations specified in Section 3.4,
including the payment of additional premiums detailed in Appendix A.'
3.  **Legal Terminology**:  These sentences are challenging to
understand due to intricate phrasing or technical legal vocabulary.
Such complexity can blur the intended meaning, increasing the
likelihood of misinterpretation.  - Potential linguistic indicators
for reference:  Any legal term or phrase that depicts complex meaning
to understand.  - Example:  'In accordance with the provisions
outlined in this agreement, any claims for reimbursement submitted by
the policyholder must be accompanied by all requisite documentation,
including, but not limited to, certified copies of original receipts,
detailed proof of loss forms, and corroborating evidence from
third-party service providers, failing which the insurer reserves the
right, at its sole discretion, to deny said claims without further
obligation to notify the policyholder of deficiencies.'
4.  **Ambiguity in Expression**:  This class refers to sentences or
clauses in policy documents that are unclear or open to multiple
interpretations due to vague or imprecise language.  - Vague Terms
like 'some', 'many', or 'a few', which lack precise meaning and can
be interpreted differently by different readers.  - Modal Verbs like
'may', 'could', 'might, or 'should' that introduce uncertainty and
flexibility in the interpretation of the statement.  - Potential
```

```
linguistic indicators for reference: may, might, can, could, would,
likely, possible, anyone, certain, everyone, numerous, some, most,
few, much, many, various, including but not limited to - Example:
'The refund will be processed soon, but it may take a while.'
### Instructions: - **Do not repeat or refer to the example
sentences** provided under each category. These examples are meant to
illustrate the categories and should not be used to classify the input
text. Instead, focus on analyzing the language and structure of the
**input text** itself. - **Provide detailed and unique explanations**
for each classification based on the actual language used in the
input text. Do not simply paraphrase or repeat the examples. -
**First, identify key linguistic indicators** in the given text, such
as conditionals (e.g., 'may', 'could', 'shall'), cross-references
(e.g., 'Section 2.1', 'Appendix A'), or complex legal terms (e.g.,
'provisions', 'corroborating evidence'). - **The above mentioned
linguistic indicators are for reference only and there could be
more that applies to each category**. These are meant as guidelines
only. - **Use the Local Context result and Global Context** to help
determine which category fits best. The **local context result**
provides specific clues for the current text, while the **global
context** can help with broader interpretation, especially for complex
or ambiguous language. - **Do not infer information that isn't
present in the provided text**. Only use the provided input text and
the accompanying context for classification. - **Focus on how the
categories apply to the input text**, and explain why each category is
applicable, based on the specific words, phrases, and structure of the
input text.
### Text for Classification: {input_text}
### Local Context result: {result_of_local_context}
### Global Context: {global_context}
### Output Format: - Return two lists, One is applicable categories
from the above mentioned categories only and the other is explanation
for the same in the same order of first list in the exact format
below. Do not add additional text. - "'labels list - ["Category A",
"Category B"]\n reasons list - ["Category A : Detailed reason for why
Category A applies to the given text","Category B : Detailed reason
for why Category B applies to the given text"] "' - If no categories
apply to the given text for classification, return "'labels list - []
\n reasons list - [] "'
- Strictly follow this format in all cases. The **number of elements
in both lists must be the same**. If there are 2 categories, then
there must be 2 reasons. The reason for each category should be in
the following format: "Category Name: Reason for why the category
applies". - Ensure that the punctuation in the **reasons list**
is correct so that it will not disrupt the Python list format when
loaded. Each reason should not have stray commas or unclosed quotes.
Use **double quotes** '"' only for values in both lists, not single
quotes ''. - The maximum number of categories in the labels list
should not exceed **4**, and the reasons list should have the **same
number** of elements. - labels list should only **Categories that are
mentioned above**.
```

Classification Prompt for Rectification Task

```
Task: You are a legal expert, and you are provided with a
potentially misinterpreted paragraph extracted from a policy document.
```

Additionally, you will receive a set of clarification questions along with the answers to those questions, which are designed to help guide your understanding of the paragraph. Your task is to use these inputs to rectify the paragraph, ensuring that the final output accurately reflects the original policy intent, without misinterpretations or ambiguities.
Inputs: 1. Potentially Misinterpreted Paragraph (Input 1): {input_paragraph}
2. Clarification Questions and Answers (Input 2 & Input 3): Clarification Question 1: {question_1} Clarification Answer 1: {answer_1} Clarification Question 2: {question_2} Clarification Answer 2: {answer_2} ...and so on for additional pairs...
Task Instructions: - Using the clarification questions and their answers, carefully analyze the potentially misinterpreted paragraph.
- Rectify the paragraph to make sure it conveys the intended meaning with clarity, aligning it with the context provided in the clarification answers. - Ensure that the rectified paragraph maintains accuracy, is free of ambiguity, and reflects the true policy intent.
Output: A rectified version of the potentially misinterpreted paragraph, corrected based on the clarification inputs provided.