Explanation of Dataset Comprehensiveness

1. **Diverse Topics and Sections**:
    - The dataset includes a wide range of topics such as liability cover, making a claim, policy amendments, and exclusions. It covers various aspects of the insurance policy, ensuring the chatbot can handle queries from different sections of the document.
2. **Variety of Query Types**:
    - The queries range from specific coverage details (e.g., "What is covered under liability?") to procedural questions (e.g., "How do I make a claim?"). This variety ensures the chatbot is tested on different types of questions, including definitions, processes, and conditions.
3. **Realistic Scenarios**:
    - The queries reflect realistic questions a policyholder might ask, ensuring practical applicability. This includes questions about policy renewal, coverage specifics, and exclusions, which are common concerns for insurance customers.
4. **Comprehensive Coverage**:
    - The dataset spans various key features of the policy, including specific coverages (e.g., fire and theft, medical expenses), policy features (e.g., no claim discount protection, courtesy car provision), and procedural aspects (e.g., making a claim, policy cancellation).
5. **Balanced Distribution**:
    - The queries are spread across different sections and pages of the document, avoiding concentration on any single section. This ensures that the chatbot's performance is evaluated on the entire document rather than a subset of it.
6. **Inclusion of Detailed and Simple Queries**:
    - The dataset includes both detailed queries (e.g., "What does the legal expenses cover include?") and simpler ones (e.g., "Is personal belongings cover included?"). This tests the chatbot's ability to handle both in-depth and straightforward questions.
7. **Testing Contextual Understanding**:
    - Some queries require the chatbot to understand context and provide relevant information based on the document (e.g., "Are additional drivers covered?" or "What is the excess waiver policy?"). This tests the chatbot's comprehension and information retrieval capabilities.
8. **Potential Edge Cases**:
    - The dataset includes potential edge cases and less commonly asked questions (e.g., "Can the policy be suspended?" or "Are vehicle modifications covered?"). This ensures the chatbot can handle a broad spectrum of inquiries, including those that might not be present in the policy document.

Conclusion

This dataset is comprehensive because it tests the chatbot on a wide array of realistic, varied, and distributed queries. It ensures the chatbot's responses are accurate, relevant, and contextually appropriate across different types of insurance-related questions. This thorough testing helps gauge the chatbot's overall performance, identify any weaknesses, and improve its accuracy and reliability in answering user queries based on the insurance policy document.

How and Why I Chose These Evaluation Metrics

1. Factual Accuracy:

- **Reason**: The primary goal of the chatbot is to provide factually correct answers based on the information in the policy document. Evaluating factual accuracy ensures that the chatbot's responses are reliable and trustworthy.
- **Metric**: Each answer is graded as CORRECT or INCORRECT based on its factual accuracy compared to the ground truth.

2. Detailed Feedback:

- **Reason**: Simply marking an answer as CORRECT or INCORRECT is not sufficient for understanding how to improve the model. Detailed feedback helps identify specific errors and areas for improvement.
- **Metric**: Detailed explanations are provided for why an answer is correct or incorrect, which helps in understanding the model's shortcomings.

What Did I Try to Improve the Accuracy

1. Prompt Engineering:

- **Approach**: Carefully designed the prompt template to ensure that the evaluation model understands how to grade responses based on factual accuracy, ignoring minor differences in phrasing and punctuation.
- **Effect**: Improved the reliability of the grading process, ensuring that the focus is on the correctness of the information provided.

2. High-Quality Data Preparation:

- **Approach**: Ensured that the dataset used for evaluation covers a diverse range of topics and query types from the policy document. This includes both simple and complex queries, procedural questions, and specific coverage details.
- **Effect**: Provided a comprehensive evaluation of the chatbot's performance across various aspects of the document, helping to identify any gaps in the chatbot's knowledge.

3. Retrieval-Augmented Generation (RAG) Model:

- **Approach**: Combined a retriever and a generator to ensure that the chatbot can accurately fetch relevant information from the document and generate coherent responses.
- **Effect**: Enhanced the accuracy of the responses by ensuring that they are based on relevant document sections, reducing the likelihood of hallucinations (i.e., generating information not present in the document).

4. Manual Evaluation and Feedback Loop:

- **Approach**: Manually reviewed the chatbot's responses and the detailed feedback from the evaluation model to identify common errors and areas for improvement.
- **Effect**: Enabled iterative improvements to the model by refining the retriever, generator, and prompt templates based on observed issues.