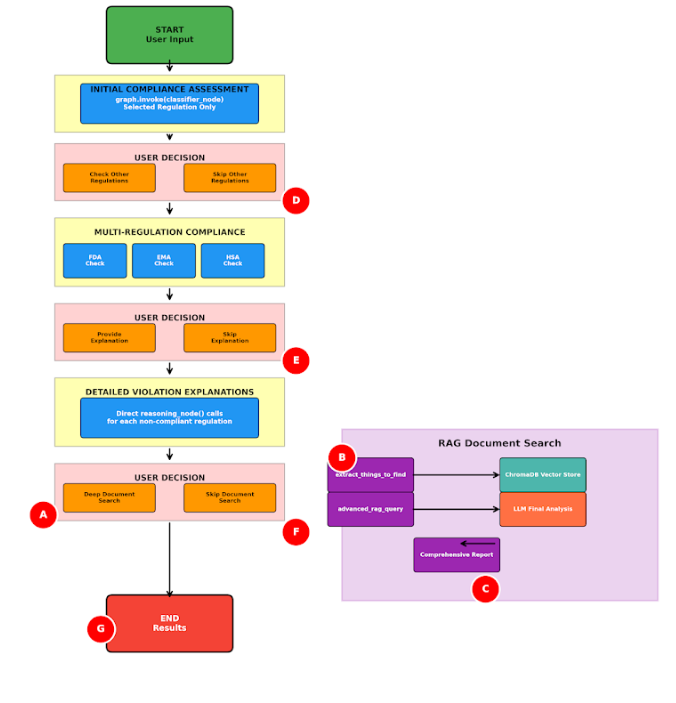
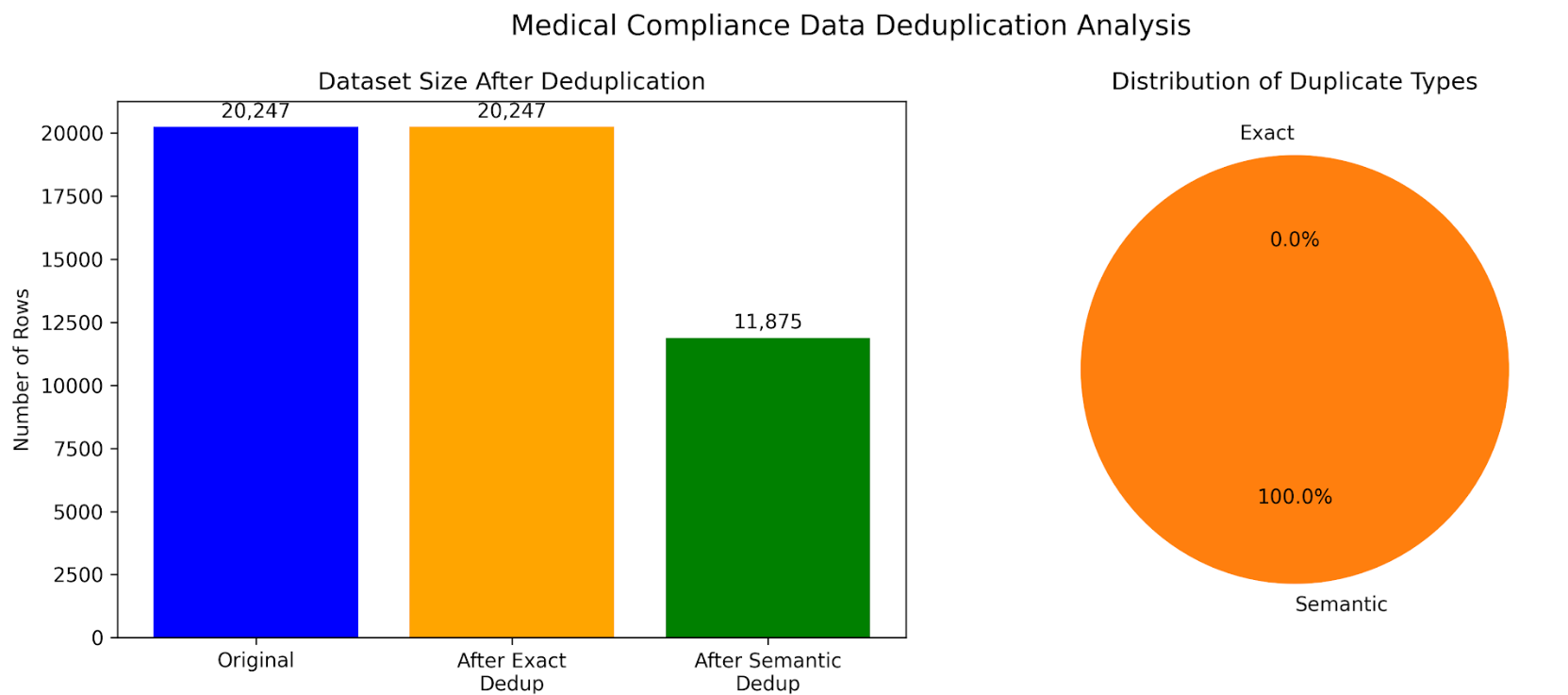
Workflow

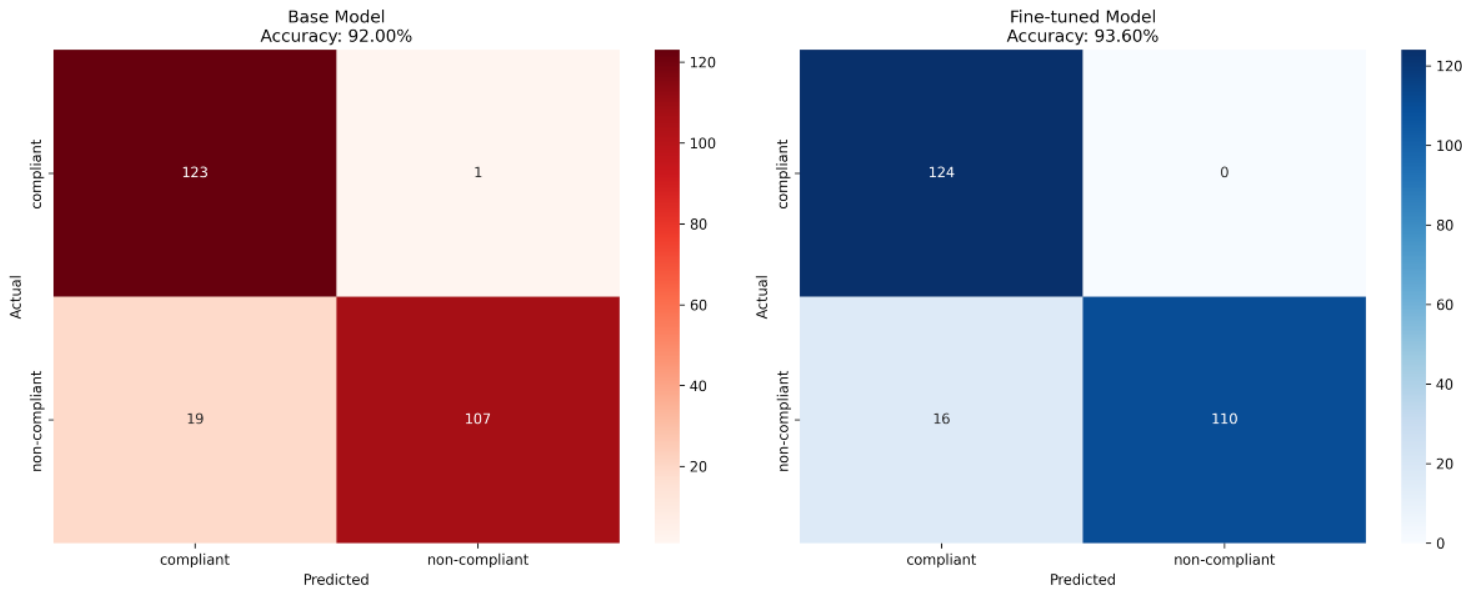


Data Deduplication



Considered datasets like rules & regulations PDF, and warning letters issued to the companies for their promotion. Extracted and processed these data to create a database providing in-depth information about domain knowledge, and later created a Q&A for fine-tuning a base model.

Evaluation Metrics



Qwen2.5-1.5B Model

Instead of using closed source model, this agent only uses open souced model (Qwen). Other base models can also be used from Ollama or other frameworks without changing the flow.\

The evaluation was done on test.json (a 250-text-class pair dataset with classes as compliance & non-compliance, which can be tested with more)

Key improvements which can be done:-

1. Make the chatbot flow better to consider the internet search as well (for more updated information extraction in the deep research section, instead of relying on the old documentation of rules and regulations).

**Why RAG?**

The rag was used in a more processed way:-

1. Finding what to search → 2) Query transformation → 3) extract for each query (say 3 for each) → 4) rerank them to select the best according to the user question (semantic similarity/LLM decision) → 4) Distill the selected docs' knowledge → 5) provide the final selected answer

Since this is a highly sensitive task, where a single rule may affect the whole product thus we need to ensure that the rules and regulations are thoroughly checked, because solely depending on LLM will open ways to hallucination and unreliable answers, there for we need some critic method/ RAG based retrieval/ an automated oracle system which can help llm to answer exact things according to rule. So we should use LLM at those places where it needs to frame an answer/ explain, etc. Instead of using them in a sensitive/decision-making process.