Qwen-based Virtual Assistant for Surplus Lines Insurance Tax & Compliance Guidance

Hemanth Sai Kosari
018223049
https://github.com/hemanthsai126/NLP-Project

1.Introduction

Surplus lines insurance regulation in the United States is highly decentralized, with each state defining its own tax rates, filing deadlines, penalties, affidavit requirements, stamping office rules, and procedural obligations for brokers and insureds. For compliance professionals, keeping track of these differences involves reviewing lengthy manuals, navigating multiple Department of Insurance (DOI) websites, and interpreting regulatory bulletins that frequently change. These manual processes create risk for errors, delays, and inconsistent reporting.

To address this challenge, our project introduces a Qwen-based Virtual Assistant designed to support users with structured, accurate, and consistently formatted compliance guidance across all U.S. states. The system integrates curated regulatory data, retrieval-augmented reasoning, fine-tuned LLM models, and controlled web search capabilities. It functions as an intelligent layer that abstracts away manual regulatory lookup while preserving factual accuracy and avoiding legal advisory behaviour. The Virtual Assistant focuses on generating educational, state-specific summaries that help users understand surplus lines obligations without replacing professional compliance processes.

2. System Architecture

The system architecture is designed as a hybrid compliance intelligence framework that combines deterministic database retrieval with advanced LLM reasoning. This approach ensures that all outputs remain factually grounded, numerically consistent, and aligned with the evolving regulatory landscape. Each component of the architecture contributes to accuracy, safety, and performance, collectively enabling the Virtual Assistant to generate reliable surplus lines compliance guidance.

2.1 Dataset Curation and Preprocessing

We extracted comprehensive regulatory information from the Surplus Lines Manual and supplemented it with web-scraped content from state DOIs and stamping offices. All regulatory text was normalized, cleaned, and converted into structured schemas suitable for Q&A fine-tuning. GPT-3.5 was then used to generate initial training templates that established a consistent answer format for the Virtual Assistant. The entire dataset is fully organized and version-controlled to ensure reproducibility and traceability.

- Extracted tax rates, filing deadlines, penalty schedules, affidavit requirements, and stamping fees from the **Surplus Lines Tax Manual 2025**, ensuring the most current regulatory coverage.
- Scraped additional compliance details from Department of Insurance (DOI) websites and Surplus Lines Association pages to ensure coverage of state-specific variations.
- Normalized and cleaned the raw text into machine-readable tables suitable for model training and database ingestion.
- Generated structured question-and-answer samples using GPT-3.5 to create a standardized fine-tuning format.
- Organized all curated data in a version-controlled directory optimized for the Unsloth training pipeline.
- Crawled through multiple regulatory and stamping office websites to capture edge-case rules, supplemental filing notes, and state-specific exceptions not explicitly documented in the manual.
- Consolidated all extracted and crawled text into unified Q&A sets, ensuring that every state's compliance rules were represented uniformly in the fine-tuning corpus.
- Verified each Q&A pair for numerical accuracy and textual clarity to reduce hallucinations during model training and evaluation.

This work ensures the Virtual Assistant is trained on a complete, accurate, and well-organized set of surplus lines regulatory rules.

2.2 Model Training

Initial model training was conducted using Unsloth fine-tuning on the curated Q&A dataset. While the model demonstrated strong conceptual understanding of regulations, early testing revealed weaknesses in numeric stability—particularly in recalling tax percentages and filing intervals. These baseline assessments established the foundation for subsequent improvements in prompting, dataset design, and tuning strategies.

- Performed full fine-tuning cycles using the structured Q&A dataset aligned with regulatory schemas.
- Established baseline accuracy metrics across tax rules, penalties, deadlines, and obligation-based reasoning.
- Identified recurrent numeric issues such as inconsistent percentages and misinterpreted date intervals.
- Tested multiple prompt variations to determine which formats yielded the most stable regulatory answers.
- Documented accuracy benchmarks to guide future iterations and tuning strategies.

These evaluations validated the overall model direction and provided concrete targets for refinement.

2.3 Prompt Engineering and Response Structuring

A specialized prompt engineering framework was developed to ensure structured, consistent, and low-variance outputs from the Virtual Assistant. Because surplus lines compliance relies heavily on precise numeric values and rule-based logic, prompts were designed to enforce formatting, guide extraction, and suppress unsupported reasoning.

- Designed structured Q&A templates for regulatory summaries, numeric lookups, and obligation-specific guidance.
- Implemented numeric extraction prompts to reduce hallucinations in tax rates, penalties, and deadlines.
- Applied chain-of-thought suppression techniques to avoid fabricated reasoning steps while preserving factual correctness.
- Created role-aware instructions for broker-based, insured-based, and combined queries to ensure state-accurate interpretations.
- Standardized answer formatting to improve consistency and readability across states.

This prompt framework significantly improves reliability and minimizes errors in regulatory responses.

2.4 Real-Time Web Search Integration

To keep regulatory knowledge current, a real-time web search module was integrated into the system. This tool retrieves updates from state DOIs, Surplus Lines Associations, and regulatory bulletins, ensuring the Virtual Assistant remains aware of changes without relying solely on static sources.

- Developed predefined, sanitized search templates that safely query regulatory terms (e.g., tax updates, bulletins, filing notices).
- Implemented automated retrieval pipelines for recent tax changes, compliance alerts, and new filing requirements.
- Filtered search results for relevance and integrated them into the LLM's context for up-to-date reasoning.
- Ensured the search layer complements the static database, offering both historical and newly updated regulatory insights.
- Added safeguards to prevent unsafe or adversarial search queries.

This integration enables the system to remain dynamically aligned with current surplus lines regulations.

3. Results and Observations

During testing, I evaluated the Virtual Assistant using both the smaller Qwen-3B model and the larger Qwen-14B model to assess their performance across numerical reasoning, multi-step compliance logic, formatting consistency, and safety behaviour. The goal was to understand how each model handled state-specific surplus lines rules and to identify the strengths and limitations of each model in real compliance scenarios. The evaluation revealed clear differences between the two models, particularly in reasoning depth, numeric accuracy, stability, and safety adherence.

3.1 Performance of the Small Model (Qwen-3B)

The Qwen-3B model performed well on simple, fact-based queries that required retrieving a single value from the SQLite database. It consistently returned accurate tax rates, affidavit requirements, and straightforward filing rules. Its smaller size resulted in faster response times and lower inference cost, making it suitable for high-volume, low-complexity questions.

- Handled simple questions such as tax rates and affidavit requirements with high accuracy.
- Returned responses quickly due to its smaller parameter size and lower computational load.
- Struggled with multi-state comparisons, nuanced penalty structures, and conditional logic.
- Produced incomplete or partially correct numeric reasoning, especially when multiple regulatory conditions were involved.
- Required stronger prompting to maintain structured formatting and complete explanations.

3.2 Performance of the Larger Model (Qwen-14B)

The Qwen-14B model showed significantly stronger performance on complex compliance tasks. It produced multi-paragraph summaries, retained context across extended prompts, and interpreted multi-state regulations with fewer errors. The model also demonstrated improved numeric stability and cleaner formatting consistency.

- Delivered detailed explanations for conditional penalties and cross-state comparisons.
- Maintained numeric accuracy across tax rates, penalties, and deadline structures more effectively than the smaller model.
- Followed structured Q&A templates consistently without requiring heavy prompting.
- Demonstrated stronger understanding of role-based distinctions (broker vs. insured).

3.3 Security and Safety Observations

An important part of the evaluation involved testing how each model behaved when given potentially unsafe, adversarial, or legally sensitive prompts. The results clearly differentiated the two models in terms of safety robustness.

- The Qwen-14B model refused or safely deflected all unsafe regulatory, legal-advice, and injection-style prompts during testing.
- The Qwen-3B model answered all unsafe prompts, including attempts to elicit prohibited legal instructions or bypass compliance constraints.
- The smaller model showed vulnerability to prompt injection and hallucinated regulatory interpretations when given adversarial questions.
- The larger model correctly followed safety-aligned prompting, demonstrating stronger alignment and internal policy adherence.

These results indicate that the large model is significantly more secure and resistant to misuse, while the small model remains vulnerable without additional safety layers.

3.4 Differences between models

Category	Qwen-3B (Small Model)	Qwen-14B (Large Model)
Accuracy vs. Complexity	Accurate on simple, single-state factual queries	Accurate on complex regulatory logic and multi-step reasoning
Numerical Reasoning	More prone to misreporting percentages and filing deadlines	Maintains high numeric accuracy and consistent values
Context Handling	Loses context in longer or multi-part prompts	Retains context across multiple states and conditional rules
Formatting Reliability	Needs strong prompting to maintain structure	Follows templates reliably with minimal formatting drift
Role-Specific Reasoning	Less consistent distinguishing broker vs. insured duties	Correctly interprets and separates role-based obligations
Handling Exceptions	Misses edge cases and regulatory exceptions	Captures exceptions and nuanced state differences more accurately

3.5 Potential Reasons for Performance Differences

The observed performance gap aligns with established LLM scaling behaviour. Larger models generally outperform smaller ones because of greater representation capacity and more extensive pretraining.

- The 14B model stores richer representations of regulatory logic, numeric structures, and formatting behaviour.
- Larger models generalize better on conditional reasoning, making them more reliable for interpreting penalties, filing rules, and exceptions.
- The 14B model can track multiple states, deadlines, and penalties within a single prompt without losing coherence.
- Small models hallucinate values more often due to limited internal precision; the 14B model showed much greater numeric reliability.
- The 14B model adhered more strictly to structured templates, whereas the 3B model drifted or produced incomplete answers.

These reasons explain why the larger model consistently outperforms the smaller one across all complex tasks and safety evaluations.

4. Conclusion and Future Work

In this project, I designed and implemented a comprehensive Virtual Assistant capable of providing surplus lines insurance compliance guidance through a hybrid architecture that integrates a structured regulatory database, dual-model LLM inference, real-time web search, and a multi-layer safety framework. The system successfully retrieves accurate tax rates, filing deadlines, penalties, and obligations while maintaining strong formatting consistency and regulatory grounding. Evaluation results demonstrated that while the smaller Qwen-3B model performs well on simple lookup tasks, the larger Qwen-14B model is essential for complex reasoning, numeric stability, and safe response behaviour. Together, these components form a reliable and extensible foundation for delivering educational compliance support.

Future development will focus on enhancing accuracy, scalability, and robustness. Planned improvements include expanding the dataset with additional regulatory sources, applying advanced hyperparameter tuning, and integrating a full retrieval-augmented generation (RAG) pipeline to further reduce hallucination. Additional work will also include strengthening safety filters, improving adversarial resistance, and optimizing routing mechanisms between the two models. With these enhancements, the Virtual Assistant will be able to deliver even more precise, up-to-date, and trustworthy guidance across diverse regulatory scenarios.