REGULAILTION

PROJECT:

Regulatory texts are lengthy, highly technical, and require strict factual accuracy. General-purpose LLMs often hallucinate and fail to capture legal nuance, making them unreliable for regulatory reasoning.

Goal: accurate, grounded Q&A over banking regulations.

Solution: Domain-aware RAG + FT models.

The Data:

- Official Bank of Israel banking & financial regulations.
- Synthetic Q&A generation to ensure coverage and consistency.
- Questions designed to test precision, grounding, and edge cases.

MODELS & METRICS

Models Evaluated:

- **Llama + MiniLM** – strong general baseline.
- **Llama + LegalBERT** – domain-aware embeddings.
- **SaulLM + MiniLM** – legal LLM with generic retrieval.
- **SaulLM + LegalBERT** – full legal-domain stack.

Metrics Used:

- **Verdict Accuracy** – correct regulatory conclusion.
- **Hit@K** – relevant source retrieved in top-K.
- **Citation Accuracy -** verifies that the model's answer is correctly supported by the cited regulatory source.
- **Hallucination Rate** – unsupported or fabricated claims.

NOVELTY & ACHIEVEMENTS

Project Novelty:

- Systematic comparison of legal vs general model stacks in a regulatory RAG setting.
- Introduced citation-level evaluation for regulatory correctness.
- Combined synthetic legal Q&A generation with real regulatory texts.
- Focused on trustworthiness, not just language fluency.

Project Achievements:

- Built a full end-to-end RAG system for regulatory question answering.
- Benchmarked general vs legal LLMs and embeddings under identical conditions.
- Designed a quantitative evaluation framework beyond generic accuracy.
- Demonstrated measurable reductions in hallucination using domain-aware retrieval.

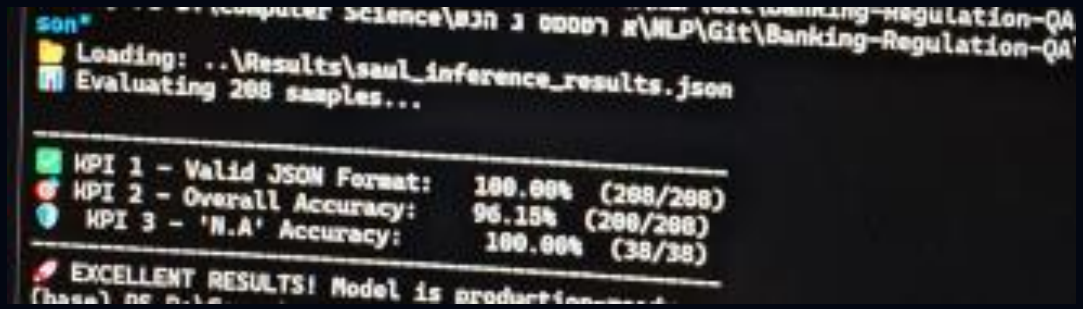
METHODOLOGY

Part 1
Fine Tuning:

Updating the model itself to optimize it for the task.

- **Domain Specialization**:- Enhanced the model's regulatory expertise
- **Robustness to Negatives** - Trained the model on hard and soft negative samples to improve its ability to identify irrelevant information and avoid false positives.
- **Structured Output Optimization:** Fine-tuned the model to ensure consistent and valid JSON formatting

- **Model Selection:** Evaluated and compared **Llama 3.1** and **Saul-7B (**pre-trained on the legal domain).
- **Dataset Construction:** Compiled a dataset of **2,000+ examples**, including **15-20% negative samples** to enhance robustness and prevent hallucinations.
- **FT KPI'S:**



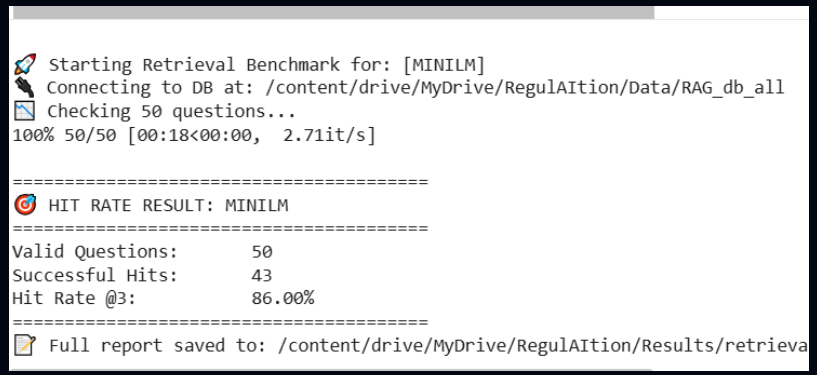
METHODOLOGY

Part 2
RAG:

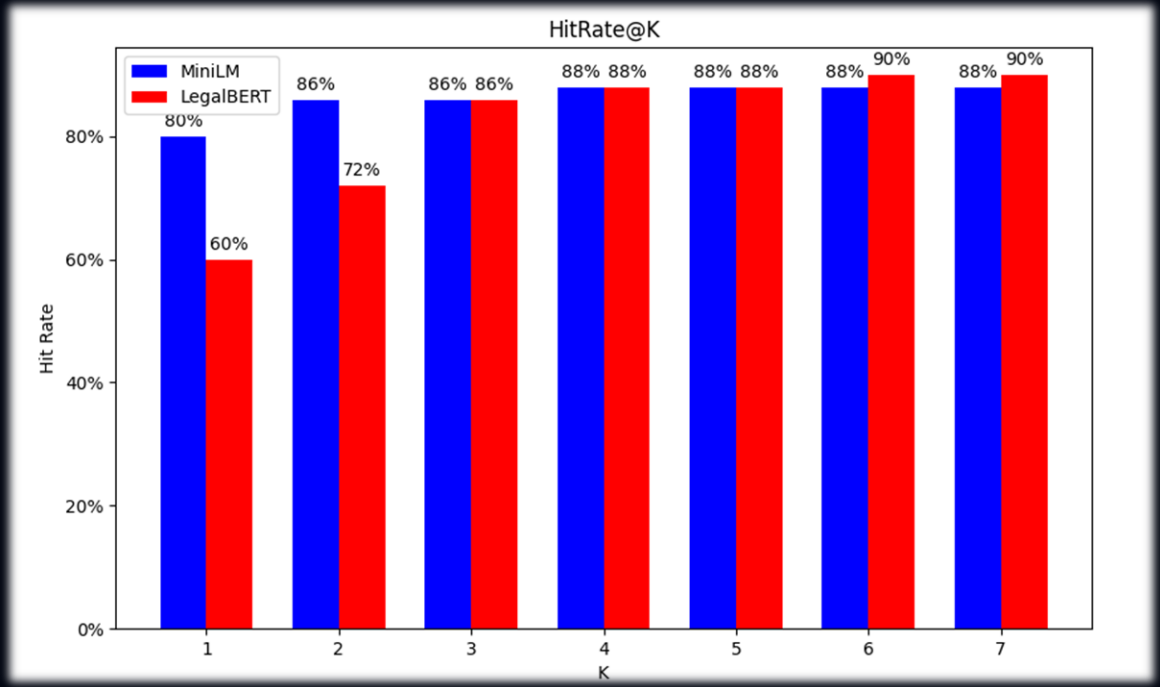
Providing the model with external data to ensure factual accuracy.

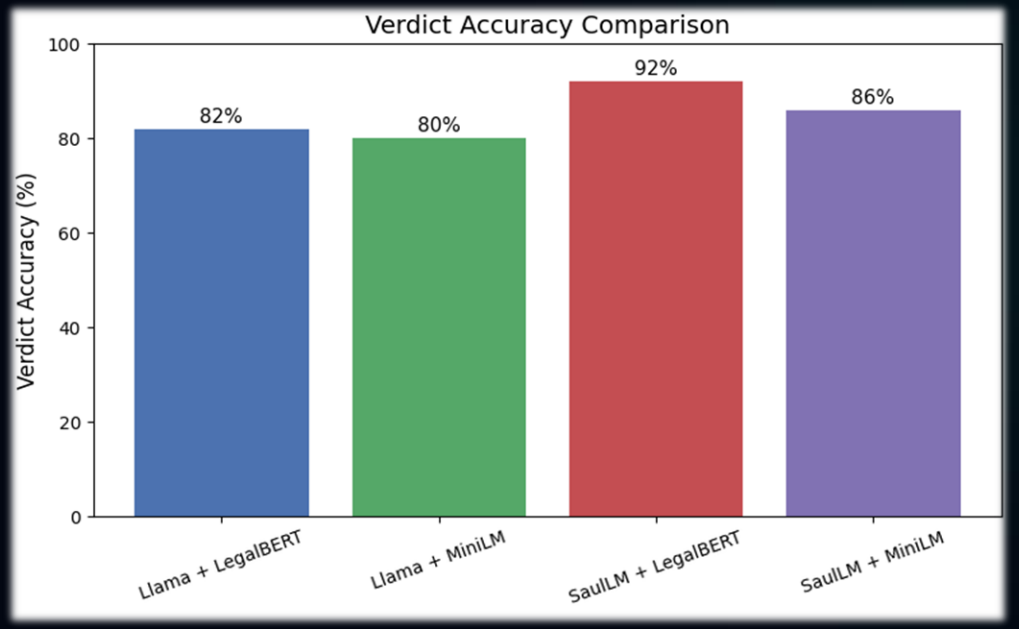
- **Context Injection:** Supplies relevant regulatory text for every query.
- **Evidence-Based:** Grounds all answers in **official sources** (B.O.I) to eliminate hallucinations.
- **Dynamic Knowledge:** Accesses the latest documents without the need for retraining.

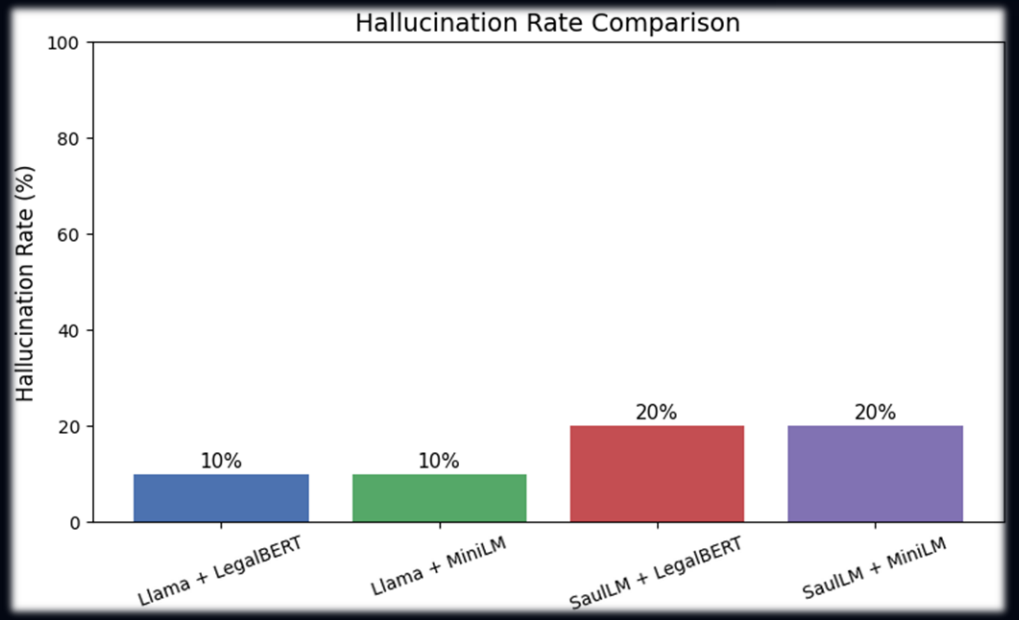
- **Vector Database:** Engineered a vector database using **ChromaDB**, consisting of **518 text chunks** extracted from official regulations.
- **Dual-Encoder Strategy:** Compared **all-MiniLM** and **LegalBERT** to determine the most effective embedding model for banking and legal terminology.
- **Retrieval Evaluation:** Measured **Hit@K (HitRate)**



RESULTS


HitRate@K
MiniLM
LegalBERT
Hit Rate
K
0%
20%
40%
60%
80%
1
2
3
4
5
6
7
80%
60%
86%
72%
86% 86%
88% 88%
88% 88%
88% 90%
88% 90%


Verdict Accuracy Comparison
Verdict Accuracy (%)
0
20
40
60
80
100
82%
80%
92%
86%
Llama + LegalBERT
Llama + MiniLM
SaulLM + LegalBERT
SaulLM + MiniLM


Hallucination Rate Comparison
Hallucination Rate (%)
0
20
40
60
80
100
10%
10%
20%
20%
Llama + LegalBERT
Llama + MiniLM
SaulLM + LegalBERT
SaulLM + MiniLM

CONCLUSION

Looks Good… But Can Be Better:

- Minor hallucinations are still present.
- The model identifies the correct chunk in most cases, but consistency is not guaranteed.
- In a high-stakes domain like this, even small errors are unacceptable.

Things We (Painfully) Learned:

- Each model comes with its own pros and cons.
- Hallucination handling needs to be significantly stricter.
- A diverse set of KPIs is needed to properly evaluate model quality.

Where Do We Go From Here?

- Experiment with additional models and architectures.
- Introduce more rigorous validation and consistency checks in the RAG pipeline.
- Add a faithfulness metric using an LLM judge to verify citation-explanation alignment.