Experimentation, Metrics, and Results

1. Purpose

This section provides a detailed overview of how experiments were conducted, how model outputs were evaluated, how scores were computed and aggregated, and how results were analyzed. It serves both as a technical reference for reproducibility and as a research record of the evaluation pipeline and findings.

2. Experimental Setup

Models Evaluated

Five large language models (LLMs) were benchmarked in this study:

- **Llama 3.3 70B (Versatile)**
- **Llama 3.1 8B (Instant)**
- **Llama 4 Maverick 17B**
- **OpenAI GPT-120B (OSS variant)**
- **Qwen 3.2 32B**

Each model was tested under three distinct **prompting strategies**:

- **Zero-shot** – direct question answering without any example.
- **Few-shot**—inclusion of a few annotated examples as contextual guidance.
- **Chain-of-thought (CoT)**—structured prompts that encourage explicit reasoning before providing answers.

3. High-Level Experimental Pipeline

1. **Data ingestion:** A curated dataset of legal contracts was transformed into evaluation examples. Each example consisted of a question and its corresponding expected field or clause.
2. **Model inference:** Every model–strategy pair was applied to each example to generate a predicted answer. Response time and metadata were recorded.
3. **Per-field evaluation:** Each predicted field was compared to its reference answer using multiple comparison metrics (exact match, containment, semantic similarity).
4. **Per-row aggregation:** Field-level scores were averaged to produce a single average_score for each evaluation example.
5. **Global aggregation:** All examples were grouped by model and prompting strategy to compute summary statistics (mean, median, standard deviation, and counts).
6. **Analysis and visualization:** Results were visualized through heatmaps and bar charts to compare model accuracy, response time, task-wise performance, and strategy effectiveness.

4. Scoring Methodology

Each target field was evaluated using three complementary metrics:

Metric	Description	Weight
Exact Match Score	Binary match after normalization (1 = exact match, 0 = no match)	0.4
Contains Score	Partial match if the reference text is contained within prediction or vice versa	0.3
Semantic Similarity	Cosine similarity between embedding vectors of prediction and reference	0.3

Combined field-score formula: field_score=0.4(exact_match)+0.3(contains)+0.3(semantic_similarity)

All metrics are normalized within a 0–1 range, where higher values indicate better performance.

5. Per-Row average_score Computation

Each evaluation example contains multiple fields.
If an example has *N* target fields with individual field scores s1,s2,...,sN, the **average per-row score** is defined as:

$$\text{average\_score} = \frac{1}{N}\sum_{i=1}^{N} s_i$$

This value represents the mean quality of a single model's prediction for one example. It is stored in results/experiment_results.csv alongside all metadata

6. Aggregation of Overall Results

The **Overall Results** table is obtained by grouping all per-row average_score values by **model** and **strategy**, then computing the arithmetic mean for each combination:

grouped_mean=mean(average_score for model × strategy)

Additional descriptive statistics (standard deviation, median, sample count) were computed to support deeper analysis.
These aggregated results provide the foundation for all summary tables and plots.

7. Interpreting the Numbers

- Scores range approximately from **0 to 1**, where higher values indicate greater agreement with reference answers.
- A difference of **0.01–0.02** may represent a meaningful gap depending on dataset size.
- **Spread matters:** similar means can conceal variability; therefore, standard deviations and counts should always be considered.

- The best-performing model–strategy pair balances **accuracy** and **response latency**.
- Pairwise significance tests (e.g., Wilcoxon or bootstrap confidence intervals) are recommended for robust comparison between strategies.

8. Quantitative Results

Overall Performance by Model and Strategy

Model	Chain-of-Thought	Few-Shot	Zero-Shot
Llama 3 70B	0.178	0.150	0.145
Llama 3 8B	0.178	0.183	0.108
Llama 4 Maverick	0.184	0.180	0.136
OpenAI GPT 120B	0.170	0.132	0.114
Qwen 32B	**0.184**	**0.194**	0.142

Key Observations

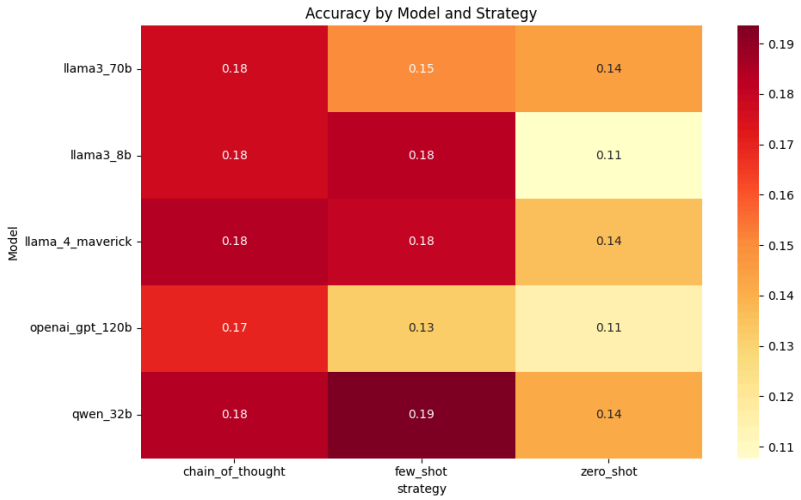
- The **Qwen 32B (Few-Shot)** configuration achieved the highest average score (≈ 0.194).
- **Llama 4 Maverick (Chain-of-Thought)** closely followed with ≈ 0.184.
- Across all models, **Zero-Shot** prompting consistently yielded the lowest performance, confirming the importance of reasoning cues or contextual examples.

9. Visual Analysis

(All plots referenced below correspond to figures generated during analysis.)

9.1 Accuracy by Model and Strategy

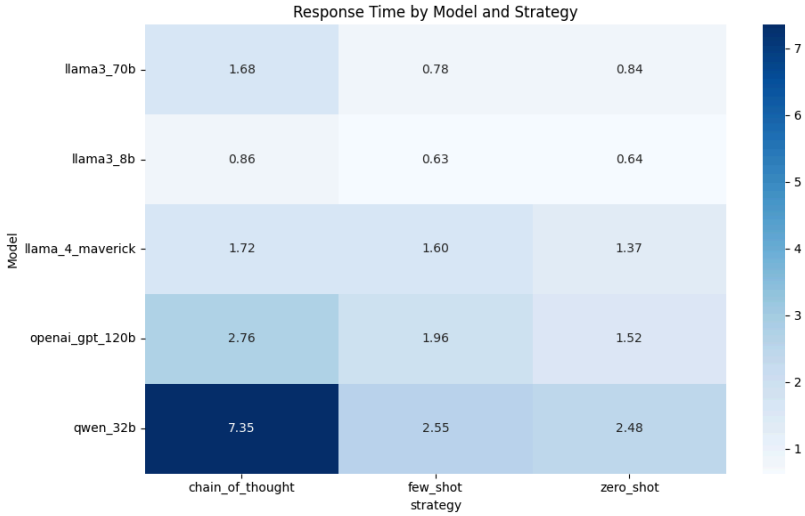
Heatmap visualizations show that **few-shot** and **Chain-of-Thought** strategies outperform **zero-shot** across all models. Qwen 32B and Llama 4 Maverick stand out as the most consistent performers.



9.2 Response Time Analysis

Response-time heatmaps reveal a trade-off:

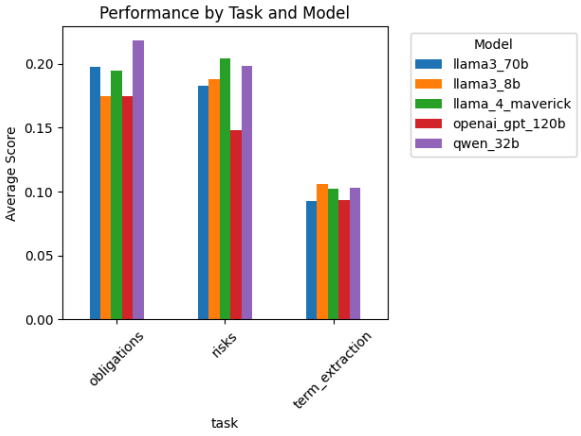
- **Qwen 32B** achieved top accuracy but was the slowest (up to 7 s per CoT response).
- **Llama 3 models** were the fastest (< 1 s for Few-Shot) with only a modest performance penalty.



9.3 Task-Wise Performance

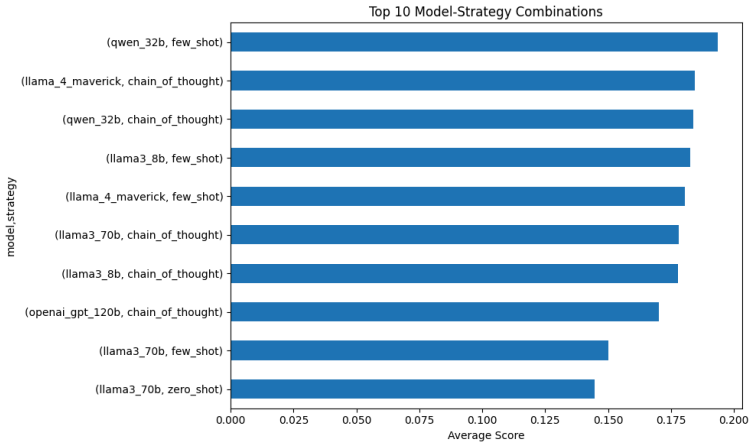
Bar charts across **obligations**, **risks**, and **term extraction** tasks show that:

- Models excelled in **risk identification**, where clause phrasing is relatively standardized.
- **Term extraction** tasks had the lowest scores, likely due to text variability and ambiguity.



9.4 Top Model–Strategy Combinations

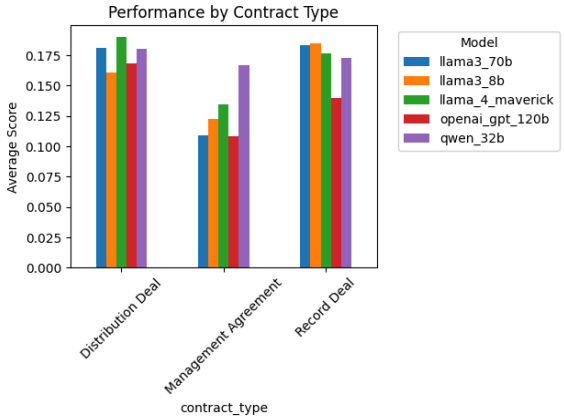
Ranking plots of the top 10 combinations confirm that **Qwen 32B (Few-Shot)** leads overall, followed closely by **Llama 4 Maverick (CoT)** and **Qwen 32B (CoT)**.
This underscores the positive effect of structured reasoning and contextual examples.



9.5 Performance by Contract Type

Across contract categories:

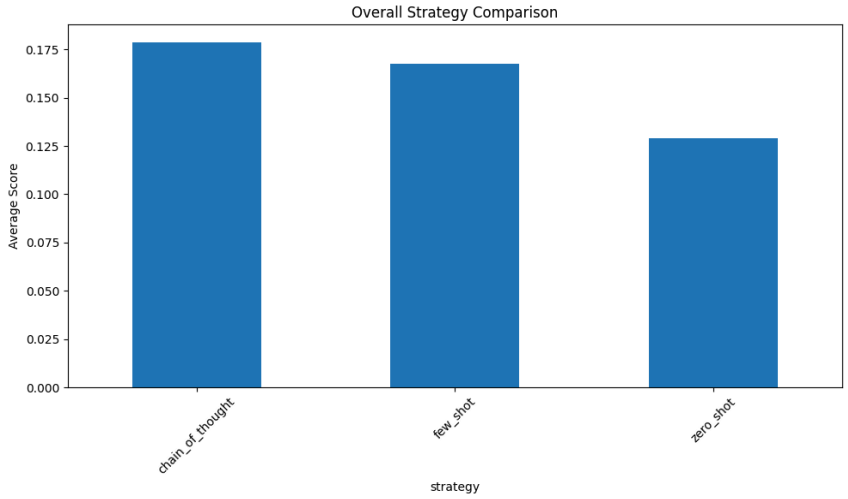
- **Management Agreements** yielded the highest accuracy, possibly due to consistent clause structure.
- **Record Deals** exhibited greater variation, suggesting that non-standard phrasing increases difficulty.



9.6 Overall Strategy Comparison

Aggregated across all models:

- **Chain-of-Thought** (≈ 0.175) and **Few-Shot** (≈ 0.172) perform nearly equally well.
- **Zero-Shot** (≈ 0.13) trails behind, indicating that reasoning guidance—explicit or example-based—significantly improves performance.



10. Discussion

1. **Prompting strategy drives performance:** The gap between Few-Shot/CoT and Zero-Shot strategies indicates that **reasoning scaffolding** is more influential than model scale.
2. **Model efficiency trade-offs:** Although Qwen 32B attained the highest accuracy, its longer response times suggest potential optimization for deployment scenarios.
3. **Task sensitivity:** Extraction-based tasks remain challenging. Future work should explore span-based evaluation metrics or hybrid approaches combining rule-based and LLM inference.
4. **Cross-model generality:** Consistent ranking patterns across models suggest the evaluation pipeline is reliable and captures genuine differences rather than random fluctuations.

11. Reproducibility and Verification

All experiments can be reproduced using the existing pipeline by re-running the analysis scripts provided in the repository.
Verification scripts confirm that each stored average_score value equals the arithmetic mean of its constituent field scores, ensuring data integrity and consistency across runs.
To guarantee reproducibility, environment versions, random seeds, and data preprocessing steps should be logged and documented.

12. Suggested Future Analyses

- Produce **per-field performance tables** to identify weak points for each model.
- Report **standard deviation, median, and confidence intervals** in future summaries.
- Conduct **pairwise significance tests** to quantify whether observed differences are statistically meaningful.
- Incorporate **weighted field aggregation** if certain clauses are legally more critical.
- Explore **Pareto analysis** combining accuracy and latency to identify optimal trade-offs.
- Integrate **OCR quality checks** for scanned or low-resolution documents.

13. Caveats and Limitations

- Consistency in embedding models for semantic similarity is crucial; changing embedding models alters comparability.
- Poor document parsing or OCR failures can depress scores and must be checked prior to evaluation.
- If datasets contain private or proprietary documents, all sensitive data must be anonymized before sharing.
- Small dataset size or unbalanced contract types may limit statistical significance.

14. Conclusion

This experimental framework demonstrates a reproducible methodology for evaluating large language models on structured legal-text understanding tasks.
Results show that **Few-Shot** and **Chain-of-Thought** prompting substantially enhance performance compared with Zero-Shot, regardless of model size.
The **Qwen 32B Few-Shot** configuration achieved the overall best results, while **Llama 4 Maverick (CoT)** offered a balanced trade-off between accuracy and inference time.
Future extensions will include field-level diagnostics, statistical testing, and integration of latency-accuracy optimization for production settings.