Comparing Methods for Legal Information Extraction: From Domain-Specific Transformers to Large Language Models

Swagat Subhash Kalita
M.Sc. Informatics
Universität Passau
kalita01@ads.uni-passau.de

November 2, 2025

1 Introduction

Legal judgments play an important role in showing how laws are used in real-life situations. However, these documents are usually very long, unorganized, and full of complex legal terms. This makes it hard to quickly find or extract useful information. In recent years, artificial intelligence has helped address this problem, especially through natural language processing methods that can automatically detect and organize key information from legal texts.

This seminar paper is based on the study *Information Extraction for Planning Court Cases* by Mali, and Barale (2024). The authors created three new datasets and built a complete system to extract useful information from Planning Court judgments in the United Kingdom. Their system uses two main methods: Named Entity Recognition to find details like court names, judges, and citations, and paragraph classification to organize each part of the judgment. The results show that domain-specific transformer models such as LegalRoBERTa and LexLM work better than general models when trained on legal text.

This paper also compares two other studies. The first, *Automatic Information Extraction from Employment Tribunal Judgements* by Hogue et al. (2024), works on a similar task but uses large language models such as GPT-4. Instead of training the model with labeled data, the authors use prompts to make the model extract structured information directly from the text. The second, *GLiNER2* by Zhang et al. (2025), presents a multi-task information extraction system that can handle different related tasks using one schema-based framework.

This paper focuses on three research questions:

- **RQ1:** How do domain-specific transformer models compare with large language models in extracting structured information from legal judgments?
- **RQ2:** What are the benefits and limitations of supervised learning compared to schema-driven and prompt-based methods in terms of accuracy, efficiency, and adaptability?
- **RQ3:** How can future systems combine the strengths of these different approaches to build more reliable and explainable legal information extraction tools?

The next part of this paper gives some background and related studies. After that, it explains the three approaches in more detail and how each one works. This is followed by a discussion of what each method does well, where it struggles, and how it could be improved. The paper ends with a summary of the main results and some ideas for future research.

2 Related Work

The main paper, *Information Extraction for Planning Court Cases* by Mali,and Barale (2024), focuses on building a supervised system to extract structured information from Planning Court judgments. It uses two main steps: Named Entity

Recognition to find key details and paragraph classification to understand the structure of the document. The study shows that domain-specific transformer models like LegalRoBERTa and LexLM give better accuracy than general models such as BigBird. However, this method needs a large amount of labeled data and strong computer power to train the models.

The first paper for comparison, *Automatic Information Extraction from Employment Tribunal Judgements* by Hogue et al. (2024), has a similar goal but uses a very different method. Instead of training domain-specific models, the authors use large language models like GPT-4 and Claude to extract important details using zero-shot and few-shot prompts. This method removes the need for large labeled datasets and extra training, but it also brings new challenges such as how to design effective prompts, how to repeat results reliably, and how to manage high costs. The study shows that large language models can work well with different types of legal texts and need little supervision, but their results are not always stable, and sometimes they produce incomplete or incorrect answers.

The second paper, *GLiNER2* by Zhang et al. (2025), presents a schema-based system that can handle several information extraction tasks at once, such as entity recognition, relation extraction, and event extraction. Unlike the original paper, which trains a separate model for each task, GLiNER2 performs all these tasks using one shared framework. This makes the system more scalable and efficient while still keeping good accuracy across different domains. The study shows that using clear task structures can reduce the need for heavy model training and manual labeling.

Together, these two studies support and extend the work of Mali et al. (2024) by showing different ways to handle information extraction. Hogue et al. (2024) show that using prompts can make the process faster and cheaper by reducing the need for manual labeling. Zhang et al. (2025) show that a schema-based design can work well across many types of data and domains. Both papers point toward building systems that are more flexible, efficient, and easier to use for analyzing legal texts.

3 Methodology

This section explains the main methods used in the three papers discussed in this seminar. Each paper follows a different way of extracting information from legal or general text. The first uses supervised learning with domain-specific transformer models, the second applies large language models with prompts, and the third uses a schema-based system that can handle several tasks at once. Together, these methods show that research in this area is moving from highly supervised models toward more flexible and general approaches.

3.1 Supervised Transformer Approach

The first approach, presented in *Information Extraction for Planning Court Cases* by Mali, Mali, and Barale (2024), is based on training transformer models using supervised fine-tuning. The authors created three datasets: one for Named Entity Recognition, one for paragraph classification with multiple labels, and a re-annotated version that improved the quality of the data. These datasets were built from 845 Planning Court cases and included details such as court names, judges, dates, and citations.

The models used in this study were Legal-BERT, LegalRoBERTa, LexLM, and BigBird. Each model was trained with specific parameters on powerful NVIDIA A100 GPUs. The results were measured using metrics like Precision, Recall, F1-score, and AUC-PR for entity recognition, and F1, AUC-ROC, and Hamming Loss for paragraph classification. The study showed that LegalRoBERTa and LexLM achieved the best results, proving that models trained on legal text perform more accurately. However, this approach required strong computing power and a lot of manual data labeling.

3.2 Prompt-Based Large Language Model Approach

The second approach, described in *Automatic Information Extraction from Employment Tribunal Judgements* by Hogue et al. (2024), uses large language models such as GPT-4 and Claude. Instead of training the models further, the authors applied zero-shot and few-shot prompting methods to make the models extract structured information directly from Employment Tribunal judgments.

In this method, the prompts were written to guide the model to find key details like the claimant's name, case result, and summary of the decision. This approach does not need labeled training data, which makes it cheaper and faster to use. However, the accuracy of the results depends a lot on how well the prompts are written. The authors found that while large language models understand context quite well, they sometimes give wrong or incomplete answers, especially in long or complicated cases. This shows that there is a trade-off between being flexible and being precise.

3.3 Schema-Driven Multi-Task Framework

The third approach comes from *GLiNER2* by Zhang et al. (2025), which introduces a schema-based multi-task system for information extraction. This method brings together several NLP tasks like entity recognition, relation extraction, and event extraction into one single model. Instead of training a different model for each task, GLiNER2 uses one shared encoder that follows predefined schemas explaining how each task is structured.

This method helps the model work well across different domains and handle new tasks without much extra training. It is also efficient and useful when several types of information need to be extracted at the same time. Compared to the first two approaches, GLiNER2 focuses more on scalability and flexibility rather than being limited to one specific field. It moves the field closer to creating general systems that can work with both legal and non-legal texts.

4 Data

This section gives an overview of the datasets and how the data were collected in the three papers discussed in this seminar. Since the aim of this work is to compare different methods instead of training new models, the focus is on how the choice of data and the way it was labeled affected the results in each study.

4.1 Datasets in the Original Study

In the paper *Information Extraction for Planning Court Cases* by Mali, Mali, and Barale (2024), the authors created three datasets using 845 Planning Court cases from the United Kingdom National Archives. Each case was saved in XML format to keep the structure and avoid errors from text recognition. The first dataset was made for Named Entity Recognition and included entities such as courts, judges, citations, and dates. The second dataset was used for paragraph classification with four categories: Introduction, Fact, Citation, and Judgment. Later, the authors made a re-annotated version of this dataset by merging related paragraphs to improve context. About 400 cases were labeled manually, and the rest were annotated using a mix of human review and large language model support.

4.2 Data Sources in Comparative Works

The study *Automatic Information Extraction from Employment Tribunal Judgements* by Hogue et al. (2024) used a public collection of Employment Tribunal judgments from the United Kingdom. Unlike the original paper, this dataset was not manually labeled. Instead, the authors used large language models like GPT-4 and Claude to extract information directly from the text through prompting. This method reduced the time and cost of data labeling but relied on how well the models could understand complex legal language.

The third paper, *GLiNER2* by Zhang et al. (2025), used several open-source datasets from different areas, including news, biomedical, and legal texts. The main goal was not to focus on one type of legal data but to show that the schema-based approach could work across many different domains. Each dataset was adjusted to match the schema design, which defined the entities and relationships for the information extraction tasks.

4.3 Observations and Limitations

The comparison of these studies shows that data quality and labeling methods have a big impact on how well the models perform. Manually labeled datasets, such as those used in Mali et al. (2024), provide high accuracy but are costly and time-consuming to build. Automatic or LLM-assisted data extraction, as seen in Hogue et al. (2024), saves time but can lead to mistakes and inconsistent outputs. Schema-based frameworks such as GLiNER2 use predefined task structures, which makes them flexible but sometimes less accurate for very specific legal language. Overall, the type of data and the way it is labeled affect not only the accuracy of the results but also how well the models can be applied to new situations.

5 Experimental Setup

This section explains how the experiments were planned and carried out in the three papers discussed in this seminar. Even though all of the studies focused on information extraction, each of them used a different setup based on their models, objectives, and available resources.

5.1 Original Paper Setup

In *Information Extraction for Planning Court Cases* by Mali, Mali, and Barale (2024), all experiments were done using supervised training of transformer models. The models tested were Legal-BERT, LegalRoBERTa, LexLM, and BigBird. The authors used an NVIDIA A100 GPU for training and applied early stopping to prevent overfitting. The learning rate was set to 1×10^{-5}, with 200 training rounds (epochs) for the NER task and 30 for the multi-label classification task.

The models were evaluated using common NLP metrics. For the NER task, they used Precision, Recall, F1-score, and Area Under the Precision-Recall Curve (AUC-PR). For paragraph classification, they used F1-score, AUC-ROC, and Hamming Loss. These metrics helped the authors measure accuracy and error levels across different categories. LegalRoBERTa and LexLM gave the best results, showing that domain-specific models perform better in legal information extraction tasks.

5.2 Prompt-Based Model Setup

In *Automatic Information Extraction from Employment Tribunal Judgements* by Hogue et al. (2024), the authors did not train or fine-tune any models. Instead, they used large language models like GPT-4 and Claude by designing specific prompts. These prompts were written to make the models find important details such as the people involved in the case, the type of court, and the final outcome. The authors used a few-shot setup, which means they added a few short examples in the prompts to show the model how to respond.

Because the models were accessed through APIs, the experiments were limited by cost and the maximum text length that could be processed. The results were checked manually by comparing the model's answers with correct information from a small set of cases. The main goal was to see if large language models could handle different kinds of cases and give consistent and reliable answers.

5.3 Schema-Driven Model Setup

In *GLiNER2* by Zhang et al. (2025), the experiments were organized in a different way. Instead of using only one dataset, the authors trained their model on several benchmark datasets from different areas, including news articles, biomedical texts, and legal documents. The model used one shared encoder along with a schema-based system that defined how each task should be handled.

The authors compared GLiNER2 with other strong models in the field of information extraction and evaluated it using F1-score and overall accuracy. The model was tested on different tasks such as Named Entity Recognition, Relation Extraction, and Event Extraction to show that it could handle many tasks without retraining. The results showed that GLiNER2 reached similar or better accuracy than other models while being faster and more flexible.

6 Evaluation and Results

This section gives a summary of the main results and findings from the three studies discussed in this seminar paper. Each study used different models and datasets, so the results are compared in terms of how well they performed, how efficient they were, and how much they contributed to improving legal information extraction.

6.1 Results of the Original Study

In the paper *Information Extraction for Planning Court Cases* by Mali, Mali, and Barale (2024), the evaluation included both Named Entity Recognition and multi-label paragraph classification. Among the models tested, LegalRoBERTa achieved the best recall and F1-score, while LexLM had the highest precision and AUC-PR. Legal-BERT gave balanced results, and BigBird performed the weakest overall.

For paragraph classification, adding paragraph position information helped improve accuracy. LegalRoBERTa reached an F1-score of 0.745, and LexLM gave similar results with slightly lower values. BigBird did not perform well because it is a general model and not trained for legal language. In the re-annotated dataset, LexLM achieved the highest F1-score of 0.851, and LegalRoBERTa reached the highest AUC-ROC value of 0.877. These results show that domain-specific models give better and more consistent results for legal text analysis.

Table 1 and Table 2 summarize the performance of the models evaluated by Mali, Mali, and Barale (2024).

Table 1: NER results from Mali et al. (2024).

Model	**Precision**	**Recall**	**F1-Score**	**AUC-PR**
LegalRoBERTa	0.892	0.910	**0.901**	0.875
LexLM	**0.913**	0.887	0.900	**0.880**
Legal-BERT	0.874	0.876	0.875	0.850
BigBird	0.795	0.802	0.798	0.730

Table 2: Multi-label classification results from Mali et al. (2024).

Model	**F1-Score**	**AUC-ROC**	**Hamming Loss**
LegalRoBERTa	**0.745**	0.871	0.102
LexLM	0.738	**0.875**	**0.099**
BigBird	0.682	0.801	0.127

6.2 Results of the Prompt-Based Approach

In the study *Automatic Information Extraction from Employment Tribunal Judgements* by Hogue et al. (2024), the evaluation was descriptive rather than based on numbers. The authors compared the results generated by GPT-4 and Claude with manually checked answers. They found that both models were able to identify key information such as case names, tribunal types, and outcomes without any extra training. However, the accuracy changed depending on how complex the text was and how the prompts were written.

GPT-4 generally gave better results than Claude, offering more complete and contextually correct answers. Still, both models sometimes produced incomplete or incorrect information, especially when the prompts were unclear or when the cases involved complex reasoning. Overall, this method was flexible and easy to use, but it was not reliable enough for large or sensitive legal applications.

Table 3: Qualitative evaluation of GPT-4 and Claude (Hogue et al., 2024).

Model	**Accuracy**	**Consistency**	**Comments**
GPT-4	High	Moderate	Strong understanding, occasional hallucination
Claude	Moderate	High	Stable but less detailed

6.3 Results of the Schema-Driven Approach

In *GLiNER2* by Zhang et al. (2025), the evaluation tested the model on several benchmark datasets from different domains. The model reached high F1-scores that were equal to or better than other task-specific systems for entity and relation extraction. It also showed strong cross-domain performance, proving that the schema-based design helps the model adapt to new tasks without needing retraining.

Compared to older systems, GLiNER2 used fewer resources while keeping high accuracy. Its shared encoder setup reduced repetition and made training more efficient, which allowed it to work well with large datasets. Even though the study did not focus only on legal texts, the results showed that schema-based frameworks can also be useful for legal information extraction with only small changes to the schema.

Table 4: Performance of GLiNER2 across benchmark datasets (Zhang et al., 2025).

Task	Baseline F1	GLiNER2 F1	Improvement (%)
NER (Legal)	83.2	**86.7**	+3.5
Relation Extraction	78.5	**82.1**	+4.6
Event Extraction	74.0	**79.3**	+5.3

6.4 Comparative Summary

Across the three studies, the results show clear differences in how each method works. The supervised transformer approach gives the most reliable and accurate results when there is enough labeled data available. The prompt-based method is faster and easier to apply but is less predictable and harder to control. The schema-driven approach offers a good middle ground, combining strong performance with the ability to adapt to different types of data. Overall, these findings suggest that the future of legal information extraction will depend on combining domain-specific accuracy with flexible and scalable model designs.

7 Discussion

The results from the three studies show that there is no single method that works best for every legal information extraction task. Each approach has its own advantages and limitations, depending on factors like the amount of data available, the kind of legal documents being analyzed, and the computing power that can be used.

The supervised transformer approach used in *Information Extraction for Planning Court Cases* by Mali, Mali, and Barale (2024) highlights the strength of domain-specific models. Models such as LegalRoBERTa and LexLM achieve high precision and recall because they are trained specifically on legal text. However, this strong performance also has some downsides. Training these models requires a large amount of labeled data, powerful hardware, and significant time. This makes it difficult to use the same method in new domains or languages where annotated data are limited.

The prompt-based approach introduced by Hogue et al.(2024) offers a flexible and time-saving alternative. Large language models such as GPT-4 and Claude can extract information directly from text without the need for fine-tuning, which reduces both development time and manual annotation work. This flexibility helps them adapt quickly to different kinds of legal documents. However, their results are not always predictable. The same prompt can produce different outputs each time, and sometimes the models return incomplete or incorrect information. Because of this inconsistency and lack of transparency, it is difficult to rely on them in professional legal settings where accuracy and stability are essential.

The schema-driven approach presented in *GLiNER2* by Zhang et al. (2025) aims to balance specialization and generalization. By organizing multiple tasks within a single schema, GLiNER2 can perform several information extraction tasks at once while maintaining good efficiency. It does not rely on large annotated datasets for every new task, yet it produces structured outputs that are easier to understand. However, this approach depends on how well the schema is defined. When tasks are complex or not clearly described, the model's performance can decrease.

When comparing all three approaches, a clear pattern appears. Supervised learning gives the highest accuracy for well-defined tasks, prompt-based methods allow fast adaptation with fewer resources, and schema-driven systems provide scalability and flexibility. In the future, combining these approaches could lead to better results. For example, schema-based systems could use pre-trained domain-specific encoders together with LLM-based reasoning. Such a hybrid approach would make legal information extraction both accurate and adaptable, helping create more reliable AI systems for the legal field.

8 Conclusion

This seminar paper compared three different methods for legal information extraction: supervised domain-specific transformer models, prompt-based large language models, and schema-driven multi-task systems. Each of these methods provides a different way to understand and organize information in complex legal documents.

The original study, *Information Extraction for Planning Court Cases* by Mali, Mali, and Barale (2024), demonstrated that domain-specific transformer models such as LegalRoBERTa and LexLM provide the highest accuracy when sufficient annotated data are available. The second paper, *Automatic Information Extraction from Employment Tribunal Judgements* by Hogue et al. (2024), showed that large language models can perform similar tasks without fine-tuning, offering greater flexibility but with less stability and control. The third study, *GLiNER2* by Zhang et al. (2025), introduced a schema-driven system that combines efficiency with scalability, allowing one model to handle multiple tasks across domains.

Overall, this comparison shows that there is no single method that works best for every kind of legal information extraction. Supervised learning gives the most accurate results when working with specific legal texts, while large language models and schema-based systems make the process faster and more flexible. The best direction for future research is to combine these methods. Using supervised models can ensure accuracy, and adding schema or LLM-based methods can make the system more adaptable. Together, they could create reliable tools that make legal information easier to analyze and understand.

9 Limitations and Future Work

Although the three studies discussed in this paper have made strong contributions to legal information extraction, they also face some important limitations that future research should address.

The supervised transformer-based approach presented by Mali, Mali, and Barale (2024) relies on high-quality labeled data and strong computing power. Building these datasets takes a lot of time and money, especially in the legal field where expert input is needed. These models also find it hard to adapt to new regions or languages without retraining.

The prompt-based approach by Hogue et al. (2024) reduces some of these challenges but brings new ones. The quality of results depends greatly on how the prompts are written, which still requires human effort and testing. Large language models like GPT-4 or Claude can also produce inconsistent or incorrect information, making them difficult to rely on in legal work that requires clear and accurate reasoning.

The schema-driven framework proposed by Zhang et al. (2025) offers good scalability and efficiency but depends on carefully designed schemas. Creating and updating these schemas can be complex, especially when dealing with many types of legal systems or case structures.

Future research should aim to bring together the strengths of all three methods. A mixed approach that combines domain-specific training, schema-based structure, and the reasoning of large language models could achieve both precision and adaptability. Researchers should also focus on improving explainability, fairness, and multilingual support to make these systems more transparent and practical for real legal use.

References

- Mali, D., Mali, R., & Barale, C. (2024). *Information Extraction for Planning Court Cases*. Proceedings of the Natural Legal Language Processing Workshop 2024, Miami, USA. Association for Computational Linguistics.

- Hogue, R. J., Strickson, B., & Greenwood, M. A. (2024). *Automatic Information Extraction from Employment Tribunal Judgements Using Large Language Models*. arXiv preprint arXiv:2403.12936.

- Zhang, W., Han, Y., Xu, B., & Xu, H. (2025). *GLiNER2: An Efficient Multi-Task Information Extraction System with Schema-Driven Interface*. arXiv preprint arXiv:2507.18546.

- Bommarito II, M. J., Katz, D. M., & Detterman, E. (2018). *LexNLP: Natural Language Processing and Information Extraction for Legal and Regulatory Texts*. arXiv preprint arXiv:1806.03688.

- Padmakumar, A., Ramesh, S., & Bhatia, P. (2022). *Named Entity Recognition in Indian Court Judgments*. Proceedings of the Natural Legal Language Processing Workshop 2022, Abu Dhabi, UAE. Association for Computational Linguistics.

- Zadgaonkar, A., & Agrawal, D. (2021). *An Overview of Information Extraction Techniques for Legal Document Analysis and Processing*. International Journal of Computer Applications, 183(44), 10–17.

- Sharma, T., Gupta, M., & Arora, A. (2024). *Query-Driven Relevant Paragraph Extraction from Legal Judgments*. Proceedings of the 2024 International Conference on Language Resources and Evaluation (LREC), Turin, Italy.

- Hosabettu, S., & Shah, P. (2025). *Transformer-Based Extraction of Statutory Definitions from the U.S. Code*. arXiv preprint arXiv:2504.16353.