SemEval-2024 Task 5: Argument Reasoning in Civil Procedure

Lena Held[1] and **Ivan Habernal**[2]
Trustworthy Human Language Technologies
[1] Department of Computer Science, Technical University of Darmstadt
[2] Department of Computer Science, Paderborn University
lena.held@tu-darmstadt.de, ivan.habernal@uni-paderborn.de
www.trusthlt.org

Abstract

This paper describes the results of SemEval-2024 Task 5: Argument Reasoning in Civil Procedure, consisting of a single task on judging and reasoning about the answers to questions in U.S. civil procedure. The dataset for this task contains question, answer and explanation pairs taken from *The Glannon Guide To Civil Procedure* (Glannon, 2018). The task was to classify in a binary manner if the answer is a correct choice for the question or not. Twenty participants submitted their solutions, with the best results achieving a remarkable 82.31% F_1-score. We summarize and analyze the results from all participating systems and provide an overview over the systems of 14 participants.

1 Introduction

"Arguing a legal case is an essential skill that aspiring lawyers must master. This skill requires not only knowledge of the relevant area of law, but also advanced reasoning abilities, such as using analogy arguments or finding implicit contradictions." – (Bongard et al., 2022)

In order to test these abilities, we organized the SemEval-2024 Task 5: Argument Reasoning in Civil Procedure. By basing our dataset on an established textbook in the domain of U.S. civil procedure (*The Glannon Guide To Civil Procedure*, (Glannon, 2018)), we ensure that we can leverage the high quality and refined content aimed at law students to create a challenging task in the competition. The book follows the philosophy, that learning about civil procedure can be achieved by reading about a given topic and answering questions afterwards. Therefore, each chapter is accompanied by a set of hard reasoning problems formulated as multiple-choice questions. As a teaching resource, the book provides a thorough analysis for each answer candidate. This enables the student to learn by example.

We frame our task in a simple manner: classifying whether the given answer is a correct solution to the question or not. With this task, we want to put the legal reasoning capabilities of various state-of-the-art models to the test and provide a reliable benchmark.

2 Related work

As the task is based upon our previous paper (Bongard et al., 2022), we refer to the detailed related work section in there. In a nutshell, legal question answering is an inherently difficult task because it requires both reasoning skills and expertise. Legal question datasets in NLP are scarce and vary in terms of the specific topics covered, such as the U.S. Multistate Bar Examination (Fawei et al., 2016), Tax Law (Holzenberger et al., 2020), and Japanese Bar Exams (Kano et al., 2019; Rabelo et al., 2022). Although existing datasets focus on finding the correct answer to the question posed, the reasoning behind a correct or incorrect answer is often ignored. More recently, LLMs have found their way into legal question answering, demonstrating their potential in this area (Katz et al., 2023) by solving complex legal questions at a level comparable to humans. But these circumstances also highlight the need for appropriate tasks to evaluate such systems (Guha et al., 2023).

3 Dataset

The dataset was collected by parsing *The Glannon Guide To Civil Procedure* (Glannon, 2018) which was done in our previous work (Bongard et al., 2022). The details of the data collection and baseline methods are also outlined there. Instead of treating the questions from the book as multiple choice queries, we decided to pair each answer with its question and attach a binary label for a correct or incorrect conclusion. Because there are usually multiple incorrect answers to a question, the dataset is highly imbalanced towards incorrect answers. A question can either be a stand-alone sentence or in cloze text form. To make the context

Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024), pages 2027–2038
June 20-21, 2024 ©2024 Association for Computational Linguistics

Question 7. A switch in time. Yasuda, from Oregon, sues Boyle, from Idaho, on a state law unfair competition claim, seeking $250,000 in damages. He sues in state court in Oregon. Ten days later (before an answer is due in state court), Boyle files a notice of removal in federal court. Five days after removing, Boyle answers the complaint, including in her answer an objection to personal jurisdiction. Boyle's objection to personal jurisdiction is

Answer not waived by removal. The court should dismiss if there is no personal jurisdiction over Boyle in Oregon, even though the case was properly removed.

Solution 1

Analysis D is the correct answer. Boyle has not waived his objection to personal jurisdiction. If the federal court lacks jurisdiction over Boyle, it should dismiss the case, even though it was properly removed.

Complete Analysis There are so many ways to go astray on this issue [...].

Introduction My students always get confused about the relationship between removal to federal court and personal jurisdiction [...].

Figure 1: Example data point

of most of the questions clear, there is an introduction text which provides background information to the question. In addition, Glannon has written further explanatory texts which justify why the answer was a correct choice or not. Each data point in the dataset consists of *question, answer candidate, solution, analysis (answer), complete analysis (all answers to the question), introduction*. An example data point is presented in Figure 1.

However, the dataset version used in the competition differs slightly from the original version. To correct errors in the initial version of the dataset, we removed a mistakenly included chapter of the book. Additionally, we corrected two instances in which the explanation text was missing. Although the dataset size changed, the partitions are still based on the paradigm used in (Bongard et al., 2022), resulting in a training partition (666), development partition (84), and testing partition (98). The *rational data split* is meant to sort questions which appear later in each chapter into the test set, assuming that these questions are harder to answer than earlier ones. To conceal the labels in the test set, we eliminated both fields *label* and *analysis* in that partition.

3.1 Potential question leakage from dev to test

When splitting the dataset partitions, we created some unwanted potential leakage. In particular, some questions that appear in the test partition may have already been part of the development partition with a different answer candidate. This occurred because each partition should contain questions from each chapter and data points were not considered as questions with multiple answer candidates, but rather as question-answer pairs. Because some dataset requests had already been answered, we chose not to readjust the partitioning. The training partition is not affected by this. About 27 of 98 data points in the test partition are affected and due to the small size of the dataset, we chose not the remove the data points either.

Instead, we take this opportunity to analyze if the behavior of the participating systems differs in regards to the leaked questions. The details of this additional analysis are presented in section 6.2. However, a future version of the dataset will contain a modified split that fixes the issue.

4 Task description

Reasoning is still one of the hardest task state-of-the-art models and techniques can face. Simply understanding language is certainly not enough to understand expert legal questions, much less answer them correctly. The task is meant to probe the capacity of methods for understanding complex legal topics and applying them in exemplary scenarios. However, to avoid over-complicating the output and evaluation, the task is formulated as a simple yes or no question. By default this approach also makes the task harder, because there is no option to find one correct answer by process of elimination. The task remains the same as introduced by Bongard et al. (2022):

Task Given a question with a possible correct answer and a short introduction to the topic of the question, identify if the answer candidate is correct or incorrect.

Although systems may use the analysis that is provided in the training and development partitions for enhancement, they should be able to produce a prediction based on introduction, question and answer candidate alone.

4.1 Evaluation methods

Due to the simplicity of the task itself, we consider standard metrics to be best suited to evaluate the submissions. We calculate the macro F_1-score to account for the dataset imbalance between correct and incorrect answers. We evaluate the accuracy as well as an additional point of comparison. The F_1-score is the relevant evaluation metric for the competition ranking.

As a baseline, we provide a simple majority baseline which predicts each answer as incorrect and achieves an F_1-score of 42.69%.

4.2 Organization

We setup the competition on the CodaLab platform.[1] Participants needed to register first and acquire the dataset by filling out the required form as agreed with the publisher of the book[2]. We sent out the training and development partitions of the dataset first. The practice phase of the competition was officially accessible from November 28th, 2023 to allow participants to get accustomed to the submission platform and upload their scores for the development set. The test partition was sent out on January 9th, 2024 via email to those who had previously requested the dataset. Between January 10th, 2024 and February 1st, 2024 (00:00:00 UTC), participants could upload up to 5 submissions in total. After the end of the evaluation phase, participants could still upload contrastive runs in the post-evaluation phase with the same evaluation script.

5 Participant systems

During the competition period, we received 59 requests for the dataset. Of the 55 participants who registered on the CodaLab platform, 20 submitted results in the evaluation phase. We summarize and evaluate the 14 teams that submitted system papers.

[1] https://codalab.lisn.upsaclay.fr/competitions/14817

[2] https://github.com/trusthlt/legal-argument-reasoning-task

Rank	Participant	Acc.	F_1
1	HW-TSC	0.8673	0.8231
2	MAINDZ	0.8265	0.7747
3	SU-FMI	0.8367	0.7728
4	qiaoxiaosong	0.8163	0.7644
5	UTSA-NLP	0.7959	0.7315
6	kubapok	0.7857	0.6971
7	LegalSense	0.7449	0.6599
8	hrandria	0.6939	0.6327
9	Yuan_Lu	0.6327	0.6000
10	PengShi	0.6735	0.5910
11	Mistral	0.5714	0.5597
12	Hwan_Chang	0.5918	0.5556
13	kriti7	0.6020	0.5511
14	woody	0.6633	0.5510
15	odysseas_aueb	0.6122	0.5143
16	SCaLAR Group, NITK Surathkal	0.6224	0.4966
17	lhoorie	0.5000	0.4957
18	yms	0.7245	0.4827
19	U_201060	0.6633	0.4503
20	langml	0.4490	0.4375
21	majority baseline	0.7449	0.4269

Table 1: Official Leaderboard, counting the last submission made by a participant.

In addition to the descriptions, we present a brief summary of the key features of the proposed systems in Table 3.

5.1 Leaderboard results

We allowed participants to make up to 5 submissions in the evaluation phase to encourage them to try out several approaches. For the official leaderboard, which is taken from CodaLab, only the last valid submission is counted, resulting in the ranking shown in Table 1. We have also created a leaderboard that counts the best submission instead of the last one. This leaderboard variant is shown in Table 2. The differences between the leaderboard rankings are minimal. Both leaderboards are available on the competition webpage[3].

5.2 System descriptions

The systems mostly rely on established LLMs like GPT-4 (OpenAI, 2023), Llama (Touvron et al., 2023a) or Llama 2 (Touvron et al., 2023b), Mistral (Jiang et al., 2023) or Mixtral (Jiang

[3] https://trusthlt.github.io/semeval24/

Rank	Participant	Acc.	F_1
1	HW-TSC	0.8673	0.8231
2	MAINDZ	0.8265	0.7747
3	SU-FMI	0.8367	0.7728
4	qiaoxiaosong	0.8163	0.7644
5	UTSA-NLP	0.8061	0.7341
6	kubapok	0.7857	0.6971
7	LegalSense	0.7449	0.6599
8	hrandria	0.6939	0.6327
9	PengShi	0.6837	0.6166
10	Yuan_Lu	0.6327	0.6000
10	Hwan_Chang	0.6735	0.6000
12	Mistral	0.5714	0.5597
13	kriti7	0.6020	0.5511
14	woody	0.6633	0.5510
15	SCaLAR Group, NITK Surathkal	0.6429	0.5238
16	odysseas_aueb	0.6122	0.5143
17	lhoorie	0.5000	0.4957
18	yms	0.7245	0.4827
19	U_201060	0.6633	0.4503
20	langml	0.4490	0.4375
21	majority baseline	0.7449	0.4269

Table 2: Leaderboard, counting the best submission made by a participant.

et al., 2024), Zephyr (Tunstall et al., 2023) or Flan-T5 (Longpre et al., 2023). Other popular models are Legal-BERT (Chalkidis et al., 2020), RoBERTa (Liu et al., 2019), Longformer (Beltagy et al., 2020) and Big Bird (Zaheer et al., 2020). Many teams explore different strategies to prompt the LLMs, for instance using Chain-of-Thought (Wei et al., 2022).

Rank 1: HW-TSC – Self-Eval? A Confident LLM System for Auto Prediction and Evaluation for the Legal Argument Reasoning Task (Zhao et al., 2024) This team uses different GPT-4 prompt designs and strategies alongside a self evaluation approach leveraging a confidence score. Their best-performing system remodels the task into a multiple-choice question answering task and uses an ensemble of 3 runs. The authors' experiments show that prompting the LLM for a confidence score improves the performance in all tested settings. Their results also highlight that remodeling the task into a multiple-choice question answer task improves the performance significantly.

Rank 2: MAINDZ – CLUEDO - Choosing Legal oUtcome by Explaining Decision through Oversight (Benedetto et al., 2024) This team took an interesting approach by employing a two-stage decision process. In the first step, an ensemble of three models is fine-tuned with all available information (introduction, questions, answer cast as multiple-choice task) and not only generates the correct predictions, but also the explanations. In the second step, these generated candidates are evaluated by another zero-shot system (a 'detective') which chooses the final solution (given the labels and the explanations).

Rank 3: SU-FMI – From BERT Fine-Tuning to LLM Prompt Engineering - Approaches in Legal Argument Reasoning (Krumov et al., 2024) The authors experimented with a large number of approaches, starting with fine-tuning BERT-based models, adding external fine-tuning data, over to utilizing commercial LLMs with prompt engineering. The best results were achieved by utilizing GPT-4 and legal prompt engineering (prompts tailored for legal reasoning tasks). This team also provides a thorough comparison with other, partly open-source models.

Rank 5: UTSA-NLP – Prompt Ensembling for Argument Reasoning in Civil Procedures with GPT4 (Schumacher and Rios, 2024) This team uses the analysis part as a Chain-of-Thought mechanism in in-context learning. In particular, they prompt GPT-4 which, given the intro, question, and the answer candidate at test time, also generates the analysis part and the final label. The final system is an ensemble model combining several variants of the base models. The authors also provide an error analysis, showing that longer introductions tend to confuse the models.

Rank 7: NLP at UC Santa Cruz – Legal Answer Validation using Few-Shot Multi-Choice QA (Pahilajani et al., 2024) This team analyzed several fine-tuning strategies based on BERT models, or the effects of integrating additional CaseHold data, but concludes that multi-choice QA few-shot prompting on GPT-4 was the most effective method in their experiments.

Rank 9: 0x.Yuan – Enhancing Legal Argument Reasoning with Structured Prompts (Lu and Kao, 2024) The team investigates several prompting strategies on Mixtral-8x7B in a zero-shot manner which make use of established legal reasoning methodologies like the IRAC (Issue, Rule, Application, Conclusion) analysis. The authors note that prompt designs tailored to legal reasoning methods outperform Chain-of-Thought strategies and direct prompting.

Rank 10: YNU-HPCC – Regularized Legal-BERT for Legal Argument Reasoning Task in Civil Procedure (Shi et al., 2024) The approach by this team employs fine-tuning of Legal-BERT and other BERT models and overcomes the input limitations by applying sliding window approaches. On top of comparing several losses (Cross-Entropy, Focal, Dice), they also compare the use of Regularized Dropout and Supervised Contrastive Learning for data augmentation and imbalances.

Rank 11: Mistral – Mistral 7B for argument reasoning in Civil Procedure (Siino, 2024) This team tested the pre-trained LLM Mistral-7B in a zero-shot prompting manner to classify a given question-answer pair.

Rank 13: Transformers – Legal Argument Reasoning Task in Civil Procedure using RoBERTa (Singhal and Bedi, 2024) The approach proposed by this team fine-tunes a pre-trained RoBERTa model with all input fields available in the training data and further uses minority sampling to counter the dataset imbalances.

Rank 14: ignore – A Legal Classification Model with Summary Generation and Contrastive Learning (Sun and Zhou, 2024) The team uses a Legal-BERT classifier with a contrastive learning approach. They additionally shorten the introduction text by summarizing it with GPT-3.5 and augment the training data by concatenating parts of the input in different ways. The authors note that generative summarization proves feasible to handle the introduction text and the contrastive loss improves the robustness of the model.

Rank 15: Archimedes-AUEB – LLM explains Civil Procedure (Chlapanis et al., 2024) This team proposes extending the training data by synthetic data generated by GPT-3, where the generated data resemble Chain-of-Thought reasoning. The authors also fine-tune a student model, an open-source Llama-2-7b, with QLoRA and provide an expert-based analysis, which reveals some shortcomings in explanations of the models.

Rank 16: SCaLAR NITK – Towards Unsupervised Question Answering system with Multi-level Summarization for Legal Text (Prabhu et al., 2024) The team tried various approaches using Word2Vec, GloVe and Legal-BERT embeddings to identify the most likely answer in a multiple-choice setup based on similarity scores. Additionally, they employ a segment-wise summarization of the introduction text with T5 and investigate the differences in similarity scores between the summarized and original input. The approach relies on open-source models and is reproducible.

Rank 17: eagerlearners – The Legal Argument Reasoning Task in Civil Procedure (Sabzevari et al., 2024) This team experimented with different designs for prompting GPT-3.5, Gemini and Copilot in a zero-shot manner. In additional experiments, the authors find that among some BERT-family models, a fine-tuned Legal-BERT exhibits the best potential, outperforming Longformer and Big Bird.

Rank 18: DUTh – A multi-task learning approach for the Legal Argument Reasoning Task in Civil Procedure (Maslaris and Arampatzis, 2024) This team compared the Legal-BERT model with a multi-task Flan-T5 model, which eventually performed on par. The authors relied mostly on fully open-source models and make their approach reproducible.

6 Analysis

6.1 Error analysis

We take a closer look how individual instances in the test set were classified. For this, we cluster the instances by the chapter they appear in and sort the chapters by the average performance (see Figure 2). With the goal of identifying the questions that were more challenging for the systems to answer, we cross-check the chapter titles and content of the best and worst-performing chapters. Chapters 6, 12, and 7 were the best-performing and cover the topics “More Personal Jurisdiction: General In Personam Jurisdiction and In Rem Jurisdiction”, “Two Ways to Run a Railroad: Substance and Procedure After York, Byrd, and Hanna” and “More than an Afterthought: Long-arm Statutes as a Limit on Personal Jurisdiction”. Legal expertise would be required to carefully assess why some chapters appear more difficult than others. Throughout our analysis, we could not identify a clear common

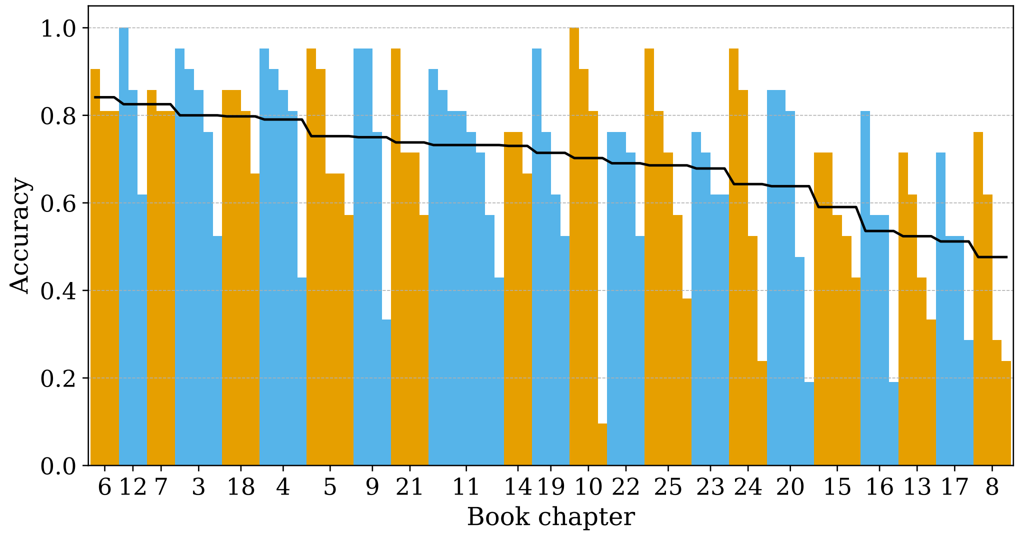

Figure 2: Prediction accuracy of all systems on all questions individually, grouped by the chapter the questions appear in *The Glannon Guide To Civil Procedure*. The line indicates the average accuracy per chapter. The alternating colors serve to delimit the individual chapters.

factor for difficult and easy instances. This can be attributed to the small sample size of the test partition and the carefully designed questions. Please refer to Table 5 for a full list of chapter titles.

Another important distinction is between question-answer pairs with a correct answer and those with an incorrect answer. As expected, because of the imbalance of the dataset, correct answers were much harder to classify correctly, as shown in Figure 3 (highlighted in green). On average, only 48.76% of these instances were classified correctly by all participants. For incorrect answers, 76.25% were classified correctly.

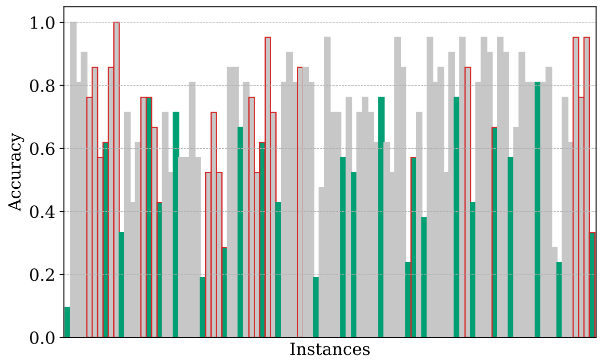

Figure 3: Prediction accuracy of all instances in the test set. Green instances mark questions-answer pairs with a true answer. Indicated by red boxes are instances that could have potential leakage of the question from the dev set.

6.2 Potentially leaked data points

Furthermore, we want to investigate the impact of our potentially leaked data points. We compare the performance on non-leaked questions to that on potentially leaked questions in Figure 3 (indicated by a red border) and find that the performance remains almost identical for incorrect answers (76.69% for leaked vs. 76.10% for non-leaked), but shows a slight increase for correct answers (53.57% for leaked vs. 46.50% for non-leaked).

Table 4 also displays the difference in the final score that would result from removing potentially leaked data points for each participant. While the ranking may change for some teams, the gains and losses are minimal and do not follow a discernible pattern.

All in all, we could not detect a strong impact of the potentially leaked data points. This could also be due to the very limited use of fine-tuning or training with the provided data, since many models simply use zero-shot prompting or similar methods that do not require the training data at all.

6.3 Findings

The best-performing systems all use GPT-4, either with a double-checking mechanism (prompting more than once), tailoring the prompt to a legal reasoning method, or using ensembling to achieve optimal results. Domain-specific models, such as the popular Legal-BERT, which were explored in several approaches, are consistently outperformed by systems using GPT-4 and could not demonstrate their advantages. The authors of some systems also noted that task performance improved when the task was remodeled as a multiple-choice task. Although this was not prohibited, it undermines the idea of the task and should be taken into account in a potential future iteration. Lastly, additional data was rarely used and did not contribute to the best results. Although the focus of the best submissions was on leveraging the power of LLMs, the techniques used to acquire a label from the prompts were creative, diverse and tailored to the legal domain.

7 Conclusion

In this paper we presented an overview of Task 5 of the SemEval-2024 competition, a task on argument reasoning in civil procedure. The dataset and the problems related to data leakage due to partitioning were briefly outlined. The submitted systems were described and summarized, and insights into the achieved results were provided. The submitted solutions indicate that LLMs, specifically GPT-4, are surprisingly decent in handling argument reasoning in civil procedure. Although Legal-BERT and other older domain-specific models can still solve the task to some extent, they are outperformed by a significant margin. The average performance of older or simpler techniques also suggests that this task is a suitable benchmark for evaluating legal reasoning in civil procedure. Although the top-performing systems still have room for improvement, the submitted solutions demonstrate that performance can be enhanced using various techniques. This task is far from solved. A future iteration of this competition could also utilize the mostly unused *analysis* field. This could alleviate the dataset's shortcoming of lacking traceable reasoning steps in the solution to further boost the emphasis on the reasoning aspect of the task.

Limitations

In theory, the dataset should not have leaked to a large language model yet, because the book is not freely available online. Consequently, the dataset should contain mostly new and unseen questions for the NLP community, while also having limited risk of leakage into a large language model. However, especially because of the use of closed LLMs and the lack of knowledge about the training corpora used for them, we can not be entirely sure that our dataset has not been seen by the LLMs used in the systems.

Although some of the answers to the questions can be argued about and might even be outdated in terms of applicable laws and statutes (the basis for the dataset is the 4th edition of the book), we can consider them correct, because they were answered by an expert – the author of the book.

Acknowledgements

We would like to thank John Glannon and Aspen Publishing for their support. This work has been funded by the German Research Foundation as part of the ECALP project (HA 8018/2-1).

References

Iz Beltagy, Matthew E. Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv:2004.05150*.

Irene Benedetto, Alkis Koudounas, Lorenzo Vaiani, Eliana Pastor, Luca Cagliero, and Francesco Tarasconi. 2024. MAINDZ at SemEval-2024 task 5: CLUEDO - Choosing Legal oUtcome by Explaining Decision through Oversight. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 986–994, Mexico City, Mexico. Association for Computational Linguistics.

Leonard Bongard, Lena Held, and Ivan Habernal. 2022. The legal argument reasoning task in civil procedure. In *Proceedings of the Natural Legal Language Processing Workshop 2022*, pages 194–207, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. 2020. LEGAL-BERT: The Muppets straight out of Law School. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2898–2904, Online. Association for Computational Linguistics.

Odysseas Chlapanis, Ion Androutsopoulos, and Dimitrios Galanis. 2024. Archimedes-AUEB at SemEval-2024 task 5: LLM explains civil procedure. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 1617–1632, Mexico City, Mexico. Association for Computational Linguistics.

Biralatei Fawei, Adam Wyner, and Jeff Pan. 2016. Passing a USA national bar exam: a first corpus for experimentation. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3373–3378, Portorož, Slovenia. European Language Resources Association (ELRA).

Joseph W Glannon. 2018. *The Glannon Guide To Civil Procedure: Learning Civil Procedure Through*

Multiple-Choice Questions and Analysis, 4th edition. Aspen Publishing.

Neel Guha, Julian Nyarko, Daniel E. Ho, Christopher Ré, Adam Chilton, Aditya Narayana, Alex Chohlas-Wood, Austin Peters, Brandon Waldon, Daniel N. Rockmore, Diego Zambrano, Dmitry Talisman, Enam Hoque, Faiz Surani, Frank Fagan, Galit Sarfaty, Gregory M. Dickinson, Haggai Porat, Jason Hegland, Jessica Wu, Joe Nudell, Joel Niklaus, John Nay, Jonathan H. Choi, Kevin Tobia, Margaret Hagan, Megan Ma, Michael Livermore, Nikon Rasumov-Rahe, Nils Holzenberger, Noam Kolt, Peter Henderson, Sean Rehaag, Sharad Goel, Shang Gao, Spencer Williams, Sunny Gandhi, Tom Zur, Varun Iyer, and Zehua Li. 2023. Legalbench: A collaboratively built benchmark for measuring legal reasoning in large language models.

Nils Holzenberger, Andrew Blair-Stanek, and Benjamin van Durme. 2020. A dataset for statutory reasoning in tax law entailment and question answering. In *Proceedings of the Natural Legal Language Processing Workshop 2020 co-located with the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD 2020)*.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7B.

Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, Lélio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, Szymon Antoniak, Teven Le Scao, Théophile Gervet, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2024. Mixtral of Experts.

Yoshinobu Kano, Mi-Young Kim, Masaharu Yoshioka, Yao Lu, Juliano Rabelo, Naoki Kiyota, Randy Goebel, and Ken Satoh. 2019. COLIEE-2018: Evaluation of the competition on legal information extraction and entailment. In *New Frontiers in Artificial Intelligence*, volume 11717 of *Lecture Notes in Computer Science*, pages 177–192, Cham. Springer International Publishing.

Daniel Martin Katz, Michael James Bommarito, Shang Gao, and Pablo Arredondo. 2023. GPT-4 passes the Bar Exam.

Kristiyan Krumov, Svetla Boytcheva, and Ivan Koytchev. 2024. SU-FMI at SemEval-2024 task 5: From BERT fine-tuning to llm prompt engineering - approaches in legal argument reasoning. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 1662–1668, Mexico City, Mexico. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized bert pretraining approach.

Shayne Longpre, Le Hou, Tu Vu, Albert Webson, Hyung Won Chung, Yi Tay, Denny Zhou, Quoc V. Le, Barret Zoph, Jason Wei, and Adam Roberts. 2023. The Flan collection: Designing data and methods for effective instruction tuning.

Yu-An Lu and Hung-Yu Kao. 2024. 0x.Yuan at SemEval-2024 task 5: Enhancing legal argument reasoning with structured prompts. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 385–390, Mexico City, Mexico. Association for Computational Linguistics.

Ioannis Maslaris and Avi Arampatzis. 2024. DUTh at SemEval 2024 task 5: A multi-task learning approach for the legal argument reasoning task in civil procedure. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 1042–1046, Mexico City, Mexico. Association for Computational Linguistics.

OpenAI. 2023. GPT-4 technical report.

Anish Pahilajani, Samyak Rajesh Jain, and Devasha Trivedi. 2024. NLP at UC Santa Cruz at SemEval-2024 task 5: Legal answer validation using few-shot multi-choice QA. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 1298–1303, Mexico City, Mexico. Association for Computational Linguistics.

M Manvith Prabhu, Haricharana Srinivasa, and Anand Kumar M. 2024. SCaLAR NITK at SemEval-2024 task 5: Towards unsupervised question answering system with multi-level summarization for legal text. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 193–199, Mexico City, Mexico. Association for Computational Linguistics.

Juliano Rabelo, Randy Goebel, Mi-Young Kim, Yoshinobu Kano, Masaharu Yoshioka, and Ken Satoh. 2022. Overview and discussion of the competition on legal information extraction/entailment (COLIEE) 2021. *The Review of Socionetwork Strategies*, 16(1):111–133.

Hoorieh Sabzevari, Mohammadmostafa Rostamkhani, and Sauleh Eetemadi. 2024. eagerlearners at SemEval2024 task 5: The legal argument reasoning task in civil procedure. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 909–913, Mexico City, Mexico. Association for Computational Linguistics.

Dan Schumacher and Anthony Rios. 2024. Team UTSA-NLP at SemEval 2024 task 5: Prompt ensembling for argument reasoning in civil procedures with GPT4. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 1283–1290, Mexico City, Mexico. Association for Computational Linguistics.

Peng Shi, Jin Wang, and Xuejie Zhang. 2024. YNU-HPCC at SemEval-2024 task 5: Regularized Legal-BERT for legal argument reasoning task in civil procedure. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 743–748, Mexico City, Mexico. Association for Computational Linguistics.

Marco Siino. 2024. Mistral at SemEval-2024 task 5: Mistral 7B for argument reasoning in civil procedure. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 155–162, Mexico City, Mexico. Association for Computational Linguistics.

Kriti Singhal and Jatin Bedi. 2024. Transformers at SemEval-2024 task 5: Legal argument reasoning task in civil procedure using RoBERTa. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 956–959, Mexico City, Mexico. Association for Computational Linguistics.

Binjie Sun and Xiaobing Zhou. 2024. ignore at SemEval-2024 task 5: A legal classification model with summary generation and contrastive learning. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 517–522, Mexico City, Mexico. Association for Computational Linguistics.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023a. LLaMA: Open and efficient foundation language models.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023b. Llama 2: Open foundation and fine-tuned chat models.

Lewis Tunstall, Edward Beeching, Nathan Lambert, Nazneen Rajani, Kashif Rasul, Younes Belkada, Shengyi Huang, Leandro von Werra, Clémentine Fourrier, Nathan Habib, Nathan Sarrazin, Omar Sanseviero, Alexander M. Rush, and Thomas Wolf. 2023. Zephyr: Direct distillation of lm alignment.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. 2022. Chain-of-Thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 35, pages 24824–24837. Curran Associates, Inc.

Manzil Zaheer, Guru Guruganesh, Kumar Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, and Amr Ahmed. 2020. Big Bird: Transformers for longer sequences. In *Advances in Neural Information Processing Systems*, volume 33, pages 17283–17297. Curran Associates, Inc.

Xiaofeng Zhao, Xiaosong Qiao, Kaiwen Ou, Min Zhang, Su Chang, Mengyao Piao, Yuang Li, Yinglu Li, Ming Zhu, Yilun Liu, Feiyu Yao, shimin tao, Hao Yang, and Yanfei Jiang. 2024. HW-TSC at SemEval-2024 task 5: Self-eval? a confident llm system for auto prediction and evaluation for the legal argument reasoning task. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 1817–1821, Mexico City, Mexico. Association for Computational Linguistics.

A Participants systems

#	Team	LLM	Prompting	Fine-tuning	Inputs	+Data	MC
1	HW-TSC	GPT-4	custom	–	Q, A, E	–	✓
2	MAINDZ	Flan T5 XXL, Llama 13B, Zephyr 7B, Mistral 7B, GPT-4	zero-shot	✓	Q, A, E	–/✓	✓
3	SU-FMI	GPT-4	custom	–	Q, A, E	–	–
5	UTSA-NLP	GPT-4	CoT	–	Q, A, E	–	–
7	UC Santa Cruz	GPT-4	zero-shot	–	Q, A, E	–/✓	✓
9	0x.Yuan	Mixtral-8x7B	CoT	–	Q, A, E	–	–
10	YNU-HPCC	Legal-BERT	–	✓	Q, A, E	–	–
11	Mistral	Mistral 7B Instruct	zero-shot	–	Q, A	–	–
13	Transformers	RoBERTa	–	✓	Q, A, E, An.	–	–
14	ignore	Legal-BERT, GPT-3.5	–	✓	Q, A, E, An.	–	–
15	Archimedes-AUEB	GPT family, Llama2 7B	CoT	✓	Q, A, E	–	–
16	SCaLAR NITK	Legal-BERT, T5	–	–	Q, A, E	–	✓
17	eagerlearners	Longformer, Big Bird, Legal-RoBERTa, GPT-3.5, Gemini, Copilot	CoT, zero-shot	✓	Q, A, E	–	–
18	DUTh	Legal-BERT, Flan T5	–	✓	Q, A, E	–	–

Table 3: Summarized features of the submitted systems.

B Leaderboard accounting for leaked data points

Rank	Participant	F_1	Diff
1	SU-FMI	0.8143	0.0415
2	HW-TSC	0.7829	-0.0403
2	MAINDZ	0.7829	0.0082
4	qiaoxiaosong	0.7535	-0.0109
5	UTSA-NLP	0.7464	0.0149
5	kubapok	0.7464	0.0493
7	hrandria	0.6048	-0.0279
8	LegalSense	0.6019	-0.0580
8	odysseas_aueb	0.6019	0.0875
10	Mistral	0.5824	0.0227
11	Hwan_Chang	0.5750	0.0195
12	PengShi	0.5594	-0.0316
13	kriti7	0.5177	-0.0335
14	Yuan_Lu	0.5127	-0.0873
15	yms	0.5071	0.0244
16	lhoorie	0.5007	0.0050
17	woody	0.4970	-0.0541
18	SCaLAR Group, NITK Surathkal	0.4779	-0.0187
19	langml	0.4510	0.0135
20	majority baseline	0.4320	0.0051
21	U_201060	0.4283	-0.0219

Table 4: Performance of the systems on data points that have not potentially leaked from dev, compared to the original score with potentially leaked data points.

C The Glannon Guide to Civil Procedure – Chapters

Chapter	Title
3	Federal Claims and Federal Cases
4	Removal Jurisdiction: The Defendant Chooses the Forum
5	Personal Jurisdiction: Myth and Minimum Contact
6	More Personal Jurisdiction: General In Personam Jurisdiction and In Rem Jurisdiction
7	More than an Afterthought: Long-arm Statutes as a Limit on Personal Jurisdiction
8	Home and Away: Litigating Objections to the Court's Jurisdiction
9	Due Process and Common Sense: Notice and Service of Process
10	Venue and Transfer: More Limits on the Place of Suit
11	State Law in Federal Courts: Basics of the Erie Doctrine
12	Two Ways to Run a Railroad: Substance and Procedure After York, Byrd, and Hanna
13	The Scope of the Action: Joinder of Claims and Parties Under the Federal Rules
14	Of Hooks and Nuclei: Supplemental Jurisdiction over State Law Claims
15	Sufficient Allegations: Pleading Under the Federal Rules
16	Change over Time: Amending the Pleadings Under Rule 15
17	Never Forget Rule 11: Representations to the Court
18	Technicalities, Technicalities: Pre-answer Motions Under the Federal Rules
19	Probing to the Limits: The Scope of Discovery Under the Federal Rules
20	The Basic Tools of Discovery in Federal Court
21	Dispositive Motions: Dismissal for Failure to State a Claim and Summary Judgment
22	Judgment as a Matter of Law in the Federal Courts
23	Second Time Around: The Grounds and Procedure for Motions for New Trial
24	The Quest for Finality: Claim Preclusion Under the Second Restatement of Judgments
25	Collateral Estoppel, Issue Preclusion, Whatever

Table 5: Chapter titles