MAINDZ at SemEval-2024 Task 5: CLUEDO-Choosing Legal oUtcome by Explaining Decision through Oversight

MAINDZ at SemEval-2024 Task 5: CLUEDO - Choosing Legal oUtcome by Explaining Decision through Oversight

Irene Benedetto[1,2] **Alkis Koudounas**[1] **Lorenzo Vaiani**[1]

Eliana Pastor[1] **Luca Cagliero**[1] **Francesco Tarasconi**[2]

[1] Politecnico di Torino, {name.surname}@polito.it
[2] MAIZE, {name.surname}@maize.io

Abstract

Large language models (LLMs) have recently obtained strong performance on complex reasoning tasks. However, their capabilities in specialized domains like law remain relatively unexplored. We present CLUEDO, a system to tackle a novel legal reasoning task that involves determining if a provided answer correctly addresses a legal question derived from U.S. civil procedure cases. CLUEDO utilizes multiple collaborator models that are trained using multiple-choice prompting to choose the right label and generate explanations. These collaborators are overseen by a final "detective" model that identifies the most accurate answer in a zero-shot manner. Our approach achieves an F1 macro score of 0.74 on the development set and 0.76 on the test set, outperforming individual models. Unlike the powerful GPT-4, CLUEDO provides more stable predictions thanks to the ensemble approach. Our results showcase the promise of tailored frameworks to enhance legal reasoning capabilities in LLMs.

1 Introduction

Recent improvements in large language models are leading to a rethinking of legal practices, particularly in the United States (Frankenreiter and Nyarko, 2022; Hoffman and Arbel, 2023; Glaze et al., 2021). This can potentially transform time-consuming tasks such as brief writing and corporate compliance (Guha et al., 2023; Benedetto et al., 2023a). This could also contribute to alleviating the access-to-justice crisis (Corporation, 2017; Tito, 2017). The unique properties of LLMs, including their ability to learn from limited labeled data and proficiency in complex reasoning tasks, make them appealing for legal applications (Zheng et al., 2021; Guha et al., 2023; Benedetto et al., 2023b, 2024).

However, enthusiasm is tempered by concerns about the risks associated with LLMs, such as generating offensive, misleading, or factually incorrect content (Engstrom and Gelbach, 2020; Bender et al., 2021). These issues could have significant consequences, particularly affecting marginalized or under-resourced populations (Surden, 2020; Volokh, 2023; Koudounas et al., 2023, 2024).

To address safety implications, there is a pressing need to evolve and enhance legal reasoning capabilities in LLMs. Despite this urgency, practitioners face challenges in assessing LLMs' legal reasoning capabilities, as existing legal benchmarks are limited and often fail to capture the diverse aspects of legal tasks (Guha et al., 2023).

In this direction, the organizers of SemEval-2024 Task 5 introduce a novel Natural Language Processing (NLP) task and dataset derived from the U.S. civil procedure domain (Bongard et al., 2022). Each dataset instance comprises a case introduction, a specific question, and a potential solution argument, along with an in-depth analysis justifying the argument's applicability to the case. When provided with a topic introduction, a question, and a potential answer, the objective of the proposed task is to determine whether the given answer is accurate or not.

To tackle this task, we initially transform the dataset into a multiple-choice question answering problem using the multiple-choice prompting (MCP) approach (Robinson et al., 2023). We experimented with various open-source language models on this modified dataset, including Flan T5 XXL (Wei et al., 2021; Chung et al., 2022), LLama 7B and 13B (Touvron et al., 2023), Zephyr 7B (Touvron et al., 2023), and Mistral 7B (Jiang et al., 2023). Specifically, we trained these models to solve legal problems while also providing an explanation for the predicted outcome, leveraging the analysis provided. We thus introduce the *CLUEDO* approach, which stands for "**C**hoosing **L**egal o**U**tcome by **E**xplaining **D**ecisions through **O**versight". This framework utilizes multiple collaborative models to synthesize the final outcome based on each model's predictions. Each individual

Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024), pages 997–1005

model is trained to predict the label of the correct candidate answer and generate an explanation accordingly. The final *"detective"* model operates in a zero-shot manner, relying upon the outputs of the collaborators. The model processes the answers and the explanations of all collaborators and deduces the ultimate answer.

The results on the challenge dataset demonstrate that our proposed methodology surpasses the performance of single models trained with standard fine-tuning. Furthermore, our approach achieved the second-place position in the public competition, achieving a final test F1 macro score of 0.77[1].

Research Questions. We investigate the following research questions (RQs):

- **RQ1.** Is the multiple-choice setting more effective than the single-choice one?
- **RQ2.** Does including the analysis in the training and generation process improve performance?
- **RQ3.** Is our detective model CLUEDO more effective than individual collaborators in a zero-shot setting? Are CLUEDO results more stable?

2 Related Work

In the legal domain, the advent of Legal LLMs has reshaped how legal professionals approach case analysis, decision-making, and document generation processes (Lai et al., 2023). LLMs possess logical reasoning capabilities that enable legal professionals to comprehend case processes, aid judges in decision-making, swiftly identify similar cases through language comprehension, analyze and condense essential case details, and utilize automated content generation to draft repetitive legal documents (Guha et al., 2023). Researchers have recently started exploring whether large language models have the capability to carry out legal reasoning. Unlike BERT-based models, LLMs are evaluated on their ability to learn tasks in-context, primarily through prompting (Liu et al., 2022). Studies have explored the role of prompt-engineering for Legal Judgment Prediction (Jiang and Yang, 2023), statutory reasoning (Blair-Stanek et al., 2023) legal exams (Yu et al., 2023). Several case studies (Nay et al., 2023; Drápal et al., 2023; Savelka, 2023; Savelka et al., 2023; Westermann et al., 2023) highlight the potential and the limitations of GPT models in real use cases. However, to the best of our knowledge, limited effort has been devoted to analyzing the effectiveness of smaller and open-source language models (e.g., Llama 2 (Touvron et al., 2023)) in this domain (Guha et al., 2023), and how they can effectively be employed in conjunction with closed-source foundational models, such as GPT-4 (OpenAI et al., 2023).

3 Dataset and Task Description

Bongard et al. (2022) present a new dataset from the U.S. civil procedure domain. This dataset is derived from a book intended for law students, suggesting its complexity and suitability for benchmarking modern legal language models. Each instance of the dataset consists of:

- *General introduction to the case*: an overview of the case to set the context.
- *Particular question*: a specific legal question related to the case is presented.
- *Possible solution argument*: a potential answer associated with the question is provided.
- *Annotated label*: it defines if the possible solution is correct (1) or not (0).
- *Detailed analysis*: Accompanying each solution argument is a thorough analysis explaining why the argument applies to the case in question.

The task is structured as a binary classification task where the goal is to predict the correctness of the answer provided, i.e., the label provided together with the textual information. The analysis and the labels are not available during test time.

4 System Overview

This section provides a comprehensive overview of the proposed methodology. Firstly, we outline the approach to the multiple-choice question-answering problem and how we adapt it to our scenario. Secondly, we introduce the *CLUEDO* framework, along with details about the competitors incorporated into our study.

[1]Code available at https://github.com/irenebenedetto/PoliToHFI-SemEval2024-Task5

Table 1: **Zero-shot** models on dev set. The best performance (in terms of F1 macro) for each model family is in bold. The multiple-choice approach leads to higher performance in five out of six cases.

Model	Classification task	Prec	Rec	F1	Acc
Flan T5 XXL	Multiple choice	0.60	0.67	**0.59**	0.64
Flan T5 XXL	Single choice	0.54	0.53	0.32	0.32
GPT-4	Multiple choice	0.66	0.73	**0.66**	0.57
GPT-4	Single choice	0.40	0.50	0.44	0.80
Llama 2 13B	Multiple choice	0.64	0.58	**0.59**	0.79
Llama 2 13B	Single choice	0.55	0.58	0.54	0.61
Llama 2 7B	Multiple choice	0.51	0.51	0.51	0.74
Llama 2 7B	Single choice	0.53	0.52	**0.52**	0.73
Mistral v0.1 7B	Multiple choice	0.55	0.59	**0.54**	0.61
Mistral v0.1 7B	Single choice	0.55	0.58	0.52	0.57
Zephyr beta 7B	Multiple choice	0.54	0.56	**0.50**	0.69
Zephyr beta 7B	Single choice	0.40	0.50	0.44	0.80

Table 2: **Trained** models performance on dev set. All models are trained to generate both labels and analysis, following the multiple-choice setting.

Model	Prec	Rec	F1	Acc
Llama 2 7B	0.57	0.60	0.56	0.64
Mistral v0.1 7B	0.61	0.63	0.62	0.73
Zephyr beta 7B	0.62	0.65	0.63	0.73
Llama 2 13B	0.65	0.69	**0.66**	0.75

Multiple-choice. Following the intuition of Robinson et al. (2023), we convert the dataset into a multiple-choice question answering problem and adopt multiple choice prompting (MCP) (Robinson et al., 2023). In MCP, the language model is presented not only with the question but also with a set of candidate answers, akin to a multiple-choice test. Each answer is linked to a symbol such as *"A," "B,"* or *"C."* This approach enables the model to compare answer choices explicitly and diminishes computational expenses for a generation. In cases where there is only one candidate answer, the system automatically generates the alternative *"None of the above is true"*. These additional answers are not accounted in the test and validation metrics.

In our experiments, we evaluate whether the multi-choice approach is indeed more effective than a single-choice approach. In the single-choice setting, we prompt a single choice, and the model should directly predict whether it is correct.

CLUEDO. To tackle the task of the challenge, we introduce the *CLUEDO* framework, which stands for "Choosing Legal Outcome by Explaining Decisions through Oversight." In a nutshell, multiple collaborative models are trained to predict the correct label for a candidate answer that addresses the legal question. These models generate their analysis as part of their training. The final model, operating in a zero-shot manner, utilizes the responses and explanations from the set of collaborators to identify the most accurate final answer, considering their collective performance. More in detail, the *CLUEDO* system is structured as follows:

- *N collaborative models*: given the introduction, the legal question, and the candidate answers, these models are trained to predict the label of the candidate answer that correctly responds to the legal question and generate an explanation. We fix the number of collaborators equal to three. We select the collaborators based on their results on the dev set.

- *The final "detective" model*: this model is employed in a zero-shot manner. Based on the responses from the collaborators and their corresponding explanations, this model must identify the most accurate final answer, overseeing the collaborators' performance. The final model is also provided with the introduction, legal questions, and candidate answers.

Example of prompts for collaborative and detective models are reported in Table 3.

Competitors. To assess the strength of the proposed CLUEDO approach, we compare the results with a set of alternatives on the final test set: the best collaborator chosen based on the results achieved on the dev set (that we call *Best collaborator*), and the correction of collaborator models based on consensus (after named *Collaborators agreement*). The latter approach involves taking the predictions of the top-performing collaborator (on the dev set) and rectifying instances where both the second and third collaborators mutually confirm inaccuracies. We finally employ the zero-shot final model without any collaborators to test its generalization capabilities, namely *Zero-shot detective model*.

5 Experimental Setup

Models. We evaluated various open-source models, employing both zero-shot and fine-tuning methodologies. Our analysis covered Flan T5 XXL (Wei et al., 2021; Chung et al., 2022), LLama 7B (Touvron et al., 2023) and 13B, Zephyr 7B (Touvron et al., 2023), and Mistral 7B (Jiang et al., 2023), selected for their unique features and performance metrics. Furthermore, we integrated into our assessment GPT-4 (OpenAI et al., 2023) in a zero-shot context.

Training procedure. We employed a Supervised Fine-Tuning (SFT) approach, implementing precision enhancement with 8-bit quantization. The models were trained for three epochs utilizing Parameter-Efficient Fine-Tuning (PEFT) (Mangrulkar et al., 2022), with a batch size set at 4 and a learning rate of 5e-5. The sequences were processed with a context length of 4096, optimizing the model's ability to capture long-range dependencies in the data.

Hardware. We run the experiments on a machine equipped with Intel® Core™ i9-10980XE CPU, 1 × Nvidia® Tesla T4 GPU, 16 GB of RAM running Ubuntu 22.04 LTS.

6 Results

To illustrate the efficacy of the multiple-choice setting and model selection criteria, we conduct individual tests for each configuration and present the obtained results on the development set. The following paragraphs address the research questions previously presented.

RQ1: Impact of the multiple-choice setting. Table 1 shows the zero-shot models' performance on the development set. For each model family, the multiple-choice question-answering approach consistently outperforms the single-choice approach in terms of F1 Macro. There is variability in the performance of different models within the same family. In general, larger models tend to exhibit stronger generalization capabilities than smaller ones.

RQ2: Impact of analysis inclusion in model training. In Table 4, we highlight the impact of including the analysis in the models' training process. To examine outcomes across various model sizes and classification tasks, we fixed the model family (Llama 2 from Meta). In both the 7B and 13B models, including the analysis (✓) consistently leads to higher performance for multiple-choice tasks. In particular, including the analysis during training leads to more balanced precision and recall metrics, resulting in an overall improvement in the F1 Macro score. For both Llama 2 7B and Llama 2 13B, the F1 Macro scores in single-choice tasks do not show significant improvement with the inclusion of the analysis. This may indicate that these models are less sensitive to additional analysis in single-choice tasks.

Additionally, the training of Llama 2 13B with the analysis allows for an additional $+0.07$ F1 score compared to its zero-shot counterpart, while for the 7B models, the training deteriorates the performance.

RQ3: CLUEDO results. The selection of collaborative models is guided by the results obtained on the development set as shown in Table 2. All models are configured to generate both labels and analysis, following the multiple-choice setting. Among the models, Llama 2 13B stands out with the highest F1 Macro score, indicating robust performance across multiple evaluation metrics, followed by Mistral and Zephyr models. For the supervisor model, we choose GPT-4, the best performer in the zero-shot setting (see Table 1).

Results on the test set are summarized in Table 5. Applying corrections based on the consensus of the second and third collaborators (Mistral and Zephyr) slightly reduces the F1 Macro to 0.65 on both development and test sets. This suggests that the initial collaborator's predictions were already quite accurate. The zero-shot model without collaborators

Table 3: **Example of prompts** for collaborative models and our CLUEDO approach.

Approach	Example Prompt
Collaborative Models	<s>[INST] <<SYS>>Given the following explanation and the question, which of the candidate answers is correct? The correct answer is the one that is true according to the explanation. <</SYS>> <explanation>Although discovery usually extends to all evidence relevant to claims and defenses in the action, Rule 26(b)(1) expressly carves out one [...] </explanation> <question>4. Confidential chat. Shag, a budding rock star with no business experience, enters into a five-year exclusive contract with Fringe Records, after [...] </question> <candidate_answers> 1 - Shag will not have to answer any of the interrogatories, because all three were discussed in å confidence with Rivera in the course of his representation. 2 - Shag will have to answer the first interrogatory, but not the other two. 3 - Shag will have to answer all three interrogatories, because [...] 5 - None of the above is true. </candidate_answers> [/INST] <correct_answer>5 </correct_answer> <analysis>Let's start by eliminating A. It proceeds on the premise that all three items are subject to discovery, because all [...] </analysis>
CLUEDO	You are a legal supervisor tasked with resolving legal queries. You are working alongside three artificial intelligence models, named m1, m2, and m3. Given an introductory context, a question, and a set of candidate answers, these three models must choose the correct answer and provide justification for their choice. Your responsibility is to assess the models' responses and determine whether they are correct or not. To do so, you must read the context (enclosed within the tags <context></context>), the question (within <question></question>tags), and the candidate answers (within <candidate_answers></candidate_answers>tags), and identify the correct answer among them (using the <supervisor_answer>tag). Additionally, you must provide reasoning for your choice (using the <supervisor_explanation>tag). While collaborating with the models and considering their advice, the ultimate decision rests with you. For each response, use the following format: <supervisor_answer>SUPERVISOR ANSWER</supervisor_answer> <supervisor_explanation>SUPERVISOR ANSWER</supervisor_explanation> <context>Although discovery usually extends to all evidence relevant to claims and defenses in the action, Rule 26(b)(1) expressly carves out one [...] </context> <question>4. Confidential chat. Shag, a budding rock star with no business experience, enters into a five-year exclusive contract with Fringe Records, after [...] </question> <candidate_answers> 1 - Shag will not have to answer any of the interrogatories, because all three were discussed in å confidence with Rivera in the course of his representation. 2 - Shag will have to answer the first interrogatory, but not the other two. 3 - Shag will have to answer all three interrogatories, because [...] 5 - None of the above is true. </candidate_answers> <m1_answer>1</m1_answer> <m1_explanation>[...] </m1_explanation> <m2_answer>1</m2_answer> <m2_explanation>[...] </m2_explanation> <m3_answer>2</m3_answer> <m3_explanation>[...] </m3_explanation> <supervisor_answer>

Table 4: **Trained** models on dev set. The best results (in terms of F1 Macro) are in bold. The generation of the analysis leads to higher performance for both 7B and 13B models.

Model	Classification task	Analysis included	Prec	Rec	F1	Acc
Llama 2 7B	Multiple choice	x	0.49	0.48	0.47	0.56
Llama 2 7B	Multiple choice	✓	0.57	0.60	**0.56**	0.64
Llama 2 7B	Single choice	x	0.40	0.50	0.44	0.80
Llama 2 7B	Single choice	✓	0.40	0.50	0.44	0.80
Llama 2 13B	Single choice	x	0.55	0.58	0.52	0.57
Llama 2 13B	Multiple choice	✓	0.65	0.69	**0.66**	0.75

Table 5: **Final Results** on dev and test sets: the best collaborator, collaborative agreements, and collaborators within CLUEDO are trained to generate the analysis along with the labels and adopt the MCP approach.

	Dev		Test	
Method	**F1**	**Acc**	**F1**	**Acc**
Best collaborator	0.66 (± 0.001)	0.75 (± 0.001)	0.69 (± 0.001)	0.75 (± 0.001)
Collaborators agreement	0.65 (± 0.001)	0.75 (± 0.001)	0.65 (± 0.001)	0.75 (± 0.001)
Zero-shot detective model	0.63 (± 0.038)	0.71 (± 0.024)	**0.77** (± 0.022)	0.83 (± 0.016)
CLUEDO	**0.74** (± 0.017)	0.78 (± 0.017)	**0.77** (± 0.017)	0.82 (± 0.013)

(GPT-4) performs well on the development set with an F1 score of 0.63. However, it surpasses all other methods on the test set with a notable F1 Macro of 0.77, showcasing its robust generalization capabilities. The CLUEDO model outperforms other methods with the highest F1 Macro on the development set (0.74) while achieving the second-highest score on test data. To assess the stability of predictions, we experimented five times on the validation set and test set and measured the performance of the models. Even with a greedy decoding strategy, small discrepancies regarding floating point operations lead to divergent generations, especially for larger models (Gawlikowski et al., 2021). It is known that this issue primarily concerns GPT-4[2]. Therefore, even though the temperature is set to 0 for all experiments, users have often reported significant variations in the output.

Although the predictions of trained models remained consistent, notable differences were observed in GPT-4 predictions, particularly when used without collaborators (the temperature is set to zero with no sampling). The results are presented in Table 5. With the proposed CLUEDO approach, the standard deviation is reduced by half. Additionally, the error estimate on the development set aligns with the one obtained on the test set. In conclusion, even though CLUEDO may not outperform others on test data, it ensures higher stability in predictions.

7 Conclusion

This paper presents a novel solution to the SemEval 2024 - Legal Reasoning Task, which introduced a challenge for evaluating contemporary legal language models. We transform the original dataset into a multiple-choice question-answering problem using the multiple-choice prompting approach and propose an original system, namely *CLUEDO*, that utilizes multiple collaborative LLMs and employs a final *"detective"* model to predict the outcome. Results show that our framework outperforms individual models in the public competition while returning more stable predictions, securing second place in the public competition.

[2]Here some discussion of the OpenAI community on models variability: https://community.openai.com/t/why-the-api-output-is-inconsistent-even-after-the-temperature-is-set-to-0/329541, https://community.openai.com/t/run-same-query-many-times-different-results/140588

References

Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the

dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 610–623.

Irene Benedetto, Luca Cagliero, Francesco Tarasconi, Giuseppe Giacalone, and Claudia Bernini. 2023a. *Benchmarking Abstractive Models for Italian Legal News Summarization*.

Irene Benedetto, Alkis Koudounas, Lorenzo Vaiani, Eliana Pastor, Elena Baralis, Luca Cagliero, and Francesco Tarasconi. 2023b. PoliToHFI at SemEval-2023 task 6: Leveraging entity-aware and hierarchical transformers for legal entity recognition and court judgment prediction. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 1401–1411, Toronto, Canada. Association for Computational Linguistics.

Irene Benedetto, Alkis Koudounas, Lorenzo Vaiani, Eliana Pastor, Luca Cagliero, Francesco Tarasconi, and Elena Baralis. 2024. Boosting court judgment prediction and explanation using legal entities. *Artificial Intelligence and Law*.

Andrew Blair-Stanek, Nils Holzenberger, and Benjamin Van Durme. 2023. Can gpt-3 perform statutory reasoning?

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.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2022. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*.

Legal Services Corporation. 2017. The justice gap: Measuring the unmet civil legal needs of low-income americans.

Jakub Drápal, Hannes Westermann, and Jaromir Savelka. 2023. Using large language models to support thematic analysis in empirical legal studies.

David Freeman Engstrom and Jonah B Gelbach. 2020. Legal tech, civil procedure, and the future of adversarialism. *U. Pa. L. Rev.*, 169:1001.

Jens Frankenreiter and Julian Nyarko. 2022. Natural language processing in legal tech. *Legal Tech and the Future of Civil Justice (David Engstrom ed.) Forthcoming*.

Jakob Gawlikowski, Cedrique Rovile Njieutcheu Tassi, Mohsin Ali, Jongseok Lee, Matthias Humt, Jianxiang Feng, Anna M. Kruspe, Rudolph Triebel, Peter Jung, Ribana Roscher, Muhammad Shahzad, Wen Yang, Richard Bamler, and Xiao Xiang Zhu. 2021. A survey of uncertainty in deep neural networks. *CoRR*, abs/2107.03342.

Kurt Glaze, Daniel E Ho, Gerald K Ray, and Christine Tsang. 2021. Artificial intelligence for adjudication: The social security administration and ai governance.

Neel Guha, Julian Nyarko, Daniel E Ho, Christopher Ré, Adam Chilton, Aditya Narayana, Alex Chohlas-Wood, Austin Peters, Brandon Waldon, Daniel N Rockmore, et al. 2023. Legalbench: A collaboratively built benchmark for measuring legal reasoning in large language models. *arXiv preprint arXiv:2308.11462*.

David A Hoffman and Yonathan A Arbel. 2023. Generative interpretation. *Available at SSRN*.

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.

Cong Jiang and Xiaolei Yang. 2023. Legal syllogism prompting: Teaching large language models for legal judgment prediction.

Alkis Koudounas, Eliana Pastor, Giuseppe Attanasio, Vittorio Mazzia, Manuel Giollo, Thomas Gueudre, Luca Cagliero, Luca de Alfaro, Elena Baralis, and Daniele Amberti. 2023. Exploring subgroup performance in end-to-end speech models. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5.

Alkis Koudounas, Eliana Pastor, Giuseppe Attanasio, Vittorio Mazzia, Manuel Giollo, Thomas Gueudre, Elisa Reale, Luca Cagliero, Sandro Cumani, Luca de Alfaro, Elena Baralis, and Daniele Amberti. 2024. Towards comprehensive subgroup performance analysis in speech models. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*.

Jinqi Lai, Wensheng Gan, Jiayang Wu, Zhenlian Qi, and Philip S. Yu. 2023. Large language models in law: A survey.

Haokun Liu, Derek Tam, Mohammed Muqeeth, Jay Mohta, Tenghao Huang, Mohit Bansal, and Colin Raffel. 2022. Few-shot parameter-efficient fine-tuning is better and cheaper than in-context learning.

Sourab Mangrulkar, Sylvain Gugger, Lysandre Debut, Younes Belkada, Sayak Paul, and Benjamin Bossan. 2022. Peft: State-of-the-art parameter-efficient fine-tuning methods. `https://github.com/huggingface/peft`.

John J. Nay, David Karamardian, Sarah B. Lawsky, Wenting Tao, Meghana Bhat, Raghav Jain, Aaron Travis Lee, Jonathan H. Choi, and Jungo Kasai. 2023. Large language models as tax attorneys: A case study in legal capabilities emergence.

OpenAI, :, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mo Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2023. Gpt-4 technical report.

Joshua Robinson, Christopher Michael Rytting, and David Wingate. 2023. Leveraging large language models for multiple choice question answering.

Jaromir Savelka. 2023. Unlocking practical applications in legal domain: Evaluation of gpt for zero-shot semantic annotation of legal texts. In *Proceedings of the Nineteenth International Conference on Artificial Intelligence and Law*, ICAIL 2023. ACM.

Jaromir Savelka, Kevin D. Ashley, Morgan A. Gray, Hannes Westermann, and Huihui Xu. 2023. Explaining legal concepts with augmented large language models (gpt-4).

Harry Surden. 2020. The ethics of artificial intelligence in law: Basic questions. *Forthcoming chapter in Oxford Handbook of Ethics of AI*, pages 19–29.

Joel Tito. 2017. How ai can improve access to justice.

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. 2023. Llama 2: Open foundation and fine-tuned chat models.

Eugene Volokh. 2023. Chatgpt coming to court, by way of self-represented litigants. *The Volokh Conspiracy*.

Jason Wei, Maarten Bosma, Vincent Y Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. 2021. Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652*.

Hannes Westermann, Jaromir Savelka, and Karim Benyekhlef. 2023. Llmediator: Gpt-4 assisted online dispute resolution.

Fangyi Yu, Lee Quartey, and Frank Schilder. 2023. Exploring the effectiveness of prompt engineering for legal reasoning tasks. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13582–13596, Toronto, Canada. Association for Computational Linguistics.

Lucia Zheng, Neel Guha, Brandon R Anderson, Peter Henderson, and Daniel E Ho. 2021. When does pretraining help? assessing self-supervised learning for law and the casehold dataset of 53,000+ legal holdings. In *Proceedings of the eighteenth international conference on artificial intelligence and law*, pages 159–168.