A Comparative Study of Prompting Strategies for Legal Text Classification

Ali Hakimi Parizi[1*] and **Yuyang Liu**[1*] and **Prudhvi Nokku**[2*]
Sina Gholamian[1*] and **David B. Emerson**[3]
[1]Thomson Reuters Labs, Toronto, Canada
[2]Thomson Reuters Labs, Bangalore, India
[3]Vector Institute, Toronto, Canada
[1,2]{name.surname}@thomsonreuters.com
[3]david.emerson@vectorinstitute.ai

Abstract

In this study, we explore the performance of large language models (LLMs) using different prompt engineering approaches in the context of legal text classification. Prior research has demonstrated that various prompting techniques can improve the performance of a diverse array of tasks done by LLMs. However, in this research, we observe that professional documents, and in particular legal documents, pose unique challenges for LLMs. We experiment with several LLMs and various prompting techniques, including zero/few-shot prompting, prompt ensembling, chain-of-thought, and activation fine-tuning and compare the performance on legal datasets. Although the new generation of LLMs and prompt optimization techniques have been shown to improve generation and understanding of generic tasks, our findings suggest that such improvements may not readily transfer to other domains. Specifically, experiments indicate that not all prompting approaches and models are well-suited for the legal domain which involves complexities such as long documents and domain-specific language.

1 Introduction

Large Language Models (LLMs) have gained significant popularity recently due to their remarkable versatility across a wide range of natural language processing tasks. These models, such as OpenAI's GPT-3 (Brown et al., 2020) and Facebook's OPT (Zhang et al., 2022), have exhibited notable ability to generate coherent and fluent text, resulting in their adaptation in various domains (Artetxe et al., 2022; Xu et al., 2022; Chintagunta et al., 2021). In addition, through prompting, LLMs have demonstrated significant generalizability (Touvron et al., 2023; Wei et al., 2022).

Prompt engineering refers to techniques used to steer LLM generations by providing specific instructions or augmenting a model's input context, which leads to improved responses and better performance on downstream tasks. LLMs have achieved impressive results for a wide variety of tasks, and prior research has shown the effectiveness of different prompt engineering approaches to further improve their performance (van der Meer et al., 2022; Ma et al., 2022; Reif et al., 2022). However, their applicability to specific domains, particularly legal texts, requires more exploration.

Previous research has investigated the potential of various prompting techniques, such as zero- and few-shot prompting (Brown et al., 2020; Agrawal et al., 2022), prompt ensembling (Pitis et al., 2023), chain-of-thought (Wei et al., 2022), and activation fine-tuning (Azaria and Mitchell, 2023; Li and Liang, 2021), in enhancing the performance of LLMs across different tasks, including text-classification. In addition, text-classification is a central focus of parameter efficient finetuning (PEFT) research (Hu et al., 2022; Lester et al., 2021). However, these studies consider general NLP settings rather than the legal domain.

In this paper, we explore various prompting approaches to utilize pre-trained LLMs for legal document classification. The datasets considered here represent a strong test of LLMs' out-of-domain generalization. We begin with zero-shot learning and compare with more advanced approaches such as prompt ensembling and chain-of-thought. Our findings indicate that legal documents present unique challenges for LLMs, such as complex terminologies and long documents, that hinder the transferability of prompting techniques to this domain. Recently, Toma et al. (2023) observed similar findings in leveraging LLMs for medical text.

2 Methodology

2.1 Models

For the experiments, we utilize the OPT family of language models (Zhang et al., 2022). OPT175B

[*]These authors contributed equally to this work.

Proceedings of the Natural Legal Language Processing Workshop 2023, pages 258–265
December 7, 2023 ©2023 Association for Computational Linguistics

is chosen to represent very large foundation models, and a smaller variant, OPT6.7B, is selected to provide a path to compare the impact of model scale on each prompting approach. The OPT family is chosen because they are fully open-source, come in various sizes, and have shown equivalent capabilities to GPT-3 (Brown et al., 2020). The pretraining data for OPT is drawn from the RoBERTa, The Pile, and PushShift.io Reddit datasets (Zhang et al., 2022). Experiments are conducted using the Kaleidoscope library (Sivaloganathan et al., 2022).

2.2 Datasets

We use two legal classification datasets, ECHR (Chalkidis et al., 2019) and SCOTUS (Chalkidis et al., 2022), both of which are English language datasets. ECHR contains several tasks, one of which is binary classification to determine if a human rights article has been violated. The SCOTUS dataset is a multi-class classification task where Supreme Court case decisions are classified into 14 distinct issue areas. This task is quite challenging as its class distribution is long-tailed. Due to compute and resource limitations, the datasets are randomly down-sampled to 10% and 5% respectively, in a stratified fashion to maintain the original distribution of classes. Additional details are presented in Appendix A.

Since the average token length of SCOTUS documents is substantially larger than the context length of OPT, which is 2048 tokens, each case is first summarized into a smaller paragraph. We use GPT4 (OpenAI, 2023) for this purpose and generate summaries with at most 512 words. Prior to applying this approach to all SCOTUS documents, several summaries were checked for quality and hallucinations. Note that summarization is not necessary for ECHR as its average text length is not higher than the OPT context length.

2.3 Prompt Templates

Each dataset leverages a unique prompt template. For ECHR, we use the prompt template:

> *Complete the sentence with Yes or No*
> **{Few-shot legal case text with label}**
> **{Legal case text to classify}**
> *Has any human rights article or protocol been violated?*

Following (Trautmann et al., 2022), we formulate the prompt as question answering by first providing an instruction and possible answers. The context is then provided, along with the key question to be answered by the LLM. For the SCOTUS dataset, the prompt is structured as follows.

> *Classify the given document with _only one_ label from the following list:*
> **[A list of labels in SCOTUS]**
> **{Few-shot legal case text with label}**
> **{Legal case text to classify}**
> *The label is*

Here, the prompt is formulated as a completion task. We first provide instructions with a set of possible classes and, at the end, provide an unfinished sentence for the LLM to complete. Additional details are presented in Appendix D.

2.4 Prompt Techniques

We investigate four prompting techniques: *Zero-shot & Few-shot*, *Prompt Ensembling*, *Chain-of-Thought (CoT)*, and *Activation Fine-Tuning*. Each approach is reviewed briefly below.

2.4.1 Zero-shot and Few-shot

As observed in Brown et al. (2020); Dong et al. (2023), LLMs trained on massive corpora of texts have shown an ability to perform new tasks from instructions or from a few examples. In zero-shot prompting, the model is tasked with making predictions for a previously unseen problem without training examples, leveraging only prior knowledge obtained during pre-training. While such prompts do not leverage labeled data, they often incorporate task-specific instructions. Few-shot learning offers a middle ground, where the model is provided with a handful of examples ("shots") relevant to the task at hand. Note that few-shot prompting is also often referred to as in-context learning. Effective incorporation of few-shot prompts tends to require models of sufficient size (Kaplan et al., 2020).

2.4.2 Prompt Ensembling

LLMs are known to be sensitive to prompt design, and their task utility can be unpredictable. Ensemble methods have improved deep learning models by pooling classifiers (Ganaie et al., 2022). This has also been successful in prompt-based LLM settings. Pitis et al. (2023) utilized a boosting technique to iteratively construct a group of few-shot prompts of "hard" examples. In this work, we experimented with majority voting classifiers composed of three few-shot prompts.

2.4.3 Chain-of-Thought (CoT)

CoT prompting was initially introduced to elicit internal reasoning in LLMs when generating responses to prompts (Wei et al., 2022). This approach incorporates a series of intermediate examples to steer the LLM to think in a step-by-step manner and, thereby, better achieve the desired outputs. It has been shown to significantly improve multi-step reasoning for LLMs (Yu et al., 2022). In the experiments to follow, we use the zero-shot CoT approach of Kojima et al. (2023). When incorporating few-shot examples, we prepend the CoT prompt with three examples and their labels.

2.4.4 Activation Fine-Tuning

In activation fine-tuning, for each prompt, hidden states for the last non-pad token of the input to the LLM are extracted. Two forms of input to the LLM are considered. In the first, activations are generated by simply passing the case text as input to the LLM. In the second, a three-shot prompt is created as input to generate the activations. This is similar in concept to tuning a classifier on a frozen LM, as is common for models such as BERT (Devlin et al., 2019). In this work, activations from the last feed-forward layer of the transformer are used, along with the label of each sample, to train a small auxiliary feed-forward network. This approach has been effective for challenging classification tasks (Azaria and Mitchell, 2023; Lester et al., 2021). Training details for the classifier are found in Appendix B.

2.5 Prompt Styles

Because OPT models are not instruction fine-tuned (Chung et al., 2022) or trained for question-answering tasks, we expect completion-type prompts, as discussed in Section 2.3, to perform better than question-answering ones. As such, we also experimented with different prompt styles to evaluate this scenario. The question-answering prompt template for SCOTUS, in which the final completion sentence is removed, is shown below.

> *Can you classify the given "Text" with only one class from the following list?*
> **[A list of labels in SCOTUS]**
> **{Legal case text to classify}**

2.6 Classification Predictions

Predictions are produced by processing responses to the prompts generated by OPT. Without special treatment, the OPT generations do not always match a label from the desired label space. Therefore, to extract classification predictions from the LLMs on these two datasets, the log-likelihood of prompt completions associated with labels in the taxonomy is used. For ECHR, the labels are **Yes** or **No** to indicate if the legal case contains a violation of human rights articles or protocol. Prompts are presented to the OPT models and the probability distribution over the vocabulary for the last token is computed. Since we are only interested in the probabilities of the two classes, we select the one with the highest value after applying softmax. A similar approach is employed for the multi-class SCOTUS setting. Use of the vocabulary probability distribution ensures that the role of hallucinations, a common issue in legal applications, is minimized.

3 Results and Discussion

Table 1 summarizes the experimental results. Additional figures are shown in Appendix C. Across all approaches, prompts and other variables are kept the same to ensure the results are comparable across legal classification tasks. The results are compared with state-of-the-art HIER-BERT (Chalkidis et al., 2019) and Legal-BERT (Chalkidis et al., 2020) models for ECHR and SCOTUS, respectively. These are fully supervised encoder-only transformers, specialized for the target tasks.

In all cases, we observe notable improvements going from zero-shot to few-shot prompting, which implies that more context and task demonstration are quite helpful. However, performance remains well below the supervised approaches. This suggests that the OPT models, despite being large, general purpose language models with similar capabilities to GPT-3, struggle to perform the specialized legal tasks considered here without fine-tuning or domain-adaptation techniques. Results comparing question-answering to completion-style prompts for SCOTUS are shown in Appendix C, Table 3.

Furthermore, we note that performance for CoT is lower than zero-shot prompting. We find that the OPT output/answer generations are not of high quality, and the generations actually add noise to the prompt, likely causing the degradation. This is an interesting finding as it illustrates that CoT improves performance only in cases where the LLM is capable of generating relevant reasoning. There is a growing body of evidence that LLMs whose pre-training corpus meaningfully includes code are

Model		**ECHR**						**SCOTUS**					
	Approach	ma-P	ma-R	ma-F1	we-P	we-R	we-F1	ma-P	ma-R	ma-F1	we-P	we-R	we-F1
OPT6.7**B**	Zero-shot	0.50	0.50	0.43	0.55	0.65	0.54	0.12	0.12	0.07	0.16	0.05	0.05
	Few shot	0.63	0.62	0.55	0.69	0.55	0.55	0.21	0.15	0.14	0.26	0.17	0.15
	Zero-shot + CoT	0.45	0.47	0.44	0.51	0.59	0.53	0.01	0.09	0.01	0.01	0.04	0.02
	Few-shot + CoT	0.46	0.49	0.43	0.52	0.62	0.54	0.01	0.08	0.02	0.03	0.16	0.05
	Few-shot + Ensemble	0.55	0.54	0.54	0.59	0.62	0.60	0.21	0.20	0.16	0.40	0.23	0.19
	Activation finetuning	0.59	0.59	0.59	0.63	0.62	0.62	0.23	0.25	0.23	0.30	0.30	0.29
	Activation finetuning + prompt	**0.69**	**0.67**	**0.68**	**0.72**	**0.73**	**0.72**	**0.42**	**0.51**	**0.44**	**0.58**	**0.61**	**0.58**
OPT175**B**	Zero-shot	0.33	0.50	0.40	0.43	0.66	0.52	0.07	0.16	0.07	0.09	0.09	0.07
	Few shot	0.61	0.51	0.44	0.62	0.66	0.55	0.32	0.27	0.17	0.56	0.30	0.23
	Zero-shot + CoT	0.44	0.49	0.41	0.51	0.64	0.53	0.11	0.16	0.09	0.14	0.09	0.09
	Few-shot + CoT	0.40	0.49	0.40	0.48	0.64	0.52	0.07	0.10	0.07	0.10	0.14	0.09
	Few-shot + Ensemble	0.61	0.57	0.44	0.68	0.46	0.41	0.11	0.20	0.11	0.15	0.16	0.13
	Activation finetuning	0.68	0.67	0.67	0.71	0.71	0.71	0.11	0.10	0.09	0.30	0.21	0.23
	Activation finetuning + prompt	**0.79**	**0.73**	**0.75**	**0.79**	**0.79**	**0.78**	**0.53**	**0.57**	**0.54**	**0.67**	**0.71**	**0.68**
Supervised	-	0.85	0.78	0.80	-	-	-	-	-	0.66	-	-	-

Table 1: Results from experiments using different prompting techniques. Supervised results are from (Chalkidis et al., 2019) and (Chalkidis et al., 2022) for ECHR and SCOTUS, respectively, using all the available data. The prefixes ma- and we- denote macro and weighted metrics, respectively. R denotes recall and P, precision.

often better at reasoning and responding to CoT prompting (Li et al., 2023; Liang et al., 2022). This may be part of why OPT, which is not explicitly pre-trained on code, does not benefit from CoT prompt structures. It is also possible that reasoning hallucinations degrade the effectiveness of such prompts in this domain. However, qualitative assessments suggest that poor reasoning quality is the major contributing factor.

For OPT6.7B, few-shot learning with ensemble is the best performing approach, excluding activation tuning, for both the ECHR and SCOTUS datasets. For the larger OPT175B model, the results for ensembling are weaker, generally failing to surpass standard few-shot prompting for both datasets. These results suggest that ensembled prompts can be beneficial for this task, but require careful construction and potentially the use of more advanced ensemble approaches.

Activation finetuning shows the best performance among all the methods analyzed. Similar to the findings of Azaria and Mitchell (2023), we observe that the internal states of the pre-trained LLM contain rich information on which to train a classifier. Moreover, few-shot prompting the LLM further improves performance in this context. We see increases in macro-$F1$ percentage points of 9 and 8 for ECHR dataset with OPT6.7B and OPT175B, respectively, by including informative prompts. In the case of SCOTUS, the difference is even more prominent, with 21 and 45 point improvements for the same models.

While the results remain below the fully supervised models for both tasks, only a small proportion of available training data is used and this approach is quite competitive, even for a difficult long-tail classification task. In addition, pure prompting approaches require no task-specific training and activation fine-tuning only trains a small number of parameters. Finally, the activation finetuning results strongly suggest that pre-trained LLMs have the potential to perform these tasks well with proper treatment.

4 Conclusions

In this research, we performed a comparative study of different prompting techniques to assess their utility in the legal domain. The considered prompting techniques are generally helpful, but their applicability to specialized domains requires further exploration, in line with findings for medical texts (Toma et al., 2023). Among the approaches tested, activation finetuning improves performance the most, especially when accompanied by a few-show prompt, even with limited labeled data. For OPT175B, this approach improves the macro-F1 score zero-shot baselines by **35** and **47** points for ECHR and SCOTUS, respectively. Of the pure prompting approaches, few-shot prompts with and without ensembling perform best, while CoT prompting does not work well in this context.

In future work, we plan to explore instruction-tuned LLMs, such as OPT-IML (Iyer et al., 2023), and experiment with PEFT methods (Hu et al., 2022). We are also interested in assessing whether current LLMs exhibit notable biases in their legal decision processes with respect to protected groups (Liang et al., 2022). Finally, we aim to explore how recent advances in reasoning-based prompting can be used to produce good task performance and more interpretable responses (Tian et al., 2023).

Limitations

We have experimented using OPT model variations and several prompting techniques. Our findings might not generalize to other LLMs or other prompting strategies. In addition, we focused on legal domain datasets, and our analysis might not readily transfer to documents from other domains. Foundation models, like OPT, are not instruction fine-tuned, and they may not respond reliably to instruction-style prompts. Therefore, we presume the results we have observed are also limited to the prompt templates that we have leveraged and might not generalize to different prompt templates.

References

Monica Agrawal, Stefan Hegselmann, Hunter Lang, Yoon Kim, and David Sontag. 2022. Large language models are few-shot clinical information extractors. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1998–2022, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Mikel Artetxe, Shruti Bhosale, Naman Goyal, Todor Mihaylov, Myle Ott, Sam Shleifer, Xi Victoria Lin, Jingfei Du, Srinivasan Iyer, Ramakanth Pasunuru, Giridharan Anantharaman, Xian Li, Shuohui Chen, Halil Akin, Mandeep Baines, Louis Martin, Xing Zhou, Punit Singh Koura, Brian O'Horo, Jeffrey Wang, Luke Zettlemoyer, Mona Diab, Zornitsa Kozareva, and Veselin Stoyanov. 2022. Efficient large scale language modeling with mixtures of experts. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 11699–11732, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Amos Azaria and Tom Mitchell. 2023. The internal state of an llm knows when its lying. *arXiv preprint arXiv:2304.13734*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Ilias Chalkidis, Ion Androutsopoulos, and Nikolaos Aletras. 2019. Neural legal judgment prediction in English. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4317–4323, Florence, Italy. 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.

Ilias Chalkidis, Abhik Jana, Dirk Hartung, Michael Bommarito, Ion Androutsopoulos, Daniel Katz, and Nikolaos Aletras. 2022. LexGLUE: A benchmark dataset for legal language understanding in English. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4310–4330, Dublin, Ireland. Association for Computational Linguistics.

Bharath Chintagunta, Namit Katariya, Xavier Amatriain, and Anitha Kannan. 2021. Medically aware GPT-3 as a data generator for medical dialogue summarization. In *Proceedings of the Second Workshop on Natural Language Processing for Medical Conversations*, pages 66–76, Online. 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, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling instruction-finetuned language models.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Zhiyong Wu, Baobao Chang, Xu Sun, Jingjing Xu, Lei Li, and Zhifang Sui. 2023. A survey on in-context learning.

M.A. Ganaie, Minghui Hu, A.K. Malik, M. Tanveer, and P.N. Suganthan. 2022. Ensemble deep learning: A review. *Engineering Applications of Artificial Intelligence*, 115:105151.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Srinivasan Iyer, Xi Victoria Lin, Ramakanth Pasunuru, Todor Mihaylov, Daniel Simig, Ping Yu, Kurt Shuster, Tianlu Wang, Qing Liu, Punit Singh Koura, Xian Li, Brian O'Horo, Gabriel Pereyra, Jeff Wang, Christopher Dewan, Asli Celikyilmaz, Luke Zettlemoyer, and Ves Stoyanov. 2023. Opt-iml: Scaling language model instruction meta learning through the lens of generalization.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2023. Large language models are zero-shot reasoners.

Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The power of scale for parameter-efficient prompt tuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3045–3059. Association for Computational Linguistics.

Raymond Li, Loubna Ben Allal, Yangtian Zi, Niklas Muennighoff, Denis Kocetkov, Chenghao Mou, Marc Marone, Christopher Akiki, Jia Li, Jenny Chim, Qian Liu, Evgenii Zheltonozhskii, Terry Yue Zhuo, Thomas Wang, Olivier Dehaene, Mishig Davaadorj, Joel Lamy-Poirier, João Monteiro, Oleh Shliazhko, Nicolas Gontier, Nicholas Meade, Armel Zebaze, Ming-Ho Yee, Logesh Kumar Umapathi, Jian Zhu, Benjamin Lipkin, Muhtasham Oblokulov, Zhiruo Wang, Rudra Murthy, Jason Stillerman, Siva Sankalp Patel, Dmitry Abulkhanov, Marco Zocca, Manan Dey, Zhihan Zhang, Nour Fahmy, Urvashi Bhattacharyya, Wenhao Yu, Swayam Singh, Sasha Luccioni, Paulo Villegas, Maxim Kunakov, Fedor Zhdanov, Manuel Romero, Tony Lee, Nadav Timor, Jennifer Ding, Claire Schlesinger, Hailey Schoelkopf, Jan Ebert, Tri Dao, Mayank Mishra, Alex Gu, Jennifer Robinson, Carolyn Jane Anderson, Brendan Dolan-Gavitt, Danish Contractor, Siva Reddy, Daniel Fried, Dzmitry Bahdanau, Yacine Jernite, Carlos Muñoz Ferrandis, Sean Hughes, Thomas Wolf, Arjun Guha, Leandro von Werra, and Harm de Vries. 2023. Starcoder: may the source be with you!

Xiang Lisa Li and Percy Liang. 2021. Prefix-tuning: Optimizing continuous prompts for generation. *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, abs/2101.00190.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, Benjamin Newman, Binhang Yuan, Bobby Yan, Ce Zhang, Christian Cosgrove, Christopher D. Manning, Christopher Ré, Diana Acosta-Navas, Drew A. Hudson, Eric Zelikman, Esin Durmus, Faisal Ladhak, Frieda Rong, Hongyu Ren, Huaxiu Yao, Jue Wang, Keshav Santhanam, Laurel Orr, Lucia Zheng, Mert Yuksekgonul, Mirac Suzgun, Nathan Kim, Neel Guha, Niladri Chatterji, Omar Khattab, Peter Henderson, Qian Huang, Ryan Chi, Sang Michael Xie, Shibani Santurkar, Surya Ganguli, Tatsunori Hashimoto, Thomas Icard, Tianyi Zhang, Vishrav Chaudhary, William Wang, Xuechen Li, Yifan Mai, Yuhui Zhang, and Yuta Koreeda. 2022. Holistic evaluation of language models.

Ruotian Ma, Xin Zhou, Tao Gui, Yiding Tan, Linyang Li, Qi Zhang, and Xuanjing Huang. 2022. Template-free prompt tuning for few-shot NER. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5721–5732, Seattle, United States. Association for Computational Linguistics.

OpenAI. 2023. Gpt-4 technical report.

Silviu Pitis, Michael Ruogu Zhang, Andrew Wang, and Jimmy Ba. 2023. Boosted prompt ensembles for large language models. *ArXiv*, abs/2304.05970.

Emily Reif, Daphne Ippolito, Ann Yuan, Andy Coenen, Chris Callison-Burch, and Jason Wei. 2022. A recipe for arbitrary text style transfer with large language models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 837–848, Dublin, Ireland. Association for Computational Linguistics.

J. Sivaloganathan, M. Coatsworth, J. Willes, M. Choi, and G. Shen. 2022. *Kaleidoscope*. Vector Institute for Artificial Intelligence, Toronto, Canada.

Jacob-Junqi Tian, Omkar Dige, David Emerson, and Faiza Khan Khattak. 2023. Interpretable stereotype identification through reasoning.

Augustin Toma, Patrick R Lawler, Jimmy Ba, Rahul G Krishnan, Barry B Rubin, and Bo Wang. 2023. Clinical camel: An open-source expert-level medical language model with dialogue-based knowledge encoding. *arXiv preprint arXiv:2305.12031*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. LLaMA: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Dietrich Trautmann, Alina Petrova, and Frank Schilder. 2022. Legal prompt engineering for multilingual legal judgement prediction. *arXiv preprint arXiv:2212.02199*.

Michiel van der Meer, Myrthe Reuver, Urja Khurana, Lea Krause, and Selene Baez Santamaria. 2022. Will it blend? mixing training paradigms & prompting for argument quality prediction. In *Proceedings of the 9th Workshop on Argument Mining*, pages 95–103, Online and in Gyeongju, Republic of Korea. International Conference on Computational Linguistics.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed Chi, Quoc Le, and Denny Zhou. 2022. Chain of thought prompting elicits reasoning in large language models. *arXiv preprint arXiv:2201.11903*.

Hanzi Xu, Slobodan Vucetic, and Wenpeng Yin. 2022. OpenStance: Real-world zero-shot stance detection. In *Proceedings of the 26th Conference on Computational Natural Language Learning (CoNLL)*, pages 314–324, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Fangyi Yu, Lee Quartey, and Frank Schilder. 2022. Legal Prompting: Teaching a Language Model to Think Like a Lawyer. In *Proceedings of the Natural Legal Language Processing Workshop (NLLP)*.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. 2022. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*.

A ECHR and SCOTUS Dataset Details

Table 2 presents the details of each dataset after down-sampling. Note that the down-sampling process respected the original train, validation, and test splits, such that no cross-split pollution occurred. The analysis in the table is at the word level, rather than the token level. All text sequences are truncated after tokenization, if required. For ECHR, the task considered is binary classification determining whether a human rights article has been violated based on the provided context. We specifically use the anonymized version of the ECHR dataset. For SCOTUS, the task is to categorize Supreme Court case decisions into one of 14 issue areas. These issue areas include topics such as Criminal Procedure, Civil Rights, and Economic Activity.

Dataset	Train	Val.	Test	Avg. # of Words
ECHR	710	138	299	1947
SCOTUS	250	72	70	10095

Table 2: The number of documents in the ECHR and SCOTUS training, validation, and test datasets. The average number of words are calculated using white-space separation in the test splits for each dataset.

B Activation Fine-Tuning Setup

The same hyperparameters and training procedure are used to train the feed-forward network for both the ECHR and SCOTUS datasets, {*batch_size*: 16, *lr*: 0.0001, *weight_decay*: 0.01, *hidden_size*: 128 *patience*: 3}. In the case of SCOTUS, weighted cross entropy is used to address class imbalance.

C Prompting Strategy Results: Visualization

Figures 1 and 2 provide visualizations of results shown in Table 1 associated with the various prompting strategies discussed in Section 2.4. Solid and dashed lines correspond to macro- and weighted-$F1$ scores, respectively. Horizontal dashed lines in each figure represent the macro-$F1$ scores of HIER-BERT (Chalkidis et al., 2019) for ECHR and Legal-BERT (Chalkidis et al., 2022) for SCOTUS. For the prompt strategies, "Ens." stands for ensembling, "Act." denotes activation, and "FT" is short for finetuning.

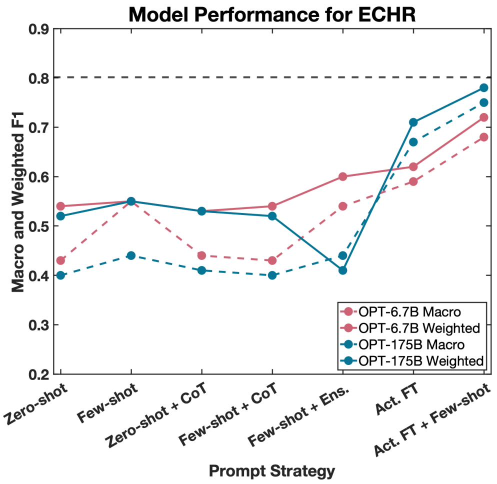

Figure 1: Macro- and weighted-$F1$ scores for the OPT models on the ECHR task. The dashed horizontal line represents the macro-$F1$ score for HIER-BERT in (Chalkidis et al., 2019).

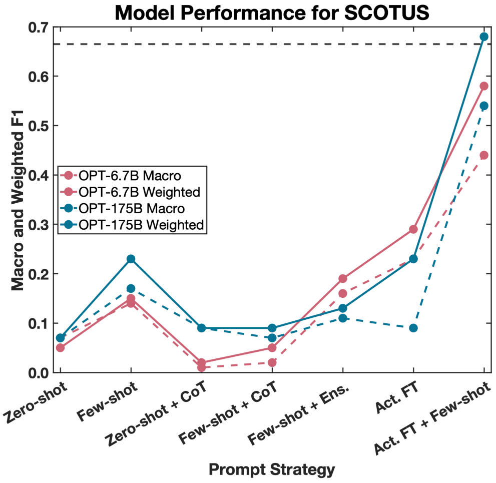

Figure 2: Macro- and weighted-$F1$ scores for the OPT models on the SCOTUS task. The dashed horizontal line represents the macro-$F1$ score for Legal-BERT in (Chalkidis et al., 2022).

Table 3 summarizes the performance differences for prompts following a question-answer or completion-style for the SCOTUS dataset. The results suggest that, when prompting OPT in this domain and potentially other LLMs without special fine-tuning, it is better to formulate the task in a completion-style format rather than a question-answering one. These findings are in line with Touvron et al. (2023) on effective prompt design.

Model	Prompt type	ma-P	ma-R	ma-F1
OPT6.7**B**	Question	0.02	**0.12**	0.03
	Completion	**0.12**	**0.12**	**0.07**
OPT175**B**	Question	0.04	0.13	0.05
	Completion	**0.07**	**0.16**	**0.07**

Table 3: Details of SCOTUS dataset results with different prompt styles in a zero-shot setting. Completion-style prompts perform better for both model sizes.

D Prompt Templates

To illustrate the appearance of the tasks within the context of the prompts, we provide examples of completed prompt templates for ECHR and SCOTUS below. As the few-shot prompts are quite long, only completed zero-shot templates are shown.

Complete the sentence with Yes or No
NORP The applicant, Mr PERSON LOC, is a NORP national who was born on DATE and lives in GPE. From DATE to DATE the applicant was the director general of a NORP limited liability company, ORG (hereinafter "the company"). On DATE ORG of GPE declared the company bankrupt and opened liquidation proceedings. The proceedings were closed on DATE. On DATE ORG of GPE considered a request from ORG to hold the applicant jointly liable for the company's debts because of the company's inability to cover all its debts with its remaining assets. Although duly notified, the applicant was not present at the hearing. ORG request was granted and the applicant was found liable to pay MONEY (RUB, MONEY). On DATE the company officially went into liquidation. The applicant only became aware of the outcome of the proceedings against him in DATE when money was seized from his bank account. The applicant appealed against the decision of ORG of GPE of DATE. The CARDINALth ORG of Appeal considered the appeal on the merits and on DATE discontinued the appeal proceedings. With reference to the position of ORG, the appeal court concluded that it was impossible to adjudicate bankruptcy-related disputes concerning a company that had already gone into liquidation. The operative part of the appeal decision stated that the applicant could file a cassation appeal. On DATE ORG of GPE dismissed a cassation appeal of the applicant
Has any human rights article or protocol been violated?

Can you classify the given "Text" with only one class from the following list? [Attorneys, Civil Rights, Criminal Procedure, Due Process, Economic Activity,Fed Taxation, Federalism, First Amendment, Interstate Relations, Judicial Power, Miscellaneous, Privacy, Unions]
In Kaupp v. Texas, 538 U.S. 626 (2003), the Supreme Court vacated the judgment of the Texas Court of Appeals, holding that Kaupp was arrested within the meaning of the Fourth Amendment before the detectives began to question him. Kaupp, then 17 years old, was implicated in the murder of a 14-year-old girl by the confession of the girl's half brother. Detectives attempted but failed to obtain a warrant to question Kaupp. They instead went to his house at 3 a.m., awakened and handcuffed him, and took him to the sheriff's headquarters, where they advised him of his rights under Miranda v. Arizona, 384 U. S. 436. Once presented with the brother's confession, Kaupp admitted to having a part in the crime. Kaupp moved unsuccessfully to suppress his confession as the fruit of an illegal arrest. On appeal, the Supreme Court held that Kaupp was arrested before he was questioned and that the State did not have probable cause to detain him. The Court found that Kaupp's detention was "in important respects indistinguishable from a traditional arrest" and required probable cause or judicial authorization to be legal. The State's failure to demonstrate that Kaupp's confession was "an act of free will [sufficient] to purge the primary taint of the unlawful invasion" required suppression of the confession. The Supreme Court vacated the judgment of the Texas Court of Appeals and remanded the case for further proceedings consistent with its opinion.