Harnessing GPT-3.5-turbo for Rhetorical Role Prediction in Legal Cases

Anas Belfathi, Nicolas Hernandez, Laura Monceaux

September 2023

Harnessing GPT-3.5-turbo for Rhetorical Role Prediction in Legal Cases

Anas BELFATHI [a], Nicolas HERNANDEZ [a] and Laura MONCEAUX [a]
[a] *Nantes Université, École Centrale Nantes, CNRS, LS2N, UMR 6004, France*

Abstract.
We propose a comprehensive study of one-stage elicitation techniques for querying a large pre-trained generative transformer (GPT-3.5-turbo) in the rhetorical role prediction task of legal cases. This task is known as requiring textual context to be addressed. Our study explores strategies such as zero-few shots, task specification with definitions and clarification of annotation ambiguities, textual context and reasoning with general prompts and specific questions. We show that the number of examples, the definition of labels, the presentation of the (labelled) textual context and specific questions about this context have a positive influence on the performance of the model. Given non-equivalent test set configurations, we observed that prompting with a few labelled examples from direct context can lead the model to a better performance than a supervised fined-tuned multi-class classifier based on the BERT encoder (weighted F1 score of ≈72%). But there is still a gap to reach the performance of the best systems ≈86%) in the LegalEval 2023 task which, on the other hand, require dedicated resources, architectures and training.

Keywords. rhetorical role prediction, legal domain, case law, in-context learning, prompt engineering, generative large language model, gpt-3.5-turbo

1. Introduction

Large Language Models (LLMs) have proved effective for a variety of applications, but adapting them to a task or a specialized domain remains a major challenge. In recent years, prompting Generative LLMs has become a dominant paradigm as a first approach to solving various downstream tasks [1]. However, few studies have focused on the task of rhetorical role prediction using generative approaches, in particular in legal cases [2]. Kalamkar et al. showed that labelling sentences of legal cases with rhetorical roles, such as Facts, Arguments or Analysis, improve performance on the tasks of summarization and legal judgment prediction [3].

This research paper focuses on evaluating the potential of generative pre-trained transformers (GPT), specifically OpenAI's GPT-3.5-turbo [1,4], to autonomously conduct rhetorical analysis on sentences extracted from legal cases. Generative approaches in the legal domain are appealing because adapting a LLM by fine-tuning it requires annotated data, which is expensive to produce for each court in each country. In addition, predicting the rhetorical label of a sentence requires taking into account its textual context, and even the text as a whole. But the capacity of the state-of-the-art LLMs does not always allow them to take into account the entirety of a legal case [5]. The state-of-the-art systems define the problem as a sequence labelling task [6]. Because of the genera-

tive aspect of the GPT-3.5-turbo model and inspired by Savelka et al. [2], we define the problem as a multi-classification task through experimentation with several prompting strategies, such as Zero-Few shot prompting, Chain-of-Thought reasoning, specialized legal knowledge specification and textual context prompting.

We propose the following research questions: (RQ1) To what extent can GPT-3.5-turbo successfully perform labelling tasks using classical prompting techniques by giving *zero, one or a few examples*? (RQ2) To what extent does providing the *label definitions* influence the efficiency and accuracy of GPT-3.5-turbo, and how does the model benefit from the inclusion of *clarification of ambiguities between labels*? (RQ3) What are the implications of utilizing *textual context* in the label prediction of a sentence? (RQ4) How does prompting the model to *think step by step and to explain its reasoning* without specifying particular expectations affect its performance and response quality in comparison to asking *precise questions about the textual context*?

2. Related works

Prompt engineering, also known as *in-context prompting*, refers to techniques aiming at steering the Generative LLM's behavior towards a particular outcome without updating the model's parameters [7,8]. The most basic technique, called *zero-shot prompting*, consists of feeding the model with a request and asking for completion. This technique can be enhanced by offering one or a few examples of input–output pairs in the prompt that guide the model to carry out the task; The technique is so called *few-shot prompting*. Brown et al. [1] and Wei et al. [9,10] demonstrated the ability of LLMs with more than 100 billion parameters (such as 175B GPT-3) to respond successfully to such requests for several tasks, with even better results when the models were fine-tuned to respond to instructions (such as 175B GPT-3.5-turbo). By investigating GPT-3 on few-shot classification tasks, Zhao et al. [11] demonstrated that the choice of the prompt format, the training examples, and the order of the examples can affect the accuracy of the results. To select the examples, Liu et al. [12] recommend to retrieve examples that are semantically similar to the test example and Diao et al. [13] supplement by showing that examples with high disagreement or entropy (from a set of candidate examples) are among the most important and useful. Lu et al. [14] observed that generative models (like 175B GPT-3) are sensitive to the examples ordering whatever the model size or the number of examples. In their approach, called In-Context Instruction Learning (ICIL), Ye et al. [15] showed that providing a fixed prompt with multiple cross-task demonstrations[1] as context of a third-party task query enhance the model performance on several tasks. The authors suggest that effectiveness comes from 1) selecting classification tasks that include explicit answer choice in the instruction and 2) retrieving demonstrations that are similar to the target task. Recent works have shown that explaining the reasoning or *Chain-of-Thought* (CoT), required to solve a task, increases the performance of generative models in solving the task [16]. Reasoning can be seen as decomposing a problem into a sequence of sub-problems either iteratively or recursively [17]. Surprisingly, simply encouraging the model to reason (by adding "Let's think step by step" before an answer) can also improve the generation [18]. Ye and Durrett [19] have shown that GPT-3.5 benefits substantially

[1] Where each demonstration is a concatenation of an instruction, input, and output instance of a task.

from prompting with explanations for reasoning over text (question answering and entailment). Fu et al. [20] have shown that prompts with higher reasoning complexity (i.e. chains with more reasoning steps) achieve better performance than simple prompts on math word reasoning tasks.

Generative LLM for legal tasks. Measuring the influence of generative models on legal tasks has become one of the main concerns of NLP researchers working in this field [21, 2,22]. On a legal entailment task (question answering task based on a legal article of a few sentences), Yu et al. [22] showed that giving the article in the prompt and asking a GPT-3 model to analyse it according to a given rhetorical schema (corresponding to a legal reasoning approach) improved performance compared to few-shot examples techniques or a zero-shot CoT [18] strategy. Savelka et al. [2] questioned the use of GPT-4 for multi-class sentence classification tasks on US court opinions. They experimented with prompts containing annotation guidelines originally designed for human annotators, with clarifications of ambiguities between labels and with requests for explanations. They did not measure the contribution of the annotation guidelines but they observed that the model performance is comparable to that of the best-performing law student annotators. They showed that disambiguation of labels enhance the performance. Eventually they found that asking the model to explain its choice of label reduces performance.

The rhetorical role prediction task The SemEval 2023 LegalEval shared task [6] provides a good insight of the dominant approach in addressing the rhetorical role prediction task. Most of the participants defined the problem as a sequence sentence classification task and adopted a system architecture based on the Hierarchical Sequential labelling Network (HSLN) [3,23], denoted as *SciBERT-HSLN*. The best system [24], denoted *AntContentTech*, equipped *SciBERT-HSLN* with domain-adaptive pretraining, data augmentation strategies, as well as auxiliary-task learning techniques. On the LegalEval test dataset, *SciBERT-HSLN* obtained a weighted F1 score of 0.79 while *AntContentTech* obtained a score of 0.8593. For an indicative comparison, [5][2] reported a score of 0.65 with a simple BERT fine-tuned for single sentence classification (hereinafter denoted as *BERT*), and between 0.75-0.77 with architecture adaptations to take into account the local context of the sentence to label (denoted as *BERT*+*local_context*).

3. Exploring Various GPT Prompting Strategies

Our prompts are inspired from [2]. The template is made of four parts: `PREAMBLE`, `EXAMPLES-or-CoT`, `INPUT` and `REQUEST`. The `PREAMBLE` is common to all prompts and takes 211 tokens (See Figure 1 (a)). It sets the persona and the domain and the task definition of the forthcoming `REQUEST`. The `EXAMPLES-or-CoT` is the more volatile part of the prompt. It will be presented in detail in the following sections. The `INPUT` part has a simple form: "`SENTENCE:` $\backslash n$‘‘‘{`sentence`}’’’". And so has the `REQUEST` part: "`EXPECTED OUTPUT FORMAT:` $\backslash n$`Label: <label>`". In practice, this part may be more specified depending on the experiment. The presentation of the task, the label definitions and the bootstrapping examples come from the LegalEval 2023 task [3].

[2]Since the authors did not have access to LegalEval test dataset, training, validation and test were performed on 80/10/10 splits from the concatenation of the train and dev LegalEval datasets.

(a)	You are a specialized system focused on semantic annotation of court opinion. \n RHETORICAL ROLE: Rhetorical roles in legal writing refer to the distinct functions or purposes that different parts of a document, such as a legal opinion, serve in conveying information, persuading the reader, and constructing a coherent argument. These roles encompass various elements like factual background, legal principles, arguments, counter arguments, and conclusions, each contributing to the document's overall persuasive and informative structure. \n labelling TASK: Please label each sentence in the document with one of the following predefined rhetorical roles: 'Preamble', 'Facts', 'Ruling by Lower Court', 'Issues', 'Argument by Petitioner', 'Argument by Respondent', 'Analysis', 'Statute', 'Precedent Relied', 'Precedent Not Relied', 'Ratio of the decision', 'Ruling by Present Court', 'NONE'. Assign the role that best describes the purpose or function of each sentence in the context of the legal opinion.
(b)	EXAMPLES: SENTENCE: IN THE COURT OF THE IV ADDL SESSIONS JUDGE, CHENNAI. Dated this the 10th day of September 2023. LABEL: Preamble SENTENCE: SUPREME COURT OF INDIA. Dated this the 5th day of June 2022. This judgment pertains to the case of John Doe versus Jane Smith. LABEL: Preamble\ n [...]
(c)	ANNOTATION GUIDELINES: - 'Preamble':A typical judgement would start with the court name, the details of parties, lawyers and judges' names, Headnotes. This section typically would end with a keyword like (JUDGEMENT or ORDER etc.). Some supreme court cases also have HEADNOTES, ACTS section. They are also part of Preamble. - 'Issues': Some judgements mention the key points on which the verdict needs to be delivered. Such Legal Questions Framed by the Court are ISSUES. \ n E.g. "he point emerge for determination is as follow:- (i) Whether on 06.08.2017 the accused persons in furtherance of their common intention intentionally caused the death of the deceased by assaulting him by means of axe ?"\ n [...]
(d)	ANNOTATORS QUALITY ASSESSMENT: It is important to note that during the annotation process, certain patterns emerged in annotators' assessments: ∗ High Agreement: Amongst annotators, high agreement was observed for 'Preamble', 'Ruling by Present Court', 'NONE', and 'Issues'. ∗ Medium Agreements: For 'Facts', 'Ruling by Lower Court', 'Analysis', 'Precedent Relied', and 'Argument by Petitioner' and 'Argument by Respondent', medium agreements were noted.\ n [...]
(e)	CONTEXT SENTENCES: SENTENCE: "'It entered into transactions in the nature of forward transactions with parties at Bhatinda (in the Patiala State outside the taxable territories of British India) in which it suffered losses."' LABEL: Preamble SENTENCE: "'The assessee claimed deduction of these losses in the computation of its income."' LABEL: Preamble \ n [...]
(f)	EXPECTED OUTPUT FORMAT(Give your response in a json format. Stick to less than 30 words): {{\n "Let's think step by step": <reasoning why particular label should be assigned >\n "Label": <label >\n }}
(g)	EXPECTED OUTPUT FORMAT(Give your response in a json format. Stick to less than 30 words): {{\n "Label": <label >, \n "Relative position": "<return the relative position corresponding to the most sentence presented in the context (NEXT SENTENCES) that impact more on the decision>", \n "Sentence": "<return the full text of the impacted sentence corresponding to the relative position>", \n "Terms": "<List up to 5 words from the context and the predicted sentence that significantly influence the decision, in array format>"\n }} """

Figure 1. Various possible parts of the prompts: PREAMBLE (a); `EXAMPLES-or-CoT` with Few-Shot Prompting (b), with Label definitions (c), with Clarification of label ambiguities (d) in extension of (c), with labelled textual context (e); `REQUEST` to encourage the model to reason (f), with specific questions (g).

3.1. Zero-Few Shot Prompting (RQ1)

Our first experiment was to assess the proficiency of GPT in performing our task using zero-few-shot prompting. For *zero-shot* prompting, we left the `EXAMPLES-or-CoT` part empty. For one and more shot prompting, we left the problem of selecting significant examples for future work. Instead we asked GPT-3.5 to generate examples by taking inspiration from the examples given in the explanatory Figure about the Rhetorical Roles of [3]. We assumed that they were representative and that the model would better understand something that it had generated itself. The prompt we used for generate the examples was: `Given these examples of each Rhetorical Roles label, generate four representative sentences for each label`. We limited the generation to four sets of examples due to the input length limits of the model

(one shot was about 850 tokens). Figure 1 (b) shows an illustration of the use of examples generated for a two-shot prompting.

3.2. Label Definitions and Clarification Between Labels (RQ2)

In this experiment, we sought to explore the impact of providing label definitions, possibly supplemented by the clarification from the annotator errors (denoted as *definition+clarifications*). So the `EXAMPLE-or-CoT` part of the prompt was first fed with the label definitions provided by Kalamkar et al. [3] in table of an appendix of the online resources[3] (See Figure 1 (c)). Subsequently, to address the consequences of introducing clarifications about the annotator errors and ambiguities between labels to the GPT model, we extended the label definitions with the content of the "Annotation Quality Assessment" section of [3] (See Figure 1 (d)). To ensure alignment with the narrative of our prompt message, we made some minor modifications, including the removal of references and the organization of the paragraph into distinct points, each addressing separate ambiguities. In order to have a complete view, we also have combined the definitions with four-shot examples.

3.3. Textual Context Enrichment (RQ3)

The main objective of this experiment was to examine the impact of presenting the textual context of an input sentence to the model. The coherence of a text is expressed by the fact that consecutive sentences are linked by thematic and rhetorical relationships. The underlying idea is to get the model to exploit this information. We also discuss the fact of providing the labels of the sentences in this context. Indeed the experiment can be seen as a variant of few-shot learning where examples are selected for certain reasons (their belonging to the textual context of the target sentence), possibly coming without labels. Based on Belfathi et al. [5], we studied both the direction and the size of the context window to consider. In practice, we experimented with adding 2 or 8 preceding or following sentences in the `EXAMPLE-or-CoT` part of the prompt (See Figure 1 (e)).

3.4. Encouraging General or Specific Reasoning (RQ4)

The aim of this experiment was to observe the behaviour of the model with general reasoning questions compared with specific questions. General questions were implemented by adding the expression "Let's think step by step" [18] into the `REQUEST` part of the prompt and by specifying we were expecting an explanation about the choice of a label by the model (See Figure 1 (f))[4]. To probe the model's reasoning capacity with specific questions we targeted questions about the textual context. Based on the configuration that obtained the best results in the experiment described in Section 3.3 (i.e. inserting the following 8 labelled sentences), we asked the model about the relative position of the most influential sentence within the context and the relevant legal terms that impact the decision (See Figure 1 (g))[4].

[3] `https://github.com/Legal-NLP-EkStep/rhetorical-role-baseline`

[4] We instructed the model to provide the results in a JSON format to facilitate the evaluation process, and limited the token generation to 30 words to manage costs associated with the API usage.

4. Experimental setting

Table 1. Statistical Distribution: Percentage for Each Rhetorical Role in the LegalEval Corpus and Their Segments Used in Our Experimental Subset (%)

Dataset	Analysis	FACTS	PRMBL	NONE	PRE-R.	ARG-PET	RPC	RLC	RATIO	ARG-RES	STA	ISSUE	PRE-NOT-R.
LegalEval	36.65	19.84	14.67	5.06	4.93	4.34	3.67	2.72	2.33	2.30	1.59	1.30	0.53
Experimental	30.33	18.34	18.43	6.90	7.81	1.63	3.63	6.63	1.72	1.18	1.36	1.72	0.27

Data We utilized the data provided by Sub-task A, "Rhetorical Roles Prediction," of the SemEval 2023 Task 6, "LegalEval - Understanding Legal Texts" challenge [3][5]. This dataset consists of Indian legal data extracted from court judgments, featuring 13 distinct rhetorical roles (RRs). and averaging 117.31 sentences per document. To prepare for our experiments, we randomly selected 10 documents (1,101 sentences) from the validation data to manage costs associated with using the GPT API. We observed that there was less variation between the original LegalEval dataset and the segments chosen for our experimentation (See Table 1).

Model parameters We experimented the `gpt-3.5-turbo` model (`-0613` snapshot from June 13th 2023) which extends `text-davinci-003` (175B GPT-3 LLM trained on code-completion tasks and fine-tuned on natural language instruction tasks) with optimization for chat[6]. Its maximum input length is 4,096 tokens. To ensure that its completion was deterministic, we set the temperature for all experiments to 0. Other parameters were set to their default values (Top P=1, Frequency penalty = 0, Presence penalty = 0). The cost of all experiments was 68 euros.

Measures The performance of the NLP models for the rhetorical roles task is assessed using Weighted-Precision ($_wP$), Weighted-Recall ($_wR$), Accuracy (A), Weighted-F1 ($_wF1$) and Macro F1 ($_MF1$) scores based on the hidden test set. The weighted F1 score considers both precision and recall, and it is calculated by taking into account the class-wise F1 scores weighted by the number of samples in each class.

5. Results and discussion

5.1. Zero-Few Shot Prompting (RQ1)

In this experiment, we examined GPT-3.5's efficiency in rhetorical role prediction within the legal domain utilizing *zero* to 4-shot prompting (See zero- and[*1-4*]-*shot* examples in Table 2). The low Macro-F1 score indicates that the Zero-Few prompts encountered challenges in label recognition, often leading to confusion between different rhetorical role labels. We can see that by increasing the number of examples, the Weighted-F1 score increases, but on the other hand the Macro-F1 score slightly decreases. As confirmed by Table 3, this means that the addition of examples mainly benefits certain classes, and that these are well represented in the corpus.

[5] `https://sites.google.com/view/legaleval`

[6] See `https://platform.openai.com/docs/model-index-for-researchers` and `https://platform.openai.com/docs/models` for more details.

Table 2. Performance of Prompting Strategies Ordered by Research Question (RQ). Reported results for *BERT* and *BERT+local_context* [5], *SciBERT-HSLN* [3], and *AntContentTech* [24] are given for information. All were trained on the same dataset source (LegalEval 2023) but with various splits and amounts of data.

RQ	Model	$_wP$	$_wR$	A	$_wF1$	$_MF1$
1	zero-shot example	0.42	0.34	0.34	0.33	0.29
	one-shot example	0.45	0.33	0.33	0.33	0.30
	2-shot example	0.45	0.34	0.34	0.36	0.29
	4-shot example	0.46	0.35	0.35	0.37	0.28
2	definition	0.46	0.42	0.42	0.42	0.33
	definition+clarification	0.46	0.41	0.41	0.41	0.32
	definition+examples	0.49	0.41	0.41	0.42	0.33
3	context-2	0.45	0.39	0.39	0.39	0.32
	context-8	0.45	0.36	0.37	0.36	0.29
	context+2	0.46	0.43	0.43	0.42	0.36
	context+8	0.43	0.39	0.39	0.38	0.31
	labelled_context-2	0.66	0.63	0.63	0.63	0.50
	labelled_context-8	0.71	0.68	0.68	0.68	0.50
	labelled_context+2	0.69	0.66	0.66	0.66	0.51
	labelled_context+8	0.72	0.70	0.70	0.70	0.53
4	zero-shot-cot	0.46	0.29	0.29	0.31	0.27
	cot-by-queries	**0.77**	**0.71**	**0.71**	**0.72**	**0.61**
	BERT [5]				0.65	
	BERT+local_context [5]				0.75-0.77	
	SciBERT-HSLN [3]				0.79	
	AntContentTech [24]				0.8593	

Table 3. Performance Measurement (F1 Score) of Models for Each Rhetorical Role Across All Experimentation Prompts. The blue cells signify the highest performance for each label.

	Analysis	ARG-PET	ARG-RES	FACTS	ISSUE	NONE	PRMBL	PRE-NOT	PRE_R.	RATIO	RLC	RPC	STA
zero-shot example	0.36	0.21	0.29	0.46	0.44	0.27	0.27	0.29	0.22	0.00	0.21	0.50	0.26
one-shot example	0.32	0.12	0.32	0.48	0.46	0.27	0.26	0.31	0.32	0.00	0.26	0.51	0.23
2-shot example	0.36	0.20	0.24	0.45	0.45	0.22	0.40	0.22	0.29	0.00	0.26	0.47	0.21
4-shot example	0.40	0.17	0.20	0.47	0.45	0.25	0.38	0.12	0.29	0.00	0.25	0.48	0.23
definition	0.45	0.26	0.26	0.54	0.49	0.13	0.50	0.20	0.22	0.09	0.30	0.58	0.22
definition+clarification	0.47	0.21	0.15	0.55	0.44	0.17	0.41	0.25	0.31	0.08	0.27	0.62	0.26
definition+examples	0.46	0.23	0.22	0.53	0.49	0.23	0.46	0.17	0.33	0.04	0.28	0.62	0.20
context-2	0.43	0.26	0.31	0.55	0.48	0.22	0.36	0.24	0.33	0.07	0.20	0.46	0.24
context-8	0.38	0.10	0.28	0.52	0.51	0.24	0.36	0.10	0.29	0.11	0.23	0.42	0.25
context+2	0.47	0.22	0.25	0.55	0.49	0.24	0.44	0.44	0.30	0.09	0.25	0.53	0.34
context+8	0.43	0.19	0.19	0.52	0.41	0.23	0.39	0.36	0.27	0.09	0.19	0.49	0.30
labelled_context-2	0.67	0.32	0.70	0.74	0.47	0.65	0.71	0.18	0.50	0.10	0.37	0.61	0.48
labelled_context-8	0.74	0.34	0.67	0.79	0.49	0.54	0.76	0.00	0.65	0.00	0.55	0.54	0.45
labelled_context+2	0.71	0.30	0.61	0.80	0.49	0.55	0.77	0.33	0.58	0.10	0.45	0.53	0.38
labelled_context+8	0.78	0.27	0.44	0.79	0.59	0.58	0.78	0.50	0.64	0.09	0.58	0.50	0.38
cot-by-queries	0.79	0.36	0.55	0.75	0.61	0.57	0.84	0.86	0.75	0.35	0.60	0.46	0.42

5.2. Label Definitions and Clarification Between Labels (RQ2)

With globally far fewer tokens (1,063), our results indicated a higher performance when employing the label definitions (See *definition* in Table 2) than providing just the examples. This suggests that the model learns better from definitions than from examples because the classes are difficult to explain with examples in our case. This opens a big question about heuristics addressed to the process of selection of examples. The impact of introducing clarifications about annotator errors and label ambiguities (*definition+clarification*), or 4-shot examples (*definition+examples*), into the model to the definition does not bring any global improvements. Some classes seem to benefit but to the detriment of others. As reported by [2], the model does not appear to assimilate

knowledge from annotators' mistakes and the inherent ambiguity, at least when they are presented as mere declarations.

5.3. Textual Context Enrichment (RQ3)

This experiment starts from the *definition* configuration and studies the impact of adding textual context to the target sentence. By adding unlabelled sentences from the context (*context*$[-+]\backslash d$ in Table 2), we can see that the performance is deteriorating. However we see that any configuration with contextual sentences give better performance than zero-few shot examples, and that contextual sentences augmented with labels (*labelled_c*$[-+]\backslash d$) outperform any of our experimented prompts. A possible explanatory hypothesis may come from the degree of similarity shared by these various types of sentences with the target sentence [12,15]. Indeed few-shot (i.e. labelled examples that do not come from the document), unlabelled contextual sentences and labelled[7] contextual sentences can be seen as three types of examples with increasing similarity and precision. Notably, across all the contextual experiments, we observed that our results consistently improved when adding the following context compared to preceding context.

5.4. Encouraging General or Specific Reasoning (RQ4)

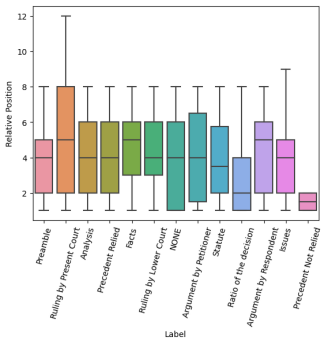

Figure 2. Box Plot Illustrating the Distribution of Relation Positions by Labels

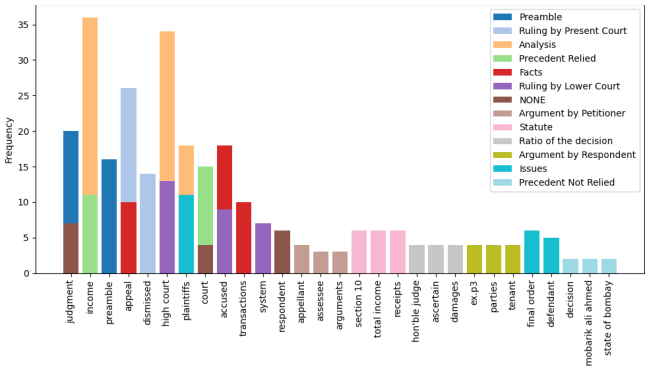

Figure 3. Top 3 terms For each Rhetorical Roles

Regarding the experimentation with a general reasoning instruction, 'Let's think step by step' (*zero-shot-cot*), added to the *definition* configuration, the model performs even worse than a zero-shot prompting. These results confirm what was discussed in [2], which indicated that GPT-3.5 struggles with correctly interpreting the annotation guidelines. When we targeted questions about the best prompt with context (*labelled_context+8*), we achieved higher performance compared to all the prompt strategies with an F1 score of 0.72 (*cot-by-queries*). In the analysis of the targeted question about the relative position of the most influential sentence (Figure 2), over 50% of sentences in the RPC, FACTS, and ARG-RES roles are impacted by sentences located at

[7] On average, 70% of the sentences that make up the 8 sentences preceding a sentence have the same label as that sentence. 80% for 2 preceding sentences.

least at position 5 within the context. However, the RATIO and PRE-NOT-RELIED roles have a lower median sensitivity, suggesting that they can be effectively recognized with shorter context sentences. Furthermore, as shown in Figure 3, certain terms, such as "income," occurring in both the Analysis and Precedent roles, and "Plaintiffs," appearing in both Analysis and ISSUES roles, led to confusion, as discussed before in [3,25]. Additionally, terms like "preamble" and "dismissed" were found to be specialized for specific roles (PREAMBLE and RPC).

5.5. Comparison with state-of-the-art

The results we report from the state-of-the-art systems (See the bottom 4 rows of Table 2) concern systems which were fine-tuned with at least 25,800 pairs of examples. Through our experiments (in particular the labelled context ones) we show that a generative LLM, prompted in one stage, can outperform a supervised fine-tuned multi-class classifier based on the Transformer encoder model (*BERT*). Although artificial, it opens the way to research. However it seems difficult for such a generative system fed with classical prompts to beat a fine-tuned system with a context representation (i.e. *BERT+local_context*, *SciBERT-HSLN* and *AntContentTech*).

6. Conclusion and Future work

This study assessed the capabilities of GPT-3.5 in analyzing legal cases for the task of rhetorical roles prediction. We show that the number of examples, the definition of labels, the presentation of the textual context and specific questions about this context have a positive influence on the performance of the model. In an artificial experiment, we observed that prompting with a few labelled examples from direct context can lead the model to a better performance than a supervised fined-tuned multi-class classifier based on the BERT encoder (weighted F1 score of $\approx$72%). But there is still a gap to reach the performance of the best systems $\approx$86%) in the LegalEval 2023 task which, on the other hand, require dedicated resources, architectures and training.

Acknowledgments

This research was funded, in whole or in part, by l'Agence Nationale de la Recherche (ANR), project ANR-22-CE38-0004.

References

[1] Brown T, Mann B, Ryder N, Subbiah M, Kaplan JD, Dhariwal P, et al. Language Models are Few-Shot Learners. In: Advances in Neural Information Processing Systems; 2020. p. 1877-901.

[2] Savelka J, Ashley KD, Gray MA, Westermann H, Xu H. Can GPT-4 Support Analysis of Textual Data in Tasks Requiring Highly Specialized Domain Expertise? In: Proceedings of the Sixth Workshop on Automated Semantic Analysis of Information in Legal Text; 2023. p. 1-12.

[3] Kalamkar P, Tiwari A, Agarwal A, Karn S, Gupta S, Raghavan V, et al. Corpus for Automatic Structuring of Legal Documents. In: Proceedings of the Thirteenth Language Resources and Evaluation Conference; 2022. p. 4420-9.

[4] Ouyang L, Wu J, Jiang X, Almeida D, Wainwright C, Mishkin P, et al. Training language models to follow instructions with human feedback. In: Advances in Neural Information Processing Systems; 2022. p. 27730-44.

[5] Belfathi A, Hernandez N, Monceaux L. Enhancing Pre-Trained Language Models with Sentence Position Embeddings for Rhetorical Roles Recognition in Legal Opinions. In: Proceedings of the 6th Workshop on Automated Semantic Analysis of Information in Legal Text; 2023. p. 19-27.

[6] Modi A, Kalamkar P, Karn S, Tiwari A, Joshi A, Tanikella SK, et al. SemEval-2023 Task 6: LegalEval - Understanding Legal Texts. In: Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023). ACL; 2023. p. 2362-74.

[7] Huang J, Chang KCC. Towards Reasoning in Large Language Models: A Survey. In: Findings of the Association for Computational Linguistics: ACL 2023; 2023. p. 1049-65.

[8] Qiao S, Ou Y, Zhang N, Chen X, Yao Y, Deng S, et al. Reasoning with Language Model Prompting: A Survey. In: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics; 2023. p. 5368-93.

[9] Wei J, Bosma M, Zhao VY, Guu K, Yu AW, Lester B, et al. Finetuned Language Models Are Zero-Shot Learners. CoRR. 2021.

[10] Wei J, Tay Y, Bommasani R, Raffel C, Zoph B, Borgeaud S, et al. Emergent Abilities of Large Language Models. Transactions on Machine Learning Research. 2022. Survey Certification.

[11] Zhao Z, Wallace E, Feng S, Klein D, Singh S. Calibrate Before Use: Improving Few-shot Performance of Language Models. In: Proceedings of the 38th International Conference on Machine Learning; 2021. p. 12697-706.

[12] Liu J, Shen D, Zhang Y, Dolan B, Carin L, Chen W. What Makes Good In-Context Examples for GPT-3? In: Proceedings of Deep Learning Inside Out. ACL; 2022. p. 100-14.

[13] Diao S, Wang P, Lin Y, Zhang T. Active Prompting with Chain-of-Thought for Large Language Models; 2023.

[14] Lu Y, Bartolo M, Moore A, Riedel S, Stenetorp P. Fantastically Ordered Prompts and Where to Find Them: Overcoming Few-Shot Prompt Order Sensitivity. In: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics; 2022. p. 8086-98.

[15] Ye S, Hwang H, Yang S, Yun H, Kim Y, Seo M. In-Context Instruction Learning; 2023.

[16] Wei J, Wang X, Schuurmans D, Bosma M, Ichter B, Xia F, et al. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. In: Advances in Neural Information Processing Systems; 2022. p. 24824-37.

[17] Mialon G, Dessi R, Lomeli M, Nalmpantis C, Pasunuru R, Raileanu R, et al. Augmented Language Models: a Survey. Transactions on Machine Learning Research. 2023. Survey Certification.

[18] Kojima T, Gu SS, Reid M, Matsuo Y, Iwasawa Y. Large Language Models are Zero-Shot Reasoners. In: Advances in NIPS; 2022. p. 22199-213.

[19] Ye X, Durrett G. The Unreliability of Explanations in Few-shot Prompting for Textual Reasoning. In: Advances in Neural Information Processing Systems; 2022. p. 30378-92.

[20] Fu Y, Peng H, Sabharwal A, Clark P, Khot T. Complexity-Based Prompting for Multi-step Reasoning. In: The Eleventh International Conference on Learning Representations; 2023. .

[21] Savelka J. 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. p. 447-51.

[22] Yu F, Quartey L, Schilder F. Exploring the Effectiveness of Prompt Engineering for Legal Reasoning Tasks. In: Findings of the Association for Computational Linguistics: ACL 2023; 2023. p. 13582-96.

[23] Brack A, Hoppe A, Buschermöhle P, Ewerth R. Cross-domain multi-task learning for sequential sentence classification in research papers. In: Proceedings of the 22nd ACM/IEEE Joint Conference on Digital Libraries; 2022. p. 1-13.

[24] Huo J, Zhang K, Liu Z, Lin X, Xu W, Zheng M, et al. AntContentTech at SemEval-2023 Task 6: Domain-adaptive Pretraining and Auxiliary-task Learning for Understanding Indian Legal Texts. In: Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023). ACL; 2023. p. 402-8.

[25] Malik V, Sanjay R, Guha SK, Hazarika A, Nigam S, Bhattacharya A, et al. Semantic Segmentation of Legal Documents via Rhetorical Roles. In: Proceedings of the Natural Legal Language Processing Workshop 2022. ACL; 2022. p. 153-71.