

Can GPT-4 Support Analysis of Textual Data in Tasks Requiring Highly Specialized Domain Expertise?

Jaromir Savelka^{1,*}, Kevin D. Ashley², Morgan A. Gray², Hannes Westermann³ and Huihui Xu²

¹Computer Science Department, Carnegie Mellon University, Pittsburgh, PA, USA

²Intelligent Systems Program, University of Pittsburgh, PA, USA

³Cyberjustice Laboratory, Faculté de droit, Université de Montréal, Montréal, Canada

Abstract

We evaluated the capability of generative pre-trained transformers (GPT-4) in analysis of textual data in tasks that require highly specialized domain expertise. Specifically, we focused on the task of analyzing court opinions to interpret legal concepts. We found that GPT-4, prompted with annotation guidelines, performs on par with well-trained law student annotators. We observed that, with a relatively minor decrease in performance, GPT-4 can perform batch predictions leading to significant cost reductions. However, employing chain-of-thought prompting did not lead to noticeably improved performance on this task. Further, we demonstrated how to analyze GPT-4's predictions to identify and mitigate deficiencies in annotation guidelines, and subsequently improve the performance of the model. Finally, we observed that the model is quite brittle, as small formatting related changes in the prompt had a high impact on the predictions. These findings can be leveraged by researchers and practitioners who engage in semantic/pragmatic annotations of texts in the context of the tasks requiring highly specialized domain expertise.

Keywords

GPT-4, legal analysis, court opinions, annotation guidelines, chain-of-thought prompting, batch predictions, model brittleness, semantic annotation, generative pre-trained transformers

1. Introduction

This paper assesses the capability of generative pre-trained transformers (GPT), specifically OpenAI's GPT-4, to automatically perform semantic analysis of sentences extracted from court opinions [1] to support interpretation of legal concepts as used in statutory law. The multi-label sentence classification task requires highly specialized legal domain expertise. We use selected parts of an existing manually labeled data set¹ to assess the effectiveness of GPT-4, comparing it to the performance of human annotators. Further, we explore the implications of processing the data in batches as a cost effective alternative to analyzing one data point at a time. We also report the results of our prompt engineering efforts aimed at improving the effectiveness of the system on the task. These include general techniques, such as chain of thought prompting (CoT) [2], as well as task specific

tweaking of annotation guidelines. Finally, we assess GPT-4's predictions in terms of their robustness.

Early systematic efforts of applying empirical methods of computational linguistics to semantic, discourse-related, and/or pragmatic aspects of textual data date back to the mid 1990s [3]. Such efforts require annotated resources which have traditionally relied on subjective human judgement. In the legal domain, the approach has been embraced in many practical workflows in eDiscovery or contract review as well as in the research field of empirical legal analysis. Machine learning (ML) methods enabled approaches where humans needed to annotate only a part of the corpus. The remainder is analyzed automatically via a ML system trained on the manually annotated portion of the data.

Recently, a new paradigm has emerged where a large language model (LLM) is employed in zero/few-shot settings, using carefully crafted natural language prompts (akin to human readable instructions) [4]. This paradigm could be valuable because it may enable generation of high quality annotations with less demand for human annotators—an expensive resource, especially in tasks that require highly specialized domain expertise. Such expertise is often required in analysis of legal documents such as court opinions or statutory provisions. The cost of human labor required to annotate large legal data sets has been an important bottleneck in carrying out certain types of research in the field of AI & Law.

To investigate the capability of GPT-4 to analyze court opinions in the context of the task focused on interpre-

Proceedings of the Sixth Workshop on Automated Semantic Analysis of Information in Legal Text (ASAIL 2023), June 23, 2023, Braga, Portugal.

*Corresponding author.

✉ jsavelka@cs.cmu.edu (J. Savelka); ashley@pitt.edu

(K. D. Ashley); mag454@pitt.edu (M. A. Gray);

hannes.westermann@umontreal.ca (H. Westermann);

huihui.xu@pitt.edu (H. Xu)

🌐 <https://www.cs.cmu.edu/~jsavelka/> (J. Savelka)

📞 0000-0002-3674-5456 (J. Savelka); 0000-0002-3800-2103

(M. A. Gray); 0000-0002-4527-7316 (H. Westermann)

© 2023 Copyright for this paper by its authors. Use permitted under Creative Commons License

Attribution 4.0 International (CC BY 4.0).

CEUR Workshop Proceedings (CEUR-WS.org)

¹Statutory Interpretation Data Set. Available at: https://github.com/jsavelka/statutory_interpretation [Accessed 2023-05-01]

tation of legal concepts from statutory law, we analyzed the following research questions:

- (RQ1) How successfully can GPT-4 perform the task as compared to human annotators?
- (RQ2) Can GPT-4 perform the task as batch prediction, i.e., analyzing multiple data points at the same time?
- (RQ3) Does the accuracy of GPT-4’s predictions improve when the model is forced to provide explanations (akin to CoT)?
- (RQ4) What are the effects of modifying the annotation guidelines based on the identified shortcomings?
- (RQ5) How robust (i.e., stable) are the predictions of GPT-4 against changes of the prompt that are not related to the task definition?

By carrying out this work, we provide the following contributions to the AI & Law research community. As far as we know, this is the first study that, in the context of a task requiring highly specialized legal expertise:

- (C1) Benchmarks the performance of human annotators to the performance of GPT-4 prompted with an (almost) exact copy of annotation guidelines.
- (C2) Compares the performance of GPT-4 on batch prediction to the performance of analyzing a single data point at a time.
- (C3) Reports and discusses results of diverse prompt engineering efforts aimed at improving task specific performance of GPT-4.
- (C4) Analyzes the robustness of GPT-4’s predictions.

2. Related Work

LLMs have shown promising results in various text analysis tasks. Wang et al. [5] and Ding et al. [6] explored the use of GPT-3 for data labeling in tasks such as text entailment, sentiment analysis, topic classification, summarization, question generation, or named entity recognition. Multiple studies demonstrated that ChatGPT outperforms crowd-workers in text annotation tasks [7, 8]. At the same time, researchers caution about issues with reliability of ChatGPT in such tasks [9]. There are several studies employing various GPT models to analyze texts within tasks that require specialized domain expertise. For example, Kuzman et al. examined ChatGPT on the task of automatic genre identification [10]. Huang et al. investigated the strengths and limitations of ChatGPT in annotating implicit hate speech [11]. Ziems et al. discussed the potential of LLMs to transform computational social science and the role they could play in social science analysis [12]. Zhu et al. explored ChatGPT’s capabilities in reproducing human-generated label annotations in social computing tasks [13]. Our study explores

the efficacy of GPT-4 for analysis of texts of court opinions in the context of the task focused on interpretation of legal concepts from statutory law.

This work explores the use of GPT-4 to support semantic analysis of legal texts. There has been a growing interest in exploring capabilities of GPT models in such applications. Yu et al. applied GPT-3 to the COLIEE legal entailment task that is based on the Japanese Bar exam, substantially improving over the state-of-the-art result [14]. Similarly, Bommarito and Katz utilized GPT-3.5 for the Multistate Bar Examination [15]. Later, Katz et al. applied GPT-4 to the entire Uniform Bar Examination (UBE) and observed the system passing the exam [16]. Other use cases involve assessment of trademark distinctiveness [17], legal reasoning [18, 19], including statutory interpretation [20], U.S. Supreme court judgment modeling [21], providing legal information [22], annotation of legal documents [23], and online dispute resolution [24].

A steady line of work in AI & Law focuses on making the text analysis effort (i.e., annotation) more effective. Westermann et al. proposed and assessed a method for building strong, explainable classifiers in the form of Boolean search rules [25], as well as a method based on sentence semantic similarity [26]. Savelka and Ashley evaluated the effectiveness of an approach where a user labels the documents by confirming (or correcting) the prediction of a ML algorithm [27]. The application of active learning has been explored in the context of classification of statutory provisions [28] and eDiscovery [29, 30]. Hogan et al. proposed and evaluated a human-aided computer cognition framework for eDiscovery [31]. In this study, we evaluate the zero-shot capabilities of GPT-4 to support the analysis.

3. Data

To investigate the research questions listed above, we use a subset of the data set released in [32] focused on interpretation of legal concepts from statutory provisions. Statutory and regulatory provisions are difficult to understand because the rules they express must account for diverse situations, even those not yet encountered. When the application of a general rule is not straightforward a lawyer must present arguments as to why a provision should be applied in a particular way. In doing so the lawyer must often defend a specific account of the meaning of one or more terms (i.e. “phrase of interest”). A thorough analysis of the past treatment of the phrase of interest is foundational to formation of an adequate argument. The treatment consists of past mentions and uses of the phrase in sentences from documents such as court decisions, legislative histories, or journal articles.

The ability to sift through large amounts of legal documents and distill the content, that could be subsequently

29 U.S. CODE §203. DEFINITIONS 1

(1) "Enterprise" means the related activities performed (either through unified operation or common control) by any person or persons for a **common business purpose**. 2 includes all such activities whether performed in one or more establishments or by one or more corporate or other organizational units including departments of an establishment operated through leasing arrangements, but shall not include the related activities performed for such enterprise by an independent contractor. Within the meaning of this subsection, a retail or service establishment which is under independent ownership shall not be deemed to be so operated or controlled as to be other than a separate and distinct enterprise by reason of any arrangement, which includes, but is not necessarily limited to, an agreement, (A) that it will sell, or sell only, certain goods specified by a particular manufacturer, distributor, or advertiser, or (B) that it will join with other such establishments in the same industry for the purpose of collective purchasing, or (C) that it will have the exclusive right to sell the goods or use the brand name of a manufacturer, distributor, or advertiser within a specified area, or by reason of the fact that it occupies premises leased to it by a person who also leases premises to other retail or service establishments.

SHOWING RESULTS FOR **COMMON BUSINESS PURPOSE**.

3 The **common business purpose** of this enterprise was framing construction in the construction of single and multi-family homes. [Ann McLAUGHLIN, Plaintiff, v. STINECO, INC., et al., Defendants →](#) 4

The Fifth Circuit has held that the profit motive is a **common business purpose** if shared. [George P. SHULTZ, Plaintiff, v. William P. MORRIS, et al., Defendants →](#)

Appellants **common "business purpose"** is the operation of an institution primarily engaged in the care of the sick or aged. [Elizabeth H. DOLE, Plaintiff-Appellee, v. ODD FELLOWS HEB, Defendants-Appellants →](#)

The "**common business purpose**" requirement is not defined in the Act. [Peter J. BRENNAN, Plaintiff-Appellee, v. VETERANS CS, Defendants-Appellants →](#)

The utilization of a common service does not by itself establish a **common business purpose** shared by the owners of separate businesses. [James D. HODGSON, Appellant v. ARNHEIM AND NEELY, INC., Intervenor →](#)

Figure 1: A mock-up interface with an example statutory provision on the left (1). The user indicated that they are interested in the meaning of the "common business purpose" phrase as used in the provision (2). The system responds with a list of sentences that are deemed useful for explaining the meaning of the phrase (e.g., 3). The user may follow the link to the full-text of an opinion to view the sentence in its original context (4).

used in argumentation about the meaning of a phrase, is an important part of any lawyer's skill set. To understand the value of a sentence that uses the phrase of interest one may need to answer questions such as:

- Does a sentence provide additional information to what is already known from the statutory provision?
- Does the sentence content provide solid grounds for understanding some useful facets of the meaning of the phrase of interest?
- Is the meaning of the phrase used in the sentence the same as the meaning of the phrase of interest?

Given a text of a single statutory provision (i.e., the source provision) and the phrase of interest (i.e., one or more words in whose meaning we are interested), the task is to evaluate sentences' as to their *explanatory value* [33]. The sentences come from case decisions responsive to a query in the form of the phrase of interest (e.g. "common business purpose"). A sentence should be labeled with one of the following categories [34]:

- **High value:** This label is reserved for sentences that explicitly elaborate on the meaning of the phrase of interest.
- **Certain value:** The system should select this label if the sentence does not explicitly elaborate on the meaning of the phrase of interest, yet

the sentence still provides grounds to draw some (even modest or quite vague) conclusions about the meaning of the phrase of interest.

- **Potential value:** This label is appropriate if the sentence does not appear to be useful for elaboration on the meaning of the phrase of interest but the sentence provides some additional information (even quite marginal) over what is known from the source provision.
- **No value:** This label should be selected if the sentence does not provide any additional useful information over what is already known from the source provision.

This type of text analysis may enable training of ML models supporting, e.g., a legal information retrieval system focused on legal concepts interpretation such as the one shown in Figure 3 [35, 36, 37].

The original data set was annotated by domain experts—11 law students and 2 legal scholars with law degrees. The law students performed the first pass of the annotations and the scholars were responsible for the second pass resulting in the consensus labels. The agreement between the students' annotations and the consensus labels, measured in terms of Krippendorff's α [38], was $0.1 < \alpha < 0.6$ (see Figure 8) while the inter-annotator agreement between the two scholars was $\alpha = 0.79$ [39].

Table 1

Data set descriptive statistics showing the distribution of sentence labels per phrase of interest (the first column). NV – No value, PV – Potential value, CV – Certain value, HV – High value.

Phrase of interest	NV	PV	CV	HV	Total
Accommodation trade	4	48	10	7	69
Cybercrime sentence	4	54	11	2	71
Digital musical recording	6	13	11	13	43
Semiconductor chip product	2	9	12	2	25
Unduly disrupt the operations	7	36	2	3	48
Total	23	160	46	27	256

Hence, clearly this is a very demanding text analysis task requiring highly specialized domain expertise.

The original data set consists of 42 queries (i.e., phrases of interest) associated with 26,959 labeled sentences from 20 different areas of legal regulation (e.g., intellectual property, criminal law). Considering the non-negligible cost of large numbers of requests to the GPT-4 API, we decided to work with a small subset of the original data set. We selected 5 phrases of interest associated with 256 sentences. While limited, the sample of this size is sufficient to support the experiments in this work. The distribution of labels within the data set is reported in Table 1.

4. Model

In our experiments, we use the GPT-4 model. As of the writing of this paper, GPT-4 is by far the most advanced model released by OpenAI. The model is focused on dialog between a user and a system (i.e., an assistant). The original GPT model [40] is a 12-layer decoder-only transformer [41] with masked self-attention heads. Its core capability is fine-tuning on a downstream task. The GPT-2 model [42] largely follows the details of the original GPT with a few modifications, such as layer normalization moved to the input of each sub-block, additional layer-normalization after the first self-attention block, and a modified initialization. Compared to the original model it displays remarkable multi-task learning capabilities [42]. The third generation of GPT models [43] uses almost the same architecture as GPT-2. The only difference is that it alternates dense and locally banded sparse attention patterns in the layers of the transformer. The main focus of [43] was to study the dependence of performance and model size where eight differently sized models were trained (from 125 million to 175 billion parameters). The largest of these models is commonly referred to as GPT-3. The interesting property of these models is that they appear to be very strong zero- and few-shot learners. This ability appears to improve with

the increasing size of the model [43]. The technical details of the recently released GPT-4 model have not been disclosed due to concerns about potential misuses of the technology as well as a highly competitive market for generative AI [44].

We set the temperature of the model to 0.0, which corresponds to no randomness. The higher the temperature the more creative the output but it can also be less factual. As the temperature approaches 0.0, the model becomes deterministic and can be repetitive. We set `max_tokens` to various values depending on the expected size of the output (a token roughly corresponds to a word) as this parameter controls the maximum length of the completion (i.e., the output). For a single data point classification task where we only expect a single label as a completion the setting of 50 is sufficient. For a batch classification with a CoT prompt, a much larger size of output is expected (1,500 tokens). Note that GPT-4 has an overall token length limit of 8,192 tokens, comprising both the prompt and the completion.² We set `top_p` to 1, as is recommended when temperature is set to 0.0. This parameter is related to temperature and also influences creativity of the output. We set `frequency_penalty` to 0, which allows repetition by ensuring no penalty is applied to repetitions. Finally, we set `presence_penalty` to 0, ensuring no penalty is applied to tokens appearing multiple times in the output.

5. Experimental Design

5.1. GPT-4 Text Analysis (RQ1)

The first experiment was focused on answering RQ1, i.e., how successfully GPT-4 can perform the annotation task, as compared to human annotators. To that end we used the annotation guidelines³ originally designed for the human annotators and turned them into a system prompt for GPT-4. The system prompt is typically used to steer the system (i.e., the GPT-4 model) towards performing the desired task. We introduced only minimal changes to the annotation guidelines to ensure close mapping between the original task performed by human annotators and the task performed by GPT-4 automatically. We left out pieces of the annotation guidelines related to the specifics of the annotation environment used by humans, as these would have made no sense in the GPT-4’s prompt, e.g.:

At the top of each sheet there is a cell with a light yellow background that contains a text of a single statutory provision [...]

²There is also a variant of the model that supports up to 32,768 tokens.

³Annotation Guidelines for Evaluating Sentences for Argumentation about the Meaning of Statutory and Regulatory Terms. Available at: https://github.com/jsavelka/statutory_interpretation/blob/master/annotation_guidelines_v2.pdf [Accessed 2023-04-30]


```

You are a specialized system focused on semantic annotation
of court opinions. ①

BACKGROUND ②
Statutory and regulatory provisions are difficult to
[3,300 characters ...]

ANNOTATION TASK ③
The system is provided with a text of a single statutory
[1,508 characters ...]

RULES FOR SENTENCE EVALUATION ④
The system should evaluate the sentence using the procedure
[5,648 characters ...]

```

Figure 2: The system Prompt is populated with annotation guidelines as shown above. The typical preamble (1) is followed by the Background section (2) describing the context of the text analysis task to be performed. The Annotation Task section (3) provides more specific information about the mechanics of the task. Finally, the Rules for Sentence Evaluation section (4) contains the fine-grained instructions on how to categorize retrieved sentences. The grey tokens inform about the size of the parts of the prompt not shown in the figure. The prompt is a sizeable text spanning multiple pages.

Furthermore, we replaced references to “students” with a reference to a “system”. The guidelines contained a visual diagram, encoding the workflow of annotation rules which we translated into a list of questions. Finally, we omitted several examples in order to fit the annotation guidelines within the prompt and leave sufficient space for the output. The overall structure of the system prompt (i.e., the annotation guidelines) is shown in Figure 2. Note that this sizeable piece of text is much longer than what is typically used as a system prompt with GPT-4.

Each data point was provided to the system as a message coming from a user. The message contained the phrase of interest, citation to the source provision, the text of the source provision, as well as a retrieved sentence that should have been labeled with one of the categories described in Section 3. The exact layout and formatting of the message is provided in Figure 3. GPT-4 was expected to return a message (coming from an assistant) containing the predicted label. In this experiment we set the `max_tokens` parameter to 50 as this was sufficient for this type of completion.

We inserted each data point from the data set into the template from Figure 3 and submitted it individually to OpenAI’s GPT-4 API, together with the system prompt. Note that this approach, despite the limited size of the data set of 256 samples, incurred a non-negligible cost exceeding \$20. The cost was, of course, lower than the cost of equivalent human labor on the same task. We extracted the predicted labels from the GPT-4 responses and compared them to the gold labels (Section 6).

```

PHRASE OF INTEREST: {{phrase_of_interest}}

SOURCE PROVISION:
{{source_provision_citation}}
{{source_provision_text}}

SENTENCE:
{{sentence}}

EXPECTED OUTPUT FORMAT: #
Label: <label>

```

Figure 3: User message template for a single sentence prediction. The tokens shown in blue and surrounded by double curly braces are replaced with the corresponding data elements. Hence, the message is typically a somewhat longer text. Note the Expected Output Format section (#) instructing the model as to the expected format of the response.

5.2. Batch Prediction (RQ2)

The next experiment was focused on answering RQ2, that is, whether GPT-4 can perform the task as batch prediction. To this end we used the same system prompt as in the preceding experiment (Figure 2). We modified the user message as shown in Figure 4. Instead of a single data point (i.e., sentence), we inserted multiple sentences. Correspondingly, the expected output part of the message was changed to reflect that GPT-4 should have returned more than one prediction. We constructed the batches dynamically to fit as many sentences as possible using the `tiktoken` Python library⁴ to determine the size of the prompt before sending it to the GPT-4 API. Hence, the size of each batch is determined by the length of the submitted sentences. Typically, several tens of sentences were submitted within a single batch. For this experiment, we increased the `max_tokens` parameter to 1,000 to accommodate lengthier completions. Note that this approach was significantly cheaper than the one presented earlier.

5.3. Explanations – CoT (RQ3)

To explore RQ3, i.e., the effects of requiring the model to explain its predictions, we first modified the user message submitted to GPT-4 as shown in Figure 5. This experiment was similar to the first one. The only difference was that we asked the model to first spell out an explanation regarding the predicted label, and to provide the prediction after that. This was inspired by the work on chain of thought (CoT) prompting that has been shown to improve performance of the models on diverse tasks [2], including those in the legal domain [14]. For this experiment, we set the `max_tokens` parameter to 500 to

⁴tiktoken. Available at: <https://github.com/openai/tiktoken> [Accessed: 2023-04-30]

```
[...]
SENTENCES:
Sentence 1: {{sentence_1}}
Sentence 2: {{sentence_2}}
[...]

EXPECTED OUTPUT FORMAT: #
Sentence 1: <label>
Sentence 2: <label>
Sentence 3: <label>
```

Figure 4: Excerpt from the user message template for batch prediction. The top part of the template that is omitted from the figure is the same as that shown in Figure 3. The tokens shown in blue and surrounded by double curly braces are replaced with the corresponding data elements. Note the Expected Output Format section (#) instructing GPT-4 how to output multiple labels related to the submitted sentences.

```
[...]
EXPECTED OUTPUT FORMAT: #
Explanation: <reasoning why particular label should be
              assigned>
Label: <label>
```

Figure 5: Excerpt from the user message template requiring explanation before prediction. The top part of the template that is omitted from the figure is the same as that shown in Figure 3. Note the Expected Output Format section (#) instructing GPT-4 how to output the explanation before the prediction.

accommodate the expected completions. Given the increased size of the completion, this approach was even costlier than the one presented as the first experiment.

To further explore RQ3, we modified the user message as shown in Figure 6. Here, we tested the effects of requiring explanations in the batch predictions task. Since full-blown natural language explanations, as in the preceding experiment, would have drastically decreased the size of the batch that could have been submitted to the API, we opted for schematic explanations encoding the answers of the model to the individual questions from the annotation guidelines stemming from the visual workflow (see Section 5.1). For this experiment, we increased the `max_tokens` parameter to 1,500 tokens. Note that this experiment was slightly more costly than the original batch prediction experiment. This was because GPT-4’s completions cost more than the tokens submitted to the API. However, the cost was still significantly reduced when compared to the two experiments where the data points are submitted one by one.

```
[...]
EXPECTED OUTPUT FORMAT: 1
Sentence 1: <explanation> => <label>
Sentence 2: <explanation> => <label>
Sentence 3: <explanation> => <label>

EXAMPLE: 2
Sentence 1: Q1 No => no value
Sentence 2: Q1 Yes -> Q2 No -> Q4 Yes => high value
Sentence 3: Q1 Yes -> Q2 No -> Q4 No -> Q5 No => potential
              value
...
```

Figure 6: Excerpt from the user message template requiring explanations before predictions (batch). The top part of the template that is omitted from the figure is the same as that shown in Figure 4. Note the Expected Output Format section (1) instructing GPT-4 how to output the schematic explanations before the predictions as well as the Examples section (2).

5.4. Prompt (Annotation Guidelines) Modification (RQ4)

The next experiment was focused on answering RQ4, i.e., analyzing the effects of modifying the annotation guidelines. In a typical annotation workflow where human annotators are involved, the early stages are dedicated to the training of the human annotators as well as to the refinement of annotation guidelines. Note that GPT-4 provides means for similar types of interventions. The training of the human annotators is akin to augmenting GPT-4’s prompt with labeled examples (i.e., few-shot settings) or fine-tuning the model. The refinement of annotation guidelines translates into modifications of the system prompt containing the guidelines. In this work, we focused on exploring the refinement of annotation guidelines (i.e., the prompt), leaving the exploration of few-shot learning and fine-tuning as open questions for future work.

Based on the results of the preceding four experiments, we identified a prominent weakness in the predictions of the GPT-4 model. We modified the system prompt (i.e., the annotation guidelines) with the aim of mitigating the issue. In order to answer RQ4, we analyzed the effects of the changes on the performance of the model. Specifically, we repeated all the preceding experiments with the modified prompt, and observed the changes in performance.

5.5. Robustness (RQ5)

The final experiment was focused on answering RQ5, that is, analyzing the robustness of the GPT-4 annotations. The preceding experiments yielded multiple sets of labels over the same data points. Each version of the

Table 2

Experimental Results. The Instructions column encodes if the original or updated annotation guidelines were used in GPT-4’s system prompt. The Annotation Modality column describes the experimental setting. The remaining columns report the performance metrics computed against the gold labels.

Instructions	Annotation Modality	Precision	Recall	F1-score	Accuracy	α
Original	Single – Labels Only (RQ1)	.63	.46	.53	.46	.51
	Batch – Labels Only (RQ2)	.61	.45	.52	.45	.42
	Single – Labels & Explanation (RQ3)	.69	.40	.51	.40	.44
	Batch – Labels & Explanation (RQ3)	.52	.29	.37	.29	.19
Updated	Single – Labels Only (RQ4)	.60	.55	.57	.55	.53
	Batch – Labels Only (RQ4)	.57	.46	.51	.46	.42
	Single – Labels & Explanation (RQ4)	.58	.57	.57	.57	.48
	Batch – Labels & Explanation (RQ4)	.48	.46	.47	.46	.27

annotation guidelines, that is, the original system prompt and the updated one, was associated with four labels for each data point—two from the single sentences experiments (labels only and labels with explanations), and two from the batch predictions. While these experiments differed in the form of how the model was prompted (i.e., with one or multiple sentences, and with or without an explanation), the annotation instructions remained the same. Therefore, this experiment explored how the form of the prompting affects the results. Specifically, we were interested in assessing stability of predictions across the four labels produced within different experiments relying on the same annotation guidelines.

6. Results and Discussion

6.1. GPT-4 Text Analysis (RQ1)

The results of the experiment focused on GPT-4’s performance on the text analysis task as compared to the human annotators (RQ1) are reported in Table 2 under the Original instructions and Single – Labels Only entry. The overall $F_1 = .53$ suggests that GPT-4 is able to successfully analyze the texts while at the same time leaving ample room for improvement. Additional insight is provided by the confusion matrix in the upper left corner of Figure 7. There, we can see that the system struggled with the Potential value label where many instances of this class were either predicted as No value or Certain value.

It is important to recall that the task is very challenging even for human annotators and requires highly specialized domain expertise. Hence, we are interested in how the performance of GPT-4 compares to that of the human annotators. Figure 8 benchmarks the agreement, in terms of Krippendorff’s α , of GPT-4 with the consensus labels to the agreement of the law students’ labels with the consensus. In Figure 8, we can clearly recognize two groups of annotators, i.e., those whose agreements are $> .5$ and

those whose agreements are $< .4$. This significant gap quite likely distinguishes between well-performing and less well-performing human annotators. GPT-4’s performance is on par with the well-performing law student annotators.

6.2. Batch Prediction (RQ2)

The results of the experiment focused on GPT-4’s performance on batch prediction (RQ2) are also reported in Table 2 under the Original instructions and Batch – Labels Only entry. The overall $F_1 = .52$ is a slight decrease in performance as compared to the prediction performed on one data point at a time. The significantly lower cost of this approach may justify the difference in performance. However, while the overall performance remained similar, the performance on the individual labels changed to a larger extent, as can be seen in the corresponding confusion matrix shown in Figure 7 (first row, second from the left). While the performance on the sentences with the Potential label is improved, the model performed less well on the sentences from the other three classes.

6.3. Explanations – CoT (RQ3)

The results of the experiment focused on GPT-4’s performance when providing explanations in addition to the predictions (RQ3) are reported in Table 2 under the Original instructions and Single – Labels & Explanation entry. Interestingly, we observe a decrease in performance as compared to the single sentence prediction experiment. The overall F_1 went from 0.53 to 0.51 and accuracy from 0.46 to 0.40. Further insight is provided by the confusion matrix in Figure 7 (first row, second from the right). Apparently, the issue of predicting Potential value sentences as Certain value is even more pronounced than before. This strongly suggests that GPT-4 struggles with correctly interpreting the annotation guidelines when it comes to distinguishing between the two classes. Note

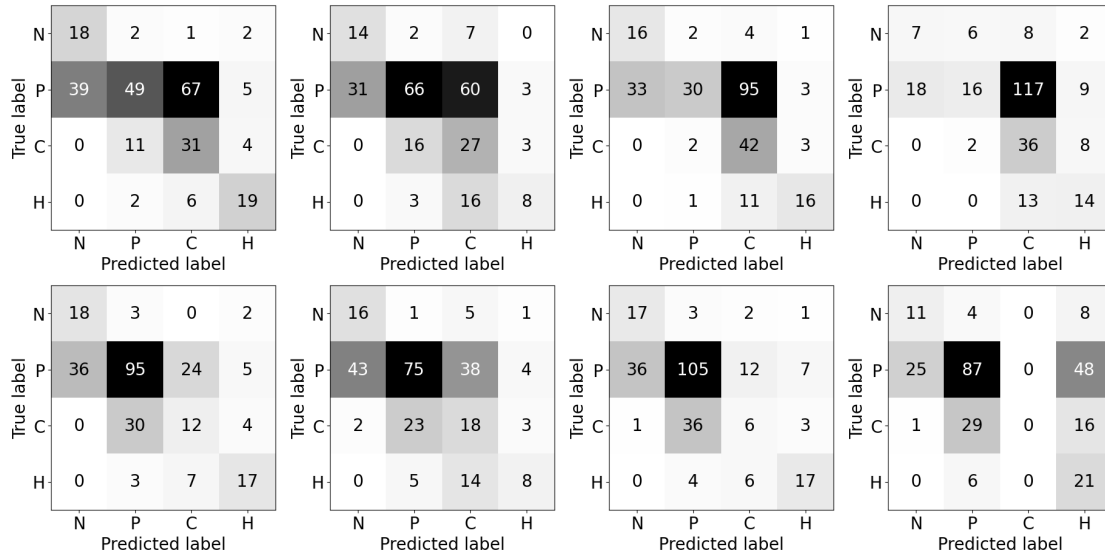


Figure 7: Confusion matrices for the original annotation guidelines (top row) and updated annotation guidelines (bottom row). From left to right the matrices describe the following experimental conditions: Single – Labels Only, Batch – Labels Only, Single – Labels & Explanation, Batch – Labels & Explanation. The labels: N – No value, P – Potential value, C – Certain value, H – High value.

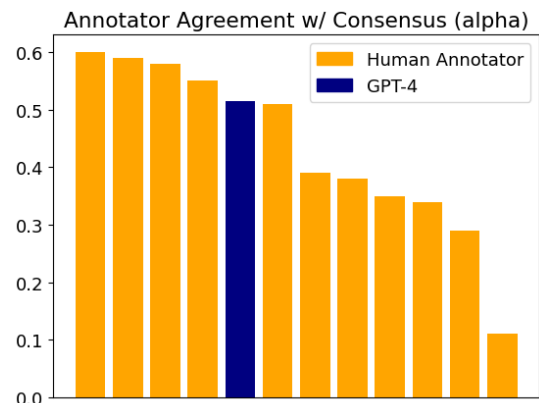


Figure 8: The annotator agreement (Krippendorff’s α) of the manually created annotations and GPT-4 predictions computed against the consensus (gold) labels. GPT-4 performs comparably to human annotators (law students).

that this is contrary to the expectations of improving the performance by having GPT-4 explain its predictions since this is akin to CoT prompting which often leads to improvements in performance on a task.

The provided explanations often appear to be in agreement with the predicted labels but this is not always the case. For example, the following explanation is provided for a sentence that is correctly predicted as having No

value:

The sentence is a verbatim citation of the source provision and does not provide any additional information about the meaning of the phrase “digital musical recording.”

The following explanation is attached to a sentence that is wrongly predicted as Certain value:

The sentence provides an explanation of what would not qualify under the basic definition of a digital musical recording, which is useful for understanding the boundaries of the phrase of interest.

The sentence should have been predicted as High value and the explanation is aligned with such a prediction. Interestingly, it did not help to steer the prediction towards assigning the High value label.

The results of the batch experiment focused on further explorations of RQ3 are reported in Table 2 under the Original instructions and Batch – Labels & Explanation entry. We observe a complete degradation of the performance under this condition. As apparent from the corresponding confusion matrix in the upper right corner of Figure 7, a large portion of the sentences were mislabeled as having Certain value. This suggests that the definition of the Certain value class may be too broad. Interestingly, the schematic explanations are generally

in agreement with the predicted labels irrespective of the prediction being correct or not. Below are example predictions with explanations from the batch experiment:

Q1 Yes -> Q2 No -> Q4 Yes => High value
Q1 Yes -> Q2 No -> Q4 No -> Q5 Yes => Certain value
Q1 Yes -> Q2 No -> Q4 No -> Q5 No => Potential value
Q1 No => No value

Recall that the Q# refer to the questions from the annotation guidelines an annotator is supposed to consider in order to correctly label a sentence.

6.4. Prompt (Annotation Guidelines) Modification (RQ4)

The preceding experiments identified a potential issue with the definition of the Certain value class: it may be too broad. Hence, we use this particular issue as the test bed for investigating RQ4. Specifically, we modify the guidelines with the aim to mitigate the issue, i.e., improve the performance of the GPT-4 model on the task. The annotation guidelines contain the following definition of the Certain value class:

The system should select this label if the sentence does not explicitly elaborate on the meaning of the phrase of interest, yet the sentence still provides grounds to draw some (even modest or quite vague) conclusions about the meaning of the phrase of interest.

Furthermore, the guidelines direct an annotator to consider the below question after ruling out the High value and No value labels:

Does the sentence provide useful context with respect to the elaboration of the meaning of the phrase of interest?

A positive answer to that question should result in annotating the respective sentence with the Certain value label. A negative answer directs the annotator to assign the Positive value label. Indeed, the experiments focused on explanations clearly show that the system often tends to answer the question in positive. Consider the following example of an explanation in natural language:

The sentence [...] does not explicitly elaborate on the meaning of the phrase “cybercrime” [...] However, it provides useful context by mentioning a convention that deals with cybercrime [...]

Similarly, the following chain of reasoning is predominantly used in the batch prediction with explanation experiment (see Figure 6 to understand the format of the below):

Q1 Yes -> Q2 No -> Q4 No -> Q5 Yes

Question 5 (Q5) is the one that directs an annotator to assign the sentence the Certain value label in case it is answered in positive.

Based on the above analysis, our aim is to modify the annotation guidelines to make the system less likely to annotate a sentence as Certain value and opt for a different label. To achieve this goal, we replaced the above definition of the Certain value class with a more restrictive one:

The system should select this label if the sentence elaborates on the meaning of the phrase of interest implicitly.

The definition follows up on the definition of the High value class where an *explicit* elaboration is required.

The results of the experiment focused on the effects of modifying the prompt (RQ4) are reported in Table 2 under the Updated instructions section. The overall $F_1 = .57$ for the Single – Labels Only condition is a noticeable improvement over the $F_1 = .53$ performance with the original guidelines. The corresponding confusion matrix shown in the bottom left of Figure 7 reveals that the issue of over-predicting the Certain class at the expense of the Potential value class has been addressed effectively. On the other hand, it appears that the system now errs on the other side, being reluctant to label a sentence as having Certain value. Nevertheless, the overall performance of the system appears to be improved.

Furthermore, application of the CoT prompting, i.e., asking the model to provide explanations alongside the predictions, no longer leads to dramatic deterioration of performance with the updated annotation guidelines. While we can still observe a slight decrease in performance of the CoT prompt for the batch prediction, it is quite small compared to the decrease observed with the original annotation guidelines.

6.5. Robustness (RQ5)

The results of the experiment focused on the robustness of GPT-4’s predictions (RQ5) are reported in Table 3. The table shows inter-annotator agreement (Krippendorff’s α) among the predictions from the earlier experiments. Interestingly, the agreement appears to be relatively low considering the fact that we are comparing systems based on the identical annotation guidelines. While further investigation is needed, it appears that small changes in the expected format of the output can dramatically affect the predictions.

Table 3

The inter-annotator agreement (Krippendorff’s α) between the predictions from the experiments (RQ5): S-LO: Single – Labels Only, S-LE: Single – Labels & Explanation, B-LO: Batch – Labels Only, B-LE: Batch – Labels & Explanation

	Original				Updated			
	S-LO	S-LE	B-LO	B-LE	S-LO	S-LE	B-LO	B-LE
S-LO	1.0	.78	.58	.36	1.0	.83	.55	.37
S-LE		1.0	.48	.44		1.0	.50	.27
B-LO			1.0	.44			1.0	.58
B-LE				1.0				1.0

7. Limitations

In this study, we focused on a single specific task requiring highly specialized domain expertise, which may limit the generalizability of our findings. The task was selected based on the assumption that it represents the complex nature of tasks that may arise in specialized domains. However, it is possible that the performance of GPT-4 in other tasks requiring domain expertise might differ significantly. Moreover, the relatively small data set used in our analysis might not capture the full range of complexities and nuances associated with tasks requiring specialized knowledge. Consequently, the results obtained in this study should be interpreted with caution and not generalized to all tasks requiring domain expertise.

Another limitation concerns the general issues of reproducible experiments with proprietary OpenAI’s GPT models. As access to these models is limited and often subject to certain terms and conditions, it can be challenging for independent researchers to replicate the experiments and validate the findings. This raises concerns about the reproducibility and robustness of the results, which are essential aspects of scientific research. Furthermore, any changes or updates to the GPT models by OpenAI might affect the performance and outcomes of experiments, making it difficult to establish a consistent baseline for comparison across studies. Therefore, it is crucial to address these concerns and develop strategies to promote reproducibility and robustness in future studies involving GPT models.

8. Conclusions and Future Work

This study assessed the capabilities of GPT-4 in analyzing textual data in the context of a task focused on interpretation of legal concepts. Our findings indicate that GPT-4 can perform at a level comparable to well-trained law student annotators. The fact that the model is able to take a multi-page document, understand the instructions contained therein, and apply these instructions to com-

plex real-world textual data demonstrates the impressive performance of GPT-4. Further, this could have a significant impact on research in domains where complex annotation tasks are performed, such as the legal domain. Being able to utilize GPT-4, instead of hiring and training human annotators over extended periods of time could enable many types of research efforts, and open the door to novel large-scale research or data science projects.

We demonstrated that GPT-4 can be effectively utilized for batch predictions, offering significant cost reductions without a major decline in performance. On the other hand, CoT prompting did not yield a noticeable improvement in performance. We showcased an example of analyzing GPT-4’s predictions to identify and address deficiencies in annotation guidelines, leading to improvements in the model’s performance. However, the study also highlighted the model’s brittleness, as minor formatting changes in the prompt had a substantial impact on the predictions. Researchers and practitioners can leverage these findings to effectively employ GPT-4 in semantic and pragmatic annotation tasks within specialized domains, while being mindful of the limitations.

Future work should focus on evaluation of GPT-4’s capabilities across a broader range of tasks and domains, involving larger data sets, that require highly specialized expertise. Additionally, exploring methods to improve the model’s robustness and resilience to minor formatting changes in the prompts would be valuable, ensuring more consistent and reliable performance. Furthermore, investigating alternative prompting techniques or fine-tuning strategies could potentially lead to enhanced performance in specialized tasks.

Acknowledgments

This work was supported in part by a National Institute of Justice Graduate Student Fellowship (Fellow: Jaromir Savelka) Award # 2016-R2-CX-0010, “Recommendation System for Statutory Interpretation in Cybercrime,” a University of Pittsburgh Pitt Cyber Accelerator Grant entitled “Annotating Machine Learning Data for Interpreting Cyber-Crime Statutes,” and the National Science Foundation, grant no. 2040490, FAI: Using AI to Increase Fairness by Improving Access to Justice.

References

- [1] J. Savelka, V. R. Walker, M. Grabmair, K. D. Ashley, Sentence boundary detection in adjudicatory decisions in the united states, *Traitement automatique des langues* 58 (2017) 21.
- [2] J. Wei, X. Wang, D. Schuurmans, M. Bosma, E. Chi, Q. Le, D. Zhou, Chain of thought prompting elicits

- reasoning in large language models, arXiv preprint arXiv:2201.11903 (2022).
- [3] R. Artstein, M. Poesio, Inter-coder agreement for computational linguistics, *Computational linguistics* 34 (2008) 555–596.
- [4] P. Liu, W. Yuan, J. Fu, Z. Jiang, H. Hayashi, G. Neubig, Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing, *ACM Computing Surveys* 55 (2023) 1–35.
- [5] S. Wang, Y. Liu, Y. Xu, C. Zhu, M. Zeng, Want To Reduce Labeling Cost? GPT-3 Can Help, 2021. URL: <http://arxiv.org/abs/2108.13487>, arXiv:2108.13487 [cs].
- [6] B. Ding, C. Qin, L. Liu, L. Bing, S. Joty, B. Li, Is GPT-3 a Good Data Annotator?, 2022. URL: <http://arxiv.org/abs/2212.10450>, arXiv:2212.10450 [cs].
- [7] F. Gilardi, M. Alizadeh, M. Kubli, ChatGPT Outperforms Crowd-Workers for Text-Annotation Tasks, 2023. URL: <http://arxiv.org/abs/2303.15056>, arXiv:2303.15056 [cs].
- [8] P. Törnberg, ChatGPT-4 Outperforms Experts and Crowd Workers in Annotating Political Twitter Messages with Zero-Shot Learning, 2023. URL: <http://arxiv.org/abs/2304.06588>, arXiv:2304.06588 [cs].
- [9] M. V. Reiss, Testing the Reliability of ChatGPT for Text Annotation and Classification: A Cautionary Remark, 2023. URL: <http://arxiv.org/abs/2304.11085>, arXiv:2304.11085 [cs].
- [10] T. Kuzman, I. Mozetič, N. Ljubešić, ChatGPT: Beginning of an End of Manual Linguistic Data Annotation? Use Case of Automatic Genre Identification, 2023. URL: <http://arxiv.org/abs/2303.03953>, arXiv:2303.03953 [cs].
- [11] F. Huang, H. Kwak, J. An, Is ChatGPT better than Human Annotators? Potential and Limitations of ChatGPT in Explaining Implicit Hate Speech, in: *Companion Proceedings of the ACM Web Conference 2023*, 2023, pp. 294–297. URL: <http://arxiv.org/abs/2302.07736>. doi:10.1145/3543873.3587368, arXiv:2302.07736 [cs].
- [12] C. Ziems, W. Held, O. Shaikh, J. Chen, Z. Zhang, D. Yang, Can large language models transform computational social science?, 2023. arXiv:2305.03514.
- [13] Y. Zhu, P. Zhang, E.-U. Haq, P. Hui, G. Tyson, Can ChatGPT Reproduce Human-Generated Labels? A Study of Social Computing Tasks, 2023. URL: <http://arxiv.org/abs/2304.10145>, arXiv:2304.10145 [cs].
- [14] F. Yu, L. Quartey, F. Schilder, Legal prompting: Teaching a language model to think like a lawyer, 2022. URL: <https://arxiv.org/abs/2212.01326>. doi:10.48550/ARXIV.2212.01326.
- [15] M. Bommarito, D. M. Katz, Gpt takes the bar exam, arXiv preprint arXiv:2212.14402 (2022).
- [16] D. M. Katz, M. J. Bommarito, S. Gao, P. Arredondo, Gpt-4 passes the bar exam, Available at SSRN 4389233 (2023).
- [17] J. Goodhue, Y. Wei, Classification of trademark distinctiveness using openai gpt 3.5 model, Available at SSRN 4351998 (2023).
- [18] A. Blair-Stanek, N. Holzenberger, B. Van Durme, Can gpt-3 perform statutory reasoning?, arXiv preprint arXiv:2302.06100 (2023).
- [19] H.-T. Nguyen, R. Goebel, F. Toni, K. Stathis, K. Satoh, How well do sota legal reasoning models support abductive reasoning?, arXiv preprint arXiv:2304.06912 (2023).
- [20] J. Savelka, K. Ashley, M. Gray, H. Westermann, H. Xu, Explaining legal concepts with augmented large language models (gpt-4), in: *AI4Legs 2023: AI for Legislation*, 2023.
- [21] S. Hamilton, Blind judgement: Agent-based supreme court modelling with gpt, arXiv preprint arXiv:2301.05327 (2023).
- [22] J. Tan, H. Westermann, K. Benyekhlef, Chatgpt as an artificial lawyer?, in: *Artificial Intelligence for Access to Justice (AI4AJ 2023)*, 2023.
- [23] J. Savelka, Unlocking practical applications in legal domain: Evaluation of gpt for zero-shot semantic annotation of legal texts, arXiv preprint arXiv:2305.04417 (2023).
- [24] H. Westermann, J. Savelka, K. Benyekhlef, Llmmediator: Gpt-4 assisted online dispute resolution, in: *Artificial Intelligence for Access to Justice (AI4AJ 2023)*, 2023.
- [25] H. Westermann, J. Savelka, V. R. Walker, K. D. Ashley, K. Benyekhlef, Computer-assisted creation of boolean search rules for text classification in the legal domain., in: *JURIX*, 2019, pp. 123–132.
- [26] H. Westermann, J. Savelka, V. R. Walker, K. D. Ashley, K. Benyekhlef, Sentence embeddings and high-speed similarity search for fast computer assisted annotation of legal documents, in: *Legal Knowledge and Information Systems: JURIX 2020: The Thirty-third Annual Conference*, Brno, Czech Republic, December 9-11, 2020, volume 334, IOS Press, 2020, p. 164.
- [27] J. Šavelka, G. Trivedi, K. D. Ashley, Applying an interactive machine learning approach to statutory analysis, in: *Legal Knowledge and Information Systems*, IOS Press, 2015, pp. 101–110.
- [28] B. Wärtl, J. Muhr, I. Glaser, G. Bonczek, E. Scepánková, F. Matthes, Classifying legal norms with active machine learning., in: *JURIX*, 2017, pp. 11–20.
- [29] G. V. Cormack, M. R. Grossman, Scalability of continuous active learning for reliable high-recall text classification, in: *Proceedings of the 25th ACM international on conference on information and knowledge management*, 2016, pp. 1039–1048.

- [30] G. V. Cormack, M. R. Grossman, Autonomy and reliability of continuous active learning for technology-assisted review, arXiv preprint arXiv:1504.06868 (2015).
- [31] C. Hogan, R. Bauer, D. Brassil, Human-aided computer cognition for e-discovery, in: Proceedings of the 12th International Conference on Artificial Intelligence and Law, 2009, pp. 194–201.
- [32] J. Šavelka, K. D. Ashley, Discovering explanatory sentences in legal case decisions using pre-trained language models, in: Findings of the Association for Computational Linguistics: EMNLP 2021, 2021, pp. 4273–4283.
- [33] J. Šavelka, K. D. Ashley, On the role of past treatment of terms from written laws in legal reasoning, *New Developments in Legal Reasoning and Logic: From Ancient Law to Modern Legal Systems* (2022) 379–395.
- [34] J. Šavelka, K. D. Ashley, Extracting case law sentences for argumentation about the meaning of statutory terms, in: Proceedings of the third workshop on argument mining (ArgMining2016), 2016, pp. 50–59.
- [35] J. Šavelka, H. Xu, K. D. Ashley, Improving sentence retrieval from case law for statutory interpretation, in: Proceedings of the seventeenth international conference on artificial intelligence and law, 2019, pp. 113–122.
- [36] J. Šavelka, K. D. Ashley, Learning to rank sentences for explaining statutory terms., in: ASAIL@ JURIX, 2020.
- [37] J. Šavelka, K. D. Ashley, Legal information retrieval for understanding statutory terms, *Artificial Intelligence and Law* (2021) 1–45.
- [38] K. Krippendorff, *Computing krippendorff’s alpha-reliability* (2011).
- [39] J. Šavelka, *Discovering sentences for argumentation about the meaning of statutory terms*, Ph.D. thesis, University of Pittsburgh, 2020.
- [40] A. Radford, K. Narasimhan, T. Salimans, I. Sutskever, *Improving language understanding by generative pre-training* (2018).
- [41] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, *Advances in neural information processing systems* 30 (2017).
- [42] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, *Language models are unsupervised multitask learners* (2019).
- [43] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al., Language models are few-shot learners, *Advances in neural information processing systems* 33 (2020) 1877–1901.
- [44] OpenAI, *Gpt-4 technical report*, 2023. arXiv:2303.08774.