Legal Knowledge and Information Systems
G. Sileno et al. (Eds.)

doi:10.3233/FAIA230961

Can GPT Alleviate the Burden of Annotation?

Morgan GRAY [a,1], Jaromir SAVELKA [c] Wesley OLIVER [d] and Kevin ASHLEY [a,b]
[a] *Intelligent Systems Program, University of Pittsburgh, USA*
[b] *School of Law, University of Pittsburgh, USA*
[c] *School of Computer Science, Carnegie Mellon University, USA*
[d] *School of Law, Duquesne University, USA*
ORCiD ID: Morgan Gray https://orcid.org/0000-0002-3800-2103, Jaromir SAVELKA https://orcid.org/0000-0002-3674-5456, Wesley OLIVER https://orcid.org/0000-0002-3873-8479, Kevin ASHLEY https://orcid.org/0000-0002-3873-8479

Abstract. Manual annotation is just as burdensome as it is necessary for some legal text analytic tasks. Given the promising performance of Generative Pretrained Transformers (GPT) on a number of different tasks in the legal domain, it is natural to ask if it can help with text annotation. Here we report a series of experiments using GPT-4 and GPT 3.5 as a pre-annotation tool to determine whether a sentence in a legal opinion describes a legal factor. These GPT models assign labels that human annotators subsequently confirm or reject. To assess the utility of pre-annotating sentences at scale, we examine the agreement among gold-standard annotations, GPT's pre-annotations, and law students' annotations. The agreements among these groups support that using GPT-4 as a pre-annotation tool is a useful starting point for large-scale annotation of factors.

Keywords. GPT-4, Annotation, Interrater Agreement, Generative LLMs

1. Introduction

State-of-the-art large language models (LLMs) have been shown to have remarkable zero-shot performance on many diverse tasks. Prior work shows that LLMs can approach or even surpass the performance of human experts on some legal tasks. In legal practice or empirical legal studies (ELS), there is often a need to classify or annotate large amounts of texts. Since this is costly in terms of experts' time and labor, it is natural to ask if and how LLMs can alleviate the costs of manual annotation. In the case of our task, identifying factors, stereotypical patterns of facts that strengthen or weaken a legal claim, we observe that both `gpt-4` and `gpt-3.5-turbo` somewhat under-perform expert human annotators. Hence, replacing human annotators with LLMs does not seem feasible at this point. Nevertheless, LLMs' automated predictions may support human annotators, improving the quality or the efficiency of their performance. LLMs could

[1] Corresponding Author: Morgan Gray, Learning Research and Development Center: 3420 Forbes Ave. Pittsburgh, PA 15260, USA; Email: mag454@pitt.edu

assign “pre-annotation” labels that indicate whether a sentence is likely to contain any factors and which factors in particular. Human annotators could subsequently confirm or reject the labels after reviewing the model’s suggestion. A concern, however, is whether the annotators would over-rely on the model’s predictions, potentially harming their performance.

Table 1. Factor Type System

1 Occupant Appearance or Behavior	**4 Vehicle**
Furtive Movement	Expensive Vehicle
Nervous Behavior or Appearance	Vehicle License Plate or Registration
Suspicious or Inconsistent Answers	Unusual Vehicle Ownership
2 Occupant Status	**5 Vehicle Status**
Motorist License	Indicia of Hard Travel
Driver Status	Masking Agent
Refused Consent	Vehicle Contents Suggest Drugs
Legal Indications of Drug Use	Suspicious Communication Device
Motorist’s Appearance Related to Drug Use	Suspicious Storage
3 Travel Plans	**6 Other Annotation Labels**
Possible Drug Route	Suspicion Found
Unusual Travel Plans	Suspicion Not Found

To assess the usefulness of GPT models assigning pre-annotation label suggestions to sentences that potentially describe legally relevant factors, we apply the models in the context of a study on Drug-Interdiction Auto-Stop (DIAS) cases. In [1], we identified legal factors that are important in determining whether an officer has reasonable suspicion to detain a motor vehicle on suspicion of drug trafficking. See Table 1 for a listing of DIAS factors. A number of law students annotated 211 legal opinions to identify which sentences in the opinion described any of the factors. This annotation task engaged the seven hired law students over two months at a cost of several thousand dollars.

The time and expense of this manual annotation effort illustrates a bottleneck that limits the number of cases and data points an empirical legal study can address. To alleviate these burdens, we examine the potential of leveraging GPT models to pre-annotate sentences and to suggest labels for human annotators to confirm or change. To investigate using GPT models for pre-annotation, we analyzed these research questions:

(RQ1) What is the zero-shot capability of GPT to identify legal factors in sentences and to identify sentences that are “factor-like”?

(RQ2) Do law student annotators reach a meaningful level of agreement with GPT pre-annotations, or do they over rely on GPT pre-annotations?

(RQ3) Do law student annotators using GPT pre-annotations achieve higher agreement with gold-standard annotations?

In our experiments, we employed two of OpenAI’s generative LLMs, `gpt-4`, the latest and most advanced model, and a variation of its predecessor, `gpt-3.5-turbo-16k`. The `gpt-4` model supports a dialog between the user and the system serving as a kind of assistant. OpenAI has so far withheld technical details of the `gpt-4` model reportedly due to concerns about potential misuses of the technology and the highly competitive market for generative AI [2]. As the size of the successive GPT models has increased [3], however, so, apparently, have their abilities as strong zero- and few-shot learners. However, gpt-4 is slower and more expensive. Therefore, we investigate the capability of the faster and less expensive gpt-3.5-turbo-16k.

This work contributes to AI & Law research in the following ways: First, we show that GPT pre-annotations achieve substantial agreement with gold-standard annotations. Moreover, GPT has an even stronger ability to identify sentences that are "factor-like". Second, we show that law student annotators, using GPT pre-annotations do not overrely on the models' suggestions. Agreement between law student annotators and GPT pre-annotations is similar to agreement between GPT pre-annotations and gold-standard annotations. Lastly, we provide evidence that annotators achieve the same level of agreement with gold-standard annotators regardless of their use of GPT pre-annotations, but they are able to annotate faster.

2. Related Work

This work explores using gpt-4 and gpt-3.5-turbo-16k to support semantic analysis of legal cases by improving annotation of case texts. Some examples of prior work include [4], in which GPT-3.5 generated case summaries using argumentative case segmentation and [5], where LegalBERT generated abstractive case summaries based on argument structure labels. In both, the argumentative structures were based on supervised and semi-supervised annotation of human-prepared case summaries in terms of issues, conclusions, and reasons. Recently, the LegalBench project [6] has assessed how well large language models like OpenAI's GPT-4 can perform a wide range of basic legal reasoning tasks.

Considerable work in AI & Law has focused on making case annotation more effective. Westermann, et al. employed sentence semantic similarity to improve case annotation consistency and efficiency [7]. Researchers have developed annotation pipelines to extract factor-related information from trade secret opinions [8] or from case summaries prepared by law students [9] and to classify trade secret misappropriation opinions by applicable factors [10].

Legal factors are employed in diverse areas of law including copyright fair use, works made for hire, trademark infringement and dilution, assessing spousal support, or determining violations of the right to a speedy trial [11,12,13,14]. In [15], factor values were automatically extracted from divorce cases using rules, augmented with word embeddings. Shaikh, et al. [16] created an ML model to explain outcomes of murder cases based on legal factors. In the SCALE project, the authors employed semi-supervised case annotation in training an ML program to pair legal issues in WIPO domain name dispute cases with applicable factors [17].

Unlike the SCALE project, the DIAS cases we address involve more varied judicial styles and factual circumstances. As the LegalBench authors note on p. 28, "an example of a reasoning ability which is not currently evaluated in LegalBench would be analogical reasoning grounded in case law." In particular, the LegalBench project does not address a legal reasoning task to which we have here applied GPT-4: classifying sentences in opinion texts by factors to support case-based reasoning. In our auto stop domain, GPT-4 needs to classify sentences in cases by eighteen categories of factor types. To our best knowledge, we are the first researchers to do so.

Table 2. Dataset Statistics

Data	No. Sentences	Mean	Median	Min	Max	Avg. Sentence Length
Base 346	64275	186	159	15	1042	128.0
Coarse 346	49670	144	123	8	785	133.0
Annotation Set (100)	14396	141	136	15	414	134.0

3. Dataset

We build on the annotated dataset described in [1], which contained cases relevant to whether an officer had reasonable suspicion to detain a motorist to find evidence of drug trafficking. Here, we employed two law students to search the Harvard Law School Case Law Access Project (HCAP. `https://case.law/`.) for DIAS cases. The students employed the following search queries: "'reasonable suspicion' and 'canine' and 'detention'" and "'reasonable suspicion' and 'detention' and 'k-9' or 'K9'". Having retrieved a relevant case, they also retrieved relevant cases cited therein. Ultimately, this procedure returned 346 cases. Table 2 provides a statistical summary of the sentences in these cases. Based on our experience with the corpus in [1,18], we found that roughly 90% of the corpus contains sentences that do not describe any factor. We implemented a rule based classifier to filter out sentences that clearly do not describe a factor.[2] As shown in Table 2, rows 1 and 2, this coarse classifier reduced the number of sentences by about 23%.

Of the remaining sentences, we randomly selected 100 cases of varying lengths and split them into two groups of 50 for annotation. The first group, called "base" cases, was annotated by law students without the help of GPT and used to measure the effect, if any, of GPT assistance. The second group was split into two groups of 25, one set pre-annotated by `gpt-4` and the other by `gpt-3.5-turbo-16k`. We used these to compare the performance between `gpt-3.5-turbo-16k`, the less expensive and faster model, and `gpt-4` the slower, more expensive, but more powerful model.

4. Experiments

4.1. Pre-Annotation with GPT

For both models, we set the `temperature` of the model to 0.0. This parameter controls randomness. Higher `temperature` values cause more creative output but it can also be less factual. Temperatures closer to 0.0, cause the model to be deterministic and repetitive. For the task of pre-annotation, where the same set of labels are going to be applied consistently, we desired deterministic and repetitive output. We set `max_tokens` on the models' output to between 3000 and 3400 depending on the size of the prompt. A token corresponds roughly to a word. GPT-4 has an overall token length limit of 8,192 tokens, comprising both the prompt and the completion. 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 creativeness of the output. We set `frequency_penalty` to 0, which ensures no penalty is applied to repitious language. Finally, we set `presence_penalty` to 0, ensuring no penalty is applied to tokens appearing multiple times in the output.

[2] The classifier works by identifying sentences that clearly do not describe a relevant type. These include legal citations, case information, document headers, etc.

Table 3. Average Input, Completion, and Total Tokens for each GPT model.

Average	Input Tokens	Completion Tokens	Total Tokens
`gpt-3.5-turbo-16k`	3667	3729	7466
`gpt-4`	3730	3856	7741

Table 4. Abridged Guideline-Prompt Example.

1. TASK In this task, we are attempting to label sentences to assess whether they contain important information. [81 characters...]

2. We are interested in highway drug-interdiction. This occurs when a motorist is stopped by police [2694 characters...]

3. YOU will assess sentences to and determine whether they belong to any of the following categories ...
Furtive Movement - Use this label if the driver or passenger in the vehicle makes a suspicious movement, [5759 characters...]

4. You are also to follow these specific rules:
Typically, a sentence will describe a single factor, however, in some cases, a single sentence may include more than one factor [1738 characters...]

5. You should apply a label for each sentence. Here are some examples:

   1. Sentence: Officer Guthrie testified that while the above exchanges were taking place, he noticed that the driver, Arturo Tapia, seemed nervous, and that his hands were shaking.
   Label: Physical Appearance of Nervousness [502 characters...]

6. Label all of these *n* sentences: [*n* Sentences]

4.2. *Prompt Development*

Pre-annotation with both GPT models was carried out with a single prompt. Generally, we pre-annotate by prompting the model to label a number of sentences based on certain rules, definitions, and instructions. To develop the prompt, we follow [19], and provide the model with almost an exact copy of the annotation guidelines provided to annotators in [1] (cf. [20] where only excerpts are used). We call this "guideline-prompting". As in [19], we developed the final version of the prompts over 7 iterations of testing. At each iteration, the performance of the prompt was evaluated, with commensurate edits to the prompt to improve the models' performance. This procedure is described in more detail in [21]. There, we showed that about 90% of the time, `gpt-4` provided a label comparable to a reasonable annotator's. Table 4 shows an outline of the prompt we used. We introduce the model to the task (item 1), provide a description of the legal problem (item 2), and describe in detail each factor (item 3). Next, in item 4, we provide the model with specific instructions, similar to those that would be provided to annotators. We then provide the model with example sentence-label pairs (item 5) and sentences for the model to label (item 6). In Table 1 we show the type-system available both to GPT as a pre-annotator and to the human annotators. Generally, we follow the same type system as defined in [1]. When prompting both models we include the basic guideline prompt and

40 sentences to label. As shown in Table 3 the prompt and 40 sentences, with completion, comprises between 7,400 and 7,700 tokens on average. Although `gpt-3.5-turbo-16k` can handle up to 16,000 tokens, to maintain comparability of the models' performance, we stayed below the `gpt-4` threshold of 8,192 tokens. The token counts in Table 3 were calculated with the Python `tiktoken` package.

4.3. Human Annotation with Pre-Annotated Sentences

The guideline-prompt instructed the models to return each of the 40 sentences with the model's suggested label. Here is an example of the pre-annotations produced by GPT-4:

> **Sentence 15**: Throughout the encounter Adrienne, rather than Angela, the driver, did almost all the talking, which Krause said can be a sign of nervousness.
> **Label:** Nervous Behavior or Appearance

Two second year law students were trained during three initial sessions. A first session introduced the annotators to the factors they would be searching for. The students then annotated a handful of training cases. In a second session, they received feedback on their labeling of the training cases. They then annotated about 25 cases for a third session in which they received additional instruction based on the quality of their annotations.

After initial training, the students annotated two sets of legal opinions, segmented into sentences, each sentence on a new line. For one set of legal opinions, a GPT model had pre-annotated the sentences and a suggested label was included below each sentence as shown in the example. For both sets of opinions, annotators were instructed to mark up sentences that describe 1) any of the 18 factors identified in Table 1 and 2) the court's conclusion as to whether the reasonable suspicion standard was satisfied. The dataset's limited size and the cost of creating it are significant limitations that we address in this work. The annotators were instructed to read *each* sentence in the opinion, including sentences where the model suggested "No Factor", decide if the model's suggested label was correct, and choose the appropriate label. Annotators received an equal mix of sentences annotated by `gpt-4` and by `gpt-3.5-turbo-16k`. The other set of legal opinions had not been pre-annotated by a GPT model. The annotators received the opinions segmented into sentences in the same format, except each "Label" suggestion was set to "None". The annotators were instructed to apply the label they believed was appropriate.

5. Results & Discussion

5.1. RQ1: GPT's zero-shot capability to pre-annotate cases and identify factor-like sentences

Understanding how either GPT model performs on the pre-annotation task as compared to gold-standard human annotations is key to understanding whether it would be a useful starting point for other human annotators. Gold standard annotations were performed on the base cases without the assistance of GPT pre-annotation and were performed by legal expert familiar with the anntation task and legal domain. For gold-standard comparisons, we randomly sampled 60 out of our 100 law-student annotated cases for gold-standard annotation (30 to compare base annotations, and 30 to compare pre-annotations). To ad-

Table 5. Agreement between GPT pre-annotations and gold standard

Pre-annotation	% Agreement	Cohen's κ	Gwet's AC1	Per case
`gpt-3.5-turbo-16k`	0.76	0.39	0.73	0.56
`gpt-4`	0.80	0.41	0.77	0.63

Table 6. GPT pre-annotation ability to identify factor-like sentences.

Model	% GPT-Gold Standard	GPT-Human Annotation
`gpt-3.5-turbo-16k`	0.85	0.87
`gpt-4`	0.89	0.95

dress RQ1, we first measure the agreement between pre-annotated cases with `gpt-4` and `gpt-3.5-turbo-16k`. First, we examine the overall percentage of agreement at the sentence level. The models register scores of 0.76 and 0.80, for `gpt-3.5-turbo-16k` and `gpt-4`, respectively. Next, we examine the Cohen's κ between GPT pre-annotations and gold-standard annotations. With `gpt-4`, the pre-annotations and gold-standard annotations share a Cohen's κ score of 0.41, which according to [22] corresponds to moderate agreement. In [1], on the same task, annotators achieved moderate agreement with gold-standard annotations with a κ of 0.57. Although the score in [1] is higher, both scores are in the moderate range as defined by [22].

The Cohen's κ statistic is known to suffer from a "paradox". Where one class has a very high prevalence in comparison to others, Cohen's κ can be too low [23]. Such is the case with our data, where the "No Factor" type consists of roughly 90% of the labels [18]. To address this we apply Gwet's AC1, which like Cohen's κ measures inter-rater agreement and accounts for chance agreement, but has been found resistant the paradox just described [24]. Between `gpt-4` and gold-standard annotations Gwet's AC1 was 0.77. As the authors in [24] did, we rank this score using the scale promulgated by [22] and conclude that this score corresponds to substantial agreement.

The above calculations are based on agreement as to the labels for each individual *sentence*. We are also interested in the agreement *per case*, that is, to what extent do the GPT pre-annotations agree with the factors that have been assigned to each gold standard case. For a factor to be assigned to a case, an annotator needs to have found at least one sentence is an instance of that factor. To measure this, for each case, we divided the cardinality of the intersection of the set of factors identified by the annotators by that of the union of the set of factors identified by the annotators. In Table 5 under the column "Per case," with `gpt-4` pre-annotations and gold-standard annotations agree 0.63 of the time on what factors are present in each case.

We estimated both models' capability to identify whether a sentence is "factor-like". We considered any sentence to which an annotator assigned a factor as a factor-like sentence regardless of whether it was the correct factor. We measured how many of these sentences also were assigned a factor by GPT in pre-annotation. In other words, our focus is on whether the annotators deem a sentence worthy of any label, rather than the specificity of the label they assign. The percentage of factor-like sentences identified appears to be high.

With respect to RQ1, based on the results in Table 5, we see that GPT pre-annotations under-perform human expert annotation and are not sufficient for the task

Table 7. Agreement between annotators and both GPT models.

Model	% Agreement	Cohen's κ	Gwet's AC1	Per case
`gpt-3.5-turbo-16k`	0.78	0.39	0.74	0.53
`gpt-4`	0.86	0.53	0.84	0.63

of identifying factors in legal opinions. On the other hand, pre-annotating opinions with GPT is successful enough as to present a reasonable starting place for attempting to improve the quality or efficiency of human annotation. We note that using pre-annotations provided by `gpt-4` improves annotator performance across all metrics, when compared to using `gpt-3.5-turbo-16k`. Since the more powerful `gpt-4` is more expensive and slower, `gpt-3.5-turbo-16k` may be preferable depending on one's resources. In our study, there were 20 possible factor labels for sentences. In Table 6, we were not concerned with which specific factor label was applied, but rather if any factor label was applied. This metric evaluates the agreement between GPT pre-annotations and human annotators on identifying these "factor-like" sentences.

5.2. RQ2: Is there meaningful agreement between GPT pre-annotations and law student annotations, or do law student annotators over rely on the model's suggestions?

A major concern with the use of GPT pre-annotations, is that law student annotators would be tempted to follow the model's suggestions without much scrutiny. Before addressing this, we examined the agreement between GPT pre-annotations and law student annotators. The results of this experiment are shown in Table 7. As with the gold-standard v. GPT pre-annotations shown in Table 5, agreement with `gpt-3.5-turbo-16k` is generally lower than with `gpt-4`. As for `gpt-4`, agreement between law student annotators and the model's pre-annotations is generally higher across all metrics and is comparable to the agreement between gold-standard and GPT pre-annotations, which confirms that GPT pre-annotations are a reasonable starting point for law student annotators.

Importantly, when viewing the sets of results presented in Table 5 and 7, we can conclude that these law student annotators did **not** "blindly" follow the pre-annotations provided by GPT. We note that the agreement between GPT pre-annotations and human annotators is not much higher than the agreement between pre-annotations and the gold standard. If the law student annotators had frequently adopted the labels suggested by GPT, one would expect nearly perfect agreement. That is not what we see here. Although there is a bump in agreement between pre-annotations and human annotators, when compared to gold-standard annotations, it is not extreme. Thus, we conclude that human annotators may be slightly more likely to follow GPT pre-annotations but this does not seem to harm their annotation performance.

5.3. RQ3: Does pre-annotation with GPT improve manual annotation?

Table 8 shows the comparison between the gold standard and law student annotators with and without the use of pre-annotations. As indicated, the results are very similar and sometimes identical. We can conclude that, while the use of pre-annotations does not improve the quality of their performance, it does not harm performance either. That

Table 8. Agreement between gold-standard and law students without pre-annotations and between gold-standard and law students with pre-annotations.

Gold-Standard v. Law Student with Pre-Annotations				
Pre-annotation	% Agreement	Cohen's κ	Gwet's AC1	Per case
`gpt-3.5-turbo-16k`	0.901	0.557	0.878	0.734
`gpt-4`	0.901	0.569	0.883	0.792
Gold-Standard v. Law Student Without Pre-Annotations				
Base v. Gold Standard	0.913	0.565	0.894	0.729

is important if, as the annotators report, using pre-annotations made their work go more quickly, especially at the beginning. Perhaps one of the most important measures is if law student annotators and gold standard annotations are agreeing on what factors are present *per case*. As shown in Table 8 law student-annotators and gold-standard annotations agree more on what individual factors are present in each case with the help of GPT pre-annotations. Not only are law student annotators not over-relying on pre-annotations, but the pre-annotations are helping agreement on what factors are present. This is important for pipelines which predict the outcome of suspicion cases, like that in [18], where cases are represented as binary vectors indicating what factors were identified in a case. This insight, and the showing that GPT pre-annotations and law students agree on what sentences contain a factor, show a meaningful utility for using pre-annotations.

6. Conclusions and Future Work

In this work, we have applied `gpt-4`'s and `gpt-3.5-turbo-16k`'s capability for zero-shot performance to the task of automatically identifying sentences in legal opinions that describe factors of interest. First, we have observed that direct application of `gpt-4` somewhat under-performs the human experts, which is to be expected. We have shown that pre-annotation with `gpt-4` does not harm the quality of law students' annotations and confirmed that the students do not appear to over-rely on the pre-annotations. We have found anecdotal evidence that pre-annotation speeds up the law students' annotation work. In addition, we have shown that `gpt-4` can effectively identify factor-like sentences, which could focus human annotators more quickly on those sentences in the opinion that are most likely to be worth their attention. While it does not appear that pre-annotation can notably improve the quality of the law students' annotations, it has the potential to make the process more efficient, preserving the quality achieved by law students annotating the texts independently.

In future work, we will investigate whether GPT pre-annotations enable humans to annotate more efficiently. While our cases and factors pertain to drug interdiction auto stops, there is no apparent reason why our techniques would not apply to other legal domains that involve reasoning with factors.

References

[1] Gray M, Savelka J, Oliver W, Ashley K. Toward Automatically Identifying Legally Relevant Factors. In: Legal Knowledge and Information Systems. IOS Press; 2022. p. 53-62.

[2] OpenAI. Gpt-4 technical report. arXiv preprint arXiv:230308774. 2023.

[3] Brown T, Mann B, Ryder N, Subbiah M, Kaplan JD, Dhariwal P, et al. Language models are few-shot learners. Advances in neural information processing systems. 2020:1877-901.

[4] Xu H, Ashley K. Argumentative Segmentation Enhancement for Legal Summarization. ASAIL 2023 preprint arXiv:230705081. 2023.

[5] Elaraby M, Litman D. ArgLegalSumm: Improving Abstractive Summarization of Legal Documents with Argument Mining. Proc 29th Int'l Conf on Computational Linguistics. 2022:6187-94.

[6] Guha N, Nyarko J, Ho DE, Ré C, Chilton A, Narayana A, et al. LegalBench: A Collaboratively Built Benchmark for Measuring Legal Reasoning in Large Language Models. arXiv:230811462. 2023.

[7] Westermann H, Savelka J, Walker VR, Ashley KD, Benyekhlef K. Sentence embeddings and high-speed similarity search for fast computer assisted annotation of legal documents. In: Legal Knowledge and Information Systems: JURIX 2020: 33d Annual Conference, Brno, Czech Republic, December 9-11, 2020. vol. 334. IOS Press; 2020. p. 164.

[8] Wyner A, Peters W. Towards annotating and extracting textual legal case factors. In: SPLeT-2012; 2010. p. 36-45.

[9] Ashley KD, Brüninghaus S. Automatically classifying case texts and predicting outcomes. Artificial Intelligence and Law. 2009;17(2):125-65.

[10] Falakmasir M, Ashley K. Utilizing Vector Space Models for Identifying Legal Factors from Text. In: JURIX 2017. vol. 302. IOS Press; 2017. p. 183-92.

[11] Beebe B. An empirical study of the multifactor tests for trademark infringement. Calif L Rev. 2006;94:1581.

[12] Beebe B. An empirical study of US copyright fair use opinions, 1978-2005. U Pa L Rev. 2007;156:549.

[13] Beebe B. An Empirical Study of US Copyright Fair Use Opinions Updated, 1978-2019. NYU J Intell Prop & Ent L. 2020;10:1.

[14] Rempell S. Factors. Buffalo L Rev. 2022;70:1755. Available from: http://dx.doi.org/10.2139/ssrn.4095435.

[15] Li J, Zhang G, Yu L, Meng T. Research and design on cognitive computing framework for predicting judicial decisions. Journal of Signal Processing Systems. 2019;91:1159-67.

[16] Shaikh RA, Sahu TP, Anand V. Predicting outcomes of legal cases based on legal factors using classifiers. Procedia Computer Science. 2020;167:2393-402.

[17] Branting LK, Pfeifer C, Brown B, Ferro L, Aberdeen J, Weiss B, et al. Scalable and explainable legal prediction. Artificial Intelligence and Law. 2021;29(2):213-38.

[18] Gray M, Savelka J, Oliver W, Ashley K. Automatic Identification and Empirical Analysis of Legally Relevant Factors. In: Int'l Conf. Artificial Intelligence and Law. ACM Press; 2023. p. 53-62.

[19] Savelka J, Ashley K, Gray M, Westermann H, Xu H. Can GPT-4 Support Analysis of Textual Data in Tasks Requiring Highly Specialized Domain Expertise? In: Automatic Semantic Analysis of Information in Legal Text (ASAIL); 2023. .

[20] Savelka J, Ashley KD. The Unreasonable Effectiveness of Large Language Models in Zero-shot Semantic Annotation of Legal Texts. Frontiers in Artificial Intelligence. 2023;6:1279794.

[21] Gray M, Savelka J, Oliver W, Ashley K. [In review]. Empirical Legal Analysis Simplified: Reducing Complexity through Automatic Identification and Evaluation of Legally Relevant Factors. To appear in Philosophical Transactions A. 2023.

[22] Landis JR, Koch GG. The measurement of observer agreement for categorical data. biometrics. 1977;33(1):159-74.

[23] Zec S, Soriani N, Comoretto R, Baldi I. High Agreement and High Prevalence: The Paradox of Cohen's Kappa. The Open Nursing Journal. 2017:221-18.

[24] Wongpakaran N, Wongpakaran T, Wedding D, Gwet K. A comparison of Cohen's Kappa and Gwet's AC1 when calculating inter-rater reliability coefficients: a study conducted with personality disorder samples. BMC Med Res Methodol. 2013.