CaseSumm: A Large-Scale Dataset for Long-Context Summarization from U.S. Supreme Court Opinions

Mourad Heddaya
University of Chicago
mourad@uchicago.edu

Kyle Macmillan
University of Chicago
macmillan@uchicago.edu

Chenhao Tan
University of Chicago
chenhao@uchicago.edu

Anup Malani
University of Chicago Law School
amalani@uchicago.edu

Hongyuan Mei
Toyota Technological Institute at Chicago
hongyuan@ttic.edu

Abstract

This paper introduces a novel dataset, CASESUMM, for long-context summarization in the legal domain, addressing the need for longer and more complex datasets for summarization evaluation. We collect 25,611 U.S. Supreme Court (SCOTUS) opinions and their official summaries (called syllabuses). Our dataset is the first to include summaries of SCOTUS decisions dating back to 1815, resulting in the largest open legal case summarization dataset.

We also present a comprehensive evaluation of LLM-generated summaries using both automatic metrics and expert human evaluation, revealing discrepancies between these assessment methods. Our evaluation shows Mistral 7b, a smaller open-source model, outperforms larger models on automatic metrics and successfully generates syllabus-like summaries. In contrast, human expert annotators indicate that Mistral summaries contain hallucinations and annotators consistently rank GPT-4 summaries as clearer and exhibiting greater sensitivity and specificity. Our analysis identifies specific hallucinations in generated summaries, such as precedent citation errors and misrepresentations of case facts. These findings demonstrate the limitations of current automatic evaluation methods for legal summarization and underscore the critical role of human evaluation in assessing summary quality, particularly in complex, high-stakes domains.

1 Introduction

Although large language models (LLMs) are claimed to handle long contexts (GPT-4 Team, 2024; Bubeck et al., 2023; Claude Team, 2024), including summarizing very long inputs, how well they perform long-context summarization is an open question.

Evaluating long-context summarization is challenging for several reasons. First, human ground-truth summaries are often not available (Cao et al., 2024; Chang et al., 2024). Moreover, it's unclear whether we should trust human abilities to even create ground-truth summaries. Second, the desiderata remain opaque. What makes a good summary in one domain may not generalize to another domain. For example, what's relevant in a legal text is different than what's relevant in a novel. Lastly, identifying salient information in complex domains often requires expertise. Our work addresses these challenges by introducing a new dataset where "ground-truth" summaries are available and conducting a comprehensive human evaluation to benchmark existing models.

We build a legal case summarization dataset focused on U.S. Supreme Court cases, CASESUMM. Our dataset consists of 25,611 U.S. Supreme Court cases and their official summaries, called syllabuses, from the period 1815-2019. Syllabuses are written by an attorney employed by the Court and approved by the Justices. The syllabus is therefore the gold standard for summarizing majority opinions, and ideal for evaluating other summaries of the opinion. We obtain the opinions from Public Resource Org's archive[1] and extract syllabuses from the official opinions published in the U.S. Reporter and hosted by the Library of Congress. Our dataset is at least 25% larger, covers 3 times as many years, and is publicly available with fewer copyright restrictions than similar legal datasets (Fang et al., 2023; Trivedi et al., 2024), representing a rich resource for the research community.

Beyond the legal domain, several datasets have been introduced to improve evaluation of long-context summarization (Kryściński et al., 2022; Sharma et al., 2019; Eidelman, 2019; Huang et al., 2021). CASESUMM continues the trend of larger datasets with both longer source and summary texts, where the summaries represent high quality ground-truths. In contrast to existing work, however, our

[1] https://public.resource.org/

dataset spans over two centuries, demonstrating unique variation in the lengths and compression rates of summaries, while also reflecting a high-stakes and useful domain for summarization.

To highlight the opportunities and challenges of our dataset, we present both automatic and human expert evaluations of LLM-generated summaries of SCOTUS opinions and include two "control" human-written summaries from Westlaw and Oyez. According to both human and automatic metrics, fine-tuning Mistral successfully guides the model to more accurately mimic the official syllabuses and reflect the lexical and semantic content within them, than other much larger models.

However, we find that automatic and human metrics disagree: syllabuses and fine-tuned Mistral summaries perform highly on automatic evaluation but rank lower according to human evaluators, whereas GPT-4 is reliably ranked highly in human evaluation despite only average performance on automated metrics. Furthermore, GPT-4-generated summaries often outperform human-written ones, including official syllabuses, but not on factual correctness. These findings challenge the notion of human-written ground-truth summaries.

Finally, we conduct an error analysis of hallucinations in GPT-4- and Mistral- generated summaries and identify factual errors ranging from precedent citation errors to misrepresentations of the facts of the case and procedural history as recounted in the source opinions.

In sum, we make the following contributions:

- We introduce a new large-scale dataset for long-context summarization in the legal domain, consisting of 25,611 U.S. Supreme Court cases and their official syllabuses from 1815-2019.

- We present a comprehensive evaluation of LLM-generated summaries using both automatic metrics and expert human evaluation, revealing discrepancies between these assessment methods.

- We provide a comparative analysis of summaries generated by fine-tuned models and larger, general-purpose models, offering insights into their relative strengths and weaknesses in legal summarization tasks.

We will release the dataset and code upon publication.

2 Related Work

Evaluation for summarization. ROUGE (Lin, 2004) has been the dominant summarization metric, despite criticism of its high lexical dependence (Schluter, 2017; Cohan and Goharian, 2016). Newer metrics like BERTScore (Zhang et al., 2019) and BARTScore (Yuan et al., 2021) aim to capture semantic similarities. However, automatic metrics often don't correlate well with human judgments (Yuan et al., 2021; Fabbri et al., 2021; Bhandari et al., 2020). We focus on high-stakes long-context summarization, showing the need for better metrics persists despite LLM progress. Chang et al. (2024) extended LLM-based evaluation to book-length summaries, but this approach doesn't consider how experts weigh the importance of including or omitting certain information in a summary, while also being slow and costly. Cao et al. (2024) developed a framework for characterizing LLM summaries of financial documents. Our work extends this research by evaluating and comparing model- and human-generated summaries in the legal domain. Addressing factual discrepancies in model-generated summaries, recent work has developed automatic methods for evaluating faithfulness in summarization (Krishna et al., 2023; Chang et al., 2024; Falke et al., 2019; Laban et al., 2022; Wang et al., 2020; Fabbri et al., 2022).

NLP and summarization in the Legal Domain. Natural language processing has been applied to various legal tasks, including summarization (Bauer et al., 2023), discovery (Zou and Kanoulas, 2020), redaction (Garat and Wonsever, 2022), case outcome prediction (Medvedeva et al., 2023; Cui et al., 2023), and Bar Exam performance (Katz et al., 2023a). For comprehensive surveys of NLP in the legal domain, see Katz et al. (2023b) and Kapoor et al. (2024).

Datasets in the legal domain. Our dataset is unique in providing U.S. Supreme Court opinions with syllabuses, unlike other datasets that lack syllabuses (Chalkidis et al., 2022; Henderson et al., 2022) or provide only ancillary data (Law, 2024). Fang et al. (2023) present Super-SCOTUS, a dataset of Supreme Court documents, including a subset of syllabus (scraped from online websites and not validated) and opinion pairs and highlight its contribution to political and social science research. In contrast, our CASESUMM dataset is comprised of pre-processed and cleaned syllabuses and opinions that extend further back to 1815 and are extracted directly from source opinions, providing the community with a readily available summarization resource with fewer copyright restrictions.

3 Dataset

When the Supreme Court resolves a case, it publishes a majority "opinion" announcing the outcome and reasoning for their decision. In addition, the Court will disseminate a summary of the opinion called the "syllabus", which is written by an attorney employed by the Court and approved by the Justices. The syllabus must include the main elements of the opinion: the facts of the case, the procedural history, the legal question to be decided, and the answer to that question. Accurately summarizing each of these sections requires (1) understand sophisticated legal reasoning and (2) identify the most salient aspects of the case.

As one of the longest standing institutions in U.S. history, the Supreme Court has published thousands of opinions and syllabuses over the past 200 years. Looking at cases between 1815 and 2019, we collect 25,611 pairs of opinions and syllabuses for our dataset, to be available under a CC BY-NC 4.0 license.

Dataset construction We compile our dataset from multiple sources. Opinions published in U.S. Reports Volume 15-546 (years 1815-2005) and Volumes 546-591 (2005 through *Trump v. Vance* (2019)) are obtained from Public Resource Org's online archive (Public Resource Org, 2024) and the Super-SCOTUS data set (Fang et al., 2023), respectively. We extract syllabuses from PDFs of the opinions hosted on the Library of Congress's website (Library of Congress, 2024).

Extracting syllabuses from the original PDFs is challenging for several reasons. First, identifying the start and end of the syllabus is complicated because the formatting and style of SCOTUS decisions have changed over time. Low quality scans of 19th and 20th century documents make the extraction task even more difficult. Together, these issues constrain the kinds of rules, or signals, we can leverage to reconstitute the structure of the text in the PDFs, requiring us to devise creative alternatives. For example, while syllabuses have a smaller font-size than the rest of the decision and would be a straight-forward heuristic to leverage, this information is often incorrectly encoded in OCR data.

To ensure accurate syllabus extraction, we process the PDFs in multiple ways. First, we design a set of regular expressions to identify the start of the syllabus, providing coverage of decisions with different styles. Then, we develop an algorithm based on open-source computer vision software (Bradski, 2000) to identify continuous lines, allowing us to distinguish footers from the main text of a page. Finally, we take advantage of differences in line density, a measure that is more robust to OCR and scan quality, combined with regular expressions to determine when the syllabus ends.

Since we build a new dataset, there are no accessible ground-truths to automatically evaluate our technique for extracting syllabuses from PDFs. Instead, we randomly sample 100 cases and manually evaluate the extracted syllabus by comparing them to the original PDFs. We find that 96 of the 100 are perfect extractions while the remaining 4 syllabuses are partially truncated. These results highlight the quality of our dataset as a rich resource for long-context summarization.

Descriptive statistics To demonstrate the value of our dataset as a resource for abstractive summarization, we compare the lengths of the opinions and their syllabuses. The average Supreme Court opinion is 2,612 words long. The average syllabus is 314 words long, about 21.8% the length of the opinion it is summarizing. Figure 1 shows lengths have risen over time. Since 1980, opinions and syllabuses average 4,151 and 731 words, respectively, nearly double the average for the entire 1815-2019 period. Although compression rates, defined as the ratio of words in a syllabus to words in an opinion have been relatively stable over time, averaging 21.8% from 1815-2019, the Pearson correlation between the length of an opinion and its syllabus, while variable, has increased over time. Whereas this correlation was just 0.46 before 1920, it has been 0.68 since then. Given the changes in opinion and syllabus lengths and in the correlation between syllabus and opinion lengths, this data set is a valuable resource for modeling and evaluating expert summaries, especially in the legal domain.

4 Experiment Setup

In this section, we introduce our summarization task setup and evaluation strategies.

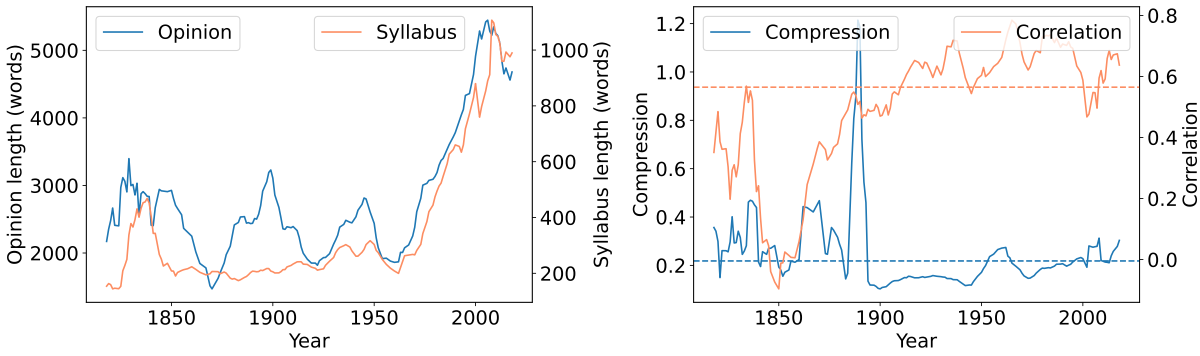

Figure 1: Opinion and syllabus lengths, compression rates by syllabuses, and correlations between opinion and syllabus lengths, 1815-2019. Dashed blue and orange lines give average compression rate and correlation. Lines are smoothed with 5-year moving-average.

4.1 Data and Modeling

Data preprocessing and splits. We use syllabuses as a supervision signal in our summarization modeling experiment and as reference summaries for evaluating the human and model-generated summaries.

As discussed in §3, the substance and style of syllabuses have changed over time. Therefore, the supervision signal has changed over time. The motivating use case in our summarization task is a legal professional conduction research. For such a professional, while concision has value, comprehensiveness is more valuable. By manually studying summaries, we determine that more comprehensive syllabuses begin with a summary of the facts of the case, followed by a new section—marked by the text "Held:"—containing details about the issues, analyses, and conclusions that the opinion commented on. Modern syllabuses consistently adhere to this structure.

Therefore, we filter our dataset to include only opinion/syllabus pairs where the syllabus contains the pattern "Held:". We call this subset of the dataset "structured". We find that the length of structured syllabuses is more strongly correlated with the length of their respective opinions ($r = 0.65$) than the length of unstructured syllabuses is with the length of their opinions ($r = 0.46$). Furthermore, structured syllabuses are on average 2.5x longer than the unstructured syllabuses. Overall, the structured dataset contains 6,683 case/syllabus pairs. We split these into a training set (n=5,455), validation set (n=606), and test set (n=622).

Modeling. We pursue and test two approaches for completing our legal case summarization task. The first approach is zero-shot prompting with proprietary and with open-source LLMs. The propriety LLM we employ is GPT-4 Turbo (`gpt-4-1106-preview`) (GPT-4 Team, 2024). The open-source LLM we employ is Mistral 7b Instruct (`Mistral-7b-Instruct-v0.1`) (Jiang et al., 2023). The opinions in our dataset have 4,983 tokens on average, and the syllabuses average 755 tokens. The second approach is instruction fine-tuning (Wei et al., 2021) the open-source model, Mistral 7b Instruct, using the syllabuses in our training data set. We will refer to models used in a zero-shot setting by model name: Mistral Base and GPT-4, and to the fine-tuned Mistral model as Mistral FT.

For Mistral in both settings, we design a prompt following best practices suggested by its authors.[2] For GPT-4, we optimize prompt-selection using DSPy (Khattab et al., 2023) with 10 opinion/syllabus pairs from the training set and ROUGE-2 as the optimization metric.

For fine-tuning, our input consists of a short instruction, the opinion, and the syllabus. We do standard auto-regressive language modeling but only backpropagate the language modeling loss for the syllabus. We use LoRA-based Parameter-Efficient Fine-Tuning (PEFT) (Hu et al., 2021) to train a subset of the parameters.

4.2 Evaluation strategies

"Control group" summaries. We benchmark our three machine-generated summaries (Mistral Base, Mistral FT, and GPT-4 Turbo) along with two additional human-generated sources for purposes of having a control group of human-written summaries not explicitly intended to mimic syllabuses. First, we collect public Oyez summaries from the Super-SCOTUS dataset (Fang et al., 2023). Oyez

[2] https://huggingface.co/mistralai/Mistral-7B-Instruct-v0.1

Method	ROUGE-1 ↑			ROUGE-2 ↑			ROUGE-L ↑			BERTScore ↑		
	P	R	F1	P	R	F1	P	R	F1	P	R	F1
GPT-4 Turbo	71.2	37.1	45.1	31.1	15.5	19.2	34.9	18.1	21.9	**67.4**	62.2	**64.6**
Mistral Base	64.3	13.4	20.0	23.4	4.6	7.0	41.1	7.8	11.8	61.3	48.5	54.0
Mistral FT	63.3	**43.1**	**48.1**	30.1	**20.5**	**23.0**	34.9	**23.6**	**26.4**	66.0	**64.4**	**65.1**
Oyez	64.0	35.1	41.6	28.5	15.0	18.1	34.4	18.6	22.1	64.2	61.8	62.9
Westlaw	**71.5**	20.5	29.4	**32.7**	9.1	13.2	**42.3**	11.8	17.0	65.0	55.7	59.9

Table 1: Automatic evaluation of model-generated and human-written summaries, where official syllabuses are the reference summaries. Sample includes 622 Supreme Court cases. There are 622 observations on each type of summary except Westlaw, for which we only have 156 observations. For each metric, we report precision (P), recall (R), and F1-score (F1). For each metric, we **bold** the best score(s) and underline the second best score(s).

summaries are composed of three sections: Facts of the Case, Question, and Conclusion. Second, we collect Westlaw's commercial summaries of cases via their online interface.[3] Because manual download is slow, our sample size for Westlaw downloads was smaller: whereas our test set has 622 instances of model-generated summaries and Oyez summaries, we have 156 Westlaw summaries.[4]

Automatic Evaluation. Following recent work on summarization (Koh et al., 2022), we use ROUGE and BERTScore (Lin, 2004; Zhang et al., 2019) as our automated metrics for evaluating generated summaries against the reference syllabuses. With this, we assess the *relevance* of the summaries. We breakdown each of the metrics by their precision, recall, and F1-score, highlighting how models balance trade-offs between coverage and concision. We also experimented with BARTscore (Yuan et al., 2021) (see Appendix B.3) but exclude it from our main analysis due to its sensitivity to whether text is in- or out-of- distribution relative to the scoring model. Since we compare Mistral after fine-tuning on syllabuses to models that were not fine-tuned, we expect unreliable results.

To further characterize the summaries, we compare the summaries based on *compression rate*, defined as the number of words in a syllabus over the number of words in an opinion, and the *correlation* between opinion lengths and summary lengths. We use compression as a measure of brevity and correlation as a measure of how responsive summaries are to changes in the amount of content in the opinions.

Human Evaluation. For the human evaluation, we recruited and paid[5] second- and third-year law students to read several opinions and 5 summaries of each opinion (Mistral FT[6], GPT-4, official syllabus, Westlaw, and Oyez). We asked students to rank each summary (from 1 to 5) on several metrics: did the summary contain all relevant information from the opinion (*sensitivity*), did it exclude irrelevant information (*specificity*), was the summary clear (*clarity*), and did the summary have a style suggesting it was written by an experienced attorney? (*style*) Finally, we asked students to report the number of facts in the summary that were false based on their reading of the opinion (*error*). Students were not told the source of each summary.[7] See Appendix C for additional details on the annotation interface and procedure.

In total, students read 57 opinions. Our sample of opinions and summaries included 33 unique cases, and the median student read 5 cases. Given that we ask students to rank opinions from 1 to 5 (implying a mean of 3 and variance of 2), our minimum detectable effect, with 95% confidence and 80% power, was 0.52 rank points.

5 Results

Our results indicate consistencies and discrepancies in the outcomes of automatic and human evaluations. On the one hand, model-generated

[3]We obtain these manually to avoid legal risks under our Westlaw subscription license.

[4]We initially included summaries from Justia, another publicly available legal resource, as a human baseline but, after manually inspecting 5 randomly sample summaries, we determine that they were largely derivative of the Court syllabuses and copied significant quantities of text from them. This was further validated by finding that Justia summaries achieved 0.97 ROUGE-1 score, which is exceedingly alike in a long-form summarization task such as this.

[5]Participants were paid $20/hr, $4 more than RA minimum. See Appendix C.3 for instructions & consent.

[6]We exclude Mistral base from our comparisons because it has rather poor performance overall on automatic metrics, helping us reduce the cognitive load on our participants.

[7]This evaluation was deemed exempt from IRB review by our institution's IRB (IRB24-0277).

summaries largely outperform the control human-written summaries on automatic measures of relevance, while also matching or exceeding them in our human evaluation. On the other hand, automatic metrics prefer Mistral FT summaries over GPT-4 ones, whereas expert humans most commonly rank GPT-4 over Mistral FT. Furthermore, we show that all summaries are shorter than their reference syllabuses and do not correlate as strongly with opinion lengths. Despite this, humans prefer GPT-4 summaries, revealing that its summaries may represent a more desirable trade-off between concision and comprehensiveness.

5.1 Automated Evaluation Favors Fine-tuned Mistral Summaries

We start by looking at the results in Table 1 of automatic evaluation between summaries and official syllabuses for the three generated summaries (Mistral Base, Mistral FT, and GPT-4) and for two human summaries. Overall, we find that fine-tuning Mistral is particularly effective at improving the recall scores across all the metrics: ROUGE recall scores increase by an average of 21 points, BERTScore recall by 15 points. However, effects of fine-tuning on precision are weaker and more mixed. Perhaps fine-tuning sacrifices brevity for inclusion of more words in a syllabus, i.e., improves the sensitivity of summaries at a cost to specificity.

Control summaries help highlight effects of style differences on automatic metrics. By comparing against the two control human-written summaries, we can clearly see that Westlaw is an outlier. While GPT-4 and Mistral FT scores mostly resemble Oyez, Westlaw's recall scores are particularly low, only surpassing Mistral base. This poor performance on recall, but strong performance on precision, may be a product of how short those summaries are.

5.2 Summaries do not Scale with Opinion Length as much as Official Syllabuses

A unique aspect of CASESUMM is that it provides coverage of SCOTUS cases that unfold over more than two centuries. This breadth enables researchers to investigate summaries from many different angles. In this subsection, we characterize candidate summaries through the lens of length and compression and explore how these variables may affect summary quality over time.

Method	Length	Compression	Correlation
Opinion	6640	-	-
Syllabus	750	0.176	0.676^{***}
GPT-4	321	0.092	0.088^{*}
Mistral Base	126	0.034	0.151^{***}
Mistral FT	447	0.121	0.179^{***}
Oyez	332	0.096	0.094^{*}
Westlaw	142	0.044	0.025

Table 2: Descriptive statistics on summaries in test set (n=622). Length is number of words. Compression rate is ratio of words in syllabus to words in opinion. Smaller number is more compression. Opinion included as reference. ($^{*}p < 0.05$ $^{**}p < 0.01$, $^{***}p < 0.001$)

Length & Compression. In our dataset, both the opinion and syllabus lengths systematically co-vary across time. Table 2 shows that syllabuses in our sample have an average compression rate of 17.6%, meaning they are typically about one-sixth the length of the original opinions. We find that in generated summaries, Mistral FT produces summaries closest in length to these syllabuses, even outperforming GPT-4, which was also prompt-optimized to mimic syllabuses. Westlaw produced the shortest summaries, followed by Mistral without fine-tuning.

Regarding the correlation between summary and opinion lengths, syllabuses demonstrate the strongest relationship with opinion lengths: doubling opinion length increases syllabus length by nearly 2/3. In contrast, Mistral FT summaries show a weaker correlation, with doubling opinion length increasing summary length by only 18%. Westlaw summaries exhibit almost no correlation with opinion length, maintaining a consistent target length of approximately 150 words. These findings highlight our dataset as a rich resource for future work in investigating how automatic summarization methods may adapt to varying source document lengths, ensuring that all salient information is captured regardless of length.

Precision & recall diverge over time. We use the Supreme Court Data Base[8], which contains metadata on SCOTUS cases, to see if any particular metadata can explain variation in summarization quality. While we do not find notable variation across most of these features, we observe one exception: the divergence between recall and precision across all summaries increases over time. Fig-

[8] We obtain data on features of cases by downloading case metadata from Washington University Law School's Supreme Court Data Base (SCDB).

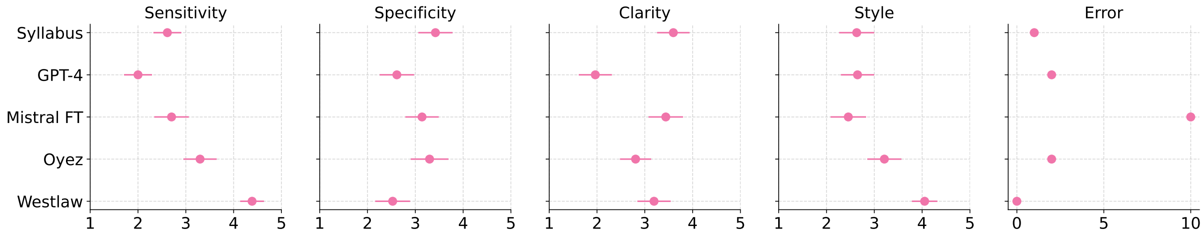

Figure 2: Human evaluation of model-generated and human summaries. x-axis is a rank, where 1 is best and 5 is worst. For **Error**, x-axis shows counts of the total number of errors identified by participants for each summary method. See Appendix C.1 for explanation of each dimension.

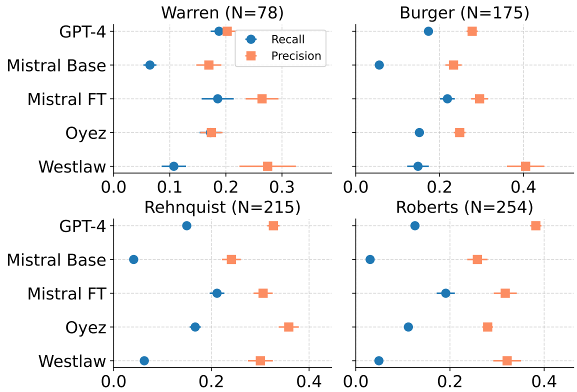

Figure 3: ROUGE-2 evaluation of model-generated and human summaries, by Chief Justice of SCOTUS when the opinion was written. Markers are means and whiskers are 95% confidence intervals.

ure 3 illustrates this trend, comparing summaries for opinions based on the Chief Justice of the Supreme Court at the time an opinion was issued. Summaries of earlier opinions, e.g., under the Warren Court, have greater parity between recall and precision compared to summaries from later opinions. One possible explanation for this trend is that opinions and syllabuses have become longer over time (Figure 1 and Table 2), while the summaries we evaluate show a growing disparity between their lengths and opinion/syllabus lengths over time.

5.3 Human Evaluation Disagrees with Automatic Evaluation

The results of our human evaluation, presented in Figure 2 are distinctly different than those of our automatic evaluation. Whereas under automatic evaluation, Mistral FT outperforms other models as well as the control human-written summaries, we find that humans most commonly prefer GPT-4 summaries. GPT-4 particularly excels on *clarity*, a crucial yet difficult to measure desideratum for the summarization task. Nonetheless, Mistral FT remains an above-average performer, successfully matching the original opinion syllabuses on every dimension except, importantly, number of errors.

Evaluators report that roughly 20% of Mistral FT summaries have at least 1 factual error, with a total of 10 errors identified across all evaluations. However, we see that these factual errors, or *hallucinations*, are not necessarily a product of using LLMs, as GPT-4 has performance on par with syllabuses and Oyez in terms of factual correctness.

Surprisingly, the human evaluation also revealed that all three human-written summaries, including the official syllabuses, often performed worse than GPT-4. Westlaw summaries, despite being a paid service designed for legal professionals, ranked below average on sensitivity, clarity, and style. Even more intriguingly, the official syllabuses only matched or under-performed the LLM-generated summaries on all metrics except, crucially, factual correctness (*error*). This result both challenges the assumption that human-written summaries are inherently superior, while also revealing opportunities and challenges in using LLMs for generating concise, correct, and accessible summaries.

5.4 Error Analysis

Mistral hallucinates more conspicuously than GPT-4. We conduct further analysis of each summary flagged as containing factual errors according to the participants in the human evaluation. We compare each such summary to the original opinion to identify specific factual errors. Recent work has often referred to errors of this type as "hallucinations" (Huang et al., 2023).

Table 3 presents example errors. Fine-tuned mistral contained the most errors in its summaries. Furthermore, these errors were more egregious than any produced in the GPT-4 Turbo summaries. These errors include simple factual errors (examples 1), incorrect citations (example 2), temporal understanding errors (example 3), as well as procedural history outcome errors (examples 4).

In contrast, GPT-4 Turbo errs in a more subtle way, failing to properly convey the legal analysis

Hallucination in Summary	Explanation
Fine-tuned Mistral	
1. "Petitioner, a Negro, applied for admission to the University of Washington Law School, a state-operated institution."	The opinion indicates the petitioner is not a member of a "favored group" nor a "minority applicant". This strongly implies the petitioner is white and rules out the petitioner to be "a negro". 416 U.S. 312, 332, 325 (1974).
2. "Sherbert v. Velelline, 416 U. S. 456"	The opinion cites "Sherbert v. Verner, 374 U.S. 398" 494 U.S. 872, 875 (1990).
3. "He was held incommunicado for some five or seven days after signing the statement."	Petitioner Haynes testified he was held incommunicado until some five or seven days after his arrest. 373 U.S. 503, 504 (1963).
4. "The District Court ultimately entered judgment for petitioner, holding that the Texas death penalty scheme was unconstitutional."	While the District Court initially stayed the execution pending judgement, it ultimately "filed its findings and conclusions, rejecting each of the several grounds asserted by petitioner. The writ was accordingly denied; also, the stay of petitioner's death sentence was vacated." 463 U.S. 880, 885 (1983).
GPT-4 Turbo	
5. "The Court rejected the State's interest in [...] preserving the flag as an unalloyed symbol of the nation"	The Supreme Court did not reject the State's interest. They "assume[d], arguendo, that it is [valid]" but found "[t]he statute is nonetheless unconstitutional as applied to appellant's activity". 418 U.S. 405, 414 (1974).
6. "The PCAOB is a regulatory body that oversees the audits of public companies"	The PCAOB was created to govern the entire industry of accounting, including "hiring and professional development, promotion, supervision of audit work, the acceptance of new business and the continuation of old, internal inspection procedures, professional ethics rules, and 'such other requirements as the Board may prescribe.'" 561 U.S. 477, 485 (2010).

Table 3: Comparison of model hallucinations and their explanations.

presented in the opinion (example 5) or misrepresenting background details (example 6). While the opinion indeed reverses the judgement of the court below, it does not reject its reasoning. Rather, the ruling is reversed due to a superseding issue of constitutionality. The summary generated by GPT-4 Turbo is thus incorrect in its characterization of the Supreme Courts decision.

Lexical variation. We define *lexical variation* as the percentage of unique words in the summary not present in the opinion and consider it a measure of summary style. Mirroring our comparison of compression rates, syllabuses are shown to exhibit the lowest percentage of lexical variation from the original opinion. Surprisingly, the fine-tuned Mistral summaries have the highest average percentage of lexical variation at 41.7%, even surpassing those written by Oyez (37.9%). This is unexpected because Mistral FT is trained on legal syllabuses, while Oyez summaries are written for a general audience and might borrow less from the opinion. The high lexical variation rate of Mistral FT may be related to its higher rate of factual errors.

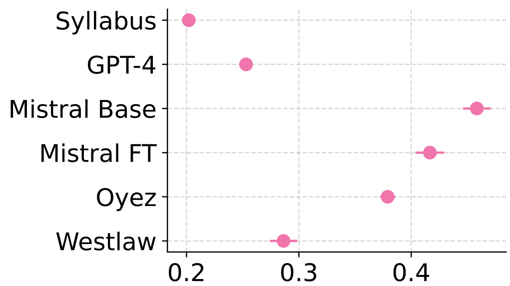

Figure 4: Lexical Variation. Measures the fraction of words in summary that are not in opinion.

6 Conclusion

This paper introduces CASESUMM, a novel dataset for long-context summarization in the legal domain, comprising 25,611 U.S. Supreme Court opinions and their official syllabuses. Our comprehensive evaluation of LLM-generated summaries, using both automatic metrics and expert human evaluation, reveals discrepancies between these assessment methods. While fine-tuned Mistral 7b outperforms larger models on automatic metrics, human experts rank GPT-4 summaries higher in clarity and accuracy. Our human evaluation also showed that GPT-4 summaries often outperformed human-written summaries, including official syllabuses and professional services, in several metrics except factual correctness. Our findings highlight the limitations of current automatic evaluation methods for legal summarization and underscore the importance of human evaluation in assessing summary quality, particularly in complex, high-stakes domains like law. This work contributes to the ongoing dialogue about evaluation methodologies in NLP and opens avenues for research in legal text summarization.

7 Limitations

Our analyses has several limitations. Our automatic evaluations use ROUGE, but ROUGE has important limitations (Conroy et al., 2011; Cohan and Goharian, 2016). We address some of these using human evaluations, but our sample size limits the conclusions we can draw. Second, while we are able to offer insight into the value of fine-tuning, at least with respect to the open-source Mistral model, we are unable to estimate the value of prompt-engineering even the GPT-4 model because we do not have a natural benchmark, non-optimized prompt for that model. A related weakness is that our evaluation of fine-tuning Mistral does not tell us the value of fine-tuning other models, such as GPT-4. It is possible that the benefit to fine-tuning the latter may be lower than the former because GPT-4 is trained on more data and has far more estimated parameters. Finally, while we demonstrate through a manual evaluation that our PDF extraction procedure is largely accurate (96%), it is not perfect. A fraction of syllabuses, particularly those extracted from low-quality scans from SCOTUS opinions in the early 1800s, may not be fully correct.

References

Emmanuel Bauer, Dominik Stammbach, Nianlong Gu, and Elliott Ash. 2023. Legal extractive summarization of us court opinions. *arXiv preprint arXiv:2305.08428*.

Manik Bhandari, Pranav Narayan Gour, Atabak Ashfaq, Pengfei Liu, and Graham Neubig. 2020. Re-evaluating evaluation in text summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9347–9359, Online. Association for Computational Linguistics.

G. Bradski. 2000. The OpenCV Library. *Dr. Dobb's Journal of Software Tools*.

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, Harsha Nori, Hamid Palangi, Marco Tulio Ribeiro, and Yi Zhang. 2023. Sparks of artificial general intelligence: Early experiments with gpt-4. *Preprint*, arXiv:2303.12712.

Tianyu Cao, Natraj Raman, Danial Dervovic, and Chenhao Tan. 2024. Characterizing multimodal long-form summarization: A case study on financial reports. *Preprint*, arXiv:2404.06162.

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

Yapei Chang, Kyle Lo, Tanya Goyal, and Mohit Iyyer. 2024. Booookscore: A systematic exploration of book-length summarization in the era of llms. *Preprint*, arXiv:2310.00785.

Claude Team. 2024. Introducing the next generation of claude.

Arman Cohan and Nazli Goharian. 2016. Revisiting summarization evaluation for scientific articles. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 806–813, Portorož, Slovenia. European Language Resources Association (ELRA).

John M. Conroy, Judith D. Schlesinger, and Dianne P. O'Leary. 2011. Squibs: Nouveau-ROUGE: A novelty metric for update summarization. *Computational Linguistics*, 37(1):1–8.

Junyun Cui, Xiaoyu Shen, and Shaochun Wen. 2023. A survey on legal judgment prediction: Datasets, metrics, models and challenges. *IEEE Access*.

Vladimir Eidelman. 2019. Billsum: A corpus for automatic summarization of us legislation. In *Proceedings of the 2nd Workshop on New Frontiers in Summarization*, page 48–56. Association for Computational Linguistics.

Alexander Fabbri, Chien-Sheng Wu, Wenhao Liu, and Caiming Xiong. 2022. QAFactEval: Improved QA-based factual consistency evaluation for summarization. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2587–2601, Seattle, United States. Association for Computational Linguistics.

Alexander R. Fabbri, Wojciech Kryściński, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir Radev. 2021. SummEval: Re-evaluating Summarization Evaluation. *Transactions of the Association for Computational Linguistics*, 9:391–409.

Tobias Falke, Leonardo F. R. Ribeiro, Prasetya Ajie Utama, Ido Dagan, and Iryna Gurevych. 2019. Ranking generated summaries by correctness: An interesting but challenging application for natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2214–2220, Florence, Italy. Association for Computational Linguistics.

Biaoyan Fang, Trevor Cohn, Timothy Baldwin, and Lea Frermann. 2023. Super-scotus: A multi-sourced dataset for the supreme court of the us. In *Proceedings of the Natural Legal Language Processing Workshop 2023*, pages 202–214.

Diego Garat and Dina Wonsever. 2022. Automatic curation of court documents: Anonymizing personal data. *Information*, 13(1):27.

GPT-4 Team. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Peter Henderson, Mark Krass, Lucia Zheng, Neel Guha, Christopher D Manning, Dan Jurafsky, and Daniel Ho. 2022. Pile of law: Learning responsible data filtering from the law and a 256gb open-source legal dataset. *Advances in Neural Information Processing Systems*, 35:29217–29234.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *Preprint*, arXiv:2106.09685.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. 2023. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *Preprint*, arXiv:2311.05232.

Luyang Huang, Shuyang Cao, Nikolaus Parulian, Heng Ji, and Lu Wang. 2021. Efficient attentions for long document summarization. *Preprint*, arXiv:2104.02112.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

Sayash Kapoor, Peter Henderon, and Arvind Narayanan. 2024. Promises and pitfalls of artificial intelligence for legal applications. *Journal of Cross-Disciplinary Research in Computational Law*.

Daniel Martin Katz, Michael James Bommarito, Shang Gao, and Pablo Arredondo. 2023a. Gpt-4 passes the bar exam. *Available at SSRN 4389233*.

Daniel Martin Katz, Dirk Hartung, Lauritz Gerlach, Abhik Jana, and Michael J Bommarito II. 2023b. Natural language processing in the legal domain. *arXiv preprint arXiv:2302.12039*.

Omar Khattab, Arnav Singhvi, Paridhi Maheshwari, Zhiyuan Zhang, Keshav Santhanam, Sri Vardhamanan, Saiful Haq, Ashutosh Sharma, Thomas T. Joshi, Hanna Moazam, Heather Miller, Matei Zaharia, and Christopher Potts. 2023. Dspy: Compiling declarative language model calls into self-improving pipelines. *Preprint*, arXiv:2310.03714.

Huan Yee Koh, Jiaxin Ju, Ming Liu, and Shirui Pan. 2022. An empirical survey on long document summarization: Datasets, models, and metrics. *ACM Computing Surveys*, 55(8):1–35.

Kalpesh Krishna, Erin Bransom, Bailey Kuehl, Mohit Iyyer, Pradeep Dasigi, Arman Cohan, and Kyle Lo. 2023. LongEval: Guidelines for human evaluation of faithfulness in long-form summarization. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1650–1669, Dubrovnik, Croatia. Association for Computational Linguistics.

Wojciech Kryściński, Nazneen Rajani, Divyansh Agarwal, Caiming Xiong, and Dragomir Radev. 2022. Booksum: A collection of datasets for long-form narrative summarization. *Preprint*, arXiv:2105.08209.

Philippe Laban, Tobias Schnabel, Paul N. Bennett, and Marti A. Hearst. 2022. SummaC: Re-Visiting NLI-based Models for Inconsistency Detection in Summarization. *Transactions of the Association for Computational Linguistics*, 10:163–177.

Washington University Law. 2024. The supreme court database.

Library of Congress. 2024. United states reports (official opinions of the u.s. supreme court).

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Masha Medvedeva, Martijn Wieling, and Michel Vols. 2023. Rethinking the field of automatic prediction of court decisions. *Artificial Intelligence and Law*, 31(1):195–212.

Public Resource Org. 2024. United states reports.

Natalie Schluter. 2017. The limits of automatic summarisation according to ROUGE. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 41–45, Valencia, Spain. Association for Computational Linguistics.

Eva Sharma, Chen Li, and Lu Wang. 2019. BIGPATENT: A large-scale dataset for abstractive and coherent summarization. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2204–2213, Florence, Italy. Association for Computational Linguistics.

Pawan Trivedi, Digha Jain, Shilpa Gite, Ketan Kotecha, Anant Bhatt, and Nithesh Naik. 2024. Indian legal corpus (ilc): A dataset for summarizing indian legal proceeding using natural language. *Engineered Science*, 27:1022.

Alex Wang, Kyunghyun Cho, and Mike Lewis. 2020. Asking and answering questions to evaluate the factual consistency of summaries. In *Proceedings of the*

58th Annual Meeting of the Association for Computational Linguistics, pages 5008–5020, Online. Association for Computational Linguistics.

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

Weizhe Yuan, Graham Neubig, and Pengfei Liu. 2021. Bartscore: Evaluating generated text as text generation. *Advances in Neural Information Processing Systems*, 34:27263–27277.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Jie Zou and Evangelos Kanoulas. 2020. Towards question-based high-recall information retrieval: Locating the last few relevant documents for technology-assisted reviews. *ACM Transactions on Information Systems (TOIS)*, 38(3):1–35.

A Fine-tuning and Generation Implementation Details

In our fine-tuning experiments, we use a batch size of 56. We select the best performing learning rate out of $\{2e-5, 2e-4, 2e-3\}$ and early stop based on dev loss convergence. We conduct our experiments on 7 A100 80GB GPUs, with each fine-tuning run taking approximately 2 hours. During summary generation, we don't use sampling and set max tokens to 1500. We truncate opinions which exceed Mistral's 32768 context-length limit. In approximately 10% of Mistral generations, the generation stops due to the length limit, rather than an `<eot>` token being generated. In such cases, we fallback to sampling a generation with a repetition penalty of 1.3 and a top_p value of 0.9. This ensures a complete summary is produced and reduces degenerated summaries from the model.

B Automatic Evaluation

B.1 ROUGE Implementation Details

We use the ROUGE implementation from the HuggingFace `evaluate` Python package. We set `use_stemmer = True` and `use_aggregator = True`.

B.2 BERTScore implementation Details

We use the `bert-score` PyPI package. We use the default `bert-base-uncased` scoring model and all other default settings.

B.3 BARTScore

See Table B.3.

Method	BARTScore ↓ P	R	F1
GPT-4 Turbo	**256.0**	<u>335.1</u>	**289.5**
Mistral Base	<u>277.0</u>	380.1	316.8
Mistral FT	297.9	**312.3**	<u>298.1</u>
Oyez	346.5	<u>334.6</u>	339.3
Westlaw	307.8	345.2	323.5

Table 4: BARTScores of model-generated and human-written summaries, where official syllabuses are the reference summaries. Sample includes 622 Supreme Court cases. There are 622 observations on each type of summary except Westlaw, for which we only have 156 observations. We report precision (P), recall (R), and F1-score (F1). BARTScores are negative log-likelihoods, so lower scores are better. We **bold** the best score(s) and <u>underline</u> the second best score(s). For the scoring model, we use `facebook/bart-large-cnn`, the default model used in Yuan et al. (2021)

C Human Evaluation

C.1 Dimensions of Summary Quality

Sensitivity: Does this summary include all relevant information required to understand the facts, judgment and reasoning? Outcome is a rank, where 1 is best, rank 5 is worst. Ranks are mutually exclusive: only one case per rank.

Specificity: Does this summary exclude irrelevant information that is not required to understand the facts, judgment and reasoning? Rank from 1 to 5.

Clarity: Is this summary clear and easy to read? Rank from 1 to 5.

Style: Does this summary have a legal style, defined as something written by a well-trained lawyer? Rank from 1 to 5. For all measures where the outcome is rank, we mark the mean rank identically 3) with a red dashed line.

Factuality: Does this summary contain any factual errors? (Yes/No).

C.2 Annotation Interface

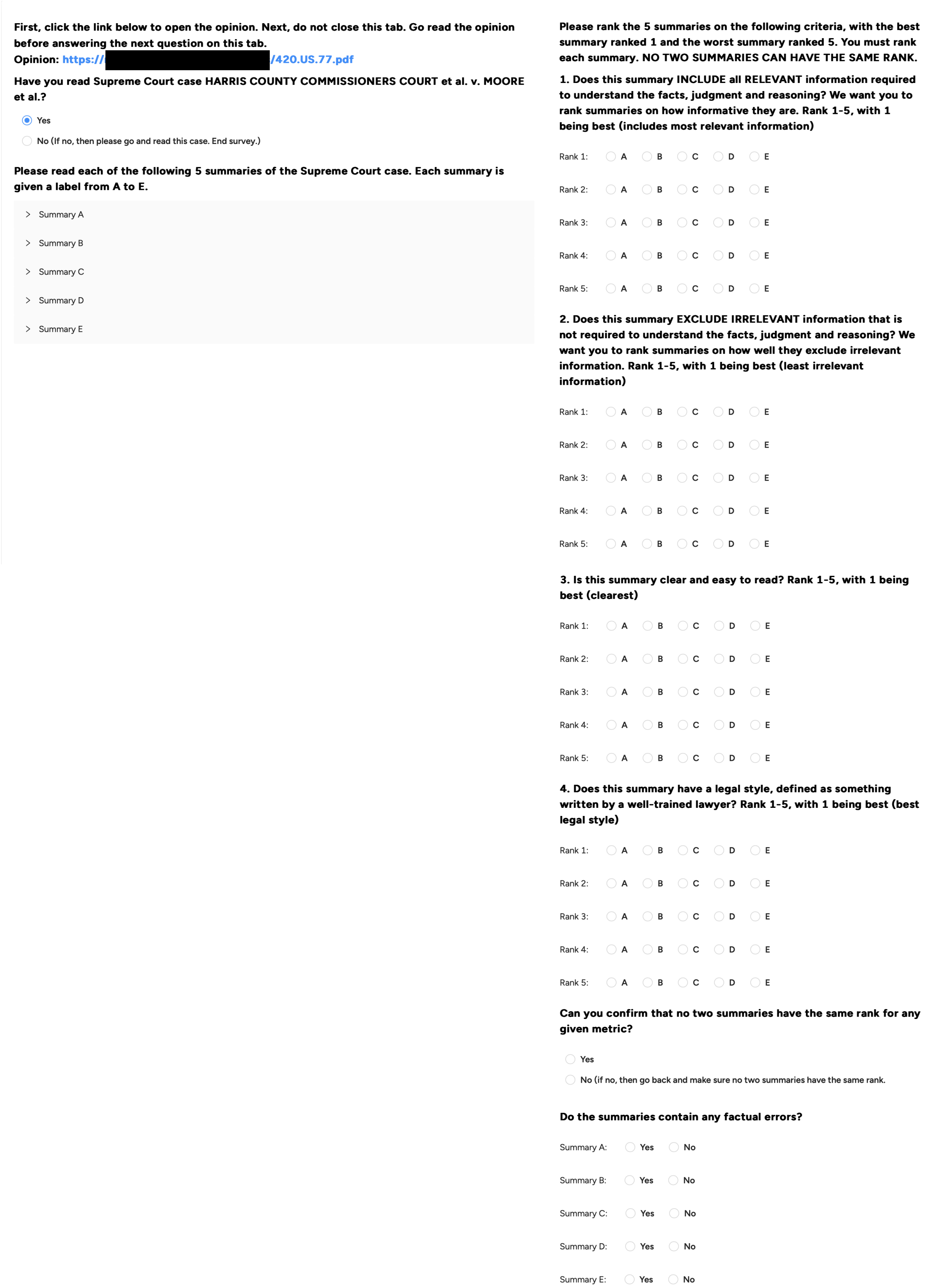

Figure 5: Labelstudio Annotation Interface

C.3 Instruction & Consent Materials for Participants

Online Consent Form for Research Participation

Study Title: Investigating the quality of automatic legal case summarization
Researcher(s): [redacted]

Description: We are researchers at [redacted] doing a research study that uses artificial intelligence (large-language models or LLMs) to summarize US legal cases. The purpose of this research is to develop tools to facilitate legal research. One step in this research is to assess the quality of these summaries. We are recruiting research assistants to do that. We will ask research assistants to (a) read a Supreme Court case that we have summarized, read 5 summaries of this cases, and (c) rank the 5 summaries on accuracy, clarity, and legal style. Together these 3 steps are called a "case summary analysis". We will not ask any personal questions about you, except your name, so that we can pay you for participating. Each case summary analysis will take roughly 45 minutes. We may do as many cases analyses as you like, though we request you complete a minimum of 4. You are eligible to be a research assistant on this study if you are a 2L or 3L law student at [redacted] Law School. Your participation is voluntary.

Incentives: We will pay you $15 per case summary, which equals $20/hour because each case summary analysis should take 45 minutes. You will be paid per case summary analysis completed. You will not be paid for partial case analyses as these are not usable for the research study.

Risks and Benefits: Your participation in this study does not involve any risk to you beyond that of everyday life.

Confidentiality: We will maintain a file that includes your name and a randomly generated ID number. You will use that ID number to log into a server that contains your case assignment, the 5 summaries associated with that case, and a web-form for ranking the 5 summaries. Each completed case analyses and your associated ID will be shared with Professor [redacted] who will connect your ID to your name for the purpose of paying you for the "case summary analysis". The data connecting your ID to your name will be maintained on the secure, encrypted [redacted] Box server. We will not use any identifiable data from you for the research analysis. Such data is only used for purposes of paying you compensation for your research assistance. We will not use data or identifiers from any case analysis summaries that are incomplete. You may stop working as a research assistant at any time. If you stop working as a research assistant, data collected up until the point of withdrawal may still be included in the analysis. No identifiable data will be used in future research. De-identified data may be used in future research and shared with other researchers for future research without additional informed consent.

Contacts & Questions:
If you have questions or concerns about the study, you can contact the researchers by reaching out to [redacted], [redacted].

If you have any questions about your rights as a participant in this research, feel you have been harmed, or wish to discuss other study-related concerns with someone who is not part of the research team, you can contact the [redacted]

Institutional Review Board (IRB) Office by phone at [redacted] or by email at [redacted].

Consent:
Participation is voluntary. Refusal to participate or withdrawing from the research will involve no penalty or loss of benefits to which you might otherwise be entitled.

By clicking "Agree" below, you confirm that you have read the consent form, are at least 18 years old, and agree to participate in the research. Please print or save a copy of this page for your records.

☐ I agree to participate in the research
☐ I do NOT agree to participate in the research

Dear X,

Thanks for helping with our AI legal summarization project! As a reminder, you are going to evaluate the quality of the summaries produced by our AI-based tool for summarizing Supreme Court decisions. We hope that you are able to begin this work as soon as possible, ideally even today. Moreover, we request that you complete at least 4 evaluations, though we would be very happy for you to complete as many as you can. Remember, we pay $20/hour: because we expect each case evaluation to take you 45 minutes, this means we will pay $15 for every case evaluation you complete.

There are 3 steps to begin work as an RA.

1/ Please read and answer the attached consent form. While you are an RA, computer science projects have a norm of registering human evaluations with the IRB. Please email me back your completed consent form.

2/ Once you do that, I will send you a link to a custom survey that will give you a case, some summaries, and ask some questions. Once you complete one survey, you will be given a another one. Again, please try to complete at least 4 surveys. However, you may complete more.

3/ Once you are done, do let us know. The system will tabulate how many forms you have completed, and I will have the law school issue an RA payment to you. (The law school may need some additional paperwork, but it should not be onerous. Just the usual RA paperwork.)

If you have any questions, about the project, the consent form, or the survey, do reach out. The best way to reach me is via email at [redacted].

Figure 7: Email with instructions sent to participants..