Adapting Abstractive Summarization to Court Examinations in a Zero-Shot Setting: A Short Technical Paper

Maya Epps[1], Lucille Njoo[2], Chéla Willey[1] and Andrew Forney[1]

[1]*Loyola Marymount University, 1 LMU Dr., Los Angeles, CA, 90045, USA.*
[2]*University of Washington, 1900 Commerce Street, Tacoma, WA, 98402, USA.*

Abstract

Automated summarization of court trial transcripts can enable lawyers to review and understand cases much more efficiently, but it is challenging for pre-trained large language models (LLMs) in zero-shot settings due to the uniqueness and noisiness of legal dialogue. This is further complicated by the high-stakes of errors, which can mislead readers in a domain where factuality and impartiality are paramount. In this short technical paper, we apply summarization methods to this new domain and experiment with manipulating the transcript text to reduce model errors and generate higher-quality summaries. With human evaluations of metrics like factuality and completeness, we find that zero-shot summarization of trial transcripts is possible with preprocessing, but it remains a challenging task. We observe several open problems in summarizing court dialogue and discuss future directions for addressing them.

Keywords

summarization, court transcripts, dialogue preprocessing

1. Introduction

Transcripts of court trials can be lengthy, sometimes spanning thousands of pages, making them time-consuming and mentally taxing to read in-full. Lawyers whose work centers around review of these transcripts thus face challenges of understanding, retaining, and finding details nested in court dialogue that may have occurred in their distant past or that comes from other attorneys. As collaborators on the present endeavor, lawyers at the Innocence Project (IP) [1] must read through many such transcripts as part of their work to exonerate convicts who have been wrongfully incarcerated. The IP has a rapidly growing queue of clients waiting to have their cases reviewed for evidence of a mistrial and other mitigating factors, but the IP's limited staff are unable to keep up due to the time and effort each lengthy transcript requires.

In this work, we explore how language technologies can be used to *automatically summarize examinations* in trial transcripts in order to provide lawyers with a concise overview of important points. Summaries that are *factually accurate and preserve relevant details* could enable lawyers to review transcripts more efficiently and holistically, significantly accelerating their trial review process and enabling the IP to serve more clients. The IP's social justice work is one example of a high-impact humanitarian effort that would benefit from summarization tools, but this would also be useful to other stakeholders who process long cases, such as litigators and law students.

Summarization of many types of text has been made possible by recent advancements in natural language processing (NLP), particularly the rise of large language models (LLMs): neural models pretrained on vast amounts of text [2, 3]. Previous studies have endeavored to summarize legal text using both LLMs and others in several settings, including abstractive summaries to make legal jargon approachable to laypeople [4], summarizing case outcomes [5], and performing information extraction from legal texts [6]. However, summarization has not yet been applied to the domain of individual examinations in trial transcripts, and doing so presents technical challenges that the current introductory work hopes to explore.

Though LLMs are very powerful, most of their training data comes from the Web and does not resemble the language, cadence, and procedural nature of dialogue spoken in court. Additionally, we only have access to a limited number of raw transcripts and do not have gold standard summarization examples with which to finetune a model for this new domain. Thus, we focus on summarization in a *zero-shot* setting: adapting existing LLMs to trial transcripts to generate helpful summaries without additional training. In doing so, we experiment with different ways of manipulating the transcript text to make them sound more natural and understandable to pretrained LLMs.

Summarization in this domain is also challenging because of the unique characteristics of trial transcripts. [7]

Workshop on Artificial Intelligence for Access to Justice (AI4AJ 2023), June 19, 2023, Braga, Portugal

mepps@lion.lmu.edu (M. Epps); lnjoo@cs.washington.edu (L. Njoo); chela.willey@lmu.edu (C. Willey); andrew.forney@lmu.edu (A. Forney)

Not only is legal discourse linguistically different from text scraped from the Web, but trial transcripts also carry all the nuances and noisiness of *spoken* dialogue, and they are furthermore formatted in ways that may seem unnatural to LLMs. Such out-of-domain inputs can exacerbate language generation problems like factuality errors and social biases. In such a high-stakes domain, tools with errors can be more harmful than helpful, such as by causing readers to miss important details or influencing their interpretation of the actual text. Because of the gravity of these potential errors, we rely not only on automatic metrics like perplexity, but also on manual human evaluation to judge whether generated summaries are truthful and relevant.

This short paper shares some empirical findings in pursuit of addressing the above, and specifically contributes the following:

- Assesses the out-of-box performance of a popular LLM dialogue summarizer on a selection of real court transcript examinations.
- Provides human-labeled evaluations of summarizer outputs on measures of factuality, completeness, and overall quality.
- Reports on the effects of several dialogue preprocessing techniques on these metrics.
- Shares qualitative insights on the summaries that may pave the way for future explorations.

Although zero-shot summarization of longform documents remains an open challenge, we show that factual, complete, and helpful summarization of court examinations is *possible* with appropriate preprocessing techniques that manipulate rigidly formatted trial transcripts to sound more like natural language.

2. Background and Related Work

Trial Transcripts. Trial transcripts in United States courts follow a consistent high-level structure, though the text formatting often varies across cases. In general, transcripts primarily consist of dialogue, typically written in all capital letters as a speaker's name followed by their spoken line, interspersed with descriptive text. Much of this dialogue is comprised of *examinations*, where a witness is called to the stand and interrogated by a prosecution or defense lawyer. Examinations' formatting switches to a Q/A pattern: rather than referring to the examiner and witness by name, they are instead introduced at the beginning of the examination and subsequently referred to as *Q* and *A* respectively. These examinations can be of any length—from a few sentences to several dozen pages—and are the portions of dialogue that we aim to summarize.

Challenges of NLP in High-Stakes Real-World Domains. LLMs pretrained on vast amounts of Web data have been used to analyze and generate text in a variety of high-stakes domains [2]. However, it remains a challenge to apply language technologies to real-world settings that are often very noisy and may differ from the data the models were trained on. In the absence of readily available training data for new domains, prior works have experimented with modifying text inputs to optimize *zero-shot* model performance without additional training [8]. For example, prompt tuning has emerged as a popular way to improve model outputs for a wide variety of tasks [9]. However, these works focus on manipulating relatively short prompts, whereas we experiment with high-level text patterns to make longform court dialogue more understandable to models. Aside from the difficulties of handling out-of-domain text, text generated by LLMs is prone to problems like social biases, where models perpetuate stereotypes about gender, race, or other aspects of identity [10], and factuality errors, where models hallucinate false information [11]. Our results demonstrate these common pitfalls, and we explore how preprocessing can be used to minimize them and discuss avenues for future work.

Summarization in NLP. The goal of *summarization* is to distill the most important information from long passages of text. With the rise of neural language models, summarization models have shifted from extractive (identifying important sentences in the original text) to abstractive (generating the summary from scratch) and have made extraordinary performance improvements in summarizing documents ranging from news articles [12] to novels [13]. Most prior work in summarization has focused on model design and training, but our work is a zero-shot setting and particularly focuses on dialogue. *Dialogue* adds new challenges to summarization because, unlike text written by a single author, it involves multiple participants, frequent coreferences, and a less structured discussion flow, with some related recent work summarizing written dialogues like chats and email threads [14, 15]. However, many datasets and benchmarks for summarization are constructed in artificial settings: for example, the SAMSum Corpus contains abstractive summaries of chats between linguists who were aiming to emulate conversations in a messenger app [16]. Spoken conversations in the real world are studied much more sparsely and are even noisier, but a small number of recent works have begun to explore it [17]. Our work builds on this by attempting to apply summarization methods to spoken dialogue in US courts.

3. Method

3.1. Data

The IP lawyers collaborating on this project furnished 5 trial transcripts from which 59 examinations were extracted. The transcripts were provided as scanned PDFs from court proceedings. For each transcript, we use the Google Tesseract library to perform Optical Character Recognition (OCR) and recreate the lines of the transcript as plain-text. The beginnings and ends of examinations are clearly marked on trial transcripts due to a standardized format of court transcripts. Examinations ranged in length from 42 to 6511 words ($M = 1563, SD = 1369$).

3.1.1. Sanitization.

Because of small imperfections in the OCR plain-text conversion, we first sanitized the data by fixing any mistakes manually, including the addition of multiple spaces or newlines where inconsistent. We also removed most procedural text that was secondary to the examination dialogue, typically found following an examiner's statement of "nothing further" or "no further questions" and which dealt only in court logistics like taking recesses.

3.1.2. Preprocessing.

Preprocessing techniques were applied as interventions on the sanitized data and serve as the chief independent variables in this study. We hypothesized that transforming the unique structure of trial transcript dialogue into a format more akin to the language that LLMs tend to be trained on could lead to improvements in summarization clarity. In particular, our compared conditions included:

- *Control.* Nothing about the examination was changed before it was summarized; any Q/A tags remained as is, and each speaker's dialogue ended with a newline.
- *Speaker.* In an effort to give the summarizer more information about the speaker, we replaced the Q/A tags with the participant's role in the examination—"The Examiner" or "The Witness" respectively— resulting in a format of "<Role>: <Their dialogue>". (Occasionally, other speakers may interject during the back-and-forth between the examiner/Q and the witness/A; we left those speakers as is. We also omitted any initial parenthesized text stating the examiner's name, which sometimes appeared before their first spoken line.)
- *No quote.* Since the LLM we used was finetuned on news articles (see section 3.2), we attempted to preprocess the examinations to mimic quotes in news articles. We once again replaced Q/A tags with roles ("The Examiner" or "The Witness"), but this time, for all speakers, we added the word "says" between the speaker and their dialogue, resulting in a format of "<Role or Name> says <Their dialogue>". The preprocessed lines were concatenated together without newlines into a long paragraph. (Again, we omitted any initial parenthesized text stating the examiner's name.)
- *Quote.* This condition was identical to the "No quote" preprocessing above, except that we enclosed all spoken dialogue in quotation marks, resulting in a format of "<Name> says "<Their dialogue>"". We wanted to see whether the summarizer would understand speech better when it was enclosed in quotations, as is commonly seen in books and articles, which comprise much of LLMs' training data.

Many of the trial transcripts that were furnished were entirely uppercased. Because LMs account for casing when tokenizing text, they treat uppercased tokens as separate tokens from the lowercased versions. LMs tend to see much more lowercased text in their training data, so summarizers tend to do better on lowercased than uppercased text. For all interventions except the control, we lowercased all examinations that were not already truecased before applying any preprocessing techniques.

3.2. Procedure

We used a large version of BART fine-tuned on CNN data[1] as the primary summarizer model for evaluation [3]. We chose to use a model in the BART family because of their popularity and ubiquity on natural language generation tasks, and this particular fine-tuned model is one of the most widely used for the task of summarization. As this model is already fine-tuned for summarization, we did not engineer any prompt to accompany the text passed in from an examination. Other models exist that have previously performed well on summarization, which we briefly compare: T5[2] [18] and BART[3] [3], both finetuned on the SAMSum corpus. However, there did not seem to be drastic differences between the summaries and perplexities of the BART-CNN model and others, so we chose to focus primarily only BART-CNN for this paper and the effects of differing preprocessing techniques. We leave experiments with additional models for future work.

Setting summary lengths. For examinations shorter than twice the model's maximum output summary length of 142 tokens, the maximum summary length was set to half the length of the examination and the minimum summary length was set to a quarter of the length of

[1] "facebook/bart-large-cnn" on HuggingFace
[2] "philschmid/flan-t5-base-samsum" on HuggingFace
[3] "philschmid/bart-large-cnn-samsum" on HuggingFace

the examination to prevent the generation of summaries that were of a similar length or longer than the examinations themselves. For examinations longer than the summarizer's 1024 token input maximum, the examination was split into "chunks" just below the summarizer's maximum input length without splitting a sentence. The very last "chunk" of text was prefixed with text from the previous chunk to provide context for short inputs and prevent summaries that were longer than their inputs. Each chunk was then summarized individually and concatenated together. [4]

For the particularly long examinations, this "chunking" method resulted in very long summaries, so any summaries over 400 tokens in length were repeatedly re-summarized until they were under 400 tokens. This was not common, and when it was necessary it almost always only took one re-summarization. Pursuant to our goals with these summaries, we hoped this would produce summaries that were brief enough to provide a quick overview of the examination's content that a lawyer could read quickly.

Generating and evaluating summaries. For each extracted examination under each preprocessing condition, the summarizer was applied with the above constraints on summary length. We compiled all generated summaries, and each examination along with its 4 summaries was assigned to two human judges. The human judges were asked to rate summaries based on the metrics described in the following section.

3.3. Analyses

Summaries produced in each of the control and preprocessing conditions were assessed using the following metrics and comparative statistical tests.

3.3.1. Metrics.

Two standard, objective, automatically-generated descriptive metrics were recorded for each summary:

- *Perplexity,* assessed first comparing the summaries of the BART-CNN model with the perplexity computation from GPT-2 [19] using a sliding window technique with a stride of 512 tokens, and again using the perplexity computed from each summarizer variant (i.e., BART-CNN, BART-CNN-SAMSum [abbreviated to BART-SAMSum], and T5) [20]. Perplexity is typically used to evaluate language models, but it can also be used to get an idea for the quality of generated text by quantifying how "confused" a typical LLM would be about the text.
- *Lexical Overlap,* assessed by finding the lexical overlap between the summary and the top 20% most frequently occurring tokens (excluding stopwords) in each examination. We report this as a ratio of words that were retained in the summary over the number of frequently occurring tokens. In principle, this metric could assess the balance the summarizer struck between being abstractive vs. extractive, as well as how true the summarizer stays to the examination's language and most common discussion points.

Central to validation of summarizers in the domain of court transcript review, we also examined several aspects of summary quality that required human examination:

- *Factuality,* a Boolean assessment of whether or not all of the summary's stated accounts of the examination are faithful to the original text. If even a single statement, attribution, name, or pronoun ran counter to fact, that summary was not considered factual.
- *Completeness,* a Boolean assessment of whether or not the summary mentioned all of the important events in the examination. If even a single essential detail of the examination was omitted, that summary was considered incomplete.
- *Overall quality,* a Boolean assessment of whether or not the summary was interpretable enough to obtain a gist for the examination. It was possible for a summary to be factual and complete, but e.g., discuss additional non-sequiturs or arrange the sentence structure poorly so that meaning was obscured, and would thus be perceived as poor quality.

For each summary generated from the examinations, two human judges provided their subjective assessment on the three metrics above. They were asked to first read the unsummarized examination in full and then read/rate each summary created from it so that the examination's details would be fresh in-mind.

3.3.2. Statistical Tests.

Because the same examination was used as input to each of the summary conditions, we performed a 4-way repeated measures ANOVA for each of the dependent variables (Perplexity, Lexical Overlap, Factuality, Completeness, and Overall Quality) to detect differences between groups and performed Bonferroni correction for multiple comparisons (p_{crit} = .008). For the metrics from human judges (Factuality, Completeness, and Overall Quality), we first converted Boolean answers of True/False and

[4] The tokenizer used for computing examination lengths was loaded from HuggingFace's "facebook/bart-base" to match the tokenizer used by the summarizer model. To determine the length of summaries, we used SpaCy's tokenizer.

Good/Not Good to 1/0, respectively, and then took the average rating for each summary. To examine the degree to which subjective interpretation of the summaries affected perceptions of quality, we also computed Cohen's Kappa (κ) as the standard metric of interrater reliability, which describes the proportion of agreement between raters above and beyond chance [21].

4. Results

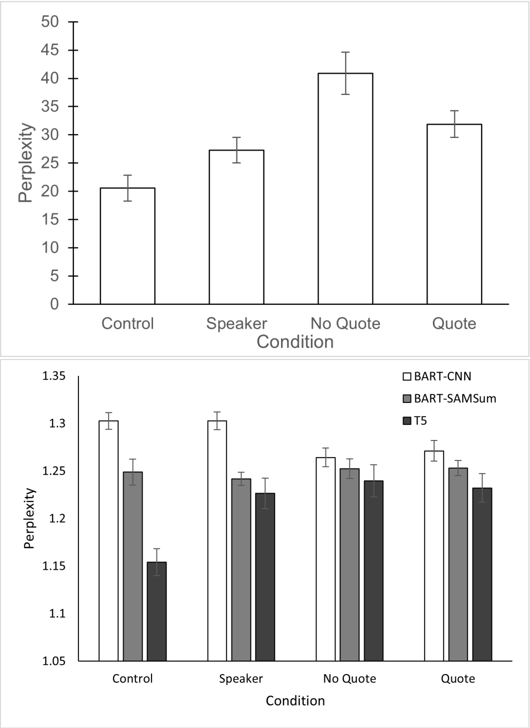

Figure 1: Perplexity compared between 4 preprocessing conditions for (Top) the BART-CNN model with perplexity computed using GPT-2 and (Bottom) the summarizer model variants with perplexity computed against themselves. Error bars represent standard error about the mean.

Perplexity (BART-CNN using GPT-2 Perplexity). There were significant differences in perplexity scores between conditions, $F(3, 174) = 24.96, p < .001, \eta_p^2 = .30$. After Bonferroni correction, all conditions were significantly different from one another, $p < .001$ (see Fig. 1, Top).

Perplexity (Summarizer Variant Comparison). One-way ANOVAs were conducted on each of the three models to compare across conditions. Within BART-CNN, there was a main effect of condition, $F(3, 174) = 5.89, p < .001, \eta_p^2 = .09$. Specifically, the no quote condition had significantly lower perplexity scores than both the control and speaker conditions. The quote condition was also significantly lower than the speaker condition. Within in BART-SAMSum, there were no significant differences between condition in perplexity scores$F(3, 174) = 0.42, p = .736$. Within T5, there was a significant main effect of condition in perplexity scores $F(3, 174) = 8.36, p < .001, \eta_p^2 = .13$. Specifically, the control condition had significantly lower perplexity scores than all other conditions (see Fig. 1, Bottom).

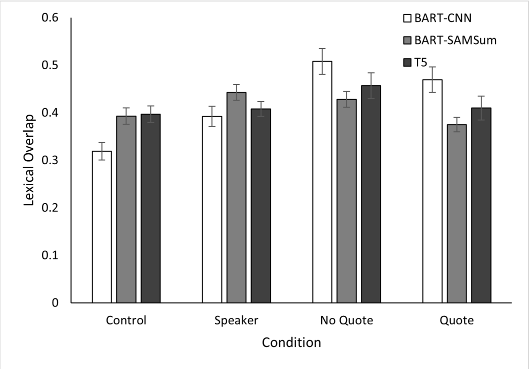

Figure 2: Lexical overlap compared between 4 preprocessing conditions for the BART-CNN model versus variants of BART-SAMSum and T5. Error bars represent standard error about the mean.

Lexical Overlap. There were significant differences in lexical overlap between preprocessing conditions within the BART-CNN model, $F(3, 174) = 25.65, p < .001, \eta_p^2 = .31$. After Bonferroni correction, all, but one comparison, were significantly different from one another, $p < .001$, (see Fig. 2). The difference in lexical overlap between the No Quote and the Quote condition was not significant, $p = .018$.

A 3 (Summarizer Variant) x 4 (Condition) ANOVA showed no significant main effect of summarizer variant in lexical overlap, $F(2, 116) = 0.42, p = .66$. However, there was an interaction effect between summarizer variant and condition such that the difference between models was greatest in the control condition and the speaker condition, $F(6, 348) = 12.73, p < .001, \eta_p^2 = .18$. Specifically within T5 model, the no quote condition had significantly higher lexical overlap compared to all other conditions, though this was no significantly different after Bonferroni corrections. Additionally, in the BART-CNN model, all comparisons were shown to mirror the effects described previously. However, again due to the number of comparisons after Bonferroni corrections, none of these effects would be significant in this particular analysis.

The remaining results examine the subjective rater scores on the BART-CNN summaries alone.

Table 1 provides the calculated Cohen's Kappa for each of the three ratings described previously across the two

	Factual	**Complete**	**Quality**
Control	.688 (< .001)	.309 (.017)	.522 (< .001)
Speaker	.361 (.002)	.216 (.081)	.316 (.014)
No Quote	.535 (< .001)	.157 (.207)	.302 (.010)
Quote	.187 (.148)	.176 (.174)	.256 (.049)

Table 1
Interrater reliability (Cohen's κ (sig.)) of summary ratings from two human judges on dependent measures of Factuality, Completeness, and Overall Quality, by condition.

independent reviewers. The following statistical analyses were conducted using the average rating of the reviewers for each condition.

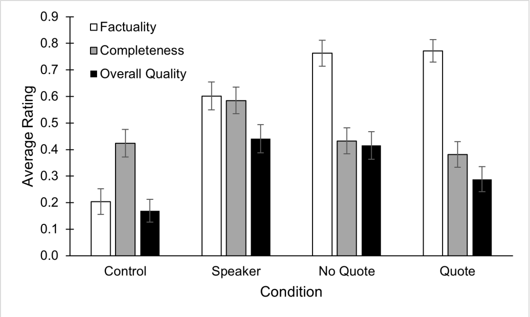

Figure 3: Averages of human ratings on Factuality, Completeness, and Overall Quality of summaries from the BART-CNN model in each of the 4 conditions. Error bars represent standard error about the mean.

Factuality. There were significant differences in factuality ratings across conditions, $F(3, 174) = 46.89, p < .001, \eta_p^2 = .45$. After Bonferroni correction, all but two comparisons were significantly different from one another, $p < .002$, (see Fig. 3). Specifically, factuality ratings in the speaker condition were only marginally lower than the no quote condition, $p = .009$. Additionally, there were no significant differences in factuality ratings between the no quote and the quote conditions, $p = .874$.

Completeness ratings. There were significant differences in completeness ratings across conditions, $F(3, 174) = 3.69, p = .13, \eta_p^2 = .060$. After Bonferroni correction, only two comparisons demonstrated significant differences. Specifically, the speaker condition had significantly higher completeness ratings than the quote ($p = .003$) and the no quote ($p = .008$) conditions.

Overall Quality ratings. There were significant differences in overall quality ratings across conditions, $F(3, 174) = 7.88, p < .001, \eta_p^2 = .120$. After Bonferroni correction, only two comparisons demonstrated significant differences. Specifically, the control condition had significantly lower overall quality ratings than the speaker ($p < .001$) and the no quote ($p = .001$) conditions.

Qualitative Reports. Although lacking by way of an objective report, we discovered several themes in summary quality that bear mentioning, and may be of use for future studies.

Exemplar Summary. Many summaries provided excellent synopses of the dialogue's contents, including the following that condensed an examination that was 590 words:

> "The witness is a senior criminalist with the orange county sheriff's crime lab. The witness is asked to examine a knife found at the scene of a murder. The knife is a buck-style knife with a brown plastic piece on either side of it. The Witness says he did not find any trace elements of blood or bodily fluids."

However, although the above summary accurately depicts the contents, it does misrepresent the gender of the witness, leading to a pervasive mistake:

Gender Bias. Through qualitatively studying generated summaries, we observed an explicit male-gender bias: many summaries defaulted to assuming actors were men rather than women, even when the original examination text was explicit in referring to an actor with feminine titles like "ma'am." This asymmetrical representation of men and women is not a novel phenomenon; gender bias has been well-documented in many LLMs [10].

Repetition. Sharing a snippet from a summary that was marked as factually accurate and complete, the output still lacks some readability due to repetition of actor nouns:

> "The Witness says he has known the boy since he was in his mother's womb. He says he knows the boy because he knows his family. The Witness says the boy is not in a gang. The witness says he's never heard of the boy being a gang member. The witness says he knows the victim from church. He says the victim is not in a gang."

Hallucinations. Hallucinations that obviously misrepresent the examination content are arguably of less concern for users because they are more likely to be caught by readers compared to subtle perturbations of court facts. The following examples demonstrate the absurdity of such dramatic hallucinations:

> "A man was shot in the head by a colleague in a New York City office. The shot was fired by a member of the jury in the trial. The gunman was standing in the same position as the shooter. A man

> was taken to jail for a photo shoot. He saw a photo of a man he thought looked like him."

Some hallucinations also demonstrate sensitivities to the fine-tuning training set and the effects of hyper-compression from re-summarizing long examinations, with the following example mentioning a commonly-referenced figure in the contemporary news who was plainly not a party to the case being summarized:

> "A fight broke out between Edward Snowden and a group of friends after he tried to leave the house."

Other hallucinations are almost understandable consequences of the quirks of spoken language, herein producing a summary mentioning two characters with nicknames "Rock" and "Blue Dog:"[5]

> "The court asks the witness if he or she has ever made an arrest of a dog. The witness tells the court he has never seen a dog in his life. The court asks if the witness has ever seen a rock in his or her life. He says he has, but he doesn't know if it was a dog or a rock."

As Figure 3 demonstrates, there is much room to improve these summaries, and repairing the qualitative issues above may likewise improve factuality, completeness, and perceived overall quality.

5. Discussion

Effect of Preprocessing. Factuality errors were present to some degree in all four preprocessing conditions, but all forms of preprocessing helped to improve factuality, completeness, and overall summary quality over the control. This is likely because preprocessed examinations more closely resembled the text that BART was trained on, and suggests that manipulating the input text may be a way to boost summarization quality. However, there seems to be a tradeoff between factuality and completeness: the Quote and No Quote conditions' propensity to produce more extractive than abstractive summaries led to improved ratings of factuality, but suffered in terms of completeness compared to the Speaker condition.

Challenges with Evaluation Metrics. Measuring summarization quality is challenging because neither quantitative nor qualitative metrics are perfect, and they sometimes contradict each other. Although the preprocessing conditions significantly increased the perplexity compared to the control, these approaches led to significant improvements in factuality, completeness, and overall quality, showing that perplexity is not necessarily a reflection of summary quality. Additionally, interrater reliability was fairly uniform in condemning the quality of the control condition's summaries, but was surprisingly lower across preprocessing conditions. This highlights yet another difficulty of assessment for summaries in the court dialogue domain: subjective disagreements over what constitutes good summaries and/or omission of key details.

Limitations and Future Directions. This study's human raters were not lawyers, who may have had feedback on the subjective measures and better expertise on how helpful a summary would be in practice. Future work should iterate with lawyers to develop more fine-grained criteria for what makes a summary "good" or "bad," and by providing continuous, rather than binary, ratings of success; e.g., determining how *many* important facts were omitted rather than whether or not *any* were.

Additionally, we only performed subjective rater comparisons on summaries from one model, BART-CNN. Though we briefly tried out other models, including T5 and a BART model fine-tuned on the SAMSum corpus, we found only minor perplexity differences and little tangibly different in summary outputs; however, experimenting with models with different architectures and training datasets could improve zero-shot summarization performance, especially with more modern generative models. In the big picture, this work demonstrates that while LLMs are powerful, they may not be able to keep track of facts reliably. This motivates work on NLP approaches that can store information in a more consistent and interpretable way than black-box LLMs, such as with maintaining state graphs and more recent chain-of-thought techniques [22].

Lastly, this work may expand avenues for novel application of court dialogue summary, including: as a learning tool for law-students to either evaluate or produce summaries, as an avenue for increasing public literacy of court proceedings by providing summaries stripped of legal procedure, and as a possible novel benchmark for domain-specific LLM adaptations in preserving the factuality and completeness of summarized text.

6. Conclusion

Our empirical results suggest that automated summarization of raw legal examinations yields poor quality summaries, but that this can be improved by preprocessing the court dialogue to better resemble the natural language that LLMs were pretrained on. These approaches still leave large gaps in the factuality and completeness of

[5] Note: this summary comes from an output lacking any preprocessing; in each of the preprocessing conditions, the nickname ambiguity was avoided.

summaries, and their perceived quality is volatile. Nevertheless, this work may serve as a motivating recipe for manipulating court examinations to achieve reasonable summarizations in a zero-shot setting, an approach that may be practical due to the domain's sparsity of finetuning data and could potentially make lengthy transcripts easier for lawyers to review.

Acknowledgments

We would like to extend a special thanks to Michael Petersen (a lawyer working with the Loyola Law School's Project for the Innocent) for his feedback and guidance on this project, as well as the efforts of several research assistants working on the Briefcase[6] team who aided with the subjective summary quality metrics: Saad Salman, Tanya Nobal, Jennifer Siao, and Evan Sciancelapore.

References

[1] I. Project, About, 2023. URL: https://innocenceproject.org/about/.

[2] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, D. Amodei, Language models are few-shot learners, 2020. URL: https://arxiv.org/abs/2005.14165. doi:10.48550/ARXIV.2005.14165.

[3] M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, L. Zettlemoyer, BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension, CoRR abs/1910.13461 (2019). URL: http://arxiv.org/abs/1910.13461. arXiv:1910.13461.

[4] O. Salaün, A. Troussel, S. Longhais, H. Westermann, P. Langlais, K. Benyekhlef, Conditional abstractive summarization of court decisions for laymen and insights from human evaluation, in: Legal Knowledge and Information Systems, IOS Press, 2022, pp. 123–132.

[5] H. Xu, J. Savelka, K. D. Ashley, Toward summarizing case decisions via extracting argument issues, reasons, and conclusions, in: Proceedings of the eighteenth international conference on artificial intelligence and law, 2021, pp. 250–254.

[6] C. Uyttendaele, M.-F. Moens, J. Dumortier, Salomon: automatic abstracting of legal cases for effective access to court decisions, AI & L. 6 (1998) 59.

[7] D. Jain, M. D. Borah, A. Biswas, Summarization of legal documents: Where are we now and the way forward, Computer Science Review 40 (2021) 100388.

[8] A. Schofield, M. Magnusson, L. Thompson, D. Mimno, Pre-processing for latent dirichlet allocation, 2017.

[9] Y. Yao, B. Dong, A. Zhang, Z. Zhang, R. Xie, Z. Liu, L. Lin, M. Sun, J. Wang, Prompt tuning for discriminative pre-trained language models, 2022. URL: https://arxiv.org/abs/2205.11166. doi:10.48550/ARXIV.2205.11166.

[10] S. L. Blodgett, S. Barocas, H. Daumé, H. Wallach, Language (technology) is power: A critical survey of "bias" in nlp, 2020. URL: https://arxiv.org/abs/2005.14050. doi:10.48550/ARXIV.2005.14050.

[11] Z. Ji, N. Lee, R. Frieske, T. Yu, D. Su, Y. Xu, E. Ishii, Y. Bang, A. Madotto, P. Fung, Survey of hallucination in natural language generation, ACM Comput. Surv. (2022). URL: https://doi.org/10.1145/3571730. doi:10.1145/3571730, just Accepted.

[12] H. Lin, V. Ng, Abstractive summarization: A survey of the state of the art, Proceedings of the AAAI Conference on Artificial Intelligence 33 (2019) 9815–9822. URL: https://ojs.aaai.org/index.php/AAAI/article/view/5056. doi:10.1609/aaai.v33i01.33019815.

[13] J. Wu, L. Ouyang, D. M. Ziegler, N. Stiennon, R. Lowe, J. Leike, P. Christiano, Recursively summarizing books with human feedback, 2021. URL: https://arxiv.org/abs/2109.10862. doi:10.48550/ARXIV.2109.10862.

[14] X. Feng, X. Feng, B. Qin, A survey on dialogue summarization: Recent advances and new frontiers, 2021. URL: https://arxiv.org/abs/2107.03175. doi:10.48550/ARXIV.2107.03175.

[15] Y. Zhang, A. Ni, T. Yu, R. Zhang, C. Zhu, B. Deb, A. Celikyilmaz, A. H. Awadallah, D. Radev, An exploratory study on long dialogue summarization: What works and what's next, 2021. URL: https://arxiv.org/abs/2109.04609. doi:10.48550/ARXIV.2109.04609.

[16] B. Gliwa, I. Mochol, M. Biesek, A. Wawer, SAMSum corpus: A human-annotated dialogue dataset for abstractive summarization, in: Proceedings of the 2nd Workshop on New Frontiers in Summarization, Association for Computational Linguistics, 2019. URL: https://doi.org/10.18653%2Fv1%2Fd19-5409. doi:10.18653/v1/d19-5409.

[17] Y. Zou, L. Zhao, Y. Kang, J. Lin, M. Peng, Z. Jiang, C. Sun, Q. Zhang, X. Huang, X. Liu, Topic-oriented spoken dialogue summarization for customer service with saliency-aware topic modeling, Proceedings of the AAAI Conference on Artificial Intelligence 35 (2021) 14665–14673.

[6] See briefcaselaw.com for an application related to this paper's goals.

URL: https://ojs.aaai.org/index.php/AAAI/article/view/17723. doi:10.1609/aaai.v35i16.17723.

[18] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, P. J. Liu, Exploring the limits of transfer learning with a unified text-to-text transformer, The Journal of Machine Learning Research 21 (2020) 5485–5551.

[19] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, et al., Language models are unsupervised multitask learners, OpenAI blog 1 (2019) 9.

[20] H. Face, Perplexity of fixed-length models, 2023. URL: https://huggingface.co/docs/transformers/perplexity.

[21] J. R. Landis, G. G. Koch, An application of hierarchical kappa-type statistics in the assessment of majority agreement among multiple observers, Biometrics (1977) 363–374.

[22] B. Wang, X. Deng, H. Sun, Iteratively prompt pretrained language models for chain of thought, in: Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, 2022, pp. 2714–2730.