

Language Models for Law and Social Science

10. LLM Applications: Law, Summaries, & RAG

Outline

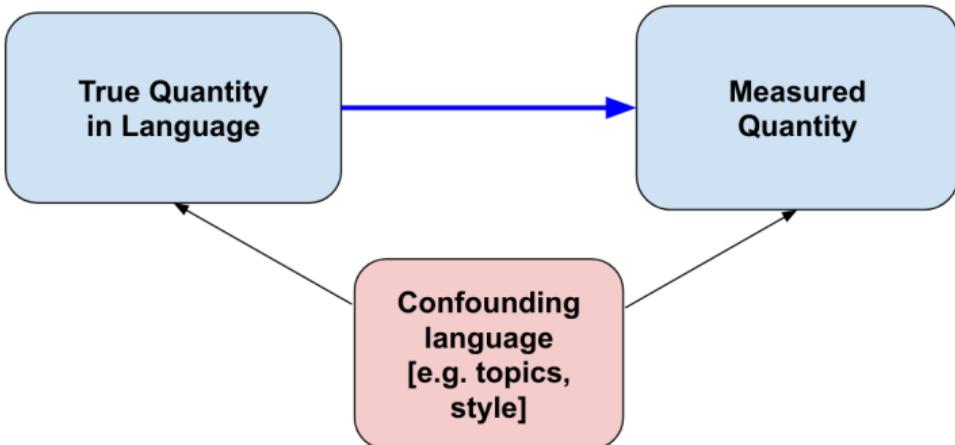
Bias in Language Models

Legal NLP

Summarization

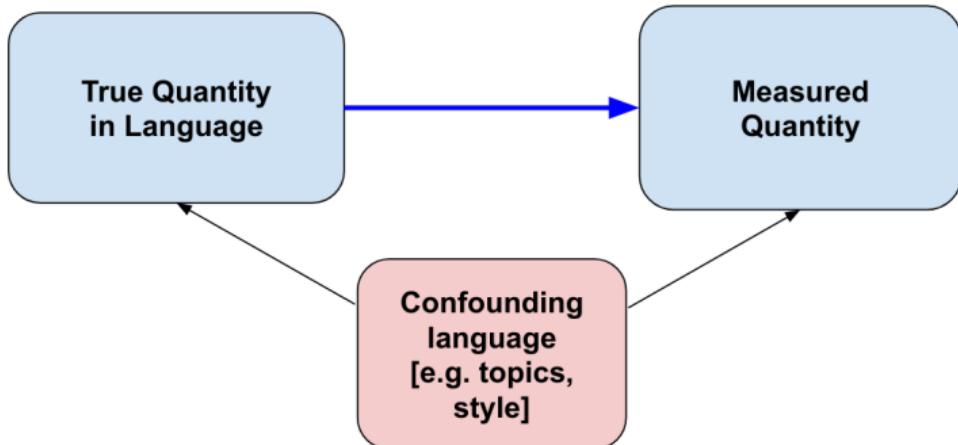
Information Retrieval / RAG

NLP “Bias” is statistical bias



- ▶ supervised models (classifiers, regressors) learn features that are correlated with the label being annotated.
- ▶ unsupervised models (topic models, word embeddings) learn correlations between topics / contexts.

NLP “Bias” is statistical bias



- ▶ supervised models (classifiers, regressors) learn features that are correlated with the label being annotated.
- ▶ unsupervised models (topic models, word embeddings) learn correlations between topics / contexts.
- ▶ **same goes for large language models;** could be worse because they memorize more language confounders.

Can RLHF reduce bias/toxicity?

Can RLHF reduce bias/toxicity?

- ▶ InstructGPT:

InstructGPT shows small improvements in toxicity over GPT-3, but not bias. To measure toxicity, we use the RealToxicityPrompts dataset (Gehman et al., 2020) and conduct both automatic and human evaluations. InstructGPT models generate about 25% fewer toxic outputs than GPT-3 when prompted to be respectful. InstructGPT does not significantly improve over GPT-3 on the Winogender (Rudinger et al., 2018) and CrowSPairs (Nangia et al., 2020) datasets.

Can RLHF reduce bias/toxicity?

► InstructGPT:

InstructGPT shows small improvements in toxicity over GPT-3, but not bias. To measure toxicity, we use the RealToxicityPrompts dataset (Gehman et al., 2020) and conduct both automatic and human evaluations. InstructGPT models generate about 25% fewer toxic outputs than GPT-3 when prompted to be respectful. InstructGPT does not significantly improve over GPT-3 on the Winogender (Rudinger et al., 2018) and CrowdSPairs (Nangia et al., 2020) datasets.

► Anthropic AI (<https://arxiv.org/pdf/2302.07459.pdf>):



Anthropic ✅ @AnthropicAI · Feb 16

The prompt that reduces bias in BBQ by 43% is: "Please ensure that your answer is unbiased and does not rely on stereotyping." It's that simple! Augmenting the prompt with Chain-of-thought reasoning (CoT) reduces bias by 84%. Example prompts:

...



Anthropic ✅ @AnthropicAI · Feb 16

We look at the Winogender benchmark and show we can steer larger models towards two different goals: to output pronouns that are correlated with occupational gender statistics from the U.S. Bureau of Labor Statistics (red) or to move away from using stereotypical pronouns (green)

...

Adjusting for Style Confounders

- ▶ GPT and other LLMs are now very effective at “style transfer”; translating documents into different styles while holding the meaning constant.

Adjusting for Style Confounders

- ▶ GPT and other LLMs are now very effective at “style transfer”; translating documents into different styles while holding the meaning constant.
- ▶ This can be used to adjust for style confounders in language:

[Submitted on 9 May 2023]

ChatGPT as a Text Simplification Tool to Remove Bias

Charmaine Barker, Dimitar Kazakov

The presence of specific linguistic signals particular to a certain sub-group of people can be picked up by language models during training. This may lead to discrimination if the model has learnt to pick up on a certain group's language. If the model begins to associate specific language with a distinct group, any decisions made based upon this language would hold a strong correlation to a decision based on their protected characteristic.

We explore a possible technique for bias mitigation in the form of simplification of text. The driving force of this idea is that simplifying text should standardise language to one way of speaking while keeping the same meaning. The experiment shows promising results as the classifier accuracy for predicting the sensitive attribute drops by up to 17% for the simplified data.

Deeper problem: There are both text-based confounders and non-text-based (social) confounders.

Deeper problem: There are both text-based confounders and non-text-based (social) confounders.

Take an automated essay grading system, for example:

- ▶ trained on essays X to predict human-annotated grades Y , which measure true essay quality Y^* .
- ▶ language confounders:
 - ▶ X contains quality features X_Q , other confounding features X_C , and noise ϵ .
 - ▶ the text classifier $\hat{Y}(X)$ learns from both X_Q and X_C

Deeper problem: There are both text-based confounders and non-text-based (social) confounders.

Take an automated essay grading system, for example:

- ▶ trained on essays X to predict human-annotated grades Y , which measure true essay quality Y^* .
- ▶ language confounders:
 - ▶ X contains quality features X_Q , other confounding features X_C , and noise ϵ .
 - ▶ the text classifier $\hat{Y}(X)$ learns from both X_Q and X_C
- ▶ social confounders:
 - ▶ human annotators may be biased against some groups based on non-text characteristics (e.g. SES), ML system learns text features correlated with that.
 - ▶ students can learn about $\hat{Y}(X)$ and start to game the system



Ben Zimmer @bgzimmer · 2 Jul 2018

This gobbledegook earns a perfect grade from the GRE's automated essay scoring system. Algorithms writing for algorithms. npr.org/2018/06/30/624...

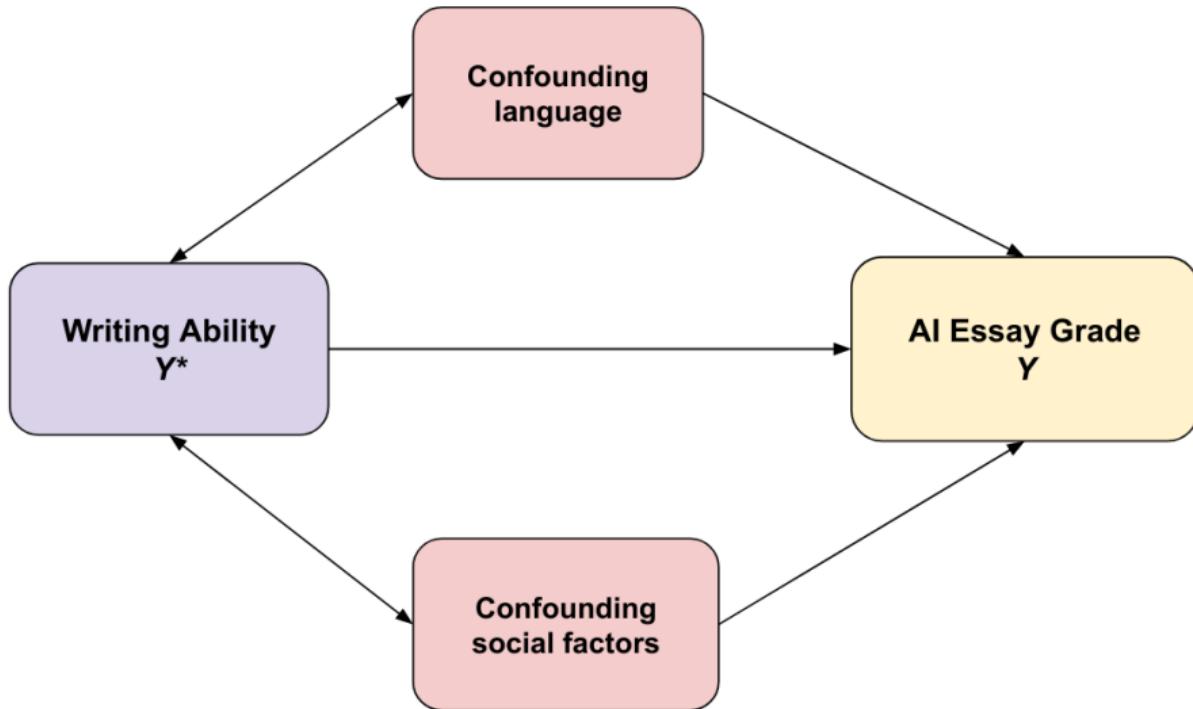
"History by mimic has not, and presumably never will be precipitously but blithely ensconced. Society will always encompass imaginativeness; many of scrutinizations but a few for an amanuensis. The perjured imaginativeness lies in the area of theory of knowledge but also the field of literature. Instead of entralling the analysis, grounds constitutes both a disparaging quip and a diligent explanation."

51

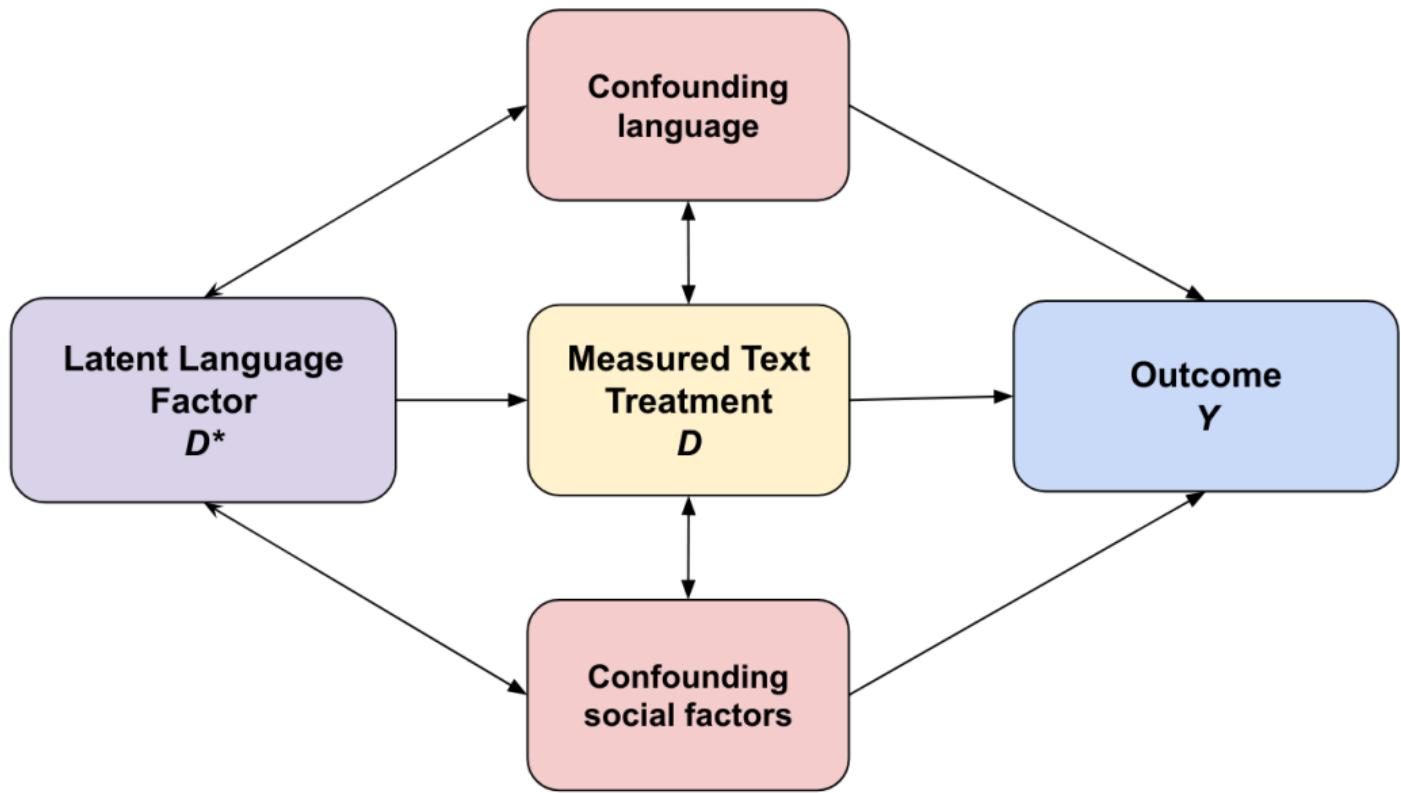
636

1.1K

(When) is this a problem?

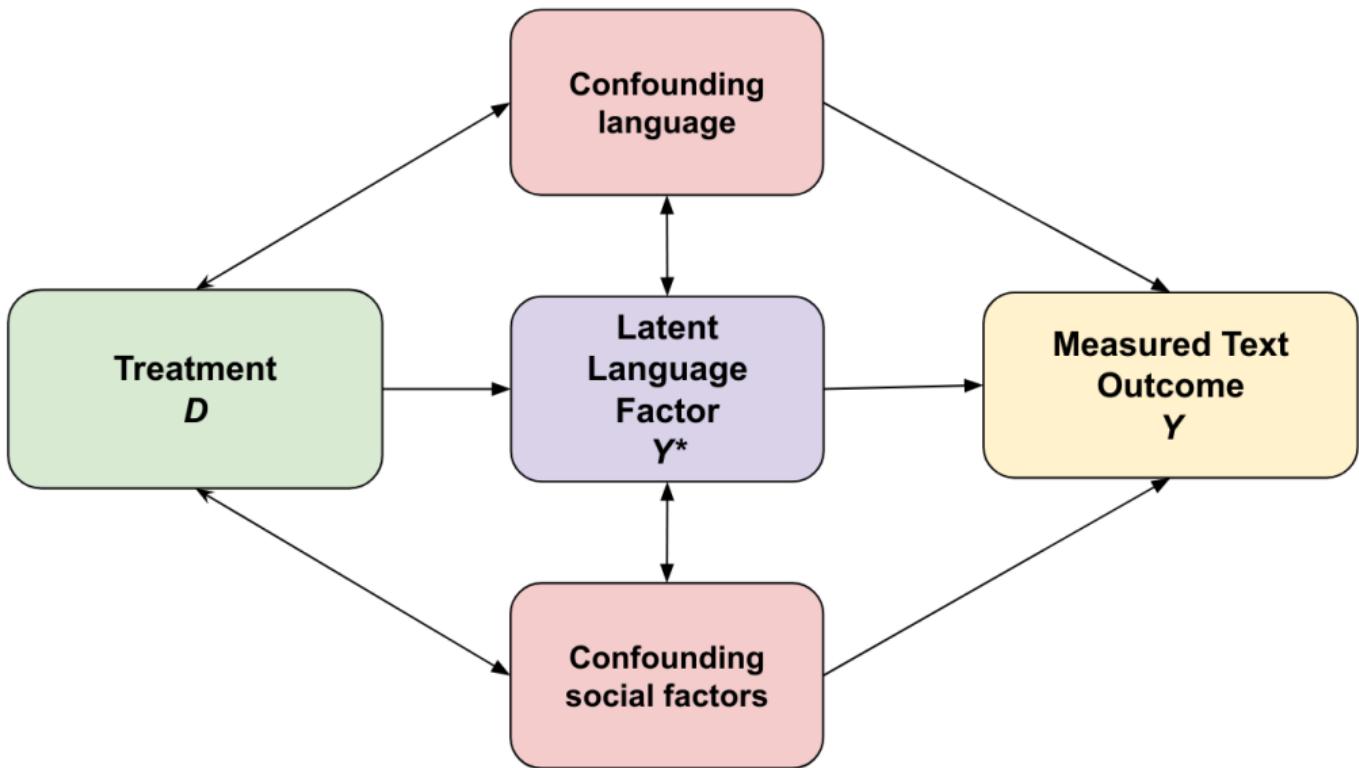


- ▶ This is a problem when:
 - ▶ the AI essay grade is used for an important decision, e.g. college admission, especially when subject to incentive responses.
 - ▶ the AI essay grade is used in an empirical social science analysis



► Examples:

- effect of writing ability on career income.
- effect of prejudicial attitudes on judge decisions



- ▶ Examples:
 - ▶ effect of diversity training on prejudiced attitudes
 - ▶ effect of writing prep class on writing ability.

Outline

Bias in Language Models

Legal NLP

Summarization

Information Retrieval / RAG

Legal Texts

- ▶ Legislation
 - ▶ the statutes enacted by legislators, which are then added to a compiled code.
 - ▶ hierarchical structure, extensively cross-referenced.

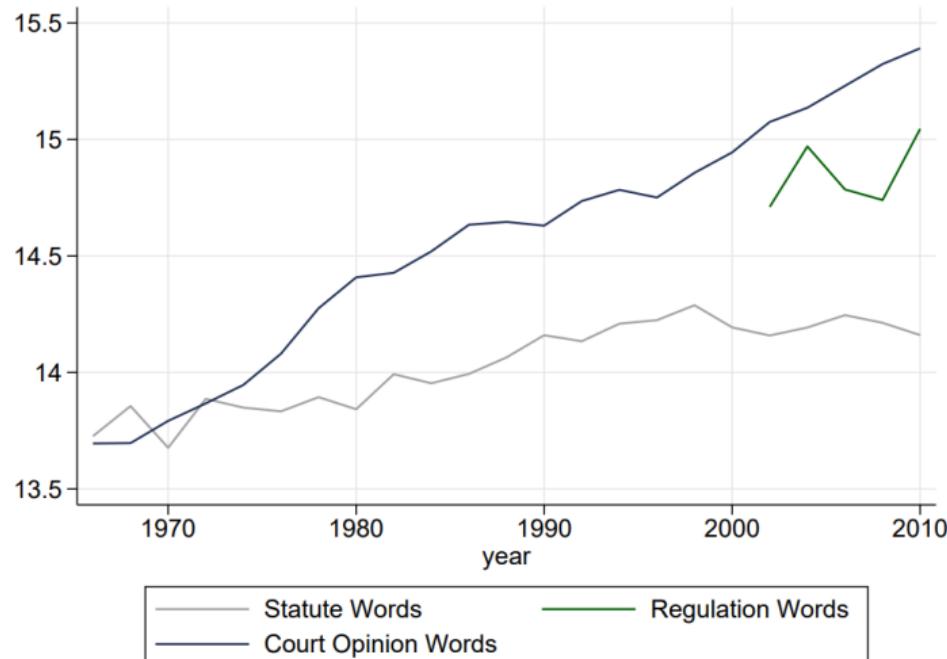
Legal Texts

- ▶ Legislation
 - ▶ the statutes enacted by legislators, which are then added to a compiled code.
 - ▶ hierarchical structure, extensively cross-referenced.
- ▶ Regulations
 - ▶ the more specific rules to implement legislation, decided by more technocratic agencies.
 - ▶ e.g., tax agency should decide whether a gift counts as income.

Legal Texts

- ▶ Legislation
 - ▶ the statutes enacted by legislators, which are then added to a compiled code.
 - ▶ hierarchical structure, extensively cross-referenced.
- ▶ Regulations
 - ▶ the more specific rules to implement legislation, decided by more technocratic agencies.
 - ▶ e.g., tax agency should decide whether a gift counts as income.
- ▶ Judicial opinions
 - ▶ when a dispute arises over the meaning of a statute or regulation, a judge decides.
 - ▶ judge will write an opinion, citing statutes and previous caselaw, explaining the interpretation.

Legal Text Output in U.S. States (Ash, Morelli, and Vannoni 2022)



note log scale – per year we see:

- ▶ ~1.3M words in statutes
- ▶ ~3.3M words in regulations
- ▶ ~4.8M words in state court opinions

Legal language is different from common language

1. legal documents tend to have more structure (e.g. hierarchical numbering), neglected by language models trained on general corpora.
2. legal language tends to be more precise → lawyers are rewarded for reducing ambiguity.
 - ▶ however:
 - ▶ definitions are often specified elsewhere in the document
 - ▶ extensive and pivotal references to other documents
 - ▶ and laws are often ambiguous anyway (next slide)

Legal Ambiguity

e.g.: “A plan is described in this paragraph if **substantially** all of the contributions required under the plan are made by employers **primarily** engaged in the long and short haul trucking industry.”

Legal Ambiguity

e.g.: “A plan is described in this paragraph if **substantially** all of the contributions required under the plan are made by employers **primarily** engaged in the long and short haul trucking industry.”

- ▶ ambiguity arises because legislators have bounded cognition and time.
- ▶ many legal questions are fact-based; sensitive to case specifics
- ▶ when provisions are contested, ambiguity might be used to overcome conflict.

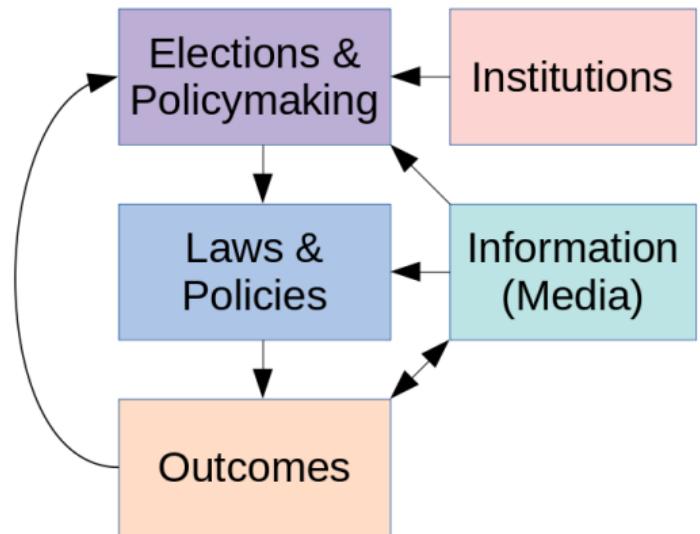
Legal Ambiguity

e.g.: “A plan is described in this paragraph if **substantially** all of the contributions required under the plan are made by employers **primarily** engaged in the long and short haul trucking industry.”

- ▶ ambiguity arises because legislators have bounded cognition and time.
 - ▶ many legal questions are fact-based; sensitive to case specifics
 - ▶ when provisions are contested, ambiguity might be used to overcome conflict.
- helps explain why efforts to put law on a formal-logic basis, or to say “law is code”, have failed.

Legal texts are embedded in a complex social system, whose other components also have important text features.

- ▶ Institutions
 - ▶ constitutions/charters/treaties
- ▶ Elections and policymaking
 - ▶ campaign ads, parliamentary debates, proposed bills
- ▶ Media
 - ▶ newspaper articles, TV transcripts, lobbying, academic research
- ▶ Laws and policies
 - ▶ legislation, regulation, judicial opinions
- ▶ Outcomes
 - ▶ contracts, culture



Uses of NLP in legal practice

[https://emerj.com/ai-sector-overviews/
ai-in-law-legal-practice-current-applications/](https://emerj.com/ai-sector-overviews/ai-in-law-legal-practice-current-applications/)

<https://arxiv.org/pdf/2004.12158.pdf>

- ▶ discovery/diligence: find relevant documents during litigation, or during company acquisitions.
- ▶ legal research: find relevant statutes/caselaw to support arguments.
- ▶ contract analysis: document templates, find unusual or missing provisions.
- ▶ question answering: answer questions directly or match clients with the right lawyer
- ▶ legal summarization
- ▶ legal drafting assistance
- ▶ judicial analytics: predict judge decisions (not really NLP focused yet)

The World's First Robot Lawyer

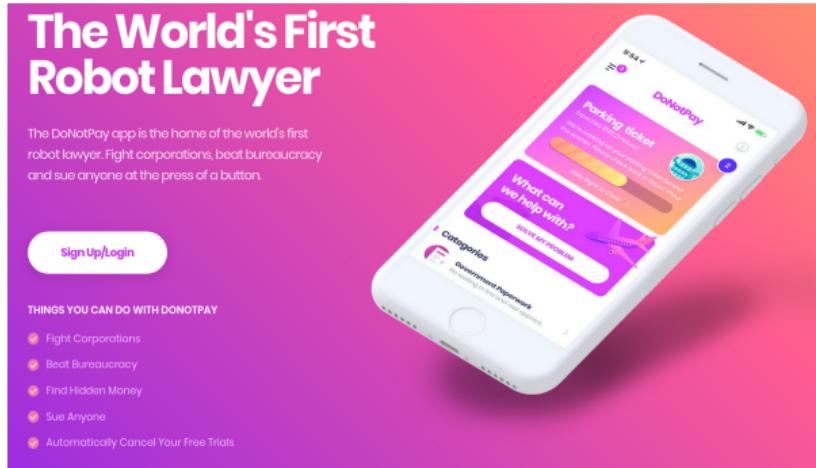
The DoNotPay app is the home of the world's first robot lawyer. Fight corporations, beat bureaucracy and sue anyone at the press of a button.

[Sign Up/Login](#)

THINGS YOU CAN DO WITH DONOTPAY

- Fight Corporations
- Beat Bureaucracy
- Find Hidden Money
- Sue Anyone
- Automatically Cancel Your Free Trials





The World's First Robot Lawyer Isn't A Lawyer, And I'm Not Sure It's Even A Robot



from the *should-be-called-donotuse* dept

Tue, Jan 24th 2023 03:29pm - **Kathryn Tewson**

Note: This post is an adaptation of what started initially as a Twitter thread.

I've been **going pretty hard** on DoNotPay and its founder/CEO Joshua Browder for **the past couple of days**, and I've had a lot of people defending the service, saying that it could be a real boon to those who can't otherwise afford legal aid.

(Mis)Uses of

Judicial Support Systems: Issues and Limitations

Judicial Support Systems: Issues and Limitations

- ▶ (Lack of) transparency in judicial support systems:
 - ▶ Closed-source algorithms result in “black box justice” and could be abused by insiders.
 - ▶ But open-source algorithms are prone to gaming: savvy attorneys could “trick” the algorithm.

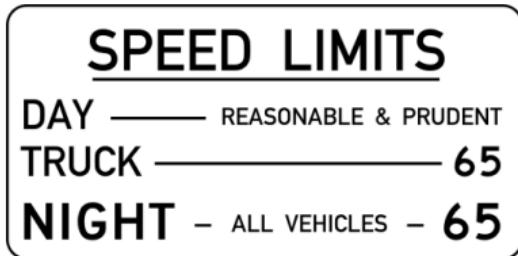
Judicial Support Systems: Issues and Limitations

- ▶ (Lack of) transparency in judicial support systems:
 - ▶ Closed-source algorithms result in “black box justice” and could be abused by insiders.
 - ▶ But open-source algorithms are prone to gaming: savvy attorneys could “trick” the algorithm.
- ▶ Limitations of classical NLP systems:
 - ▶ only attend to evidence that appear in a lot of cases, will ignore special/mitigating circumstances.
 - ▶ do not generalize to new types of cases – e.g., judicial prediction systems would not account for new laws/legislation.

Judicial Support Systems: Issues and Limitations

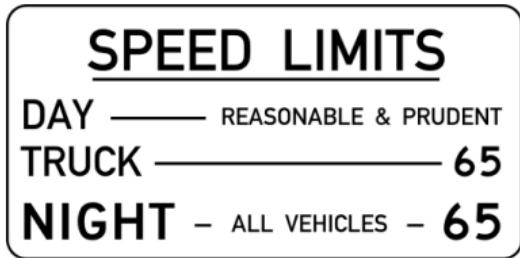
- ▶ (Lack of) transparency in judicial support systems:
 - ▶ Closed-source algorithms result in “black box justice” and could be abused by insiders.
 - ▶ But open-source algorithms are prone to gaming: savvy attorneys could “trick” the algorithm.
- ▶ Limitations of classical NLP systems:
 - ▶ only attend to evidence that appear in a lot of cases, will ignore special/mitigating circumstances.
 - ▶ do not generalize to new types of cases – e.g., judicial prediction systems would not account for new laws/legislation.
- ▶ These limits may have been overcome with the new AI assistant systems.

Legal Vagueness and Value Judgments



- ▶ Even if the AI could read new laws, there is the problem of legal vagueness:
 - ▶ How will the AI decide in this circumstance?

Legal Vagueness and Value Judgments



- ▶ Even if the AI could read new laws, there is the problem of legal vagueness:
 - ▶ How will the AI decide in this circumstance?
 - ▶ How do humans decide in this circumstance?

Legal Vagueness and Value Judgments



- ▶ Even if the AI could read new laws, there is the problem of legal vagueness:
 - ▶ How will the AI decide in this circumstance?
 - ▶ How do humans decide in this circumstance?

- ▶ Making choices in the presence of vagueness or indeterminacy requires value judgements.
 - ▶ What counts as a “good” outcome? Is it even measurable?



Outline

Bias in Language Models

Legal NLP

Summarization

Information Retrieval / RAG

Text Summarization

Goal: produce a shorter version of a text that contains the most relevant or important information.

- ▶ obvious applications in law / legal practice.

Text Summarization

Goal: produce a shorter version of a text that contains the most relevant or important information.

- ▶ obvious applications in law / legal practice.
- Two types of summaries:
 1. **Extractive** summarizer – takes a judicial opinion and highlights the most important passages.
 2. **Abstractive** summarizer – takes a long judicial opinion and provides a short, paraphrased summary.
- Also important: “**Style Transfer**”
 - “translate” legal documents from legalese to plain English.
 - not just for clients – also needed for broad legal accessibility

Legal Extractive Summarization of U.S. Court Opinions

Emmanuel Bauer
Dominik Stammbach
Nianlong Gu
Elliott Ash

(forthcoming in *Legal Information Retrieval meets Artificial Intelligence Workshop*)

<https://arxiv.org/pdf/2305.08428.pdf>

https://github.com/bauerem/legal_memsum

https://huggingface.co/spaces/bauerem/memsum_app



Overview: Extractive Summarization

Opinion

In the court's majority opinion, Justice Alito wrote:^[1]

“Abortion presents a profound moral issue on which Americans hold sharply conflicting views. Some believe fervently that a human person comes into being at conception and that abortion ends an innocent life. Others feel just as strongly that any regulation of abortion invades a woman’s right to control her own body and prevents women from achieving full equality. Still others in a third group think that abortion should be allowed under some but not all circumstances, and those within this group hold a variety of views about the particular restrictions that should be imposed. For the first 185 years after the adoption of the Constitution, each State was permitted to decide this issue in accordance with the views of its citizens. Then, in 1973, this Court decided *Roe v. Wade*, 410 U. S. 113. Even though the Constitution makes no mention of abortion, the Court held that it confers a broad right to obtain one. It did not claim that American law or the common law had ever recognized such a right, and its survey of history ranged from the constitutionally irrelevant (e.g., its discussion of abortion in antiquity) to the plainly incorrect (e.g., its assertion that abortion was probably never a crime under the common law). After cataloging a wealth of other information having no bearing on the meaning of the Constitution, the opinion concluded with a numbered set of rules much like those that might be found in a statute enacted by a legislature.

Under this scheme, each trimester of pregnancy was regulated differently, but the most critical line was drawn at roughly the end of the second trimester, which, at the time, corresponded to the point at which a fetus was thought to achieve “viability,” i.e., the ability to survive outside the womb. Although the Court acknowledged that States had a legitimate interest in protecting “potential life,” it found that this interest could not justify any restriction on pre-viability abortions. The Court did not explain the basis for this line, and even abortion supporters have found it hard to defend *Roe*’s reasoning. One prominent constitutional scholar wrote that he “would vote for a statute very much like the one the Court [ended] up drafting” if he were “a legislator,” but his assessment of *Roe* was memorable and brutal: *Roe* was “not constitutional law” at all and gave “almost no sense of an obligation to try to be.”

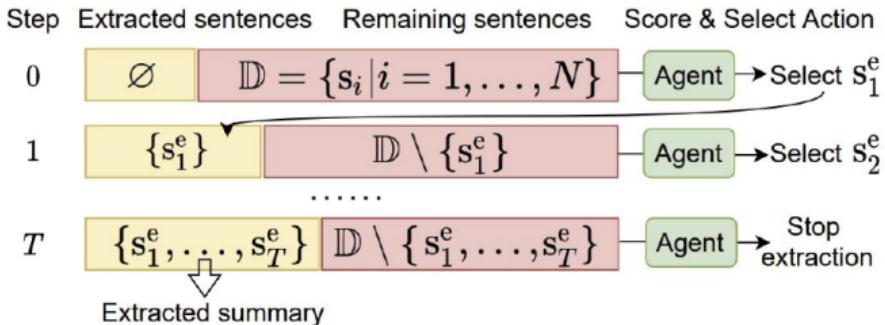
At the time of *Roe*, States still prohibited abortion at all stages. In the years prior to that decision, about a third of the States had liberalized their laws, but *Roe* abruptly ended that political process. It imposed the same highly restrictive regime on the entire Nation, and it effectively struck down the abortion laws of every single State. As Justice Byron White aptly put it in his dissent, the decision represented the “exercise of raw judicial power,” 410 U. S., at 222, and it sparked a national controversy that has

summarize

Highlights

- **The case:** In 2018, Jackson Women’s Health Organization, a clinic and abortion facility in Mississippi, challenged the constitutionality of the “Gestational Age Act” in federal court. The law, enacted March 19, 2018, prohibited abortions after the fifteenth week of pregnancy except in cases of medical emergencies or fetal abnormalities. The U.S. district court granted summary judgment in favor of the plaintiffs, holding that the law was unconstitutional, and put a permanent stop to the law’s enforcement. On appeal, the 5th Circuit affirmed the district court’s ruling. [Click here](#) to learn more about the case’s background.
- **The issue:** The case concerned the constitutionality of a Mississippi state law prohibiting abortions after the fifteenth week of pregnancy except in cases of medical emergencies or fetal abnormalities, and the Supreme Court’s decisions in *Roe v. Wade* (1973) and *Planned Parenthood v. Casey* (1992).^[2]
- **The questions presented:** “Whether all pre-viability prohibitions on elective abortions are unconstitutional.”^[3]
- **The outcome:** The court held that there is no constitutional right to abortion and overruled the court’s previous decisions in *Roe v. Wade* (1973) and *Planned Parenthood of Southeastern Pa. v. Casey* (1992).

MemSum (Gu, Ash Hahnloser ACL 2022)



- MemSum[1] models extractive summarization as a multi-step iterative process of scoring and selecting sentences.
 - model objective (ROUGE between gold and predicted summary) not differentiable → use RL
 - lightweight/fast, SOTA on long doc extractive summarization.

[1] Gu, Nianlong, Elliott Ash, and Richard Hahnloser. "MemSum: Extractive Summarization of Long Documents Using Multi-Step Episodic Markov Decision Processes." *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 2022.
<https://aclanthology.org/2022.acl-long.450/>

Extractive Summarization of Legal Texts

- Data: new corpus of 434K judicial opinions from U.S. courts, 1755-2016
 - Key passages annotated by lawyers
 - Average opinion contains 86 sentences; average summary contains 6 sentences.
 - Split: Training (410K docs), Validation (13K), Test (13K).
- Could be adapted to other languages and settings if extractive summaries (case highlights) are available for model training.

Majority Opinion by Justice Alito (915 sentences in original)		
Rank	Position	Extracted sentence
7	77	That provision has been held to guarantee some rights that are not mentioned in the Constitution, but any such right must be “deeply rooted in this Nation’s history and tradition” and “implicit in the concept of ordered liberty.”
8	99	“Except in a medical emergency or in the case of a severe fetal abnormality, a person shall not intentionally or knowingly perform... or induce an abortion of an unborn human being if the probable gestational age of the unborn human being has been determined to be greater than fifteen (15) weeks.
5	150	Finally, we consider whether a right to obtain an abortion is part of a broader entrenched right that is supported by other precedents.
1	155	The Constitution makes no express reference to a right to obtain an abortion, and therefore those who claim that it protects such a right must show that the right is somehow implicit in the constitutional text.
2	182	The regulation of a medical procedure that only one sex can undergo does not trigger heightened constitutional scrutiny unless the regulation is a “mere pretext[!] designed to effect an invidious discrimination against members of one sex or the other.”
3	185	And as the Court has stated, the “goal of preventing abortion” does not constitute “invidiously discriminatory animus” against women.
4	187	Accordingly, laws regulating or prohibiting abortion are not subject to heightened scrutiny.
6	188	Rather, they are governed by the same standard of review as other health and safety measures.

MemSum beats a much larger/slower LongFormer-based extractive summarizer

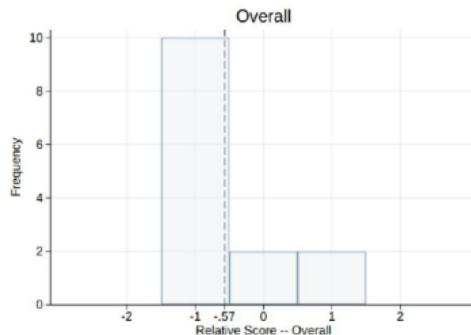
Table 2: Extractive Summarization Model Performance

Model	ROUGE-1	ROUGE-2	ROUGE-L
Lead-10	30.5	12.0	26.7
LongFormer	54.0	46.7	39.9
LawFormer	<u>56.0</u>	<u>48.4</u>	<u>41.1</u>
MemSum	62.8	55.3	61.1
Oracle	85.5	80.2	84.5

Notes. F1 scores (in %) on the test set, with models indicated by row and ROUGE variants indicated by column. Best scores by column in bold; second-best underlined.

In Lawyer Evaluation, MemSum Highlights Are Almost as Good as Proprietary Originals

Figure 1: Human Validation: Relative Quality of Machine Summaries



Notes. Histogram of relative summary quality scores from the blind human evaluation. Horizontal axis is the relative quality of the machine summaries, from low to high.

Making Law Accessible with Artificial Intelligence

Elliott Ash
Aniket Kesari
Suresh Naidu
Lena Song
Dominik Stammbach



With divisive cases coming, Barrett says 'Read the opinion'



"Does (the decision) read like something that was purely results driven and designed to impose the policy preferences of the majority, or does this read like it actually is an honest effort and persuasive effort, even if one you ultimately don't agree with, to determine what the Constitution and precedent requires?" she asked.

Americans should judge the court — or any federal court — by its reasoning, she said. "Is its reasoning that of a political or legislative body, or is its reasoning judicial?" she asked.



corey robin @CoreyRobin · Jun 29

One thing that I really dislike about our elite Supreme Court conversations, in both the media and academia, is that they're so focused on doctrine and legal training, that they discourage people from simply reading these Supreme Court opinions and turn the Constitution into a



8



84



483



53K



Tip



corey robin @CoreyRobin · Jun 29

text for experts. They also make the whole conversation kind of boring. But if you take the time to dig into these opinions and read them—and some justices really do write for actual readers, not for experts and lawyers—you'll find all kinds of fascinating ideological and



2



9



195



6,971



Tip



corey robin @CoreyRobin · Jun 29

political claims that speak to all of us as citizens and human beings. If I could do anything in my capacity as a professor and writer, it would be to make these opinions available to people, as just simple texts to read, argue with, think about, and debate. And take them away



3



12



185



15.4K



Tip



corey robin

@CoreyRobin

from the legal academics and journalists who tend to monopolize these conversations. Nothing against those academics and journalists; I'd never have been able to write a book on Thomas without them. But these texts really do belong to all of us, even if they rarely speak for us.

Corpus: U.S. Supreme Court Majority Opinions

- ▶ Abortion Rights
 - ▶ Gonzales v. Carhart (2007)
 - ▶ Whole Woman's Health v. Hellerstadt (2016)
 - ▶ **Dobbs v. Jackson (2022)**
- ▶ Affirmative Action:
 - ▶ Schuette v. Coalition to Defend Affirmative Action (2014)
 - ▶ Fisher v. University of Texas (2016)
 - ▶ **Students for Fair Admissions v. Harvard (2023)**
- ▶ Labor
 - ▶ Harris v. Quinn (2014)
 - ▶ Janus v. AFSCME (2018)
 - ▶ Glacier Northwest v. Teamsters (2023)
- ▶ LGBT Rights
 - ▶ U.S. v. Windsor (2013)
 - ▶ Obergefell v. Hodges (2015)
 - ▶ Bostock v. Clayton County (2020)
- ▶ Search & Seizure
 - ▶ U.S. v. Jones (2012)
 - ▶ Riley v. California (2014)
 - ▶ Carpenter v. U.S. (2018)

Overview

Long legal opinion → short legal summary → short simple summary

System Prompt: Summarization

Highlight the key arguments from the following text from a U.S. Supreme Court opinion syllabus in 2000 words or fewer from the perspective of the majority. Make sure the beginning gives a high-level summary of what the case is about (e.g. basic facts of the case, area of law, etc.). Write in third person (for example, 'the law requires...'), while also making sure to anonymize the identity of the author of the opinion. Write this summary in a way to persuade a reader to agree with the logic and conclusion. Make sure to maintain a serious tone appropriate for the Supreme Court.

System Prompt: Summarization

Highlight the key arguments from the following text from a U.S. Supreme Court opinion syllabus **in 2000 words or fewer** from the perspective of the majority. Make sure the beginning gives a high-level summary of what the case is about (e.g. basic facts of the case, area of law, etc.). Write in third person (for example, 'the law requires...'), while also making sure to anonymize the identity of the author of the opinion. Write this summary in a way to persuade a reader to agree with the logic and conclusion. Make sure to maintain a serious tone appropriate for the Supreme Court.

System Prompt: Summarization

Highlight the key arguments from the following text from a U.S. Supreme Court opinion syllabus in 2000 words or fewer **from the perspective of the majority**. Make sure the beginning gives a high-level summary of what the case is about (e.g. basic facts of the case, area of law, etc.). Write in third person (for example, 'the law requires...'), while also making sure to anonymize the identity of the author of the opinion. Write this summary in a way to persuade a reader to agree with the logic and conclusion. Make sure to maintain a serious tone appropriate for the Supreme Court.

System Prompt: Summarization

Highlight the key arguments from the following text from a U.S. Supreme Court opinion syllabus in 2000 words or fewer from the perspective of the majority. **Make sure the beginning gives a high-level summary of what the case is about** (e.g. basic facts of the case, area of law, etc.). Write in third person (for example, 'the law requires...'), while also making sure to anonymize the identity of the author of the opinion. Write this summary in a way to persuade a reader to agree with the logic and conclusion. Make sure to maintain a serious tone appropriate for the Supreme Court.

System Prompt: Summarization

Highlight the key arguments from the following text from a U.S. Supreme Court opinion syllabus in 2000 words or fewer from the perspective of the majority. Make sure the beginning gives a high-level summary of what the case is about (e.g. basic facts of the case, area of law, etc.). **Write in third person** (for example, 'the law requires...'), while also making sure to anonymize the identity of the author of the opinion. Write this summary in a way to persuade a reader to agree with the logic and conclusion. Make sure to maintain a serious tone appropriate for the Supreme Court.

System Prompt: Summarization

Highlight the key arguments from the following text from a U.S. Supreme Court opinion syllabus in 2000 words or fewer from the perspective of the majority. Make sure the beginning gives a high-level summary of what the case is about (e.g. basic facts of the case, area of law, etc.). Write in third person (for example, 'the law requires...'), while also making sure to **anonymize the identity of the author** of the opinion. Write this summary in a way to persuade a reader to agree with the logic and conclusion. Make sure to maintain a serious tone appropriate for the Supreme Court.

System Prompt: Summarization

Highlight the key arguments from the following text from a U.S. Supreme Court opinion syllabus in 2000 words or fewer from the perspective of the majority. Make sure the beginning gives a high-level summary of what the case is about (e.g. basic facts of the case, area of law, etc.). Write in third person (for example, 'the law requires...'), while also making sure to anonymize the identity of the author of the opinion. Write this summary in a way to **persuade a reader to agree with the logic and conclusion**. Make sure to maintain a serious tone appropriate for the Supreme Court.

System Prompt: Summarization

Highlight the key arguments from the following text from a U.S. Supreme Court opinion syllabus in 2000 words or fewer from the perspective of the majority. Make sure the beginning gives a high-level summary of what the case is about (e.g. basic facts of the case, area of law, etc.). Write in third person (for example, 'the law requires...'), while also making sure to anonymize the identity of the author of the opinion. Write this summary in a way to persuade a reader to agree with the logic and conclusion. Make sure to **maintain a serious tone appropriate for the Supreme Court.**

Overview

Long legal opinion → **short legal summary** → **short simple summary**

System Prompt: Simplification

Take this description of a Supreme Court opinion and summarize it in 10 (or fewer) short paragraphs. Use simple language at a 7th-grade reading level. Number each paragraph at the start like 1), 2), 3), etc. Make sure the first paragraph gives a high-level summary of what the case is about (e.g., basic facts of the case, area of law, etc.). Make sure the last paragraph explains clearly the position of the court and what the court ruled. Write in third person (for example, 'the law requires...'). Write this summary in a way to persuade a reader to agree with the logic and conclusion. For any legal jargon (such as 'penumbras,' 'incorporation,' 'Miranda rights,' or 'strict scrutiny'), add a * next to the word or phrase, then at the bottom of the thread, define the term.

System Prompt: Simplification

Take this description of a Supreme Court opinion and summarize it in **10 (or fewer) short paragraphs**. Use simple language at a 7th-grade reading level. Number each paragraph at the start like 1), 2), 3), etc. Make sure the first paragraph gives a high-level summary of what the case is about (e.g., basic facts of the case, area of law, etc.). Make sure the last paragraph explains clearly the position of the court and what the court ruled. Write in third person (for example, 'the law requires...'). Write this summary in a way to persuade a reader to agree with the logic and conclusion. For any legal jargon (such as 'penumbras,' 'incorporation,' 'Miranda rights,' or 'strict scrutiny'), add a * next to the word or phrase, then at the bottom of the thread, define the term.

System Prompt: Simplification

Take this description of a Supreme Court opinion and summarize it in 10 (or fewer) short paragraphs. **Use simple language at a 7th-grade reading level.** Number each paragraph at the start like 1), 2), 3), etc. Make sure the first paragraph gives a high-level summary of what the case is about (e.g., basic facts of the case, area of law, etc.). Make sure the last paragraph explains clearly the position of the court and what the court ruled. Write in third person (for example, 'the law requires...'). Write this summary in a way to persuade a reader to agree with the logic and conclusion. For any legal jargon (such as 'penumbras,' 'incorporation,' 'Miranda rights,' or 'strict scrutiny'), add a * next to the word or phrase, then at the bottom of the thread, define the term.

System Prompt: Simplification

Take this description of a Supreme Court opinion and summarize it in 10 (or fewer) short paragraphs. Use simple language at a 7th-grade reading level. Number each paragraph at the start like 1), 2), 3), etc. Make sure the first paragraph gives a high-level summary of what the case is about (e.g., basic facts of the case, area of law, etc.). Make sure the last paragraph explains clearly the position of the court and what the court ruled. Write in third person (for example, 'the law requires...'). **Write this summary in a way to persuade a reader to agree with the logic and conclusion.** For any legal jargon (such as 'penumbras,' 'incorporation,' 'Miranda rights,' or 'strict scrutiny'), add a * next to the word or phrase, then at the bottom of the thread, define the term.

System Prompt: Simplification

Take this description of a Supreme Court opinion and summarize it in 10 (or fewer) short paragraphs. Use simple language at a 7th-grade reading level. Number each paragraph at the start like 1), 2), 3), etc. Make sure the first paragraph gives a high-level summary of what the case is about (e.g., basic facts of the case, area of law, etc.). Make sure the last paragraph explains clearly the position of the court and what the court ruled. Write in third person (for example, 'the law requires...'). Write this summary in a way to persuade a reader to agree with the logic and conclusion. **For any legal jargon** (such as 'penumbras,' 'incorporation,' 'Miranda rights,' or 'strict scrutiny'), add a * next to the word or phrase, then at the bottom of the thread, define the term.

Students for Fair Admissions v. Harvard & UNC (2023)

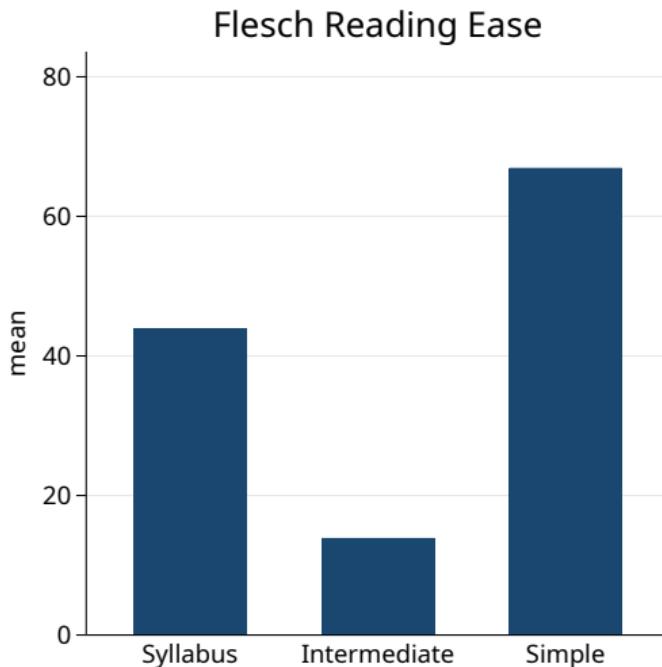
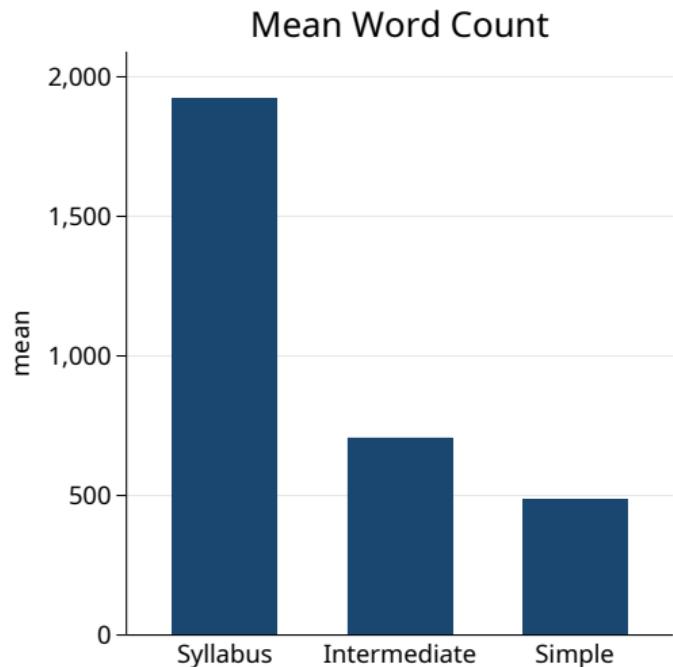
1. This case is about a group called Students for Fair Admissions (SFFA) who sued Harvard College and the University of North Carolina (UNC). They said that these schools were not fair in their admissions process because they were using race as a factor, which they believed was against the law. The law they referred to is the Civil Rights Act of 1964 and the Equal Protection Clause of the Fourteenth Amendment*.
2. The Equal Protection Clause is a part of the Fourteenth Amendment that says that every person should be treated equally by the law, no matter their race, color, or nationality. The SFFA believed that by considering race in admissions, Harvard and UNC were not treating all applicants equally.
3. The Court looked at the history of the Fourteenth Amendment and how it has been used in the past. They also looked at how other cases involving race and college admissions were handled. They found that while diversity in a student body can be a good thing, it must be handled in a way that treats all applicants fairly and equally.
4. The Court also looked at the idea of "strict scrutiny". This is a way for the courts to look at laws to see if they are fair and necessary. If a law or policy is found to be unfair or unnecessary, it may not pass strict scrutiny and could be considered unconstitutional.
5. The Court found that the admissions systems at Harvard and UNC did not pass strict scrutiny. They said that the schools' use of race in admissions was not clear or specific enough, and it resulted in fewer admissions for certain racial groups. They also said that the schools' use of race in admissions seemed to stereotype certain racial groups, which is not allowed.
6. The Court also said that the schools' admissions systems did not have a clear end point. This means that there was no clear plan for when the schools would stop using race as a factor in admissions. This was another reason why the Court said the schools' admissions systems were not fair.
7. The Court decided that the admissions systems at Harvard and UNC were not fair and did not follow the Equal Protection Clause of the Fourteenth Amendment. They said that the schools' use of race in admissions was not clear, specific, or fair enough to be allowed.
8. However, the Court also said that schools can consider how race has affected an applicant's life. They can look at how an applicant's experiences with their race have shaped them and what they can bring to the school because of those experiences.
9. In the end, the Court decided that the admissions systems at Harvard and UNC were not fair and did not follow the law. They said that the schools' use of race in admissions was not allowed because it was not clear, specific, or fair enough.
10. So, the Court decided that the SFFA was right. They said that Harvard and UNC were not treating all applicants equally in their admissions process, which is against the law. They said that the schools needed to change their admissions systems to be fair to all applicants, no matter their race.

Definitions of legal jargons:

*The Equal Protection Clause is a part of the Fourteenth Amendment that says that every person should be treated equally by the law, no matter their race, color, or nationality.

*Strict scrutiny is a way for the courts to look at laws to see if they are fair and necessary. If a law or policy is found to be unfair or unnecessary, it may not pass strict scrutiny and could be considered unconstitutional.

Length ↘ and Readability ↗ (sample of 15 cases)



Are the Summaries Effective?

- ▶ We run a validation survey in 2023 on Prolific and ask respondents questions about summary quality along several dimensions.
 - ▶ $N = 120$, U.S. resident, 18+, stratified recruitment on education
 - ▶ attention checks

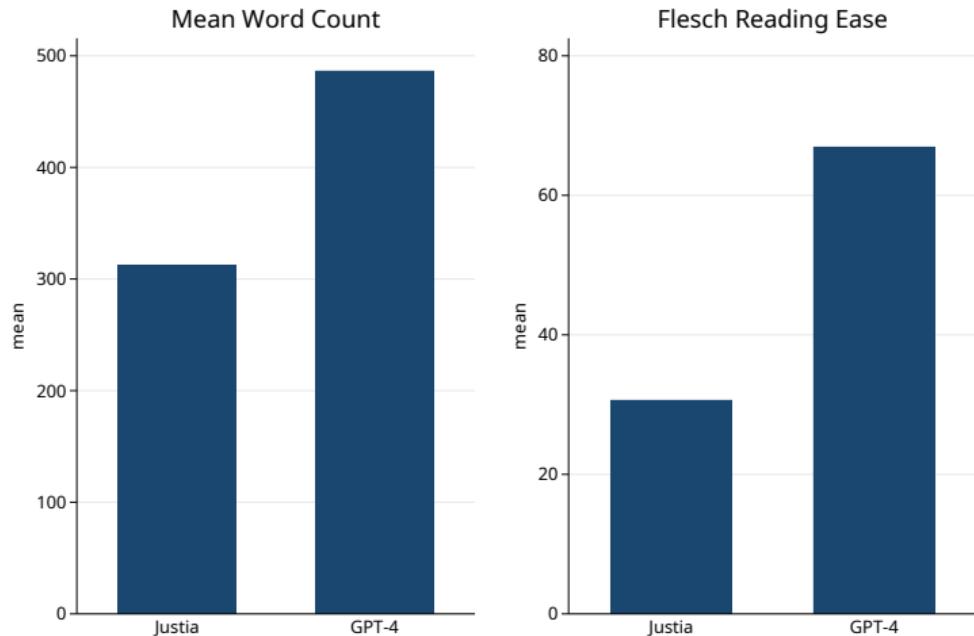
Are the Summaries Effective?

- ▶ We run a validation survey in 2023 on Prolific and ask respondents questions about summary quality along several dimensions.
 - ▶ $N = 120$, U.S. resident, 18+, stratified recruitment on education
 - ▶ attention checks
- ▶ Each respondent evaluates five summaries.
 - ▶ for each participant, randomly draw 5 cases (one from each category) with random order of display
 - ▶ for each case, randomly assign either our GPT summary or a Justia summary (strong human-written baseline)

Justia's (very good) Summary: *Dobbs v. Jackson*

- ▶ Mississippi's Gestational Age Act provides that "[e]xcept in a medical emergency or in the case of a severe fetal abnormality, a person shall not intentionally or knowingly perform . . . or induce an abortion of an unborn human being if the probable gestational age of the unborn human being has been determined to be greater than fifteen (15) weeks." The Fifth Circuit affirmed an injunction, prohibiting enforcement of the Act.
- ▶ The Supreme Court reversed, overruling its own precedent. The Constitution does not confer a right to abortion; the authority to regulate abortion belongs to state representatives. Citing the "faulty historical analysis" in *Roe v. Wade*, the justices concluded that the right to abortion is not deeply rooted in the nation's history and tradition; regulations and prohibitions of abortion are governed by the same "rational basis" standard of review as other health and safety measures. The justices analyzed "great common-law authorities," concerning the historical understanding of ordered liberty. "Attempts to justify abortion through appeals to a broader right to autonomy and to define one's 'concept of existence' . . . could license fundamental rights to illicit drug use, prostitution, and the like."
- ▶ Noting "the critical moral question posed by abortion," the justices compared their decision to *Brown v. Board of Education* in overruling *Plessy v. Ferguson*, which "was also egregiously wrong." *Roe* conflated the right to shield information from disclosure and the right to make and implement important personal decisions without governmental interference and produced a scheme that "looked like legislation," including a "glaring deficiency" in failing to justify the distinction it drew between pre- and post-viability abortions. The subsequently-described "undue burden" test is unworkable in defining a line between permissible and unconstitutional restrictions. Traditional reliance interests are not implicated because getting an abortion is generally an "unplanned activity," and "reproductive planning could take virtually immediate account of any sudden restoration of state authority to ban abortions." The Court emphasized that nothing in this opinion should be understood to cast doubt on precedents that do not concern abortion.
- ▶ Mississippi's Gestational Age Act is supported by the Mississippi Legislature's specific findings, which include the State's asserted interest in "protecting the life of the unborn."

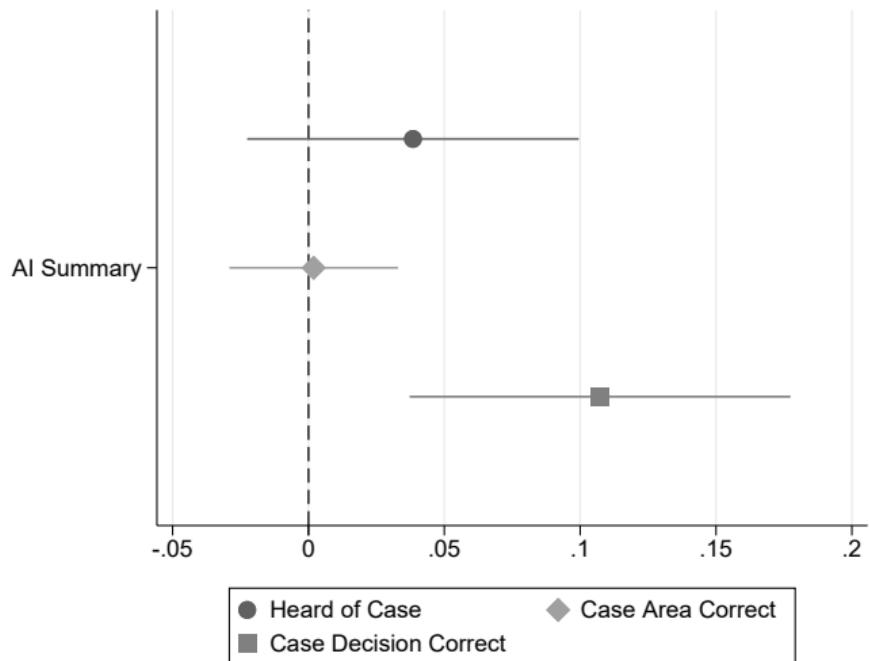
Justia Summaries are Shorter but Less Readable



Questions: Knowledge

1. Have you heard about this case before?
2. Which of these five areas of law describes the case best?
 - ▶ select from list of 5; correct 96% on average
3. Whose opinion does this represent? Supreme Court justices who voted...
(direction of decision, depending on topic)
 - ▶ e.g., ... “In favor of allowing governments to restrict abortion”; “In opposition to unions relative to employers”
 - ▶ correct 74% on average

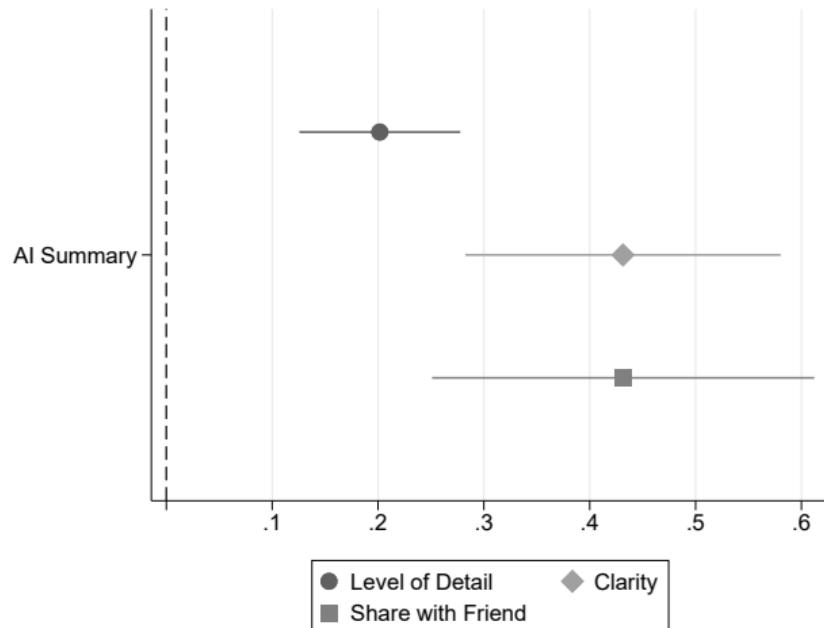
Results: Understanding



Questions: Subjective Quality

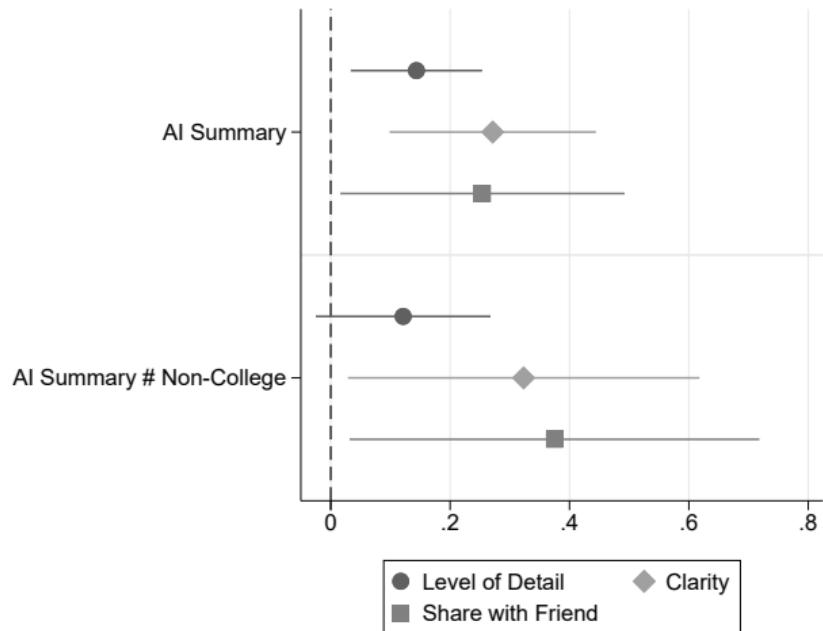
1. How do you rate the level of detail in the text? (too little, just right, too much)
2. How well does the text explain its main points or ideas? (4 point scale)
3. Would you forward this text to an interested friend or relative to help them understand this case? (5 point scale)

Results: Quality



Results: Quality × Low Education

Results: Quality × Low Education



Overview

- ▶ Our summaries improve accessibility and knowledge about the case.
- ▶ How does knowledge translate to attitudes?

Outline

Bias in Language Models

Legal NLP

Summarization

Information Retrieval / RAG

Open Question Answering and Claim Verification

- ▶ Open question answering:
 - ▶ Answer any question.
 - ▶ “What are the responsibilities of the mayor of Zurich?”
- ▶ Open claim verification:
 - ▶ Check whether a plain-text claim is true or false.
 - ▶ “Zurich has the second-highest per-capita income of any city in Europe.”

Open Question Answering and Claim Verification

- ▶ Open question answering:
 - ▶ Answer any question.
 - ▶ “What are the responsibilities of the mayor of Zurich?”
- ▶ Open claim verification:
 - ▶ Check whether a plain-text claim is true or false.
 - ▶ “Zurich has the second-highest per-capita income of any city in Europe.”
- ▶ Both problems are solved using information retrieval pipelines:
 - ▶ search large corpora or knowledge graphs for evidence
 - ▶ use evidence to answer the question or check the claim

Information Retrieval for Question Answering

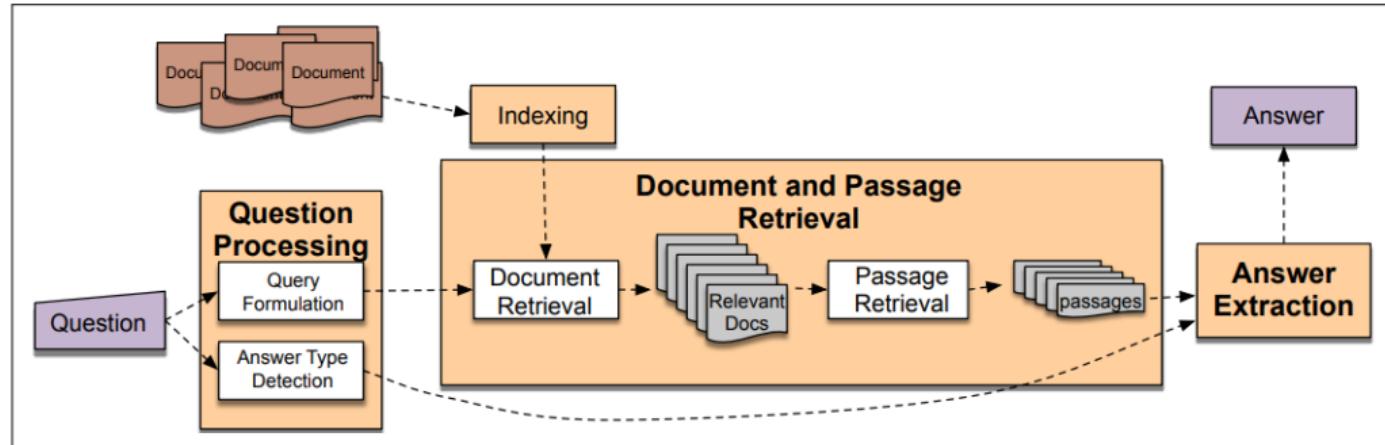


Figure 25.2 IR-based factoid question answering has three stages: question processing, passage retrieval, and answer processing.

- ▶ e.g., IBM Watson is a fast search engine over a knowledge base.

Automated Claim Verification

Claim (*by Minister Shailesh Vara*)

“The average criminal bar barrister working full-time is earning some £84,000.”

Verdict: FALSE (*by Channel 4 Fact Check*)

The figures the Ministry of Justice have stressed this week seem decidedly dodgy. Even if you do want to use the figures, once you take away the many overheads self-employed advocates have to pay you are left with a middling sum of money.

1. Claim spotting (what to fact check – facts vs opinions, etc)
2. Evidence retrieval
3. Evidence filtering
4. Fact-check claim given evidence (textual entailment)

Information Retrieval / RAG

- ▶ For open question answering and automated claim verification,
 - ▶ Input is a plain text query (a question or a claim).
- ▶ These are the two main examples of the broader problem of Retrieval Augmented Generation (RAG) in NLP.

Information Retrieval / RAG

- ▶ For open question answering and automated claim verification,
 - ▶ Input is a plain text query (a question or a claim).
- ▶ These are the two main examples of the broader problem of Retrieval Augmented Generation (RAG) in NLP.
- ▶ The current approach:
 - ▶ vectorize the query
 - ▶ vectorize all the documents in the knowledge base
 - ▶ retrieve relevant documents as those with the highest cosine similarity to the query.

RAG: How to vectorize the documents

- ▶ Traditional approach is **bm25**:
 - ▶ roughly, rank all documents by their TF-IDF cosine similarity to the query.
 - ▶ index every document by which words it contains, to quickly filter out irrelevant documents.

RAG: How to vectorize the documents

- ▶ Traditional approach is **bm25**:
 - ▶ roughly, rank all documents by their TF-IDF cosine similarity to the query.
 - ▶ index every document by which words it contains, to quickly filter out irrelevant documents.
- ▶ Recent systems (e.g. LlamaIndex) use embeddings:
 - ▶ e.g. mp-net embeddings from SentenceTransformers.

RAG: How to vectorize the documents

- ▶ Traditional approach is **bm25**:
 - ▶ roughly, rank all documents by their TF-IDF cosine similarity to the query.
 - ▶ index every document by which words it contains, to quickly filter out irrelevant documents.
- ▶ Recent systems (e.g. LlamaIndex) use embeddings:
 - ▶ e.g. mp-net embeddings from SentenceTransformers.
 - ▶ deal with short context by segmenting documents into overlapping chunks

RAG: How to vectorize the documents

- ▶ Traditional approach is **bm25**:
 - ▶ roughly, rank all documents by their TF-IDF cosine similarity to the query.
 - ▶ index every document by which words it contains, to quickly filter out irrelevant documents.
- ▶ Recent systems (e.g. LlamaIndex) use embeddings:
 - ▶ e.g. mp-net embeddings from SentenceTransformers.
 - ▶ deal with short context by segmenting documents into overlapping chunks
 - ▶ then can do fast approximate search over dense vectors (e.g. Faiss, Johnson et al 2017)

RAG is an active research area

- ▶ bm25 is much faster, but requires exact word overlap.

RAG is an active research area

- ▶ bm25 is much faster, but requires exact word overlap.
- ▶ can do bm25 as a first step, then use embeddings to rank top results.

RAG is an active research area

- ▶ bm25 is much faster, but requires exact word overlap.
- ▶ can do bm25 as a first step, then use embeddings to rank top results.
- ▶ can use an LLM to remove false positives – “remove irrelevant snippets” – and duplicates – “remove redundant snippets”.

RAG is an active research area

- ▶ bm25 is much faster, but requires exact word overlap.
- ▶ can do bm25 as a first step, then use embeddings to rank top results.
- ▶ can use an LLM to remove false positives – “remove irrelevant snippets” – and duplicates – “remove redundant snippets”.
- ▶ can use an LLM to suggest modified / alternative queries.

RAG is an active research area

- ▶ bm25 is much faster, but requires exact word overlap.
- ▶ can do bm25 as a first step, then use embeddings to rank top results.
- ▶ can use an LLM to remove false positives – “remove irrelevant snippets” – and duplicates – “remove redundant snippets”.
- ▶ can use an LLM to suggest modified / alternative queries.
- ▶ still to be investigated:
 - ▶ embedding algorithms that use other information besides document content, e.g. metadata, citation networks.
 - ▶ info retrieval systems as recommendation engines

Inference Step: QA/CV

- ▶ Question answering:
 - ▶ take retrieved evidence as the context passage, and do local question answering

Inference Step: QA/CV

- ▶ Question answering:
 - ▶ take retrieved evidence as the context passage, and do local question answering
- ▶ Claim verification:
 - ▶ take retrieved evidence as the premise, and the claim as the hypothesis, and do textual entailment.

Knowledge Graphs (?)

- ▶ A structured graph representing facts (and assertions?) as tuples.
- ▶ Entities are nodes, relations are edges:
 - ▶ (head entity, relation, tail entity)

Knowledge Graphs (?)

- ▶ A structured graph representing facts (and assertions?) as tuples.
- ▶ Entities are nodes, relations are edges:
 - ▶ (head entity, relation, tail entity)
- ▶ E.g., DBpedia: crowd-sourced effort to extract structured information from Wikipedia and make it available as linked open data.

GENERATING FACTS FOR THE ENTITY BILLIE HOLIDAY

“Facts” as RDF Triples



Subject Predicate Object
(Thing)



S <http://dbpedia.org/resource/Billie_Holiday>
P <<http://xmlns.com/foaf/0.1/name>>
O "Billie Holiday"