Logical Fallacies in Supreme Court Rulings

Haley Johnson
haleyej@umich.edu

1 Introduction

Institutional trust in the Supreme Court of the United States (SCOTUS) is at historic lows. Recent polling from Gallup found that just 49% of American report having "a great deal" or a "fair amount" of trust in the nation's highest court, down from 68% just one year ago (Brenan, 2024). Recent reporting on the Justice Clarence Thomas's failure to disclose his financial relationship with billionaire Harlan Crow (Kaplan et al., 2023), Justine Samuel Alito's relationship with Republican Donors (Elliott et al., 2023), and perceived political motivations behind recent rulings (Benshoff, 2023) have all contributed to declining public confidence.

Supreme Court decisions set precedent, establish lines of reasoning, and develop legal tests that are subsequently used by lower courts. For instance, the case *Miranda v. Arizona* effectively expanded the protections of the Fifth Amendment by ruling that a person must be notified of their constitutional rights prior to an interrogation (Schrock et al., 1978). In another notable instance, attorneys representing then-presidential candidate George W. Bush argued for the Independent State Legislator Theory in *Bush v. Gore*, a controversial legal theory that holds that state legislators have complete authority of over elections in that state and cannot be overseen by state courts or governors. In a concurrence, three justices partially endorsed the Independent State Legislator Theory, raising the ideas profile in conservative legal circles (Marisam, 2022). The Independent State Legislator Theory was the basis for several of Donald Trump's challenges to the 2020 election and plot to prevent electors from casting ballots on January 6th (Marisam, 2022). Consequentially, the evidence and reasoning used in SCOTUS opinions has significant implications for jurisprudence, the practical application of the law, and for which ideas are considered legitimate legal theories.

This also makes the justice's reasoning a prime concern for those invested in the court's legitimacy and public perception. Motivated or illogical reasoning — whether real or perceived — contribute to the growing distrust in the courts. For instance, some critics have argued that the courts landmark 1973 decision *Roe v. Wade* "invented" a right to an abortion, while others have responded that this argument relies on an overly narrow reading of the Fourteenth Amendment's substantive due process clause (Mitchell, 2023).

1.1 Project Goals

A logical fallacy is an argument "that appears correct and may even be extremely persuasive, but which proves upon closer inspection to be logically invalid" (McClurg, 2010). In Supreme Court opinions, logically fallacies are a strong indicator that a justice's argument may be inconsistent with existing jurisprudence or the facts of a case.

This project seeks to investigate 1) the prevalence of logical fallacies and motivated reasoning in SCOTUS decisions and 2) if logical fallacies occur more frequently in decisions pertaining to controversial or political divisive issues, such as abortion, the Second Amendment, voting rights, religious freedom, or LGBT marriage. This work has implications for individuals interested in the evolution of the court, how partisan pressures influence decisions, and the overall quality of jurisprudence.

Additionally, I also aim to improve upon existing methods for detecting logical fallacies and evaluating reasoning ability in natural language. These improvements can be leveraged for other NLP tasks, such as detecting bias in news or propaganda in political communications.

2 Task Definition

This project will use transfer learning to evaluate logical reasoning in Supreme Court opinions.

Transfer learning is a "a technique where a neural network is fine-tuned on a specific task after being pre-trained on a general task" Malte and Ratadiya (2019)

Transfer learning offers several advantages. First, extensive work has been done on logical fallacy detection in other domains, such as climate denial (Jin et al., 2022), propaganda detection (Oliinyk et al., 2020), and hyper partisan news detection (Kiesel et al., 2019). Secondly, the Supreme Court typically decides on just 50-70 cases per year, creating a relatively small corpus. Transfer learning will allow me to train a logical fallacy detection model using data from other domains, and then fine tune is on a dataset of SCOTUS opinions.

In particular, this project will use multiple fine-tuning on a pre-trained model (e.g. BERT). First, the model will be fine-tuned to learn to detect logical reasoning. Then, it will be fine-tuned to identify logical fallacies. I'll use low-rank adaptation of large language models (LoRA), a procedure developed by Hu et al. (2021) that allows for fine-tuning without diverging too much from the original model. By combining logical reasoning and logical fallacy detection, I believe I can improve upon existing methods for evaluating reasoning abilities in large language models.

Finally, I'll apply the resulting model to a corpus of Supreme Court opinions to examine 1) the overall prevalence of logical fallacies and 2) if fallacies are more prevalent in opinions from certain justices or about certain issues. The corpus of Supreme Court opinions includes majority opinions, concurrences (when justices sign onto the majority's conclusion, but use different reasoning or want to highlight different jurisprudence), and dissents (when justices disagree with the majority and explain their disagreement).

3 Data

3.1 Logical Reasoning Data

This project leverage two datasets to evaluate logical reasoning in large language models: LogicNLI, a dataset of 30,000 statements and if they hold logically developed by (Tian et al., 2021), and a dataset of 1,700 logical fallacies developed by (Jin et al., 2022).

LogicNLI contains examples of logical entailment, contradiction, neural statements (where one statement does entail or imply another), and paradoxes. Individual training examples in LogicNLI include a mix of facts, rules, and statements Tian et al. (2021). The dataset is avaliable for download on github and has already been split into train and test sets by the original authors.

Logical fallacy examples were scrapped from educational websites and include 13 types of reasoning errors:

- faulty generalizations
- ad hominems
- ad populums
- false causality
- circular claims
- appeal to emotion
- fallacy of relevance
- deductive fallacies
- intentional fallacies
- fallacies of extensions
- false dilemmas
- fallacies of credibility
- equivocation

The dataset is available for download on github. The original authors have already separated the data into testing, training and development sets. The test set will be used in the evaluation portion of this project.

3.2 Supreme Court Data

While there are many Supreme Court databases available, I was not able to identify any dataset that included 1) the full text of all opinions related to a case, 2) the authors of all opinions 3) the core issues the case was deciding on (e.g. property law, the equal protection clause).

JUSTIA maintains a free, comprehensive database of Supreme Court opinions. I'll manually download text from JUSTIA. Since it is infeasible to download data for every case myself, I'll restrict my SCOTUS corpus to cases from 1950 to the present. For each court term, I'll select up to 5 of the most notable cases and 10 randomly chosen ones. This will ensure my dataset includes all major decisions from the modern court and a representative sample of other cases. I will use the labels from JUSTIA to categorize cases by issue. I have already began identifying which cases I'll download from each term and anticipate it will only take a few hours to go through JUSTIA and manually download opinions.

4 Related Work

4.1 Evaluating Logical Claims using NLP

Jin et al. (2022) adopt a "structure aware approach" for detecting logical fallacies about climate change. Their approach, which draws inspiration from philosophy and the study of logic, holds that the *structure* of the argument, rather than the content, is the best way to identify faulty reasoning. The author's selected several pre-trained baseline models and fine-tuned them on data about logical fallacies, achieving 57% accuracy. Similar to this project, the authors were working with a small dataset of 1,700 logical fallacy examples and 1,000 fallacious claims about climate change.

Tian et al. (2021) fine-tune three state of the art language models (BERT, RoBERTa, Xl-Net) on the LogicNLI corpus. They find that all three models perform better than random guessing, but are worse than human annotators. RoBERTa performed the best, achieving 68.3% accuracy, while BERT achieved 55.9% accuracy. All three model's performance significantly degraded when more irrelevant information was introduced. This indicates that these models may struggling to pick out relevant information when evaluating claims in longer or more complex texts.

Notably, the authors found that training the models to detect paradoxical claims significantly improved first-order logical reasoning and that the resulting models were less likely to pick up spurious correlations in the data. Therefore, fine-tuning a model on LogicNLI and then performing additional fine-tuning on (Jin et al., 2022)'s logical fallacy corpus may result in a more robust model that's better able to evaluate complex arguments.

4.2 Legal Scholarship

Legal writing is highly technical and contains domain specific language. As a result, pre-trained models may struggle to generalize to legal writing and other specialized contexts. Chalkidis et al. systemically explore strategies for adapting pre-trained to legal context and perform an extensive hyper-parameter search to create LEGAL-BERT, a fine tunned version of BERT design for computational law. LEGAL-BERT shows significant performance gains over BERT fine-tuned on legal documents in a variety of tasks. While neither of the evaluation tasks were specifically related to legal reasoning, the author's results indicate that LEGAL-BERT may perform better than language models that aren't fine tuned on legal tasks. I may explore using LEGAL-BERT in my project.

Logical fallacies and the rhetoric of the court has been written about extensively in the legal sphere (see McClurg (2010) and Saunders (1993)). However, these analyses have relied on the author's domain and are limited of scope. To my knowledge, there has been no systematic review of fallacies or illogical reasoning in a substantial corpus of SCOTUS opinions. While current NLP systems cannot replace the expertise of judicial scholars, they can allow researchers to study more extensive corpus's of opinions and illuminate potential inconsistencies, issues, or misapplications of jurisprudence that would have otherwise gone unnoticed.

To evaluate the accuracy of my model on the SCOTUS corpus, I'll create a small test set of the logical fallacies identified by McClurg (2010). The author has already identified the fallacious claims and where they occur in various opinions — all I'll need to do is go through the text and create a dataset. While this is a relatively small corpus of logical fallacies in SCOTUS opinions, these are instance identified by legal scholars, making these higher quality annotations than ones I could generate myself.

5 Evaluation

5.1 Evaluation Metrics

The LogicNLI and logical fallacies datasets both have test sets created by the original authors. I'll use these to evaluate my model. To test how well this method performs on the SCOTUS corpus, I'll use the small test set detailed in the previous section.

Both datasets contain examples of different kinds of logical reasoning / errors. However, this project is just interested in if a claim is illogical. Therefore, I will be evaluating a binary classification task and using accuracy and F1 scores are my evaluation metrics.

5.2 Baselines

I'll compare my model's performance against two baselines: random performance and a simple naive bayes classifier.

Additionally, I am trying to asses if a second fine-tuning step focused on logical errors will improve the model's performance. Therefore, I'll pre-train one language model on just the LogicNLI dataset and another language model on just the logical

fallacy dataset, and then asses if the model trained using both datasets achieves higher accuracy.

6 Work Plan

This is a rough timeline for the remainder of the semester

- **By February 21:** Finish downloading SCOTUS dataset from JUSTIA
- **By February 28:** Pre-train model on LogicNLI dataset, evaluate results
- **By March 6:** Pre-train model on logical fallacy dataset, evaluate results
- **By March 16:** Implement LoRA to pre-train model on LogicNLI and logical fallacies dataset, evaluate results
- **By March 23:** Work on analysis of Supreme Court corpus
- **By April 3:** Continue work on analysis of Supreme Court corpus
- **April 8:** Project update due
- **April 15:** Write report
- **April 20:** Flex time in case other components of the project take longer than anticipated, continue to work on report
- **April 26:** Project report due

References

Laura Benshoff. 2023. Most americans say overturning roe was politically motivated, npr/ipsos poll finds.

Megan Brenan. 2024. Views of supreme court remain near record lows.

Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. 2020. LEGAL-BERT: The muppets straight out of law school. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2898–2904, Online. Association for Computational Linguistics.

Justin Elliott, Joshua Kaplan, and Alex Mierjeski. 2023. Alito took unreported luxury trip with gop donor paul singer.

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

Zhijing Jin, Abhinav Lalwani, Tejas Vaidhya, Xiaoyu Shen, Yiwen Ding, Zhiheng Lyu, Mrinmaya Sachan, Rada Mihalcea, and Bernhard Schoelkopf. 2022. Logical fallacy detection. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 7180–7198, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Joshua Kaplan, Justin Elliott, and Alex Mierjeski. 2023. Clarence thomas secretly accepted luxury trips from gop donor.

Johannes Kiesel, Maria Mestre, Rishabh Shukla, Emmanuel Vincent, Payam Adineh, David Corney, Benno Stein, and Martin Potthast. 2019. SemEval-2019 task 4: Hyperpartisan news detection. In *Proceedings of the 13th International Workshop on Semantic Evaluation*, pages 829–839, Minneapolis, Minnesota, USA. Association for Computational Linguistics.

Aditya Malte and Pratik Ratadiya. 2019. Evolution of transfer learning in natural language processing. *ArXiv*, abs/1910.07370.

Jason Marisam. 2022. The dangerous independent state legislature theory.

Andrew Jay McClurg. 2010. Logical fallacies and the supreme court: A critical analysis of justice rehnquist's decisions in criminal procedure cases.

Jonathan F. Mitchell. 2023. Why was roe v. wade wrong?

Vitaliia-Anna Oliinyk, Victoria Vysotska, Yevhen Burov, Khrystyna Mykich, and Vítor Basto Fernandes. 2020. Propaganda detection in text data based on nlp and machine learning. In *MoMLeT+DS*.

Kevin W. Saunders. 1993. Informal fallacies in legal argumentation.

Thomas S Schrock, Robert C Welsh, and Ronald Collins. 1978. Interrogational rights: Reflections on miranda v. arizona. *S. Cal. L. Rev.*, 52:1.

Jidong Tian, Yitian Li, Wenqing Chen, Liqiang Xiao, Hao He 0007, and Yaohui Jin. 2021. Diagnosing the first-order logical reasoning ability through logicnli. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 3738–3747. Association for Computational Linguistics.