SCOTUS-Net

by F. A. Vagnoni
(fvagnoni.ieu2021@student.ie.edu)

September, 2025

Table of Contents

1. Executive Summary

In response to challenges in predicting the outcomes of cases before the Supreme Court of the United States (SCOTUS), we introduce SCOTUS-Net, a zero-shot legal learner designed to overcome the major obstacles faced by prior models: Data drift and Cold-start. Unlike existing approaches, which are optimized on historical voting patterns, SCOTUS-Net leverages biographical information to infer judicial tendencies. This enables accurate predictions even when facing newly appointed Justices with no prior voting record, or in hypothetical Court compositions.

The system integrates a dual-pipeline architecture, based on sentence transformers, to process both case descriptions and Justice biographies, with a pretraining stage ensuring that truncated biographies encode tenure-relevant information. Experiments across more than nine thousand cases from the period (1946-2023) demonstrate that SCOTUS-Net maintains robustness under distribution shifts in the legal and political landscape. While limitations remain, particularly in explainability and per-Justice granularity, our results suggest that biographical embeddings offer a powerful and scalable foundation for predictive legal modeling.

2. Motivation

In June 2025 The Economist introduced SCOTUSbot [1], a deep learning model that predicts Supreme Court (SCOTUS) case outcomes from a case description. Its architecture assigns a per-Justice score and aggregates these into a Court-level prediction. A useful property of this design is interpretability: Extreme scores (v.g.: 0/10 or 10/10) indicate higher model certainty for a Justice, while mid-range scores (v.g.: 5/10) reflect uncertainty. This enables per-Justice analysis of tendencies and the identification of swing-vote scenarios.

However, SCOTUSbot faces two major challenges: Data drift and cold-start. Because it learns to predict the votes of specific Justices, any change in the Court composition shifts the conditional distribution the model relies on, a Court-level drift. At the Justice level, a newly appointed Justice lacks vote history, making immediate retraining unfeasible. Consequently, a new appointment can disable the model until sufficient new data accrues.

To address these challenges, we propose an architecture capable of predicting case outcomes involving previously unseen Justices, that is, a zero-shot approach. Beyond newly appointed Justices, this architecture allows the evaluation of what-if scenarios (v.g.: Testing how the case outcome shifts by including or omitting a specific Justice, even if that person has never served in SCOTUS). The architecture is based on the hypothesis:

> “The biography of a SCOTUS Justice includes relevant information that will determine the outcome of his vote.”

Intuitively, background and professional experience shape a person’s views. Knowing a Justice’s biography should therefore guide general positions and the perspective brought to a case.

Formally, SCOTUS-Net and SCOTUSbot would not be trained on the same problem, but to learn different conditional distributions. The former is trained to learn the probability of the court outcome (O_C) given the conjoined distribution of the case description (C) and the set of Justice biographies ($\{j_i\}_{i=0}^{i=J}$) (Formula 2.1); while the latter is trained to learn the probability of justice outcome (O_j) given the conjoined distribution of the case description (C) and the Justice identity (I) (Formula 2.2), which limits the model to learn from this Justice’s historical votes.

SCOTUS-Net Objective: $\mathrm{P}(O_C|(C,\{j_i\}_{i=0}^{i=J}))$ (Formula 2.1)

SCOTUSbot Objective: $\mathrm{P}(O_j|(C,I))$ (Formula 2.2)

3. Data Pipeline

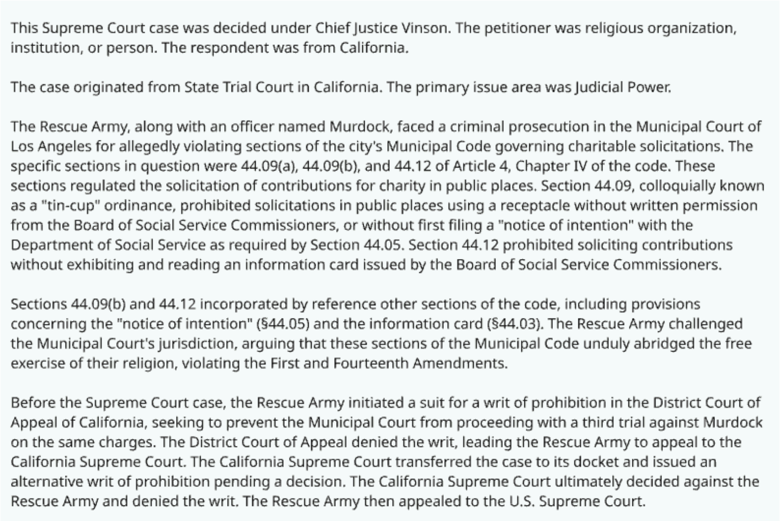
3.1. Sources

The data sources include:

- **Justice biographies**: extracted from each Justice’s Wikipedia page [2]. Includes all Justices who have served on SCOTUS up to 2025.
- **The Supreme Court Database (SCDB)**, by Washington University [3]: used to construct a dataset with case outcomes, identifiers, and participating Justices; covers cases from 1946 to 2023.
- **Justia**, by the U.S. Supreme Court [4]: provides case opinions from which case descriptions are extracted.

3.2. Collection & Cleaning

All data pass through a two-stage pipeline: raw to processed. Raw biographies are scraped from Wikipedia and saved as .txt files. Raw opinions are retrieved from Justia using the SCDB identifier. Biographies are then truncated to avoid data leakage (details in §Splits & Leakage Control). Case descriptions are created by prompting the Gemini 2.5 API [5] with minimal temperature to reduce hallucinations. Figure 3.1 contains an example of a case description generated by this method.



This Supreme Court case was decided under Chief Justice Vinson. The petitioner was religious organization, institution, or person. The respondent was from California.

The case originated from State Trial Court in California. The primary issue area was Judicial Power.

The Rescue Army, along with an officer named Murdock, faced a criminal prosecution in the Municipal Court of Los Angeles for allegedly violating sections of the city's Municipal Code governing charitable solicitations. The specific sections in question were 44.09(a), 44.09(b), and 44.12 of Article 4, Chapter IV of the code. These sections regulated the solicitation of contributions for charity in public places. Section 44.09, colloquially known as a "tin-cup" ordinance, prohibited solicitations in public places using a receptacle without written permission from the Board of Social Service Commissioners, or without first filing a "notice of intention" with the Department of Social Service as required by Section 44.05. Section 44.12 prohibited soliciting contributions without exhibiting and reading an information card issued by the Board of Social Service Commissioners.

Sections 44.09(b) and 44.12 incorporated by reference other sections of the code, including provisions concerning the "notice of intention" (§44.05) and the information card (§44.03). The Rescue Army challenged the Municipal Court's jurisdiction, arguing that these sections of the Municipal Code unduly abridged the free exercise of their religion, violating the First and Fourteenth Amendments.

Before the Supreme Court case, the Rescue Army initiated a suit for a writ of prohibition in the District Court of Appeal of California, seeking to prevent the Municipal Court from proceeding with a third trial against Murdock on the same charges. The District Court of Appeal denied the writ, leading the Rescue Army to appeal to the California Supreme Court. The California Supreme Court transferred the case to its docket and issued an alternative writ of prohibition pending a decision. The California Supreme Court ultimately decided against the Rescue Army and denied the writ. The Rescue Army then appealed to the U.S. Supreme Court.

Figure 3.1.: Case Description of 331 U.S. 549

3.3. Label

Three labels are constructed from SCDB votes: In favor (F) of the petitioner, against (A) the petitioner, and absent (B). Labels are represented as a three-component probability vector that sums to one, modeling a discrete distribution over the three outcomes.

The mapping from SCDB [3] case disposition/outcome into the labels was:

- In favour of the petitioner (F):
  Stay, petition, or motion granted (1); reversed (3); reversed and remanded (4); vacated and remanded (5); affirmed and reversed/vacated in part (6); affirmed and reversed/vacated in part and remanded (7); vacated (8).
- Against the petitioner (A):
  Affirmed (2); petition denied or appeal dismissed (9).
- Unclear/other (B):
  Certification to/from a lower court (10); no disposition (11).

Two notes:

1. The (B) label is necessary because Justices are sometimes absent (v.g.: Retirement, death) or recuse themselves (v.g.: Conflicts of interest), and it is also used as default label in unclear case outcomes.
2. Votes are not uniformly distributed. In our sample (1946-2025), around 70% of the time the Court rules for the petitioner, yielding a relatively imbalanced dataset. This will drive certain decisions and considerations.

3.4. Splits & Leakage Control

There are numerous potential leakage routes when structuring the problem and the data. The following are the ones we have identified and their proposed mitigations:

1. **Opinions contain rulings**: Training on opinions would leak the ground truth. We therefore use opinions only to extract a non-leaky case description via Gemini 2.5 [5], restricting summaries to case presentation (v.g.: Petitioner, cause, etc) and not the ruling.
2. **Biographies include SCOTUS tenure facts**: Unavailable at inference time, especially under cold-start. We truncate biographies at the point of appointment to the Court.
3. **Temporal references and precedent**: Newer cases tend to cite older rulings in Common Law systems. To avoid training on information about test-era cases, we use time-based splits in which training is strictly older than validation, with a hold-out test set covering the most recent twenty-five years.
4. **Biographies that cite SCOTUS cases**: Even truncated biographies may reference SCOTUS decisions if the Justice's previous legal activity involved citing SCOTUS rulings as precedents, notably frequent in Common Law systems. Time-based splitting limits this risk for validation and test because the model trains on older Justices less likely to reference newer cases.

3.5. Ethical Note

Beyond acknowledging data providers, note that data were scraped. None of the sources explicitly prohibit scraping, but some discourage it. Given the small scope of the project, we proceed under the assumption of minimal impact, while respecting attribution and responsible access.

4. Model Design

4.1. SCOTUS-Net Architecture

Dual parallel pipelines, Justice-Bio and Case-Description, process inputs separately for each data point: a set of *n* Justice biographies and a single case description.

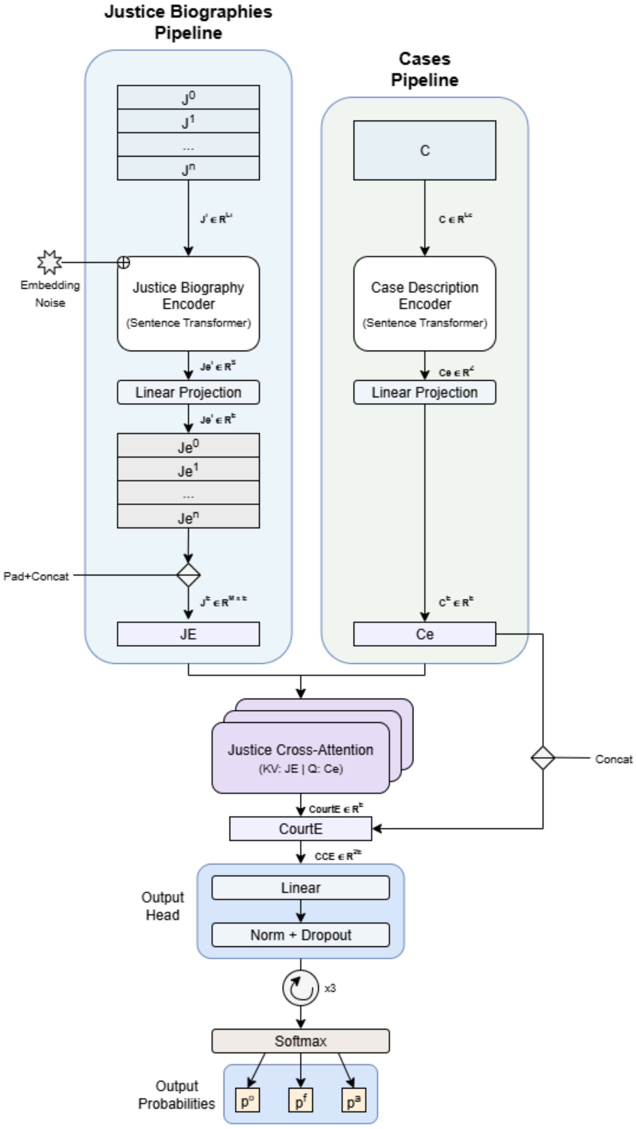


Figure 4.1.: SCOTUS-Net Architecture

- **Justice-Bio pipeline**: Input is a set {J_0, J_1, …, J□} of varying-length biographies. A pretrained sentence transformer [6][7] produces fixed-size vectors after the application of a NEFTune regularization at the embedding layer [8]. Each vector is then projected linearly to embedding dimension E. Concatenation yields an n × E tensor.

- **Case-Description pipeline**: The case description is encoded by a sentence transformer into length Z, then linearly projected to E, while the weights are not shared with the Justice projection.
- **Justice cross-attention**: Given the Justice tensor J^E and case embedding C^E, we apply cross-attention [9] using C^E as query to retrieve biography features most relevant to the case. The output is a Court-Embedding contextualized by the case. This is concatenated with C^E, a residual-style connection [10], to preserve case information.

The resulting 2E-dimensional vector is fed to a fully connected block with three output neurons and Softmax activation to predict the outcome distribution.

4.2. Justice Encoder Pretraining

SCOTUS-Net relies on two pretrained sentence transformers, one for the Justice-Bio pipeline and one for the Case-Description pipeline. Because zero-shot performance depends critically on the Justice encoder, we introduce a pretraining stage designed to ensure that a truncated biography encodes tenure-relevant information.

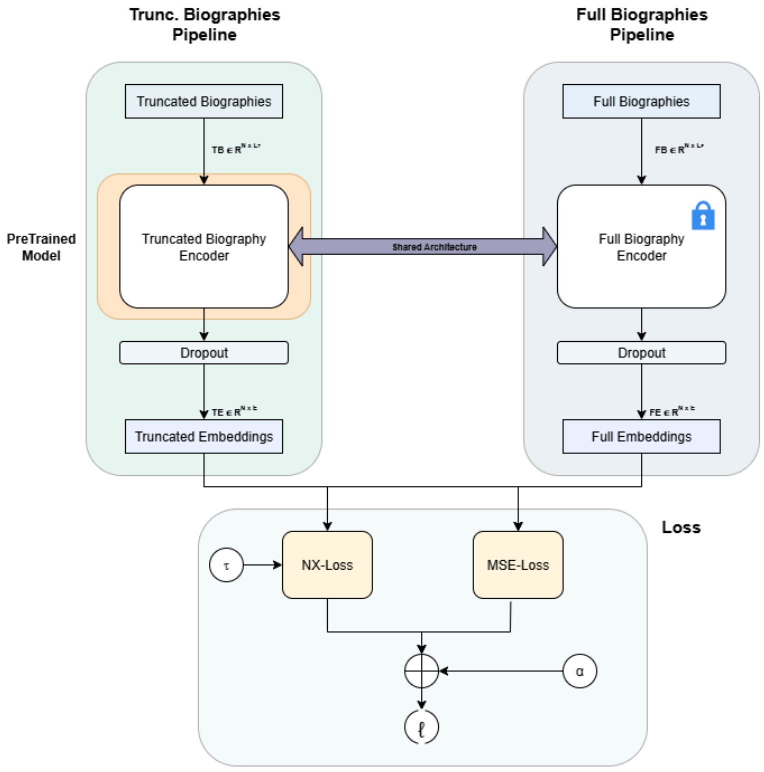


Figure 4.2.: Justice-Encoder Pretraining Task

The setup (Figure 4.2) involves two encoders with shared architecture and initialization:

- **T-Encoder (Truncated)**, which ingests biographies truncated at the time of appointment.
- **F-Encoder (Full)**, which ingests the complete biographies. Its weights are frozen.

The T-Encoder is optimized and later transferred into the Justice pipeline, while the F-Encoder serves as a fixed reference. The training objective is to bring the embeddings of the T-Encoder closer to those of the F-Encoder, under the assumption that the full biography contains relevant tenure information absent from the truncated version.

This is achieved by combining two losses: NT-Xent [11] and a similarity-based term, weighted together (details in §Loss Functions). In this way, the truncated encoder is explicitly encouraged to approximate the richer representational space of the full biography.

4.3. Loss Functions

SCOTUS-Net is optimized using a KL-divergence loss (Formula 4.1) between the predicted probability distribution and the true label distribution. This choice reflects the probabilistic nature of the task, where each label, (F, A, B), is modeled as a component of a three-class distribution. By minimizing KL-divergence, the model learns to align its output distribution with the ground-truth distribution rather than simply predicting the most likely class. However, to explore other possibilities and establish the usefulness of this loss function implementation, it was decided to use a MSE loss as baseline (More details §Experiments).

$$L_{KL}(p|q) = \frac{1}{B}\sum_{i=0}^{B}\sum_{j=0}^{3} q_i(j)\left(log\, q_i(j) - log\, p_i(j)\right) \quad \text{(Formula 4.1)}$$

For evaluation, we report F1-Macro across the output of the cases. This metric gives equal importance to each class, ensuring that minority outcomes are not overshadowed by the majority class. This is especially relevant given the previously stated class imbalance.

The Justice encoder pretraining employs the contrastive objective described §Justice Encoder Pretraining. Specifically, the T-Encoder is encouraged to approximate the embedding space of the F-Encoder through a combined loss (Formula 4.4): an NT-Xent contrastive term [11] to separate positive from negative pairs (Formula 4.2), and a similarity maximization term (v.g.: cosine similarity or MSE) (Formula 4.3). These are weighted together by hyperparameter α, which balances the contribution of alignment and similarity. This setup ensures that truncated biographies retain and express tenure-relevant information despite their reduced content.

$$L_{NT} = \frac{1}{B}\sum_{i=0}^{B}\left[- log\left(\frac{e^{s_i}}{e^{s_i}+(1-p)\sum_{j\neq i} e^{s_{ij}}}\right)\right] \quad \text{(Formula 4.2)}$$

$$L_{cos} = 1 - \frac{1}{B}\sum_{i=0}^{B} e_i^{(t)} \cdot e_i^{(f)} \quad \text{(Formula 4.3)}$$

$$L_{pretrain} = \alpha \cdot L_{NT} + (1 - \alpha) \cdot L_{cos} \quad \text{(Formula 4.4)}$$

4.4. Implementation & Training Details

For both pretraining and training, the virtual environment set up was Ubuntu 22.04 with Python 3.10 installed, and CUDA version 12.1.0 with access to a NVIDIA T4 Tensor Core GPU. Also, models were built utilising the last version of PyTorch [12] and optimized using AdamW [13], as well as leveraging HuggingFace's sentencetransformers library [14]. Finally, all tuning was carried out through Bayesian Search via the Optuna [15] package.

It must be noted that budgetary and time constraints on GPU usage limited the model training to 15 epochs and the hyperparameter optimization to 100 trials, both for pretraining and the main task. This means that there is potential for scaling the performance of both the Justice Encoder and SCOTUS-Net if more investment is made.

5. Experiments

5.1. Baseline for the Justice Encoder Pretraining

To assess the pretrained Justice encoder, we compare it against the un-finetuned sentence transformer. The fine-tuned encoder should outperform this baseline if it is indeed learning to anticipate tenure-relevant signals from truncated biographies.

5.2. Ablations & Sensitivities

To study the sensibility of the model to the presence of Justice metadata (v.g.: Nominated president, era of tenure, State), a version of the biographies was prepared for which this data was ablated. This would allow us to understand what this information's contribution to the overall test-time performance is.

Additionally, two training experiments were conducted: First, training the model with or without the pretrained Justice Encoder; and, secondly, training on different losses. As mentioned in §Model Design, both MSE and KL functions were taken as objective. The hypothesis is that the KL Divergence-trained model would outperform the MSE-trained one since it would teach the conditional distribution $\mathrm{P}(O_C|(C,\{J_i\}_{i=0}^{i=J}))$ better.

6. Results & Analysis

6.1. Pretraining

(Data scarcity was a central challenge: Only 116 individuals have served on SCOTUS. Training had fewer than 70 samples, and the test set contained only 20 biographies. Therefore, we emphasize 95% confidence intervals, as point estimates are less trustworthy at this scale.)

Contrastive pretraining of the Justice encoder yielded Mean Reciprocal Rank (MRR) ≈ 0.7 at test, considered high for a hard retrieval task. However, assuming a t-distribution over MRRs, the 95% CI, [0.5, 0.87], overlaps with the baseline's [0.22, 0.63], so we cannot claim a statistically significant improvement (Figure 6.1).

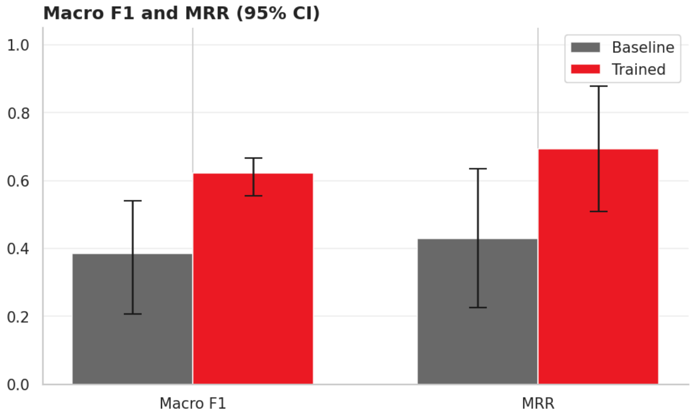


Figure 6.1

To explore information content beyond MRR, we trained a simple logistic regression on the encoded embeddings of both models to predict ideology (v.g.: Liberal, Moderate, Conservative; manually labelled). Because embeddings share dimensionality, this provides a fair proxy for usefulness in SCOTUS-Net. The Justice encoder outperformed the baseline in F1-Macro and Recall@k for k ≥ 6; for k ∈ {1,…,5} the point estimate favored the main encoder but CIs overlapped (Figure 6.3). Notably, F1-Macro CIs did not overlap, indicating a statistically significant difference, with category-level improvements of +0.56 (Liberal) and +0.16 (Conservative); the Moderate class remained at 0 for both (Figure 6.2).

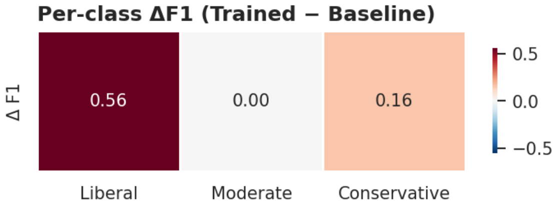


Figure 6.2

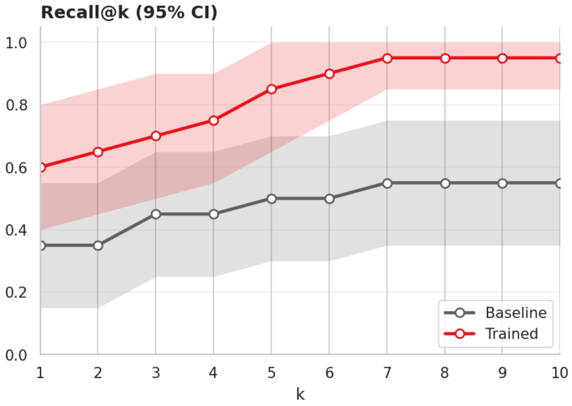


Figure 6.3

6.2. SCOTUS-Net

SCOTUS-Net was evaluated on 1,900 cases spanning the period (1997-2023). This ensured that the model was tested both on unseen cases and unseen Justices. Of the fifteen Justices appearing in the test set, only seven overlapped with the training set (v.g.: J. P. Stevens, S. D. O'Connor, W. Rehnquist, A. Scalia, A. Kennedy, C. Thomas, R. B. Ginsburg). The evaluation further included significant distributional shifts in the social and political landscape, such as the arc of equality of the LGBT movement and the onset of the War on Terror. These factors together constitute a particularly challenging test environment.

Model Experiments	MSE Error			KL Divergence			F1-Macro		
	Pretrained	No-Pretrained	Ablated	Pretrained	No-Pretrained	Ablated	Pretrained	No-Pretrained	Ablated
KL-Loss	**0.07**	0.1	0.8	**0.34**	0.36	0.35	**0.67**	0.65	0.65
MSE-Loss	0.08	0.09	0.09	0.38	0.4	0.41	0.59	0.55	0.58

Table 6.1.

Table 6.1 reports the results for the three experimental setups (v.g.: including the Pre-Trained Encoder; excluding the Pre-Trained Encoder; including the Pre-Trained Encoder while ablating the Justices' metadata), each optimized with either KL-Divergence Loss or MSE Loss. The models trained with KL-Divergence showed a slight advantage in learning the underlying conditional distribution, though the gap was not substantial. Still, across all three metrics, the configurations that integrated the Pre-Trained Encoder consistently achieved superior performance. And the model generally showed resilience when facing the ablated data, meaning that most information acquired is not derived simply from the Justices' metadata, but from the actual biographical features.

7. Limitations & Improvements

SCOTUS-Net, at its current implementation, has numerous limiting factors and fields of improvement. The following is a non-exhaustive list with proposals to address each.

Lack of Explainability: Despite the experiments carried out, SCOTUS-Net and the pretrained Justice Encoder remain mostly black-box models. This does not disable them, but makes their contributions limited. Improving the understanding of their weights and interactions is essential for building a more useful system. For example, the Justice Encoder's embeddings should be studied in detail to uncover the most relevant dimensions and to explore potential Justice-clusters or distances in the representational space.

Lack of Justice Granularity: As discussed in §Motivation, one of the advantages of SCOTUSbot [1] was its granular per-Justice predictions, where each Justice received a score. SCOTUS-Net, instead, was trained to predict overall Court rulings and vote distributions, but it lacks the ability to model uncertainty or specific Justice-level outcomes, as SCOTUSbot did. To achieve this, new data would have to be collected and the model retrained on that objective. It would be expected that performance would remain comparable, or potentially improve, under this formulation.

Better Baseline for SCOTUS-Net: The baseline model used in the current experiments is not trained on the same amount of data as SCOTUS-Net, which limits comparability. A stronger baseline would be to replace the Justice-Bio pipeline with a simple TF-IDF [16] vectorizer. This would allow testing the hypothesis that the more complex architecture of SCOTUS-Net indeed extracts more effective information than a vocabulary-based approach.

Potential Biography-Based Biases: SCOTUS-Net is trained to extract relevant facts about a Justice's biography for the given case. This creates a risk of systematic biases tied to Justice backgrounds. For example, the model could incorrectly generalize that Catholic Justices always vote (A) against the petitioner in sexual liberties cases. Auditing such risks is necessary to ensure the model does not reproduce or amplify spurious correlations.

Better Hard Negative Examples for Pretraining: The Justice Encoder was trained mainly with a contrastive objective, NT-Xent loss [11]. For each truncated and full biography pair, the negatives were all other full biographies. This design treats even very dissimilar bios as equivalent negatives, which weakens contrastive learning. The model would likely benefit from better crafted negatives, for instance by integrating a proxy similarity measure (v.g.: TF-IDF [16]) to weigh or select them more effectively.

8. Conclusions

This research introduced SCOTUS-Net, a new architecture for predicting Supreme Court outcomes that shifts the learning signal from historical votes to Justice biographies. This design enables zero-shot generalization to unfamiliar Justices and hypothetical Court compositions.

Experiments demonstrated that signals of Justices’ behavior can be extracted from biographical features, combined with case descriptions. New applications are thus enabled, such as sensibility studies and the inquiry on how backgrounds influence decisions at the highest level of the legal system. Limitations yet remain, mainly around interpretability, biases, and lack of per-Justice granularity.

Altogether, SCOTUS-Net progresses in the field of predictive legal modeling beyond outcome forecasting, offering a tool to explore judicial dynamics and institutional change.

References

[1]The Economist, “Meet SCOTUSbot, our AI tool to predict Supreme Court rulings,” *The Economist*, Jun. 04, 2025. https://www.economist.com/united-states/2025/06/04/meet-scotusbot-our-ai-tool-to-predict-supreme-court-rulings

[2]Wikipedia, “Wikipedia,” *Wikipedia.org*, Jan. 15, 2001. https://www.wikipedia.org/

[3]“The Supreme Court Database,” *scdb.wustl.edu*. http://scdb.wustl.edu/

[4]Justia, “US Supreme Court Center,” *Justia Law*, Mar. 20, 2019. https://supreme.justia.com/

[5]G. Comanici *et al.*, “Gemini 2.5: Pushing the Frontier with Advanced Reasoning, Multimodality, Long Context, and Next Generation Agentic Capabilities,” *arXiv.org*, 2025. https://arxiv.org/abs/2507.06261

[6]N. Reimers and I. Gurevych, “Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks,” *arXiv.org*, 2019. https://arxiv.org/abs/1908.10084

[7]“sentence-transformers/all-roberta-large-v1 · Hugging Face,” *Huggingface.co*, 2015. https://huggingface.co/sentence-transformers/all-roberta-large-v1

[8]A. Yadav, “Understanding and Improving Noisy Embedding Techniques in Instruction Finetuning,” *Openreview.net*, 2025. https://openreview.net/forum?id=AHhDpMMXtf#discussion

[9]A. Vaswani *et al.*, “Attention is All You Need,” *Advances in Neural Information Processing Systems*, vol. 30, pp. 5998–6008, 2017, Available: https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html

[10]K. He, X. Zhang, S. Ren, and J. Sun, “Deep Residual Learning for Image Recognition,” *openaccess.thecvf.com*, 2016. https://openaccess.thecvf.com/content_cvpr_2016/html/He_Deep_Residual_Learning_CVPR_2016_paper.html

[11]T. Chen, S. Kornblith, M. Norouzi, and G. Hinton, “A Simple Framework for Contrastive Learning of Visual Representations,” *proceedings.mlr.press*, Nov. 21, 2020. https://proceedings.mlr.press/v119/chen20j.html

[12]B. Steiner *et al.*, “PyTorch: An Imperative Style, High-Performance Deep Learning Library,” *openreview.net*, Sep. 2019, Available: https://openreview.net/forum?id=Byef6EBl8B

[13]I. Loshchilov and F. Hutter, “Decoupled Weight Decay Regularization,” arxiv.org, Nov. 2017, Available: https://arxiv.org/abs/1711.05101

[14]“sentence-transformers (Sentence Transformers),” *huggingface.co*. https://huggingface.co/sentence-transformers

[15]T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama, “Optuna: A Next-generation Hyperparameter Optimization Framework,” Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, Jul. 2019, doi: https://doi.org/10.1145/3292500.3330701.

[16]K. SPARCK JONES, “A STATISTICAL INTERPRETATION OF TERM SPECIFICITY AND ITS APPLICATION IN RETRIEVAL,” *Journal of Documentation*, vol. 28, no. 1, pp. 11–21, Jan. 1972, doi: https://doi.org/10.1108/eb026526.