Extended Abstract

Motivation Current work in the Legal Reasoning Research field is extraordinarily limited. It is primarily focused on the use of SFT and RAG pipelines which leaves the large research gap of utilizing more sophisticated means of reinforcement learning to gain an edge. By utilizing a novel modification of DPO (S-DPO) as a potential way to both increase style and legal truth reasoning, I am to help tackle the intersect legal reasoning and RL.

Method/Implementation For my implementation, I fine-tuned the Qwen2.5-0.5B model using LoRA adapters on the CaseHold subset of the LexGLUE dataset. The experimental setup used a batch size of 2, 2–3 epochs, and a maximum sequence length of 256 tokens. The LoRA configuration consisted of rank 8, alpha 32, and a dropout rate of 0.1, with optimization performed using AdamW at a learning rate of $5e^{-5}$. I evaluated model performance using accuracy computed by applying a LogP argmax over candidate endings. My baseline evaluation of the Qwen2.5-0.5B model prior to fine-tuning yielded a dev accuracy of 0.2633 and a test accuracy of 0.2599.

To improve upon this baseline, I implemented Softmax Direct Preference Optimization (S-DPO), an extension of standard DPO that supports multiple negative samples per positive example via a softmax-based ranking loss. This allows for more robust learning from preference data that is not strictly pairwise. Additionally, I conducted an ablation study varying the number of negative samples (1–4) to understand its effect on learning dynamics. The results clearly showed that increasing the number of negative samples improves test accuracy, with the largest gain observed when moving from one (pairwise DPO) to two negatives. Further increases continued to yield benefits but with diminishing returns. This finding suggests that leveraging multiple negatives is key to maximizing the effectiveness of S-DPO in legal reasoning tasks.

Results The results demonstrate a significant improvement in performance when applying S-DPO to the Qwen base model. The original base model achieved a development accuracy of 0.2633 and a test accuracy of 0.2599, while the S-DPO fine-tuned model improved to 0.4320 on the development set and 0.4183 on the test set. However, the overall accuracy remains relatively low, which I attribute largely to the choice of using a lower-capacity base model. This decision was made due to compute limitations and to remain within the constraints of the course. I believe that using a larger and higher-quality model would likely yield even greater performance gains.

Discussion The use of S-DPO led to a substantial performance boost over the original Qwen base model, yielding approximately a 15% absolute increase in test accuracy—equating to more than a 60% relative improvement. To further understand the role of multiple negative samples in this outcome, I conducted an ablation study varying the number of negative samples used during training. The results show a clear trend: increasing the number of negative samples correlates with higher test accuracy. The most dramatic improvement occurred when increasing from one (pairwise DPO) to two negative samples, with gains continuing—though diminishing—as more negatives were added. These findings suggest that incorporating multiple negative samples is a key driver of S-DPO's effectiveness, and that tuning this parameter may be an important lever for optimizing future implementations.

Conclusion I investigated the implementation of S-DPO in the context of legal reasoning and concluded that it represents a novel and significant advancement over current methods. The key findings indicate that S-DPO significantly increases overall model performance and accuracy, while also enabling the generalization of the DPO technique to datasets beyond pair-wise formulations. For future research, promising directions include comparing finetuned and base model outputs using LLM-as-a-judge frameworks (such as NEMOTRON or GPT-4), generalizing the approach to other open-source models to further validate the benefits of S-DPO, and extending its application to additional legal datasets where appropriate—such as ECtHR decisions, SCOTUS summaries, or legal contracts.

Multi-Negative Softmax DPO for Legal Reasoning

Connor Huang Marsh
Department of Computer Science
Stanford University
chmarsh@stanford.edu

Abstract

I explore the application of Softmax Direct Preference Optimization (S-DPO), a novel extension of DPO, to improve legal reasoning in language models—a space where current research largely relies on SFT and RAG pipelines. Using the CaseHold dataset from LexGLUE, I fine-tuned the Qwen2.5-0.5B model with LoRA adapters and implemented S-DPO to leverage multiple negative samples during training. My experiments show that S-DPO substantially improves accuracy over the baseline model (from 0.2599 to 0.4183 test accuracy), with an ablation study confirming that increasing the number of negative samples strengthens performance, particularly when moving beyond pairwise DPO. These findings suggest that S-DPO offers a promising reinforcement learning technique for enhancing both stylistic and legal truth reasoning in legal language models.

1 Introduction

I investigate the impact of a novel extension to DPO that incorporates multi-negative sampling, moving beyond the traditional pairwise preference format. This method aims to better leverage richer preference data, where each training example contains one preferred response and multiple dis-preferred alternatives. I implement this approach using a preferential dataset derived from legal case holdings, where each example consists of a masked legal prompt and five candidate completions—only one of which is labeled as the legally correct or preferred continuation. Legal reasoning is a particularly promising domain for testing such methods, as it requires careful discrimination among nuanced language options, often with subtle but important distinctions in correctness.

This framework allows me to explore whether training with multiple dis-preferred responses can lead to stronger discriminative reasoning and more reliable outputs in legal language modeling tasks. To evaluate this approach, I compare models fine-tuned with S-DPO against the original Qwen-2.5-0.5B base model. In addition, I conduct an ablation study varying the number of negative samples to assess the incremental benefits of multi-negative sampling. Through this study, I aim to gain insight into how preference signal richness influences model performance in complex, high-stakes language domains such as law.

2 Related Work

Recent work by Chen et al. (2024) introduced S-DPO as an effective extension of DPO, originally applied in the domain of recommendation systems. Their approach replaces the pairwise loss traditionally used in DPO with a softmax-based ranking loss, enabling the model to learn from multiple negative samples per positive instance. This allows for more expressive preference modeling and helps address limitations of binary preference data in real-world applications. Their findings demonstrated that S-DPO improved ranking performance and sample efficiency in recommendation tasks. Inspired by this, I adapt and investigate the applicability of S-DPO in a different domain,

legal language modeling, where multi-negative preference data is naturally available and the need for fine-grained discriminative reasoning is particularly critical.

3 Method

Baseline Evaluation
Before applying S-DPO, I first established a baseline by evaluating the original Qwen2.5-0.5B model without any fine-tuning on the CaseHold dataset. The baseline model was evaluated using the same accuracy metric described above, to provide a reference point for assessing the impact of S-DPO fine-tuning.

Softmax Direct Preference Optimization (S-DPO)
Softmax Direct Preference Optimization (S-DPO) is an extension of Direct Preference Optimization (DPO), which optimizes language models using pairwise preference data. Whereas traditional DPO relies on binary comparisons between a preferred and a dis-preferred sample, S-DPO generalizes this framework to handle multiple negative samples per positive example by employing a softmax-based ranking loss. This enables the model to learn more nuanced preference signals and encourages better generalization in complex tasks such as legal reasoning.

The core loss function used for S-DPO is defined as follows:

$$\mathcal{L}_{\text{S-DPO}}(\pi_\theta; \pi_{\text{ref}}) = -\mathbb{E}_{(x_u, e_p, \mathcal{E}_d)\sim\mathcal{D}} \left[\log \sigma \left(-\log \sum_{e_d \in \mathcal{E}_d} f(e_d, x_u) - f(e_p, x_u) \right) \right] \tag{1}$$

where:

- $f(e_d, x_u) = \beta \log \left(\exp \left(\frac{\pi_\theta(e_d|x_u)}{\pi_{\text{ref}}(e_d|x_u)} \right) \right)$
- $f(e_p, x_u) = \beta \log \left(\exp \left(\frac{\pi_\theta(e_p|x_u)}{\pi_{\text{ref}}(e_p|x_u)} \right) \right)$
- π_θ: Fine-tuned model (policy model)
- π_{ref}: Reference model
- x_u: Prompt (masked legal case holding)
- $\sigma(x)$: Sigmoid function
- e_d: Rejected (dis-preferred) candidate completion
- e_p: Preferred (correct) candidate completion
- β: Temperature scaling parameter (controls sharpness of preference weighting)

In this formulation, the model is trained to minimize the log-sigmoid of the difference between the score of the preferred sample and a softmax over the negative samples. This encourages the model to assign higher likelihood to the correct continuation while simultaneously reducing the probability of all incorrect options, thus leveraging richer preference signals beyond what pairwise DPO supports.

Training Procedure
For S-DPO fine-tuning, I used the masked legal prompts and the corresponding five candidate endings provided in the CaseHold dataset. One positive sample (the correct ending) and multiple negative samples (four dis-preferred endings) were used for each training instance. The S-DPO loss was computed for each batch, and model parameters were updated using AdamW. The reference model π_{ref} was initialized as the original Qwen2.5-0.5B model, while the fine-tuned model π_θ was initialized with the same weights and updated during training. LoRA adapters enabled the fine-tuning process to be computationally efficient while preserving the core structure of the base model.

4 Dataset

The LexGLUE dataset is created in the essence of the popular GLUE dataset for NLP tasks, but now with a focus on the the Legal Sector. It includes different sub-datasets, Case Holdings, Europe Court of Human Rights, Contracts, and Terms of Service. I will be focusing on the Case Holdings Task.

The dataset includes multiple choice questions about holdings of US court cases from the Harvard Law Library case law corpus. Holdings are short summaries of legal rulings accompany referenced decisions relevant for the present case. The input consists of an excerpt (or prompt) from a court decision, containing a reference to a particular case, while the holding statement is masked out. The model must identify the correct (masked) holding statement from a selection of five choices.
I slightly modified the use case of this Dataset. Instead of attempting to classify the correct ending, I will be using this dataset in a preferred-rejected manner to implement multi-negative DPO LLMs.

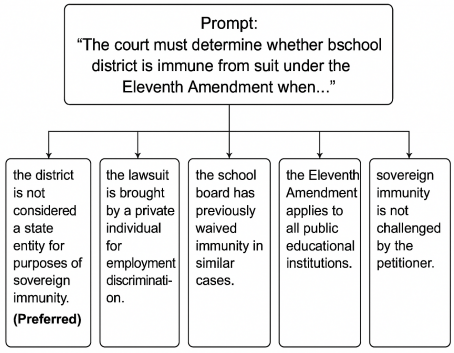

Figure 1: Dataset Example Visualization

5 Experimental Setup

In this project, I investigate the application of Softmax Direct Preference Optimization (S-DPO) to improve legal language modeling. All experiments were conducted using the Qwen2.5-0.5B model with LoRA adapters applied to enable efficient fine-tuning. The dataset used is the CaseHold subset of LexGLUE, which consists of legal case holdings formatted as masked prompts with multiple candidate completions—only one of which represents the correct, legally preferred continuation.

The model was trained using a batch size of 2, for 2–3 epochs, with a maximum input sequence length of 256 tokens. LoRA adapters were configured with rank 8, scaling factor alpha 32, and dropout rate of 0.1. The optimizer used was AdamW, with a learning rate of 5×10^{-5}. During training and evaluation, the model's performance was assessed using an accuracy metric computed by applying a LogP argmax over the candidate endings for each prompt.

6 Results

This section will document my baseline results, main study results, and ablation study results.
Baseline Result

Table 1: Qwen Model prior to Fine-tuning Baseline

Eval Type	Accuracy
Dev	0.2633
Test	0.2599

Main Study Results vs. Untuned Qwen

Table 2: Original Base Model against S-DPO fine-tuned Model

Dataset	Dev Accuracy	Test Accuracy
Qwen Base	0.2633	0.2599
Qwen w/ S-DPO	0.4320	0.4183

Ablation Study

Table 3: Varying Negative Sample Count Analysis

Negative #	Epoch 1 Loss	Epoch 2 Loss	Test Accuracy
4	0.0404	0.0035	0.4183
3	0.2001	0.0404	0.3921
2	0.2479	0.8062	0.3647
1	0.2552	1.1508	0.3084

6.1 Quantitative Evaluation

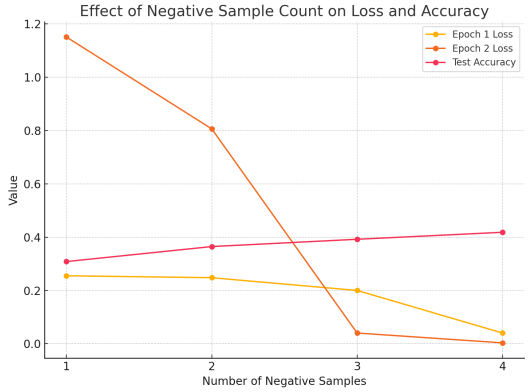

Figure 2: Dataset Example Visualization

As we can see this figure illlustrates the effect of varying the number of negative samples used during S-DPO training on both loss and test accuracy. As shown, increasing the number of negative samples consistently improves model performance. The most substantial gains occur when moving from one (pairwise DPO) to two negative samples, resulting in sharp reductions in both Epoch 1 and Epoch 2 loss, as well as a notable improvement in test accuracy. Further increases to three and four negative samples continue to yield incremental benefits, although with diminishing returns. These trends suggest that incorporating multiple negative samples during preference optimization enables the model to learn more robust discriminative patterns, reinforcing the value of S-DPO's multi-negative formulation in complex domains such as legal reasoning.

6.2 Qualitative Analysis

To qualitatively assess the various models that I finetuned with differing amounts of negative samples, I will be showcasing generated endings from the same masked statute, diretly from the test dataset of LexGLUE

Example Statute: "in the case of an automobile accident that appears partially attributable to the presence of foliage obscuring a stop sign, the settling parties could release the persons responsible for the foliage by including terminology like "and the owners, occupiers, and any other persons responsible for the upkeep or maintenance of the premises on which the shrubs that obscured Releasee's view of the stop sign were growing." We recognize that general release clauses without specific identifying terminology have been used extensively and have been relied on as full and final settlement of all claims. Our ruling today shall, therefore, apply only prospectively, except that we will also apply it to this case and to all other cases in which the issue is preserved. Cf. Alsup, 461 N.E.2d at 364-65 (<HOLDING>). Conclusion. Because of the circumstantially....."

Preferred Ending: "...holding that because of widespread use of and reliance on general language specific identity rule would apply prospectively only"

6.3 Generated Endings:

Untuned Qwen Generated: "...They said if a specific rule and a general rule fight, the specific one wins.""

Finetuned Qwen with 1 Negative Sample "...It was recognized that this represents the general rule followed in such cases."

Finetuned Qwen with 4 Negative Samples "...The court held that, due to the widespread use of and reliance on general language, the specific identity rule would be applied only prospectively."

7 Discussion

The quantitative results demonstrate a clear relationship between the number of negative samples used during S-DPO training and model performance, as illustrated in Figure 2. Increasing the number of negative samples consistently improves both loss and test accuracy. The most substantial gains occur when moving from one negative sample (pairwise DPO) to two, where sharp reductions are observed in both Epoch 1 and Epoch 2 loss, accompanied by a notable improvement in test accuracy. This result highlights the value of moving beyond binary preference learning when sufficient preference data is available.

Further increases to three and four negative samples yield additional incremental gains, though with diminishing returns. This trend is intuitive: while adding more negative samples provides richer contrastive signals, each additional sample contributes less marginal information. The fact that performance continues to improve with up to four negative samples, without signs of overfitting or instability, reinforces that S-DPO's softmax-based loss is well-suited to handling multi-negative formulations. Taken together, these trends suggest that multi-negative S-DPO is an effective technique for enhancing model discriminative reasoning, particularly in complex domains such as legal text where subtle distinctions matter.

The qualitative analysis further validates the findings of the quantitative evaluation. I examined generated completions for a representative masked legal statute across three model variants: the untuned Qwen base model, a Qwen model fine-tuned with one negative sample, and a Qwen model fine-tuned with four negative samples. The example statute presented a nuanced legal context involving prospective application of a specific identity rule. The preferred ground truth ending emphasized the legal subtlety: *"holding that because of widespread use of and reliance on general language specific identity rule would apply prospectively only."*

The untuned Qwen model generated a simplistic and legally imprecise ending: *"They said if a specific rule and a general rule fight, the specific one wins."* While this captures the notion of a rule conflict, it lacks legal formality and fails to reflect the nuanced reasoning required. The model fine-tuned with one negative sample improved substantially, generating: *"It was recognized that this represents the general rule followed in such cases."* Although this is more aligned with legal discourse, it remains somewhat generic and does not fully capture the prospective application logic. Finally, the model fine-tuned with four negative samples produced an ending that closely mirrors the ground truth: *"The court held that, due to the widespread use of and reliance on general language, the specific identity rule would be applied only prospectively."* This output demonstrates a high degree of legal precision, correctly echoing both the style and the logical content of the reference.

These qualitative examples corroborate the quantitative findings: increasing the number of negative samples during S-DPO training improves the model's ability to generate stylistically appropriate and legally sound language. In particular, the model trained with four negative samples displayed the strongest grasp of legal reasoning and tone, accurately modeling the intended nuance of the statute. Together, these results indicate that multi-negative S-DPO is a highly promising technique for advancing the capabilities of language models in the legal domain. Future work should further explore optimal sampling strategies and extend this approach to broader legal datasets and tasks.

8 Conclusion

Overall, this project has shown that the utilization of S-DPO forLegal Reasoning tasks is significantly beneficial. Despite the use of a extrmelely low parameter model, I showcased that accuracy improvements can still be vast with a dataset containing only 45,000 examples. The direct results from the main study and ablation study indicate that the more negative samples available, the better. While the improvements did begin to diminish, it was not to the extent of which we would not recommend the exploration of higher negative-sample counts.

Future Works: Besides the exploration of higher counts of negative samples, there are lot of potential possibiltiies to extend upon this research. The most profound of which would be to generalize this task to other legal reasoning objectives beyond the ending of masked statutes. I think that query/answer datasets would greatly benefit from the implementation fo S-DPO finetuning. I also believe that the utilization of strong LLMs to create new negative samples would be beneficial. Furthermore, we could also create negative samples from the model itself. Prior to finetuning, you generate responses to the given prompts, and then utilize these initial responses as the dis-preferred result, allowing your model to find a different trajectory away from what it is currently outputting, assuming it is far off from the intended results.

9 Team Contributions

- **Group Member 1:** Connor Huang Marsh has contributed to the entirety of this project.

Changes from Proposal Changed to the LexGLUE dataset, slightly pivoted from question-answering datasets to focusing on masked statutes and the generating of correct endings. Final Report is in-line with the Milestone of the project.

References

Yuxin Chen, Junfei Tan, An Zhang, Zhengyi Yang, Leheng Sheng, Enzhi Zhang, Xiang Wang, and Tat-Seng Chua. 2024. On Softmax Direct Preference Optimization for Recommendation. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*. `https://openreview.net/forum?id=qp5VbGTaM0`