Augmenting Legal Decision Support Systems with LLM-based NLI for Analyzing Social Media Evidence

Ram Mohan Rao Kadiyala *
University of Maryland
rkadiyal@umd.edu

Siddartha Pullakhandam *
University of Wisconsin
pullakh2@uwm.edu

Kanwal Mehreen
Traversaal.ai
kanwal@traversaal.ai

Subhasya Tippareddy
University of South Florida
subhasyat@usf.edu

Ashay Srivastava
University of Maryland
ashays06@umd.edu

Abstract

This paper presents our system description and error analysis of our entry for NLLP 2024 shared task on Legal Natural Language Inference (L-NLI) (Hagag et al., 2024). The task required classifying these relationships as entailed, contradicted, or neutral, indicating any association between the review and the complaint. Our system emerged as the winning submission, significantly outperforming other entries with a substantial margin and demonstrating the effectiveness of our approach in legal text analysis. We provide a detailed analysis of the strengths and limitations of each model and approach tested, along with a thorough error analysis and suggestions for future improvements. This paper aims to contribute to the growing field of legal NLP by offering insights into advanced techniques for natural language inference in legal contexts, making it accessible to both experts and newcomers in the field.

1 Introduction

In today's digital age, vast amounts of information circulate online, creating an overwhelming stream of text that spans news articles, social media, and user-generated content. Within this unstructured data, legal violations often remain hidden, blending into the surrounding noise. Legal violations frequently leave behind data traces. To identify these traces and detect violations, prior research in Legal NLI (Koreeda and Manning, 2021) has typically utilized specialized models designed for particular domain applications (Silva et al., 2020) (Yu et al., 2020). Uncovering these violations is not only crucial for upholding individual rights and ethical standards, but also for maintaining societal justice in an increasingly digital world. Addressing this challenge requires more than traditional methods. While existing models have proven effective within their specialized domains, they lack the flexibility needed to tackle the complex and varied nature of legal violations found in diverse online contexts. Our work seeks to bridge this gap by leveraging advanced language models for the nuanced task of Legal Natural Language Inference (L-NLI), as part of the NLLP 2024 shared task. The aim was to classify relationships between legal complaints and reviews as either entailed, contradicted, or neutral. In this study, we implemented a range of techniques, including multi-layered fine-tuning and alignment strategies, to enhance text classification. We experimented with several LLMs, such as Gemma (Team, 2024), Phi3 (Abdin, 2024), Zephyr (Tunstall et al., 2023), LLaMA (Dubey et al., 2024), Mistral (Jiang et al., 2023), OpenHermes (Teknium, 2023) and Qwen (Yang et al., 2024) refining each model for optimal performance. These approaches proved highly effective, with our system outperforming other entries by a large margin. Beyond technical achievements, we present a thorough error analysis, highlighting where the models excelled / struggled. Through our findings, we aim to advance the field of legal NLP, making complex legal analysis accessible to a wider audience, while pushing the boundaries of NLI in legal domain. The code and models used in the official submission and the later found best model can be found here. [1] [2]

* equal contribution

[1] https://github.com/1-800-SHARED-TASKS/EMNLP-2024-NLLP

[2] https://huggingface.co/collections/1-800-SHARED-TASKS/

Proceedings of the Natural Legal Language Processing Workshop 2024, pages 318–325
November 16, 2024 ©2024 Association for Computational Linguistics

2 Dataset

The dataset for the NLI task consists of a legal premise (a summary of resolved class-action cases) and a corresponding hypothesis (an online media text). The training and test splits of the dataset consist of 312 and 84 samples. For the initial fine-tuning, the test and validation subsets of the SNLI dataset (Bowman et al., 2015) were used consisting of 20000 samples. The distributions of each of the training sets and the test set can be seen in Table 1. The original dataset (Bernsohn et al., 2024) used had just 312 rows, the aggregation of datasets is explained in detail in Appendix. The length of the texts are both mostly 4-7 sentences long in both the premise and hypothesis.

	Train-1	Train-2	Test
Entailed	34.0%	32.7%	47.6%
Neutral	33.1%	33.9%	34.5%
Contradict	32.9%	33.3%	17.9%

Table 1: Distributions of each class in each data split

* Train-1 is a subset of SNLI dataset , Train-2 is the NLLP dataset

3 System Description

Various LLMs were tested with and without additional training data or additional training stages. They were also tested with various alignment approaches in various configurations. The metrics obtained on the test set with each of these approaches/models can be seen in Table 2. The official metric used was Macro F1 score [F1]. Additionally accuracy [A], precision [P] and recall [R] were also reported.

3.1 Multi-stage Learning

Given the small size of the existing training dataset (312 samples), we have additionally tested multi-stage learning by first fine-tuning over a subset of 20000 rows from the SNLI dataset to first let the models adapt to generic NLI tasks with a lower learning rate and then further fine-tuned the resultant models on the NLLP training samples with a higher learning rate. Additionally we have tested using additional training data from previous works (more in Appendix). Both of these approaches did result in better performance. An overview of the process can be seen in Figure 1.

emnlp-2024-nllp-66e7af534b7e708a36db02df

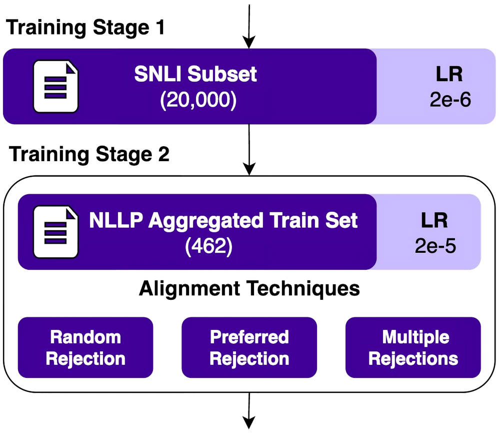

Figure 1: Multi-stage Training Overview

3.2 Alignment approaches used

We have tested using ORPO (Hong et al., 2024) during fine-tuning using various LLMs in 3 different configurations i.e the rejected sample being a) random, b) preferred and c) multiple rejected samples. The usage of ORPO did improve the performance over all of the domains in any of the configurations.

3.2.1 Random Rejection

In this approach, the actual label being the accepted response would lead to the rejected response being a random class form the remaining two. The results did improve compared to not using ORPO but by a very slight margin.

3.2.2 Preferred Rejection

In cases where the actual label is Neutral, a random label is chosen as the rejected sample among the other two. We chose 'Neutral' as the rejected response when the actual label is either Entailed or Contradict. The reason being all of the errors being one of the other two classes being labelled as 'Neutral or vice versa. This did improve the performance significantly by reducing the mis-classified samples between Neutral and the other classes.

3.2.3 Multiple Rejections

In this approach, while the label class would be the accepted class, both the other two classes were added as the rejected samples. Although this was computationally expensive, the results were close to those from preferred rejection approach.

4 Error Analysis

We were able to completely avoid Type-1 errors i.e classification of 'Entailed' as 'Contradict' and

LLM Used	Trained on	Alignment approach	A	P	R	F1
GEMMA-2-27B	NLLP*	None	0.857	0.871	0.894	0.871
GEMMA-2-27B	NLLP	None	0.857	0.859	0.891	0.865
Mistral-8x7B	NLLP*	None	**0.869**	0.877	**0.902**	0.881
QWEN-2-7B	NLLP*	None	0.833	0.828	0.868	0.839
QWEN-2-7B	NLLP	None	0.821	0.852	0.869	0.842
Phi-3-Medium	NLLP*	None	0.821	0.853	0.813	0.820
OpenHermes-13B	NLLP*	None	0.774	0.820	0.832	0.803
GEMMA-2-27B	SNLI, NLLP*	None	**0.869**	0.866	0.899	0.874
GEMMA-2-27B	SNLI, NLLP	None	0.821	0.828	0.862	0.831
GEMMA-2-27B	SNLI, NLLP*	ORPO Random	0.845	0.852	0.882	0.855
GEMMA-2-27B	NLLP*	ORPO Multiple	0.833	0.842	0.860	0.840
GEMMA-2-27B	SNLI, NLLP*	ORPO Preferred	**0.869**	**0.885**	**0.902**	**0.887**
Mistral-NEMO	NLLP*	ORPO Multiple	**0.869**	0.867	0.890	0.877
Phi-3-Medium	NLLP*	ORPO Multiple	0.845	0.872	0.833	0.838
Zephyr-7B	NLLP*	ORPO Multiple	0.810	0.838	0.858	0.832
Phi-3-Medium‘	NLLP‘	ORPO Multiple‘	*0.845*‘	*0.884*‘	*0.844*‘	*0.853*‘
baseline	-	-	-	-	-	0.807

Table 2: Metrics on the test set with some of the approaches/models tested
* Indicated aggregated train set of NLLP (more in appendix)
‘ indicates official submission

vice versa, limiting the error cases to Type-2 errors i.e classification of ’Neutral’ as another and vice versa. Confusion matrix of our models’ predictions on the test set can be seen in Figure 2 and Figure 3.

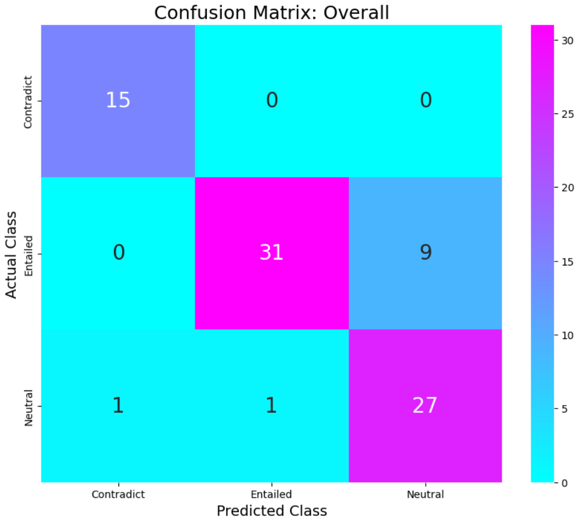

Figure 2: Confusion Matrix : Our system’s (best) predictions over the test set

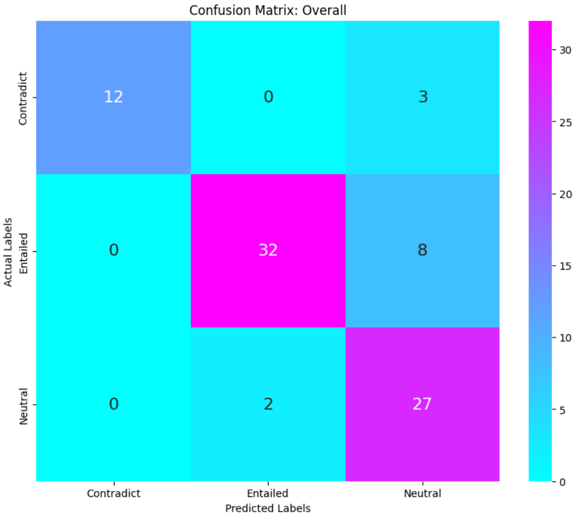

Figure 3: Confusion Matrix : Our system’s (submission) predictions over the test set

It can be observed from both Figure 2 and Figure 3 that most common case of errors was those being mis-classified among Neutral and Entailed. We found these to be cases where the hypothesis consisted of multiple sentences which entail the premise followed by a vague / unrelated statement, while some are to be labelled as ’Entailed’ and rest as ’Neutral’ based on the perceived tone/feeling of the user, it would be likely that there might not be consensus among human annotators as well in many such cases. It is worth looking into the performance of models trained on not just the labels, but also the reasoning of the annotators on why a certain label was chosen, as it might help the model learn better.

Legal act	in Train set	Domain	in Test set	A	P	R	F1
Privacy	229	BIPA	22	0.73	0.80	0.86	0.77
		Data-Breach	20	0.95	0.96	0.95	0.95
		VPPA	6	1.00	1.00	1.00	1.00
TCPA	111	TCPA	9	0.89	0.89	0.93	0.90
Consumer	102	Consumer	8	0.88	0.92	0.92	0.90
WAGE	20	WAGE	19	0.89	0.80	0.92	0.83
Overall(best)	-	-	-	**0.87**	**0.89**	**0.90**	**0.89**
Overall(submission)	-	-	-	**0.85**	**0.89**	**0.84**	**0.85**

Table 3: Performance of our models on the test set : Domain wise

4.1 Performance on each Domain

The performance of our system on each domain in the test set can be seen in Table 3. The metrics obtained on most of the domains were significantly higher than that of the baseline. The system worked well on all domains, however comparatively weaker on BIPA which was imbalanced in the training set.

5 Scope For Improvement

As seen in Table 3 the performance across each domain varied by a significant margin. However, the domains over which some models underperformed, some other performed well. It is likely that using ensembles can improve the performance by a considerable margin.

5.1 Low training data

Some cases did get misclassified too often especially those whose domain data was less represented in the training dataset. From what was observed from comparison of performance over original and aggregated datasets and the models with and without SNLI fine-tuning step involved, It can be determined that more training data would improve the performance considerably especially the domains with less data.

5.2 Individual Annotations availability

In models built using Preferred Rejection, cases with Neutral as the label had used a random label from the other two as the rejected sample. However availability of individual annotations might provide more info on what choice of rejected label might lead to better results compared to choosing a rejected label at random.

6 Conclusion

Compared to the well known SNLI dataset which consist of premise and hypothesis pair which are usually one or two sentences long, the current dataset has texts (both premise and hypothesis) which are roughly four times longer leading to more complexity. Since, the SNLI dataset has a 98% consensus and 58% unanimous annotation among 5 annotators, it can be expected that a human annotation on the current dataset can lead to even less proportion of texts where a consensus or unanimous vote can be reached. Yet, our models were able to provide a reliable performance completely avoiding Type-1 errors, performing better than human annotations expected from those with domain knowledge, hinting at a potential of practical applicability.

Limitations

Due to computational resource limitations, the base models of LLMs were initially loaded in 4-bit precision, It is likely that a larger model used in full-precision might perform better. Since the test dataset used in the task is relatively small, the LLMs/approaches that might perform better in practical scenarios may vary from those found to be better on the current dataset.

Ethics Statement

Automating the identification of legal violations may inadvertently generate false positives or negatives, potentially impacting individual rights and the integrity of the legal system. Therefore, we emphasize that our models are intended to complement, not replace, legal professionals. It is critical that any use of our models is approached with caution, recognizing the inherent limitations and biases that automated systems may present.

References

et al. Abdin, Marah. 2024. Phi-3 technical report: A highly capable language model locally on your phone.

Dor Bernsohn, Gil Semo, Yaron Vazana, Gila Hayat, Ben Hagag, Joel Niklaus, Rohit Saha, and Kyryl Truskovskyi. 2024. Legallens: Leveraging llms for legal violation identification in unstructured text.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference.

William Bruno and Dan Roth. 2022. Lawngnli: A long-premise benchmark for in-domain generalization from short to long contexts and for implication-based retrieval.

Alexis Conneau, Douwe Kiela, Holger Schwenk, Loic Barrault, and Antoine Bordes. 2018. Supervised learning of universal sentence representations from natural language inference data.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Neel Guha, Julian Nyarko, Daniel Ho, Christopher Ré, Adam Chilton, Aditya K, Alex Chohlas-Wood, Austin Peters, Brandon Waldon, Daniel Rockmore, Diego Zambrano, Dmitry Talisman, Enam Hoque, Faiz Surani, Frank Fagan, Galit Sarfaty, Gregory Dickinson, Haggai Porat, Jason Hegland, Jessica Wu, Joe Nudell, Joel Niklaus, John Nay, Jonathan Choi, Kevin Tobia, Margaret Hagan, Megan Ma, Michael Livermore, Nikon Rasumov-Rahe, Nils Holzenberger, Noam Kolt, Peter Henderson, Sean Rehaag, Sharad Goel, Shang Gao, Spencer Williams, Sunny Gandhi, Tom Zur, Varun Iyer, and Zehua Li. 2023. Legalbench: A collaboratively built benchmark for measuring legal reasoning in large language models. In *Advances in Neural Information Processing Systems*, volume 36, pages 44123–44279. Curran Associates, Inc.

Ben Hagag, Liav Harpaz, Gil Semo, Dor Bernsohn, Rohit Saha, Pashootan Vaezipoor, Kyryl Truskovskyi, and Gerasimos Spanakis. 2024. Legallens shared task 2024: Legal violation identification in unstructured text.

Jiwoo Hong, Noah Lee, and James Thorne. 2024. Orpo: Monolithic preference optimization without reference model.

John Hudzina, Kanika Madan, Dhivya Chinnappa, Jinane Harmouche, Hiroko Bretz, Andrew Vold, and Frank Schilder. 2020. Information extraction/entailment of common law and civil code. In *JSAI International Symposium on Artificial Intelligence*, pages 254–268. Springer.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Daniel Martin Katz, Dirk Hartung, Lauritz Gerlach, Abhik Jana, and Michael J. Bommarito II au2. 2023. Natural language processing in the legal domain.

Yuta Koreeda and Christopher D. Manning. 2021. Contractnli: A dataset for document-level natural language inference for contracts.

Alice Kwak, Gaetano Forte, Derek E Bambauer, and Mihai Surdeanu. 2023. Transferring legal natural language inference model from a us state to another: What makes it so hard? In *Proceedings of the Natural Legal Language Processing Workshop*.

Alice Saebom Kwak, Jacob O. Israelsen, Clayton T. Morrison, Derek E. Bambauer, and Mihai Surdeanu. 2022. Validity assessment of legal will statements as natural language inference.

Paulo Silva, Carolina Gonçalves, Carolina Godinho, Nuno Antunes, and Marilia Curado. 2020. Using nlp and machine learning to detect data privacy violations. In *IEEE INFOCOM 2020 - IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*, pages 972–977.

Joe Stacey, Pasquale Minervini, Haim Dubossarsky, and Marek Rei. 2022. Logical reasoning with span-level predictions for interpretable and robust nli models.

et al. Team, Gemma. 2024. Gemma 2: Improving open language models at a practical size.

Teknium. 2023. Openhermes-13b.

Lewis Tunstall, Edward Beeching, Nathan Lambert, Nazneen Rajani, Kashif Rasul, Younes Belkada, Shengyi Huang, Leandro von Werra, Clémentine Fourrier, Nathan Habib, Nathan Sarrazin, Omar Sanseviero, Alexander M. Rush, and Thomas Wolf. 2023. Zephyr: Direct distillation of lm alignment.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, et al. 2024. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.

Zhanye Yang. 2022. Legalnli: natural language inference for legal compliance inspection. In *International Conference on Advanced Algorithms and Neural Networks (AANN 2022)*, volume 12285, pages 144–150. SPIE.

Yaoquan Yu, Yuefeng Guo, Zhiyuan Zhang, Mengshi Li, Tianyao Ji, Wenhu Tang, and Qinghua Wu. 2020. Intelligent classification and automatic annotation of violations based on neural network language model. In *2020 International Joint Conference on Neural Networks (IJCNN)*, pages 1–7.

A Training Data Aggregation

Due to training dataset provided being not large enough, we have used additional training data which include the dataset from the LegalLens paper. The aggregated training dataset used is what was obtained by merging both the datasets, upon removal of duplicates.

- Current Dataset huggingface.co/datasets/darrow-ai/LegalLensNLI-SharedTask : 312 training samples
- Additional Dataset huggingface.co/datasets/darrow-ai/LegalLensNLI : 312 training samples
- Aggregated Dataset huggingface.co/datasets/1-800-SHARED-TASKS/EMNLP-2024-NLLP : 462 training samples

B System Replication

We have used each of the LLMs tested by loading them in 4bit precision before fine-tuning on each dataset in both the training stages using LoRA. The hyper parameters used in each of the training stages can be seen in Table 4. The hyper parameters not specified below were used with their default values in both stages. The code used can be found here : github.com/1-800-SHARED-TASKS/EMNLP-2024-NLLP.

parameter	Stage-1 (SNLI)	Stage-2 (NLLP)
Learning Rate	2e-6	2e-5
Max Length (tokens)	1024	2048
LoRA alpha	32	16
LoRA dropout	0	0
beta	0.1	0.1
random state	1024	1024
number of epochs	1	3
loaded prev. model as	fp4	fp32

Table 4: Hyperparameters used in each training stage

C Models used / SNLI version of LLMs

The models used in the paper including the best performing model and the one used in the official submission can be found here :

- Best performing model : huggingface.co/1-800-SHARED-TASKS/EMNLP-NLLP-NLI-GEMMA2-27B-withSNLI-withORPO
- Model used for submission : huggingface.co/1-800-SHARED-TASKS/EMNLP-NLLP-NLI-PHI3-medium-withoutSNLI-withORPO

Additionally the models obtained after fine-tuning LLMs used on the SNLI dataset can be found here :

- GEMMA NLI : huggingface.co/1-800-SHARED-TASKS/GEMMA2-27B-NLI-16bit
- PHI3 NLI : huggingface.co/1-800-SHARED-TASKS/PHI3-Medium-NLI-16bit

D Performance of both models : domain wise

The performance of our best performing model (GEMMA-2-27B-SNLI) can be seen below followed by those from our submission model (PHI-3-SNLI).

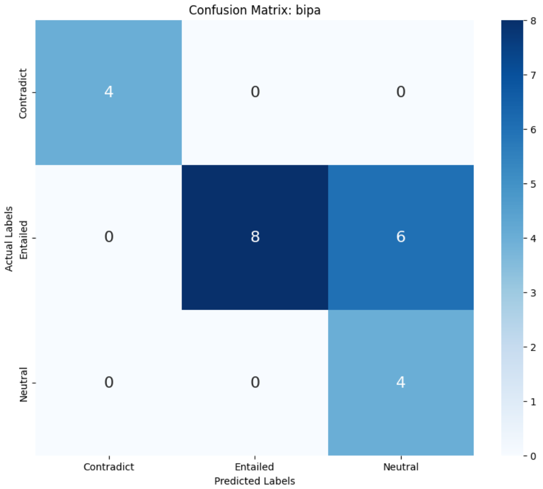

Figure 4: performance on test set : GEMMA2-SNLI : BIPA

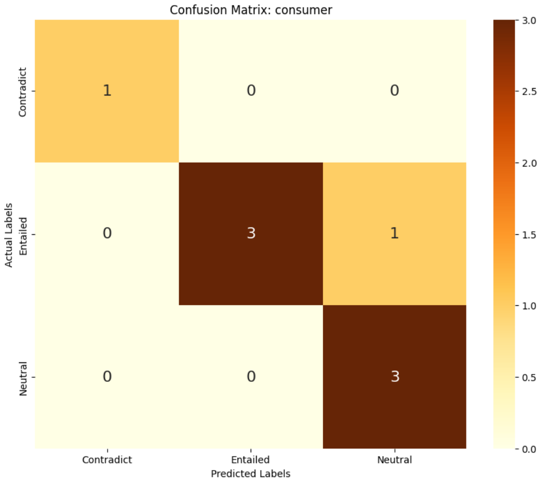

Figure 5: performance on test set : GEMMA2-SNLI : Consumer

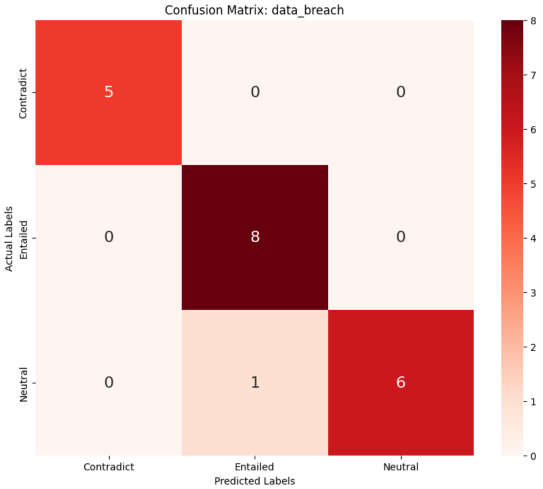

Figure 6: performance on test set : GEMMA2-SNLI : Data Breach

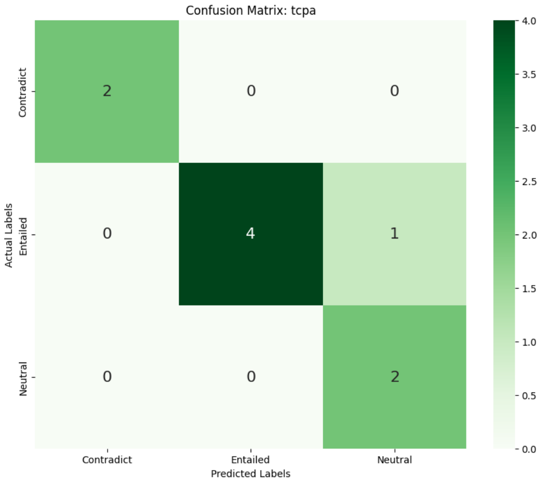

Figure 7: performance on test set : GEMMA2-SNLI : TCPA

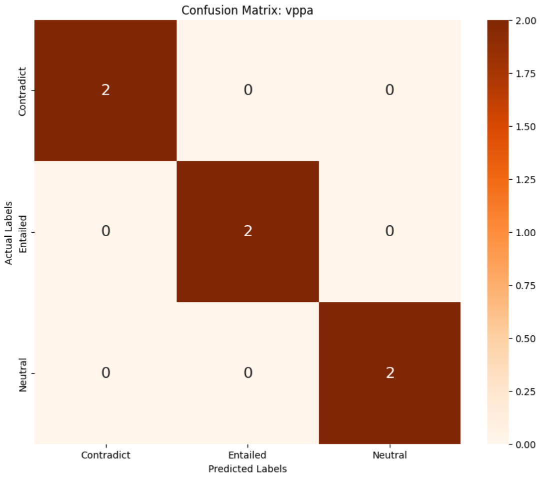

Figure 8: performance on test set : GEMMA2-SNLI : VPPA

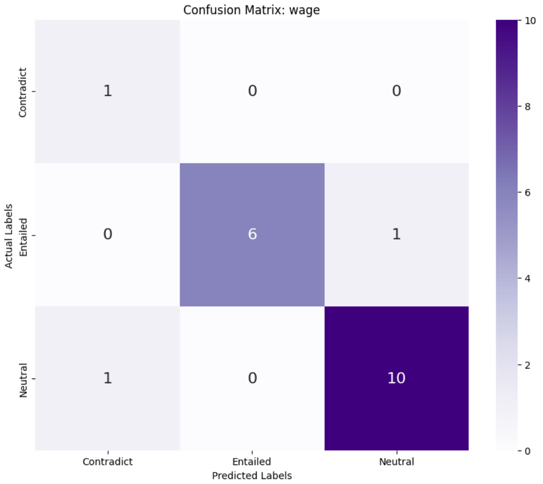

Figure 9: performance on test set : GEMMA2-SNLI : WAGE

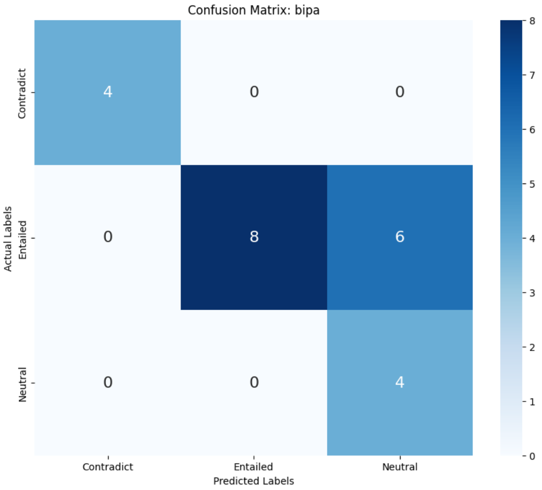

Figure 10: performance on test set : PHI3-SNLI : BIPA

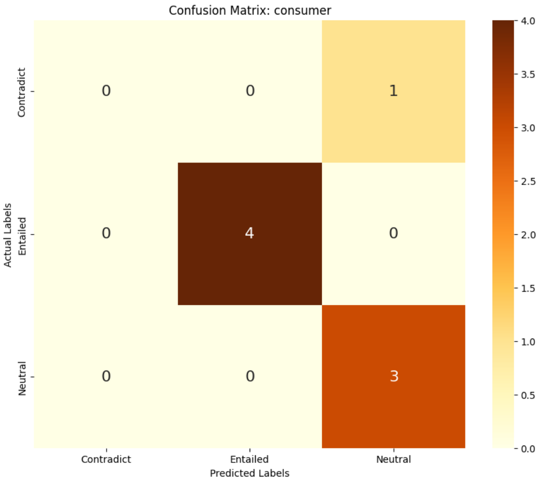

Figure 11: performance on test set : PHI3-SNLI : Consumer

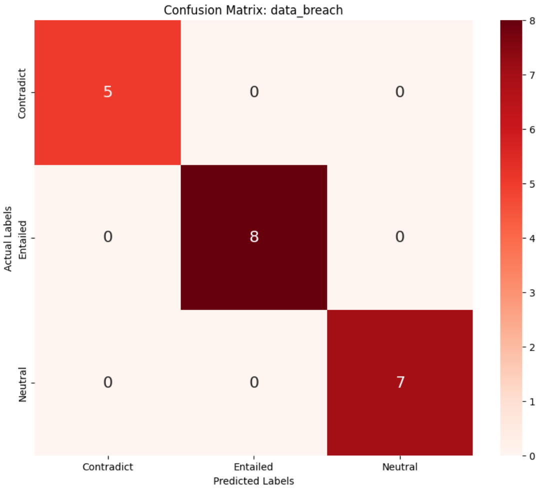

Figure 12: performance on test set : PHI3-SNLI : Data Breach

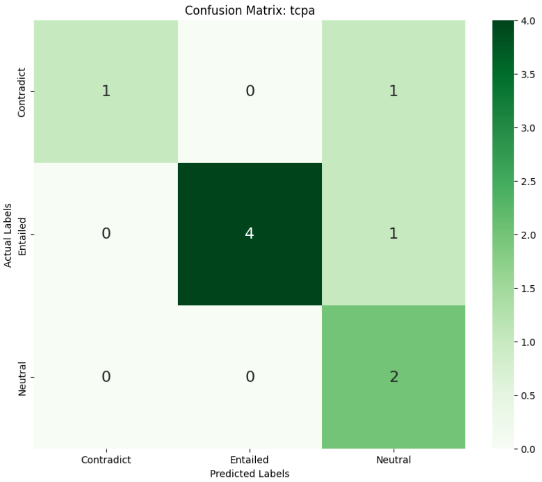

Figure 13: performance on test set : PHI3-SNLI : TCPA

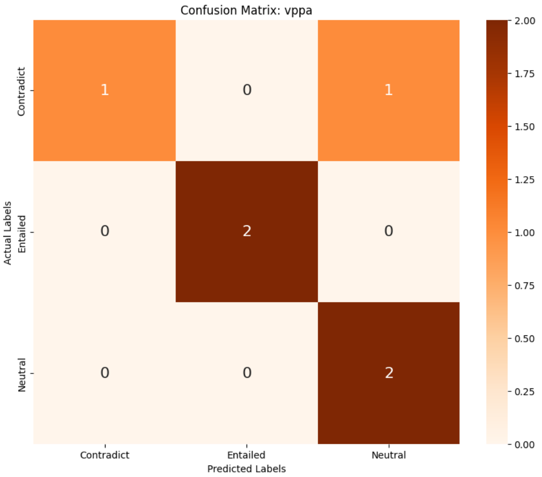

Figure 14: performance on test set : PHI3-SNLI : VPPA

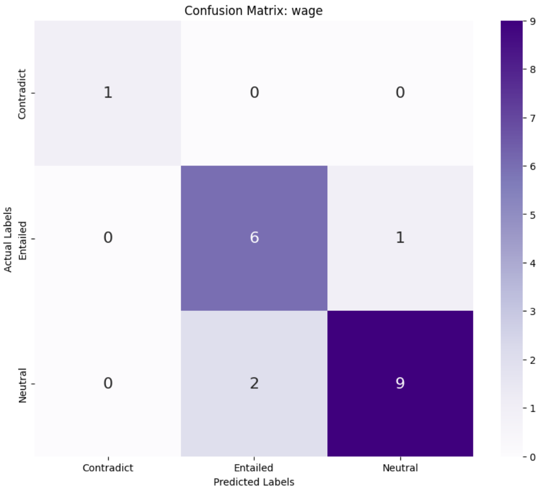

Figure 15: performance on test set : PHI3-SNLI : WAGE