UC-85

HELPR: Helping Extrapolate Labels for Police Reports using Large Language Models

Abstract

We Introduce HELPR: a LLM powered Human-AI Teaming framework for classifying Behavioral Health Reports from Police Narratives. Police officers spend many hours a week documenting their findings when reporting to a 911 call. There is so much detail in these reports that they remain an untapped resource for future data analytics by the police department. To assist the experts and reduce the time spent reading and analyzing, we propose the use of large language models (LLMs) to tag police reports based on their content. We introduce two fine-tuned models Mistral-7B and TinyLlama fine-tuned to provide both a Behavioral Health classification and rationale behind the classification of the report. We find that through our framework we can train models that not only outperform the base models, but yield a high level of agreement with human annotators.

Introduction

Kennesaw State University has been given four years of police reports and the corresponding data to use for research. There are approximately 196,000 reports. 297 reports have been annotated by social workers to identify tags such as "Mental Health", "Substance Abuse", "Domestic Social", "NonDomestic Social", and "Other". We then trained the models using the annotated data and performed Human-AI teaming.

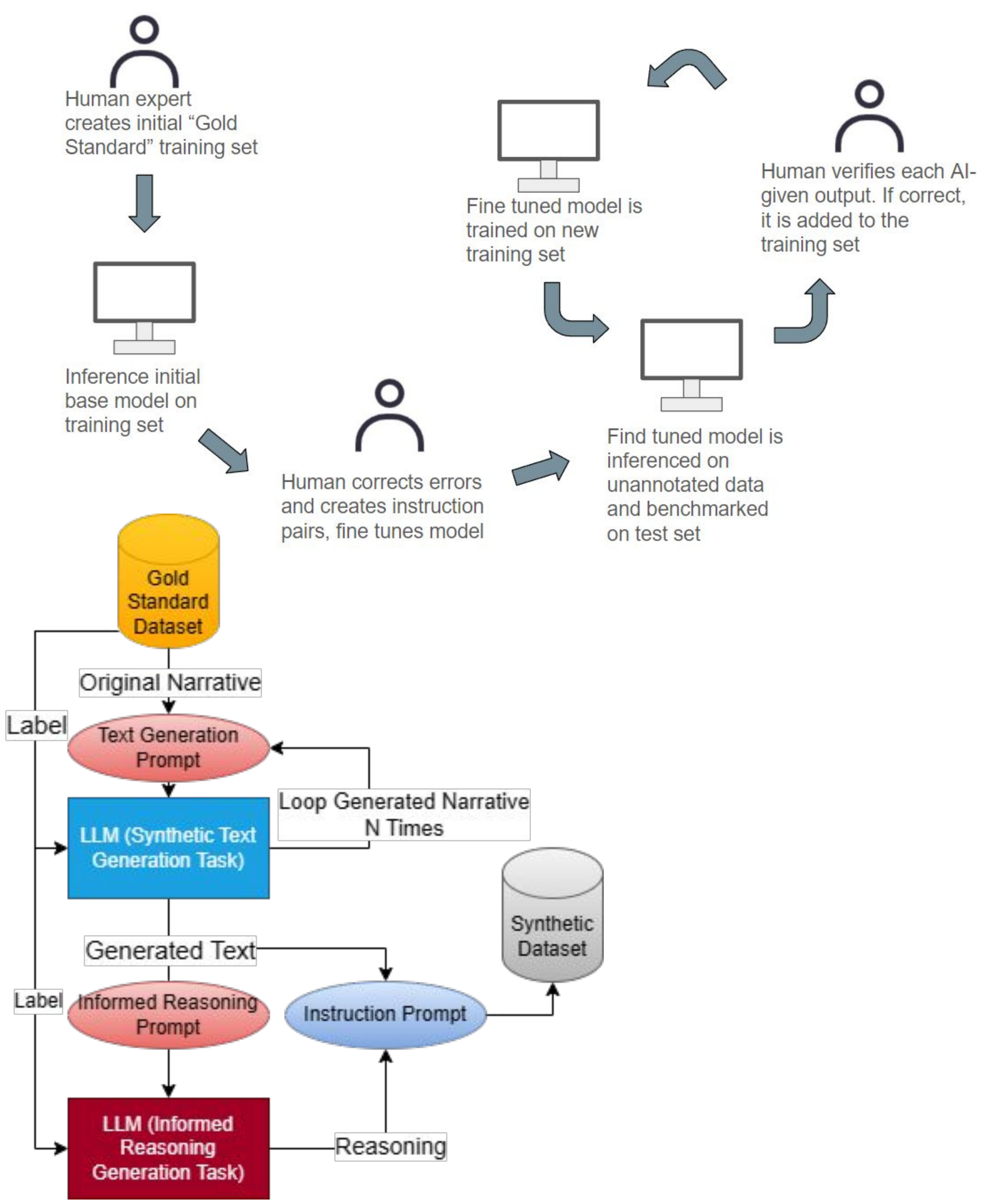
Because of the sensitive nature of the data, we use a VPN and a secure server. The weights for all the models trained are open-source and can be hosted and stored locally, maintaining confidentiality of public information. We have both completed the CITI training for Human Experimentation to ensure that we know the laws and potential issues that can occur when using personal data, such as police reports.

Research Question(s)

Using Local LLMs, can we accurately predict the Behavioral Health classification as well as give coherent reasoning for the model output? Can we develop a usable framework for iterative development through Human-AI teaming?

Methodology

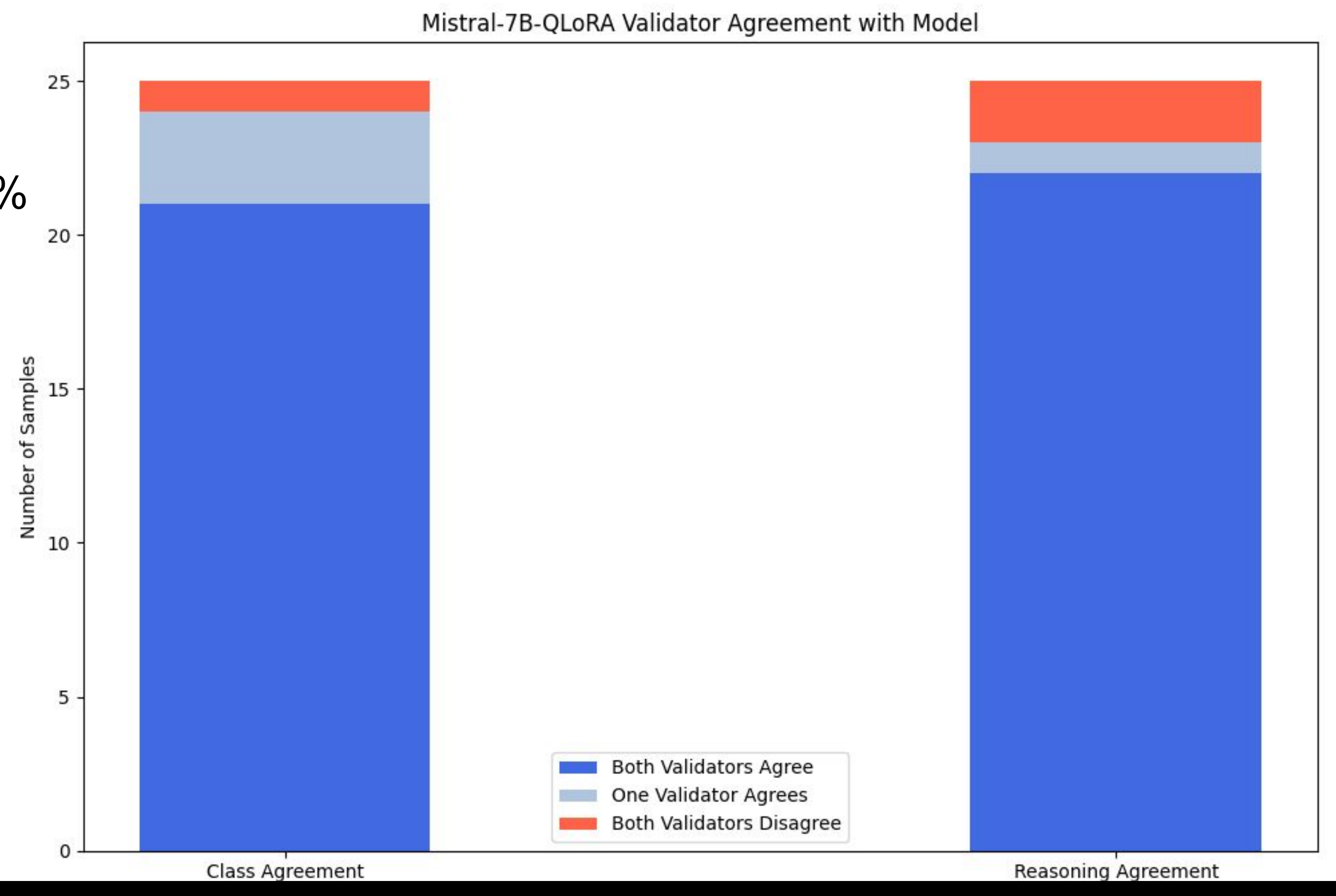
- We utilized a Gold Standard Dataset of 297 samples for Benchmarking and for training the Initial model for Human-AI Teaming.
- First, you have to inference the base model on your test set and create instruction pairs that contain a prompt, an input, and an output.
- We then train the model on this test set of instruction pairs.
- After that, we fine-tune the model and inference it on the unannotated data (no instruction pairs) to create a new training set.
- We go through each AI-created answer and determine if the output is correct or not.
- We can then train the new model on the human plus AI made training set and benchmark the test set.
- To meet the data requirements of training LLMs we implemented a Synthetic Data Generation Strategy using Mixtral-8x7B. We inference the model on the Gold Standard dataset twice on two separate tasks by pre-informing the LLM of the label in our prompt. To guarantee independent samples, we recursively pass the previous generation's output as the input narrative for the next generation task.
- Through HELPR we can create not only Instruction Pairs for QLoRA training but also we can create chosen:rejected pairs for DPO Training, which is a Reinforcement Learning From Human Feedback (RLHF) Technique, through this we can better correct mistakes and guide model behavior.
- NEFTune adds noise to the embeddings during the forward pass of fine-tuning NEFTune has been proven through research to improve instruction accuracy and reduce overfitting, in this work we use a alpha of 10.
- We perform additional testing on the models on unannotated data with human annotators to measure the level of agreement on unseen data.


Human expert creates initial "Gold Standard" training set
Inference initial base model on training set
Human corrects errors and creates instruction pairs, fine tunes model
Find tuned model is inferenced on unannotated data and benchmarked on test set
Fine tuned model is trained on new training set
Human verifies each AI-given output. If correct, it is added to the training set
Gold Standard Dataset
Original Narrative
Label
Text Generation Prompt
Loop Generated Narrative N Times
LLM (Synthetic Text Generation Task)
Generated Text
Synthetic Dataset
Label
Informed Reasoning Prompt
Instruction Prompt
LLM (Informed Reasoning Generation Task)
Reasoning

Results

Fine Tuning Results:
Mistral-7B categorical Accuracy: 79%
Tiny Llama categorical accuracy: 95.6%

Mistral-7B binary accuracy: 83%
Tiny Llama binary accuracy: 98%



Conclusions

During this project, we have found that TinyLlama models are more accurate for categorization, but they offer very poor text generation capabilities. Mistral models have a slightly lower accuracy for categorization, it has much reasoning capabilities, evidenced by 92% dual annotator agreement and 96% single annotator agreement on the subjectivity test. Having quality text generation for the reasonings allows for text generation that can expand the training set and increase accuracy over time.

The choice of model depends on future needs. If the police only need the category, we should further develop the TinyLlama model. However, if the police were to use our models as a training program for future administrators, we should further develop the Mistral model as its text generation will help people learn more easily.

Through our process we have created 6,500 more samples for training.

Acknowledgments

Professor Sharon Perry, Department of Computer Science, KSU
Drs. Md Abdullah Al Hafiz Khan; Monica Nandan; Dominic Thomas and Yong Pei, KSU
Joshua Sprung, Expert Annotator
Cobb County Police Department

Contact Information

Project Website: https://sites.google.com/view/hkllm

Hailey Walker's LinkedIn: www.linkedin.com/in/hailey-walker-0103a1244
William Stigall's LinkedIn: www.linkedin.com/in/william-stigall-94055823a

References

[1] M. Brown, M. A. A. H. Khan, D. Thomas, Y. Pei and M. Nandan, "Detection of Behavioral Health Cases from Sensitive Police Officer Narratives," 2023 IEEE 47th Annual Computers, Software, and Applications Conference (COMPSAC), Torino, Italy, 2023, pp. 1398-1403, doi: 10.1109/COMPSAC57700.2023.00213. keywords: {Training;Law enforcement;Annotations;Computational modeling;Manuals;Mental health;Predictive models;Mental Health;Deep Learning;CNN-LSTM;Active Learning;Prodigy;Machine Learning;NLP;Behavioral Health},
[2] Neel Jain, Ping yeh Chiang, Yuxin Wen, John Kirchenbauer, Hong-Min Chu, Gowthami Somepalli, Brian R. Bartoldson, Bhavya Kailkhura, Avi Schwarzschild, Aniruddha Saha, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Neftune: Noisy embeddings improve instruction finetuning, 2023.
[3] Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, L´elio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timoth´ee Lacroix, and William El Sayed. Mistral 7b, 2023.
[4] Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, L´elio Renard Lavaud, Lucile Saulnier, Marie- Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, Szy- mon Antoniak, Teven Le Scao, Th´eophile Gervet, Thibaut Lavril, Thomas Wang, Timoth´ee Lacroix, and William El Sayed. Mixtral of experts, 2024.
[5] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2023.
[6] Peiyuan Zhang, Guangtao Zeng, Tianduo Wang, and Wei Lu. Tinyllama: An open-source small language model, 2024.
[7] Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. Qlora: Efficient finetuning of quantized llms, 2023.
[8] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models, 2021.
[9] Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model, 2023

KENNESAW STATE UNIVERSITY
COLLEGE OF COMPUTING AND SOFTWARE ENGINEERING

Authors: Hailey Walker, William Stigall
Advisor: Dr. Md Abdullah Al Hafiz Khan