Fine-tuning Large Language Models for Argument Stance Detection in Unseen Domains

Antonio Mari[1], Matteo Santelmo[1], and Jakhongir Saydaliev[1]

[1]Department of Computer Science, EPFL, Switzerland

***Abstract*—Stance detection holds a substantial significance in applications involved across various domains, ranging from social media analysis to political discourse and beyond. In our work, we fine-tune various Large Language Models (LLMs) through Low-Rank Adaptation (LoRA) on SEMEVAL2016 and IBM-DEBATER, testing how well they generalize on unseen datasets for argument stance detection. We find that MISTRAL-7B outperforms other LLMs, as well as the reported baselines, and that fine-tuning on SEMEVAL2016 and extrapolating by predicting on IBM-DEBATER might lead to better results than directly fine-tuning on the latter.**

I. INTRODUCTION

Stance detection is a classification task, aimed to discern whether an argument is in favor, against, or neutral towards a given topic. Such task may empower diverse applications, such as misinformation recognition and polarization assessment. Thus, the implications of a proficient stance detection model extend far and wide, fostering a deeper understanding of societal viewpoints and aiding in the extraction of valuable insights from vast pools of textual information.

As the described task requires a deep understanding of arguments within textual data. For this reason, leveraging Large Language Models is a compelling approach. LLMs possess an extensive understanding of language nuances and contextual relationships. Moreover, they are suitable for fine-tuning on specific downstream tasks and employing parameter-efficient strategies (section II-A) offers a promising approach for stance detection.

Our work is part of a bigger project, CommPass, which aim is to create awareness by media producers and aggregators for the polarity of media contents. To this end, CommPass provides media readers with visualizations showing where the content is situated in the "space" of a given media event (ex. Russia-Ukraine war, Covid-19, etc.).

II. METHODS

A. Models & Efficient Fine-Tuning

We explored the performances of three LLMs on our task, MISTRAL-7B [1], LLAMA-2-7B [2] and PHI-1.5 [3]. We experiment with such models to test their peculiarities. PHI-1.5, for example, is a "small" model (counting 1.3 billion parameters) trained mainly on textbooks data; on the other hand, LLAMA-2-7B and MISTRAL-7B count around

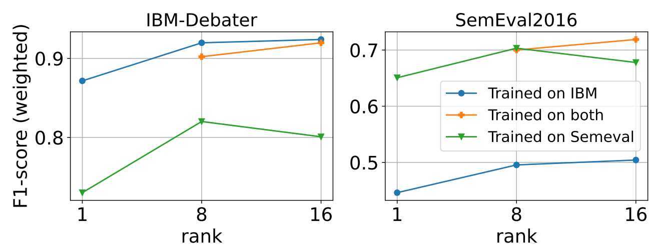

Figure 1: **F1-score of MISTRAL-7B** trained on IBM-DEBATER, SEMEVAL2016 or on both datasets. Models with LoRA rank=1 are trained on 10% of the training set(s), whereas those with rank 8 and 16 are trained on 70% of the training set(s). The values are computed on the test sets, and are averaged over three different random seeds.

7 billion parameters and their architectures present more complex features, like Grouped Query Attention (GQA) and Sliding Window Attention (SWA). As described in section III, we perform various sessions of experiments in order to select the best performing model for our task.

Parameter-Efficient Transfer Learning [4] for natural language processing (NLP) presents an approach geared towards achieving robust performance across multiple downstream tasks while introducing minimal additional parameters. It involves employing adapter modules inserted between pre-trained network layers to ease fine-tuning of large pre-trained models. In particular, PEFT achieves computational resources and time efficiency by freezing some of the layers of the pre-trained model and only fine-tuning the last few layers that are specific to the downstream task.

Low Rank Adaptation [5] stands out as an efficient adaptation technique. By incorporating trainable rank decomposition matrices into selected layers of the Transformer architecture and freezing pre-trained model weights, LoRA significantly diminishes the number of trainable parameters, rendering it more feasible to customize large pre-trained models. For this reason, we chose LoRA for performing the LLMs fine-tuning.

B. Data Overview

We used two datasets to train and evaluate our models. Our aim is achieving a model with a good generalizability over different kinds of data.

The first dataset originates from the SemEval2016's [6] dataset for stance detection task. Collecting data extracted from Twitter, it focuses on six different targets ranging from politics to religion. Each tweet is labeled as either in "Favor", "Neutral" or "Against" the target.

The second dataset employed comes from IBM-debater [7] project and contains claims and evidence manually collected from hundreds of Wikipedia articles targeting 33 controversial topics. This dataset only allows two labels (PRO and CON, mapped respectively to 1 and -1).

An important difference between the two datasets stands in the target formulation. while in SemEval targets are made of one or a few words, IBM-debater uses small sentences. Moreover, IBM-debater collects more information about the sentiment of claims and the relationship between each claim and its target. The stance of a claim x_c towards a target x_t is given as:

$$\text{Stance}(x_c, x_t) = s_c \cdot R(x_t, x_c) \cdot s_t \tag{1}$$

where $s_c \in \{-1, 1\}$ is the claim's sentiment, $s_t \in \{-1, 1\}$ the target's sentiment and $R(x_t, x_c) \in \{-1, 1\}$ the contrast relation between the claim and the target.

By using two datasets, we could assess the generalization of models on unseen domain (section III-C).

Finally, to qualitatively evaluate our best model, we used a dataset consisting of articles mentioning the murder of Samuel Paty, a French school teacher that was killed and beheaded in a suburb of Paris by an Islamist terrorist. This news event from French chronicle was chosen as closely related to debates concerning sensitive topics such as religious fundamentalism and the issues of laws against Islamic separatism. This unlabeled dataset served as an example of a real-world application section III-D of our best model, targeting the assessment of newspaper articles' polarization.

C. Baseline models

For SEMEVAL2016, we considered a collections of models based on BERTweet, and another one based on RoBERTa [8]. These collections are made of multiple models, each of which was trained specifically on one of the six targets of SEMEVAL2016. We analyze results only on five targets, excluding "Donald Trump".

For IBM-DEBATER, StanceBERTa was chosen as the baseline model. This model (also based on RoBERTa architecture [9]) does not focus on a specific target but predicts the stance towards some entity mentioned in the text. This peculiarity of StanceBERTa derives from the fact this model acts more as a sentiment predictor, which fits better the data available in IBM-debater. For this particular case, by keeping in mind eq. (1), we can compute the stance of a claim x_c towards a target x_t:

$$\text{Stance}(x_c, x_t) = \hat{s_c} \cdot R(x_t, x_c) \cdot s_t \tag{2}$$

where $\hat{s_c}$ is the claim's sentiment predicted by StanceBERTa, s_t is the given sentiment of the target and $R(x_t, x_c)$ is the relation between the claim and the target.

Baseline results are reported in table I

D. Zero-shot evaluation

Zero-shot evaluation is a pivotal methodology employed to assess the generalizability and adaptability of a pre-trained language model across tasks, often domain specific, that were not explicitly included in their training phase. The rationale behind conducting zero-shot evaluation lies in measuring the model's inherent understanding, pointing out its ability to extrapolate knowledge for novel tasks.

We evaluated LLAMA-2-7B-CHAT-HF and MISTRAL-7B-INSTRUCT-V0.1, which are alternative versions of LLAMA-2-7B and MISTRAL-7B, already fine-tuned with instructions so that they are better suited to follow instructions contained in the user prompt. The zero-shot evaluation was performed employing the same prompt subsequently used in the training phase, see appendix B. Results are reported in table II.

III. EXPERIMENTS

In section III-A, we assess the role of the LoRA rank hyper-parameter; in section III-B we analyze the behavior of the LLMs in different data regimes. We find that MISTRAL-7B consistently outperforms LLAMA-2-7B and PHI-1.5. Finally, in section III-C we select the best model after training and evaluating on many combinations of datasets.

A. Assessing the rank of LoRA

Firstly, we fine-tuned the models over SEMEVAL2016, using different values of rank $r \in \{1, 2, 4, 8, 16, 32, 64\}$ for LoRA, in order to assess the effect of this hyper-parameter on our task.

We fine-tuned on the full training set and compared the models on the test set. Refer to appendix A for the experimental details.

The results are summarized in fig. 2: whilst PHI-1.5 has lower F1-score, MISTRAL-7B slightly outperforms LLAMA-2-7B. We cannot easily recognize a trend in the F1-score in function of the rank. The subsequent experiments are performed by only choosing few values for the rank, e.g. $r \in \{1, 8, 16\}$.

B. Fine-tuning on different data regimes

We experiment fine-tuning of the LLMs under various data regimes, which means using different portions of the training set in the fine-tuning. We test the following volumes of data (SEMEVAL2016):

- 1% of the training set (29 samples);
- 10% of the training set (291 samples);
- 50% of the training set (1457 samples).

We train each model using rank=1, since it proved effective at least as much as bigger rank values in a full training data

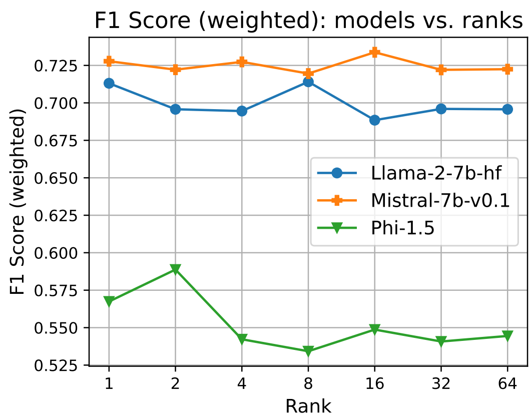

Figure 2: **Assessing the role of LoRA rank**: we cannot find significant differences between models with lower/higher ranks. Each model is trained for 10000 iterations, evaluating the checkpoint with the lowest test loss.

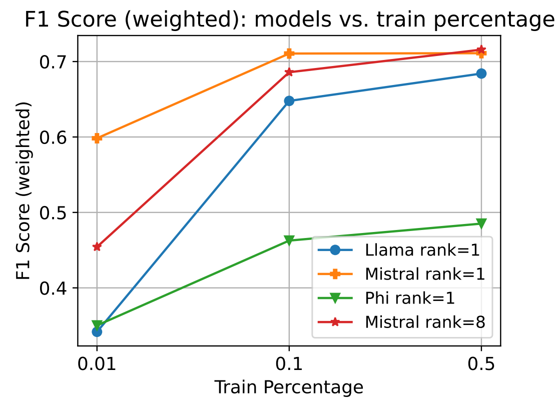

Figure 3: **Fine-tuning LLMs in different data regimes**: each model is fine-tuned for 5000 iterations. Then, the checkpoint with the lowest validation loss is selected and evaluated.

regime. As showed in fig. 3, also in this case MISTRAL-7B outperforms the other models. In particular, it is well-suited for fine-tuning when few training data is available.

Based on this result, we also included MISTRAL-7B rank=8 to make a comparison with its counterpart with rank 1. Our intake is that a higher rank easily overfits few data, so the "optimal performance" will occur using a higher volume of data than a lower rank. We can observe this on the plot: the "rank 8" model improves on 50% of data, while the performance of "rank 1" degrades between 10% and 50% of the data.

Our main conclusion from these experiments is that a good rank choice depends on how much data we fine-tune the model on.

C. Selecting the best settings for Transfer Learning

The main result we obtained so far is that MISTRAL-7B is the best performing model for our task. At this point, we still want to tune the best settings for our model.

To this end, we fine-tune models with rank 1 (10% of training set) 8 and 16 (on 70% of training set). Each model is trained either on SEMEVAL2016, IBM-DEBATER or on both datasets. For each setting, the fine-tuning is repeated three times, using different random seeds. We present the averaged results in fig. 1, in which all the models are separately evaluated on both test sets.

Firstly, we note that training only on SEMEVAL2016 leads to overfit both test sets using a higher rank (16), while this does not happens when adding (or training only on) IBM-DEBATER. We note that models fine-tuned only on SEMEVAL2016 generalize well on IBM-DEBATER, highly outperforming the baselines.

The results of rank=16 models, evaluated on SEMEVAL2016, are detailed in fig. 4. Surprisingly, we observe

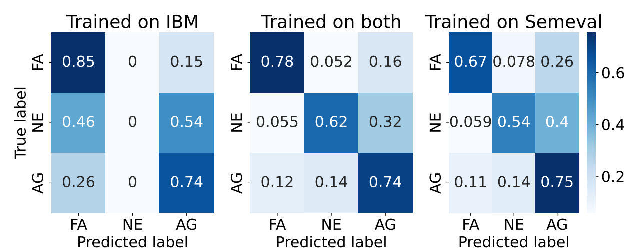

Figure 4: **Confusion matrices for MISTRAL-7B LoRA-rank=16 on SEMEVAL2016**: models are fine-tuned on IBM-DEBATER, SEMEVAL2016 or both. Values are normalized by rows.

that training on both datasets improves recall for the neutral class even if IBM-DEBATER does not provide any example labeled as neutral. Overall, the combination of the two training sets consistently leads to better results, and the extrapolation results are promising. In fact, we note a better performance on the positive class of SEMEVAL2016 with models trained only on IBM-DEBATER.

The performance of the single best model (rank=16, trained on both datasets) are reported in table I. It surpassed the baselines on four topics out of five for the SEMEVAL2016 task and critically outperformed the baseline on IBM-DEBATER. Zero-shot performances for MISTRAL-7B-INSTRUCT-V0.1 are also reported to appreciate the beneficial effect of fine-tuning. It is also noticeable in the zero-shot comparison (table II) that, despite poor performances, MISTRAL-7B-INSTRUCT-V0.1 still outperformed LLAMA-2-7B-CHAT-HF, remarking the suitability of Mistral AI's model for this task.

Table I: **Weighted F1 scores on SemEval2016 and IBM-Debater test sets**. With *Mistral-7B fine-tuned we refer to the best model obtained, i.e. Mistral-7B fine-tuned using LoRA with rank 16 on both SemEval2016 and IBM-Debater (70% of each training set).

	Abortion	Atheism	Climate change	Feminist movement	Hillary Clinton	SemEval2016 (weighted avg)	IBM-Debater (weighted avg)
BERTweet (baseline)	0.65	0.76	0.79	0.65	0.69	0.70	-
RoBERTa (baseline)	0.54	**0.79**	0.80	0.64	0.71	0.68	-
StanceBERTa (baseline)	-	-	-	-	-	-	0.61
Mistral-7B-Instruct-v0.1 (zero-shot)	0.54	0.33	0.55	0.57	0.66	0.54	0.44
Mistral-7B fine-tuned (**ours**)*	**0.71**	0.73	**0.84**	**0.76**	**0.80**	**0.76**	**0.92**

	SemEval2016		IBM-debater	
	F1 score	acc.	F1 score	acc.
Llama-2-7b-chat-hf	0.40	0.41	0.35	0.27
Mistral-7B-Instruct-v0.1	**0.54**	0.53	**0.44**	0.37

Table II: **Results of zero-shot evaluation** on instructed models. F1 score is intended as weighted over the classes (2 for IBM-debater, 3 for SemEval2016).

D. Stance detection on Samuel Paty articles

As the results obtained on unseen datasets are promising, we deployed our best model to predict the stance in a real-world scenario, to qualitatively evaluate its performances. The dataset used is composed of arguments extracted from articles mentioning the murder of Samuel Paty, using RoBERTArg, a RoBERTa-based model trained to detect whether a sentence constitutes an argument or not[10].

We visually inspected some samples and questioned ourselves about which strengths and weaknesses our system could have. In general, we detected improvements on our model in predicting the Favor/Against class, while the baseline predicted "Neutral" 63% of the times.

1) The short target: The Samuel Paty dataset contains articles that were clustered semantically using sentence embeddings. The keyword list for these articles is: separatism, Islamism, macron, Muslim, extremism, France, law. For this reason, we test the general target of "Islam", which is itself very broad, thus supportive and hostile arguments come in many heterogeneous nuances.

After inspecting some predictions, our main conclusions are the following:

- The stance towards this topic is asymmetric: arguments against this target are very explicit, whilst supportive ones might be implicit, therefore difficult to identify.
- Understanding the meaning of an argument with an implicit subject might be difficult. In this case, we assume that the subject matches with our topic. However, even a human reader may often not be able to contextualize an argument without access to the whole article.
- Some arguments are actually punctual facts and can be perceived as positive/negative even if they are not claims but statements. This issue, related to the argument extraction process, might shift the proportion of our prediction "Against Islam".

2) The long target: We could ease the job of the fine-tuned LLM by giving a more detailed or expressive target. We chose "Islamism and fundamentalism are bad for French safety and society" as it clearly takes a side, making more clear also the meaning of a supportive stance. From the visual inspection, though, it seemed that the detailed topic leads to worse results than just "Islam". For example, reports of terrorist attacks are classified as against the given detailed topic, which is clearly a contradiction (see appendix C).

3) Combination of targets: Finally, we consider the possibility of combining the predictions of the stance towards multiple targets. We introduce the target "Equality", according to our assumption that predicting the stance of any argument towards it might be an easier task.

In principle, the combination of "Islam" and "Equality" might help differentiate between various nuances of being in favor "Islam" (e.g. "Every religious community should be integrated in the society" versus "Islam is superior to all the other religions").

Pragmatically, we find it hard to assess if this strategy leads to a significant improvement than only using "Islam" as topic. Further investigations are due to future developments.

IV. Conclusion

In our experiments, we prove Mistral-7B to be the best performing LLM for stance detection, working well even in low data regimes. The most interesting finding is its generalizability on an unseen datasets, even outperforming models fine-tuned on SemEval2016 itself for some classes. Nevertheless, further improvements of the models are left to future work, as well as establishing a pipeline to employ our model with unlabeled datasets, as in the case of our articles about Samuel Paty. Here, we proposed some promising directions, but further investigation is required to properly address the issues found during our qualitative evaluation.

V. Ethical Risk Assessment

In assessing the ethical considerations of our project, particularly its application in online social, media, and news environments, several concerns arise that demand careful attention and mitigation strategies. Firstly, a significant ethical concern revolves around the model's ability to impartially identify both supportive and hostile arguments with equal correctness. Biased inclinations of the model towards a particular stance may lead to biased outcomes, potentially mischaracterizing a balanced article as polarized.

This challenge is often domain or topic-specific, emphasizing the importance of labeled data. Detailed assessments relying on target-specific distributions and metrics become feasible when labeled data is available, allowing to compare recall and precision across different stances (labels) and providing insights into the model performance in identifying diverse perspectives. However, in scenarios lacking labeled data for a particular target or when the model extends its application to entirely new domains, such as in real-world experiments involving knowledge transfer, evaluating predictions using metrics becomes notably challenging. In these instances, the comprehensive engineering of targets becomes imperative. Crafting diverse and well-rounded targets becomes a crucial strategy to compensate for the absence of labeled data, ensuring a more robust and nuanced representation of possible stances across various domains or new contexts. This point of view has been explored in the real-world example application of the model in section III-D3.

Another critical ethical consideration pertains to the potential misuse of the stance detection model, particularly in amplifying polarization. While designed to identify polarization, there is a risk that its application could inadvertently fuel polarization if integrated with recommender systems. Recommending only polarized content aligned with a user's viewpoint might exacerbate echo chambers, reinforcing extreme beliefs and widening ideological divides.

These ethical concerns highlight the necessity for responsible deployment and strategic precautions in the application of our stance detection model. Mitigation strategies involve meticulous topic formulation and careful consideration of potential repercussions when integrated into recommendation algorithms. This approach aims to harness the model insights responsibly, avoiding inadvertent amplification of polarization and promoting balanced discourse within online environments.

References

[1] A. Q. Jiang, A. Sablayrolles, A. Mensch, C. Bamford, D. S. Chaplot, D. de las Casas, F. Bressand, G. Lengyel, G. Lample, L. Saulnier, L. R. Lavaud, M.-A. Lachaux, P. Stock, T. L. Scao, T. Lavril, T. Wang, T. Lacroix, and W. E. Sayed, "Mistral 7b," 2023.

[2] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, A. Rodriguez, A. Joulin, E. Grave, and G. Lample, "Llama: Open and efficient foundation language models," 2023.

[3] Y. Li, S. Bubeck, R. Eldan, A. Del Giorno, S. Gunasekar, and Y. T. Lee, "Textbooks are all you need ii: phi-1.5 technical report," *arXiv preprint arXiv:2309.05463*, 2023.

[4] N. Houlsby, A. Giurgiu, S. Jastrzebski, B. Morrone, Q. de Laroussilhe, A. Gesmundo, M. Attariyan, and S. Gelly, "Parameter-efficient transfer learning for nlp," 2019.

[5] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "Lora: Low-rank adaptation of large language models," 2021.

[6] S. Mohammad, S. Kiritchenko, P. Sobhani, X. Zhu, and C. Cherry, "SemEval-2016 task 6: Detecting stance in tweets," in *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, S. Bethard, M. Carpuat, D. Cer, D. Jurgens, P. Nakov, and T. Zesch, Eds. San Diego, California: Association for Computational Linguistics, Jun. 2016, pp. 31–41. [Online]. Available: https://aclanthology.org/S16-1003

[7] R. Bar-Haim, I. Bhattacharya, F. Dinuzzo, A. Saha, and N. Slonim, "Stance classification of context-dependent claims," in *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, M. Lapata, P. Blunsom, and A. Koller, Eds. Valencia, Spain: Association for Computational Linguistics, Apr. 2017, pp. 251–261. [Online]. Available: https://aclanthology.org/E17-1024

[8] F. Barbieri, J. Camacho-Collados, L. Espinosa Anke, and L. Neves, "TweetEval: Unified benchmark and comparative evaluation for tweet classification," in *Findings of the Association for Computational Linguistics: EMNLP 2020*, T. Cohn, Y. He, and Y. Liu, Eds. Online: Association for Computational Linguistics, Nov. 2020, pp. 1644–1650. [Online]. Available: https://aclanthology.org/2020.findings-emnlp.148

[9] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized bert pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019.

[10] C. Stab, T. Miller, B. Schiller, P. Rai, and I. Gurevych, "Cross-topic argument mining from heterogeneous sources," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, E. Riloff, D. Chiang, J. Hockenmaier, and J. Tsujii, Eds. Brussels, Belgium: Association for Computational Linguistics, Oct.-Nov. 2018, pp. 3664–3674. [Online]. Available: https://aclanthology.org/D18-1402

Appendix

A. Experiment settings

We fine-tune LLMs using working with the python package `lit-gpt`[1] and use the default settings for all the LoRA hyperparameters, except for the rank, as we explained. Specifically, here the detailed hyper-parameter settings:

- learning_rate = 3e-4
- batch_size = 128
- micro_batch_size = 4
- weight_decay = 0.01
- lora_alpha = 16
- lora_dropout = 0.05
- warmup_steps = 100

The weights that are LoRA fine-tuned are the Query and Value matrices.

B. Prompts

The prompt that we used for rank and data regimes experiment:

"Below is an instruction that describes a task. Write a response that appropriately completes the request.
Instruction: Check whether the Article content is in favor, against or neutral respect to the Target. If in favor, your Response will be FAVOR; if against, your response will be AGAINST; if neutral, your response will be NEUTRAL.
Article:
{article}
Target:
{target (topic)}
Response:{stance}"

To select and train the best model for transfer learning, we used a more detailed prompt (the same we use for zero-shot evaluation):

"You are a helpful, respectful and honest assistant for stance detection for a given target always always answer one of the possible options given below as helpfully as possible.

Stance detection is the process of determining whether the author of a article is in support of or against a given target. The target may not always be explicitly mentioned in the text, and the article's stance can be conveyed implicitly through subtext, regional and cultural references, or other implicit meanings.

Please analyze the following article, which is in the political domain, deeply. Consider any subtext, regional and cultural references, or implicit meanings to determine the stance towards the target {target}".

The possible stances are:
- FAVOR: The article has a positive or supportive attitude towards the target, either explicitly or implicitly.
- AGAINST: The article opposes or criticizes the target, either explicitly or implicitly.
- NEUTRAL: The article is neutral or doesn't have a stance towards the target.
Article: {article}
Stance towards the target {target}"

[1] lit-gpt library: https://github.com/Lightning-AI/lit-gpt/tree/main

C. Examples of Predictions on Samuel Paty

These examples are not cherry-picked and give an idea about some good/bad predictions of our model.

1) Topic target: "Islam": generally good for Favor and Against (apart the general problems residing in the data, as we already pointed out), we notice some issues in the Neutral class.

Favor:

- The defence of human life, liberty and freedom of expression is a moral and legal obligation of a rules-based society, not a phobia against Islam.
- It is a cardinal liberal principle that no one should be targeted for being a member of a particular community.
- "I can understand that this creates anger, including murderous anger," says Blom.

Against:

- France, many would argue, is moving toward s a centralised state where civil liberties are threatened.
- In this attack, 12 people were killed.
- The French believe that, at their core, the tenets of religions could be naturally offensive to the adherents of other faiths.

Neutral:

- ""By harming it and harming its beneficiaries and infrastructures, the government is committing an injustice that will endanger millions of people," it added.
- It has been denounced by the United Nations and Amnesty International, with the latter describing the Global Security Bill as "a huge threat to human rights".
- The decision to ban BarakaCity comes as French authorities have launched a crackdown following the murder of Samuel Paty, a teacher who was beheaded by an 18-year-old Chechen refugee after showing his students a caricature of the Prophet Muhammad during a class on freedom of expression.

2) Topic target: "Islamism and fundamentalism are bad for French safety and society": we can note issues with Against and Neutral samples

Favor:

- A universalist principle that once stood for progress has become a defensive partisan slogan.
- On July 14, 2016, a terrorist rammed his truck on a crowded street in Nice, killing 84 people.
- The government counters that the threat is real, pointing to repeated terror attacks and what Macron called the development of a "counter-society" that rejects secularism, equality and other French values and laws.

Against:

- Seventeen people were killed over three days of attacks in January 2015, beginning with the massacre of 12 people at satirical weekly Charlie Hebdo, which had published cartoons of the Prophet Mohammed.
- Christian leaders have also voiced fear the bill will impose undue limits on basic freedoms.
- This allows the individual to freely worship or to abandon their faith without consequence.

Neutral:

- The Muslim community needs a strategy to break the vicious cycle of discrimination creating resentment and isolation, which in turn creates a fertile ground for extremist ideology.
- "[The police] can designate something Muslim as problematic even if it is not violent, they can do the same with something activist that is calling to protest."
- French lawmakers debated a bill on Monday they hope will uproot radical Islam in the country, beliefs that authorities maintain are creeping into public services, associations, some schools and online with the goal of undermining national values.