GreyBox at SemEval-2024 Task 4: Progressive Fine-tuning (for Multilingual Detection of Propaganda Techniques)

Nathan Roll
University of California, Santa Barbara
nroll@ucsb.edu

Calbert Graham
University of Cambridge
crg29@cam.ac.uk

Abstract

We introduce a novel fine-tuning approach that effectively primes transformer-based language models to detect rhetorical and psychological techniques within internet memes. Our end-to-end system retains multilingual and task-general capacities from pretraining stages while adapting to domain intricacies using an increasingly targeted set of examples– achieving competitive rankings across English, Bulgarian, and North Macedonian. We find that our monolingual post-training regimen is sufficient to improve task performance in 17 language varieties beyond equivalent zero-shot capabilities despite English-only data. To promote further research, we release our code publicly on GitHub: github.com/Nathan-Roll1/GreyBox.

1 Introduction & Background

The digital age has radically transformed the nature of propaganda and disinformation, requiring innovative detection mechanisms attuned to these shifts (DeCook, 2018; Macdonald, 2006; Sparkes-Vian, 2019).

Previous work on propaganda detection (Li et al., 2019) leveraged a logistic regression model to determine whether or not a given passage was propagandistic using vectors based on Linguistic Inquiry and Word Count (LIWC), TF-IDF, BERT, and sentence features. These researchers have reported an F1 score of 66.16%, which significantly outperformed their baseline model. Oliinyk et al. (2020) used a similar architecture on the task, achieving improved performance by replacing manual feature selection with induced sentence-level and article-level vectors. Elhadad et al. (2020) used a variety of machine learning models, including logistic regression, to create an ensemble classifier for COVID-19 misinformation. More recently, there has been an emergence of work focusing on detection of propaganda in memes, with Dimitrov et al. (2021) releasing a corpus of memes, hand-labeled with one of 22 propaganda techniques, and utilizing a fusion of large language models (LLMs) to successfully identify labels for a shared task: "Multilingual Detection of Persuasion Techniques in Memes" (*SemEval 2024 Task 4*).

The purpose of the shared task is to foster the development of systems which detect rhetorical and psychological devices, often propagandistic in nature, from memes (a more comprehensive explanation is available in Dimitrov et al., 2024). It contains the following subtasks:

- *Subtask 1*: Given exclusively the text extracted from a given meme, identify the specific persuasion technique(s) utilized (if any).
- *Subtask 2*: Given both the text and image of a meme, identify the specific technique(s) being utilized (*Subtask 2a*), and whether or not the meme contains any propagandistic techniques (*Subtask 2b*).

Our system primarily tackles *Subtask 1*, using the text of a given meme to identify which, if any, of the devices are present. Our approach builds on Dimitrov et al. (2021), in tackling the challenge by leveraging the comprehensive pretraining of large language models (LLMs) and fine-tuning it with human-annotated examples.

The multilingual and multi-task capabilities of LLMs have been well established, however low-resource languages and tasks often require additional data to meet or exceed human level performance. Given that fine-tuning generally degrades baseline model capabilities (Zhai et al., 2023), this reality presents obstacles when available language data does not extend to desired task contexts or vice-versa. Through interative refinement, we discover that successive fine-tuning rounds – encompassing increasing task-specific data – result in models which better adapt to our specific task while

Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024), pages 888–893
June 20-21, 2024 ©2024 Association for Computational Linguistics

also retaining sufficient multilingual capabilities. Our approach to split the post-training regimen into multiple steps finds support in prior research. Xu et al. (2021) found that multi-stage fine-tuning has downstream benefits, particularly in low-resource settings. ValizadehAslani et al. (2022) examined the challenge of class imbalance by introducing a two-stage fine-tuning strategy in which they initially adjusted the model with a class-balanced 'reweighting' loss to ensure that underrepresented classes are not overlooked.

Our system makes use of the provided English meme data, manually labeled according to the requirements of the corresponding task. A total of 18,650 training examples generated from 11,111 unique memes were provided across the training, development, and validation splits.

This paper describes our system and explores how progressive fine-tuning learns the syntactic and semantic properties of memes, with potential future applications in a variety of tasks. For more details, please see the task paper Dimitrov et al. (2024).

2 System Overview

Our system leverages a novel, multi-stage fine-tuning process which progressively adapts a pretrained LLM (GPT 3.5-Turbo[1]) to the task of identifying persuasion techniques in memes. This process consists of two distinct fine-tuning steps (see figure 1):

1. **Priming for meaning**: Expose the LLM to all released data in the train and validation splits to understand the context, intention, and implied meanings in memes.

2. **Structural adaptation**: Undergo an additional fine-tuning round on only *Subtask 1* data to align to the specific structural requirements of the output.

2.1 Data preparation

Each of the provided .json files were parsed into Python dictionaries, and reformatted into chat-like training examples with the text of the meme as the "user" and the label(s) as the "assistant"[2]. Memes which appeared in multiple subtask train/validation sets (based on the id field) were filtered to only include a single instance of each. The reformatted chat examples were saved as .jsonl files and programmatically uploaded to the OpenAI fine-tuning API[3] for usage.

Language	Rank	F_h	Pr_h	Rec_h
English	5/32	0.670	0.652	0.688
Bulgarian	7/19	0.476	0.438	0.521
N. Macedonian	8/19	0.434	0.440	0.430

Table 1: Official performance on *Subtask 1* languages.

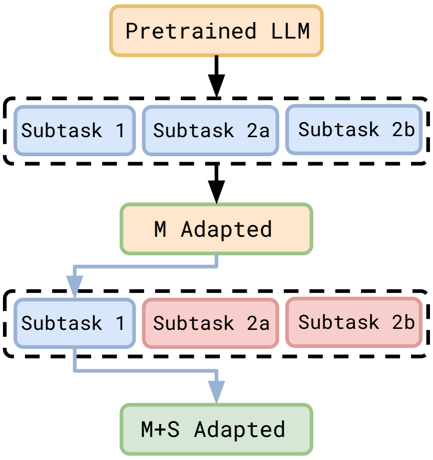

Figure 1: Our implementation of progressive fine-tuning on the *SemEval 2024 Task 4* data. Meaning-based (M) fine-tuning on broader data precedes a more targeted structural (S) fine-tuning step.

2.2 Fine-tuning

2.2.1 Step 1: Priming for Meaning

The priming stage of our fine-tuning process leveraged the train and validation splits across all subtasks. Given that each subtask has a distinct labeling methodology, the purpose of the priming stage is to impart task-specific knowledge (in terms of relevant tokens and their relationship to human-generated labels). Three epochs of fine-tuning were performed on GPT 3.5 Turbo with the training set, using the validation split to ensure that no overfitting was occurring during training. A total of 2.9M tokens were processed during the priming stage.

[1] For zero-shot evaluation, fine-tuning, and experiments we use the gpt-3.5-turbo-1106 model with a context window of 16,385 tokens and a maximum output length of 4,096 tokens.

[2] We leave the system prompt blank in our fine-tuning pipeline to avoid excess costs from input redundancy.

[3] The GPT-3.5 family model weights can only be interacted with using OpenAI's API.

	Avg. F_h	GPT 3.5 Turbo			llama-2-70b-chat			mixtral-8x7b-instruct			Baseline[4]
		F_h	Pr_h	Rec_h	F_h	Pr_h	Rec_h	F_h	Pr_h	Rec_h	F_h
English	**0.276**	**0.281**	0.194	**0.512**	**0.270**	0.180	**0.538**	**0.277**	0.185	**0.556**	0.358
Spanish	0.265	0.275	**0.194**	0.470	0.257	0.176	0.481	0.264	0.179	0.503	" "
French	0.264	0.268	0.187	0.472	0.250	0.170	0.466	0.274	**0.186**	0.524	" "
Haitian Creole	0.258	0.259	0.193	0.393	0.256	0.181	0.438	0.259	0.176	0.492	" "
Ukrainian	0.257	0.265	0.189	0.443	0.246	0.166	0.469	0.261	0.176	0.500	" "
Turkish	0.253	0.264	0.190	0.432	0.239	0.165	0.432	0.257	0.176	0.478	" "
Finnish	0.253	0.264	0.192	0.426	0.231	0.160	0.414	0.263	0.180	0.488	" "
Chinese (Simp.)	0.251	0.259	0.184	0.439	0.243	0.168	0.441	0.251	0.172	0.461	" "
Chinese (Trad.)	0.251	0.265	0.191	0.436	0.246	0.172	0.435	0.241	0.166	0.440	" "
Swahili	0.250	0.250	0.183	0.395	0.237	0.166	0.418	0.262	0.181	0.476	" "
Hindi	0.248	0.254	0.183	0.415	0.250	**0.183**	0.397	0.239	0.160	0.469	" "
Arabic	0.246	0.264	0.188	0.445	0.233	0.174	0.352	0.241	0.165	0.447	" "
Yoruba	0.223	0.216	0.183	0.263	0.221	0.154	0.388	0.234	0.162	0.420	" "
Tamil	0.221	0.214	0.183	0.259	0.222	0.162	0.352	0.226	0.156	0.411	" "
Burmese	0.216	0.187	0.194	0.181	0.247	0.175	0.424	0.214	0.148	0.390	" "
Amharic	0.196	0.143	0.141	0.146	0.227	0.157	0.406	0.219	0.147	0.423	" "
Mean	*0.246*	*0.246*	*0.185*	*0.383*	*0.242*	*0.169*	*0.428*	*0.249*	*0.170*	*0.467*	***0.358***

Table 2: Zero-shot performance on the *Subtask 1* development set varies by model and source language.

2.2.2 Step 2: Structural Adaptation

Model finalization involved an additional two epochs of fine-tuning on the pragmatically-primed model, using only data specific to *Subtask 1*. Two epochs of training were performed, however we encourage further study on the impact of hyperparameters on downstream performance.

2.3 Evaluation Metrics

To capture the hierarchical nature of propaganda techniques, we utilize three metrics which weight errors based on their similarity to each other via higher order categories: hierarchical precision (Pr_h), hierarchical recall (Rec_h), and hierarchical F1 score (Pr_h) (Silla and Freitas, 2011). While the official evaluation of the task does not require leaf-node predictions, our system is not designed to output broader categories in cases of ambiguity. Further justification for the usage of these metrics, along with the exact hierarchy, is provided in Dimitrov et al. (2024).

2.3.1 Hierarchical Precision

Hierarchical precision (Pr_h) measures, in aggregate, the quality of each prediction. This metric is defined as the weighted sum of the predicted classes and their ancestors in the hierarchy, normalized by the total weight of the predicted classes across all test examples. It is given by:

$$Pr_h = \frac{\sum_i |P_i \cap T_i|}{\sum_i |P_i|}$$

Where P_i is the set consisting of the most classes predicted for each test example i, and all of its ancestor classes; T_i is the set consisting of the true most specific class(es) of test example i, and all ancestor classes.

2.3.2 Hierarchical Recall

Similar to hierarchical precision, hierarchical recall (Rec_h) measures the total capture of correct predictions. It is expressed as:

$$Rec_h = \frac{\sum_i |P_i \cap T_i|}{\sum_i |T_i|}$$

2.3.3 Hierarchical F1 Score

The hierarchical F-1 score (F_h) combines both hierarchical precision and recall (using a harmonic mean) to provide a single measure of model performance. It is computed as:

$$F_h = \frac{2 * Pr_h * Rec_h}{Pr_h + Rec_h}$$

This is also the official evaluation metric used to rank performance in *Subtask 1* and *Subtask 2a*.

3 Analysis

We benchmark the performance of our progressively fine-tuned model, and its intermediates, on the Subtask 1 development set. To further explore multilingual capabilities across post-training steps, we create 16 translated versions[5] of the held-out data encompassing a wide variety of languages.

[5] Translation was performed by the Google Translate API: `cloud.google.com/translate`

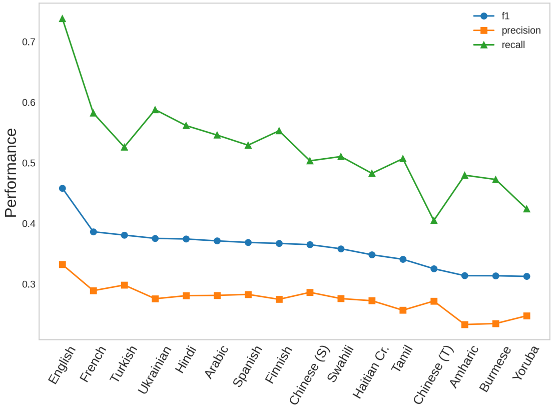

Figure 2: **Primed model (intermediate)**: Recall remains higher than precision in the intermediate model, generally indicating over-prediction. We also find that the relative performance of language varieties shift substantially.

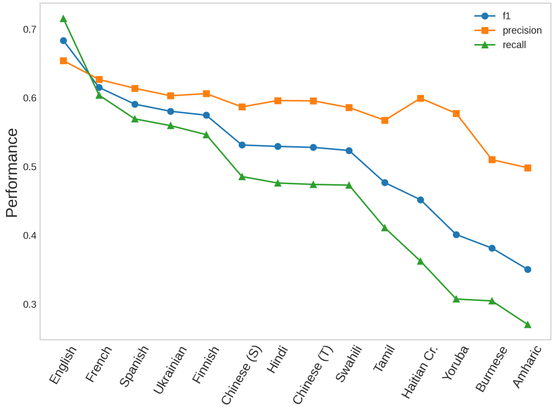

Figure 3: **Final Model**: The structure-tuned model exhibits the highest performance for most languages, included English.

3.1 Zero-Shot

We evaluate the capabilities of three popular out-of-the-box LLMs: OpenAI's GPT 3.5 Turbo, Meta's Llama 2 70B Chat model, and Mistral AI's Mixtral 8x7B instruct mixture of experts (MoE) model. Despite some variation in training data and architecture (see table 2), our tests reveal a consistent bias towards more-common languages (or those closely related to common languages). Furthermore, we find that multilingual capabilities do extend, at least in part, to the propaganda detection task.

3.2 Intermediate Model

After the first fine-tuning step (see section 2.2.1), we again evaluate how the 'meaning-primed' LLM performs in a multilingual setting in fig. 2. Despite English-only fine-tuning data, we find within-language performance improvements in nearly all settings. Our results also indicate that this step also improved some languages more than others, however these shifts do not have any clear syntactic, orthographic, or morphological basis.

3.3 Final Model

After the structural fine-tuning step described in section 2.2.2 , hierarchical F1, precision, and recall demonstrate further gains (see fig. 3). Again, despite English-only data, most languages[6] outperform zero-shot and intermediate counterparts. This is the version of the model which produced our official submissions for *Subtask 1*.

3.4 Multilingual Gains

In addition to producing the highest overall scores (likely a consequence of English-dominant pretraining and fine-tuning data), English also demonstrated the highest gain from additional data, as summarized in fig. 4. While both the priming and structural adaptation phases contributed positively, our results show that the latter was generally more impactful. We hypothesize that labeling differences across related subtask data prevented further performance increases between zero-shot and intermediate evaluation contexts. However, the minor modifications to the evaluation function which allowed for non-exact Python syntax and technique capitalization in the intermediate step would likely only serve to boost reported metrics.

4 Conclusion

Our work highlights the challenges inherent in adapting language models to tasks where relevant information deviates in format and/or linguistic scope from that of the desired output. Our results indicate that progressive fine-tuning offers a promising method for bridging this gap. By tailoring a standard LLM to effectively identify persuasion techniques within multilingual memes, we demonstrate the potential for decoupling syntactic requirements from task-specific 'understanding'. Although monolingual in post-training, this method yielded performance gains across all evaluated lan-

[6] Due to orthographic complications, we were unable to perform a final analysis on Arabic and Turkish.

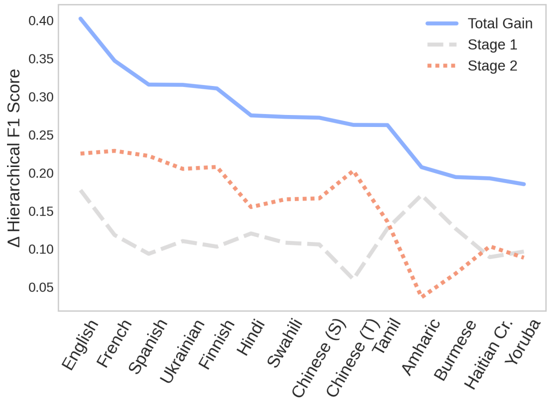

Figure 4: **Relative Performance**: English produced the highest overall increase between zero-shot and final performance, with the highest delta (in percentage points) coming from Stage 2 of the fine-tuning process.

guages compared to zero-shot settings, implying similar capabilities across a wide variety of use cases.

Nevertheless, this work prompts further questions regarding the interplay between pre-training corpora, post-training regimes, and the nature of evaluation data. Our results also call for further work in understanding how the data integration process impacts downstream performance– specifically in comparing our progressive fine-tuning approach to more common single-stage methods. Crucially, our findings reinforce the urgent need to investigate and mitigate biases in LLMs (Lai et al., 2023; Navigli et al., 2023) that impact their performance across varied language communities and use cases.

5 Acknowledgments

We thank Simon Todd for his input and suggestions. We also acknowledge the task organizers for designing such a topical and valuable challenge.

References

Julia R DeCook. 2018. Memes and symbolic violence:# proudboys and the use of memes for propaganda and the construction of collective identity. *Learning, Media and Technology*, 43(4):485–504.

Dimitar Dimitrov, Firoj Alam, Maram Hasanain, Abul Hasnat, Fabrizio Silvestri, Preslav Nakov, and Giovanni Da San Martino. 2024. Semeval-2024 task 4: Multilingual detection of persuasion techniques in memes. In *Proceedings of the 18th International Workshop on Semantic Evaluation*, SemEval 2024, Mexico City, Mexico.

Dimitar Dimitrov, Bishr Bin Ali, Shaden Shaar, Firoj Alam, Fabrizio Silvestri, Hamed Firooz, Preslav Nakov, and Giovanni Da San Martino. 2021. Detecting Propaganda Techniques in Memes. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6603–6617, Online. Association for Computational Linguistics.

Mohamed K. Elhadad, Kin Fun Li, and Fayez Gebali. 2020. Detecting Misleading Information on COVID-19. *IEEE access: practical innovations, open solutions*, 8:165201–165215.

Viet Dac Lai, Nghia Trung Ngo, Amir Pouran Ben Veyseh, Hieu Man, Franck Dernoncourt, Trung Bui, and Thien Huu Nguyen. 2023. Chatgpt beyond english: Towards a comprehensive evaluation of large language models in multilingual learning. *arXiv preprint arXiv:2304.05613*.

Jinfen Li, Zhihao Ye, and Lu Xiao. 2019. Detection of Propaganda Using Logistic Regression. In *Proceedings of the Second Workshop on Natural Language Processing for Internet Freedom: Censorship, Disinformation, and Propaganda*, pages 119–124, Hong Kong, China. Association for Computational Linguistics.

Scot Macdonald. 2006. *Propaganda and Information Warfare in the Twenty-First Century: Altered images and deception operations*. Routledge.

Roberto Navigli, Simone Conia, and Björn Ross. 2023. Biases in large language models: Origins, inventory and discussion. *ACM Journal of Data and Information Quality*.

Vitaliia-Anna Oliinyk, Victoria Vysotska, Yevhen Burov, Khrystyna Mykich, and Vítor Basto Fernandes. 2020. Propaganda Detection in Text Data Based on NLP and Machine Learning. In *MoMLeT+DS*.

Carlos N Silla and Alex A Freitas. 2011. A survey of hierarchical classification across different application domains. *Data mining and knowledge discovery*, 22:31–72.

Cassian Sparkes-Vian. 2019. Digital propaganda: The tyranny of ignorance. *Critical sociology*, 45(3):393–409.

Taha ValizadehAslani, Yiwen Shi, Jing Wang, Ping Ren, Yi Zhang, Meng Hu, Liang Zhao, and Hualou Liang. 2022. Two-Stage Fine-Tuning: A Novel Strategy for Learning Class-Imbalanced Data. ArXiv:2207.10858 [cs].

Haoran Xu, Seth Ebner, Mahsa Yarmohammadi, Aaron Steven White, Benjamin Van Durme, and Kenton Murray. 2021. Gradual Fine-Tuning for Low-Resource Domain Adaptation. ArXiv:2103.02205 [cs].

Yuexiang Zhai, Shengbang Tong, Xiao Li, Mu Cai, Qing Qu, Yong Jae Lee, and Yi Ma. 2023. Investigating the catastrophic forgetting in multimodal large language models. *arXiv preprint arXiv:2309.10313*.

A Appendix

```
[{...}, {
      "id": "125",
      "text": "I HATE TRUMP\n\nMOST TERRORIST DO",
      "labels": [
        "Loaded Language",
        "Name calling/Labeling"
    ],
    "link": "https://..."
    }, {...}...]
```

Listing 1: Pre-formatting .json snippet

```
{...}, {
    "messages": [
        {"role": "system", "content": ""},
        {"role": "user", "content": "I HATE TRUMP\n\
     nMOST TERRORIST DO"},
        {"role": "assistant", "content": "["Loaded
     Language","Name calling/Labeling"]"}
    ]
}, {...}
```

Listing 2: Post-formatting .jsonl snippet

GPT 3.5-Turbo zero-shot prompt

System Prompt: Respond only with a python list, nothing more. Identify which, if any, of the following propoganda labels apply to the given meme: ['Name Calling','Doubt','Smears','Reductio ad Hitlerum','Bandwagon','Glittering Generalities','Exaggeration','Loaded Language','Flag Waving','Appeal to Fear','Slogans','Repetition','Intentional Vagueness','Straw Man','Red Herring','Whataboutism','Causal Oversimplification','Black & White Fallacy','Thought Terminating Cliché']

Input: `<MEME TEXT>`

Mixtral 8x7b zero-shot prompt

Input: Respond only with a python list, nothing more. Identify which, if any, of the following propaganda labels apply to the given meme: ['Name Calling','Doubt','Smears','Reductio ad Hitlerum','Bandwagon','Glittering Generalities','Exaggeration','Loaded Language','Flag Waving','Appeal to Fear','Slogans','Repetition','Intentional Vagueness','Straw Man','Red Herring','Whataboutism','Causal Oversimplification','Black & White Fallacy','Thought Terminating Cliché']. Meme: `<MEME TEXT>`

Llama 2 70b zero-shot prompt

Input: Respond only with a python list, nothing more. Identify which, if any, of the following propaganda labels apply to the given meme: ['Name Calling','Doubt','Smears','Reductio ad Hitlerum','Bandwagon','Glittering Generalities','Exaggeration','Loaded Language','Flag Waving','Appeal to Fear','Slogans','Repetition','Intentional Vagueness','Straw Man','Red Herring','Whataboutism','Causal Oversimplification','Black & White Fallacy','Thought Terminating Cliché']. Meme: `<MEME TEXT>`