Transforming Causal LLM into MLM Encoder for Detecting Social Media Manipulation in Telegram

Anton Bazdyrev Ivan Bashtovyi Ivan Havlytskyi
Oleksandr Kharytonov Artur Khodakovskyi
National Technical University of Ukraine "Igor Sikorsky Kyiv Polytechnic Institute"

Abstract

We participated in the Fourth UNLP shared task on detecting social media manipulation in Ukrainian Telegram posts (Kyslyi et al., 2025), addressing both multilabel technique classification and token-level span identification. We propose two complementary solutions: for classification, we fine-tune the decoder-only model with class-balanced grid-search thresholding and ensembling. For span detection, we convert causal LLM into a bidirectional encoder via masked language modeling pretraining on large Ukrainian and Russian news corpora before fine-tuning. Our solutions achieve SOTA metric results on both shared task track. Our work demonstrates the efficacy of bidirectional pretraining for decoder-only LLMs and robust threshold optimization, contributing new methods for disinformation detection in low-resource languages.

1 Introduction

1.1 Motivation & Context

Disinformation on social media poses significant threats to public discourse and democratic processes. In the Ukrainian context, Telegram is a primary channel for news dissemination and propaganda, where rhetorical manipulation techniques can influence opinions without factual support. Accurate detection of these techniques at both the document and span levels is crucial for fact-checking, media literacy, and automated moderation.

1.2 Shared Task Overview

The Fourth UNLP workshop, held alongside ACL 2025, hosted a shared task on detecting social media manipulation in Ukrainian Telegram posts. Participants addressed two subtasks: multilabel classification of manipulation techniques per post and char-level identification of manipulative spans. The dataset comprises 9,500 posts annotated by media experts.

1.3 Contributions

We make 2 key contributions:

1. We demonstrate that threshold optimization via grid search regularized with respect to the class balance improves F scores for both shared task tracks.

2. We introduce a bidirectional pretraining procedure for converting a decoder-only LLM into an encoder via masked language modeling on large Ukrainian and Russian corpora, yielding superior span detection performance.

2 Related Work

2.1 Disinformation & Propaganda Detection

One of the very first works to address the task of detecting manipulative techniques in texts in detail was written by Da San Martino et al. (2019). It introduces the task of fine-grained propaganda analysis, which involves identifying specific text fragments that contain propaganda techniques and classifying them by type. The issue of manipulation and propaganda in the media is also explored in the context of the Ukrainian media space, especially in Telegram channels. For example, in the study by Steblyna (2022), pro-Kremlin propaganda in popular Odessa-based Telegram channels is detected using frame analysis. This topic is highly important due to the ongoing Russo-Ukrainian war.

An additional challenge in manipulation detection in social media is domain shift, especially when it comes to specific sources like Telegram. In a recent study by (Bazdyrev, 2025), it is shown that Telegram channel data containing manipulative content related to the Russia-Ukraine war significantly differs from more general news and social texts. The author conclude that domain-adaptive pretraining of models on Telegram corpora is necessary. Given that our task is situated in a similar domain, we likewise apply pretraining on Telegram

Proceedings of the Fourth Ukrainian Natural Language Processing Workshop (UNLP 2025), pages 112–119

posts to improve the model's robustness to source-specific characteristics and the stylistics of manipulative content.

2.2 LLMs in Low-Resource Languages

Applying large language models (LLMs) to low-resource languages presents a significant challenge, owing to the lack of high-quality training data and their limited representation in existing pre-training corpora. Researchers have investigated several remedies, including further pre-training on synthetic corpora generated with machine translation (Joshi et al., 2024) and the injection of structured linguistic knowledge through adapters and knowledge graphs (Gurgurov et al., 2024).

While the field remains challenging, recent initiatives such as Meta's No Language Left Behind project (Costa-Jussà et al., 2022) and multilingual evaluation benchmarks like XTREME (Hu et al., 2020) have pushed companies to invest more seriously in improving multilingual coverage. Nevertheless, performance in truly low-resource settings is still lagging, especially in tasks requiring domain adaptation or fine-grained understanding.

2.3 Adapting Decoder Models for Encoder-Specific Tasks

Recent studies have explored methods to adapt decoder-only models for encoder-specific tasks by addressing their causal, unidirectional attention limitations. Proposed solutions range from training-free to complex multiple stage pretraining pipelines.

Training-free methods enhance models without further training. Springer et al. (2024) showed that repeating input text (echo) improves embeddings. Fu et al. (2024) proposed feeding each layer's decoded sentence embedding to the beginning of the sentence in the next layer's input for pseudo-bidirectional context.

Another line of work explores modifying attention behavior during fine-tuning to enable bidirectional context. Li et al. (2023) removed causal masks entirely when fine-tuning LLaMA2 for tasks like classification and named entity recognition (NER). Li and Li (2023) enabled bidirectional attention only in the final layer to improve sentence embeddings. Dukić and Šnajder (2024) extended this idea across multiple layers for NER and chunking tasks. Extending this line of work, Suganthan et al. (2025) made a in-depth evaluation of different causal unmasking strategies across a wide set of tasks.

Incorporating additional pretraining, BehnamGhader et al. (2024) introduced LLM2Vec, a method that applies two stage pretraining before fine-tuning.

3 Dataset

3.1 Data Source & Annotation

The UNLP shared task dataset[1] is a multilingual annotated collection of social media posts, mainly in the context of the ongoing war in Ukraine. It is annotated for the presence of manipulation and the corresponding manipulative spans. A single dataset is used for both tasks. For the classification task, the goal is to predict the binary manipulative label. For the span detection task, the model must also identify character spans (i.e., trigger_words) responsible for manipulation. Annotation guidelines are available at the shared task repository.

3.2 Structure & Target Format

Each data sample in the dataset includes the following fields:

- `id`: A unique identifier for the message.
- `content`: The full text of the social media post.
- `lang`: The language code of the post (e.g., `uk` for Ukrainian, `ru` for Russian).
- `manipulative`: A binary label indicating whether the content is manipulative (`True`) or not (`False`).
- `techniques`: A list of manipulation techniques used in the message (e.g., `loaded_language`, `euphoria`, `cherry_picking`).
- `trigger_words`: A list of character-span indices identifying the positions of manipulative text segments within the content. This enables fine-grained span-level supervision for models.

The dataset provides distinct target formats for the two subtasks:

[1] https://github.com/unlp-workshop/unlp-2025-shared-task

1. **Classification:** The target is a multi-label binary vector over 10 manipulation categories.

2. **Span Identification:** The target consists of character-level spans for each sample where manipulative content occurs.

3.3 Data Splits & Stratification

Since the dataset is shared between the classification and span identification tasks, the same split is suitable for both. This approach ensures consistency across tasks and maintains label balance.

We divided the dataset into 5 folds using multi-label stratified K-Fold cross-validation. One of the folds was selected as the validation set, while the remaining four folds were used for training. The test set corresponds to the official leaderboard data provided by the competition organizers and was not used during training or validation.

Split	Posts	Avg. Chars
Train	3,058	612
Val	764	588
LB	5,735	590

Table 1: Dataset Statistics

3.4 Pretraining Corpora

We also prepared a pretraining news corpora, constructed by merging two publicly available datasets:

- Ukrainian news: 200K documents[2]
- Russian news: QA pairs[3]

4 Evaluation Metric and Threshold Optimization

4.1 Evaluation Metric: F_1 Score

The F_1 score is a widely used metric for evaluating classification models, particularly under class imbalance, as it balances precision and recall.

We evaluated our tasks with F_1, but with different levels of aggregation. For more detailed information, see Table 2.

Given the multi-label nature of the classification task and the imbalance between classes, we focused on optimizing the F1-score during training and postprocessing. To convert predicted probabilities into binary decisions, we performed a class-specific threshold search. This approach allowed us to handle both frequent and rare classes more effectively, rather than relying on a fixed threshold.

Task	Evaluation Metric
Classification	Macro-averaged F_1
Span detection	Character-level F_1

Table 2: Evaluation metrics used for each task.

4.2 F_1-Maximizing Grid Search

For each class, we perform an independent grid search over $t \in [0, 1]$ to find the threshold that maximizes validation F_1:

$$t_{\text{gs}} = \arg\max_t \ F_{1\,\text{val}}(t).$$

While this yields the highest F_1 on local cross-validation, it risks overfitting to validation idiosyncrasies.

4.3 Class-Balance Regularization

To counteract overfitting, we select a threshold that matches the predicted positive rate to the true class prevalence. Denote by r^* the true positive rate and by $r(t)$ the predicted positive rate at threshold t. We choose

$$t_{\text{cb}} = \arg\min_t |\, r(t) - r^*|.$$

This ensures the classifier's output distribution mirrors the dataset's class balance, enhancing stability.

4.4 Alternative Method

We also evaluated the thresholding method of Lipton (Lipton et al., 2014), but found its performance inferior to hybrid the F_1-maximizing and class-balance approach in our setting.

4.5 Hybrid Threshold

We average the two thresholds to obtain

$$t_{\text{final}} = \alpha\, t_{\text{gs}} + \beta\, t_{\text{cb}},$$

where the weights are defined as

$$\alpha = \beta = \frac{1}{2}.$$

Thereby combining peak F_1 performance with distributional robustness.

[2] https://huggingface.co/datasets/zeusfsx/ukrainian-news

[3] https://huggingface.co/datasets/AIR-Bench/qa_news_ru

5 Experimental Setup

5.1 Technique Classification

We conducted a series of experiments[4] with such models as Aya-Expanse (Dang et al., 2024), LLaMA3 (AI@Meta, 2024), and Mistral-Large (Mistral AI team, 2024) on held-out validation data, evaluating our competiton metric. Gemma2 consistently outperformed all alternatives, demonstrating superior capacity to capture nuanced patterns in the text. Accordingly, Gemma2-27B was adopted as the core architecture for our classification pipeline.

5.1.1 Performance Summary

Results in Table 3 confirm that scaling to larger decoder-only architectures and combining F_1-maximizing grid search with class-balance regularization [4.5] yields solid performance and robust generalization across public and private leaderboards.

5.2 Span Identification

The nature of the sequence labeling task requires models to be capable of bidirectional contextual understanding. Consequently, our experiments were primarily focused on encoder-only architectures, including models such as mBERT (Devlin et al., 2018), XLM-RoBERTa (Conneau et al., 2019), EuroBERT (Boizard et al., 2025), mDeBERTaV3 (He et al., 2021), Aya-101 (encoder) (Üstün et al., 2024).

We also investigated whether large-scale architectures with robust pretraining could overcome their inherent unidirectional limitations. We experimented with decoder-only architectures, including Mistral (Mistral AI team, 2024), Phi4 (Abdin et al., 2024), LLaMA3 (AI@Meta, 2024), Gemma2 (Gemma Team, 2024), Gemma3 (Gemma Team, 2025). Among these, Gemma models performed competitively, achieving results comparable to encoder-only models.

5.2.1 Bidirectional Pretraining

Given Gemma's promising performance despite its unidirectional attention, we explored strategies to enhance its bidirectional capabilities. Motivated by approaches outlined in related literature [2.3], we adopted a two-stage training pipeline:

1. *Causal Unmasking via Masked Language Modeling (MLM)*: We conducted MLM pretraining on domain-related corpra [3.4] to improve Gemma2's bidirectional context modeling capabilities, which resulted to what we call the **biGemma2** encoder model.

2. *Span Identification Fine-tuning:* Subsequently, we fine-tuned the model specifically for span identification, optimizing its ability to detect token-level manipulation.

5.2.2 Performance Summary

We employed F1-Maximizing Grid Search [4.2] for threshold selection. While we experimented with Class-Balance Regularization [4.3, 4.5], we found it less effective as our data splits were stratified by classification labels, resulting in different span distributions and more balanced classes compared to the classification task.

[4] https://github.com/AntonBazdyrev/unlp2025_shared_task

Model	Local Validation	Public LB	Private LB
Gemma2-27b (ensemble)	-	**0.474**	**0.494**
Gemma2-27b	**0.500**	0.460	0.481
Gemma2-9b	0.496	0.440	0.480
Gemma3-27b	0.483	0.439	0.468
Gemma2-27b (Lipton)	0.493	0.428	0.457
Gemma2-2b (translated)	0.413	0.375	0.370
Aya-Expanse-8b	0.419	0.389	0.414
Aya-101	0.307	-	-
LLaMA3.2-3b translated texts	0.410	0.334	0.357
Phi-4	0.412	-	-
Mistral-Large-123b	0.458	-	-

Table 3: Technique Classification Performance (Macro-F_1)

Our bidirectional Gemma2-27B[5] achieves Char-F_1 of 0.640, outperforming both encoder-only and decoder-only baselines. Table 4 presents performance metrics across models.

6 Alternative Approaches

In addition to our primary architectures, we explored several complementary strategies. Although these methods offered conceptual advantages, none outperformed our main models during evaluation.

6.1 Technique Classification

6.1.1 Translation-Based Methods

To leverage mature English-language LLMs, we translated Ukrainian posts into English and applied LLaMA3 and Gemma2 for multilabel technique classification. Despite the strong performance of these models in English, translation-induced noise and domain mismatch significantly degraded their macro-F_1 scores compared to models trained directly on Ukrainian text. This translation approach is applicable only to the classification task since span detection requires precise character-level alignment with the original text.

[5] https://huggingface.co/ABazdyrev/bigemma-2-27b-lora

6.1.2 Zero-Shot Classification & Annotation Consistency

In a zero-shot evaluation, GPT-4o achieved an F1 score of 0.32 for identifying manipulation techniques. Introducing a chain-of-thought prompting strategy raised the score to 0.36, but this remained far below the performance obtained via fine-tuning, suggesting potential issues with label reliability. To assess annotation consistency, three experts independently re-annotated a small sample of the dataset according to the original guidelines. The resulting inter-annotator disagreements exposed overlapping class definitions and ambiguous labels, which likely impose an upper bound on model performance. We therefore recommend (1) combining multiple independent estimators—such as diverse human annotators and complementary automated models—and (2) refining and enforcing stricter label definitions. Although these methods have not yet been applied at scale, we anticipate they will improve both the consistency of annotations and the accuracy of social-media manipulation classification and detection.

6.2 Span Identification

6.2.1 LLaDA

We explored LLaDA (Nie et al., 2025), an 8-billion-parameter bidirectional text diffusion model, for token-level span detection. Although its architec-

Model	Local Validation	Public LB	Private LB
biGemma2-27b/Aya-101/mDeBERTa-v3 (ensemble)	-	**0.646**	**0.642**
biGemma2-27b (ensemble)	-	0.646	0.641
biGemma2-27b	**0.650**	0.641	0.640
biGemma2-9b	0.646	0.632	0.637
Gemma3-27b	0.633	0.615	0.613
Gemma2-27b	0.627	0.610	0.611
biLLaMA3.1-8b	0.611	0.615	0.614
LLaMA3.3-70b	0.547	-	-
LLaMA3.1-8b	0.581	0.570	0.572
LLaDA-8b	0.553	0.540	0.542
Mistral-Large-123b	0.599	-	-
Aya-101 (encoder)	0.628	0.611	0.613
mDeBERTa-v3	0.624	0.610	0.612
EuroBERT-2b	0.566	-	-
mT5	0.572	-	-
No ML solution	0.396	0.393	0.389

Table 4: Span Detection Performance (Char-F_1)

ture and scale suggested potential advantages over smaller encoder-only or unidirectional decoder-only models, LLaDA underperformed both mDeBERTa and Gemma2 – likely due to language and domain adaptation challenges.

6.2.2 Two-Stage Positive-Only Pipeline

To mitigate errors in span predictions on non-manipulative posts, we devised a two-stage framework: a binary classifier to detect manipulative posts, followed by a dedicated span identifier applied only to positive instances. This approach reduced spurious spans on clean posts but suffered from error propagation, ultimately yielding lower char-level F_1 than our end-to-end sequence labeling baseline.

6.3 Combining Both Tasks With Auxiliary Loss

Recognizing the potential synergy between tasks, we implemented a dual-head fine-tuning strategy on mDeBERTa and Gemma2, combining a multi-label classification head with a token-level span detection head via an auxiliary loss. Although training remained stable, joint optimization introduced task interference: neither classification macro-F_1 nor span-level char-F_1 improved over separate single-task models.

7 Conclusions & Future Work

7.1 Summary of Findings

Our experiments demonstrate that incorporating bidirectional context into the encoder is essential for accurately identifying span boundaries, yielding a marked improvement over unidirectional baselines. Moreover, we find that naively applied thresholds can exacerbate performance degradation in the presence of class imbalance; instead, class-aware threshold selection consistently maintains precision–recall balance. Finally, out-of-fold ensembling offers a dependable mechanism to smooth out idiosyncratic errors across folds, thereby substantially enhancing model robustness. Collectively, these results underscore the importance of carefully calibrated architectural and post-processing strategies in low-resource settings.

7.2 Broader Impacts

Beyond raw performance gains, our methodological advances have tangible applications for fact-checking and misinformation detection in Ukrainian media ecosystems. By demonstrating transferability of bidirectional pretraining, we pave the way for adoption in other under-resourced languages, where annotated data are scarce and annotation consistency remains a concern. In doing so, we believe this work establishes a new state of the art for a broad array of Ukrainian-language downstream tasks.

7.3 Future Directions

Scaling masked language model pretraining to vastly larger Ukrainian text corpora is an important direction for enriching contextual representations. Equally critical is the establishment of a formal annotation-consistency framework—comprising inter-annotator agreement studies, iterative guideline refinement, and automated label-overlap detection. Together, these measures help ensure cleaner training signals and drive model performance closer to its theoretical upper bound.

Limitations

Despite our advances, this study remains limited by the relatively small and unevenly distributed annotated corpora available for a Ukrainian language, as well as variability in the consistency and quality of disinformation labels.

Acknowledgements

We acknowledge organizers of the UNLP shared task: Nataliia Romanyshyn, Oleksiy Syvokon, Volodymyr Kyrylov, Roman Kyslyi, Volodymyr Sydorskyi.

We also gratefully acknowledge Dr. Pavlo O. Kasyanov, Professor, Head of the Scientific Department of System Mathematics IASA and Dr. Nataliya Dmytrivna Pankratova, Professor, Deputy Director for Scientific Research at the IASA, for their extensive academic support.

References

Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J. Hewett, Mojan Javaheripi, Piero Kauffmann, James R. Lee, Yin Tat Lee, Yuanzhi Li, Weishung Liu, Caio C. T. Mendes, Anh Nguyen, Eric Price, Gustavo de Rosa, Olli Saarikivi, and 8 others. 2024. Phi-4 technical report. *arXiv preprint arXiv:2412.08905*.

AI@Meta. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Anton Bazdyrev. 2025. Russo-ukrainian war disinformation detection in suspicious telegram channels. *arXiv preprint arXiv:2503.05707*.

Parishad BehnamGhader, Vaibhav Adlakha, Marius Mosbach, Dzmitry Bahdanau, Nicolas Chapados, and Siva Reddy. 2024. Llm2vec: Large language models are secretly powerful text encoders. *arXiv preprint arXiv:2404.05961*.

Nicolas Boizard, Hippolyte Gisserot-Boukhlef, Duarte M. Alves, André Martins, Ayoub Hammal, Caio Corro, Céline Hudelot, Emmanuel Malherbe, Étienne Malaboeuf, Fanny Jourdan, Gabriel Hautreux, João Alves, Kevin El-Haddad, Manuel Faysse, Maxime Peyrard, Nuno M. Guerreiro, Patrick Fernandes, Ricardo Rei, and Pierre Colombo. 2025. Eurobert: Scaling multilingual encoders for european languages. *arXiv preprint arXiv:2503.05500*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

Marta R Costa-Jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, and 1 others. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

Giovanni Da San Martino, Seunghak Yu, Alberto Barrón-Cedeño, Rostislav Petrov, and Preslav Nakov. 2019. Fine-grained analysis of propaganda in news articles. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5636–5646, Hong Kong, China. Association for Computational Linguistics.

John Dang, Shivalika Singh, Daniel D'souza, Arash Ahmadian, Alejandro Salamanca, Madeline Smith, Aidan Peppin, Sungjin Hong, Manoj Govindassamy, Terrence Zhao, Sandra Kublik, Meor Amer, Viraat Aryabumi, Jon Ander Campos, Yi-Chern Tan, Tom Kocmi, Florian Strub, Nathan Grinsztajn, Yannis Flet-Berliac, and 26 others. 2024. Aya expanse: Combining research breakthroughs for a new multilingual frontier. *Preprint*, arXiv:2412.04261.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

David Dukić and Jan Šnajder. 2024. Looking right is sometimes right: Investigating the capabilities of decoder-only llms for sequence labeling. *arXiv preprint arXiv:2401.14556*.

Yuchen Fu, Zifeng Cheng, Zhiwei Jiang, Zhonghui Wang, Yafeng Yin, Zhengliang Li, and Qing Gu. 2024. Token prepending: A training-free approach for eliciting better sentence embeddings from llms. *arXiv preprint arXiv:2412.11556*.

Gemma Team. 2024. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*.

Gemma Team. 2025. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*.

Daniil Gurgurov, Mareike Hartmann, and Simon Ostermann. 2024. Adapting multilingual llms to low-resource languages with knowledge graphs via adapters. *arXiv preprint arXiv:2407.01406*.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2021. Debertav3: Improving deberta using electra-style pretraining with gradient-disentangled embedding sharing. *Preprint*, arXiv:2111.09543.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. Xtreme: A massively multilingual multi-task benchmark for evaluating cross-lingual generalisation. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 4411–4421. PMLR.

Raviraj Joshi, Kanishk Singla, Anusha Kamath, Raunak Kalani, Rakesh Paul, Utkarsh Vaidya, Sanjay Singh Chauhan, Niranjan Wartikar, and Eileen Long. 2024. Adapting multilingual llms to low-resource languages using continued pre-training and synthetic corpus. *arXiv preprint arXiv:2410.14815*.

Roman Kyslyi, Nataliia Romanyshyn, and Volodymyr Sydorskyi. 2025. The unlp 2025 shared task on detecting social media manipulation. *ACL 2025*, page to appear.

Xianming Li and Jing Li. 2023. Bellm: Backward dependency enhanced large language model for sentence embeddings. *arXiv preprint arXiv:2311.05296*.

Zongxi Li, Xianming Li, Yuzhang Liu, Haoran Xie, Jing Li, Fu-lee Wang, Qing Li, and Xiaoqin Zhong. 2023. Label supervised llama finetuning. *arXiv preprint arXiv:2310.01208*.

Zachary C. Lipton, Charles Elkan, and Balakrishnan Narayanaswamy. 2014. Thresholding classifiers to maximize f1 score. *arXiv preprint arXiv:1402.1892*.

Mistral AI team. 2024. Large enough. `https://mistral.ai/news/mistral-large-2407`. Blog post; Accessed: 2025-04-17.

Shen Nie, Fengqi Zhu, Zebin You, Xiaolu Zhang, Jingyang Ou, Jun Hu, Jun Zhou, Yankai Lin, Ji-Rong Wen, and Chongxuan Li. 2025. Large language diffusion models. *arXiv preprint arXiv:2502.09992*.

Jacob Mitchell Springer, Suhas Kotha, Daniel Fried, Graham Neubig, and Aditi Raghunathan. 2024. Repetition improves language model embeddings. *arXiv preprint arXiv:2402.15449*.

Natalya Steblyna. 2022. Pro-russian propaganda detection in the most popular telegram channels of odesa region (frame analysis). *Bulletin of Lviv Polytechnic National University: journalism*, 1:80–88.

Paul Suganthan, Fedor Moiseev, Le Yan, Junru Wu, Jianmo Ni, Jay Han, Imed Zitouni, Enrique Alfonseca, Xuanhui Wang, and Zhe Dong. 2025. Adapting decoder-based language models for diverse encoder downstream tasks. *arXiv preprint arXiv:2503.02656*.

Ahmet Üstün, Viraat Aryabumi, Zheng-Xin Yong, Wei-Yin Ko, Daniel D'souza, Gbemileke Onilude, Neel Bhandari, Shivalika Singh, Hui-Lee Ooi, Amr Kayid, Freddie Vargus, Phil Blunsom, Shayne Longpre, Niklas Muennighoff, Marzieh Fadaee, Julia Kreutzer, and Sara Hooker. 2024. Aya model: An instruction finetuned open-access multilingual language model. *arXiv preprint arXiv:2402.07827*.