7-2023

BERTastic at SemEval-2023 Task 3: Fine-Tuning Pretrained Multilingual Transformers – Does Order Matter?

Tarek Mahmoud
Mohamed Bin Zayed University of Artificial Intelligence

Preslav Nakov
Mohamed Bin Zayed University of Artificial Intelligence

BERTastic at SemEval-2023 Task 3: Fine-Tuning Pretrained Multilingual Transformers – Does Order Matter?

Tarek Mahmoud[†◇], **Preslav Nakov**[†]
[†]Mohamed Bin Zayed University of Artificial Intelligence, [◇] Presight.ai
{tarek.mahmoud,preslav.nakov}@mbzuai.ac.ae

Abstract

The naïve approach for fine-tuning pretrained deep learning models on downstream tasks involves feeding them mini-batches of randomly sampled data. In this paper, we propose a more elaborate method for fine-tuning Pretrained Multilingual Transformers (PMTs) on multilingual data. Inspired by the success of curriculum learning approaches, we investigate the significance of fine-tuning PMTs on multilingual data in a sequential fashion language by language. Unlike the curriculum learning paradigm where the model is presented with increasingly complex examples, we do not adopt a notion of "easy" and "hard" samples. Instead, our experiments draw insight from psychological findings on how the human brain processes new information and the persistence of newly learned concepts. We perform our experiments on a challenging news-framing dataset that contains texts in six languages. Our proposed method outperforms the naïve approach by achieving improvements of $\mathbf{2.57}\%$ in terms of F1 score. Even when we supplement the naïve approach with recency fine-tuning, we still achieve an improvement of $\mathbf{1.34}\%$ with a $\mathbf{3.63}\%$ convergence speed-up. Moreover, we are the first to observe an interesting pattern in which deep learning models exhibit a human-like *primacy-recency effect*.

1 Introduction

Deep learning models are state-of-the-art (SOTA) in many fields including natural language processing (NLP). In NLP, the current SOTA models (Wang et al., 2019) are based on transformers, which are deep-learning models with attention mechanism (Vaswani et al., 2017). While many transformers are monolingual, there has been increased research and public interest in multilingual transformers (Doddapaneni et al., 2021). Notably, pretrained transformers, require huge amounts of training data, no matter the domain (Devlin et al., 2019; Dosovitskiy et al., 2021).

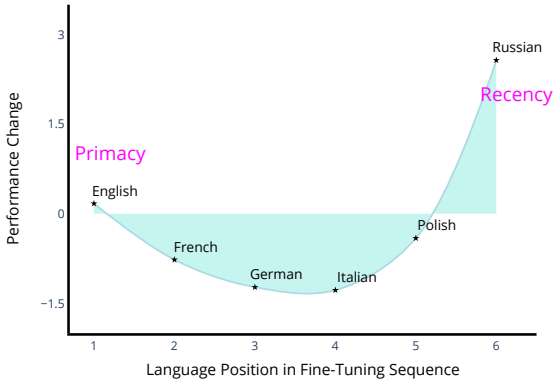


Figure 1: **Human-like primacy-recency effect in deep learning models:** imposing an order on the training data in which PMTs are fine-tuned sequentially language by language matters. PMTs tend to perform relatively better on languages at either end of the training sequence. The performance improvement is measured against the naïve approach of language-agnostic uniform sampling.

Consequently, this results in a substantial carbon footprint (Strubell et al., 2019), which is against global sustainability objectives.[1] There are many approaches to address the AI carbon footprint concerns ranging from using more carbon-efficient energy sources to applying more efficient AI models and training algorithms. Curriculum learning (CL) encompasses a specific class of efficient training strategies for deep learning models. On the one hand, the naïve approach of fine-tuning pretrained deep learning models on downstream tasks involves feeding them mini-batches of randomly sampled subsets of the available training data. On the other hand, in curriculum learning, the idea is to fine-tune the model with a sequence of progressively more challenging examples. This is motivated by and mimics the way humans learn, where we start with simpler concepts and gradually build up more complex ones. Curriculum learning research (Soviany et al., 2022) shows that such a strategy helps the model achieve better performance and converge faster.

[1] https://sdgs.un.org/goals

Proceedings of the The 17th International Workshop on Semantic Evaluation (SemEval-2023), pages 58–63

Contributions Following the success of CL approaches, and inspired by cognitive science, we propose an approach to fine-tuning PMTs on multilingual data. We investigate the significance of doing this in a sequential fashion *language by language*. Unlike CL where the model is presented with increasingly complex examples, we do not adopt a notion of "easy" and "hard" examples. Instead, our experiments draw insight from psychological findings on how the human brain processes new information and the persistence of newly learned concepts (Murdock, 1962).

We perform our experiments on a multi-label text classification dataset (Piskorski et al., 2023) that contains text in six languages: English, French, German, Italian, Polish, and Russian. Our proposed method outperforms the naïve approach by achieving an F1 score gain of $\mathbf{2.57}\%$. Even when we supplement the naïve approach with recency fine-tuning, it achieves an F1 score gain of $\mathbf{1.34}\%$ with a $\mathbf{3.63}\%$ average convergence speed-up. Moreover, we observe an interesting pattern in which deep learning models exhibit a human-like *primacy-recency effect*, which is also commonly referred to as the *serial-position effect* (Murdock, 1962). The effect describes the human tendency to remember the first and the last items in a list more accurately than the ones in the middle. Our contributions are as follows:

- We propose and evaluate fine-tuning PMTs on multilingual data in a sequential fashion language by language.
- We find that a deep learning model exhibits a human-like primacy-recency effect.
- We compare the performance of PMTs fine-tuned on monolingual data versus multilingual data.
- We examine the use of translation for data augmentation and analyze the performance of monolingual versus multilingual pretrained language models.

2 Background

2.1 SemEval Task Description

We perform our experiments on data from the second subtask of task 3 of SemEval-2023 on "Detecting the Genre, the Framing, and the Persuasion Techniques in Online News in a Multi-Lingual Setup" (Piskorski et al., 2023).

Language	Train (%)		Dev (%)		Test (%)	
English	433	*76.0%*	83	*14.6%*	54	*9.5%*
French	158	*60.5%*	53	*20.3%*	50	*19.2%*
German	132	*58.1%*	45	*19.8%*	50	*22.0%*
Italian	227	*62.4%*	76	*20.9%*	61	*16.8%*
Polish	145	*60.2%*	49	*20.3%*	47	*19.5%*
Russian	143	*54.4%*	48	*18.3%*	72	*27.4%*
Spanish	-	-	-	-	30	*100%*
Greek	-	-	-	-	64	*100%*
Georgian	-	-	-	-	29	*100%*
Total	1238	*60.4%*	354	*17.3%*	457	*22.3%*

Table 1: **News framing dataset:** Note that for Spanish, Greek, and Georgian, training data is not provided.

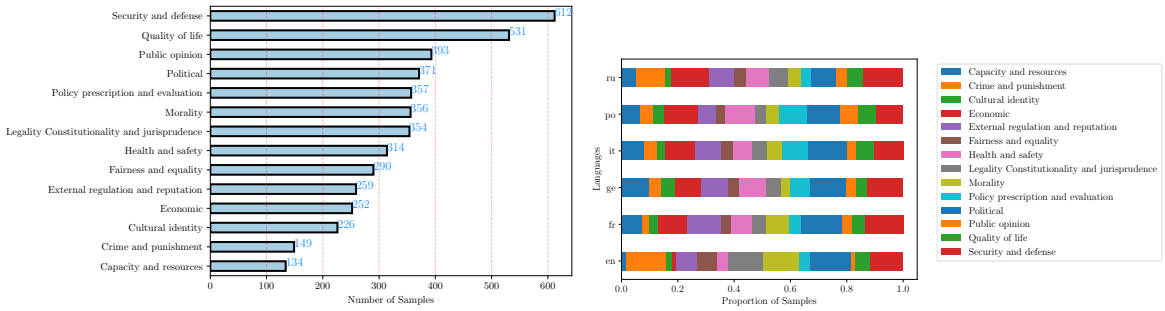
This is a challenging task and it is more nuanced than mere topic classification (Card et al., 2015), e.g., while the topic of a news article may be COVID-19, the framing could be from an economic, political, or/and health perspective(s). In concrete terms, the news framing subtask can be formulated as a multilabel text classification problem. Given a news article, $T_i = [w_1, ..., w_{N_i}]$, where w_i denotes the i^{th} token, and N_i denotes the number of tokens in T_i, the goal is to learn a mapping $f : T \rightarrow S$ where $S = [a_1, ..., a_M]$ and $a_j \in True, False$ denotes whether T_i contains the j^{th} framing label and M denotes the total number of framing labels, which is fourteen in this subtask.

Table 1 shows in detail the $train/dev/test$ splits per language as provided by Piskorski et al. (2023). The dataset has many challenges:

- The classes are imbalanced (Figure 2a).
- The frames' proportions across languages are unequal (Figure 2b).
- Most examples, namely 80.51%, contain more than 512 tokens, which is the maximum sequence length for BERT-like models.
- The number of examples per language is small and so is the collective dataset size.
- Three of the test languages are surprise languages for which no training or development data was provided.

2.2 Related Work

Curriculum Learning A large body of research on CL (Soviany et al., 2022) investigates efficient training strategies for deep learning models. A core idea in CL research is the notion of "easy" and "hard" examples.



(a) (b)

Figure 2: (a) Frames histogram for the training data. (b) Frames proportion by language in the training data.

Qualifying the difficulty of an example requires creating a complexity evaluation method, which could be challenging. Previous evaluation methods in NLP ranged from simply looking at sentence length (Cirik et al., 2016) to examining specialized linguistic features (Jafarpour et al., 2021).

Primacy-Recency Effect Nikishin et al. (2022) showed that deep reinforcement learning models exhibit a primacy bias. Wang et al. (2023) applied a machine learning model to study primacy-peek-recency effect in human subjects. Both papers, however, do not tackle, nor do they observe a primacy-recency effect in deep learning models. To the best of our knowledge, we are the first to propose and to evaluate fine-tuning PMTs on multilingual data in a sequential fashion language by language. We are also the first to observe the human-like primacy-recency effect in deep learning models in general and in PMTs in particular.

3 System Overview

We perform all of our experiments using two transformer-based models, XLM-R with 278M parameters (Conneau et al., 2020) and uncased BERT with 110M parameters (Devlin et al., 2019), using the setup depicted in Figure 3. We illustrate in the upcoming subsections the different training strategies we adopt in our experiments.

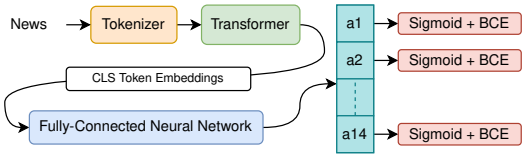


Figure 3: System overview.

3.1 XLM-R-BL: Baseline

The baseline XLM-R-BL is the pretrained XLM-R model without fine-tuning (i.e., it has a randomly initialized classification head). Note that "BL" stands for "baseline." This baseline will be helpful later on to examine how fine-tuning for a particular language contributes to the performance for other languages.

3.2 XLM-R-S: Naïve Approach of Shuffling

We denote by XLM-R-S the naïve approach of shuffling all languages and randomly sampling during fine-tuning of XLM-R-BL. Here, "S" stands for "shuffling."

3.3 XLM-R-S-FT*: Shuffle then Fine-Tune on the Target Language

For a fair comparison of the recency effect contribution of the sequential training approach (section 3.4) against the naïve approach (section 3.2), we fine-tune XLM-R-S six times once on each of the six target languages. Here,"FT" denotes "fine-tuning," and the asterisk denotes a wild-card. We have six target languages, and thus this means we independently fine-tune XLM-R-S six times, once on each of the six respective target languages, resulting in six distinct models. For example, one of the six models that fall under XLM-R-S-FT* is XLM-R-S-FT-EN, which denotes XLM-R-S fine-tuned on English.

3.4 XLM-R-O*: Fine-Tune Sequentially Language by Language

Fine-tuning XLM-R-BL sequentially by imposing an ordered structure on our batches based on the language the examples belong to yields XLM-R-O*, where "O" denotes "order," and, again, the asterisk is a wild-card.

We have six languages, and thus the number of possible fine-tuning sequences is 7! = 720 from which we select six:

- O_1 =[English, French, German, Italian, Polish, Russian]
- O_3 =[Italian, Russian, English, Polish, German, French]
- O_5 =[German, French, Russian, English, Italian, Polish]

The remaining three sequences O_2, O_4, and O_6 are the reversed counterparts of O_1, O_3, and O_5, respectively. Observe that the six sequences were selected so that each of the six languages appears once either at the start or at the end of a sequence.

3.5 BERT-EN and BERT-TR: Data Augmentation via Translation

We investigate the performance of monolingual against multilingual models by fine-tuning BERT on English training data to obtain BERT-EN. We also investigate translation using Google Translate as a means for data augmentation by translating training data from all six languages to English then fine-tuning BERT to obtain *BERT-TR* where "TR" denotes "translation." Needless to say, we also translate test data to English before inference. Note that we find that BERT-TR outperforms all other approaches on the dev set, and thus our *official submission* on the test set is based on BERT-TR.

3.6 Contribution of Each Language to All Other Languages

We study how fine-tuning XLM-R-BL on any single one of the six languages affects the performance of the model on all the other languages, which we did not fine-tune the model on. We use XLM-R-FR, for instance, to denote fine-tuned XLM-R-BL on only French data.

4 Experimental Setup

For all of our experiments, we split the task's training set using stratified sampling[2] to deal with class imbalance. We adopt a $70/30$ $train/dev$ split, and we use the official development set as a test set. In sections 3.2, 3.3, and for BERT-TR from section 3.5, we do the split on the combined dataset from all languages, while for the other sections, we do this for each language independently.

We use an NVIDIA Quadro RTX 6000 GPU to fine-tune our models. We use WordPiece tokenizer for BERT and SentencePiece tokenizer for XLM-R.[3] After tokenization, we truncate the input to 512 tokens. The CLS contextualized embeddings from either BERT or XLM-R are fed to a fully-connected neural network for multilabel classification with a binary cross entropy loss for each of the fourteen labels. We use a learning rate of $1e-5$, and a batch size of 16. We report micro-F1 scores which is a suitable metric for examining fine-tuning sequence ordering impact and is also the task's official metric. We also optimize for the threshold used in the decision functions after the Sigmoids for all models across all experiments. That is, after fine-tuning, we use the training split to find the optimal threshold in terms of micro-F1 score.

We fine-tune epoch by epoch and converge with early stopping with a patience of three epochs. Here, it is important to explain that for XLM-R-S-FT*, we first fine-tune XLM-R-BL on the shuffled dataset and converge with early stopping to obtain XLM-R-S. Next, we fine-tune XLM-R-S on any of the six target languages, say, English, again with early stopping, to obtain XLM-R-S-FT-EN. This process is repeated to obtain five models that cover the five remaining languages.

As for XLM-R-O_1, we fine-tune XLM-R-BL on English with early stopping, then take the resulting model and fine-tune on French with early stopping, and so on all the way till Russian to obtain XLM-R-O_1. In other words, XLM-R-O_1 involves six fine-tuning steps with six early stopping convergences. The same methodology is repeated for all other XLM-R-O* models with the only difference being the different order of languages when fine-tuning.

5 Results

5.1 Order Matters

Looking at Table 2, we see that XLM-R-S-O* has the best overall performance achieving an average F1 score gain of 2.57% and 1.34% over XLM-R-S and XLM-R-S-FT*, respectively. Note that XLM-R-S-FT* outperforms XLM-R-O* only on German. Even though XLM-R-S-FT* exhibits the recency effect by virtue of fine-tuning XLM-R-S on each target language, we can see that the recency effect of XLM-R-O* is stronger. Not only does XLM-R-O* yield superior performance, but it also converges 3.63% quicker as shown in Table 3.

[2] https://github.com/trent-b/iterative-stratification

[3] https://huggingface.co/

Method	English	French	German	Italian	Polish	Russian	Average
XLM-R-BL	47.11	36.39	49.70	43.44	52.00	32.03	43.45
XLM-R-S	63.98	55.45	61.99	**56.33**	63.07	46.01	57.81
XLM-R-S-FT*	65.91	57.4	**63.25**	52.53	65.11	50.00	59.03
XLM-R-O*	**70.21**	**58.51**	61.27	53.75	**65.96**	**52.53**	60.37
$\Delta_{XLM-R-S}$	6.23	3.06	−0.72	−2.58	2.89	6.52	2.57
$\Delta_{XLM-R-S-FT*}$	4.30	1.11	−1.98	1.22	0.85	2.53	1.34

Table 2: **Order matters (dev):** Here, we report the micro-F1 scores on the development set. As discussed, XLM-R-S-FT* consists of six models as does XLM-R-O*. In the table, the goal is to compare the recency effect between XLM-R-S-FT* and XLM-R-O*. Thus, if we consider English F1 scores, the table shows the results of XLM-R-S-FT-EN for XLM-R-S-FT* and XLM-R-O_2 for XLM-R-O* because both XLM-R-S-FT-EN and O_2 have English as the last (most recent) language in the fine-tuning sequence. In a similar fashion, the table displays the results of the respective models according to recency against the corresponding language in the column.

XLM-R-O*	Number of Examples	XLM-R-S-FT*	Number of Examples	Speed Up
XLM-R-O1	15522	XLM-S-FT-RU	15405	−0.76%
XLM-R-O2	15458	XLM-S-FT-EN	18963	18.48%
XLM-R-O3	14336	XLM-S-FT-FR	15450	7.21%
XLM-R-O4	15385	XLM-S-FT-IT	15884	3.14%
XLM-R-O5	15684	XLM-S-FT-PO	15556	−0.82%
XLM-R-O6	16071	XLM-S-FT-GE	15240	−5.45%
			Average Speed Up	3.63%

Table 3: **Convergence speed-up:** We compare each of the six XLM-R-O* models to the respective XLM-R-S-FT* model according to the last language the models were fine-tuned on, and we show the number of examples (i.e., the number of optimization steps multiplied by the batch size) till convergence.

We also make a striking observation where PMTs exhibit a human-like primacy-recency effect as shown in Table 4 and in Figure 1.

5.2 Data Augmentation via Translation

Recall that training data for three of the test languages is not available (Table 1). For this reason, we explore translation as a form of data augmentation, and report our results in Table 5. We find that this approach yields better performance than PMTs, so the system we adopt for the task submission is BERT-TR. We fine-tune BERT on English

Language	Δ_{O_1}	Δ_{O_2}	Δ_{O_3}	Δ_{O_4}	Δ_{O_5}	Δ_{O_6}	Language Position	Δ_{avg}
English	−4.61	+6.23	+0.10	+4.33	−1.38	+11.14	1 (Primacy)	+0.17%
French	+1.12	+1.08	+3.06	+3.70	+0.68	+2.70	2	−0.77%
German	−1.54	−0.55	−1.34	+0.93	+0.68	−0.72	3	−1.23%
Italian	−4.72	−8.34	+0.16	−2.58	+0.64	+0.51	4	−1.28%
Polish	−0.76	−4.83	−2.79	−3.80	+2.89	+0.65	5	−0.41%
Russian	+6.52	+0.46	−3.03	−4.78	−4.91	−2.57	6 (Recency)	+2.57%

Table 4: **Primacy-recency effect:** The table shows the performance difference, Δ, between the naïve approach (XLM-R-S) and the sequential approach (XLM-R-O*). The first six columns show Δ for all six sequences, O_1 to O_6. In the last column, we show Δ averaged by language position in the fine-tuning sequence.

Language	BERT-EN (dev)	BERT-TR (dev)	BERT-TR (test)
English	67.96	**72.68**	51.23
French	-	53.69	53.69
German	-	60.81	60.33
Italian	-	**57.92**	54.50
Polish	-	61.81	58.70
Russian	-	46.77	39.27
Spanish	-	-	47.66
Greek	-	-	52.58
Gregorian	-	-	55.17

Table 5: **Data augmentation via translation (dev+test):** We report micro-F1 scores on the development set. We also show BERT-TR results on the test set from the official leaderboard.

Method	English	French	German	Italian	Polish	Russian
XLM-R-BL	47.11	36.39	49.70	43.44	52.00	32.03
XLM-R-EN	**67.34**	**51.57**	53.41	49.72	51.69	38.27
XLM-R-FR	52.50	50.94	54.88	46.32	54.67	40.42
XLM-R-GE	60.00	49.53	**63.57**	50.08	56.85	37.78
XLM-R-IT	59.85	50.46	55.72	**53.83**	56.66	**42.15**
XLM-R-PO	56.25	45.21	61.95	52.54	**58.32**	33.89
XLM-R-RU	58.01	45.81	53.44	45.44	48.28	41.86

Table 6: **Contribution of each language to all other languages (dev):** We report micro-F1 scores on XLM-R model trained on data from a single language. An untrained XLM-R baseline is also shown for comparison.

data only to study the effect of data augmentation and notice a ~5% gain in performance in BERT-TR over BERT-EN.

5.3 Contribution of Each Language to All Other Languages

In Table 6, we see that fine-tuning XLM-R-BL on any of the six languages benefits all other languages. We also observe that French and Russian seem to benefit negligibly more from out-of-domain English and Italian fine-tuning, respectively, than from in-domain fine-tuning.

6 Conclusion and Future Work

We are the first to illustrate the significance of fine-tuning PMTs sequentially language by language. We show that this not only yields sizable performance gains of $\mathbf{2.57}\%$ over the naïve approach, but it also converges $\mathbf{3.63}\%$ faster when supplemented with recency fine-tuning. Moreover, we are the first to observe a human-like primacy-recency effect in deep learning models in general and PMTs in particular. We also perform other experiments to study data augmentation via translation, and we study how fine-tuning on any single language benefits all other languages.

In the future, we plan to experiment with larger datasets, more languages, multiple runs with different random seeds, different downstream tasks, and more variants of PMTs.

This work has the potential to give us a better understanding of the inner workings of deep learning models by drawing insight from psychology. It also has the potential to improve the way deep learning models in general, and PMTs in particular are pretrained and fine-tuned.

In our work, we looked at only six out of 720 possible sequences of languages for fine-tuning. It would be interesting to investigate other sequences. For example, it is interesting to study keeping languages from the same family closer in the sequence or farther apart. It would also be interesting to frame this research as a CL problem and define a notion of "easy" and "hard" languages.

Acknowledgements

The authors would like to thank Timothy Baldwin for his valuable insights and helpful discussion.

References

Dallas Card, Amber E. Boydstun, Justin H. Gross, Philip Resnik, and Noah A. Smith. 2015. The media frames corpus: Annotations of frames across issues. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, NAACL-HLT'15, pages 438–444, Beijing, China. Association for Computational Linguistics.

Volkan Cirik, Eduard H. Hovy, and Louis-Philippe Morency. 2016. Visualizing and understanding curriculum learning for long short-term memory networks. *CoRR*, abs/1611.06204.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Sumanth Doddapaneni, Gowtham Ramesh, Anoop Kunchukuttan, Pratyush Kumar, and Mitesh M. Khapra. 2021. A primer on pretrained multilingual language models. *CoRR*, abs/2107.00676.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Borna Jafarpour, Dawn Sepehr, and Nick Pogrebnyakov. 2021. Active curriculum learning. In *Proceedings of the First Workshop on Interactive Learning for Natural Language Processing*, pages 40–45, Online. Association for Computational Linguistics.

Bennet B Murdock. 1962. The serial position effect of free recall. *Journal of Experimental Psychology*, 64(5):482–488.

Evgenii Nikishin, Max Schwarzer, Pierluca D'Oro, Pierre-Luc Bacon, and Aaron Courville. 2022. The primacy bias in deep reinforcement learning. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 16828–16847. PMLR.

Jakub Piskorski, Nicolas Stefanovitch, Giovanni Da San Martino, and Preslav Nakov. 2023. SemEval-2023 task 3: Detecting the category, the framing, and the persuasion techniques in online news in a multi-lingual setup. In *Proceedings of the 17th International Workshop on Semantic Evaluation*, SemEval'23, Toronto, Canada.

Petru Soviany, Radu Tudor Ionescu, Paolo Rota, and Nicu Sebe. 2022. Curriculum learning: A survey. *Int. J. Comput. Vision*, 130(6):1526–1565.

Emma Strubell, Ananya Ganesh, and Andrew McCallum. 2019. Energy and policy considerations for deep learning in NLP. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3645–3650, Florence, Italy. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2019. SuperGLUE: A stickier benchmark for general-purpose language understanding systems. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Ping Wang, Hanqin Yang, Jingrui Hou, and Qiao Li. 2023. A machine learning approach to primacy-peak-recency effect-based satisfaction prediction. *Information Processing Management*, 60(2):103196.