From Standard to Slay: Adapting LLMs for Gen Z Slang

Matei Gabriel Cosa **Florian Ranbir Vaid Daefler**

Emilija Milanovic **Kassym Mukhanbetiyar** **Changchen Yu**

Abstract

Large Language Models (LLMs) operate on textual data that naturally exhibit stylistic characteristics which are often correlated with their domain. Changing the style of text may affect LLMs' performance on downstream tasks, especially when the new style is not typically associated with the original domain. This phenomenon occurs when using Gen Z slang in a non-colloquial context. We provide a dataset of standard and Gen Z semantically equivalent pairs of English text from five different domains and use it to test LLM performance on several classification tasks. We showcase the positive effects of fine-tuning popular encoder models on Gen Z speech identification, sentiment analysis, and paraphrase detection. Finally, we fine-tune a decoder model for style transfer, and demonstrate a simple and effective benchmark for evaluating the quality of generated text using the feedback from the fine-tuned classifiers. Our analysis yields a series of interesting findings relating model performance to linguistic peculiarities of Gen Z speech.

1 Introduction

Gen Z slang can be identified with a set of informal expressions, abbreviations, and repurposed words that reflect the social identity, humor, and digital culture of Generation Z. Linguistically, this type of jargon is characterized by several noteworthy properties, including semantic shift ("spill the tea" = gossip), clipping ("fr" = for real), or references to pop culture and memes ("Karen" = pejorative term for entitled, discriminatory people) (Puspita and Ardianto, 2024; Kulkarni and Wang, 2017). Moreover, these linguistic peculiarities are commonly associated with online speech that is colloquial in nature and differs significantly from the style present in scientific publications or news reports.

Given this linguistic complexity as well as limited domain use, it is important to assess wether language models are able to effectively identify, understand, and generate Gen Z slang. For instance, chatbot providers may be interested in analyzing how their systems respond to questions about political or scientific topics written in Gen Z slang. To investigate this issue, we put together a custom dataset of around 5000 pairs of semantically equivalent standard English and Gen Z text spanning five different domains. We believe this in itself could prove valuable for future studies.

Starting from our dataset, we approach a sequence of tasks of increasing complexity. Firstly, we test wether Gen Z text is detrimental to the efficiency of the tokenization process, a starting point in any modern LLM pipeline. Next, we investigate how difficult it is to distinguish Gen Z speech from standard English, identify the sentiment, and check if one piece of text is a paraphrase of another. These problems provide us with valuable insights about the stylistic robustness of LLMs across several domains. We then build a style transfer model that takes standard text and outputs a Gen Z version. To evaluate the output of this model, we use our fine-tuned classifiers from the previous tasks.

2 Experiments

2.1 Dataset

We combine manual data collection with existing sources to build a dataset called ***PairZ*** with 5,002 entries. Each entry consists of a short text (from one sentence to a paragraph) in standard English paired with its Gen Z translation. These documents span a variety of domains: ***Humanities***, ***News***, ***Science***, ***Fiction***, and ***Colloquial*** speech.

Texts from the *Science* and *Humanities* domains were sourced from Wikipedia. For the *News* category, we gathered excerpts from articles published by various news outlets. Fictional content was drawn from well-known series such as *Harry Potter*, *The Lord of the Rings*, and *The Chronicles*

of Narnia. From these sources, we selected clear, plain-English sentences or short paragraphs and we prompted several LLMs to generate the Gen Z translations. All the pairs were inspected by the authors and adjustments were made when necessary. The *Humanities* section was further expanded with text from Gen-Z-Bible, whereas *Colloquial* texts were entirely sourced from the *MLB-Trio/gen_z_slang_dataset* by Park et al. (2024).

2.2 Tokenization

We evaluate three common tokenization methods: ***SentencePiece*** (Park et al., 2024), BERT's ***WordPiece*** (Song et al., 2021), and ***Byte Pair Encoding*** (BPE) (Zouhar et al., 2023). In particular, we are interested in comparing tokenization efficiency (i.e., tokens per words) for Gen Z and standard text across several domains for all tokenizers.

As an artifact of the dataset creation process, lentghs of Gen Z documents may vary widely relative to the corresponding standard text. To control for this phenomenon, we build a regression model (Equation 1) that explains differences in tokenized text length based on domains and differences in the number of words.

Let $standard_i$ and $genz_i$ denote the documents in the ith pair of our dataset. The independent variables include the word count difference Δ^i_{word} (i.e., number of words in $genz_i$ minus number of words in $standard_i$) and domain dummy variables d^i. The dependent variable is the token count difference Δ^i_{token} (i.e., number of tokens of $genz_i$ minus number of tokens of $standard_i$).

$$\Delta^i_{token} = \beta_0 + \beta_1 \Delta^i_{word} + \beta_{d^i} + \epsilon^i \quad (1)$$

2.3 Classification

Style Detection. We fine-tune a sentence-transformer based on ***MPNET*** (Song et al., 2020) for sequence classification. Text written in Gen Z style is assigned to the positive class, while text in standard English is assigned to the negative class.

As a baseline, we first evaluate a simple rule-based classifier. For each input text, we check for the presence of any Gen Z slang terms listed in Park et al. (2024)—properly stemmed and augmented to capture semantic variations such as singular/plural forms and verb conjugations.

We then compare the performance of this baseline against two versions of the fine-tuned model: one where only the classification head is trained, and another where we use ***LoRA*** (Low-Rank Adaptation) (Hu et al., 2021) to also fine-tune the internal attention layers of the transformer.

Sentiment Analysis. We use ***Twitter-roBERTa-base*** (Camacho-collados et al., 2022; Loureiro et al., 2022) to provide ground truth labels (*positive, negative, neutral*) for the documents written in standard English. Our task is to predict the sentiment from Gen Z documents.

We start from a simple discrete representation and regularized logistic regression approach. This naïve baseline is then compared with the pretrained *Twitter-roBERTa-base*. Finally, we perform fine-tuning on 1) the classification head, and 2) the entire model using *LoRA*.

Paraphrase Detection. We take a pair of documents, one in standard English and one in Gen Z slang, and predict whether they are semantically equivalent. The positive samples naturally coincide with our original dataset.

To construct negative examples, we select pairs that are semantically similar, so the task is not trivial, but that are not true paraphrases. Specifically, we embed only the standard English texts using the pre-trained *MPNET* model and, for each sample $standard_i$, retrieve its nearest neighbor $standard_j$ in embedding space. We then form the negative pair by combining $standard_i$ with $genz_j$. In this way, we obtain 5000 negative samples.

We evaluate a fine-tuned *MPNET* model against a sparse baseline. Additionally, we test the classification model on datasets constructed using both the 1-nearest and the 5-nearest neighbors for negative pair creation to assess the impact of semantic proximity on classification difficulty.

2.4 Style Transfer

Models. We prompt two popular decoder LLMs, ***Mistral-7B*** (Jiang et al., 2023) and ***Llama-2-7B*** (Touvron et al., 2023) to generate Gen Z equivalents of standard English documents. Based on preliminary human evaluation, we identify pre-trained *Mistral-7B* to be the better performing model and decide to fine-tune it. We employ parameter-efficient fine-tuning via *LoRA* on the 4-bit quantized model using *bitsandbytes* (Dettmers et al., 2022) for memory optimization.

Benchmark. We also introduce ***ZScore***, a quantitative benchmark for Gen Z Slang style transfer models. We take a subset of 1000 standard English documents from *PairZ*, ensuring a balanced distribution across domains. Gen Z text generated start-

Tokenizer	Overall	Colloquial	Fiction	Humanities	News	Science
SentencePiece Standard	1.48	1.65	1.42	1.39	1.51	1.41
SentencePiece Gen Z	1.56	1.73	1.50	1.44	1.60	1.53
BERT Standard	**1.36**	1.66	**1.34**	**1.25**	**1.26**	**1.31**
BERT Gen Z	1.44	1.65	1.45	1.30	1.37	1.40
BPE Standard	1.42	**1.60**	1.36	1.34	1.45	1.34
BPE Gen Z	1.49	1.68	1.42	1.37	1.52	1.45

Table 1: Average number of the tokens per word for each domain.

ing from each standard document is then scored by our best classifiers for slang identification, sentiment prediction, and paraphrase detection. The goal is to assess 1) if the generated text is written in Gen Z style, 2) if the sentiment is preserved between standard and Gen Z pairs, and 3) if the output is a paraphrase of the input.

3 Results

3.1 Tokenization

Our main finding is that for every tokenizer, Gen Z text results in a higher tokens to words ratio (Table 1). *BERT* is the most efficient overall and across almost every domain, while *SentencePiece* is the least efficient both at the global and domain levels.

Across domains, standard text requires fewer tokens per word than Gen Z with one exception: *BERT* for *Colloquial* speech. This domain has the highest tokens per word ratio for every tokenizer. On the other hand, *Humanities* appears to provide the most efficient tokenization.

	SentencePiece	BERT	BPE
Constant	1.32	1.65	0.99
Δ_{word}	1.32	1.23	1.27
Colloquial	-1.49	-2.53	-1.12
Fiction	0.75	1.26	0.74
Humanities	<u>-0.36</u>	<u>-0.11</u>	-0.52
News	<u>-0.27</u>	1.30	-0.56
Science	2.70	1.73	2.44
R^2	0.90	0.91	0.91

Table 2: Regression Model Summary for Equation 1 for all three tokenizers. All coefficients are signifacant at 0.99 confidence except the underlined ones.

Our regression analysis (Table 2) shows that, on average, for every surplus word $genz_i$ has w.r.t. $standard_i$, tokenized $genz_i$ will have around 1.3 additional tokens compared to tokenized $standard_i$. At domain level, *Colloquial* has a large negative coefficient, likely due to the large number of abbreviations present in the Gen Z documents. At the opposite extreme, *Science* has a large positive coefficient, likely due to scientific jargon. These patterns are consistent across tokenizers.

3.2 Classification

Style Detection. Our word-presence baseline achieves an accuracy of 0.70 on the test set. The model where only the classification head is fine-tuned reaches an accuracy of 0.84, while the fully fine-tuned model achieves an accuracy of 0.94 as shown in Table 3.

We also analyze the shift in embeddings before and after fine-tuning using *t-SNE* (van der Maaten and Hinton, 2008). In the pre-trained model, embeddings of Gen Z and standard English texts overlap considerably, while a clear clustering by domain is observed. This suggests that the pre-trained model captures general semantic similarity and domain structure, but not stylistic variations such as Gen Z and standard English.

After fine-tuning, Gen Z and standard texts become clearly separated in the embedding space, indicating that the model has learned to capture stylistic properties that were previously absent in the representation, as shown in Figure 1. At the same time, domain separation mostly disappears, except for *Colloquial* documents.

Sentiment Anlysis. Our simple baseline achieves an accuracy of 0.64 on the test set, while the pre-trained *Twitter-roBERTa-base* improves to 0.80. The latter indicates there is a notable drop in performance from standard to Gen Z text, since the same model provided the ground truth labels.

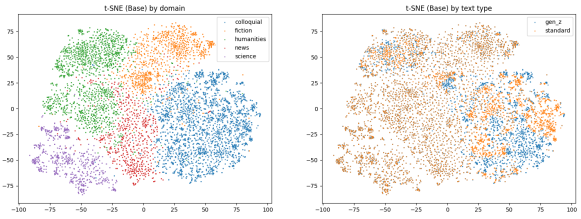
Fine-tuning the classification head raises accuracy to 0.82, while fine-tuning the entire model brings accuracy up to 0.83 (Table 4). Besides having three possible labels, we hypothesize that this task is harder than the other classification problems

Model	Accuracy	Precision	Recall	F1 Score
Word-Presence Model	0.69	0.70	0.69	0.69
MPNET(fine-tuned classification head)	0.84	0.84	0.84	0.84
MPNET (fine-tuned with LoRA)	**0.95**	**0.94**	**0.95**	**0.95**

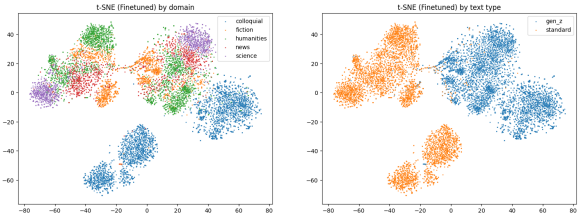
Table 3: Style Detection Model Performance Comparison on Test Set.
Note: Precision, Recall & F1 score are weighted averages.

Model	Accuracy	Precision	Recall	F1 Score
TF-IDF + logistic regression	0.63	0.64	0.63	0.64
Twitter-roBERTa-base (pretrained)	0.80	0.81	0.80	0.80
Twitter-roBERTa-base (fine-tuned classification head)	0.82	**0.83**	0.82	0.82
Twitter-roBERTa-base (fine-tuned with LoRA)	**0.83**	**0.83**	**0.83**	**0.83**

Table 4: Sentiment Analysis Model Performance Comparison on Test Set.
Note: Precision, Recall & F1 score are weighted averages.



(a) Pretrained – Domain (b) Pretrained – Label



(c) Finetuned – Domain (d) Finetuned – Label

Figure 1: *t-SNE* projections of text embeddings from the pre-trained and fine-tuned *MPNET* model. Each point represents a text sample, colored by *domain* or *style*. Fine-tuning yields clearer separation by style, while preserving domain structure.

we consider due to the highly emotional nature of Gen Z slang. This explanation is supported by the linguistic literature (see Section 4) and empirical results (see Figure 3).

Paraphrase Detection. Our baseline model achieves an accuracy of 0.83 when evaluated on the nearest-neighbor dataset., while the fine-tuned model has an accuracy of 0.94 on the same dataset. When we test both models on a new set of 5,000 text pairs constructed using the fifth-nearest neighbor (in theory, simpler and more dissimilar pairs), we observe that the baseline model's accuracy remains similar. In contrast, the fine-tuned model's accuracy further increases to 0.98. This suggests that the fine-tuned model is more robust to variations in input pairs. The takeaway here is that training on more challenging examples generalizes better to easier settings.

Model	Nearest Neighbor	5th Neighbor
Baseline	0.83	0.81
Fine-tuned	0.94	0.98

Table 5: Paraphrase Detection Test Accuracy: baseline and fine-tuned models on the nearest and 5th-nearest neighbor datasets.

3.3 Style Transfer

Fine-tuned *istral-7B* demonstrates strong stylistic adaptation, consistently generating outputs aligned with the tone and lexical patterns of Gen Z slang. Quantitative evaluations performed on *ZScore* highlight that fine-tuning on a relatively small dataset helped the model learn to generate stylistically and semantically accurate paraphrases.

We observe that stylistic accuracy and semantic consistency are high across all domains. Preserving the sentiment proves to be more difficult for all domains but *Science*, where the majority of samples are *neutral*. *News* gives the lowest affective cohesion score, followed by *Fiction* and *Humanities*.

Standard metrics such as *BLEU, ROUGE* and *METEOR* are comparable with the insights provided by *ZScore*, while lacking the granularity of

Model	Gen Z Style	Paraphrase	Sentiment	BLEU	ROUGE-L	METEOR
LLaMA-7B (P)	0.52	0.73	0.82	12.96	34.98	21.20
Mistral-7B (P)	0.23	0.25	0.58	2.10	11.26	3.98
Mistral-7B (F)	**0.98**	**0.95**	**0.83**	**75.87**	**89.71**	**86.28**

Table 6: Evaluation Metrics: *ZScore*, BLEU, ROUGE-L, METEOR. (P) stands for pretrained and (F) for fine-tuned.

Domain	Gen Z Style	Paraphrase	Sentiment	BLEU	ROUGE-L	METEOR
Fiction	0.92	0.93	0.79	75.85	87.36	88.66
News	0.97	**0.98**	0.73	78.41	88.59	89.56
Colloquial	**1.00**	0.95	0.87	**95.17**	**96.46**	**96.85**
Humanities	0.98	0.94	0.81	81.21	88.72	89.22
Science	**1.00**	0.93	**0.95**	58.47	75.79	76.32

Table 7: Evaluation Metrics by Domain (Mistral-7B fine-tuned): *ZScore*, BLEU, ROUGE-L, METEOR.

our benchmark. There is an exception to to this pattern: *Science* yields significantly lower performance according to traditional metrics, while scoring very high on *ZScore*. We suspect this is because of the very different vocabulary used in scientific literature compared to Gen Z slang and the model's capability of re-phrasing scientific topics in colloquial, simple terms. This is a strength of the model that our benchmark is able to pick up on.

4 Related Work

Wegmann et al. (2025) found that stylistic changes and language variation have significant effects for the tokenizer quality. To the best of our knowledge, there is no other study on tokenizers in the context of Gen Z slang.

Slang detection has been studied with pre-Transformer models (Vaswani et al., 2017) with moderate success (Pei et al., 2019). Improvements have been made by Sun et al. (2024), who use both zero-shot decoder models, as well as BERT-style models to reach 0.95 accuracy for slang recognition tasks, a level of performance matched by our fine-tuned classifier.

Emotional intensity and the way in which it is conveyed in Gen Z slang has been studied by Keidar et al. (2022); Kulkarni and Wang (2017). These studies found key characteristics: affective compression ("slay" = strong approval), exaggeration ("this sent me" = made me laugh hard) and humor ("delulu is the solulu" = hope is used as a coping mechanism). These properties lower the performance of sentiment classifiers by making models over-predict *positive* and *negative* classes for *neutral* documents, a phenomenon we also encounter.

Paraphrase detection and style transfer have been explored together by Krishna et al. (2020). The authors propose an unsupervised approach for style transfer, as well as a benchmark that measures style accuracy, semantic similarity to the original text and fluency. We swap fluency for sentiment matching, as it is more relevant for our problem. Moreover, we use a simpler pipeline for style transfer that relies on a small dataset for fine-tuning.

5 Conclusion

Our work demonstrates that the performance of pretrained LLMs and their tokenizers may suffer in the context of Gen Z Slang. Most notably, we remark that while these models seem to be robust w.r.t. stylistic changes for some tasks, they can misinterpret and over-exaggerate the emotional dimension of text. This problem remains present after fine-tuning, which emphasizes the linguistic properties of Gen Z slang' affective component.

On the bright side, fine-tuning, even on our relatively small dataset, showed remarkable decoder model results on *ZScore*, indicating that the models are capable of quickly learning stylistic patterns.

Finally, we conclude that text written in a style not typically associated with its domain creates problems for LLMs. This motivates further study in the area of adapting models for style changes. Exploring more types of slang, dialects, and linguistic variations in a multi-domain context may yield more meaningful results that can help LLM providers better serve a more diverse range of users and backgrounds.

References

Jose Camacho-collados, Kiamehr Rezaee, Talayeh Riahi, Asahi Ushio, Daniel Loureiro, Dimosthenis Antypas, Joanne Boisson, Luis Espinosa Anke, Fangyu Liu, and Eugenio Martínez Cámara. 2022. TweetNLP: Cutting-edge natural language processing for social media. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–49, Abu Dhabi, UAE. Association for Computational Linguistics.

Tim Dettmers, Mike Lewis, Younes Belkada, and Luke Zettlemoyer. 2022. Llm.int8(): 8-bit matrix multiplication for transformers at scale. *arXiv preprint arXiv:2208.07339*.

Gen-Z-Bible. Gen z bible. `https://genz.bible/`.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *Preprint*, arXiv:2106.09685.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

Daphna Keidar, Andreas Opedal, Zhijing Jin, and Mrinmaya Sachan. 2022. Slangvolution: A causal analysis of semantic change and frequency dynamics in slang. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1422–1442, Dublin, Ireland. Association for Computational Linguistics.

Kalpesh Krishna, John Wieting, and Mohit Iyyer. 2020. Reformulating unsupervised style transfer as paraphrase generation. *ArXiv*, abs/2010.05700.

Vivek Kulkarni and William Yang Wang. 2017. Tfw, damngina, juvie, and hotsie-totsie: On the linguistic and social aspects of internet slang. *Preprint*, arXiv:1712.08291.

Daniel Loureiro, Francesco Barbieri, Leonardo Neves, Luis Espinosa Anke, and Jose Camacho-collados. 2022. TimeLMs: Diachronic language models from Twitter. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 251–260, Dublin, Ireland. Association for Computational Linguistics.

Seoyeon Park, Junghyun Kim, and Shim Jiyoung. 2024. Mlbtrio/genz-slang-dataset. `https://huggingface.co/datasets/MLBtrio/genz-slang-dataset`.

Zhengqi Pei, Zhewei Sun, and Yang Xu. 2019. Slang detection and identification. In *Conference on Computational Natural Language Learning*.

Vinka Ganita Puspita and Ardik Ardianto. 2024. Code-switching and slang: An analysis of language dynamics in the everyday lives of generation z. *Linguistics Initiative*, 4(1):76–87.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. 2020. Mpnet: masked and permuted pre-training for language understanding. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA. Curran Associates Inc.

Xinying Song, Alex Salcianu, Yang Song, Dave Dopson, and Denny Zhou. 2021. Fast WordPiece tokenization. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2089–2103, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Zhewei Sun, Qian Hu, Rahul Gupta, Richard Zemel, and Yang Xu. 2024. Toward informal language processing: Knowledge of slang in large language models. In *North American Chapter of the Association for Computational Linguistics*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, and 49 others. 2023. Llama 2: Open foundation and fine-tuned chat models. *Preprint*, arXiv:2307.09288.

Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(86):2579–2605.

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

Anna Wegmann, Dong Nguyen, and David Jurgens. 2025. Tokenization is sensitive to language variation. *Preprint*, arXiv:2502.15343.

Vilém Zouhar, Clara Meister, Juan Gastaldi, Li Du, Tim Vieira, Mrinmaya Sachan, and Ryan Cotterell. 2023. A formal perspective on byte-pair encoding. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 598–614, Toronto, Canada. Association for Computational Linguistics.

A Tokenizers

A.1 Plot of Tokens-per-Word Ratio

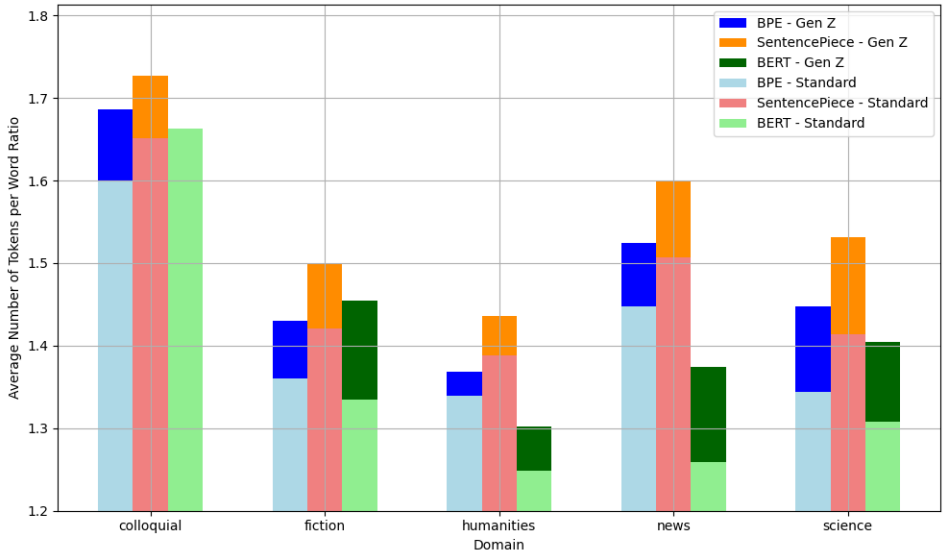


Figure 2: Average value of the tokens per word ratio for each domain. Gen Z text generally results in a higher average tokens-per-word ratio compared to standard text, indicating tokenization inefficiency. There is a single exception for BERT model in the colloquial domain. We note that the y-axis is shifted to start from 1.2. This was done to better visualize results.

A.2 Regression Tables

	Coefficient	**t**	**P>|t|**
Constant	1.3260	19.064	0
$\delta_{wordcount}$	1.3249	201.741	0
Colloquial	-1.4852	-12.466	0
Fiction	0.7455	4.481	0
Humanities	-0.3640	-2.634	0.008
News	-0.2743	-1.479	0.139
Science	2.7039	14.317	0

Table 8: Regression Model Summary for SentencePiece.
R-squared value is high (0.90), suggesting that the model fits the data well and explains a significant portion of the variation in $\delta_{tokencount}$.

	Coefficient	t	P>|t|
Constant	1.6520	27.932	0
$\delta_{wordcount}$	1.2307	220.389	0
Colloquial	-2.5256	-24.932	0
Fiction	1.2623	8.924	0
Humanities	-0.1115	-0.949	0.342
News	1.2955	8.219	0
Science	1.7313	10.781	0

Table 9: Regression Model Summary for BERT.
R-squared value is high (0.91), suggesting that the model fits the data well and explains a significant portion of the variation in $\delta_{tokencount}$.

	Coefficient	t	P>|t|
Constant	0.9915	15.666	0
$\delta_{wordcount}$	1.2665	211.947	0
Colloquial	-1.1199	-10.331	0
Fiction	0.7448	4.920	0
Humanities	-0.5173	-4.115	0
News	-0.5622	-3.333	0
Science	2.4461	14.235	0

Table 10: Regression Model Summary for BPE.
R-squared value is high (0.91), suggesting that the model fits the data well and explains a significant portion of the variation in $\delta_{tokencount}$.

B Additional Figure for Sentiment Analysis

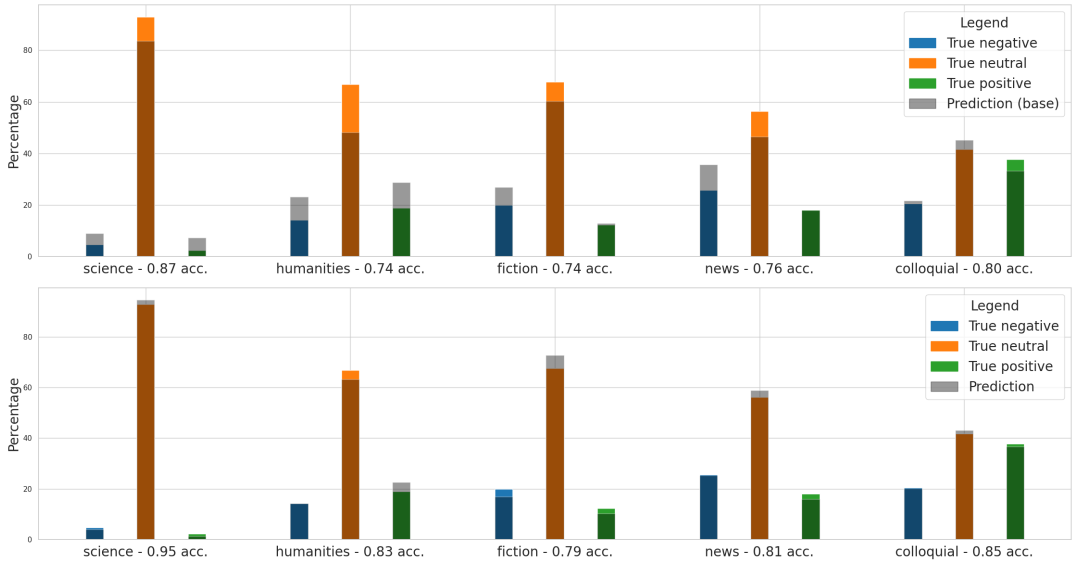


Figure 3: Classification performance per domain of Twitter-roBERTa-base: (top) evaluation of the pretrained base model; (bottom) evaluation of the fine-tuned model. Predictions are made on the entire dataset, i.e. including train & test. The goal is to highlight the change in the amount of *neutral* predictions from the pretrained to the fine-tuned version.