Interpreting the Fine-Tuning Results of Language Models (Transformers)

Tuesta Llaja, Juan Carlos

April 15, 2025

Abstract

Transformer-based language models have transformed NLP, excelling across tasks and languages. Yet, challenges remain in interpreting their internal states and generating coherent paraphrases. This work proposes a method to describe Transformer states interpretably, linking them to model performance. We evaluate BERT- and GPT-based models, both monolingual and multilingual, under low-frequency and syntactically complex inputs. Planned experiments include techniques like AESOP [3], which applies syntactic control, and unsupervised paraphrasing methods [1]. Evaluation includes BLEU, ROUGE, and LLM explainability tools [5]. Preliminary findings highlight the role of syntax-aware control and transparency, laying a foundation for future work on interpretable and robust NLP models.

1 Introduction

Transformer models like BERT and GPT have reshaped NLP, excelling in translation, summarization, and text generation [4]. However, their internal logic remains opaque, raising concerns about trust and interpretability. Paraphrase generation also struggles to preserve meaning and structure, especially with rare or complex inputs.

Studies such as [2] suggest monolingual models better capture syntactic detail than multilingual ones. Methods like AESOP [3] control syntax during generation, while unsupervised approaches [1] leverage implicit syntax without labeled data. Yet, deeper understanding of model behavior under these conditions is still needed.

This work aims to: (1) interpret Transformer internal states and their performance impact, and (2) assess modern paraphrasing methods. We use explainability tools and standard metrics to test various models under challenging conditions. Techniques like AESOP and unsupervised generation help analyze how syntax affects model behavior.

Our findings support the need for syntactic control and transparency in multilingual and sensitive NLP applications.

2 Related Work

2.1 Language Models and Multilingual Adaptation

The Transformer [4] enabled models like BERT and GPT. [2] showed that monolingual models outperform multilingual ones in capturing syntactic detail, highlighting the trade-off between generalization and specialization, especially for low-resource languages.

2.2 Paraphrase Generation

Paraphrasing has moved from supervised to unsupervised methods. AESOP [3] uses syntactic control to ensure structure and meaning. Unsupervised methods [1] exploit pretrained syntactic knowledge to generate fluent variants. Both approaches must balance semantic fidelity and generative flexibility.

2.3 Interpretability and Model Evaluation

Explaining LLMs is increasingly critical. Techniques include attention visualization [4] and saliency maps. [5] reviews methods tailored to complex morphologies. BLEU and ROUGE are common metrics but often miss deeper semantic shifts, prompting interest in model-based alternatives.

2.4 Contribution of This Work

This study proposes a framework linking Transformer internal states with performance metrics. We combine explainability tools with paraphrasing techniques—syntactic control [3] and unsupervised methods [1]—to explore the dynamics between syntax and adaptability. We also assess explainability in multilingual contexts, extending prior work [5] and addressing gaps noted in [2]. The systematic evaluation of multilingual and monolingual model robustness—including BERT-based and GPT-based architectures—under low-frequency and syntactically complex inputs lays a foundation for developing more transparent and resilient language models.

3 Methodology

To evaluate the robustness of multilingual (ruBERT-base, BERT multilingual) and monolingual (RoBERTa-base, DistilBERT) models, as well as GPT-based models, against low-frequency data and syntactic modifications, an experimental framework was designed, encompassing data preprocessing, training configuration, and model evaluation. The following sections detail the main components, including planned experiments for paraphrase generation.

3.1 Datasets and Preprocessing

Two Russian textual genres were processed with:

Frequency-based reduction Six filtering levels with occurrence thresholds $\tau \in \{3, 5, 9, 29, 49, 100\}$:

$$D_\tau = \{x \in D \mid \text{freq}(w_i) \geq \tau \ \forall w_i \in x\} \quad (1)$$

Syntactic modification Parts-of-speech reduction:

Table 1: Modification scheme by genre

Removed category	Genre 1	Genre 2
Adverbs	✓	×
Verbs	✓	×
Verbs + Adverbs	✓	×
Adjectives	✓	✓

3.2 Training Configuration

Parameters varied by frequency threshold (Table 2):

Table 2: Hyperparameters by frequency threshold

Threshold	Length	Batch	Epochs	LR
>100	52	128	3	2e-5
>49	60	128	4	3e-5
>29	51	128	4	3e-5
>9	45	128	5	4e-5
>5	150	32	6	5e-5
>3	150	32	6	5e-5

3.3 Models Evaluated

- ruBERT-base (180M parameters)
- BERT multilingual (179M parameters)
- DistilBERT (66M parameters)
- RoBERTa-base (125M parameters)
- GPT-based model (e.g., GPT-2 or a Russian variant, parameters TBD)

3.4 Future Experiments

To complement current experiments, paraphrase generation techniques will be integrated. Specifically:

- **AESOP**: The method of [3] will be applied to generate Russian paraphrases with adaptive syntactic control, evaluating its ability to preserve structural coherence and semantic similarity across BERT-based and GPT-based models.
- **Unsupervised Paraphrasing under Syntax Knowledge**: The approach of [1] will be implemented, leveraging implicit syntactic knowledge in pre-trained models to generate paraphrases without labeled data, including tests with GPT-based models.

Both techniques will be evaluated using standardized metrics such as BLEU and ROUGE, alongside interpretability metrics (e.g., attention visualization or saliency maps) to analyze how syntactic modifications impact performance. These results will be integrated into future iterations of the study.

4 Results

4.1 Preliminary and Future Results

Current experiments have yielded preliminary results showing trends in the robustness of multilingual and monolingual models against low-frequency data and syntactic modifications. For instance, multilingual models like ruBERT and mBERT maintain high precision across most configurations, while monolingual models like DistilBERT show improvements with specific syntactic modifications (e.g., removal of verbs and adverbs). Future experiments will extend these findings by incorporating AESOP [3], unsupervised methods [1], and GPT-based models to compare their performance and interpretability. Full details of these results will be presented in a future version of this work.

4.2 Frequency-based Modification

Key findings:

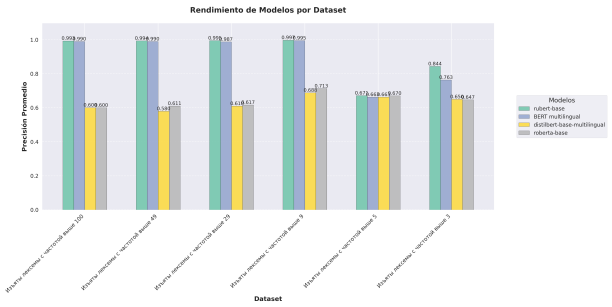

Figure 1: Performance by lexical frequency threshold

- Large models: 99.9% precision for $\tau \in [9, 100]$
- Performance drop: $\sim$25% for $\tau \leq 5$
- Non-Russian models: Peak 65% at $\tau = 9$

4.3 Syntactic Modification

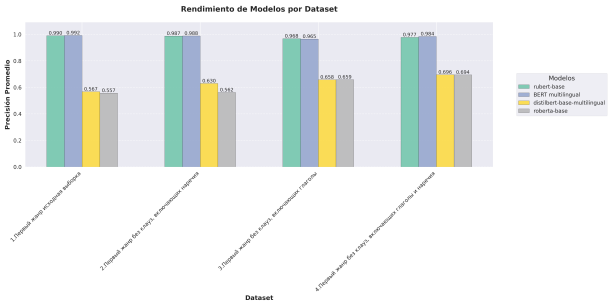

Figure 2: Impact of grammatical category removal

Notable results:

- ruBERT/mBERT: Stable at 99% for all variants
- DistilBERT: +1% with verb+adverb removal
- Adjectives: Minor impact (1% decrease)

5 Discussion

- **Fragmentation**: Thresholds $\leq$5 produce sequences >800 tokens, exceeding BERT's 512-token limit

- **Optimal length**: 45-60 tokens maximizes performance
- **Syntactic noise**: Non-core element removal benefits small models

6 Conclusion

Preliminary results indicate that multilingual models exhibit robustness to syntactic changes, while monolingual models benefit from simplified syntax. Frequency filtering must be tailored to model size to optimize performance. Future experiments with AESOP [3], unsupervised methods [1], and GPT-based models will complement these findings, providing deeper insights into the relationship between syntax, adaptability, and transparency in language models.

References

[1] Tianyuan Liu, Yanan Sun, Jun Wu, Xinyu Xu, Yuwei Han, Chen Li, and Bo Gong. Unsupervised paraphrasing under syntax knowledge. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 13273–13281, 2023.

[2] Debora Nozza, Federico Bianchi, and Dirk Hovy. What the [mask]? making sense of language-specific bert models. *arXiv preprint arXiv:2003.02912*, 2020.

[3] Jiaao Sun, Xuezhe Ma, and Nanyun Peng. Aesop: Paraphrase generation with adaptive syntactic control. In *EMNLP 2021*, pages 5176–5186, 2021.

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

[5] Haiyan Zhao, Hanjie Chen, Fan Yang, Ninghao Liu, Huiqi Deng, Hengyi Cai, Shuaiqiang Wang, Dawei Yin, and Mengnan Du. Explainability for large language models: A survey. *arXiv preprint arXiv:2309.01029*, 2023.