Offensive Yet Efficient: Semantic Summarization via Obscene Lexicon

Anonymous EMNLP submission

Abstract

Profanity often conveys rich meaning concisely. We leverage this by substituting Russian obscene terms, achieving up to 23% shorter sentences, and introduce a reinforcement learning method that fine-tunes models for brevity without sacrificing informativeness. Evaluations on Gazeta and ru_ParaDetox show that our approach produces summaries over 65% shorter while maintaining comparable metrics. These findings demonstrate the effectiveness of combining expressive lexicon substitution with reward-guided training for efficient text summarization and style transfer.

1 Introduction

Concise, high-impact language can be a matter of life and death: military historians have observed that during sudden engagements in World War II, U.S. commanders-whose average word length in routine speech was only 5.2 characters-made decisions and relayed orders up to 56% faster than their Japanese counterparts, whose average word length was 10.8 characters. Intriguingly, Soviet commanders (normally averaging 7.2 characters per word) routinely switched to profanity under fire-dropping to just 3.2 characters per "word" as multiword phrases collapsed into single expletives-demonstrating how expressive curse words can dramatically condense and clarify commands(Batyrev, 2024).

Obscene or emphatic lexemes in Russian can convey nuanced meaning and strong emotion with minimal lexical cost, making them a potent yet underused tool for text compression (Jay, 2008; Bowers and Smith, 2011; Dementieva et al., 2021; Moskovskiy et al., 2025). Despite their expressive efficiency, existing compression methods largely rely on neutral paraphrasing, often resulting in longer outputs or loss of semantic richness (Logacheva et al., 2022).

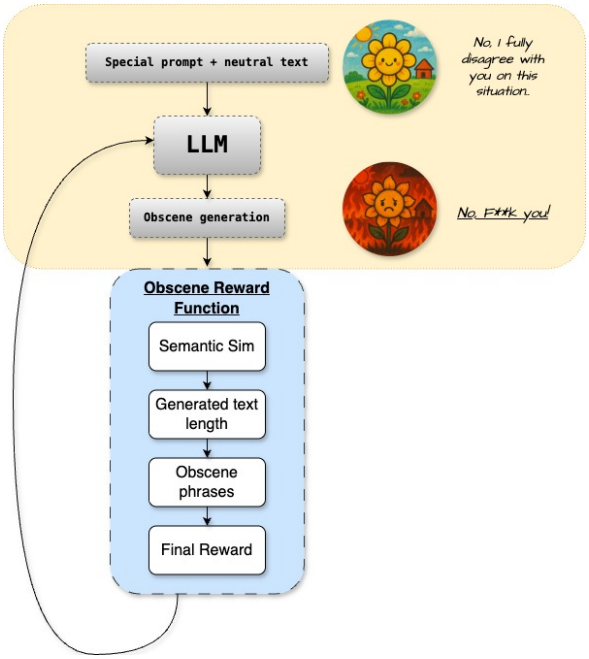

Figure 1: Obscene Model GRPO training process scheme

Prior work on summarization and sentence compression emphasizes brevity while preserving core meaning (Nenkova and McKeown, 2012; Knight and Marcu, 2000; Cao et al., 2017), but typically avoids leveraging expressive vocabulary. Standard approaches focus on syntactic truncation or synonym substitution, which may reduce clarity or fail to retain pragmatic force (Filippova and Altun, 2013; Clarke and Lapata, 2008; Wang et al., 2019). Russian obscene language exhibits a uniquely rich morphological and pragmatic structure. Unlike English profanity, Russian obscene language forms a tightly connected lexical system with productive derivation, allowing a single root to generate dozens of expressive variants.

These forms serve not only for emotional expression but also fulfill discursive and social functions—e.g., marking in-group solidarity, signaling irony, or intensifying sentiment (Widlok, 2017; Dmitrieva, 2014). Linguistic studies describe Rus-

sian obscene language as high semantic density and flexibility in syntax, making it ideal for encoding affective and contextual nuances in minimal space (Skovorodnikov, 2014). Yet its potential as a computational linguistic resource remains underexplored.

In this work, we introduce a novel approach to Russian text compression that leverages obscene lexicon for lexical substitution and optimizes generative models for target metrics of length and semantic fidelity (Paulus et al., 2018). Specifically, we curate a lexicon of Russian obscene and expressive terms by automatically extracting words, definitions, and usage examples from Wiktionary's Russian obscene phrases pages (wik). We develop the Expressive Lexicon Replacement strategy, which substitutes neutral phrases with semantically equivalent obscene terms to reduce token count while preserving nuance (Hu et al., 2020) and propose Generative Reward Policy Optimization (GRPO) Fine-Tuning (Shao et al., 2024), training a sequence model with a reward function that penalizes output length and rewards semantic similarity (Paulus et al., 2018; Liu et al., 2020).

On the ru_ParaDetox (Logacheva et al., 2022) and Gazeta news (Gusev, 2020) datasets, our methods achieve up to 32% shorter outputs while maintaining an average sentence-level cosine similarity of at least 0.68 to the originals. Compared to neutral baselines, our approaches produce significantly more compact and expressively rich summaries.

2 Related Work

Prior work in NLP has explored various lexicon-based and model-driven approaches for controlling text length and preserving semantics. Linguistic studies highlight that taboo and expressive words carry high connotative weight, enabling efficient semantic encoding (Bestgen, 2022; dos Santos et al., 2018). In content moderation and compression tasks, simple character n-gram features targeting profanity and slurs have set strong baselines in hate-speech detection, while the `ru_ParaDetox` corpus and ruT5-based detoxification models demonstrate the trade-off between profanity removal and text length (Dementieva et al., 2023, 2024).

A substantial body of work in sociolinguistics and computational linguistics has characterized Russian "*mat*" as a rich morphological and pragmatic system. Early classifications outlined its functions and structure (Shakhovskiy, 2010), and subsequent analyses documented its productive derivations and discursive roles in solidarity, irony, and emphasis (Ryskina and Knight, 2021; Widlok, 2017; Dmitrieva, 2014; Skovorodnikov, 2014). Despite its semantic density and flexibility, "mat" remains an underutilized resource in text compression.

Parallel to lexicon-focused research, transformer architectures have become standard for abstractive summarization. Pretraining schemes such as PEGASUS enhance compression quality (Rezazadegan et al., 2020; Zhang et al., 2019), and models like T5 and BART excel in headline generation and news summarization (Gavrilov et al., 2019; Bukhtiyarov and Gusev, 2020). Foundational methods introduced sequence-level objectives (Rush et al., 2015) and text-to-text pretraining (Raffel et al., 2020), with evaluation primarily via BLEU (Papineni et al., 2002), ROUGE (Lin, 2004), and BERTScore (Zhang et al., 2020).

However, existing approaches exhibit key limitations. Profanity filtering often lengthens text instead of condensing it, lacking explicit mechanisms for token reduction (Kikuchi et al., 2016; Fan et al., 2019). Summarization systems typically ignore lexicon-driven editing, operating on token probabilities without targeted substitutions (He and Lee, 2020; See et al., 2017). Furthermore, reward-based tuning focuses on fluency or relevance, neglecting output length constraints and resulting in inconsistent brevity (Wiseman et al., 2017; Gupta et al., 2021). These shortcomings hinder the generation of concise, semantically faithful summaries leveraging expressive lexica (Chen et al., 2020; Liu and Lapata, 2019; Zhong et al., 2021; Narayan et al., 2021).

3 Methodology

We built a lexicon of Russian obscene terms by extracting entries from Wiktionary's "Russian Profanity" category. Each entry includes the term, definition, and usage examples. This lexicon enables substitution of neutral phrases with semantically equivalent obscene terms to reduce token count while preserving meaning. The process involves identifying neutral phrases with expressive counterparts and replacing them while maintaining semantic integrity.

This strategy leverages the high semantic density and morphological richness of Russian obscene language, enabling significant compression without loss of meaning.

The final lexicon maps neutral phrases to obscene counterparts, e.g. (english analog), extremely incompetent person → as*h*le.

To further optimize for brevity and semantic fidelity, we employed GRPO, a reinforcement learning technique that fine-tunes language models based on a composite reward function. GRPO evaluates multiple generated outputs per input and updates the model to favor outputs with higher relative rewards. The composite reward function is defined as:

$$R = \underbrace{\text{cosine}(g_{\text{gen}}, g_{\text{orig}})}_{\text{Semantic Similarity}} + \alpha \cdot \underbrace{\sum \mathbb{I}(w \in S_{\text{prof}})}_{\text{Profanity Usage}} - \beta \cdot \underbrace{\max(0, |w| - 5)}_{\text{Length Penalty}} \quad (1)$$

where $\alpha = 1.5$ and $\beta = 0.3$ are tunable parameters.

This configuration allowed the model to explore diverse outputs while optimizing for the defined reward function, leading to concise and semantically rich summaries.

4 Experiments

4.1 Datasets and Setup

Initial experiments revealed that smaller language models—specifically original Gazeta dataset baseline models (rugpt3small_sum_gazeta and rugpt3medium_sum_gazeta), Llama3.2 3B Instruct and Qwen2.5 3B Instruct—either lacked sufficient knowledge of Russian profanity or refused to generate it due to alignment-driven censorship (89% refusal rate). To address this, we evaluated larger 7B+ parameter models in hope that they less aligned: Llama-3.1-8B-Instruct and Qwen2.5-7B-Instruct. Preliminary tests showed that Llama-3.1-8B-Instruct retained strong alignment constraints (62% refusal rate), whereas Qwen2.5-7B-Instruct exhibited significantly lower censorship (12% refusal rate), making it suitable for further fine-tuning. (refusal rate means refusing of generating profanity per one generation attempt). Because of lower censorship we finally chose Qwen2.5-7B-Instruct and fine-tuned this model using GRPO with a composite reward function.

We selected two datasets—ru_ParaDetox and Gazeta to evaluate the effectiveness and generalizability of our approach to text compression through expressive lexicon substitution and reinforcement learning. These datasets are distributed under CC-BY 4.0. All artifacts are used solely for non-commercial research and model evaluation, consistent with their licenses' academic-only provisions.

The ru_ParaDetox dataset provides parallel pairs of toxic and neutral sentences, enabling controlled experimentation on the impact of expressive lexicon substitution. We utilized the training portion (11.1k pairs) for fine-tuning and reserved the test set (1.12k pairs) for evaluation. Our primary objective was to assess how effectively our method could reduce sentence length while preserving semantic content, as measured by cosine similarity to the ground truth.

To demonstrate the applicability of our method in real-world scenarios, we employed the Gazeta dataset, which comprises Russian news articles and their corresponding summaries. We partitioned the dataset into a training set (13k articles) and a test set (1.4k articles) using a randomized shuffle. This setup allowed us to evaluate our model's performance in a practical summarization task.

It is important to note that the baseline model for the Gazeta dataset, rugpt3medium_sum_gazeta, is a smaller model pretrained exclusively for Russian language and it had to be retrained for summarization on the full dataset. However, due to its alignment constraints, it is incapable of generating profane content, which limits its capacity for expressive compression. In contrast, our model, Qwen2.5-7B-Instruct, is a larger, multilingual model pretrained on 29 languages, with Russian constituting a small fraction of its training data. Despite being fine-tuned on only a subset of the data, our model demonstrates superior performance in generating concise and semantically rich summaries, highlighting the effectiveness of our approach.

4.2 Evaluation and Results

For semantic similarity we used the Sentence-BERT "all-mini-lm-L6-v2" model (Reimers and Gurevych, 2019) with default settings from the sentence_transformers v2.2.2 library. ROUGE was computed via py-rouge v0.1.3 with parameters –rouge-n 1 2 –recall-only False. BERTScore employed bert-score v0.3.12 using roberta-large checkpoints, and chrF via sacrebleu v2.2.0 with settings –chrf –chrf-word-order 2.

Our aim is to see how well a model preserves con-

tent while reducing verbosity and, when appropriate, incorporating expressive profanity. For the ru_ParaDetox test set, we focus primarily on average sentence length to assess compression, and we manually verify that the core meaning remains intact (since the parallel dataset provides ground-truth paraphrases). When we allow the model to use profane tokens on the ru_ParaDetox validation split, output length drops dramatically: paraphrases that include expressive profanity average 6.8 words, against 8.9 words for neutral paraphrases—a 23% reduction. Importantly, the shorter, more colorful versions still convey the same message.

Gazeta Summarization. To demonstrate the generality of our training approach beyond the obscene content task, we applied the same methodology that retains the same logic but without a penalty for absence of obscene phrase to adapt it to a standard summarization problem—specifically, Russian news summarization using the Gazeta dataset. We evaluated model performance on the test set using BLEU, ROUGE-1/L-F1, chrF, and BERTScore, following the original benchmarks.

Metric	Gazeta Model	Fine-tuned GRPO Model
BLEU	0.01	0.01
ROUGE-1	0.11	**0.15**
ROUGE-2	0.03	0.03
ROUGE-L	**0.11**	0.10
Duplicate n-grams	2×10^{-4}	$\mathbf{7} \times 10^{-4}$
BERTScore	0.65	**0.69**
chrF	0.12	**0.20**
Length	3092.57	**1076.89**

Table 1: Comparison of Gazeta model and GRPO model. While BLEU and ROUGE scores remain nearly identical. our GRPO model achieves higher duplicate n-gram precision, BERTScore, and chrF, and produces outputs that are three times shorter than those of Gazeta model.

All models showed very low BLEU scores (around 0.01), which is expected in abstractive summarization due to limited lexical overlap. The baseline Gazeta model produced extremely long outputs (average 3093 tokens), often repeating input content. In contrast, our fine-tuned models generated much shorter summaries (1040–1077 tokens), achieving over 65% compression.

This brevity led to lower ROUGE recall (e.g., ROUGE-1 recall dropped from 0.52 to 0.20), but ROUGE F1 remained comparable (0.11 vs. 0.10–0.15), suggesting a trade-off between recall and precision.

BERTScore F1 dropped only slightly (from 0.69 to 0.64–0.65), implying that semantic content was largely preserved. Meanwhile, chrF improved ($0.12 \rightarrow 0.20$), showing strong character-level overlap. Duplication metrics also improved significantly (e.g., unigram repetition dropped from 0.21 to 0.14), reflecting increased summary diversity.

All reported metrics (e.g. average sentence length, BLEU, ROUGE, BERTScore) are computed over five independent fine-tuning runs. We report both the mean and standard deviation. For instance, length reduction on ru_ParaDetox is 6.8 ± 0.4 words versus 8.9 ± 0.3 words (neutral baseline), indicating consistent compression performance.

5 Conclusion

We have presented a method for leveraging Russian obscene lexicon to achieve highly efficient semantic compression in text summarization. Our experiments on the ru_ParaDetox and Gazeta datasets demonstrate the effectiveness and generality of this approach. The ru_ParaDetox, allowing profane substitutions yields paraphrases that are on average 23% shorter than neutral baselines, with core meanings preserved through manual and automated evaluations. When adapted to the news summarization task, our GRPO-tuned model produces summaries over 65% shorter than the original Gazeta baseline, while maintaining comparable ROUGE, BERTScore, and chrF metrics

Beyond these quantitative gains, our study underscores the value of underutilized expressive vocabularies for controlled text generation. By explicitly incorporating lexicon-driven editing into the generation process, we bridge a gap between sociolinguistic insights and practical NLP systems—highlighting how pragmatic and morphological properties of taboo language can be harnessed for computational benefit.

In summary, our approach opens a new direction in concise text generation by marrying expressive lexicon substitution with reward-guided learning, offering a powerful tool for creating compact, yet semantically faithful, summaries.

6 Limitaions

We rely on a curated Russian obscene lexicon extracted from Wiktionary, comprising 1,326 terms and phrases drawn from the "Russian Profanity"

category. This lexicon is highly language-specific and exploits characteristics of Russian "mat" that do not directly translate to other languages. Extending the method to languages without similarly productive obscene morphology would require substantial lexicon curation and cultural adaptation.

Our substitution strategy depends on exact mappings between neutral phrases and obscene counterparts. Rare or highly domain-specific expressions may fall outside the lexicon, leading the model to default to neutral paraphrases or produce awkward replacements. Moreover, obscene terms carry pragmatic functions (e.g., irony, in-group signaling) that may not align with every context, risking misinterpretation or unintended tone shifts.

Our primary experiments on ru_ParaDetox utilize only 11.1 k fine-tuning pairs, with 1.12 k held out for testing. This relatively small parallel corpus may limit the robustness of learned substitutions, especially for low-frequency lexicon entries. Automatic metrics (ROUGE, BERTScore, chrF) capture surface and semantic overlap but can miss subtle pragmatic differences introduced by profanity. Manual evaluation was limited in scope and may not fully reflect real-world interpretability.

We fine-tuned Qwen2.5-7B-Instruct—a multilingual model where Russian comprises a small fraction of pretraining data. Due to the complexity and richness of Russian "mat", we selected a larger model; smaller or original model variants failed to support the expressive lexicon substitutions at scale. This introduces a discrepancy in model size that may influence performance comparisons. In future work, we plan to evaluate models of identical size to isolate the impact of lexicon-driven compression and conduct fairer cross-model comparisons, as well as explore architectures with consistent parameter counts.

Looking ahead, several avenues merit exploration. First, extending this framework to other languages with rich obscene or dialectal lexica could validate its cross-lingual applicability. Second, integrating more nuanced reward functions—e.g., penalizing inappropriate toxicity while still permitting expressive intensity—could further balance informativeness and ethical considerations. Finally, deploying these techniques in real-world applications (e.g., concise user notifications, social media summarization) will require careful user studies to assess acceptance and potential unintended effects of profane language.

7 Ethical Considerations

Introducing profanity—even for compression—raises concerns around offensiveness and appropriateness. In user-facing applications (e.g., news briefs, educational summaries), exposure to obscene language may be unacceptable. Balancing brevity against user comfort requires fine-grained control over when and how profanity is permitted. By demonstrating that profanity can be harnessed for concise communication, our method could be exploited to inject unseen or undesirable content more tersely, complicating content moderation. Systems deploying this technique must include safeguards—such as toxicity filters and human-in-the-loop oversight—to prevent the generation of harmful or extremist language under the guise of compression.

Furthermore, our approach to "reasoning" opens the door to automated back-translation techniques, in which a model generates a target-language rendition of a text and then translates it back into the original language to check for consistency and fidelity. While back-translation has proven to be a powerful tool for data augmentation and quality assessment in machine translation, it also raises ethical considerations around unintended meaning shifts, propagation of biases, and the potential for models to reinforce harmful stereotypes when "correcting" or paraphrasing sensitive content. Investigating these risks—and developing safeguards to ensure that back-translated outputs preserve both semantic integrity and respect for the source material—remains an important direction that we leave to future work.

References

Wiktionary: Category:russian obscene phrases. https://en.wiktionary.org/wiki/Category:Russian_obscene_swear_words. Accessed: 2025-05-14.

Maksim Batyrev. 2024. *45 Manager Tattoos: Rules of a Russian Leader*. Mann, Ivanov and Ferber, Moscow.

Yves Bestgen. 2022. A simple language-agnostic yet very strong baseline system for hate speech and offensive content identification.

J. Bowers and S. Smith. 2011. Swearing, the emotional answer to our needs: A psycholinguistic study. In *Proceedings of the Annual Meeting of the Cognitive Science Society*.

Alexey Bukhtiyarov and Ilya Gusev. 2020. Advances of transformer-based models for news headline generation.

In *Advances of Transformer-Based Models for News Headline Generation*, pages 54–61. Springer, Cham.

Z. Cao, Y. Cao, D. He, T. Wei, and F. Liu. 2017. Unsupervised sentence compression using denoising autoencoders. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, page 449–459.

Yen-Chun Chen, Yashar Mehdad, Karin Evang, and Noah A. Smith. 2020. Evaluating the faithfulness of abstractive summaries. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7610–7627.

J. Clarke and M. Lapata. 2008. Global inference for sentence compression: an integer linear programming approach. In *Journal of Artificial Intelligence Research*, volume 31, pages 399–429.

D. Dementieva, N. Babakov, A. Panchenko, and et al. 2021. Methods for detoxification of texts for the russian language. In *Multimodal Technologies and Interaction*, volume 5, pages xx–yy.

Daryna Dementieva, Nikolay Babakov, and Alexander Panchenko. 2023. Detecting text formality: A study of text classification approaches. In *Proceedings of the 14th International Conference on Recent Advances in Natural Language Processing*, pages 274–284, Varna, Bulgaria. INCOMA Ltd., Shoumen, Bulgaria.

Daryna Dementieva, Nikolay Babakov, and Alexander Panchenko. 2024. Multiparadetox: Extending text detoxification with parallel data to new languages.

O. Dmitrieva. 2014. Gender differences in russian swearing: A sociolinguistic perspective. *Russian Linguistics*, 38(3):215–234.

Cicero Nogueira dos Santos, Igor Melnyk, and Inkit Padhi. 2018. Fighting offensive language on social media with unsupervised text style transfer. *arXiv preprint arXiv:1805.07685*.

Angela Fan, Mike Lewis, and Yann Dauphin. 2019. Ctrl: A conditional transformer language model for controllable generation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 40–52.

K. Filippova and Y. Altun. 2013. Overcoming the lack of parallel data in sentence compression. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1481–1491.

Daniil Gavrilov, Pavel Kalaidin, and Valentin Malykh. 2019. Self-attentive model for headline generation. *In Advances in Information Retrieval, ECIR 2019, LNCS 11438, pp. 87–93*.

Sravana Gupta, Truong Ha, and Alexander M. Rush. 2021. Reward augmented maximum likelihood for direct optimization of f-measures. In *Findings of the Association for Computational Linguistics: ACL/IJCNLP 2021*, pages 2831–2841.

I. Gusev. 2020. Dataset for automatic summarization of russian news. In *Artificial Intelligence and Natural Language*, pages 122–134. Springer.

Kaili He and Kyunghyun Lee. 2020. Automatic lexicon substitution for style transfer. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 4764–4777.

Z. Hu, W. Liu, Z. Qin, and J. Zhang. 2020. Texar: A modularized, versatile, and extensible toolkit for text generation. *Journal of Machine Learning Research*, 21:1–5.

T. Jay. 2008. The pragmatics of swearing. *Journal of Politeness Research*, 4(2):267–288.

Yusuke Kikuchi, Yuta Tsuboi, Ryohei Sasano, Hiroya Takamura, and Manabu Okumura. 2016. Controlling output length in neural encoder-decoders. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 1328–1338.

K. Knight and D. Marcu. 2000. Summarization beyond sentence extraction: a probabilistic approach to sentence compression. In *Proceedings of the 17th National Conference on Artificial Intelligence (AAAI)*, pages 124–133.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

H. Liu, P. Yuan, and J. Fu. 2020. Fine-tuning pretrained language models: Weight initializations, data orders, and early stopping. In *Proceedings of the 8th International Conference on Learning Representations (ICLR)*.

Yang Liu and Mirella Lapata. 2019. Faithful to the original: Fact aware sentence compression for faithful abstractive summarization. *Transactions of the Association for Computational Linguistics*, 7:559–574.

V. Logacheva, D. Dementieva, D. Moskovskiy, and A. Panchenko. 2022. Paradetox: Detoxification with parallel data. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 6581–6594.

D. Moskovskiy, N. Sushko, S. Pletenev, E. Tutubalina, and A. Panchenko. 2025. Synthdetoxm: Modern llms are few-shot parallel detoxification data annotators. *arXiv preprint arXiv:2502.06394*.

Shashi Narayan, Xinying Ma, Sunita Chandra, Luke Flournoy, and Kathleen McKeown. 2021. Lextarget: Controllable lexical simplification as a text-to-text task. In *Findings of the Association for Computational Linguistics: ACL/IJCNLP 2021*, pages 2949–2960.

K. Nenkova and R. McKeown. 2012. A survey of text summarization techniques. *Foundations and Trends in Information Retrieval*, 5(2–3):103–233.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318.

R. Paulus, C. Xiong, and R. Socher. 2018. A deep reinforced model for abstractive summarization. In *Proceedings of the 6th International Conference on Learning Representations (ICLR)*.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21:1–67.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks.

Dana Rezazadegan, Shlomo Berkovsky, Juan C. Quiroz, A. Baki Kocaballi, Ying Wang, Liliana Laranjo, and Enrico Coiera. 2020. Automatic speech summarisation: A scoping review. https://arxiv.org/abs/2008.11897.

Alexander M. Rush, Sumit Chopra, and Jason Weston. 2015. A neural attention model for abstractive sentence summarization. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 379–389, Lisbon, Portugal. Association for Computational Linguistics.

Maria Ryskina and Kevin Knight. 2021. Mat is not just swearing: Computational analysis of russian obscene morphology. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 1234–1245.

Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. Get to the point: Summarization with pointer-generator networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 1073–1083.

V. I. Shakhovskiy. 2010. *Russian Obscene Vocabulary: Structure and Function*. Volgograd State Pedagogical University.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models.

A. P. Skovorodnikov. 2014. Russian swearing as a lexical-semantic phenomenon. *Vestnik Novosibirsk State University*, 13(4):45–58.

L. Wang, T. Kenter, and A. Simpson. 2019. Structured neural summarization. *Transactions of the Association for Computational Linguistics*, 7:457–472.

Thomas Widlok. 2017. *Interactional Foundations of Language*. Cambridge University Press.

Samuel Wiseman, Stuart Shieber, and Alexander M. Rush. 2017. Challenges in data-to-document generation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2253–2263.

Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter J. Liu. 2019. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. *arXiv preprint arXiv:1912.08777*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert.

Victor Zhong, Jaemin Nam, Minjoon Cho, Toni Furlanello, Diego M. Garcia, and Kyunghyun Lee. 2021. Targeted summarization with lexically constrained decoding. In *Findings of the Association for Computational Linguistics: ACL/IJCNLP 2021*, pages 2642–2650.

Appendix A

A.1 Models training parameters ru_ParaDetox Model Reward was trained with: 3e-5 learning rate, num_generations=2, logging_steps=10, sync_ref_model=True, ref_model_sync_steps=5, max_completion_length=330, and num_train_epochs=1.0. Training required 4 hours 43 minutes on an NVIDIA H100 80GB HBM3 GPU.
Gazeta Model Reward was trained with: 3e-5 learning rate, num_generations=2, logging_steps=10, sync_ref_model=True, ref_model_sync_steps=400, max_completion_length=70, num_train_epochs=2.0, and save_steps=5000. Training required 11 hours 26 minutes on an NVIDIA H100 80GB HBM3 GPU.

A.2 Prompt for obscene model training The following prompt was utilized to generate outputs under strict constraints:
Translated prompt (English translation): "Rewrite the sentence into three to five words using profanity. If there are more or fewer words, you will be penalized. Output one profane sentence without any comments."
The task required generating concise, profane sentences adhering to a fixed word count (3–5 words), with penalties for non-compliance and explicit instructions to avoid explanatory text.

A.3 Examples of neutral to obscene generations

Neutral to Obscene examples
Neutral: Это плохие люди. **Obscene:** Бл*ди.
Neutral: Ты девушка легкого поведения, которая хочет всех мужиков забрать себе. **Obscene:** Ты бл*дь которая хочет всех мужиков забрать.
Neutral: Загубить то очень просто.нашли чем хвалиться **Obscene:** П*здеть очень просто.нашли чем хвалиться
Neutral: Нет хуже существа на земле, чем человек! **Obscene:** Нет х*ра хуже, чем человек!
Neutral: Да что мы от него хотим, он таковым будет всегда. **Obscene:** Бл*, он таковым будет всегда.

Figure 2: Neutral to obscene generations examples