Team The Toxinators 2000 at TextDetox CLEF 2025/Multilingual Text Detoxification 2025: The Evolution of Methods for Text Detoxification: The Role of Language in Method Selection

Notebook for the PAN at CLEF 2025

Andrei Totok[1,*], Artemiy Ermolaev[1], Anastasia Izyumova[1,2] and Evgeniy Finogeev[1,3]

[1]*Ecom.tech, Moscow, Russia*
[2]*Institute of Precision Mechanics and Optics (ITMO), Saint Petersburg, Russia*
[3]*National Research Nuclear University (MEPhI), Moscow, Russia*

Abstract

This paper explores various approaches to the task of text detoxification, including the use of lexical resources (such as toxic word dictionaries), methods based on deep learning algorithms (T5, VAE, etc.) and modern large language models (LLMs). The study is conducted on data from PAN: Multilingual Text Detoxification (TextDetox) 2025 — with the aim of identifying the most effective strategies for handling toxic language depending on linguistic specifics. It is shown that for low-resource languages, like Tatar, dictionary-based methods show the highest effectiveness. For widely spoken languages, such as English, deep learning methods show better quality. The results highlight the importance of considering linguistic features when selecting a detoxification method and open up possibilities for the further development of adaptive multilingual content filtering systems.

Keywords

detoxification, pretrained models, dictionaries, PAN 2025

1. Introduction

The rapid growth of online communication platforms has led to an increasing amount of user-generated content, which often includes toxic, aggressive, or otherwise harmful language. This phenomenon poses a serious challenge for maintaining safe and inclusive digital environments. One of the key solutions to this issue lies in the development of automated systems capable of detecting and neutralizing toxic language — a task commonly referred to as text detoxification.

Text detoxification involves transforming a given text in such a way that its toxic or offensive elements are removed or softened, while preserving the original meaning and stylistic coherence. The field of NLP has seen rapid development in recent years, enabling diverse approaches to address the problem of text detoxification.

Each of these approaches comes with its own strengths and limitations. Lexical methods, such as filtering based on lists of toxic words, offer simplicity and speed but often fail to account for context, sarcasm, or subtle forms of toxicity. Deep learning models provide better contextual understanding when trained on high-quality annotated datasets[1], yet they may struggle with generalization across domains and languages[2]. LLMs, particularly those fine-tuned for generation tasks, show good results in rewriting toxic sentences into non-toxic equivalents while maintaining fluency and intent. However, performance can vary significantly depending on the linguistic structure and data availability for each specific language.

This paper presents a comparative analysis of different detoxification strategies across multiple languages, emphasizing how linguistic characteristics influence the effectiveness of each method. We

CLEF 2025 Working Notes, 9 – 12 September 2025, Madrid, Spain
*Corresponding author.
atotok@ecom.tech (A. Totok); artermolaev@ecom.tech (A. Ermolaev); aizyumova@ecom.tech (A. Izyumova); efinogeev@ecom.tech (E. Finogeev)

argue that no single approach is universally optimal and that the choice of method should be guided by the specific properties of the target language. This study utilizes data from PAN: Multilingual Text Detoxification (TextDetox) 2025[3, 4] to explore the most effective strategies for managing toxic language, taking into account linguistic and cultural differences. We use the Joint metric from PAN, which consists of three components: Style Transfer Accuracy, Content Preservation, and Fluency. Each component ranges from 0 to 1, and higher values of both individual components and the final score indicate better performance.

2. The formation of toxic words in different languages

In modern linguistics, toxic words are commonly referred to as invective vocabulary. The invective lexicon of a language is enriched through several sources: reinterpretation of existing literary units, borrowings from foreign languages, and slang. Importantly, invective vocabulary differs from slang in terms of stability: while slang undergoes frequent changes and updates, the core set of invectives tends to remain relatively stable over long periods.

As an example, consider American English, where a significant source of lexical borrowings is African-American vernacular, including such expressions as "motherf*cker" or "of*y". However, invectives should be distinguished from colloquial speech, which may appear similar but lacks the same emotionally charged connotation. Colloquial words such as "sh*t" or "a*s", although considered vulgar in formal contexts, can function as neutral terms in informal speech — especially within certain social groups[5].

Thus, the meaning and communicative function of the same word depend on the speaker's social status and context, leading to an overlap between invective and colloquial vocabulary. This process facilitates lexical exchange between these two linguistic layers.

The formation of invective vocabulary can be classified into several key directions. First, many invectives originate from colloquial, slang, or marginal vocabulary often used to express social differences and conflicts. Second, a major contribution to the development of invectives comes from pejoration — the semantic re-evaluation of literary and conversational words, during which originally neutral or positive meanings transform into negative ones.

Linguist I.I. Kremikh identifies several types of pejoration, among which metaphorical pejoration is considered the most productive. It is based on the transfer of negative meaning to an object through similarity or association. Examples include the Spanish words "globo" ("balloon"), used metaphorically to refer to a fat person, and "buzón de correos" ("mailbox"), used to describe someone with a large mouth.

Of particular importance in invective vocabulary are zoomorphisms — metaphors in which people are compared to animals based on stereotypical associations. For instance, the Spanish words "cerdo" and "cochino" (both meaning "pig") or the English "pig" and "dirty dog" are used to describe messy or rude individuals. Similarly, repurposed names of fruits and vegetables, such as "melón" ("melon") and "carrot" (used derogatorily for red-haired people), illustrate how stable phonetic forms can undergo substantial semantic transformation and shift in communicative function.

An interesting phenomenon is the positive use of invectives. In some cases, these expressions are used to convey admiration or praise, reflecting their deep integration into everyday speech and the wide range of emotions they can convey. For instance, in the Tatar language, the word "эт" ("dog"), which functions as an invective, appears in the phrase "Ах, эт, эшне ничек оста хәл иткән!" ("Wow, what a masterful way to solve the problem!"), where it serves as a compliment to one's skills. Such expressive usage of invectives is also widely employed in literature to reflect natural, vivid speech[6].

Ethnolinguistic factors play a significant role in the formation and functioning of invective vocabulary, especially evident when comparing related or contact languages. The Tatar language, with its rich historical and cultural layer, contains numerous invectives whose semantics are closely tied to collective memory and religious identity. A comparative analysis of Tatar invectives with corresponding units in Bashkir, Chuvash, Udmurt, Mari, and Turkish — a fellow Turkic language — reveals consistent patterns in the development of offensive vocabulary and its national-cultural characteristics.

For example, in Tatar, the lexemes "тарe" ("christian cross", "damned") and "чукынган" ("christened", and in an invective sense — "damned") emerged under the influence of the Christianization of Volga Tatars in the 16th–18th centuries. Originally neutral or positively connotated, these words transformed into strong invectives with clearly negative semantics within the Muslim community. The invective "чукынган" demonstrates high derivational productivity, generating forms such as "чукынчык" ("scoundrel"), "чукыныпкит" ("disappear!", "go to hell!"), "чукынды" ("all is lost"), and "чукынгыры" ("may he perish"), which can serve as insults, curses, or exclamatory interjections.

It is important to note that the strength of the invective tone of these words depends on the religious and cultural identity of the addressee: for Muslims, they carry a clearly negative nuance, while for Christians, they retain a neutral or even positive meaning. Similar semantic shifts are observed in Russian as well (e.g., "нехристь" ("heathen") and "креста на тебе нет" ("no cross, no crown"), originally used to refer to people of different faiths).

Thus, invective vocabulary serves not only as a means of expressing insult but also as a marker of deep cultural-historical layers and boundaries of linguistic consciousness, reflecting complex processes of identity interaction within society.

3. Lexical methods

One of the simplest and fastest ways to reduce the level of toxicity in text is the dictionary-based approach, which involves replacing potentially toxic words with neutral equivalents or completely removing offensive terms. The replacement can be either full (replacing the entire word) or partial (applied when a toxic word appears as part of a longer lexical unit).

We use a Multilingual Toxic Lexicon [7] which consists of toxic words and phrases for each language from the current task. Three types of processing were considered in this section:

- copy base — the text remains unchanged;
- replace — toxic words or their parts are removed or replaced;
- delete — toxic words are removed without replacement.

Copy base method consisted in the fact that the toxic was not changed in any way. In other words, we looked at the metrics of unmodified sentences. Replace method involved removing the toxic component from a word. For example, if the dictionary includes the word "f*ck", but "f*cker" not in the toxic dictionary, only the suffix "er" would remain after processing.

Delete method consisted of removing an entire word if it contained any toxic substring. For instance, if the toxic dictionary included the word "f*ck", but "f*cker" was not included, then full word "f*cker" would be deleted.

Results of lexical methods presented in Table 1. The values in the table are given for the Joint metric.

However, in some languages, such as Ukrainian and Japanese, the replace method demonstrated significantly better metrics. This effect may be attributed to specific contextual features that inadvertently amplify the negative tone of the text during word replacement. Delete method showed significantly better results for the Arabic language. Therefore, when applying automated detoxification methods, it is essential to take into account the linguistic structure and contextual usage of words.

The highest level of toxicity was observed when using the copy base method, where the input text remained unchanged. Although this approach preserves the original meaning of the message, it does not address the task of reducing toxicity and should only be used as a baseline for comparison.

The purely dictionary-based approach has several limitations, including an inability to account for context, the need for regular dictionary updates, and the risk of false positives. Hence, a promising direction for future work appears to be the combination of lexical filtering with text rewriting models, which can not only eliminate toxic elements but also restore textual structure and preserve semantic meaning.

Table 1
Results for delete, replace and copy base approaches. Values in the table are given for the Joint metric.

language	copy-base	delete	replace
en	0.353	0.473	0.430
es	0.566	0.603	0.613
de	0.572	0.586	0.577
zh	0.477	0.516	0.501
ar	0.564	0.611	0.565
hi	0.417	0.480	0.461
uk	0.442	0.581	0.694
ru	0.424	0.514	0.533
am	0.461	0.461	0.437
it	0.653	0.668	0.676
ja	0.440	0.441	0.636
he	0.425	0.436	0.418
fr	0.447	0.518	0.521
tt	0.510	0.573	0.572
am	0.419	0.425	0.449

4. Deep learning approaches

A more recent solution in the field of text detoxification involves approaches based on deep learning models, such as T5 (Text-To-Text Transfer Transformer) [8] and its variations. These models are capable not only of removing or replacing toxic words but also of rewriting the original text while preserving its semantic structure and stylistic features. Thanks to their training on parallel corpora — where each toxic sentence has a corresponding non-toxic version — these models can generate more natural and contextually appropriate outputs.

Such models demonstrate strong performance in neutralizing offensive language without significantly altering the meaning of the original message. Their effectiveness is especially evident when they are fine-tuned on high-quality detoxification datasets. A more detailed discussion of these experiments will follow in the section on large language models (LLMs).

For the current set of experiments, we used the HRQ-VAE [9] model, which has shown promising results in text paraphrasing tasks. The model is based on the Variational Autoencoder (VAE) [10] architecture and is designed for soft style transfer, including reducing the level of toxicity in text.

The experiments were conducted on the English language using both the pretrained version of the model and several versions fine-tuned on the Paradetox [11, 12] parallel dataset. We applied the models trained on three different datasets: MSCOCO [13], Paralex [14], and QQP [15], evaluating its ability to reduce toxicity in input sentences. We also explored how different training strategies affect performance whether toxic phrases are included or excluded from paraphrase clusters. A paraphrase cluster is a group containing sentences that are similar in meaning. For example, the sentences "How do I get to the nearest metro station?", "Can you tell me how to get to the subway?", and "Where is the closest metro station?" have similar meanings, so they are placed in the same cluster.

Results of different HRQ-VAE models for English presented in Table 2.The values in the table are given for the Joint metric.

Our findings show that the model's performance depends heavily on the quality of the training data and the strategy used to form paraphrase clusters. In some configurations, the model was able to significantly reduce toxicity levels, while in others, it failed to fully eliminate harmful content or even preserved negative connotations.

In summary, the HRQ-VAE model demonstrates potential for use in text detoxification, particularly when fine-tuned on carefully curated parallel datasets. However, like other methods, it requires careful tuning and contextual awareness to avoid unintended preservation or amplification of toxic tone.

Table 2
Results of different HRQ-VAE models for English. “args for no question” and “args for question” refer to training parameters used for datasets without and with questions, respectively.

Model	Result
HRQ-VAE, train on MSCOCO	0.074
HRQ-VAE, train on Paralex	0.138
HRQ-VAE, train on QQP	0.208
HRQ-VAE, clusters with toxic, args for no question	0.606
HRQ-VAE, clusters with toxic, args for question	0.592
HRQ-VAE, clusters without toxic, args for no question	0.492
HRQ-VAE, clusters without toxic, args for question	0.478

5. Large language models

A modern stage in the development of text models is represented by large language models (LLMs) such as ChatGPT [16], Llama3 [17] and others. These models possess significant capacity for understanding and generating natural language, making them promising candidates for solving complex tasks, including text detoxification. Unlike traditional approaches based on rigid rules or simple neural architectures, LLMs are capable not only of removing or replacing potentially toxic words, but also of rewriting phrases while preserving the original meaning, style, and logical coherence.

One of the key advantages of LLMs is their ability to interpret user instructions through prompts, which enables flexible control over the generation process. For example, prompts such as "Make this more neutral" or "Paraphrase this sentence" allow the model to understand that the tone of the text needs to be adjusted without sacrificing its informational content. This makes LLMs particularly suitable for use in real-world content filtering systems, where both the reduction of toxicity and the maintenance of communication quality are essential.

In the course of this study, several versions of the Flan-T5 [18] model (small, large) were tested. These models were initially trained on the Paradetox dataset and further fine-tuned on a mixed dataset that included Paradetox, Parallel Detoxification Dataset small [19], Paranmt-for-detox [20], and Filtered_paranmt [21]. The training was conducted using different prompts, such as "Make more neutral" and "Detoxify", which allowed us to assess the impact of prompt formulation on the quality of text rewriting. mT0 [22] model from the TextDetox 2024 Multilingual Text Detoxification task.

Results of different prompts presented in Table 3. The values in the table are given for the Joint metric.

Table 3
Results of different prompts in English (Joint metric). For the mT0 model, the prompt was translated to each language.

Model	Result	Prompt
flan-t5-small	0.680	“Detoxify: ”
flan-t5-small	0.686	“Make more neutral: ”
flan-t5-large	0.705	“Make more neutral: ”
mt0	0.727	“Detoxify:”
mt0	0.723	“Remove or edit toxic words:”
mt0	0.729	“Remove toxic words: ”
mt0, replace toxic	0.676	“Detoxify:”

The results demonstrated that the Flan-T5-large variant outperformed smaller versions in terms of detoxification performance. Specifically, after 4696 training steps, the joint metric reached 0.70497 when using the "Make more neutral" prompt, compared to 0.68628 for Flan-T5-small. This indicates that increasing the number of model parameters has a positive effect on its ability to reduce textual toxicity.

It is also worth noting that T5-based models can be adapted to specific languages and domains. For instance, the spivavtor [23] model, designed for the Ukrainian language, showed improved results when prompted with "Перефразуйте" ("Paraphrase"). It achieved a joint metric of 0.33832, significantly lower than that of dictionary-based approaches. This supports the conclusion that T5-based models have strong potential for application in multilingual environments.

Results for spivavtor models on Ukrainian language presented in Table 4. The values in the table are given for the Joint metric.

Table 4
Results for spivavtor models on Ukrainian language. The prompt “Перефразуйте” requests rephrasing, while “Спростіть речення” instructs sentence simplification.

Model	Result	Prompt
spivavtor-large	0.291	“Перефразуйте” (Paraphrase)
spivavtor-large	0.301	“Спростіть речення” (Simplify sentence)
spivavtor-xxl	0.338	“Перефразуйте” (Paraphrase)

Additionally, we evaluated the Llama3[24] model with varying temperature settings, different prompts, and multiple LoRA adapters [25], which allowed us to explore the impact of both generation parameters and architectural modifications on detoxification quality. In our experiments, pre-trained versions of Llama3 were enhanced with adapters trained on parallel datasets of toxic and non-toxic text.

Prompts were carefully selected to explicitly instruct the model to reduce toxicity while preserving the original meaning. Generation was performed using various temperature values (ranging from 0.01 to 0.95), allowing us to evaluate the balance between deterministic and diverse outputs.

The influence of temperature on detoxification quality proved to be inversely proportional. Lower temperature values yielded better results, as the model generated outputs that closely adhered to the original sentence structure and meaning. In contrast, higher temperature values increased output variability — although the overall meaning was preserved, the rewritten sentences often used completely different wording, sometimes reintroducing potentially harmful expressions.

We also tested several LoRA adapters that differed in size and training data. Interestingly, an adapter trained on only 50 examples performed slightly better than another trained on 200 examples. However, the difference between these configurations was minimal, and both still produced outputs that were relatively close to the base Llama3. Results of llama3-70B model presented in Table 5. The values in the table are given for the Joint metric.

Table 5
Results of llama3-70B model on English, evaluated with different temperature values and LoRA adapters (Joint metric).

Model	Result
llama3-70b, temp=0.95	0.601
llama3-70b, temp=0.05	0.626
llama3-70b, temp=0.01	0.628
llama3-70b, temp=0.01, LoRA 50 examples	0.627
llama3-70b, temp=0.01, LoRA 200 examples	0.626

6. Leaderboard overview

Our team, The Toxinators 2000, participated in the Multilingual Text Detoxification task at PAN 2025. The developed system combined dictionary-based filtering with generative rewriting using a large language model (LLMs) mT0. According to the evaluation on the platform, we ranked 5th out of 32 teams, achieving an average score of 0.675 across all languages. Additionally, on the LLM-as-Judge

leaderboard, we ranked 13th out of 32 teams, achieving an average score of 0.648 across all languages. The best results were obtained for the following languages on leaderboard:

Table 6
Results on leaderboard.

Language	Result	Result on LLM-as-Judge
English	0.727	0.842
Spanish	0.696	0.758
Deutsch	0.757	0.828
Chinese	0.543	0.715
Arabic	0.715	0.788
Hindi	0.627	0.759
Ukrainian	0.770	0.766
Russian	0.754	0.818
Amharic	0.491	0.638
Italian	0.746	0.742
Japanese	0.644	0.745
Hebrew	0.415	0.495
French	0.760	0.790
Tatar	0.617	0.611
Hindi	0.449	0.503

The full leaderboard is available on the competition page.

7. Conclusion

In the course of this study, various approaches to the task of text detoxification were examined, including the use of lexical resources (dictionaries), deep learning models, and modern large language models (LLMs). The main objective was to compare the effectiveness of these methods across several languages — English, Russian, Ukrainian, and others — and to identify optimal strategies for processing toxic content depending on linguistic characteristics and data availability.

The mT0 model , when combined with preliminary removal of toxic words and word parts, demonstrated better performance than the baseline system. For example, in Tatar, the score increased from 0.580 to 0.617 and in Japanese — from 0.582 to 0.644. At the same time, the dictionary-based method achieved the best results for Hindi, increasing the score from 0.351 to 0.449.

These findings confirm that the effectiveness of text detoxification methods is directly influenced by linguistic specificity. No single approach proves universally optimal; instead, the choice of method should be guided by the morphological complexity, cultural context, and data availability of the target language.

Thus, future research should focus on developing language-adaptive detoxification systems, which combine the strengths of dictionary filtering, deep learning, and prompt-based LLM rewriting to ensure both high-quality output and meaningful reduction of toxic content.

Declaration on Generative AI

The authors have not employed any Generative AI tools.

References

[1] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, I. Polosukhin, Attention is all you need, 2023. URL: https://arxiv.org/abs/1706.03762. arXiv:1706.03762.

[2] N. Arivazhagan, A. Bapna, O. Firat, D. Lepikhin, M. Johnson, M. Krikun, M. X. Chen, Y. Cao, G. Foster, C. Cherry, W. Macherey, Z. Chen, Y. Wu, Massively multilingual neural machine translation in the wild: Findings and challenges, 2019. URL: https://arxiv.org/abs/1907.05019. arXiv:1907.05019.

[3] J. Bevendorff, D. Dementieva, M. Fröbe, B. Gipp, A. Greiner-Petter, J. Karlgren, M. Mayerl, P. Nakov, A. Panchenko, M. Potthast, A. Shelmanov, E. Stamatatos, B. Stein, Y. Wang, M. Wiegmann, E. Zangerle, Overview of PAN 2025: Voight-Kampff Generative AI Detection, Multilingual Text Detoxification, Multi-Author Writing Style Analysis, and Generative Plagiarism Detection, in: J. C. de Albornoz, J. Gonzalo, L. Plaza, A. G. S. de Herrera, J. Mothe, F. Piroi, P. Rosso, D. Spina, G. Faggioli, N. Ferro (Eds.), Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Sixteenth International Conference of the CLEF Association (CLEF 2025), Lecture Notes in Computer Science, Springer, Berlin Heidelberg New York, 2025.

[4] D. Dementieva, V. Protasov, N. Babakov, N. Rizwan, I. Alimova, C. Brune, V. Konovalov, A. Muti, C. Liebeskind, M. Litvak, D. Nozza, S. Shah Khan, S. Takeshita, N. Vanetik, A. A. Ayele, F. Schneider, X. Wang, S. M. Yimam, A. Elnagar, A. Mukherjee, A. Panchenko, Overview of the multilingual text detoxification task at pan 2025, in: G. Faggioli, N. Ferro, P. Rosso, D. Spina (Eds.), Working Notes of CLEF 2025 – Conference and Labs of the Evaluation Forum, CEUR Workshop Proceedings, CEUR-WS.org, 2025.

[5] Z. N., Ways of formation of invective vocabulary (based on the material of the piraeus national variant of the spanish language and the american national variant of the english language), https://cyberleninka.ru/article/n/puti-formirovaniya-invektivnoy-leksiki-na-materiale-pireyskogo-natsionalno-varianta-ispanskogo-yazyka-i-amerikanskogo-natsionalnogo, 2010.

[6] D. K. Vakhitova, Invective Vocabulary of the Tatar Language: Functional and Ethnomental Aspects, Ph.d. dissertation, Kazan Federal University, Kazan, Russia, 2013.

[7] Hugging Face, Multilingual toxic lexicon, https://huggingface.co/datasets/textdetox/multilingual_toxic_lexicon, 2025. Accessed: 2025-07-06.

[8] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, P. J. Liu, Exploring the limits of transfer learning with a unified text-to-text transformer, Journal of Machine Learning Research 21 (2020) 1–67.

[9] T. Hosking, hrq-vae, https://github.com/tomhosking/hrq-vae, 2025. Accessed: 2025-07-06.

[10] D. P. Kingma, M. Welling, Auto-encoding variational bayes, https://arxiv.org/abs/1312.6114, 2013. ArXiv preprint arXiv:1312.6114.

[11] V. Logacheva, D. Dementieva, S. Ustyantsev, D. Moskovskiy, D. Dale, I. Krotova, N. Semenov, A. Panchenko, ParaDetox: Detoxification with parallel data, in: Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Dublin, Ireland, 2022, pp. 6804–6818. URL: https://aclanthology.org/2022.acl-long.469.

[12] D. Dementieva, S. Ustyantsev, D. Dale, O. Kozlova, N. Semenov, A. Panchenko, V. Logacheva, Crowdsourcing of parallel corpora: the case of style transfer for detoxification, in: Proceedings of the 2nd Crowd Science Workshop: Trust, Ethics, and Excellence in Crowdsourced Data Management at Scale co-located with 47th International Conference on Very Large Data Bases (VLDB 2021 (https://vldb.org/2021/)), CEUR Workshop Proceedings, Copenhagen, Denmark, 2021, pp. 35–49. URL: http://ceur-ws.org/Vol-2932/paper2.pdf.

[13] T.-Y. Lin, M. Maire, S. Belongie, L. Bourdev, R. Girshick, J. Hays, P. Perona, D. Ramanan, C. L. Zitnick, P. Dollár, Microsoft coco: Common objects in context, 2015. URL: https://arxiv.org/abs/1405.0312. arXiv:1405.0312.

[14] A. Fader, L. Zettlemoyer, O. Etzioni, Paraphrase-driven learning for open question answering, in: Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL), 2013, pp. 1608–1618.

[15] Z. Wang, W. Hamza, R. Florian, Bilateral multi-perspective matching for natural language sentences, in: Proceedings of the 26th International Joint Conference on Artificial Intelligence (IJCAI), 2017, pp. 4144–4150.

[16] OpenAI, Chatgpt (may 15 version) [large language model], https://openai.com/chatgpt, 2025. San Francisco, CA, USA, Accessed: 2025-07-06.

[17] A. Grattafiori, A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Vaughan, A. Yang, A. Fan, A. Goyal, A. Hartshorn, A. Yang, A. Mitra, A. Sravankumar, A. Korenev, A. Hinsvark, A. Rao, A. Zhang, A. Rodriguez, A. Gregerson, A. Spataru, B. Roziere, B. Biron, B. Tang, B. Chern, C. Caucheteux, C. Nayak, C. Bi, C. Marra, C. McConnell, C. Keller, C. Touret, C. Wu, C. Wong, C. C. Ferrer, C. Nikolaidis, D. Allonsius, D. Song, D. Pintz, D. Livshits, D. Wyatt, D. Esiobu, D. Choudhary, D. Mahajan, D. Garcia-Olano, D. Perino, D. Hupkes, E. Lakomkin, E. AlBadawy, E. Lobanova, E. Dinan, E. M. Smith, F. Radenovic, F. Guzmán, F. Zhang, G. Synnaeve, G. Lee, G. L. Anderson, G. Thattai, G. Nail, G. Mialon, G. Pang, G. Cucurell, H. Nguyen, H. Korevaar, H. Xu, H. Touvron, I. Zarov, I. A. Ibarra, I. Kloumann, I. Misra, I. Evtimov, J. Zhang, J. Copet, J. Lee, J. Geffert, J. Vranes, J. Park, J. Mahadeokar, J. Shah, J. van der Linde, J. Billock, J. Hong, J. Lee, J. Fu, J. Chi, J. Huang, J. Liu, J. Wang, J. Yu, J. Bitton, J. Spisak, J. Park, J. Rocca, J. Johnstun, J. Saxe, J. Jia, K. V. Alwala, K. Prasad, K. Upasani, K. Plawiak, K. Li, K. Heafield, K. Stone, K. El-Arini, K. Iyer, K. Malik, K. Chiu, K. Bhalla, K. Lakhotia, L. Rantala-Yeary, L. van der Maaten, L. Chen, L. Tan, L. Jenkins, L. Martin, L. Madaan, L. Malo, L. Blecher, L. Landzaat, L. de Oliveira, M. Muzzi, M. Pasupuleti, M. Singh, M. Paluri, M. Kardas, M. Tsimpoukelli, M. Oldham, M. Rita, M. Pavlova, M. Kambadur, M. Lewis, M. Si, M. K. Singh, M. Hassan, N. Goyal, N. Torabi, N. Bashlykov, N. Bogoychev, N. Chatterji, N. Zhang, O. Duchenne, O. Çelebi, P. Alrassy, P. Zhang, P. Li, P. Vasic, P. Weng, P. Bhargava, P. Dubal, P. Krishnan, P. S. Koura, P. Xu, Q. He, Q. Dong, R. Srinivasan, R. Ganapathy, R. Calderer, R. S. Cabral, R. Stojnic, R. Raileanu, R. Maheswari, R. Girdhar, R. Patel, R. Sauvestre, R. Polidoro, R. Sumbaly, R. Taylor, R. Silva, R. Hou, R. Wang, S. Hosseini, S. Chennabasappa, S. Singh, S. Bell, S. S. Kim, S. Edunov, S. Nie, S. Narang, S. Raparthy, S. Shen, S. Wan, S. Bhosale, S. Zhang, S. Vandenhende, S. Batra, S. Whitman, S. Sootla, S. Collot, S. Gururangan, S. Borodinsky, T. Herman, T. Fowler, T. Sheasha, T. Georgiou, T. Scialom, T. Speckbacher, T. Mihaylov, T. Xiao, U. Karn, V. Goswami, V. Gupta, V. Ramanathan, V. Kerkez, V. Gonguet, V. Do, V. Vogeti, V. Albiero, V. Petrovic, W. Chu, W. Xiong, W. Fu, W. Meers, X. Martinet, X. Wang, X. Wang, X. E. Tan, X. Xia, X. Xie, X. Jia, X. Wang, Y. Goldschlag, Y. Gaur, Y. Babaei, Y. Wen, Y. Song, Y. Zhang, Y. Li, Y. Mao, Z. D. Coudert, Z. Yan, Z. Chen, Z. Papakipos, A. Singh, A. Srivastava, A. Jain, A. Kelsey, A. Shajnfeld, A. Gangidi, A. Victoria, A. Goldstand, A. Menon, A. Sharma, A. Boesenberg, A. Baevski, A. Feinstein, A. Kallet, A. Sangani, A. Teo, A. Yunus, A. Lupu, A. Alvarado, A. Caples, A. Gu, A. Ho, A. Poulton, A. Ryan, A. Ramchandani, A. Dong, A. Franco, A. Goyal, A. Saraf, A. Chowdhury, A. Gabriel, A. Bharambe, A. Eisenman, A. Yazdan, B. James, B. Maurer, B. Leonhardi, B. Huang, B. Loyd, B. D. Paola, B. Paranjape, B. Liu, B. Wu, B. Ni, B. Hancock, B. Wasti, B. Spence, B. Stojkovic, B. Gamido, B. Montalvo, C. Parker, C. Burton, C. Mejia, C. Liu, C. Wang, C. Kim, C. Zhou, C. Hu, C.-H. Chu, C. Cai, C. Tindal, C. Feichtenhofer, C. Gao, D. Civin, D. Beaty, D. Kreymer, D. Li, D. Adkins, D. Xu, D. Testuggine, D. David, D. Parikh, D. Liskovich, D. Foss, D. Wang, D. Le, D. Holland, E. Dowling, E. Jamil, E. Montgomery, E. Presani, E. Hahn, E. Wood, E.-T. Le, E. Brinkman, E. Arcaute, E. Dunbar, E. Smothers, F. Sun, F. Kreuk, F. Tian, F. Kokkinos, F. Ozgenel, F. Caggioni, F. Kanayet, F. Seide, G. M. Florez, G. Schwarz, G. Badeer, G. Swee, G. Halpern, G. Herman, G. Sizov, Guangyi, Zhang, G. Lakshminarayanan, H. Inan, H. Shojanazeri, H. Zou, H. Wang, H. Zha, H. Habeeb, H. Rudolph, H. Suk, H. Aspegren, H. Goldman, H. Zhan, I. Damlaj, I. Molybog, I. Tufanov, I. Leontiadis, I.-E. Veliche, I. Gat, J. Weissman, J. Geboski, J. Kohli, J. Lam, J. Asher, J.-B. Gaya, J. Marcus, J. Tang, J. Chan, J. Zhen, J. Reizenstein, J. Teboul, J. Zhong, J. Jin, J. Yang, J. Cummings, J. Carvill, J. Shepard, J. McPhie, J. Torres, J. Ginsburg, J. Wang, K. Wu, K. H. U, K. Saxena, K. Khandelwal, K. Zand, K. Matosich, K. Veeraraghavan, K. Michelena, K. Li, K. Jagadeesh, K. Huang, K. Chawla, K. Huang, L. Chen, L. Garg, L. A, L. Silva, L. Bell, L. Zhang, L. Guo, L. Yu, L. Moshkovich, L. Wehrstedt, M. Khabsa, M. Avalani, M. Bhatt, M. Mankus, M. Hasson, M. Lennie, M. Reso, M. Groshev, M. Naumov, M. Lathi, M. Keneally, M. Liu, M. L. Seltzer, M. Valko, M. Restrepo, M. Patel, M. Vyatskov, M. Samvelyan, M. Clark, M. Macey, M. Wang, M. J. Hermoso, M. Metanat, M. Rastegari, M. Bansal, N. Santhanam, N. Parks, N. White, N. Bawa, N. Singhal, N. Egebo, N. Usunier, N. Mehta, N. P. Laptev, N. Dong,

N. Cheng, O. Chernoguz, O. Hart, O. Salpekar, O. Kalinli, P. Kent, P. Parekh, P. Saab, P. Balaji, P. Rittner, P. Bontrager, P. Roux, P. Dollar, P. Zvyagina, P. Ratanchandani, P. Yuvraj, Q. Liang, R. Alao, R. Rodriguez, R. Ayub, R. Murthy, R. Nayani, R. Mitra, R. Parthasarathy, R. Li, R. Hogan, R. Battey, R. Wang, R. Howes, R. Rinott, S. Mehta, S. Siby, S. J. Bondu, S. Datta, S. Chugh, S. Hunt, S. Dhillon, S. Sidorov, S. Pan, S. Mahajan, S. Verma, S. Yamamoto, S. Ramaswamy, S. Lindsay, S. Lindsay, S. Feng, S. Lin, S. C. Zha, S. Patil, S. Shankar, S. Zhang, S. Zhang, S. Wang, S. Agarwal, S. Sajuyigbe, S. Chintala, S. Max, S. Chen, S. Kehoe, S. Satterfield, S. Govindaprasad, S. Gupta, S. Deng, S. Cho, S. Virk, S. Subramanian, S. Choudhury, S. Goldman, T. Remez, T. Glaser, T. Best, T. Koehler, T. Robinson, T. Li, T. Zhang, T. Matthews, T. Chou, T. Shaked, V. Vontimitta, V. Ajayi, V. Montanez, V. Mohan, V. S. Kumar, V. Mangla, V. Ionescu, V. Poenaru, V. T. Mihailescu, V. Ivanov, W. Li, W. Wang, W. Jiang, W. Bouaziz, W. Constable, X. Tang, X. Wu, X. Wang, X. Wu, X. Gao, Y. Kleinman, Y. Chen, Y. Hu, Y. Jia, Y. Qi, Y. Li, Y. Zhang, Y. Zhang, Y. Adi, Y. Nam, Yu, Wang, Y. Zhao, Y. Hao, Y. Qian, Y. Li, Y. He, Z. Rait, Z. DeVito, Z. Rosnbrick, Z. Wen, Z. Yang, Z. Zhao, Z. Ma, The llama 3 herd of models, 2024. URL: https://arxiv.org/abs/2407.21783. arXiv:2407.21783.

[18] Hugging Face, Flan-t5 models, https://huggingface.co/google/flan-t5-small, https://huggingface.co/google/flan-t5-base, https://huggingface.co/google/flan-t5-large, 2025. Accessed: 2025-07-06.

[19] s-nlp, Parallel detoxification dataset (small), https://github.com/s-nlp/parallel_detoxification_dataset/blob/main/parallel_detoxification_dataset_small.tsv, 2025. Accessed: 2025-07-06.

[20] Hugging Face, Paranmt for detox, https://huggingface.co/datasets/s-nlp/paranmt_for_detox, 2025. Accessed: 2025-07-06.

[21] s-nlp, Detox releases, https://github.com/s-nlp/detox/releases/, 2025. Accessed: 2025-07-06.

[22] Hugging Face, mt0-xl-detox-orpo, https://huggingface.co/s-nlp/mt0-xl-detox-orpo, 2025. Accessed: 2025-07-06.

[23] Hugging Face, Spivavtor models, https://huggingface.co/grammarly/spivavtor-xxl, https://huggingface.co/grammarly/spivavtor-large, 2025. Accessed: 2025-07-06.

[24] Hugging Face, Meta-llama-3-70b, https://huggingface.co/meta-llama/Meta-Llama-3-70B, 2025. Accessed: 2025-07-06.

[25] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen, Lora: Low-rank adaptation of large language models, in: Proceedings of the 10th International Conference on Learning Representations (ICLR), 2022.