Записки научных
семинаров ПОМИ
Том 552, 2026 г.

**D. Galimzianova, S. Gorovaia, D. Vikhorev,
I. P. Yamshchikov, E. Zhemchuzhina**

CLEANCOMEDY: CREATING FRIENDLY HUMOR THROUGH GENERATIVE TECHNIQUES

Abstract. Humor generation is a challenging task in natural language processing due to limited resources and the quality of existing datasets. Available humor language resources often suffer from toxicity and duplication, limiting their effectiveness for training robust models. This paper proposes CleanComedy, a specialized, partially annotated toxicity-filtered corpora of English and Russian jokes collected from various sources. We study the effectiveness of our data filtering approach through a survey on humor and toxicity levels in various joke groups. In addition, we investigate advances in computational humor generation by comparing human-written jokes with jokes generated by large language models. We fine-tuned several models on the CleanComedy corpora, generated jokes using both fine-tuned and original pre-trained versions, and collected human feedback to assess the quality and humor of the outputs.

§1. Introduction

Humor is a widely recognized but inherently complex phenomenon: it is intuitively associated with amusement and laughter, and serves as a powerful tool for social bonding, communication, and creativity. Despite its apparent ubiquity, there is no universally accepted definition of humor. As noted by Nabil, "humor is a phenomenon that does not have a universally accepted definition" [30]. Its interpretation varies across linguistic, cognitive, cultural, and contextual dimensions, making it difficult to formalize. Similar observations are made by Karyakin [31], who emphasize that humor encompasses a wide range of mechanisms—from wordplay to situational and cultural references—and is a multilayered cognitive construct not easily reducible to a single computational framework.

Computational humor is a significant area of research within natural language processing. The humor analysis is complex due to its reliance on contextual dependencies. Developing AI systems capable of generating

Key words and phrases: humor generation, computational humor, large language models, dataset, toxicity filtering, text generation.

humor is important for improving human-computer interaction, making it more natural, engaging, and relatable. As AI becomes increasingly integrated into everyday life, the ability to generate appropriate and friendly humor could enhance applications ranging from education and therapy to entertainment and customer service. However, many current approaches rely heavily on large-scale models accessed through costly APIs, making them less accessible and harder to deploy in lightweight settings. Our research aims to address this gap by developing a more efficient, lightweight approach to humor generation.

Previous works in this field [13, 3, 4] have laid the baseline in humor recognition. Recent research has focused on developing humor generation solutions based on LLMs [18, 19]. LLMs capture complex linguistic patterns, cultural references, and subtle nuances of humor far better than traditional template-based or rule-based systems, which often produce rigid, repetitive, and shallow jokes that lack the richness and creativity of human humor.

In this study, we collect, filter, and prepare humor corpora in English and Russian languages. In order to further enhance the utility of the corpora, we collect human feedback for 1,000 Russian and 1,000 English jokes. We publish individual annotations for every joke without aggregating scores provided by multiple annotators.

The datasets are publicly available on GitHub [1], facilitating access for the research community.

Following, we fine-tune and align an LLM on the prepared data, generate new humor samples and ask human annotators to evaluate them. The model undergoes two training stages: **Supervised Fine-Tuning** and **Alignment** to learn the humor style and subsequently pick up what "funniness" is. We use human feedback obtained from the annotators during the alignment stage.

The human feedback is then obtained for results of the two-stage fine-tuning presented in this research. They show that this training technique might give rise to lightweight humorous models and encourage responsible and effective generative AI.

According to our results, although one could cherry-pick some fun and interesting jokes, generative humor generally remains an open research problem.

[1] `https://github.com/gorovuha/CleanComedy`

This paper presents a detailed account of our methodology for dataset curation, filtering, and annotation, and provides corpora for future research, as well as human feedback. We also describe the settings and results of experiments with LLMs fine-tuned on our dataset, evaluating their performance against existing models with annotators' feedback.

§2. Related work

Computational humor research has seen significant advancements in recent years. One of the primary challenges in this field is capturing the details and contextual dependencies that make humor effective and enjoyable.

Works in computational humor [4, 3, 13] lay the groundwork by exploring the field of humor recognition. They introduce linguistic features, which are essential in distinguishing humorous content from non-humorous text.

Humor often involves a delicate balance of positive and negative emotions, the work [2] explores the dual nature of humor and offense, highlighting the importance of sentiment analysis in humor research.

According to one of the humor theories, the element of surprise is what constitutes humor [25]. This incongruity theory has been tested with computational methods [26] and utilized to generate humorous content [27, 28].

2.1. Data. The evolution of computational humor research has produced a variety of popular datasets, which serve as the foundational data sources for training humor recognition and generation models. Table 1 provides an overview of the most commonly referenced datasets.

These datasets are broadly used in solving computational humor tasks, yet they share two major problems that hinder their usage for the training of generative humor models:

(1) **Toxicity**: Many jokes in these datasets contain offensive content, including sexist, racist, or otherwise discriminatory jokes. This does not only raise ethical concerns but also impacts the generalization capacity of the humor models trained on such data.
(2) **Redundancy**: There is a high number of duplicate jokes, both explicit when the same exact words are used, and implicit when the same joke is paraphrased multiple times. Training models on duplicate data can lead to overfitting.

№	Description	Jokes	Non-jokes
	English		
1 [13]	Humorous samples from daily joke websites	16,000	16,000
2 [2]	Short jokes scraped from various joke websites with lengths ranging from 10 to 200 characters	231,657	0
3 [2]	80% jokes from twitter and 20% jokes from Short Jokes dataset	4,932	3,068
4 [14]	Multiple sources were combined and deduplicated (the two sources for jokes are CrowdTruth and Subreddits)	96,910	173,633
5 [15]	Scrapped from Reddit and includes different types of humour	92,153	10,710
6 [16]	The original part is scraped from Reddit	14,946	0
7 [3]	Puns from Pun of the Day website	2,423	2,423
8	Ethical filtered jokes with 2-scale score	44,481	0
9	Ethical filtered jokes with human humor 5-scale score	1,000	0
	Russian		
10 [1]	One-liners from social media	46,608	46,608
11 [24]	Russian jokes from social media and online collections	156,605	156,605
12	Ethical filtered jokes with 2-scale score	40,926	0
13	Ethical filtered jokes with human humour 5-scale score	1,000	0

Table 1. Comparison of Humor Datasets: 1 – 16k One-Liners, 2 – Short Jokes, 3 – SemEval 2021 Task 7: HaHackathon, Detecting and Rating Humor and Offense, 4 – Knowledge Amalgam: Generating Jokes and Quotes Together, 5 – The Naughtyformer: A Transformer Understands Offensive Humor, 6 – Humor Detection: A Transformer Gets the Last Laugh, 7 – Pun of the Day, 8 – CleanComedy English, 9 – CleanComedy English Gold, 10 – Stierlitz, 11 – FUN, 12 – CleanComedy Russian, 13 – CleanComedy Russian Gold.

2.2. LLMs. There have been several recent attempts to explore the humor abilities of LLMs.

The authors in [18] explore the abilities of GPT-3.5 in both creating and comprehending humor. Their findings indicate that the model struggles with the generation and interpretation of jokes. For instance, GPT-3.5 produces only a limited set of unique jokes, which appear to be taken from a prepared list of appropriate jokes.

LLMs as creative writing assistants for comedians are studied in [21]. It is concluded that the models can be deployed successfully in this quality with certain restrictions, such as contextual alignment. The paper [22] introduces an adaptive humor generation model that personalizes content based on user feedback, demonstrating improved engagement through tailored humor. LLM's reasoning capabilities are explored in [29] where it is shown that the multi-step reasoning approach can improve the quality of generative humor.

§3. Dataset curation

In this section, we describe the methodology employed in collecting and processing our dataset, which includes humor content in both English and Russian languages. The approach to dataset curation is guided by the goal to create a resource that is not only large and varied, but also devoid of offensive content and redundancies that could potentially bias or undermine the performance of language models trained on it.

3.1. Data processing. First, we utilize resources reviewed in Section 2.1, see Table 1 for a quick overview. Then we unite it into one large collection and remove exact duplicates, ignoring punctuation and case. Every step of the further filtering process removes bad samples. We do not paraphrase the jokes in order to preserve the intricate semantics of humor.

The initial preprocessing of English data involves the keeping of examples that contain only Latin characters and punctuation marks to unify the joke structure. It helps get rid of bad examples comprising emojis, links and also, for example, corrupted jokes from Reddit containing *[deleted]* or *[removed]* words (these words indicate that a part of the joke was deleted before it was collected). We also noticed that excessively short or long entries have a large amount of noise (for example, repetitions, unfunny utterances), so the next step was to keep examples only between 50 and 150 characters. This window length was chosen by the series of experiments, see Table 3 in Appendix. The remaining part of English jokes after this step consists of 230k jokes, for Russian there were no such noise.

For both English and Russian, we utilized the unbiased toxicity classifier Detoxify [5] for extracting toxic jokes as this model is capable of detecting different types of toxicity like threats, obscenity, insults, and identity-based hate. As the authors of the classifier mention, words associated with swearing, insults, or profanity are likely to be classified as toxic, regardless of the tone or the intent. According to our goal, which is to collect an ethical humor dataset, we removed all the jokes that were supposed to be toxic. Due to the lower representation of Russian in multilingual models, we employ the ruBERTConv Toxic Classifier[3], a transformer model tailored for the Russian language [12, 11], to improve the accuracy of toxicity recognition. The remaining part of the English dataset contained 99k jokes, Russian – 100k.

To identify and remove duplicate jokes, we utilize the Sentence-BERT pretrained model **all-mpnet-base-v2** [6, 10]. This framework allows us to compute text embeddings and then compare their cosine similarity and find sentences with similar semantics [7]. The goal is to remove jokes with the same meaning expressed in different words or jokes where some parts are presented in a different order (see Table 5 in the Appendix). As we collect English jokes from various sources, there is a large number of duplicate instances. We consider two entries as duplicates if the cosine between their embeddings is higher than 0.7 for English and higher than 0.9 for Russian, these thresholds were chosen empirically due to a series of experiments. Thus, at this step, we removed a considerable number of 29k duplicating jokes, and the remaining part is 73k unique jokes in English and 80k in Russian.

After that, rigorous manual analysis of the dataset shows that a large part of jokes contains metaphorical insults. To make the resulting data cleaner, we label jokes with zero-shot classifier **DeBERTa-v3-large-mnli-fever-anli-ling-wanli** [8]. The set of labels this time is *politics, neutral, offending, alcohol, drugs, racist*. Thus, we removed all instances labeled as political, racist, and insulting content (see Table 6 in the Appendix) to avoid unethical expressions, with 49k jokes left in English and 52k left in Russian.

A thorough analysis of the remaining jokes reveals a significant number of texts on sensitive topics. Therefore, we utilize topic modeling to understand our large corpus better. Our objective is to have topics that are specific enough (e.g., separating soft drinks from alcohol) but not overly

[3]`https://huggingface.co/IlyaGusev/rubertconv_toxic_clf`

fragmented (e.g., having 100 classes about animals would be uninteresting). **BERTopic**, a topic modeling technique that leverages transformers and c-TF-IDF [9], is used to create clusters with easily interpretable topics while keeping important words in the topic descriptions. The visual representation of these clusters can be found in Figures 1, 2.

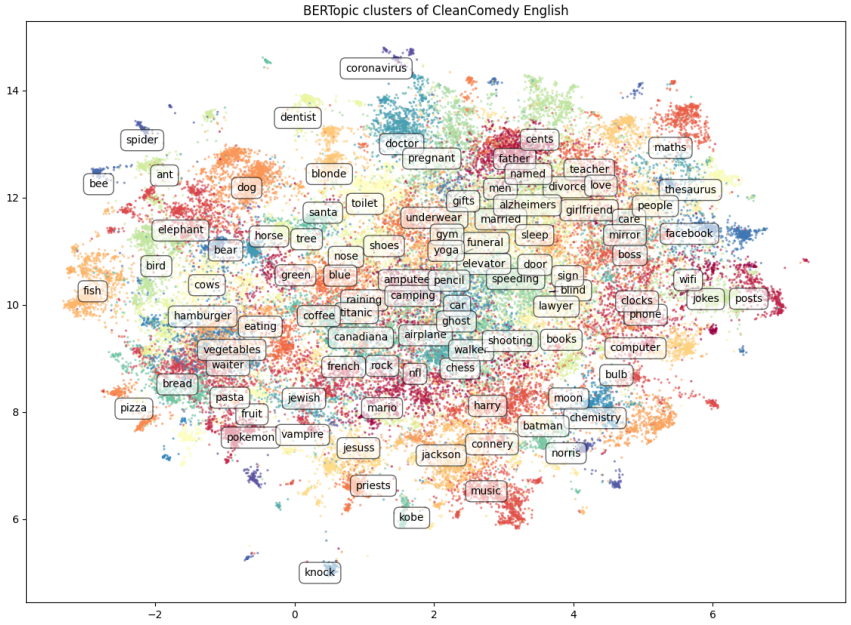

Figure 1. Topic modeling for CleanComedy English.

Cluster modeling allows us to thoroughly analyze the collected dataset in terms of diversity and content distribution. We remove sensitive jokes that belong to the clusters about *religion, funerals, bathrooms, officers, pregnancy, nations, disabilities, and divorce* ensuring an ethical and useful dataset.

The final remaining dataset, after the full set of filters has been applied, contains 44k English jokes and 40k Russian jokes. The comprehensive filtering pipeline can be found in Figures 3 and 4.

3.2. Dataset annotation. Datasets for both languages include manual human annotation of 1,000 samples for each language. We call this subset

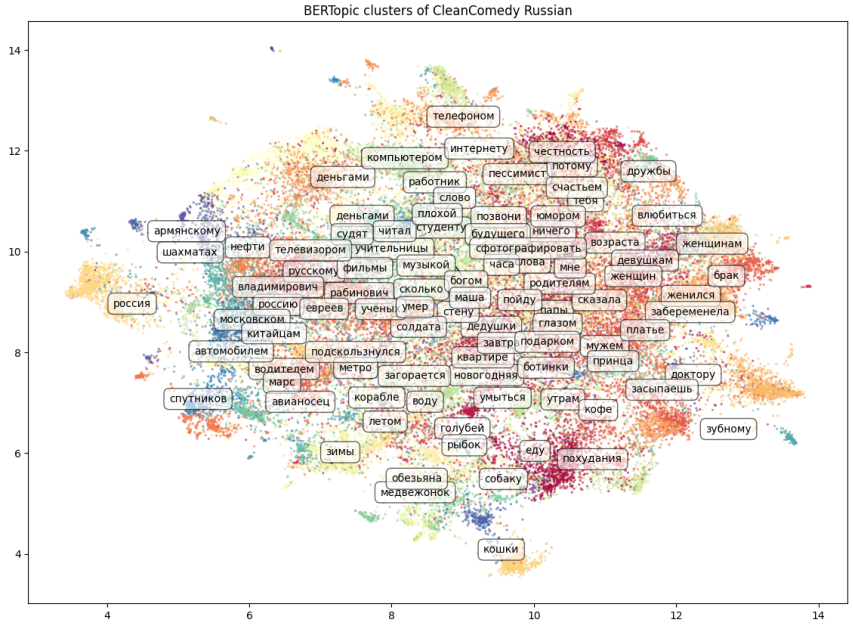

Figure 2. Topic modeling for CleanComedy Russian.

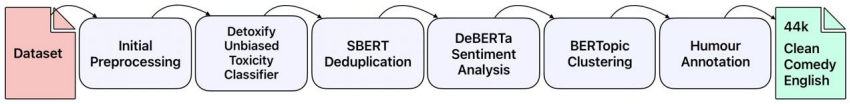

Figure 3. Full filtering pipeline for CleanComedy English.

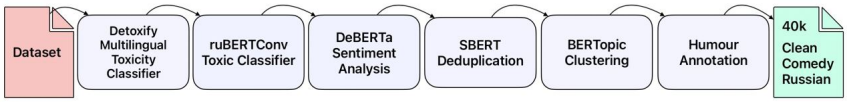

Figure 4. Full filtering pipeline for CleanComedy Russian.

of the data **CleanComedy Gold**. Volunteers have been asked to rate a

joke on a scale from 1 to 5 depending on how funny they believe the joke is. The instructions for the annotation included the following points:

- Rate how funny a joke is on a five-point scale, where 1 is not funny and 5 is very funny.
- Do you find the text of the joke vulnerable or inappropriate?

Each joke has been rated by five different annotators. The scores were collected through **Telegram** bot by crowdsourcing. The annotators volunteered their time, dedicating a maximum of one hour. For each instance, we calculated the mean score for five annotators, which is used in CleanComedy Gold datasets as the human humor feedback, see Figure 5. We also publish individual scores to facilitate further research on the personalization of generative humor.

With the consent of the annotators, we also publish their anonymized data, such as age, gender (female, male, and other), education level (secondary vocational, bachelor's degree, specialist degree, master's degree, PhD) and the language fluency level for both languages (from A1 to C2 + native).

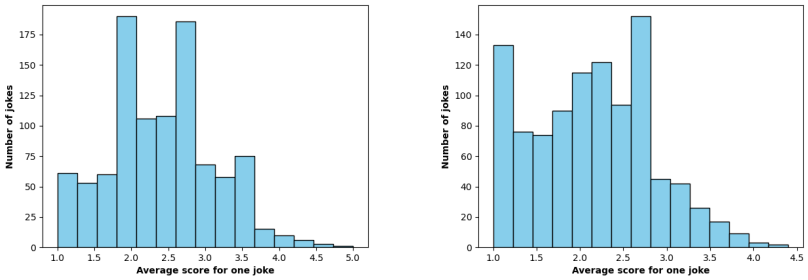

Figure 5. Average scores for **CleanComedy Gold** datasets in English (in the left picture) and Russian (in the right picture). The average score of 5 annotators was computed for each joke.

§4. Models

In our experiments, we fine-tune a pre-trained version of a multilingual large language model meta-llama/Llama-3.1-8B. We utilize the pre-trained

version of the model because we believe that in this setting the influence of our dataset on the final result increases, while using the instruction model meta-llama/Llama-3.1-8B-Instruct increases the dependence on both the model itself and the prompt selected for it.

Our prompt for fine-tuning models looks like this:

```
<|begin_of_text|>
<|reserved_special_token_{i}|>
{text}
```

where `begin_of_text` is the start of text token, `reserved_special_token_{i}` is the language token (i=0 for English and i=1 for Russian language), and `{text}` is a joke text.

For fine-tuning the model, we use LoRA [17]. Instead of updating all the weights in every layer, we only train low-rank approximations (with a rank of 4) for all layers except for the classification and embedding layers, which we train fully.

We fine-tune the model in two stages: **Supervised Fine-Tuning** and **LLM Alignment**. At the first stage, we train meta-llama/Llama-3.1-8B to imitate the jokes of our large datasets. At the second stage, we additionally train the model obtained at the previous stage to generate funny examples from our small annotated dataset and not generate unfunny ones. At both stages, optimization is performed using the Adam algorithm, coupled with a cosine learning rate scheduler. The batch size per these two stages is equal to 64.

4.1. Supervised fine-tuning. At this stage, we train meta-llama/Llama-3.1-8B using the datasets described in Data processing subsection.

The model's training employs the standard language modeling loss function. A fine-tuning phase takes 2 epochs using a maximum learning rate of 1e-05.

4.2. LLM alignment. At this stage, we additionally train the model obtained at the previous stage using the CleanComedy Gold dataset. We calculate the level of humor in a joke by averaging the annotators' ratings in this one. Since these average ratings are between 1 and 5, we apply a linear transformation to map the scores to a 0-to-1 scale to get soft labels for the sequence classification task.

The model is trained using binary cross-entropy as the loss function for binary classification tasks inspired by **DPO** [23] and **SimPO** [20]. Our method differs from these methods by using soft positive/negative labels instead of requiring a dataset of chosen-rejected answer pairs:

$$t = (s - 1)/4,$$

$$loss = -\sum_{(t,p)\in B} (t \log(p) + (1 - t) \log(1 - p)),$$

where s[4] is an average humor score of a joke from the annotated dataset and p is the probability of the same joke according to the language model that is trained.

A fine-tuning phase takes 5 epochs using a maximum learning rate of 4e-07.

§5. Results

For both languages, we sample 100 examples from each of the following six groups:

(1) **LLaMA 3.1 8B (Supervised Fine-Tuned).** See Supervised fine-tuning section for training details.
(2) **LLaMA 3.1 8B (Supervised Fine-Tuned + Aligned).** See LLM alignment section for training details.
(3) **LLaMA 3.1 8B (Instruct).** We use following prompts for two languages.

System prompt:
You are a comedian with a lot of experience. Your income level and future career depend on the level of humor in jokes. Since your audience is educated, it's best to avoid using well-known jokes.

User prompt:
Write a funny short joke for your performance. I only need the text of this joke and nothing else.

[4]The range of s varies from 1 to 5, so t falls between 0 and 1 and can be interpreted as the target probability for classification task using soft labels.

(4) **GPT-4o.** Since we don't have access to the GPT-4o API, we use the **default system prompt** and generate all 100 jokes of the class at one time, using the following prompt:

> **User prompt:**
> *Hi! Imagine that you are a comedian with a lot of experience. Write 100 funny short jokes for your performance. Your income level and future career depend on them.*

(5) **Unfiltered Dataset.** In order to understand how the levels of humor and toxicity in the dataset are changed after toxicity filters, we also take 100 samples from the dataset without toxicity filters.

(6) **Clean Dataset.** In order to understand how large language models are able to learn humor from humorous data, we also take 100 samples from the dataset as some reference for comparing humor scores.

We employ ancestral sampling with a temperature of 0.5 for all generated examples, except for the **LLaMA 3.1 8B (Instruct)** model for English, where we use a temperature of 0.9 to mitigate its tendency towards generating absolutely identical texts at lower temperatures. Additionally, for the **LLaMA 3.1 8B (Instruct)** generations, we remove semantic duplicates using the SBERT model in the same manner as during the dataset deduplication stage, applying the same cutoff threshold for identifying duplicates[5].

Each of the 600 sampled examples was evaluated by 3 individuals, following a process similar to that described in Dataset annotation subsection. In this case, in addition to rating each joke, the evaluators were also asked to assess its level of toxicity or inappropriateness[6]. The average ratings for each group can be found in Table 2. For both English and Russian, the toxicity percentage in our clean datasets is half that of the originally collected data, demonstrating the effectiveness of the aforementioned filtering process. The data on the demographics of the annotators can be found in Figures 7, 8, 9 and 10.

The evaluation of humor generation models reveals interesting trends across different setups for both English and Russian datasets.

[5]We do this for two languages.

[6]We ask: "Do you find this text vulnerable or inappropriate?" An annotator can answer "yes" or "no".

Model	Humor score	Toxicity percentage
English		
LLaMA 3.1 8B (SFT)	2.11 ± 1.2	13.38
LLaMA 3.1 8B (SFT + Aligned)	2.02 ± 1.08	15.38
LLaMA 3.1 8B (Instruct)	2.65 ± 1.23	**3.97**
GPT-4o	$\mathbf{3.02} \pm 1.3$	9.27
Unfiltered Dataset	2.72 ± 1.42	26.09
Clean Dataset	2.96 ± 1.36	11.41
Russian		
LLaMA 3.1 8B (SFT)	1.68 ± 1.09	4.01
LLaMA 3.1 8B (SFT + Aligned)	1.74 ± 1.06	**3.3**
LLaMA 3.1 8B (Instruct)	2.07 ± 1.27	4.26
GPT-4o	2.38 ± 1.23	4.67
Unfiltered Dataset	2.77 ± 1.4	20.93
Clean Dataset	$\mathbf{2.84} \pm 1.44$	11.37

Table 2. Comparison of average metrics for different groups of jokes. Standard deviations are also calculated for the humor scores. Each group includes 100 examples, with each example rated by at least 3 people. The toxicity percentage ranges from 0 to 100, while the humor score ranges from 1 to 5.

For English, among the models, GPT-4o achieves the highest humor score (3.02), closely followed by the Clean Dataset of human-written jokes (2.96). The LLaMA 3.1 8B (Instruct) model also performs reasonably well (2.65), but significantly lags behind GPT-4o, suggesting that high-quality humor generation remains a challenge for open models. In contrast, models fine-tuned on the CleanComedy dataset (SFT and SFT + Aligned) achieve noticeably lower humor scores (2.0–2.1), reflecting the difficulty of adapting LLMs to humor tasks using filtered or smaller-scale data. One possible reason for this is the bias in human annotation: the average humor scores across the dataset hover around 2.5, and the annotators may not have been fluent or culturally attuned enough to assess humor nuances in English jokes. This annotation noise likely contributed to the ineffectiveness of alignment procedures for English models.

In terms of toxicity, LLaMA 3.1 8B (Instruct) achieves the lowest toxicity percentage (3.97), even lower than the Clean Dataset (11.41), and

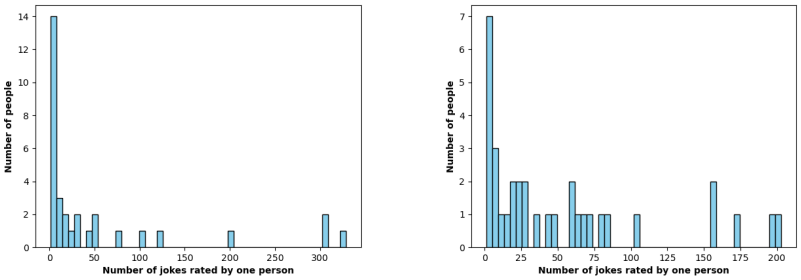

Figure 6. Number of jokes rated by one person for English (in the left picture) and Russian (in the right picture). Only annotators with at least one score are taken into account.

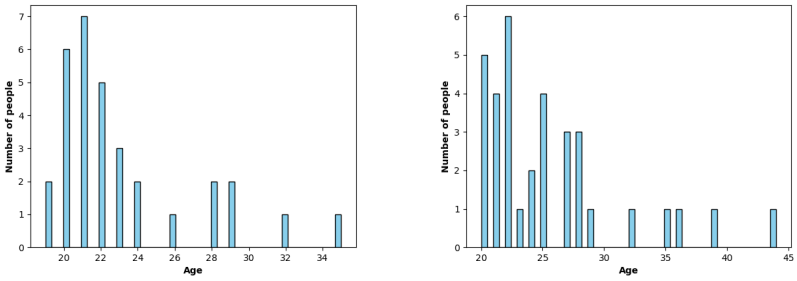

Figure 7. Age of annotators for English (in the left picture) and Russian (in the right picture). Only annotators with at least one score are taken into account.

significantly better than the unfiltered jokes (26.09). Interestingly, GPT-4o also maintains relatively low toxicity (9.27), suggesting that proprietary closed models like GPT-4o may benefit from more robust and large-scale alignment pipelines, which our lightweight methods cannot yet fully replicate. This points to a gap we aim to reduce with future improvements.

In Russian, similar trends are observed, but with more encouraging results. The Clean Dataset again achieves the highest humor score (2.84),

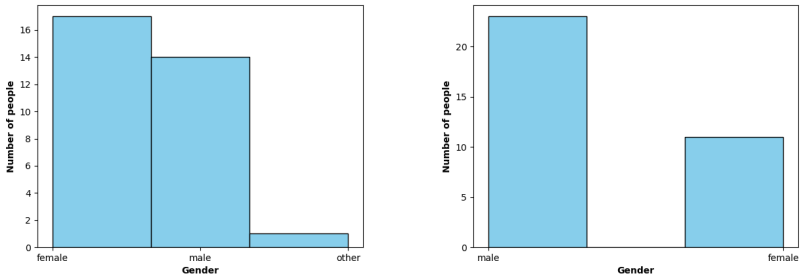

Figure 8. Gender of annotators for English (in the left picture) and Russian (in the right picture). Only annotators with at least one score are taken into account.

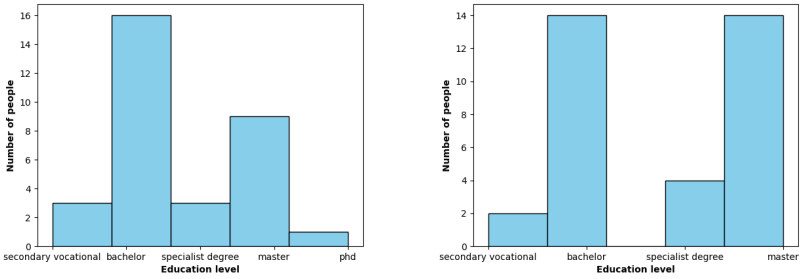

Figure 9. Education level of annotators for English (in the left picture) and Russian (in the right picture). Only annotators with at least one score are taken into account.

followed by GPT-4o (2.38). The LLaMA 3.1 8B (Instruct) model performs slightly worse (2.07), and fine-tuned models (SFT and SFT + Aligned) remain behind, consistent with the English results. However, in contrast to English, alignment in Russian does show clear benefits for reducing toxicity: toxicity drops from 4.26 (Instruct) to 4.01 (SFT) and further to 3.3 (SFT + Aligned). This confirms that our intended approach works in principle – alignment can reduce toxicity — and highlights that annotation

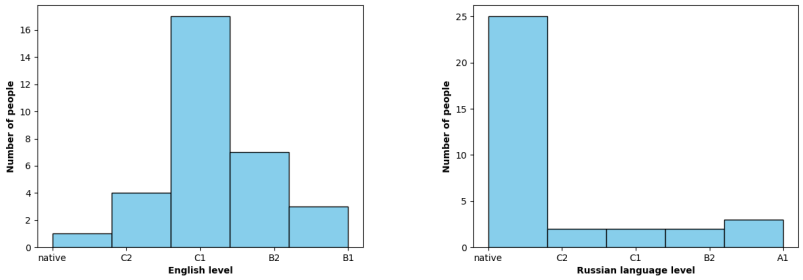

Figure 10. Language level of annotators. Only annotators with at least one evaluation point are taken into account.

quality plays a critical role in the effectiveness of such methods, particularly for subjective tasks like humor assessment.

The age histograms in Figure 7 show the majority of our annotators are aged 20–30, which skews the humor perception toward this age group. It is also important to point out that there were almost no native speakers of English among our volunteers (Figure 10), which has significantly influenced the humor scores obtained for the English jokes. We theorize that the English jokes generated by our models were rated higher than Russian due to this linguistic bias. The humor scores for the Russian generative jokes are lower for the models trained on the filtered data because they could have been deemed more boring for the annotators, as the humor scores are lower for the jokes with a lower toxicity percentage.

§6. Conclusion

In this research, we share the insights we gained while experimenting with prompting and fine-tuning models for humor generation.

This paper presents CleanComedy, a novel and ethically curated dataset for humor generation and evaluation in English and Russian. Through rigorous data filtering and annotation processes, we address the common challenges in datasets, such as toxicity and redundancy. By fine-tuning a LLM with this dataset, we demonstrated its capability to generate contextually appropriate and ethically aligned humor.

The results provided by the annotators of our experiments underscore the importance of alignment techniques in improving the quality and relevance of generated humor. While our models perform well within the scope of the dataset, challenges remain in generating humor that adapts to different languages and cultures.

This study contributes to the field by introducing methodologies that prioritize ethical considerations while advancing computational humor research. However, humor's subjective and cultural nature requires ongoing attention to ensure AI-generated content remains inclusive, engaging, and responsible.

Limitations

The English dataset faces challenges related to the quality and diversity of humor instances. While significant efforts are made to remove offensive and inappropriate content from the datasets, the filtering processes may not be ideal. Enhanced filtering techniques and ongoing monitoring are essential to mitigate this risk. Moreover, future research should focus on developing more advanced generative techniques to enhance the coherence and creativity of the outputs.

When training language models, we have only experimented with PEFT methods. Judging by the quality of the resulted generations, the models would benefit greatly from full fine-tuning. Another limitation of this study is the size of the dataset which comprised only ∼40k samples for each language. It seems obvious that collecting a bigger and higher quality humor dataset could further advance humor generation. For example, transcribing popular humor shows seems to be an interesting idea for humor data collection.

Humor is inherently connected to cultural and contextual details, which can be challenging for language models to understand and generate. The models we trained for English and Russian show different types of jokes. Thus, they might not perform well

across other cultural contexts and languages, raising the need for culturally adaptive humor generation techniques.

Ethical statement

The ethical risks that the computational humor research carries are discussed in this section to ensure responsible development of AI technologies.

Computational humor systems may mimic biases present in training data. Despite multiple stage filtering process and mitigation strategies to

prevent the spread of stereotypes and discriminatory content, generated humorous texts still can sometimes result in offensive or harmful content.

Text	Drawback
well.... [https://imgur.com/gallery/2CmdahS] (https://imgur.com/gallery/2CmdahS)	links
A veteran walks into a bar. 12 people in the bar [removed]	[removed]
Chinas president is so bad beacuse.... [Deleted]	[deleted]
lock lock me, i dare you.	not a joke (excessively short)
Test post Test	not a joke (excessively short)
meat meat meat meat meat meat meat meat meat meat meat meat meat meat meat meat meat meat meat meat meat meat meat meat meat meat meat meat meat meat meat meat meat meat meat meat meat meat meat meat meat meat meat meat meat meat meat meat meat meat meat meat meat meat meat meat meat meat meat meat meat meat meat meat meat meat meat meat meat meat meat meat meat	repetition (excessively long)
Mars is Earth's second moon Mars is Earth's second moon Mars is Earth's second moon Mars is Earth's second moon Mars is Earth's second moon... (and so on)	repetition (excessively long)

Table 3. Different types of noise before initial preprocessing of the English data. All the provided examples were removed after initial preprocessing that includes the removal of corrupted characters and excessively short or long examples (shorter than 30 or longer than 150 characters).

There is a risk that computational humor systems could be misused for malicious purposes, such as cyberbullying, harassment, or spreading misinformation, thus, we claim that our systems are not intended for any malicious applications.

The rise of computational humor may impact human writing and entertainment. It is essential to consider societal impacts and strive for a balance that supports human creativity while leveraging technological advancements.

Acknowledgments. We would like to express our gratitude to the volunteer crowdworkers who dedicated their time and effort to evaluate our generative models.

Appendix A. Supplementary material

This appendix collects the supplementary tables and figures referenced throughout the paper. Table 3 shows representative noise removed during preprocessing of the English data, and Table 4 compares toxicity classifiers on the Russian jokes. Table 5 illustrates near-duplicate jokes detected by cosine similarity, while Tables 6 and 7 report the zero-shot label and sentiment analysis used for data cleaning. Figures 6–10 summarise the annotator statistics for the human evaluation: the number of jokes rated per person and the age, gender, education level, and language level of the annotators.

Joke	ruBERT Conversational Toxicity Classifier	Multilingual Detoxify
Если бы я была принцессой, то моим замком был бы ликеро-водочный завод. *(If I were a princess, my castle would be a distillery.)*	0	0.01
Где твоя грудь? Ты ее спугнул. *(Where are your breasts? You scared them away.)*	1	0.83
Мы не будем больше встречаться... Что, тараканы в твоей голове проголосовали против меня? *(We won't be dating anymore... What, the bugs in your head voted against me?)*	0	0.73

Table 4. Different toxicity classifiers results for CleanComedy Russian, 1 – for toxic, 0 – neutral, for ruBERT-Conv we considered jokes with score more than 0.1 to be inappropriate.

Duplicate	Cosine
English	
what's a rock group with four guys that don't sing? mount rushmore	
What rock group has four guys who can't sing? Mount Rushmore.	0.922
who's an all male rock group that doesn't sing? mount rushmore .	0.8857
what do you call a rock group with no bassist, drummer, singer or guitarist? Mount Rushmore	0.741
Q: What do you call a male quartet? A: Three men and a tenor.	0.528
Russian	
а зачем в низу эскалатора бабулька в будке сидит? – она там педали крутит. *(Why is a grandma sitting in a booth at the bottom of the escalator? – She's pedalling there.)*	
А знаете, зачем в метро внизу эскалатора бабулька в будке? Она там педали крутит. *(Do you know why there is a granny in a booth in the subway at the bottom of the escalator? She's pedalling there.)*	0.928
Привет, как дела? Вот все тебе надо знать. *(Hi, how are you? You gotta know everything, don't you?)*	
– Привет! Как дела? – Порадовать тебя нечем. – Как? Неужели всё хорошо? *(Hello! How are you? – I can't please you with anything. – How come? Is everything really good?)*	0.9033

Table 5. Cosine similarity between duplicates.

Joke	Label
English	
Romeo: Your cheeks are like petals. Juliet: Really? Romeo: Yes, bicycle pedals.	neutral
What would one call a movie about meth addictions? Need for speed.	drugs
what do you call ghosts that haunt liquor stores? spirits	alcohol
What do Syrian refugees eat for breakfast? Syrial!	racist
Bought my epileptic girlfriend a strobe light for her birthday. She will have a fit when she sees it.	offending
Why can't you have Christmas dinner in the EU? Because there is no turkey.	politics
Russian	
Иногда я выпиваю стакан воды, просто для того что бы удивить свою печень. *(Sometimes I drink a glass of water, just to surprise my liver.)*	alcohol
– Знаешь, я целый месяц ходила на фитнес!... – И сколько сбросила? – 12 тысяч!!!... *(– You know, I went to fitness for a whole month!... – And how much did you lose? – 12 hundred!!!...)*	neutral
Лучше было бы, если бы народные избранники отвечали не за свои предвыборные слова, а за свои послевыборные дела. *(It would be better if the people's representatives were responsible not for their pre-election words, but for their post-election deeds.)*	politics
99 процентов моей подготовки к экзамену это курение с грустным лицом. *(99 of my exam preparation involves smoking with a sad face.)*	drugs
Бомжи вышли на демонстрацию против новых технологий. По их словам, в коробке из-под плоского телевизора вообще жить невозможно. *(Homeless people demonstrated against new technologies. According to them, it is generally impossible to live in a flat-screen TV box.)*	offending
Американский турист ходит с гидом по Лондону. — Все тут у вас такое маленькое, зажатое, — говорит он. — Это здание, например, было бы в Америке раз в десять выше. — О, конечно, сэр! Это-же психиатрическая клиника. *(An American tourist walks with a guide around London. "Everything here is so small and cramped," he says. "This building, for example, would be ten times taller in America." – Oh, of course, sir! This is a psychiatric clinic.)*	racist

Table 6. Sentiment analysis.

Joke	Cluster name
English	
Idea:A Transformers movie that can transform into a much better movie.	movie
'Why did voldermort used Twitter instead of Facebook? Because he only had followers. Not friends.	facebook
What did the Jewish man say to himself on a hot day? I should be used to being in an oven by now.	jews
What kind of bee can't be understood? A mumble bee !	bee
what do you call a dead bee?a was	bee
What happens to someone who gets attacked by bees? They get bee'd up	bee
I just tried coffee for the first time... To be honest, it wasn't my cup of tea...	coffee
What do you call a cat in love? Romeow	cat
Russian	
Вся жизнь – футбол, а ты в ней – сборная России. *(All life is football, and in it you are the Russian national team.)*	футбол *(football)*
Доктор, помогите мне. У меня проблема. Я часто ошибаюсь в людях. Я не доктор. *(Doctor, help me. I have a problem. I often make mistakes in people. I'm not a doctor.)*	доктор *(doctor)*
Поймал мужик золотую рыбку хочу машину и завод. Рыбка: хорошо но в кредит или по лизингу? Старик: так выбирай на подсолнечном или сливочном? *(A man caught a goldfish and tells it that he wants a car and a factory. Goldfish: okay, do you want to loan or to lease it? Old man: so choose oil or fat?)*	рыба *(fish)*
Я не понимаю фразу: Устал, как собака. Моя собака спит, ест и гуляет. Я бы сам от такой жизни не отказался. *(I don't understand the phrase: Tired like a dog. My dog sleeps, eats and walks. I myself would not refuse such a life.)*	собака *(dog)*
В компании со мной обычно смеются не над шуткой, а над тем, как я смеюсь над шуткой. *(People with me usually laugh not at the joke, but at the way I laugh at the joke.)*	юмор *(humour)*

Table 7. Sentiment analysis.

References

1. V. Blinov, V. Bolotova-Baranova, and P. Braslavski, *Large Dataset and Language Model Fun-Tuning for Humor Recognition*, in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2019, pp. 4027–4032.
2. J. A. Meaney, S. Wilson, L. Chiruzzo, A. Lopez, and W. Magdy, *SemEval 2021 Task 7: HaHackathon, Detecting and Rating Humor and Offense*, in: Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021), 2021, pp. 105–119.
3. D. Yang, A. Lavie, C. Dyer, and E. Hovy, *Humor Recognition and Humor Anchor Extraction*, in: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, 2015, pp. 2367–2376.
4. R. Mihalcea and C. Strapparava, *Making Computers Laugh: Investigations in Automatic Humor Recognition*, in: Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing, 2005, pp. 531–538.
5. L. Hanu and Unitary team, *Detoxify*, Github, 2020.
6. N. Reimers and I. Gurevych, *Sentence-BERT: Sentence Embeddings Using Siamese BERT-Networks*, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing, 2019, pp. 3982–3992.
7. M. Farouk, *Measuring Sentences Similarity: A Survey*, Indian Journal of Science and Technology, 2019, pp. 1–11.
8. M. Laurer, W. Van Atteveldt, A. Casas, and K. Welbers, *Less Annotating, More Classifying: Addressing the Data Scarcity Issue of Supervised Machine Learning with Deep Transfer Learning and BERT-NLI*, Political Analysis, 2024, pp. 84–100.
9. M. Grootendorst, *BERTopic: Neural Topic Modeling with a Class-Based TF-IDF Procedure*, ArXiv preprint arXiv:2203.05794 (2022).
10. N. Reimers and I. Gurevych, *Making Monolingual Sentence Embeddings Multilingual Using Knowledge Distillation*, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2020, pp. 4512–4525.
11. D. Dementieva, D. Moskovskiy, V. Logacheva, D. Dale, O. Kozlova, N. Semenov, and A. Panchenko, *Methods for Detoxification of Texts for the Russian Language*, Multimodal Technologies and Interaction, 2021.
12. Y. Kuratov and M. Arkhipov, *Adaptation of Deep Bidirectional Multilingual Transformers for Russian Language*, ArXiv preprint arXiv:1905.07213 (2019).
13. P.-Y. Chen and V.-W. Soo, *Humor Recognition Using Deep Learning*, in: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2018, pp. 113–117.
14. B. Chippada and S. Saha, *Knowledge Amalgam: Generating Jokes and Quotes Together*, ArXiv preprint arXiv:1806.04387 (2018).
15. L. Tang, A. Cai, S. Li, and J. Wang, *The Naughtyformer: A Transformer Understands Offensive Humor*, ArXiv preprint arXiv:2211.14369 (2022).

16. O. Weller and K. Seppi, *Humor Detection: A Transformer Gets the Last Laugh*, ArXiv preprint arXiv:1909.00252 (2019).
17. E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, *LoRA: Low-Rank Adaptation of Large Language Models*, ArXiv preprint arXiv:2106.09685 (2021).
18. S. Jentzsch and K. Kersting, *ChatGPT Is Fun, but It Is Not Funny! Humor Is Still Challenging Large Language Models*, in: Proceedings of the 13th Workshop on Computational Approaches to Subjectivity, Sentiment, & Social Media Analysis, 2023, pp. 325–340.
19. M. Amin and M. Burghardt, *A Survey on Approaches to Computational Humor Generation*, in: Proceedings of the 4th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature, 2020, pp. 29–41.
20. Y. Meng, M. Xia, and D. Chen, *SimPO: Simple Preference Optimization with a Reference-Free Reward*, in: Advances in Neural Information Processing Systems (NeurIPS), 2024.
21. P. Mirowski, J. Love, K. Mathewson, and S. Mohamed, *A Robot Walks into a Bar: Can Language Models Serve as Creativity Support Tools for Comedy? An Evaluation of LLMs' Humour Alignment with Comedians*, in: The 2024 ACM Conference on Fairness, Accountability, and Transparency, 2024, pp. 1622–1636.
22. P. Sunkara, *et al.*, *Adaptive Humor Generation: Enhancing Personalization with User Feedback*, ArXiv preprint arXiv:2405.07280 (2024).
23. R. Rafailov, A. Sharma, E. Mitchell, S. Ermon, C. D. Manning, and C. Finn, *Direct Preference Optimization: Your Language Model Is Secretly a Reward Model*, in: Proceedings of the 37th International Conference on Neural Information Processing Systems, 2023.
24. A. Ermilov, N. Murashkina, V. Goryacheva, and P. Braslavski, *Stierlitz Meets SVM: Humor Detection in Russian*, in: Artificial Intelligence and Natural Language — 7th International Conference, AINL 2018, 2018, pp. 178–184.
25. M. Buijzen and P. Valkenburg, *Developing a Typology of Humor in Audiovisual Media*, Media Psychology **6** (2004), 147–167.
26. S. Skalicky and S. Attardo, *Testing Humor Theory Using Word and Sentence Embeddings*, in: Proceedings of the 1st Workshop on Computational Humor (CHum), 2025, pp. 58–62.
27. H. Yamane, *Generic Joke Generation with Moral Constraints*, in: International Conference on Artificial Neural Networks, 2024, pp. 340–355.
28. H. Kariyawasam, A. Niwarthana, A. Palmer, J. Kay, and A. Withana, *Appropriate Incongruity Driven Human-AI Collaborative Tool to Assist Novices in Humorous Content Generation*, in: Proceedings of the 29th International Conference on Intelligent User Interfaces, 2024, pp. 650–659.
29. A. Tikhonov and P. Shtykovskiy, *Humor Mechanics: Advancing Humor Generation with Multistep Reasoning*, ArXiv preprint arXiv:2405.07280 (2024).

30. M. Nabil, A. Elmadany, and W. Magdy, *Humor Detection: A Survey*, Artificial Intelligence Review, 2024.
31. Y. Karyakin, O. Shapovalova, and A. Kozlov, *Machine Humor: Open Dataset and Baselines*, in: Proceedings of the Dialogue Conference, CEUR-WS, 2023.

MTS AI, Moscow, Russia
E-mail: dariagalimzianova@gmail.com

LEYA Lab, HSE University,
St.Petersburg, Russia
E-mail: sgorovaya@hse.ru
E-mail: dvikhorev@hse.ru

CAIRO, Technical University of Applied Sciences
Würzburg-Schweinfurt, Germany
E-mail: ivan.yamshchikov@thws.de

LEYA Lab, HSE University,
St. Petersburg, Russia
E-mail: ezhemchuzhina@hse.ru

Поступило 4 февраля 2026 г.