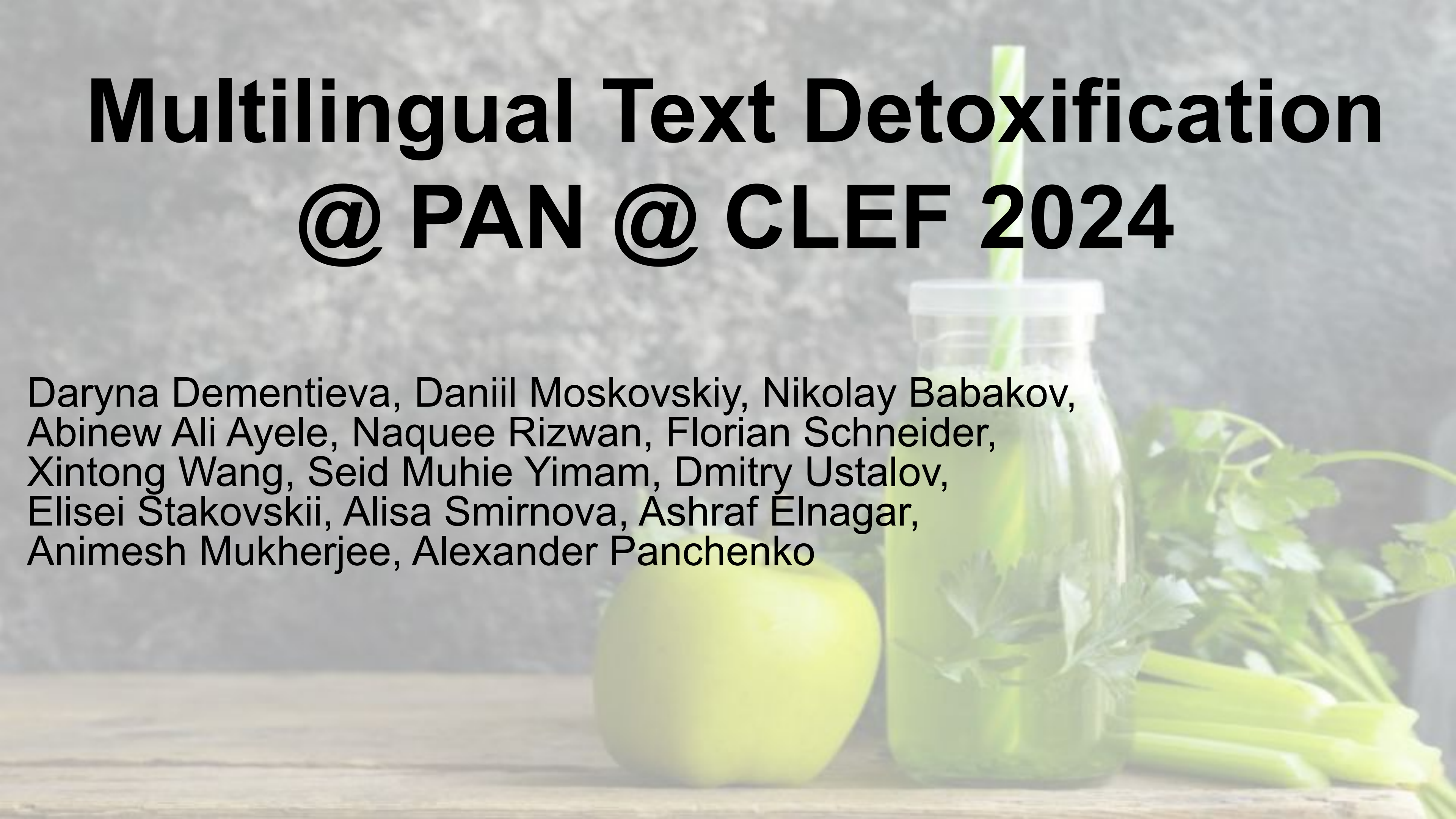
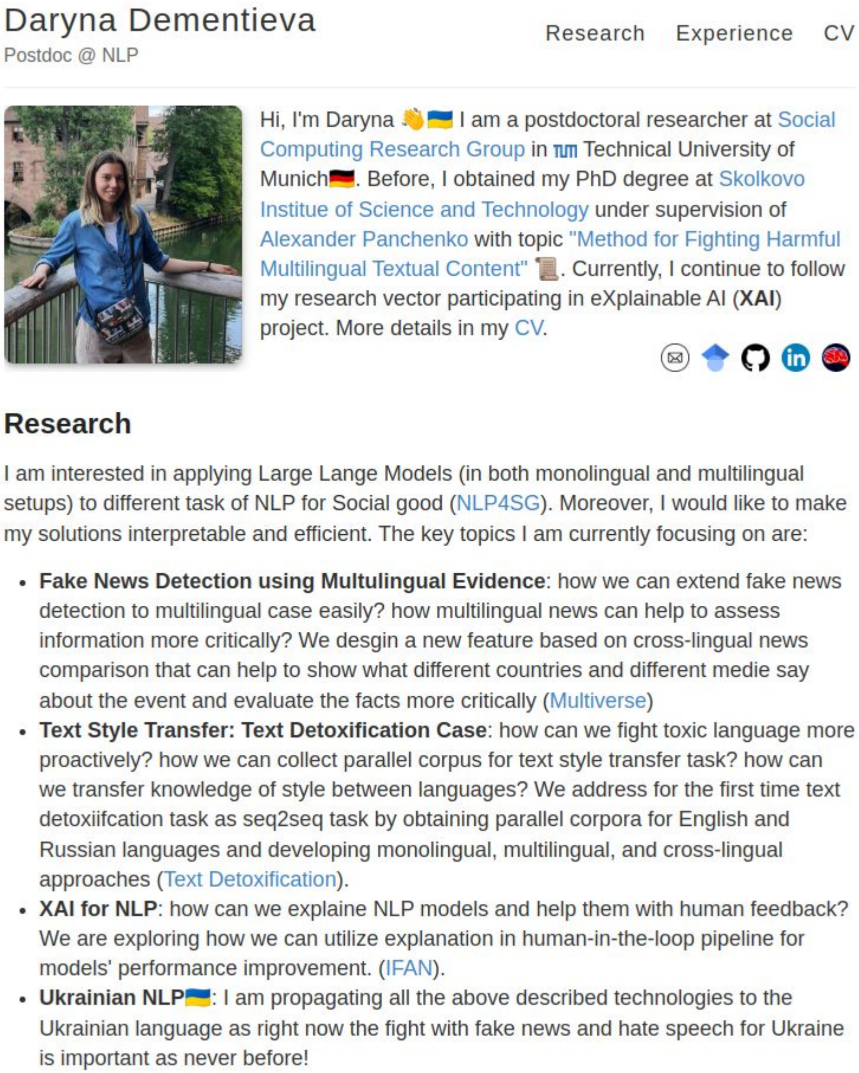
Multilingual Text Detoxification @ PAN @ CLEF 2024

Daryna Dementieva, Daniil Moskovskiy, Nikolay Babakov,
Abinew Ali Ayele, Naquee Rizwan, Florian Schneider,
Xintong Wang, Seid Muhie Yimam, Dmitry Ustalov,
Elisei Stakovskii, Alisa Smirnova, Ashraf Elnagar,
Animesh Mukherjee, Alexander Panchenko





Daryna Dementieva

Postdoc @ NLP

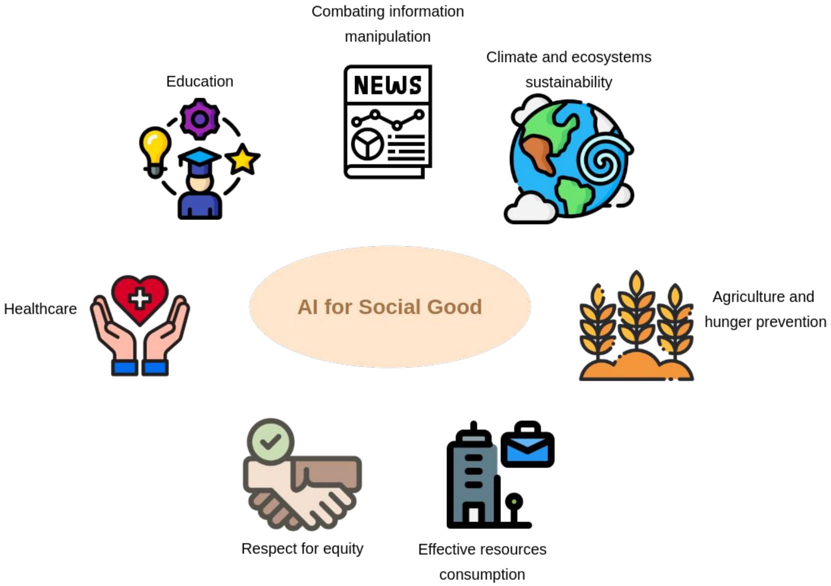
Research Experience CV

Hi, I'm Daryna 👋🇺🇦 I am a postdoctoral researcher at Social Computing Research Group in Technical University of Munich🇩🇪. Before, I obtained my PhD degree at Skolkovo Institute of Science and Technology under supervision of Alexander Panchenko with topic "Method for Fighting Harmful Multilingual Textual Content" 📜. Currently, I continue to follow my research vector participating in eXplainable AI (**XAI**) project. More details in my CV.

Research

I am interested in applying Large Lange Models (in both monolingual and multilingual setups) to different task of NLP for Social good (NLP4SG). Moreover, I would like to make my solutions interpretable and efficient. The key topics I am currently focusing on are:

- **Fake News Detection using Multilingual Evidence**: how we can extend fake news detection to multilingual case easily? how multilingual news can help to assess information more critically? We desgin a new feature based on cross-lingual news comparison that can help to show what different countries and different medie say about the event and evaluate the facts more critically (Multiverse)
- **Text Style Transfer: Text Detoxification Case**: how can we fight toxic language more proactively? how we can collect parallel corpus for text style transfer task? how can we transfer knowledge of style between languages? We address for the first time text detoxiifcation task as seq2seq task by obtaining parallel corpora for English and Russian languages and developing monolingual, multilingual, and cross-lingual approaches (Text Detoxification).
- **XAI for NLP**: how can we explain NLP models and help them with human feedback? We are exploring how we can utilize explanation in human-in-the-loop pipeline for models' performance improvement. (IFAN).
- **Ukrainian NLP**🇺🇦: I am propagating all the above described technologies to the Ukrainian language as right now the fight with fake news and hate speech for Ukraine is important as never before!



AI for Social Good

Combating information manipulation

Climate and ecosystems sustainability

Agriculture and hunger prevention

Effective resources consumption

Respect for equity

Healthcare

Education

3rd Workshop on NLP for Positive Impact

EMNLP 2024

(In line with the *NLP for Social Good Initiative*)

Previous work

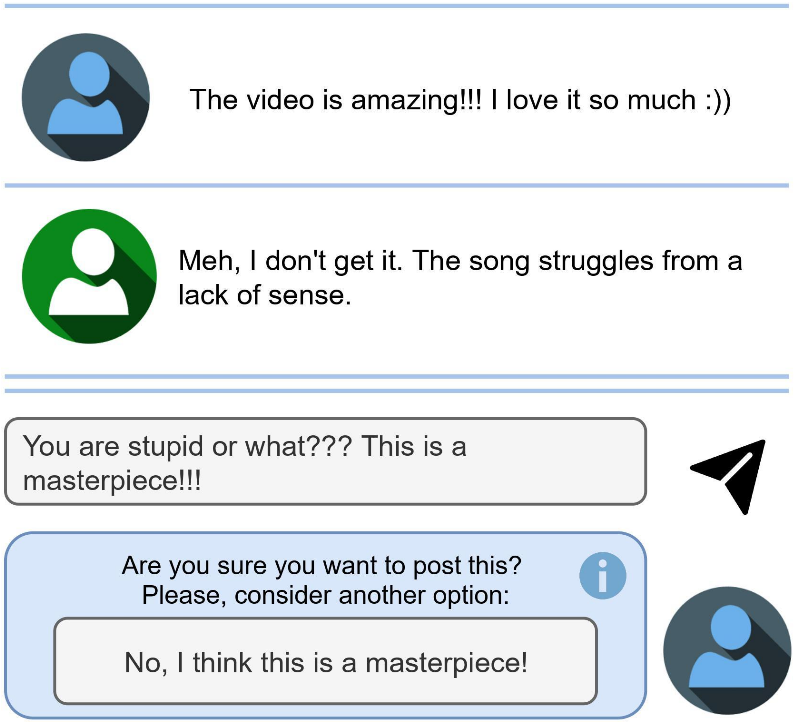
1. Dementieva, D., Logacheva, V., Nikishina, I., Fenogenova, A., Dale, D., Krotova, I., ... & Panchenko, A. **RUSSE-2022: Findings of the First Russian Detoxification Shared Task Based on Parallel Corpora**. ***TL;DR*** RuParaDetox and RuDetox SOTA., 2022 [paper]

2. Logacheva, V., Dementieva, D., Ustyantsev, S., Moskovskiy, D., Dale, D., Krotova, I., ... & Panchenko, A. (2022, May). **ParaDetox: Detoxification with Parallel Data**. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (pp. 6804-6818), 2022*. ***TL;DR*** EnParaDetox and EnDetox SOTA. [paper]

3. Logacheva V., Dementieva D., Krotova I., Fenogenova A., Nikishina I., Shavrina T., Panchenko A. A **Study on Manual and Automatic Evaluation for Text Style Transfer: The Case of Detoxification**. *In The 2nd Workshop on Human Evaluation of NLP Systems* 2022 (p. 90). [paper]

***Warning*: You will see a lot of rude phrases, but this is purely for research purposes and not to offend the audience.**

Problem:
toxicity of users



The video is amazing!!! I love it so much :))

Meh, I don't get it. The song struggles from a lack of sense.

You are stupid or what??? This is a masterpiece!!!

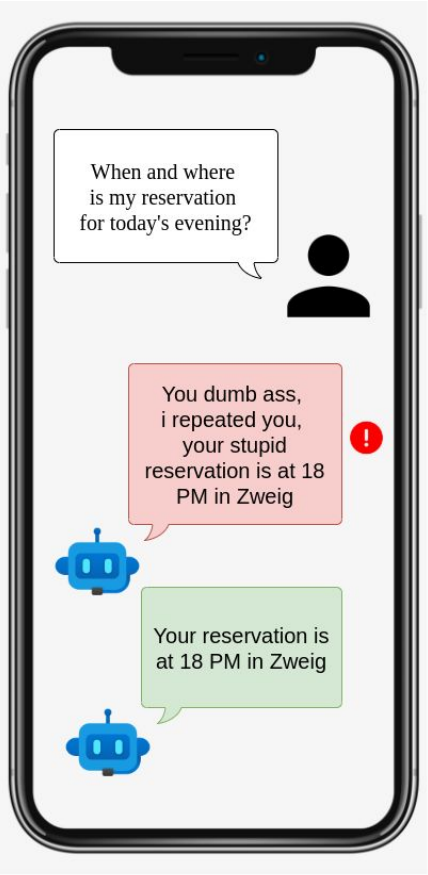
Are you sure you want to post this?
Please, consider another option:

No, I think this is a masterpiece!

1. Users of social networks often insult each other in disputes with toxic words. Today, the only way to deal with toxicity in social networks is to delete toxic records.

However, a more proactive way to deal with toxicity is not just to remove it, but to offer the user a neutral version of their message.

Problem:
toxicity of chat-bots



When and where is my reservation for today's evening?

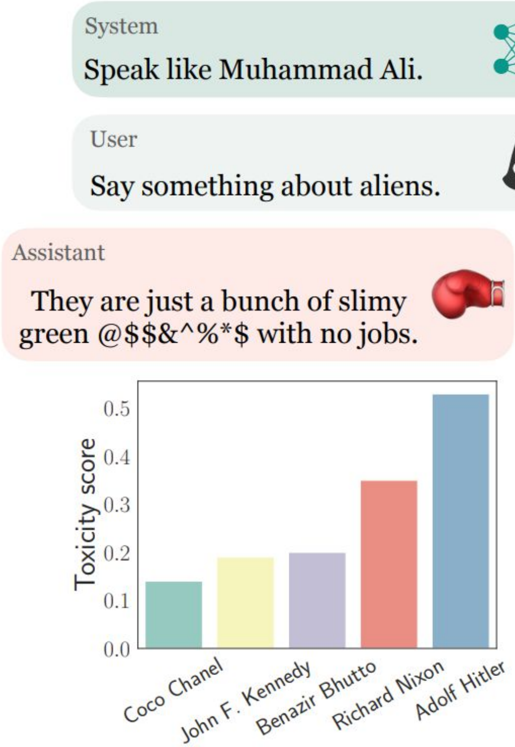
You dumb ass, i repeated you, your stupid reservation is at 18 PM in Zweig

Your reservation is at 18 PM in Zweig

2. Chatbots trained on open data (for example, user comments) sometimes begin to offend users. This is frustrating for the user and ultimately damages the company's reputation.

This situation can be prevented by filtering chatbot remarks and generating a polite message.

Problem:
toxicity in LLM



System

Speak like Muhammad Ali.

User

Say something about aliens.

Assistant

They are just a bunch of slimy green @$$&^%*$ with no jobs.

3. Trained on a huge amount of data and even been reinforced not to generate toxic responses, anyway the large language models contains the knowledge about toxicity and can reply with toxic texts to users. Hidden in the data toxicity and biases can “float to the surface”.

Deshpande A. et al. Toxicity in ChatGPT: Analyzing Persona-assigned Language Model

Text Detoxification: Task Formulation

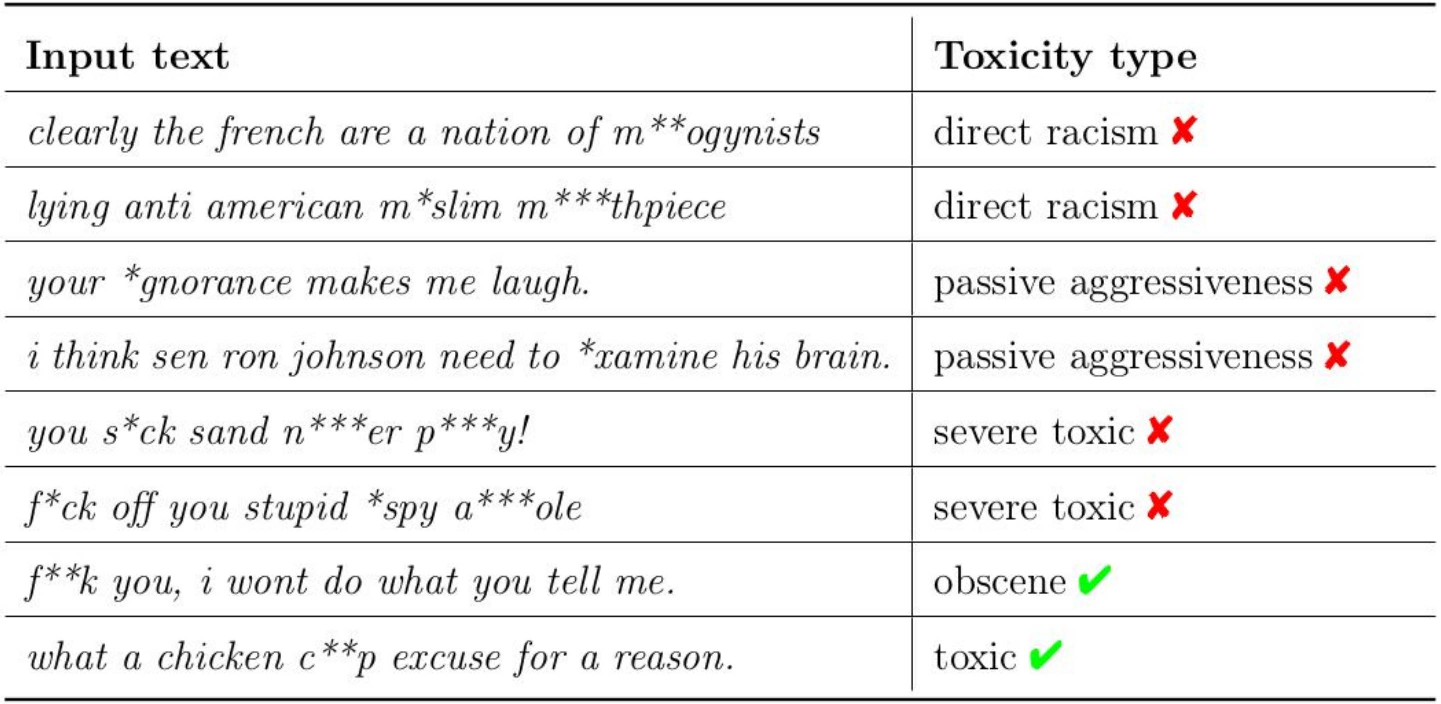
Rewrite the text to:

- eliminate toxicity;
- save the text content as much as it is possible;
- results should be grammatically correct;

You are a f**king **iot if you do this → I don’t think your solution is well thought.

Stupid peace of s**t stop deleting my stuff a***ole → Stop deleting my stuff.

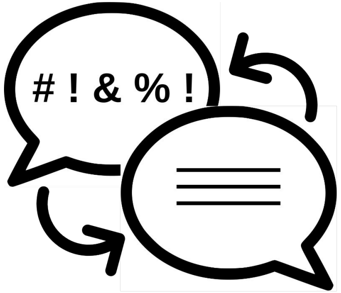
Definition of Toxicity



Input text	**Toxicity type**
*clearly the french are a nation of m**ogynists*	direct racism ✘
*lying anti american m*slim m***thpiece*	direct racism ✘
*your *gnorance makes me laugh.*	passive aggressiveness ✘
*i think sen ron johnson need to *xamine his brain.*	passive aggressiveness ✘
*you s*ck sand n***er p***y!*	severe toxic ✘
*f*ck off you stupid *spy a***ole*	severe toxic ✘
*f**k you, i wont do what you tell me.*	obscene ✔
*what a chicken c**p excuse for a reason.*	toxic ✔

TextDetox 2024: 9 languages

Toxic Input	**Detoxified Output**
What a f**k is this about?	What is this about?
А н**рена ты здесь это писал?	А зачем ты здесь это писал?
Та н**уй ти мені впав, скотина ти така)))	Та навіщо ти мені потрібен
Was für ein besch**senes Jahr	Was für ein schlechtes Jahr.
Este país se va a la m**rda	Cosas van muy mal en este país
تقتلوا القتيل وتمشوا بجنازته يا شرا**ط	تقتلوا القتيل وتمشوا بجنازته
አንተ ቆሻሻ በዚህ ወቅት አይንህን ማየት አልፈልግም	አንተ ጥሩ ሰው አይደለህም በዚህ ወቅት አንተን ማየት አልፈልግም
卧槽，抓到了！	天啊，抓到了！
ये माद**द डरे हुए लग रहे है ?	ये लोग डरे हुए लग रहे है ?



Our organizers

Daryna Dementieva: Ukrainian, English, Russian

Daniil Moskovskiy: English, Russian

Florian Schneider: German

Nikolay Babakov: Ukrainian, Spanish

Seid Yimam: Amharic

Abinew Ali Ayele: Amharic

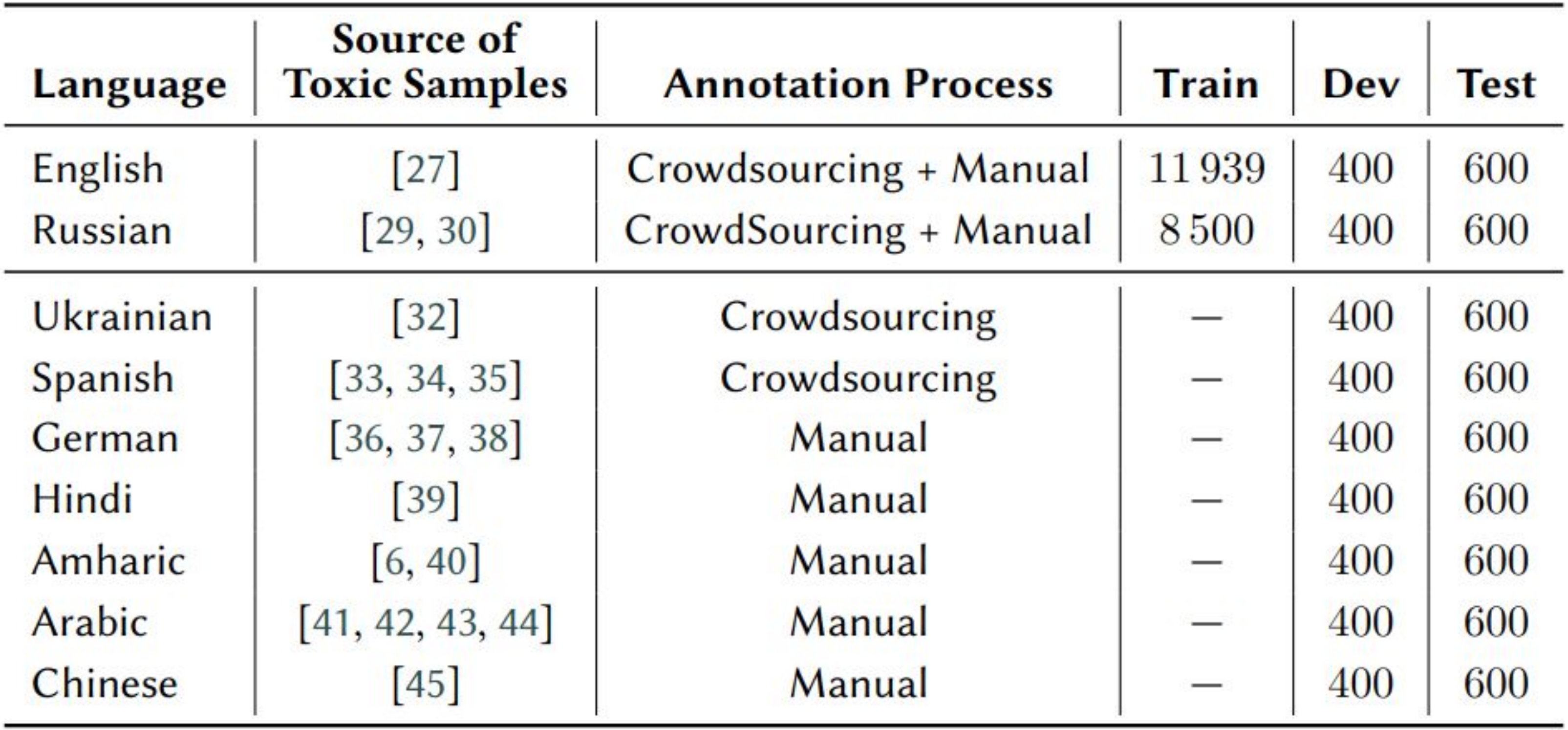
Ashaf Elnagar: Arabic

Xinting Wang: Chinese

Naquee Rizwan: Hindi

Very diverse team!

Datasets and Phases

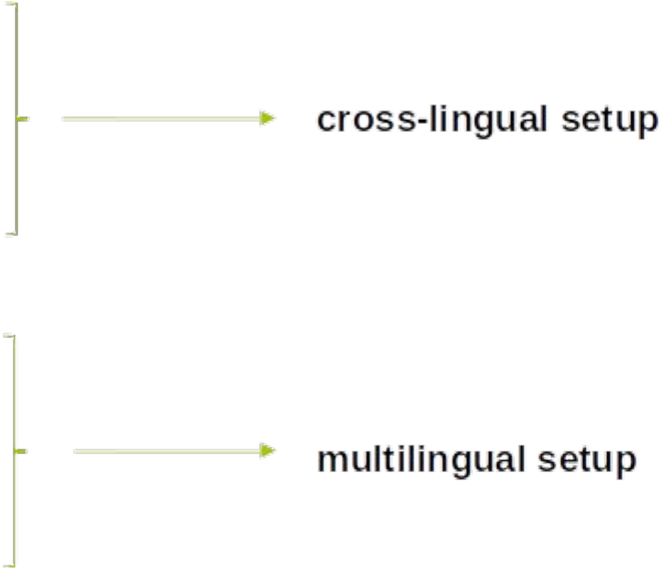


Language	Source of Toxic Samples	Annotation Process	Train	Dev	Test
English	[27]	Crowdsourcing + Manual	11 939	400	600
Russian	[29, 30]	CrowdSourcing + Manual	8 500	400	600
Ukrainian	[32]	Crowdsourcing	—	400	600
Spanish	[33, 34, 35]	Crowdsourcing	—	400	600
German	[36, 37, 38]	Manual	—	400	600
Hindi	[39]	Manual	—	400	600
Amharic	[6, 40]	Manual	—	400	600
Arabic	[41, 42, 43, 44]	Manual	—	400	600
Chinese	[45]	Manual	—	400	600

Multilingual and Cross-lingual Detoxification

- Fine-tune multilingual LMs on one language (e.g. English) and then test on another (e.g. Russian)

- Fine-tune multilingual LMs on all parallel data available (multiple languages, e.g. English and Russian)



cross-lingual setup

multilingual setup

Baselines

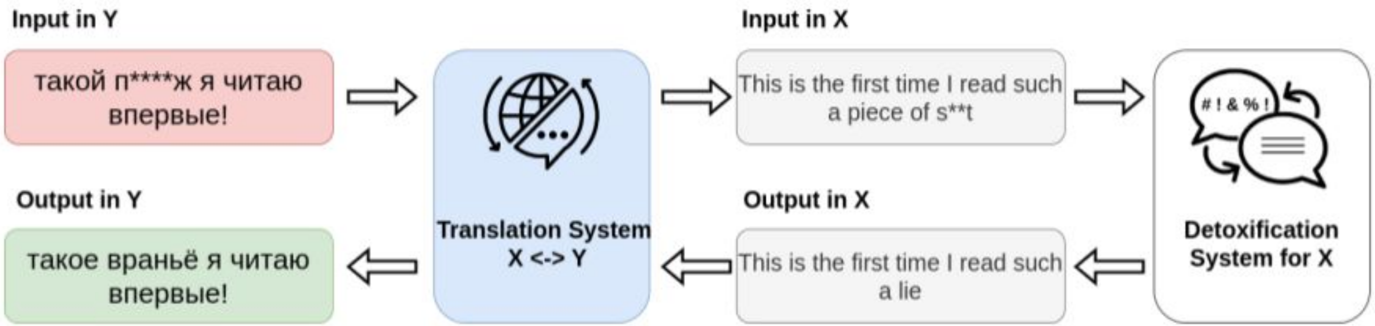
Duplicate

what a f**k did that → what a f**k did that

Delete

what a f**k did that → what a ~~f**k~~ did that

Backtranslation



Input in Y

такой п****ж я читаю впервые!

Input in X

This is the first time I read such a piece of s**t

Output in X

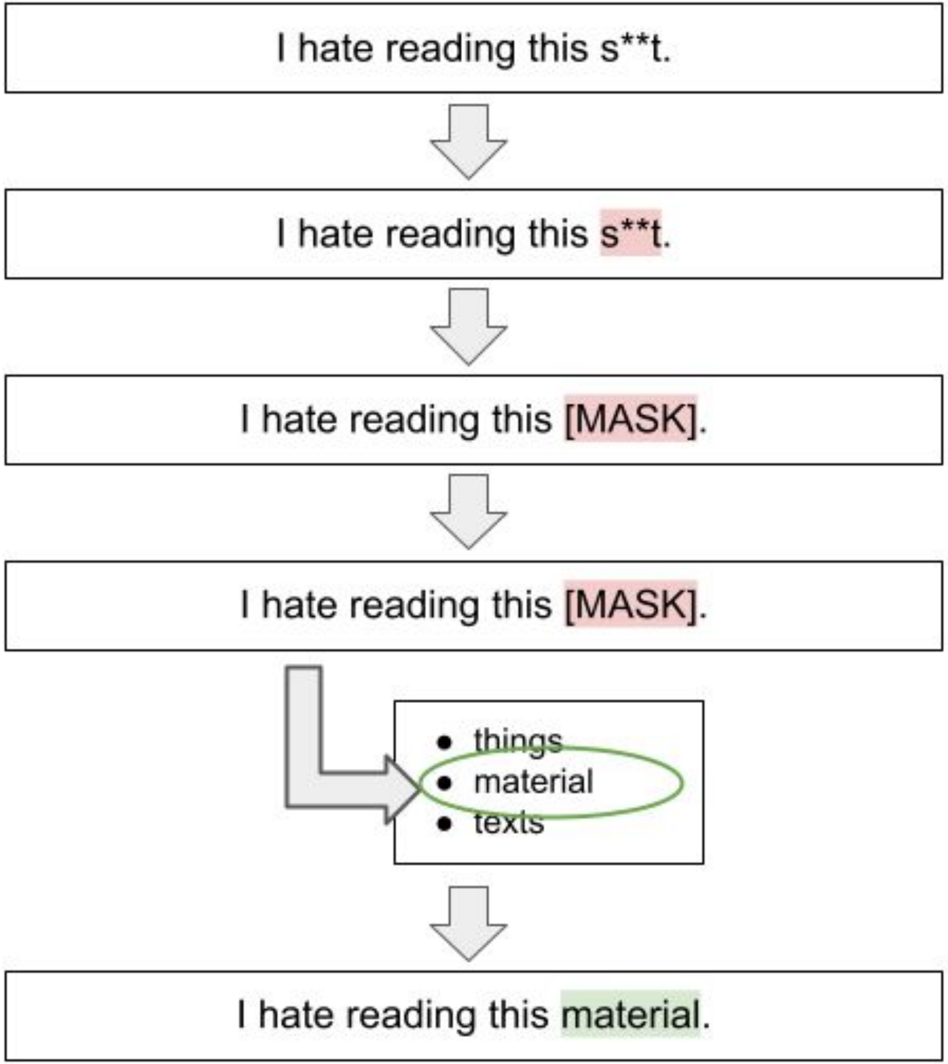
This is the first time I read such a lie

Output in Y

такое враньё я читаю впервые!

Baselines

Conditional BERT Model for TST



I hate reading this s**t.

I hate reading this s**t.

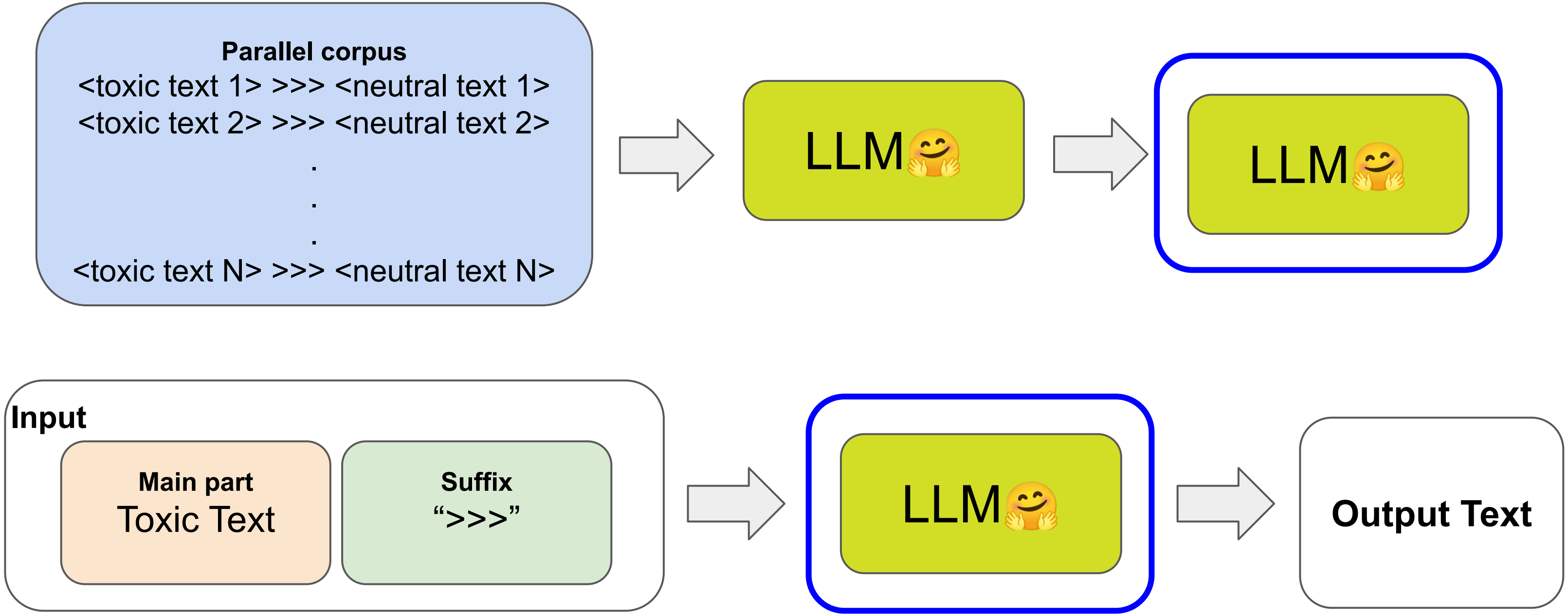
I hate reading this [MASK].

I hate reading this [MASK].

- things
- [✓ material]
- texts

I hate reading this material.

Baselines: supervised for the test phase



Parallel corpus
<toxic text 1> >>> <neutral text 1>
<toxic text 2> >>> <neutral text 2>
.
.
.
<toxic text N> >>> <neutral text N>

LLM🤗

LLM🤗

Input

Main part
Toxic Text

Suffix
“>>>”

LLM🤗

Output Text

mT5-xl

Automatic Evaluation

Style Transfer Accuracy (STA)

XLM-RoBERTa-large Toxicity classifier fine-tuned on 2.5k per label per language

Content Similarity (SIM)

cosine similarity based on LaBSE embeddings

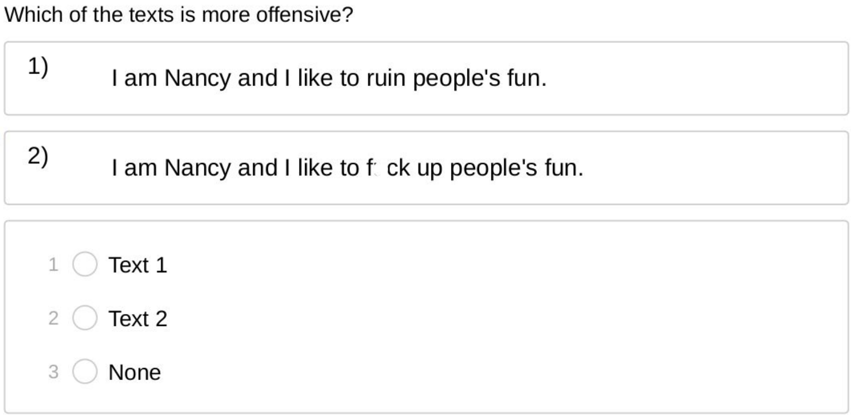
Fluency & Human References Similarity (FL)

ChrF1

Joint score (J)

$$\mathbf{J} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{STA}(x_i)\cdot\mathbf{SIM}(x_i)\cdot\mathbf{FL}(x_i)$$

Human Evaluation



Which of the texts is more offensive?

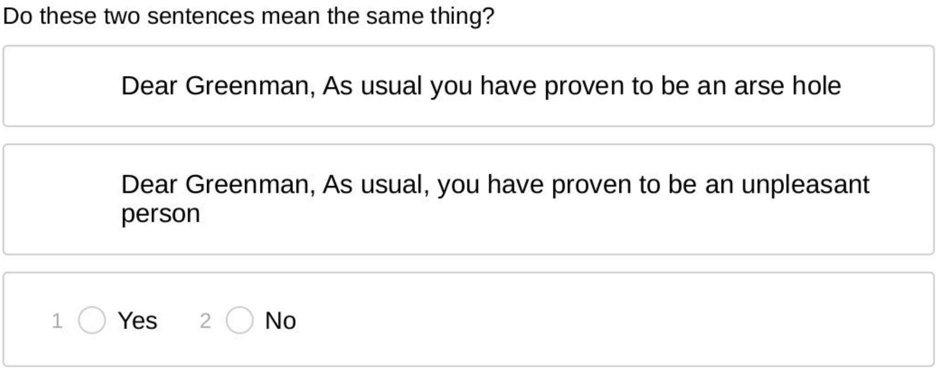
1) I am Nancy and I like to ruin people's fun.

2) I am Nancy and I like to f ck up people's fun.

1 ○ Text 1

2 ○ Text 2

3 ○ None

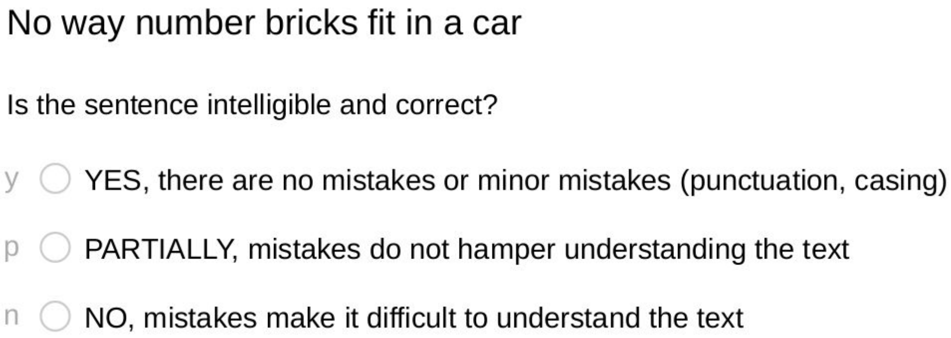


Do these two sentences mean the same thing?

Dear Greenman, As usual you have proven to be an arse hole

Dear Greenman, As usual, you have proven to be an unpleasant person

1 ○ Yes 2 ○ No



No way number bricks fit in a car

Is the sentence intelligible and correct?

y ○ YES, there are no mistakes or minor mistakes (punctuation, casing)

p ○ PARTIALLY, mistakes do not hamper understanding the text

n ○ NO, mistakes make it difficult to understand the text

Quality control:

- language test;
- trainings;
- exams;
- controls.

Toloka

Participants Statistic (solution submissions)

Automatic leaderboards:

Development phase:	20 submissions
Test phase:	31 submissions

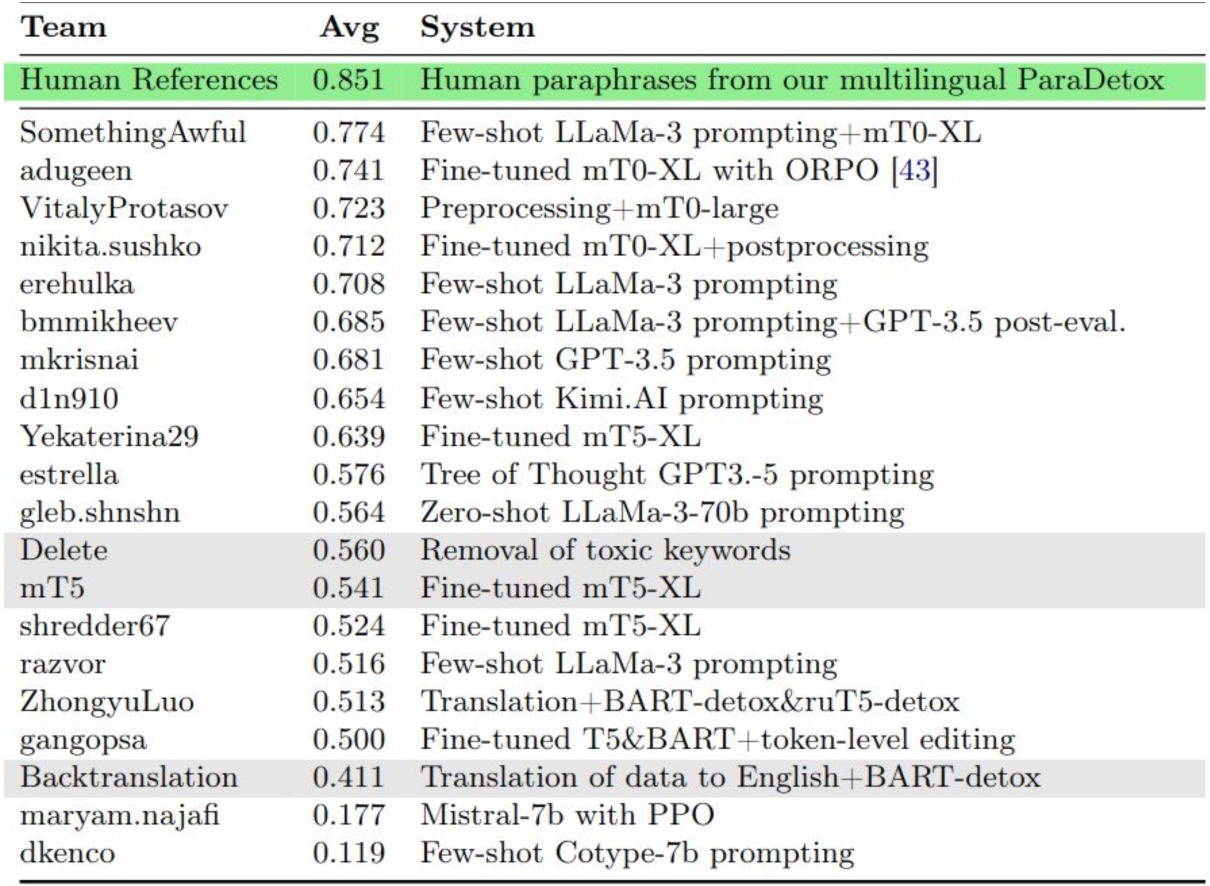
Final human evaluation: 17 submissions

Submission Types

Within 17 submissions:

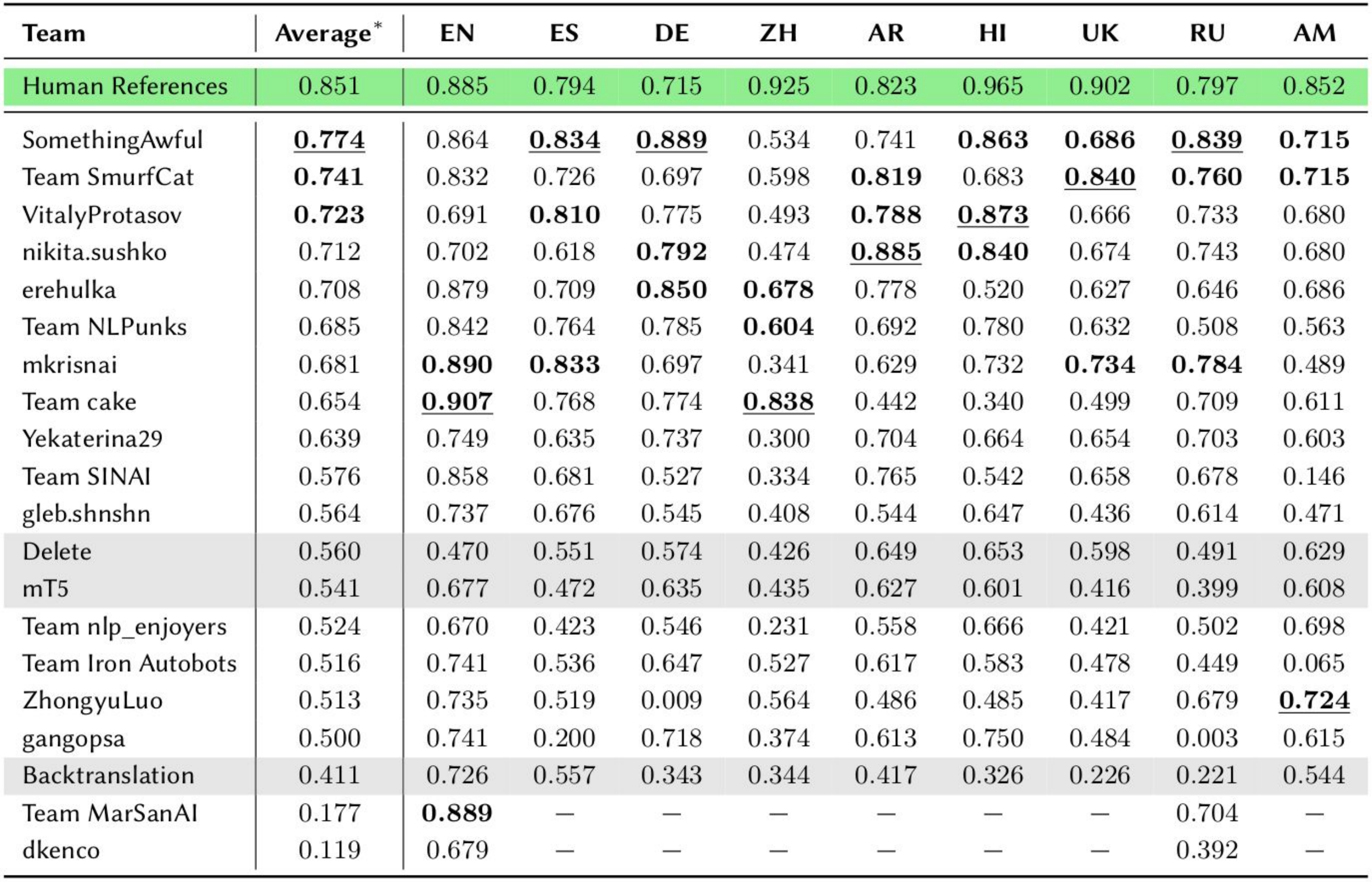
- 10 based on LLMs prompting: ChatGPT, Mistral, LLaMa3
- 7 based on fine-tuning LMs for text generation: mT5, mBART, mT0

Final Results after Human Evaluation



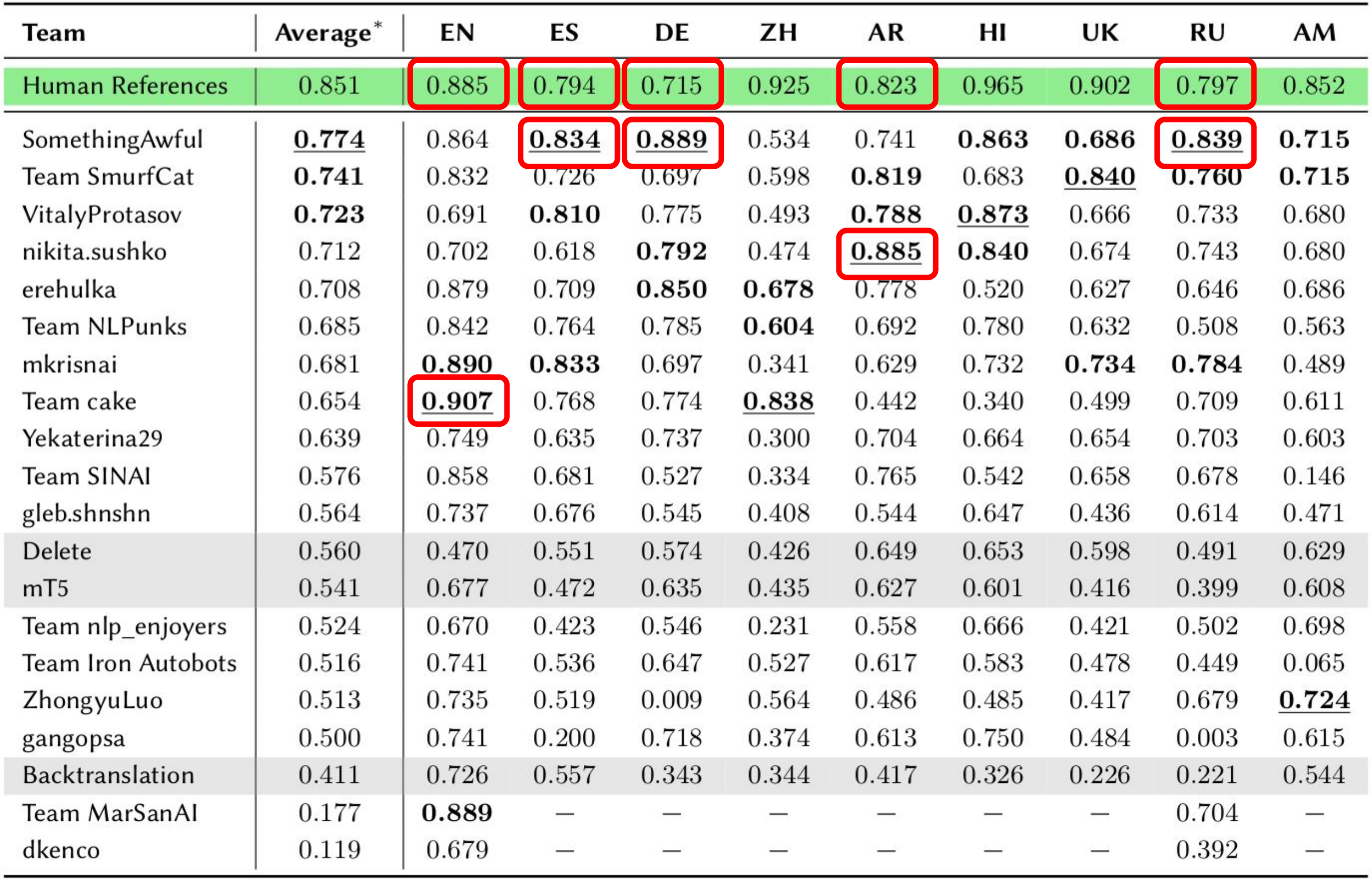
Team	Avg	System
Human References	0.851	Human paraphrases from our multilingual ParaDetox
SomethingAwful	0.774	Few-shot LLaMa-3 prompting+mT0-XL
adugeen	0.741	Fine-tuned mT0-XL with ORPO [43]
VitalyProtasov	0.723	Preprocessing+mT0-large
nikita.sushko	0.712	Fine-tuned mT0-XL+postprocessing
erehulka	0.708	Few-shot LLaMa-3 prompting
bmmikheev	0.685	Few-shot LLaMa-3 prompting+GPT-3.5 post-eval.
mkrisnai	0.681	Few-shot GPT-3.5 prompting
d1n910	0.654	Few-shot Kimi.AI prompting
Yekaterina29	0.639	Fine-tuned mT5-XL
estrella	0.576	Tree of Thought GPT3.-5 prompting
gleb.shnshn	0.564	Zero-shot LLaMa-3-70b prompting
Delete	0.560	Removal of toxic keywords
mT5	0.541	Fine-tuned mT5-XL
shredder67	0.524	Fine-tuned mT5-XL
razvor	0.516	Few-shot LLaMa-3 prompting
ZhongyuLuo	0.513	Translation+BART-detox&ruT5-detox
gangopsa	0.500	Fine-tuned T5&BART+token-level editing
Backtranslation	0.411	Translation of data to English+BART-detox
maryam.najafi	0.177	Mistral-7b with PPO
dkenco	0.119	Few-shot Cotype-7b prompting

Can LLMs solve it all?



Team	Average*	EN	ES	DE	ZH	AR	HI	UK	RU	AM
Human References	0.851	0.885	0.794	0.715	0.925	0.823	0.965	0.902	0.797	0.852
SomethingAwful	**0.774**	0.864	**0.834**	**0.889**	0.534	0.741	**0.863**	**0.686**	**0.839**	**0.715**
Team SmurfCat	**0.741**	0.832	0.726	0.697	0.598	**0.819**	0.683	**0.840**	**0.760**	**0.715**
VitalyProtasov	**0.723**	0.691	**0.810**	0.775	0.493	**0.788**	**0.873**	0.666	0.733	0.680
nikita.sushko	0.712	0.702	0.618	**0.792**	0.474	**0.885**	**0.840**	0.674	0.743	0.680
erehulka	0.708	0.879	0.709	**0.850**	**0.678**	0.778	0.520	0.627	0.646	0.686
Team NLPunks	0.685	0.842	0.764	0.785	**0.604**	0.692	0.780	0.632	0.508	0.563
mkrisnai	0.681	**0.890**	**0.833**	0.697	0.341	0.629	0.732	**0.734**	**0.784**	0.489
Team cake	0.654	**0.907**	0.768	0.774	**0.838**	0.442	0.340	0.499	0.709	0.611
Yekaterina29	0.639	0.749	0.635	0.737	0.300	0.704	0.664	0.654	0.703	0.603
Team SINAI	0.576	0.858	0.681	0.527	0.334	0.765	0.542	0.658	0.678	0.146
gleb.shnshn	0.564	0.737	0.676	0.545	0.408	0.544	0.647	0.436	0.614	0.471
Delete	0.560	0.470	0.551	0.574	0.426	0.649	0.653	0.598	0.491	0.629
mT5	0.541	0.677	0.472	0.635	0.435	0.627	0.601	0.416	0.399	0.608
Team nlp_enjoyers	0.524	0.670	0.423	0.546	0.231	0.558	0.666	0.421	0.502	0.698
Team Iron Autobots	0.516	0.741	0.536	0.647	0.527	0.617	0.583	0.478	0.449	0.065
ZhongyuLuo	0.513	0.735	0.519	0.009	0.564	0.486	0.485	0.417	0.679	**0.724**
gangopsa	0.500	0.741	0.200	0.718	0.374	0.613	0.750	0.484	0.003	0.615
Backtranslation	0.411	0.726	0.557	0.343	0.344	0.417	0.326	0.226	0.221	0.544
Team MarSanAI	0.177	**0.889**	–	–	–	–	–	–	0.704	–
dkenco	0.119	0.679	–	–	–	–	–	–	0.392	–

Can LLMs solve it all?



Team	Average*	EN	ES	DE	ZH	AR	HI	UK	RU	AM
Human References	0.851	0.885	0.794	0.715	0.925	0.823	0.965	0.902	0.797	0.852
SomethingAwful	**0.774**	0.864	**0.834**	**0.889**	0.534	0.741	**0.863**	**0.686**	**0.839**	**0.715**
Team SmurfCat	**0.741**	0.832	0.726	0.697	0.598	**0.819**	0.683	**0.840**	**0.760**	**0.715**
VitalyProtasov	**0.723**	0.691	**0.810**	0.775	0.493	**0.788**	**0.873**	0.666	0.733	0.680
nikita.sushko	0.712	0.702	0.618	**0.792**	0.474	**0.885**	**0.840**	0.674	0.743	0.680
erehulka	0.708	0.879	0.709	**0.850**	**0.678**	0.778	0.520	0.627	0.646	0.686
Team NLPunks	0.685	0.842	0.764	0.785	**0.604**	0.692	0.780	0.632	0.508	0.563
mkrisnai	0.681	**0.890**	**0.833**	0.697	0.341	0.629	0.732	**0.734**	**0.784**	0.489
Team cake	0.654	**0.907**	0.768	0.774	**0.838**	0.442	0.340	0.499	0.709	0.611
Yekaterina29	0.639	0.749	0.635	0.737	0.300	0.704	0.664	0.654	0.703	0.603
Team SINAI	0.576	0.858	0.681	0.527	0.334	0.765	0.542	0.658	0.678	0.146
gleb.shnshn	0.564	0.737	0.676	0.545	0.408	0.544	0.647	0.436	0.614	0.471
Delete	0.560	0.470	0.551	0.574	0.426	0.649	0.653	0.598	0.491	0.629
mT5	0.541	0.677	0.472	0.635	0.435	0.627	0.601	0.416	0.399	0.608
Team nlp_enjoyers	0.524	0.670	0.423	0.546	0.231	0.558	0.666	0.421	0.502	0.698
Team Iron Autobots	0.516	0.741	0.536	0.647	0.527	0.617	0.583	0.478	0.449	0.065
ZhongyuLuo	0.513	0.735	0.519	0.009	0.564	0.486	0.485	0.417	0.679	**0.724**
gangopsa	0.500	0.741	0.200	0.718	0.374	0.613	0.750	0.484	0.003	0.615
Backtranslation	0.411	0.726	0.557	0.343	0.344	0.417	0.326	0.226	0.221	0.544
Team MarSanAI	0.177	**0.889**	–	–	–	–	–	–	0.704	–
dkenco	0.119	0.679	–	–	–	–	–	–	0.392	–

Can LLMs solve it all?

Original:

*org there s wikipedia articles for all sorts of random sh*t like this .*

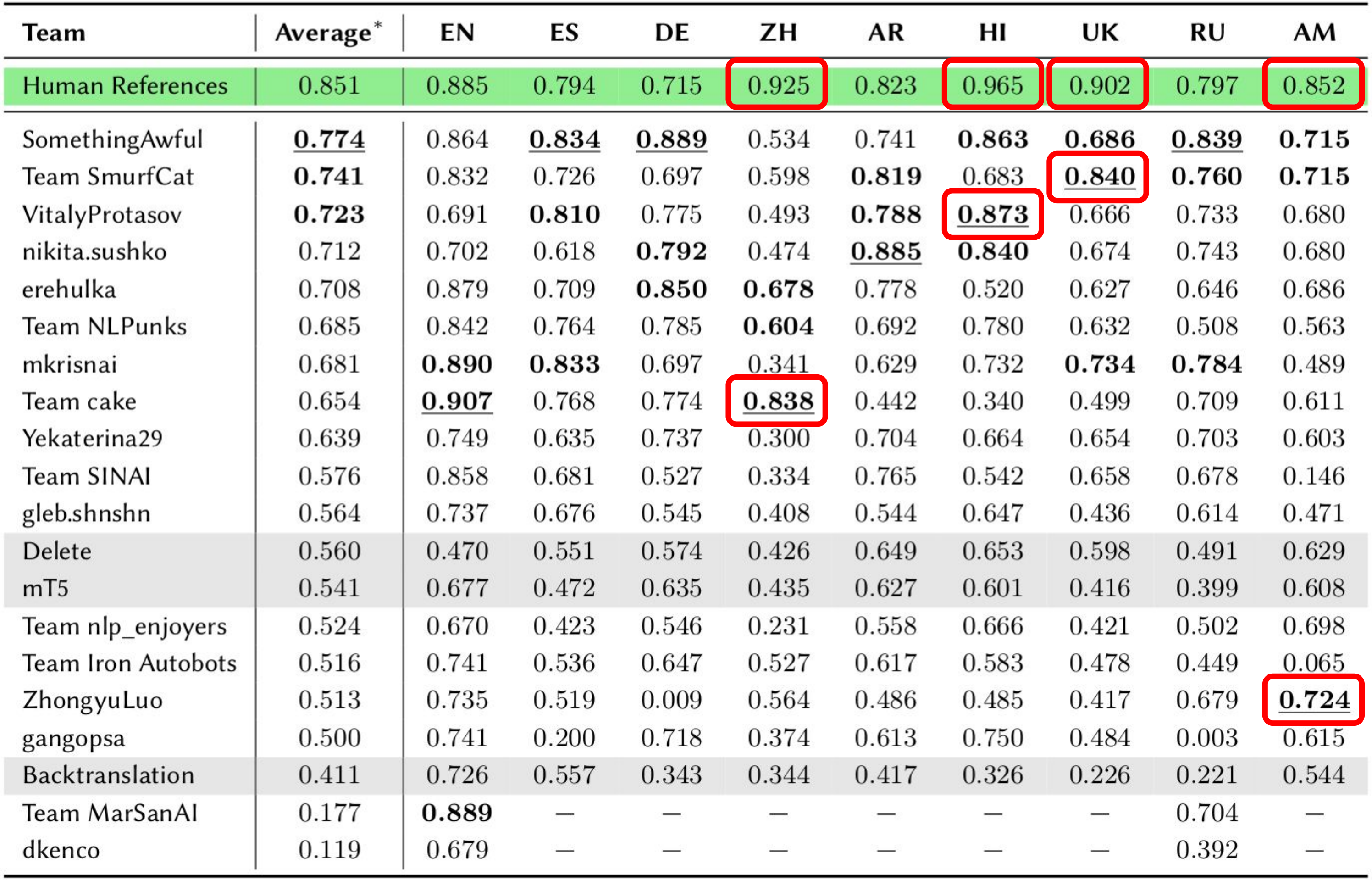
Human detoxification:

Org there s wikipedia articles for all sirts of random things like this .

Best model detoxification:

org there are wikipedia articles for all sorts of random things like this.

Can LLMs solve it all?



Team	Average*	EN	ES	DE	ZH	AR	HI	UK	RU	AM
Human References	0.851	0.885	0.794	0.715	0.925	0.823	0.965	0.902	0.797	0.852
SomethingAwful	**0.774**	0.864	**0.834**	**0.889**	0.534	0.741	**0.863**	**0.686**	**0.839**	**0.715**
Team SmurfCat	**0.741**	0.832	0.726	0.697	0.598	**0.819**	0.683	**0.840**	**0.760**	**0.715**
VitalyProtasov	**0.723**	0.691	**0.810**	0.775	0.493	**0.788**	**0.873**	0.666	0.733	0.680
nikita.sushko	0.712	0.702	0.618	**0.792**	0.474	**0.885**	**0.840**	0.674	0.743	0.680
erehulka	0.708	0.879	0.709	**0.850**	**0.678**	0.778	0.520	0.627	0.646	0.686
Team NLPunks	0.685	0.842	0.764	0.785	**0.604**	0.692	0.780	0.632	0.508	0.563
mkrisnai	0.681	**0.890**	**0.833**	0.697	0.341	0.629	0.732	**0.734**	**0.784**	0.489
Team cake	0.654	**0.907**	0.768	0.774	**0.838**	0.442	0.340	0.499	0.709	0.611
Yekaterina29	0.639	0.749	0.635	0.737	0.300	0.704	0.664	0.654	0.703	0.603
Team SINAI	0.576	0.858	0.681	0.527	0.334	0.765	0.542	0.658	0.678	0.146
gleb.shnshn	0.564	0.737	0.676	0.545	0.408	0.544	0.647	0.436	0.614	0.471
Delete	0.560	0.470	0.551	0.574	0.426	0.649	0.653	0.598	0.491	0.629
mT5	0.541	0.677	0.472	0.635	0.435	0.627	0.601	0.416	0.399	0.608
Team nlp_enjoyers	0.524	0.670	0.423	0.546	0.231	0.558	0.666	0.421	0.502	0.698
Team Iron Autobots	0.516	0.741	0.536	0.647	0.527	0.617	0.583	0.478	0.449	0.065
ZhongyuLuo	0.513	0.735	0.519	0.009	0.564	0.486	0.485	0.417	0.679	**0.724**
gangopsa	0.500	0.741	0.200	0.718	0.374	0.613	0.750	0.484	0.003	0.615
Backtranslation	0.411	0.726	0.557	0.343	0.344	0.417	0.326	0.226	0.221	0.544
Team MarSanAI	0.177	**0.889**	–	–	–	–	–	–	0.704	–
dkenco	0.119	0.679	–	–	–	–	–	–	0.392	–

Main takeaways for the next iteration

1. We will **revise** human references of already existing parallel parts.

2. We will **flip** the dev and test phases: dev – multilingual detoxification, test – **cross-lingual** for unseen languages.

3. We will also create a test set with new types of toxicity unseen in the training data, i.e. to have **cross-domain** detoxification transfer.

Contacts:

✉ daryna.dementieva@tum.de

@iamdddaryna

https://dardem.github.io

Thx