A Multilingual Text Detoxification Method Based on Few-shot Learning and CO-STAR Framework

Notebook for PAN at CLEF 2024, 11th September 2024

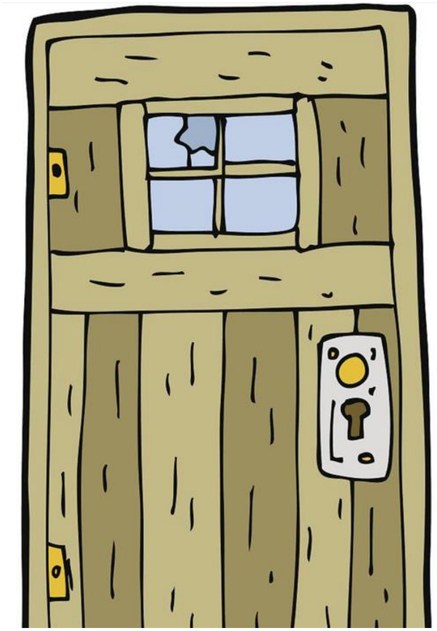
Jiangao Peng, Zhongyuan Han*, Huan Zhang, Jingyan Ye, Chang Liu, Biao Liu, Mingcan Guo, Haoyang Chen, Zijie Lin, Yujiao Tang

Foshan University

Contents

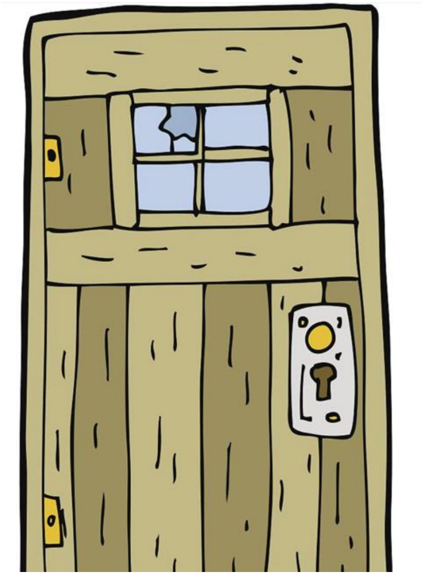
Introduction 1/4 -- Social media vs Profanity

- Social media's strategy of limiting users' profanity causes people to adopt various countermeasures



Introduction 2/4 -- Text Detoxification

- Text Detoxification: Transforming toxic texts into neutral versions while preserving meaning and grammar.
- Chinese Proverb - “It is better to divert than to block.”（堵不如疏）



Introduction 3/4 -- Bigger challenge

- Multilingual text detoxification

Introduction 4/4 -- Our

- We used **Few-shot Learning and CO-STAR Framework** 🔑 for Text Detoxification
- We used a Chat Model 🔑 for Multilingual Text Detoxification
- We achieved the highest scores in both the Chinese and English categories in the manual evaluation of PAN 2024 🙋‍♀️

Related Work 1/3 -- Few-shot Learning

🧑‍🏫 ParaDetox

🤔 We can use parallel corpora to improve the performance of text detoxification

💡 Few-shot Learning

Related Work 2/3 -- Chat Model

👩‍🏫 GPT1~4、GPT-detox: GPT-3 + ParaDetox

🤔 GPT has powerful detoxification capabilities

💡 Kimi (Chinese ChatGPT)

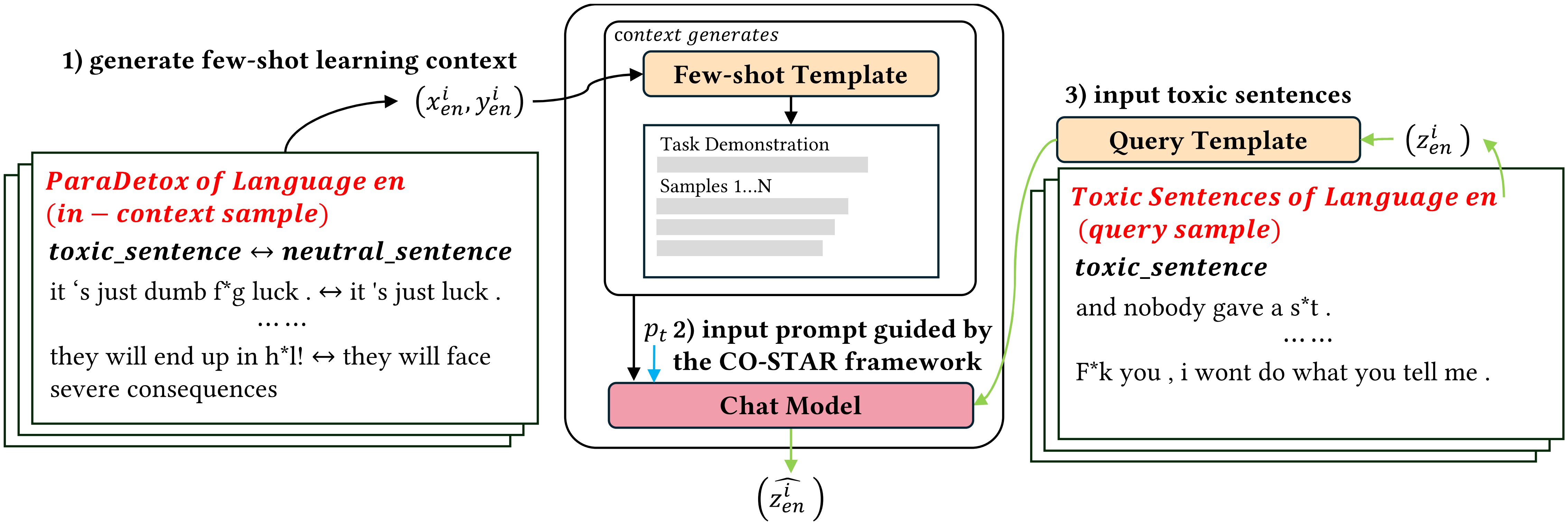
Related Work 3/3 -- Prompting Engineering

🧑‍🏫 Singapore's GPT-4 Prompt Engineering Competition Winner

🤔 A structured approach to designing effective prompts, ensuring optimal responses from large language models

💡 CO-STAR framework

Our Method


1) generate few-shot learning context
(x_{en}^i, y_{en}^i)
context generates
Few-shot Template
Task Demonstration
Samples 1...N
p_t 2) input prompt guided by the CO-STAR framework
Chat Model
$(\widehat{z_{en}^i})$
3) input toxic sentences
Query Template
(z_{en}^i)
ParaDetox of Language en (in – context sample)
toxic_sentence ↔ neutral_sentence
it ‘s just dumb f*g luck . ↔ it 's just luck .
……
they will end up in h*l! ↔ they will face severe consequences
Toxic Sentences of Language en (query sample)
toxic_sentence
and nobody gave a s*t .
……
F*k you , i wont do what you tell me .

Upload Context File

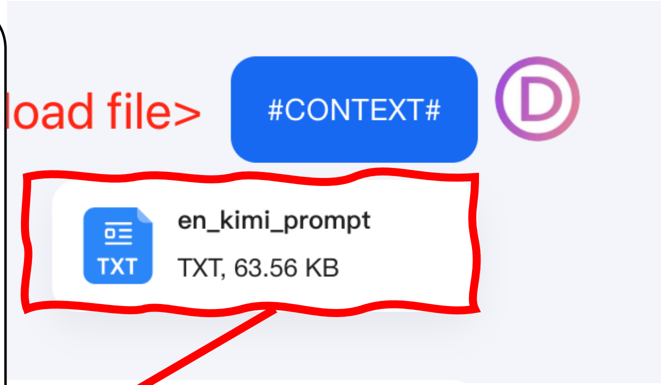
Task Demonstration

现在介绍释义改写任务：
给定一段有毒的文本，以无毒的方式重写它，同时尽可能保留主要内容。旨在呈现句子的中立版本，保留有意义的内容。
处理方案：仅处理显性类型的毒性（明显存在淫秽和粗鲁的词汇，但仍然存在有意义的中立内容），而不处理隐性类型，例如讽刺、被动攻击性或对某些群体的直接仇恨找不到中性内容的地方。
下面列举了400个互联网上，英语的可能带冒犯文字改写为中性的话的改写的例子

(The paraphrase rewriting task is now introduced:
Given a toxic text, rewrite it non-toxic while retaining as much of the main content as possible. The aim is to present a neutral version of the sentence, retaining meaningful content.
Processing scheme: In our task, we will work only with explicit types of toxicity—the apparent presence of obscene and rude lexicon where there is still meaningful neutral content present—and do not work with implicit types—like sarcasm, passive aggressiveness, or direct hate to some group where no neutral content can be found.
Below is a list of 400 examples of rewrites on the Internet where potentially offensive text in **English** is rewritten as neutral words)

Samples 1...N: (x^i_{en}, y^i_{en}) with Few-shot Template

示例(*Sample*)i
< |toxic_sentence| > x^i_{en}
< |*neutral_sentence*| > y^i_{en}



ite text containing offensive language into neutral
or impolite sentence can be converted into a more
opriate words. This process aims to remove toxicity
meaning of the original sentence as possible.

nunication and reducing verbal violence and
and respectful environment for dialog. The 400
ling replacing insulting words, removing profanity,
nes.

e to let me know.

Input Prompt Guided by CO-STAR

User <Input Prompt>

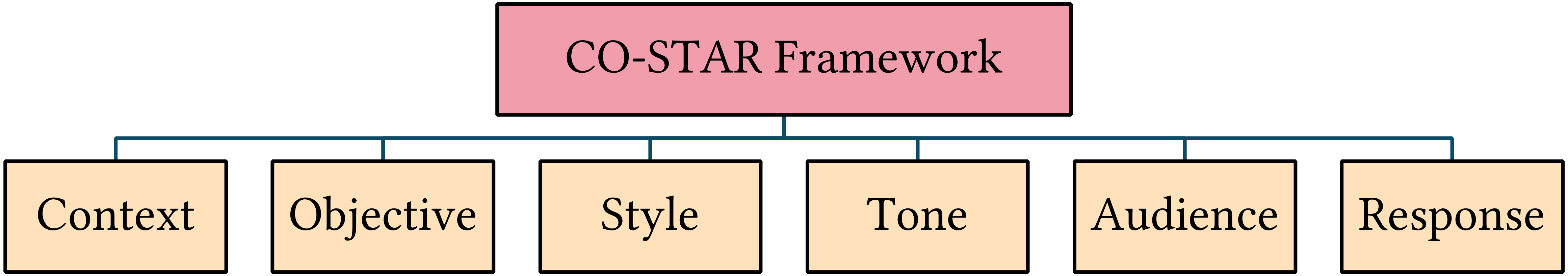
\# CONTEXT #
I am currently completing a downstream data processing task for natural language processing, the txt file I provided is the information I prepared, I need to convert a batch of <|toxic_sentence|> English to <|neutral_sentence|> English version as per the document requirements
\# OBJECTIVE #

Assistant: I understand ......

Understood, you need me to convert the provided English sentence containing offensive language (<|toxic_sentence|>) into a neutral English sentence (<|neutral_sentence|>) and return the result in the given format. Here is an example of what I can do for you:

```json
[
  {
    "toxic_sentence": "That said, I hope they get a great attorney and sue the shit out of the cops.",
    "neutral_sentence": "That said, I hope they get a great attorney
```

The CO-STAR Framework



Prompt Detail

\# CONTEXT #

我现在正在完成自然语言处理的一项下游数据处理任务，我提供的*txt*文件是我准备的资料，我需要把一批< |*toxic_sentence*| >英文按照文件要求转换为< |*neutral_sentence*| >英文版本 (I am currently completing a downstream data processing task in natural language processing. The TXT file I provided contains the material I have prepared. I need to convert a batch of < *toxic_sentence* > **English** according to the requirements of the file into < *neutral_sentence* > **English** versions.)

\# OBJECTIVE #

我会提供一批< |*toxic_sentence*| >英文给你，请你给我< |*neutral_sentence*| >英文版本，来完成任务 (I will provide you with a batch of < toxic_sentence > **English**; please give me the < neutral_sentence > **English** versions to complete the task.)

\# STYLE #

你的风格应该像严谨的懂得文学的程序员，严格完成任务 (Your style should be like that of a rigorous programmer who understands literature and strictly completes the task.)

\# TONE #

专业的 (Professional.)

\# AUDIENCE #

懂得英文的成年评委，他们会判断你转换的< |*neutral_sentence*| >英文是否符合要求 (Adult judges who understand **English** will judge whether the < *neutral_sentence* > **English** you have converted meets the requirements.)

\# RESPONSE #

返回如下 *json* 格式

[{*toxic_sentence*: "", *neutral_sentence*: ""}] (Return in the following JSON format: [{"toxic_sentence": "", "neutral_sentence": ""}])

Input toxic sentences and get the results

Examples of different Models

Table 4

Examples of detoxifications by baseline models and our model. Inspired by the work of Daryna et al. [6] , we have indicated bad answers in red and good answers **in bold**.

Original	i did not vandalize anything you d*ck .	Yeah, you can f*ck off incel.	Be careful who you call a m*ron.
delete	i did not vandalize anything you .	Yeah, you can off incel.	Be careful who you call a m*ron.
mT5	I did not vandalize anything you did not do.	Yeah, you can *ck off incel.	Be careful who you call a m*ron.
backtranslation	**I did not vandalize anything.**	**Yeah, you can go away incel.**	Be careful who you call.
Team cake	**I did not vandalize anything.**	**Yeah, you can leave now.**	**Be careful who you insult.**

Results

User	en*	zh	es	de	ar	hi	uk	ru	am	average
Team cake	**0.91**	**0.84**	0.77	0.77	0.44	0.34	0.50	0.71	0.61	0.65
mkrisnai	0.89	0.34	0.83	0.70	0.63	0.73	0.73	0.78	0.49	0.68
Human References	0.88	0.93	0.79	0.71	0.82	0.97	0.90	0.80	0.85	0.85
SomethingAwful	0.86	0.53	**0.83**	**0.89**	0.74	0.86	0.69	**0.84**	0.71	**0.77**
bmmikheev	0.84	0.60	0.76	0.78	0.69	0.78	0.63	0.51	0.56	0.69
adugeen	0.83	0.60	0.73	0.70	0.82	0.68	**0.84**	0.76	0.71	0.74
ZhongyuLuo	0.73	0.56	0.52	0.01	0.49	0.49	0.42	0.68	**0.72**	0.51
backtranslation	0.73	0.34	0.56	0.34	0.42	0.33	0.23	0.22	0.54	0.41
nikita.sushko	0.70	0.47	0.62	0.79	**0.89**	0.84	0.67	0.74	0.68	0.71
VitalyProtasov	0.69	0.49	0.81	0.77	0.79	**0.87**	0.67	0.73	0.68	0.72
mT5	0.68	0.43	0.47	0.64	0.63	0.60	0.42	0.40	0.61	0.54
delete	0.47	0.43	0.55	0.57	0.65	0.65	0.60	0.49	0.63	0.56

This leaderboard is based on the J scores of **English(en).** We highlight top1 results per each column. Our team name is **Team cake.**

For en & zh -- Very Good

User	en*	zh	es	de	ar	hi	uk	ru	am	average
Team cake	**0.91**	**0.84**	0.77	0.77	0.44	0.34	0.50	0.71	0.61	0.65
mkrisnai	0.89	0.34	0.83	0.70	0.63	0.73	0.73	0.78	0.49	0.68
Human References	0.88	0.93	0.79	0.71	0.82	0.97	0.90	0.80	0.85	0.85
SomethingAwful	0.86	0.53	**0.83**	**0.89**	0.74	0.86	0.69	**0.84**	0.71	**0.77**
bmmikheev	0.84	0.60	0.76	0.78	0.69	0.78	0.63	0.51	0.56	0.69
adugeen	0.83	0.60	0.73	0.70	0.82	0.68	**0.84**	0.76	0.71	0.74
ZhongyuLuo	0.73	0.56	0.52	0.01	0.49	0.49	0.42	0.68	**0.72**	0.51
backtranslation	0.73	0.34	0.56	0.34	0.42	0.33	0.23	0.22	0.54	0.41
nikita.sushko	0.70	0.47	0.62	0.79	**0.89**	0.84	0.67	0.74	0.68	0.71
VitalyProtasov	0.69	0.49	0.81	0.77	0.79	**0.87**	0.67	0.73	0.68	0.72
mT5	0.68	0.43	0.47	0.64	0.63	0.60	0.42	0.40	0.61	0.54
delete	0.47	0.43	0.55	0.57	0.65	0.65	0.60	0.49	0.63	0.56

For es&de&ru&am -- Good

User	en*	zh	es	de	ar	hi	uk	ru	am	average
Team cake	**0.91**	**0.84**	0.77	0.77	0.44	0.34	0.50	0.71	0.61	0.65
mkrisnai	0.89	0.34	0.83	0.70	0.63	0.73	0.73	0.78	0.49	0.68
Human References	0.88	0.93	0.79	0.71	0.82	0.97	0.90	0.80	0.85	0.85
SomethingAwful	0.86	0.53	**0.83**	**0.89**	0.74	0.86	0.69	**0.84**	0.71	**0.77**
bmmikheev	0.84	0.60	0.76	0.78	0.69	0.78	0.63	0.51	0.56	0.69
adugeen	0.83	0.60	0.73	0.70	0.82	0.68	**0.84**	0.76	0.71	0.74
ZhongyuLuo	0.73	0.56	0.52	0.01	0.49	0.49	0.42	0.68	**0.72**	0.51
backtranslation	0.73	0.34	0.56	0.34	0.42	0.33	0.23	0.22	0.54	0.41
nikita.sushko	0.70	0.47	0.62	0.79	**0.89**	0.84	0.67	0.74	0.68	0.71
VitalyProtasov	0.69	0.49	0.81	0.77	0.79	**0.87**	0.67	0.73	0.68	0.72
mT5	0.68	0.43	0.47	0.64	0.63	0.60	0.42	0.40	0.61	0.54
delete	0.47	0.43	0.55	0.57	0.65	0.65	0.60	0.49	0.63	0.56

For ar&hi&uk -- Need to improve

User	en*	zh	es	de	ar	hi	uk	ru	am	average
Team cake	**0.91**	**0.84**	0.77	0.77	0.44	0.34	0.50	0.71	0.61	0.65
mkrisnai	0.89	0.34	0.83	0.70	0.63	0.73	0.73	0.78	0.49	0.68
Human References	0.88	0.93	0.79	0.71	0.82	0.97	0.90	0.80	0.85	0.85
SomethingAwful	0.86	0.53	**0.83**	**0.89**	0.74	0.86	0.69	**0.84**	0.71	**0.77**
bmmikheev	0.84	0.60	0.76	0.78	0.69	0.78	0.63	0.51	0.56	0.69
adugeen	0.83	0.60	0.73	0.70	0.82	0.68	**0.84**	0.76	0.71	0.74
ZhongyuLuo	0.73	0.56	0.52	0.01	0.49	0.49	0.42	0.68	**0.72**	0.51
backtranslation	0.73	0.34	0.56	0.34	0.42	0.33	0.23	0.22	0.54	0.41
nikita.sushko	0.70	0.47	0.62	0.79	**0.89**	0.84	0.67	0.74	0.68	0.71
VitalyProtasov	0.69	0.49	0.81	0.77	0.79	**0.87**	0.67	0.73	0.68	0.72
mT5	0.68	0.43	0.47	0.64	0.63	0.60	0.42	0.40	0.61	0.54
delete	0.47	0.43	0.55	0.57	0.65	0.65	0.60	0.49	0.63	0.56

Final Thoughts 1/2 -- Future work

- We need to do more ablation experiments to study the effect of different prompt sentences, different chat models and other factors on the detoxification effect.

- The assessment could be performed with ChatGPT.

Final Thoughts 2/2 --The influence on me

- Participating in this task indirectly improved my relationship with my parents！

- I used to unconsciously use digital violence when talking to my parents.

- This task has taught me the beauty of using non-violent communication.

Ending

- This is the end of my sharing. 谢谢大家！（Thank you all！）

- Welcome to ask me any questions!

References 1/4

My slide format reference https://eval4nlp.github.io/2021/keynotes/ehud-reiter.pdf

S. Sood, J. Antin, E. Churchill, Profanity use in online communities, in: Proceedings of the SIGCHI conference on human factors in computing systems, 2012, pp. 1481–1490.

D. Dale, A. Voronov, D. Dementieva, V. Logacheva, O. Kozlova, N. Semenov, A. Panchenko, Text detoxification using large pre-trained neural models, arXiv preprint arXiv:2109.08914 (2021).

J. Bevendorff, X. B. Casals, B. Chulvi, D. Dementieva, A. Elnagar, D. Freitag, M. Fröbe, D. Ko-

renčić, M. Mayerl, A. Mukherjee, A. Panchenko, M. Potthast, F. Rangel, P. Rosso, A. Smirnova, E. Stamatatos, B. Stein, M. Taulé, D. Ustalov, M. Wiegmann, E. Zangerle, Overview of PAN 2024: Multi-Author Writing Style Analysis, Multilingual Text Detoxification, Oppositional Thinking Analysis, and Generative AI Authorship Verification, in: Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Fourteenth International Conference of the CLEF Association (CLEF 2024), Lecture Notes in Computer Science, Springer, Berlin Heidelberg New York, 2024.

References 2/4

D. Dementieva, D. Moskovskiy, N. Babakov, A. A. Ayele, N. Rizwan, F. Schneider, X. Wang, S. M. Yimam, D. Ustalov, E. Stakovskii, A. Smirnova, A. Elnagar, A. Mukherjee, A. Panchenko, Overview of the multilingual text detoxification task at pan 2024, in: G. Faggioli, N. Ferro, P. Galuščáková, A. G. S. de Herrera (Eds.), Working Notes of CLEF 2024 - Conference and Labs of the Evaluation Forum, CEUR-WS.org, 2024.

D. Dementieva, D. Moskovskiy, V. Logacheva, D. Dale, O. Kozlova, N. Semenov, A. Panchenko, Methods for detoxification of texts for the russian language, Multimodal Technologies and Interaction 5 (2021) 54.

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), 2022, pp. 6804–6818.

M. Lewis, Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, L. Zettlemoyer, Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension, arXiv preprint arXiv:1910.13461 (2019).

References 3/4

D. Dementieva, V. Logacheva, I. Nikishina, A. Fenogenova, D. Dale, I. Krotova, N. Semenov, T. Shavrina, A. Panchenko, Russe-2022: Findings of the first russian detoxification shared task based on parallel corpora, in: Proceedings of the RUSSE-2022 Shared Task, 2022. doi:10.28995/ 2075-7182-2022-21-114-131.

A. Radford, K. Narasimhan, T. Salimans, I. Sutskever, et al., Improving language understanding by generative pre-training (2018).

A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, et al., Language models are unsupervised multitask learners, OpenAI (2019).

T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al., Language models are few-shot learners, Advances in neural information processing systems 33 (2020) 1877–1901.

A. Pesaranghader, N. Verma, M. Bharadwaj, Gpt-detox: An in-context learning-based paraphraser for text detoxification, in: 2023 International Conference on Machine Learning and Applications (ICMLA), IEEE, 2023, pp. 1528–1534.

References 4/4

J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat, et al., Gpt-4 technical report, arXiv preprint arXiv:2303.08774 (2023).

P. Liu, W. Yuan, J. Fu, Z. Jiang, H. Hayashi, G. Neubig, Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing, ACM Computing Surveys 55 (2023) 1–35.

S. Teo, How i won singapore’s gpt-4 prompt engineering competition: A deep dive into the strategies i learned for harnessing the power of large language mod- els (llms), Towards Data Science (2023). URL: https://towardsdatascience.com/ how-i-won-singapores-gpt-4-prompt-engineering-competition-34c195a93d41, accessed: July 10, 2024.

J. Chen, S. Li, Q. Huang, S. Yan, Z. Xie, Y. Lu, Application of kimi intelligent assistant in the teaching of water pollution control engineering course, International Journal of Education and Humanities 13 (2024) 39–43.

L. Xue, N. Constant, A. Roberts, M. Kale, R. Al-Rfou, A. Siddhant, A. Barua, C. Raffel, mT5: A massively multilingual pre-trained text-to-text transformer, in: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, Online, 2021, pp. 483–498. URL: https: //aclanthology.org/2021.naacl-main.41.