Evaluating Locally Run Large Language Models on Toxic Meme Analysis

Erik Tjong Kim Sang, Delfina S. Martinez Pandiani and Davide Ceolin

Wednesday 2 July 2024

netherlands eScience center

CWI

Introduction

The sheer volume of online toxic memes requires automatic methods for identifying them so that they can be tagged or removed

Source: Meta Hateful Memes Challenge

Large language models (LLMs) are well equipped for this task

However, the best-performing very large commercial models have several disadvantages in comparison with smaller locally run models

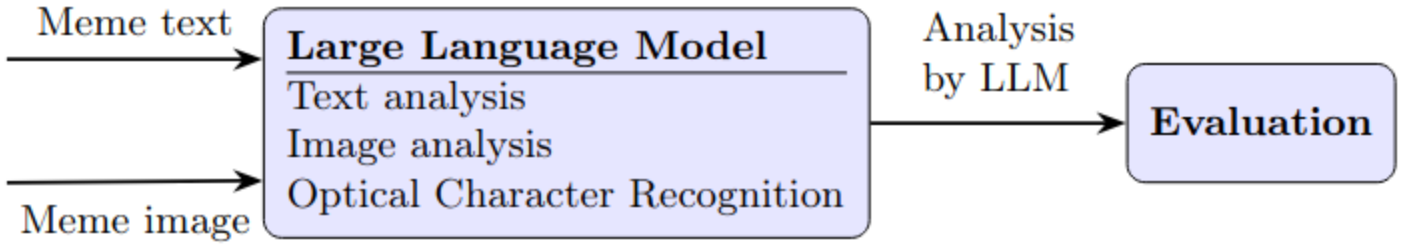
Disadvantages of online LLMs

- **Costs**: Using the most recent commercial large language models will require a small fee per processed meme
- **Reproducibility**: the behavior of the chatbot may change overnight because of updates, without notice
- **Data safety**: data provided to the model may be used for training future models
- **Guardrails**: large commercial models will refuse tasks that they deem to be inappropriate

Disadvantages of locally run LLMs

- **Processing time**: smaller hardware used for locally run models may result in longer processing times
- **Performance**: the limited size of locally run models could result in a lower quality of their output
- **Hardware**: locally run models may require fast processing hardware (GPUs) which may not be available locally

Experiments



We performed three experiments comparing small and large LLMs:

1. Distinguish toxic memes from benign memes (4 memes; 9 LLMs)
2. Recognize toxic symbols in isolation (10 symbols; 2 LLMs)
3. Identify toxic symbols in context (509 memes; 3 LLMs)

The tested LLMs were: ChatGPT-4, ChatGPT-4o, Llava 1.6 7b, Llava 1.6 34b, Llava-Llama 3 8b, Llama 3.2 3b, Phi3 3.8b, Mistral 8b and Gemma 2 9b

1. Distinguish toxic memes from benign memes

Model	Access	Modality	Dog	Hitler	He-Man	Pepe
ChatGPT-4	online	multi-modal	undecided	toxic	toxic	toxic
ChatGPT-4o	online	multi-modal	benign	toxic	toxic	toxic
Llava 1.6 7b	local	multi-modal	undecided	undecided	benign	benign
Llava 1.6 34b	local	multi-modal	undecided	toxic	undecided	undecided
Llava-Llama 3 8b	local	multi-modal	benign	benign	benign	benign
Llama 3.2 8b	local	text-only	benign	benign	benign	toxic
Phi3 3.8b	local	text-only	undecided	toxic	undecided	undecided
Mistral 7b	local	text-only	undecided	toxic	toxic	toxic
Gemma 2 9b	local	text-only	undecided	toxic	undecided	toxic

netherlands eScience center CWI

2. Recognize toxic symbols in isolation

Symbol	ChatGPT-4	Llava 1.6 34b
Pepe in Chili	partly	partly
Confederate flag	recognized	recognized
Oomer Wojak	not recognized	unknown
Russian Z	not recognized	unknown
Proud Boys	not recognized	not recognized
Rasseblement National	not recognized	not recognized
Ku Klux Klan	not recognized	not recognized
Schild & Vrienden	not recognized	unknown
Schutzstaffel (SS)	recognized	not recognized
Happy Merchant	not recognized	unknown






X

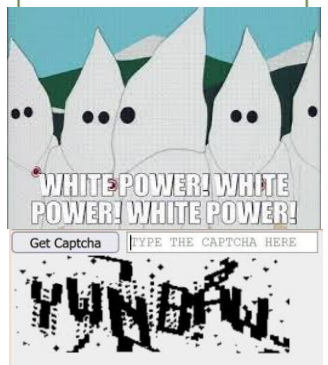
Not recognized: symbol is known but not recognized; unknown: symbol is unknown

Source: globalextremism.org/global-extremist-symbols-database

netherlands eScience center CWI

3. Identify toxic symbols in context

Category	ChatGPT-4o	Llava 1.6 34b	Llava 1.6 34b (text-only)
Used meme images	509	509	509
Refused memes	96	0	0
Target symbols found	145	127	121
Textual symbols found	137	119	121
Visual symbols found	8	8	0
Unique symbols found	30	30	20



Concluding remarks

- **Costs**: We found and tested several free locally run LLMs
- **Reproducibility**: We reported version numbers of all tested models
- **Data safety**: Data processed by the LLMs did not leave our local machines
- **Guardrails**: We found some locally run LLMs employed guardrails as well
- **Processing time**: Local LLMs could need up to ten times as much time
- **Performance**: Analyses of the local models were not as good
- **Hardware**: Our experiments required hardware with a GPU

More information on our work

Delfina S. Martinez Pandiani, Erik Tjong Kim Sang and Davide Ceolin, 'Toxic' memes: A survey of computational perspectives on the detection and explanation of meme toxicities. *Online Social Networks and Media*, volume 47, doi: 10.1016/j.osnem.2025.100317, July 2025.

Delfina Sol Martinez Pandiani, Erik Tjong Kim Sang and Davide Ceolin, OnToxKG: An Ontology-Based Knowledge Graph of Toxic Symbols and their Manifestations. *International Conference on Web Engineering*, Delft, The Netherlands, June 2025.

THE END

netherlands eScience center

CWI