Evaluating Locally Run Large Language Models on Toxic Meme Analysis

Erik Tjong Kim Sang[1], Delfina S. Martinez Pandiani[2], and Davide Ceolin[2]
e.tjongkimsang@esciencecenter.nl, dsmp@cwi.nl, davide.ceolin@cwi.nl

[1] Netherlands eScience Center, Amsterdam, The Netherlands, Europe
[2] Centrum Wiskunde & Informatica, Amsterdam, The Netherlands, Europe

1 July 2025

Introduction

Large language models work well in separating toxic memes from benign memes. While large commercial models like ChatGPT perform best on this task, smaller locally run models have several advantages over larger models

Advantages of locally run LLMs

Costs Large commercial models require subscription fees
Reproducibility Large models may change overnight
Data safety Input data may be reused by large models
Guardrails Some tasks may be refused by large models

Disadvantages of locally run LLMs

Processing time can be longer due to smaller hardware
Performance may be worse because of smaller models
Hardware may also be non-standard for smaller models

Test memes

Results

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
Lama 3.2 3b	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

Conclusions

Costs All tested locally run LLMs are freely available
Reproducibility We report all model version numbers
Data safety All data remained on our local machines.
Guardrails Some locally run LLMs also had guardrails
Processing time Models required ten times as much time
Performance Analyses from local models were worse
Hardware Tests required hardware with one GPU

More information

More test results will be presented in our talk:
Wednesday 2 July 2025, 14:50-15:00, IDE Arena