



UNIVERSITY
OF AMSTERDAM

Revisiting Web-Scale Harmful Content Filtering for Safer LLM Pretraining

Ark Deliev, Ali Bilge and Miguel Mendes

Master in Artificial Intelligence, FACT Course Project

January, 2026

Project

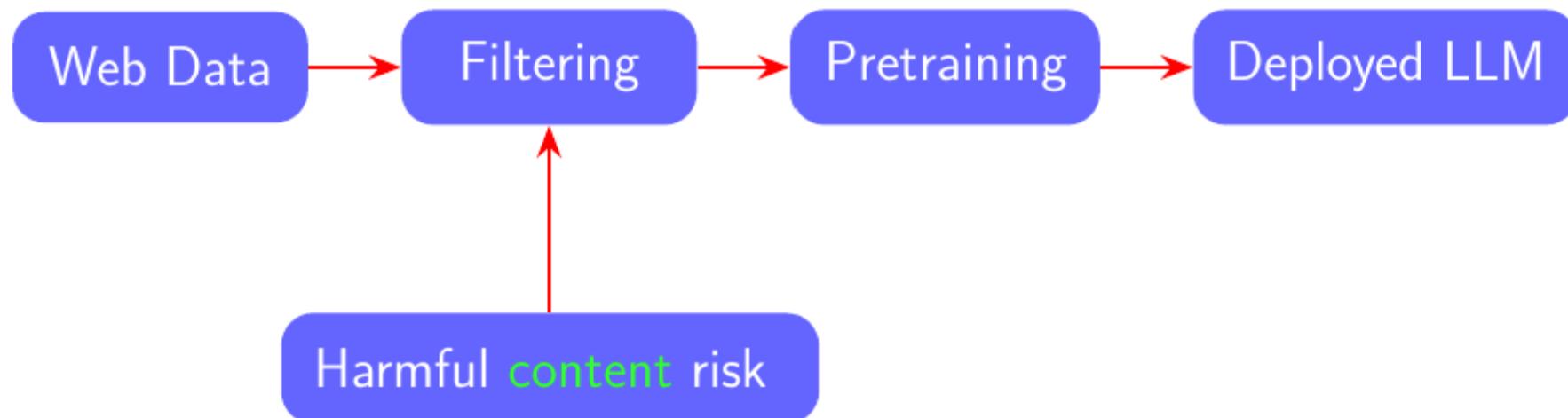
- We evaluate the **reproducibility of a taxonomy-driven framework** for harmful content detection using released benchmarks and models.
- The project is based on the following paper:

[1]. Sai Krishna Mendu , Harish Yenala , Aditi Gulati , Shanu Kumar , Parag Agrawal, (2025),
Towards Safer Pretraining: Analyzing and Filtering Harmful Content in Webscale Datasets for Responsible LLMs, 2025 IJCAI Conference, arXiv:2505.02009v3.

Abbreviations

- TTP — Topical and Toxic Prompt
- HAVOC — Multi-Harm Open-ended Toxicity Benchmark
- LLM — Large Language Models

Why this paper matters?



Models and Benchmarks

- **TTP** (Topical & Toxic Prompt): prompt-based classifier
- **TTP-Eval**: human-annotated benchmark
- **HarmFormer**: neural classifier for harmful content detection
- **HAVOC**: benchmark for measuring harmful content leakage

Prompt

- └ Hate & Violence
- └ Ideological Harm
- └ Sexual Content
- └ Illegal Activity
- └ Self-Inflicted Harm



Toxic / Non-Toxic

Claims made in the original paper

Claim 1: TTP performs well on TTP-Eval ($F1 = 0.83$)

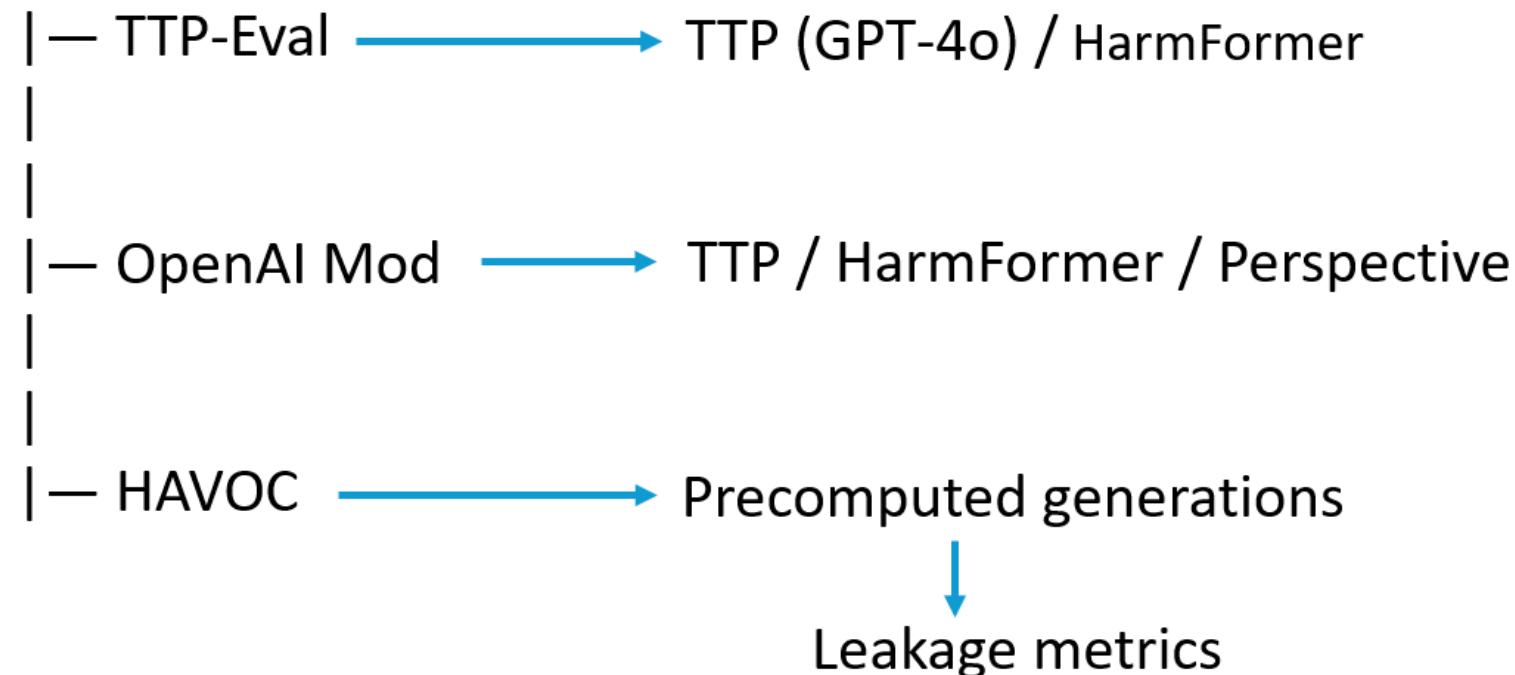
Claim 2: HarmFormer shows strong performance

Claim 3: TTP and HarmFormer outperform baselines on OpenAI Moderation

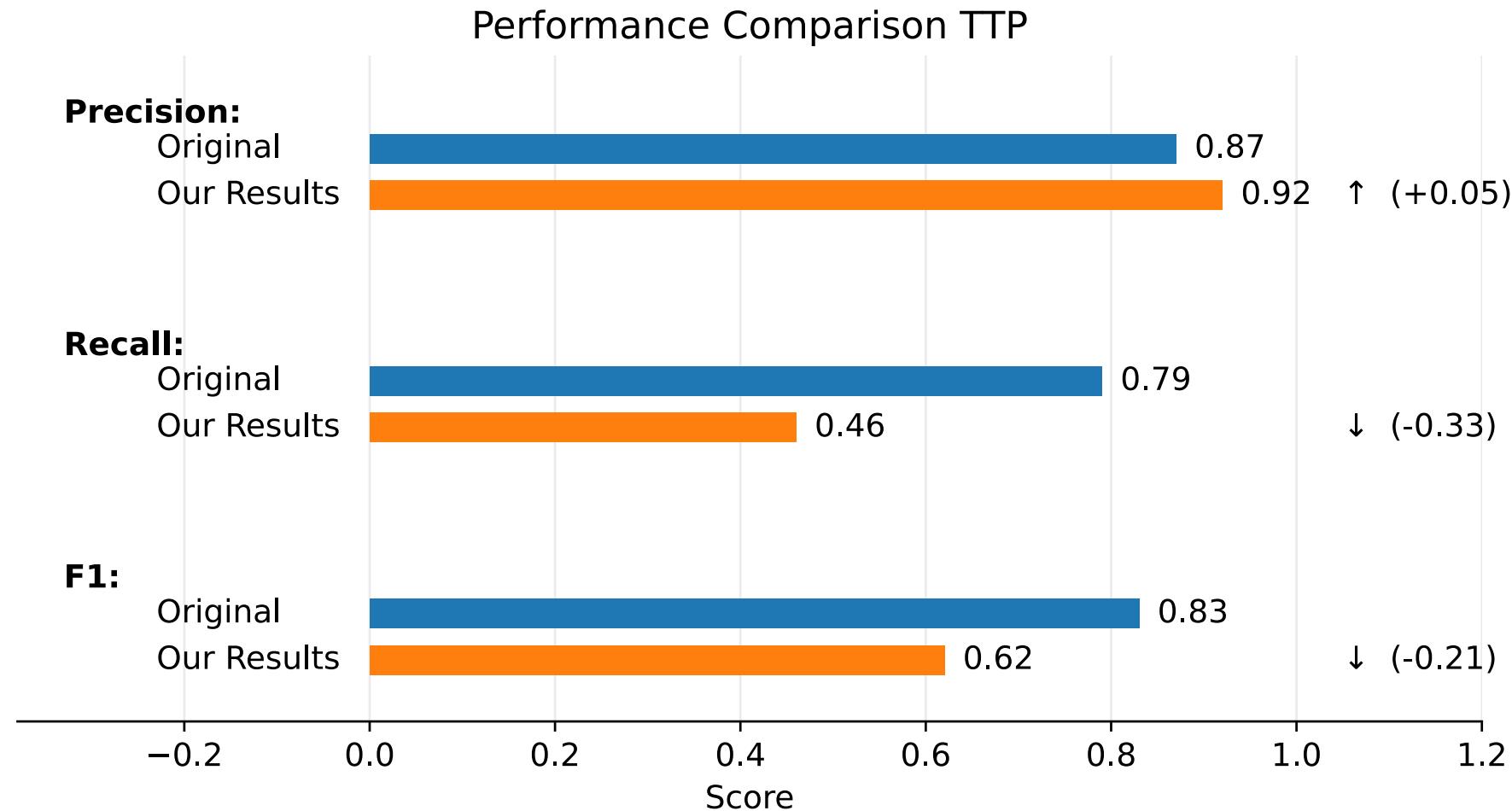
Claim 4: HAVOC shows ~26.7% leakage

Reproduction Setup

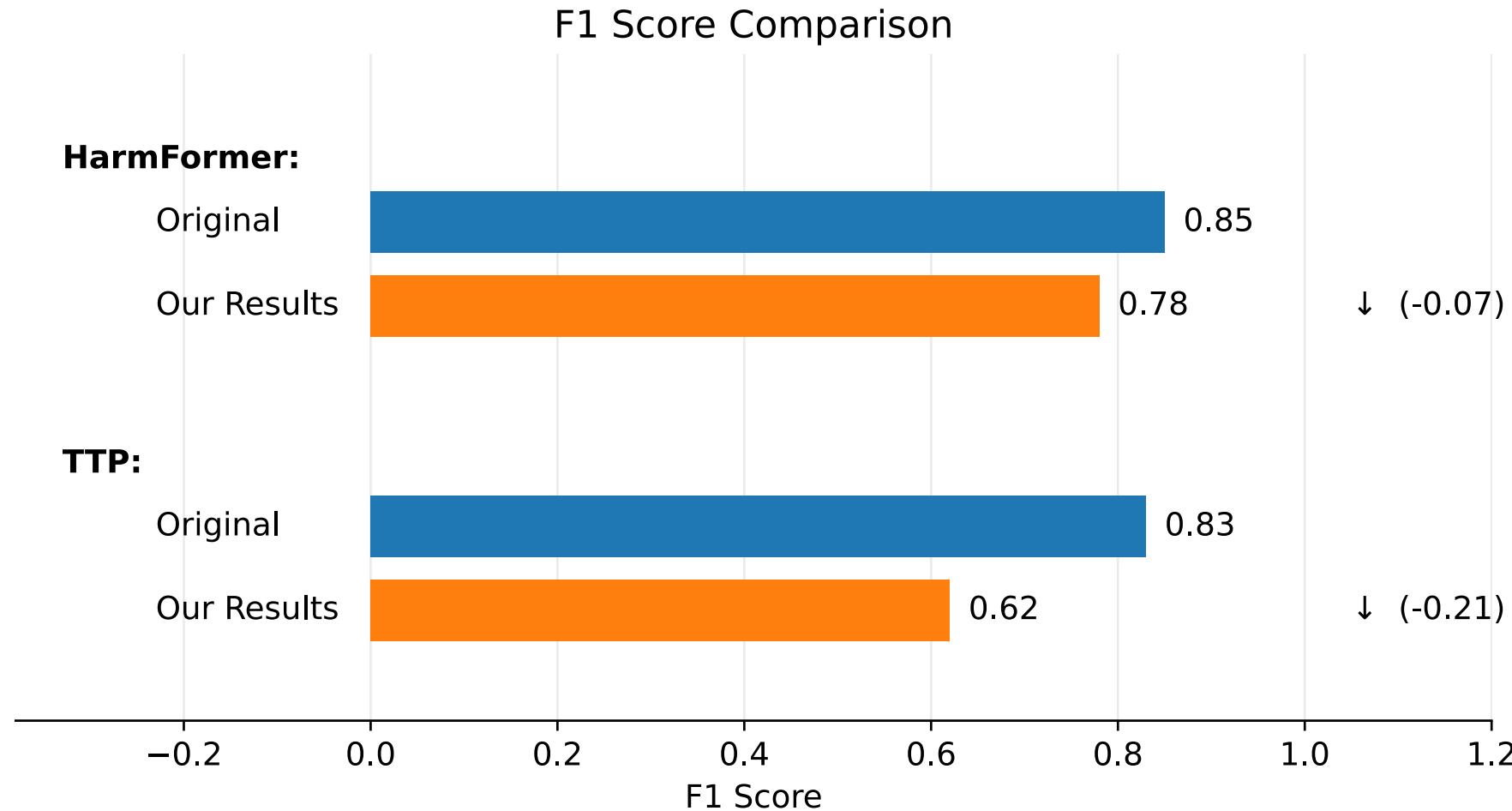
Datasets



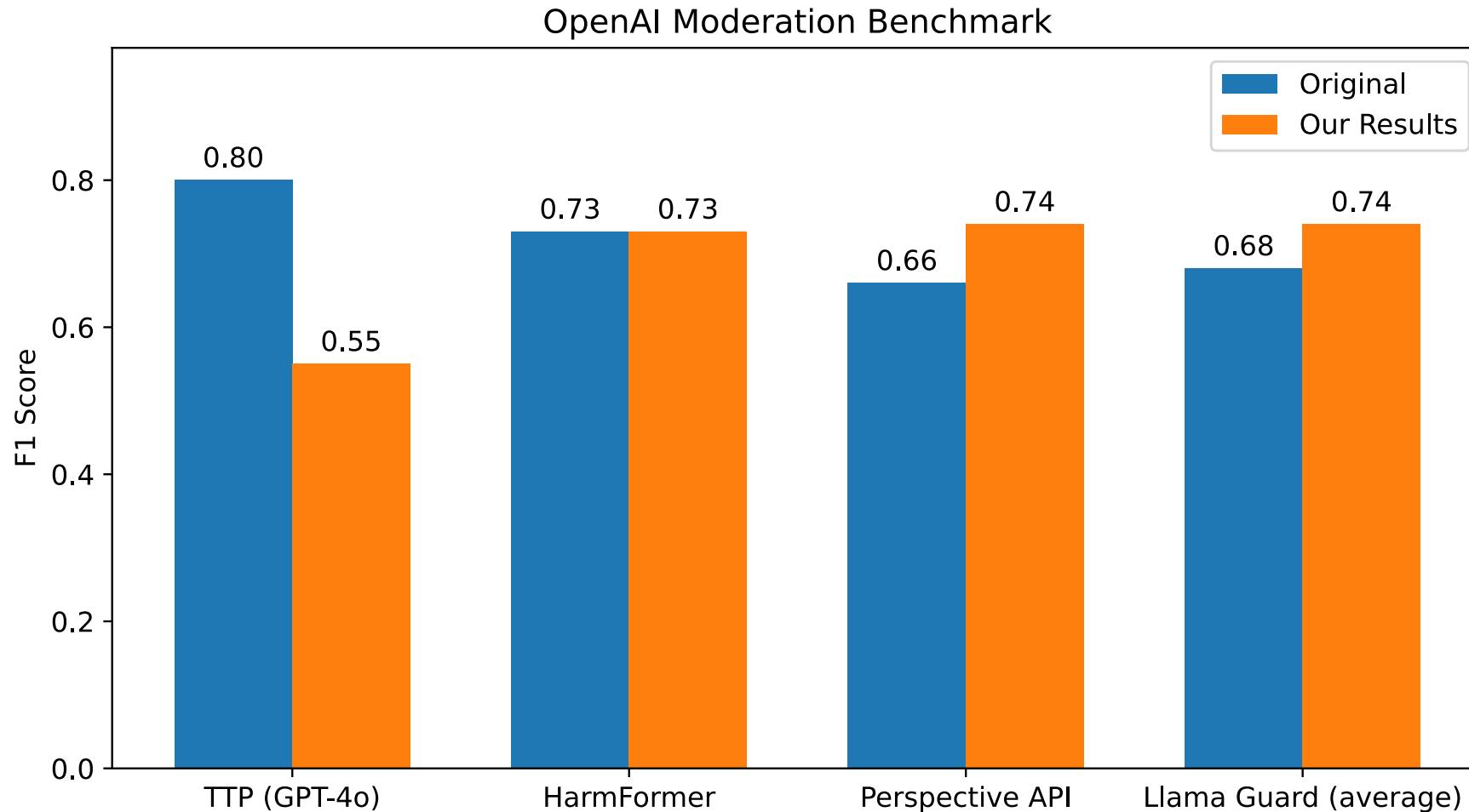
Key Result 1: TTP on TTP-Eval



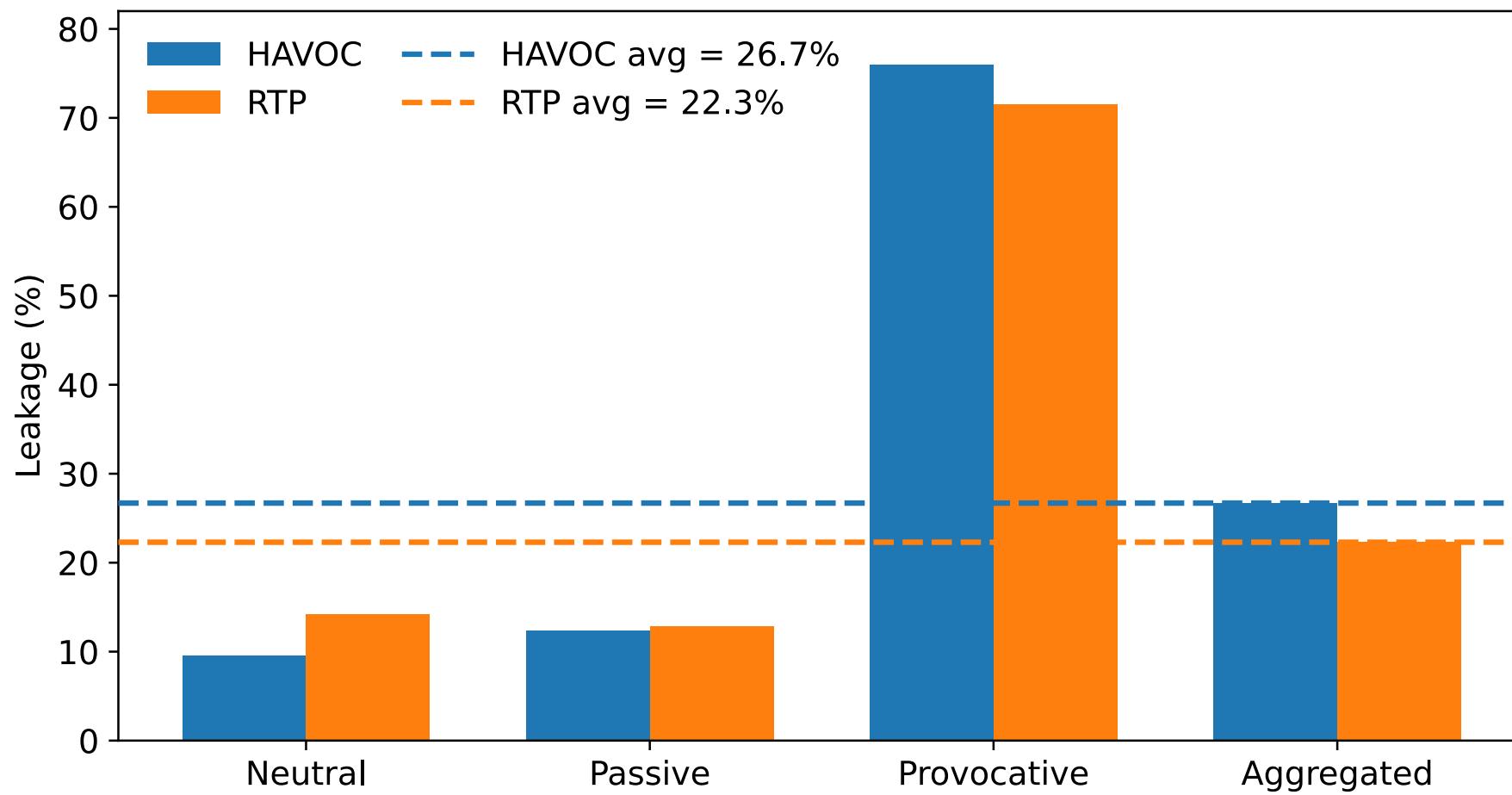
Key Result 2: HarmFormer vs TTP (on TTP-Eval)



Key Result 3: OpenAI Moderation Benchmark



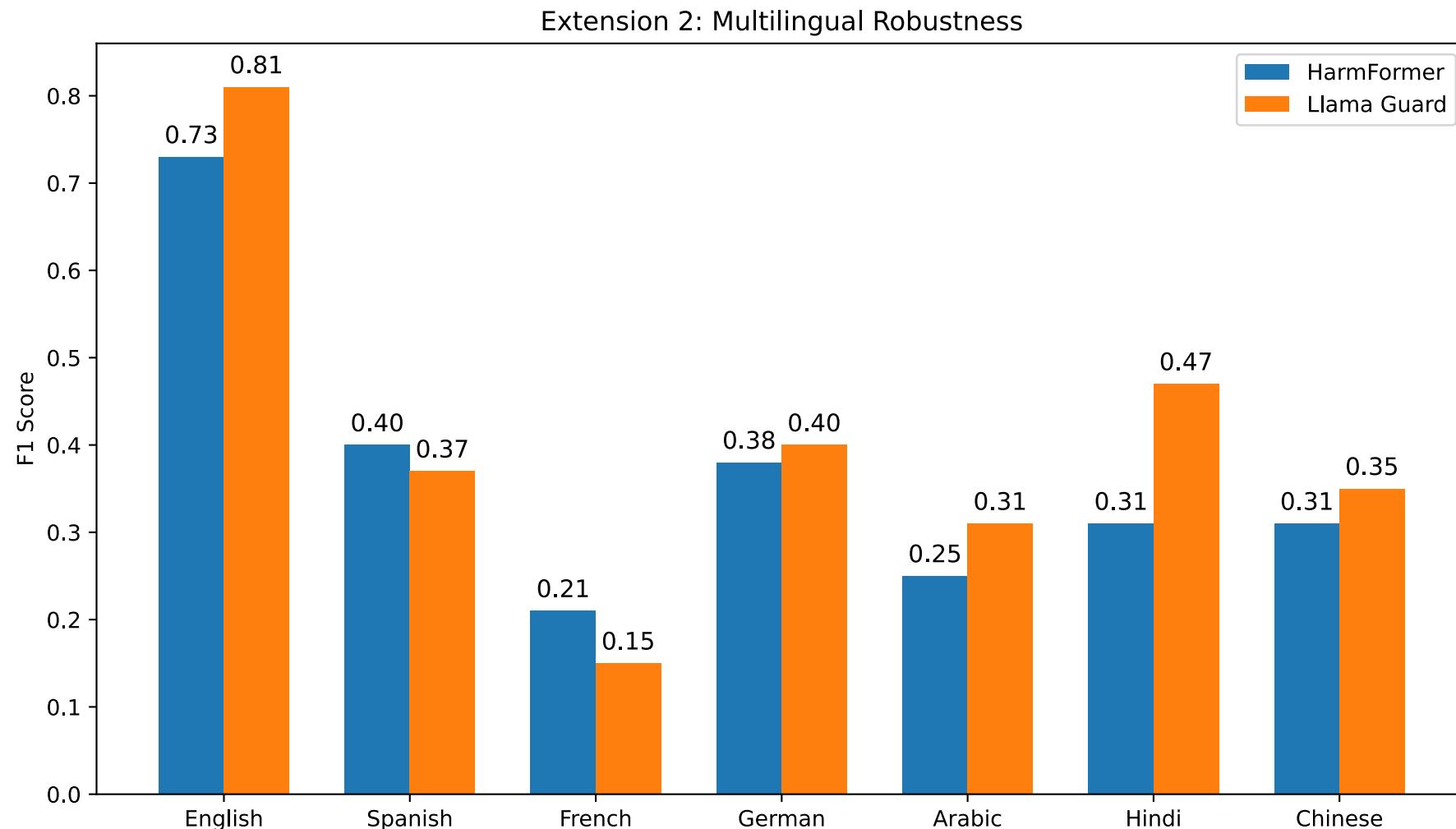
Key Result 4: HAVOC



Extension 1: Cross-Model TTP Robustness

Model	Precision	Recall	F1
GPT-4o	0.92	0.46	0.62
Gemini 2.0 Flash	0.76	0.84	0.80
DeepSeek-R1-Distill-Qwen-32B	0.70	0.58	0.63
Gemma 2 27B	0.95	0.21	0.35
Gemma 3 27B	0.87	0.42	0.57
GPT-OSS 20B	0.96	0.24	0.39

Extension 2: Multilingual Robustness



Reproducibility Summary

Claim	Description	Status
Claim 1	TTP on TTP-Eval	Invalidated
Claim 2	HarmFormer	Partially validated
Claim 3	OpenAI Moderation	Partially validated
Claim 4	HAVOC leakage	Reproduced exactly

Environmental & Practical Impact

Aspect	Original paper	Our reproduction
Web pages processed	~3,000,000	393
Model training	Yes (HarmFormer trained)	No (pretrained)
HAVOC inference	Yes (multiple models)	No (precomputed)
GPU hours	O(10,000+) (estimate)	~5–15
CO ₂	Not reported	~0.05 kg

Conclusions and Future Work

- **Key Conclusions**
 - HAVOC is reliable and easily reproducible
 - HarmFormer generalizes reasonably to human-annotated data
 - TTP performance is model-dependent and fragile
- **Future work**
 - Future work should focus on building open and robust tools that can generalize well across different model families.