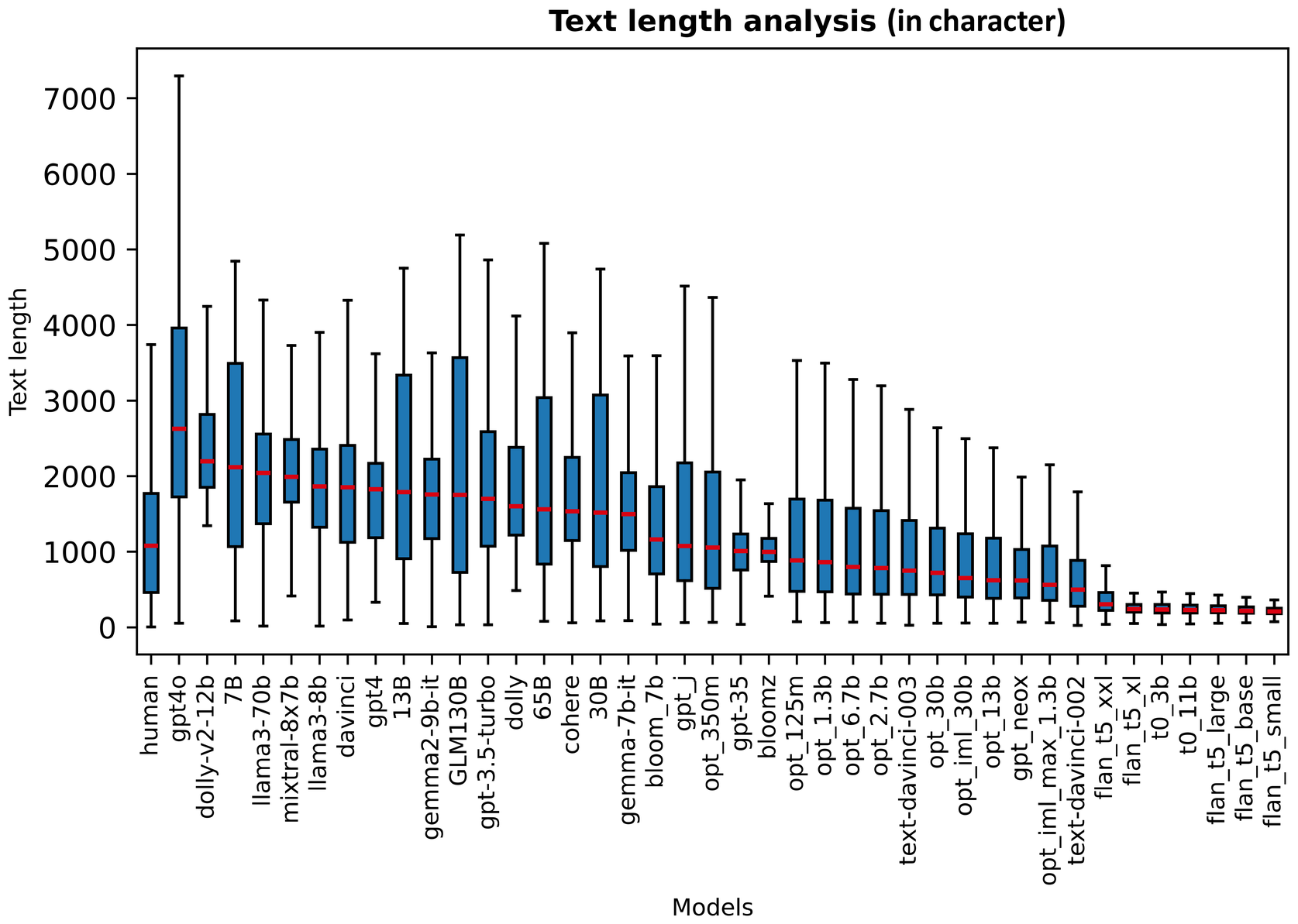
Széchenyi Plan Plus Let's build Hungary together!
Government of Hungary
Funded by the European Union NextGeneration EU
ARTIFICIAL INTELLIGENCE National Laboratory

Beyond Binary - Soft-Voting Multi-Class Classification for Binary Machine-Generated Text Detection Across Diverse Language Models

Mihály Kiss, Gábor Berend

SZTE UNIVERSITY OF SZEGED


Text length analysis (in character)
Text length
0
1000
2000
3000
4000
5000
6000
7000
human
gpt4o
dolly-v2-12b
7B
llama3-70b
mixtral-8x7b
llama3-8b
davinci
gpt4
13B
gemma2-9b-it
GLM130B
gpt-3.5-turbo
dolly
65B
cohere
30B
gemma-7b-it
bloom_7b
gpt_j
opt_350m
gpt-35
bloomz
opt_125m
opt_1.3b
opt_6.7b
opt_2.7b
text-davinci-003
opt_30b
opt_iml_30b
opt_13b
gpt_neox
opt_iml_max_1.3b
text-davinci-002
flan_t5_xxl
flan_t5_xl
t0_3b
t0_11b
flan_t5_large
flan_t5_base
flan_t5_small
Models

Text quality comparison

Flan_t5_small

A 0day bug exists in iOS 13.4.1 that allows malicious attackers to steal iMessage history. The bug allows apps to access sms.db and send it to a remote server. Steps 0day bug exists in iOS 13,4.1 that allows malicious attackers to steal iMessage history.

GPT4o

In this study, we address the challenge of distinguishing genuine gravitational-wave (GW) burst signals from spurious noise artifacts in interferometric detectors. Our primary objective is to enhance the reliability of GW detections by developing robust physical instrumental vetoes. We propose a novel methodology that leverages auxiliary channels to identify and exclude non-astrophysical triggers. By systematically analyzing correlations between GW burst triggers and instrumental noise, we establish criteria for vetoing false positives...

Experiments

Argmax – Deberta – 512 token

Experiments

- **41 class**

A straightforward approach: treat each model as a separate class.

Experiments

- 41 class
- **5/6 class (parameter-based)**

Human	Below 1 billion	1.3-5 billion	6-10 billion	11-130 billion	Above 130 billion
human	flan_t5_small, opt_125m, flan_t5_base, opt_350m, flan_t5_large	opt_1.3b, opt_iml_max_1.3b, opt_2.7b, t0_3b, flan_t5_xl	dolly, gpt_j, opt_6.7b, mixtral-8x7b, gemma-7b-it, bloom_7b, 7B, llama3-8b, gemma2-9b-it	t0_11b, flan_t5_xxl, dolly-v2-12b, 13B, opt_13b, gpt_neox, opt_iml_30b, 30B, opt_30b, 65B, llama3-70b, bloomz, GLM130B	davinci, gpt-35, text-davinci-003, text-davinci-002, gpt-3.5-turbo, cohere, gpt4o, gpt4

Experiments

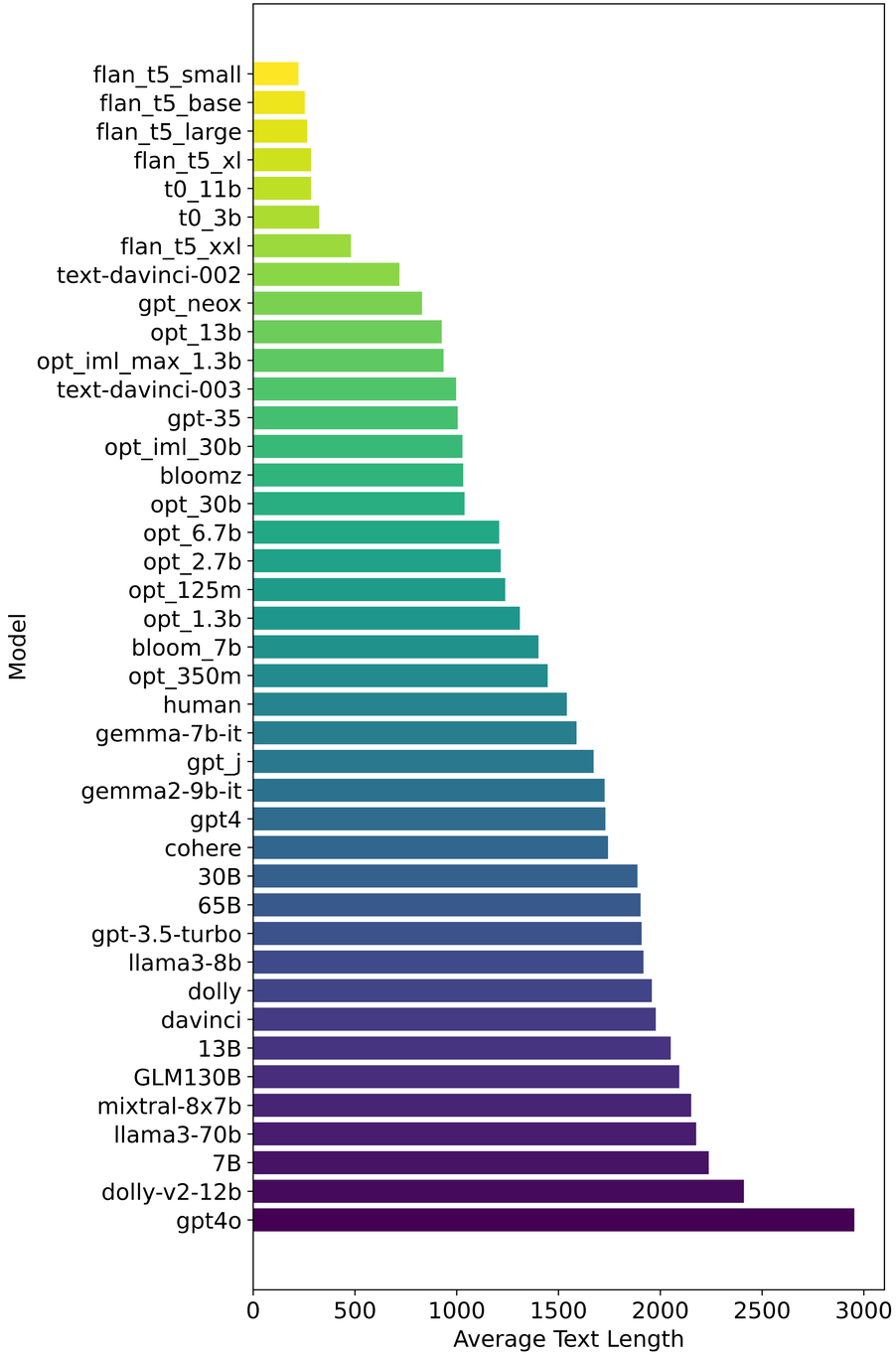
- 41 class
- 5/6 class (parameter-based)
- **3 class (strength-based)**

Human	**Weak**	**Strong**
human	Flan_t5 Opt_3b Bloom_7b ...	GPT4 Mixtral Llama3 ...

Experiments

- 41 class
- 5/6 class (parameter-based)
- 3 class (strength-based)
- **Length-based (41 class)**

3 intervals, 3 expert models



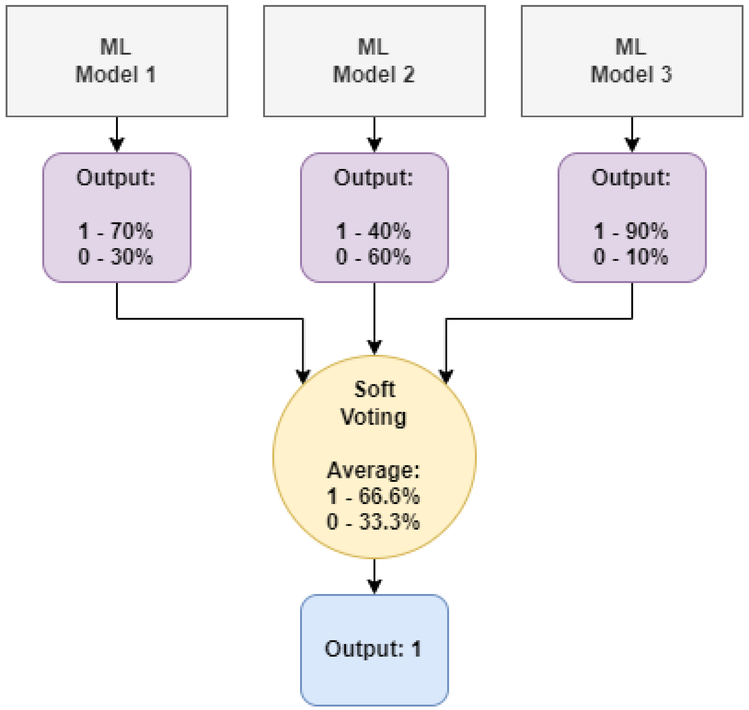
Experiments

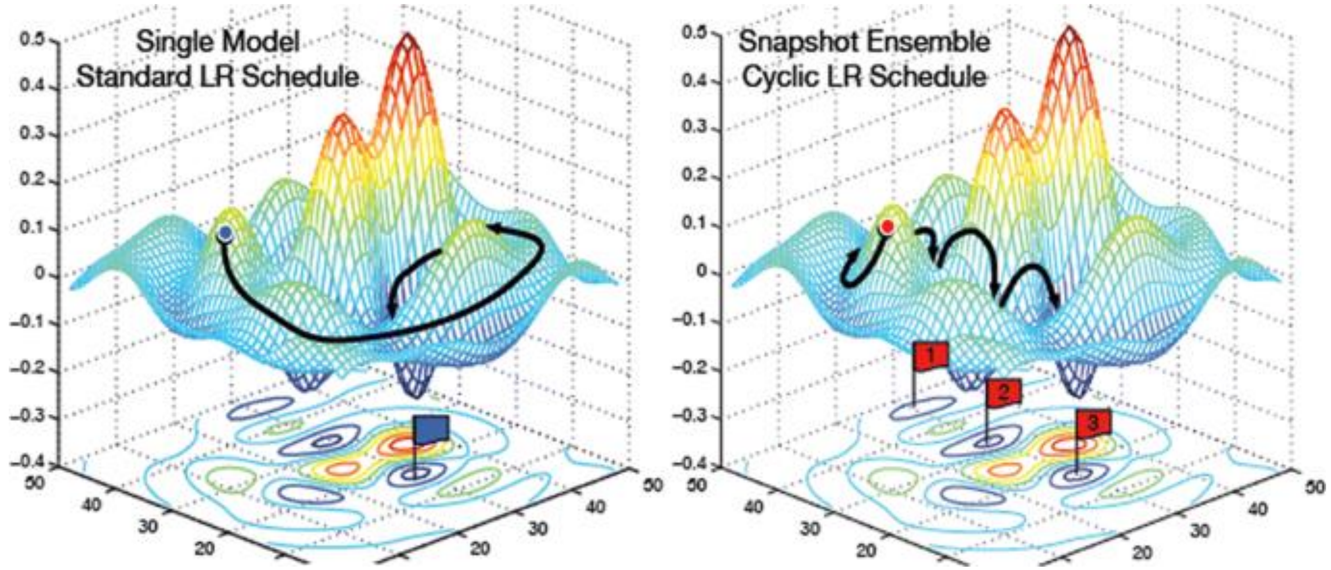
- 41 class
- 5/6 class (parameter-based)
- 3 class (strength-based)
- Length-based (41 class)

Reduced to binary (using argmax)

Ensembles

Standard and snapshot ensembles





Results

Experiment	Without Ensemble	Ensemble
Binary	0.780	0.795
3-class	0.750	0.754

Results

Experiment	Without Ensemble	Ensemble
Binary	0.780	0.795
3-class	0.750	0.754
5-class	0.790	0.796
6-class	0.771	0.791

Results

Experiment	Without Ensemble	Ensemble
Binary	0.780	0.795
3-class	0.750	0.754
5-class	0.790	0.796
6-class	0.771	0.791
41-class	0.806	0.826
Snapshot	0.796	0.810
Length-based	**0.820**	**0.827**

Results - zeroGPT

Experiment	F1-score
zeroGPT	0.715
Binary	0.797
3-class	0.767
5-class	0.814
6-class	0.805
41-class	0.836
Snapshot	0.826
Length-based	**0.845**

Results - zeroGPT

Experiment	F1-score
zeroGPT	0.715
Binary	0.797
3-class	0.767
5-class	0.814
6-class	0.805
41-class	0.836
Snapshot	0.826
Length-based	**0.845**

Experiment	F1-score
zeroGPT	0.615
Binary	0.794
3-class	0.830
5-class	**0.841**
6-class	0.838
41-class	0.831
Snapshot	0.812
Length-based	0.807

Summary

- It is worth considering the problem as non-binary
- The ensemble methods helped in all the experiments
- Length-based expert models achieved the best result

Széchenyi Plan Plus Let's build Hungary together!

Government of Hungary

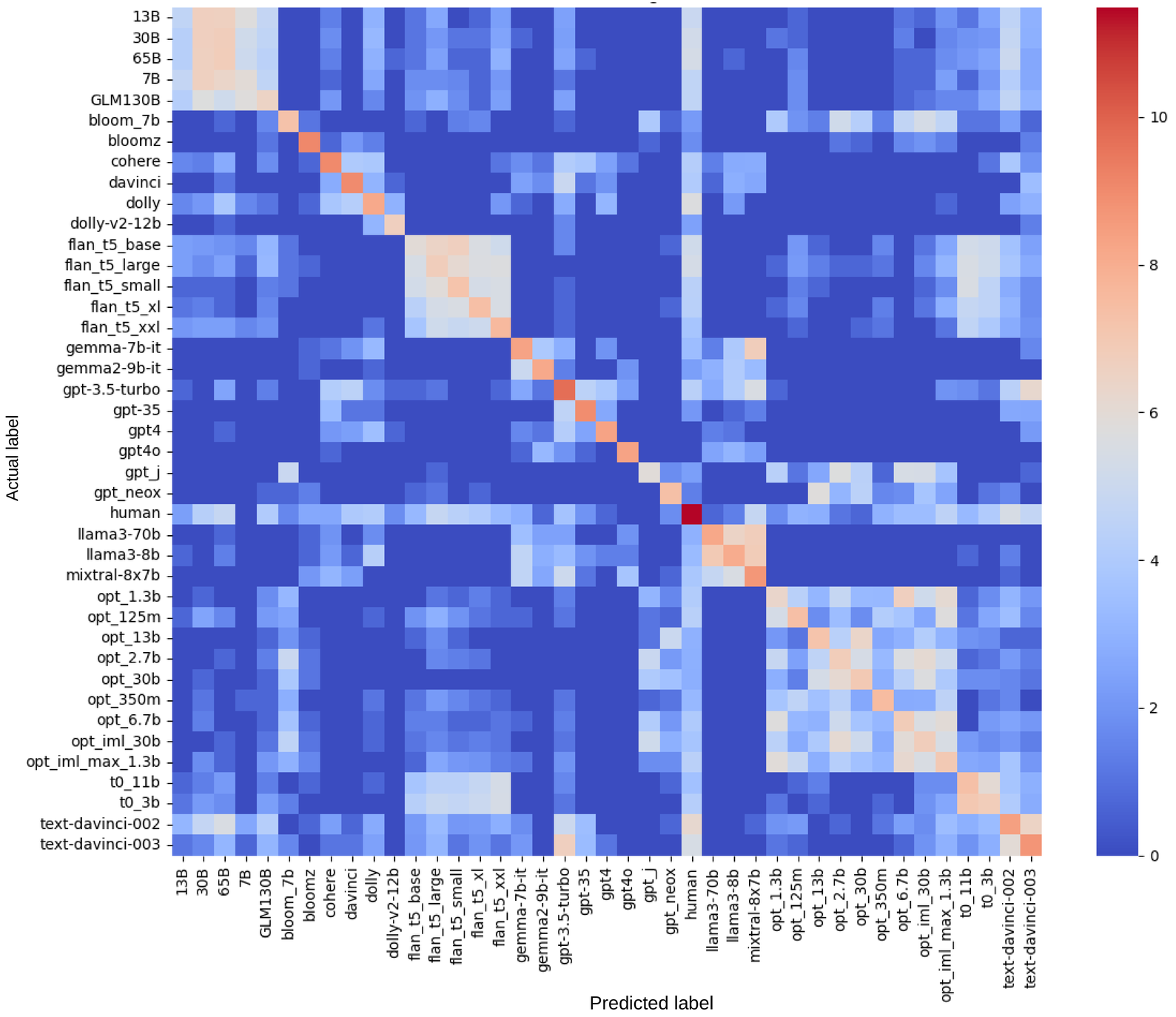
Funded by the European Union NextGeneration EU

ARTIFICIAL INTELLIGENCE National Laboratory

Thank you for your attention!

The research received support from the European Union project RRF-2.3.1-21-2022-00004 within the framework of the Artificial Intelligence National Laboratory. Additionally, we are grateful for the possibility to use ELKH Cloud (https://science-cloud.hu/) which helped us in achieving the results published in this paper.

SZTE University of Szeged


Actual label
Predicted label
13B
30B
65B
7B
GLM130B
bloom_7b
bloomz
cohere
davinci
dolly
dolly-v2-12b
flan_t5_base
flan_t5_large
flan_t5_small
flan_t5_xl
flan_t5_xxl
gemma-7b-it
gemma2-9b-it
gpt-3.5-turbo
gpt-35
gpt4
gpt4o
gpt_j
gpt_neox
human
llama3-70b
llama3-8b
mixtral-8x7b
opt_1.3b
opt_125m
opt_13b
opt_2.7b
opt_30b
opt_350m
opt_6.7b
opt_iml_30b
opt_iml_max_1.3b
t0_11b
t0_3b
text-davinci-002
text-davinci-003
0
2
4
6
8
10