Developing Hungarian language models in the context of global trends

Tamás Váradi

HUN REN | HUNGARIAN RESEARCH CENTRE FOR LINGUISTICS

HUN REN | HUNGARIAN RESEARCH CENTRE FOR LINGUISTICS

Introduction

- Leading Centre for Linguistics for 75 years
- 30 year experience in Corpus linguistics
- One of the leading centres of HU language technology
- Leading regional player since 2010
- Coordinators of several EU funded projects in HLT
- Switched to neural net NLP in 2020
- Only player in Hungary to produce HU LLM’s

Technology | Data | Compute

- Own A100 GPU Supercomputer Cluster

European Commission

A PULI család

Pretrained PULI models (LLM)

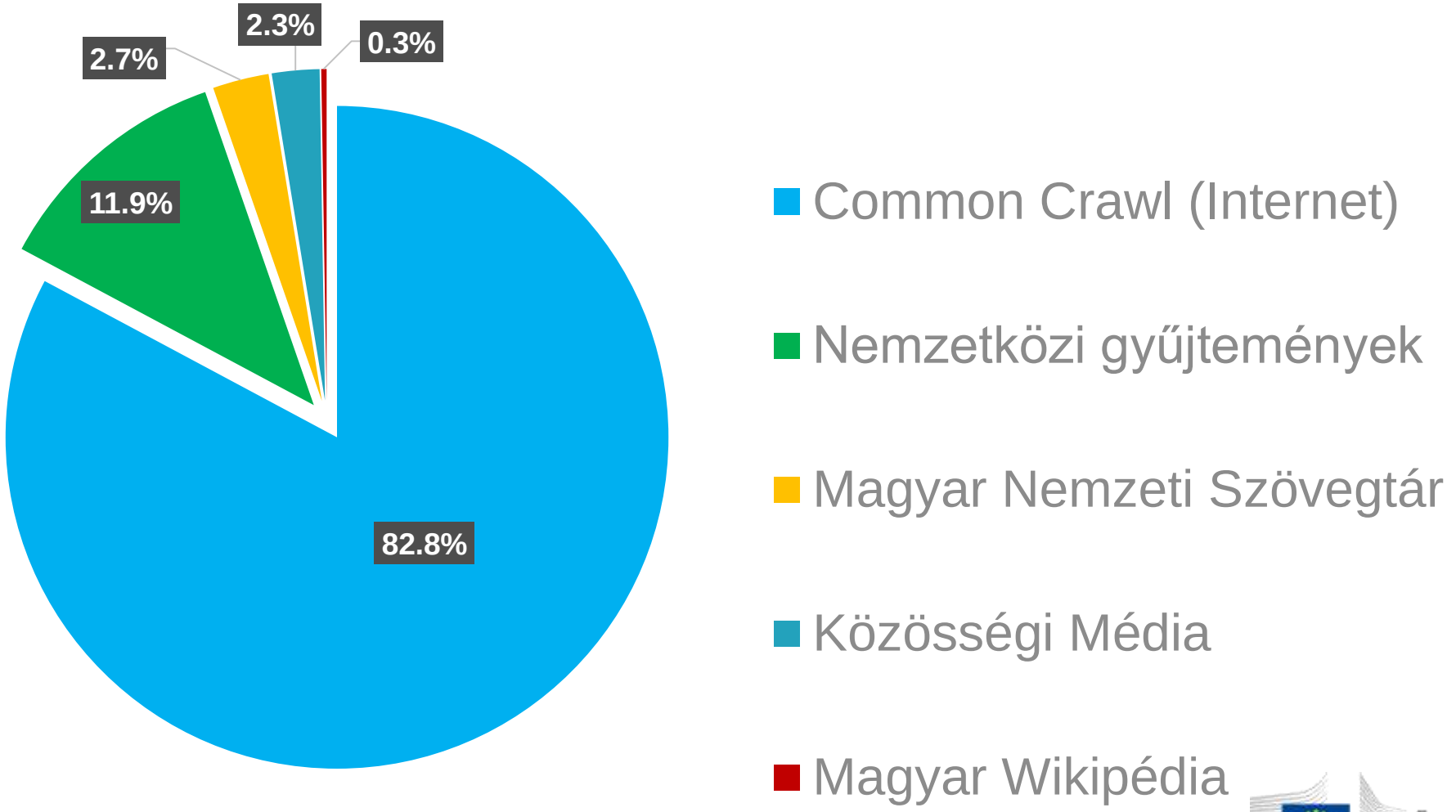
2022: PULI 3SX (a.k.a.GPT-3SX)

- 6,7 B parameters
- Trained Hungarian only
- 32 B HU token Corpus (mainly from Common Crawl text)

2023: PULI Trio (GPTrio)

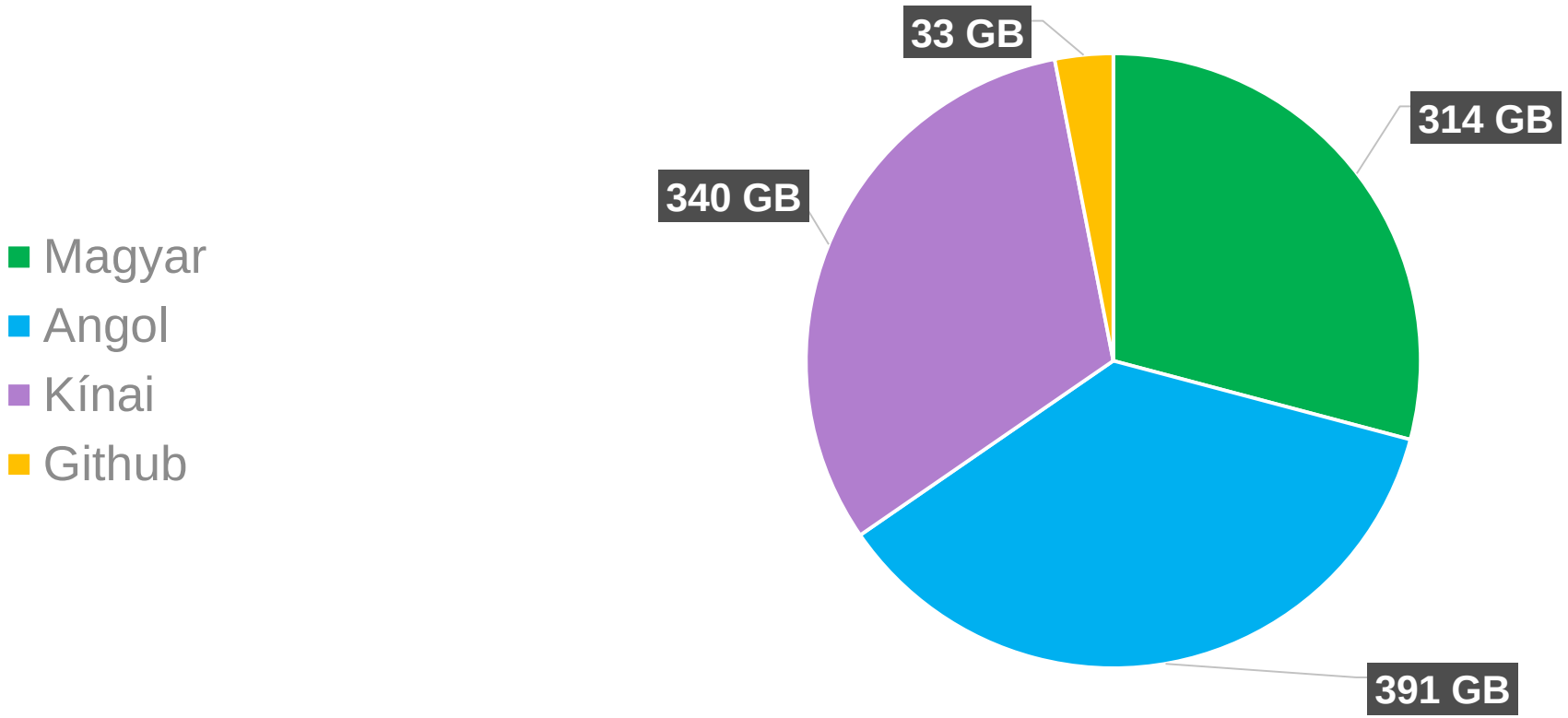
- 7,7 B parameters
- HU, EN, ZH trilingual
- 41 B HU token corpus

Hungarian training data: ~41 milliárd szó



European Commission

Trilingual Corpus: >150 B words

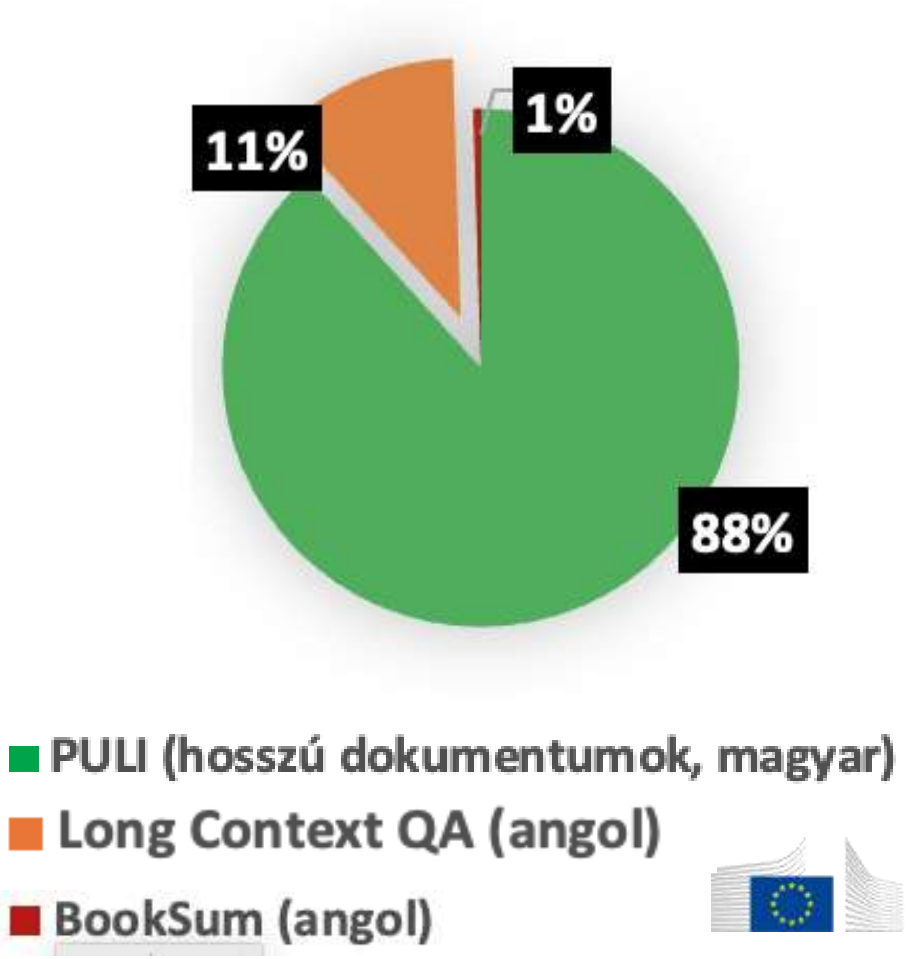


European Commission

HU tuned PULI modell (LLM)

2024: PULI Llumix 32K

- 6,7 B parameters Llama-2
- 2 B multilingual tokens including 600m? HU tokens
- 8 B HU words

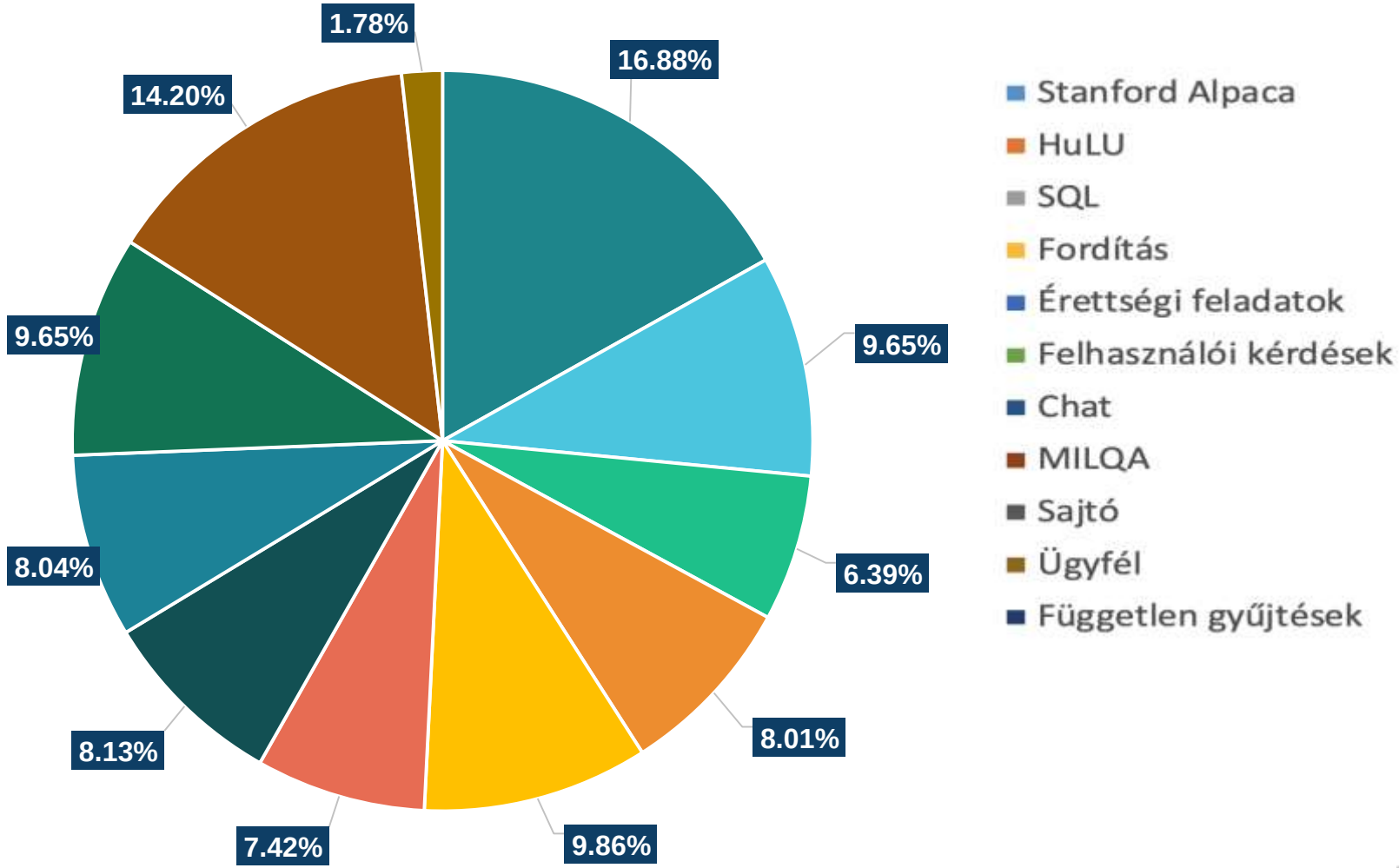


European Commission

Instruct PULI modellek

- ParancsPULI
  - 7,7 milliárd paraméteres utasításkövető (instruct) GPTrio
- PULI LlumiX 32K Instruct
  - 6,7 milliárd paraméteres utasításkövető Llama-2

PULI Llumix 32K Instruct (12440 Hungarian prompts)



European Commission

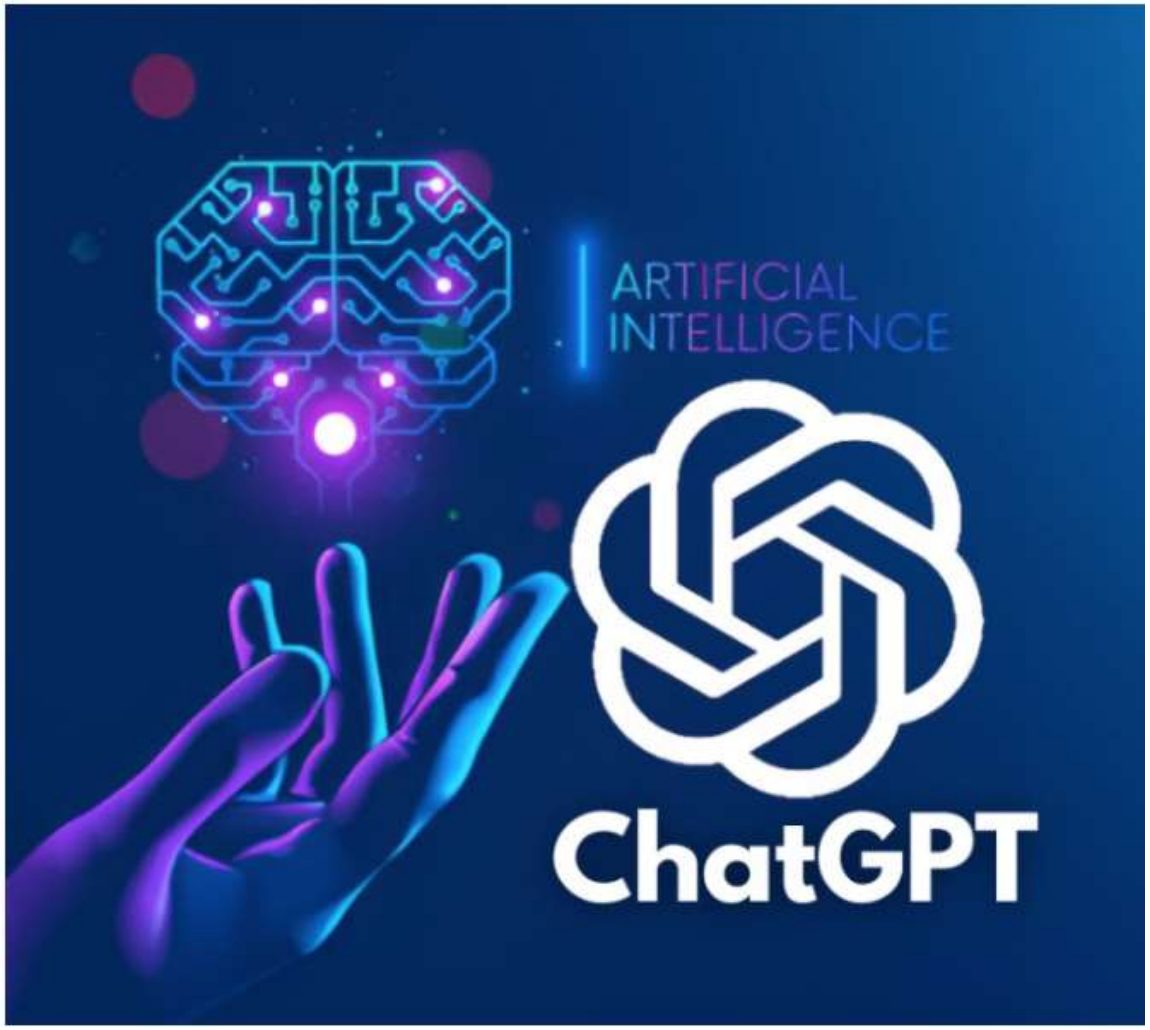
How well do LLM's know Hungarian?

ChatGPT in Hungarian

- **GPT-3 training corpus:**

Rank	Lang	No. of words	% share
1.	En	181 014 683 608	92,65
2.	Fr	3 553 061 536	1,82
…	…	…	…
19.	Hu	127 224 375	0.065

https://seo.ai/blog/how-many-languages-does-chatgpt-support



European Commission

PULI vs GPT-3

PULI
(ParancsPULI)
~billion parameter
~150 billion words
~41 billion Hu words

GPT-3
(ChatGPT)
~175 billion parameter
~200 billion words
~127 million words

European Commission

Language map of Llama-2

2 Trillion x 0,03% = 600 m Hu tokens

Language	Percent	Language	Percent
en	89.70%	uk	0.07%
unknown	8.38%	ko	0.06%
de	0.17%	ca	0.04%
fr	0.16%	sr	0.04%
sv	0.15%	id	0.03%
zh	0.13%	cs	0.03%
es	0.13%	fi	0.03%
ru	0.13%	hu	0.03%
nl	0.12%	no	0.03%
it	0.11%	ro	0.03%
ja	0.10%	bg	0.02%
pl	0.09%	da	0.02%
pt	0.09%	sl	0.01%
vi	0.08%	hr	0.01%

Language map of mC4 corpus

39 B Hu tokens

ISO Code	Language	Tokens (B)	Pages (M)	mT5 (%)
en	English	2,733	3,067	5.67
ru	Russian	713	756	3.71
es	Spanish	433	416	3.09
de	German	347	397	3.05
fr	French	318	333	2.89
it	Italian	162	186	2.43
pt	Portuguese	146	169	2.36
pl	Polish	130	126	2.15
nl	Dutch	73	96	1.98
tr	Turkish	71	88	1.93
ja	Japanese	164	87	1.92
vi	Vietnamese	116	79	1.87
id	Indonesian	69	70	1.80
cs	Czech	63	60	1.72
zh	Chinese	39	55	1.67
fa	Persian	52	54	1.67
ar	Arabic	57	53	1.66
sv	Swedish	45	49	1.61
ro	Romanian	52	46	1.58
el	Greek	43	42	1.54
uk	Ukrainian	41	39	1.51
hu	Hungarian	39	37	1.48

CulturaX Corpus: 43 B Hu tokens

Code	Language	#Documents (M)						#Tokens	
		Initial	URL Filtering	Metric Filtering	MinHash Dedup	URL Dedup	Filtering Rate (%)	(B)	(%)
en	English	5783.24	5766.08	3586.85	3308.30	3241.07	43.96	2846.97	45.13
ru	Russian	1431.35	1429.05	922.34	845.64	799.31	44.16	737.20	11.69
es	Spanish	844.48	842.75	530.01	479.65	450.94	46.60	373.85	5.93
de	German	863.18	861.46	515.83	447.06	420.02	51.34	357.03	5.66
fr	French	711.64	709.48	439.69	387.37	363.75	48.89	319.33	5.06
zh	Chinese	444.37	444.03	258.35	222.37	218.62	50.80	227.06	3.60
it	Italian	406.87	406.04	254.72	226.42	211.31	48.06	165.45	2.62
pt	Portuguese	347.47	346.76	217.21	200.11	190.29	45.24	136.94	2.17
pl	Polish	270.12	269.73	170.86	151.71	142.17	47.37	117.27	1.86
ja	Japanese	247.67	247.19	137.88	114.64	111.19	55.11	107.87	1.71
vi	Vietnamese	182.88	182.72	118.67	108.77	102.41	44.00	98.45	1.56
nl	Dutch	238.92	238.56	148.19	125.51	117.39	50.87	80.03	1.27
ar	Arabic	132.88	132.65	84.84	77.65	74.03	44.29	69.35	1.10
tr	Turkish	183.65	183.47	109.94	99.18	94.21	48.70	64.29	1.02
cs	Czech	136.91	136.44	80.38	69.01	65.35	52.27	56.91	0.90
fa	Persian	118.55	118.50	70.26	62.42	59.53	49.78	45.95	0.73
hu	Hungarian	88.59	88.21	53.29	46.89	44.13	50.19	43.42	0.69
el	Greek	100.77	100.68	61.43	54.33	51.43	48.96	43.15	0.68
ro	Romanian	89.37	89.25	45.99	42.8	40.33	54.87	39.65	0.63

CulturaX Corpus

ro	Romanian	89.37	89.25	45.99	42.8	40.33	54.87	39.65	0.63
sv	Swedish	103.04	102.76	58.67	52.09	49.71	51.76	38.49	0.61
uk	Ukrainian	81.50	81.44	50.95	47.12	44.74	45.10	38.23	0.61
fi	Finnish	59.85	59.80	36.69	32.15	30.47	49.09	28.93	0.46
ko	Korean	46.09	45.85	25.19	21.17	20.56	55.39	24.77	0.39
da	Danish	53.16	52.99	28.67	26.48	25.43	52.16	22.92	0.36
bg	Bulgarian	47.01	46.90	28.09	25.45	24.13	48.67	22.92	0.36
no	Norwegian	40.07	40.01	20.69	19.49	18.91	52.81	18.43	0.29
hi	Hindi	35.59	35.50	22.01	20.77	19.67	44.73	16.79	0.27
sk	Slovak	40.13	39.95	22.20	19.56	18.58	53.70	16.44	0.26
th	Thai	49.04	48.96	26.20	21.93	20.96	57.26	15.72	0.25
lt	Lithuanian	27.08	27.01	15.87	14.25	13.34	50.74	14.25	0.23
ca	Catalan	31.13	31.12	18.99	16.46	15.53	50.11	12.53	0.20
id	Indonesian	48.08	48.05	25.79	23.74	23.25	51.64	12.06	0.19
bn	Bangla	20.90	20.85	13.82	13.22	12.44	40.48	9.57	0.15
et	Estonian	16.20	16.15	9.69	8.45	8.00	50.62	8.81	0.14
sl	Slovenian	15.46	15.39	8.00	7.60	7.34	52.52	8.01	0.13
lv	Latvian	14.14	14.09	8.37	7.48	7.14	49.50	7.85	0.12
he	Hebrew	10.78	10.77	5.90	4.77	4.65	56.86	4.94	0.08
sr	Serbian	7.80	7.75	4.80	4.25	4.05	48.08	4.62	0.07
ta	Tamil	8.77	8.75	5.27	4.94	4.73	46.07	4.38	0.07
sq	Albanian	9.40	9.38	5.96	5.04	5.21	44.57	3.65	0.06
az	Azerbaijani	9.66	9.65	5.73	5.24	5.08	47.41	3.51	0.06
Total (42 languages)		**13397.79**	**13366.17**	**8254.28**	**7471.48**	**7181.40**	**46.40**	**6267.99**	**99.37**
Total (167 languages)		**13506.76**	**13474.94**	**8308.74**	**7521.23**	**7228.91**	**46.48**	**6308.42**	**100.00**

Conclusions so far

- The latest multilingual LLM’s know Hungarian pretty well
- Pretrained global LLM’s see enough data in Hungarian
- Continued training in HU require much less data
- Instruct tuning require smaller datasets, which require more efforts to produce

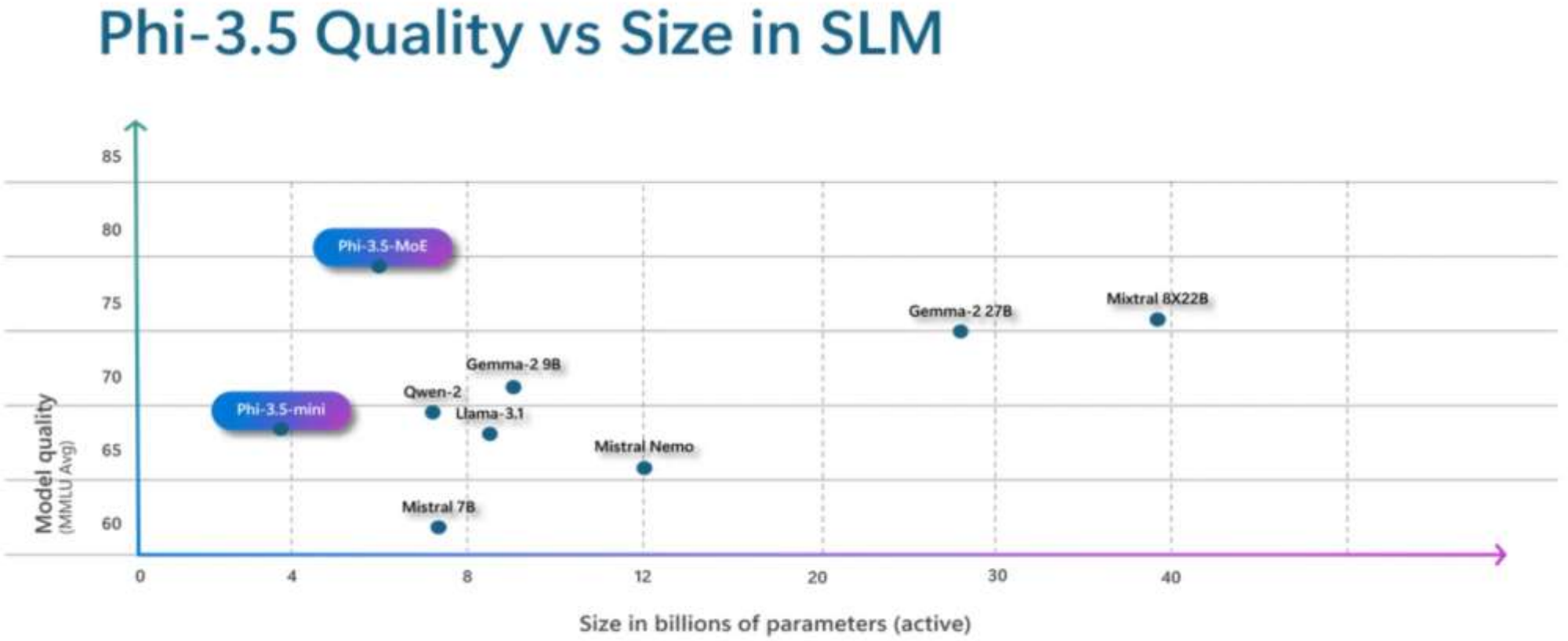
The rationale to develop Hungarian LLM’s

- Sovereignity from global tech players
- Data security
- Linguistic competence
- Cultural alignment
- Responsible, explanable AI
- Cost (!)

New global trends

European Commission

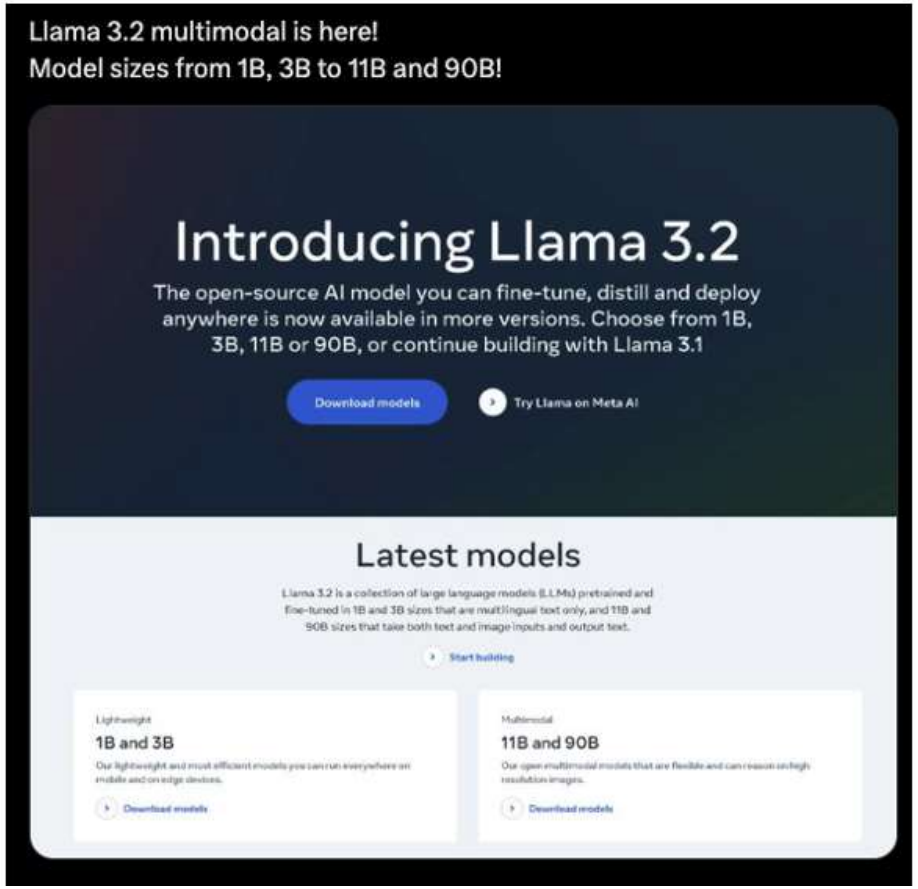
Small but smart models



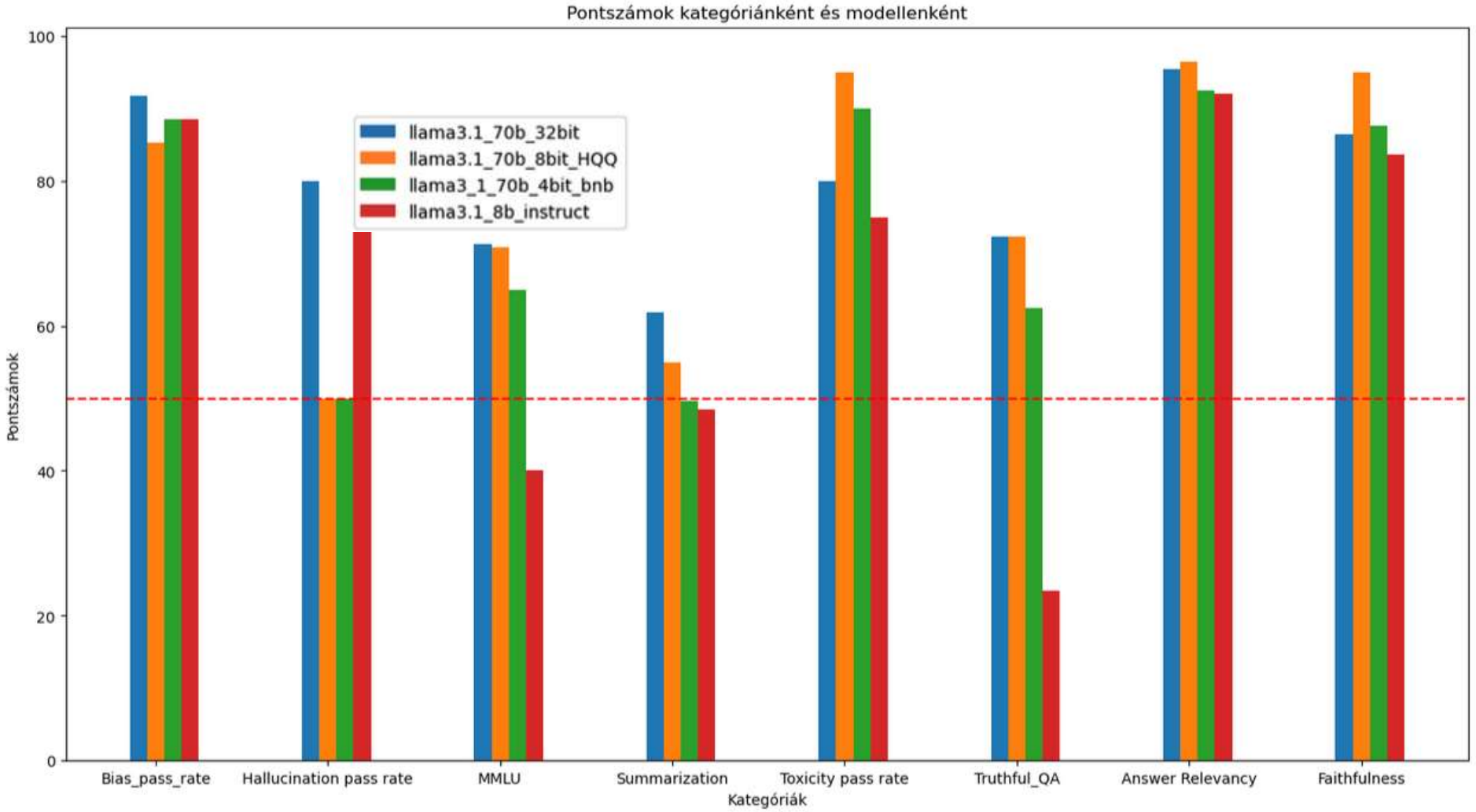
https://techcommunity.microsoft.com/t5/ai-azure-ai-services-blog/discover-the-new-multi-lingual-high-quality-phi-3-5-slms/ba-p/4225280

Open source models

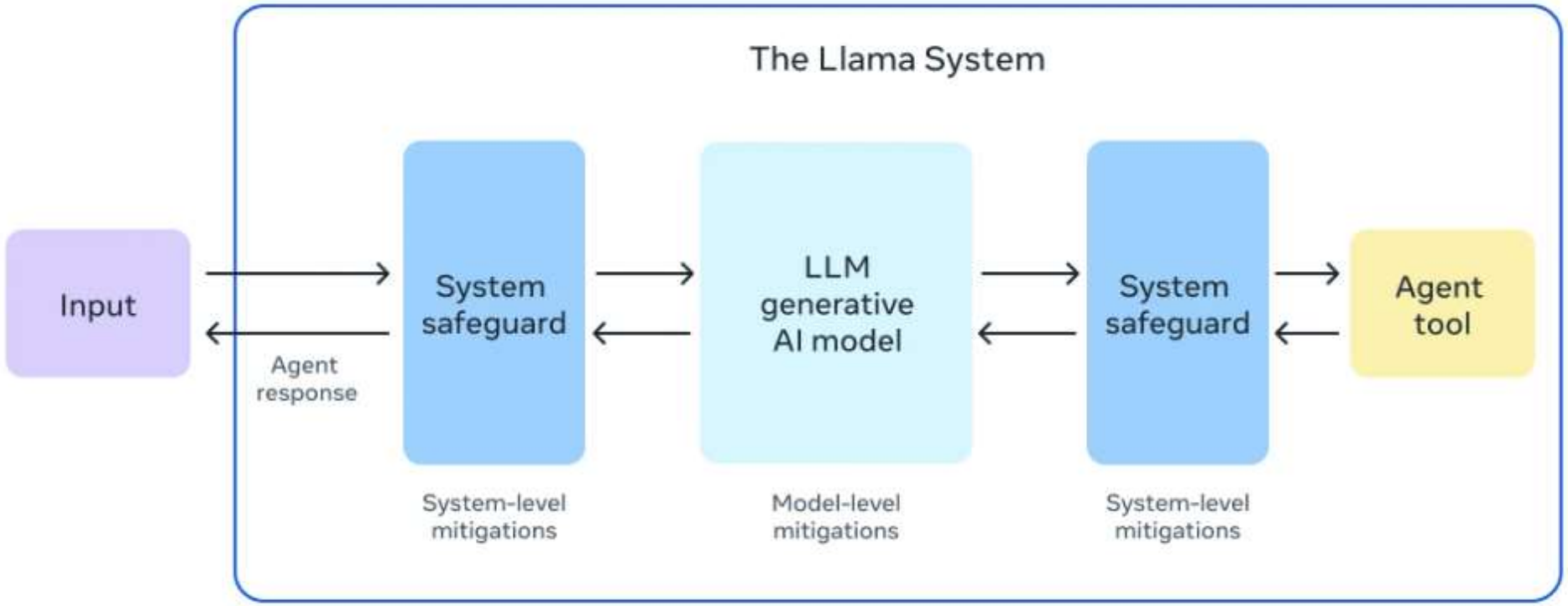
Llama!!

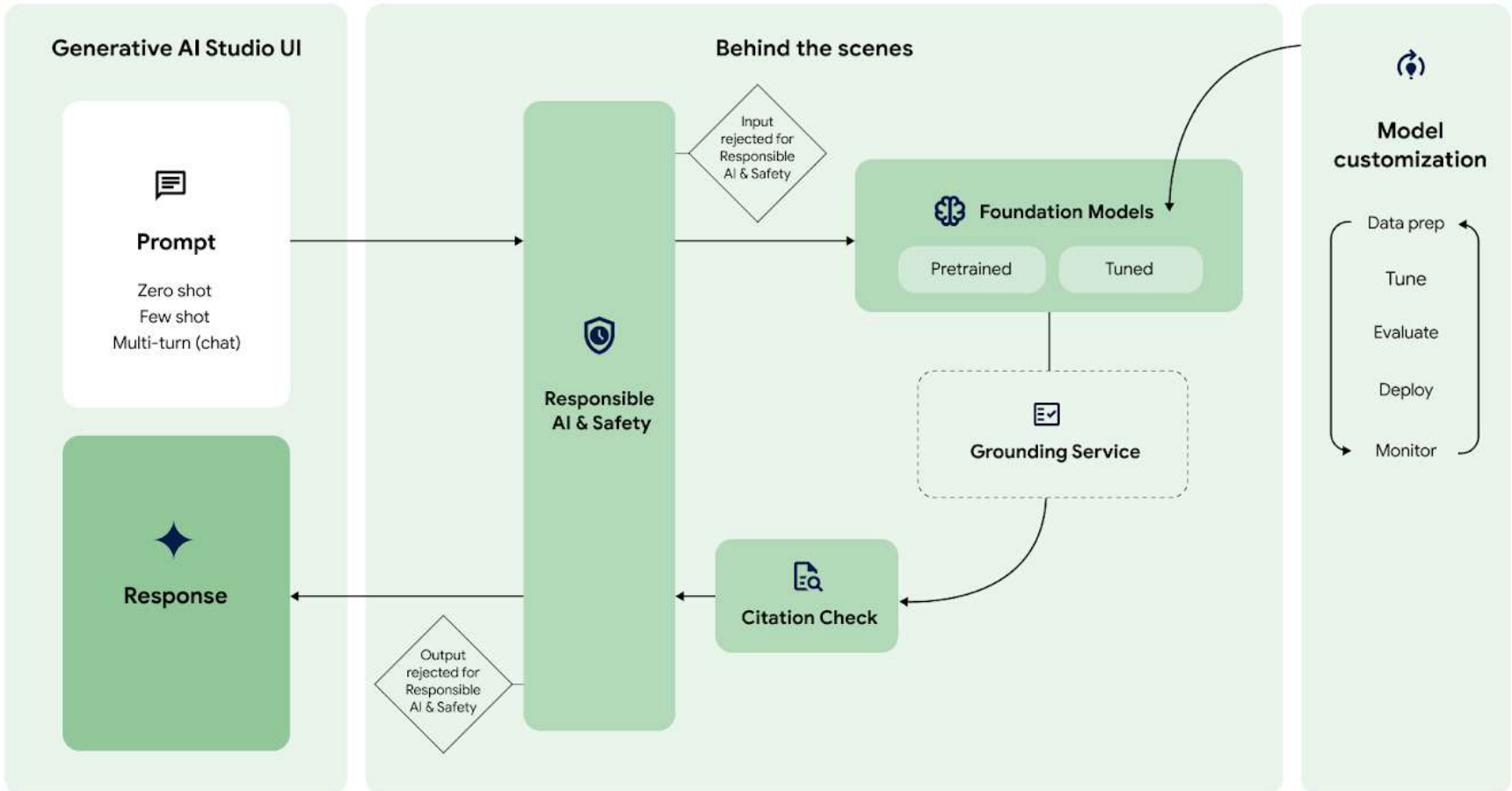


quantized models



Llama Guard 3





Evaluation

Evaluation

European Commission

What makes LLM's smart

- Pre-training (attention, "find the next word") → general linguistic competence
- Instruct tuning (RLHF, DPO) → alignment
- Cross-linguistic transfer → multilinguality

HuLU (Hungarian Language Understanding Benchmark Kit)

- Hungarian benchmark to evaluate Hungarian language models (modelled on GLUE)
- https://hulu.nytud.hu
- 7 benchmarks:
  - HuCB (A CommitmentBank Corpus magyar változata)
  - HuCOLA (Elfogadhatósági ítéletek korpusza)
  - HuCoPa (A hihető alternatívák korpusza)
  - HuRTE (Következtetések felismerésének korpusza)
  - HuSST (A Stanford Sentiment Treebank magyar változata)
  - HuWNLI (Anafora-feloldási korpusz)
  - HuRC (Reading Comprehension with Commonsense Reasoning)

Evaluation

Few-shot – base models

	HuCOLA	HuSST	HuRTE
PULI GPT-3SX	54.27	64.27	57.42
PULI GPTrio	52.71	61.58	54.54
PULI LlumiX 32K	**57.66**	**76.89**	**66.98**
SambaLingo Hungarian Base	56.96	76.55	51.25

Evaluation

Zero-shot (instruct models)

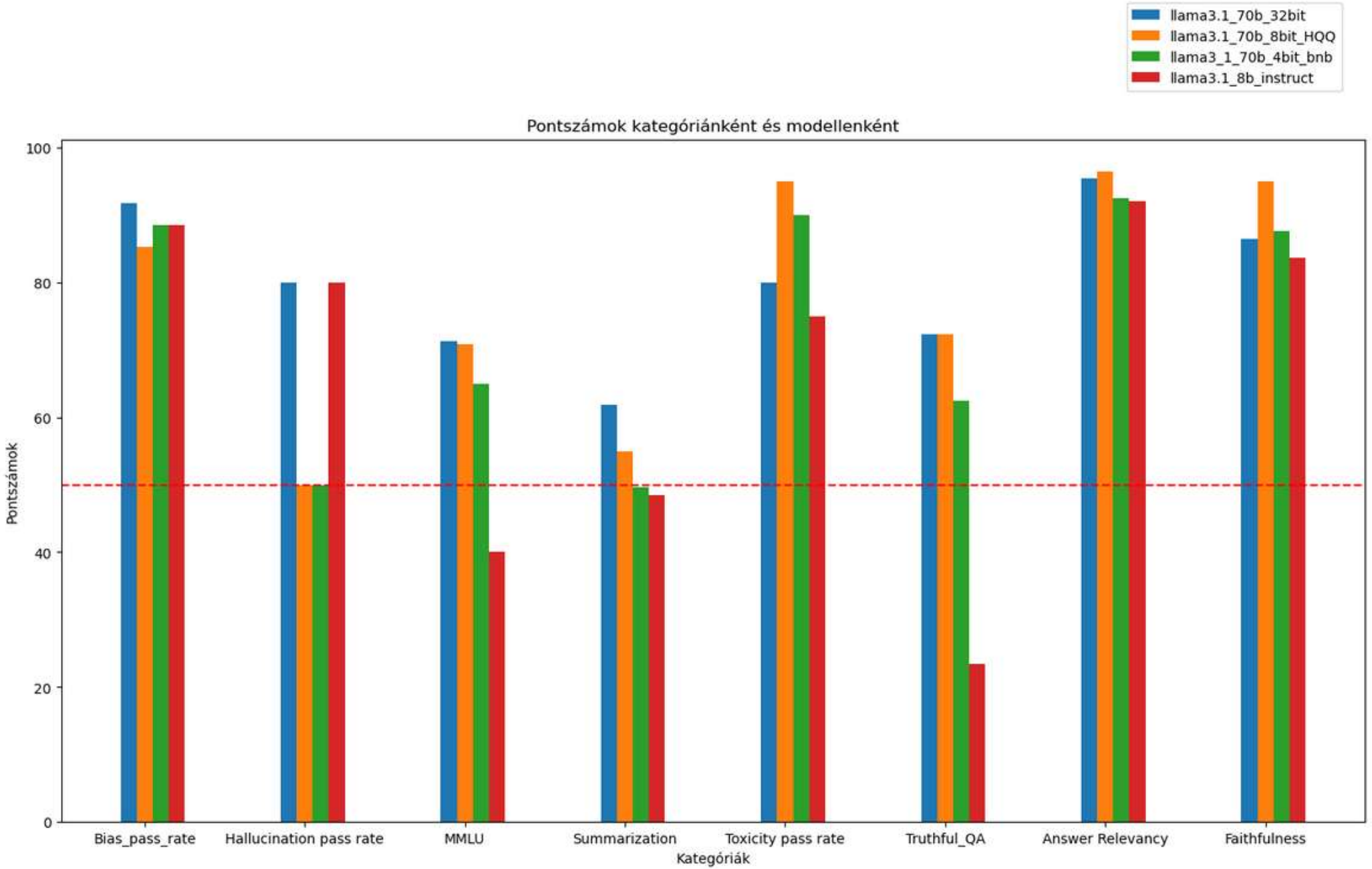
	HuCOLA	HuSST	HuRTE
PULI GPT-3SX Instruct	61.76	46.27	52.09
PULI GPTrio Instruct	52.12	59.20	58.14
PULI LlumiX 32K Instruct	**62.41**	**69.60**	**72.58**
SambaLingo-Hungarian-Chat	53.06	55.15	60.98
ChatGPT (turbo 3.5)	49.10	36.99	50.26
text-davinci-001	50.78	35.48	49.06

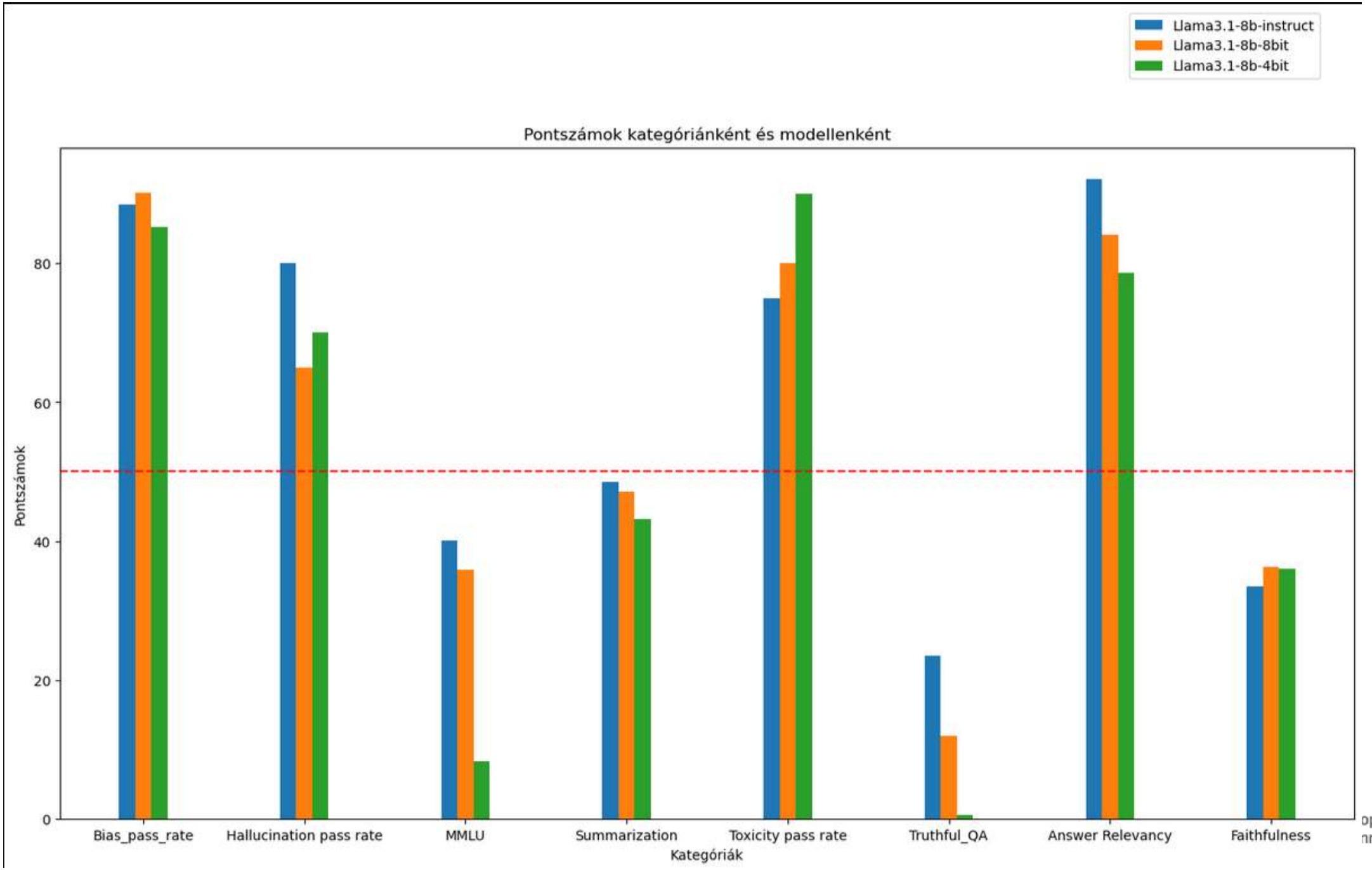
HuGME (Hungarian Generative Model Evaluation)

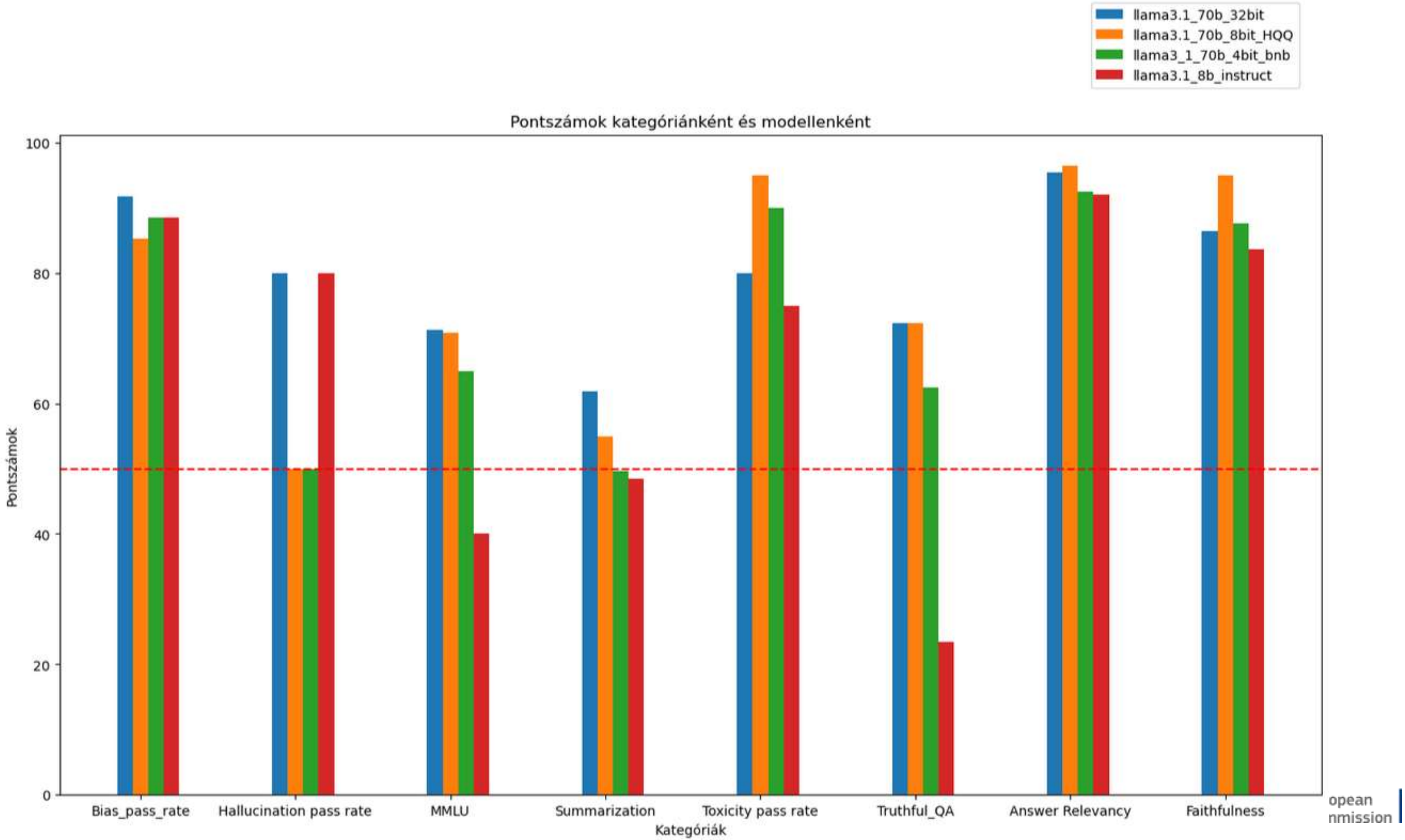
- Summarization
- Answer Relevancy
- Hallucination
- Toxicity
- Bias
- MMLU
- TruthfulQA

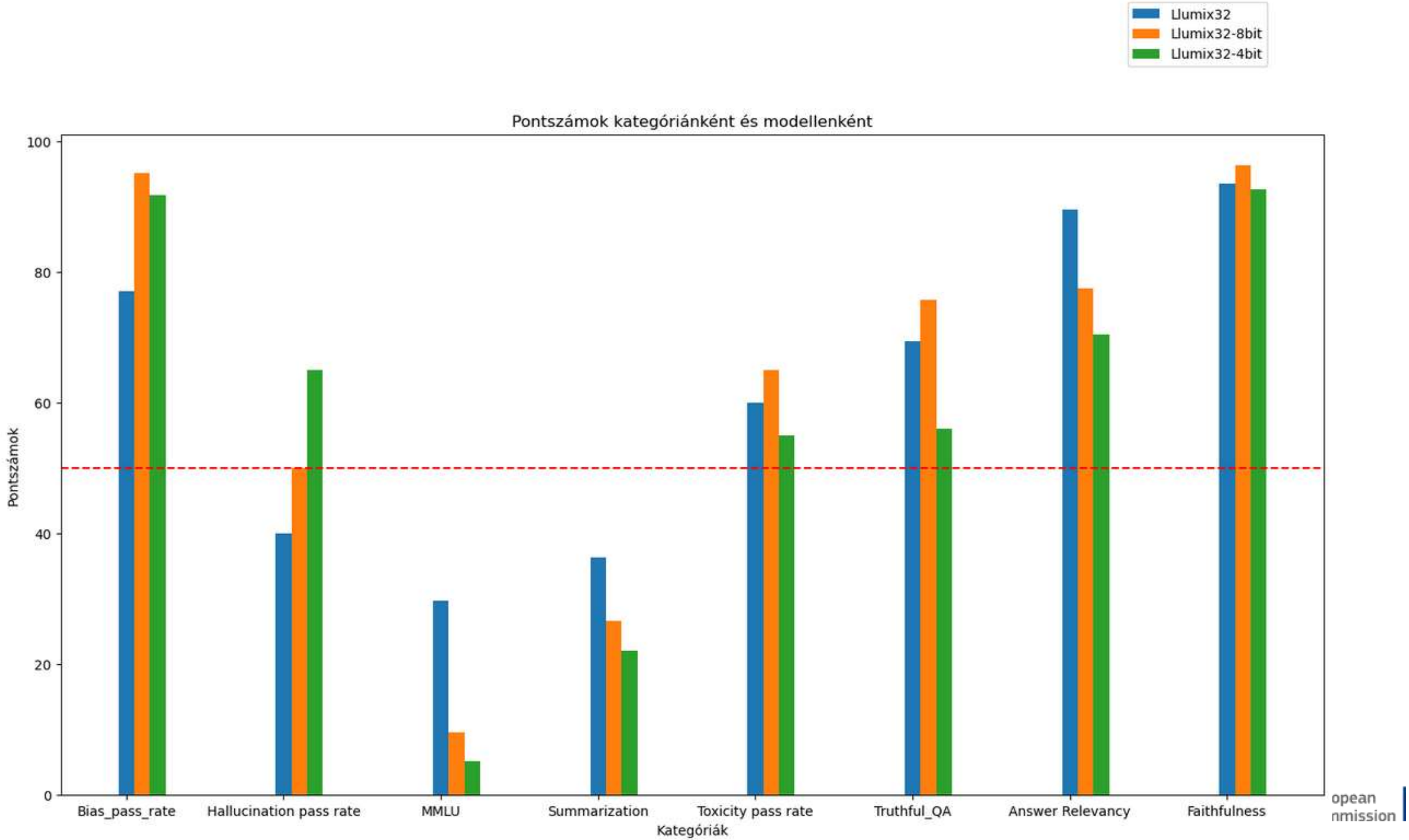
- ~6000 magyar prompt
- Based on DeepEval system
- Work in progress

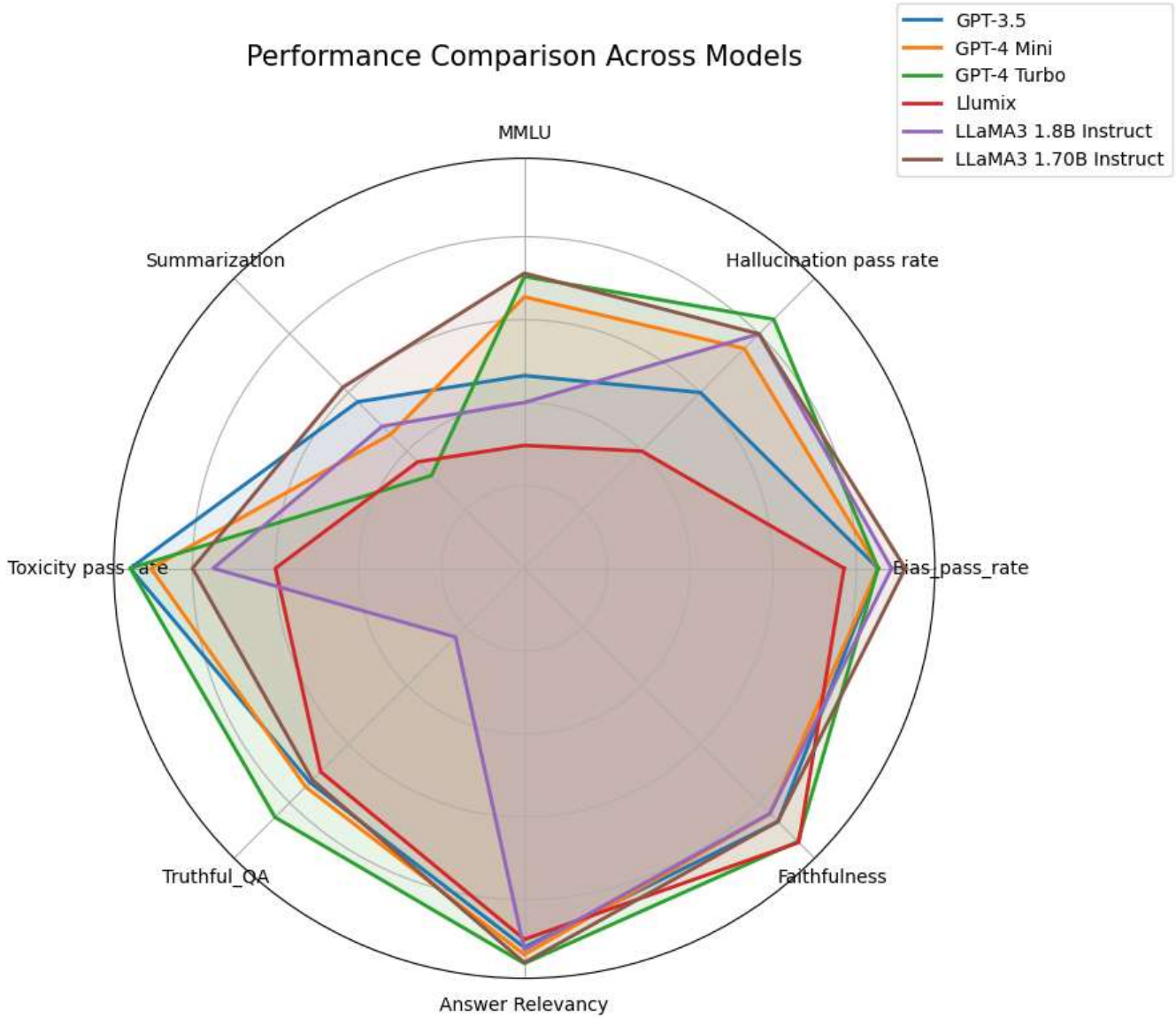
European Commission


llama3.1_70b_32bit
llama3.1_70b_8bit_HQQ
llama3_1_70b_4bit_bnb
llama3.1_8b_instruct
Pontszámok kategóriánként és modellenként
Pontszámok
0
20
40
60
80
100
Bias_pass_rate
Hallucination pass rate
MMLU
Summarization
Toxicity pass rate
Truthful_QA
Answer Relevancy
Faithfulness
Kategóriák

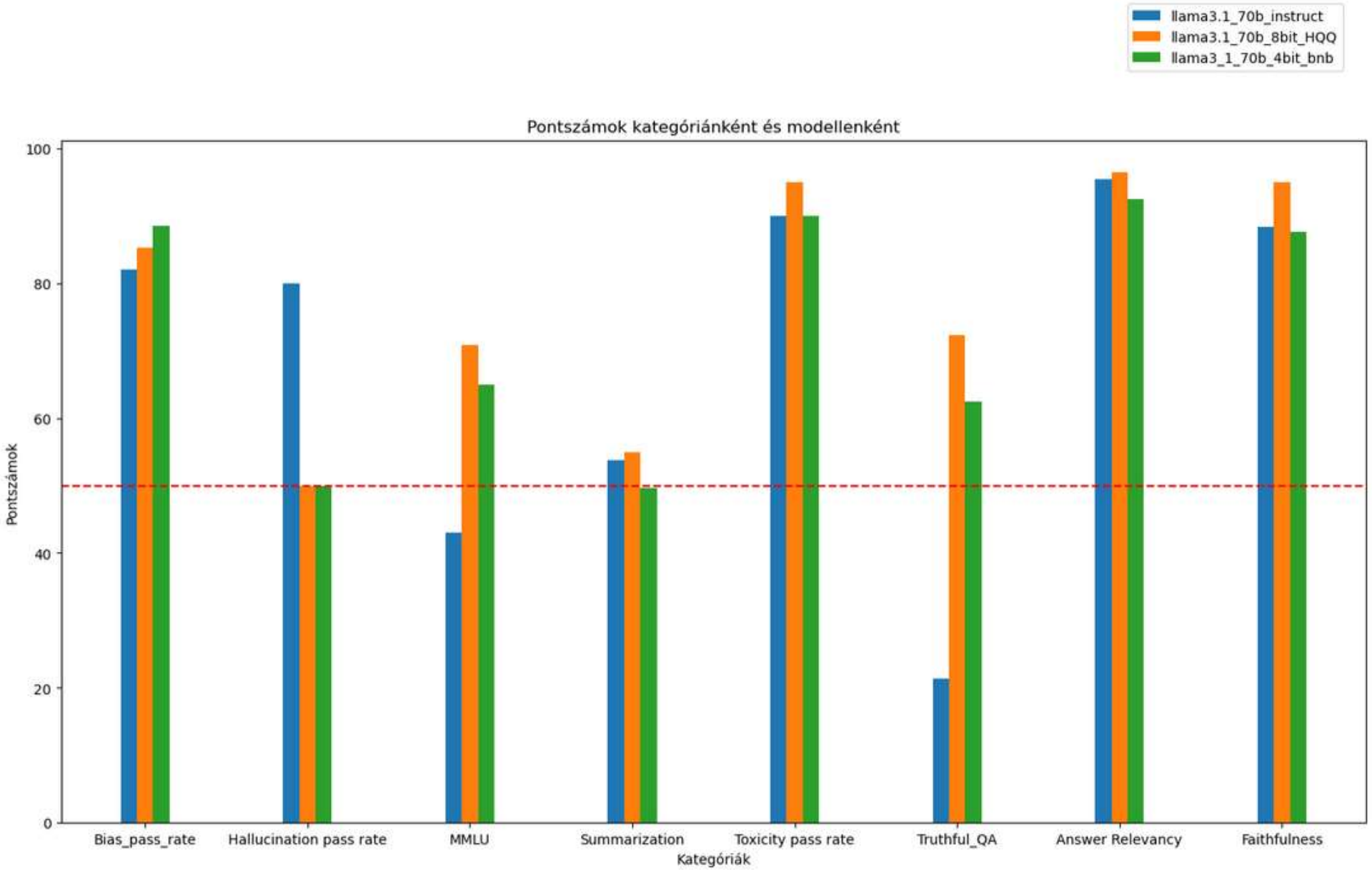

Llama3.1-8b-instruct
Llama3.1-8b-8bit
Llama3.1-8b-4bit
Pontszámok kategóriánként és modellenként
Pontszámok
0
20
40
60
80
Bias_pass_rate
Hallucination pass rate
MMLU
Summarization
Toxicity pass rate
Truthful_QA
Answer Relevancy
Faithfulness
Kategóriák


llama3.1_70b_32bit
llama3.1_70b_8bit_HQQ
llama3_1_70b_4bit_bnb
llama3.1_8b_instruct
Pontszámok kategóriánként és modellenként
Pontszámok
100
80
60
40
20
0
Bias_pass_rate
Hallucination pass rate
MMLU
Summarization
Toxicity pass rate
Truthful_QA
Answer Relevancy
Faithfulness
Kategóriák

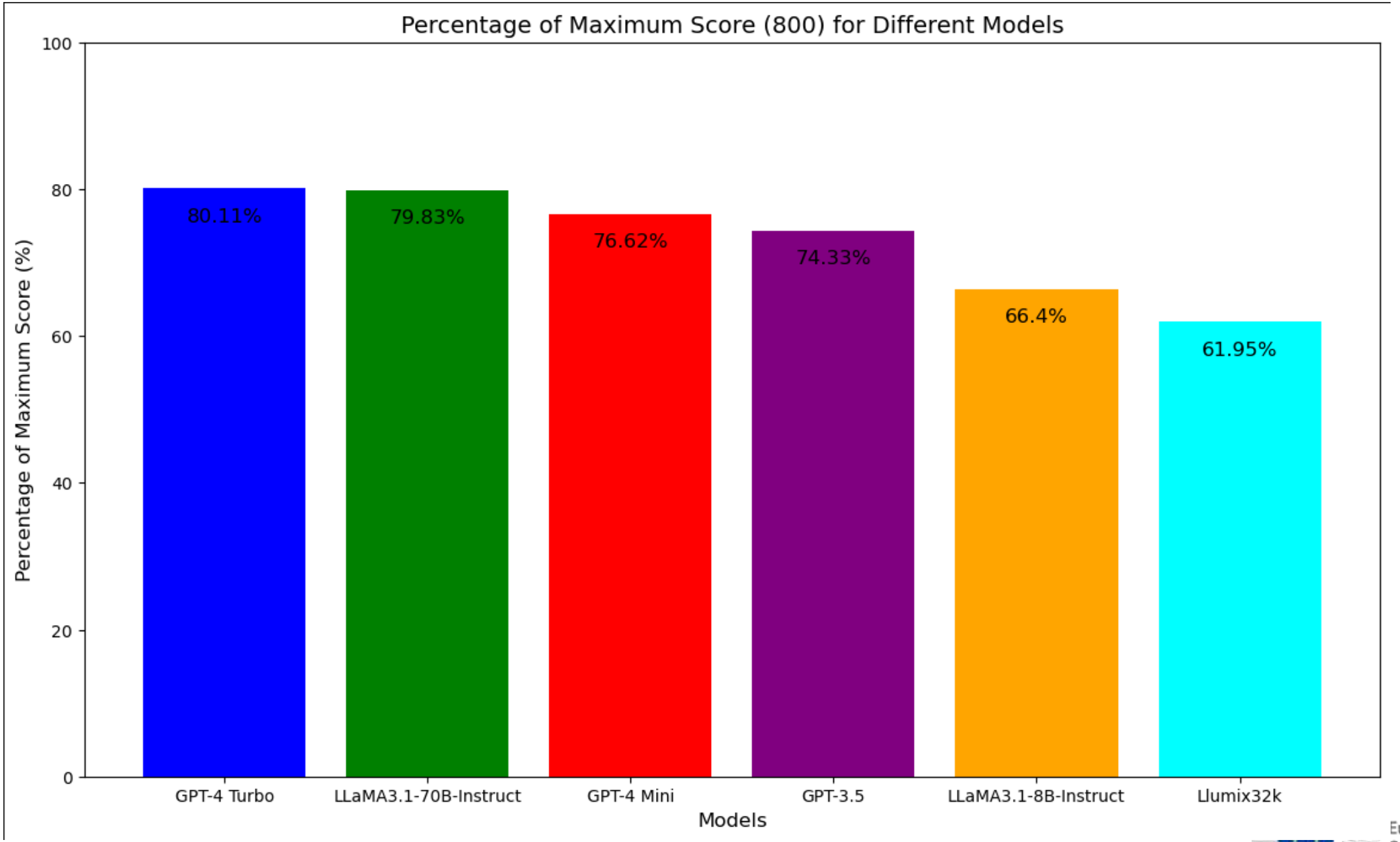

Llumix32
Llumix32-8bit
Llumix32-4bit
Pontszámok kategóriánként és modellenként
Pontszámok
100
80
60
40
20
0
Bias_pass_rate
Hallucination pass rate
MMLU
Summarization
Toxicity pass rate
Truthful_QA
Answer Relevancy
Faithfulness
Kategóriák



European Commission



European Commission


Percentage of Maximum Score (800) for Different Models
Percentage of Maximum Score (%)
100
80
60
40
20
0
80.11%
79.83%
76.62%
74.33%
66.4%
61.95%
GPT-4 Turbo
LLaMA3.1-70B-Instruct
GPT-4 Mini
GPT-3.5
LLaMA3.1-8B-Instruct
Llumix32k
Models

Overall conclusions

The eco-system has changed

- The latest multilingual LLM’s know Hungarian pretty well
- Pretrained global LLM’s see enough data in Hungarian
- Continued training in HU require much less data

The eco-system has changed

- The latest multilingual LLM’s know Hungarian pretty well
- Pretrained global LLM’s see enough data in Hungarian
- Continued training in HU require much less data

Objectives of HU model development must be reviewed

- The rationale for national language model developmen still apply (less so regarding general language competence)
- Pre-training from scratch is not viable
- Multilingual models are more robust anyway (cross-lingual transfer)
- Focus less on general linguistic capabilities than on alignment
- Instruct tuning require smaller datasets, but require more efforts to produce

New goal: bring LLM's to local devices

- Technologies to make LLM's more useful and relevant (RAG, PEFT, LORA)
- Quantized models

→ Make LLM's more accessible and adaptible locally

Thank you for your attention

varadi.tamas@nytud.hun-ren.hu

European Commission