

Collecting **cultural-specific** **open-ended** question answering dataset in Czech, Slovak and Ukrainian

Jindra Helcl

March 3



UNIVERSITY
OF OSLO

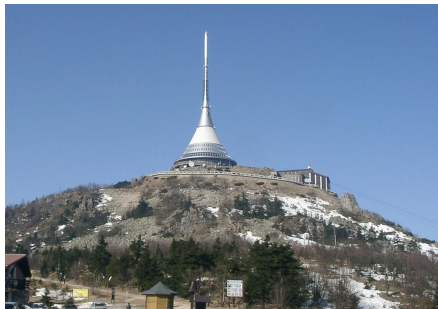


unless otherwise stated

Outline

- Data Collection
- Baselines
- Evaluation
- Lessons Learned

Taste of the Data (Visual, CZ)



Jak se nazývá tato hora?
What is the name of this mountain?



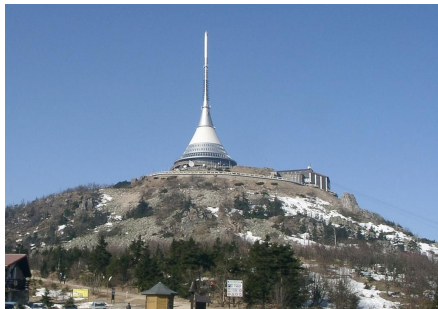
Kde se nachází toto panelové sídliště Doubravka?
Where is this Doubravka housing estate located?



Kdo je muž na obrázku?
Who is the man in the picture?



Taste of the Data (Visual, CZ)



Jak se nazývá tato hora?

What is the name of this mountain?



Ještěd



Kde se nachází toto panelové sídliště Doubravka?

Where is this Doubravka housing estate located?



Kdo je muž na obrázku?

Who is the man in the picture?



Petr Pavel



V Plzni
Pilsen

Taste of the Data (Text, SK+UA)

Kde sa narodil slovenský zbojník Juraj Jánošík?

Where was Juraj Jánošík born?



Ako sa volá najdlhší krytý most na Slovensku?

What is the name of the longest covered bridge in Slovakia?



Ako zahynul Milan Rastislav Štefánik?

How did Milan Rastislav Štefánik die?



Генеральним продюсером якого українського телеканалу свого часу був сьомий президент України Володимир Зеленський?

Volodymyr Zelenskyy, the seventh president of Ukraine, was the general producer of which Ukrainian TV channel at one time?



У якому місті народився український композитор Борис Лятошинський?

In which city was the Ukrainian composer Borys Lyatoshynsky born?



Taste of the Data (Text, SK+UA)

Kde sa narodil slovenský zbojník Juraj Jánošík?

Where was Juraj Jánošík born?



v Terchovej
in Terchová

Ako sa volá najdlhší krytý most na Slovensku?

What is the name of the longest covered bridge in Slovakia?



Kolonádový most
Kolonada bridge

Ako zahynul Milan Rastislav Štefánik?

How did Milan Rastislav Štefánik die?



Havária lietadla.
A plane crash.

Генеральним продюсером якого українського телеканалу свого часу був сьомий президент України Володимир Зеленський?

Volodymyr Zelenskyy, the seventh president of Ukraine, was the general producer of which Ukrainian TV channel at one time?



Телеканалу «Інтер».
Inter TV channel.

У якому місті народився український композитор Борис Лятошинський?

In which city was the Ukrainian composer Borys Lyatoshynsky born?



У Житомирі.
In Zhytomyr.

Phase 1 - Collecting Data

Annotators | Interface | Examples | Statistics

Annotators

- 28 annotators
 - University students (8 Czechs, 11 Slovaks, and 9 Ukrainians)
 - Each spent anywhere between 1 to 20 hours (~11.3 on average)
 - 319 hours of work in total (~140 NOK/h)

	Number of annotators	Number of women	Hours	Text questions	Visual questions	Total questions	Cost per question
Czech	8	4	89	1242	596	1838	14.53
Slovak	11	3	120	1203	329	1532	23.50
Ukrainian	9	1	110	889	740	1629	20.26

Annotation Interface

- Random wiki page with images and links preserved
 - Annotation form
 - Question/answer about the content (local language + translation into English)
 - Image URL (or check no image)
 - Question/answer about the image (local + English)
 - Skip button, links working
-
- Includes lovely features like dashboard

Annotation Interface

Regional QA

Admin

Annotation Instructions

Dashboard

Annotation

Logout jindrich

Question in Czech

Jak se za socialismu jmenovalo Stavovské divadlo?

Answer in Czech

Tylovo divadlo.

Question in English

What was the name of the Estates Theatre during socialism?

Answer in English

The Tyl Theatre.

English translation

Image URL:

Visual question in Czech

Kde se nachází socha na fotce?

Visual answer in Czech

Před Stavovským divadlem v Praze.

Visual question in English

Where is the statue in the photo?

Visual answer in English

In front of the Estates Theatre in Prague.

Comment (optional)

Delete

Submit

Stavovské divadlo

Stavovské divadlo (původní název **Hraběcí Nosticovo národní divadlo**, *Gräflich Nostitzsches Nationaltheater*, za socialismu **Tylovo divadlo**) je divadlo v Praze na **Starém Městě**, budova čp. 540/I na **Ovocném trhu** 11, v městské části **Praha 1**. Stavovské divadlo je úzce spjato s osobou **Wolfganga Amadea Mozarta** a je jediným dosud existujícím divadlem, kde skladatel působil.

Architektura

Budova

Budovu dal postavit jako scénu pro veřejnost **osvětlený vlastenec** **hrabě František Antonín z Nostic-Rienecku**. Plány uspořádání prostor divadla **naskicoval** hrabě **Kašpár Herman Künnigel**, stavbu v letech 1781–1783 projektoval a provedl dvorní stavitel **Anton Haffenecker**. Stavba jedné z prvních **klasicistních** budov pro veřejnost v Praze trvala pouhé dva



undefined

Question about image in the article



miniaturu tohoto divadla na svém letním sídle. A dokonce tyto dvě budovy měly být i propojeny

to velká půlkruhové zakončená **okna** prostupující patry. Budova byla z přední strany od **ocného trhu**. Čelní **rizalit** nebyl rozdělen do dvou pater jako dnes, nýbrž byl uvnitř jeden vysoký

tedy o patro nižší než to dnešní. Disponovalo plochým **parterem**, jak bylo v 18. století zvykem, a

výraznější z nich byla provedena od 23. září do 29. října 1834, kdy došlo (navzdory časové tísní) k a přistavěna druhá galerie. Stěny hlediště byly vyzdobeny štukovými plastikami a rámcí v červeno- do jednoho svažitého. Přizemní pořadí **lůží** bylo propojeno v jednu velkou, průběžnou **galerii**. alo novou masinérii. Mezi lóžemi byl instalováno osvětlení nástěnnými žirandolami a na strop pci které se spouštěly, dole se svíce rozsvěčely i zhasínaly. Nově otevřené divadlo zahajovalo

Druhou a neuvěřitelnější přestavbu Stavovského divadla od dubna do listopadu **1859**, tedy osm měsíců, prováděl stavitel Karel Brüst ve spolupráci s

LTG Seminar, 3 March 2026

Seed Articles

- Annotators primed by selecting local Wiki pages
 - (Skipping and clicking links still allowed!)
- Random Wiki pages selected from seed pool, containing pages
 - Existing only in local language
 - With Wikidata entry geolocated within country
 - (like <https://www.wikidata.org/wiki/Q387782>)

Data Curation

- Collection ran over 2 months
- Manual post-processing
 - Move data to a spreadsheet
 - Mark duplicates (auto+manual)
 - Minor fixes like typos (English and Czech only)
 - Mark bad examples
- Shuffle, split into dev and test
 - Roughly half-half, but each file has different number of lines

Collection Size

Czech, Text (88% kept)

Yes	976	1080
Yes with edits	104	
No	81	145
No; duplicate	64	

Czech, Visual (77% kept)

Yes	421	456
Yes with edits	35	
No	94	136
No; duplicate	42	

Slovak, Text (81% kept)

Yes	929	972
Yes with edits	43	
No	150	225
No; duplicate	75	

Slovak, Visual (73% kept)

Yes	235	238
Yes with edits	3	
No	54	88
No; duplicate	34	

Ukrainian, Text (78% kept)

Yes	748	755
Yes with edits	7	
No	144	213
No; duplicate	69	

Ukrainian, Visual (81% kept)

Yes	333	403
Yes with edits	70	
No	42	92
No; duplicate	50	

Data Augmentation

- Automatic translations for the other two languages
 - (English translations created manually)
 - Claude 3.5 Sonnet (best at WMT'24)
 - Quality confirmed by manual evaluation of three systems (GPT-4o, Google Translate)
- Automatically assigned category
 - One of {Geography, Culture, Politics, History, Sports, Other}
- Answers generated by open-weight models
- Human evaluation

Example (Czech, Textual)

```
{
  "id": 319,
  "question_orig": "Ve kterém městě v severozápadních Čechách se nachází závodní okruh?",
  "question_en": "In which town in Northwest Bohemia is the racing circuit located?",
  "answer_orig": "V Mostě",
  "answer_en": "Most",
  "wikititle": "Autodrom_Most",
  "translations": {
    "questions": {
      "sk": "V ktorom meste v severozápadných Čechách sa nachádza pretekársky okruh?",
      "uk": "У якому місті на північному заході Чехії знаходиться гоночний трек?"
    },
    "answers": {
      "sk": "V Moste",
      "uk": "У Мості"
    }
  },
  "category": "geography",
  "generated_orig": { stay tuned! },
  "generated_en": { stay tuned! },
  "human_eval_orig": { stay tuned! },
  "human_eval_en": { stay tuned! },
}
```

Content Classification

		Textual			Visual		
		CZ	SK	UA	CZ	SK	UA
Categories	Geography	39	42	32	60	50	39
	Culture	27	23	19	21	24	20
	Politics	5	9	12	3	9	9
	History	21	12	26	10	6	20
	Sports	5	8	5	2	3	5
	Other	3	6	4	4	8	5

Categories according to LLaMA 3.3 70B
(percentage, exclusive)

Content Classification

		Textual			Visual		
		CZ	SK	UA	CZ	SK	UA
Categories	Geography	39	42	32	60	50	39
	Culture	27	23	19	21	24	20
	Politics	5	9	12	3	9	9
	History	21	12	26	10	6	20
	Sports	5	8	5	2	3	5
	Other	3	6	4	4	8	5
Basic NER	Location	94	87	85	76	60	57
	Person	46	44	50	32	39	43
	Organization	33	41	45	11	19	31
	Date/Year	8	7	11	1	0	5

Categories according to LLaMA 3.3 70B
(percentage)

Entities in the dataset according to GliNER
(percentage, non-exclusive)

Content Classification

		Textual			Visual		
		CZ	SK	UA	CZ	SK	UA
Categories	Geography	39	42	32	60	50	39
	Culture	27	23	19	21	24	20
	Politics	5	9	12	3	9	9
	History	21	12	26	10	6	20
	Sports	5	8	5	2	3	5
	Other	3	6	4	4	8	5
Basic NER	Location	94	87	85	76	60	57
	Person	46	44	50	32	39	43
	Organization	33	41	45	11	19	31
	Date/Year	8	7	11	1	0	5
Fine-grained named entities	River	9	8	7	4	3	1
	Mountain	14	17	6	10	8	6
	City/Town	56	48	65	43	40	42
	Village	11	13	5	5	9	2
	Movie	3	1	0	0	0	0
	TV show	5	7	3	0	0	1
	Book	7	3	3	2	0	1
	Song	1	2	3	0	0	1
	Actor	6	3	3	1	1	1
	Writer	9	6	8	3	3	5
	Musician	3	6	4	2	10	1
	Other artist	5	6	8	2	5	4
	Politician	13	13	24	3	11	14
	Sportsperson	2	5	2	1	5	2
	Other person	32	24	32	29	27	33
	Sports event	5	9	5	0	1	3
	Political party	3	5	5	0	0	1
	Other org.	21	27	36	6	12	22
	Product/Brand	11	8	13	8	2	14
	Company	4	2	5	0	1	3

Categories according to LLaMA 3.3 70B
(percentage)

Entities in the dataset according to GliNER
(percentage, non-exclusive)

Fine-grained named entities (GliNER)
(percentage, non-exclusive)

More Content Classification (Visual)

		Location				content				medium						
		SK	CZ	UA	portrait	building	city	nature	statue	food	technology	other	has text	color photo	BW photo	painting
		23	31	25	57	30	8	8	2	1	4	18	86	10	7	4
		36	52	36	15	5	4	3	1	0	6	25	92	10	2	10
		27			5	5	4	1	0	1	4	23	79	10	10	

3.2 11B Vision

According to LLaMA 3.2 11B Vision
(percentage)

Phase 2 - Baselines

Models | Example Outputs

Baselines

- Zero-shot decoding with instruction-tuned models

Llama-3.3-70B-Instruct

Llama-3.1-8B-Instruct

EuroLLM-9B-Instruct

Mistral-7B-Instruct-v0.3

Llama-4-Scout-17B-16E-Instruct

Llama-3.2-11B-Vision-Instruct

Llama-4-Scout-17B-16E-Instruct

mayan

Idelfics3-8B-Llama3

gemma-3-12b-it

- Various system prompts / language settings

You are an experienced trivia game contestant.

Provide a short and correct answer in the same language as the question!

- No additional context
 - *Will talk about very basic RAG experiments upon request*

Baselines - Example (Text, SK)

Ako sa volá tajná služba na Slovensku?

What is the name of the secret service in Slovakia?



Llama-3.3-70B-Instruct:

Llama-3.1-8B-Instruct:

EuroLLM-9B-Instruct:

Mistral-7B-Instruct-v0.3:

Llama-4-Scout-17B-16E-Instruct:

Slovenská informačná služba (SIS).

BIS - Binnení služba.

Ako sa volá tajná služba na Slovensku? **Slovenská informačná služba (SIS).**



Já som nebol som tajná služba na Slovensku. Ako by som bol, musel by som zachovávať tajnosť.

Slovenská informačná služba.



Llama-3.3-70B-Instruct:

Llama-3.1-8B-Instruct:

EuroLLM-9B-Instruct:

Mistral-7B-Instruct-v0.3:

Llama-4-Scout-17B-16E-Instruct:

SIS (Slovak Information Service).

Boris Zala's BOR.

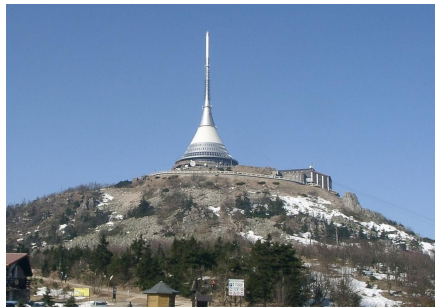
The name of the secret service in Slovakia is the Office for the Protection of Constitutional Democracy (Úrad na ochranu ústavy) in Slovakia, **which is commonly known as the ŠTB.**



The name of the secret service in Slovakia is the **Slovak Information Service (SIS).**

The secret service in Slovakia is called **Slovenská informačná služba (SIS).**

Baselines - Example (Visual, CZ)



Jak se nazývá tato hora?
What is the name of this mountain?

Llama-3.2-11B-Vision-Instruct:

Llama-4-Scout-17B-16E-Instruct:

maya:

Idefics3-8B-Llama3:

gemma-3-12b-it:

The transmitter on the photograph is called Liberec

Ještědský vysílač

Antena na fotce se nazývá "Všechny směry"

Rozhledna

Ježíškovský výštinový vysílač

Baselines - Example (Visual, CZ)



Kdo je muž na obrázku?
Who is the man in the picture?

Llama-3.2-11B-Vision-Instruct:

I'm not comfortable sharing the identity of the person in this image.

Llama-4-Scout-17B-16E-Instruct:

Donald Trump.

maya:

Muž na obrázku je John Kerry.

Idefics3-8B-Llama3:

Muž.

gemma-3-12b-it:

Je to herec John Aniston.

Baselines - Example (Visual, CZ)



Kdo je muž na obrázku?
Who is the man in the picture?

Llama-3.2-11B-Vision-Instruct:

I'm not comfortable sharing the identity of the person in this image.

Llama-4-Scout-17B-16E-Instruct:

Donald Trump.

maya:

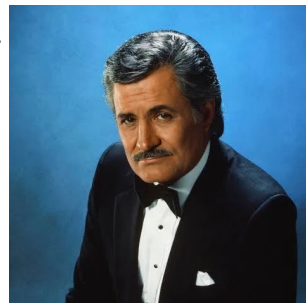
Muž na obrázku je John Kerry.

Idefics3-8B-Llama3:

Muž.

gemma-3-12b-it:

Je to herec John Aniston.



Phase 3 - Evaluation

Manual | Automatic | Results

Human Evaluation - Annotators

- 16 annotators
 - University students (5 Czechs, 7 Slovaks, and 4 Ukrainians)
 - Each spent anywhere between 30 min to 14 hours (~6.7 on average)
 - 106 hours of work in total (~140 NOK/h)
- 2 annotators per example
 - Question, reference answer, 2*5 generated answers (orig, en)
 - Human eval run on dev set, 1956 examples in total

Human Evaluation

Annotators rated each answer in terms of

1. **Correctness**

- The answer is correct given the reference (minor noise allowable)

Human Evaluation

Annotators rated each answer in terms of

1. **Correctness**

- The answer is correct given the reference (minor noise allowable)

2. **Truthfulness**

- The answer is factually true (might not answer the question)
- Refusal labeled as not true

Human Evaluation

Annotators rated each answer in terms of

1. **Correctness**

- The answer is correct given the reference (minor noise allowable)

2. **Truthfulness**

- The answer is factually true (might not answer the question)
- Refusal labeled as not true

3. **Relevance**

- The generated answer contains only information asked for (regardless of whether it is true or not)

Human Evaluation

Annotators rated each answer in terms of

1. **Correctness**

- The answer is correct given the reference (minor noise allowable)

2. **Truthfulness**

- The answer is factually true (might not answer the question)
- Refusal labeled as not true

3. **Relevance**

- The generated answer contains only information asked for (regardless of whether it is true or not)

4. **Coherence**

- The answer is grammatically correct and in the correct language

Automatic Evaluation

Metrics used for open-ended generation

- n-gram overlap
 - chrF
 - BLEU
 - ROUGE-L
- semantic
 - BertSCORE
 - BLEURT
- LLM as a judge
 - LLaMA 3.3 70B
 - M-Prometheus

Automatic Evaluation

Metrics used for open-ended generation

- n-gram overlap
 - chrF
 - BLEU
 - ROUGE-L
- semantic
 - BertSCORE
 - BLEURT
- LLM as a judge
 - LLaMA 3.3 70B
 - M-Prometheus

System-level correlations

	Textual							Visual						
	BLEU	chrF	RG	BScr	BLRT	LLM	M-P	BLEU	chrF	RG	BScr	BLRT	LLM	M-P
Correct	.733	.867	.833	.817	.883	.950	.950	.667	.550	.761	-.267	.817	1.000	.962
+ true	.753	.879	.845	.829	.879	.970	.970	.667	.550	.761	-.267	.817	1.000	.962
+ relevant	.783	.917	.883	.867	.917	.967	.967	.667	.550	.761	-.267	.817	1.000	.962
+ coherent	.783	.917	.883	.867	.917	.967	.967	.629	.695	.773	-.215	.829	.920	.858

Automatic Evaluation

Metrics used for open-ended generation

- n-gram overlap
 - chrF
 - BLEU
 - ROUGE-L
- semantic
 - BertSCORE
 - BLEURT
- LLM as a judge
 - LLaMA 3.3 70B
 - M-Prometheus

System-level correlations

	Textual							Visual						
	BLEU	chrF	RG	BScr	BLRT	LLM	M-P	BLEU	chrF	RG	BScr	BLRT	LLM	M-P
Correct	.733	.867	.833	.817	.883	.950	.950	.667	.550	.761	-.267	.817	1.000	.962
+ true	.753	.879	.845	.829	.879	.970	.970	.667	.550	.761	-.267	.817	1.000	.962
+ relevant	.783	.917	.883	.867	.917	.967	.967	.667	.550	.761	-.267	.817	1.000	.962
+ coherent	.783	.917	.883	.867	.917	.967	.967	.629	.695	.773	-.215	.829	.920	.858



n-gram metrics
decent for textual

Automatic Evaluation

Metrics used for open-ended generation

- n-gram overlap
 - chrF
 - BLEU
 - ROUGE-L
- semantic
 - BertSCORE
 - BLEURT
- LLM as a judge
 - LLaMA 3.3 70B
 - M-Prometheus

System-level correlations

BertSCORE worse than n-gram based

	Textual							Visual						
	BLEU	chrF	RG	BScr	BLRT	LLM	M-P	BLEU	chrF	RG	BScr	BLRT	LLM	M-P
Correct	.733	.867	.833	.817	.883	.950	.950	.667	.550	.761	-.267	.817	1.000	.962
+ true	.753	.879	.845	.829	.879	.970	.970	.667	.550	.761	-.267	.817	1.000	.962
+ relevant	.783	.917	.883	.867	.917	.967	.967	.667	.550	.761	-.267	.817	1.000	.962
+ coherent	.783	.917	.883	.867	.917	.967	.967	.629	.695	.773	-.215	.829	.920	.858

n-gram metrics
decent for textual

Automatic Evaluation

Metrics used for open-ended generation

- n-gram overlap
 - chrF
 - BLEU
 - ROUGE-L
- semantic
 - BertSCORE
 - BLEURT
- LLM as a judge
 - LLaMA 3.3 70B
 - M-Prometheus

System-level correlations

BertSCORE worse than n-gram based

	Textual							Visual						
	BLEU	chrF	RG	BScr	BLRT	LLM	M-P	BLEU	chrF	RG	BScr	BLRT	LLM	M-P
Correct	.733	.867	.833	.817	.883	.950	.950	.667	.550	.761	-.267	.817	1.000	.962
+ true	.753	.879	.845	.829	.879	.970	.970	.667	.550	.761	-.267	.817	1.000	.962
+ relevant	.783	.917	.883	.867	.917	.967	.967	.667	.550	.761	-.267	.817	1.000	.962
+ coherent	.783	.917	.883	.867	.917	.967	.967	.629	.695	.773	-.215	.829	.920	.858

n-gram metrics decent for textual

LLM judges are great!



Automatic Evaluation

Metrics used for open-ended generation

- n-gram overlap
 - chrF
 - BLEU
 - ROUGE-L
- semantic
 - BertSCORE
 - BLEURT
- LLM as a judge
 - LLaMA 3.3 70B
 - M-Prometheus

Answer-level correlations

	Textual					Visual				
	chrF	RG	BLRT	LLM	M-P	chrF	RG	BLRT	LLM	M-P
Correct	.594	.477	.552	.790	.707	.409	.370	.450	.807	.699
+ true	.585	.472	.543	.775	.703	.408	.370	.449	.804	.696
+ relevant	.567	.457	.524	.743	.683	.410	.370	.452	.797	.715
+ coherent	.548	.448	.511	.715	.665	.402	.376	.423	.710	.654

LLM-as-judge confusions

Legend		Average		Textual	CZ cs		CZ en		SK sk		SK en		UA uk		UA en	
TP	FN	29	4		34	3	30	5	26	3	25	6	32	3	28	7
FP	TN	5	61	2	61	8	57	2	69	7	62	3	63	9	56	
				Visual	CZ cs		CZ en		SK sk		SK en		UA uk		UA en	
		7	1		7	1	9	4	9	2	13	4	14	1		
		2	90	3	89	1	86	1	88	2	82	4	81			

Some confusions systematic
(e.g. Karel Gott), but no clear pattern

Results

Results of all models (*judged by LLaMA 3.3*)

Model	CZ		SK		UA	
	cs	en	sk	en	uk	en
LLaMA 3.3 70B Ins.	60.0	63.1	47.7	48.3	52.2	54.9
LLaMA 3.1 8B Ins.	29.6	36.5	19.6	25.6	25.7	34.1
EuroLLM 9B Ins.	39.3	43.1	30.8	30.2	26.8	28.4
Mistral 7B Ins. v0.3	22.9	30.4	14.4	22.3	24.6	27.8
LLaMA 4 Scout Ins.	41.8	45.6	34.6	34.4	45.7	45.4

LLaMA 3.2 11B Ins.	12.6	14.3	11.7	15.0	11.6	12.1
LLaMA 4 Scout Ins.	18.7	18.3	15.0	17.5	15.1	23.6
maya	1.7	3.0	1.7	2.5	2.5	4.0
idefics	4.8	7.4	6.7	9.2	6.0	8.5
gemma3	15.2	13.9	11.7	13.3	16.6	12.1

Results

Results of all models (*judged by LLaMA 3.3*)

Model	CZ		SK		UA	
	cs	en	sk	en	uk	en
LLaMA 3.3 70B Ins.	60.0	63.1	47.7	48.3	52.2	54.9
LLaMA 3.1 8B Ins.	29.6	36.5	19.6	25.6	25.7	34.1
EuroLLM 9B Ins.	39.3	43.1	30.8	30.2	26.8	28.4
Mistral 7B Ins. v0.3	22.9	30.4	14.4	22.3	24.6	27.8
LLaMA 4 Scout Ins.	41.8	45.6	34.6	34.4	45.7	45.4

← LLaMA 3.3 best open-weight
for text

LLaMA 3.2 11B Ins.	12.6	14.3	11.7	15.0	11.6	12.1
LLaMA 4 Scout Ins.	18.7	18.3	15.0	17.5	15.1	23.6
maya	1.7	3.0	1.7	2.5	2.5	4.0
idefics	4.8	7.4	6.7	9.2	6.0	8.5
gemma3	15.2	13.9	11.7	13.3	16.6	12.1

Results

Results of all models (*judged by LLaMA 3.3*)

Model	CZ		SK		UA	
	cs	en	sk	en	uk	en
LLaMA 3.3 70B Ins.	60.0	63.1	47.7	48.3	52.2	54.9
LLaMA 3.1 8B Ins.	29.6	36.5	19.6	25.6	25.7	34.1
EuroLLM 9B Ins.	39.3	43.1	30.8	30.2	26.8	28.4
Mistral 7B Ins. v0.3	22.9	30.4	14.4	22.3	24.6	27.8
LLaMA 4 Scout Ins.	41.8	45.6	34.6	34.4	45.7	45.4

LLaMA 3.3 best open-weight
for text

LLaMA 3.2 11B Ins.	12.6	14.3	11.7	15.0	11.6	12.1
LLaMA 4 Scout Ins.	18.7	18.3	15.0	17.5	15.1	23.6
maya	1.7	3.0	1.7	2.5	2.5	4.0
idefics	4.8	7.4	6.7	9.2	6.0	8.5
gemma3	15.2	13.9	11.7	13.3	16.6	12.1

LLaMA 4 best open-weight
for visual

Results

Results of all models (*judged by LLaMA 3.3*)

Model	CZ		SK		UA	
	cs	en	sk	en	uk	en
LLaMA 3.3 70B Ins.	60.0	63.1	47.7	48.3	52.2	54.9
LLaMA 3.1 8B Ins.	29.6	36.5	19.6	25.6	25.7	34.1
EuroLLM 9B Ins.	39.3	43.1	30.8	30.2	26.8	28.4
Mistral 7B Ins. v0.3	22.9	30.4	14.4	22.3	24.6	27.8
LLaMA 4 Scout Ins.	41.8	45.6	34.6	34.4	45.7	45.4

LLaMA 3.3 best open-weight
for text

LLaMA 3.2 11B Ins.	12.6	14.3	11.7	15.0	11.6	12.1
LLaMA 4 Scout Ins.	18.7	18.3	15.0	17.5	15.1	23.6
maya	1.7	3.0	1.7	2.5	2.5	4.0
idefics	4.8	7.4	6.7	9.2	6.0	8.5
gemma3	15.2	13.9	11.7	13.3	16.6	12.1

Visual much harder

LLaMA 4 best open-weight
for visual

Results

Results of all models (*judged by LLaMA 3.3*)

Slovak worst, but try asking in English

Model	CZ		SK		UA	
	cs	en	sk	en	uk	en
LLaMA 3.3 70B Ins.	60.0	63.1	47.7	48.3	52.2	54.9
LLaMA 3.1 8B Ins.	29.6	36.5	19.6	25.6	25.7	34.1
EuroLLM 9B Ins.	39.3	43.1	30.8	30.2	26.8	28.4
Mistral 7B Ins. v0.3	22.9	30.4	14.4	22.3	24.6	27.8
LLaMA 4 Scout Ins.	41.8	45.6	34.6	34.4	45.7	45.4

LLaMA 3.3 best open-weight for text

LLaMA 3.2 11B Ins.	12.6	14.3	11.7	15.0	11.6	12.1
LLaMA 4 Scout Ins.	18.7	18.3	15.0	17.5	15.1	23.6
maya	1.7	3.0	1.7	2.5	2.5	4.0
idefics	4.8	7.4	6.7	9.2	6.0	8.5
gemma3	15.2	13.9	11.7	13.3	16.6	12.1

Visual much harder

LLaMA 4 best open-weight for visual

Results

Results of all models (*judged by LLaMA 3.3*)

Slovak worst, but try asking in English

Model	CZ		SK		UA	
	cs	en	sk	en	uk	en
LLaMA 3.3 70B Ins.	60.0	63.1	47.7	48.3	52.2	54.9
LLaMA 3.1 8B Ins.	29.6	36.5	19.6	25.6	25.7	34.1
EuroLLM 9B Ins.	39.3	43.1	30.8	30.2	26.8	28.4
Mistral 7B Ins. v0.3	22.9	30.4	14.4	22.3	24.6	27.8
LLaMA 4 Scout Ins.	41.8	45.6	34.6	34.4	45.7	45.4
Gemini 2.5 Flash	78.0	75.8	64.2	61.3	68.9	63.0
Gemini 2.5 Fl.-Lite	60.0	54.2	47.9	43.3	54.3	51.9
GPT-5-2025-08-07	83.1	80.7	73.3	70.2	73.2	72.2
GPT-5-mini-2025-08-07	67.5	70.7	52.3	50.4	61.4	59.7
LLaMA 3.2 11B Ins.	12.6	14.3	11.7	15.0	11.6	12.1
LLaMA 4 Scout Ins.	18.7	18.3	15.0	17.5	15.1	23.6
maya	1.7	3.0	1.7	2.5	2.5	4.0
idefics	4.8	7.4	6.7	9.2	6.0	8.5
gemma3	15.2	13.9	11.7	13.3	16.6	12.1
Gemini 2.5 Flash	39.6	37.0	38.3	34.2	41.2	40.7
Gemini 2.5 Fl.-Lite	30.9	27.4	26.7	24.2	27.1	27.6
GPT-5-2025-08-07	46.5	42.6	44.2	41.7	37.2	36.2
GPT-5-mini-2025-08-07	33.0	38.7	24.2	26.7	26.1	31.7

Commercial models best

LLaMA 3.3 best open-weight for text

Visual much harder

LLaMA 4 best open-weight for visual

Results

Best model (*LLaMA 3.3 for Text, LLaMA 4 for Visual*)

			Human dev		Auto. dev		Auto. test	
			OK	Perf.	chrF	LLM	chrF	LLM
Textual	CZ	cs	58.7	53.4	42.4	57.4	43.7	60.0
		en	52.5	44.7	37.8	56.0	37.5	63.1
	SK	sk	46.2	40.8	39.2	45.4	35.2	47.7
		en	45.8	40.4	32.5	47.5	32.6	48.3
	UA	uk	51.9	46.2	42.4	57.4	42.6	52.2
		en	47.8	43.1	37.8	56.0	37.0	54.9
Visual	CZ	cs	14.6	11.1	21.3	14.6	21.3	14.6
		en	13.3	8.4	15.5	15.9	15.5	15.9
	SK	sk	23.7	20.3	24.4	18.6	24.4	18.6
		en	19.5	12.7	16.6	16.9	20.4	17.5
	UA	uk	29.9	29.9	21.1	26.0	20.7	15.1
		en	17.2	17.2	19.1	20.1	21.7	23.6

Results

Best model (*LLaMA 3.3 for Text, LLaMA 4 for Visual*)

			Human dev		Auto. dev		Auto. test	
			OK	Perf.	chrF	LLM	chrF	LLM
Textual	CZ	cs	58.7	53.4	42.4	57.4	43.7	60.0
		en	52.5	44.7	37.8	56.0	37.5	63.1
	SK	sk	46.2	40.8	39.2	45.4	35.2	47.7
		en	45.8	40.4	32.5	47.5	32.6	48.3
	UA	uk	51.9	46.2	42.4	57.4	42.6	52.2
		en	47.8	43.1	37.8	56.0	37.0	54.9
Visual	CZ	cs	14.6	11.1	21.3	14.6	21.3	14.6
		en	13.3	8.4	15.5	15.9	15.5	15.9
	SK	sk	23.7	20.3	24.4	18.6	24.4	18.6
		en	19.5	12.7	16.6	16.9	20.4	17.5
	UA	uk	29.9	29.9	21.1	26.0	20.7	15.1
		en	17.2	17.2	19.1	20.1	21.7	23.6

Visual much harder

Results

Best model (*LLaMA 3.3 for Text, LLaMA 4 for Visual*)

			Human dev		Auto. dev		Auto. test	
			OK	Perf.	chrF	LLM	chrF	LLM
Textual	CZ	cs	58.7	53.4	42.4	57.4	43.7	60.0
		en	52.5	44.7	37.8	56.0	37.5	63.1
	SK	sk	46.2	40.8	39.2	45.4	35.2	47.7
		en	45.8	40.4	32.5	47.5	32.6	48.3
	UA	uk	51.9	46.2	42.4	57.4	42.6	52.2
		en	47.8	43.1	37.8	56.0	37.0	54.9
Visual	CZ	cs	14.6	11.1	21.3	14.6	21.3	14.6
		en	13.3	8.4	15.5	15.9	15.5	15.9
	SK	sk	23.7	20.3	24.4	18.6	24.4	18.6
		en	19.5	12.7	16.6	16.9	20.4	17.5
	UA	uk	29.9	29.9	21.1	26.0	20.7	15.1
		en	17.2	17.2	19.1	20.1	21.7	23.6

OK vs. Perfect gap
suggests hallucinations

Visual much harder

Cross-Lingual Results

Which language to ask in? (*Judged by LLaMA 3.3*)

		Czechia				Slovakia				Ukraine			
		cs	sk	uk	en	cs	sk	uk	en	cs	sk	uk	en
Textual	EuroLLM-9B-Instruct	35.5	29.8	8.3	31.7	29.0	30.2	9.7	32.9	17.4	15.8	26.8	29.1
	Llama-3.1-8B-Instruct	28.5	17.0	8.9	31.3	21.5	18.3	7.9	25.2	18.7	15.1	23.4	34.8
	Llama-3.3-70B-Instruct	57.5	52.3	30.0	56.2	43.6	43.8	24.9	47.1	40.8	39.0	54.5	54.3
	Llama-4-Scout-17B-16E-Instruct	38.3	33.2	21.5	38.1	28.6	32.9	18.7	33.5	31.7	34.8	47.3	46.8
	Mistral-7B-Instruct-v0.3	22.6	11.9	11.7	26.8	17.2	13.6	8.9	22.7	17.4	12.2	21.8	32.5
Visual	gemma3	10.6	7.5	5.8	12.8	12.7	13.6	7.6	11.9	13.7	12.7	22.1	21.1
	idefics	7.5	3.5	4.4	6.6	6.8	7.6	8.5	10.2	7.8	4.9	8.3	10.3
	Llama-3.2-11B-Vision-Instruct	10.2	7.5	8.4	13.7	11.9	9.3	11.0	11.0	11.8	9.3	15.2	19.1
	Llama-4-Scout-17B-16E-Instruct	15.0	11.1	4.9	15.9	16.1	18.6	7.6	16.1	16.2	14.2	25.0	29.9
	maya	1.8	0.0	0.4	1.8	0.0	0.0	0.0	0.0	2.9	1.5	2.9	5.9

Cross-Lingual Results

Which language to ask in? (*Judged by LLaMA 3.3*)

Czech is good for
Czech stuff

		Czechia				Slovakia				Ukraine			
		cs	sk	uk	en	cs	sk	uk	en	cs	sk	uk	en
Textual	EuroLLM-9B-Instruct	35.5	29.8	8.3	31.7	29.0	30.2	9.7	32.9	17.4	15.8	26.8	29.1
	Llama-3.1-8B-Instruct	28.5	17.0	8.9	31.3	21.5	18.3	7.9	25.2	18.7	15.1	23.4	34.8
	Llama-3.3-70B-Instruct	57.5	52.3	30.0	56.2	43.6	43.8	24.9	47.1	40.8	39.0	54.5	54.3
	Llama-4-Scout-17B-16E-Instruct	38.3	33.2	21.5	38.1	28.6	32.9	18.7	33.5	31.7	34.8	47.3	46.8
	Mistral-7B-Instruct-v0.3	22.6	11.9	11.7	26.8	17.2	13.6	8.9	22.7	17.4	12.2	21.8	32.5
Visual	gemma3	10.6	7.5	5.8	12.8	12.7	13.6	7.6	11.9	13.7	12.7	22.1	21.1
	idefics	7.5	3.5	4.4	6.6	6.8	7.6	8.5	10.2	7.8	4.9	8.3	10.3
	Llama-3.2-11B-Vision-Instruct	10.2	7.5	8.4	13.7	11.9	9.3	11.0	11.0	11.8	9.3	15.2	19.1
	Llama-4-Scout-17B-16E-Instruct	15.0	11.1	4.9	15.9	16.1	18.6	7.6	16.1	16.2	14.2	25.0	29.9
	maya	1.8	0.0	0.4	1.8	0.0	0.0	0.0	0.0	2.9	1.5	2.9	5.9

Cross-Lingual Results

Which language to ask in? (*Judged by LLaMA 3.3*)

Czech is good for
Czech stuff

For Slovak, English
is better

		Czechia				Slovakia				Ukraine			
		cs	sk	uk	en	cs	sk	uk	en	cs	sk	uk	en
Textual	EuroLLM-9B-Instruct	35.5	29.8	8.3	31.7	29.0	30.2	9.7	32.9	17.4	15.8	26.8	29.1
	Llama-3.1-8B-Instruct	28.5	17.0	8.9	31.3	21.5	18.3	7.9	25.2	18.7	15.1	23.4	34.8
	Llama-3.3-70B-Instruct	57.5	52.3	30.0	56.2	43.6	43.8	24.9	47.1	40.8	39.0	54.5	54.3
	Llama-4-Scout-17B-16E-Instruct	38.3	33.2	21.5	38.1	28.6	32.9	18.7	33.5	31.7	34.8	47.3	46.8
	Mistral-7B-Instruct-v0.3	22.6	11.9	11.7	26.8	17.2	13.6	8.9	22.7	17.4	12.2	21.8	32.5
Visual	gemma3	10.6	7.5	5.8	12.8	12.7	13.6	7.6	11.9	13.7	12.7	22.1	21.1
	idefics	7.5	3.5	4.4	6.6	6.8	7.6	8.5	10.2	7.8	4.9	8.3	10.3
	Llama-3.2-11B-Vision-Instruct	10.2	7.5	8.4	13.7	11.9	9.3	11.0	11.0	11.8	9.3	15.2	19.1
	Llama-4-Scout-17B-16E-Instruct	15.0	11.1	4.9	15.9	16.1	18.6	7.6	16.1	16.2	14.2	25.0	29.9
	maya	1.8	0.0	0.4	1.8	0.0	0.0	0.0	0.0	2.9	1.5	2.9	5.9

Cross-Lingual Results

Which language to ask in? (*Judged by LLaMA 3.3*)

Czech is good for
Czech stuff

For Slovak, English
is better

		Czechia				Slovakia				Ukraine			
		cs	sk	uk	en	cs	sk	uk		en	cs	sk	uk
Textual	EuroLLM-9B-Instruct	35.5	29.8	8.3	31.7	29.0	30.2	9.7	32.9	17.4	15.8	26.8	29.1
	Llama-3.1-8B-Instruct	28.5	17.0	8.9	31.3	21.5	18.3	7.9	25.2	18.7	15.1	23.4	34.8
	Llama-3.3-70B-Instruct	57.5	52.3	30.0	56.2	43.6	43.8	24.9	47.1	40.8	39.0	54.5	54.3
	Llama-4-Scout-17B-16E-Instruct	38.3	33.2	21.5	38.1	28.6	32.9	18.7	33.5	31.7	34.8	47.3	46.8
	Mistral-7B-Instruct-v0.3	22.6	11.9	11.7	26.8	17.2	13.6	8.9	22.7	17.4	12.2	21.8	32.5
Visual	gemma3	10.6	7.5	5.8	12.8	12.7	13.6	7.6	11.9	13.7	12.7	22.1	21.1
	idefics	7.5	3.5	4.4	6.6	6.8	7.6	8.5	10.2	7.8	4.9	8.3	10.3
	Llama-3.2-11B-Vision-Instruct	10.2	7.5	8.4	13.7	11.9	9.3	11.0	11.0	11.8	9.3	15.2	19.1
	Llama-4-Scout-17B-16E-Instruct	15.0	11.1	4.9	15.9	16.1	18.6	7.6	16.1	16.2	14.2	25.0	29.9
	maya	1.8	0.0	0.4	1.8	0.0	0.0	0.0	0.0	2.9	1.5	2.9	5.9

For Ukrainian, it
depends on the
model

Cross-Lingual Results

Which language to ask in? (*Judged by LLaMA 3.3*)

Czech is good for
Czech stuff

For Slovak, English
is better

		Czechia				Slovakia				Ukraine			
		cs	sk	uk	en	cs	sk	uk		en	cs	sk	uk
Textual	EuroLLM-9B-Instruct	35.5	29.8	8.3	31.7	29.0	30.2	9.7	32.9	17.4	15.8	26.8	29.1
	Llama-3.1-8B-Instruct	28.5	17.0	8.9	31.3	21.5	18.3	7.9	25.2	18.7	15.1	23.4	34.8
	Llama-3.3-70B-Instruct	57.5	52.3	30.0	56.2	43.6	43.8	24.9	47.1	40.8	39.0	54.5	54.3
	Llama-4-Scout-17B-16E-Instruct	38.3	33.2	21.5	38.1	28.6	32.9	18.7	33.5	31.7	34.8	47.3	46.8
	Mistral-7B-Instruct-v0.3	22.6	11.9	11.7	26.8	17.2	13.6	8.9	22.7	17.4	12.2	21.8	32.5
Visual	gemma3	10.6	7.5	5.8	12.8	12.7	13.6	7.6	11.9	13.7	12.7	22.1	21.1
	idefics	7.5	3.5	4.4	6.6	6.8	7.6	8.5	10.2	7.8	4.9	8.3	10.3
	Llama-3.2-11B-Vision-Instruct	10.2	7.5	8.4	13.7	11.9	9.3	11.0	11.0	11.8	9.3	15.2	19.1
	Llama-4-Scout-17B-16E-Instruct	15.0	11.1	4.9	15.9	16.1	18.6	7.6	16.1	16.2	14.2	25.0	29.9
	maya	1.8	0.0	0.4	1.8	0.0	0.0	0.0	0.0	2.9	1.5	2.9	5.9

For visual questions, English is a safe choice
(but quality is very poor anyway)

For Ukrainian, it
depends on the
model

Cross-Lingual Results

Which language to ask in? (*Judged by LLaMA 3.3*)

Czech is good for
Czech stuff

For Slovak, English
is better

		Czechia				Slovakia				Ukraine			
		cs	sk	uk	en	cs	sk	uk		en	cs	sk	uk
Textual	EuroLLM-9B-Instruct	35.5	29.8	8.3	31.7	29.0	30.2	9.7	32.9	17.4	15.8	26.8	29.1
	Llama-3.1-8B-Instruct	28.5	17.0	8.9	31.3	21.5	18.3	7.9	25.2	18.7	15.1	23.4	34.8
	Llama-3.3-70B-Instruct	57.5	52.3	30.0	56.2	43.6	43.8	24.9	47.1	40.8	39.0	54.5	54.3
	Llama-4-Scout-17B-16E-Instruct	38.3	33.2	21.5	38.1	28.6	32.9	18.7	33.5	31.7	34.8	47.3	46.8
	Mistral-7B-Instruct-v0.3	22.6	11.9	11.7	26.8	17.2	13.6	8.9	22.7	17.4	12.2	21.8	32.5
Visual	gemma3	10.6	7.5	5.8	12.8	12.7	13.6	7.6	11.9	13.7	12.7	22.1	21.1
	idefics	7.5	3.5	4.4	6.6	6.8	7.6	8.5	10.2	7.8	4.9	8.3	10.3
	Llama-3.2-11B-Vision-Instruct	10.2	7.5	8.4	13.7	11.9	9.3	11.0	11.0	11.8	9.3	15.2	19.1
	Llama-4-Scout-17B-16E-Instruct	15.0	11.1	4.9	15.9	16.1	18.6	7.6	16.1	16.2	14.2	25.0	29.9
	maya	1.8	0.0	0.4	1.8	0.0	0.0	0.0	0.0	2.9	1.5	2.9	5.9

For visual questions, English is a safe choice
(but quality is very poor anyway)

Czech-Slovak mutually somewhat useful
(but English always better than nonlocal)

For Ukrainian, it
depends on the
model

Cross-Lingual Results

Which language to ask in? (*Judged by LLaMA 3.3*)

Czech is good for
Czech stuff

For Slovak, English
is better

		Czechia				Slovakia				Ukraine			
		cs	sk	uk	en	cs	sk	uk		en	cs	sk	uk
Textual	EuroLLM-9B-Instruct	35.5	29.8	8.3	31.7	29.0	30.2	9.7	32.9	17.4	15.8	26.8	29.1
	Llama-3.1-8B-Instruct	28.5	17.0	8.9	31.3	21.5	18.3	7.9	25.2	18.7	15.1	23.4	34.8
	Llama-3.3-70B-Instruct	57.5	52.3	30.0	56.2	43.6	43.8	24.9	47.1	40.8	39.0	54.5	54.3
	Llama-4-Scout-17B-16E-Instruct	38.3	33.2	21.5	38.1	28.6	32.9	18.7	33.5	31.7	34.8	47.3	46.8
	Mistral-7B-Instruct-v0.3	22.6	11.9	11.7	26.8	17.2	13.6	8.9	22.7	17.4	12.2	21.8	32.5
Visual	gemma3	10.6	7.5	5.8	12.8	12.7	13.6	7.6	11.9	13.7	12.7	22.1	21.1
	idefics	7.5	3.5	4.4	6.6	6.8	7.6	8.5	10.2	7.8	4.9	8.3	10.3
	Llama-3.2-11B-Vision-Instruct	10.2	7.5	8.4	13.7	11.9	9.3	11.0	11.0	11.8	9.3	15.2	19.1
	Llama-4-Scout-17B-16E-Instruct	15.0	11.1	4.9	15.9	16.1	18.6	7.6	16.1	16.2	14.2	25.0	29.9
	maya	1.8	0.0	0.4	1.8	0.0	0.0	0.0	0.0	2.9	1.5	2.9	5.9

For visual questions, English is a safe choice
(but quality is very poor anyway)

Czech-Slovak mutually somewhat useful
(but English always* better than nonlocal)

For Ukrainian, it
depends on the
model

Lessons Learned

Lessons Learned

- Making benchmarks is fun!
 - You get to be creative and laugh at silly LLM mistakes.
 - But you also need to make a lot of arbitrary decisions.
- Annotators need to be watched closely (at least at the beginning)
- It is hard to scientifically interpret the results
- ACL reviewers need a lot of experimental results
- Real time spent on this will not match your expectations (ever).

Future Work

Make the dataset useful (not only) for OpenEuroLLM

1. Add support into LM Evaluation Harness
2. Include other languages (plans for Norwegian, South Slavic)
3. Propose a shared task to SemEval
4. Suggestions?

Summary

- We have a dataset of
 - regional-specific questions and answers
 - in three regions and four languages
- Open-ended QA of this sort can be evaluated automatically
- Language of the question matters even for multilingual models

arxiv.org/abs/2507.22752

huggingface.co/datasets/ufal/cus-qa