BenCzechMark

Czech-centric Multitask and Multimetric Benchmark for Language Models with Duel Scoring Mechanism

Presentation Author: Martin Fajčík (**BUT**)

Work authors: Martin Fajcik, Martin Docekal, Jan Dolezal, Karel Ondrej, Karel Benes, Jan Kapsa, Alexander Polok, Michal Hradis, **as BUT**
Zuzana Neverilova, Ales Horak, Michal Stefanik, as **MUNI**
Adam Jirkovsky, David Adamczyk, Jan Hula, Jan Sedivy, as **CVUT**
Hynek Kydlicek as **Hugging Face**

November 15, 2024

Venue: PROTECT 2024

Aims

- Provide **Fair** Comparison on how LLMs perform on Czech, reflecting how they fare on
  - On cultural-specific aspects
    - A/B/C/D/E .. school tests, mimicking Hendrycks et al.’s MMLU-like Benchmarks

Hendrycks, Dan, et al. "Measuring Massive Multitask Language Understanding." International Conference on Learning Representations.

Aims

- Provide **Fair** Comparison on how LLMs perform on Czech, reflecting how they fare on
  - On cultural-specific aspects
    - A/B/C/D/E .. school tests, mimicking Hendrycks et al.’s MMLU-like Benchmarks
  - On traditional NLP tasks
    - Entity Recognition, Sentiment

Hendrycks, Dan, et al. "Measuring Massive Multitask Language Understanding." International Conference on Learning Representations.

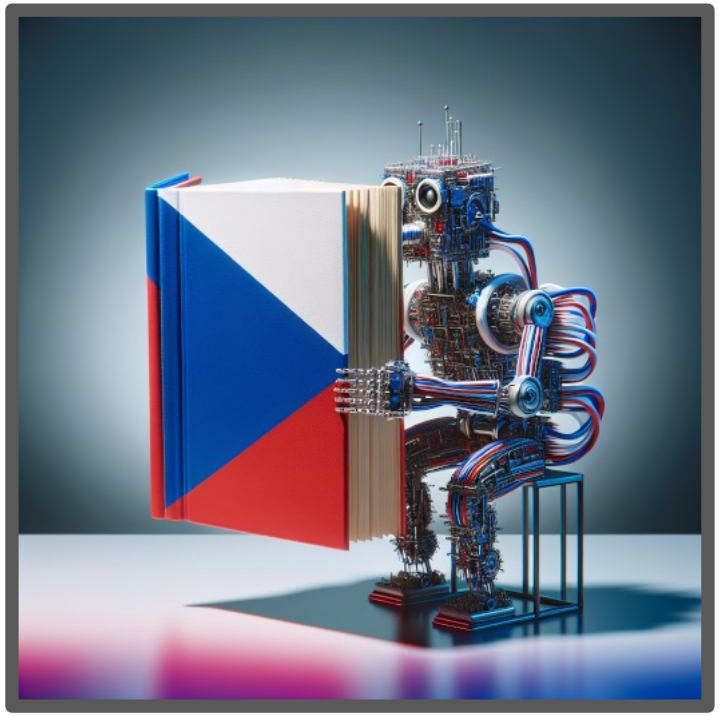
Aims

- Provide **Fair** Comparison on how LLMs perform on Czech, reflecting how they fare on
  - On cultural-specific aspects
    - A/B/C/D/E .. school tests, mimicking Hendrycks et al.’s MMLU-like Benchmarks
  - On traditional NLP tasks
    - Entity Recognition, Sentiment
  - Fair =
    - a specific metric for every task,
    - threshold-free metric to allow for an uncalibrated comparison,
    - standardized code-base (extending ElutherAI’s Language Model Harness),
    - Metric comparison/aggregation with statistical testing

Hendrycks, Dan, et al. "Measuring Massive Multitask Language Understanding." International Conference on Learning Representations.

Aims

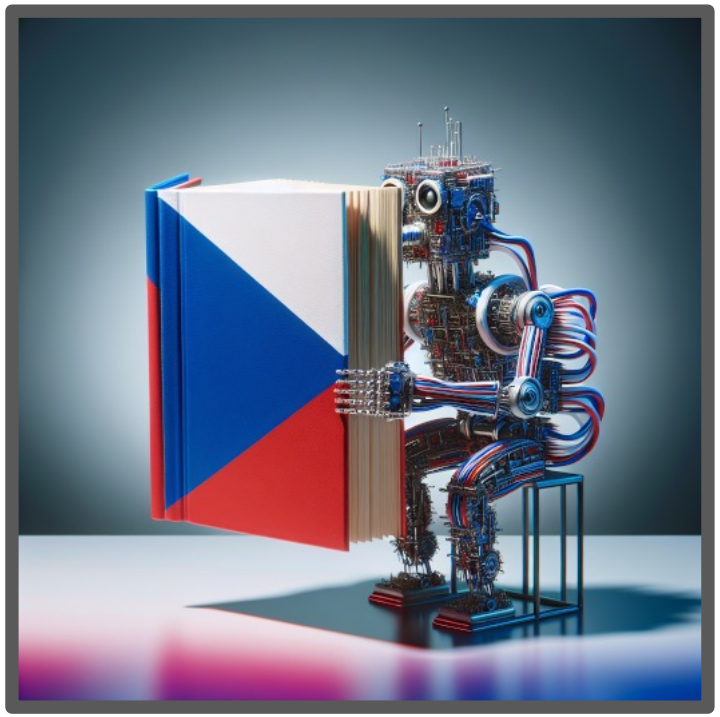
- Provide **Fair** Comparison on how LLMs perform on Czech, reflecting how they fare on
- An Excuse for Training **Czech-centric LLM**



Hendrycks, Dan, et al. "Measuring Massive Multitask Language Understanding." International Conference on Learning Representations.

Czech Centric LLM

- We trained **BUT LM Family** (available on huggingface.co/)
  - BUT-FIT/CSTinyLlama-1.2B (1.2B)
  - BUT-FIT/Czech-GPT-2-XL-133k (1.5B)
  - BUT-FIT/csmpt7b (6.7B)
- Results on a nutshell
  - Great for language modelling
    (ASR,OCR, Phone Keyboard, Spelling Error Detection)
  - Limited few-shot learning ability
    (usually worse then similarly sized multilingual models)
  - Baselines for BCM



Hendrycks, Dan, et al. "Measuring Massive Multitask Language Understanding." International Conference on Learning Representations.

Few-shot learning?

You are an assistant that specializes in converting sentences into their corresponding opposites. Based on the provided examples, transform each input sentence into its opposite meaning.
Input: "The weather is sunny and warm."
Output: "The weather is cloudy and cold."

Input: "I enjoy waking up early to start my day."
Output: "I dislike waking up early and prefer to sleep in."

Input: "She is always on time for her appointments." **Output:**

Hendrycks, Dan, et al. "Measuring Massive Multitask Language Understanding." International Conference on Learning Representations.

Contents

Czech Language Understanding
- CERMAT - Czech Language Tests (MC/OPEN/TF)
- Grammar Error Correction

Czech Math Understanding
- CERMAT - Czech Math Tests (MC/OPEN)
- Klokan QA
- Umimeto.cz Math

Factual Knowledge
- Umimeto.cz (Biology, Chemistry, History, Informatics, Physics)
- TriviaQA-CZ (*Automatic Translation (AT)*)
- Natural Questions-CZ (*AT*)

Language Modeling
- Czech National Corpus (Dialect, Essays, Fiction, Karel Havlicek News, Correspondence, Spoken)
- SEMANT Historical Corpus
- HellaSwag-CZ (*Automatic Translation*)

BenCzechMark

50 tasks
8 categories
4 metrics

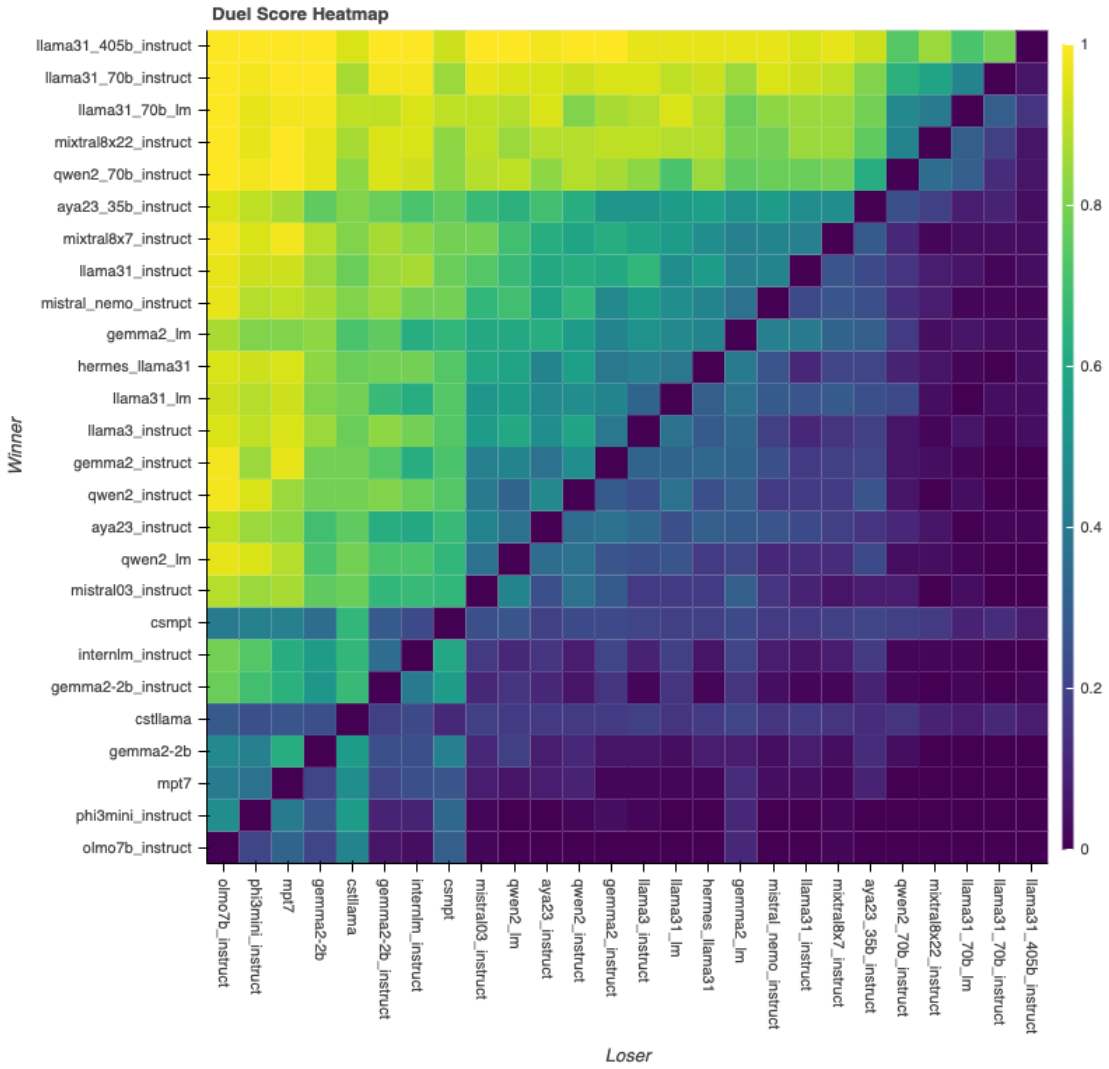
Reading Comprehension
- Belebele
- SQAD3.2
- HistoryIR

Named Entity Recognition
- Czech Court Decisions
- Czech Named Entity Corpus 2.0

Natural Language Inference
- CTKFacts
- CSFEVER (AT)
- Czech SNLI (*AT + Manual Verification*)
- Propaganda Dataset Tasks (Argumentace, Démonizace, Emoce, Fabulace, Nálepkování, Lokace, Zaměření, Názor, Relativizace, Rusko, Strach, Vina, Žánr)

Sentiment Analysis
- Czech Sentiment (Facebook, Mall, CSFD)
- Subjectivity

Evaluation in a Nutshell



- On each BCM task, we improvement of each model over each with significance test (alpha 5%).
- We compute how many times model X improves significantly over everyone else (**duel win score**).
- Average inside categories (**category duel win score**)
- Average across categories (**overall duel win score**)

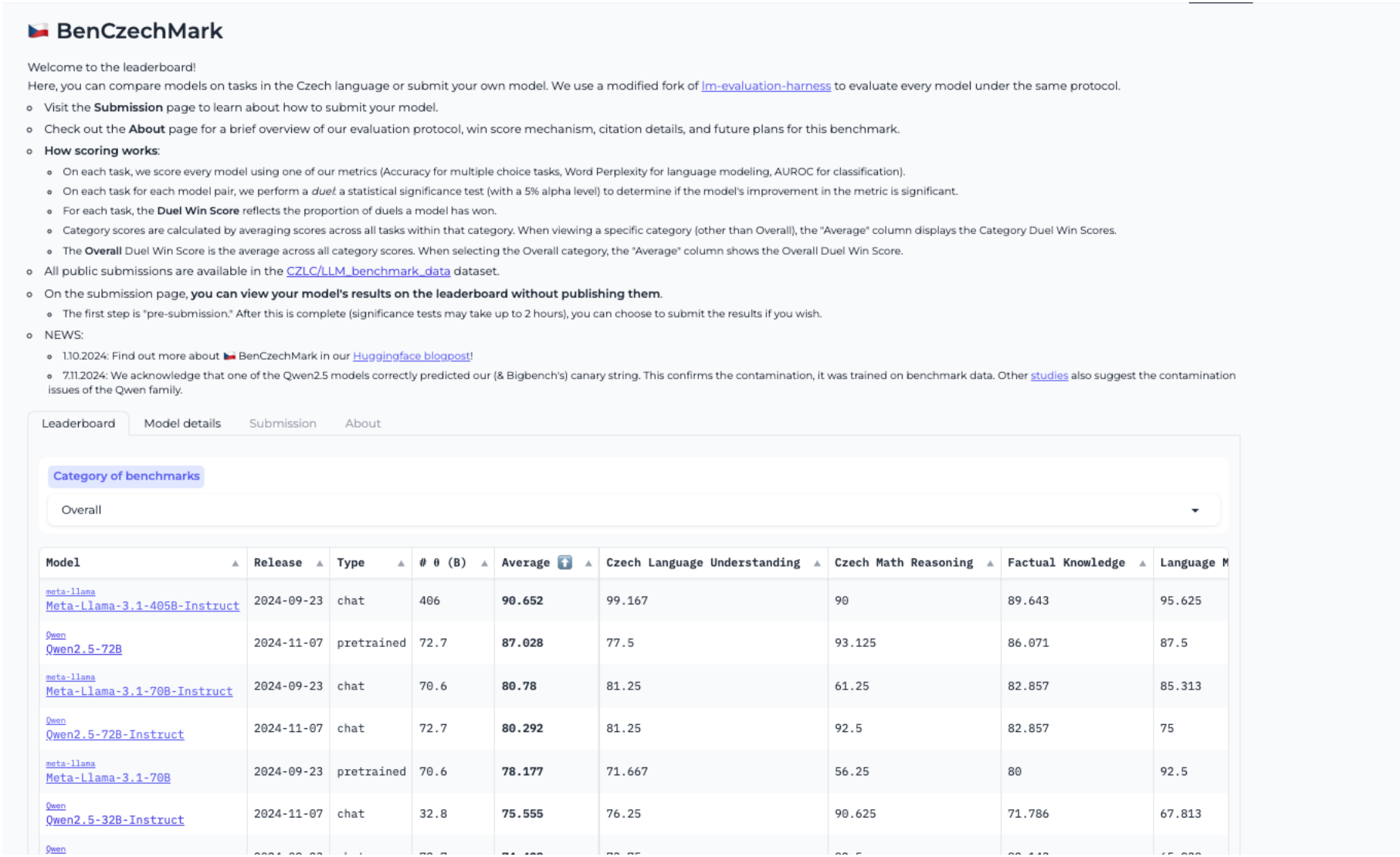
Leaderboard

Available at https://huggingface.co/spaces/CZLC/BenCzechMark

40 models currently

1B to 406B param. models

3-shot, 2048 tokens max



BenCzechMark

Welcome to the leaderboard!

Here, you can compare models on tasks in the Czech language or submit your own model. We use a modified fork of lm-evaluation-harness to evaluate every model under the same protocol.

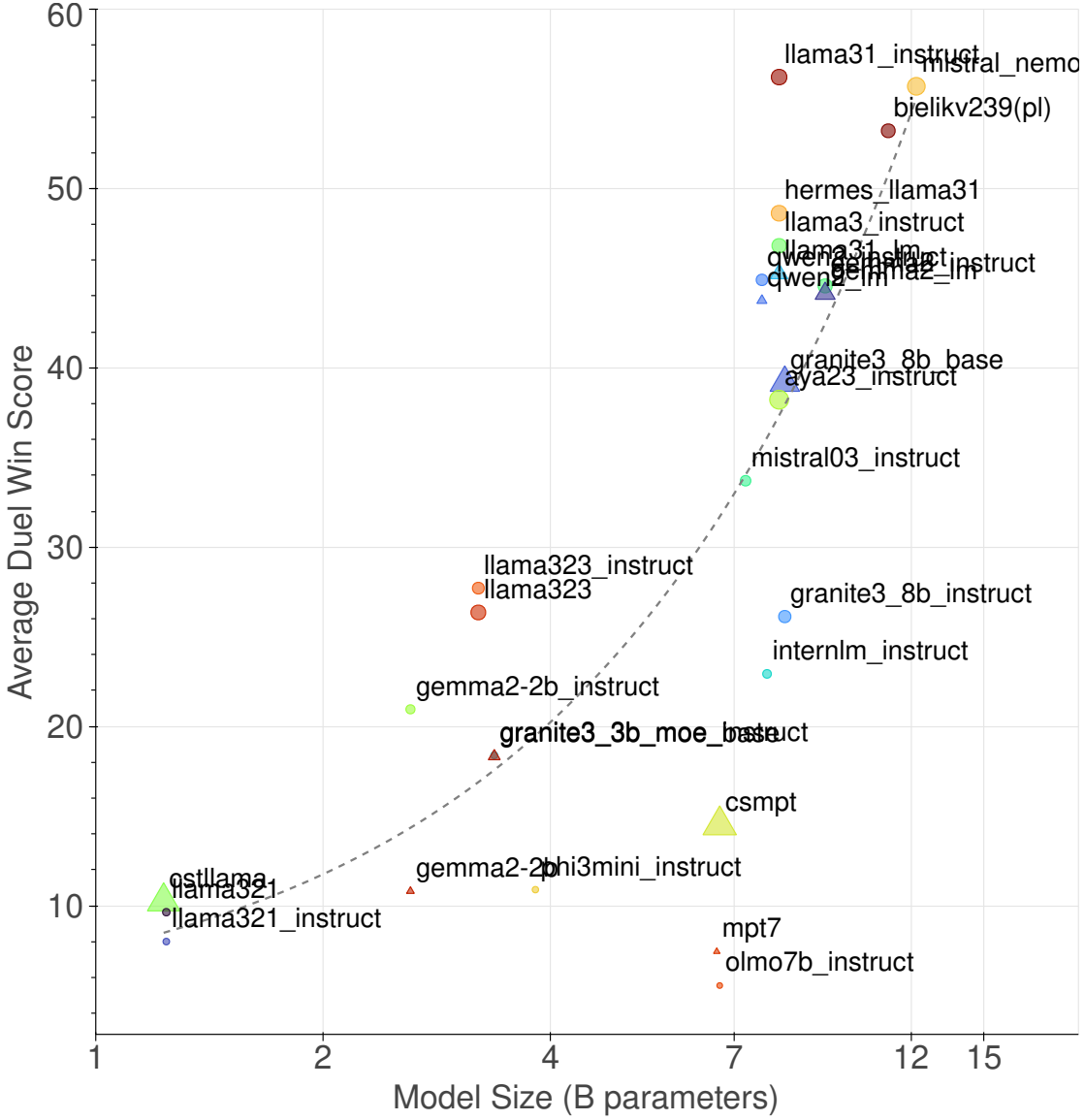
- Visit the **Submission** page to learn about how to submit your model.
- Check out the **About** page for a brief overview of our evaluation protocol, win score mechanism, citation details, and future plans for this benchmark.
- **How scoring works**:
  - On each task, we score every model using one of our metrics (Accuracy for multiple choice tasks, Word Perplexity for language modeling, AUROC for classification).
  - On each task for each model pair, we perform a *duel*: a statistical significance test (with a 5% alpha level) to determine if the model's improvement in the metric is significant.
  - For each task, the **Duel Win Score** reflects the proportion of duels a model has won.
  - Category scores are calculated by averaging scores across all tasks within that category. When viewing a specific category (other than Overall), the "Average" column displays the Category Duel Win Scores.
  - The **Overall** Duel Win Score is the average across all category scores. When selecting the Overall category, the "Average" column shows the Overall Duel Win Score.
- All public submissions are available in the CZLC/LLM_benchmark_data dataset.
- On the submission page, **you can view your model's results on the leaderboard without publishing them**.
  - The first step is "pre-submission." After this is complete (significance tests may take up to 2 hours), you can choose to submit the results if you wish.
- NEWS:
  - 1.10.2024: Find out more about BenCzechMark in our Huggingface blogpost!
  - 7.11.2024: We acknowledge that one of the Qwen2.5 models correctly predicted our (& Bigbench's) canary string. This confirms the contamination, it was trained on benchmark data. Other studies also suggest the contamination issues of the Qwen family.

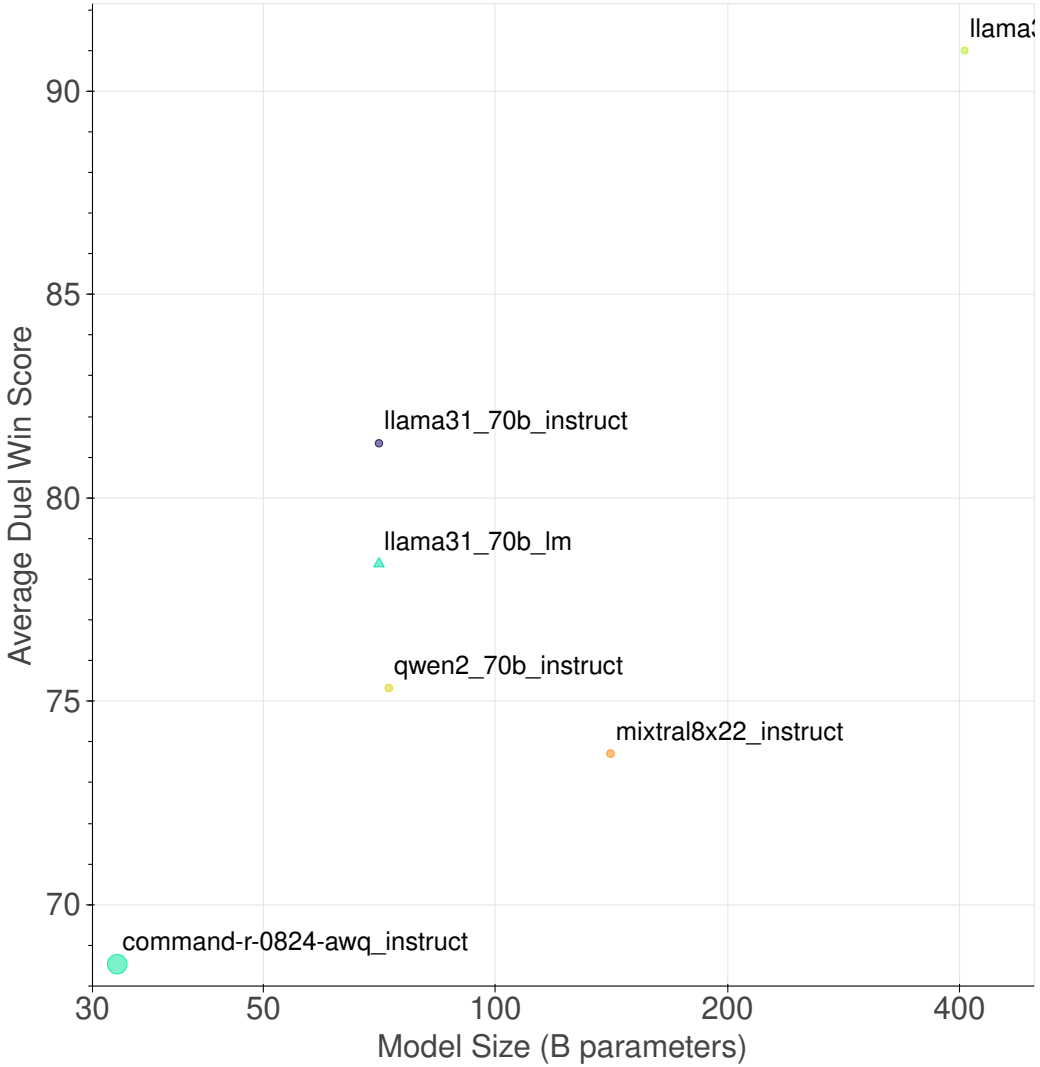
Leaderboard | Model details | Submission | About

Category of benchmarks: Overall

Model	Release	Type	# 0 (B)	Average	Czech Language Understanding	Czech Math Reasoning	Factual Knowledge	Language M
meta-llama Meta-Llama-3.1-405B-Instruct	2024-09-23	chat	406	**90.652**	99.167	90	89.643	95.625
Qwen Qwen2.5-72B	2024-11-07	pretrained	72.7	**87.028**	77.5	93.125	86.071	87.5
meta-llama Meta-Llama-3.1-70B-Instruct	2024-09-23	chat	70.6	**80.78**	81.25	61.25	82.857	85.313
Qwen Qwen2.5-72B-Instruct	2024-11-07	chat	72.7	**80.292**	81.25	92.5	82.857	75
meta-llama Meta-Llama-3.1-70B	2024-09-23	pretrained	70.6	**78.177**	71.667	56.25	80	92.5
Qwen Qwen2.5-32B-Instruct	2024-11-07	chat	32.8	**75.555**	76.25	90.625	71.786	67.813
Qwen								

Results


Average Duel Win Score
Model Size (B parameters)
60
50
40
30
20
10
1
2
4
7
12
15
llama31_instruct
mistral_nemo
bielikv239(pl)
hermes_llama31
llama3_instruct
llama31_lm
qwen2_lm
granite3_8b_base
aya23_instruct
mistral03_instruct
llama323_instruct
llama323
granite3_8b_instruct
internlm_instruct
gemma2-2b_instruct
csmpt
gemma2-2b
llama321
llama321_instruct
mpt7
olmo7b_instruct


Average Duel Win Score
Model Size (B parameters)
90
85
80
75
70
30
50
100
200
400
llama31_70b_instruct
llama31_70b_lm
qwen2_70b_instruct
mixtral8x22_instruct
command-r-0824-awq_instruct

Dataset Contamination

Dataset	C [%]
SumeCzech	97.07
CzechNews	96.53
Propaganda Datasets	95.90
Czech Court Decisions	95.10
CNC - Karel Čapek	83.04
CNC - Speeches	62.35
HistoryIR	54.88
Umimeto (math)	45.00
Subjectivity	42.32
CERMAT - Czech Language Tests	23.40
CNC - Karel Havlicek News	19.40
Umimeto (chemistry)	18.00
Umimeto (physics)	18.00
CERMAT Tests - Czech Math Tests	17.46
CNC Corpora - Fiction (FicTree)	17.18
Umimeto (informatics)	17.00
Umimeto (biology)	15.00
Subjectivity - Agree	13.34
Grammar Error Collection	13.20
Czech Sentiment - CSFD	12.10
Czech Sentiment - Mall.cz	11.87
Klokan QA	11.65
Czech Sentiment - Facebook Comments	11.40
Umimeto (history)	9.00
SQAD3.2	8.34
CNEC 2.0	7.84
CERMAT Tests - Czech Language Tests	7.09
Umimeto (czech)	7.00
CERMAT Tests - Czech Language Tests	6.49
CTKFacts-NLI	3.23
CNC Corpora - Essays	2.95
CSFever-NLI	1.82
CERMAT Tests - Czech Math Tests	1.57
Natural-Questions-CZ	0.78
CNC Corpora - Spoken	0.69
TriviaQA-CZ	0.51
CNC Corpora - Private Correspondence	0.27
Belebele	0.23
CNC Corpora - Dialect Corpus	0.10
HellaSwag-CZ	0.10
Czech SNLI	0.01

Table 2: Contamination Analysis.

- We released our LLM training data at huggingface.co under name BUT-FIT/BUT-LCC (320G of clean text).
- We analyzed 13-gram overlap* between sample texts & the corpus, and report proportion of samples matched in the corpus.
- We removed 4 tasks with high contamination

* (not in cases where sample did not had 13 tokens)

Model Contamination

- We ‘ve hidden BigBench’s **canary string** in our dataset
  - Canary string is a long unique hash, LLM has no chance to predict unless it saw the data

- **A sufficient condition**: If LLM predicts (the rest of) canary string
  - It was trained on (potentially ours) benchmark data

We request that you **do not reveal examples from this dataset in plain text or images online**, to reduce the risk of leakage into model training corpora. Additionally, we incorporate this canary string `26b5c67b-`[redacted] into the dataset, to make it easier for training corpora to filter out our benchmark. **All publications that include excerpts from BenCzechMark tasks should include the canary string**, so they do not become part of future model training corpora. Our canary string is identical with Srivastava et al. (2023).

Suspect?

Incident of 7.11. 24

- We’ve added evaluation of Qwen2.5 LLM family to leaderboard:
  - Qwen is a line of models released by **Qwen Team from Alibaba Group**
  - Previous generation of Qwen was working quite well on Czech
- 32B version (of base, not instruction-aligned) model correctly predicted **canary string!**
  - Others models could be contaminated too (the condition wasn’t necessary)

Qwen Team. *Qwen2.5: A Party of Foundation Models*. September 2024, https://qwenlm.github.io/blog/qwen2.5/.

A Natural Extension of Contamination Analysis

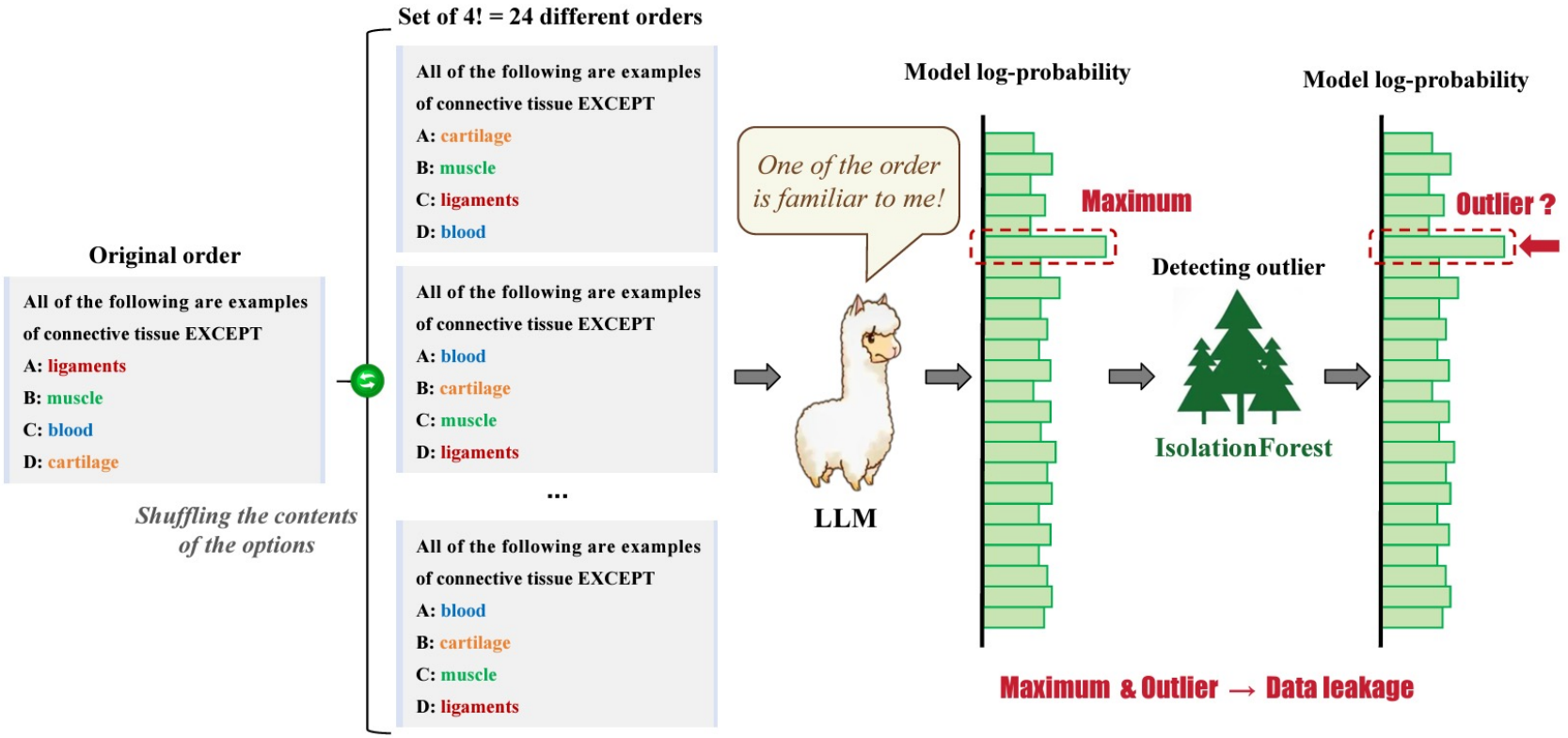


Figure 2: The order with the largest probability value, which is an outlier, indicates that the data in that order was pre-trained.

Figure borrowed from
Ni, Shiwen, et al. "Training on the Benchmark Is Not All You Need." *arXiv preprint arXiv:2409.01790* (2024).

Question Time?