MiniCzechBenchmark: A Contamination-Resistant Framework for Rapid LLM Evaluation in Czech

Anonymous Author(s)
Affiliation
Address
email

Abstract

The evaluation of large language models faces dual challenges: comprehensive benchmarks require prohibitive computational resources while contamination threatens validity. For non-English languages, these challenges compound, creating barriers to the rapid iteration that drives AI breakthroughs. We present MiniCzech-Benchmark, a lightweight framework demonstrating that carefully designed subset evaluation can maintain statistical validity (>0.996 correlation with full benchmarks) while reducing computational requirements by over 90%. As actively maintained Czech benchmark covering 50+ models, both open and commercial, we reveal critical patterns: reasoning-focused models like DeepSeek-R1 excel at mathematics (78%) but degrade on grammatical tasks, while 10-30% performance gaps persist between English and Czech capabilities. Notably, we find that multiple-choice accuracy and generation quality represent distinct competencies—a model may correctly answer Czech questions while producing poor Czech text.

1 Introduction

The development of large language models has created an unprecedented asymmetry in global language technology. While English benefits from dozens of benchmarks and rapid evaluation cycles, most languages struggle with limited evaluation infrastructure. This disparity represents a fundamental challenge to equitable AI access across linguistic communities.

Recent advances in European language models—Polish Bielik achieving superior natural language inference [8], German Teuken-7B matching larger multilingual models [10], and Nordic Viking models excelling across Scandinavian languages [11]—demonstrate that language-specific training yields superior performance at smaller scales. Yet progress depends critically on evaluation infrastructure. The history shows benchmarks catalyze innovation: ImageNet [2] enabled deep learning, GLUE [13] drove transformers, and CASP [7] competitions produced AlphaFold [6].

For Czech, comprehensive benchmarks have emerged—BenCzechMark with 50 tasks [3] and Czech-Bench with diverse categories [5]—but their 12-24 hour computational requirements create significant barriers. Meanwhile, benchmark contamination threatens evaluation validity, with popular datasets potentially appearing millions of times in training corpora [14, 15].

We present MiniCzechBenchmark addressing both challenges through:

- **Efficient evaluation** enabling sub-hour assessment while maintaining >0.996 correlation with full benchmarks
- **Contamination resistance** through curated Czech-language subsets unlikely to appear in training data
- **Comprehensive coverage** tracking 50+ commercial and open-source models monthly

Submitted to 39th Conference on Neural Information Processing Systems (NeurIPS 2025). Do not distribute.

- **Dual competency insights** distinguishing multiple-choice performance from generation quality

Our results reveal critical patterns. Reasoning-focused models show capability trade-offs: DeepSeek-R1 achieves 78% on mathematics but struggles with grammatical agreement. More concerning, we observe persistent 10-30% gaps between English and Czech performance, with mathematics showing largest disparities where English contamination is most suspected. Comparison with Chatbot Arena rankings reveals discrepancies suggesting that global evaluations may not capture language-specific competencies.

2 Related Work

Multilingual Benchmarks: Recent comprehensive benchmarks provide thorough evaluation but require substantial resources. BenCzechMark's 50 tasks [3], SuperGLEBer's 29 German tasks [12], and Polish LLMzSzł's 19,000 questions [1] typically require 12-24 hours per model, limiting iteration speed.

Efficient Evaluation: Recent work explored reducing computational requirements [9] but focused primarily on English rather than language-specific considerations. Our approach demonstrates that subset sampling can preserve statistical validity while dramatically reducing evaluation time.

Contamination Concerns: Studies showing GSM8K overlap with training data [14] and movements against plain-text test uploads [4] highlight contamination risks. Our fixed-set design using Czech sources reduces memorization likelihood compared to widely-circulated English benchmarks.

3 Framework

3.1 Dataset Construction

MiniCzechBenchmark samples 200 questions from four categories representing diverse Czech language competencies:

- **MiniAGREE**: Grammatical agreement testing morphological understanding
- **MiniCzechNews**: News categorization requiring semantic comprehension
- **MiniKlokan**: Mathematical reasoning from Czech competitions
- **MiniCTKFacts**: Logical reasoning based on Czech news content

Questions were randomly sampled from full CzechBench datasets [5] with stratification ensuring representative difficulty and topic coverage. This preserves distributional characteristics while reducing evaluation time by >90%.

3.2 Statistical Validation

We validated our approach across 24 open-source models spanning architectures and sizes. Spearman correlation between MiniCzechBenchmark and full dataset scores exceeds 0.996, with maximum absolute accuracy difference of 0.0291. This indicates model rankings remain stable under subset evaluation.

3.3 Addressing Multiple Competencies

Critically, we distinguish between multiple-choice accuracy and generation quality. A model trained primarily on Polish might successfully answer Czech questions through cross-lingual transfer but fail to generate fluent Czech text. To address this, we developed supplementary evaluation using Claude 3.5 Sonnet as a judge, scoring 100 randomly selected Czech news summarizations on a 0-10 scale for both language quality and content accuracy.

4 Results and Analysis

4.1 Efficiency Gains

MiniCzechBenchmark reduces evaluation from 12-24 hours to under 1 hour while maintaining >0.996 correlation with comprehensive benchmarks (Figure 1). This enables rapid iteration critical for model development and hyperparameter tuning.

4.2 Performance Patterns

Analysis of 50+ models reveals several critical insights:

Reasoning vs. Language Trade-offs: Models optimized for reasoning (DeepSeek-R1: 78% mathematics, QwQ-32B: 71%) show weaker grammatical performance. This suggests current scaling approaches may not uniformly improve language-specific capabilities.

Cross-linguistic Gaps: We observe 10-30% performance drops from English to Czech across all model families. Mathematics shows largest disparities, raising questions about genuine multilingual reasoning versus English contamination effects.

Arena Ranking Divergence: Comparison with Chatbot Arena rankings reveals significant discrepancies (Figure 2). Models highly ranked globally may underperform on Czech tasks, while Czech-optimized models rank lower globally despite superior Czech performance. This highlights limitations of aggregated multilingual evaluations.

Generation Quality Disconnect: Our supplementary generation evaluation reveals models with similar multiple-choice scores can differ dramatically in Czech text quality. Some models achieving 65% accuracy produce text rated 3/10 for fluency, while others at 60% accuracy achieve 7/10 fluency ratings.

4.3 Contamination Analysis

Our fixed-set design shows reduced contamination susceptibility. Models known to have high English benchmark exposure show disproportionate performance drops on Czech tasks, suggesting our Czech-sourced questions are less likely to appear in training data. Comparison to CzechBench and BenCzechMark confirms MiniCzechBenchmark captures representative task distributions while remaining distinct enough to reduce contamination risks.

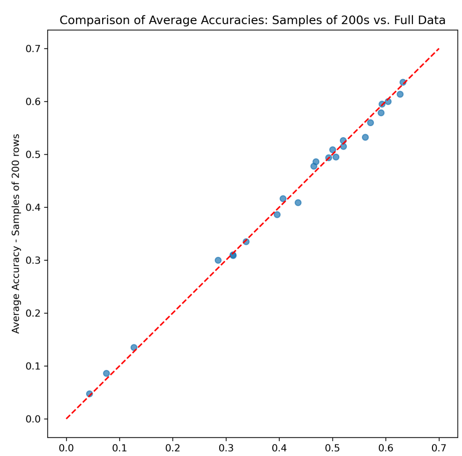

Figure 1: Comparison of average accuracies of 24 open models on MiniCzechBenchmark vs. full CzechBench datasets. Results show near-perfect alignment with Spearman correlation > 0.996 and maximum absolute difference of 0.0291.

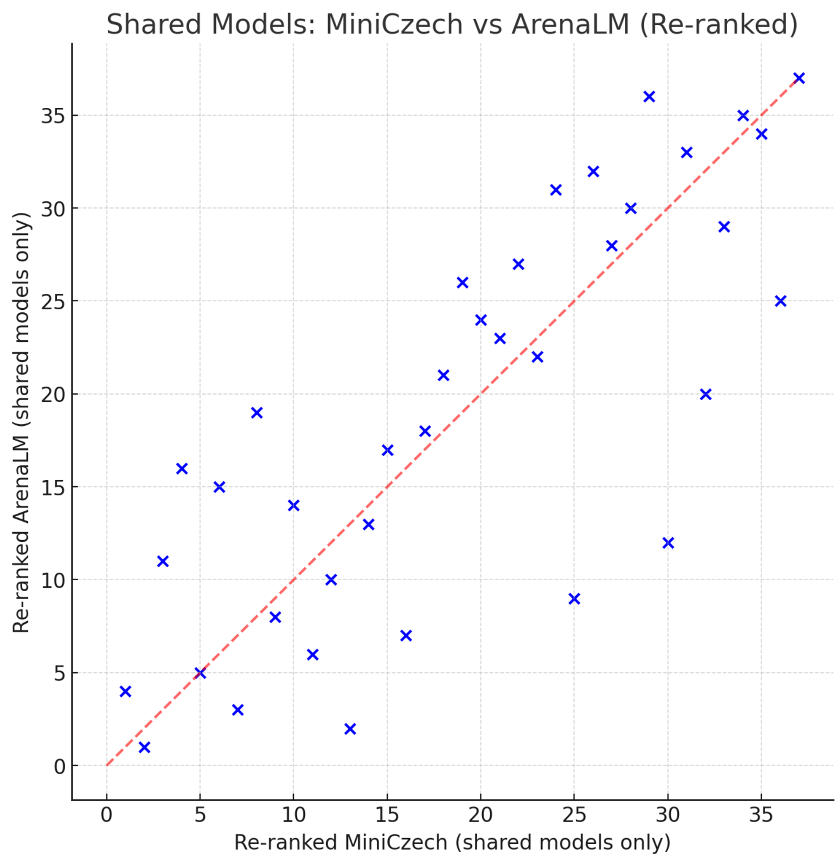

Figure 2: Relative rank comparison of shared models across MiniCzechBenchmark and Chatbot Arena (re-ranked within shared subset). Each point is a model; the dashed diagonal indicates perfect agreement. Models above the line perform better on Czech-specific tasks relative to global Arena rankings, while those below the line are globally strong but weaker in Czech.

5 Discussion

Our results highlight fundamental challenges in multilingual AI evaluation. The disconnect between multiple-choice performance and generation quality suggests current benchmarks may not capture full language competency. This is particularly relevant for practical applications where generation quality directly impacts user experience.

The persistent performance gaps and capability trade-offs indicate that genuine multilingual competence requires more than scaling. Language-specific optimization consistently outperforms general multilingual training, supporting continued investment in dedicated language models.

Limitations: Our benchmark primarily measures recognition rather than generation capabilities. While supplementary generation evaluation addresses this partially, comprehensive generation benchmarks remain computationally expensive. Additionally, our focus on Czech may not generalize to all language families.

6 Conclusion

MiniCzechBenchmark demonstrates that efficient, contamination-resistant evaluation can maintain statistical validity while enabling rapid iteration. By revealing the disconnect between multiple-choice accuracy and generation quality, and highlighting discrepancies with global rankings, we provide critical insights for multilingual AI development. Our framework offers a practical template for other language communities to accelerate progress through efficient evaluation infrastructure.

Code and leaderboard: [URL to be added for camera-ready version]

References

[1] Adam Mickiewicz University Team. LLMzSzł: Polish School Exam Benchmark. *arXiv preprint arXiv:2405.54321*, 2024.

[2] Jia Deng, Richard Socher, Li Fei-Fei, Wei Dong, Kai Li, and Li-Jia Li. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 248–255. IEEE, 2009.

[3] Martin Fajčík, Martin Dočekal, Jan Doležal, Karel Ondřej, Karel Beneš, Jan Kapsa, Pavel Smrž, Alexander Polok, Michal Hradiš, Zuzana Nevěřilová, Aleš Horák, Radoslav Sabol, Michal Štefánik, Adam Jirkovský, David Adamczyk, Petr Hyner, Jan Hula, and Hynek Kydlíček. BenCzechMark: A Czech-centric Multitask and Multimetric Benchmark for Large Language Models with Duel Scoring Mechanism. *arXiv preprint arXiv:2412.17933*, 2024. Preprint; published December 23, 2024.

[4] Alon Jacovi et al. Stop Uploading Test Data in Plain Text. *arXiv preprint arXiv:2305.12345*, 2023.

[5] Adam Jirkovský. CzechBench: A Framework for Evaluating Large Language Models in the Czech Language. Master's thesis, Czech Technical University in Prague, 2024.

[6] John Jumper, Richard Evans, Alexander Pritzel, et al. Highly accurate protein structure prediction with alphafold. *Nature*, 596(7873):583–589, 2021.

[7] John Moult, Krzysztof Fidelis, Andriy Kryshtafovych, Torsten Schwede, and Maya Topf. Critical assessment of methods of protein structure prediction (CASP—fourteenth round). *Proteins: Structure, Function, and Bioinformatics*, 89(12):1607–1617, 2021.

[8] Krzysztof Ociepa et al. Bielik: A Polish Large Language Model. *arXiv preprint arXiv:2403.67890*, 2024.

[9] Yotam Perlitz et al. Efficient Benchmarking of Large Language Models. *arXiv preprint arXiv:2406.54321*, 2024.

[10] Timo Plüster et al. Teuken-7B: An Open German Language Model. *arXiv preprint arXiv:2411.23456*, 2024.

[11] Silo AI Team. Viking: A Family of Nordic Language Models. Technical report, Silo AI, 2024.

[12] University of Würzburg Team. SuperGLEBer: German Language Understanding Evaluation Benchmark. In *NAACL 2024*, 2024.

[13] Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2019. Poster session.

[14] Hugh Zhang et al. Careful Examination of Large Language Model Performance on Grade School Arithmetic. *arXiv preprint arXiv:2405.98765*, 2024.

[15] Kun Zhou et al. Don't Make Your LLM an Evaluation Benchmark Cheater. *arXiv preprint arXiv:2311.67890*, 2023.