sciendo
DOI 10.2478/jazcas-2025-0022

FINANCIAL QUESTION-ANSWERING DATASET FOR SLOVAK LANGUAGE MODEL EVALUATION

DANIEL HLÁDEK[1] – KRISTIÁN SOPKOVIČ[2] – JÁN STAŠ[3] – ZUZANA SOKOLOVÁ[4] – MATÚŠ PLEVA[5]

[1]Department of Electronics and Multimedia Communications, Faculty of Electrical Engineering and Informatics, Technical University, Košice, Slovakia
(ORCID: 0000-0003-1148-3194)

[2]Department of Electronics and Multimedia Communications, Faculty of Electrical Engineering and Informatics, Technical University, Košice, Slovakia
(ORCID: 0009-0007-0835-3491)

[3]Department of Electronics and Multimedia Communications, Faculty of Electrical Engineering and Informatics, Technical University, Košice, Slovakia
(ORCID: 0000-0001-7403-0235)

[4]Department of Electronics and Multimedia Communications, Faculty of Electrical Engineering and Informatics, Technical University, Košice, Slovakia
(ORCID: 0000-0002-2337-8749)

[5]Department of Electronics and Multimedia Communications, Faculty of Electrical Engineering and Informatics, Technical University, Košice, Slovakia
(ORCID: 0000-0003-4380-0801)

HLÁDEK, Daniel – SOPKOVIČ, Kristián – STAŠ, Ján – SOKOLOVÁ, Zuzana – PLEVA, Matúš: Financial Question-answering Dataset for Slovak Language Model Evaluation. Journal of Linguistics, 2025, Vol. 76, No 1, pp. 247 – 257.

Abstract: The limited availability of language resources for Slovak presents a significant challenge for the development and evaluation of language models. In this paper, we introduce a multiple-choice question-answering dataset specifically designed for the financial domain in Slovak. The dataset contains 1,334 questions, each with one correct answer and four incorrect ones. It is systematically organized by topic and difficulty level to facilitate structured evaluation. Using this dataset, we assess the performance of several Slovak generative language models and compare their results against a general question-answering dataset to analyze domain-specific model capabilities. The best-performing model is a monolingual Slovak model. Furthermore, the observed performance differences between financial-domain and general question-answering tasks suggest that domain-specific language modeling requires further research.

Keywords: question answering, financial domain, large language model, evaluation, Slovak language resource

1 INTRODUCTION

The development of high-performing natural language processing (NLP) models heavily depends on the availability of high-quality datasets and evaluation benchmarks. While significant progress has been made in creating language resources for widely spoken languages such as English, low-resource languages, including Slovak, remain under-represented. This lack of resources poses challenges in training, fine-tuning, and evaluating Slovak generative language models, particularly for specialized domains like finance and law. Without domain-specific benchmarks, it is difficult to measure model performance accurately and ensure its practical applicability.

Existing Slovak language models are often evaluated on machine-translated or general-purpose datasets that do not sufficiently capture the complexity of real-world applications. Financial and legal texts, for example, involve specialized terminology and structured reasoning, which may not be well-represented in commonly available corpora.

To address these challenges, we introduce a question-answering dataset specifically designed for the legal and financial domain in the Slovak language. The dataset is structured according to topic and difficulty level, allowing for targeted assessment and benchmarking of language models. By providing a structured and domain-specific evaluation resource, we enable more precise measurement of model capabilities and facilitate further advancements in Slovak NLP. The dataset contains 1,334 questions from the financial domain. Each question has 5 possible answers; exactly one is correct. The language model can calculate the probability of each question-answer pair and select the best. This method of evaluation does not require specific fine-tuning; thus it is useful for the assessment of foundation models, trained only on unannotated data. Using the *lm-evaluation-harness* framework (Sutawika 2025), we evaluate multiple Slovak generative language models on our dataset and compare their results with those obtained on a general fact question-answering dataset.

2 STATE OF THE ART

There are multiple surveys of language model evaluation. The recent "Survey on evaluation of Large Language Models" (LLMs) (Chang 2024) claims that, as LLMs are becoming larger with more emergent abilities, existing evaluation protocols may not be enough to evaluate their capabilities and potential risks.

Guo (2023) categorizes the evaluation of LLMs into three major groups:

1. knowledge and capability evaluation,
2. alignment evaluation,
3. and safety evaluation.

In addition, it collates a compendium of evaluations pertaining to LLM performance in specialized domains, and discusses the construction of comprehensive evaluation platforms that cover LLM evaluations on capabilities, alignment, safety, and applicability.

2.1 General language model benchmarks

The Holistic Evaluation of Language Models (HELM) (Liang 2023) is a comprehensive framework developed to enhance the transparency and understanding of large language models (LLMs). HELM addresses the vast array of potential use cases and evaluation metrics by establishing a taxonomy that identifies and categorizes these scenarios and desiderata. By acknowledging existing gaps and under-represented areas, HELM provides a more inclusive and thorough assessment of LLMs. It evaluates LLMs across seven key metrics: accuracy, calibration, robustness, fairness, bias, toxicity, and efficiency. In a large-scale evaluation, HELM assessed 30 prominent LLMs across 42 scenarios, half of which were novel to mainstream evaluation.

The best language model in a general evaluation benchmark might not be the best Slovak language model. Lai et al. (2023) introduced Okapi, a system with instruction-tuned LLMs based on reinforcement learning with human feedback (RLHF) in 26 diverse languages to facilitate experiments and the development of future multilingual research. It also created benchmark datasets to enable the evaluation of generative LLMs in multiple languages including Slovak. The Okapi evaluation benchmark machine-translated is widely used English datasets. It leverages four datasets in the HuggingFace Open LLM Leaderboard, i.e., AI2 Reasoning Challenge (ARC) (Clark 2018), HellaSwag (Zellers 2019), MMLU (Hendrycks 2020), TruthfulQA (Lin 2022) to evaluate multilingual fine-tuned LLMs.

The *lm-evaluation-harness* framework (Sutawika 2025) is a widely used open-source tool designed to systematically evaluate language models across various tasks. It provides a standardized methodology for assessing model performance by enabling direct comparisons across different architectures, datasets, and evaluation metrics. The framework supports a diverse range of question-answering, reasoning, and text generation tasks, making it particularly useful for benchmarking generative language models. By using *lm-evaluation-harness*, researchers can ensure consistency in evaluation, reducing biases introduced by ad-hoc testing procedures. Its compatibility with multiple pre-trained models allows for seamless integration and fair performance assessment across various NLP applications.

2.2 Financial language model benchmarks

General language model benchmarks often fail to accurately reflect a model's performance within specialized domains, such as finance. Besides that, they are often specific to the English-speaking cultural domain. To address this limitation, several financial domain-specific benchmarks have been developed. For instance, the Financial Language Understanding Evaluation (FLUE) (Shah 2022) offers a comprehensive suite of tasks tailored to financial contexts, including sentiment analysis and named entity recognition. Labrak (2023) evaluates four state-of-the-art instruction-tuned large language models on a set of 13 real-world clinical and biomedical natural NLP tasks in English, such as named-entity recognition, question-answering or relation extraction.

Z. Guo et al. (2023) present FinLMEval, a framework for Financial Language Model Evaluation, comprising nine datasets designed to evaluate the performance of language models in English. X. Guo et al. (2023) present FinEval – a benchmark to evaluate Chinese language models in the financial domain. The dataset contains 8,351 multiple choice questions categorized into four different key areas: Financial Academic Knowle.g. Financial Industry Knowle.g. Financial Security Knowle.g. and Financial Agent.

2.3 Language model benchmarks with Slovak support

There are only a couple of publicly available monolingual Slovak datasets suitable for fine-tuning or evaluation of a generative language model.

GEST (Pikuliak 2024) is a manually created dataset designed to measure gender-stereotypical reasoning in language models and machine translation systems. GEST contains samples for 16 gender stereotypes about men and women in the English language and 9 Slavic languages, including Slovak.

The largest manually annotated set of questions and answers from Wikipedia in Slovak is SkQuAD (Hládek 2023). It consists of more than 91k factual questions and answers from various fields. Each question has an answer marked in the corresponding paragraph. It also contains negative examples in the form of "unanswered questions" and "plausible answers".

The same authors automatically translated the original SQUAD (Rajpurkar 2018) into Slovak (Staš 2023). The dataset was automatically translated from the original English SQuAD v2.0 using the Marian neural machine translation together with the Helsinki-NLP Opus English-Slovak model.

SlovakSum is a Slovak news summarization dataset consisting of over 200,000 news articles with titles and short abstracts obtained from multiple Slovak newspapers. The abstractive approach, including mBART and mT5 models, was used to evaluate various baselines in by Ondrejová (2024).

Annotation of the named entities is one of the tasks common for generative model evaluation, for example in mT5 family of models (Xue 2021). WikiGoldSK (Šuba 2023) is a dataset consisting of 10,000 manually annotated named entities in over 400 Wikipedia pages.

2.4 Generative models with Slovak support

Mistral (Jiang 2023) is a series of advanced language models developed by Mistral AI, designed to handle complex multilingual tasks with robust reasoning capabilities. The flagship model, Mistral Large, is fluent in multiple languages, including Slovak. With a context window of 32,000 tokens, it can accurately recall information from extensive documents, facilitating precise and contextually relevant text generation, advanced reasoning and agentic capabilities. Slovak Mistral (mistral-sk-7b) is a Slovak language model created by full fine-tuning of the Mistral-7B-v0.1 large language model (Jiang 2023) with data from the Araneum Slovacum VII Maximum web corpus (Benko 2024). The model was developed in collaboration with the Technical University of Košice and the Slovak Academy of Sciences. Presently, the model is not fine-tuned to follow instructions.

LLaMA (Large Language Model Meta AI) is a series of open-source language models developed by Meta, designed to advance natural language understanding and generation. These models are pre-trained on a diverse range of languages, enabling them to perform effectively across multiple linguistic contexts (Grattafiori 2024).

Qwen (Yang 2024), developed by Alibaba, is another prominent series of language models emphasizing multilingual proficiency. These models have demonstrated strong performance in multilingual tasks, making them suitable for applications requiring cross-lingual understanding and generation.

RWKV (Peng 2023) is a language model architecture different from the classic transformer encoder-only models. It combines the strengths of recurrent neural networks (RNNs) and transformers, aiming to offer efficient training and inference capabilities. Its design inherently supports sequential data processing, which can be advantageous for modeling languages with complex syntactic structures.

3 THE PROPOSED DATASET

The proposed dataset consists of 1,334 questions from the financial advisor certification of the Slovak National Bank, according to § 22 Act. no. 186/2009 Z. z., valid until 5.8.2023. The questions are published in XML and PDF format on the website of the Slovak National Bank (NBS). We parsed the test, extracted meta-information about the difficulty level and the category. The test evaluates knowledge of the applicant in the areas in Tab. 1. Tab. 2 displays the detailed table of contents of the dataset within formation about the question category, difficulty level and identification number; the example questions and answers are in Tab. 3 and Tab. 4.

Acronym	**Name**	**Translation**
VSE	*Všeobecná časť*	General questions
PaZ	*Sektor poistenia alebo zaistenia*	Insurance or reinsurance
KT	*Sektor kapitálového trhu*	Capital market
Vkl	*Sektor prijímania vkladov*	Deposits
Uv	*Sektor poskytovania úverov*	Credit granting
DDS	*Sektor doplnkového dôchodkového sporenia*	Supplementary pension
DSS	*Sektor starobného dôchodkového sporenia*	Retirement pension savings

Tab. 1. The categories of the dataset

Acronym	**Topic**	**Difficulty Level**	**Last ID**	**Count**
VSE	General questions	1	236	236
PaZ	Insurance or reinsurance	2	373	137
PaZ	Insurance or reinsurance	3	438	65
KT	Capital market	2	606	168
KT	Capital market	3	646	40
Vkl	Deposits	2	792	146
Vkl	Deposits	3	850	58
Uv	Credit granting	2	1005	135
Uv	Credit granting	3	1032	27
DDS	Supplementary pension	2	1171	139
DDS	Supplementary pension	3	1228	57
SDS	Retirement pension savings	2	1309	81
SDS	Retirement pension savings	3	1334	25

Tab. 2. Detailed table of contents of the dataset

Question	*Blízkou osobou v priamom rade je:*	A close person in the direct line is:
A	*Bratranec*	Cousin
B	*Otcov brat*	Father’s brother
C	*Druh-družka*	Spouse
D	***Syn***	**Son**
E	*Neter*	Niece

Tab. 3. Example general question and answers (The correct answer is D.)

Question	*Národná evidencia vozidiel je:*	The National Vehicle Registry is:
A	*Evidencia všetkých poistených vozidiel v poistnom kmeni poisťovne vedená Národnou bankou Slovenska*	A registry of all insured vehicles in the insurance portfolio of an insurance company maintained by the National Bank of Slovakia
B	*Evidencia všetkých vozidiel, ktoré predajca predal v kalendárnom roku vedená Slovenskou obchodnou inšpekciou*	A registry of all vehicles sold by a seller in a calendar year maintained by the Slovak Trade Inspection
C	***Informačný systém o motorových vozidlách evidovaných v Slovenskej republike evidovaný Ministerstvom vnútra Slovenskej republiky***	**An information system on motor vehicles registered in the Slovak Republic maintained by the Ministry of the Interior of the Slovak Republic**
D	*Evidencia vozidiel ktoré sú v premávke na pozemných komunikáciách v Slovenskej republike vedený Ministerstvom vnútra Slovenskej republiky*	A registry of vehicles in traffic on land roads in the Slovak Republic maintained by the Ministry of the Interior of the Slovak Republic

E	*Elektronický informačný systém o vlastníkoch motorových vozidiel v Slovenskej republike ktorý spravuje Slovenská kancelária poisťovateľov.*	An electronic information system on motor vehicle owners in the Slovak Republic administered by the Slovak Insurers' Office.

Tab. 4. Example question and answers from the insurance category (The correct answer is C.)

4 EXPERIMENTS

In this study, we evaluate the performance of widely used LLMs across two distinct datasets. To ensure comparability and reproducibility, we select open-source models with approximately 7 billion parameters. The foundation models are trained solely for next-token prediction and are not further adapted to understand instructions. In contrast, instruction fine-tuned models are evaluated to assess the impact of fine-tuning on task performance. The selected generative models are used as-is, without additional fine-tuning.

The primary research questions in this study are:

1. Does performance on the financial-domain dataset correlate with that on the SKQuad dataset?
2. Does instruction fine-tuning improve performance in both tasks?

We investigate the extent to which instruction fine-tuning enhances both multiple-choice and generative question-answering tasks, providing insights into its effectiveness across different evaluation settings.

4.1 Evaluation metrics

The models are tested on the presented financial multiple-choice dataset, as well as on a general open-domain question-answering dataset, SKQuad. The financial dataset consists of manually curated multiple-choice questions where the model is tasked with selecting the most probable answer from a set of options. In contrast, the SKQuad dataset follows a generative question-answering paradigm, where the model generates an answer after reading a provided context and question. The generated response is then compared to the expected answer to evaluate accuracy.

To measure model performance, we employ different evaluation metrics tailored to each dataset type. For the multiple-choice financial dataaset, we compute normalized accuracy to account for the length of the possible answer. The evaluation system takes the question and possible answer together and calculates the probability of an answer.

The answer with the highest probability is chosen as the generated answer. Furthermore, the answer probability is divided by the number of its words to mitigate too long answers.

In the SKQuad dataset, performance is measured using standard text similarity metrics such as F1-score, ensuring a fair comparison between generated and expected answers. The question is used as a prompt and the language model generates the answer. The generated and expected answer can consist of several words. The F1 metric calculates the overlap between the generated and expected answers. This metric is considered the standard for this dataset.

4.2 Evaluation results

Results of the evaluation of the foundation models are presented in Tab. 5; the instruction models in Tab. 6. According to the experiments, the correlation between performance on the financial-domain dataset and the SKQuad dataset appears weak. While some models, such as Gemma 7B, maintain relatively strong performance across both datasets, others, such as Slovak Mistral 7B, achieve high accuracy in the financial domain but comparatively lower scores in SKQuad. This fact shows that the proposed financial dataset evaluates different abilities of the language model rather than the database of general facts.

Instruction fine-tuning of multilingual LLMs does not show a clear and consistent improvement across both tasks. These findings suggest that instruction fine-tuning has variable effects depending on the model architecture and task, highlighting the need for task-specific tuning strategies.

The best model for answering questions from the financial domain is fine-tuned with a large corpus of the Slovak web data. Its normalized accuracy is 46.6, which is much better than pure random selection, but the model still answers more than half of the questions incorrectly. Taking the current rules into the account, the best Slovak language model still can not become a certified financial advisor. For that, we would need a better model and more data for the model fine-tuning.

Dataset and metric	Slovak Financial Normalized Acc	SKQuad F1
Gemma 7B	40.70	42.52
Qwen 2.5 7B	33.58	48.62
LLama 3.2 3B	34.63	38.40
Slovak Mistral 7B	46.62	41.01
Mistral 7B 0.3	33.80	40.27

Tab. 5. Foundation models evaluation

Dataset and metric	Slovak Financial Normalized Acc	SKQuad F1
Gemma 7B	32.95	48.76
Mistral 7B 0.3	34.63	45.09
Qwen 2.5 7B	34.63	35.57
RWKV-6-finch-7B	39.80	25.07

Tab. 6. Instruct models evaluation

ACKNOWLEDGEMENTS

The research was partially supported by the Ministry of Education, Development, Research, and Youth of the Slovak Republic under the projects KEGA 049TUKE-4/2024, by the Slovak Research and Development Agency under the projects APVV-22-0414 & APVV-22-0261, as well as by the Erasmus+ MetaCog 2024-1-FI01-KA220-HED-000247530 project.

References

Benko, V. (2024). The Aranea Corpora Family: Ten+ Years of Processing Web-Crawled Data. Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 15048 LNAI, pp. 55–70. Accessible at: https://doi.org/10.1007/978-3-031-70563-2_5/TABLES/4.

Chang, Y., Wang, X. U., Yi, X., Wang, Y., Ye, W., Yu, P. S., Chang, Y., et al. (2024). A Survey on Evaluation of Large Language Models. Journal of the ACM, 37(3), 39 p. Accessible at https://doi.org/10.1145/3641289.

Clark, P., Cowhey, I., Etzioni, O., Khot, T., Sabharwal, A., Schoenick, C., and Tafjord, O. (2018). Think you have Solved Question Answering? Try ARC, the AI2 Reasoning Challenge. Accessible at: https://arxiv.org/abs/1803.05457v1.

Grattafiori, A., Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Vaughan, A., Yang, A., Fan, A., Goyal, A., Hartshorn, A., Yang, A., Mitra, A., Sravankumar, A., Korenev, A., Hinsvark, A., … and Ma, Z. (2024). The Llama 3 Herd of Models. Accessible at: https://arxiv.org/abs/2407.21783v3.

Guo, X., Xia, H., Liu, Z., Cao, H., Yang, Z., Liu, Z., Wang, S., Niu, J., Wang, C., Wang, Y., Liang, X., Huang, X., Zhu, B., Wei, Z., Chen, Y., Shen, W., and Zhang, L. (2023). FinEval: A Chinese Financial Domain Knowledge Evaluation Benchmark for Large Language Models. Accessible at: https://arxiv.org/abs/2308.09975v2.

Guo, Y., Xu, Z., and Yang, Y. (2023). Is ChatGPT a Financial Expert? Evaluating Language Models on Financial Natural Language Processing. Accessible at: https://arxiv.org/abs/2310.12664v1.

Guo, Z., Jin, R., Liu, C., Huang, Y., Shi, D., Supryadi, Yu, L., Liu, Y., Li, J., Xiong, B., and Xiong, D. (2023). Evaluating Large Language Models: A Comprehensive Survey. Accessible at: https://arxiv.org/abs/2310.19736v3.

Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D., and Steinhardt, J. (2020). Measuring Massive Multitask Language Understanding. ICLR 2021 – 9th International Conference on Learning Representations. Accessible at: https://arxiv.org/abs/2009.03300v3.

Hládek, D., Staš, J., Juhár, J., and Koctúr, T. (2023). Slovak Dataset for Multilingual Question Answering. IEEE Access, 11, pp. 32869–32881. Accessible at: https://doi.org/10.1109/ACCESS.2023.3262308.

Staš J., Hládek, D., and Koctúr, T. (2023). Slovak Question Answering Dataset Based on the Machine Translation of the SQuAD v2.0. Jazykovedný Časopis, 74 (1), pp. 381–390.

Jiang, A. Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D. S., Casas, D. de las, Bressand, F., Lengyel, G., Lample, G., Saulnier, L., Lavaud, L. R., Lachaux, M.-A., Stock, P., Scao, T. le, Lavril, T., Wang, T., Lacroix, T., and Sayed, W. el. (2023). Mistral 7B. Accessible at: https://arxiv.org/abs/2310.06825v1.

Labrak, Y., Rouvier, M., and Dufour, R. (2023). A Zero-shot and Few-shot Study of Instruction-Finetuned Large Language Models Applied to Clinical and Biomedical Tasks. 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation, LREC-COLING 2024 – Main Conference Proceedings, pp. 2049–2066. Accessible at: https://arxiv.org/abs/2307.12114v3.

Lai, V. D., van Nguyen, C., Ngo, N. T., Nguyen, T., Dernoncourt, F., Rossi, R. A., and Nguyen, T. H. (2023). Okapi: Instruction-tuned Large Language Models in Multiple Languages with Reinforcement Learning from Human Feedback. EMNLP 2023–2023 Conference on Empirical Methods in Natural Language Processing, Proceedings of the System Demonstrations, pp. 318–327. Accessible at: https://doi.org/10.18653/v1/2023.emnlp-demo.28.

Liang, P., Bommasani, R., Lee, T., Tsipras, D., Soylu, D., Yasunaga, M., Zhang, Y., Narayanan, D., Wu, Y., Kumar, A., New-Man, B., Yuan, B., Yan, B., Zhang, C., Cosgrove, C., Manning, C. D., Ré, C., Acosta-Navas, D., Hudson, D. A., … and Koreeda, Y. (2023). Holistic Evaluation of Language Models. Accessible at: https://doi.org/10.48550/arXiv.2211.09110.

Lin, S., Hilton, J., and Evans, O. (2022). TruthfulQA: Measuring How Models Mimic Human Falsehoods. Proceedings of the Annual Meeting of the Association for Computational Linguistics, 1, pp. 3214–3252. Accessible at: https://doi.org/10.18653/V1/2022.ACL-LONG.229.

NBS National Bank of Slovakia. Accessible at: https://regfap.nbs.sk/static/otazky/otazky-2023-08-05.pdf.

Ondrejová, V., and Šuppa, M. (2024). SlovakSum: A Large Scale Slovak Summarization Dataset, pp. 14916–14922. Accessible at: https://aclanthology.org/2024.lrec-main.1298/.

Open LLM Leaderboard – a Hugging Face Space by open-llm-leaderboard. (n.d.). Accessible at: https://huggingface.co/spaces/open-llm-leaderboard/open_llm_leaderboard#/quote [24/02/2025].

Pikuliak, M., Hrčková, A., Oreško, S., and Šimko, M. (2023). Women Are Beautiful, Men Are Leaders: Gender Stereotypes in Machine Translation and Language Modeling. Accessible at: https://arxiv.org/abs/2311.18711v3.

Peng, B., Alcaide, E., Anthony, Q., Albalak, A., Arcadinho, S., Biderman, S., Cao, H., Cheng, X., Chung, M., Du, X., Grella, M., Kranthi Kiran, G. v., He, X., Hou, H., Lin, J., Kazienko, P., Kocon, J., Kong, J., Koptyra, B., … and Zhu, R. J. (2023). RWKV: Reinventing RNNs for the Transformer Era. Findings of the Association for Computational Linguistics: EMNLP 2023, pp. 14048–14077. Accessible at: https://doi.org/10.18653/v1/2023.findings-emnlp.936.

Rajpurkar, P., Jia, R., and Liang, P. (2018). Know what you don't know: Unanswerable questions for SQuAD. ACL 2018 – 56th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference (Long Papers), 2, pp. 784–789. Accessible at: https://doi.org/10.18653/v1/p18-2124.

Shah, R. S., Chawla, K., Eidnani, D., Shah, A., Du, W., Chava, S., Raman, N., Smiley, C., Chen, J., and Yang, D. (2022). When FLUE Meets FLANG: Benchmarks and Large

Pretrained Language Model for Financial Domain. Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, pp. 2322–2335. Accessible at: https://doi.org/10.18653/V1/2022.EMNLP-MAIN.148.

Šuba, D., Šuppa, M., Kubík, J., Hamerlik, E., and Takáč, M. (2023). WikiGoldSK: Annotated Dataset, Baselines and Few-Shot Learning Experiments for Slovak Named Entity Recognition. Accessible at: https://arxiv.org/abs/2304.04026v1.

Sutawika L, Schoelkopf H. , Gao L, Abbasi B. , Biderman S., Tow J. et al. (2025). 'Eleutherai/lm-evaluation-harness: V0.4.8'. Zenodo (March 5, 2025). Accessible at: https://doi.org/10.5281/zenodo.14970487.

Xue, L., Constant, N., Roberts, A., Kale, M., Al-Rfou, R., Siddhant, A., Barua, A., and Raffel, C. (2020). mT5: A massively multilingual pre-trained text-to-text transformer. NAACL-HLT 2021 – 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Conference, pp. 483–498. Accessible at: https://doi.org/10.18653/v1/2021.naacl-main.41.

Yang, A., Yang, B., Hui, B., Zheng, B., Yu, B., Zhou, C., Li, C., Li, C., Liu, D., Huang, F., Dong, G., Wei, H., Lin, H., Tang, J., Wang, J., Yang, J., Tu, J., Zhang, J., Ma, J., … and Group, A. (2024). Qwen2 Technical Report. Accessible at: https://arxiv.org/abs/2407.10671v4.

Zellers, R., Holtzman, A., Bisk, Y., Farhadi, A., and Choi, Y. (2019). HellaSwag: Can a Machine Really Finish Your Sentence? ACL 2019 – 57th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference, pp. 4791–4800. Accessible at: https://doi.org/10.18653/V1/P19-1472.