Tracking the evolution of LLM capabilities for Belarusian with OpenAI Evals

Vladislav Poritski[1], **Oksana Volchek**[1], **Maksim Aparovich**[2],
Volha Harytskaya[1], **Pavel Smrz**[2]
[1] Independent researcher, Vilnius, Lithuania
[2] Brno University of Technology, Brno, Czech Republic
Correspondence: belarusianglue@gmail.com

Abstract

We examine how the capabilities of large language models (LLMs) have evolved on eight Belarusian language tasks contributed in 2023 to OpenAI's Evals framework. We evaluate state-of-the-art models both on the original development sets and newly created test sets. Results demonstrate significant but non-uniform progress over this period: some tasks are almost saturated, while others show minor improvement beyond trivial baselines. Error analysis shows that certain challenges haven't yet been addressed, e.g. misidentification of non-words as legitimate vocabulary, or conversion from modern to classical orthography. We release the datasets and the generated completions.[1]

1 Introduction

Large language models (LLMs) are evolving at a rapid pace, that's why it is important to measure how their performance changes over time. While holistic indicators, such as the Chatbot Arena leaderboard (Chiang et al., 2024), are tracking overall improvement, comprehensive benchmarks like BIG-bench (BIG-bench authors, 2023) and tools like lm-evaluation-harness (Gao et al., 2024) cover many specific facets of LLM performance. Evals,[2] a framework introduced by OpenAI in March 2023, is among such tools. Launched as a collaborative effort, it quickly incorporated many diverse tasks intended to be hard for then-current generation of LLMs, including tasks based on the data of low-resource languages. Linguistic challenges present in these tasks make them valuable for evaluating the generalization capabilities of models, revealing weaknesses that can be masked by the sheer volume of training data available for English.

This study revisits a set of eight tasks based on the data of Belarusian, an East Slavic language. All instances in these tasks were marked as dev set data and have been publicly available since 2023. Years later, evaluating on the original instances is insufficient: there is a risk of data contamination (Aiyappa et al., 2023), as these sets may have been incorporated into the training corpora of newer models. To address this, we create new, non-overlapping test sets that closely follow the original author's design.

Beyond simply measuring performance, these tasks can be used as diagnostic probes to analyze failure modes of state-of-the-art LLMs. For instance, syllable counting tests a model's phonological awareness, its ability to handle character-level rules, and back-translation between closely related languages (Belarusian and Russian) can be designed as an identity mapping, making any deviation a potential signal of negative transfer (Bardovi-Harlig and Sprouse, 2018).

Our contribution in this paper is twofold: (1) we evaluate recent LLMs on eight Belarusian language tasks, using both the original development sets and our newly created test sets; (2) we perform an in-depth error analysis to identify weaknesses in areas such as phonology, lexicon, and grammar.

2 Related work

Multilingual LLM evaluation has been driven by large-scale, comprehensive benchmarks like MEGA (Ahuja et al., 2023), GlotEval (Luo et al., 2025) that assess a wide array of capabilities across dozens or even hundreds of languages. While these benchmarks provide essential breadth, their coverage of any single language is often limited.

A significant body of research focuses on creating targeted diagnostic datasets to assess specific linguistic competencies. It is common in the literature that a diagnostic task is first proposed for English and then expanded to other languages. For instance, BLiMP (Warstadt et al., 2020) evaluates

[1] https://doi.org/10.5281/zenodo.18163825
[2] https://github.com/openai/evals

models' understanding of various aspects of English syntax via grammatical acceptability judgments, and MultiBLiMP (Jumelet et al., 2025) is a multilingual dataset of the same type. The word analogy dataset introduced by Mikolov et al. (2013) to evaluate semantic and grammatical knowledge captured by word vectors has since been translated to multiple European languages by Ulčar et al. (2020). Tasks proposed in PhonologyBench (Suvarna et al., 2024) to test the phonological skills of LLMs in English, such as rhyme word generation, grapheme-to-phoneme conversion, inspired creating similar tasks in Cantonese (Cheng et al., 2025). Unlike these examples, some of the tasks in Evals have been designed to assess capabilities specific to Belarusian and thus don't have any direct English counterparts.

While Belarusian is considered low-resource, an effort has recently been made to create a dedicated evaluation resource – BelarusianGLUE (Aparovich et al., 2025), a language understanding benchmark with five tasks that cover polarity classification, linguistic acceptability, word sense disambiguation, and two kinds of textual entailment. Our present work is complementary to this benchmark, as we're tracking performance longitudinally on a pre-existing set of challenges, trying to detect the impact of data contamination. Also, BelarusianGLUE isn't adversarial by design, while in Evals only those community-contributed tasks got accepted that used to be challenging for mainstream LLMs in 2023.

3 Task descriptions

Below we briefly introduce the tasks. Table 1 provides statistics on the dev and test size for each task. Further explanations of the data creation process can be found in Appendix A; prompts and sample instances are shown in Appendix B. Except minor differences that are highlighted below, we constructed the test sets following the procedures outlined by the original author.

Syllable count. Given a single Belarusian word, the model must output its number of syllables.

Numerals. Given a Belarusian cardinal numeral in text form, the model must output its corresponding representation in digits. The test set has less instances than the dev set, since small integers (0. . . 19) have already been covered.

Lexicon. The model is prompted with a single word and must determine if it is a legitimate word that exists in the Belarusian language (`Y`) or not (`N`).

Task	Topic	No. instances	
		dev	test
Syllable count	phonology	100	100
Numerals	numeral / number conversion	100	80
Lexicon	lexicology	300	300
Orthography	orthography	250	220
Rhyme	phonology, word-level translation	100	100
Inflectional analogies	morphology	300	300
Grammar	morphology, syntax	100	100
Translation	sentence-level translation	100	100

Table 1: A summary of tasks.

Orthography. Given a Belarusian word in the official modern orthography, the model must convert it to the classical orthography *taraškievica*, and vice versa.[3] The test set has less instances than the dev set, because certain phenomena have already been extensively covered, thus exhausting the pool of available examples. Also, the test set is slightly harder as it includes a few words that the rule-based tool *Taraskevizatar*[4] fails to process correctly.

Rhyme. The input is a pair of English nouns. The model must provide a pair of their Belarusian translations that rhyme, or output `NONE` if no such rhyming pair exists. In the dev set, all rhyming words are ending in *-a*; in the test set we lifted this restriction to obtain more diverse rhyming pairs.

Inflectional analogies. This task presents an analogy problem in the format "*A* is to *B* as *C* is to ?". The model must produce the correct inflected form *D* that completes the analogy, based on the morphological relation between *A* and *B*.

Grammar. The input is a single Belarusian sentence, which the model must classify as either grammatically well-formed (`Y`) or ungrammatical (`N`). The dataset is constructed from minimal pairs of correct and incorrect sentences.

Translation. This task detects deviations in translation between Belarusian and a closely related high-resource language, Russian. Given a sentence, the model must perform a round-trip translation: Belarusian ⇒ Russian ⇒ Belarusian. Instances are specifically constructed so that a perfect, lossless translation results in an output identical to the input.

In all tasks, the evaluation metric is accuracy. This allows to contextualize the LLM scores by

[3] On the two orthographic standards of Belarusian see e.g. Cychun (2002), Siwirska (2024).

[4] `https://gooseob.github.io/taraskevizatar`

providing trivial baseline solutions. Intuitively, a trivial baseline for a specific task might be defined as a very short program for solving this task, e.g. a Python one-liner, like `lambda x: "Y"` in Lexicon and Grammar (accuracy 0.5 because half of the instances are acceptable), `lambda x: "NONE"` in Rhyme (accuracy 0.5 because half of the instances aren't rhyming), `lambda x: x` in Translation (accuracy 1.0 because all outputs are the same as the inputs). A slightly less trivial baseline of Syllable count, `lambda x: sum(1 for c in x.lower() if c in "аоуэыяёюеі")`, has accuracy 1.0. For the tasks Numerals, Orthography, and Inflectional analogies, solutions of similar complexity would yield near-zero accuracy.

Due to perfect-quality baselines, it can be argued that Syllable count and Translation, unlike all other tasks, are testing the models' ability to capture simple deterministic rules, rather than language-specific capabilities. Nevertheless, as shown below, they can still help identify certain language-specific failure modes, such as negative Russian ⇒ Belarusian transfer in Translation.

4 Experiments

Between May and July 2025, we've evaluated recent commercial models with a focus on OpenAI products: the reasoning model o4-mini, the general purpose models GPT-4.1 (including mini and nano versions) and GPT-4o (including mini version). Two competing offerings – Claude 3.7 Sonnet and Gemini 2.5 Flash – were evaluated via OpenRouter API.[5] Additionally, in October 2025 we evaluated open-weight models Gemma 3 27B (Gemma Team et al., 2025), gpt-oss-120b (OpenAI et al., 2025) and Qwen3-235B-A22B (Yang et al., 2025). In o4-mini and gpt-oss-120b we set reasoning effort to "low". As a historic reference, we also report the scores of GPT-3.5 Turbo (2023 state-of-the-art).

Except o4-mini, which doesn't expose temperature T as a hyperparameter, two runs are performed for each model: with $T = 0.0$ and 1.0, and the best accuracy is reported. For the two cheapest models, Gemma and Qwen, we performed additional runs with $T = 0.2, \ldots, 0.8$, so that the best accuracy is estimated over the total of six runs.

Dev and test accuracy scores are shown in Table 2.[6] Where applicable, the trivial baseline is provided. Highlighted in bold are the best overall scores per task and the best scores of general purpose models (all but o4-mini and gpt-oss-120b).

Most tasks have approached saturation two years after they've been introduced, but the improvements aren't uniform. For example:

- In Grammar, most current models beat the baseline by a wide margin.
- In Numerals and Syllable count, the strongest models deliver near-perfect performance, and the outsiders are lagging far behind.
- In Lexicon, general purpose models improve only slightly over the trivial baseline.
- In Translation, Gemini and Claude are ahead of the competing offerings.

Looking at the differences between dev and test set scores, we find no clear evidence of data contamination, suggesting that the dev sets haven't been memorized by any of the evaluated models.[7] All models underperform on the test set of Orthography; as explained in Section 3, it is slightly harder than the dev set, due to instances crafted adversarially against a popular rule-based tool. Many models also underperform on the test set of Rhyme, which is likely related to higher diversity of rhyme endings in the test set, as noted in Section 3. One thing we cannot explain is the drop in GPT-4.1 mini, GPT-4.1 nano performance on Numerals.

Syllable count. LLMs are known to be bad at counting and manipulating numbers (Ball et al., 2024). In line with this, errors are more frequent in longer words.[8] Stronger models are less sensitive to the input length, while weaker models undercount syllables in longer words, rather then overcount. Asyllabic clitics are often predicted to have 1 syllable. In equally long words, errors are more frequent when the word contains the semivowel *ў* (IPA: `/w/`).

Numerals. On average, larger numbers are harder to produce.[9] However, instances of similar numeric magnitude aren't equally hard: e.g., all models correctly process *дванаццаць тысяч трыста дваццаць пяць* '12325.NOM', though 4/11 current models cannot process *пятнаццацю тысячамі ста двума* '15102.INS'. Two factors contribute to this. First, the digit most frequently removed or added in wrong answers vs. the ground

[5] https://openrouter.ai

[6] For the scores at $T = 0.0$, see Table 4 in Appendix C.

[7] Indeed, denote the dev set accuracy by x and the test set accuracy by y, then the least-squares approximation would be $y \approx 1.01x - 0.03$, $r^2 \approx 0.94$, suggesting almost perfect fit.

[8] See Figure 1 in Appendix C.

[9] See Figure 2 in Appendix C.

Model ID	dev								test							
	syl	**num**	**lex**	**ort**	**rhm**	**inf**	**grm**	**trn**	**syl**	**num**	**lex**	**ort**	**rhm**	**inf**	**grm**	**trn**
o4-mini-2025-04-16	**0.96**	**1**	**0.84**	0.56	0.85	0.78	**0.93**	0.36	**0.98**	0.99	**0.81**	0.49	0.66	0.78	**0.92**	0.47
gpt-4.1-2025-04-14	**0.85**	0.83	0.59	0.80	0.64	0.86	**0.89**	0.05	0.88	0.84	0.60	0.64	0.56	0.86	0.85	0.03
gpt-4.1-mini-2025-04-14	0.79	0.64	0.55	0.57	0.33	0.76	0.85	0.27	0.78	0.54	0.59	0.50	0.18	0.76	0.82	0.34
gpt-4.1-nano-2025-04-14	0.40	0.36	0.56	0.42	0.35	0.57	0.74	0.12	0.31	0.23	0.57	0.32	0.39	0.54	0.67	0.14
gpt-4o-2024-08-06	0.83	0.86	0.56	0.85	0.73	0.81	0.76	0.09	0.86	0.86	0.55	0.66	0.66	**0.87**	0.71	0.07
gpt-4o-mini-2024-07-18	0.64	0.48	0.53	0.54	0.50	0.86	0.80	0.19	0.63	0.44	0.53	0.43	0.51	0.83	0.80	0.17
claude-3.7-sonnet	0.84	0.97	**0.60**	**0.92**	**0.88**	**0.87**	0.88	0.74	**0.92**	0.99	0.59	0.79	**0.78**	0.85	**0.88**	0.79
gemini-2.5-flash	0.76	**0.99**	0.59	0.90	0.41	0.79	0.75	**0.86**	0.79	**1**	**0.61**	**0.83**	0.32	0.76	0.75	**0.88**
gemma-3-27b-it	0.46	0.84	0.57	0.70	0.61	0.70	0.64	0.29	0.50	0.80	0.58	0.52	0.54	0.65	0.60	0.29
gpt-oss-120b	0.95	0.86	0.72	0.36	0.65	0.59	0.90	0.48	0.91	0.85	0.73	0.31	0.55	0.56	0.86	0.56
qwen3-235b-a22b-2507	0.68	0.98	0.54	0.58	0.50	0.50	0.81	0.04	0.79	0.91	0.54	0.36	0.51	0.48	0.78	0.04
gpt-3.5-turbo-0125	0.63	0.54	0.51	0.36	0.50	0.35	0.51	0.17	0.66	0.35	0.54	0.26	0.50	0.34	0.53	0.12
Trivial baseline	1		0.5		0.5		0.5	1	1		0.5		0.5		0.5	1

Table 2: Dev and test set accuracy scores. Task name abbreviations: **syl** = Syllable count, **num** = Numerals, **lex** = Lexicon, **ort** = Orthography, **rhm** = Rhyme, **inf** = Inflectional analogies, **grm** = Grammar, **trn** = Translation.

truth is `0`. Second, presenting the numeral in nominative case, rather than any of the oblique cases, results in higher prediction quality.

Lexicon. Misidentification of non-words as words is a major issue: in the outputs of all general purpose models, false positives are much more frequent than false negatives.[10] Among non-words recognized by all models as words, there are common orthographic errors (*масштабны* instead of *маштабны* ‘large-scale’), structurally coherent potential wordforms (*злучанне*, *прывабляюць* instead of *злучэнне* ‘connection’, *прываблiваюць* ‘attract.PRS.3PL’), and wordforms not attested in modern Belarusian that turn out to coincide with archaic accentual variants (*звычаёў* instead of *звычаяў* ‘custom.GEN.PL’). This behavior may stem from vague phrasing of the prompt; more generally, LLMs might be biased to accept non-words due to limited amount and low quality of Belarusian texts in the training and fine-tuning data.

Orthography. Classical ⇒ modern spelling conversion is more accurate than the modern ⇒ classical spelling conversion. This is true for all models, though the absolute difference in accuracy varies.[11] A possible reason is that classical Belarusian orthography is much less widespread in the available training data. On average, instances with larger Levenshtein distance between the classical and the modern spellings are harder to solve.[12]

Rhyme. Two steps are required to arrive at the correct answer: translate both words in all possible ways, then find a matching pair if it exists. Failure at the translation step leads to false negatives: e.g., none of the models find the rhyme *гультай* ‘slacker’ – *небакрай* ‘horizon (a rare poetic word)’. Failure at the matching step leads to false negatives (not identifying a pair of translations as a rhyme) or to false positives: pairs matching graphically but not accentually, like *ба́цькаўшчына* ‘fatherland’ – *жанчы́на* ‘woman’, or not matching at all. Among the expected negative instances, o4-mini found one plausible rhyming pair overlooked by us in the test set: *змей* ‘dragon’ – *чарадзей* ‘sorcerer’.

Inflectional analogies. If the predicted wordform D' differs from the expected wordform D, the Levenshtein distance $d(C, D') < d(C, D)$ on average, that is, hypothesized inflection patterns are often simpler than the real ones, like *салавей* ‘nightingale.NOM.SG’ : **салавея* ‘nightingale.GEN.SG’ in place of *салавей* : *салаўя*. This holds for all three parts of speech in the task.[13] Wrong answers are often preceded by a longer chain of thought.

Grammar. Same as in Lexicon, the models treat some clearly unacceptable items as if they were acceptable: e.g., **Такія дзеці спрыяюць росквi**це** будучыні краіны* ‘Such children contribute to the prosperity.LOC of the country’s future’, with locative instead of dative *росквi**ту*** ‘prosperity.DAT’, is ungrammatical but accepted by all general purpose models. Errors in case government tend to be harder to recognize than errors in gender or number agreement, although the evidence is inconclusive.

Translation. Most models don’t use the “identity mapping” trick: instead, they apply non-obligatory lexical and grammatical transformations. For instance, only 4% of GPT-4.1’s back-

[10] See Table 5 in Appendix C.

[11] See Table 6 in Appendix C.

[12] See Figure 3 in Appendix C.

[13] See Table 7 in Appendix C.

translations are the same as the inputs, even though the meanings are essentially preserved. The transformed utterances show the influence of Russian in lexical choice (*квіткі* ⇒ *білеты* ‘tickets’) and, occasionally, grammar (**вялікі дзюба* ‘large.M beak.F’). Models favor frequent lexical items and neutral structural options, such as unmarked word order, no omitted constituents, and no discourse markers. While exact match accuracy is a very crude measure in this task, BLEU scores[14] are strongly correlated with it. Qwen often misinterprets the prompt and returns the intermediate Russian translation, resulting in low BLEU score.

5 Conclusion

While the multilingual capabilities of mainstream LLMs, including Belarusian language support, have vastly improved since 2023, our evaluation shows that some of the challenges observed back then haven’t yet been addressed, e.g., misidentification of non-words as legitimate vocabulary, or modern ⇒ classical orthography conversion. The performance disparities between tasks hint at the influence of higher-resource languages, primarily Russian, which may explain why more general tasks (Syllable count, Numerals) show stronger performance than language-specific tasks. In some of the latter (e.g., Rhyme, Translation), we observe a preference for higher-frequency variants, common to more than one language, and a tendency toward simpler, more regular structures.

Limitations

Our evaluation is restricted to a small set of mainstream commercially available and open-weight LLMs. Currently, cutting-edge reasoning models aren’t covered, which limits the scope of our analysis. Agentic tool use (Plaat et al., 2025, §3.2) is also beyond the scope of the present contribution: we focus on measuring raw model capabilities, rather than their augmentation with external tools, such as a spell checker in Lexicon or an orthography converter in Orthography. Due to limited budget, we haven’t performed more extensive testing for data contamination, e.g. with prompt perturbation (Deng et al., 2024, §4.4), leaving it to future work.

Acknowledgments

The authors thank Ilia Afanasev and three anonymous reviewers for their valuable feedback.

[14] See Table 8 in Appendix C.

References

Kabir Ahuja, Harshita Diddee, Rishav Hada, Millicent Ochieng, Krithika Ramesh, Prachi Jain, Akshay Nambi, Tanuja Ganu, Sameer Segal, Maxamed Axmed, Kalika Bali, and Sunayana Sitaram. 2023. MEGA: Multilingual evaluation of generative AI. *Preprint*, arXiv:2303.12528.

Rachith Aiyappa, Jisun An, Haewoon Kwak, and Yongyeol Ahn. 2023. Can we trust the evaluation on ChatGPT? In *Proceedings of the 3rd Workshop on Trustworthy Natural Language Processing (TrustNLP 2023)*, pages 47–54, Toronto, Canada. Association for Computational Linguistics.

Maksim Aparovich, Volha Harytskaya, Vladislav Poritski, Oksana Volchek, and Pavel Smrz. 2025. BelarusianGLUE: Towards a natural language understanding benchmark for Belarusian. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 511–527, Vienna, Austria. Association for Computational Linguistics.

Thomas Ball, Shuo Chen, and Cormac Herley. 2024. Can we count on LLMs? The fixed-effect fallacy and claims of GPT-4 capabilities. *Preprint*, arXiv:2409.07638.

Kathleen Bardovi-Harlig and Rex A. Sprouse. 2018. Negative versus positive transfer. *The TESOL Encyclopedia of English Language Teaching*, pages 1–6.

BIG-bench authors. 2023. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *Transactions on Machine Learning Research*.

Tsz Chung Cheng, Chung Shing Cheng, Chaak Ming Lau, Eugene Tin-Ho Lam, Chun Yat Wong, Hoi On Yu, and Cheuk Hei Chong. 2025. HKCanto-Eval: A benchmark for evaluating cantonese language understanding and cultural comprehension in llms. *Preprint*, arXiv:2503.12440.

Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Hao Zhang, Banghua Zhu, Michael Jordan, Joseph E. Gonzalez, and Ion Stoica. 2024. Chatbot arena: An open platform for evaluating LLMs by human preference. *Preprint*, arXiv:2403.04132.

Hienadź Cychun. 2002. Weißrussisch. In Miloš Okuka and Gerald Krenn, editors, *Lexikon der Sprachen des europäischen Ostens*, pages 563–579. Wieser Verlag, Klagenfurt.

Chunyuan Deng, Yilun Zhao, Yuzhao Heng, Yitong Li, Jiannan Cao, Xiangru Tang, and Arman Cohan. 2024. Unveiling the spectrum of data contamination in language model: A survey from detection to remediation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 16078–16092, Bangkok, Thailand. Association for Computational Linguistics.

Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac'h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, and 5 others. 2024. The language model evaluation harness.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, and 197 others. 2025. Gemma 3 technical report. *Preprint*, arXiv:2503.19786.

Jaap Jumelet, Leonie Weissweiler, Joakim Nivre, and Arianna Bisazza. 2025. MultiBLiMP 1.0: A massively multilingual benchmark of linguistic minimal pairs. *Preprint*, arXiv:2504.02768.

Hengyu Luo, Zihao Li, Joseph Attieh, Sawal Devkota, Ona de Gibert, Shaoxiong Ji, Peiqin Lin, Bhavani Sai Praneeth Varma Mantina, Ananda Sreenidhi, Raúl Vázquez, Mengjie Wang, Samea Yusofi, and Jörg Tiedemann. 2025. GlotEval: A test suite for massively multilingual evaluation of large language models. *Preprint*, arXiv:2504.04155.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *Preprint*, arXiv:1301.3781.

OpenAI, Sandhini Agarwal, Lama Ahmad, Jason Ai, Sam Altman, Andy Applebaum, Edwin Arbus, Rahul K. Arora, Yu Bai, Bowen Baker, Haiming Bao, Boaz Barak, Ally Bennett, Tyler Bertao, Nivedita Brett, Eugene Brevdo, Greg Brockman, Sebastien Bubeck, Che Chang, and 107 others. 2025. gpt-oss-120b & gpt-oss-20b model card. *Preprint*, arXiv:2508.10925.

Aske Plaat, Max van Duijn, Niki van Stein, Mike Preuss, Peter van der Putten, and Kees Joost Batenburg. 2025. Agentic large language models, a survey. *Preprint*, arXiv:2503.23037.

Anna Berenika Siwirska. 2024. Czy we współczesnej białoruszczyźnie istnieją dwie normy językowe? *Studia Białorutenistyczne*, 18:269–283.

Ashima Suvarna, Harshita Khandelwal, and Nanyun Peng. 2024. PhonologyBench: Evaluating phonological skills of large language models. In *Proceedings of the 1st Workshop on Towards Knowledgeable Language Models (KnowLLM 2024)*, pages 1–14, Bangkok, Thailand. Association for Computational Linguistics.

Matej Ulčar, Kristiina Vaik, Jessica Lindström, Milda Dailidėnaitė, and Marko Robnik-Šikonja. 2020. Multilingual culture-independent word analogy datasets. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4074–4080, Marseille, France. European Language Resources Association.

Alex Warstadt, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang, and Samuel R. Bowman. 2020. BLiMP: A benchmark of linguistic minimal pairs for English. In *Proceedings of the Society for Computation in Linguistics 2020*, pages 409–410, New York, New York. Association for Computational Linguistics.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

A Task details

In addition to the 8 tasks submitted by one contributor who described how the instances were produced, 3 tasks by other contributors are available for Belarusian in Evals. They are documented less extensively, so that it is harder to ensure that any newly-created instances would be comparable with the original ones. Therefore we opted not to construct the test sets for these tasks, instead skipping their evaluation.

The new test sets were constructed by three experts – native speakers of Belarusian with degrees in linguistics (M.A. or Ph.D.). Each instance produced by one expert was then manually validated by at least one other expert.

Syllable count. Frequent wordforms were sampled from Belarusian texts on the web, with syllable counts ranging between 0 (in asyllabic clitics) and 9 (in compound adjectives and nouns).

Numerals. Belarusian cardinal numerals in various (randomly assigned) grammatical forms, representing integers between 0 and 999999999, were generated programmatically. The dev set has 100 instances: each of the numbers 0. . . 19, ten random numbers from the range 20. . . 99, and ten more numbers sampled randomly from each subsequent decimal order: $[10^2, 10^3), \ldots, [10^8, 10^9)$. The test set has 80 instances, i.e. except 0. . . 19.

Lexicon. The inputs constitute pairs: one item in each pair is a hallucinated non-word (either meaningless in Belarusian or violating orthographic and phonetic rules of the standard language), and another item is an actual Belarusian word with similar spelling. Non-words were produced by GPT-3.5

Turbo in the dev set, and by multilingual open-weight LLMs (including Llama 3.1, Mistral Nemo, Gemma 3 and others) in the test set.

Orthography. Today most speakers of Belarusian are using the official modern orthography, as taught in school, and some speakers prefer the classical orthography *taraškievica*, which is in fact an alternative literary norm. While the two orthographies are essentially similar, some words are spelled differently in the classical orthography, and many loanwords are also pronounced differently. The dataset contains Belarusian words that represent a wide range of discrepancies between the two orthographies. All instances were sampled from the two Belarusian Wikipedias[15] and aligned using the rule-based tool *Taraskevizatar*.[16] The inputs constitute pairs: given a word w in the modern spelling, the task is to produce the corresponding word w' in the classical spelling, and vice versa. In the dev set, w and w' always differ; the test set includes 15 pairs of instances with $w' = w$ that used to be handled incorrectly by *Taraskevizatar*.

Rhyme. The dataset contains pairs of English nouns whose Belarusian translations rhyme, and different pairs consisting of the same nouns but reordered, so that in each of those pairs there aren't any Belarusian translations that rhyme. The rhyming pairs have been picked manually, and most of them contain at least one word distinctive of Belarusian, i.e. not attested in closely related Russian and Ukrainian languages. The dev and test sets each contain 50 rhyming and 50 non-rhyming instances. In the dev set, all rhyming words are ending in *-a*, in the test set there is no such restriction. Note that the task isn't open-ended, unlike Rhyme Word Generation (Suvarna et al., 2024) in which any valid rhyme for the given word is accepted: instead, the rhyming pair of translations either is unique or doesn't exist.

Inflectional analogies. The dataset contains word analogy problems in the field of Belarusian inflectional morphology. A word analogy (Mikolov et al., 2013) is a statement "*A* is to *B* as *C* is to *D*", where the relation between the words *A* and *B* is the same as the relation between *C* and the held-out word *D*. Mikolov et al.'s dataset includes, among others, "syntactic" relations, which could be more appropriately called morphological, e.g. singular / plural (*bird* : *birds* :: *child* : *children*), adjective / adverb (*apparent* : *apparently* :: *obvious* : *obviously*), etc. Unlike English, Belarusian is a morphologically rich language, so the types of relations between inflected wordforms are quite diverse, with one or few examples per each relation type. To make the task even more challenging, it was ensured that none of the instances can be solved by straightforward string manipulation, such as adding *-ly* in English adjective / adverb analogies. The dev set has 300 instances representing three open-class parts of speech (noun, adjective, verb). The test set is derived from the dev set by replacing a half of each analogy (either *A* : *B* or *C* : *D* but not both) with a structurally similar pair picked from the grammatical database of Belarusian.

Grammar. The inputs constitute pairs: one Belarusian sentence in each pair, taken from LLM output with minor adjustments, is ungrammatical (i.e. has mismatches in gender / number / case agreement, incorrect verbal forms, etc.), and another sentence is its counterpart corrected by a native speaker. All words in all sentences are valid Belarusian words. Ungrammatical sentences were produced by GPT-3.5 Turbo in the dev set, and by multilingual open-weight LLMs in the test set. Except the data source, this task is essentially similar to BelaCoLA in BelarusianGLUE (Aparovich et al., 2025).

Translation. The dataset contains Belarusian sentences from Common Voice,[17] filtered with a 2-step procedure:

1. A batch of randomly sampled sentences was processed with Belazar,[18] a rule-based be ⇔ ru translation tool, and only those sentences were kept in consideration which get mapped to themselves after be ⇒ ru ⇒ be translation.
2. Several hundred of these sentences were independently manually verified by two human experts: one did the Russian translations, then the other one did back-translations into Belarusian without accessing the originals. This continued until accumulating 100 sentences whose manual back-translations are confirmed to be the same as the original sentences.

[15] https://be.wikipedia.org and https://be-tarask.wikipedia.org

[16] https://gooseob.github.io/taraskevizatar

[17] https://github.com/common-voice/common-voice/tree/main/server/data/be

[18] http://belazar.info

B Prompts and sample instances

Task	Prompt	Input	Output
Syllable count	You will be prompted with a single Belarusian word. Your output must be the number of syllables in this word (a single digit). Return only this number and nothing else.	аднак	2
		апублікаваў	5
		археалагічныя	7
Numerals	You will be prompted with a Belarusian cardinal numeral, representing an integer number between 0 and 999999999. Your task is to convert this numeral to its respective number. Don't use any delimiters, output only the digits and nothing else.	пяці тысячах чатырохстах дзевяноста адным	5491
Lexicon	You will be prompted with a single word. Does this word exist in Belarusian language? Answer Y or N.	абоўязак	N
		абавязак	Y
Orthography	You will be prompted with a single Belarusian word written in the classical orthography, also known as Taraškievica. Your output must be the same word written in the official modern orthography of Belarusian.	ісьляндзкі	ісландскі
	You will be prompted with a single Belarusian word written in the official modern orthography. Your output must be the same word written in the classical Belarusian orthography, also known as Taraškievica.	ісландскі	ісьляндзкі
Rhyme	For each pair of words, determine whether some of their Belarusian translations rhyme. If they do, output the pair of rhyming words in Belarusian. If not, output NONE.	food, tower	ежа, вежа *or* вежа, ежа
		church, tower	NONE
Inflectional analogies	A word analogy problem has structure A : B :: C : ?, where A, B, C are some words and the relation between A and B is the same as between C and a missing word, represented by a question mark. The task is to fill in the missing word. You will be prompted with a word analogy problem in the field of Belarusian inflectional morphology. That is, A and B are two distinct Belarusian wordforms belonging to the same lemma, and C is a Belarusian wordform belonging to another lemma but sharing all its morphological characteristics with A. The missing wordform, represented by a question mark, should belong to the same lemma as C and share all its morphological characteristics with B. Your task is to output this wordform. Provide the chain of thought before answering, and enclose the final answer in square brackets, like this: [word]	журавель : жураўля :: салавей : ?	[салаўя]
		вострым : вастрэйшым :: глыбокім : ?	[глыбейшым]
		жыў : жывеш :: працаваў : ?	[працуеш]
Grammar	You will be prompted with a sentence. Is this sentence grammatically well-formed in Belarusian language? Answer Y or N.	Мы зайшлі ў кавярню для павячэраць.	N
		Мы зайшлі ў кавярню павячэраць.	Y
Translation	You will be prompted with a sentence in Belarusian. Your output must be a single Belarusian sentence, the result of Belarusian => Russian => Belarusian translation of the input. Please note: it must not be Russian, it must be the final result of translating into Russian and then back into Belarusian.	Вялікая адказнасць ляжыць на старшыні калгаса.	Вялікая адказнасць ляжыць на старшыні калгаса.

Table 3: Prompts and sample instances.

C Detailed evaluation results

Model ID	dev								test							
	syl	num	lex	ort	rhm	inf	grm	trn	syl	num	lex	ort	rhm	inf	grm	trn
o4-mini-2025-04-16	**0.96**	**1**	**0.84**	0.56	0.85	0.78	**0.93**	0.36	**0.98**	0.99	**0.81**	0.49	0.66	0.78	**0.92**	0.47
gpt-4.1-2025-04-14	**0.85**	0.81	0.59	0.80	0.64	**0.86**	**0.89**	0.05	0.85	0.84	0.59	0.64	0.56	0.84	0.84	0.03
gpt-4.1-mini-2025-04-14	0.75	0.64	0.55	0.57	0.33	0.76	0.85	0.27	0.78	0.54	0.59	0.50	0.18	0.74	0.82	0.34
gpt-4.1-nano-2025-04-14	0.36	0.36	0.53	0.42	0.35	0.52	0.72	0.12	0.31	0.23	0.57	0.32	0.33	0.54	0.67	0.14
gpt-4o-2024-08-06	0.83	0.86	0.55	0.85	0.73	0.81	0.76	0.09	0.86	0.86	0.54	0.66	0.66	**0.87**	0.70	0.07
gpt-4o-mini-2024-07-18	0.64	0.48	0.52	0.54	0.49	0.85	0.80	0.19	0.62	0.43	0.53	0.43	0.50	0.81	0.80	0.17
claude-3.7-sonnet	0.84	0.97	**0.60**	**0.92**	**0.88**	**0.86**	0.88	0.74	**0.92**	0.99	0.59	0.79	**0.78**	0.84	**0.88**	0.79
gemini-2.5-flash	0.76	**0.99**	0.59	0.90	0.41	0.78	0.75	**0.86**	0.79	**1**	**0.61**	**0.83**	0.25	0.76	0.75	**0.88**
gemma-3-27b-it	0.46	0.83	0.57	0.68	0.61	0.68	0.63	0.29	0.50	0.80	0.58	0.52	0.54	0.64	0.60	0.29
gpt-oss-120b	0.95	0.86	0.71	0.36	0.65	0.59	0.90	0.48	0.91	0.81	0.72	0.31	0.53	0.56	0.82	0.56
qwen3-235b-a22b-2507	0.68	0.98	0.52	0.58	0.50	0.50	0.81	0.04	0.79	0.91	0.54	0.36	0.50	0.48	0.78	0.04
gpt-3.5-turbo-0125	0.63	0.54	0.51	0.35	0.50	0.33	0.51	0.17	0.66	0.34	0.54	0.26	0.50	0.32	0.52	0.12
Trivial baseline	1		0.5		0.5		0.5	1	1		0.5		0.5		0.5	1

Table 4: Dev and test set accuracy scores with $T = 0.0$ (all except o4-mini that doesn't expose temperature). Task name abbreviations are the same as in Table 2.

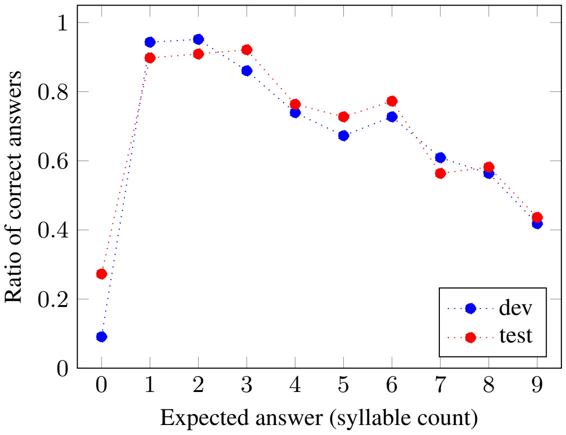

Figure 1: Syllable counting accuracy, averaged over the best-accuracy runs of all current models.

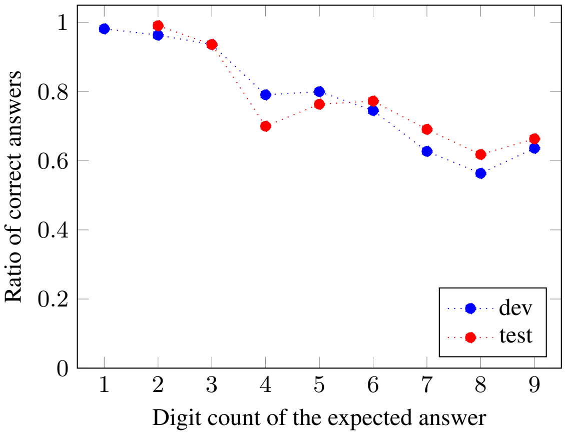

Figure 2: Numeral to number conversion accuracy, averaged over the best-accuracy runs of all current models.

Model ID		dev		test	
		N	Y	N	Y
o4-mini-2025-04-16	N	134	16	126	24
	Y	32	118	32	118
gpt-4.1-2025-04-14	N	29	121	36	114
	Y	1	149	5	145
gpt-4.1-mini-2025-04-14	N	25	125	31	119
	Y	9	141	5	145
gpt-4.1-nano-2025-04-14	N	22	127	25	125
	Y	5	145	5	145
gpt-4o-2024-08-06	N	19	131	16	134
	Y	2	148	2	148
gpt-4o-mini-2024-07-18	N	9	141	9	141
	Y	1	149	0	150
claude-3.7-sonnet	N	53	97	45	105
	Y	23	127	19	131
gemini-2.5-flash	N	28	122	35	115
	Y	1	149	2	148
gemma-3-27b-it	N	37	113	36	114
	Y	14	136	10	140
gpt-oss-120b	N	103	47	100	50
	Y	37	113	31	119
qwen3-235b-a22b-2507	N	19	131	16	134
	Y	7	143	5	145
gpt-3.5-turbo-0125	N	17	133	23	127
	Y	14	136	11	139

Table 5: Contingency tables of the best-accuracy runs in Lexicon (expected answers in rows, predicted answers in columns).

Model ID	dev	test
o4-mini-2025-04-16	0.89 / 0.24	0.72 / 0.25
gpt-4.1-2025-04-14	0.92 / 0.67	0.78 / 0.50
gpt-4.1-mini-2025-04-14	0.86 / 0.29	0.71 / 0.29
gpt-4.1-nano-2025-04-14	0.72 / 0.14	0.47 / 0.16
gpt-4o-2024-08-06	0.94 / 0.77	0.78 / 0.55
gpt-4o-mini-2024-07-18	0.78 / 0.29	0.60 / 0.27
claude-3.7-sonnet	0.94 / 0.90	0.84 / 0.75
gemini-2.5-flash	0.94 / 0.86	0.85 / 0.81
gemma-3-27b-it	0.86 / 0.54	0.59 / 0.45
gpt-oss-120b	0.68 / 0.05	0.49 / 0.14
qwen3-235b-a22b-2507	0.74 / 0.42	0.51 / 0.25
gpt-3.5-turbo-0125	0.57 / 0.16	0.32 / 0.20

Table 6: Best accuracy scores of classical ⇒ modern and modern ⇒ classical orthography conversion.

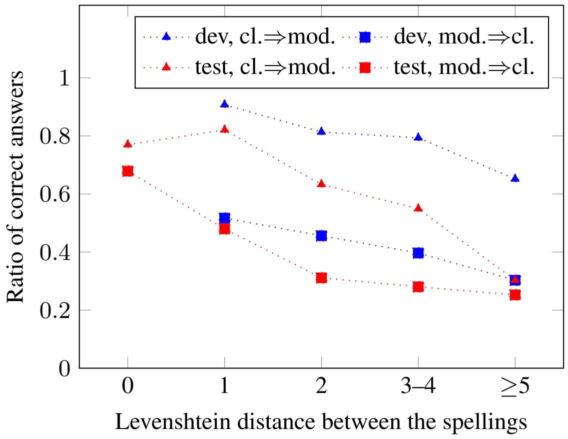

Figure 3: Orthography conversion accuracy, averaged over the best-accuracy runs of all current models.

PoS	dev		test	
	$d(C, D)$	$d(C, D')$	$d(C, D)$	$d(C, D')$
noun	3.525	2.575	3.502	2.551
adjective	4.239	3.587	4.175	3.964
verb	5.550	5.112	5.439	4.923

Table 7: Levenshtein distances between the wordform C, the expected answer D and the predicted wrong answer D' in Inflectional analogies, averaged over the best-accuracy runs of all current models.

Model ID	dev	test
o4-mini-2025-04-16	58.14	69.94
gpt-4.1-2025-04-14	27.08	29.02
gpt-4.1-mini-2025-04-14	57.89	66.88
gpt-4.1-nano-2025-04-14	36.60	43.68
gpt-4o-2024-08-06	44.17	50.90
gpt-4o-mini-2024-07-18	50.13	53.08
claude-3.7-sonnet	88.36	89.70
gemini-2.5-flash	93.72	95.11
gemma-3-27b-it	57.54	58.88
gpt-oss-120b	68.60	79.95
qwen3-235b-a22b-2507	19.07	17.67
gpt-3.5-turbo-0125	44.58	43.83

Table 8: Best BLEU scores of round-trip translation.