Evaluating Large Language Models on Lithuanian Grammatical Cases

Urtė Jakubauskaitė
University of Amsterdam
Institute for Logic, Language and Computation
urte.jakubauskaite@student.uva.nl

Raquel G. Alhama
University of Amsterdam
Institute for Logic, Language and Computation
rgalhama@uva.nl

Abstract

We present a systematic evaluation of large language models (LLMs) on Lithuanian grammatical case marking, a task that has received little prior attention. Lithuanian is a relatively low-resource language, with rich morphology and explicit marking. To enable fine-grained syntactic and morphological assessment, we introduce a novel dataset of 305 minimal sentence pairs contrasting correct and incorrect case usage. Our results show that case marking is challenging for current models, with overall accuracy ranging from 0.662 to 0.852. A monolingual Lithuanian LLM consistently outperforms multilingual counterparts, highlighting the value of language-specific training over model size. Performance varies across cases: genitive and locative forms are generally better handled, while rarer constructions and subtle functional distinctions remain difficult. The dataset and analysis provide a resource for future work, supporting the development of more robust LLMs and targeted evaluation benchmarks for morphologically rich, low-resource languages.

1 Introduction

In recent years, large language models (LLMs) have demonstrated impressive performance across a wide range of natural language processing (NLP) tasks, including machine translation, question answering, and text generation, among many others (Minaee et al., 2025). Despite these advances, the extent to which such models possess genuine linguistic competence, particularly with respect to syntactic knowledge, remains an open question.

To investigate this issue, targeted evaluation benchmarks have been developed, among which BLiMP (Warstadt et al., 2019) has become a widely used resource for assessing syntactic knowledge in English. BLiMP targets the evaluation of language models using minimal pairs: pairs of sentences that differ only in grammaticality. Each pair consists of one grammatical and one ungrammatical sentence, and a model is considered to perform correctly if it assigns a higher probability to the grammatical sentence. The benchmark tests twelve distinct linguistic phenomena, enabling fine-grained analysis of model behavior.

Originally developed for English, BLiMP has inspired similar benchmarks for other languages, including Dutch (BLiMP-NL; (Suijkerbuijk et al., 2025), and Pestel et al. (2025)), Italian (BLiMP-IT; (Barbini et al., 2025)), Irish (Irish-BLiMP; (McGiff et al., 2025)), Turkish (TurBLiMP; (Başar et al., 2025)), Japanese (JBLiMP; (Someya and Oseki, 2023)), Urdu (UrBLiMP; (Adeeba et al., 2025)), and the multilingual MultiBLiMP benchmark, which covers 101 languages (Jumelet et al., 2025). Nevertheless, many less-studied languages remain underrepresented in such evaluations.

Although MultiBLiMP includes Lithuanian, its coverage is limited to two linguistic phenomena (subject-verb agreements), both of which rely heavily on overt morphological cues. Given the rich case-marking system of Lithuanian, these tasks are highly predictable and therefore yield artificially high model performance (ranging from 0.832 to 0.985 across the models), offering limited insight into deeper syntactic understanding.

This study addresses this gap by making two main contributions. First, we introduce a newly created dataset of minimal sentence pairs targeting Lithuanian grammatical case usage, covering a range of error types. Second, we use this dataset to conduct a systematic evaluation of multiple large language models, analyzing overall performance as well as performance by grammatical case and error type.

2 Methodology

We evaluate large language models' knowledge of Lithuanian grammatical cases using a newly cre-

ated dataset of minimal sentence pairs that consists of correct and incorrect case usage. Three model families, varying in size and training setup, are evaluated.

For tasks that target grammatical knowledge, comparing the likelihoods of sentences is often employed as an evaluation method. Approaches that compare sentence-based metrics are generally more reliable than prompting for explicit grammaticality judgments, since prompting produces categorical outputs, whereas likelihood directly evaluates the model's internal probability distribution over sequences, providing a finer-grained measure that is less sensitive to the wording of the prompt. While perplexity is a commonly used metric for evaluating language models, it is not well-suited for assessing syntactic competence, as it primarily rewards models for predicting frequent collocations (Linzen et al., 2016; Marvin and Linzen, 2018; Tran et al., 2018).

For each sentence pair, we compute the negative log-likelihood (NLL) of both the grammatical and ungrammatical sentences in each model. A per-token NLL is used to normalize for sentence length. The model's grammaticality assessment of a sentence pair is based on a direct comparison of NLL values: the sentence with the lower NLL is considered more acceptable, and the model is deemed correct if this sentence is the grammatical one.

Model performance is evaluated using overall accuracy across all sentence pairs. In addition, accuracy is computed separately for each grammatical case and for each error type, allowing for fine-grained analysis of model behavior and identification of systematic strengths and weaknesses in handling Lithuanian case morphology.

3 Data

To our knowledge, prior to this study no Lithuanian dataset was available for evaluating minimal pairs involving correct and incorrect case usage. To address this gap, we construct and release a dataset consisting of 305 minimal sentence pairs, which we make available for future research[1].

The primary source of the dataset is the Valstybinė lietuvių kalbos komisija (Vlkk; in English, *The State Commission of the Lithuanian Language*) (2023), a state institution accountable to the Seimas (the Lithuanian Parliament). The commission consists of 17 members appointed for five-year terms and is responsible for regulating the Lithuanian language, including language policy, standardization, and the approval of authoritative language resources such as dictionaries and textbooks.

The official VLKK website provides a list entitled *Didžiųjų kalbos klaidų sąrašas* (in English, *List of Major Language Errors*), which documents the most common errors in Lithuanian language use. The list was compiled on the basis of language norm violations attested in major daily newspapers and other press outlets over several years (Urnėžiūtė, 2014). One section of this list focuses specifically on *Linksnių vartojimas* (in English, *Use of Cases*). The majority of the dataset, 174 sentence pairs, are taken directly from this list. Some other sentence pairs (49 in total) are collected from other Lithuanian-language websites, primarily those of city governments, which provide additional examples of incorrect case usage (see Appendix A).

In addition, part of the dataset is semi-automatically generated using a Lithuanian text corpus (Vytauto Didžiojo universitetas (VDU), 2013) (3 sentence pairs) and ChatGPT-4o (OpenAI, 2024) (79 sentence pairs). Specifically, ChatGPT is prompted with descriptions of particular case usage errors and several illustrative examples, and is asked to generate additional minimal sentence pairs exhibiting similar patterns with lexical and syntactic variation (see Appendix A). Due to a relatively high rate of erroneous or linguistically invalid outputs, all automatically generated sentence pairs are manually reviewed and corrected by a native speaker. ChatGPT is therefore used as a source of inspiration rather than as a direct data generator, ensuring linguistic correctness and consistency while minimizing author bias in data creation.

The final dataset includes errors divided across six grammatical cases: nominative (further split into two error types), genitive (two error types), dative (four error types), accusative (three error types), instrumental (five error types), and locative (six error types) (see Appendix A, Table 3).

For example, *3.1.1 Nominative cases should not be used to express an indefinite quantity of things or a part of something (instead of the partitive genitive) with certain verbs:*

[1] https://github.com/urtuteja/Evaluating-Large-Language-Models-on-Lithuanian-Grammatical-Cases

- **Correct:** Mieste atsirado valkataujančių šunų. (In English, *In the city appeared wandering dogs.*)
- **Incorrect:** Mieste atsirado valkataujantys šunys. (In English, *In the city appeared wandering dogs.*)

One error type within the nominative group also involves the vocative case, where the nominative is incorrectly used instead of the vocative. As a result, the dataset indirectly covers all seven grammatical cases present in the Lithuanian language. Each error type includes at least 10 examples, with some error types containing up to 55 examples. Full dataset statistics are provided in Appendix A, Figure 1.

Importantly, 11 sentence pairs in the dataset include exceptions - cases in which the observed case usage does not follow the standard grammatical rule, yet is still correct.

4 Models

For this study, we select three LLM families: Neurotechnology LLM (Nakvosas et al., 2024), EuroLLM (Martins et al., 2024), and Qwen3 (Yang et al., 2025). Neurotechnology LLM is currently the only available monolingual Lithuanian LLM and is built on the Llama 2 architecture. EuroLLM supports 24 European languages, including Lithuanian. The Qwen3 model family covers 119 languages and dialects worldwide, also including Lithuanian.

As Neurotechnology LLM is the only monolingual Lithuanian LLM currently available, we include both of its released variants, with 7 and 13 billion parameters each. For EuroLLM, we evaluate all available versions, namely the 1.7, 9, and 22 billion–parameter models.

For Qwen3, we select all available dense model variants, specifically those with 0.6, 1.7, 4, 8, and 14 billion parameters. In addition, we include a smaller Mixture-of-Experts (MoE) variant, consisting of 30 billion parameters with 3 billion parameters activated per token. The larger MoE version of Qwen3 is excluded from this study due to its higher computational requirements, which exceed the resources available for our experimental setup.

Importantly, for each model we evaluate both the *base* and *instruct* variants. This allows us to compare pre-trained models with fine-tuned models that have undergone additional instruction tuning, and to analyze the effects of language coverage and model size.

5 Results

Model	Version	Acc.	Nom	Gen	Dat	Acc	Ins	Loc
Neurotechnology 7B	Base	0.852	0.786	0.846	0.911	0.833	0.758	0.909
	Instruct	0.852	0.786	0.885	0.889	0.833	0.723	0.900
Neurotechnology 13B	Base	0.826	0.643	0.846	0.822	0.867	0.758	0.900
	Instruct	0.800	0.821	0.923	0.8	0.733	0.697	0.845
EuroLLM 1.7B	Base	0.770	0.429	0.846	0.867	0.6	0.727	0.823
	Instruct	0.770	0.323	0.769	0.889	0.6	0.712	0.9
EuroLLM 9B	Base	0.810	0.607	0.885	0.844	0.733	0.788	0.864
	Instruct	0.790	0.5	0.846	0.778	0.833	0.712	0.891
EuroLLM 22B	Base	0.800	0.607	0.846	0.867	0.633	0.803	0.855
	Instruct	0.767	0.571	0.846	0.867	0.7	0.712	0.809
Qwen3 0.6B	Base	0.682	0.464	0.769	0.644	0.333	0.667	0.836
	Instruct	0.679	0.393	0.577	0.467	0.467	0.742	0.882
Qwen3 1.7B	Base	0.672	0.393	0.846	0.733	0.333	0.621	0.8
	Instruct	0.685	0.5	0.808	0.733	0.4	0.742	0.727
Qwen3 4B	Base	0.682	0.464	0.654	0.578	0.367	0.697	0.864
	Instruct	0.702	0.429	0.770	0.644	0.467	0.773	0.8
Qwen3 8B	Base	0.728	0.464	0.846	0.778	0.633	0.652	0.818
	Instruct	0.698	0.464	0.769	0.733	0.5	0.652	0.809
Qwen3 14B	Base	0.711	0.357	0.846	0.667	0.533	0.682	0.855
	Instruct	0.718	0.5	0.846	0.711	0.367	0.773	0.809
Qwen3 30B	Base	0.666	0.357	0.808	0.756	0.633	0.591	0.727
	Instruct	0.662	0.464	0.692	0.733	0.6	0.652	0.7

Table 1: Comparison of overall accuracy and accuracy across grammatical cases. Green highlights indicate the best performance among the models for a given grammatical case, while red highlights indicate the worst performance.

Tables 1 and 2 present overall accuracy as well as accuracy broken down by grammatical case and error type. Across models, overall accuracy ranges from 0.662 to 0.852, indicating that Lithuanian case marking remains a non-trivial challenge for large language models.

Model family has a stronger impact on performance than model size. Neurotechnology models, which are explicitly fine-tuned for Lithuanian, consistently achieve the highest overall accuracy, with the 7B model reaching 0.852 and outperforming several substantially larger models. In contrast, Qwen3 models exhibit lower performance, and neither increasing the parameter count nor introducing a Mixture-of-Experts (MoE) architecture leads to consistent improvements. This suggests that architectural choices and, crucially, training data are more important than scale alone. Notably, within the EuroLLM and Qwen3 families, models in the mid-range of approximately 8B–13B parameters tend to outperform both smaller and larger variants.

Instruction tuning has mixed effects. While instruct variants occasionally improve performance for specific grammatical cases or error types, they often fail to yield consistent gains and in some instances lead to decreased accuracy. This pattern

Model	Version	Accuracy	Nominative		Genitive		Dative				Accusative			Instrumental					Locative					
			Err1	Err2	Err1	Err2	Err1	Err2	Err3	Err4	Err1	Err2	Err3	Err1	Err2	Err3	Err4	Err5	Err1	Err2	Err3	Err4	Err5	Err6
Neurotechnology 7B	Base	0.852	0.824	0.727	0.8	0.875	1	0.9	0.727	1	0.7	1	0.8	0.818	0.727	0.7	1	0.667	1	0.8	0.909	1	0.923	0.8
	Instruct	0.852	0.824	0.727	0.9	0.875	1	0.9	0.727	0.929	0.7	1	0.8	0.818	0.636	0.7	1	0.75	1	0.8	0.927	1	0.769	0.8
Neurotechnology 13B	Base	0.826	0.706	0.545	0.9	0.813	1	0.8	0.727	0.786	0.7	1	0.9	0.818	0.636	0.6	1	0.75	1	0.7	0.891	1	0.923	0.9
	Instruct	0.8	0.824	0.818	1	0.875	1	0.8	0.636	0.786	0.3	1	0.9	0.636	0.636	0.5	1	0.708	0.8	0.6	0.873	1	0.769	0.9
EuroLLm 1.7B	Base	0.77	0.588	0.182	0.8	0.875	0.9	0.9	0.727	0.929	0.6	0.5	0.7	0.727	0.636	0.7	1	0.667	0.8	0.8	0.836	1	0.923	1
	Instruct	0.77	0.529	0.182	0.7	0.813	1	0.9	0.727	0.929	0.6	0.5	0.7	0.545	0.636	0.8	1	0.667	0.9	0.9	0.855	1	0.923	1
EuroLLM 9B	Base	0.81	0.765	0.364	0.8	0.938	1	0.9	0.727	0.786	0.6	0.9	0.7	0.909	0.455	0.9	1	0.75	0.8	0.9	0.818	1	0.923	0.9
	Instruct	0.79	0.706	0.182	0.8	0.875	0.9	0.8	0.545	0.857	0.8	0.9	0.8	0.727	0.364	0.8	1	0.708	0.6	1	0.891	1	0.923	0.9
EuroLLM 22B	Base	0.8	0.882	0.182	0.8	0.875	1	0.7	0.727	1	0.5	0.8	0.6	0.909	0.545	0.9	1	0.75	1	0.7	0.836	1	0.846	0.8
	Instruct	0.767	0.882	0.091	0.9	0.813	0.9	0.9	0.636	1	0.6	0.7	0.8	0.818	0.545	0.6	1	0.667	0.8	0.9	0.782	1	0.769	0.7
Qwen3 0.6B	Base	0.682	0.647	0.182	0.9	0.688	0.9	0.6	0.364	0.714	0.4	0.4	0.2	0.909	0.727	0.4	1	0.5	0.9	0.7	0.8	1	0.846	0.9
	Instruct	0.679	0.529	0.182	0.9	0.375	0.6	0.4	0.182	0.643	0.5	0.5	0.4	0.909	0.727	0.8	0.9	0.583	0.9	0.7	0.855	1	1	0.9
Qwen3 1.7B	Base	0.672	0.529	0.182	0.8	0.875	0.7	1	0.455	0.786	0.5	0.3	0.2	0.636	0.818	0.1	1	0.583	0.7	0.8	0.782	0.833	0.846	0.9
	Instruct	0.685	0.647	0.273	0.8	0.813	0.9	0.9	0.364	0.786	0.5	0.1	0.6	0.727	1	0.5	1	0.625	0.7	0.5	0.727	0.917	0.615	0.9
Qwen3 4B	Base	0.682	0.588	0.273	0.6	0.688	0.7	0.6	0.364	0.643	0.3	0.2	0.6	0.818	0.818	0.3	1	0.625	0.8	0.5	0.909	1	0.846	0.9
	Instruct	0.702	0.471	0.364	0.8	0.75	0.7	0.7	0.273	0.857	0.5	0.4	0.5	0.909	0.909	0.6	0.9	0.667	0.8	0.5	0.782	1	0.846	0.9
Qwen3 8B	Base	0.728	0.529	0.364	0.7	0.938	0.9	1	0.455	0.786	0.6	0.7	0.6	0.909	0.727	0.2	1	0.542	0.8	0.5	0.836	1	0.769	0.9
	Instruct	0.698	0.588	0.273	0.7	0.813	1	0.9	0.364	0.714	0.5	0.6	0.4	0.727	0.818	0.2	1	0.583	0.7	0.7	0.818	0.917	0.769	0.9
Qwen3 14B	Base	0.711	0.529	0.091	0.7	0.938	0.9	0.9	0.273	0.643	0.4	0.7	0.5	0.818	0.727	0.6	0.9	0.542	0.8	0.8	0.836	1	0.846	0.9
	Instruct	0.718	0.647	0.273	0.6	1	0.9	0.8	0.364	0.786	0.4	0.5	0.2	0.909	0.909	0.6	1	0.625	0.8	0.7	0.782	1	0.769	0.9
Qwen3 30B	Base	0.666	0.412	0.273	0.7	0.875	1	0.9	0.364	0.786	0.6	0.5	0.8	0.364	0.636	0.4	1	0.583	0.7	0.9	0.673	0.833	0.923	0.5
	Instruct	0.662	0.588	0.273	0.6	0.75	0.9	1	0.273	0.786	0.6	0.5	0.7	0.545	0.818	0.2	1	0.667	0.7	1	0.636	0.75	0.846	0.5

Table 2: Comparison of overall accuracy and accuracy across error types. Green highlights indicate the best performance among the models for a given error type, while red highlights indicate the worst performance.

suggests that instruction tuning may interfere with the probability-based acceptability judgments required for minimal-pair evaluation.

6 Error Analysis

Performance varies considerably across grammatical cases. The genitive and locative cases are generally handled well, whereas the nominative and accusative cases show lower accuracy across most models. This pattern likely reflects differences in morphological transparency and syntactic ambiguity. For example, the nominative case sometimes shares forms with the vocative depending on a word's number and gender: *Ieva* (female name) and *šunys* (in English, *dogs*) do not change when used as a vocative, which makes nominative less morphologically transparent and can create ambiguity. While such cases were avoided in this study, the models could have been exposed to them during training. Similarly, the accusative case is challenging because whether it is required depends on syntactic context. In the sentence *Mačiau katę* (in English, *I saw a cat*), the accusative is used, but after negation (*Nemačiau katės* – in English, *I did not see a cat*), the genitive is required, as also illustrated by Error 2 in accusative. These factors likely contribute to why some cases are easier for LLMs to handle than others.

An analysis by error type reveals substantial variability within each grammatical case. Some error types achieve near-ceiling performance, while others approach chance level even for otherwise high-performing models. For example, Error 2 yielded the worst performance in the nominative case category, specifically *3.1.2 Nominative case should not be used to express direct address*. Only the Neurotechnology models achieved moderate to good results (0.545–0.727), while all other models and their versions scored between 0.091 and 0.364. A likely reason for the weaker performance of models not fine-tuned for Lithuanian is that the vocative case, which should be used instead, is relatively rare and, in many other languages, coincides with the nominative form (Ambrazas, 2026). Furthermore, it differs from other declensions in that it does not indicate a direct syntactic or semantic connection to other word forms (Ambrazas, 2026). Consequently, multilingual models may overlook such infrequent patterns, resulting in low performance on this error type.

Another particularly challenging error was Error 3 in the dative case, *3.3.3 Dative case should not be used for indicating a specific time limit or moment when purpose is not being expressed*. The Qwen3 model family performed poorly on this error, with accuracy ranging from 0.182 to 0.455. One possible explanation is overgeneralization, as the rule is highly specific. In fact, in a closely related context, the dative is acceptable: *The dative is used to indicate the duration and purpose of an action*. Subtle distinctions like this may confuse models, which often struggle with fine-grained, rule-based restrictions that require reasoning about the function of a case beyond surface form.

7 Conclusions

This work evaluates several LLMs on Lithuanian grammatical case marking. The results show an

average accuracy of 0.742 (range: 0.662–0.852), substantially lower than the 0.939 average accuracy (range: 0.832–0.985) previously reported for Lithuanian in MultiBLiMP (Jumelet et al., 2025). This gap suggests that the newly created dataset captures more challenging Lithuanian phenomena and is better suited for evaluating a morphologically rich language.

Our analysis highlights persistent difficulties for current LLMs, particularly with rare constructions such as the vocative case and nuanced functional contrasts. These findings indicate that multilingual pretraining alone is often insufficient to capture language-specific grammatical phenomena, and that larger model size does not necessarily lead to better performance.

For future work, we are currently expanding the dataset to include additional linguistic phenomena, such as prepositions, grammatical forms, and sentence coordination, evaluate a broader range of LLM families, and incorporate graded human judgments from native Lithuanian speakers to ensure alignment with human intuitions about grammaticality.

8 Limitations

This study has several limitations. First, the dataset is relatively small, and some error types are represented by as few as ten examples, which limits the generalizability of the results. Second, expanding the original sentence pairs without compromising dataset quality proved challenging. Large language models such as ChatGPT-4o, used in this study to generate additional examples, do not reliably produce linguistically valid sentences for all error types. As a result, all automatically generated sentences had to be manually reviewed and corrected to ensure both linguistic correctness and variability. This process is time-consuming and introduces the potential for human biases in dataset creation. Finally, the study does not evaluate all LLMs that support Lithuanian. Consequently, the findings are limited to the specific models included in this analysis.

References

Farah Adeeba, Brian Dillon, Hassan Sajjad, and Rajesh Bhatt. 2025. UrBLiMP: A Benchmark for Evaluating the Linguistic Competence of Large Language Models in Urdu. arXiv:2508.01006.

Vytautas Ambrazas. 2026. Šauksmininkas. Visuotinė lietuvių enciklopedija.

Matilde Barbini, Maria Letizia Piccini Bianchessi, Veronica Bressan, Achille Fusco, Sofia Neri, Sarah Rossi, Tommaso Sgrizzi, and Cristiano Chesi. 2025. BLiMP-IT: Harnessing Automatic Minimal Pair Generation for Italian Language Model Evaluation. In *Proceedings of the Eleventh Italian Conference on Computational Linguistics (CLiC-it 2025)*, pages 64–71, Cagliari, Italy. CEUR Workshop Proceedings.

Ezgi Başar, Francesca Padovani, Jaap Jumelet, and Arianna Bisazza. 2025. TurBLiMP: A Turkish Benchmark of Linguistic Minimal Pairs. arXiv:2506.13487.

Jaap Jumelet, Leonie Weissweiler, Joakim Nivre, and Arianna Bisazza. 2025. MultiBLiMP 1.0: A Massively Multilingual Benchmark of Linguistic Minimal Pairs. arXiv:2504.02768.

Tal Linzen, Emmanuel Dupoux, and Yoav Goldberg. 2016. Assessing the Ability of LSTMs to Learn Syntax-Sensitive Dependencies. *Transactions of the Association for Computational Linguistics*, 4:521–535.

Pedro Henrique Martins, Patrick Fernandes, João Alves, Nuno M. Guerreiro, Ricardo Rei, Duarte M. Alves, José Pombal, Amin Farajian, Manuel Faysse, Mateusz Klimaszewski, Pierre Colombo, Barry Haddow, José G. C. de Souza, Alexandra Birch, and André F. T. Martins. 2024. EuroLLM: Multilingual Language Models for Europe. arXiv:2408.12963.

Rebecca Marvin and Tal Linzen. 2018. Targeted Syntactic Evaluation of Language Models. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1192–1202, Brussels, Belgium. Association for Computational Linguistics.

Josh McGiff, Khanh-Tung Tran, William Mulcahy, Dáibhidh Ó Luinín, Jake Dalzell, Róisín Ní Bhroin, Adam Burke, Barry O'Sullivan, Hoang D. Nguyen, and Nikola S. Nikolov. 2025. Irish-BLiMP: A Linguistic Benchmark for Evaluating Human and Language Model Performance in a Low-Resource Setting. arXiv:2510.20957.

Shervin Minaee, Tomas Mikolov, Narjes Nikzad, Meysam Chenaghlu, Richard Socher, Xavier Amatriain, and Jianfeng Gao. 2025. Large Language Models: A Survey. arXiv:2402.06196.

Artūras Nakvosas, Povilas Daniušis, and Vytas Mulevičius. 2024. Open Llama2 Model for the Lithuanian Language. arXiv:2408.12963.

OpenAI. 2024. Hello GPT-4o.

Julia Pestel, Jelke Bloem, and Raquel G. Alhama. 2025. Evaluating Dutch Speakers and Large Language Models on Standard Dutch: a grammatical Challenge Set based on the Algemene Nederlandse Spraakkunst. *Computational Linguistics in the Netherlands Journal*, 14:555–582.

Taiga Someya and Yohei Oseki. 2023. JBLiMP: Japanese Benchmark of Linguistic Minimal Pairs. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 1581–1594, Dubrovnik, Croatia. Association for Computational Linguistics.

Michelle Suijkerbuijk, Zoë Prins, Marianne de Heer Kloots, Willem Zuidema, and Stefan L. Frank. 2025. BLiMP-NL: A Corpus of Dutch Minimal Pairs and Acceptability Judgments for Language Model Evaluation. *Computational Linguistics*.

Ke Tran, Arianna Bisazza, and Christof Monz. 2018. The Importance of Being Recurrent for Modeling Hierarchical Structure. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4731–4736, Brussels, Belgium. Association for Computational Linguistics.

Rita Urnėžiūtė. 2014. Taisyklingos kalbos niekas neatšaukė. *Gimtoji kalba*, (2):18–26.

Valstybinė lietuvių kalbos komisija (Vlkk). 2023. Commission.

Vytauto Didžiojo universitetas (VDU). 2013. Corpus of Contemporary Lithuanian Language.

Alex Warstadt, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang, and Samuel R. Bowman. 2019. BLiMP: A Benchmark of Linguistic Minimal Pairs for English. arXiv:1912.00582.

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. arXiv:2505.09388.

A Data

The following links were used as sources for additional sentence pairs in the dataset:

- https://alytus.lt/uploads/documents/files/LT/savivaldybes-administracija/administracine-informacija/kalbos%20taisyklingumas/AdministracinÄŪs%20kalbos%20atmintinÄŪ%202019%2002%2018.pdf
- https://www.kurstoti.lt/s/13594/dazniausios-sintakses-sakiniu-sandaros-klaidos-abiturientu-vbe-rasiniuose
- https://www.nsa.smsm.lt/wp-content/uploads/2025/07/VERTINIMO-GAIRES-2022-koreguotos.pdf
- https://vlkk.lt/konsultacijos/11178-ne-mazesnis-nemaziau-ne-maziau-nemaziau
- https://vlkk.lt/konsultacijos/1138-ne-ilgiau-ko
- https://vlkk.lt/konsultacijos/978-galininkas-kilmininkas
- https://salcininkai.lt/valstybine-kalba/linksniu-vartojimo-klaidos/1004

The following prompt was used for sentence pair generation:

> I will provide examples of sentence pairs consisting of a grammatically correct sentence and a corresponding incorrect sentence. I will also specify a particular grammatical error illustrated by these pairs. Please generate 10 additional sentence pairs that contain the same error type. The generated pairs should display both lexical and syntactic diversity and must not reuse vocabulary from the original examples.
>
> *[Sentence pairs and error description inserted here.]*

Error Code	Error Description
3.1.1	Nominative cases should not be used to express an indefinite quantity of things or a part of something (instead of the partitive genitive) with certain verbs.
3.1.2	Nominative case should not be used to express direct address.
3.2.1	Genitive case should not be used to denote the object of an action with certain verbs.
3.2.2	Genitive case should not be used to express comparative quantity with the comparative adverbs (ne) daugiau, (ne) mažiau, (ne) ilgiau, (ne) vėliau, (ne) anksčiau.
3.3.1	Dative case should not be used to denote the object of an action with certain expressions.
3.3.2	Dative case should not be used to describe a thing/object when purpose is not being expressed.
3.3.3	Dative case should not be used for indicating a specific time limit or moment when purpose is not being expressed.
3.3.4	Dative should not be used with verbs of motion to express purpose.
3.4.1	Accusative case should not be used to express an indefinite quantity of things or a part of something (instead of the partitive genitive).
3.4.2	Accusative case should not be used to express the direct object next to a negative verb (instead of the genitive).
3.4.3	Accusative case should not be used with verbs of motion to express purpose or aim when the accusative cannot stand alone without the infinitive.
3.5.1	Instrumental case should not be used to express the object with verbs denoting fullness or increase.
3.5.2	Instrumental case should not be used to express content of quality with adjectives denoting abundance.
3.5.3	Instrumental case should not be used to express the agent or cause of a state (but not the instrument) with passive participles.
3.5.4	Instrumental case should not be used with forms of the verb būti to express a permanent (unchanging) state.
3.5.5	The instrumental of adjectives (and words used adjectivally) should not be used to express a state.
3.6.1	Locative case should not be used to express the experiencer of a state.
3.6.2	Locative case should not be used to express the domain of an action, state, or quality (but not a place).
3.6.3	Locative case should not be used to express the state, condition, or characteristic of a thing.
3.6.4	Locative case should not be used to express the cause or basis of a state.
3.6.5	Locative case should not be used to express the manner or timing of an action.
3.6.6	Locative case should not be used to express a time period as a preposition or postposition.

Table 3: List of Lithuanian grammatical case errors.

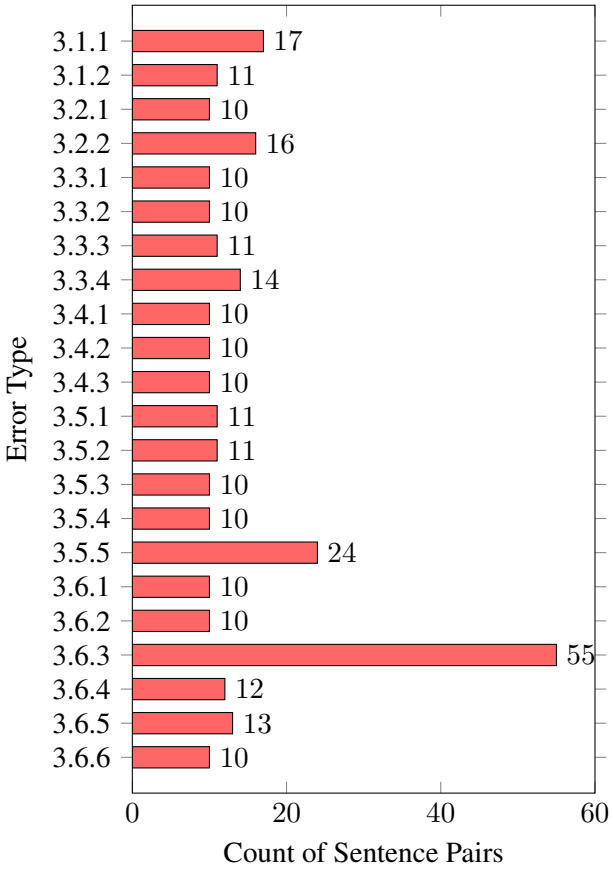

Figure 1: Number of sentence pairs per Lithuanian grammatical case error type.