# Evaluating The Syntactic Knowledge of Language Models on Lithuanian

**MSc Thesis** *(Afstudeerscriptie)*

written by

**Urtė Jakubauskaitė**

under the supervision of **Dr. Ing. Raquel Garrido Alhama**, and submitted to the Examinations Board in partial fulfillment of the requirements for the degree of

**MSc in Logic**

at the *Universiteit van Amsterdam.*

| **Date of the public defense:** | **Members of the Thesis Committee:** |
|---|---|
| *26$^{th}$ June, 2026* | Dr. Balder ten Cate (chair) |
| | Dr. Ing. Raquel Garrido Alhama (supervisor) |
| | Dr. Jelke Bloem |
| | Prof. Dr. Jeannette Schaeffer |

Institute for Logic, Language and Computation

**Abstract**

Despite substantial recent advances in language models (LMs), their syntactic competence remains largely unexplored in low-resource languages. Lithuanian presents a particularly interesting test case due to its rich morphology and the scarcity of evaluation resources. The only existing benchmark covers a limited range of syntactic phenomena, making it difficult to assess how well language models capture the complexities of Lithuanian grammar.

To address this gap, this thesis introduces a new minimal-pair benchmark for Lithuanian syntax. The dataset consists of grammatical and ungrammatical sentence pairs targeting 31 linguistic phenomena and 64 error types derived from attested language use. The benchmark is used to evaluate 78 monolingual and multilingual language models. In addition, human acceptability judgments are collected to enable comparisons between model predictions and native speaker intuitions.

The results reveal large variation in performance across models and syntactic constructions. Monolingual Lithuanian models achieve the strongest results, and performance generally improves with model size. Moreover, the proposed benchmark proves considerably more challenging than the existing multilingual benchmark including Lithuanian, exposing weaknesses that remained undetected in previous research. Overall, this work extends the available evaluation resources for Lithuanian and contributes to a more comprehensive assessment of syntactic competence in low-resource, morphologically rich languages.

**Keywords**: language models, syntactic competence, low-resource languages, Lithuanian, morphology, minimal-pair evaluation.

# Acknowledgements

To me, this thesis represents more than the final step of my Master's degree. Earlier in life, I was used to travelling from A to Z along relatively straightforward paths. The journey toward completing this degree, however, took a few zigzags that taught me to become highly adaptable. Along the way, I completed another Master's degree, presented my work at both the national and international levels, and, most importantly, discovered where my true passions and strengths lie. Looking back, the journey was challenging at times but ultimately truly rewarding.

Of course, none of this would have been the same without the people who accompanied me along the way. Given the linguistic focus of this thesis, it feels fitting to borrow an expression from Lithuanian to express my gratitude. In my native language, exceptional respect is sometimes conveyed by describing someone as being *with a capital letter* (in Lithuanian, *iš didžiosios raidės*), meaning they truly embody the qualities associated with their role.

That is how I think of my supervisor, Raquel G. Alhama. To me, she is an Educator with a capital *E*. Her expertise, kindness, and sincere enthusiasm for my research made this project enjoyable and helped me overcome the self-doubt I often struggled with.

I would also like to thank Paul Dekker, another remarkable Educator, for his genuine care and dedication to students.

My gratitude also goes to Tanja Kassenaar, the Study Advisor for the Master of Logic programme, who helped me navigate my academic path and encouraged me to reflect on difficult questions, as well as to Dick de Jongh, my Mentor, for his constant support.

I am grateful to my fellow student and friend, Elynn Weijland, whose friendship made this journey much easier and more fun.

I want to thank my parents, my sister, and my partner: Remigijus, Daiva, Agnė, and Adel. No matter what I do, they are always rooting for me so strongly that I sometimes cannot believe how lucky I am to be part of this family.

Finally, although I cannot mention everyone by name, I would like to thank all the staff members of the ILLC and the fellow students I had the pleasure of meeting and working with. From my very first day, I was captivated by the enthusiasm for learning within this community. I do not think I have ever met so many people in one place who genuinely love what they do, and their passion continues to inspire me.

# Contents

# 1 Introduction

In this chapter, I present the study by outlining its motivation and aim, formulating the research questions and hypotheses, providing an overview of the approach, and discussing the contributions to the field.

## 1.1 Motivation

Language models (LMs) continue to demonstrate increasingly impressive capabilities, including in the domain of linguistic competence (Braga et al., 2025). To better understand the extent of these abilities, researchers have developed targeted evaluation benchmarks. One such benchmark, BLiMP (Warstadt, Parrish, et al., 2019), assesses models' syntactic knowledge using minimally different sentence pairs that contrast in grammaticality. However, this and most other widely used benchmarks focus on major world languages, particularly English, which remains predominant in global research (Rao et al., 2025; X. Zhang et al., 2023). As a result, English benefits from both extensive training data and abundant evaluation resources.

In contrast, low-resource and less-studied languages face a scarcity of both data and evaluation benchmarks. Lithuanian is one such language. With regard to syntactic evaluation, only one study to date has addressed Lithuanian, namely MultiBLiMP (Jumelet et al., 2025). While MultiBLiMP provides a multilingual syntactic benchmark for numerous underrepresented languages, it covers only a limited set of linguistic phenomena, specifically two types of subject–verb agreement. In Lithuanian, these phenomena are marked by strong morphological cues, which may allow models to achieve high performance without demonstrating a deeper understanding of syntactic structure.

This limitation highlights the need for a more comprehensive and linguistically informed evaluation of language models on Lithuanian syntax.

## 1.2 Research Aim, Questions and Hypotheses

### 1.2.1 Aim

The present study has two main aims: (i) to develop a dataset suitable for evaluating language models on Lithuanian, and (ii) to assess the extent to which LMs capture the syntactic properties of Lithuanian using this novel dataset. Specifically, the study investigates whether LMs can reliably distinguish between grammatical and ungrammatical sentences across a range of syntactic phenomena.

### 1.2.2 Research Questions

The following research questions (RQs) guide this study:

#### Confirmatory RQs

1. To what extent do large language models correctly distinguish between grammatical and ungrammatical sentences in Lithuanian?
2. How does the performance of globally oriented multilingual models compare to that of European-focused multilingual models? How do both compare to monolingual (Lithuanian) models?
3. How does model size affect performance?
4. How do different model versions (base vs. instruct) compare in performance?

#### Exploratory RQs

5. To what extent do model judgments align with native speaker acceptability judgments?
6. Which syntactic phenomena and error types pose the greatest challenges for these models?

### 1.2.3 Hypotheses

The following hypotheses are formulated for the confirmatory research questions, whereas the exploratory questions are left open. The numbering corresponds to the research questions presented in Section 1.2.2.

1. Although accuracy for Lithuanian in MultiBLiMP (Jumelet et al., 2025) is relatively high (ranging from 0.832 to 0.985 across models), it is expected that the dataset developed in this study, which is specifically designed to target syntactic phenomena that are challenging even for native speakers, will yield a more realistic evaluation, resulting in lower model accuracy.
2. Following previous research on Dutch LMs (Vlantis & Bloem, 2025), language-specific models tend to outperform multilingual models on various linguistic tasks. Therefore, it is expected that monolingual Lithuanian models will outperform European-focused multilingual models, which in turn will outperform globally oriented multilingual models.
3. The influence of model size tends to depend on the task. While larger models often outperform smaller ones on complex tasks (Luo et al., 2025), small language models (SLMs) can sometimes achieve competitive or even superior performance on simpler tasks due to their efficiency (Subramanian et al., 2025). Therefore, for the relatively constrained task of grammaticality judgment on short Lithuanian sentences, it is expected that larger and smaller models will perform similarly.
4. Finally, prior studies suggest that the relative performance of base versus instruct models also depends on the task. For example, in domain-specific tasks such as mathematics, instruct models do not always outperform base models (Munjal et al., 2026). Moreover, the authors claim that instruct models tend to outperform base versions in few-shot settings. Given that the current task is also domain-specific and does not require interactive capabilities, it is expected that base models will either outperform instruct models or yield similar results.

## 1.3 Contributions

This thesis contributes to the fields of natural language processing (NLP) and Lithuanian language technology in several ways:

1. The development of a novel syntactic evaluation dataset specifically designed for Lithuanian, grounded in linguistic evidence and native speaker error patterns. Unlike benchmarks based on automatically extracted data, such as MultiBLiMP (Jumelet et al., 2025), this dataset is manually constructed and carefully curated, thereby avoiding errors that may arise from automatic extraction. Moreover, in comparison to MultiBLiMP, which focuses on two types of agreement phenomena, this dataset includes 31 linguistic phenomena further divided into 64 error types, thus providing a more comprehensive and challenging evaluation setting.

2. A systematic quantitative evaluation of a wide range of language models (LMs) on Lithuanian syntax, including comparisons between multilingual and monolingual models, as well as across different model sizes and versions.

3. A qualitative linguistic analysis of the resulting error patterns, identifying the syntactic phenomena that are handled robustly by current models and those that continue to pose difficulties.

4. The design and collection of human acceptability judgments for Lithuanian minimal sentence pairs, alongside a comparison with model predictions. This contributes to the broader discussion on whether large language models exhibit human-like linguistic competence.

# 2 Background and Literature Review

This chapter presents the theoretical background and related work relevant to this study. It begins with an introduction to language models (LMs), followed by an overview of their key architectures and characteristics. In particular, it discusses distinctions between base, instruction-tuned, and reasoning models, as well as between dense and sparse architectures. It also covers additional aspects such as model size, context length, and language coverage.

The chapter then examines the use of LMs in linguistic evaluation tasks, providing an overview of grammaticality judgments, previous research in this area, and the challenges associated with evaluating low-resource and morphologically rich languages. Next, it reviews prior work on the evaluation of Lithuanian language processing and linguistic competence. Finally, the chapter discusses human acceptability judgments and their importance as a benchmark for assessing language model performance on linguistic tasks.

## 2.1 Language Models

Language models (LMs) are computational systems trained on large amounts of text data. By being exposed to extensive linguistic input, they learn syntactic and semantic relationships that enable them to perform various of natural language processing (NLP) tasks, such as sequence labeling, named entity recognition, question answering, text summarization, and machine translation (Jurafsky & Martin, 2026).

Modern language models are typically based on Transformer architectures (Vaswani et al., 2017), which represent a major milestone in artificial intelligence (Ghaseminejad Raeini, 2025). One of the most important properties of these architectures is the self-attention mechanism, which allows the model to learn the relationships between all tokens in an input sequence. In this way, the model can dynamically weigh the importance of different tokens when computing contextual representations (Jurafsky & Martin, 2026).

Transformer models are commonly categorized into three main architectural families: encoder-only, decoder-only, and encoder–decoder models. These architectures differ in how they process text. Encoder-only models, such as BERT (Devlin et al., 2018), are trained using the masked language modeling (MLM) objective, where selected tokens in an input sequence are masked and the model learns to predict them using both left and right context. This bidirectional modeling approach is particularly effective for token-level NLP tasks, such as sequence labeling and named entity recognition (Jurafsky & Martin, 2026).

However, encoder-only models are less suitable for open-ended text generation tasks, such as text generation, question answering, and text summarization, where the model must produce new sequences rather than predict missing tokens within a fixed input. This limitation is addressed by decoder-only

models. Decoder-only models, such as GPT (Radford et al., 2018), are trained using the causal language modeling (CLM) objective, where each token is predicted based only on preceding tokens in a left-to-right manner (Jurafsky & Martin, 2026).

Finally, encoder–decoder models combine two Transformer components: an encoder, which processes the input sequence and produces contextual representations, and a decoder, which generates an output sequence conditioned on these representations. Unlike encoder-only and decoder-only models, encoder–decoder architectures are designed for tasks in which the input and output sequences may differ substantially in both content and length (Jurafsky & Martin, 2026). Consequently, encoder–decoder models, such as T5 (Raffel et al., 2019), are particularly well suited for sequence-to-sequence tasks, such as machine translation.

However, this study focuses on decoder-only models, as they have become the dominant architecture in contemporary language modeling following the success of the GPT family of models (Radford et al., 2018). Their simpler training procedure compared to encoder–decoder architectures has further contributed to their widespread adoption (B. Zhang et al., 2025).

## 2.2 Language Model Characteristics

Language models can be characterized by several key attributes, including their version, architecture, size, context length, and language support. These characteristics are discussed in the following subsections.

### 2.2.1 Base, Instruct, and Reasoning Models

The first step in developing a language model is a process known as pre-training, during which the model is trained on large amounts of data (Jurafsky & Martin, 2026). The resulting models are referred to as base models. While these models acquire general knowledge from the training data, they are not specifically optimized to follow user instructions or preferences.

To better align language models with human expectations, additional post-training techniques are applied. Two common forms of alignment are instruction tuning and preference alignment (Jurafsky & Martin, 2026). In instruction tuning, a base model is fine-tuned on a dataset consisting of instruction-response pairs, enabling it to follow user instructions and perform a wide range of tasks more effectively. This process is typically carried out using Supervised Fine-Tuning (SFT) and results in an instruct model (Jurafsky & Martin, 2026).

Preference alignment aims to make model outputs more consistent with human preferences. In this approach, common methods include Reinforcement Learning from Human Feedback (RLHF) and Direct Preference Optimization (DPO) (Jurafsky & Martin, 2026).

A third category of language models is reasoning models. These models are typically created by applying SFT followed by additional Reinforcement Learning (RL) aimed at improving reasoning capabilities. For example, this training pipeline was used in the development of `DeepSeek-R1` (DeepSeek-AI, 2025). Reasoning models are designed to perform complex multi-step reasoning and generally achieve stronger performance on tasks involving mathematics, logic, coding, and problem solving.

However, this study employs only base and instruct models. The task considered in this work is straightforward and does not require the advanced reasoning and problem-solving capabilities offered by reasoning models.

### 2.2.2 Dense vs. Sparse (Mixture-of-Experts) Models

Language models also differ in their underlying architectures. Two major architectural categories are dense models and sparse models, more formally referred to as Mixture-of-Experts (MoE) models (Mu & Lin, 2025). In dense models, all model parameters are involved in processing every input token. In contrast, MoE models activate only a subset of their parameters for each token, thereby reducing the amount of computation required during inference.

MoE models consist of multiple specialized expert networks and a routing mechanism that determines which experts should process a given token. This architecture enables the model to increase its total number of parameters and specialize in different domains while keeping the computational cost per token relatively low.

This study includes both dense and sparse models.

### 2.2.3 Model Size

Another important characteristic of language models is the number of parameters. This quantity is typically expressed in millions or billions, for example as 125M or 14B (Wu & Tang, 2024). As presented in Section 2.2.2, dense models use all of their parameters when processing an input, whereas in sparse models only a portion of parameters is activated for each input.

For MoE models, it is therefore common to distinguish between the total number of parameters and the number of active parameters used during inference (Wu & Tang, 2024). For example, `LLaMA 4 Scout` from the `LLaMA 4` model family (Meta AI, 2025) has a total of 109B parameters, while approximately 17B parameters are activated per token.

### 2.2.4 Context Length

When discussing language models, another important factor to consider is context length, sometimes also referred to as the context window. This term refers to the maximum number of tokens that a model can process at once (Jurafsky & Martin, 2026). Tokens are defined by a tokenizer and represent subword units of text (Jurafsky & Martin, 2026).

The size of the context window varies across language models and is often used as an indicator of their ability to handle long input sequences. A longer context length allows a model to process longer texts and capture long-range dependencies (An et al., 2024).

However, in this study, although context length is reported for each model to provide a broader understanding of their general capabilities, this characteristic is not meaningful for the task at hand. The task, namely evaluating grammaticality in minimal sentence pairs, involves very short sentences and does not require modeling long-range dependencies.

### 2.2.5 Language Support

Language models also differ in the languages they support. While many language models are trained primarily on English data, others are pre-trained on multilingual corpora and are therefore capable of processing and generating text in multiple languages (Ozsoy, 2024).

However, even among multilingual models, the degree of language support varies (Ozsoy, 2024). For example, a model may be trained on both a high-resource language, such as English, and a low-resource

language, such as Lithuanian. Because the amount of training data available for English is typically much larger, the model may exhibit stronger capabilities in English than in Lithuanian.

In this study, language support is an important characteristic because the evaluated models represent different language coverage strategies. This study includes monolingual Lithuanian models, which are specifically designed to perform well on Lithuanian data, as well as models with a European focus, such as `EuroLLM`, which is trained on all official languages of the European Union. Finally, the study also includes models with a global language scope, meaning that they are trained on a diverse set of languages from around the world.

## 2.3 Linguistic Evaluation of Language Models

### 2.3.1 Grammaticality Judgments

As language models (LMs) become increasingly capable of solving a wide range of NLP tasks, it is crucial to develop methods to assess the extent of their linguistic knowledge. Grammaticality judgment tasks serve this purpose: through either binary classification or comparative judgments using minimal sentence pairs, they are designed to evaluate whether LMs can successfully distinguish ill-formed from well-formed sentences in natural languages, based on syntactic and, in some cases, semantic criteria.

Several benchmarks have been proposed for grammaticality judgment tasks. One of them is *Targeted Syntactic Evaluation* (TSE) (Marvin & Linzen, 2018). This work is notable for its use of minimal sentence pairs to assess language models' understanding of specific syntactic phenomena, thereby enabling a more fine-grained analysis of their strengths and weaknesses. Minimal sentence pairs consist of two sentences that differ only in a small, controlled linguistic property. In this evaluation setting, models are presented with a grammatical and an ungrammatical version of an otherwise identical sentence and must determine which one is more acceptable. However, TSE focuses on only three linguistic phenomena: subject–verb agreement, reflexive anaphora, and negative polarity items.

Another influential benchmark is the *Corpus of Linguistic Acceptability* (CoLA) (Warstadt, Singh, & Bowman, 2019), which forms part of the *General Language Understanding Evaluation* (GLUE) benchmark (A. Wang et al., 2018). CoLA is a binary classification task consisting of 10,657 grammatical and ungrammatical English sentences, which are further categorized into 15 broad classes of syntactic and semantic phenomena. Models are presented with each sentence in isolation and must determine whether it is grammatically acceptable or unacceptable. Unlike minimal-pair benchmarks, CoLA evaluates acceptability judgments at the level of individual sentences, allowing direct comparison between model predictions, human acceptability judgments, and theoretical claims in generative linguistics.

A more comprehensive benchmark based on minimal sentence pairs is the *Benchmark of Linguistic Minimal Pairs* (BLiMP) (Warstadt, Parrish, et al., 2019). BLiMP extends the minimal-pair methodology of TSE by substantially increasing the coverage of linguistic phenomena while also adopting a different evaluation setting than CoLA. Since CoLA is formulated as a supervised acceptability classification task, evaluating language models on it typically requires training a classifier on labeled acceptability judgments. Consequently, it can be difficult to determine whether a model's performance reflects grammatical knowledge already encoded in its representations or information acquired during task-specific supervised training. BLiMP avoids this issue by evaluating language models directly on minimal sentence pairs, without requiring additional acceptability supervision. The benchmark covers 12 linguistic phenomena, which are further divided into 67 paradigms (referred to in this study as *error types*). Each paradigm

contains 1,000 minimal sentence pairs that are automatically generated from handcrafted templates. As in TSE, models are evaluated based on their ability to assign a higher probability to the grammatical sentence in each pair (Marvin & Linzen, 2018).

Since the creation of BLiMP, such evaluation paradigms have become a standard method for assessing the syntactic knowledge of LMs and have inspired numerous language-specific adaptations, including Chinese (CBLiMP; (Xiang et al., 2021)), Japanese (JBLiMP; (Someya & Oseki, 2023)), Russian (RuBLiMP; (Taktasheva et al., 2024)), Icelandic (Ármannsson et al., 2025), Urdu (UrBLiMP; (Adeeba et al., 2025)), Italian (BLiMP-IT; (Barbini et al., 2025)), Irish (Irish-BLiMP; (McGiff et al., 2025)), Turkish (TurBLiMP; (Başar et al., 2025)), and Dutch (BLiMP-NL; (Suijkerbuijk et al., 2025), (Pestel et al., 2025)). In addition, the multilingual MultiBLiMP benchmark covers 101 languages (Jumelet et al., 2025).

### 2.3.2 Challenges in the Evaluation of Low-Resource and Morphologically Rich Languages

Nevertheless, while there has been a rise in BLiMP-style benchmarks (see Section 2.3.1) that have been developed not only for widely spoken languages but also for low-resource languages such as Icelandic, most well-known benchmarks remain English-centric. Consequently, underrepresented languages often include both low-resource languages and languages that differ typologically from English, such as those with rich morphological systems.

This underrepresentation is particularly important because evaluating language models only on high-resource languages such as English provides a limited understanding of their true linguistic capabilities. Even when models are multilingual and explicitly include low-resource languages, their performance on English often exceeds that on other languages (Martinez et al., 2024), largely due to the substantially larger amount of training data available (Nag et al., 2025).

Morevoer, morphologically rich languages often encode grammatical information through a wide range of inflectional markings, such as case marking, agreement, and verbal inflections (Kondratyuk et al., 2018). As a result, data available for training language models in these languages is typically sparser, and models are required to generalize across a large number of surface forms that correspond to the same underlying lexical items (Singh et al., 2023). Therefore, models that are primarily trained on English may struggle to demonstrate linguistic competence when evaluated on morphologically rich languages.

### 2.3.3 Evaluation of Lithuanian in Language Models

One such low-resource and morphologically rich language is Lithuanian, a Baltic language belonging to the Indo-European language family. It is spoken by approximately three million speakers in Lithuania, as well as by fewer than one million speakers outside the country. Interestingly, Lithuanian is considered one of the most archaic living Indo-European languages (Hogan-Brun et al., 2005). For this reason, studying it is of particular importance for linguists.

Lithuanian exhibits a rich inflectional system, including seven grammatical cases (nominative, genitive, dative, accusative, instrumental, locative, and vocative) and an additional case, the illative, which is still used in dialectal varieties (Ambrazas, 2026b). It has two grammatical genders (masculine and feminine), while neuter forms are largely restricted to adjectives and participles used substantively (Holvoet, 2026). The language also has two productive numbers (singular and plural), although remnants

of the dual number are preserved in certain noun forms and dialects (Ambrazas, 2026d). In addition, Lithuanian verbs exhibit extensive inflection, encoding categories such as person, number, tense, mood, aspect, and reflexivity (Ambrazas, 2026a).

Although Lithuanian has a preferred subject–verb–object (SVO) word order (Dryer, 2013), other word orders are also fully grammatical, particularly when used for stylistic or pragmatic emphasis (Ambrazas, 2026e). Consequently, morphological marking plays a crucial role in expressing grammatical relations, in contrast to English, where such relations are largely encoded through relatively rigid word order (Kahane et al., 2023). As a result, evaluation resources developed for English cannot be directly transferred to Lithuanian without careful adaptation.

### Evaluation of General NLP Tasks

As Lithuanian is a low-resource language, research on it remains somewhat limited. However, researchers continuously work to address existing gaps, and studies covering various NLP tasks are steadily emerging.

For example, Kostiuk et al. (2025b) evaluates multilingual large language models (LLMs) on multiple-choice question answering tasks, focusing on Lithuanian and general knowledge domains, including history. Another study by Kostiuk et al. (2025a) evaluates LLMs on short-question answering tasks in Lithuanian and Latvian, two languages belonging to the Baltic language family.

In his Master's thesis, *Large Language Models for Lithuanian Language*, Plesevičius (2025) evaluates LLMs on several widely used benchmarks: TruthfulQA (Lin et al., 2021), which measures factual accuracy; ARC (Clark et al., 2018), which assesses the ability to answer science-related questions; WinoGrande (Sakaguchi et al., 2019), which evaluates commonsense reasoning and word-sense disambiguation; MMLU (Hendrycks et al., 2020), which measures performance across a broad range of subjects, including mathematics and history; HellaSwag (Zellers et al., 2019), which evaluates commonsense reasoning capabilities; and GSM8K (Cobbe et al., 2021), which assesses mathematical problem-solving skills.

Mandravickaitė et al. (2025) evaluates LLMs on a text summarization task in which models are instructed to rewrite administrative texts containing domain-specific jargon into plain-language summaries that are understandable to a broader audience. Similarly, Kuoraitė and Gružauskas (2025) investigates the ability of LLMs to simplify Lithuanian texts, making them more accessible to individuals with cognitive challenges as well as learners of the Lithuanian language.

Finally, Bryskina et al. (2025) evaluates how effectively open-source LLMs can detect biases related to race, gender, religion, and physical appearance.

### Syntactic Evaluation Tasks

Although there is currently no BLiMP-style benchmark specifically developed for Lithuanian, the recently introduced MultiBLiMP benchmark (A Massively Multilingual Benchmark of Linguistic Minimal Pairs; Jumelet et al., 2025) includes Lithuanian among its 101 evaluated languages, with 1,180 minimal sentence pairs. MultiBLiMP reports the performance of eleven language models, including the multilingual models `LLaMA 3` (8B base, 70B base, and 70B instruct) (Llama Team, 2024), `Aya-expanse` (32B instruct) (Dang et al., 2024), `Gemma 3` (27B base and instruct) (Gemma Team, 2025), `OLMo2` (32B base and instruct) (Team OLMo, 2024), `Qwen3` (14B instruct) (Qwen Team, 2025), and `EuroLLM` (9B base) (Martins et al., 2024), as well as the monolingual Lithuanian model `Goldfish` (Chang et al., 2024).

Despite its importance as a large-scale multilingual syntactic benchmark to include numerous low-resource languages, MultiBLiMP has several limitations. First, the minimal sentence pairs used in the benchmark are generated automatically from Universal Dependencies (UD) treebanks[1] and UniMorph (UM) resources[2]. Consequently, the resulting datasets are not manually validated and may contain annotation errors and unnatural sentence constructions.

Second, the range of linguistic phenomena covered by the benchmark is relatively narrow. The evaluation focuses exclusively on two types of subject-verb agreement: subject-finite-verb agreement and subject-participle agreement, tested across the grammatical categories of number, person, and gender. These phenomena represent only a small subset of a grammar of any language. Moreover, in morphologically rich languages, agreement relations are typically expressed through overt morphological marking. As a result, language models can often rely on surface-level morphological cues when making grammaticality judgments, reducing the need for deeper syntactic generalization.

Therefore, while MultiBLiMP (Jumelet et al., 2025) provides an important foundation for evaluating the syntactic knowledge of language models in low-resource languages, its results should not be interpreted as a comprehensive assessment of linguistic competence in Lithuanian. The evaluated models achieved accuracies ranging from 83.2% (`OLMo2`, a model primarily trained on English-language data) to 98.5% (`EuroLLM`, a model specifically designed for European languages) (see Table 2.1). Such high scores suggest that the benchmark may not be sufficiently challenging to reveal finer-grained differences in the models' grammatical knowledge. Nevertheless, all eleven models included in MultiBLiMP are also evaluated in the present study, enabling direct comparison with previously reported results.

| **Model Family** | **Version** | **Accuracy** |
|---|---|---|
| LLaMA 3 | 8B Base | 93.6 |
| | 70B Base | 96.6 |
| | 70B Instruct | 96.1 |
| Aya-expanse | 32B Instruct | 92.3 |
| Gemma 3 | 27B Base | 98.1 |
| | 27B Instruct | 95.5 |
| OLMo2 | 32B Base | 85.1 |
| | 32B Instruct | 83.2 |
| Qwen3 | 14B Instruct | 96.3 |
| EuroLLM | 9B Base | 98.5 |
| Goldfish | 125M Base | 97.9 |

Table 2.1: Accuracy of language models on Lithuanian in the MultiBLiMP benchmark, reported as percentages in the original study. The lowest and highest accuracies among the evaluated models are highlighted in red and green, respectively.

To the best of my knowledge, research on the syntactic evaluation of language models for Lithuanian remains scarce beyond MultiBLiMP (Jumelet et al., 2025). Related work includes syntactic role labeling based on Universal Dependencies treebanks, which evaluates models' ability to recover argument structure rather than their sensitivity to grammatical errors such as incorrect case marking (Temesgen

[1] https://universaldependencies.org
[2] https://unimorph.github.io

et al., 2025). Consequently, there remains a substantial gap in the evaluation of Lithuanian language models on a broader range of syntactic and morphosyntactic phenomena.

## 2.4 Human Acceptability Judgments

Human acceptability judgments play a crucial role in syntactic evaluation, as they provide a direct measure of how native speakers perceive well-formed and ill-formed sentences in a given language (Chomsky, 1965; Schütze, 1996). Such measures may reveal more than purely grammatical descriptions and can therefore serve as a valuable benchmark when evaluating language models. In this way, human acceptability judgments help provide a more grounded assessment of whether models actually possess the linguistic knowledge in question. Moreover, this approach is of particular value for low-resource languages (such as Lithuanian), which often have limited availability of high-quality linguistic resources.

For example, human acceptability judgments could be a valuable addition to benchmarks such as MultiBLiMP. Due to its limited coverage of syntactic phenomena, it is important to verify whether the tested constructions are genuinely challenging. Moreover, when using generated datasets, there is a risk of including sentences that are grammatically correct but unnatural to native speakers, thereby introducing bias into the evaluation.

For this reason, the present study also includes human acceptability ratings. However, this approach is not novel; rather, it is inspired by prior work in the field. For instance, the BLiMP benchmark (Warstadt, Parrish, et al., 2019) employed 20 native English speakers to produce 6,700 judgments in total, in order to verify whether sentence pairs reliably contrast grammatical and ungrammatical constructions. In this setup, human participants were given a task similar to that performed by language models: a forced-choice task in which they were presented with minimal sentence pairs and asked to select the more acceptable sentence. Similar forced-choice paradigms have also been used in earlier benchmarks, such as CoLA (Warstadt, Singh, & Bowman, 2019) and TSE (Marvin & Linzen, 2018).

Furthermore, the forced-choice format is also used in most BLiMP adaptations, with exceptions such as BLiMP-NL (Suijkerbuijk et al., 2025) and TurBLiMP (Başar et al., 2025), which was inspired by the Dutch benchmark. These studies adopt gradient grammaticality judgments due to evidence that human judgments of grammaticality are often gradient rather than binary (Lau et al., 2017).

Therefore, the present study follows this latter approach as well, employing a 7-point Likert scale (Likert, 1932) to collect human acceptability ratings. Joshi et al. (2015) argue that this scale is preferable to a 5-point scale, as it provides participants with greater granularity rather than forcing them to round their judgments to the nearest category.

# 3 Methodology Overview

In this chapter, I present the methodological approach used in this study. Specifically, I describe the processes of dataset creation, model evaluation, and the collection of acceptability ratings. I then outline the approach used for discussing the results and conducting the error analysis (see Figure 3.1).

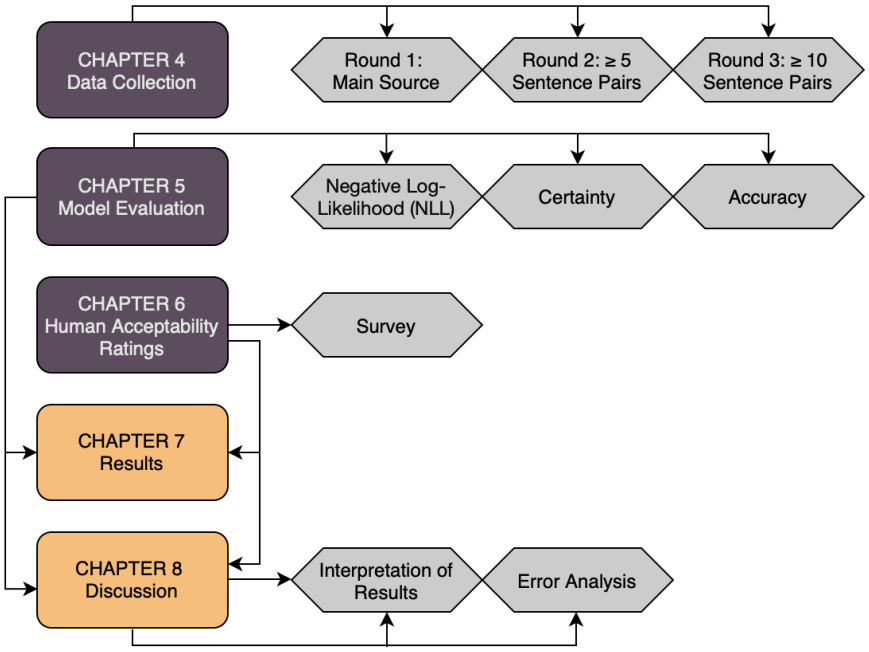


Figure 3.1: Flowchart depicting the methodological steps undertaken to address the research questions.

## 3.1 Data

To evaluate the syntactic knowledge of Lithuanian in language models (LMs), I present the models with minimal sentence pairs. For this purpose, I construct a dataset of Lithuanian minimal pairs covering four phenomenon types: *Use of Cases*, *Use of Prepositions*, *Use of Forms*, and *Coordination of Sentence Elements and Clauses*. These phenomenon types are further subdivided into individual phenomena and error types (see Sections 4.4.1 and 4.4.2).

The dataset is constructed in three rounds. In each round, additional sentences from different sources (see Section 4.2) are incorporated until the dataset reaches a final size of 777 minimal pairs. The dataset creation process is described in detail in Section 4.3.

## 3.2 Model Evaluation

The next step in the study involves selecting models for evaluation (see Chapter 5). In total, the study includes 13 language model (LM) families, covering different model sizes and versions, resulting in 78 models overall. The selection comprises both monolingual and multilingual models, with varying language coverage, including models focused on European languages as well as more globally oriented ones.

The selected models are evaluated using minimal sentence pairs (see Section 5.3). For each sentence, negative log-likelihood (NLL) values are calculated. These values are then used to determine which sentence in each pair is preferred by the model, as well as the model's confidence in its choice (certainty). Finally, model performance is assessed by computing accuracy across different phenomena, error types, and the dataset as a whole.

## 3.3 Human Acceptability Ratings

Another important component of the study is the design and collection of sentence acceptability ratings from native speakers of Lithuanian (see Chapter 6). These ratings provide insight into whether the selected phenomena and error types are sufficiently challenging, and they serve as a human benchmark for evaluating the models' performance. I collect these ratings through a survey conducted using Qualtrics[1] (see Section 6.1).

I then present the results using both descriptive statistics and inferential analyses based on linear mixed-effects models (see Section 6.4). This approach allows me to test whether the observed effects hold while accounting for variability across participants and items.

## 3.4 Results

While human acceptability ratings are discussed in Chapter 6, the results of the model evaluation are presented both separately and in comparison with human judgments in Chapter 7.

Here, I provide quantitative insights based on model accuracy (see Sections 7.1-7.3), followed by logistic mixed-effects models to analyse the influence of model characteristics such as model size and language support (see Section 7.4). Finally, model predictions are compared with human responses using Spearman rank correlations (see Sections 7.5 and 7.6).

## 3.5 Discussion

The results are discussed in two parts. First of all, they are interpreted by a research question and the corresponding hypothesis (see Section 8.1). Then, the qualitative error analysis focuses on identifying systematic patterns in the errors made by both the models and human participants, as well as comparing these patterns (see Section 8.2). It aims to determine which phenomena and error types are particularly challenging for the models and whether similar difficulties are observed among native speakers of Lithuanian. In addition, this analysis provides insight into the quality of the dataset and highlights potential directions for future work.

[1] https://www.qualtrics.com

# 4 Data Collection

This chapter presents the structure and content of the dataset used in the present study. It also describes the steps taken during the dataset creation process.

## 4.1 Motivation and Design Principles

To achieve the goal of this thesis, namely, to evaluate the syntactic knowledge of language models on Lithuanian, it was essential to construct an appropriate dataset. This proved to be one of the main challenges at the outset of the research. To the best of my knowledge, such an evaluation has not previously been conducted for Lithuanian, and therefore no suitable dataset was available.

Before creating the dataset, several considerations were established as guiding principles. First, the dataset needed to be sufficiently diverse and include more than only those errors that are highly transparent due to explicit morphological marking, as such markings may serve as clear cues for the models. Second, it was important to focus on relevant errors, namely, those that frequently occur in native speakers' usage. Finally, the selected errors had to be sufficiently challenging, meaning that they should not be trivially easy even for native speakers.

## 4.2 Sources

The dataset consists of 777 sentence pairs. The most important source is the *Valstybinė lietuvių kalbos komisija* (VLKK) (Valstybinė lietuvių kalbos komisija (VLKK), 2023), which provided the largest number of sentence pairs and also determined the dataset structure presented in Section 4.4. In addition, four other sources (see Table 4.1), described in detail below, were used to expand the base dataset and ensure a sufficient number of sentence pairs for each error type.

### 4.2.1 Valstybinė lietuvių kalbos komisija

Following the design principles outlined above (see Section 4.1), the primary source for the dataset was selected to be the *Valstybinė lietuvių kalbos komisija* (VLKK; in English, *The State Commission of the Lithuanian Language*) (Valstybinė lietuvių kalbos komisija (VLKK), 2023). The VLKK is a state institution that reports to the Seimas (the Lithuanian Parliament). It is composed of 17 members appointed for five-year terms and is tasked with safeguarding the official status of the Lithuanian language. In addition, the Commission oversees language policy and standardization, and approves authoritative linguistic resources, including dictionaries and textbooks.

The official VLKK website provides a list entitled *Didžiųjų kalbos klaidų sąrašas* (in English, *List of Major Language Errors*), which documents some of the most frequent errors in Lithuanian language

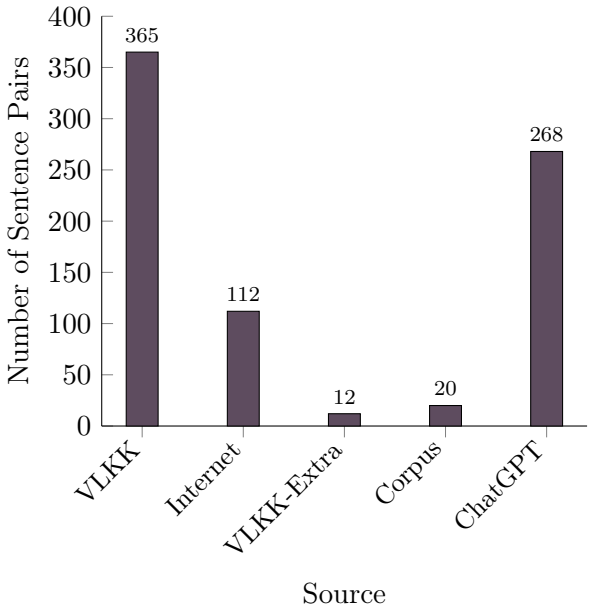


Figure 4.1: Number of sentence pairs by source.

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). It was first published in 1997 and, for more than twenty years, until 2019, had legal force. During that period, failure to comply with the prescribed norms could result in administrative fines, although in practice these were typically replaced by official warnings (Jačauskas, 2019).

Although the list no longer has legal authority, the errors described in it are still considered non-normative. According to Aurelija Dvylytė, Head of the Consultation Group of the State Commission of the Lithuanian Language, while some of the listed errors may be less salient today, they continue to occur in everyday usage. This is particularly noticeable in the current sociolinguistic context of Lithuania, where the number of non-native speakers has increased, especially among speakers whose first language is Russian[1].

Therefore, this list aligns well with the design principles of the dataset: the errors it documents are still attested in contemporary language use and remain sufficiently challenging, even for native speakers. The dataset includes 365 sentence pairs from this source.

### 4.2.2 Additional Websites

When expanding the dataset originally sourced from VLKK (Valstybinė lietuvių kalbos komisija (VLKK), 2023), additional Lithuanian-language websites were used to ensure dataset variability and reduce potential bias during the creation process. Whenever possible, sentences were collected from sources such as municipal government websites and exam preparation resources, resulting in 112 sentence pairs. The list of these websites is provided in Appendix A.1.

Furthermore, 12 sentence pairs were obtained from the official VLKK website, particularly from the *Konsultacijų bankas* (in English, *Consultation Bank*).

[1]Personal communication with A. Dvylytė via email on 12th February, 2025.

### 4.2.3 Corpus

The third source used for dataset construction was the Lithuanian language corpus, specifically the Corpus of Contemporary Lithuanian (Vytauto Didžiojo universitetas (VDU), 2013). When searching for additional sentence pairs, words that are typically associated with errors were selected and used as search queries in the corpus. The retrieved correct examples were then manually modified to produce corresponding erroneous examples.

However, this source proved to be challenging to use. First, the selection of search words risked introducing bias, as choosing specific query words could influence the types of errors collected. Second, despite the large size of the corpus (approximately 140.9 million words), the number of relevant search results was often very limited, making it difficult to obtain high-quality examples. Therefore, only 20 examples were collected using corpus.

### 4.2.4 Semi-Automatically Generated Sentence Pairs

The final resource used for dataset expansion involved generating sentence pairs using `ChatGPT-5.1` (OpenAI, 2025), followed by manual correction. This approach was employed to increase lexical variability and reduce potential selection bias.

When generating sentence pairs, `ChatGPT` was provided with an error description along with several examples of correct and incorrect sentences. The model was then instructed to generate ten additional sentence pairs similar to the provided examples, while maintaining lexical diversity (see Appendix A.2).

However, many of the examples generated by `ChatGPT` contained errors, were repetitive, or were otherwise unsuitable for inclusion in the dataset. Therefore, all generated examples were manually reviewed, and only corrected and reliable sentence pairs were selected, adding additional 268 sentence pairs to the dataset.

## 4.3 Dataset Creation Process

The dataset was created in three rounds. In the first round, only examples from the VLKK *List of Major Language Errors* (Valstybinė lietuvių kalbos komisija (VLKK), 2023) were collected, resulting in 365 sentence pairs. In the second round, the dataset was expanded to ensure that each error type was represented by at least five sentence pairs. After this stage, the dataset contained 445 sentence pairs. In the final, third round, the dataset was further expanded to include at least ten sentence pairs per error type, resulting in a total of 777 sentence pairs. The distribution of sentence pairs across different language phenomena after each round is presented in Appendix A.3: Figures A.1, A.2, and A.3.

## 4.4 Dataset Structure

The final dataset consists of 777 sentence pairs, divided into four language phenomenon types: *Use of Cases*, *Use of Prepositions*, *Use of Forms*, and *Coordination of Sentence Elements and Clauses* (see Table 4.2).

Each phenomenon type is further subdivided into individual phenomena, 31 in total: *Linksnių vartojimas* (in English, *Use of Cases*) comprises 6 phenomena, *Prielinksnių vartojimas* (in English, *Use of Prepositions*) 10, *Formų vartojimas* (in English, *Use of Forms*) 11, and *Sakinio dalių ir sakinių jungimas* (in English, *Coordination of Sentence Elements and Clauses*) 4.

At the most fine-grained level, these phenomena are further divided into specific error types. In total, the dataset includes 22 error types within *Use of Cases*, 20 within *Use of Prepositions*, 18 within *Use of Forms*, and 4 within *Coordination of Sentence Elements and Clauses* (see Appendix A.4).

Finally, for error types that concern only specific words, the dataset includes a separate column identifying the source of the problem, namely the problematic word.

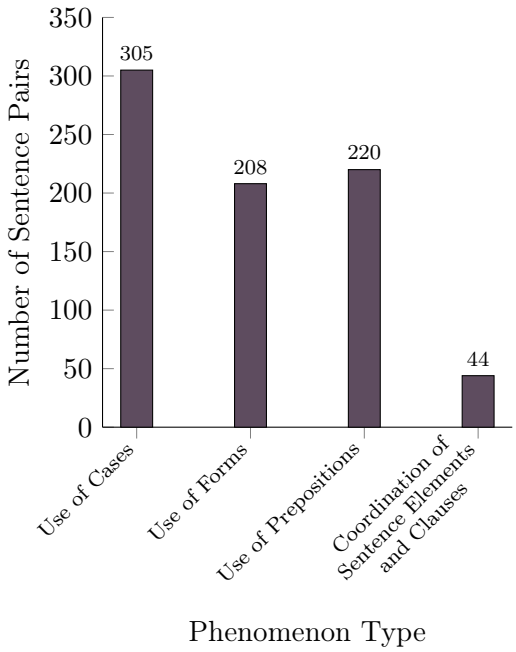


Figure 4.2: Distribution of sentence pairs by phenomenon type.

### 4.4.1 Linguistic Phenomena

The dataset comprises 31 linguistic phenomena. Within *Use of Cases*, six phenomena are distinguished, corresponding to the six core grammatical cases in Lithuanian, with the exception of the vocative. These include the nominative, genitive, dative, accusative, instrumental, and locative cases. Therefore, the vocative, the seventh grammatical case in Lithuanian, is not represented as a separate phenomenon. As described on the official website of the Municipality of Marijampolė (Marijampolės savivaldybė, 2017), this case is relatively rare and its use is largely restricted to directly addressing people. Consequently, native speakers rarely make errors involving the vocative, with exceptions occurring mainly in writing, where incorrect case endings may be used, or in pronunciation, where stress may be placed on the wrong syllable.

However, one specific error type (3.1.2) under the *Nominative* phenomenon concerns the incorrect use of the nominative in contexts where the vocative should be used. Thus, although the vocative is not treated as a separate phenomenon, it is indirectly represented in the dataset. Consequently, all seven grammatical cases of Lithuanian are covered.

The *Use of Prepositions* category comprises ten linguistic phenomena corresponding to the prepositions *ant* (in English, *on*), *apie* (*about*), *į* (*into/to*), *iki* (*until/up to*), *iš* (*from/out of*), *pas* (*at/to someone's place*), *po* (*after/under*), *prie* (*near/by*), *prieš* (*before/against/in front of*), and *už* (*behind/for*).

The third linguistic phenomenon type, *Use of Forms*, comprises eleven linguistic phenomena. These

include *Gender Forms*; *abu, abi* (the masculine and feminine forms corresponding to the English *both*); *Ordinal numbers denoting decades*; *Infinitive*; *Negative predicate forms*; *Simple forms of the subjunctive mood*; *Reflexive forms*; *Participles*; *Half-participles*; *Adverbial participles*, and *Adverbs*. Overall, this category covers a broad range of challenging form-related phenomena in the language.

Finally, the smallest category *Coordination of Sentence Elements and Clauses* comprises four phenomena: *gi* (a discourse particle used for emphasis), *ir* (in English, *and*), *sykį* (*once*), and *vietoj to, kad* (*instead of*).

### 4.4.2 Types of Errors

Across the 31 linguistic phenomena included in the dataset, 64 distinct error types are identified. While some phenomena involve only a single error type, such as all phenomena within the *Coordination of Sentence Elements and Clauses* category, others comprise multiple error types. In some cases, the number of errors reaches up to nine, as observed for the *apie* phenomenon within the *Use of Prepositions* category.

Below, I provide an example of each phenomenon type to illustrate the dataset. All examples are glossed according to the Leipzig Glossing Rules (Max Planck Institute for Evolutionary Anthropology, 2015). Ungrammatical sentences are marked with an asterisk. Moreover, the minimal differences are colour-coded (green for grammatical and red for ungrammatical sentence parts). The list of abbreviations is provided in Table 4.1.

| **Abbreviation** | **Meaning** |
|---|---|
| 2 | Second person |
| 3 | Third person |
| ACC | Accusative |
| DAT | Dative |
| F | Feminine |
| GEN | Genitive |
| M | Masculine |
| NEG | Negation |
| NOM | Nominative |
| PASS | Passive |
| PL | Plural |
| PFV | Perfective |
| PRS | Present |
| PST | Past |
| PTCP | Participle |
| REFL | Reflexive |
| SG | Singular |

Table 4.1: Glossing abbreviation conventions.

***Use of Cases - 3.3 Dative case - 3.3.4 Dative should not be used with verbs of motion to express purpose.***

(1) *Su-si-rin-ko* *į* *posėd-į.*
PFV-REFL-gather-PST.3PL to meeting-M.ACC.SG

'They gathered for a meeting.'

(2) *Su-si-rink-o* *posėdž-iui.*
PFV-REFL-gather-PST.3PL meeting-M.DAT.SG

'They gathered for a meeting.'

***Use of Prepositions - 4.3 Į - 4.3.1 Į should not be used to express the object of an action with certain words.***

(3) *Vadov-as* *nu-rod-ė* *kelet-ą* *darb-o* *trūkum-ų.*
manager-M.NOM.SG PFV-show-PST.3SG several-ACC.PL job-M.GEN.SG shortcoming-M.GEN.PL

'The manager pointed out several shortcomings in the work.'

(4) **Vadov-as* *nu-rod-ė* *į* *kelet-ą* *darb-o* *trūkum-ų.*
manager-M.NOM.SG PFV-show-PST.3SG to several-ACC.PL job-M.GEN.SG shortcoming-M.GEN.PL

'The manager pointed out several shortcomings in the work.'

***Use of Forms - 5.7 Reflexive forms - 5.7.1 Reflexive forms of certain verbs should not to be used with a passive meaning if the action cannot occur by itself.***

(5) *Nosin-ė* *čia* *ne-raš-om-a.*
*nosinė*-F.NOM.SG here NEG-write-PRS.PASS.PTCP-F.NOM.SG

'The diacritic *nosinė* is not used here.'

(6) **Nosin-ė* *čia* *ne-si-raš-o.*
*nosinė*-F.NOM.SG here NEG-REFL-write-PRS.3SG

'The diacritic *nosinė* is not used here.'

***Coordination of Sentence Elements and Clauses - 6.2 Ir - 6.2.1 Ir should not be used to link words in intensifying constructions with adjectives and adverbs.***

(7) *Buv-ome* *labai* *labai* *laiming-i.*
to.be-PST.2PL very very happy-M.NOM.PL

'We were very, very happy.'

(8) **Buv-ome* *labai* *ir* *labai* *laiming-i.*
to.be-PST.2PL very and very happy-M.NOM.PL

'We were very, very happy.'

### 4.4.3 Sources of Errors

Finally, some errors occur only in combination with specific words. In such cases, the dataset includes a separate column, `problem_source`, which specifies the particular word associated with the error.

For example, error 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*) applies to specific verbs. One such verb is *įvykti* (in English, *to happen*). Accordingly, the ninth sentence pair in the dataset contains this verb:

***Use of Cases - 3.1 Nominative case - 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.***

(9) *Per t-ą laik-ą į-vyk-o ir nelaim-ių.*
during that-ACC time-M.ACC.SG happen-PST.3PL and accident-F.GEN.PL

'During that time, accidents also happened.'

(10) **Per t-ą laik-ą į-vyk-o ir nelaim-ės.*
during that-ACC time-M.ACC.SG happen-PST.3PL and accident-F.NOM.PL

'During that time, accidents also happened.'

### 4.4.4 Exceptions

Importantly, out of the 777 sentence pairs in the dataset, 20 represent exceptions. In these cases, the sentences do not follow the general rule; therefore, sentences that would otherwise be considered "incorrect" are treated as "correct". For example, when referring to a person, the title should agree in gender with the person being described. When referring to a woman, the feminine form should be used, such as *literatūros kritikė Ambraškaitė* (see Example 11) instead of *literatūros kritikas Ambraškaitė* (see Example 12). However, there are exceptions. In particular, masculine nouns are used when referring to academic degrees. Therefore, it is more appropriate to write *Absolventei Rutkutei įteiktas chemijos dėstytojo diplomas* (see Example 13) rather than *Absolventei Rutkutei įteiktas chemijos dėstytojos diplomas* (see Example 14).

***Use of Forms - 5.1 Gender forms - 5.1.1 Masculine-gender noun forms are not to be used to refer to women when describing professions, positions, academic titles.***

(11) *Literatūr-os kritik-ė Ambraškait-ė pri-stat-ė nauj-ą*
literature-F.GEN.SG critic-F.NOM.SG Ambraškaitė-F.NOM.SG PFV-present-PST.3SG new-ACC.SG
*poezij-os rinktin-ę ir ap-tar-ė šiuolaikin-ių*
poetry-F.GEN.SG anthology-F.ACC.SG and PFV-discuss-PST.3SG contemporary-GEN.PL
*autor-ių tendencij-as.*
author-GEN.PL trend-F.NOM.PL

'Literary critic Ambraškaitė presented a new poetry anthology and discussed trends among contemporary authors.'

(12) **Literatūr-os* *kritik-as* *Ambraškait-ė* *pri-stat-ė* *nauj-ą*
literature-F.GEN.SG critic-M.NOM.SG Ambraškaitė-F.NOM.SG PFV-present-PST.3SG new-ACC.SG
*poezij-os* *rinktin-ę* *ir* *ap-tar-ė* *šiuolaikin-ių*
poetry-F.GEN.SG anthology-F.ACC.SG and PFV-discuss-PST.3SG contemporary-GEN.PL
*autor-ių* *tendencij-as.*
author-GEN.PL trend-F.NOM.PL

'Literary critic Ambraškaitė presented a new poetry anthology and discussed trends among contemporary authors.'

**EXCEPTION:** ***Use of Forms - 5.1 Gender forms - 5.1.1 Masculine-gender noun forms are not to be used to refer to women when describing professions, positions, academic titles; however, they are used when referring to academic degrees.***

(13) *Absolvent-ei* *Rutkut-ei* *į-teik-t-as* *chemij-os*
graduate-F.DAT.SG Rutkutė-F.DAT.SG PFV-give-PAST.PASS.PTCP-M.NOM.SG chemistry-F.GEN.SG
*dėstytoj-o* *diplom-as.*
teacher-M.GEN.SG diploma-M.NOM.SG

'Graduate Rutkutė was awarded a chemistry teacher diploma.'

(14) **Absolvent-ei* *Rutkut-ei* *į-teik-t-as* *chemij-os*
graduate-F.DAT.SG Rutkutė-F.DAT.SG PFV-give-PAST.PASS.PTCP-M.NOM.SG chemistry-F.GEN.SG
*dėstytoj-os* *diplom-as.*
teacher-F.GEN.SG diploma-M.NOM.SG

'Graduate Rutkutė was awarded a chemistry teacher diploma.'

## 4.5 Sentence Length

It is worth noting that the sentences provided by VLKK (Valstybinė lietuvių kalbos komisija (VLKK), 2023) were generally relatively short, most often consisting of a single clause without additional subordination. The average sentence length was 5.26 words per sentence, with a median sentence length of 5 words. This was not considered a disadvantage. On the contrary, keeping sentences relatively short helps limit the influence of additional lexical material or syntactic structures that are unrelated to the target error (Salhan, 2025).

Accordingly, when adding further sentence pairs in Rounds 2 and 3, preference was also given to shorter sentences to maintain comparability across sources. As a result, the average sentence length of sentences from additional sources was 6.53 words (median = 6), while the overall average sentence length of the dataset was 5.93 words. Therefore, most sentences in the dataset contain approximately 4–8 words.

## 4.6 Minimality in Sentence Pairs

An important notion when presenting the dataset is minimality. Although the dataset consists of minimal sentence pairs, the type of minimal change between grammatical and ungrammatical sentences varies depending on the linguistic phenomenon and the corresponding error type. The dataset creation process started from the aforementioned VLKK (Valstybinė lietuvių kalbos komisija (VLKK), 2023) examples; therefore, in rounds 2 and 3, newly constructed sentences followed the structure of the original ones.

In some cases, the difference between sentence pairs is as small as a word part, such as a word ending in phenomena related to case usage (see Examples 1 and 2). In other cases, the difference involves an entire lexical item, as illustrated in Examples 5 and 6. Sentence pairs may also differ in the number of words. For instance, in Examples 7 and 8, the ungrammatical sentence contains an additional conjunction *ir* (in English, *and*).

For some error types, correcting an ungrammatical sentence requires more than a single modification. For example, in Examples 15 and 16, the sentences differ both in lexical choice and in the number of words, as the ungrammatical sentence contains an additional word. In another sentence pair (Examples 17 and 18), the correction involves replacing *pas ją* with *jos*. Both constructions attempt to express possession by *ji* (in English, *she*), in this case indicating that *she has beautiful hair*, but only the latter is grammatical.

In all cases, however, the corrections are designed to involve the smallest possible change required to make the sentence grammatical.

***Use of Prepositions - 4.1 Ant - 4.1.8 Ant should not be used to express the nature/mode/condition of an action or state.***

(15) *Nuomoj-u* *but-ą.*
rent-PRS.1SG apartment-M.ACC.SG

'I am renting an apartment.'

(16) **Gyven-u* *ant* *but-o.*
live-PRS.1SG on apartment-M.GEN.SG

'I am renting an apartment.'

***Use of Prepositions - 4.6 Pas - 4.6.1 Pas should not be used to indicate ownership, belonging, or possession.***

(17) *J-os* *graž-ūs* *plauk-ai.*
she-F.GEN.SG beautiful-M.NOM.PL hair-M.NOM.PL

'She has beautiful hair.'

(18) **Pas* *j-ą* *graž-ūs* *plauk-ai.*
at she-F.ACC.SG beautiful-M.NOM.PL hair-M.NOM.PL

'She has beautiful hair.'

# 5 Model Evaluation

This chapter provides an overview of the models evaluated in this study, including a description of the language model families and their main characteristics, as well as the corresponding evaluation process.

## 5.1 Language Model Families

This study involves 13 language model (LM) families: `Aya-expanse` (Dang et al., 2024), `EuroLLM` (Martins et al., 2024), `Phi 3` (Microsoft Team, 2024), `Goldfish` (Chang et al., 2024), `Gemma 3` (Gemma Team, 2025), `LLaMa 3-3.2` (Llama Team, 2024), `LLaMa 4` (Meta AI, 2025), `Neurotechnology` (Nakvosas et al., 2024), `OLMo2` (Team OLMo, 2024), `Qwen3` (Qwen Team, 2025), `Salamandra` (Gonzalez-Agirre et al., 2025), `Teuken` (Ali et al., 2024), and `TildeOpen` (Bergmanis et al., 2026) (see Table 5.1). Notably, the `Neurotechnology` family consists of continual pre-training adaptations of the `LLaMa 2` models (Meta GenAI, 2023), specifically the 7B and 13B parameter variants.

In selecting models for this study, priority was given to the most recently released families. Nevertheless, several `LLaMA` models from the 3.1-3.3 versions were also included, as `LLaMA 4` only provides very large models, of which only `LLaMA 4 Scout` was included. The `LLaMA 4 Maverick` variant was excluded due to computational constraints. Including these mid-sized predecessors allows for more meaningful comparisons within the same model lineage.

| Model Family | Size | Version | Supported Languages | LT |
|---|---|---|---|---|
| Aya-expanse | 8B, 32B | Instruct | 23 languages | No |
| EuroLLM | 1.7B, 9B, 22B | Base/Instruct | European languages | Yes |
| Phi 3 | 3.8B, 7B, 14B (Context: 4K/8K/128K) | Instruct | 23 languages | No |
| Goldfish | 39M (Data: 5MB/10MB)<br>125M (Data: 100MB/1GB) | Base | Monolingual (350 languages) | Yes |
| Gemma 3 | 270M, 1B, 4B, 12B, 27B | Base/Instruct | > 140 languages | Yes[a] |
| LLaMa 3.1–3.3 | 1B, 3B, 8B, 70B | Base/Instruct | 8 languages | No |
| LLaMa 4 | 17B-16E | Base/Instruct + MoE | 12 languages | No |
| Neurotechnology | 7B, 13B | Base/Instruct | Monolingual (Lithuanian) | Yes |
| OLMo2 | 1B, 7B, 13B, 32B | Base/Instruct | Primarily English | No |
| Qwen3 | 0.6B, 1.7B, 4B, 8B, 14B, 32B, 30B, 235B | Base/Instruct + MoE | 119 languages | Yes |
| Salamandra | 2B, 7B, 40B | Base/Instruct | European languages | Yes |
| Teuken | 7B (Versions: 0.4/0.6) | Base/Instruct | European languages | Yes |
| TildeOpen | 30B | Base | 34 languages | Yes |

[a] Lithuanian is not officially supported by this model (see Section 5.2.4).

Table 5.1: Language models (LMs) included in this study. The table presents the LM families, model sizes, and available versions. It also lists the supported languages and indicates whether Lithuanian (LT) is among them.

Each family includes one or more model sizes: `Teuken` and `TildeOpen` have only a single size, whereas some families, such as `Qwen3`, offer multiple sizes (`Qwen3` includes eight). In addition to size variations, most LM families provide both base and instruct versions, while some offer only one of these

options. Overall, this study evaluates 78 LMs.

These LM families also differ in terms of language support. When selecting models, I focused primarily on those families that officially support Lithuanian. However, I later included additional model families that are frequently mentioned in evaluation contexts related to the Lithuanian language or culture (see Section 2.3.3). Two models, `Neurotechnology` and `Goldfish`, are monolingual, while the others target European languages or have a global scope. Notably, seven out of the thirteen LM families explicitly report support for Lithuanian, whereas the remaining models either do not support it or do not specify whether it is supported.

Importantly, although all models from the `Phi-3` family are advertised as small language models (SLMs) rather than full-scale large language models (LLMs), in this study I adopt the common convention of referring to models with $\leq$ 10B parameters as SLMs (Fu et al., 2023). While models with more than 100B parameters are sometimes classified exclusively as LLMs (Fu et al., 2023), for simplicity I treat all models larger than 10B parameters as LLMs. As a result, some model families consist entirely of SLMs, others entirely of LLMs, and some include both SLM and LLM variants. For example, following this categorization, the aforementioned `Phi-3` family includes two SLMs (3.8B and 7B parameters) as well as one LLM (14B parameters).

## 5.2 Model Characteristics

This section describes the key characteristics of the models, including their size, versions, context lengths, and language coverage (see Figure 5.1).

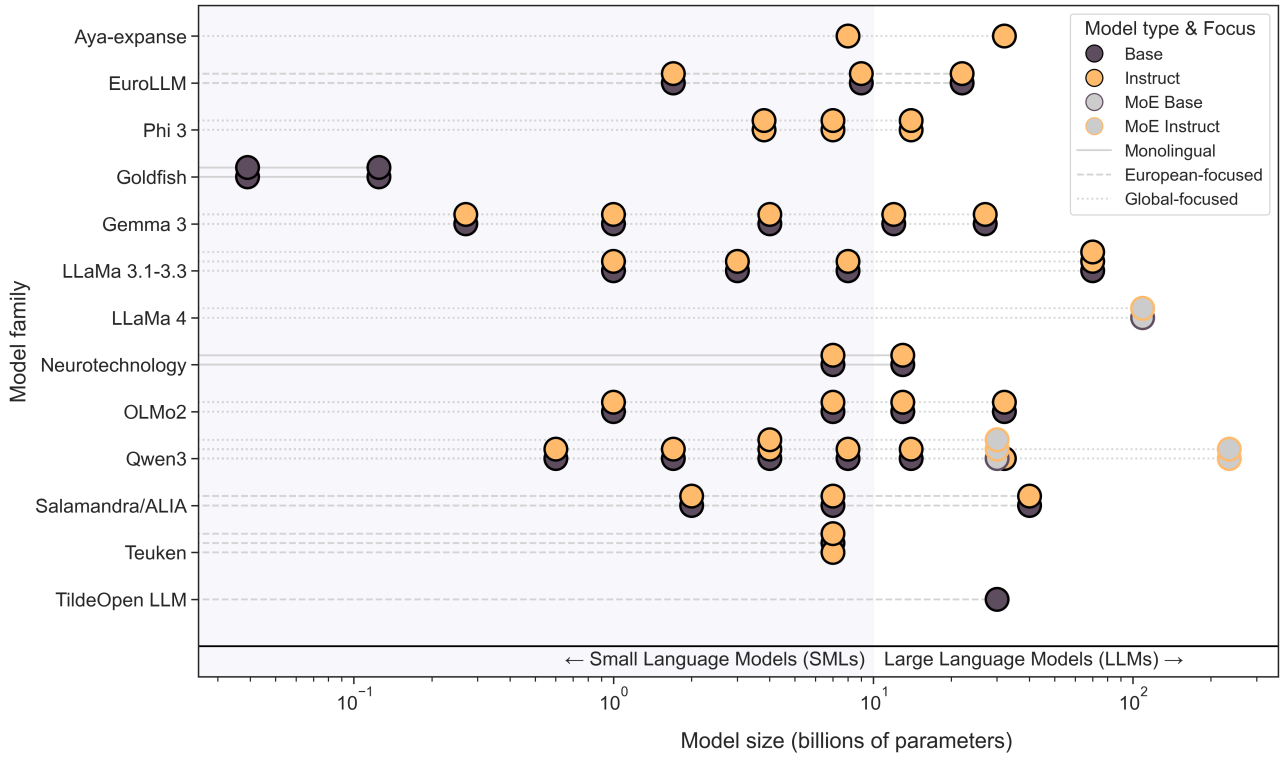


Figure 5.1: Presentation of model sizes and variants on a logarithmic scale. The darker background on the left represents small language models ($\leq$ 10B parameters), while the lighter background on the right represents large language models.

### 5.2.1 Model Sizes

The language models evaluated in this study cover a broad range of sizes in terms of total parameters (see Figure 5.1). Of the 78 evaluated models, 47 have $\leq$ 10B parameters and are classified as small language models (SLMs), while the remaining 31 exceed 10B parameters and are categorized as large language models (LLMs). A complete overview of all models, including their sizes, versions, context lengths, and supported languages, is provided in Appendix (see Table B.1). Among the SLMs, 14 models are extremely small, with $\leq$ 1B parameters. These include all four models from the `Goldfish` family (39M and 125M parameters), all of which are base models that differ in the amount of training data. For Lithuanian, two 39M models are trained on 5MB and 10MB of data, while two 125M models are trained on 100MB and 1GB. This variation allows for exploration of how training data scale may influence model behaviour. In addition, `Gemma 3` provides models with 270M and 1B parameters, both available in base and instruct variants. Similarly, the `LLaMa 3.2` and `OLMo2` families each include a 1B-parameter model, while `Qwen 3` offers a 0.6B-parameter model, again with both base and instruct versions. `Teuken`, another SLM family, is available in two versions, `v0.4` and `v0.6`, which differ in the amount of pre-training data: the older `v0.4` was trained on 4T tokens, while `v0.6` was trained on 6T tokens. `v0.4` is available only as an instruct version, while `v0.6` has both base and instruct versions. The remaining 33 SLMs have sizes between 1B and 10B parameters.

At the upper end of the spectrum, only three LM families include models with $\geq$ 100B parameters. All of these models employ the Mixture-of-Experts (MoE) architecture (see Section 2.2.2). These families include `LLaMa 4`, which provides the `Scout` (109B parameters), available in base and instruct versions, as well as `Qwen 3`, which includes two 235B-parameter instruct models released at different times.

Although the largest models in this study are MoE models, `Qwen 3` also provides a smaller MoE model with 30B parameters. This model is available in a base version as well as two instruct versions, one of which is a more recent release.

MoE models differ both in the number of activated parameters and in the number of experts. `LLaMa 4 Scout` has 17B activated parameters, while the `Qwen3 30B` and 235B models have 3B and 22B activated parameters, respectively. In terms of experts, `LLaMa 4 Scout` has 16 experts, whereas `Qwen3 30B`, and `Qwen3 235B` have 128 experts each.

Overall, out of the 78 evaluated models, 71 are dense (see Section 2.2.2) language models. Among these, 47 fall into the SLM category with $\leq$ 10B parameters, while 24 are LLMs with sizes between 10B and 100B parameters. The remaining seven models are MoE LMs, with total sizes ranging from 30B to 235B parameters.

### 5.2.2 Model Versions

All models evaluated in this research are causal language models (CLMs) (see Section 2.1). Among them, 34 are base LMs, meaning that they are trained on large text datasets to learn linguistic patterns but are not specifically trained to follow instructions (see Section 2.2.1). The remaining 44 models are instruct versions, meaning that they are fine-tuned variants of their original base models. These models are designed to answer questions and follow prompts more effectively (see Section 2.2.1). Figure 5.1 presents a visual overview of base and instruct models across families and sizes.

While some LM families offer only base versions of their models, such as `Goldfish` and `TildeOpen`,

other families provide only instruct versions (for example, `Aya-expanse` and `Phi 3`). Moreover, the `Qwen 3` family, in addition to providing instruct versions for several models (the dense 4B model and the MoE 30B-A3B and 235B-A22B models), has also released newer versions of these models. Consequently, this study includes both earlier and more recent instruct model releases.

It is worth mentioning that a third category of LMs exists, namely reasoning versions. These models are fine-tuned from the original base models to perform tasks that require logical, multi-step problem-solving (see Section 2.2.1). However, this category is excluded from the study, as such complex reasoning tasks are outside its scope.

### 5.2.3 Context Lengths

In the present study, the context length of the evaluated models is largely irrelevant (see Section 2.2.4), as the minimal-pair sentences used as inputs are relatively short, typically comprising four to eight words (see Section 4.5). Consequently, all inputs comfortably fit within the context window of every model.

Nevertheless, reviewing the models by context length provides a comprehensive overview of their capabilities. The `Goldfish` family exhibits the shortest context window of 512 tokens across all four variants. Several families, including `EuroLLM`, `Phi 3`, `Neurotechnology`, `OLMo2`, and `Teuken`, feature models with a 4K-token context window. Models from `Aya-expanse`, `Salamandra/ALIA`, and `TildeOpen` support 8K tokens, whereas `EuroLLM`, `Gemma 3`, `Qwen 3`, and `Salamandra/ALIA` include models with 32K tokens. Context windows of 128K tokens are available in several families, notably `Aya-expanse`, `Phi 3`, `Gemma 3`, and `LLaMa 3.1-3.3`.

The Mixture-of-Experts (MoE) models provide the longest context windows. Specifically, `Qwen 3` models accommodate 256K tokens, and `LLaMa Scout` 10M tokens. For a complete overview of context lengths across all individual models, refer to Table B.1 in the Appendix.

### 5.2.4 Language Coverage

The evaluated models differ in their scope of language support. As mentioned in Section 5.1, seven LM families explicitly report Lithuanian among their supported languages. In addition, `Gemma 3` is described as covering more than 140 languages worldwide; however, the full list of supported languages is not publicly available. In this study, Lithuanian is assumed to be included, based on an inspection of the tokenizer, which shows that the character *ė* (unique to Lithuanian) exists as a single token (see Appendix B.1). While this does not guarantee strong model performance on Lithuanian text, it suggests that the character appeared in the training data.

Consequently, eight of the thirteen LM families are considered to include Lithuanian, corresponding to 51 of the 80 evaluated models. A complete list of individual models and their reported language coverage is provided in Appendix Table B.1.

Two families, `Neurotechnology` and `Goldfish`, consist of monolingual models, resulting in eight monolingual systems overall. Several families, namely `EuroLLM`, `Salamandra/ALIA`, `Teuken`, and `TildeOpen`, focus primarily on European languages, including Lithuanian, comprising a total of sixteen models. In contrast, twenty-seven models that support Lithuanian have a broader, global scope, belonging to the `Gemma 3` and `Qwen 3` families.

The remaining five LM families do not officially report Lithuanian support. They are included due to their frequent use in related evaluation tasks or their popularity in the field, aiming to provide

a comprehensive comparison (see Sections 2.3.3 and 2.3.3). These families are `Aya-expanse`, `Phi 3`, `LLaMa 3.1-3.3`, `LLaMa 4`, and `OLMo2`.

## 5.3 Evaluation Process

All evaluations were conducted on Snellius[1] using NVIDIA H100 GPUs[2]. The models used in this study are publicly available on Hugging Face[3], although many are gated and require authentication. The evaluation script therefore authenticates using a Hugging Face access token prior to loading the models.

The script takes as input a CSV file containing sentence pairs and allows selection of the computational device. During initial testing on a personal laptop, Apple MPS was prioritised. On Snellius, however, the script prioritises CUDA-enabled GPUs, using the CPU only as a fallback. Although a primary device is selected for tensor operations, the model itself is loaded with `device_map="auto"`, which automatically distributes its components across available GPU and CPU resources and helps prevent memory overflow. After device selection, the tokenizer and the chosen causal language model are loaded. Additionally, the script uses `torch_dtype=torch.float16`, meaning that computations are performed using 16-bit floating point precision. This reduces memory usage compared to 32-bit precision and allows larger models to be processed.

The evaluation proceeds on a sentence-by-sentence basis rather than in batches to simplify the evaluation procedure. For each sentence, the script computes the negative log-likelihood (NLL) under a standard causal language modeling objective (Jurafsky & Martin, 2026). A causal language model factorises the probability of a token sequence $w_{1:n}$ as:

$$P(w_{1:n}) = P(w_1)P(w_2 \mid w_1)P(w_3 \mid w_{1:2})...P(w_n \mid w_{1:n-1}) = \prod_{i=1}^{n} P(w_i \mid w_{<i}).$$

The corresponding negative log-likelihood is defined as:

$$\text{NLL}(w_{1:n}) = -\log\left(\prod_{i=1}^{n} P(w_i \mid w_{<i})\right) = -\sum_{i=1}^{n} \log P(w_i \mid w_{<i}).$$

In practice, the model conditions on each preceding token in the sequence to predict the next token, and the loss is computed internally by the `transformers` library using the input sequence as labels. The resulting loss reflects how unlikely the sentence is according to the model, with lower NLL values corresponding to higher probabilities. To avoid bias due to differences in sentence length, the implementation uses the mean NLL per token rather than the total NLL:

$$\text{mean NLL}(w_{1:n}) = -\frac{1}{n}\sum_{i=1}^{n} \log P(w_i \mid w_{<i}).$$

For each sentence pair, the script computes the mean NLL for both the grammatical and ungrammatical variants. It considers the model's decision correct if the grammatical sentence receives a lower

[1] https://www.surf.nl/en/services/compute/snellius-the-national-supercomputer
[2] https://www.nvidia.com/en-us/data-center/h100/
[3] https://huggingface.co

mean NLL than the ungrammatical one:

$$\hat{y} = \begin{cases} \text{correct} & \text{if mean NLL}(w_{\text{grammatical}}) < \text{mean NLL}(w_{\text{ungrammatical}}), \\ \text{incorrect} & \text{otherwise.} \end{cases}$$

The script also computes a certainty score as the difference between the mean NLL values of the incorrect and correct sentences, indicating how strongly the model prefers one sentence over the other:

$$\text{certainty} = \text{mean NLL}(w_{\text{ungrammatical}}) - \text{mean NLL}(w_{\text{grammatical}}).$$

Finally, the script appends the results (mean NLL scores, certainty, and the model's decision) to new columns in the dataset and saves the updated data to a new CSV file. It also generates a summary report that includes overall accuracy, the number of correct and incorrect decisions, and accuracy broken down by linguistic phenomena and error types:

$$\text{accuracy} = \frac{\#\text{correct decisions}}{\#\text{total sentence pairs}}.$$

The script computes this summary both for the full dataset and for a filtered version excluding rows marked as exceptions. As the two summaries show minimal variation, this thesis reports only the full results (including exceptions), while the complete outputs are made available on GitHub[4].

[4] https://github.com/urtuteja/Evaluating-The-Syntactic-Knowledge-of-Language-Models-on-Lithuanian

# 6 Human Acceptability Ratings

In this chapter, I describe in detail how the human acceptability ratings were collected, including the experimental procedure and the characteristics of the survey, as well as how it was administered. Moreover, I provide information about the participants and the experimental items. Finally, I present the results obtained from the survey.

## 6.1 Experimental Design and Procedure

The survey was conducted using Qualtrics[1]. Participants first received an information sheet describing the study and were asked to provide their informed consent. Those who declined were automatically redirected to the end of the survey. Participants who consented then completed six background questions:

1. What is your age (in years)?
2. Do you currently reside in Lithuania? If not, how many years have you been living abroad?
3. Do you have any reading difficulties?
4. Is Lithuanian your only native language? If not, please list all your native languages.
5. What is the highest level of education you have completed?
6. Have you studied linguistics or any related field? If yes, please specify the field(-s).

All questions were mandatory, so participants could not proceed to the survey without providing responses.

After completing the background questions, participants were randomly assigned to one of two survey groups corresponding to different sets of error types (see Section 6.3). Within each group, participants were further randomly directed to one of six survey versions, resulting in a total of 12 possible surveys.

The Qualtrics randomization mechanism counted surveys as "presented" even if participants did not finish them. To ensure that each survey version received an approximately equal number of completed responses, survey counts were monitored and adjusted as needed during data collection.

Once in the survey, participants received brief instructions explaining the task. They were asked to rate the grammatical acceptability of each sentence on a 7-point Likert scale (see Section 2.4), where 1 indicated a completely unacceptable sentence and 7 indicated a fully acceptable sentence. Participants

[1] https://www.qualtrics.com

were instructed to rely on their intuitive judgment, without considering word choice or style, and without consulting any external sources. They were also informed that there were no right or wrong answers. The complete instructions provided to participants are included in Appendix C.2.

All sliders had to be adjusted before proceeding, as the `Next` button only appeared once every sentence had been rated. The 32 sentences were presented in randomized order, and each slider was initially light grey, turning dark grey when moved to indicate which sentences had been rated. These interface features, including slider colors and next-button activation, were implemented using JavaScript.

Finally, participants had the option to leave optional comments. One participant reported the Likert scale as redundant, rating all sentences as either 1 or 7. Otherwise, participants indicated that the survey was clear and easy to complete.

## 6.2 Participants

To collect the human acceptability ratings, a total of 123 participants were recruited. Of these, 120 responses were considered valid and suitable for further analysis. Two participants were automatically directed to the end of the survey because they did not agree to participate. Additionally, one participant reported having reading difficulties and was therefore excluded from the analysis.

The participants vary in age, with the youngest being 18 years old and the oldest 81 years old (M = 43.26, median = 42). The distribution of participant ages is shown in Figure 6.1.

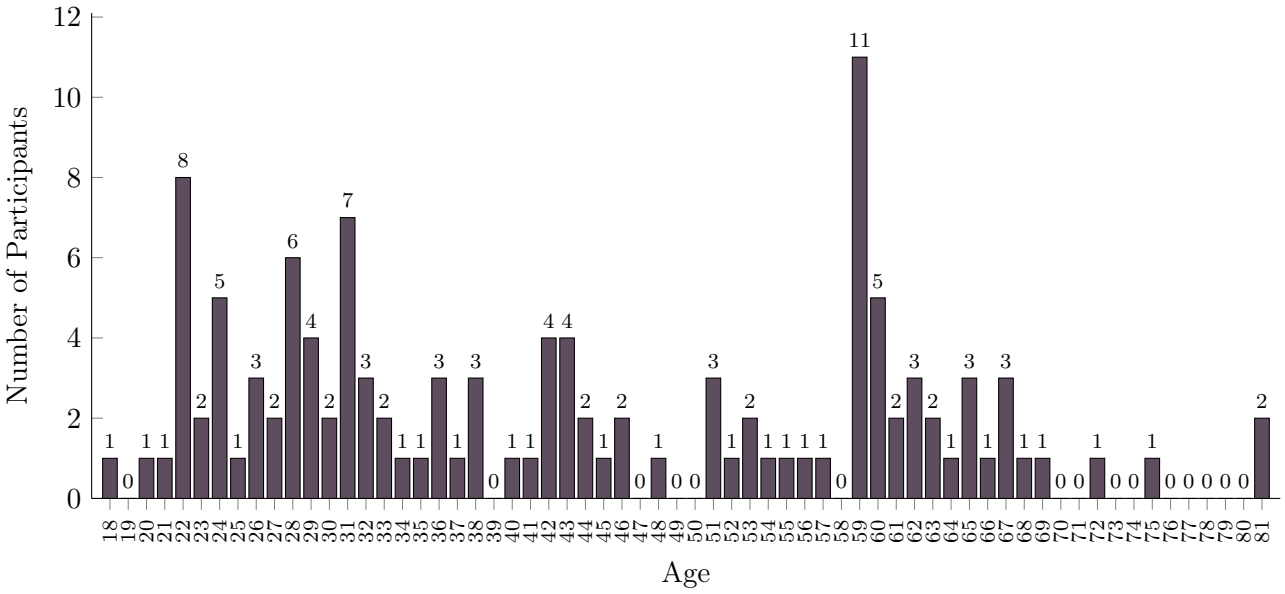


Figure 6.1: Number of participants by age.

Out of the 120 participants included in the analysis, 111 live in Lithuania, while the remaining nine live abroad. The duration of residence abroad ranges from 1 to 19 years (1, 3, 5, 7, 12, 16, 16, 17, and 19 years; M = 10.67, median = 12).

Most participants reported Lithuanian as their only native language. Specifically, 115 participants identified as monolingual Lithuanian speakers, while five participants reported having multiple native languages, including English, Russian, and Polish.

Participants were also asked to indicate their level of education. The distribution is skewed toward higher levels of education: 92 participants hold either an undergraduate or a graduate degree. The full

distribution of education levels is presented in Figure 6.2. For an overview of the education system in Lithuania, see Eurydice, which provides comprehensive information on European education systems[2].

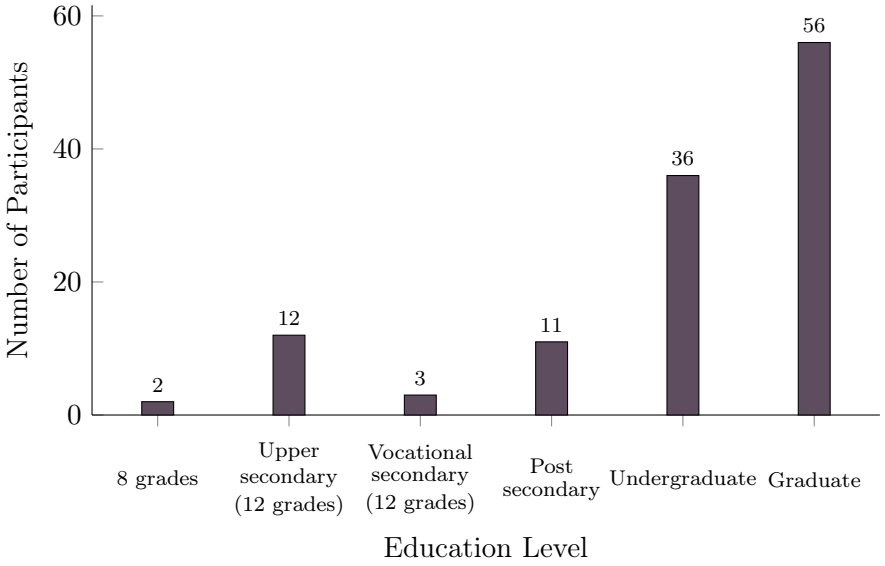


Figure 6.2: Number of participants by education level.

Finally, participants were asked whether they have studied linguistics or related fields. While 113 participants indicated that they have not, seven participants reported familiarity with fields related to linguistics, including philology (Lithuanian, Italian, and English), journalism, and communication science.

Participants were recruited through personal contacts and were encouraged to further distribute the survey to others. Additional participants were recruited through Facebook groups dedicated to sharing online surveys.[3]

## 6.3 Experimental Materials

Testing all 777 minimal sentence pairs was not feasible, as it would have made the survey too long or required a very large number of participants. The survey was designed to take approximately 10 minutes, since recent studies suggest that this is an optimal length for most types of surveys. Moreover, surveys longer than 20 minutes can be exhausting for participants and may even lead them to skip questions (Revilla & Höhne, 2020; Revilla & Ochoa, 2017). If the survey had been kept short while still testing all sentence pairs, it would have required 486 participants, which was considered too many to recruit, particularly given the logistical constraints associated with conducting participant recruitment between the Netherlands and Lithuania.

Therefore, for the survey, sentence pairs from the dataset were systematically sampled to obtain a smaller but still representative subset. Three minimal sentence pairs were randomly selected for each error type, resulting in a total of 192 sentence pairs, as the dataset contains 64 different error types (see Section 4.4.2). Randomly selecting three pairs per error type ensured coverage of the full range of errors in the dataset.

[2] https://eurydice.eacea.ec.europa.eu/eurypedia/lithuania/overview

[3] Three Facebook groups were used to recruit participants: https://www.facebook.com/groups/376071086148311/, https://www.facebook.com/groups/apklausos/, and https://www.facebook.com/groups/339591789517219/.

The 192 sentence pairs were distributed among participants to keep the survey relatively short (approximately 10 minutes) and maintain participant engagement. To achieve this, the sentence pairs were divided across several survey versions. This approach follows a design used in previous studies with similar experimental setups (Pestel et al., 2025; Warstadt, Parrish, et al., 2019).

First, the sentence pairs were divided into two sets based on error type. Consequently, half of the participants evaluated sentence pairs representing 32 error types, while the other half evaluated sentence pairs representing the remaining 32 error types. Each set was then further divided into six survey versions, resulting in a total of 12 survey versions in the experiment.

Each minimal pair consists of two sentences: a grammatical sentence and an ungrammatical sentence. Within each survey, participants evaluated only one sentence from each minimal pair. The sentences were distributed so that participants alternated between grammatical and ungrammatical examples across error types. In addition, the three sentence pairs associated with each error type were rotated across the different survey versions so that all pairs were eventually evaluated. A detailed overview of the sentence distribution across survey versions is provided in Appendix E.2, Table C.1.

Since the experiment included 120 participants and 12 survey versions, each survey version was completed by 10 participants. As a result, each minimal sentence pair was evaluated 20 times in total, and each individual sentence (grammatical or ungrammatical) was evaluated 10 times.

## 6.4 Results

Overall, grammatical sentences received substantially higher acceptability ratings than ungrammatical sentences based on 192 sentence pairs (M = 4.90, median = 5.20, SD = 1.72). In contrast, ungrammatical sentences were rated lower (M = 2.84, median = 2.48, SD = 1.67). Despite these differences, both grammatical and ungrammatical sentences exhibited notable variability: some grammatical sentences were rated as low as 1 (*totally unacceptable*), whereas some ungrammatical sentences were rated as high as 7 (*totally acceptable*).

To test whether this difference holds statistically while accounting for variation across participants and items, a linear mixed-effects model was fitted in R[4] using Posit Cloud[5]. The model included grammaticality as a fixed effect and random intercepts for both participants and items, with ungrammatical sentences serving as the reference (baseline) category. The analysis confirmed a robust effect of grammaticality: grammatical sentences received significantly higher ratings than ungrammatical sentences ($\beta = 2.07$, SE = 0.13, $p < 0.001$). The estimated baseline rating was 2.84 (SE = 0.11, $p < 0.001$). Random effects indicated substantial variability across sentences (SD = 1.12) and participants (SD = 0.70), as well as residual variability (SD = 1.64), suggesting that both sentence- and participant-level differences contributed meaningfully to rating variation.

### 6.4.1 Acceptability per Phenomenon

Figure 6.3 presents human acceptability ratings for each linguistic phenomenon. As shown in the figure, for the majority of phenomena (17 out of 18), grammatical sentences were rated higher than ungrammatical ones. An exception was phenomenon *5.2 Abu, abi*, where ungrammatical sentences

[4] https://www.r-project.org
[5] https://posit.cloud

received higher ratings (M = 4.27) than grammatical sentences (M = 3.73), indicating that participants found this phenomenon particularly challenging.

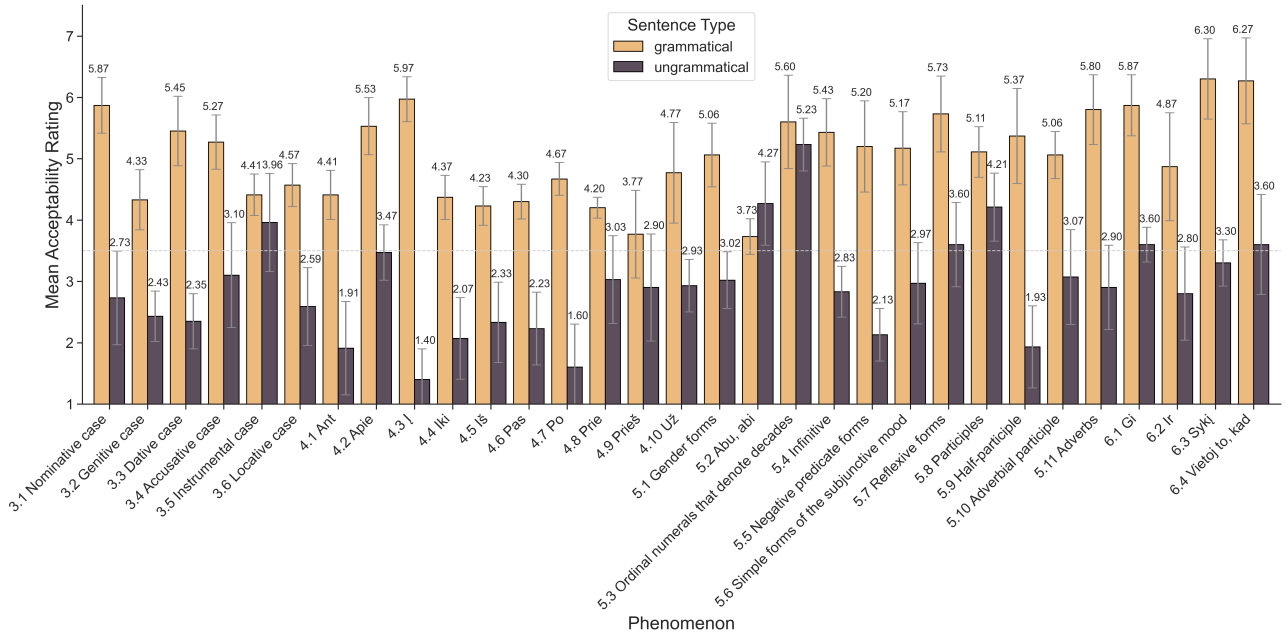


Figure 6.3: Mean acceptability ratings for each linguistic phenomenon. A horizontal dashed line at 3.5 indicates the neutral midpoint of the rating scale, and the darker grey vertical lines on each bar represent 95% confidence intervals.

Although grammatical sentences were generally rated above the midpoint of the Likert scale, seven of the 18 ungrammatical sentences also received mean ratings above the midpoint. This indicates that, while grammatical sentences were typically preferred, participants were sometimes reluctant to categorize certain ungrammatical sentences as clearly unacceptable.

These patterns are consistent with the 95% confidence intervals shown in Figure 6.3, which generally overlap more when grammatical and ungrammatical mean ratings are close and show clearer separation when larger differences between conditions are observed. A similar pattern is observed in the subsequent analysis by error type, where the overall trends remain consistent across conditions.

The highest rating for ungrammatical sentences occurred for phenomenon *5.3 Ordinal numerals that denote decades*, with an average score of 5.23, compared to 5.60 for grammatical sentences. This indicates that both grammatical and ungrammatical versions were perceived as relatively acceptable. In contrast, the phenomenon perceived as easiest by participants was *4.3 Į*, where the distinction between grammatical and ungrammatical sentences was most pronounced: grammatical sentences received an average rating of 5.97, while ungrammatical sentences received an average of 1.40.

### 6.4.2 Acceptability per Error Type

In this section, I present human acceptability ratings for each error type. As the dataset comprises 64 error types (see Section 4.4.2), the results are presented by grouping them into four broader categories by phenomenon type: *Use of Cases*, *Use of Prepositions*, *Use of Forms*, and *Coordination of Sentence Elements and Clauses*.

### Phenomenon Type 1: Use of Cases

Among the 22 error types associated with the phenomenon type *Use of Cases*, two stand out in that ungrammatical sentences were rated higher than their grammatical counterparts. These are *3.5.2 Instrumental case should not be used to express content of quality with adjectives denoting abundance* and *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* (see Figure 6.4).

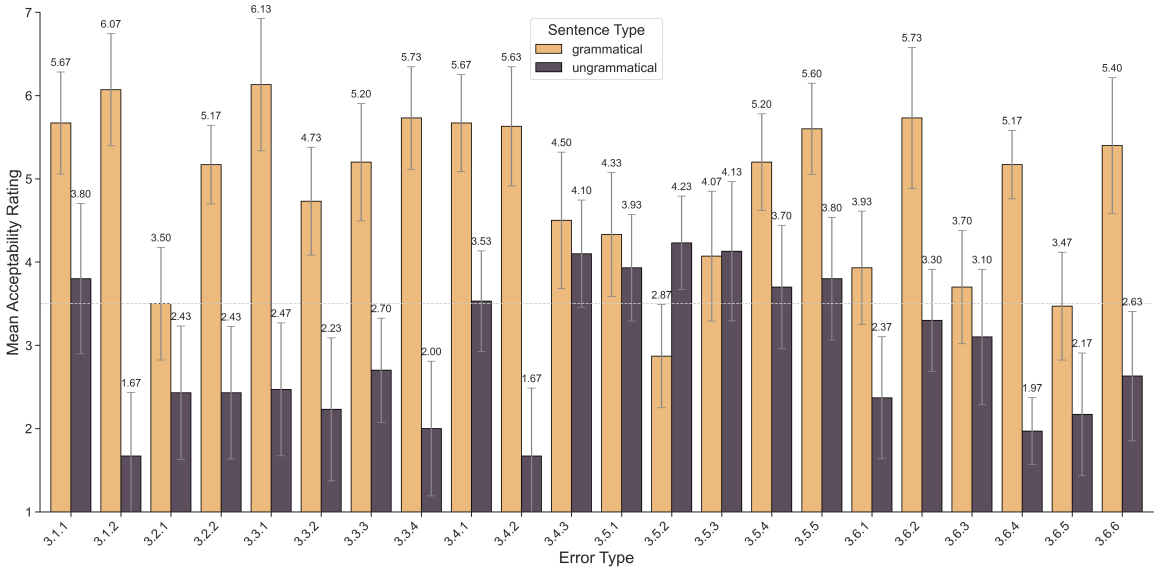


Figure 6.4: Mean acceptability ratings for each error type in *Use of Cases* phenomenon type. A horizontal dashed line at 3.5 indicates the neutral midpoint of the rating scale, and the darker grey vertical lines on each bar represent 95% confidence intervals.

Furthermore, for two additional error types, *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* and *3.5.1 Instrumental case should not be used to express the object with verbs denoting fullness or increase*, grammatical sentences were only slightly higher rated than ungrammatical ones (4.50 vs. 4.10 and 4.33 vs. 3.93, respectively).

The results also indicate that the instrumental case posed the greatest difficulty for participants. All five error types associated with this case (*3.5.1–3.5.5*) have ungrammatical sentences with mean ratings above the midpoint of 3.5, suggesting that these sentences were generally perceived as more acceptable than unacceptable.

A similar pattern, though less pronounced, can be observed for two other error types: *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* and *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)*. In these cases, ungrammatical sentences were also rated slightly above the midpoint (3.80 and 3.53, respectively). However, their grammatical counterparts received substantially higher ratings (5.67 in both cases), indicating a clearer distinction between the two sentence types.

Finally, it is noteworthy that two error types exhibit relatively low acceptability even for grammatical sentences. Specifically, *3.2.1 Genitive case should not be used to denote the object of an action with certain verbs* received a mean rating of exactly 3.5, while *3.6.5 Locative case should not be used to*

*express the manner or timing of an action* received a mean rating of 3.47.

### Phenomenon Type 2: Use of Prepositions

The 20 error types associated with the phenomenon type *Use of Prepositions* appear to have been less challenging for participants overall (see Figure 6.5). Only one error type, *4.8.2 Prie should not be used to express exchange ratios or conversion rates*, showed no distinction between grammatical and ungrammatical sentences, with both receiving an identical mean rating of 2.87. This indicates that participants perceived both sentence types as relatively unacceptable.

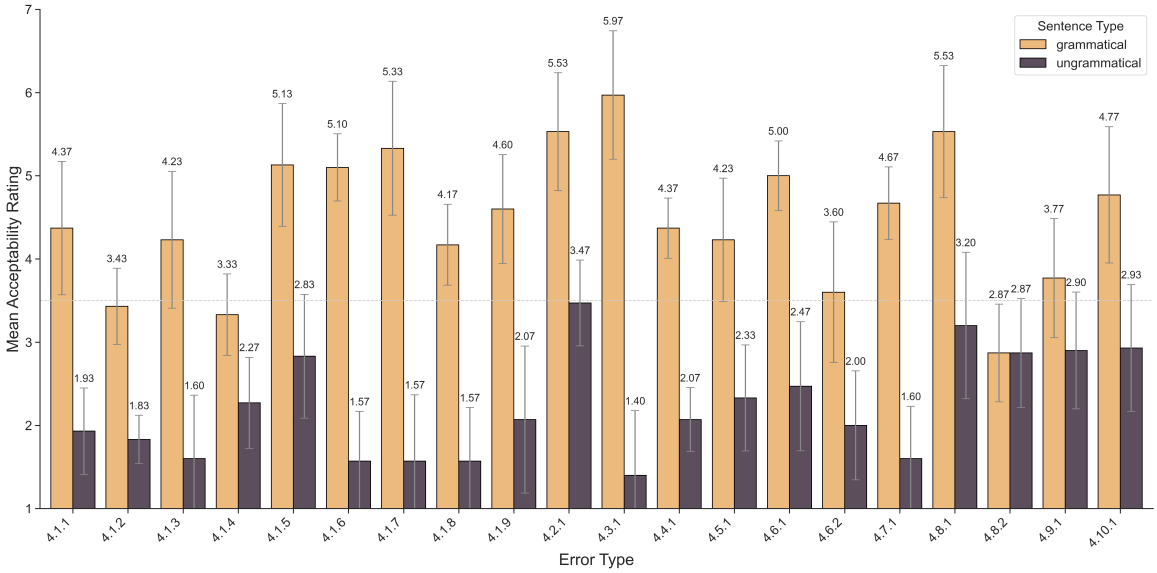


Figure 6.5: Mean acceptability ratings for each error type in *Use of Prepositions* phenomenon type. A horizontal dashed line at 3.5 indicates the neutral midpoint of the rating scale, and the darker grey vertical lines on each bar represent 95% confidence intervals.

Another comparatively challenging error type was *4.2.1 Apie should not be used to express the object of an action with certain words*, where ungrammatical sentences were rated close to the midpoint of the scale (3.47), suggesting some uncertainty in participants' judgments.

However, for several error types in this group, participants demonstrated a clear distinction between grammatical and ungrammatical sentences. The most prominent example is *4.3.1 Į should not be used to express the object of an action with certain words*, where grammatical sentences received a high mean rating of 5.97, while ungrammatical sentences were rated much lower, at 1.40.

### Phenomenon Type 3: Use of Forms

A total of 18 error types fall under the phenomenon type *Use of Forms*. Overall, these errors appear to have been challenging for participants. In particular, *5.2.1 The basic quantitative numeral forms from 2 to 9 and the pronoun abu, abi are not to be used with plural-only nouns (nouns that exist only in the plural)* (already discussed in Section 6.4.1) and *5.8.1 Active present participle should not be used with nouns that do not denote an agent (nouns that cannot perform an action)* received higher ratings for ungrammatical sentences than for grammatical ones (see Figure 6.6).

In total, eight error types had mean ratings for ungrammatical sentences above the neutral midpoint of 3.5, indicating that participants tended to perceive these sentences as relatively acceptable. Moreover,

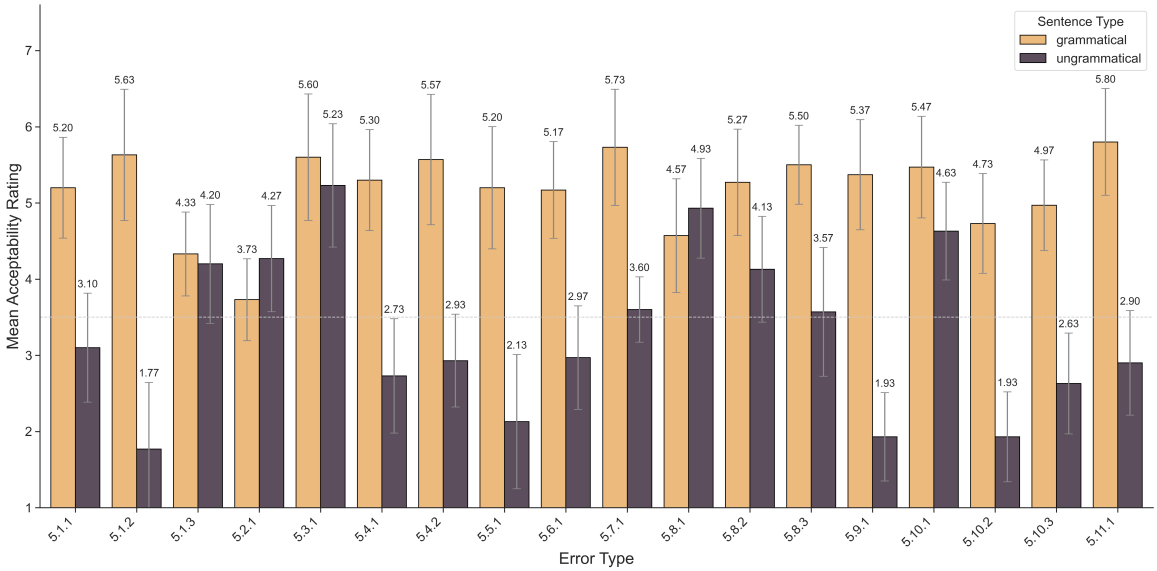


Figure 6.6: Mean acceptability ratings for each error type in *Use of Forms* phenomenon type. A horizontal dashed line at 3.5 indicates the neutral midpoint of the rating scale, and the darker grey vertical lines on each bar represent 95% confidence intervals.

two error types, *5.1.3 The masculine accusative and instrumental forms of adjectives and of participles used with adjectival meaning are not to be used in the so-called "indefinite gender" meaning* and *5.3.1 The pronominal forms (the forms with possessive endings) of ordinal numerals that denote decades are not to be used to indicate decades*, showed very similar ratings for grammatical and ungrammatical sentences, suggesting difficulty in distinguishing between the two.

While most problematic error types correspond to a single, distinct phenomenon (for example, *5.2.1* and *5.3.1*), the three error types *5.8.1–5.8.3* all fall under the broader category of *5.8 Participles*. This pattern highlights that not only individual error types, but the entire phenomenon of participles, poses challenges for native speakers of Lithuanian.

### Phenomenon Type 4: Coordination of Sentence Elements and Clauses

The final phenomenon type, *Coordination of Sentence Elements and Clauses*, comprises only four error types, each corresponding to a distinct phenomenon (see Figure 6.7). Although, in all cases, grammatical sentences were rated higher than ungrammatical ones, the primary challenge is reflected in the relatively high ratings assigned to ungrammatical sentences, many of which cluster around the midpoint of the Likert scale (3.5).

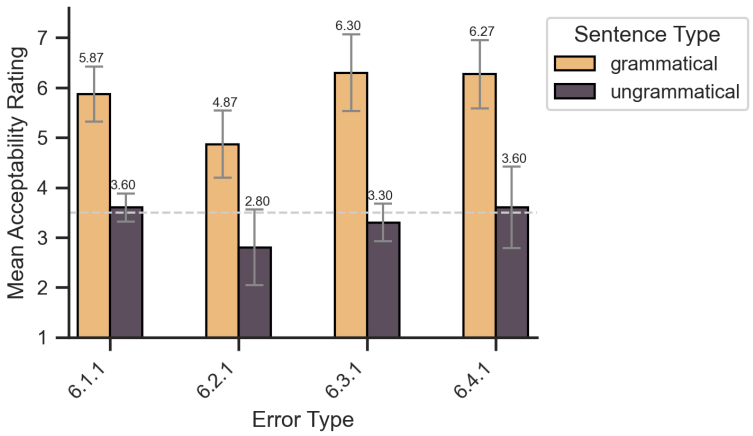


Figure 6.7: Mean acceptability ratings for each error type in *Coordination of Sentence Elements and Clauses* phenomenon type. A horizontal dashed line at 3.5 indicates the neutral midpoint of the rating scale, and the darker grey vertical lines on each bar represent 95% confidence intervals.

In particular, ungrammatical sentences for *6.1.1 Gi must not be used instead of a coordinating conjunction expressing contrast* and *6.4.1 Vietoj to, kad should not be used with an infinitive or participle in place of užuot* both received a mean rating of 3.6. This suggests that participants perceived these sentences as largely neutral, or even slightly acceptable.

The remaining two error types, *6.2.1 Ir should not be used to link words in intensifying constructions with adjectives and adverbs* and *6.3.1 Sykį should not be used to introduce subordinate conditional clauses*, received somewhat lower ratings for ungrammatical sentences (2.8 and 3.3, respectively). While these scores fall below the midpoint, they are still relatively high, indicating that participants were not consistently confident in judging these sentences as unacceptable.

# 7 Results

In this chapter, I present the results of the study. I begin by addressing the first research question, which investigates the extent to which language models (LMs) correctly discriminate between correct and incorrect sentences. To do this, I examine the accuracies of the 78 evaluated models. I present the overall accuracies, as well as accuracies by phenomenon and error type. These results are accompanied by 95% confidence intervals. In addition, I rank the models from best-performing to worst-performing based on their overall accuracies.

Next, I present the results related to the following three research questions: the effects of global/European/monolingual scope, model size, and model version. These questions are examined using a logistic mixed-effects model.

I then compare the results of the models and human participants, focusing specifically on their correlation. In this section, I present the results of the Spearman correlation test (Spearman, 1904) overall, followed by separate results for each phenomenon type (1-4).

Finally, I provide a brief summary of the main findings.

## 7.1 Overview of Model Performance

Overall, the evaluated models exhibited substantial variation in performance, with accuracies ranging from 0.47 to 0.816 (see Figure 7.1). While the highest-performing models demonstrated a strong ability to discriminate between correct and incorrect items, some models performed close to or slightly below chance level, which, for this binary-choice task, was 50%.

The best-performing models overall were the four `Neurotechnology` models (7B and 13B, base and instruct variants), which are monolingual Lithuanian models fine-tuned on `LLaMA 2` (Meta GenAI, 2023). These models achieved accuracies ranging from 0.773 to 0.816.

In contrast, the lowest-performing model was the European-scope `TildeOpen` model (30B parameters), which achieved an accuracy below chance level (0.47). Two other low-ranked models were the smallest instruct variants from the `Gemma 3` family, with 270M and 1B parameters, achieving accuracies of 0.525 and 0.515, respectively.

The monolingual `Goldfish` model family, consisting of four small language models (with 39M and 125M parameters and training data ranging from 5MB to 1000MB), showed highly variable performance. The model trained on the largest amount of data ranked seventh overall, placing behind the monolingual `Neurotechnology` models and two large-scale MoE models from the `LLaMA 4` and `Qwen3` families. In contrast, the model trained on the smallest dataset ranked fourth from the bottom.

In addition to reporting accuracies, Wilson 95% confidence intervals (CIs) (Wilson, 1927) were calculated to assess the precision and reliability of the observed accuracy scores. These results are also presented in Figure 7.1, where the models are ranked by overall accuracy, with the best-performing

models shown at the top. The dark purple lines represent the confidence intervals, while the dots at their centres indicate the corresponding accuracy scores. The numbers to the left of each line show the accuracy of each model, while the column to the right of the plot displays the CI widths, which range from 0.054 to 0.07. Considering the sample size of 777 sentence pairs, these relatively narrow confidence intervals suggest that the estimated accuracies are reasonably stable and only moderately sensitive to sampling variability.

## 7.2 Model Performance by Linguistic Phenomenon

This study investigates 31 linguistic phenomena, and the results for all evaluated models are broken down by phenomenon in Figure 7.2. The figure shows that the five easiest phenomena for the models all belong to Phenomenon Types 3 and 4, namely *Use of Forms* and *Coordination of Sentence Elements and Clauses*, with accuracies ranging from 0.743 to 0.944. In contrast, the five most challenging phenomena come from three Phenomenon Type groups: 2, 3, and 4, namely *Use of Prepositions*, *Use of Forms*, and *Coordination of Sentence Elements and Clauses*, with accuracies ranging from 0.300 to 0.437, all below chance level.

These findings suggest that, unlike Phenomenon Type 1, *Use of Cases*, which contains phenomena of moderate difficulty relative to the other groups, phenomena from the *Use of Prepositions* group tend to range from moderately difficult to very difficult for the models. By contrast, the other two groups, *Use of Forms* and *Coordination of Sentence Elements and Clauses*, are less uniform, containing both phenomena that the models handled relatively easily and others that proved extremely challenging. A summary of the easiest and most difficult phenomena is provided in Table 7.1.

Moreover, after examining the confidence intervals (CIs) per phenomenon, no outliers were observed. The CI widths ranged from 0.104 for phenomena with the largest number of sentence pairs (110 items), up to 0.527 for very small subsets containing as few as 10 items. As wider confidence intervals are expected for small datasets, these results are consistent with statistical expectations and do not indicate any anomalies. The full results can be found in Appendix D.2, Figure D.2.

| **Phenomenon Code** | **Phenomenon Description** | **Mean Accuracy** |
|---|---|---|
| *Easiest Phenomena* | | |
| 6.1. | *Gi* | 0.944 |
| 5.6. | Simple forms of the subjunctive mood | 0.902 |
| 5.9. | Participles | 0.841 |
| 5.1. | Gender forms | 0.777 |
| 6.3. | *Sykį* | 0.743 |
| *Most Challenging Phenomena* | | |
| 6.2. | *Ir* | 0.300 |
| 4.9. | *Prieš* | 0.302 |
| 5.2. | *Abu, abi* | 0.319 |
| 4.3. | *Į* | 0.398 |
| 4.5. | *Iš* | 0.437 |

Table 7.1: Five easiest and five most challenging phenomena for the models based on mean accuracy across all 78 models.

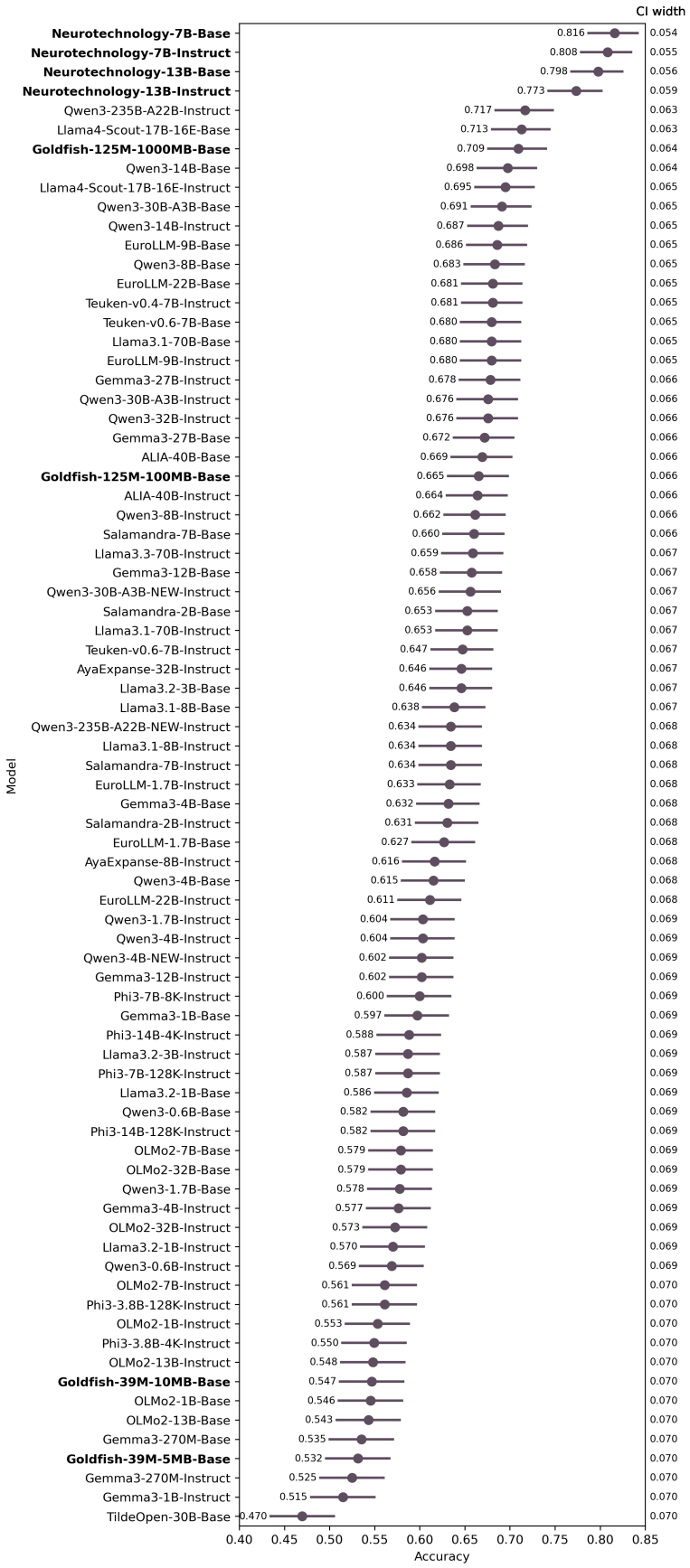


Figure 7.1: Overall accuracies across models, with accuracy scores marked by dots and corresponding Wilson 95% confidence intervals shown as dark purple lines. The numbers to the left of the lines indicate the accuracies, while the column on the right shows the widths of the confidence intervals. Models are ranked by overall accuracy, and monolingual models are highlighted in bold.

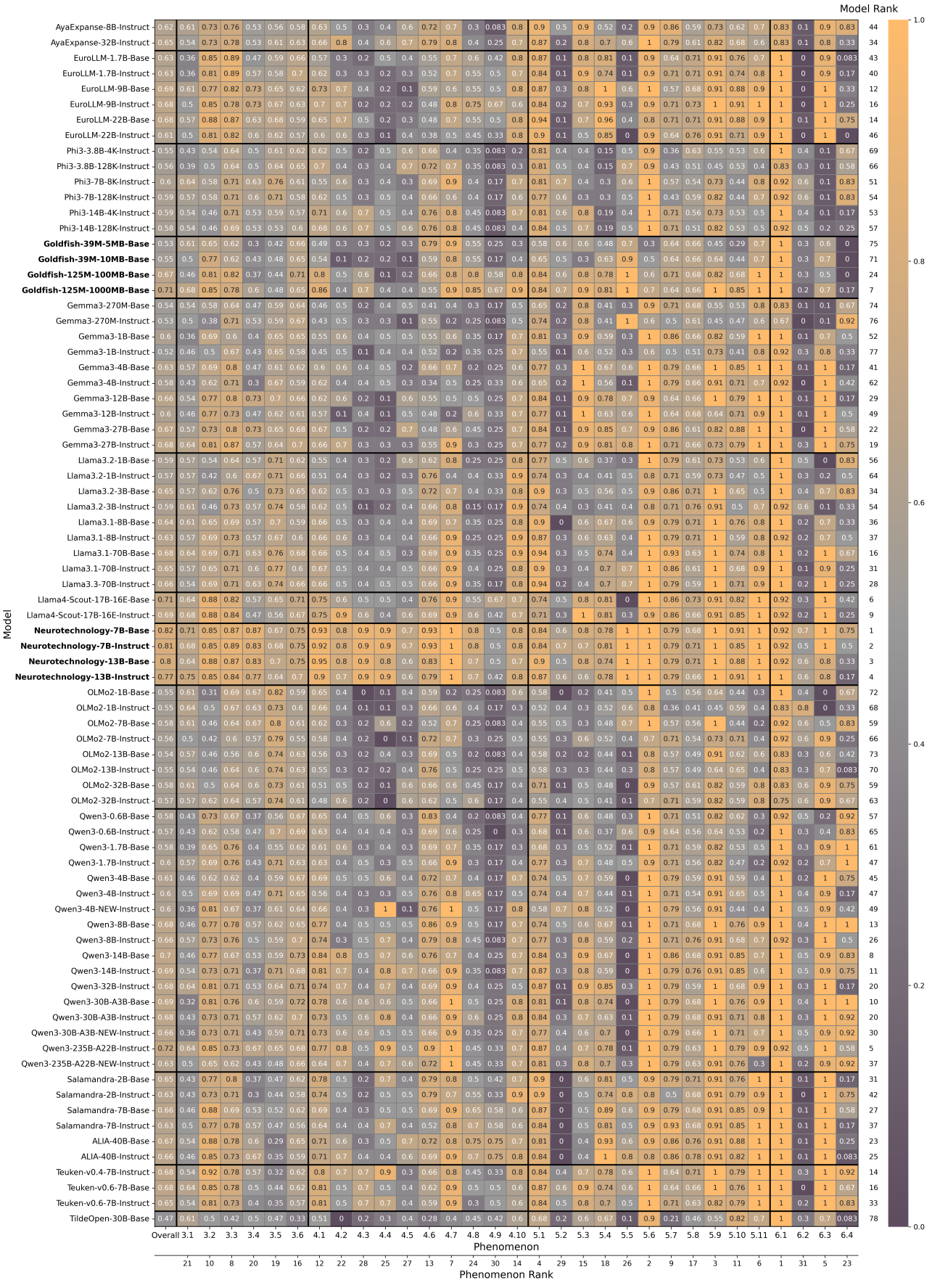


Figure 7.2: Overall and per-phenomenon accuracies across models. Monolingual models are marked in bold. Horizontal bold lines separate different model families, while vertical bold lines separate phenomenon types.

## 7.3 Model Performance by Error Type

While Section 7.2 presents model accuracies across the 31 linguistic phenomena investigated in this study, this section further breaks the results down into 64 error types. For clearer visualization, the results are not presented all at once but are instead divided into four groups based on phenomenon type. Furthermore, Table 7.2 summarizes the five easiest and five most challenging error types overall.

Moreover, the confidence intervals (CIs) per error type align with statistical expectations given the dataset sizes (0.194-0.527 for Phenomenon Types 1 and 2, 0.204-0.527 for Phenomenon Type 3, and 0.242-0.527 for Phenomenon Type 4). The full results are provided in Appendix D.2, Figures D.3-D.6.

| **Error Code** | **Error Description** | **Mean Accuracy** |
|---|---|---|
| *Easiest Error Types* | | |
| 6.1.1 | *Gi* should not replace contrastive coordinating conjunctions. | 0.944 |
| 5.6.1 | Simple subjunctive forms should not replace compound forms for past hypothetical actions. | 0.902 |
| 4.1.3 | *Ant* should not be used to express object of action with verbs. | 0.876 |
| 3.3.1 | Dative case should not be used to denote the object of an action with certain expressions. | 0.872 |
| 3.5.4 | Instrumental case should not be used with forms of the verb *būti* to express a permanent state. | 0.866 |
| *Most Challenging Error Types* | | |
| 6.2.1 | *Ir* should not be used in intensifying adjective/adverb constructions. | 0.300 |
| 4.9.1 | *Prieš* should not express object of action with certain words. | 0.302 |
| 5.2.1 | Numerals 2–9 and *abu, abi* should not be used with plural-only nouns. | 0.319 |
| 3.1.2 | Nominative case should not be used to express direct address. | 0.364 |
| 4.8.2 | *Prie* should not express exchange ratios or conversion rates. | 0.375 |

Table 7.2: Five easiest and five most challenging error types for the models based on mean accuracy across all 78 models.

### 7.3.1 Phenomenon Type 1: Use of Cases

A closer look at error types within Phenomenon Type 1, *Use of Cases*, (see Figure 7.3) shows that the easiest errors are spread across phenomena, including *3.3.1 Dative case should not be used to denote the object of an action with certain expressions* and *3.3.4 Dative should not be used with verbs of motion to express purpose* (phenomenon 3), *3.5.4 Instrumental case should not be used with forms of the verb būti to express a permanent (unchanging) state* (phenomenon 5), and *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.* (phenomenon 6). The most difficult error types are similarly distributed, including *3.1.2 Nominative case should not be used to express direct address* (phenomenon 1), *3.6.6 Locative case should not be used to express a time period as a preposition or postposition* (phenomenon 6), *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)* (phenomenon 4), and *3.3.3 Dative case should not be used for indicating a specific time limit or moment*

*when purpose is not being expressed* (phenomenon 3). Overall, the mix of easier and more challenging error types within Phenomenon Type 1 explains the mid-range performance at the phenomenon level.

What stands out, however, are differences between models. The best-performing `Neurotechnology` models reach 0.64-0.73 accuracy on the hardest error type (*3.1.2 Nominative case should not be used to express direct address*), while other families such as `EuroLLM`, `Gemma3`, `OLMo2`, and `Qwen3` drop to 0.091. `Neurotechnology` also clearly outperforms others on *3.4.2 Accusative case should not be used to express the direct object next to a negative verb (instead of the genitive)* (1.0 vs. lows of 0.2). In contrast, for *3.5.2 Instrumental case should not be used to express content of quality with adjectives denoting abundance* it is near chance (0.45-0.55), while `OLMo2` reaches 1.0 and `LLaMA 3` ranges between 0.82 and 1.

Another notable case is *3.6.6 Locative case should not be used to express a time period as a preposition or postposition*, the second-lowest in this group, where only the `Goldfish` family reaches 0.8-1.0, while others perform worse; `Neurotechnology` is below chance at around 0.4. Therefore, these cases warrant further investigation in Error Analysis 8.2.

### 7.3.2 Phenomenon Type 2: Use of Prepositions

Phenomenon Type 2, *Use of Prepositions*, includes phenomena ranging from mid to high difficulty for the models. As shown in Figure 7.4, the easiest phenomenon overall, ranked $12^{th}$ out of 31, is phenomenon *4.1 Ant*. It consists of nine error types, four of which are among the easiest in this group, while the rest show mid-level difficulty.

Most other phenomena in this group (*4.2 Apie*, *4.3 Į*, *4.4 Iki*, *4.5 Iš*, *4.7 Po*, *4.9 Prieš*, and *4.10 Už*) contain only one error type each. Among these, *4.7.1 Po (with the instrumental case) should not be used to express the mode or basis of an action* (phenomenon 7) is among the easiest error types, while *4.9.1 Prieš should not be used to express the object of an action with certain words* (phenomenon 9), *4.3.1 Į should not be used to express the object of an action with certain words* (phenomenon 3), *4.5.1 Iš should not be used with certain words to indicate the date of writing* (phenomenon 5), and *4.4.1 Iki should not be used to express a state in an impersonal sentence* (phenomenon 4) are among the most difficult, together with *4.8.2 Prie should not be used to express exchange ratios/conversion rates* (from phenomenon 8 with two error types).

Overall, there are no clear extreme outliers. The `Neurotechnology` family performs worst on *4.9.1. Prieš should not be used to express the object of an action with certain words* (0.42–0.50) and *4.8.2 Prie should not be used to express exchange ratios/conversion rates* (0.40–0.60), consistent with these being the most difficult error types in this group.

### 7.3.3 Phenomenon Type 3: Use of Forms

Figure 7.5 presents the results for error types in Phenomenon Type 3. The five easiest error types are spread across phenomena: *5.6.1 Simple forms of the subjunctive mood should not to be used instead of compound forms to express past actions (often hypothetical or unrealized)* (phenomenon 6) and *5.9.1 Half-participle should not be used to indicate a secondary action that does not coincide in time at any point with the main verb's action* (phenomenon 9) are single error types within their phenomena, along with *5.8.3 Passive past participle should not be used to express a property or state that has arisen spontaneously (without external influence or deliberate intervention)* (phenomenon 8), *5.1.2 The masculine genitive forms of ordinal numerals and of the pronoun kelintas, -a must not be used to*

*denote the day of the month*, and *5.1.3 The masculine accusative and instrumental forms of adjectives and of participles used with adjectival meaning are not to be used in the so-called "indefinite gender" meaning* (phenomenon 1). The most challenging error types are similarly distributed, including *5.2.1 The basic quantitative numeral forms from 2 to 9 and the pronoun abu, abi are not to be used with plural-only nouns (nouns that exist only in the plural)* (phenomenon 2) and *5.5.1 Negative predicate forms should not to be used in subordinate clauses expressing concession, when actual negation is not meant* (phenomenon 5; both single error types), as well as *5.10.1 Adverbial participle should not be used to indicate a secondary action of the same agent in personal sentences (instead of a half-participle or participle with proper agreement)* (phenomenon 10), *5.4.1 The infinitive should not be used with the conjunction jei/jeigu to express a condition* (phenomenon 4), and *5.8.1 Active present participle should not be used with nouns that do not denote an agent (nouns that cannot perform an action)* (phenomenon 8).

The second most challenging error type, *5.5.1 Negative predicate forms should not to be used in subordinate clauses expressing concession, when actual negation is not meant*, is particularly notable. While the `Neurotechnology` models and the two largest `Goldfish` models achieve perfect accuracy (1.0), 20 models from the `EuroLLM`, `Gemma 3`, `LLaMA 4`, `OLMo2`, `Qwen3`, and `TildeOpen` families score 0.0 or 0.1. In contrast, the easiest error type, *5.6.1 Simple forms of the subjunctive mood should not to be used instead of compound forms to express past actions (often hypothetical or unrealized)*, is solved perfectly by 41 models across families, but is harder for the `Goldfish` family, where four models score between 0.3 and 0.7. These outliers should be further investigated in Error Analysis 8.2.

### 7.3.4 Phenomenon Type 4: Coordination of Sentence Elements and Clauses

Phenomena in Phenomenon Type 4 consist of only one error type per phenomenon. Therefore, they correspond directly to the per-phenomena results and are shown in the last four columns of Figure 7.2. A separate table providing a less crowded overview is included in Appendix D.1, Figure D.1.

This phenomenon type includes both the easiest and most challenging error types overall, namely *6.1.1 Gi must not be used instead of a coordinating conjunction expressing contrast* (phenomenon 1) and *6.2.1 Ir should not be used to link words in intensifying constructions with adjectives and adverbs* (phenomenon 2). For *6.1.1 Gi must not be used instead of a coordinating conjunction expressing contrast*, all 78 models perform above chance level, with accuracies ranging from 0.67 (`Gemma3-270M-Instruct`) to 1.0 for 51 models across different families. In contrast, *6.2.1 Ir should not be used to link words in intensifying constructions with adjectives and adverbs* is highly challenging, with only 11 models achieving above-chance accuracy (0.6 or higher). The remaining error types show a wide performance range, from 0.0 to 1.0.

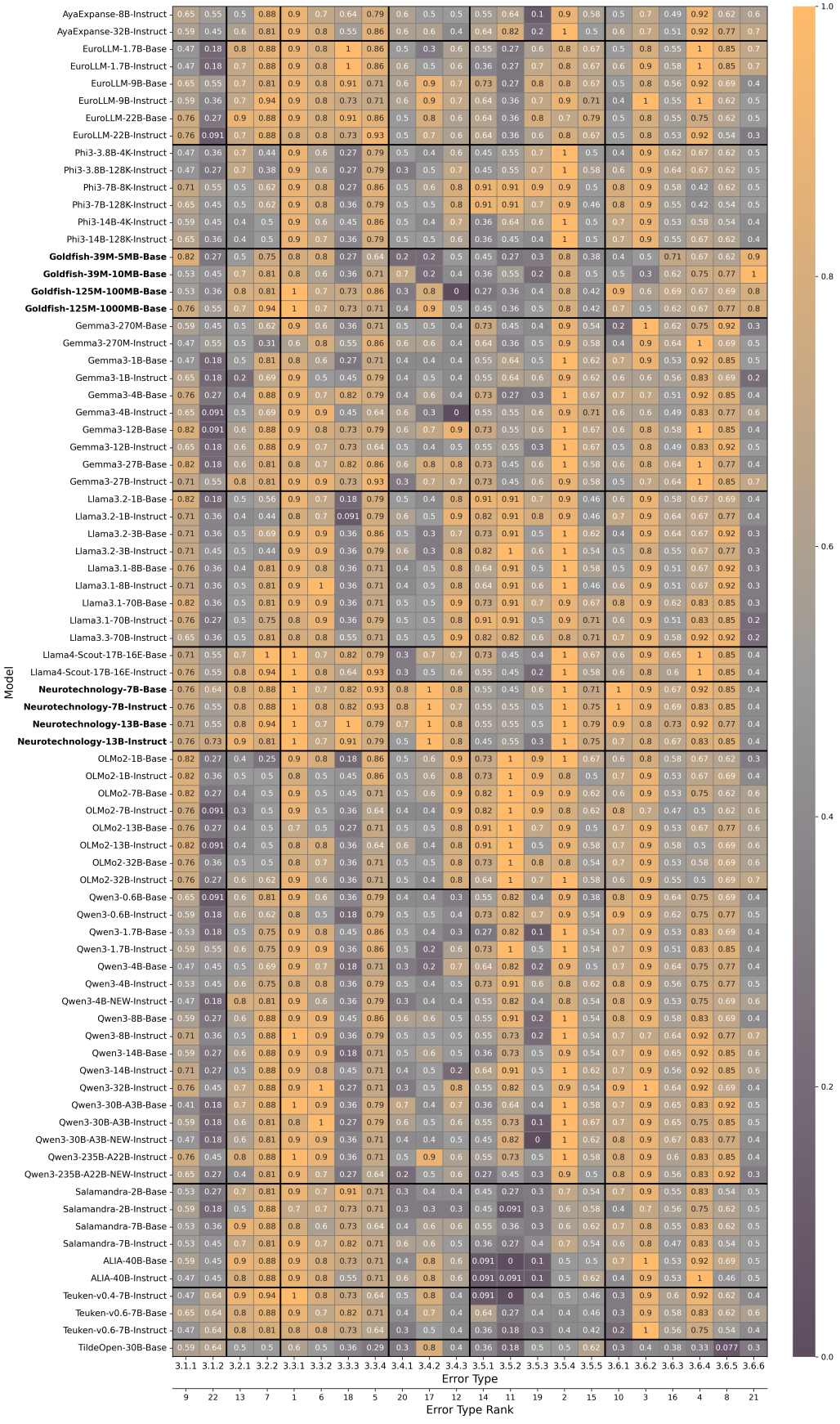


Figure 7.3: Per-error accuracies across models for Phenomenon Type 1 (Use of Cases). Monolingual models are marked in bold. Horizontal bold lines separate different model families, while vertical bold lines separate phenomenon types.

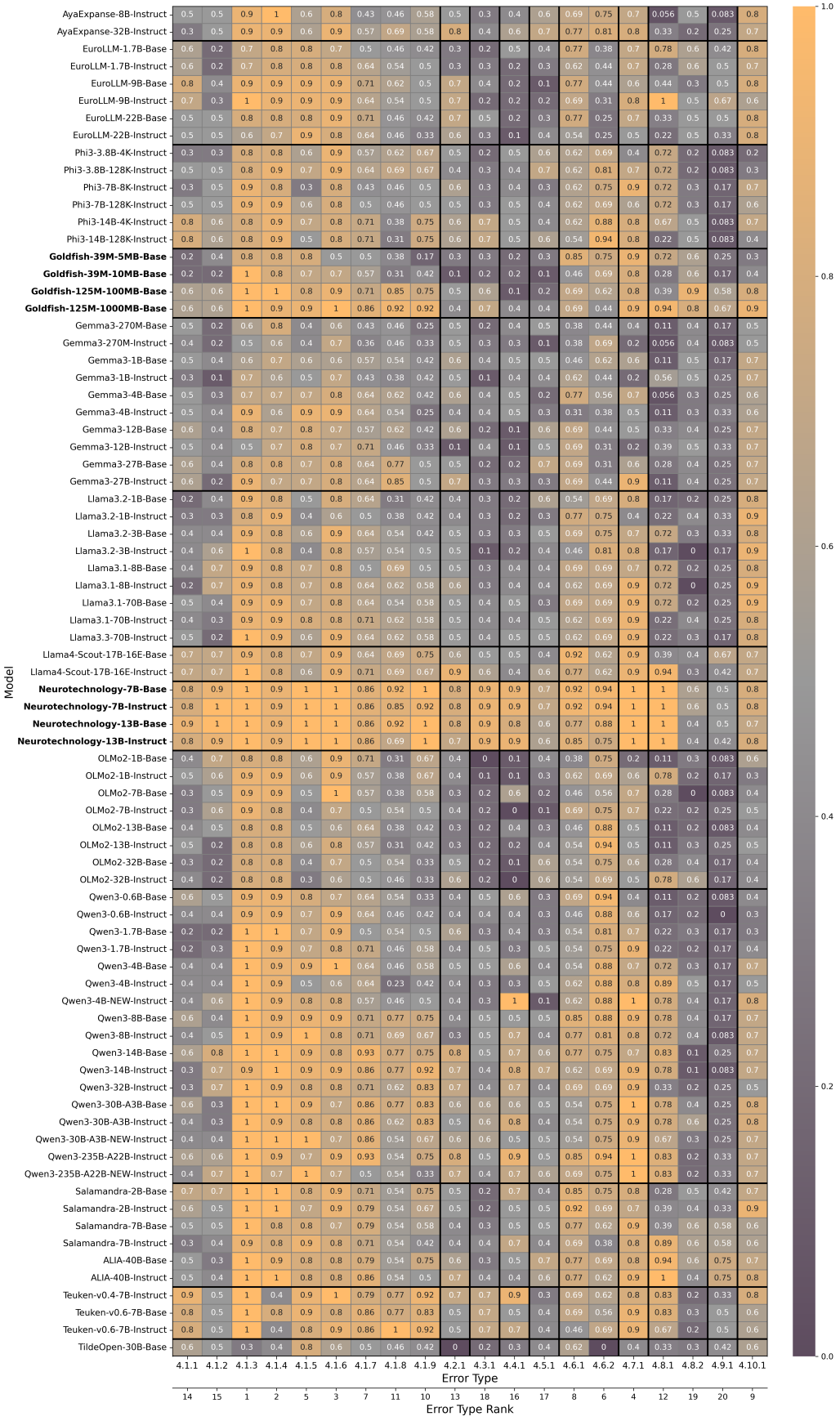


Figure 7.4: Per-error accuracies across models for Phenomenon Type 2 (Use of Prepositions). Monolingual models are marked in bold. Horizontal bold lines separate different model families, while vertical bold lines separate phenomenon types.

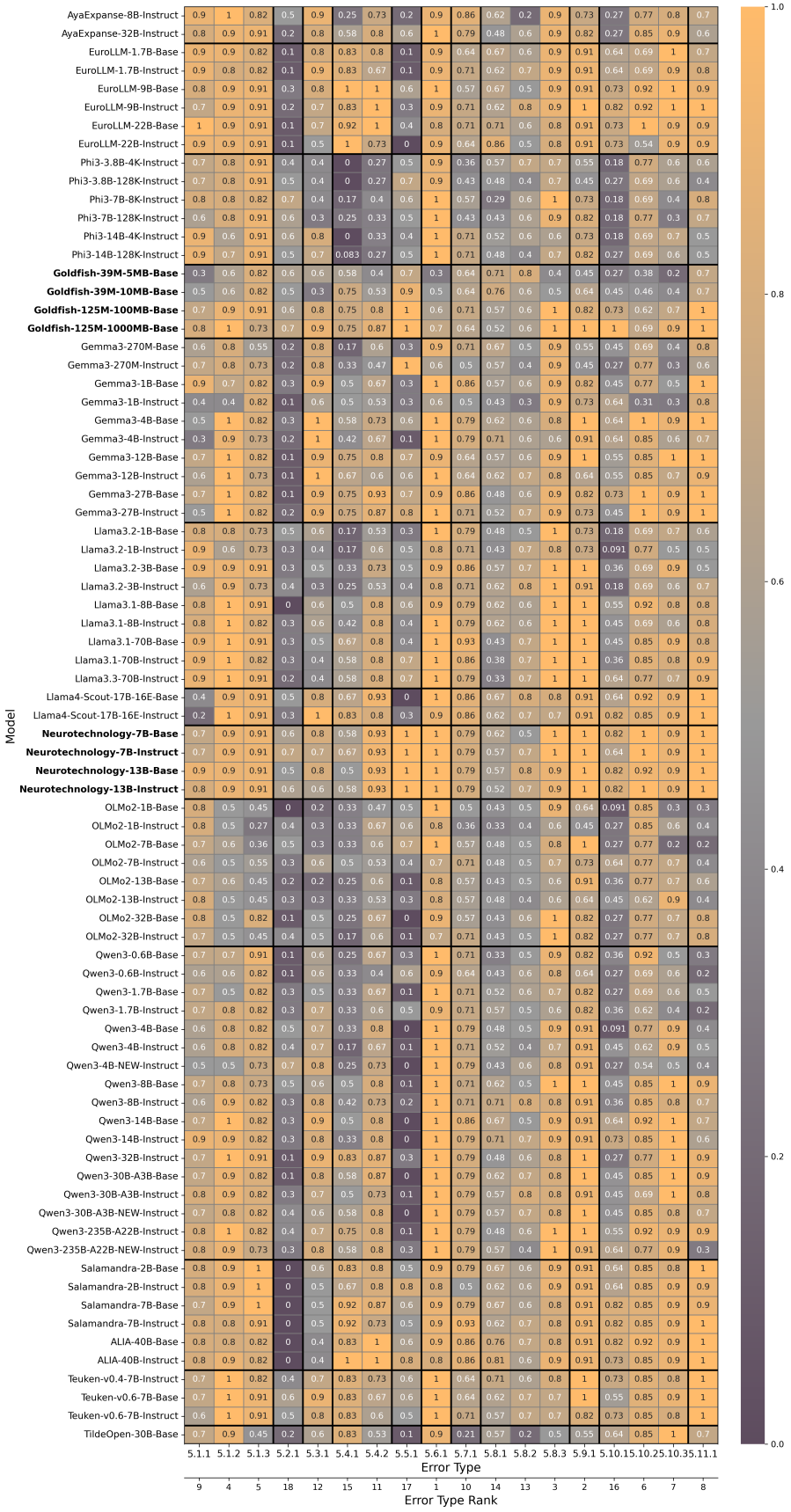


Figure 7.5: Per-error accuracies across models for Phenomenon Type 3 (Use of Forms). Monolingual models are marked in bold. Horizontal bold lines separate different model families, while vertical bold lines separate phenomenon types.

## 7.4 Effects of Model Characteristics on Decision Accuracy

In this section, I present the results obtained by fitting logistic mixed-effects models in R[1] using Posit Cloud[2]. Initially, the model included model size, architecture, version, language scope, and Lithuanian support as fixed effects. However, this specification produced a warning indicating multicollinearity among the predictors. Further investigation showed that language scope and Lithuanian support were highly collinear, meaning that the two variables were strongly associated and therefore provided overlapping information in the model. Consequently, Lithuanian support was removed, as it was not required to address the research questions (see Section 1.2.2).

The final model therefore included log-transformed model size, architecture, version, and language scope as fixed effects. Model size was log-transformed because it varied across several orders of magnitude (approximately 6,000 times difference between smallest and largest models), allowing the effect to be interpreted proportionally rather than in absolute terms.

The results showed that model size was positively associated with the likelihood of a correct decision ($\beta = 0.15$, SE $= 0.022$, $z = 6.95$, $p < 0.001$). In practical terms, larger models were more likely to produce a correct decision. The corresponding odds ratio indicated that a one-unit increase in log-transformed model size increased the odds of a correct decision by approximately 16% (OR $= 1.16$, 95% CI [1.11, 1.21]). These findings address Research Question 3 (see Section 1.2.2).

Among the categorical predictors, the reference category for language scope was European. Relative to this baseline, models with a monolingual scope had significantly higher odds of producing a correct decision ($\beta = 0.765$, SE $= 0.130$, $z = 5.88$, $p < 0.001$), with odds approximately 2.15 times higher than those of European scope models (95% CI [1.67, 2.77]).

In contrast, the global scope category showed a small but statistically significant negative effect relative to the European reference category ($\beta = -0.176$, SE $= 0.081$, $z = -2.18$, $p = 0.029$), corresponding to slightly reduced odds of a correct decision (OR $= 0.84$, 95% CI [0.72, 0.98]). No statistically significant effects were observed for architecture type or model version ($p > 0.05$). These findings relate to Research Questions 2 and 4 (see Section 1.2.2).

The random-effects structure indicated considerable variability across items (SD $= 1.58$), whereas variability across models was comparatively smaller (SD $= 0.26$). This suggests that differences between individual items contributed more to the variation in outcomes than differences between models themselves. The model explained a relatively small proportion of variance through the fixed effects alone (marginal $R^2 = 0.017$), whereas the inclusion of random effects substantially increased the explained variance (conditional $R^2 = 0.448$).

## 7.5 Correlations Between Model and Human Judgments

To evaluate how closely language model judgments align with human acceptability judgments, Spearman rank correlations (Spearman, 1904) were computed between model-derived certainty scores and human rating differences for the Lithuanian minimal sentence pairs. Model certainty was calculated as the difference between the mean negative log-likelihood (NLL) assigned to the ungrammatical and grammatical sentence variants:

[1] https://www.r-project.org
[2] https://posit.cloud

$$\text{certainty} = \text{mean NLL}(w_{\text{ungrammatical}}) - \text{mean NLL}(w_{\text{grammatical}}).$$

Human preference strength was quantified as the difference between the ratings assigned to the grammatically correct and incorrect sentence variants:

$$\text{rating difference} = \text{rating}(w_{\text{grammatical}}) - \text{rating}(w_{\text{ungrammatical}}).$$

Because the human ratings were collected using a Likert scale (see Section 2.4), the data are ordinal and the relationship between model predictions and human judgments may be non-linear. For this reason, Spearman rank correlation coefficients were used as the primary evaluation metric. This choice is also robust to outliers and does not assume linear relationships, making it suitable for the relatively small datasets (De Winter et al., 2024).

Higher positive correlations indicate stronger agreement between model predictions and human judgments, where both assign higher preference to grammatical over ungrammatical sentences. Correlations were computed separately for each model across the 192 sentence pairs, and the results are presented in Figure 7.6.

The results show that across the evaluated models, correlations with human judgments were generally positive but varied substantially in strength. The strongest correlations were observed for monolingual Lithuanian models and multilingual models that explicitly support Lithuanian. The highest Spearman correlation was obtained by `Goldfish-125M-1000MB-Base` ($\rho = 0.39$, $p < 0.001$), followed by `Teuken-v0.4-7B-Instruct` ($\rho = 0.37$, $p < 0.001$) and `Neurotechnology-13B-Base` ($\rho = 0.37$, $p < 0.001$).

When grouping by model family, the strongest performance was observed for the two `Goldfish` models, namely the variants trained on the most extensive datasets, as well as all evaluated models from `Neurotechnology`. The `Teuken` family also showed strong performance, with all three models yielding significant correlations ranging from $\rho = 0.27$ to $\rho = 0.37$. Another notable group was the Salamandra/ALIA family, whose six models achieved significant correlations between $\rho = 0.23$ and $\rho = 0.31$.

In contrast, many English-centric models exhibited weak or non-significant correlations with human judgments. Most `LLaMA 3`, `Phi 3`, and `OLMo2` models produced insignificant correlations close to zero, indicating limited alignment with human acceptability judgments on Lithuanian minimal sentence pairs.

For several model families, larger variants outperformed smaller ones. For example, within the `Qwen3` family, performance improved with scale. Smaller variants showed near-zero, non-significant correlations, whereas larger models such as `Qwen3-14B-Base` and `Qwen3-235B-A22B-Instruct` achieved weak to moderate correlations (up to $\rho = 0.25$, $p < 0.001$).

Overall, the results suggest that models with explicit Lithuanian support or monolingual training align more closely with human acceptability judgments. However, even the best-performing systems achieve only moderate correlations, indicating that the task remains challenging for current language models.

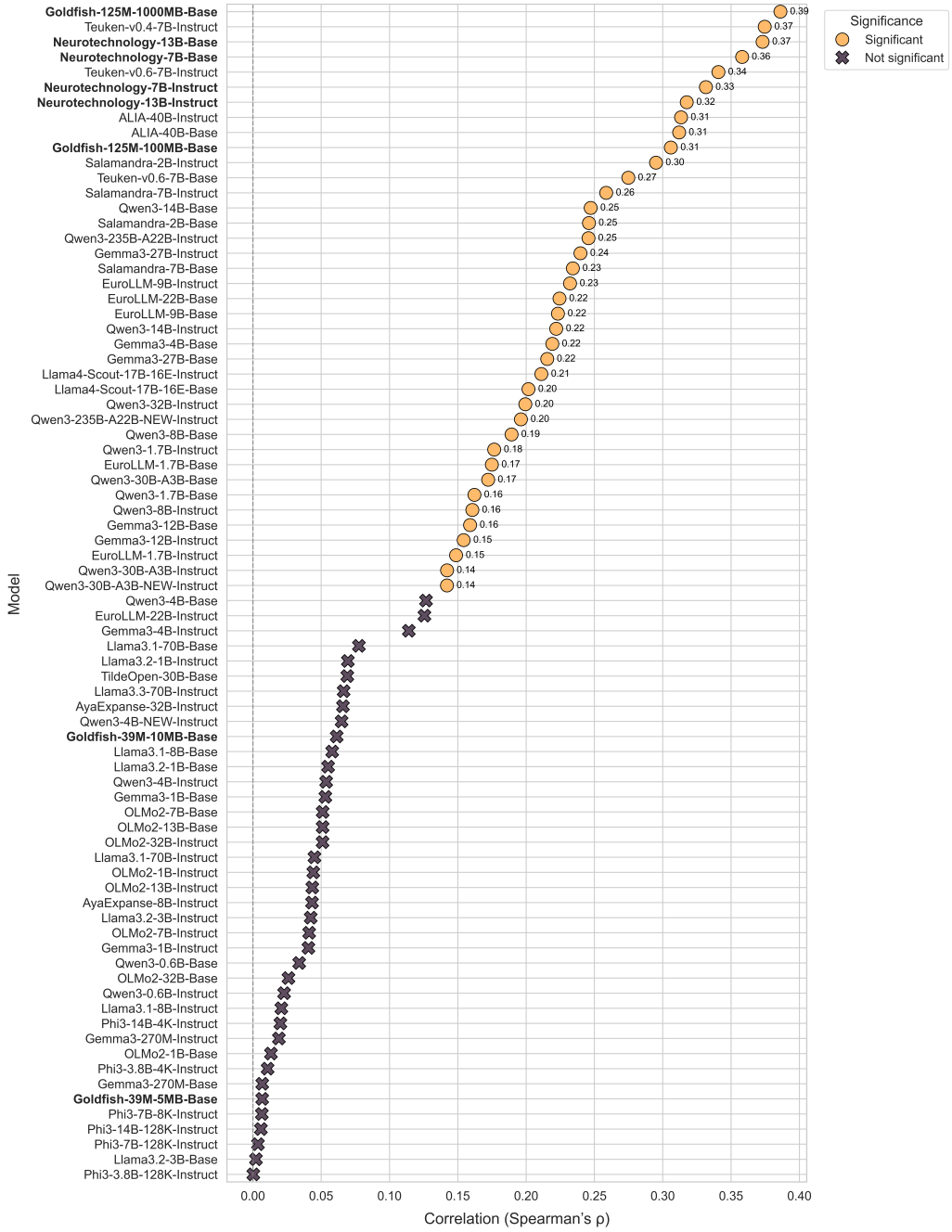


Figure 7.6: Spearman correlation coefficients between model predictions and human judgments per model. The models are ordered from best to worst performance. Monolingual models are marked in bold.

## 7.6 Correlations Between Model and Human Judgments by Phenomenon Type

As human participants evaluated only three sentence pairs for each of the 64 error types, there was insufficient data to perform statistical tests at the error-type or phenomenon level. Therefore, to obtain more fine-grained insights, Spearman correlations (Spearman, 1904) were computed across four phenomenon groups (Phenomenon Type 1–4). The results are presented in Figure 7.7.

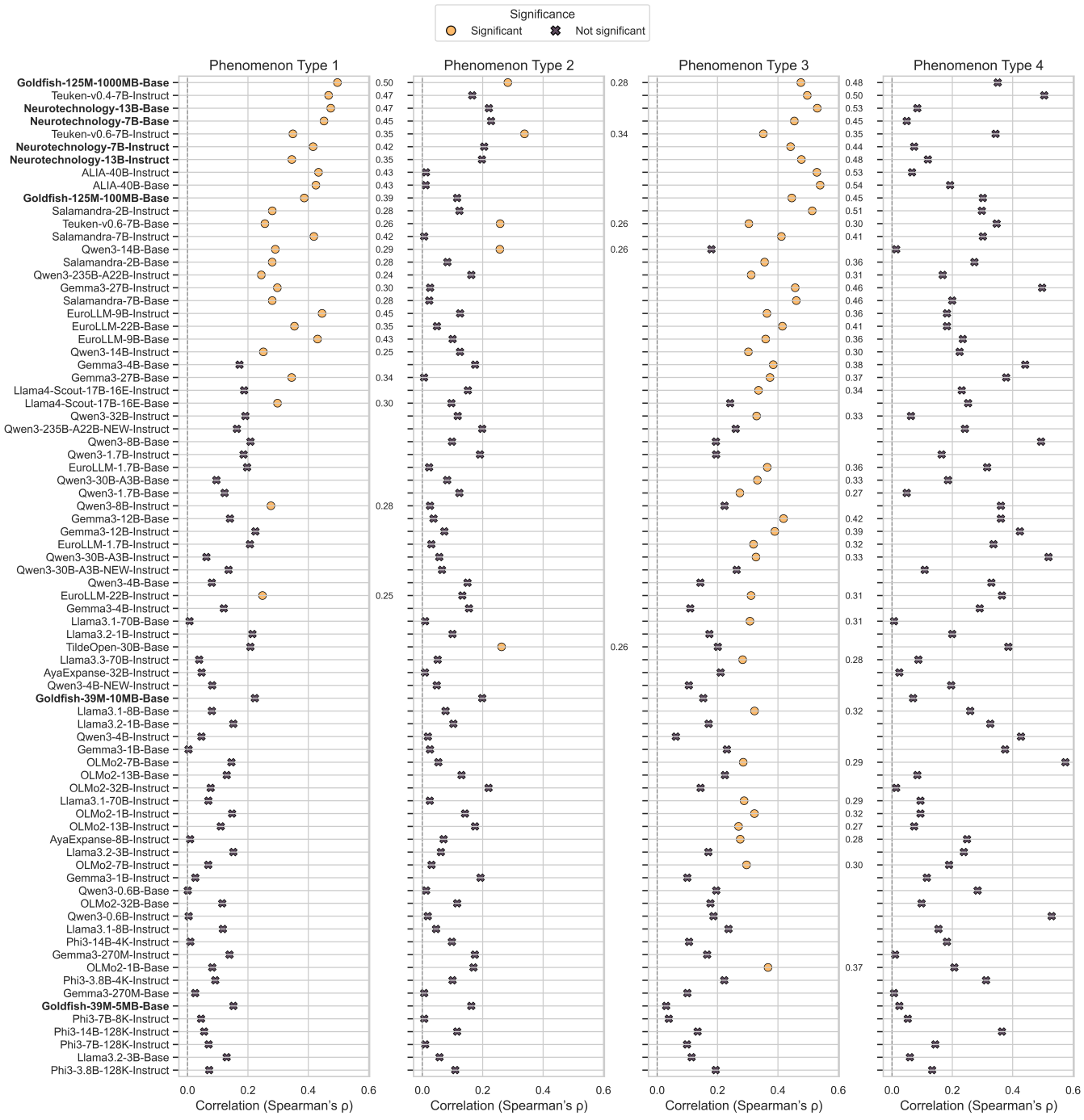


Figure 7.7: Spearman correlation coefficients between model predictions and human judgments per model and phenomenon type. The models are ordered from best to worst performance overall. Monolingual models are marked in bold. Correlation values are shown on the right side of each plot for statistically significant results.

### 7.6.1 Phenomenon Type 1: Use of Cases

This group consists of 22 error types, resulting in 66 observations between each of the 78 models and human judgments. For Phenomenon Type 1, 26 models achieved significant correlations, ranging from weak ($\rho = 0.24$ for `Qwen3-235B-A22B-Instruct`) to strong ($\rho = 0.50$ for `Goldfish-125M-1000MB-Base`) (see the left panel in Figure 7.7). The statistically significant models come from eight LM fami-

lies: Goldfish (monolingual), Teuken (European-focused multilingual), Neurotechnology (monolingual), Salamandra/ALIA (European-focused multilingual), Qwen3 (global-focused multilingual), Gemma 3 (global-focused multilingual), EuroLLM (European-focused multilingual), and LLaMA 4 (global-focused multilingual).

Overall, this phenomenon type is comparatively well captured by current language models, although correlations are predominantly weak to moderate, with only one model reaching a strong correlation.

### 7.6.2 Phenomenon Type 2: Use of Prepositions

Phenomenon Type 2, namely *Use of Prepositions*, contains 60 sentence pairs that were evaluated by both models and human participants. In this group, five models show significant correlations, ranging from $\rho = 0.26$ to $\rho = 0.34$ (see Figure 7.7). These models are Teuken-v0.6-7B-Instruct, Goldfish-125M-1000MB-Base, Teuken-v0.6-7B-Base, Qwen3-14B-Base, and TildeOpen-30B-Base. All five models officially support Lithuanian: Goldfish is monolingual, Teuken and TildeOpen are European-focused multilingual models, and Qwen3 is a globally trained multilingual model.

Overall, the results indicate that models capture aspects of prepositional usage to a limited extent, although even for significant models, correlation values remain weak to moderate.

### 7.6.3 Phenomenon Type 3: Use of Forms

Phenomenon Type 3, with 18 error types and 54 sentence pairs evaluated by both models and human participants, exhibited the best performance out of the four phenomenon types. Here, more than half of the models (43) achieved significant correlations, ranging from weak ($\rho = 0.27$) to strong ($\rho = 0.54$) (see Figure 7.7). In fact, five models reached strong correlations for this phenomenon type, namely ALIA-40B-Base, Neurotechnology-13B-Base, ALIA-40B-Instruct, Salamandra-2B-Instruct, and Teuken-v0.4-7B-Instruct. All of these models, with the exception of Neurotechnology-13B-Base, which is monolingual, are European-focused multilingual models.

Overall, Phenomenon Type 3 shows the strongest and most consistent alignment with human judgments, suggesting that this phenomenon is quite well captured by the evaluated models.

### 7.6.4 Phenomenon Type 4: Coordination of Sentence Elements and Clauses

Phenomenon Type 4 consists of four phenomena, which contain a single error type each. As each error type includes three sentence pairs evaluated by both models and human participants, this results in a very small sample size of 12 observations for this group. This severely limits statistical power, and all per-model correlations in this group are non-significant (see the right panel in Figure 7.7).

## 7.7 Summary of Key Findings

Overall, the results show considerable variation in model performance across phenomena, error types, and models (as well as their families). In the comparative judgments task, where models had to determine the more grammatical sentence, accuracies ranged from near chance level to strong performance (0.47-0.82), with monolingual Lithuanian models (particularly the Neurotechnology family) consistently achieving the highest scores.

Across linguistic phenomena and error types, model performance was highly uneven. The analyses confirm that difficulty is highly fine-grained, with strong performance on some constructions and near-chance or below-chance performance on others.

The logistic mixed-effects analysis showed that larger models perform better overall, and that monolingual models significantly outperform European and global multilingual models. Language scope emerged as a key predictor of performance, while architecture and model version did not show statistically significant effects. Item-level variation explained more variance than model-level differences, indicating substantial heterogeneity in linguistic difficulty across stimuli.

In the alignment analysis with human judgments, Spearman correlations were generally significant but modest. Monolingual Lithuanian and Lithuanian-supporting multilingual models showed the strongest alignment, with the best models reaching moderate correlations and some even reaching strong correlations. However, many English-centric models showed little to no alignment. Overall, model-human agreement was strongest for Phenomenon Type 3, moderate for Type 1, weak for Type 2, and non-significant for Type 4 (likely due to limited data).

Taken together, the results indicate that while current language models capture some aspects of Lithuanian grammar, performance is highly dependent on phenomenon and error type, training data coverage, and model scale.

# 8 Discussion

In this chapter, I discuss the results presented in Chapter 7. It is organized into two main sections. The first section addresses the research questions (see Section 1.2.2) and the extent to which the findings support the hypotheses introduced in Section 1.2.3. The second section presents a comprehensive qualitative linguistic error analysis, offering further insight into the strengths and limitations of the evaluated models.

## 8.1 Interpretation of Results

This section addresses five out of six research questions (see Section 1.2.2). The remaining research question, namely *Which syntactic phenomena and error types pose the greatest challenges for these models?*, is addressed in the error analysis (see Section 8.2).

### 8.1.1 Overall Syntactic Performance

The first research question asked: *To what extent do large language models correctly distinguish between grammatical and ungrammatical sentences in Lithuanian?* The results indicate substantial variation in performance across models. The worst-performing model was `TildeOpen-30B-Base`, which achieved an accuracy of 0.47, slightly below chance level. In contrast, the four best-performing models all belonged to the `Neurotechnology` family. These Lithuanian-adapted `LLaMA 2` models achieved accuracies ranging from 0.773 to 0.816, with `Neurotechnology-7B-Base` obtaining the highest overall score.

As expected, these results are considerably lower than those reported by MultiBLiMP (Jumelet et al., 2025), where model accuracies ranged from 0.832 to 0.985. Notably, even the best-performing model in the present study failed to match the performance of the lowest-scoring model evaluated by MultiBLiMP (see Table 8.1). This finding supports the hypothesis associated with Research Question 1, namely that the dataset developed in this study would provide a more challenging and realistic evaluation of Lithuanian syntactic competence than the currently available benchmark.

| Model Family | Version | Previous Accuracy | Current Accuracy |
|---|---|---|---|
| LLaMA 3 | 8B Base | 0.94 | 0.64 |
| | 70B Base | 0.97 | 0.68 |
| | 70B Instruct | 0.96 | 0.65 |
| Aya-expanse | 32B Instruct | 0.92 | 0.65 |
| Gemma 3 | 27B Base | 0.98 | 0.67 |
| | 27B Instruct | 0.96 | 0.68 |
| OLMo2 | 32B Base | 0.85 | 0.58 |
| | 32B Instruct | 0.83 | 0.57 |

| Model Family | Version | Previous Accuracy | Current Accuracy |
|---|---|---|---|
| Qwen3 | 14B Instruct | 0.96 | 0.69 |
| EuroLLM | 9B Base | 0.99 | 0.69 |
| Goldfish | 125M Base | 0.98 | 0.71 |

Table 8.1: Accuracies of language models on Lithuanian in the MultiBLiMP benchmark (*previous accuracy*) and in the current study (*current accuracy*). The lowest accuracies among the evaluated models are highlighted in red, while the highest accuracies are highlighted in green.

### 8.1.2 Multilingual and Monolingual Models

The second research question examined the following issues: *How does the performance of globally oriented multilingual models compare to that of European-focused multilingual models? How do both compare to monolingual (Lithuanian) models?* The results indicate that language coverage has a significant effect on overall model performance. Monolingual language models exhibited significantly higher odds of selecting the grammatical sentence than European-focused multilingual models. Furthermore, globally oriented multilingual models performed significantly worse than European-focused models, although the magnitude of this difference was relatively small.

These findings suggest that increasing the linguistic specialization of a model benefits its ability to process Lithuanian syntax. Models trained exclusively on Lithuanian appear to be better equipped to capture the language's morphological and syntactic regularities than multilingual models, whose capacity must be distributed across multiple languages. At the same time, the superior performance of European-focused models over globally oriented models indicates that reducing the number of languages represented during training may also be beneficial, even when Lithuanian itself is not the sole focus.

Overall, the results support the hypothesis associated with Research Question 2, namely that monolingual models would outperform European-focused multilingual models, which in turn would outperform globally oriented multilingual models.

### 8.1.3 The Effect of Model Size

The third research question asked: *How does model size affect performance?* The statistical results indicate that model size has a measurable effect on performance. In particular, larger models show a higher probability of selecting the grammatical sentence in each pair.

This result contradicts the initial hypothesis, which expected no substantial differences between model sizes due to the relatively constrained nature of the task, which does not require long-context reasoning or extensive world knowledge.

### 8.1.4 Base and Instruct Models

The fourth research question asked: *How do different model versions (base vs. instruct) compare in performance?* The results of the logistic mixed-effects models show no statistically significant differences between the two model types. Consequently, the hypothesis that base models would perform as well as or better than instruct models is not supported.

### 8.1.5 Alignment with Human Judgments

The fifth research question asked: *To what extent do model judgments align with native speaker acceptability judgments?* The results indicate that the alignment between model predictions and human judgments is generally modest and varies considerably across model types and linguistic phenomena.

Monolingual Lithuanian models, as well as multilingual models with strong Lithuanian or European language coverage, show the highest levels of correlation with human acceptability judgments. In contrast, global-scope models exhibit little to no meaningful alignment. This suggests that language coverage and specialization play an important role in how closely model judgments reflect native speaker intuitions.

At the level of linguistic phenomenon types, alignment is not uniform. Some phenomenon types show moderate to strong agreement with human judgments, whereas others exhibit weak or non-significant correlations, likely reflecting differences in data availability and the intrinsic difficulty of the constructions involved.

## 8.2 Error Analysis

This section presents a detailed error analysis of the model results, thereby addressing the final research question: *Which syntactic phenomena and error types pose the greatest challenges for these models?* First, I examine the five most challenging error types for the models based on their mean accuracy and discuss possible interpretations of why the models, the human participants, or both struggled with these particular error types. I then analyse additional cases in which the models' performance was particularly noteworthy during the analysis of the results (see Chapter 7). Finally, I discuss interesting cases involving human acceptability judgements, which were already presented in Chapter 6.

For each case, I provide a brief overview of the corresponding model results as well as the human acceptability ratings in order to give a self-contained discussion that does not require constant reference to previous chapters. Moreover, I include a glossed example of one sentence pair for each error type to illustrate the error more clearly. The abbreviations used for glosses in this chapter are provided in Table 8.2.

### 8.2.1 Most Challenging Error Types

#### I. 6.2.1 *Ir* should not be used to link words in intensifying constructions with adjectives and adverbs.

The most challenging error type for the models, namely *6.2.1 Ir should not be used to link words in intensifying constructions with adjectives and adverbs*, belongs to Phenomenon Type 4. Across all evaluated models, this error type achieved a mean accuracy of only 0.30. Interestingly, it did not pose comparable difficulty for human participants. In the human evaluation, grammatical sentences received an average acceptability rating of 4.87 out of 7 on the Likert scale, whereas ungrammatical sentences received an average rating of 2.80 (see Section 6.4.2). Thus, the results aligned with the expected distinction between acceptable and unacceptable constructions, with grammatical sentences rated above 3.5 and ungrammatical sentences below 3.5.

Among the evaluated models, the best individual performance on this error type was achieved by `OLMo2-1B-Instruct` (see Section 7.3.4). At the model-family level, however, the strongest results were

| Abbreviation | Meaning |
|---|---|
| 2 | Second person |
| 3 | Third person |
| ACC | Accusative |
| COND | Conditional |
| DAT | Dative |
| F | Feminine |
| FUT | Future |
| GEN | Genitive |
| IMP | Imperative |
| INF | Infinitive |
| INS | Instrumental |
| ITER | Iterative |
| LOC | Locative |
| M | Masculine |
| NEG | Negation |
| N | Neuter |
| NOM | Nominative |
| PASS | Passive |
| PL | Plural |
| PFV | Perfective |
| PRED | Predicative |
| PRS | Present |
| PST | Past |
| PTCP | Participle |
| REFL | Reflexive |
| SG | Singular |
| VOC | Vocative |

Table 8.2: Glossing abbreviation conventions.

obtained by the `Neurotechnology` models, whose accuracies ranged from 0.50 to 0.70. In contrast, another monolingual model family, `Goldfish`, performed substantially worse, with accuracies ranging from 0.20 to 0.30.

Research on multilingual language models shows that they learn shared internal representations across languages, which enables cross-lingual transfer even in zero-shot settings (H. Wang et al., 2024). This suggests that linguistic patterns are not learned in isolation for each language but may instead interact within a shared representational space. Beyond model performance, human acceptability judgments may likewise be shaped by cross-linguistic influence, which refers to humans applying knowledge of one language to another (Jarvis & Pavlenko, 2008). In post-Soviet countries such as Lithuania, Russian has historically functioned as an important contact language, particularly for older generations educated during the Soviet period. While the degree of exposure varied across individuals and regions, Russian was present in many domains of everyday life, which likely resulted in at least some degree of passive or active bilingual exposure among parts of the population (Kreusler, 1961). The present error type reflects such cross-linguistic interference from Russian (Miliūnaitė, Rita, 2003).

Against this background, one possible explanation for the generally poor model performance is that the corresponding ungrammatical Lithuanian construction is acceptable in Russian, which is the $11^{th}$ most spoken language worldwide in 2026 (Ethnologue, 2026). As a result, the models may assign higher plausibility to a structure that is acceptable in a structurally similar high-resource language, even though it is ungrammatical in Lithuanian. Consequently, Lithuanian alternatives that correctly avoid the intensifying use of *ir* may be relatively underrepresented in multilingual training data, which may limit the model's ability to reliably learn this constraint.

For example, the sentence *Turėjome peržiūrėti tūkstančius ir tūkstančius bylų* (in English, *We had to*

*review thousands and thousands of files*) is ungrammatical in Lithuanian (see Example 19), because, as the error type states, the conjunction *ir* should not be used in intensifying adjective or adverb constructions. However, such a construction is allowed in Russian, as shown in Example 21[1]. In Lithuanian, depending on the context, alternative constructions should be used. In this example, *tūkstančius ir tūkstančius* should be replaced with *tūkstančių tūkstančius*, thereby removing the conjunction *ir* and changing the first word to the genitive case instead of the accusative (see Example 20).

The comparatively stronger performance of the `Neurotechnology` models may be related to their access to extensive training data and additional fine-tuning on Lithuanian-specific resources. In contrast, the weaker performance of the monolingual `Goldfish` models could potentially be explained by the comparatively small size of their training corpora (5–1000 MB), which, despite being Lithuanian-language datasets, may provide insufficient exposure to such relatively uncommon constructions.

(19) **Lithuanian**

| **Turėj-ome* | *per-žiūrė-ti* | *tūkstanč-ius* | *ir* | *tūkstanč-ius* | *byl-ų.* |
|---|---|---|---|---|---|
| have-PST.2PL | PFV-look-INF | thousand-M.ACC.PL | and | thousand-M.ACC.PL | file-F.GEN.PL. |

'We had to review thousands and thousands of files.'

(20) **Lithuanian**

| *Turėj-ome* | *per-žiūrė-ti* | *tūkstanč-ių* | *tūkstanč-ius* | *byl-ų.* |
|---|---|---|---|---|
| have-PST.2PL | PFV-look-INF | thousand-M.GEN.PL | thousand-M.ACC.PL | file-thousand-F.GEN.PL. |

'We had to review thousands and thousands of files.'

(21) **Russian**

| *Nam* | *pri-shl-o-s'* | *pro-smotre-t'* | *tysiach-i* | *i* | *tysiach-i* |
|---|---|---|---|---|---|
| we.DAT.PL | PFV-fall-on.PST-N-REFL | PFV-look-INF | thousand-F.NOM.PL | and | thousand-F.NOM.PL |

*fail-ov.*
file-M.GEN.PL.

'We had to review thousands and thousands of files.'

### II. 4.9.1 *Prieš* should not be used to express the object of an action with certain words.

The second most challenging error type, which also achieved a mean accuracy of 0.30 among the evaluated models (see Section 7.3.2), is another example of cross-linguistic influence, namely literal translations involving sentence-construction borrowings from other languages (in this case, Russian) that are ungrammatical in Lithuanian. For example, the sentence *Teko atsiprašyti prieš draugą* (see Example 22), which translates to *(Someone) had to apologize to a friend*, is incorrect because the verb *atsiprašyti* (in English, *to apologize*) cannot be used with the preposition *prieš*. Instead, the sentence should be formulated as *Teko atsiprašyti draugą* (see Example 23).

However, such a construction is acceptable in Russian: *Prishlos' izvinit'sia pered drugom* (see Example 24), where *izvinit'sia pered* corresponds literally to *atsiprašyti prieš* (in English, *to apologize to*).

[1]These and the following Russian examples were proofread by a native speaker, Inna Oleksiichuk.

(22) **Lithuanian**

**Tek-o* *at-si-prašy-ti* *prieš* *draug-ą.*
fall.to-PST.3 PFV-REFL-ask-INF in.front.of friend-M.ACC.SG.

'(Someone) had to apologize to a friend.'

(23) **Lithuanian**

*Tek-o* *at-si-prašy-ti* *draug-ą.*
fall.to-PST.3 PFV-REFL-ask-INF friend-M.ACC.SG.

'(Someone) had to apologize to a friend.'

(24) **Russian**

*Pri-shl-o-s'* *iz-vini-t'-sia* *pered* *drug-om.*
PFV-fall-on.PST-N-REFL PFV-blame-INF-REFL in.front.of friend-M.INS.SG.

'(Someone) had to apologize to a friend.'

However, cross-lingual transfer may not be the only reason why the models failed to achieve better performance. Looking at the human acceptability ratings, it is evident that while participants rated grammatical sentences higher than ungrammatical ones (3.77 vs. 2.90) (see Section 6.4.2), the grammatical sentences are only marginally above the 3.5 threshold, and the difference between the two ratings (0.87) is not particularly large. One possible explanation for this is the quality of the dataset.

Returning to the previously discussed examples, *Teko atsiprašyti draugą* (see Example 23) is a grammatical alternative to the ungrammatical *Teko atsiprašyti prieš draugą* (see Example 22). However, it is not the only acceptable alternative. Another grammatical variant, mentioned by VLKK (Valstybinė lietuvių kalbos komisija (VLKK), 2023), is *Teko atsiprašyti draugo* (see Example 25), where the accusative case used for *friend* is replaced with the genitive case.

When constructing the dataset, I opted for the first variant in order to keep the sentence pairs as minimal as possible. In my selected sentence, the case marking on *friend* did not change, resulting in only one modification in the sentence pair instead of two. However, it is possible that the chosen variant is less frequently used in the language; therefore, it may be less natural for native speakers and may also occur less often in the training data used by the models.

Therefore, in future work, before reusing the dataset, it would be worth conducting another survey to test which sentence construction is more acceptable to native speakers. It may also be beneficial to evaluate both variants and compare model performance across them.

(25) **Lithuanian**

*Tek-o* *at-si-prašy-ti* *draug-o.*
fall.to-PST.3 PFV-REFL-ask-INF friend-M.GEN.SG.

'(Someone) had to apologize to a friend.'

Notably, not all sentence pairs were rated consistently across participants. This particular item appeared especially difficult and may have influenced the overall score. Importantly, it was the only sentence pair within this error type in which the ungrammatical variant received a higher rating than the grammatical one (2.4 vs. 1.4). Since both ratings were below 3.5, participants appeared to

judge both variants as ungrammatical. Specifically, each variant was rated by 10 participants. The grammatical version received ratings between 1 and 3, with 7 participants assigning it a rating of 1, indicating complete unacceptability. The ungrammatical version received ratings between 1 and 5, with 3 participants assigning it a rating of 1.

These results suggest variation not only across error types, but also across individual items. Moreover, because the example was taken directly from the original VLKK (Valstybinė lietuvių kalbos komisija (VLKK), 2023) source, the possibility that the item itself was incorrect can be ruled out.

**III. 5.2.1 The basic quantitative numeral forms from 2 to 9 and the pronoun *abu*, *abi* are not to be used with plural-only nouns (nouns that exist only in the plural).**

The third most challenging error type for the models achieved a mean accuracy of 0.32 (see Section 7.3.3) and was also problematic for human participants (see Section 6.4.2). Although grammatical and ungrammatical sentences received similar ratings (3.73 vs. 4.27), the ungrammatical sentences were unexpectedly rated slightly higher. A closer inspection revealed that, out of the three evaluated sentence pairs, two were problematic. Unlike the remaining pair, these two examples were semi-automatically generated.

Apparently, the five sentence pairs taken from the original source (Valstybinė lietuvių kalbos komisija (VLKK), 2023) were correct, whereas the semi-automatically generated examples contained errors. In those cases, regular plural nouns were used instead of plural-only nouns, meaning that the intended "grammatical" sentences were themselves ungrammatical and therefore failed to instantiate the target error type properly. Instead, these examples produced cases of *hypernormalism* (Valstybinė lietuvių kalbos komisija (VLKK), 2026), which, although a distinct error type occasionally found in Lithuanian, was not relevant in this context. Therefore, all five semi-automatically generated examples for error type *5.2.1* should be replaced in future versions of the dataset.

Accordingly, this error type is analysed further using only the five original sentence pairs. One example, included in the human evaluation, contained the incorrect sentence *Dalyvavo trijose varžybose* (see Example 26) and the correct alternative *Dalyvavo trejose varžybose* (see Example 27), both meaning *[He/She/They] participated in three competitions* in English. In Lithuanian, *varžybos* is a plural-only noun; therefore, numerals must take the plural form *trejose* rather than the regular form *trijose*. However, this distinction proved difficult for the participants: ratings varied considerably and spanned the full 1-7 scale for both sentence types. The correct variant received a mean rating of 3.7, only slightly above the midpoint (3.5), while the ungrammatical sentence received a mean rating of 3.5, indicating substantial uncertainty and low agreement among participants. Although these findings are based on only one sentence pair and are therefore not conclusive, they align with previous observations that this is among the most frequent error types in Lithuanian (Bielinskienė et al., 2014).

(26) **Dalyvav-o* *trij-ose* *varžyb-ose.*
participate-PST.3 three-F.LOC.SG competition.PL-F.LOC.PL.

'[He/She/They] participated in three competitions.'

(27) *Dalyvav-o* *trej-ose* *varžyb-ose.*
participate-PST.3 three-F.LOC.PL competition.PL-F.LOC.PL.

'[He/She/They] participated in three competitions.'

The model results were recalculated using only the five original sentence pairs, excluding the semi-automatically generated items. Figure 8.1 presents both the original model results (dark purple) and the recalculated scores (gold). These results should be interpreted with caution, as the recalculated scores are based on only five sentence pairs; consequently, each sentence pair accounts for 20% of the final score instead of 10%, as in the original evaluation.

For most models, performance worsened after recalculation. In the original results, eight out of 78 models scored 0.0, whereas after recalculation 29 models scored 0.0. For the majority of models, accuracy decreased by 0.1 or 0.2. Particularly notable was `OLMo2-7B-Base`, whose accuracy dropped from 0.5 to 0.0, suggesting that the model, contrary to previous assumptions, genuinely struggles with this error type.

Although most results worsened, 14 models improved and another 14 models (including the eight that originally scored 0.0) remained unchanged. This suggests that the erroneous semi-automatically generated examples may previously have confused or misled some models. Interestingly, the best-performing model family overall, `Neurotechnology`, improved for two models after recalculation, while the remaining two models maintained their previous scores. Another monolingual model family, `Goldfish`, did not improve after recalculation; its models either remained unchanged or performed worse. Notably, two `Teuken` models also improved, with `Teuken-v0.6-7B-Base` and `Teuken-v0.6-7B-Instruct` joining `Neurotechnology-7B-Instruct` as the best-performing models after recalculation, each achieving an accuracy of 0.8. By contrast, several previously high-performing models, including `Qwen3-4B-NEW-Instruct` and `Phi3-7B-8K-Instruct`, showed lower scores after recalculation, indicating that their earlier strong performance had been inflated by the flawed semi-automatically generated examples.

Overall, although the results are not definitive due to the small sample size, they suggest that this error type is genuinely challenging for language models, with the mean accuracy across models dropping to 0.23 after recalculation.

One possible reason for the difficulty of this error type for the human participants is the low frequency of plural-only, also known as *pluralia tantum* (Corbett, 2019), nouns. Low-frequency words generally require longer processing times (Baayen et al., 1997), and *pluralia tantum* nouns, in particular, exhibit a complex relationship between grammatical form and conceptual number. While they are typically morphologically plural, their association with conceptual plurality is not always consistent and may differ from that of regular plural nouns (Nenonen & Niemi, 2010).

Language models are likely affected by the same rarity issue. Both plural-only nouns and the plural numeral forms that agree with them are relatively infrequent in training corpora, meaning that models encounter them too rarely to develop robust generalisations.

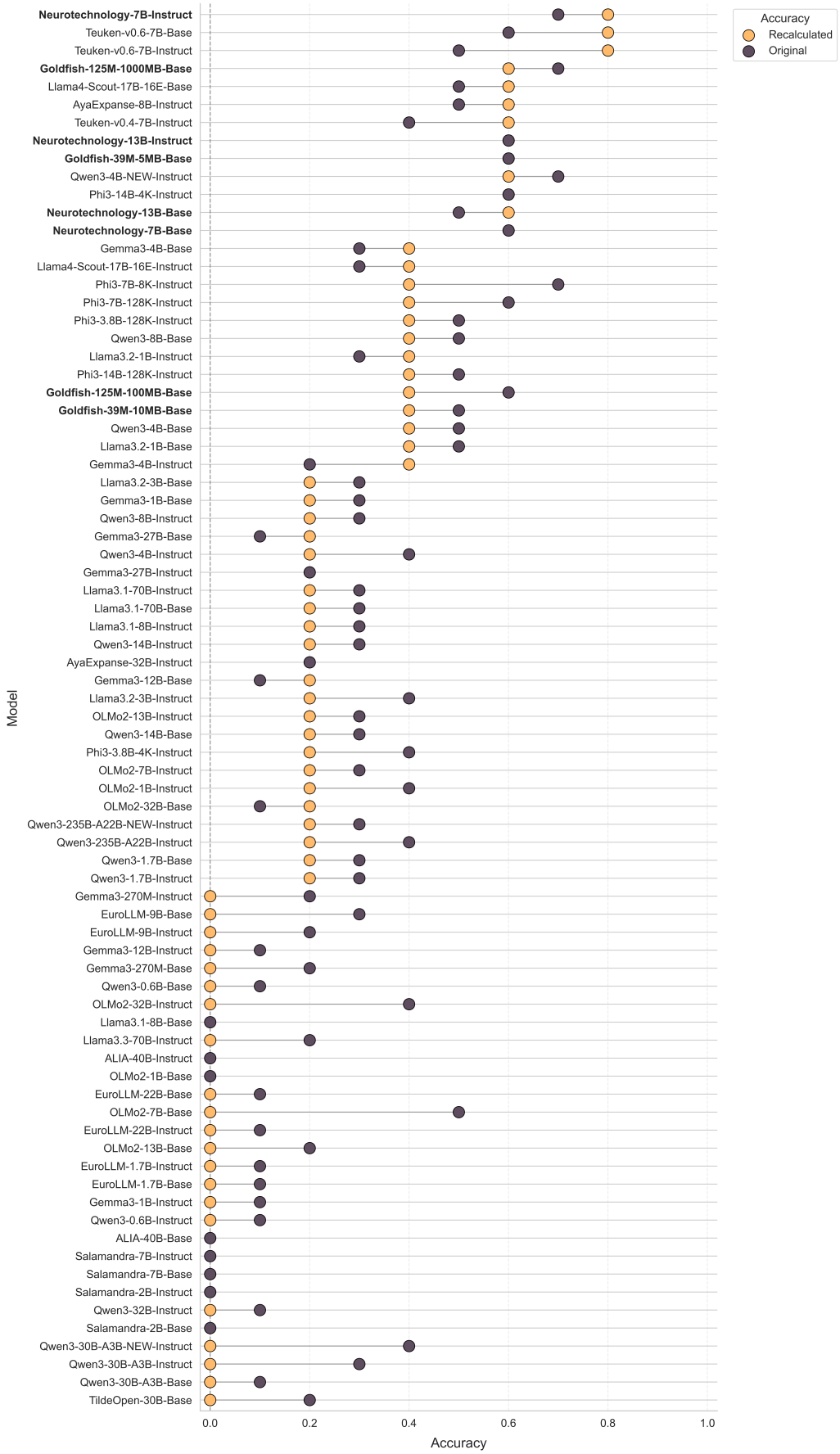


Figure 8.1: Comparison of model accuracies for error type *5.2.1* before recalculating the scores and after. The models are ordered from best to worst performance based on the recalculated scores. Monolingual models are marked in bold.

### IV. 3.1.2 Nominative case should not be used to express direct address.

Another challenging error type for the models was *3.1.2 Nominative case should not be used to express direct address*. Here, the models achieved an average accuracy of 0.36 (see Section 7.3.1). However, it is important to note that the `Neurotechnology` family models performed substantially better than the other models, reaching accuracies between 0.55 and 0.73, which indicates performance above chance level.

Interestingly, this error type was not challenging for the human participants (see Section 6.4.2). The grammatical sentences were rated as clearly grammatical, with a mean score of 6.07 across the three sentences belonging to this error type. The ungrammatical sentences were also identified as clearly ungrammatical, receiving an average score of 1.67. Thus, Lithuanian native speakers were able to distinguish grammatical from ungrammatical variants very clearly, with a mean rating difference of 4.4 points.

Although this error type is categorized under nominative case errors, it actually concerns the vocative case, which is relatively rare and, in many languages, formally identical to the nominative case (Ambrazas, 2026c). In this construction, the grammatical variants require the vocative form, as shown in Example 28, whereas the ungrammatical variants incorrectly use the nominative form instead (see Example 29). This type of error is often influenced by other languages, such as English, which does not have a productive vocative case (Budreikienė, 2018). As demonstrated in Examples 28 and 29, both Lithuanian forms, *vairuotojau* and *vairuotojas*, are translated into English simply as *driver*.

(28) *Vairuotoj-au, ati-daryk-it dur-is.*
driver-M.VOC.SG PFV-open-IMP.2PL door-F.ACC.PL.

'Driver, open the door.'

(29) **Vairuotoj-as, ati-daryk-it dur-is.*
driver-M.NOM.SG PFV-open-IMP.2PL door-F.ACC.PL.

'Driver, open the door.'

One possible explanation for why the models found this error type particularly challenging is a bias toward the nominative case, which occurs very frequently in training data. This interpretation is also supported by research on other languages. For example, in an evaluation of LLMs on Georgian, researchers found that when the dative case was the correct choice but the models failed to select it, they chose the nominative case in 83.2% of the errors (Gallagher & Heyer, 2026). This suggests that language models tend to overgeneralize the nominative case because of its high frequency in training corpora. This issue may be especially pronounced in multilingual models, where the vocative case could be underrepresented due to its rarity. In contrast, monolingual language models appear to handle such constructions more successfully, as evidenced by the performance of the `Neurotechnology` family models. Another monolingual model family, `Goldfish`, achieved accuracies ranging from 0.27 to 0.55. Although most of these models still performed below chance level, performance improved with larger training datasets, further supporting this explanation.

### V. 4.8.2 *Prie* should not be used to express exchange ratios/conversion rates.

The fifth error type, which was especially difficult for the models, was *4.8.2 Prie should not be used to express exchange ratios/conversion rates.* Here, the models demonstrated an average accuracy of 0.38 (see Section 7.3.2). Only the two largest `Goldfish` family models achieved impressive accuracies of 0.8-0.9.

Human participants also struggled with this error type (see Section 6.4.2). Both the grammatical and ungrammatical variants received identical average accuracies of 2.87, indicating that participants rated both versions as equally ungrammatical. Moreover, two out of the three evaluated sentence pairs appeared among the examples where humans assigned higher ratings to the ungrammatical sentences than to the grammatical ones. However, all three sentences were semi-automatically generated.

This error type states that the preposition *prie* should not be used to express ratios. However, native speakers sometimes make mistakes by using *prie* in sentences such as *Pinigai bus keičiami santykiu vienas prie šimto* (see Example 30) instead of *Pinigai bus keičiami santykiu vienas už šimtą* (see Example 31), meaning *Money will be exchanged at a rate of one for one hundred.* This example from the original source was not included in the human evaluation, making it difficult to determine whether the poor performance of both human participants and language models resulted from problems with the semi-automatically generated examples. To address this concern, the grammaticality of the semi-automatically generated sentences was independently reviewed by Rita Miknevičienė, a Lithuanian language teacher at Juozas Balčikonis Gymnasium in Panevėžys, Lithuania. She confirmed that the sentences were grammatically correct, suggesting that the observed difficulties are unlikely to stem from errors in the examples themselves and may instead be attributable to other factors.

(30) **Pinig-ai* *bus* *keiči-am-i* *santyk-iu* *vien-as*
money-M.NOM.PL be.FUT.3 change-PRS.PTCP.PASS-M.NOM.PL ratio-M.INS.SG one-M.NOM.SG
*prie* *šimt-o.*
at hundred-M.GEN.SG.

'Money will be exchanged at a rate of one for one hundred.'

(31) *Pinig-ai* *bus* *keiči-am-i* *santyk-iu* *vien-as*
money-M.NOM.PL be.FUT.3 change-PRS.PTCP.PASS-M.NOM.PL ratio-M.INS.SG one-M.NOM.SG
*už* *šimt-ą.*
for hundred-M.ACC.SG.

'Money will be exchanged at a rate of one for one hundred.'

Additionally, another possible issue with the dataset is that, in the original source, *vienas prie šimto* can be correctly replaced not only by *vienas už šimtą*, but also by *vienas su šimtu* and *šimtas į vieną*. Thus, the preposition *prie* may be replaced by three different alternatives. In this study, all three options were used randomly in the examples. However, after verifying the correctness of the semi-automatically generated sentences, future research could benefit from conducting a smaller follow-up study in which all three variants are evaluated instead of selecting one at random. This would make it possible to investigate whether one variant sounds more natural to native Lithuanian speakers and, if so, which variant is preferred.

### 8.2.2 Additional Observations on Model Performance

#### Most Challenging Phenomena

Section 8.2.1 discusses the five most challenging error types for the models, and the first three correspond to (or form part of) the three most challenging phenomena. However, the remaining two challenging phenomena, namely *4.3 Į* and *4.5 Iš*, are discussed in this subsection.

**IV. 4.3 *Į*** The fourth most challenging linguistic phenomenon for the models is *4.3 Į*, with a mean accuracy of 0.39 across the models (see Section 7.2). This linguistic phenomenon includes only one error type; therefore, the phenomenon and the corresponding error type *4.3.1 Į should not be used to express the object of an action with certain words* are identical.

This error type ranked $6^{th}$ among the most challenging error types (see Appendix E.2, Table E.1). However, it was not challenging for the human participants (see Section 6.4.2). None of the three sentence pairs received reversed mean ratings (where the ungrammatical variants were rated higher than the grammatical ones), and participants clearly identified the grammatical sentences as correct, with a mean rating of 5.97, and the ungrammatical ones as incorrect, with a mean rating of 1.4. This resulted in a difference of 4.57 between the two conditions.

This error type is once again likely influenced by the models' exposure to multilingual training data. For example, Russian requires a preposition in similar constructions, whereas Lithuanian does not. Lithuanian speakers may therefore incorrectly use the preposition *į* in sentences such as *Orai labai veikia į žmonių nuotaikas* (see Example 32) instead of the grammatical version without a preposition, *Orai labai veikia žmonių nuotaikas* (see Example 33); both sentences mean *The weather has a huge impact on people's moods.* This tendency may be influenced by Russian, where a preposition is required, as in *Pogoda okazyvaet ogromnoe vliianie na nastroenie liudei* (see Example 34).

(32) **Lithuanian**

| **Or-ai* | *labai* | *veik-ia* | *į* | *žmon-ių* | *nuotaik-as.* |
|---|---|---|---|---|---|
| weather-M.NOM.PL | very | affect-PRS.3 | in | people-M.GEN.PL | mood-F.ACC.PL. |

'The weather has a huge impact on people's moods.'

(33) **Lithuanian**

| *Or-ai* | *labai* | *veik-ia* | *žmon-ių* | *nuotaik-as.* |
|---|---|---|---|---|
| weather-M.NOM.PL | very | affect-PRS.3 | people-M.GEN.PL | mood-F.ACC.PL. |

'The weather has a huge impact on people's moods.'

(34) **Russian**

| *Pogod-a* | *okazyva-et* | *ogromn-oe* | *vliian-ie* | *na* | *nastroen-ie* |
|---|---|---|---|---|---|
| weather-F.NOM.SG | have.PRS.3SG | huge-N.ACC.SG | influence-N.ACC.SG | on | mood-N.ACC.SG |

*liud-ei.*
people-M.GEN.PL.

'The weather has a huge impact on people's moods.'

**V. 4.5 *Iš*** Another phenomenon that appeared among the most challenging ones was *4.5 Iš*. This phenomenon also consists of only one error type and is therefore identical to *4.5.1 Iš should not be used with certain words to indicate the date of writing*. For this error type, the models reached an average accuracy of 0.44 (see Section 7.3.2), while human participants did not find it particularly challenging (see Section 6.4.2). There were no sentences with reversed ratings (where the ungrammatical version was rated higher than the grammatical one), and, in general, participants rated the grammatical sentences as correct with a mean rating of 4.23 and the ungrammatical ones as incorrect with a mean rating of 2.33. This error type ranked $7^{th}$ across the most challenging error types for the models (see Appendix E.2, Table E.1).

This error type is again likely influenced by other languages, such as Russian. In Russian, it is grammatical to use the preposition *ot* in sentences such as *Neobkhodimo proverit' protokol zasedaniia ot 5 marta* (see Example 35), meaning *The protocol of March 5th needs to be checked*, whereas in Lithuanian it is ungrammatical to use the preposition *iš*, as in *Reikia patikrinti protokolą iš kovo 5 dienos* (see Example 36). Instead, the preposition should be omitted, resulting in *Reikia patikrinti kovo 5 dienos protokolą* (see Example 37).

(35) **Russian**

*Neobkhodim-o* *proveri-t'* *protokol* *zasedani-ia* *ot* *5* *mart-a.*
necessary-PRED check-INF protocol-M.ACC.SG meeting-N.GEN.SG from 5 March-M.GEN.SG.

'The protocol of March 5th need to be checked.'

(36) **Lithuanian**

**Reik-ia* *pa-tikrin-ti* *protokol-ą* *iš* *kov-o* *5* *dien-os.*
need.PRS.3 PFV-check-INF protocol-M.ACC.SG from March-M.GEN.SG 5 day-F.GEN.SG.

'The protocol of March 5th need to be checked.'

(37) **Lithuanian**

*Reik-ia* *pa-tikrin-ti* *kov-o* *5* *dien-os* *protokol-ą.*
need.PRS.3 PFV-check-INF March-M.GEN.SG 5 day-F.GEN.SG protocol-M.ACC.SG.

'The protocol of March 5th need to be checked.'

### Other Noteworthy Error Types

This subsection briefly examines five additional error types that were selected for analysis not because of their ranking among the most challenging error types for the models, but because of uneven performance patterns across the models.

**3.4.2 Accusative case should not be used to express the direct object next to a negative verb (instead of the genitive).** Error type *3.4.2 Accusative case should not be used to express the direct object next to a negative verb (instead of the genitive)* is worth discussing in more detail. This error type was relatively easy for the human participants: grammatical sentences received an average rating of 5.63, while ungrammatical sentences received an average rating of 1.67 (see Section 6.4.2). Although this error type was of moderate difficulty for the models and ranked $21^{st}$ among the most

challenging error types (see Appendix E.2, Table E.1), with a mean accuracy of 0.56 across systems (see Section 7.3.1), it is notable that only the four monolingual `Neurotechnology` models achieved perfect accuracy (1.0).

This error type involves a linguistic phenomenon known as *genitive of negation*, in which the genitive case is used with a negated verb instead of the accusative, which would be used in affirmative contexts (Talmina & Linzen, 2020). Thus, when the verb is negated, the genitive is required, as in *Nedarinėk durų* (see Example 38), which translates into English as *Don't keep opening the door.* However, speakers sometimes produce the accusative instead, as in Example 39: *Nedarinėk duris.* By contrast, in affirmative contexts, the accusative is expected, as in *Darinėk duris* (see Example 40).

(38) *Ne-dari-nė-k* *dur-ų!*
NEG-open-ITER-IMP.2SG door-F.GEN.PL!

'Don't keep opening the door!'

(39) **Ne-dari-nė-k* *dur-is!*
NEG-open-ITER-IMP.2SG door-F.ACC.PL!

'Don't keep opening the door!'

(40) *Dari-nė-k* *dur-is!*
open-ITER-IMP.2SG door-F.ACC.PL!

'Keep opening the door!'

This syntactic alternation is attested across the Balto-Slavic language family and is also present in Lithuanian (Pirnat, 2015). However, its realization differs across languages. For instance, in Russian, *genitive of negation* can be influenced by semantic factors, whereas in Lithuanian it is generally considered a purely syntactic phenomenon (Sigurðsson & Šereikaitė, 2023). This distinction may help explain why monolingual models from the `Neurotechnology` family perform particularly well on this error type: during training, they may more easily learn the language-specific syntactic trigger for genitive marking under negation. More broadly, this may suggest that the models acquire stable syntactic patterns more easily than semantic or context-dependent constraints.

**3.5.2 Instrumental case should not be used to express content or quality with adjectives denoting abundance.** Another error type, namely *3.5.2 Instrumental case should not be used to express content or quality with adjectives denoting abundance*, was not particularly difficult for the models, as it ranked $31^{st}$ among all error types (see Appendix E.2, Table E.1), with an average accuracy of 0.62 (see Section 7.3.1). However, this error type stands out because the monolingual models, `Goldfish` and `Neurotechnology`, achieved only moderate performance, with a maximum accuracy of 0.55, whereas global-scope models, such as those from the `OLMo` and `LLaMA 3` families, reached near-perfect or perfect accuracies (0.82–1.00).

An additional surprising observation is that several of these high-performing models ranked in the lower half of the overall accuracy table across all linguistic phenomena. Moreover, these model families do not officially support Lithuanian.

One possible explanation is that this error type could be influenced by patterns observed in other morphologically rich languages, particularly Slavic languages such as Polish and Russian, where similar

adjective–case alternations occur (Lietuvių kalbos išteklių informacinė sistema „E. kalba“, 2026). However, no direct claim can be made that these languages influenced the models, since they are not explicitly listed among their training languages.

Furthermore, according to the Collins English Dictionary (Collins English Dictionary, 2026), the expression *rich with* denotes being *well supplied* or *abundant*. The preposition *with* corresponds to the Lithuanian preposition *su*, which governs the instrumental case. Therefore, it is unlikely that the observed pattern in English-centric models is due to direct transfer from English. In particular, a straightforward transfer account would predict a stronger preference for instrumental constructions in Lithuanian; however, the results instead show a consistent preference for genitive alternatives in this error type.

Another important observation is that all three sentence pairs used in the human acceptability study exhibited a reversed pattern, where the ungrammatical sentences were rated higher than the grammatical ones (see Section 6.4.2). One of these pairs was taken from the primary source (Valstybinė lietuvių kalbos komisija (VLKK), 2023), which reduces the likelihood that the dataset itself is incorrect. In this case, participants rated the ungrammatical sentence *Jis dosnus pažadais* (see Example 41), which translates into English as *He is generous with promises*, as more acceptable (average rating: 5.4) than the grammatical sentence *Jis dosnus pažadų* (see Example 42) (average rating: 3.2).

(41) **J-is* *dosn-us* *pažad-ais.*
he-M.NOM generous-M.NOM.SG promise-M.INS.PL.

'He is generous with promises.'

(42) *Jis* *dosn-us* *pažad-ų.*
he-M.NOM generous-M.NOM.SG promise-M.GEN.PL.

'He is generous with promises.'

Overall, human acceptability judgements suggest that this error type is still highly relevant and systematically challenging for human speakers. However, the fact that several non-Lithuanian-specialized models perform exceptionally well remains unexpected and needs further investigation.

**3.6.6 Locative case should not be used to express a time period as a preposition or postposition.** Error type *3.6.6 Locative case should not be used to express a time period as a preposition or postposition* was also challenging for the models; it ranked $11^{th}$ among the most difficult error types (see Appendix E.2, Table E.1), with a mean accuracy across the models of 0.48 (see Section 7.3.1). Within its phenomenon type, namely Phenomenon Type 1 concerning the *Use of Cases*, this error type was the second most challenging. However, for the participants, this error type was not difficult. The grammatical sentences received an average rating of 5.4, while the ungrammatical ones received an average rating of 2.63 (see Section 6.4.2).

Native Lithuanian speakers sometimes make mistakes by incorrectly saying *Sprendimas gali būti apskundžiamas dešimt dienų laike* (see Example 43), where the locative case is used with *laikas* (in English, *time*). However, the grammatical alternative would be *Sprendimas gali būti apskundžiamas per dešimt dienų* (see Example 44), meaning *The decision can be appealed within ten days.* This error type is another example of interference from other languages, such as Russian. In Russian, the equivalent sentence would be *Reshenie mozhet byt' obzhalovano v techenie desiati dnei*, as in Example 45, where *v*

*techenie* governs the locative case.

(43) **Lithuanian**

**Sprendim-as* *gal-i* *bū-ti* *ap-skundž-iam-as* *dešimt*
decision-M.NOM.SG can-PRS.3 be-INF PFV-appeal-PRS.PASS.PTCP-M.NOM.SG ten
*dien-ų* *laik-e.*
day-F.GEN.PL period-M.LOC.SG

'The decision can be appealed within ten days.'

(44) **Lithuanian**

*Sprendimas* *gali* *būti* *apskundžiamas* *per* *dešimt*
decision-M.NOM.SG can-PRS.3 be-INF PFV-appeal-PRS.PASS.PTCP-M.NOM.SG during ten
*dienų.*
day-F.GEN.PL.

'The decision can be appealed within ten days.'

(45) **Russian**

*Resheni-e* *mozhe-t* *byt'* *obzhalova-n-o* *v* *techeni-e*
decision-N.NOM.SG can-PRS.3SG be-INF appeal-PST.PASS.PTCP-N.SG in course-N.LOC.SG
*desiat-i* *dnei.*
ten-GEN day-M.GEN.PL

'The decision can be appealed within ten days.'

Therefore, it is possible that the models performed poorly on this error type because they overrelied on examples from other languages in which such a construction is grammatical. However, the monolingual `Goldfish` model family clearly outperformed the other models, with accuracies ranging from 0.8 to a perfect 1.0, indicating that this model family is well equipped to handle this Lithuanian-specific error type. By contrast, it remains unclear why another monolingual model family, namely `Neurotechnology`, performed relatively poorly, with all four models achieving an accuracy of only 0.4.

**5.5.1 Negative predicate forms should not to be used in subordinate clauses expressing concession, when actual negation is not meant.** This error type, namely *5.5.1 Negative predicate forms should not be used in subordinate clauses expressing concession when actual negation is not intended*, ranked as the $8^{th}$ most challenging error type for the models (see Appendix E.2, Table E.1), with a mean accuracy across the models of 0.45 (see Section 7.3.3). However, this error type did not cause difficulties for the participants, who rated the grammatical sentences at an average of 5.2 and the ungrammatical ones at 2.13 (see Section 6.4.2).

Although this error type was generally challenging for the models, it did not cause problems for the monolingual models. All four models in the `Neurotechnology` family achieved a score of 1.0, while the performance of the `Goldfish` models was also strong: the two largest models achieved a score of 1.0, and the remaining two scored 0.7 and 0.9. These results stand out compared to the other models, clearly indicating a substantial difference between monolingual models and those trained on a broader

range of languages.

This error type is also likely influenced by other languages, such as Russian. In Russian, it is grammatical to say *Vo chto by to ni stalo, tebe nuzhno idti* (see Example 46), which, like the ungrammatical Lithuanian sentence *Kaip ten nebūtų, reikia važiuoti* (see Example 47), contains a negative predicate and translates as *No matter what, (someone) has to go.* However, in Lithuanian, negative predicates are not acceptable in such concessive constructions; instead, the correct form is *Kad ir kaip ten būtų, reikia važiuoti*, as shown in Example 48.

(46) **Russian**

*Vo chto by to ni sta-l-o, teb-e nuzhn-o id-ti.*
into what COND that NEG become-PST-N.SG you-DAT.SG necessary-N.SG go-INF.

'No matter what, (someone) has to go.'

(47) **Lithuanian**

**Kaip ten ne-bū-tų, reik-ia važiuo-ti.*
how there NEG-be-COND.3, need-PRS.3 go-INF.

'No matter what, (someone) has to go.'

(48) **Lithuanian**

*Kad ir kaip ten bū-tų, reik-ia važiuo-ti.*
that and how there be-COND.3, need-PRS.3 go-INF.

'No matter what, (someone) has to go.'

Therefore, the fact that monolingual models performed significantly better than the other models suggests that they may have been exposed to sufficient native Lithuanian data, whereas the other models may have learned cross-linguistic patterns that, when transferred into Lithuanian, result in ungrammatical constructions.

**5.6.1 Simple forms of the subjunctive mood should not to be used instead of compound forms to express past actions (often hypothetical or unrealized).** Another error type worth mentioning as particularly noteworthy is *5.6.1 Simple forms of the subjunctive mood should not be used instead of compound forms to express past actions (often hypothetical or unrealized).* This error type did not cause difficulties for the human participants, with grammatical sentences receiving a mean rating of 5.17 and ungrammatical sentences 2.97 (see Section 6.4.2). In contrast to the other error types discussed in this chapter, this was the second easiest for the models, with a mean accuracy of 0.9 across systems (see Section 7.3.3). More than half of the models (41 out of 78) achieved perfect accuracy (1.0). However, the monolingual model family `Goldfish` performed substantially worse, with accuracies ranging between 0.3 and 0.7.

It is well known that language models tend to prioritize high-frequency items (Martinez et al., 2024), and this may also be the case here. For example, in the grammatical sentence *Jeigu jie būtų važiavę mažesniu greičiu, avarija nebūtų įvykusi* (see Example 49), which translates into English as *If they had been driving at a lower speed, the accident would not have happened*, the form *būtų važiavę* is a more morphologically complex construction and therefore less frequent in usage. The `Goldfish`

language models, which are extremely small compared to the other models in this study (39M and 125M parameters, trained on 5–1000MB of data), may struggle with such morphologically complex and low-frequency constructions. Consequently, they may prefer the ungrammatical variant *Jeigu jie važiuotų mažesniu greičiu, avarija nebūtų įvykusi* (see Example 50), which contains the more frequent form *važiuotų*.

(49) **Jeigu jie bū-tų važiav-ę mažesn-iu greič-iu,*
if they be-COND.3 go-PST.PTCP.M.PL smaller-M.INS.SG speed-M.INS.SG,
*avarij-a ne-bū-tų į-vyk-usi.*
accident-F.NOM.SG NEG-be-COND.3 PFV-happen-PST.PTCP.F.SG.

'If they had been driving slower, the accident would not have happened.'

(50) *Jeigu jie važiuo-tų mažesn-iu greič-iu, avarij-a ne-bū-tų*
if they go-COND.3 smaller-M.INS.SG speed-M.INS.SG, accident-F.NOM.SG NEG-be-COND.3
*į-vyk-usi.*
PFV-happen-PST.PTCP.F.SG.

'If they had been driving slower, the accident would not have happened.'

### 8.2.3 Challenges in Human Acceptability Judgments

Out of the 192 sentence pairs evaluated by humans, 24 were rated contrary to expectations, meaning that the ungrammatical sentences in those pairs received higher ratings than the grammatical ones. Of these 24 sentence pairs, eight belonged to the error types already discussed in this chapter, which were identified as the most problematic for the models. Of the remaining 16 sentence pairs, seven were taken directly from the original source and an additional two from other online sources, indicating that these items were indeed grammatical. Therefore, the corresponding error types, namely *3.2.1 Genitive case should not be used to denote the object of an action with certain verbs*, *3.3.2 Dative case should not be used to describe a thing/object when purpose is not being expressed*, *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)*, *3.6.5 Locative case should not be used to express the manner or timing of an action*, *4.6.1 Pas should not be used to indicate ownership, belonging, or possession*, *5.1.3 The masculine accusative and instrumental forms of adjectives and of participles used with adjectival meaning are not to be used in the so-called "indefinite gender" meaning*, *5.8.1 Active present participle should not be used with nouns that do not denote an agent (nouns that cannot perform an action)*, and *5.8.2 Active past participle should not be used to indicate an action occurring simultaneously with or subsequent to the main verb*, appear to pose genuine challenges even for native speakers of Lithuanian (the full list can be found in Appendix E.2, Table E.2). Among these error types, only *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)* was also highly challenging for the models, ranking $12^{th}$ among the most difficult error types (see Appendix E.2, Table E.1), with a mean accuracy of 0.49 across the models (see Section 7.3.1).

However, one sentence pair that proved problematic for the participants, namely from error type *5.7.1 Reflexive forms of certain verbs should not to be used with a passive meaning if the action cannot occur by itself*, was problematic due to inaccuracies in the dataset itself. The dataset contains three sentence pairs (one of which was evaluated by humans) involving the verbs *seen*, *heard*, and *felt*,

which constitute exceptions to error type *5.7.1.* As a result, both the "grammatical" variant, such as the one shown in Example 51, and the "ungrammatical" variant (see Example 52) are, in fact, grammatical. Therefore, rather than preferring an ungrammatical sentence, the participants merely showed a preference for an alternative grammatical construction. Nevertheless, these examples should be excluded from future versions of the dataset or, at minimum, supplemented with additional sentence pairs that better represent this error type.

(51) *Už* *sien-os* *buvo* *girdė-t-i* *bals-ai.*
behind wall-F.GEN.SG be.PST.3 hear-PST.PASS.PTCP-M.NOM.PL voice-M.NOM.PL.

'Voices were heard behind the wall.'

(52) *Už* *sien-os* *girdėj-o-si* *bals-ai.*
behind wall-F.GEN.SG hear-PST.3-REFL voice-M.NOM.PL.

'Voices were heard behind the wall.'

Finally, six problematic sentence pairs were semi-automatically generated. These belonged to the following error types: *3.5.1. Instrumental case should not be used to express the object with verbs denoting fullness or increase*, *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*, *4.1.4 Ant should not be used to indicate dimensions or a ratio of sizes*, *5.1.3 The masculine accusative and instrumental forms of adjectives and of participles used with adjectival meaning are not to be used in the so-called "indefinite gender" meaning*, *5.3.1 The pronominal forms (the forms with possessive endings) of ordinal numerals that denote decades are not to be used to indicate decades*, and *5.10.1 Adverbial participle should not be used to indicate a secondary action of the same agent in personal sentences (instead of a half-participle or participle with proper agreement).* These semi-automatically generated examples do not appear to contain any actual errors, and it is therefore unclear why they were difficult for the human participants. Importantly, two of the corresponding error types, namely *3.5.3* and *5.10.1*, ranked $13^{th}$ and $14^{th}$ among the most challenging error types for the models (see Appendix E.2, Table E.1), both with a mean accuracy of 0.50 across the models (see Sections 7.3.1 and 7.3.3). This suggests that these examples were challenging not only for human participants but also for the models.

### 8.2.4 Summary of Key Findings

In summary, most of the error types discussed in this chapter appear in Lithuanian due to cross-linguistic interference, especially (though not exclusively) from Slavic languages. It is therefore hypothesised that the low model performance is also related to cross-linguistic transfer, whereby multilingual language models may overgeneralise based on languages with more training data, particularly for high-frequency constructions.

Moreover, the error analysis revealed several additional factors that may have significantly affected model performance. The models may prefer less complex constructions (error type *5.6.1 Simple forms of the subjunctive mood should not to be used instead of compound forms to express past actions (often hypothetical or unrealized)*) and may also show a bias towards more frequent usage patterns (error type *3.1.2 Nominative case should not be used to express direct address*).

Another observation from the error analysis concerns imperfections in dataset quality. While carefully handcrafted, the dataset was not free from human error, as illustrated by error type *5.2.1*

*The basic quantitative numeral forms from 2 to 9 and the pronoun abu, abi are not to be used with plural-only nouns (nouns that exist only in the plural).* Moreover, in some cases the dataset included only one variant out of several possible grammatical alternatives, which was not necessarily the most natural or most frequently used form (error type *4.9.1 Prieš should not be used to express the object of an action with certain words*), while in other cases it randomly selected between available variants, thereby creating unequal item difficulty (error type *4.8.2 Prie should not be used to express exchange ratios/conversion rates*).

Human acceptability ratings confirmed that, as discussed in the Results chapter (see Chapter 7), item difficulty varied significantly across the dataset. However, some model behaviour patterns remain unexplained, such as several reversed human acceptability judgements as well as error type *3.5.2 Instrumental case should not be used to express content of quality with adjectives denoting abundance.*

# 9 Conclusion

In this thesis, I investigated the syntactic competence of language models in Lithuanian, a low-resource and morphologically rich language that remains underrepresented in NLP evaluation. To this end, I introduced a manually curated minimal-pair dataset covering a broad range of syntactic phenomena and error types derived from attested patterns in human language use. The study evaluated how well current language models distinguish between grammatical and ungrammatical Lithuanian sentences and how model properties such as size, training data coverage, and instruction tuning affect performance. In addition, a human acceptability judgment study was designed and conducted to assess alignment between model predictions and native speaker intuitions. All resources produced in this work are publicly released[1].

The results show that while language models capture certain aspects of Lithuanian syntax, their performance is highly variable and strongly constrained by linguistic phenomenon, error type, and item difficulty. Monolingual models consistently outperform multilingual models, and larger models achieve higher accuracy overall. In contrast, instruction tuning does not yield statistically significant improvements. A comparison with the existing benchmark further indicates that the proposed dataset provides a substantially more challenging evaluation setting, revealing weaknesses that are not apparent in prior work. In addition, the extensive qualitative error analysis identifies the linguistic phenomena and error types that pose the greatest challenges for the models, while also highlighting potential directions for future research and dataset improvement.

Overall, this work provides the first known systematic and linguistically grounded evaluation of modern language models on Lithuanian syntax based on minimal sentence pairs, focusing on the most challenging Lithuanian-specific syntactic phenomena. It makes a substantial contribution by extending the existing foundation for more rigorous future benchmarking in low-resource, morphologically rich languages.

## 9.1 Limitations

The study is limited by the relatively small dataset size, which restricts statistical power and fine-grained analysis at the level of individual error types, as each error type is represented by only a small number of items. In addition, the evaluation is restricted to open-weight models, excluding proprietary systems that may exhibit different performance patterns. Finally, the human judgment dataset is also limited in size, with only three items per error type, which restricts more detailed human–model comparisons.

[1] https://github.com/urtuteja/Evaluating-The-Syntactic-Knowledge-of-Language-Models-on-Lithuanian

## 9.2 Future Work

Future work should primarily focus on scaling and refining the dataset. Expanding coverage through additional naturally occurring examples and expert linguistic validation would improve robustness and increase statistical power. Several issues identified in the error analysis also indicate that dataset revision is necessary before reuse.

A second priority is extending the evaluation to closed-source and state-of-the-art proprietary models, which would enable stronger and more generalizable conclusions about model capabilities. Increasing the number of human judgments per item would further improve the reliability of human-model comparisons.

Furthermore, time permitting, it would be beneficial to further expand the qualitative linguistic error analysis. Out of 64 error types, 14 achieved mean accuracy at or below chance level across the models. The current analysis focuses on nine of these, with an additional three error types that were easier for the models. In future work, it would therefore be beneficial to review the remaining error types for additional sources of error. Moreover, it would be worthwhile to conduct small-scale studies on error types that involve multiple possible grammatical variants, in order to further clarify directions for dataset improvement.

Finally, future research should move beyond isolated factors and investigate interactions between variables such as model size, language coverage, and training regime. A more comprehensive experimental design would enable a deeper understanding of which factors most strongly shape syntactic competence in language models.

# Bibliography


Adeeba, F., Dillon, B., Sajjad, H., & Bhatt, R. (2025). *UrBLiMP: A Benchmark for Evaluating the Linguistic Competence of Large Language Models in Urdu*. (Cited on page 11).

Ali, M., Fromm, M., Thellmann, K., Ebert, J., Weber, A. A., Rutmann, R., Jain, C., Lübbering, M., Steinigen, D., Leveling, J., Klug, K., Buschhoff, J. S., Jurkschat, L., Abdelwahab, H., Stein, B. J., Sylla, K.-H., Denisov, P., Brandizzi, N., Saleem, Q., … Flores-Herr, N. (2024). *Teuken-7B-Base Teuken-7B-Instruct: Towards European LLMs*. (Cited on page 26).

Ambrazas, V. (2026a). *Asmenavimas*. *Visuotinė lietuvių enciklopedija* (cited on page 12).

Ambrazas, V. (2026b). *Linksnis*. *Visuotinė lietuvių enciklopedija* (cited on page 11).

Ambrazas, V. (2026c). *Šauksmininkas*. *Visuotinė lietuvių enciklopedija* (cited on page 66).

Ambrazas, V. (2026d). *Skaičius*. *Visuotinė lietuvių enciklopedija* (cited on page 12).

Ambrazas, V. (2026e). *Žodžių tvarka*. *Visuotinė lietuvių enciklopedija* (cited on page 12).

An, C., Zhang, J., Zhong, M., Li, L., Gong, S., Luo, Y., Xu, J., & Kong, L. (2024). *Why Does the Effective Context Length of LLMs Fall Short?* (Cited on page 9).

Ármannsson, B., Ingimundarson, F. Á., & Sigurðsson, E. F. (2025). *An Icelandic Linguistic Benchmark for Large Language Models*. *Proceedings of the Joint 25th Nordic Conference on Computational Linguistics and 11th Baltic Conference on Human Language Technologies (NoDaLiDa/Baltic-HLT 2025)*, 37–47 (cited on page 11).

Baayen, R. H., Dijkstra, T., & Schreuder, R. (1997). *Singulars and plurals in Dutch: Evidence for a parallel dual-route model*. *Journal of Memory and Language*, *37*, 94–117 (cited on page 64).

Barbini, M., Piccini Bianchessi, M. L., Bressan, V., Fusco, A., Neri, S., Rossi, S., Sgrizzi, T., & Chesi, C. (2025). *BLiMP-IT: Harnessing Automatic Minimal Pair Generation for Italian Language Model Evaluation*. *Proceedings of the Eleventh Italian Conference on Computational Linguistics (CLiC-it 2025)*, 64–71 (cited on page 11).

Başar, E., Padovani, F., Jumelet, J., & Bisazza, A. (2025). *TurBLiMP: A Turkish Benchmark of Linguistic Minimal Pairs*. (Cited on pages 11, 14).

Bergmanis, T., Kronis, M., Pretkalniņš, I. J., Nicmanis, D., Jeļinska, J., Rozis, R., Vıksna, R., & Pinnis, M. (2026). *TildeOpen LLM: Leveraging Curriculum Learning to Achieve Equitable Language Representation*. (Cited on page 26).

Bielinskienė, A., Kazlauskienė, A., Rimkutė, E., & Tamošiūnaitė, A. (2014). *Lietuvių bendrinė kalba: Normos ir vartosena*. Vytauto Didžiojo universitetas. (Cited on page 63).

Braga, M., Milanese, G. C., & Pasi, G. (2025). *Investigating Large Language Models' Linguistic Abilities for Text Preprocessing*. (Cited on page 4).

Bryskina, V., Songailaitė, M., & Madravickaitė, J. (2025). *Evaluating Bias Detection in Lightweight LLMs*. *Vilnius University Open Series* (cited on page 12).

Budreikienė, J. (2018). *Lietuvių kalbos atmintinė kariams*. Generolo Jono Žemaičio Lietuvos karo akademija. (Cited on page 66).

Chang, T. A., Arnett, C., Tu, Z., & Bergen, B. K. (2024). *Goldfish: Monolingual Language Models for 350 Languages*. (Cited on pages 12, 26).

Chomsky, N. (1965). *Aspects of the theory of syntax*. MIT Press. (Cited on page 14).

Clark, P., Cowhey, I., Etzioni, O., Khot, T., Sabharwal, A., Schoenick, C., & Tafjord, O. (2018). *Think you have Solved Question Answering? Try ARC, the AI2 Reasoning Challenge*. (Cited on page 12).

Cobbe, K., Kosaraju, V., Bavarian, M., Chen, M., Jun, H., Kaiser, L., Plappert, M., Tworek, J., Hilton, J., Nakano, R., Hesse, C., & Schulman, J. (2021). *Training verifiers to solve math word problems*. (Cited on page 12).

Collins English Dictionary. (2026). *Definition of "rich"* (cited on page 71).

Corbett, G. G. (2019). *Pluralia tantum nouns and the theory of features: a typology of nouns with non-canonical number properties*. *Morphology*, *29*(1), 51–108 (cited on page 64).

Dang, J., Singh, S., D'souza, D., Ahmadian, A., Salamanca, A., Smith, M., Peppin, A., Hong, S., Govindassamy, M., Zhao, T., Kublik, S., Amer, M., Aryabumi, V., Campos, J. A., Tan, Y.-C., Kocmi, T., Strub, F., Grinsztajn, N., Flet-Berliac, Y., ... Hooker, S. (2024). *Aya Expanse: Combining Research Breakthroughs for a New Multilingual Frontier*. (Cited on pages 12, 26).

De Winter, J. C. F., Gosling, S. D., & Potter, J. (2024). *Comparing the Pearson and Spearman Correlation Coefficients Across Distributions and Sample Sizes: A Tutorial Using Simulations and Empirical Data*. (Cited on page 52).

DeepSeek-AI. (2025). *DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning*. (Cited on page 8).

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding* (cited on page 7).

Dryer, M. S. (2013). *Feature 81A: Order of Subject, Object and Verb*. *The World Atlas of Language Structures Online* (cited on page 12).

Ethnologue. (2026). *What Are the Top 200 Most Spoken Languages?* (Cited on page 60).

Fu, Y., Peng, H., Ou, L., Sabharwal, A., & Khot, T. (2023). *Specializing Smaller Language Models towards Multi-Step Reasoning*. (Cited on page 27).

Gallagher, D., & Heyer, G. (2026). *Targeted Syntactic Evaluation of Language Models on Georgian Case Alignment*. *Proceedings of the Second Workshop on Language Models for Low-Resource Languages (LoResLM 2026)*, 259–270 (cited on page 66).

Gemma Team. (2025). *Gemma 3 Technical Report*. (Cited on pages 12, 26).

Ghaseminejad Raeini, M. (2025). *The evolution of language models: From N-Grams to LLMs, and beyond*. *Natural Language Processing Journal*, *12* (cited on page 7).

Gonzalez-Agirre, A., Pàmies, M., Llop, J., Baucells, I., Da Dalt, S., Tamayo, D., Saiz, J. J., Espuña, F., Prats, J., Aula-Blasco, J., Mina, M., Pikabea, I., Rubio, A., Shvets, A., Sallés, A., Lacunza, I., Palomar, J., Falcão, J., Tormo, L., … Villegas, M. (2025). *Salamandra Technical Report*. (Cited on page 26).

Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D., & Steinhardt, J. (2020). *Measuring Massive Multitask Language Understanding*. (Cited on page 12).

Hogan-Brun, G., Ramonienė, M., & Grumadienė, L. (2005). *The language situation in Lithuania*. *Journal of Baltic Studies*, *36*(3), 345–370 (cited on page 11).

Holvoet, A. (2026). *Giminė*. *Visuotinė lietuvių enciklopedija* (cited on page 11).

Jačauskas, I. (2019). *Lietuvių kalbos komisija atsisakė Didžiųjų kalbos klaidų sąrašo*. (Cited on page 18).

Jarvis, S., & Pavlenko, A. (2008). *Crosslinguistic Influence in Language and Cognition*. Routledge. (Cited on page 60).

Joshi, A., Kale, S., Chandel, S., & Pal, D. K. (2015). *Likert Scale: Explored and Explained*. *British Journal of Applied Science  Technology*, *7*(4), 396–403 (cited on page 14).

Jumelet, J., Weissweiler, L., Nivre, J., & Bisazza, A. (2025). *MultiBLiMP 1.0: A Massively Multilingual Benchmark of Linguistic Minimal Pairs*. (Cited on pages 4–6, 11–13, 57).

Jurafsky, D., & Martin, J. H. (2026). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition with Language Models* (3rd edition). (Cited on pages 7–9, 30).

Kahane, S., Peng, Z., & Gerdes, K. (2023). *Word order flexibility: a typometric study*. *Proceedings of the Seventh International Conference on Dependency Linguistics (Depling, GURT/SyntaxFest 2023)*, 68–80 (cited on page 12).

Kondratyuk, D., Gavenčiak, T., Straka, M., & Hajič, J. (2018). *LemmaTag: Jointly Tagging and Lemmatizing for Morphologically Rich Languages with BRNNs*. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 4921–4928 (cited on page 11).

Kostiuk, Y., Vitman, O., Gagała, Ł., & Kiulian, A. (2025a). *The Veln(ia)s is in the Details: Evaluating LLM Judgment on Latvian and Lithuanian Short Answer Matching*. (Cited on page 12).

Kostiuk, Y., Vitman, O., Gagała, Ł., & Kiulian, A. (2025b). *Towards Multilingual LLM Evaluation for Baltic and Nordic languages: A study on Lithuanian History*. (Cited on page 12).

Kreusler, A. (1961). *Bilingualism in Soviet Non-Russian Schools*. *The Elementary School Journal*, *62*(2), 94–99 (cited on page 60).

Kuoraitė, S., & Gružauskas, V. (2025). *Simplifying lithuanian text into easy-to-read language using large language models*. *Proceedings of the 1st Workshop on Artificial Intelligence and Easy and Plain Language in Institutional Contexts (AI  EL/PL)*, 30–37 (cited on page 12).

Lau, J. H., Clark, A., & Lappin, S. (2017). *Grammaticality, Acceptability, and Probability: A Probabilistic View of Linguistic Knowledge*. *Cognitive Science*, *41*(5), 1149–1418 (cited on page 14).

Lietuvių kalbos išteklių informacinė sistema „E. kalba“. (2026). *Turtingas, -a* (cited on page 71).

Likert, R. (1932). *A Technique for the Measurement of Attitudes*. *Archives of Psychology*, *140*, 1–55 (cited on page 14).

Lin, S., Hilton, J., & Evans, O. (2021). *TruthfulQA: Measuring How Models Mimic Human Falsehoods*. (Cited on page 12).

Llama Team. (2024). *The Llama 3 Herd of Models*. (Cited on pages 12, 26).

Luo, J., Zhang, W., Yuan, Y., Zhao, Y., Yang, J., Gu, Y., Wu, B., Chen, B., Qiao, Z., Long, Q., Tu, R., Luo, X., Ju, W., Xiao, Z., Wang, Y., Xiao, M., Liu, C., Yuan, J., Zhang, S., … Zhang, M. (2025). *Large Language Model Agent: A Survey on Methodology, Applications and Challenges*. (Cited on page 5).

Mandravickaitė, J., Rimkienė, E., Kapkan, D. K., Kalinauskaitė, D., Čenys, A., & Krilavičius, T. (2025). *Automatic Text Simplification for Lithuanian: Transforming Administrative Texts into Plain Language*. *Mathematics*, *13*(3), 465 (cited on page 12).

Marijampolės savivaldybė. (2017). *Kaip kreiptis? Kreipinio vartojimo atmintinė* (cited on page 20).

Martinez, D., Goriely, Z., Caines, A., Buttery, P., & Beinborn, L. (2024). *Mitigating Frequency Bias and Anisotropy in Language Model Pre-Training with Syntactic Smoothing*. *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 5999–6011 (cited on pages 11, 73).

Martins, P. H., Fernandes, P., Alves, J., Guerreiro, N. M., Rei, R., Alves, D. M., Pombal, J., Farajian, A., Faysse, M., Klimaszewski, M., Colombo, P., Haddow, B., de Souza, J. G. C., Birch, A., & Martins, A. F. T. (2024). *EuroLLM: Multilingual Language Models for Europe*. (Cited on pages 12, 26).

Marvin, R., & Linzen, T. (2018). *Targeted Syntactic Evaluation of Language Models*. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 1192–1202 (cited on pages 10, 11, 14).

Max Planck Institute for Evolutionary Anthropology. (2015). *The Leipzig Glossing Rules: Conventions for interlinear morpheme-by-morpheme glosses* (cited on page 21).

McGiff, J., Tran, K.-T., Mulcahy, W., Dáibhidh Ó Luinín, Dalzell, J., Róisín Ní Bhroin, Burke, A., O'Sullivan, B., Nguyen, H. D., & Nikolov, N. S. (2025). *Irish-BLiMP: A Linguistic Benchmark for Evaluating Human and Language Model Performance in a Low-Resource Setting*. (Cited on page 11).

Meta AI. (2025). *The Llama 4 herd: The beginning of a new era of natively multimodal AI innovation* (cited on pages 9, 26).

Meta GenAI. (2023). *Llama 2: Open Foundation and Fine-Tuned Chat Models*. (Cited on pages 26, 41).

Microsoft Team. (2024). *Phi-3 Technical Report: A Highly Capable Language Model Locally on Your Phone*. (Cited on page 26).

Miliūnaitė, Rita. (2003). *Lietuvių kalbos gramatikos norminimo pagrindai*. Lietuvių kalbos instituto leidykla. (Cited on page 60).

Mu, S., & Lin, S. (2025). *A Comprehensive Survey of Mixture-of-Experts: Algorithms, Theory, and Applications*. (Cited on page 9).

Munjal, P., Christophe, C., Rajan, R., & Kanithi, P. (2026). *Do instruction-tuned models always perform better than base models? evidence from math and domain-shifted benchmarks*. (Cited on page 5).

Nag, A., Chakrabarti, S., Mukherjee, A., & Ganguly, N. (2025). *Efficient Continual Pre-training of LLMs for Low-resource Languages*. *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 3: Industry Track)*, 304–317 (cited on page 11).

Nakvosas, A., Daniušis, P., & Mulevičius, V. (2024). *Open Llama2 Model for the Lithuanian Language*. (Cited on page 26).

Nenonen, M., & Niemi, J. (2010). *Mismatches between grammatical number and conceptual numerosity: A number-decision experiment on collective nouns, number neutralization, pluralia tantum, and idiomatic plurals*. *Folia Linguistica - FOLIA LINGUIST*, *44*, 103–125 (cited on page 64).

OpenAI. (2025). *GPT-5.1: A smarter, more conversational ChatGPT*. (Cited on page 19).

Ozsoy, M. G. (2024). *Multilingual Prompts in LLM-Based Recommenders: Performance Across Languages* (cited on page 9).

Pestel, J., Bloem, J., & Alhama, R. G. (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 (cited on pages 11, 35).

Pirnat, Ž. (2015). *Genesis of the Genitive of Negation in Balto-Slavic and Its Evidence in Contemporary Slovenian*. *Slovenski jezik – Slovene Linguistic Studies*, *10*, 3–52 (cited on page 70).

Plesevičius, D. (2025). *Large language models for lithuanian language* [Master's thesis]. Vilnius University, Faculty of Mathematics and Informatics [Supervisor: Prof. Aistis Raudys; Reviewer: Dr. Tomas Plankis]. (Cited on page 12).

Qwen Team. (2025). *Qwen3 Technical Report*. (Cited on pages 12, 26).

Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). *Improving Language Understanding by Generative Pre-Training* (cited on page 8).

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., & Liu, P. J. (2019). *Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer* (cited on page 8).

Rao, V. C. S., Rao, L. T., & Kumar, V. P. (2025). *English As the Language of Research and Worldwide Academic Journals*. *Journal for Research Scholars and Professionals of English Language Teaching*, *9*(47) (cited on page 4).

Revilla, M., & Höhne, J. K. (2020). *How Long Do Respondents Think Online Surveys Should Be? New Evidence from Two Online Panels in Germany*. *International Journal of Market Research*, *62*(5), 538–545 (cited on page 34).

Revilla, M., & Ochoa, C. (2017). *Ideal and Maximum Length for a Web Survey*. *International Journal of Market Research*, *59*(5), 557–565 (cited on page 34).

Sakaguchi, K., Bras, R. L., Bhagavatula, C., & Choi, Y. (2019). *WinoGrande: An Adversarial Winograd Schema Challenge at Scale*. (Cited on page 12).

Salhan, S. (2025). *Evaluating the Cross-Lingual Syntactic Capabilities of Language Models* [Unpublished lecture notes]. (Cited on page 24).

Schütze, C. T. (1996). *The Empirical Base of Linguistics: Grammaticality Judgments and Linguistic Methodology*. University of Chicago Press. (Cited on page 14).

Sigurðsson, E. F., & Šereikaitė, M. (2023). *The dual face of structural object case: on Lithuanian genitive of negation*. *Journal of Linguistics*, *60*(1), 161–212 (cited on page 70).

Singh, T. J., Singh, S. R., & Sarmah, P. (2023). *Subwords to word back composition for morphologically rich languages in neural machine translation*. *Proceedings of the 37th Pacific Asia Conference on Language, Information and Computation*, 691–700 (cited on page 11).

Someya, T., & Oseki, Y. (2023). *JBLiMP: Japanese Benchmark of Linguistic Minimal Pairs*. *Findings of the Association for Computational Linguistics: EACL 2023*, 1581–1594 (cited on page 11).

Spearman, C. (1904). *The Proof and Measurement of Association between Two Things*. *The American Journal of Psychology*, *15*, 72–101 (cited on pages 41, 51, 53).

Subramanian, S., Elango, V., & Gungor, M. (2025). *Small Language Models (SLMs) Can Still Pack a Punch: A Survey*. (Cited on page 5).

Suijkerbuijk, M., Prins, Z., de Heer Kloots, M., Zuidema, W., & Frank, S. L. (2025). *BLiMP-NL: A Corpus of Dutch Minimal Pairs and Acceptability Judgments for Language Model Evaluation*. *Computational Linguistics* (cited on pages 11, 14).

Taktasheva, E., Bazhukov, M., Koncha, K., Fenogenova, A., Artemova, E., & Mikhailov, V. (2024). *RuBLiMP: Russian Benchmark of Linguistic Minimal Pairs*. *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 9268–9299 (cited on page 11).

Talmina, N., & Linzen, T. (2020). *Neural network learning of the Russian genitive of negation: optionality and structure sensitivity*. *Society for Computation in Linguistics*, *3*(1), 199–208 (cited on page 70).

Team OLMo. (2024). *2 OLMo 2 Furious*. (Cited on pages 12, 26).

Temesgen, T. K., Marco, M. D., & Fraser, A. (2025). *Extracting Linguistic Information from Large Language Models: Syntactic Relations and Derivational Knowledge*. *Proceedings of the 2025*

*Conference on Empirical Methods in Natural Language Processing*, 27210–27226 (cited on page 13).

Urnėžiūtė, R. (2014). *Taisyklingos kalbos niekas neatšaukė*. *Gimtoji kalba*, (2), 18–26 (cited on page 18).

Valstybinė lietuvių kalbos komisija (VLKK). (2023). *Commission*. (Cited on pages 17–19, 24, 25, 62, 63, 71).

Valstybinė lietuvių kalbos komisija (VLKK). (2026). *Kada „trys“, o kada – „trejos (treji)“?* (Cited on page 63).

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). *Attention Is All You Need* (cited on page 7).

Vlantis, D., & Bloem, J. (2025). *Intrinsic Evaluation of Mono- and Multilingual Dutch Language Models*. *Computational Linguistics in the Netherlands Journal*, *14*, 525–553 (cited on page 5).

Vytauto Didžiojo universitetas (VDU). (2013). *Corpus of Contemporary Lithuanian Language*. (Cited on page 19).

Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., & Bowman, S. R. (2018). *GGLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding* (cited on page 10).

Wang, H., Minervini, P., & Ponti, E. M. (2024). *Probing the Emergence of Cross-lingual Alignment during LLM Training*. (Cited on page 60).

Warstadt, A., Parrish, A., Liu, H., Mohananey, A., Peng, W., Wang, S.-F., & Bowman, S. R. (2019). *BLiMP: A Benchmark of Linguistic Minimal Pairs for English*. (Cited on pages 4, 10, 14, 35).

Warstadt, A., Singh, A., & Bowman, S. R. (2019). *Neural Network Acceptability Judgments* (cited on pages 10, 14).

Wilson, E. B. (1927). *Probable Inference, the Law of Succession, and Statistical Inference*. *Journal of the American Statistical Association*, *22*(158), 209–212 (cited on page 41).

Wu, C., & Tang, R. (2024). *Performance Law of Large Language Models*. (Cited on page 9).

Xiang, B., Yang, C., Li, Y., Warstadt, A., & Kann, K. (2021). *CLiMP: A Benchmark for Chinese Language Model Evaluation*. *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, 2784–2790 (cited on page 11).

Zellers, R., Holtzman, A., Bisk, Y., Farhadi, A., & Choi, Y. (2019). *Hellaswag: Can a machine really finish your sentence?* (Cited on page 12).

Zhang, B., Cheng, Y., Shakeri, S., Wang, X., Ma, M., & Firat, O. (2025). *Encoder-Decoder or Decoder-Only? Revisiting Encoder-Decoder Large Language Model* (cited on page 8).

Zhang, X., Li, S., Hauer, B., Shi, N., & Kondrak, G. (2023). *Don't trust ChatGPT when your question is not in English: A study of multilingual abilities and types of LLMs*. *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 7915–7927 (cited on page 4).

# A Data Collection

## A.1 Additional Data Sources

The following sources were used to collect additional sentence pairs for the dataset.

1. Additional VLLK Resources
    - https://vlkk.lt/konsultacijos/978-galininkas-kilmininkas
    - https://vlkk.lt/konsultacijos/6875-jei-su-padalyviu
    - https://vlkk.lt/konsultacijos/6873-jeigu-su-bendratimi
    - https://vlkk.lt/konsultacijos/1138-ne-ilgiau-ko
    - https://vlkk.lt/konsultacijos/11178-ne-mazesnis-nemaziau-ne-maziau-nemaziau
    - https://vlkk.lt/konsultacijos/3345-neveikiamosios-rusies-dalyvis
    - https://vlkk.lt/konsultacijos/6756-vietoj-to-kad-bendratis
2. Municipal Government Websites
    - https://alytus.lt/uploads/documents/files/LT/savivaldybes-administracija/administracine-informacija/kalbos%20taisyklingumas/AdministracinÄŮs%20kalbos%20atmintinÄŮ%202019%2002%2018.pdf
    - https://www.mazeikiai.lt/savivalda/administracine-informacija/veiklos-sritys/kalbos-tvarkyba/rekomendacijos
    - https://salcininkai.lt/valstybine-kalba/linksniu-vartojimo-klaidos/1004
3. Educational Websites
    - https://cris.mruni.eu/server/api/core/bitstreams/b491751c-15a9-4d9d-bf95-347ef006c635/content
    - https://etalpykla.lituanistika.lt/object/LT-LDB-0001:J.04~2014~1516887943445/J.04~2014~1516887943445.pdf
    - https://www.kurstoti.lt/s/13594/dazniausios-sintakses-sakiniu-sandaros-klaidos-abiturientu-vbe-rasiniuose
    - https://mokslai.lietuviuzodynas.lt/lietuviu-kalba/kalbos-kulturos-taisykliu-savadas
    - https://www.nsa.smsm.lt/wp-content/uploads/2025/07/VERTINIMO-GAIRES-2022-koreguotos.pdf
    - https://talpykla.elaba.lt/elaba-fedora/objects/elaba:4622103/datastreams/MAIN/content
    - https://vki.lrv.lt/lt/naujienos/birzelio-ir-liepos-men-spaudos-leidiniuose-pastebetos-kalbos-klaidos/?lang=fr
    - http://web.vu.lt/flf
4. Other Resources
    - https://www.balticmaster.lt/priedai-ir-komplektuojanciosios-dalys/ (Ungrammatical title on a company's website)

- https://www.facebook.com/muzikantai.uk/videos/kaip-ten-nebÅńtÅşbet-dainos-turi-skambÄŮtiaÄĐiÅń-uÅ¿-nuostabÅş-vakarÄĚaÄĐiÅń-uÅ¿-gÄŮles-ir-d/1332936084366782/ (Ungrammatical Facebook post)
- https://www.santaka.info/index.php/2012/04/11/idomu-260/?srsltid=AfmBOopq-fLzMpuUYvd00K3YmEkuDkOW2uTkeNxWqYHTDSCKmxkMbadc (News article containing examples of grammatical and ungrammatical sentences)
- https://www.tv3.lt/naujiena/lietuva/kalbos-specialistai-pataria-sakykime-astuntojo-desimtmecio-mada-o-ne-70-uju-metu-n1098893 (News article containing examples of grammatical and ungrammatical sentences)

## A.2 Data Generation Prompt

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.]*

## A.3 Dataset Construction Rounds

The figures below (A.1, A.2, and A.3) show the distribution of sentence pairs across different phenomena after each of the three dataset creation rounds.

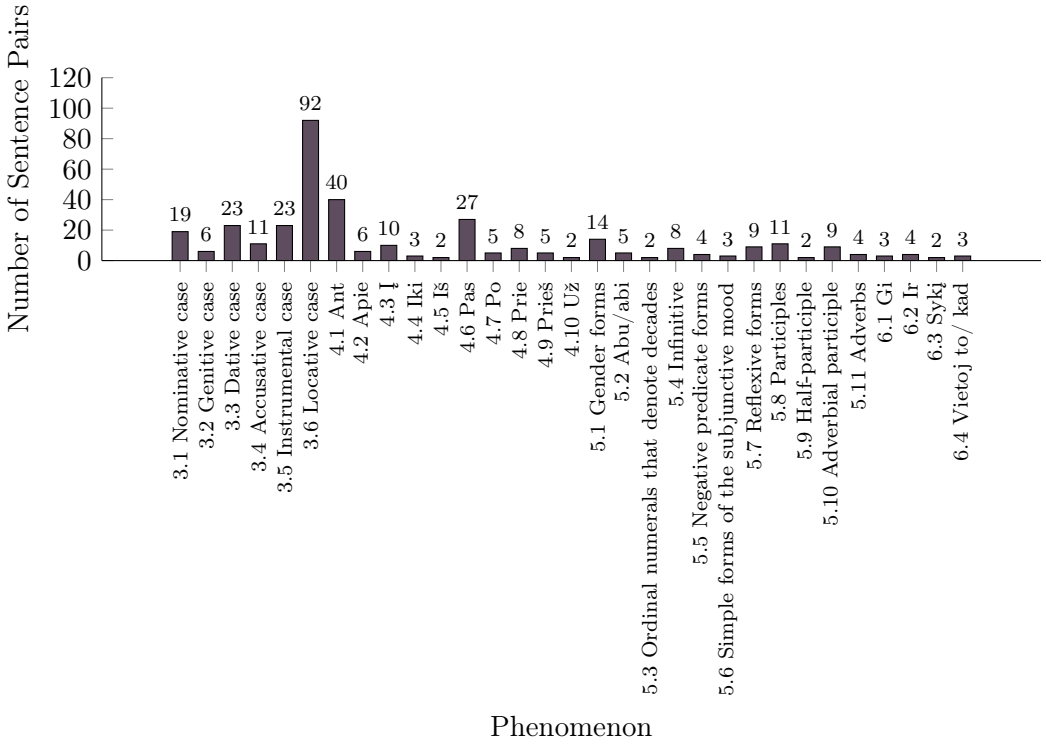


Figure A.1: Count of sentence pairs in Round 1 (original VLKK list).

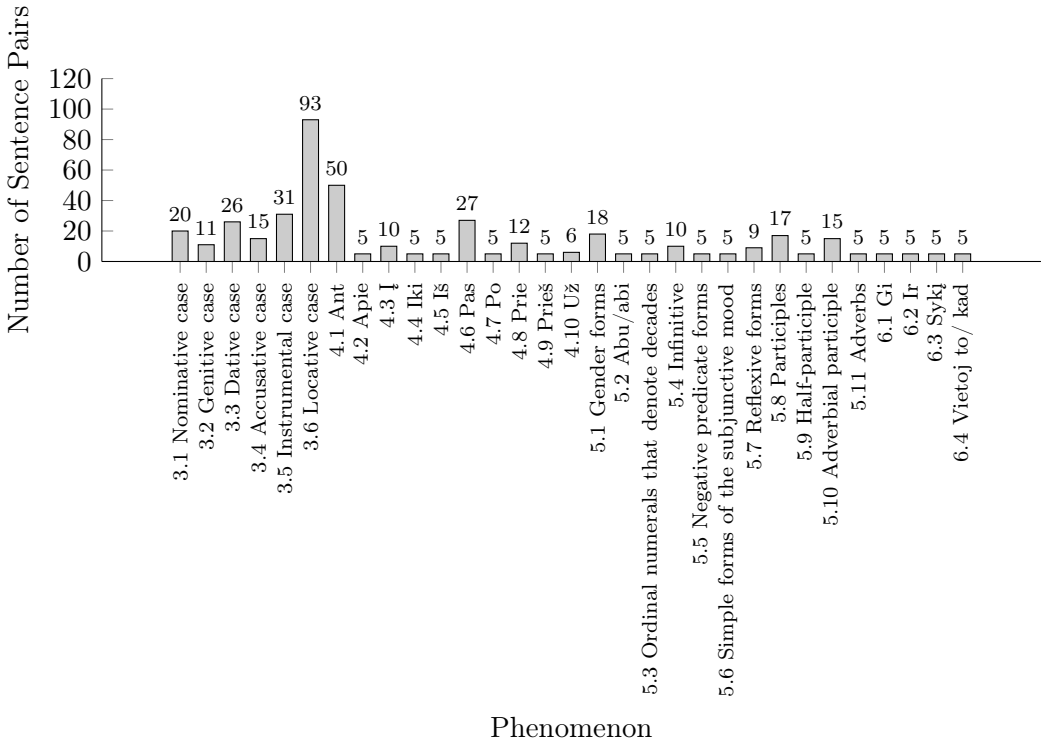


Figure A.2: Count of sentence pairs in Round 2 (each error has at least 5 sentence pairs).

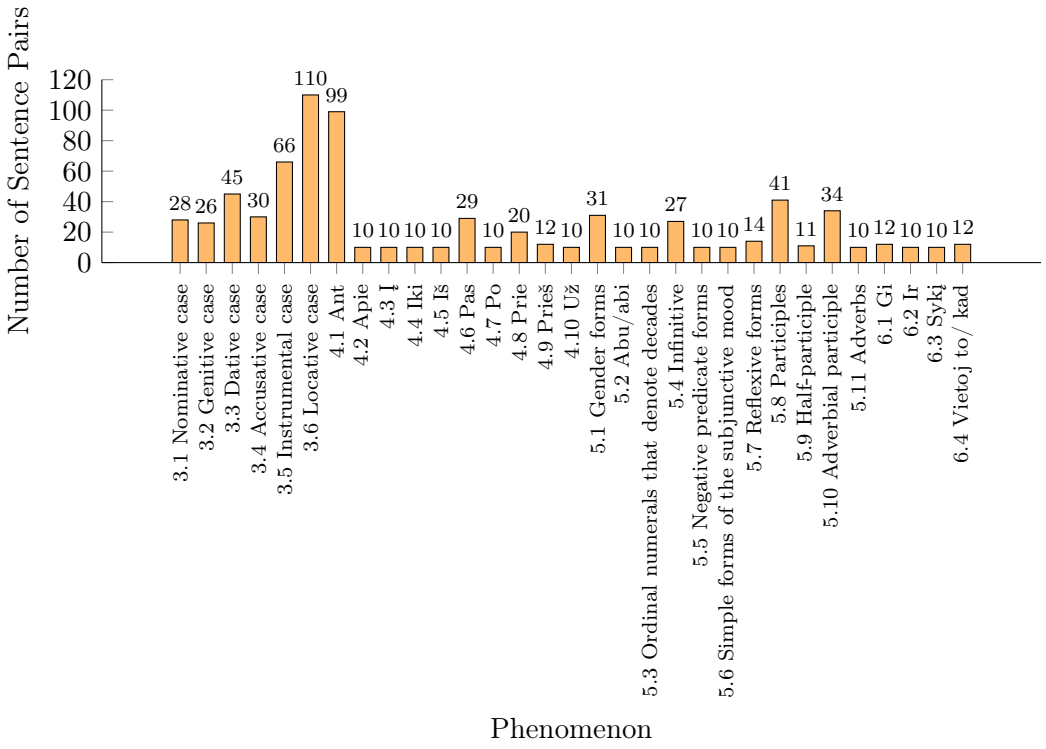


Figure A.3: Count of sentence pairs in Round 3 (each error has at least 10 sentence pairs).

## A.4 List of Phenomena and Errors

The following list presents the phenomena and corresponding error types included in the dataset.

| **Phenomenon Code & Description** | **Error Code & Description** |
|---|---|
| 3.1 Nominative case | 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 Nominative case | 3.1.2 Nominative case should not be used to express direct address. |
| 3.2 Genitive case | 3.2.1 Genitive case should not be used to denote the object of an action with certain verbs. |
| 3.2 Genitive case | 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 Dative case | 3.3.1 Dative case should not be used to denote the object of an action with certain expressions. |
| 3.3 Dative case | 3.3.2 Dative case should not be used to describe a thing/object when purpose is not being expressed. |
| 3.3 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. |
| 3.3 Dative case | 3.3.4 Dative should not be used with verbs of motion to express purpose. |
| 3.4 Accusative case | 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 Accusative case | 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 Accusative case | 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. |

| **Phenomenon Code & Description** | **Error Code & Description** |
|---|---|
| 3.5 Instrumental case | 3.5.1. Instrumental case should not be used to express the object with verbs denoting fullness or increase. |
| 3.5 Instrumental case | 3.5.2 Instrumental case should not be used to express content of quality with adjectives denoting abundance. |
| 3.5 Instrumental case | 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 Instrumental case | 3.5.4 Instrumental case should not be used with forms of the verb *būti* to express a permanent (unchanging) state. |
| 3.5 Instrumental case | 3.5.5 The instrumental of adjectives (and words used adjectivally) should not be used to express a state. |
| 3.6 Locative case | 3.6.1 Locative case should not be used to express the experiencer of a state. |
| 3.6 Locative case | 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 Locative case | 3.6.3 Locative case should not be used to express the state, condition, or characteristic of a thing. |
| 3.6 Locative case | 3.6.4 Locative case should not be used to express the cause or basis of a state. |
| 3.6 Locative case | 3.6.5 Locative case should not be used to express the manner or timing of an action. |
| 3.6 Locative case | 3.6.6 Locative case should not be used to express a time period as a preposition or postposition. |
| 4.1 *Ant* | 4.1.1 *Ant* should not be used to to express the instrument/means of an action when it is not related to location. |
| 4.1 *Ant* | 4.1.2 *Ant* should not be used to express the object with adjectives. |
| 4.1 *Ant* | 4.1.3 *Ant* should not be used to express the object of an action with verbs. |
| 4.1 *Ant* | 4.1.4 *Ant* should not be used to indicate dimensions or a ratio of sizes. |
| 4.1 *Ant* | 4.1.5 *Ant* should not be used to express a characteristic of a thing not related to location. |
| 4.1 *Ant* | 4.1.6 *Ant* should not be used to express place when what is meant is not a surface, but a direction into something or being inside something. |
| 4.1 *Ant* | 4.1.7 *Ant* should not be used to express time. |
| 4.1 *Ant* | 4.1.8 *Ant* should not be used to express the nature/mode/condition of an action or state. |
| 4.1 *Ant* | 4.1.9 *Ant* should not be used to express purpose/intended use/aim of an action. |
| 4.2 *Apie* | 4.2.1 *Apie* should not be used to express the object of an action with certain words. |
| 4.3 *Į* | 4.3.1 *Į* should not be used to express the object of an action with certain words. |
| 4.4 *Iki* | 4.4.1 *Iki* should not be used to express a state in an impersonal sentence. |
| 4.5 *Iš* | 4.5.1 *Iš* should not be used with certain words to indicate the date of writing. |
| 4.6 *Pas* | 4.6.1 *Pas* should not be used to indicate ownership, belonging, or possession. |
| 4.6 *Pas* | 4.6.2 *Pas* should not be used to express the recipient or source of an action, not related to location, with verbs. |
| 4.7 *Po* | 4.7.1 *Po* (with the instrumental case) should not be used to express the mode or basis of an action. |
| 4.8 *Prie* | 4.8.1 *Prie* should not be used to express conditions of an action in certain phrases. |
| 4.8 *Prie* | 4.8.2 *Prie* should not be used to express exchange ratios/conversion rates. |
| 4.9 *Prieš* | 4.9.1 *Prieš* should not be used to express the object of an action with certain words. |
| 4.10 *Už* | 4.10.1 *Už* should not be used in certain constructions. |
| 5.1 Gender forms | 5.1.1 Masculine-gender noun forms are not to be used to refer to women when describing professions, positions, academic degrees. |
| 5.1 Gender forms | 5.1.2 The masculine genitive forms of ordinal numerals and of the pronoun *kelintas*, *-a* must not be used to denote the day of the month. |
| 5.1 Gender forms | 5.1.3 The masculine accusative and instrumental forms of adjectives and of participles used with adjectival meaning are not to be used in the so-called "indefinite gender" meaning. |

| **Phenomenon Code & Description** | **Error Code & Description** |
|---|---|
| 5.2 *Abu, abi* | 5.2.1 The basic quantitative numeral forms from 2 to 9 and the pronoun *abu*, *abi* are not to be used with plural-only nouns (nouns that exist only in the plural). |
| 5.3 Ordinal numerals that denote decades | 5.3.1 The pronominal forms (the forms with possessive endings) of ordinal numerals that denote decades are not to be used to indicate decades. |
| 5.4 Infinitive | 5.4.1 The infinitive should not be used with the conjunction *jei/jeigu* to express a condition. |
| 5.4 Infinitive | 5.4.2 The infinitive should not be used with the conjunction *kad* (also *idant*) to express purpose. |
| 5.5 Negative predicate forms | 5.5.1 Negative predicate forms should not to be used in subordinate clauses expressing concession, when actual negation is not meant. |
| 5.6 Simple forms of the subjunctive mood | 5.6.1 Simple forms of the subjunctive mood should not to be used instead of compound forms to express past actions (often hypothetical or unrealized). |
| 5.7 Reflexive forms | 5.7.1 Reflexive forms of certain verbs should not to be used with a passive meaning if the action cannot occur by itself. |
| 5.8 Participles | 5.8.1 Active present participle should not be used with nouns that do not denote an agent (nouns that cannot perform an action). |
| 5.8 Participles | 5.8.2 Active past participle should not be used to indicate an action occurring simultaneously with or subsequent to the main verb. |
| 5.8 Participles | 5.8.3 Passive past participle should not be used to express a property or state that has arisen spontaneously (without external influence or deliberate intervention). |
| 5.9 Half-participle | 5.9.1 Half-participle should not be used to indicate a secondary action that does not coincide in time at any point with the main verb's action. |
| 5.10 Adverbial participle | 5.10.1 Adverbial participle should not be used to indicate a secondary action of the same agent in personal sentences (instead of a half-participle or participle with proper agreement). |
| 5.10 Adverbial participle | 5.10.2 Past tense aorist adverbial participle must not be used to express purpose with the conjunctions *kad* or *idant*. |
| 5.10 Adverbial participle | 5.10.3 Past tense aorist adverbial participle must not be used to express a condition with the conjunction *jeigu*. |
| 5.11 Adverbs | 5.11.1 Adverbs ending in *-ai* (formed from adjectives with the suffix -iškas) should not be used to indicate the quality of a thing. |
| 6.1 *Gi* | 6.1.1 *Gi* must not be used instead of a coordinating conjunction expressing contrast. |
| 6.2 *Ir* | 6.2.1 *Ir* should not be used to link words in intensifying constructions with adjectives and adverbs. |
| 6.3 *Sykį* | 6.3.1 *Sykį* should not be used to introduce subordinate conditional clauses. |
| 6.4 *Vietoj to, kad* | 6.4.1 *Vietoj to, kad* should not be used with an infinitive or participle in place of *užuot*. |

Table A.1: List of the Lithuanian language phenomena and errors.

# B Models

## B.1 Lithuanian Support in `Gemma 3`

While the `Gemma 3` language model family is reported to support over 140 languages, no official list specifies which languages are included. Lithuanian, however, contains a unique character, *ė*, which does not appear in any other language. To verify whether this model family supports Lithuanian, I checked whether the tokenizer includes this special character[1]. For this purpose, I locally loaded the smallest model in the family, `Gemma 3 270M`. The test was positive: *ė* was recognized as a single token in the tokenizer (see Figure B.1).

```
(llms) (base) urtejakubauskaite@Mac Thesis % python is_lithuanian_supported.py
tokenizer_config.json: 1.16MB [00:00, 37.2MB/s]
tokenizer.model: 100%|████████████████████| 4.69M/4.69M [00:00<00:00, 5.39MB/s]
tokenizer.json: 100%|████████████████████| 33.4M/33.4M [00:00<00:00, 79.7MB/s]
added_tokens.json: 100%|████████████████████| 35.0/35.0 [00:00<00:00, 140kB/s]
special_tokens_map.json: 100%|████████████████| 662/662 [00:00<00:00, 2.86MB/s]
Character: ė
Token IDs: [237965]
Tokens: ['ė']
The character exists as a single token.
```

Figure B.1: The `Gemma 3 270M` tokenizer recognizes the Lithuanian character *ė* as a single token.

## B.2 List of Models

The following table (B.1) lists all models evaluated in this study, including their families, sizes, variants, context lengths, and supported languages.

| No. | Model Family | Size | Version | Context | Supported Languages | LT |
|---|---|---|---|---|---|---|
| 1 | **Aya-expanse** | 8B | Instruct | 8K | 23 languages | No |
| 2 | | 32B | Instruct | 128K | 23 languages | No |
| 3 | **EuroLLM** | 1.7B | Base | 4K | European languages | Yes |
| 4 | | 1.7B | Instruct | 4K | European languages | Yes |
| 5 | | 9B | Base | 4K | European languages | Yes |
| 6 | | 9B | Instruct | 4K | European languages | Yes |
| 7 | | 22B | Base | 32K | European languages | Yes |
| 8 | | 22B | Instruct | 32K | European languages | Yes |
| 9 | **Phi 3** | Mini (3.8B) | Instruct | 4K | 23 languages | No |
| 10 | | Mini (3.8B) | Instruct | 128K | 23 languages | No |
| 11 | | Small (7B) | Instruct | 8K | 23 languages | No |
| 12 | | Small (7B) | Instruct | 128K | 23 languages | No |
| 13 | | Medium (14B) | Instruct | 4K | 23 languages | No |

[1]This implementation was inspired by the following community discussion: https://huggingface.co/google/gemma-2-9b/discussions/33.

| No. | Model Family | Size | Version | Context | Supported Languages | LT |
|---|---|---|---|---|---|---|
| 14 | | Medium (14B) | Instruct | 128K | 23 languages | No |
| 15 | **Goldfish** | 39M: 5MB | Base | 512 | Monolingual (350 languages) | Yes |
| 16 | | 39M: 10MB | Base | 512 | Monolingual (350 languages) | Yes |
| 17 | | 125M: 100MB | Base | 512 | Monolingual (350 languages) | Yes |
| 18 | | 125M: 1000MB | Base | 512 | Monolingual (350 languages) | Yes |
| 19 | **Gemma 3** | 270M | Base | 32K | >140 languages | Yes |
| 20 | | 270M | Instruct | 32K | >140 languages | Yes |
| 21 | | 1B | Base | 32K | >140 languages | Yes |
| 22 | | 1B | Instruct | 32K | >140 languages | Yes |
| 23 | | 4B | Base | 128K | >140 languages | Yes |
| 24 | | 4B | Instruct | 128K | >140 languages | Yes |
| 25 | | 12B | Base | 128K | >140 languages | Yes |
| 26 | | 12B | Instruct | 128K | >140 languages | Yes |
| 27 | | 27B | Base | 128K | >140 languages | Yes |
| 28 | | 27B | Instruct | 128K | >140 languages | Yes |
| 29 | **LLaMa 3.1-3.3** | 3.2 1B | Base | 128K | 8 languages | No |
| 30 | | 3.2 1B | Instruct | 128K | 8 languages | No |
| 31 | | 3.2 3B | Base | 128K | 8 languages | No |
| 32 | | 3.2 3B | Instruct | 128K | 8 languages | No |
| 33 | | 3.1 8B | Base | 128K | 8 languages | No |
| 34 | | 3.1 8B | Instruct | 128K | 8 languages | No |
| 35 | | 3.1 70B | Base | 128K | 8 languages | No |
| 36 | | 3.1 70B | Instruct | 128K | 8 languages | No |
| 37 | | 3.3 70B | Instruct | 128K | 8 languages | No |
| 38 | **LLaMa 4** | Scout 17B-16E (109B total)(MoE) | Base | 10M | 12 languages | No |
| 39 | | Scout 17B-16E (109B total)(MoE) | Instruct | 10M | 12 languages | No |
| 40 | **Neurotechnology** | 7B | Base | 4K | Monolingual (Lithuanian) | Yes |
| 41 | | 7B | Instruct | 4K | Monolingual (Lithuanian) | Yes |
| 42 | | 13B | Base | 4K | Monolingual (Lithuanian) | Yes |
| 43 | | 13B | Instruct | 4K | Monolingual (Lithuanian) | Yes |
| 44 | **OLMo2** | 1B | Base | 4K | Primarily English | No |
| 45 | | 1B | Instruct | 4K | Primarily English | No |
| 46 | | 7B | Base | 4K | Primarily English | No |
| 47 | | 7B | Instruct | 4K | Primarily English | No |
| 48 | | 13B | Base | 4K | Primarily English | No |
| 49 | | 13B | Instruct | 4K | Primarily English | No |
| 50 | | 32B | Base | 4K | Primarily English | No |
| 51 | | 32B | Instruct | 4K | Primarily English | No |
| 52 | **Qwen3** | 0.6B | Base | 32K | 119 languages | Yes |
| 53 | | 0.6B | Instruct | 32K | 119 languages | Yes |
| 54 | | 1.7B | Base | 32K | 119 languages | Yes |
| 55 | | 1.7B | Instruct | 32K | 119 languages | Yes |
| 56 | | 4B | Base | 32K | 119 languages | Yes |
| 57 | | 4B | Instruct | 32K | 119 languages | Yes |
| 58 | | 4B | Instruct (NEW) | 32K | 119 languages | Yes |
| 59 | | 8B | Base | 32K | 119 languages | Yes |
| 60 | | 8B | Instruct | 32K | 119 languages | Yes |
| 61 | | 14B | Base | 32K | 119 languages | Yes |
| 62 | | 14B | Instruct | 32K | 119 languages | Yes |
| 63 | | 32B | Instruct | 32K | 119 languages | Yes |
| 64 | | 30B-A3B (MoE) | Base | 256K | 119 languages | Yes |
| 65 | | 30B-A3B (MoE) | Instruct | 256K | 119 languages | Yes |
| 66 | | 30B-A3B (MoE) | Instruct (NEW) | 256K | 119 languages | Yes |
| 67 | | 235B-A22B (MoE) | Instruct | 256K | 119 languages | Yes |
| 68 | | 235B-A22B (MoE) | Instruct (NEW) | 256K | 119 languages | Yes |
| 69 | **Salamandra/ALIA** | 2B | Base | 8K | European languages | Yes |
| 70 | | 2B | Instruct | 8K | European languages | Yes |
| 71 | | 7B | Base | 8K | European languages | Yes |
| 72 | | 7B | Instruct | 8K | European languages | Yes |
| 73 | | 40B | Base | 32K | European languages | Yes |

| No. | Model Family | Size | Version | Context | Supported Languages | LT |
|---|---|---|---|---|---|---|
| 74 | | 40B | Instruct | 32K | European languages | Yes |
| 75 | **Teuken** | v0.4: 7B | Instruct | 4K | European languages | Yes |
| 76 | | v0.6: 7B | Base | 4K | European languages | Yes |
| 77 | | v0.6: 7B | Instruct | 4K | European languages | Yes |
| 78 | **TildeOpen** | 30B | Base | 8K | 34 languages | Yes |

Table B.1: Detailed overview of all language models (LMs) included in this study. Each row lists a single model variant, its size, version, context length, supported languages, and whether Lithuanian (LT) is supported.

# C Human Acceptability Ratings

## C.1 Survey Construction

The table below shows how the 32 sentence pairs were distributed across six survey versions. Each survey presents the same set of sentences in a different order: Survey 2 starts from the second sentence of Survey 1, Survey 3 from the third sentence, and so on. This rotation ensures counterbalancing across participants. The experiment includs two full sets of these six surveys (12 surveys in total).

| Item | Survey 1 | Survey 2 | Survey 3 | Survey 4 | Survey 5 | Survey 6 |
|---|---|---|---|---|---|---|
| 1 | 1G | 1U | 2G | 2U | 3G | 3U |
| 2 | 1U | 2G | 2U | 3G | 3U | 1G |
| 3 | 2G | 2U | 3G | 3U | 1G | 1U |
| 4 | 2U | 3G | 3U | 1G | 1U | 2G |
| 5 | 3G | 3U | 1G | 1U | 2G | 2U |
| 6 | 3U | 1G | 1U | 2G | 2U | 3G |
| 7 | 1G | 1U | 2G | 2U | 3G | 3U |
| 8 | 1U | 2G | 2U | 3G | 3U | 1G |
| 9 | 2G | 2U | 3G | 3U | 1G | 1U |
| 10 | 2U | 3G | 3U | 1G | 1U | 2G |
| 11 | 3G | 3U | 1G | 1U | 2G | 2U |
| 12 | 3U | 1G | 1U | 2G | 2U | 3G |
| 13 | 1G | 1U | 2G | 2U | 3G | 3U |
| 14 | 1U | 2G | 2U | 3G | 3U | 1G |
| 15 | 2G | 2U | 3G | 3U | 1G | 1U |
| 16 | 2U | 3G | 3U | 1G | 1U | 2G |
| 17 | 3G | 3U | 1G | 1U | 2G | 2U |
| 18 | 3U | 1G | 1U | 2G | 2U | 3G |
| 19 | 1G | 1U | 2G | 2U | 3G | 3U |
| 20 | 1U | 2G | 2U | 3G | 3U | 1G |
| 21 | 2G | 2U | 3G | 3U | 1G | 1U |
| 22 | 2U | 3G | 3U | 1G | 1U | 2G |
| 23 | 3G | 3U | 1G | 1U | 2G | 2U |
| 24 | 3U | 1G | 1U | 2G | 2U | 3G |
| 25 | 1G | 1U | 2G | 2U | 3G | 3U |
| 26 | 1U | 2G | 2U | 3G | 3U | 1G |
| 27 | 2G | 2U | 3G | 3U | 1G | 1U |
| 28 | 2U | 3G | 3U | 1G | 1U | 2G |
| 29 | 3G | 3U | 1G | 1U | 2G | 2U |
| 30 | 3U | 1G | 1U | 2G | 2U | 3G |
| 31 | 1G | 1U | 2G | 2U | 3G | 3U |
| 32 | 1U | 2G | 2U | 3G | 3U | 1G |

Table C.1: Distribution of sentence pairs across six survey versions. The numbers indicate the sentence pair (1, 2, or 3), while "G" denotes a grammatical sentence and "U" an ungrammatical sentence. Each survey shifts the starting sentence to ensure counterbalancing.

## C.2 Instructions to Participants

The instructions below were presented to participants prior to the grammaticality judgment task. For completeness, both the original Lithuanian version and its English translation are provided.

**Lithuanian (original)**

> Įvertinkite žemiau pateiktų sakinių gramatiškumą. **1** reiškia, kad sakinys Jums atrodo *visiškai nepriimtinas gramatiškai*, o įvertinimas **7** – sakinys *visiškai priimtinas gramatiškai*. Nekreipkite dėmesio į žodžių ar stiliaus pasirinkimą. Teisingų ar neteisingų atsakymų nėra – kviečiame remtis savo kalbine intuicija ir nesinaudoti jokia pašaline pagalba. Pasirinkite atsakymą, kuris pirmiausia ateina į galvą. Būtinai pakoreguokite visus slankiklius, nes tik tuomet ekrano apačioje bus rodomas mygtukas, kuris jus nuves į apklausos pabaigą.

**English (translation)**

> Please evaluate the grammatical acceptability of the sentences below. A rating of **1** means that the sentence is *completely grammatically unacceptable* to you, while a rating of **7** means that the sentence is *fully grammatically acceptable*. Please ignore word choice and stylistic aspects. There are no right or wrong answers - rely on your linguistic intuition and do not use any external assistance. Select the response that first comes to mind. Make sure to adjust all sliders, as the button leading to the end of the survey will only appear once all sliders have been moved.

# D Results

## D.1 Model Results

Figure D.1 presents model accuracies across error types in the Phenomenon Type 4 group.

## D.2 Confidence Intervals

Figure D.2 presents the widths of the confidence intervals (CIs) across different phenomena, while Figures D.3-D.6 present the same information at the error type level.

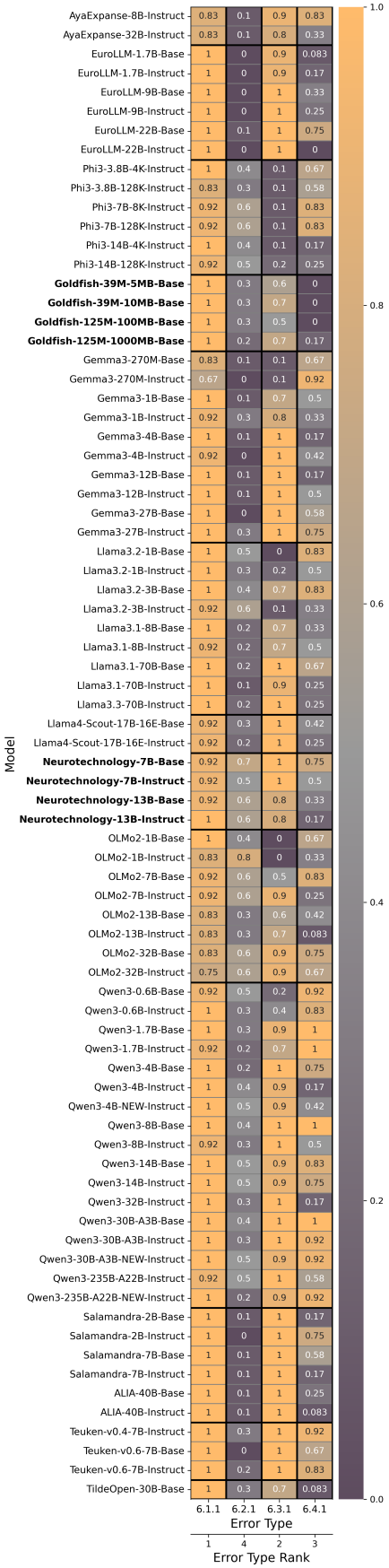


Figure D.1: Per-error accuracies across models for Phenomenon Type 4 (Coordination of Sentence Elements and Clauses). Monolingual models are marked in bold. Horizontal bold lines separate different model families, while vertical bold lines separate phenomenon types.


Figure D.2: Per-phenomenon confidence interval widths across models. Monolingual models are marked in bold. Horizontal bold lines separate different model families, while vertical bold lines separate phenomenon types.


Figure D.3: Per-error confidence interval widths across models for Phenomenon Type 1 (Use of Cases). Monolingual models are marked in bold. Horizontal bold lines separate different model families, while vertical bold lines separate phenomenon types.


Figure D.4: Per-error confidence interval widths across models for Phenomenon Type 2 (Use of Prepositions). Monolingual models are marked in bold. Horizontal bold lines separate different model families, while vertical bold lines separate phenomenon types.

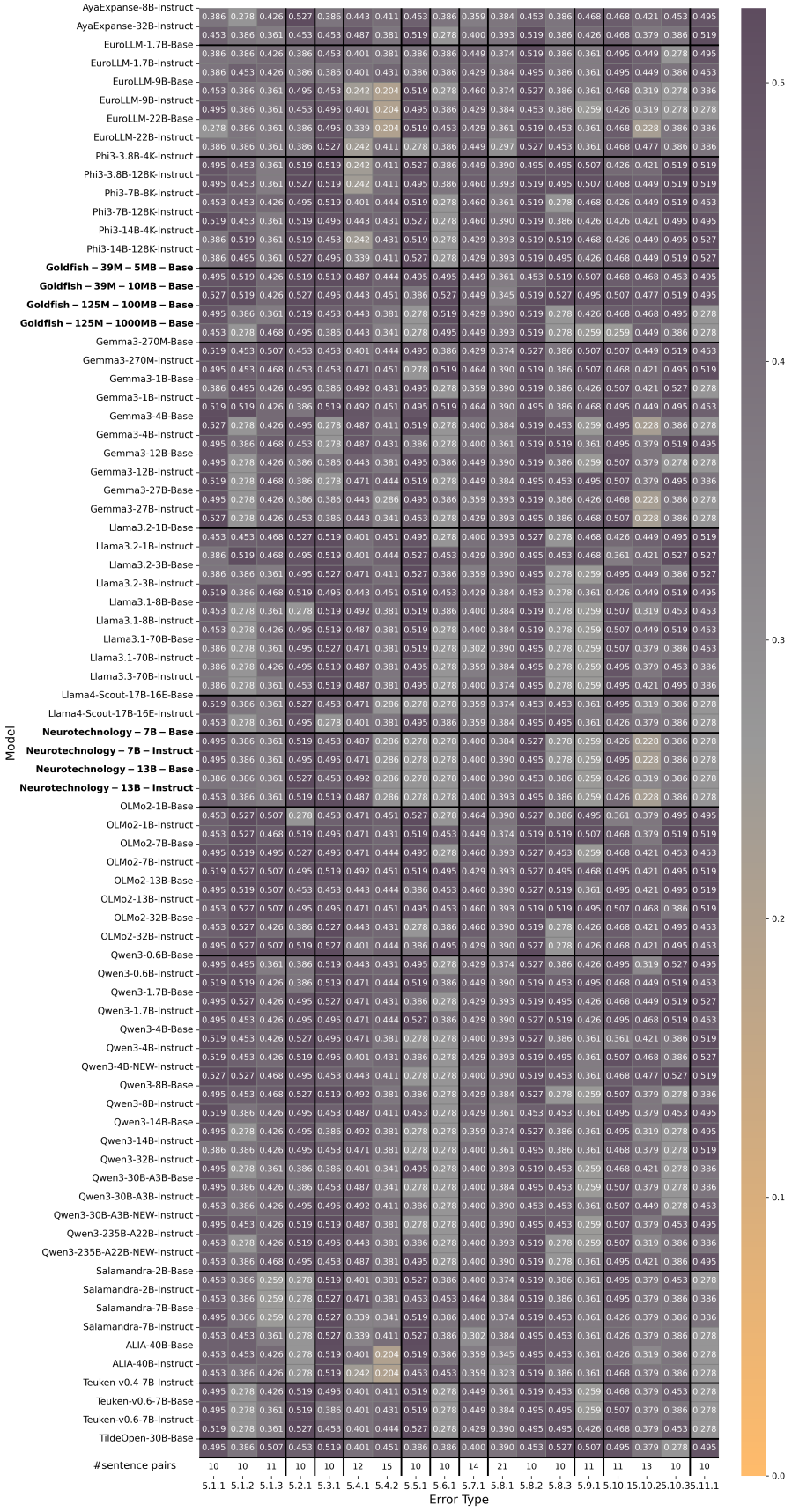


Figure D.5: Per-error confidence interval widths across models for Phenomenon Type 3 (Use of Forms). Monolingual models are marked in bold. Horizontal bold lines separate different model families, while vertical bold lines separate phenomenon types.

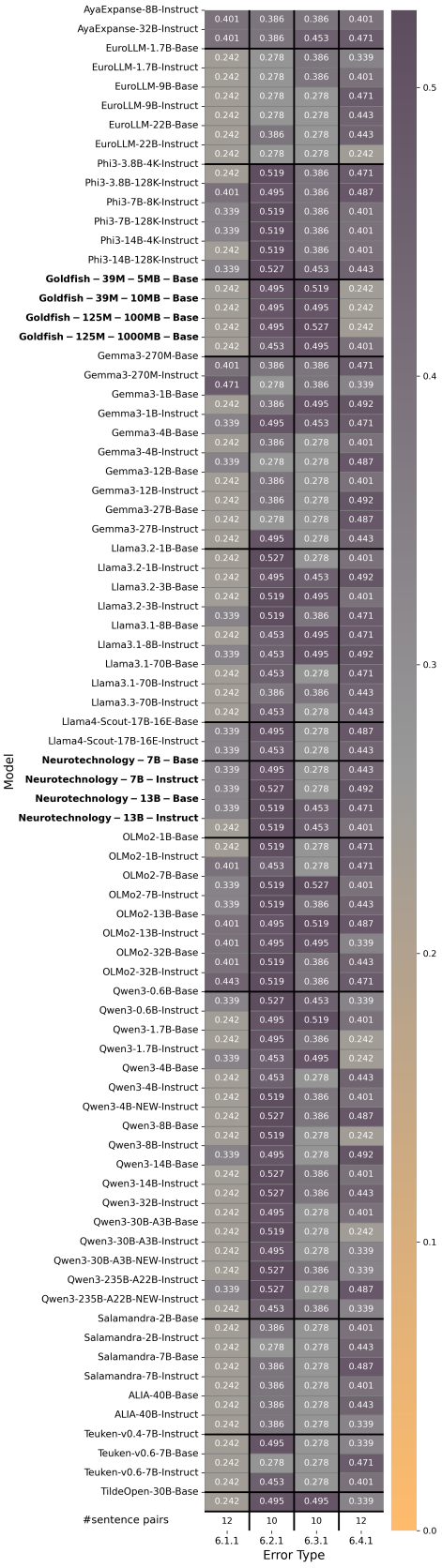


Figure D.6: Per-error confidence interval widths across models for Phenomenon Type 4 (Coordination of Sentence Elements and Clauses). Monolingual models are marked in bold. Horizontal bold lines separate different model families, while vertical bold lines separate phenomenon types..

# E Error Analysis

## E.1 Error Type Difficulty

Table E.1 shows the ordering of error types based on their difficulty for the language models. The error types are ranked from the most difficult to the easiest.

| Difficulty Rank | Phenomenon | Error Type | Mean Accuracy | Number of Items |
|---|---|---|---|---|
| 1 | 6.2 | 6.2.1 | 0.300 | 780 |
| 2 | 4.9 | 4.9.1 | 0.302 | 936 |
| 3 | 5.2 | 5.2.1 | 0.319 | 780 |
| 4 | 3.1 | 3.1.2 | 0.364 | 858 |
| 5 | 4.8 | 4.8.2 | 0.375 | 780 |
| 6 | 4.3 | 4.3.1 | 0.398 | 780 |
| 7 | 4.5 | 4.5.1 | 0.437 | 780 |
| 8 | 5.5 | 5.5.1 | 0.445 | 780 |
| 9 | 4.4 | 4.4.1 | 0.449 | 780 |
| 10 | 4.1 | 4.1.2 | 0.473 | 780 |
| 11 | 3.6 | 3.6.6 | 0.483 | 780 |
| 12 | 3.4 | 3.4.1 | 0.485 | 780 |
| 13 | 3.5 | 3.5.3 | 0.497 | 780 |
| 14 | 5.10 | 5.10.1 | 0.498 | 858 |
| 15 | 4.1 | 4.1.1 | 0.501 | 780 |
| 16 | 6.4 | 6.4.1 | 0.508 | 936 |
| 17 | 3.3 | 3.3.3 | 0.516 | 858 |
| 18 | 4.2 | 4.2.1 | 0.521 | 780 |
| 19 | 5.4 | 5.4.1 | 0.530 | 936 |
| 20 | 4.8 | 4.8.1 | 0.539 | 780 |
| 21 | 3.4 | 3.4.2 | 0.555 | 780 |
| 22 | 5.8 | 5.8.1 | 0.564 | 1638 |
| 23 | 4.1 | 4.1.8 | 0.568 | 1014 |
| 24 | 4.1 | 4.1.9 | 0.578 | 936 |
| 25 | 3.6 | 3.6.3 | 0.579 | 4290 |
| 26 | 3.5 | 3.5.5 | 0.581 | 1872 |
| 27 | 3.5 | 3.5.1 | 0.583 | 858 |
| 28 | 5.8 | 5.8.2 | 0.593 | 780 |
| 29 | 3.2 | 3.2.1 | 0.604 | 780 |
| 30 | 3.4 | 3.4.3 | 0.604 | 780 |
| 31 | 3.5 | 3.5.2 | 0.619 | 858 |
| 32 | 3.6 | 3.6.1 | 0.632 | 780 |
| 33 | 5.3 | 5.3.1 | 0.648 | 780 |
| 34 | 4.10 | 4.10.1 | 0.649 | 780 |
| 35 | 3.1 | 3.1.1 | 0.649 | 1326 |

| Difficulty Rank | Phenomenon | Error Type | Mean Accuracy | Number of Items |
|---|---|---|---|---|
| 36 | 4.6 | 4.6.1 | 0.650 | 1014 |
| 37 | 4.1 | 4.1.7 | 0.659 | 1092 |
| 38 | 4.6 | 4.6.2 | 0.667 | 1248 |
| 39 | 5.4 | 5.4.2 | 0.702 | 1170 |
| 40 | 5.7 | 5.7.1 | 0.703 | 1092 |
| 41 | 5.1 | 5.1.1 | 0.717 | 780 |
| 42 | 4.1 | 4.1.5 | 0.719 | 780 |
| 43 | 3.6 | 3.6.5 | 0.725 | 1014 |
| 44 | 4.7 | 4.7.1 | 0.729 | 780 |
| 45 | 5.11 | 5.11.1 | 0.742 | 780 |
| 46 | 3.2 | 3.2.2 | 0.742 | 1248 |
| 47 | 3.3 | 3.3.2 | 0.743 | 780 |
| 48 | 6.3 | 6.3.1 | 0.743 | 780 |
| 49 | 5.10 | 5.10.3 | 0.749 | 780 |
| 50 | 3.3 | 3.3.4 | 0.756 | 1092 |
| 51 | 5.10 | 5.10.2 | 0.782 | 1014 |
| 52 | 3.6 | 3.6.4 | 0.785 | 936 |
| 53 | 5.1 | 5.1.3 | 0.796 | 858 |
| 54 | 4.1 | 4.1.6 | 0.804 | 780 |
| 55 | 5.1 | 5.1.2 | 0.818 | 780 |
| 56 | 4.1 | 4.1.4 | 0.826 | 780 |
| 57 | 3.6 | 3.6.2 | 0.828 | 780 |
| 58 | 5.8 | 5.8.3 | 0.831 | 780 |
| 59 | 5.9 | 5.9.1 | 0.841 | 858 |
| 60 | 3.5 | 3.5.4 | 0.866 | 780 |
| 61 | 3.3 | 3.3.1 | 0.872 | 780 |
| 62 | 4.1 | 4.1.3 | 0.876 | 780 |
| 63 | 5.6 | 5.6.1 | 0.902 | 780 |
| 64 | 6.1 | 6.1.1 | 0.944 | 936 |

Table E.1: Error types ranked by average model performance (mean accuracy across all models). Lower accuracy indicates higher difficulty for the models.

## E.2 Human Acceptability Ratings

Table E.2 lists all minimal-pair items for which human participants assigned higher ratings to ungrammatical sentences than to grammatical ones. To fit the table, the full minimal sentence pairs are not provided; instead, sentence identifiers corresponding to lines in the main dataset are given.

| Phenomenon | Error Type | Sentence Line in the Dataset | Grammatical Sentence Rating | Ungrammatical Sentence Rating |
|---|---|---|---|---|
| 3.2 | 3.2.1 | 31 | 1.9 | 2.1 |
| 3.3 | 3.3.2 | 67 | 2.7 | 2.8 |
| 3.4 | 3.4.1 | 102 | 4.7 | 5.0 |
| 3.5 | 3.5.1 | 140 | 2.2 | 3.9 |
| 3.5 | 3.5.2 | 142 | 3.2 | 5.4 |
| 3.5 | 3.5.2 | 147 | 3.1 | 4.0 |

| Phenomenon | Error Type | Sentence Line in the Dataset | Grammatical Sentence Rating | Ungrammatical Sentence Rating |
|---|---|---|---|---|
| 3.5 | 3.5.2 | 151 | 2.3 | 3.3 |
| 3.5 | 3.5.3 | 161 | 3.5 | 6.6 |
| 3.6 | 3.6.5 | 293 | 2.2 | 2.3 |
| 4.1 | 4.1.4 | 345 | 1.5 | 2.4 |
| 4.6 | 4.6.2 | 459 | 1.3 | 1.8 |
| 4.8 | 4.8.2 | 496 | 1.8 | 2.6 |
| 4.8 | 4.8.2 | 503 | 2.3 | 2.8 |
| 4.9 | 4.9.1 | 505 | 1.4 | 2.4 |
| 5.1 | 5.1.3 | 547 | 4.8 | 4.9 |
| 5.1 | 5.1.3 | 556 | 2.5 | 3.2 |
| 5.2 | 5.2.1 | 563 | 3.2 | 4.2 |
| 5.2 | 5.2.1 | 566 | 4.3 | 5.1 |
| 5.3 | 5.3.1 | 576 | 4.0 | 4.6 |
| 5.7 | 5.7.1 | 636 | 5.7 | 6.3 |
| 5.8 | 5.8.1 | 654 | 3.7 | 4.7 |
| 5.8 | 5.8.1 | 656 | 4.4 | 4.9 |
| 5.8 | 5.8.2 | 661 | 2.6 | 3.9 |
| 5.10 | 5.10.1 | 700 | 4.8 | 5.0 |

Table E.2: Minimal sentence pairs that received flipped ratings from human participants, meaning that grammatical sentences were rated lower on average than ungrammatical ones.