Evaluating LLM Judgment on Latvian and Lithuanian Short Answer Matching

Yevhen Kostiuk[1,2]
[1]ARG-tech,
University of Dundee
[2]OpenBabylon
ykostiuk001@dundee.ac.uk

Oxana Vitman
University of Bremen

Łukasz Gagała
Georg-August
Universität Göttingen

Artur Kiulian
OpenBabylon

Abstract

In this work, we address the challenge of evaluating large language models (LLMs) on the short answer matching task for Latvian and Lithuanian languages. We introduce novel datasets consisting of 502 Latvian and 690 Lithuanian question-answer pairs. For each question-answer pair, we generated matched and non-matched answers using a set of alteration rules specifically designed to introduce small but meaningful changes in the text. These generated answers serve as test cases to assess the ability of LLMs to detect subtle differences in matching of the original answers. A subset of the datasets was manually verified for quality and accuracy. Our results show that while larger LLMs, such as QWEN2.5 72b and LLaMa3.1 70b, demonstrate near-perfect performance in distinguishing matched and non-matched answers, smaller models show more variance. For instance, LLaMa3.1 8b and EuroLLM 9b benefited from few-shot examples, while Mistral Nemo 12b underperformed on detection of subtle text alteration, particularly in Lithuanian, even with additional examples. QWEN2.5 7b and Mistral 7b were able to obtain a strong and comparable performance to the larger 70b models in zero and few shot experiments. Moreover, the performance of Mistral 7b was weaker in few shot experiments. The code and the dataset are available on our GitHub[1].

1 Introduction

In educational domain, open-ended questions are commonly used and can be defined as questions that require a more elaborate response than simple yes-no or selection of a correct choice. These questions help to encourage a discussion, share ideas and provide more freedom for a student.

Evaluation of responses to the open-ended question is a time-consuming and difficult task that requires an evaluator to carefully read each answer and compare it with the correct answers, ensuring they match. Automating this process makes it easier for evaluators to provide a feedback and analyze errors faster (Pillai et al., 2018; Sreevidhya and Narayanan, 2021).

The automatic short answer matching task addresses this challenge. The goal of the task is to predict whether an answer to the question is matching a correct answer. With the introduction of LLMs, reasonable performance was achieved on English and other high-resource languages for this problem (Ivanova and Handschuh, 2024). On the other hand, when it comes to low-resource settings, LLMs demonstrated weaker results, as well as displayed biases (Hackl et al., 2023; Lai et al., 2023).

In this work, we focus on Latvian and Lithuanian answer matching task, specifically on a detection of correct and incorrect responses that are similar to a set of reference "gold" answers, but differ in the key detail(s) to the question.

We automatically generated open-ended question-answer datasets for these languages based on Wikipedia. For this task, we do not focus on the factual correctness of the answers. Each element of the dataset contains a question and its reference answer. Then we generated a set of answers that are matched with the reference answer and a set of non-matched answers. The non-matched answers are created as similarly as possible to the reference answers with respect to the words inclusion, but with the key words changed to make it incorrect. To generate the answers, we formulated different text *alteration*

---

[1] https://github.com/OpenBabylon/NoDaLiDa2025-Latvian-Lithuanian-SAM

Proceedings of the 1st NB-REAL Workshop 2025 - Nordic-Baltic Responsible Evaluation and Alignment of Language models, pages 12–41
March 2, 2025 ©2025 Association for Computational Linguistics

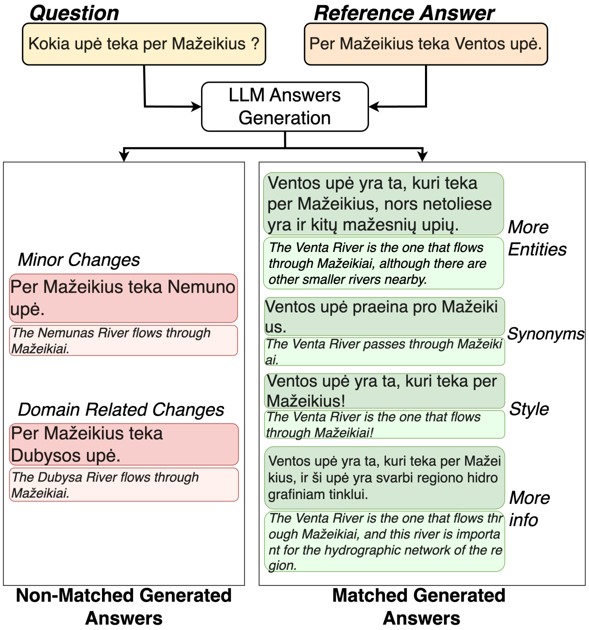

Question
Kokia upė teka per Mažeikius ?
Reference Answer
Per Mažeikius teka Ventos upė.
LLM Answers Generation
Minor Changes
Per Mažeikius teka Nemuno upė.
The Nemunas River flows through Mažeikiai.
Domain Related Changes
Per Mažeikius teka Dubysos upė.
The Dubysa River flows through Mažeikiai.
Ventos upė yra ta, kuri teka per Mažeikius, nors netoliese yra ir kitų mažesnių upių.
The Venta River is the one that flows through Mažeikiai, although there are other smaller rivers nearby.
More Entities
Ventos upė praeina pro Mažeikius.
The Venta River passes through Mažeikiai.
Synonyms
Ventos upė yra ta, kuri teka per Mažeikius!
The Venta River is the one that flows through Mažeikiai!
Style
Ventos upė yra ta, kuri teka per Mažeikius, ir ši upė yra svarbi regiono hidrografiniam tinklui.
The Venta River is the one that flows through Mažeikiai, and this river is important for the hydrographic network of the region.
More info
Non-Matched Generated Answers
Matched Generated Answers

Figure 1: Example of the element from Lithuanian generated dataset.

rules (**AR**) that are minor when it comes to a text change, but semantically are major. For each rule, the different LLMs with a few shot generation process were used. Finally, to ensure the quality, we manually evaluated a sample of the data and filtered the final dataset based on it. We expect the models to obtain high, almost perfect, results on this task.

We formulated the following research questions in this paper.

Q1: Are LLMs capable of correctly identifying matched and non-matched answers with the proposed alteration rules?

Q2: Is there a difference between few-shot and zero-shot inference for different LLMs for this task?

Our contributions are the following:

- We automatically generated a dataset of 502 Latvian and 690 Lithuanian question-answer pairs based on Wikipedia. We defined and generated a list of matched and non-matched answers to each pair of question-answer, resulting in 3,012 and 4,830 elements for Latvian and Lithuanian respectively, and partially manually evaluated samples of the datasets.
- We evaluated LLaMa3.1 (8b and 70b) (Dubey et al., 2024), Mistral Nemo 12b and Mistral 7b (Jiang et al., 2023), EuroLLM 9b (Martins et al., 2024), and QWEN2.5 (7b and 72b) (Team, 2024; Yang et al., 2024) models and compared their achieved accuracy scores per AR and overall.
- We evaluated the models in zero-shot and few-shot settings and their performance based on different ARs of matched and non-matched answers.

Our findings showed that larger LLMs, such as QWEN2.5 72b and LLaMa3.1 70b, consistently performed well across both Latvian and Lithuanian datasets, effectively distinguishing matched and non-matched answers in both zero shot and few shot experiments. However, smaller models demonstrated variation in their results. LLaMa3.1 8b and EuroLLM 9b showed improved performance with few-shot examples, while Mistral Nemo 12b showed limitations, particularly in Lithuanian. QWEN2.5 7b and Mistral 7b were able to obtain a similar performance to the larger 70b models, with Mistral 7b showing weaker performance in few shot experiments.

2 Related Work

Answer matching task can be viewed as a subtask of the automatic short answer grading (ASAG). The definition of what is a short answer and if it is acceptable can vary depending on the domain (Burrows et al., 2015; Bonthu et al., 2021). Nevertheless, all the definitions involve high semantic similarity between the correct answer(s) and predicted answers. The grading scale is also can be domain dependent (Zhang et al.; Divya et al., 2023; Krithika and Narayanan, 2015).

With the development of deep learning methods, they were widely used for the task, as they provide better robustness towards syntactic changes of the text rather than other methods (Bonthu et al., 2021), utilizing RNNs (Cai, 2019), CNNs (Chen and Zhou, 2019), transformers (Sung et al., 2019; Willms and Padó, 2022) and so on. Some of the suggested methods are aimed to not only grade an answer, but to explain its flows and inaccuracies (Tornqvist et al., 2023).

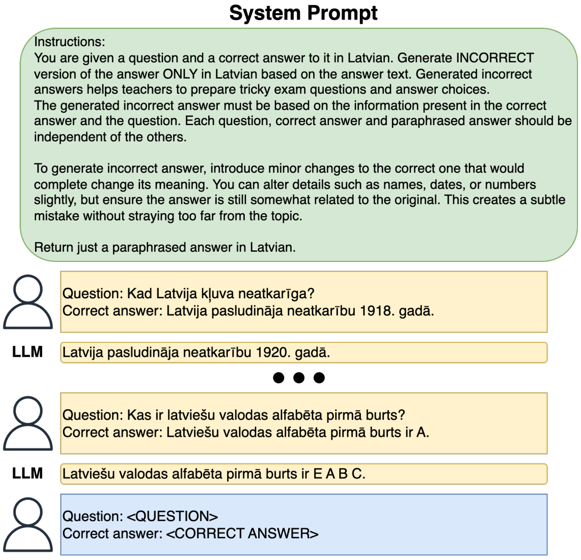
With the rise of generative large language models (LLMs), they were applied for ASAG as well (Metzler et al., 2024; Ivanova and Handschuh, 2024; Chu et al., 2024; Schneider et al., 2023; Grévisse, 2024; Yancey et al., 2023; Yoon, 2023). Analysis of LLMs for this task showed that they are capable of predicting consistent ratings for English (Hackl et al., 2023; Mizumoto and Eguchi, 2023). However, studies showed that the LLMs' performance on the non-English datasets is weaker (Lai et al., 2023; Darģis et al., 2024).

On the other hand, as any other NLP task, there is a gap in the ASAG resources for low-resource languages, including Nordic and Baltic. This area lacks high-quality datasets for these languages. The GPT-3.5 and GPT-4 models were evaluated on Finnish ASAG (Chang and Ginter, 2024) on the dataset of students' answers in Finnish for multiple subjects. The study demonstrated that the models assigned higher scores to the students' answers than the human annotator and achieving Quadratic Weighted Kappa (QWK) score of 0.44. In (Chang et al., 2022), the authors considered ASAG task as a paraphrase retrieval task, evaluating classical methods (TF-IDF) and different transformer methods.

In (Darģis et al., 2022), the self-assessment platform for Latvian language learners was proposed and developed. The authors generated exercises automatically based on data from multiple corpora (Levāne-Petrova et al., 2023; Darģis et al., 2022). In (Stefanovič et al., 2024), the research on detecting AI generated answers in Lithuanian was conducted, producing a dataset with student answers, GPT generated answers and its paraphrased versions. The authors (Weegar and Idestam-Almquist, 2024) created a dataset of student answers in Swedish in programming languages, networking and the Internet, and data abstractions and manipulations. The authors examined different machine learning methods to tackle the task. In (Klevstuen, 2022), the use of information retrieval and text mining methods were investigated to evaluate the content of Norwegian exam answers in Computer Science. In our work, we release multi-domain publicly available datasets as well as benchmark results for some of the open-source multilingual LLMs.

3 Datasets

To generate answer matching datasets, the three-stage pipeline was implemented.



Instructions:
You are given a question and a correct answer to it in Latvian. Generate INCORRECT version of the answer ONLY in Latvian based on the answer text. Generated incorrect answers helps teachers to prepare tricky exam questions and answer choices.
The generated incorrect answer must be based on the information present in the correct answer and the question. Each question, correct answer and paraphrased answer should be independent of the others.

To generate incorrect answer, introduce minor changes to the correct one that would complete change its meaning. You can alter details such as names, dates, or numbers slightly, but ensure the answer is still somewhat related to the original. This creates a subtle mistake without straying too far from the topic.

Return just a paraphrased answer in Latvian.

Question: Kad Latvija kļuva neatkarīga?
Correct answer: Latvija pasludināja neatkarību 1918. gadā.

Latvija pasludināja neatkarību 1920. gadā.

Question: Kas ir latviešu valodas alfabēta pirmā burts?
Correct answer: Latviešu valodas alfabēta pirmā burts ir A.

Latviešu valodas alfabēta pirmā burts ir E A B C.

Question: <QUESTION>
Correct answer: <CORRECT ANSWER>

Figure 2: Example of few-shot incorporating minor changes prompt for non-matched answers generation in Latvian. $< \ldots >$ indicate the sample that requires prediction.

Firstly, we used the approach for generating question-answer Knowledge and Instruction Dataset (KID) based on Wikipedia, introduced in (Kiulian et al., 2024) and adapted it for Latvian (Lat-KID) and Lithuanian (Lit-KID). More details are provided in the Section 3.1. The generated datasets consist of pairs of a question and a reference answer (assumed to be correct and relevant to the question), as well as a factual information that supports the answer.

In the second stage, for each pair of question and answer, we defined a list of different *alteration rules* that rewrites reference answer to matched or non-matched (more details are provided in the Section 3.2). We used GPT-4o and LLaMa3 8b (see Figure 1), utilizing separate prompts for each rule. The non-matched prompts were composed in a way that preserves as much words and semantics of the reference answer as possible with changing key words of the answer, while matched prompts are more flexible.

Finally, the generated results were validated and methods were filtered based on the accept ratio (more details are provided in the Section 3.3). A limitation of the approach it that we used LLMs for the benchmark generation, which could introduce an additional bias to the final dataset.

3.1 Lat-KID and Lit-KID Question-Answering Datasets

For each language, we extracted the top 1,000 articles for each month of the last 12 month from Wikipedia, resulting in 12,000 articles. From this pool, 1,000 articles with the top cumulative counts were extracted. The articles were filtered by their relevance to the corresponding country with Gemini 1.5 Pro (Team et al., 2024). Each article was separated into paragraphs and at least 3 questions were generated for it with Gemini 1.5 Pro. The prompt contains additional fields to run a self-check on the quality of the question (standalone, in the correct language, natural sounding). The prompts are available in the project's GitHub.

The obtained Lat-KID dataset has 502 unique questions. The average number of words in the question is 9.83 and in the answer is 24.37. The total number of words in the dataset (questions and reference answers) is 17,172. The unique amount of words is 5,058.

The obtained Lit-KID dataset has 690 unique questions. The average number of words in the question is 9.88 and in the answer is 29.02. The total number of words in the dataset (questions and reference answers) is 26,849. The unique amount of words is 7,725.

3.2 Matched and Non-Matched Answers Generation

Non-Matched Answers Generation. We defined two alteration rules for non-matched answers generation: incorporating minor changes (**IMC**) and changing domain related information (**CDRI**). IMC includes changes to the text that change a couple of key words like date, name, location etc, while keeping everything else unchanged. CDRI is similar to IMC, however its objective is to change a key term to the similar from the same domain. For example, changing the name of the first president to the second one, changing the word "Parliament" to "President" etc. With the **CDRI** method, the model is prompted to generated something that seems correct and from the same domain, but it is not.

To generate non-matched answers, we utilized LLaMa3 7b[2] and GPT-4o[3]. We selected these models for benchmark creation based on their performance and multilingual capabilities (Darģis et al., 2024).

When generating IMC and CDRI answers, the model was presented with the few-shot example prompts (see Figure 2).

Matched Answers Generation. We defined the following alteration rules for matched answers generation: adding more question-related entities (**Ents**), changing words to synonyms (**Synonyms**), adding more background information (**MoreInfo**), and style swap to exclamatory (**Exclamatory**).

As previously, we used GPT-4o and LLaMa3 7b. The models were presented with different prompts per rule. The code and prompts are available in the project's GitHub repository .

Postprocessing. After generating the answers, the duplicates were removed. The resulting amount of (question, reference answer, generated answer) triplets is 3,012 (1,506 are matched and other 1,506 are non-matched) for Latvian and 4,830 (2,760 are matched and 2,070 are non-matched) for Lithuanian. The amount of matched answers is 3,697. The amount of non-matched answers is 1,809.

3.3 Manual Evaluation

We recruited two native speakers for Latvian and Lithuanian to evaluate the quality of the final generated dataset. They were presented with a random triplet of (question, reference answer, generated answer) and a description if the generated answer was generated by matched or non-matched method. Based on that, the annotators had to accept a triplet if the description fits the reference and generated answers. Otherwise, they had to reject sample. The results are presented in Appendix A. The examples of rejected samples are presented in the Appendix B.

4 Methodology

To evaluate the LLMs capabilities and an influence of the prompting strategy, we used two prompting methods per language for this task: zero shot (**ZS**) and few shot (**FS**). We set all the parameters to defaults with a random seed of 2.

In all of the methods, the models were instructed to start their output with *True* if the provided reference answer and a generated answer are matched otherwise with *False*. ZS and FS shared the same system prompt, but FS gave a model additional examples in corresponding language.

We evaluated LLaMa3.1 (8b and 70b) (Dubey et al., 2024), Mistral Nemo 12b and Mistral 7b (Jiang et al., 2023), and QWEN2.5 (7b and 72b) (Team, 2024; Yang et al., 2024) models. To

---

[2] After manual evaluation, only IMC were generation was accepted for Lat-KID and CDRI for Lit-KID.

[3] We experimented with LLaMa2 13b, however manual evaluation showed much worse results.

	LT		LV	
	ZS	FS	ZS	FS
QWEN2.5;72b	0	1	0	0
LLaMa3.1:70b	1	8	1	3
Mistral:12b	0	1	0	0
EuroLLM:9b	0	0	2,845	4
LLaMa3.1:8b	111	30	2	10
QWEN2.5:7b	0	4	0	2
Mistral:7b	0	0	0	0

Table 1: Number of samples, where the model failed to produce an acceptable (parsable) answer.

	LT		*LV*	
	ZS	FS	ZS	FS
QWEN2.5 72b	0.99	0.99	0.99	0.99
LLaMa3.1 70b	0.99	0.99	0.99	0.99
Mistral Nemo 12b	0.96	0.94	0.96	0.94
EuroLLM 9b	0.13	0.97	0.05	0.84
LLaMa3.1 8b	0.89	0.98	0.87	0.96
QWEN2.5 7b	0.98	0.98	0.97	0.97
Mistral 7b	0.95	0.91	0.95	0.91

Table 2: F1 scores of binary matching. *LT* and *LV* refer to Lithuanian and Latvian respectively. ZS and FS refer to zero shot and few shot respectively.

parse the output, we checked if the model followed instructions about the output. If it did not, we retrieved the key words: “True” or “False”. If none of the words were presented, we counted it as an incorrect prediction (see Table 1).

5 Results and Discussion

The results are presented in Table 2, and on Figures 3 and 4. Additionally, we measured the percentage of times, when model followed the provided format and started with “True” or “False”. The majority of models were able to output the correct format for 99% on Latvian samples. For Lithuanian, LLaMa3.1 8b generated text in correct format in 89% of times in ZS settings. In case of the FS, this value is 99%. Other models consistently followed the format with a rate of 99%. EuroLLM 9b was not able to follow a format at all in ZS settings for both languages, even though its results were legible, but impossible to parse. However, when presented with a few shot examples, it generated expected format.

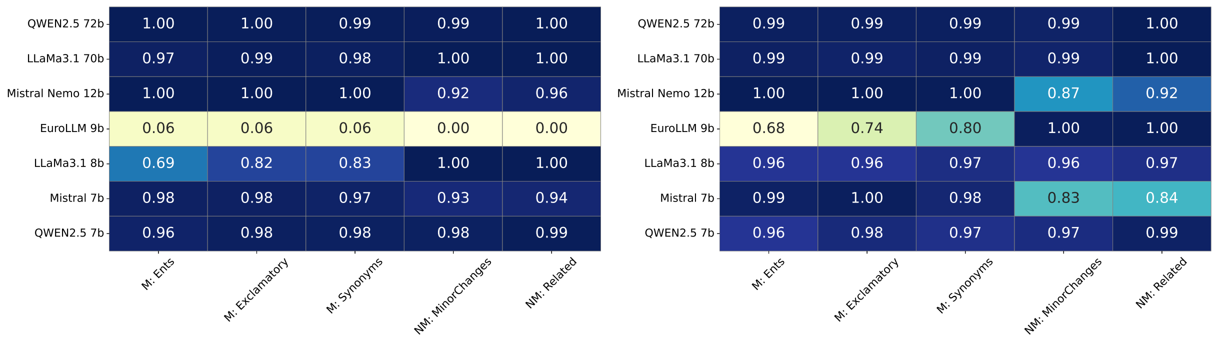
Our results demonstrated that larger LLMs (with 70b parameters) are capable of reliably detect matched and non-matched answers in Lithuanian and Latvian. We hypothesized that LLMs would output near perfect scores, however, smaller models performed differently. In the case of Mistral Nemo, there was a slight decrease of results when switched from zero shot to a few shot approach in both languages. On the contrary, LLaMa3.1 8b performed better in a few shot scenario, improving its ZS score on 9%. QWEN2.5 7b performed nearly perfectly, achieving 99 accuracy score in both settings.

Deeper analysis of results indicated that in case of Latvian, most of the models (except for LLaMa3.1 8b, MIstral 7b, and EuroLLM 9b) showed almost perfect performance on all the generated types of matched and non-matched answers. LLaMa3.1 8b was able to pick up non-matched answers in ZS and FS settings, but struggled with matched answers, demonstrating a bias towards negative answers. However, exposing it with the additional examples boosted its scores to the same level as others. EuroLLM was not able to follow instructions in zero shot prompts, therefore performing poorly. However, in the few shot settings, the model was able to detect non-matched answers, but had less success with matching answers, demonstrating bias towards negative answers. Mistral 7b perfromed well in ZS experiments, but showed a weaker performance in FS for non-match generated samples.

For Lithuanian, the least reliable model was Mistral Nemo 12b. It demonstrated a strong performance on the matched answers with more information and more entities, but was not able to effectively detect synonyms changes in both ZS and FS settings. In case of this model, providing more examples to the model did not have a noticeable effect. Interestingly, EuroLLM showed the same pattern as for Latvian in ZS, but was able to get a comparable results with the 70b groups of models in FS settings. It indicates that the model has a better understanding of Lithuanian than Latvian when it comes to this task, and can perform well when provided with examples.

Therefore, based on our observations, we can address each of the research questions we formulated.

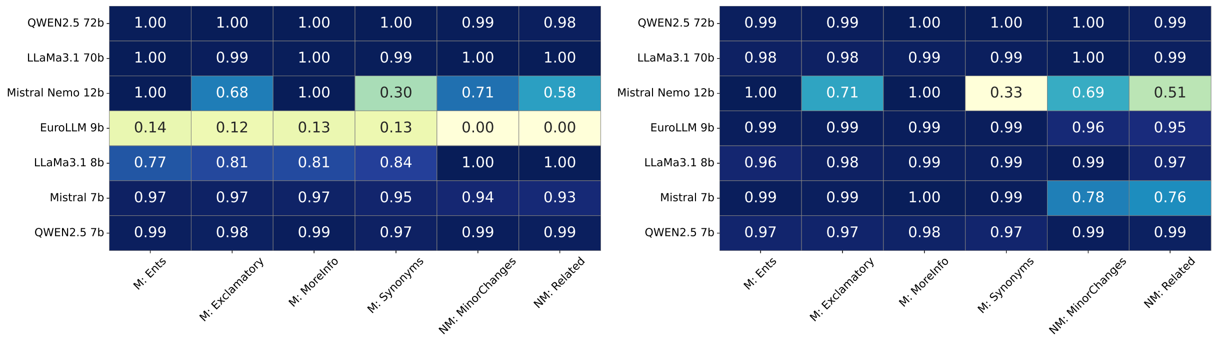
Q1: Are LLMs capable of correctly identifying matched and non-matched answers with the proposed alteration rules ? Overall, the evaluated models were able to accurately identify, which



(a) Zero shot

(b) Few shot

Figure 3: Accuracy scores per generated answer type for Latvian.



(a) Zero shot

(b) Few shot

Figure 4: Accuracy scores per generated answer type for Lithuanian.

answers are matched and which are not. LLMs with the greater number of parameters showed a very consistent performance, when smaller model can have difficulties with Latvian or Lithuanian. Specifically, LLaMa3.1 8b and EuroLLM 9b require additional examples, when QWEN2.5 7b and Mistral 7b are on par with the larger models. Moreover, we found specific types of alternation rules that models had more difficulties to pick up. Specifically LLaMa3.1 8b and EuroLLM 9b had difficulties with added entities in the text in Latvian. Mistral 7b struggled with incorporating minor changes and changing domain related information rules in Latvian FS settings. Mistral Nemo obtained weaker performance on changing words to synonyms and style swap to exclamatory (**Exclamatory**) rules in Lithuanian.

Q2: Is there a difference between few-shot and zero-shot inference for different LLMs for this task ? Our findings showed that few shot approach did not improve the scores of the larger models: they are already very high. However, it can be helpful in case of some smaller models, especially with EuroLLM 9b. In case of Mistral 7b, the perfromance was decreased with adding more examples. On the other hand, if the model struggles with a language, providing more examples will not necessarily improves its performance (e.g. Mistral Nemo in Lithuanian or Mistral 7b) for this task.

6 Conclusion

In conclusion, our findings demonstrate that large language models (LLMs) with greater parameter counts, such as QWEN2.5 72b and LLaMa3.1 70b, consistently achieve high accuracy in distinguishing matched and non-matched answers across both Latvian and Lithuanian, regardless of zero-shot or few-shot settings. Smaller models showed less robustness, with LLaMa3.1 8b and EuroLLM 9b benefiting from additional examples in few-shot scenarios. Mistral Nemo 12b struggled with detecting certain nuances, particularly in Lithuanian. QWEN2.5 7b and Mistral 7b were able to obtain a similar the performance to the larger 70b models, but in case of Mistral 7b the performance decreased in with a few shot approach. These results highlight the robustness of larger models and the potential for targeted improvements in smaller ones to address answer matching task with the defined set of alteration rules.

Acknowledgments

This chapter is a product of the research conducted in the Collaborative Research Center 1342 “Global Dynamics of Social Policy”. The center is funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation)—project number 374666841—SFB 1342. We would like to express our gratitude to the following organizations for their generous support, which made this work possible: GCP (Google Cloud) for providing credits used for model training and inference, and Tensorwave for providing AMD MI300X instance for inference and evaluations.

References

Sridevi Bonthu, S Rama Sree, and MHM Krishna Prasad. 2021. Automated short answer grading using deep learning: A survey. In *Machine Learning and Knowledge Extraction: 5th IFIP TC 5, TC 12, WG 8.4, WG 8.9, WG 12.9 International Cross-Domain Conference, CD-MAKE 2021, Virtual Event, August 17–20, 2021, Proceedings 5*, pages 61–78. Springer.

Steven Burrows, Iryna Gurevych, and Benno Stein. 2015. The eras and trends of automatic short answer grading. *International journal of artificial intelligence in education*, 25:60–117.

Changzhi Cai. 2019. Automatic essay scoring with recurrent neural network. In *Proceedings of the 3rd International Conference on High Performance Compilation, Computing and Communications*, pages 1–7.

Li-Hsin Chang and Filip Ginter. 2024. Automatic short answer grading for finnish with chatgpt. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 23173–23181.

Li-Hsin Chang, Jenna Kanerva, and Filip Ginter. 2022. Towards automatic short answer assessment for Finnish as a paraphrase retrieval task. In *Proceedings of the 17th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2022)*, pages 262–271, Seattle, Washington. Association for Computational Linguistics.

Zhiyun Chen and Yuxin Zhou. 2019. Research on automatic essay scoring of composition based on cnn and or. In *2019 2nd International Conference on Artificial Intelligence and Big Data (ICAIBD)*, pages 13–18. IEEE.

Yucheng Chu, Hang Li, Kaiqi Yang, Harry Shomer, Hui Liu, Yasemin Copur-Gencturk, and Jiliang Tang. 2024. A llm-powered automatic grading framework with human-level guidelines optimization. *arXiv preprint arXiv:2410.02165*.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

Roberts Darģis, Ilze Auziņa, Inga Kaija, Kristīne Levāne-Petrova, and Kristīne Pokratniece. 2022. Corpus based self-assessment platform for latvian language learners. *Baltic Journal of Modern Computing*.

Roberts Darģis, Ilze Auziņa, Inga Kaija, Kristīne Levāne-Petrova, and Kristīne Pokratniece. 2022. LaVA – Latvian language learner corpus. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 727–731, Marseille, France. European Language Resources Association.

Roberts Darģis, Guntis Bārzdiņš, Inguna Skadiņa, and Baiba Saulite. 2024. Evaluating open-source LLMs in low-resource languages: Insights from Latvian high school exams. In *Proceedings of the 4th International Conference on Natural Language Processing for Digital Humanities*, pages 289–293, Miami, USA. Association for Computational Linguistics.

Arunima Divya, Vivek Haridas, and Jayasree Narayanan. 2023. Automation of short answer grading techniques: Comparative study using deep learning techniques. In *2023 Fifth International Conference on Electrical, Computer and Communication Technologies (ICECCT)*, pages 1–7. IEEE.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen

Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaoqing Ellen Tan, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aaron Grattafiori, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alex Vaughan, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Franco, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, Danny Wyatt, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Firat Ozgenel, Francesco Caggioni, Francisco Guzmán, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Govind Thattai, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Karthik Prasad, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kun Huang, Kunal Chawla, Kushal Lakhotia, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Maria Tsimpoukelli, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Michael L. Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikolay Pavlovich Laptev, Ning Dong, Ning Zhang, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Rohan Maheswari, Russ Howes, Ruty Rinott, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Sungmin Cho, Sunny Virk, Suraj Subra-

manian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Kohler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vítor Albiero, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaofang Wang, Xiaojian Wu, Xiaolan Wang, Xide Xia, Xilun Wu, Xinbo Gao, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yuchen Hao, Yundi Qian, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, and Zhiwei Zhao. 2024. The llama 3 herd of models.

Christian Grévisse. 2024. Llm-based automatic short answer grading in undergraduate medical education. *BMC Medical Education*, 24(1):1060.

Veronika Hackl, Alexandra Elena Müller, Michael Granitzer, and Maximilian Sailer. 2023. Is gpt-4 a reliable rater? evaluating consistency in gpt-4's text ratings. In *Frontiers in Education*, volume 8, page 1272229. Frontiers Media SA.

Rositsa V Ivanova and Siegfried Handschuh. 2024. Evaluating llms' performance at automatic short-answer grading.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b.

Artur Kiulian, Anton Polishko, Mykola Khandoga, Oryna Chubych, Jack Connor, Raghav Ravishankar, and Adarsh Shirawalmath. 2024. From bytes to borsch: Fine-tuning gemma and mistral for the ukrainian language representation. *arXiv preprint arXiv:2404.09138*.

Pernille Andresen Klevstuen. 2022. Assisting efficient and fair grading with information retrieval and text mining techniques. Master's thesis, NTNU.

R Krithika and Jayasree Narayanan. 2015. Learning to grade short answers using machine learning techniques. In *Proceedings of the Third International Symposium on Women in Computing and Informatics*, pages 262–271.

VD Lai, NT Ngo, APB Veyseh, H Man, F Dernoncourt, T Bui, and TH Nguyen. 2023. Chatgpt beyond english: Towards a comprehensive evaluation of large language models in multilingual learning. arxiv.

Kristīne Levāne-Petrova, Roberts Darģis, Kristīne Pokratniece, and Viesturs Jūlijs Lasmanis. 2023. Balanced corpus of modern latvian (LVK2022). CLARIN-LV digital library at IMCS, University of Latvia.

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

Tim Metzler, Paul G. Plöger, and Jörn Hees. 2024. Computer-assisted short answer grading using large language models and rubrics. In *INFORMATIK 2024*, pages 1383–1393. Gesellschaft für Informatik e.V., Bonn.

Atsushi Mizumoto and Masaki Eguchi. 2023. Exploring the potential of using an ai language model for automated essay scoring. *Research Methods in Applied Linguistics*, 2(2):100050.

Lekshmi R Pillai, G Veena, and Deepa Gupta. 2018. A combined approach using semantic role labelling and word sense disambiguation for question generation and answer extraction. In *2018 Second International Conference on Advances in Electronics, Computers and Communications (ICAECC)*, pages 1–6. IEEE.

Johannes Schneider, Bernd Schenk, and Christina Niklaus. 2023. Towards llm-based autograding for short textual answers. *arXiv preprint arXiv:2309.11508*.

V Sreevidhya and Jayasree Narayanan. 2021. Short descriptive answer evaluation using word-embedding techniques. In *2021 12th international conference on computing communication and networking technologies (ICCCNT)*, pages 1–4. IEEE.

Pavel Stefanovič, Birutė Pliuskuvienė, Urtė Radvilaitė, and Simona Ramanauskaitė. 2024. Machine learning model for chatgpt usage detection in students' answers to open-ended questions: Case of lithuanian language. *Education and Information Technologies*, pages 1–23.

Chul Sung, Tejas Dhamecha, Swarnadeep Saha, Tengfei Ma, Vinay Reddy, and Rishi Arora. 2019. Pre-training bert on domain resources for short answer grading. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6071–6075.

Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, Soroosh Mariooryad, Yifan Ding, Xinyang Geng, Fred Alcober, Roy Frostig, Mark Omernick, Lexi Walker, Cosmin Paduraru, Christina Sorokin, Andrea Tacchetti, Colin Gaffney, Samira Daruki, Olcan Sercinoglu, Zach Gleicher, Juliette Love, Paul

Voigtlaender, Rohan Jain, Gabriela Surita, Kareem Mohamed, Rory Blevins, Junwhan Ahn, Tao Zhu, Kornraphop Kawintiranon, Orhan Firat, Yiming Gu, Yujing Zhang, Matthew Rahtz, Manaal Faruqui, Natalie Clay, Justin Gilmer, JD Co-Reyes, Ivo Penchev, Rui Zhu, Nobuyuki Morioka, Kevin Hui, Krishna Haridasan, Victor Campos, Mahdis Mahdieh, Mandy Guo, Samer Hassan, Kevin Kilgour, Arpi Vezer, Heng-Tze Cheng, Raoul de Liedekerke, Siddharth Goyal, Paul Barham, DJ Strouse, Seb Noury, Jonas Adler, Mukund Sundararajan, Sharad Vikram, Dmitry Lepikhin, Michela Paganini, Xavier Garcia, Fan Yang, Dasha Valter, Maja Trebacz, Kiran Vodrahalli, Chulayuth Asawaroengchai, Roman Ring, Norbert Kalb, Livio Baldini Soares, Siddhartha Brahma, David Steiner, Tianhe Yu, Fabian Mentzer, Antoine He, Lucas Gonzalez, Bibo Xu, Raphael Lopez Kaufman, Laurent El Shafey, Junhyuk Oh, Tom Hennigan, George van den Driessche, Seth Odoom, Mario Lucic, Becca Roelofs, Sid Lall, Amit Marathe, Betty Chan, Santiago Ontanon, Luheng He, Denis Teplyashin, Jonathan Lai, Phil Crone, Bogdan Damoc, Lewis Ho, Sebastian Riedel, Karel Lenc, Chih-Kuan Yeh, Aakanksha Chowdhery, Yang Xu, Mehran Kazemi, Ehsan Amid, Anastasia Petrushkina, Kevin Swersky, Ali Khodaei, Gowoon Chen, Chris Larkin, Mario Pinto, Geng Yan, Adria Puigdomenech Badia, Piyush Patil, Steven Hansen, Dave Orr, Sebastien M. R. Arnold, Jordan Grimstad, Andrew Dai, Sholto Douglas, Rishika Sinha, Vikas Yadav, Xi Chen, Elena Gribovskaya, Jacob Austin, Jeffrey Zhao, Kaushal Patel, Paul Komarek, Sophia Austin, Sebastian Borgeaud, Linda Friso, Abhimanyu Goyal, Ben Caine, Kris Cao, Da-Woon Chung, Matthew Lamm, Gabe Barth-Maron, Thais Kagohara, Kate Olszewska, Mia Chen, Kaushik Shivakumar, Rishabh Agarwal, Harshal Godhia, Ravi Rajwar, Javier Snaider, Xerxes Dotiwalla, Yuan Liu, Aditya Barua, Victor Ungureanu, Yuan Zhang, Bat-Orgil Batsaikhan, Mateo Wirth, James Qin, Ivo Danihelka, Tulsee Doshi, Martin Chadwick, Jilin Chen, Sanil Jain, Quoc Le, Arjun Kar, Madhu Gurumurthy, Cheng Li, Ruoxin Sang, Fangyu Liu, Lampros Lamprou, Rich Munoz, Nathan Lintz, Harsh Mehta, Heidi Howard, Malcolm Reynolds, Lora Aroyo, Quan Wang, Lorenzo Blanco, Albin Cassirer, Jordan Griffith, Dipanjan Das, Stephan Lee, Jakub Sygnowski, Zach Fisher, James Besley, Richard Powell, Zafarali Ahmed, Dominik Paulus, David Reitter, Zalan Borsos, Rishabh Joshi, Aedan Pope, Steven Hand, Vittorio Selo, Vihan Jain, Nikhil Sethi, Megha Goel, Takaki Makino, Rhys May, Zhen Yang, Johan Schalkwyk, Christina Butterfield, Anja Hauth, Alex Goldin, Will Hawkins, Evan Senter, Sergey Brin, Oliver Woodman, Marvin Ritter, Eric Noland, Minh Giang, Vijay Bolina, Lisa Lee, Tim Blyth, Ian Mackinnon, Machel Reid, Obaid Sarvana, David Silver, Alexander Chen, Lily Wang, Loren Maggiore, Oscar Chang, Nithya Attaluri, Gregory Thornton, Chung-Cheng Chiu, Oskar Bunyan, Nir Levine, Timothy Chung, Evgenii Eltyshev, Xiance Si, Timothy Lillicrap, Demetra Brady, Vaibhav Aggarwal, Boxi Wu, Yuanzhong Xu, Ross McIlroy, Kartikeya Badola, Paramjit Sandhu, Erica Moreira, Wojciech Stokowiec, Ross Hemsley, Dong Li, Alex Tudor, Pranav Shyam, Elahe Rahimtoroghi, Salem Haykal, Pablo Sprechmann, Xiang Zhou, Diana Mincu, Yujia Li, Ravi Addanki, Kalpesh Krishna, Xiao Wu, Alexandre Frechette, Matan Eyal, Allan Dafoe, Dave Lacey, Jay Whang, Thi Avrahami, Ye Zhang, Emanuel Taropa, Hanzhao Lin, Daniel Toyama, Eliza Rutherford, Motoki Sano, HyunJeong Choe, Alex Tomala, Chalence Safranek-Shrader, Nora Kassner, Mantas Pajarskas, Matt Harvey, Sean Sechrist, Meire Fortunato, Christina Lyu, Gamaleldin Elsayed, Chenkai Kuang, James Lottes, Eric Chu, Chao Jia, Chih-Wei Chen, Peter Humphreys, Kate Baumli, Connie Tao, Rajkumar Samuel, Cicero Nogueira dos Santos, Anders Andreassen, Nemanja Rakićević, Dominik Grewe, Aviral Kumar, Stephanie Winkler, Jonathan Caton, Andrew Brock, Sid Dalmia, Hannah Sheahan, Iain Barr, Yingjie Miao, Paul Natsev, Jacob Devlin, Feryal Behbahani, Flavien Prost, Yanhua Sun, Artiom Myaskovsky, Thanumalayan Sankaranarayana Pillai, Dan Hurt, Angeliki Lazaridou, Xi Xiong, Ce Zheng, Fabio Pardo, Xiaowei Li, Dan Horgan, Joe Stanton, Moran Ambar, Fei Xia, Alejandro Lince, Mingqiu Wang, Basil Mustafa, Albert Webson, Hyo Lee, Rohan Anil, Martin Wicke, Timothy Dozat, Abhishek Sinha, Enrique Piqueras, Elahe Dabir, Shyam Upadhyay, Anudhyan Boral, Lisa Anne Hendricks, Corey Fry, Josip Djolonga, Yi Su, Jake Walker, Jane Labanowski, Ronny Huang, Vedant Misra, Jeremy Chen, RJ Skerry-Ryan, Avi Singh, Shruti Rijhwani, Dian Yu, Alex Castro-Ros, Beer Changpinyo, Romina Datta, Sumit Bagri, Arnar Mar Hrafnkelsson, Marcello Maggioni, Daniel Zheng, Yury Sulsky, Shaobo Hou, Tom Le Paine, Antoine Yang, Jason Riesa, Dominika Rogozinska, Dror Marcus, Dalia El Badawy, Qiao Zhang, Luyu Wang, Helen Miller, Jeremy Greer, Lars Lowe Sjos, Azade Nova, Heiga Zen, Rahma Chaabouni, Mihaela Rosca, Jiepu Jiang, Charlie Chen, Ruibo Liu, Tara Sainath, Maxim Krikun, Alex Polozov, Jean-Baptiste Lespiau, Josh Newlan, Zeyncep Cankara, Soo Kwak, Yunhan Xu, Phil Chen, Andy Coenen, Clemens Meyer, Katerina Tsihlas, Ada Ma, Juraj Gottweis, Jinwei Xing, Chenjie Gu, Jin Miao, Christian Frank, Zeynep Cankara, Sanjay Ganapathy, Ishita Dasgupta, Steph Hughes-Fitt, Heng Chen, David Reid, Keran Rong, Hongmin Fan, Joost van Amersfoort, Vincent Zhuang, Aaron Cohen, Shixiang Shane Gu, Anhad Mohananey, Anastasija Ilic, Taylor Tobin, John Wieting, Anna Bortsova, Phoebe Thacker, Emma Wang, Emily Caveness, Justin Chiu, Eren Sezener, Alex Kaskasoli, Steven Baker, Katie Millican, Mohamed Elhawaty, Kostas Aisopos, Carl Lebsack, Nathan Byrd, Hanjun Dai, Wenhao Jia, Matthew Wiethoff, Elnaz Davoodi, Albert Weston, Lakshman Yagati, Arun Ahuja, Isabel Gao, Golan Pundak, Susan Zhang, Michael Azzam, Khe Chai Sim, Sergi Caelles, James Keeling, Abhanshu Sharma, Andy Swing, YaGuang Li, Chenxi Liu, Carrie Grimes

Bostock, Yamini Bansal, Zachary Nado, Ankesh Anand, Josh Lipschultz, Abhijit Karmarkar, Lev Proleev, Abe Ittycheriah, Soheil Hassas Yeganeh, George Polovets, Aleksandra Faust, Jiao Sun, Alban Rrustemi, Pen Li, Rakesh Shivanna, Jeremiah Liu, Chris Welty, Federico Lebron, Anirudh Baddepudi, Sebastian Krause, Emilio Parisotto, Radu Soricut, Zheng Xu, Dawn Bloxwich, Melvin Johnson, Behnam Neyshabur, Justin Mao-Jones, Renshen Wang, Vinay Ramasesh, Zaheer Abbas, Arthur Guez, Constant Segal, Duc Dung Nguyen, James Svensson, Le Hou, Sarah York, Kieran Milan, Sophie Bridgers, Wiktor Gworek, Marco Tagliasacchi, James Lee-Thorp, Michael Chang, Alexey Guseynov, Ale Jakse Hartman, Michael Kwong, Ruizhe Zhao, Sheleem Kashem, Elizabeth Cole, Antoine Miech, Richard Tanburn, Mary Phuong, Filip Pavetic, Sebastien Cevey, Ramona Comanescu, Richard Ives, Sherry Yang, Cosmo Du, Bo Li, Zizhao Zhang, Mariko Iinuma, Clara Huiyi Hu, Aurko Roy, Shaan Bijwadia, Zhenkai Zhu, Danilo Martins, Rachel Saputro, Anita Gergely, Steven Zheng, Dawei Jia, Ioannis Antonoglou, Adam Sadovsky, Shane Gu, Yingying Bi, Alek Andreev, Sina Samangooei, Mina Khan, Tomas Kocisky, Angelos Filos, Chintu Kumar, Colton Bishop, Adams Yu, Sarah Hodkinson, Sid Mittal, Premal Shah, Alexandre Moufarek, Yong Cheng, Adam Bloniarz, Jaehoon Lee, Pedram Pejman, Paul Michel, Stephen Spencer, Vladimir Feinberg, Xuehan Xiong, Nikolay Savinov, Charlotte Smith, Siamak Shakeri, Dustin Tran, Mary Chesus, Bernd Bohnet, George Tucker, Tamara von Glehn, Carrie Muir, Yiran Mao, Hideto Kazawa, Ambrose Slone, Kedar Soparkar, Disha Shrivastava, James Cobon-Kerr, Michael Sharman, Jay Pavagadhi, Carlos Araya, Karolis Misiunas, Nimesh Ghelani, Michael Laskin, David Barker, Qiujia Li, Anton Briukhov, Neil Houlsby, Mia Glaese, Balaji Lakshminarayanan, Nathan Schucher, Yunhao Tang, Eli Collins, Hyeontaek Lim, Fangxiaoyu Feng, Adria Recasens, Guangda Lai, Alberto Magni, Nicola De Cao, Aditya Siddhant, Zoe Ashwood, Jordi Orbay, Mostafa Dehghani, Jenny Brennan, Yifan He, Kelvin Xu, Yang Gao, Carl Saroufim, James Molloy, Xinyi Wu, Seb Arnold, Solomon Chang, Julian Schrittwieser, Elena Buchatskaya, Soroush Radpour, Martin Polacek, Skye Giordano, Ankur Bapna, Simon Tokumine, Vincent Hellendoorn, Thibault Sottiaux, Sarah Cogan, Aliaksei Severyn, Mohammad Saleh, Shantanu Thakoor, Laurent Shefey, Siyuan Qiao, Meenu Gaba, Shuo yiin Chang, Craig Swanson, Biao Zhang, Benjamin Lee, Paul Kishan Rubenstein, Gan Song, Tom Kwiatkowski, Anna Koop, Ajay Kannan, David Kao, Parker Schuh, Axel Stjerngren, Golnaz Ghiasi, Gena Gibson, Luke Vilnis, Ye Yuan, Felipe Tiengo Ferreira, Aishwarya Kamath, Ted Klimenko, Ken Franko, Kefan Xiao, Indro Bhattacharya, Miteyan Patel, Rui Wang, Alex Morris, Robin Strudel, Vivek Sharma, Peter Choy, Sayed Hadi Hashemi, Jessica Landon, Mara Finkelstein, Priya Jhakra, Justin Frye, Megan Barnes, Matthew Mauger, Dennis Daun, Khuslen Baatarsukh, Matthew Tung, Wael Farhan, Henryk Michalewski, Fabio Viola, Felix de Chaumont Quitry, Charline Le Lan, Tom Hudson, Qingze Wang, Felix Fischer, Ivy Zheng, Elspeth White, Anca Dragan, Jean baptiste Alayrac, Eric Ni, Alexander Pritzel, Adam Iwanicki, Michael Isard, Anna Bulanova, Lukas Zilka, Ethan Dyer, Devendra Sachan, Srivatsan Srinivasan, Hannah Muckenhirn, Honglong Cai, Amol Mandhane, Mukarram Tariq, Jack W. Rae, Gary Wang, Kareem Ayoub, Nicholas FitzGerald, Yao Zhao, Woohyun Han, Chris Alberti, Dan Garrette, Kashyap Krishnakumar, Mai Gimenez, Anselm Levskaya, Daniel Sohn, Josip Matak, Inaki Iturrate, Michael B. Chang, Jackie Xiang, Yuan Cao, Nishant Ranka, Geoff Brown, Adrian Hutter, Vahab Mirrokni, Nanxin Chen, Kaisheng Yao, Zoltan Egyed, Francois Galilee, Tyler Liechty, Praveen Kallakuri, Evan Palmer, Sanjay Ghemawat, Jasmine Liu, David Tao, Chloe Thornton, Tim Green, Mimi Jasarevic, Sharon Lin, Victor Cotruta, Yi-Xuan Tan, Noah Fiedel, Hongkun Yu, Ed Chi, Alexander Neitz, Jens Heitkaemper, Anu Sinha, Denny Zhou, Yi Sun, Charbel Kaed, Brice Hulse, Swaroop Mishra, Maria Georgaki, Sneha Kudugunta, Clement Farabet, Izhak Shafran, Daniel Vlasic, Anton Tsitsulin, Rajagopal Ananthanarayanan, Alen Carin, Guolong Su, Pei Sun, Shashank V, Gabriel Carvajal, Josef Broder, Iulia Comsa, Alena Repina, William Wong, Warren Weilun Chen, Peter Hawkins, Egor Filonov, Lucia Loher, Christoph Hirnschall, Weiyi Wang, Jingchen Ye, Andrea Burns, Hardie Cate, Diana Gage Wright, Federico Piccinini, Lei Zhang, Chu-Cheng Lin, Ionel Gog, Yana Kulizhskaya, Ashwin Sreevatsa, Shuang Song, Luis C. Cobo, Anand Iyer, Chetan Tekur, Guillermo Garrido, Zhuyun Xiao, Rupert Kemp, Huaixiu Steven Zheng, Hui Li, Ananth Agarwal, Christel Ngani, Kati Goshvadi, Rebeca Santamaria-Fernandez, Wojciech Fica, Xinyun Chen, Chris Gorgolewski, Sean Sun, Roopal Garg, Xinyu Ye, S. M. Ali Eslami, Nan Hua, Jon Simon, Pratik Joshi, Yelin Kim, Ian Tenney, Sahitya Potluri, Lam Nguyen Thiet, Quan Yuan, Florian Luisier, Alexandra Chronopoulou, Salvatore Scellato, Praveen Srinivasan, Minmin Chen, Vinod Koverkathu, Valentin Dalibard, Yaming Xu, Brennan Saeta, Keith Anderson, Thibault Sellam, Nick Fernando, Fantine Huot, Junehyuk Jung, Mani Varadarajan, Michael Quinn, Amit Raul, Maigo Le, Ruslan Habalov, Jon Clark, Komal Jalan, Kalesha Bullard, Achintya Singhal, Thang Luong, Boyu Wang, Sujeevan Rajayogam, Julian Eisenschlos, Johnson Jia, Daniel Finchelstein, Alex Yakubovich, Daniel Balle, Michael Fink, Sameer Agarwal, Jing Li, Dj Dvijotham, Shalini Pal, Kai Kang, Jaclyn Konzelmann, Jennifer Beattie, Olivier Dousse, Diane Wu, Remi Crocker, Chen Elkind, Siddhartha Reddy Jonnalagadda, Jong Lee, Dan Holtmann-Rice, Krystal Kallarackal, Rosanne Liu, Denis Vnukov, Neera Vats, Luca Invernizzi, Mohsen Jafari, Huanjie Zhou, Lilly Taylor, Jennifer Prendki, Marcus Wu, Tom Eccles, Tianqi Liu, Kavya Kopparapu, Francoise Beaufays, Christof Angermueller, Andreea Marzoca, Shourya Sarcar,

Hilal Dib, Jeff Stanway, Frank Perbet, Nejc Trdin, Rachel Sterneck, Andrey Khorlin, Dinghua Li, Xihui Wu, Sonam Goenka, David Madras, Sasha Goldshtein, Willi Gierke, Tong Zhou, Yaxin Liu, Yannie Liang, Anais White, Yunjie Li, Shreya Singh, Sanaz Bahargam, Mark Epstein, Sujoy Basu, Li Lao, Adnan Ozturel, Carl Crous, Alex Zhai, Han Lu, Zora Tung, Neeraj Gaur, Alanna Walton, Lucas Dixon, Ming Zhang, Amir Globerson, Grant Uy, Andrew Bolt, Olivia Wiles, Milad Nasr, Ilia Shumailov, Marco Selvi, Francesco Piccinno, Ricardo Aguilar, Sara McCarthy, Misha Khalman, Mrinal Shukla, Vlado Galic, John Carpenter, Kevin Villela, Haibin Zhang, Harry Richardson, James Martens, Matko Bosnjak, Shreyas Rammohan Belle, Jeff Seibert, Mahmoud Alnahlawi, Brian McWilliams, Sankalp Singh, Annie Louis, Wen Ding, Dan Popovici, Lenin Simicich, Laura Knight, Pulkit Mehta, Nishesh Gupta, Chongyang Shi, Saaber Fatehi, Jovana Mitrovic, Alex Grills, Joseph Pagadora, Dessie Petrova, Danielle Eisenbud, Zhishuai Zhang, Damion Yates, Bhavishya Mittal, Nilesh Tripuraneni, Yannis Assael, Thomas Brovelli, Prateek Jain, Mihajlo Velimirovic, Canfer Akbulut, Jiaqi Mu, Wolfgang Macherey, Ravin Kumar, Jun Xu, Haroon Qureshi, Gheorghe Comanici, Jeremy Wiesner, Zhitao Gong, Anton Ruddock, Matthias Bauer, Nick Felt, Anirudh GP, Anurag Arnab, Dustin Zelle, Jonas Rothfuss, Bill Rosgen, Ashish Shenoy, Bryan Seybold, Xinjian Li, Jayaram Mudigonda, Goker Erdogan, Jiawei Xia, Jiri Simsa, Andrea Michi, Yi Yao, Christopher Yew, Steven Kan, Isaac Caswell, Carey Radebaugh, Andre Elisseeff, Pedro Valenzuela, Kay McKinney, Kim Paterson, Albert Cui, Eri Latorre-Chimoto, Solomon Kim, William Zeng, Ken Durden, Priya Ponnapalli, Tiberiu Sosea, Christopher A. Choquette-Choo, James Manyika, Brona Robenek, Harsha Vashisht, Sebastien Pereira, Hoi Lam, Marko Velic, Denese Owusu-Afriyie, Katherine Lee, Tolga Bolukbasi, Alicia Parrish, Shawn Lu, Jane Park, Balaji Venkatraman, Alice Talbert, Lambert Rosique, Yuchung Cheng, Andrei Sozanschi, Adam Paszke, Praveen Kumar, Jessica Austin, Lu Li, Khalid Salama, Wooyeol Kim, Nandita Dukkipati, Anthony Baryshnikov, Christos Kaplanis, XiangHai Sheng, Yuri Chervonyi, Caglar Unlu, Diego de Las Casas, Harry Askham, Kathryn Tunyasuvunakool, Felix Gimeno, Siim Poder, Chester Kwak, Matt Miecnikowski, Vahab Mirrokni, Alek Dimitriev, Aaron Parisi, Dangyi Liu, Tomy Tsai, Toby Shevlane, Christina Kouridi, Drew Garmon, Adrian Goedeckemeyer, Adam R. Brown, Anitha Vijayakumar, Ali Elqursh, Sadegh Jazayeri, Jin Huang, Sara Mc Carthy, Jay Hoover, Lucy Kim, Sandeep Kumar, Wei Chen, Courtney Biles, Garrett Bingham, Evan Rosen, Lisa Wang, Qijun Tan, David Engel, Francesco Pongetti, Dario de Cesare, Dongseong Hwang, Lily Yu, Jennifer Pullman, Srini Narayanan, Kyle Levin, Siddharth Gopal, Megan Li, Asaf Aharoni, Trieu Trinh, Jessica Lo, Norman Casagrande, Roopali Vij, Loic Matthey, Bramandia Ramadhana, Austin Matthews, CJ Carey, Matthew Johnson, Kremena Goranova, Rohin Shah, Shereen Ashraf, Kingshuk Dasgupta, Rasmus Larsen, Yicheng Wang, Manish Reddy Vuyyuru, Chong Jiang, Joana Ijazi, Kazuki Osawa, Celine Smith, Ramya Sree Boppana, Taylan Bilal, Yuma Koizumi, Ying Xu, Yasemin Altun, Nir Shabat, Ben Bariach, Alex Korchemniy, Kiam Choo, Olaf Ronneberger, Chimezie Iwuanyanwu, Shubin Zhao, David Soergel, Cho-Jui Hsieh, Irene Cai, Shariq Iqbal, Martin Sundermeyer, Zhe Chen, Elie Bursztein, Chaitanya Malaviya, Fadi Biadsy, Prakash Shroff, Inderjit Dhillon, Tejasi Latkar, Chris Dyer, Hannah Forbes, Massimo Nicosia, Vitaly Nikolaev, Somer Greene, Marin Georgiev, Pidong Wang, Nina Martin, Hanie Sedghi, John Zhang, Praseem Banzal, Doug Fritz, Vikram Rao, Xuezhi Wang, Jiageng Zhang, Viorica Patraucean, Dayou Du, Igor Mordatch, Ivan Jurin, Lewis Liu, Ayush Dubey, Abhi Mohan, Janek Nowakowski, Vlad-Doru Ion, Nan Wei, Reiko Tojo, Maria Abi Raad, Drew A. Hudson, Vaishakh Keshava, Shubham Agrawal, Kevin Ramirez, Zhichun Wu, Hoang Nguyen, Ji Liu, Madhavi Sewak, Bryce Petrini, DongHyun Choi, Ivan Philips, Ziyue Wang, Ioana Bica, Ankush Garg, Jarek Wilkiewicz, Priyanka Agrawal, Xiaowei Li, Danhao Guo, Emily Xue, Naseer Shaik, Andrew Leach, Sadh MNM Khan, Julia Wiesinger, Sammy Jerome, Abhishek Chakladar, Alek Wenjiao Wang, Tina Ornduff, Folake Abu, Alireza Ghaffarkhah, Marcus Wainwright, Mario Cortes, Frederick Liu, Joshua Maynez, Andreas Terzis, Pouya Samangouei, Riham Mansour, Tomasz Kepa, François-Xavier Aubet, Anton Algymr, Dan Banica, Agoston Weisz, Andras Orban, Alexandre Senges, Ewa Andrejczuk, Mark Geller, Niccolo Dal Santo, Valentin Anklin, Majd Al Merey, Martin Baeuml, Trevor Strohman, Junwen Bai, Slav Petrov, Yonghui Wu, Demis Hassabis, Koray Kavukcuoglu, Jeffrey Dean, and Oriol Vinyals. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context.

Qwen Team. 2024. Qwen2.5: A party of foundation models.

Maximilian Tornqvist, Mosleh Mahamud, Erick Mendez Guzman, and Alexandra Farazouli. 2023. Exasag: Explainable framework for automatic short answer grading. In *Proceedings of the 18th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023)*, pages 361–371.

Rebecka Weegar and Peter Idestam-Almquist. 2024. Reducing workload in short answer grading using machine learning. *International Journal of Artificial Intelligence in Education*, 34(2):247–273.

Nico Willms and Ulrike Padó. 2022. A transformer for sag: What does it grade? In *Swedish Language Technology Conference and NLP4CALL*, pages 114–122.

Kevin P Yancey, Geoffrey Laflair, Anthony Verardi, and Jill Burstein. 2023. Rating short l2 essays on the cefr scale with gpt-4. In *Proceedings of the 18th*

Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2023), pages 576–584.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, Jin Xu, Jingren Zhou, Jinze Bai, Jinzheng He, Junyang Lin, Kai Dang, Keming Lu, Keqin Chen, Kexin Yang, Mei Li, Mingfeng Xue, Na Ni, Pei Zhang, Peng Wang, Ru Peng, Rui Men, Ruize Gao, Runji Lin, Shijie Wang, Shuai Bai, Sinan Tan, Tianhang Zhu, Tianhao Li, Tianyu Liu, Wenbin Ge, Xiaodong Deng, Xiaohuan Zhou, Xingzhang Ren, Xinyu Zhang, Xipin Wei, Xuancheng Ren, Yang Fan, Yang Yao, Yichang Zhang, Yu Wan, Yunfei Chu, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zhihao Fan. 2024. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.

Su-Youn Yoon. 2023. Short answer grading using one-shot prompting and text similarity scoring model. *arXiv preprint arXiv:2305.18638*.

M Zhang, S Baral, N Heffernan, and A Lan. Automatic short math answer grading via in-context meta-learning. arxiv 2022. *arXiv preprint arXiv:2205.15219*.

A Manual Evaluation

For each language, we recruited two native speakers to evaluate the outputs of LLMs on the answers generation task. Each annotator was presented with 360 random samples from the dataset. Each sample contained a question, a reference answer, a generated answer with an instruction on whether it supposed to be matched with the reference answer. If the reference answer and the generated answer are matched and they are supposed to be matched or the reference answer and the generated answer are not matched and they are not supposed to be matched, the label *accept* was assigned to the sample. Otherwise, the label *reject* was assigned. For each model (LLaMa2:13b, GPT-4o, and LLaMa3:7b) and for each matched generation type, the annotators were presented with 25 samples. For non-matched generation methods, the annotators were presented with 40 samples. The aggregated results (after cleaning the duplicates) are presented in the Table 3.

To calculate the inter-annotator agreement, we used Cohen Kappa score (Cohen, 1960) and an intersection ratio. One of the annotators per language was presented with additional 40 samples from the labeled dataset of the other annotator, including the equal coverage of models and generation methods in the data. The Cohen Kappa coefficient for Latvian language was 0.285 and the intersection score was of 0.825. The Cohen Kappa coefficient for Lithuanian language was 0.354 and the intersection score was of 0.82.

Based on the the results, we kept LLaMa3 Non-Match Relat. generation results and all of the GPT-4o generated results in the dataset for Lithuanian. Similarly, we kept LLaMa3 Non-Match Minor-Changes and GPT-4o (except for Match MoreInfo, which was excluded by mistake) in the dataset for Latvian. Our results indicate that GPT-4o is capable of generating matched and non-matched answers with different methods in these languages, when LLaMa3 and LLaMa2 struggle.

Lit R	Lit A	Lat R	Lat A	Model	Class
1	29	1	29	GPT-4o	Match Ents
0	30	4	26	GPT-4o	Match MoreInfo
1	29	2	28	GPT-4o	Match Syns
0	30	2	28	GPT-4o	Match Style
0	60	3	56	GPT-4o	Non-Match MinorCh.
3	57	2	57	GPT-4o	Non-Match Relat.
20	10	12	18	LLaMa2:13b	Match Ents
22	8	15	14	LLaMa2:13b	Match MoreInfo
13	17	10	19	LLaMa2:13b	Match Syns
12	18	14	16	LLaMa2:13b	Match Style
16	43	14	15	LLaMa2:13b	Non-Match MinorCh.
12	46	13	32	LLaMa2:13b	Non-Match Relat.
10	20	5	25	LLaMa3:7b	Match Ents
5	25	8	22	LLaMa3:7b	Match MoreInfo
8	22	3	26	LLaMa3:7b	Match Syns
13	16	14	16	LLaMa3:7b	Match Style
6	54	4	56	LLaMa3:7b	Non-Match MinorCh.
2	58	6	48	LLaMa3:7b	Non-Match Relat.
5	235	14	224	GPT-4o	All
95	142	79	144	LLaMa2:13b	All
44	195	40	193	LLaMa3:7b	All

Table 3: Annotation results. **R** and **A** indicate amount of rejected and accepted samples respectively with the language at the beginning. **Class** indicates a generation prompt that was used and whether it should match with the reference answer.

B Failure Cases Showcase

In this section, we present a sample of the samples per model and per generation rule, where the annotators marked the sample as incorrect, along with the annotator's comment on the decision if provided. The examples are provided in the subsections below.

B.1 Failure Examples for Lithuanian

Language / Rule: Lithuanian / M: Ents
Model Name: GPT-4o
Q (LT): Kokie buvo svarbiausi Kęstučio politiniai vaidmenys Lietuvos Didžiojoje Kunigaikštystėje ?
Q (EN): What were Kęstutis' most important political roles in the Grand Duchy of Lithuania?
Correct Answer (LT):Kęstutis užėmė keletą svarbių politinių vaidmenų. Jis buvo LDK submonarchas nuo 1337 ar 1345 m. iki 1381 m., o vėliau Lietuvos didysis kunigaikštis 1381–1382 m. Be to, jis valdė Gardiną ir Palenkę iki 1337 m., Žemaitiją nuo 1337 m., ir Trakus nuo 1337 ar 1345 m.
Correct Answer (EN): Kęstutis held several important political roles. He was a submonarch of the Grand Duchy of Lithuania from 1337 or 1345 to 1381, and later Grand Duke of Lithuania from 1381–1382. In addition, he ruled Grodno and Podlaskie until 1337, Samogitia from 1337, and Trakai from 1337 or 1345.
Generated (LT): Kęstutis turėjo keletą svarbių politinių pareigų Lietuvos Didžiojoje Kunigaikštystėje. Jis veikė kaip LDK submonarchas nuo 1337 arba 1345 iki 1381 metų prekybos su pagrindiniu Etijopija centru, o po to tapo Lietuvos didžiuoju kunigaikščiu nuo 1381 iki 1382 metų. Kitas svarbus postas buvo Gardino ir Palenkės valdymas iki 1337 metų, taip pat Žemaitijos valdymas nuo 1337 metų ir Trakų nuo 1337 arba 1345 metų.
Generated (EN): Kęstutis held several important political roles. He was a sub-monarch of the Grand Duchy of Lithuania from 1337 or 1345 to 1381, and later Grand Duke of Lithuania from 1381–1382. In addition, he ruled Grodno and Podlaskie until 1337, Samogitia from 1337, and Trakai from 1337 or 1345.
Comment: NA

Language / Rule: Lithuanian / M: Ents
Model Name: LLaMa3
Q (LT): Kodėl Valstiečių ir Naujosios demokratijos partijos susijungė 2001 metais?
Q (EN): On what occasions is it mandatory to sing or play the Lithuanian anthem?
Correct Answer (LT): Valstiečių ir Naujosios demokratijos partijos susijungė 2001 metais, siekdamos pagerinti savo pozicijas būsimuose rinkimuose ir tapti parlamentine partija. Valstiečių partijai grėsė likti regionine partija, turinčia įtakos tik rajonų savivaldybėse. Susijungimas pasiteisino - 2004 metais jungtinė partija gavo mandatą Europos Parlamente ir 10 mandatų Seime.
Correct Answer (EN): The Lithuanian anthem must be sung or played on various occasions, such as at the beginning and end of the Seimas sessions, at ceremonial meetings dedicated to Lithuanian holidays, when meeting foreign heads of state, when representing Lithuania abroad, on national television and radio broadcasts, and when raising the Lithuanian flag.
Generated (LT): Valstiečių ir Naujosios demokratijos partijos susijungimą 2001 metais galima aiškinti, jog tai buvo siekimas pagerinti pozicijas būsimuose rinkimuose ir tapti parlamentine partija. Valstiečių partija, kuri turėjo įtakos tik rajonų savivaldybėse, norėjo likti regionine partija, o susijungimas leido jiems dalyvauti Europos Parlamente ir Seime.
Generated (EN): The Lithuanian anthem must be sung or played when Lithuania celebrates its holidays, when the Seimas convenes, when meeting foreign heads of state, when representing the country abroad, on national television and radio broadcasts.
Comment: both wrong - grammar and logic

Language / Rule: Lithuanian / M: Ents
Model Name: LLaMa3
Q (LT): Kuo daugiausia minta baltieji gandrai?
Q (EN): What do white storks mainly eat?
Correct Answer (LT): Baltieji gandrai yra oportunistiniai "grobuonys", mintantys gausiausiais, beveik vien gyvūninės kilmės

maisto objektais. Jie gaudo vabzdžius, kitus bestuburius, žuvis, varliagyvius, ant žemės perinčių paukščių jauniklius. Kartais ryja peles, žiurkes, įvairius roplius, taip pat ir gyvates, neretai kiškių jauniklius ar net žebenkštis.
Correct Answer (EN): White storks are opportunistic "predators", feeding on the most abundant, almost exclusively animal-based food objects. They catch insects, other invertebrates, fish, amphibians, and ground-nesting bird chicks. Sometimes they devour mice, rats, various reptiles, as well as snakes, often hare chicks or even weasels.
Generated (LT): Baltieji gandrai yra panaudotinės "grobuonys", mintantys gausiausiais, beveik vien gyvūninės kilmės maisto objektais, tad jų šokas yra labai išvystytas. Jie gaudo vabzdžius, kitus bestuburius, žuvis, varliagyvius, ant žemės perinčių paukščių jauniklius.
Generated (EN): White storks are utilitarian "predators", feeding on the most abundant, almost exclusively animal-based food objects, so their shock is highly developed. They catch insects, other invertebrates, fish, amphibians, and ground-nesting bird chicks.
Comment: NA

Language / Rule: Lithuanian / M: Ents
Model Name: LlaMa2:13b
Q (LT): Kokias pareigas Algirdas Brazauskas ėjo nepriklausomoje Lietuvoje?
Q (EN): What positions did Algirdas Brazauskas hold in independent Lithuania?
Correct Answer (LT): Algirdas Brazauskas nepriklausomoje Lietuvoje ėjo prezidento (1992–1993 laikinai, 1993–1998 išrinktas) ir ministro pirmininko (2001–2006) pareigas.
Correct Answer (EN): Algirdas Brazauskas held the positions of President (1992–1993 interim, 1993–1998 elected) and Prime Minister (2001–2006) in independent Lithuania.
Generated (LT): Algirdas Brazauskas nepriklausomoje Lietuvoje ėjo prezidento ir ministro pirmininko pareigas, tokių kaip prezidentas (1992–1993 laikinai) ir ministra (2001–2006) pareigas.
Generated (EN): Algirdas Brazauskas held the positions of President and Prime Minister in independent Lithuania, such as President (1992–1993 interim) and Minister (2001–2006).
Comment: NA

Language / Rule: Lithuanian / M: Ents
Model Name: LLaMa2:13b
Q (LT): Kokios dvi seniūnaitijos priklauso Kernavės seniūnijai?
Q (EN): What two elderships belong to Kernavės eldership?
Correct Answer (LT): Kernavės seniūnijoje yra dvi seniūnaitijos: Kerniaus ir Piliakalnio.
Correct Answer (EN): There are two elderships in Kernavės eldership: Kerniaus and Piliakalnis.
Generated (LT): Kernavės seniūnijoje yra dvi seniūnaitijos: Kerniaus ir Piliakallio, tačiau tokių kaip Kernavės miestelė ir Piliakalnio savivaldybė.[Inst]
Generated (EN): There are two elderships in Kernavės eldership: Kerniaus and Piliakalnis, but such as Kernavės town and Piliakalnis municipality.
Comment: NA

Language / Rule: Lithuanian / M: Exclamatory
Model Name: LLaMa3
Q (LT): Kokias pareigas Algirdas Brazauskas ėjo nepriklausomoje Lietuvoje?
Q (EN): What positions did Algirdas Brazauskas hold in independent Lithuania?
Correct Answer (LT): Algirdas Brazauskas nepriklausomoje Lietuvoje ėjo prezidento (1992–1993 laikinai, 1993–1998 išrinktas) ir ministro pirmininko (2001–2006) pareigas.
Correct Answer (EN): In independent Lithuania, Algirdas Brazauskas held the positions of President (1992–1993 interim, 1993–1998 elected) and Prime Minister (2001–2006).
Generated (LT): Algirdas Brazauskas - tuo metu Lietuvos prezidentas ir ministras pirmininkas, atlikęs svarbias vietas nepriklausomoje šalyje!
Generated (EN): Algirdas Brazauskas - at that time the President and Prime Minister of Lithuania, who held important positions in the independent country!
Comment: NA

Language / Rule: Lithuanian / M: Exclamatory
Model Name: LLaMa3
Q (LT): Kada ir kodēl baigēsi pirmasis Kauno "Žalgirio" dominavimo laikotarpis LKL?
Q (EN): When and why did Kaunas Žalgiris' first period of dominance in the LKL end?
Correct Answer (LT): Kauno "Žalgirio" dominavimas, trukęs pirmuosius šešis sezonus (1993–1999 m.), baigēsi po jų pergalių Europos Taurēs ir Eurolygos varžybose. Kitą sezoną (1999–2000 m.) LKL čempionais tapo Vilniaus "Lietuvos rytas".
Correct Answer (EN): Kaunas Žalgiris' dominance, which lasted for the first six seasons (1993–1999), ended after their victories in the European Cup and Euroleague. The next season (1999–2000), Vilnius Lietuvos rytas became the LKL champions.
Generated (LT): Kauno "Žalgirio" dominavimas baigēsi po šešių sezučių, kai jie pergudrio Europos Taurę ir Eurolygą, o po to Vilniaus "Lietuvos ryto" tapo LKL čempionais!
Generated (EN): Kaunas Žalgiris' dominance ended after six seasons, when they won the European Cup and Euroleague, and then Vilnius Lietuvos rytas became the LKL champions!
Comment: NA

Language / Rule: Lithuanian / M: Exclamatory
Model Name: LLaMa2:13b
Q (LT): Kokia yra Gedimino pilies atsiradimo legenda?
Q (EN): What is the legend of the origin of Gediminas Castle?
Correct Answer (LT): Legenda pasakoja, kad kunigaikštis Gediminas medžiodamas užmigo Šventaragio slēnyje ir susapnavo geležinį vilką, staugiantį ant kalno prie Neries ir Vilnios santakos. Žynys Lizdeika šį sapną išaiškino kaip pranašystę apie didžio miesto atsiradimą toje vietoje.
Correct Answer (EN): The legend says that Duke Gediminas fell asleep while hunting in the Šventaragis Valley and dreamed of an iron wolf howling on a mountain at the confluence of the Neris and Vilnius rivers. The seer Lizdeika interpreted this dream as a prophecy about the emergence of a great city in that place.
Generated (LT): Gediminas pilies atsiradimo legenda – pasaulinē istorija, kuriuo kalbējo apie kunigaikščį, ką ir kiek kito, susapnavo geležinį vilką ir šventaragio slēnyje
Generated (EN): The legend of the origin of Gediminas Castle is a world story that spoke about the duke, what and how much else, dreamed of an iron wolf and in the Šventaragis Valley
Comment: NA

Language / Rule: Lithuanian / M: Exclamatory
Model Name: LLaMa2:13b
Q (LT): Kada ir kokiomis aplinkybēmis Klaipēda buvo prijungta prie Lietuvos?
Q (EN): When and under what circumstances was Klaipėda annexed to Lithuania?
Correct Answer (LT): Klaipēda buvo prijungta prie Lietuvos 1923 m. po inscenizuoto sukilimo. Prieš tai miestas buvo atskirtas nuo Vokietijos pagal Versalio taikos sutartį po Pirmojo pasaulinio karo.
Correct Answer (EN): Klaipēda was annexed to Lithuania in 1923 after a staged uprising. Before that, the city was separated from Germany under the Treaty of Versailles after World War I.
Generated (LT): Klaipēda buvo prieš 1923 m. prijungta prie Lietuvos, po inscenizuoto sukilimo, prisijungant Vokietijo pagal Versalio taikos sutartį po Pirmojo pasaulinio karo
Generated (EN): Klaipēda was annexed to Lithuania before 1923, after a staged uprising, with the accession of Germany under the Treaty of Versailles after World War I.
Comment: NA

Language / Rule: Lithuanian / M: Synonyms
Model Name: GPT-4o
Q (LT): Kada ir kokiomis aplinkybēmis Klaipēdos kraštas buvo atskirtas nuo Vokietijos?
Q (EN): When and under what circumstances was the Klaipēda region separated from Germany?
Correct Answer (LT): Klaipēdos kraštas buvo atskirtas nuo Vokietijos 1919 metais pa-

gal Versalio taikos sutartį. Ši sutartis numatė, kad šiaurinė Rytų Prūsijos dalis, kuri ir tapo Klaipėdos kraštu, būtų atskirta nuo Vokietijos ir perduota sąjungininkų valstybių žinion. Šis sprendimas buvo priimtas po Paryžiaus taikos konferencijos, kurioje Lenkijos atstovas iš pradžių reikalavo Mažąją Lietuvą sujungti su Didžiąja ir abi prijungti prie Lenkijos, bet vėliau pasiūlė prie Lietuvos prijungti tik Klaipėdos kraštą, kuris iki Lietuvos ir Lenkijos sąjungos turėtų būti Antantės kontrolėje.
Correct Answer (EN): The Klaipėda region was separated from Germany in 1919 under the Treaty of Versailles. This treaty provided that the northern part of East Prussia, which became the Klaipėda region, would be separated from Germany and transferred to the control of the Allied powers. This decision was made after the Paris Peace Conference, where the Polish representative initially demanded that Lithuania Minor be united with Lithuania Major and both be annexed to Poland, but later proposed that only the Klaipėda region be annexed to Lithuania, which should have been under the control of the Entente until the union of Lithuania and Poland.
Generated (LT): 1919 metais, remiantis Versalio taikos sutartimi, Klaipėdos kraštas buvo atskirtas nuo Vokietijos. Pagal šią sutartį šiaurinė Rytų Prūsijos dalis tapo Klaipėdos kraštu ir buvo perduota sąjungininkų valstybių kontrolei. Šis sprendimas buvo priimtas Paryžiaus taikos konferencijos metu. Lenkijos atstovas iš pradžių siekė sujungti Mažąją Lietuvą su Didžiąja Lietuva ir prijungti jas abi prie Lenkijos, bet vėliau pasiūlė Klaipėdos kraštą prijungti prie Lietuvos su sąlyga, kad jis iki Lietuvos ir Lenkijos sąjungos būtų Antantės valdymo.
Generated (EN): In 1919, according to the Treaty of Versailles, the Klaipėda region was separated from Germany. According to this treaty, the northern part of East Prussia became the Klaipėda region and was transferred to the control of the Allied powers. This decision was made during the Paris Peace Conference. The Polish representative initially sought to unite Lithuania Minor with Lithuania Greater and annex both to Poland, but later proposed annexing the Klaipėda region to Lithuania on the condition that it would be under the rule of the Entente until the union of Lithuania and Poland.
Comment: NA

Language / Rule: Lithuanian / M: Synonyms
Model Name: LLaMa3
Q (LT): Kada ir kur Jonas Žemaitis-Vytautas buvo suimtas ir kada įvykdyta mirties bausmė?
Q (EN): When and where was Jonas vZemaitis-Vytautas arrested and when was he executed?
Correct Answer (LT): Jonas Žemaitis-Vytautas buvo suimtas 1953 m. gegužės 30 d. Šimkaičių miške, Jurbarko rajone, bunkeryje. Sušaudytas 1954 m. lapkričio 26 d. Maskvos Butyrkų kalėjime.
Correct Answer (EN): Jonas Žemaitis-Vytautas was arrested on May 30, 1953 in a bunker in the Šimkaičiai Forest, Jurbarkas District. He was shot on November 26, 1954 in the Butyrka Prison in Moscow.
Generated (LT): Jonas Žemaitis-Vytautas buvo suimtas 1953 metais gegužės trečiadienį Šimkaičių miške, Jurbarko rajone, bunkeryje. Mirties bausmė įvyko 1954 metų lapkričio 26 dieną Maskvos Butyrkų kalėjime.
Generated (EN): Jonas Žemaitis-Vytautas was arrested on May 1, 1953 in a bunker in the Šimkaičia Forest, Jurbarkas District. He was executed on November 26, 1954 in the Butyrka Prison in Moscow.
Comment: NA

Language / Rule: Lithuanian / M: Synonyms
Model Name: LLaMa3
Q (LT): Kada minima Kėdainių miesto gimtadienio šventė?
Q (EN): When is the birthday of the city of Kėdainiai celebrated?
Correct Answer (LT): Kėdainių miesto gimtadienis švenčiamas rugpjūčio 8–10 dienomis.
Correct Answer (EN): The birthday of the city of Kėdainiai is celebrated on August 8-10.
Generated (LT): Kėdainių miesto jubiluojimo atšvarką švenčia rugpjūčio 8-10 dienomis.

Generated (EN): The anniversary of the city of Kēdainiai is celebrated on August 8-10.
Comment: NA

Language / Rule: Lithuanian / M: Synonyms
Model Name: LLaMa2:13b
Q (LT): Kada ir kur Marija Pečkauskaitē pradējo spausdinti savo kūrinius?
Q (EN): When and where did Marija Pečkauskaitė start publishing her works?
Correct Answer (LT): Paskatinta Povilo Višinskio, Marija Pečkauskaitē pradējo spausdinti apsakymus Lietuvos periodinēje spaudoje nuo 1896 metų.
Correct Answer (EN): Encouraged by Povilas Višinskis, Marija Pečkauskaitė started publishing her short stories in Lithuanian periodicals from 1896.
Generated (LT): Marija Pečkauskaitē pradējo spausdinti savo kūrinius iki Povilo Višinskio apsakymus Lietuvos periodinēje spaudoje nuo 1896 metų.[Inst]
Generated (EN): Marija Pečkauskaitė started publishing her works before Povilas Višinskis' short stories in Lithuanian periodicals from 1896.
Comment: NA

Language / Rule: Lithuanian / M: Synonyms
Model Name: LLaMa2:13b
Q (LT): Kuo skiriasi bobų vasara nuo paprasto vasaros sezono užsitęsimo rugsējo pradžioje?
Q (EN): What is the difference between a bob summer and a simple extension of the summer season at the beginning of September?
Correct Answer (LT): Bobų vasara negali būti pirmoje rugsējo pusēje. Jei šiuo metu ateina šiltas ir sausas oras, tai meteorologai laiko vasaros sezono užsitęsimu. Tikrasis bobų vasaros reiškinys susijęs su specifine atmosferos cirkuliacija, kurią lemia Azorų anticiklonas.
Correct Answer (EN): A bob summer cannot occur in the first half of September. If warm and dry weather arrives at this time, meteorologists consider it an extension of the summer season. The real bob summer phenomenon is related to a specific atmospheric circulation caused by the Azores anticyclone.
Generated (LT): Bobų vasara negali būti pirmoje rugsējo pusēje, nes tai meteorologiniai laiko užsitęsimai susijęs su specifine atmosferos cirkuliacija, kurią lemia Azorų anticiklonas.[Inst]
Generated (EN): A bob summer cannot occur in the first half of September, because it is a meteorological extension of time related to a specific atmospheric circulation caused by the Azores anticyclone.
Comment: NA

Language / Rule: Lithuanian / NM: Minor-Changes
Model Name: LLaMa3
Q (LT): Kokius klausimus padeda atsakyti prieveiksmiai, ir ar patys šie klausimai laikomi prieveiksmiais?
Q (EN): What questions do adverbs help answer, and are these questions themselves considered adverbs?
Correct Answer (LT): Prieveiksmiai padeda atsakyti į klausimus "kaip?", "kada?", "kur?", "kodēl?", ir "kiek?". Įdomu tai, kad patys šie klausimai taip pat yra laikomi prieveiksmiais.
Correct Answer (EN): Adverbs help answer the questions "how?", "when?", "where?", "why?", and "how much?". Interestingly, these questions themselves are also considered adverbs.
Generated (LT): Prieveiksmiai padeda atsakyti į klausimus "kaip", "kur", "kodēl?", ir "kiek?". Įdomu tai, kad patys šie klausimai laikomi užrašais.
Generated (EN): Adverbs help answer the questions "how?", "where?", "why?", and "how much?". Interestingly, these questions themselves are considered notes.
Comment: NA

Language / Rule: Lithuanian / NM: Minor-Changes
Model Name: LLaMa3
Q (LT): Kokios pagrindinēs temos vyrauja Justino Marcinkevičiaus kūryboje?
Q (EN): What are the main themes that dominate Justinas Marcinkevičius's work?
Correct Answer (LT): Pagrindinēs jo kūrybos temos buvo Lietuva, jos istorija, dabartis, gamta ir kultūra, žmogus Tēvynēje ir pasaulyje, žmogaus egzistencinē problematika: laimē, kančia, ištikimybē, pareiga,

dora.
Correct Answer (EN): The main themes of his work were Lithuania, its history, present, nature and culture, man in the Homeland and in the world, human existential problems: happiness, suffering, loyalty, duty, virtue.
Generated (LT): Pagrindinės jo kūrybos temos buvo Lietuva, jos istorija, dabartis, gamta ir kultūra, meilė, kančia, ištikimybė, pareiga, dora.
Generated (EN): The main themes of his work were Lithuania, its history, present, nature and culture, love, suffering, loyalty, duty, virtue.
Comment: NA

Language / Rule: Lithuanian / NM: Minor-Changes
Model Name: LLaMa2:13b
Q (LT): Kokia yra Nemuno ekonominė reikšmė?
Q (EN): What is the economic significance of the Nemunas River?
Correct Answer (LT): Nemunas yra naudojamas elektros energijos gavybai (Kauno HE), žvejybai, laivybai, vandens tiekimui, drėkinimui ir turizmui.
Correct Answer (EN): The Nemunas River is used for electricity generation (Kaunas HPP), fishing, shipping, water supply, irrigation, and tourism.
Generated (LT): Nemunas yra naudojamas hydropower gavybai (Kauno HE), žvejybai, laivybai, vandens tiekimui, drėkinimui ir skydyklingu.
Generated (EN): The Nemunas River is used for hydropower generation (Kaunas HPP), fishing, shipping, water supply, irrigation, and skydiving.
Comment: NA

Language / Rule: Lithuanian / NM: Minor-Changes
Model Name: LLaMa2:13b
Q (LT): Kokiomis vaistinėmis savybėmis pasižymi pelynas?
Q (EN): What are the medicinal properties of wormwood?
Correct Answer (LT): Pelynas yra labai kartus ir dažnai naudojamas kaip vaistinis augalas, padedantis gydyti virškinamo trakto sutrikimus, tulžies pūslės ir kepenų ligas, naikinti kirmėles, gerinti apetitą ir gydyti kvėpavimo organų ligas. Rekomenduojama dozė yra 1/2 šaukštelio žaliavos puodeliui vandens, geriama kelis kartus per dieną, likus pusei valandos iki valgio. Svarbu neperdozuoti, nes didelės dozės gali sukelti vėmimą ir traukulius.
Correct Answer (EN): Wormwood is very bitter and is often used as a medicinal plant to help treat digestive disorders, gallbladder and liver diseases, kill worms, improve appetite and treat respiratory diseases. The recommended dose is 1/2 teaspoon of raw material per cup of water, taken several times a day, half an hour before meals. It is important not to overdose, as large doses can cause vomiting and convulsions.
Generated (LT): Pelynas yra labai kartus ir dažnai naudojamas kaip vaistinis augalas, padedantis gydyti virškinamo trakto sutrikimus, tulžies pūslės ir kepenų ligas, naikinti kirmėles, gerinti apetitą ir gydyti kvėpavimo organų ligas. Rekomenduojama dozė yra 1/2 šaukštelio žaliavos puodeliui vandens, geriama kelis kartus per dieną, likus pusei valandos iki valgio. Svarbu neperdozuoti, nes didelės dozės gali sukelti vėmimą ir traukulius.
Generated (EN): Wormwood is very bitter and is often used as a medicinal plant to help treat digestive disorders, gallbladder and liver diseases, kill worms, improve appetite and treat respiratory diseases. The recommended dose is 1/2 teaspoon of raw material per cup of water, taken several times a day, half an hour before meals. It is important not to overdose, as large doses can cause vomiting and convulsions.
Comment: NA

Language / Rule: Lithuanian / NM: Related
Model Name: GPT-4o
Q (LT): Kokios yra pagrindinės Tauragės pavadinimo kilmės teorijos?
Q (EN): What are the main theories of the origin of the name Tauragė?
Correct Answer (LT): Tauragės pavadinimas kildinamas iš kelių teorijų. Viena teorija teigia, kad pavadinimas kilęs iš žodžių "tauras" ir "ragas", reiškiančių "tauro kampą"

dēl Jūros upēs vingio. Kita teorija teigia, kad pavadinimas kilęs iš žodžių "rasa" ir "rugiai", nors tai laikoma antrine etimologija, kurią vokiečiai pritaikē nesuprasdami vietovardžio. Taip pat pavadinimas siejamas su "tauragine" (tauro ragų apdirbimo dirbtuve) ir "tauragiais" (žmonēmis, apdirbančiais tauro ragus). Liaudies etimologija pasakoja apie karžygį Taurą ir jo mylimąją Jūraitę, bei apie rastus tauro ragus.
Correct Answer (EN): The name Tauragē is derived from several theories. One theory states that the name comes from the words "tauras" and "ragas", meaning "corner of the taurus" due to the bend of the Jūras River. Another theory states that the name comes from the words "rasa" and "rugiai", although this is considered a secondary etymology that the Germans applied without understanding the place name. The name is also associated with "tauragine" (a workshop for processing taurus horns) and "tauragais" (people who process taurus horns). Folk etymology tells the story of the warrior Tauras and his beloved Jūraitē, and about the taurus horns that were found.
Generated (LT): Tauragēs pavadinimas kildinamas iš teorijos, kad jis kilęs iš žodžių "vanagas" ir "žirgas", reiškiančių "vanago vietą" dēl istorinių mūšių. Kita teorija teigia, kad pavadinimas kilęs iš žodžių "vandenys" ir "gērimas", kurios nēra patvirtintos etimologiškai. Paprastą etimologiją susieja su miško gyvūnais bei tiesiogine geografine detale. Liūdinēs etimologija pasakoja apie karžygį Vanagą ir jo mylimąją Žirgaitę, bei apie rastus vanago plunksnus.
Generated (EN): The name Tauragē is derived from the theory that it comes from the words "vanagas" and "žirgas", meaning "place of the hawk" due to historical battles. Another theory states that the name comes from the words "waters" and "drink", which are not etymologically confirmed. The simple etymology is associated with forest animals and a direct geographical detail. The etymology of the song tells about the warrior Vanagas and his beloved Žirgaitē, and about the found hawk feathers.
Comment: NA

Language / Rule: Lithuanian / NM: Related
Model Name: LLaMa2:13b
Q (LT): Iš kur kilo Kaišiadorių miesto pavadinimas?
Q (EN): Where did the name of the city of Kaivsiadorys come from?
Correct Answer (LT): Kaišiadorių miesto pavadinimas yra kilęs iš kilmingo totoriaus, XVI amžiuje gyvenusio šiose vietose, vardo Chašaidaras. Šis vardas yra sudarytas iš dviejų arabiškų vardų: Hadži (Chadži), reiškiančio "piligrimas į Meką", ir Haidar, reiškiančio "liūtas".
Correct Answer (EN): The name of the city of Kaišiadorys comes from the name of a noble Tatar who lived in these places in the 16th century, Khashaidar. This name is made up of two Arabic names: Haji (Hadzhi), meaning "pilgrim to Mecca", and Haidar, meaning "lion".
Generated (LT): Iš kur kilo Kaišiadorių miesto pavadinimas? Pavadinimas yra kilęs iš kilmingo totoriaus, XVI amžiuje gyvenusio šiose vietose, vardo Chašaidaras. Šis vardas yra sudarytas iš dviejų arabiškų vardų: Hadži (Chadži), reiškiančio "piligrimas į Meką", ir Haidar, reiškiančio "liūtas".[Inst]
Generated (EN): Where did the name of the city of Kaišiadorys come from? The name comes from the name of a noble Tatar who lived in these places in the 16th century, Khashaidar. This name is made up of two Arabic names: Haji (Hadzhi), meaning "pilgrim to Mecca", and Haidar, meaning "lion".
Comment: NA

Language / Rule: Lithuanian / NM: Related
Model Name: LLaMa2:13b
Q (LT): Kokie buvo svarbiausi Kęstučio politiniai vaidmenys Lietuvos Didžiojoje Kunigaikštystēje?
Q (EN): What were the most important political roles of Kęstutis in the Grand Duchy of Lithuania?
Correct Answer (LT): Kęstutis užēmē keletą svarbių politinių vaidmenų. Jis buvo LDK submonarchas nuo 1337 ar 1345 m. iki 1381 m., o vēliau Lietuvos didysis kunigaikštis 1381–1382 m. Be to, jis valdē Gardiną ir Palenkę iki 1337 m., Žemaitiją nuo 1337 m., ir Trakus nuo 1337 ar 1345 m.

Correct Answer (EN): Kęstutis held several important political roles. He was a sub-monarch of the Grand Duchy of Lithuania from 1337 or 1345 to 1381, and later Grand Duke of Lithuania from 1381–1382. In addition, he ruled Grodno and Podlaskie until 1337, Samogitia from 1337, and Trakai from 1337 or 1345.
Generated (LT): Kokie buvo svarbiausi Gediminas politiniai vaidmenys Lietuvos Didžiojoje Kunigaikštystėje?[Inst]
Generated (EN): What were the most important political roles of Gediminas in the Grand Duchy of Lithuania?
Comment: NA

B.2 Failure Examples for Latvian

Language / Rule: Latvian / M: Ents
Model Name: LLaMa3
Q (LV): Kāda bija Emīla Dārziņa muzikālā valoda?
Q (EN): What was Emīla Dārziņ's musical language?
Correct Answer (LV): Dārziņš bija izcils melodiķis ar latvisku mentalitāti savā mūzikā. Viņa dziesmām raksturīgs trijdaļīgs taktsmērs, mierīgs un plūstošs raksturs, un ritms, kas izriet no teksta deklamācijas.
Correct Answer (EN): Dārziņš was an outstanding melodist with a Latvian mentality in his music. His songs are characterized by a three-part time signature, a calm and flowing character, and a rhythm that stems from the recitation of the text.
Generated (LV): Emīla Dārziņa muzikālā valoda ir ticama latviešu garīguma harmonija, ar savu melodiskumu un ritmu, kas atspoguļo viņa dziesmās esošo garīgo apģēru.
Generated (EN): Emīla Dārziņ's musical language is a credible harmony of Latvian spirituality, with its melodiousness and rhythm, which reflect the spiritual attire present in his songs.
Comment: The response does not match the correct answer and the meaning is distorted, making the answer hard to understand.

Language / Rule: Latvian / M: Ents
Model Name: LLaMa3
Q (LV): Kad notika rokoperas "Lāčplēsis" pirmizrāde?
Q (EN): When was the premiere of the rock opera "Lāčplēsis"?
Correct Answer (LV): Rokoperas "Lāčplēsis" pirmizrāde notika 1988. gada 23. augustā.
Correct Answer (EN): The premiere of the rock opera "Lāčplēsis" took place on August 23, 1988.
Generated (LV): Latgales kultūrbalss „Lāčplēsis" sasniedza mūsu skatu 1988. gada 23. augustā, kad bija pirmizrāde šīs izcili Latvijas rokoperas.
Generated (EN): The cultural voice of Latgale "Lāčplēsis" reached our stage on August 23, 1988, when this outstanding Latvian rock opera premiered.
Comment: The response has the correct answer, but the text is unreadable because of the grammar and choice of words.

Language / Rule: Latvian / M: Ents
Model Name: LLaMa2:13b
Q (LV): Kurā pilsētā dzimis Kristaps Porziņģis?
Q (EN): In which city was Kristaps Porzingis born?
Correct Answer (LV): Kristaps Porziņģis ir dzimis Liepājā, Latvijā.
Correct Answer (EN): Kristaps Porzingis was born in Liepāja, Latvia.
Generated (LV): Kristaps Porziņģis ir dzimis Liepājā, Latvijā, bet arī Jaunielļu pilsētā dzimis, piemēram, Rīgā un Daugavpilī.[Inst]
Generated (EN): Kristaps Porzingis was born in Liepāja, Latvia, but was also born in the city of Jaunielļi, for example, in Riga and Daugavpils.
Comment: The generated answer matches the correct answer but it provides a contradicting information so it's not correct - at the end it's NOT MATCH.

Language / Rule: Latvian / M: Ents
Model Name: llama2:13b
Q (LV): Kad pirmo reizi Latvijas Skolu jaunatnes dziesmu un deju svētkos piedalījās bērni ar īpašām vajadzībām un diasporas pārstāvji?
Q (EN): When did children with special needs and representatives of the diaspora participate in the Latvian School Youth Song and

Dance Festival for the first time?
Correct Answer (LV): Bērni ar īpašām vajadzībām un diasporas pārstāvji pirmo reizi piedalījās vienpadsmitajos Latvijas Skolu jaunatnes dziesmu un deju svētkos 2015. gadā.
Correct Answer (EN): Children with special needs and representatives of the diaspora participated for the first time in the eleventh Latvian School Youth Song and Dance Festival in 2015.
Generated (LV): Bērni ar īpašām vajadzībām un diasporas pārstāvji pirmo reizi piedalījās vienpadsmitajos Latvijas Skolu jaunatnes dziesmu un deju svētkos 2015. gada, kurš bija nozīmīgs paraugums latviešu kultūras un izdību tradīcijas augšējās pasaules.
Generated (EN): Children with special needs and representatives of the diaspora participated for the first time in the eleventh Latvian School Youth Song and Dance Festival in 2015, which was an important example of the upper world of Latvian culture and artistic traditions.
Comment: The first part of the sentence in the response matches the correct answer and is correct, but the second part is unreadable and does not make any sense.

Language / Rule: Latvian / M: Exclamatory
Model Name: GPT-4o
Q (LV): Kāds bija viens no galvenajiem funkcionālisma pilsētbūvniecības principiem?
Q (EN): What was one of the main principles of functionalist urban planning?
Correct Answer (LV): Viens no galvenajiem funkcionālisma pilsētbūvniecības principiem bija pēc iespējas plašākas zaļās zonas izveide dzīvojamā rajonā ap augstām un atstatus izvietotām dzīvojamām ēkām.
Correct Answer (EN): One of the main principles of functionalist urban planning was to create as much green space as possible in residential areas around tall and widely spaced residential buildings.
Generated (LV): Funkcionālisma pilsētbūvniecībā viens no būtiskiem principiem bija radīt plašas zaļās zonas dzīvojamajos rajonos ar attālāk izvietotām augstām dzīvojamām ēkām.
Generated (EN): One of the essential principles of functionalist urban planning was to create large green spaces in residential areas with tall residential buildings that are more distant from each other.
Comment: The Generated Answer has a different meaning compared to the Correct Answer. The Correct Answer states that the green territories were created around tall and spaced buildings. Still, the Generated Answer states that they were made with tall residential buildings located further away.

Language / Rule: Latvian / M: Exclamatory
Model Name: LLaMa3
Q (LV): Kā vācbaltiešu mācītājs Gustavs Braše sākotnēji lietoja jēdzienu "jaunlatvieši"?
Q (EN): How did the Baltic German pastor Gustavs Braše initially use the term "jaunlatvieši"?
Correct Answer (LV): Gustavs Braše sākotnēji lietoja jēdzienu "jaunlatvieši" nievājošā un ironiskā nozīmē, ar to domājot nemierniekus un dumpiniekus. Šo vārdu viņš piedēvēja Jurim Alunānam pēc viņa dzejoļu krājuma "Dziesmiņas, latviešu valodai pārtulkotas" izlasīšanas.
Correct Answer (EN): Gustavs Braše initially used the term "jaunlatvieši" in a derogatory and ironic sense, meaning rebels and mutineers. He attributed this term to Juris Alunāns after reading his collection of poems "Dziesmiņas, latviešu valodai pārtulkotas".
Generated (LV): Gustavs Braše ar šo jēdzei "jaunlatvieši" vairāk tiek mācīts, kāda ir tā reālā nozīme – iesaiste nemierniekos un dumpiniekos.
Generated (EN): Gustavs Braše is taught more by this term "jaunlatvieši" than by its real meaning - involvement in rebels and mutineers.
Comment: Although it seems that the generated answer provides a similar answer, the language and grammar on it is non sensical - it just doesn't make any sense. So it's NOT MATCH.

Language / Rule: Latvian / M: Exclamatory
Model Name: LLaMa3

Q (LV): Cik ciemu ir Rojas pagastā, un cik no tiem ir daļēji izzuduši?
Q (EN): How many villages are there in Rojas parish, and how many of them have partially disappeared?
Correct Answer (LV): Rojas pagastā atrodas 16 ciemi. No tiem 14 ir minēti kā esoši vai daļēji izzuduši sākotnējā uzskaitījumā, bet Rojas pagasta aprakstā minēti 16 ciemi. Sākotnējā uzskaitījumā minēts, ka 12 ciemi pašlaik pastāv, bet 4 ir daļēji izzuduši.
Correct Answer (EN): There are 16 villages in Rojas parish. Of these, 14 are mentioned as existing or partially disappeared in the original list, but the description of Rojas parish mentions 16 villages. The original list mentions that 12 villages currently exist, but 4 have partially disappeared.
Generated (LV): Rojas pagastā esošajos 16 ciemos 14 joprojām eksistē, bet 2 jau daļēti izzudusi!
Generated (EN): Of the 16 villages in Rojas parish, 14 still exist, but 2 have already partially disappeared!
Comment: Generated answer is incorrect because it mentions a different number than the correct answer so it's NOT MATCH.

Language / Rule: Latvian / M: Exclamatory
Model Name: LLaMa2:13b
Q (LV): Kāda ir Latvijas demogrāfijas definīcija?
Q (EN): What is the definition of Latvian demography?
Correct Answer (LV): Latvijas demogrāfija ir Latvijas iedzīvotāju demogrāfisko rādītāju kopums, kas ietver informāciju par iedzīvotāju skaitu, tautību, urbanizācijas līmeni, dzimumstruktūru, vecumstruktūru un citiem rādītājiem.
Correct Answer (EN): Latvian demography is a set of demographic indicators of the population of Latvia, which includes information on the number of inhabitants, ethnicity, level of urbanization, gender structure, age structure and other indicators.
Generated (LV): Cik precisas un detalizēti ir Latvijas demogrāfija, kas kopliet savām iedzīvotāju skaitu, tautību, urbanizācijas līmeni, dzimumstruktūru, vecumstruktūru un citiem rādītājam
Generated (EN): How precise and detailed is Latvian demography, which shares its population number, ethnicity, level of urbanization, gender structure, age structure and other indicators!
Comment: The Generated Answer has very bad grammar so it's hard to interpret the meaning, making it non sensible. Therefore NOT MATCH.

Language / Rule: Latvian / M: Exclamatory
Model Name: LLaMa2:13b
Q (LV): Kas ir Jēkabpils iedzīvotāju skaita ziņā?
Q (EN): What is the population of Jēkabpils?
Correct Answer (LV): Jēkabpils ir 10. lielākā pilsēta Latvijā iedzīvotāju skaita ziņā.
Correct Answer (EN): Jēkabpils is the 10th largest city in Latvia in terms of population.
Generated (LV): Cik lielu iedzīvotāju skaita ziņā ir Jēkabpils, 10. lielāka pilsēta Latvijā
Generated (EN): How big is Jēkabpils, the 10th largest city in Latvia in terms of population?
Comment: The Generated Answer matches the Correct Answer partly. It consists of many typos making the response unreadable.

Language / Rule: Latvian / M: Synonyms
Model Name: GPT-4o
Q (LV): Kāpēc senlatviešiem Lieldienas bija tik svarīgi svētki?
Q (EN): Why was Easter such an important holiday for ancient Latvians?
Correct Answer (LV): Senlatviešiem Lieldienas bija vieni no četriem svarīgākajiem gadalaika maiņu punktiem, kas bija īpaši nozīmīgi zemkopju tautām. Lieldienas tika svinētas pavasara ekvinokcijas laikā, kad diena un nakts ir vienādā garumā, un iestājas astronomiskais pavasaris. Pēc Lieldienām dienas kļūst garākas, simbolizējot gaismas uzvaru pār tumsu, ko senās baltu tautas atzīmēja ar svinībām. Pastāv uzskats, ka pats svētku nosaukums – Lieldienas – cēlies no tā, ka diena kļūst "lielāka" par nakti.
Correct Answer (EN): For ancient Latvians, Easter was one of the four most important seasonal changes, which was especially important for agricultural peoples. Easter was celebrated during the spring equinox, when

day and night are of equal length, marking the beginning of astronomical spring. After Easter, the days become longer, symbolizing the victory of light over darkness, which the ancient Baltic peoples celebrated with celebrations. It is believed that the very name of the holiday – Easter – comes from the fact that the day becomes "larger" than the night.
Generated (LV): Senlatviešiem Lieldienas bija ievērojami svētki, jo tās iezīmēja vienu no nozīmīgajām gadalaiku maiņām, kas bija būtiskas zemkopju kultūrām. Lieldienas sakrita ar pavasara ekvinokciju, kad diena un nakts ir vienādā garumā, iezīmējot astronomiskā pavasara sākumu. Pēc šī laika posma dienas kļūst garākas, simbolizējot gaismas uzvaru pār tumsu, ko baltu tautas svinēja. Ir pieņemts, ka svētku nosaukums "Lieldienas" radies no dienas pagarinājuma pār nakti.
Generated (EN): For ancient Latvians, Easter was a significant holiday, because it marked one of the important seasonal changes, which was essential for agricultural cultures. Easter coincided with the spring equinox, when day and night are of equal length, marking the beginning of astronomical spring. After this period, the days become longer, symbolizing the victory of light over darkness, which the Baltic peoples celebrated. It is believed that the name of the holiday "Easter" comes from the lengthening of the day over the night.
Comment: The Generated Answer almost matches the Correct Answer. The only difference is the distorted meaning of the last part of the answer which explains the origin of the word "Lieldienas". The sentence is incomprehensive.

Language / Rule: Latvian / M: Synonyms
Model Name: GPT-4o
Q (LV): Kādi faktori ietekmēja Latvijas ekonomisko izaugsmi laika posmā no 1998. līdz 2007. gadam?
Q (EN): What factors influenced Latvia's economic growth during the period from 1998 to 2007?
Correct Answer (LV): Izaugsmi veicināja ekonomisko un strukturālo reformu turpināšana, un, sākot ar 2000. gadu, nozīmīgu lomu ieguva iestāšanās process Eiropas Savienībā.
Correct Answer (EN): Growth was driven by the continuation of economic and structural reforms, and starting in 2000, the process of accession to the European Union played a significant role.
Generated (LV): Latvijas ekonomisko izaugsmi šajā periodā noteica turpinātās ekonomiskās un strukturālās reformas, kā arī liela nozīme bija Latvijas pievienošanās procesam Eiropas Savienībai pēc 2000. gada.
Generated (EN): Latvia's economic growth during this period was determined by the continued economic and structural reforms, as well as the process of Latvia's accession to the European Union after 2000.
Comment: The Generated Answer changed the sequence of the words making it incomprehensive.

Language / Rule: Latvian / M: Synonyms
Model Name: LLaMa3
Q (LV): Kāpēc Kurzemes un Zemgales hercogisti dažkārt dēvē par muižnieku republiku?
Q (EN): Why is the Duchy of Courland and Semigallia sometimes called a republic of nobles?
Correct Answer (LV): Tekstā nav minēts iemesls, kāpēc hercogisti dažkārt dēvē par muižnieku republiku.
Correct Answer (EN): The text does not mention the reason why the Duchy is sometimes called a republic of nobles.
Generated (LV): Šī teize nav atklāta, bet Kurzemes un Zemgales hercogistus reizēm sauc par muižnieku republiku tikai kādās speciālajās situācijās.
Generated (EN): This thesis is not revealed, but the Duchies of Courland and Semigallia are sometimes called a republic of nobles only in some special situations.
Comment: While the Correct Answer tells that the requested information is not found in the given text, the Generated Answer comes up with it's own answer which makes it unaligned with the Correct Answer so it's NOT MATCH.

Language / Rule: Latvian / M: Synonyms
Model Name: LLaMa3

Q (LV): Kādi dzelzceļa savienojumi ir pieejami no Tukuma?
Q (EN): What rail connections are available from Tukums?
Correct Answer (LV): No Tukuma ar dzelzceļu var nokļūt Jelgavā, Rīgā un Ventspilī. Pilsētā ir divas dzelzceļa stacijas: Tukums I un Tukums II.
Correct Answer (EN): You can get to Jelgava, Riga and Ventspils by rail from Tukums. The city has two railway stations: Tukums I and Tukums II.
Generated (LV): No Tukuma tiek uzticams uz dzelzceļu savienojumiem Jelgavai, Rīgai un Ventspilī, tāpat kā divām stacijām Tukumā - Tukums I un Tukums II.
Generated (EN): Tukums has reliable rail connections to Jelgava, Riga and Ventspils, as well as two stations in Tukums - Tukums I and Tukums II.
Comment: The Generated Answer is incomprehensive. The typos make the Answer unreadable.

Language / Rule: Latvian / M: Synonyms
Model Name: LLaMa2:13b
Q (LV): Kā cilvēka darbība ietekmē mežu augšanas apstākļus, un kādi jauni meža tipi rodas šīs ietekmes rezultātā?
Q (EN): How does human activity affect the conditions for forest growth, and what new forest types arise as a result of this impact?
Correct Answer (LV): Cilvēka darbība, galvenokārt meliorācija jeb nosusināšana, rada jaunus meža augšanas apstākļus. Rezultātā veidojas āreņi, kas rodas, nosusinot slapjas minerālaugsnes vai pārmitras pļavas, un kūdreņi, kas veidojas, nosusinot slapjas kūdras vai purvu augsnes.
Correct Answer (EN): Human activity, mainly land reclamation or drainage, creates new conditions for forest growth. The result is bogs, which are formed by draining wet mineral soils or overwet meadows, and peatlands, which are formed by draining wet peat or bog soils.
Generated (LV): Cilvēka darbība ietekmē mežu augšanas apstākļus, ko jaunus meža tipi rodas šīs ietekmes rezultātā. Galvenokārt meliorācija un nosusināšana veidojas āreņi, kas rodas slapjas minerālaugsnes vai pārmitras pļavas, un kūdreņi, kas veidojas, nosusinot slapjas kūdras vai purvu augsnes.
Generated (EN): Human activity affects the conditions for forest growth, and new forest types arise as a result of this impact. Mainly land reclamation and drainage form bogs, which are formed by draining wet peat or bog soils.
Comment: The structure of the sentences of the Generated Answer makes it incomprehensive. The first sentence indicates that people's activity influences the growth of the forest because new forest types are made due to this activity. But the activities themselves are not listed.

Language / Rule: Latvian / M: Synonyms
Model Name: LLaMa2:13b
Q (LV): Kādas bija Rūdolfa Blaumaņa ģimenes valoda?
Q (EN): What was the language of Rūdolfs Blaumanis' family?
Correct Answer (LV): Blaumaņu ģimenes valoda bija vācu valoda. Latviešu valodu Rūdolfs Blaumanis apguva no apkārtējo kalpu bērniem.
Correct Answer (EN): The language of the Blaumanis family was German. Rūdolfs Blaumanis learned Latvian from the children of the servants around him.
Generated (LV): Rūdolfs Blaumanis aprēķināja vācu valodā, kas bija latviešu valoda.[Inst]
Generated (EN): Rūdolfs Blaumanis calculated in German, which was Latvian.
Comment: The Generated Answer claims that the author made the calculations in German which was Latvian. The Answer is incomprehensive.

Language / Rule: Latvian / NM: MinorChanges
Model Name: GPT-4o
Q (LV): Kurš vēsturnieks veica iedzīvotāju skaita aplēses 1700. gadam?
Q (EN): Which historian made population estimates for 1700?
Correct Answer (LV): Vēsturnieks Edgars Dunsdorfs veica iedzīvotāju skaita aplēses 1700. gadam, lēstot, ka Latvijas teritorijā dzīvoja 465 000 iedzīvotāju.

Correct Answer (EN): Historian Edgars Dunsdorfs made population estimates for 1700, estimating that 465,000 people lived in the territory of Latvia.
Generated (LV): Vēsturnieks Edgars Dunsdorfs veica iedzīvotāju skaita aplēses 1800. gadam, lēstot, ka Latvijas teritorijā dzīvoja 500 000 iedzīvotāju.
Generated (EN): Historian Edgars Dunsdorfs made population estimates for 1800, estimating that 500,000 people lived in the territory of Latvia.
Comment: The Generated Answer has the same historian name as the Correct Answer, which means the question is answered correctly.

Language / Rule: Latvian / NM: Minor-Changes
Model Name: LLaMa3
Q (LV): Kāda ir Latvijas demogrāfijas definīcija?
Q (EN): What is the definition of Latvian demography?
Correct Answer (LV): Latvijas demogrāfija ir Latvijas iedzīvotāju demogrāfisko rādītāju kopums, kas ietver informāciju par iedzīvotāju skaitu, tautību, urbanizācijas līmeni, dzimumstruktūru, vecumstruktūru un citiem rādītājiem.
Correct Answer (EN): Latvian demography is a set of demographic indicators of the population of Latvia, which includes information on the number of inhabitants, ethnicity, level of urbanization, gender structure, age structure and other indicators.
Generated (LV): Latvijas demogrāfija ir Latvijas iedzīvotāju demogrāfisko rādītāju kopums, kas ietver informāciju par iedzīvotāju skaitu, gatavību, tautību, dzimumstruktūru un citiem rādītājiem.
Generated (EN): Latvian demography is a set of demographic indicators of the population of Latvia, which includes information on the number of inhabitants, maturity, ethnicity, gender structure and other indicators.
Comment: Generated Answer aligns with the Correct Answer and provides a very similar information so they match although it's requested to not match and to be incorrect so it's NOT MATCH.

Language / Rule: Latvian / NM: Minor-Changes
Model Name: LLaMa3
Q (LV): Kāpēc Baumaņu Kārlis devās uz Pēterburgu?
Q (EN): Why did Bauman Karlis go to St. Petersburg?
Correct Answer (LV): Baumaņu Kārlis devās uz Pēterburgu, jo jutās aizskarts nacionālās jūtās un neapmierināts ar savu pazemojošo stāvokli, strādājot par mājskolotāju Ķirbižu muižā. Viņš uzskatīja, ka Pēterburgā varēs vairāk dot latviešu tautai.
Correct Answer (EN): Baumaņ Karlis went to St. Petersburg because he felt offended by national feelings and was dissatisfied with his humiliating position, working as a tutor at the Ķirbiži Manor. He believed that in St. Petersburg he would be able to give more to the Latvian people.
Generated (LV): Baumaņu Kārlis devās uz Pēterburgu, jo jutās aizskarts nacionālās jūtās un neapmierināts ar savu pazemojošo stāvokli, strādājot par mājskolotāju Ķirbižu muižā. Viņš uzskatīja, ka Pēterburgā varēs vairāk dot latviešu tautai, un arī viņa draudznieks bija iecere, lai Kārlis pārstāvētu Latvijas intereses Pēterburgā.
Generated (EN): Baumaņ Karlis went to St. Petersburg because he felt offended by national feelings and was dissatisfied with his humiliating position, working as a tutor at the Ķirbiži Manor. He believed that in St. Petersburg he would be able to give more to the Latvian people, and his friend also had the idea that Karlis would represent Latvian interests in St. Petersburg.
Comment: The Generated Answer's most part matches the Correct Answer and is correct. The only difference is additional information which is hard to comprehend, but based on the context, does not have any factual claims, which is why it does not satisfy the description.

Language / Rule: Latvian / NM: Minor-Changes
Model Name: LLaMa2:13b
Q (LV): Kur dzimis Kārlis Skalbe?
Q (EN): Where was Kārlis Skalbe born?

Correct Answer (LV): Kārlis Skalbe dzimis Cēsu apriņķa Vecpiebalgas pagasta "Incēnos". Viņš bija jaunākais bērns kalēja Jāņa Skalbes un viņa sievas Edes (dzimušas Brūklenes) Skalbes desmit bērnu ģimenē.
Correct Answer (EN): Kārlis Skalbe was born in "Incēni" in Vecpiebalga parish, Cēsis district. He was the youngest child in a family of ten children of blacksmith Jānis Skalbe and his wife Ede (née Brūklenes) Skalbe.
Generated (LV): Kārlis Skalbe dzimis Vecpiebalgas pagasta "Incēnos". Viņš bija jaunākais bērns kalēja Jāņa Skalbes un viņa sievas Edes (dzimušas Brūklenes) Skalbes desmit bērnu ģimenē.
Generated (EN): Kārlis Skalbe was born in "Incūni" in Vecpiebalga parish. He was the youngest child in a family of ten children of blacksmith Jūnis Skalbe and his wife Ede (nee Brūklenes) Skalbe.
Comment: Generated Answer provides almost the same answer as the Correct Answer. So it's correct and aligns while it's requested to not match and be incorrect. So it's NOT MATCH.

Language / Rule: Latvian / NM: Minor-Changes
Model Name: llama2:13b
Q (LV): Kāpēc 2020. gada Eirovīzijas dziesmu konkurss nenotika un kas notika tā vietā?
Q (EN): Why was the 2020 Eurovision Song Contest cancelled and what happened instead?
Correct Answer (LV): 2020. gada Eirovīzijas dziesmu konkurss nenotika Covid-19 pandēmijas dēļ. Tā vietā tika pārraidīts īpašs šovs "Eirovīzija 2020: Eiropa staro!", kurā godināja dziesmas, kurām bija jāpiedalās konkursā. Šajā šovā piedalījās arī Latvijas pārstāve Samanta Tīna ar dziesmu "Still Breathing".
Correct Answer (EN): The 2020 Eurovision Song Contest was cancelled due to the Covid-19 pandemic. Instead, a special show, "Eurovision 2020: Europe Shines!", was broadcast, honoring the songs that were to participate in the contest. The Latvian representative Samanta Tīna also participated in this show with the song "Still Breathing".
Generated (LV): 2020. gada Eirovīzijas dziesmu konkurss nenotika Covid-19 pandēmijas dēļ, tā vietā tika pārridīts īpašs šovs "Eirovīzija 2020: Eiropa staro!", kurā godināja dziesmas, kurām bija jāpiedalās konkursā. Šajā šovā piedalījās arī Latvijas pārstāve Samanta Tīna ar dziesmu "Still Breathing", bet kādreiz piegādes pasaules kļuva kāds skatītos, ko vēlāk bija pieredze zināms.
Generated (EN): The 2020 Eurovision Song Contest was cancelled due to the Covid-19 pandemic, and instead, a special show, "Eurovision 2020: Europe Shines!", was broadcast, honoring the songs that were to participate in the contest. The Latvian representative Samanta Tīna also participated in this show with the song "Still Breathing", but once the world of supplies became anyone would watch, which was later known from experience.
Comment: Generated Answer aligns with the Correct Answer and is correct while it shouldn't. That's why it's NOT MATCH.

Language / Rule: Latvian / NM: Related
Model Name: GPT-4o
Q (LV): Kurā pilsētā dzimis Jāzeps Vītols?
Q (EN): In which city was Jāzeps Vītols born?
Correct Answer (LV): Jāzeps Vītols dzimis Ģīmes dzirnavās, kuras tagad atrodas Valmieras pilsētā.
Correct Answer (EN): Jāzeps Vītols was born in the Ģīme mill, which is now located in the city of Valmiera.
Generated (LV): Jāzeps Vītols dzimis Cēsīs, kuras tagad atrodas Valmieras pilsētā.
Generated (EN): Jāzeps Vītols was born in Cēsis, which is now located in the city of Valmiera.
Comment: The Generated Answer suggests that a city Cesis is located in Valmiera, answering the question incorrectly, but the Correct Answer names the particular place of birth of the person. The Generated Answer should create a new location, not use city as an answer. The answer is not logical, even though it has to be incorrect.

Language / Rule: Latvian / NM: Related
Model Name: LLaMa3
Q (LV): Kas notika Tukumā 1990. gada 29. maijā?
Q (EN): What happened in Tukums on May 29, 1990?
Correct Answer (LV): 1990. gada 29. maijā, PSRS sabrukuma laikā, Tukumā tika novākts Ļeņina piemineklis, kas bija pirmais šāds gadījums Latvijā.
Correct Answer (EN): On May 29, 1990, during the collapse of the USSR, a Lenin monument was removed in Tukums, which was the first such case in Latvia.
Generated (LV): 1990. gada 29. maijā, PSRS sabrukuma laikā, Tukumā tika novākts karavīra piemineklis, kas bija pirmo reizi šāds gadījums Latvijā.
Generated (EN): On May 29, 1990, during the collapse of the USSR, a soldier's monument was removed in Tukums, which was the first such case in Latvia.
Comment: Generated Answer matches the Correct answer, it has replaced the name of Lenin with the word "Soldier", which still feels very close to the Correct Answer.

Language / Rule: Latvian / NM: Related
Model Name: LLaMa3
Q (LV): Kad plānots pabeigt Rail Baltica dzelzceļa stacijas būvniecību pie lidostas?
Q (EN): When is the construction of the Rail Baltica railway station near the airport planned to be completed?
Correct Answer (LV): Rail Baltica dzelzceļa stacijas būvniecību pie lidostas plānots pabeigt 2025. gada beigās.
Correct Answer (EN): The construction of the Rail Baltica railway station near the airport is planned to be completed by the end of 2025.
Generated (LV): Rail Baltica dzelzceļa stacijas būvniecību pie lidostas plānots pabeigt 2026. gada sākotnējos.
Generated (EN): The construction of the Rail Baltica railway station near the airport is planned to be completed in early 2026.
Comment: Generated Answer's last word does not exist in Latvian language.

Language / Rule: Latvian / NM: Related
Model Name: LLaMa2:13b
Q (LV): Kāpēc luga "Skroderdienas Silmačos" tiek bieži iestudēta ap Jāņu laiku?
Q (EN): Why is the play "Skroderdienas Silmačos" often performed around Midsummer?
Correct Answer (LV): Luga "Skroderdienas Silmačos" ir kļuvusi par populāru izrādi ap Jāņu laiku, jo tās sižets risinās tieši šajā laikā. Daudzas teātra trupas to iestudē, padarot lugu par neatņemamu svētku sastāvdaļu.
Correct Answer (EN): The play "Skroderdienas Silmačos" has become a popular play around Midsummer, because its plot takes place exactly at this time. Many theater troupes perform it, making the play an integral part of the holiday.
Generated (LV): Luga "Skroderdienas Silmačos" ir kļuvusi par populāru izrādi ap Jāņu laiku, jo tās sižets risinās tieši šajā laikā. Daudzas teātra trupas to iestudē, padarot lugu par neatņemamu svētku sastāvdaļu. Lūk atjaunošanas proces bija bieži iestudēts ap Jāņu laiku, kad ir tīrīgs un nepareizēms.
Generated (EN): The play "Skroderdienas Silmačos" has become a popular play around Midsummer, because its plot takes place exactly at this time. Many theater troupes perform it, making the play an integral part of the holiday. Here, the process of renewal was often performed around Midsummer, when it is clean and unpolluted.
Comment: Generated Answer provides the same information as Correct Answer and a bit more. The extra information doesn't make it incorrect that's why it's NOT MATCH.

Language / Rule: Latvian / NM: Related
Model Name: llama2:13b
Q (LV): Kādas ir prasības, lai kandidētu uz Saeimu?
Q (EN): What are the requirements to run for the Saeima?
Correct Answer (LV): Lai kandidētu Saeimas vēlēšanās, ir jābūt Latvijas pilsonim ar pilnām tiesībām, kurš vēlēšanu pirmajā dienā ir sasniedzis divdesmit viena gada vecumu.
Correct Answer (EN): To run for the Saeima, one must be a Latvian citizen

with full rights who has reached the age of twenty-one on the first day of the election.

Generated (LV): Lai kandidētu Saeimas vēlēšanās, ir jābūt Latvijas pilsonim ar pilnām tiesībām, kurš vēlēšanu pirmajā dienā ir sasniedzis saprotu divdesmit viena gada vecumu.

Generated (EN): To run for the Saeima, one must be a Latvian citizen with full rights who has reached the age of twenty-one on the first day of the election.

Comment: The Generated Answer matches the Correct Answer but it shouldn't so it's NOT MATCH.