

Estonian Native Large Language Model Benchmark

Helena Grete Lillepalu, Tanel Alumäe

Department of Software Science, Tallinn University of Technology, Estonia
tanel.alumae@taltech.ee

Abstract

The availability of LLM benchmarks for the Estonian language is limited, and a comprehensive evaluation comparing the performance of different LLMs on Estonian tasks has yet to be conducted. We introduce a new benchmark for evaluating LLMs in Estonian, based on seven diverse datasets. These datasets assess general and domain-specific knowledge, understanding of Estonian grammar and vocabulary, summarization abilities, contextual comprehension, and more. The datasets are all generated from native Estonian sources without using machine translation. We compare the performance of base models, instruction-tuned open-source models, and commercial models. Our evaluation includes 6 base models and 26 instruction-tuned models. To assess the results, we employ both human evaluation and LLM-as-a-judge methods. Human evaluation scores showed moderate to high correlation with benchmark evaluations, depending on the dataset. Claude 3.7 Sonnet, used as an LLM judge, demonstrated strong alignment with human ratings, indicating that top-performing LLMs can effectively support the evaluation of Estonian-language models.

Keywords: LLM evaluation, Estonian

1. Introduction

Performance benchmarks play an important role in evaluating the capabilities of large language models (LLMs). These standardized tests enable researchers to assess various aspects of a model's language proficiency, such as text comprehension, reasoning ability, and generalization skills. For widely spoken languages, a variety of performance benchmarks have been developed to measure model performance across different tasks (Hendrycks et al., 2021; Wang et al., 2019b; Zellers et al., 2019).

While these benchmarks have provided valuable insights into the capabilities of LLMs in languages with large speaker populations, similar comprehensive evaluation frameworks are lacking for many languages with smaller speaker bases, including Estonian, a Uralic language with around one million native speakers.

We present a benchmark suite designed to evaluate the capabilities of LLMs on a variety of tasks in the Estonian language¹. The tasks include factual recall (exam and trivia questions), linguistic accuracy (word meanings, grammatical corrections, inflection), summarization, and structured information extraction from long texts (speaker name extraction). Together, these tasks offer a framework for assessing both the general and language-specific performance of LLMs in Estonian.

We evaluate a range of open- and closed-source base and instruction-tuned models. In total, we assess 6 base models and 26 instruction-tuned (chat) models. The instruction-tuned set includes many open-source models of different sizes, and 7

commercial models. Base models are evaluated in a 5-shot setting, while instruction-tuned models are evaluated in a zero-shot setting. We compare individual models across all tasks and also perform a broader comparison based on model types, specifically base models, open-source instruction-tuned models, and commercial models.

To evaluate the results, we combine human judgment with the "LLM-as-a-judge" methodology (Zheng et al., 2023). We create a set of localized open-ended questions that users might typically ask from LLMs. A custom web interface is developed to present one question alongside responses from two anonymous models, allowing users to choose the better answer, or indicate if both are equally good. For automated evaluation, we select Claude Sonnet as the LLM judge due to its strong performance in benchmark tests, and task it with evaluating the answers using the same criteria as the human evaluators.

To summarize, our contributions are:

- We present a benchmark suite for evaluating LLMs in Estonian, based on four new datasets and three existing ones, along with `lm-evaluation-harness` (Gao et al., 2024) configurations for automated testing.
- We evaluate a wide range of open and proprietary LLMs on the proposed benchmark.
- We conduct a human evaluation of chat models on a set of open-ended questions.
- We show that Claude 3.7 Sonnet can reliably judge open-ended responses with agreement comparable to human raters.
- We analyze which benchmark tasks correlate best with human evaluations.

¹<https://github.com/taltechnlp/lm-eval-harness-tasks-estonian>

2. Related work

2.1. Types of LM benchmarks

As LLMs have become more advanced, a variety of benchmarks have been developed to evaluate their capabilities across different dimensions. General-purpose benchmarks such as GLUE and SuperGLUE are widely used to assess language understanding and reasoning skills (Wang et al., 2019b,a), while others focus specifically on reading comprehension through question-answering tasks (Rajpurkar et al., 2016; Lai et al., 2017). More specialized benchmarks have also emerged. For example, MMLU evaluates subject-specific knowledge across 57 academic and professional domains using multiple-choice questions (Hendrycks et al., 2021), and HellaSwag is designed to assess common-sense reasoning (Zellers et al., 2019). In addition, several benchmarks target code generation by requiring models to write syntactically and functionally correct programs (Peng et al., 2024; Puri et al., 2021). Bias detection has also become a focal point, with benchmarks developed to measure biases in text generation across various social and demographic dimensions (Dhamala et al., 2021). Collectively, these benchmarks reflect the growing effort to comprehensively evaluate LLMs across a wide range of tasks and domains.

2.2. Benchmarking for underrepresented languages

While numerous high-quality benchmarks have been developed for high-resource languages, comparable resources for low-resource languages remain scarce. Languages such as Estonian face a notable gap in evaluation frameworks, which poses challenges for researchers and developers aiming to assess or fine-tune language models for these languages. Nevertheless, recent initiatives have begun to address this gap by developing evaluation suites tailored to low-resource languages. These efforts typically span a variety of tasks, ranging from general language understanding to more specialized applications such as reading comprehension and question answering (Etzaniz et al., 2024; Fajcik et al., 2024). In some cases, the focus has been on creating localized analogues of well-known English benchmarks, such as MMLU, adapted for specific low-resource languages (Yüksel et al., 2024; Koto et al., 2024).

A common approach to creating LLM benchmarks for low-resource languages is through manual or automatic translation. For example, the EU20 benchmark (Thellmann et al., 2024) contains benchmarks originally in the English language, such as MMLU, HellaSwag, TruthfulQA, machine-translated into 20 European languages. Similarly,

the M-MMLU benchmark (Hendrycks et al., 2021) provides professional human translations of the MMLU test set into 14 languages.

However, translated benchmarks introduce two major issues: translation noise and reduced cultural relevance. The lower the resource level of the target language, the more likely it is that machine translation will introduce errors, as translation quality tends to be poorer for such languages. This becomes particularly problematic when comparing model performance across multiple languages. Even with high-quality human translations, benchmarks may still fall short of providing reliable evaluations unless they are properly localized. Research has shown that localized benchmarks that are adapted to the culture and context of the target language correlate better with human preferences than those that are only translated. For example, while the EU20 version of MMLU is a direct translation, it lacks cultural adaptation. In contrast, language-specific benchmarks such as CMMLU (for Chinese) (Li et al., 2024) and KMMLU (for Korean) (Son et al., 2025) have been localized to reflect cultural context, resulting in more meaningful evaluations.

3. Benchmarks

To evaluate LLM performance in Estonian, we constructed a benchmark suite based on seven datasets, four created manually and three adapted from existing resources.

To convert a dataset of questions and reference answers into a LLM benchmark, we defined `lm-evaluation-harness` “tasks”. Each task specifies how a dataset item is converted into the LLM prompt, the method for extracting answers from model outputs, and the implementation of the evaluation metric.

For each individual benchmark, we implemented separate task definitions for base LLMs (using a 5-shot setting for most benchmarks) and for instruction-tuned LLMs (using a zero-shot setting).

Estonian National Exam benchmark: This benchmark is based on the Estonian National Exam dataset (https://huggingface.co/datasets/TalTechNLP/exam_et) that contains official exam questions from Estonian secondary and high school exams (2003–2024). It includes 1614 multiple-choice questions across seven subjects: Estonian as a second language, physics, chemistry, biology, geography, history, and civic studies.

The Estonian as a second language section contains 484 questions (246 secondary school, 238 middle school), primarily focused on gap-filling and reading comprehension, testing grammatical accuracy and text understanding. Other subjects con-

Prompt	Answer
Sa oled ekspert ning vastad küsimustele korrektselt. Sulle antakse küsimus koos vastusevariantidega. Anna vastus kujul: "Vastus: X", kus X on kas A, B või C. Ära kirjuta midagi muud! Vasta eesti keeles! Küsimus: Ühemunakaksikud on... Vastusevariandid: A: ühest soost B: eri soost C: nii ühest kui ka eri soost.	A
You are an expert and answer questions correctly. You are given a question with answer options. Provide the answer in the form: "Answer: X", where X is either A, B, or C. Do not write anything else! Answer in Estonian! Question: Identical twins are... Answer options: A: of the same sex B: of different sexes C: both of the same and different sexes.	A

Table 1: Question-answer pair from the Exams benchmark, together with its English translation. Question is already formatted as a prompt to an instruction-tuned LLM.

tain between 108 and 310 questions each, targeting factual knowledge and reasoning skills. The question formats include simple factual queries, gap-filling tasks, and true/false statements, all presented as multiple-choice questions with 2 to 15 answer options.

The dataset was compiled based on PDFs obtained from the state's education board. The PDFs were converted to text, edited for conversion errors and manually restructured into triplets of question, answer choices and the reference answer. An example question from this benchmark is shown in Table 1.

Trivia benchmark is based on the Trivia dataset (https://huggingface.co/datasets/TalTechNLP/trivia_et) that consists of 800 multiple-choice questions from the board game "Eesti mälumäng", covering history, science, culture, geography, sports, and more. Most questions are highly specific to Estonia or Estonian culture. The task evaluates general and Estonia-specific knowledge.

The dataset was compiled by scanning 400 individual double-sided game cards, each with two questions (one to be read up-side down), and with answers written on the opposite side of the card (see Figure 1). Scanned cards were OCR-ed, restructured using an LLM and post-edited for errors. An example sample prompt from this benchmark is shown in Table 2.

Declension benchmark uses the Declension dataset (https://huggingface.co/datasets/TalTechNLP/inflection_et) and evaluates LLMs' ability to generate correct grammatical case forms in Estonian – a morphologically rich language with 14 noun cases. The task consists of 200 adjective-noun pairs, each combined with 7 target case-number combinations

Prompt	Answer
Sa oled ekspert ning vastad küsimustele korrektselt. Sulle antakse küsimus koos vastusevariantidega. Anna vastus kujul: "Vastus: X", kus X on kas A, B, C või D. Ära kirjuta midagi muud! Vasta eesti keeles! Küsimus: Kus toimuvad reeglina Nargen Opera kontserdid? Vastusevariandid: A: Naissaarel B: Abruksal C: Aegnal D: Muhus	A
You are an expert and answer questions correctly. You are given a question with answer options. Provide the answer in the form: "Vastus: X", where X is either A, B, C, or D. Do not write anything else! Answer in Estonian! Question: Where are Nargen Opera concerts usually held? Answer options: A: Naissaar B: Abruksa C: Aegna D: Muhu	A

Table 2: Question-answer pair from the Trivia benchmark, together with translation.

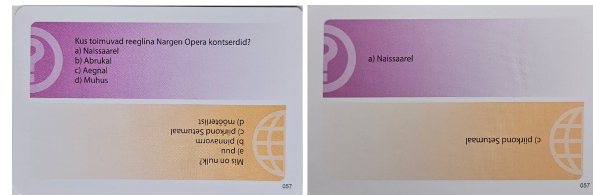


Figure 1: A game card from the quiz boardgame that was the source for the Trivia dataset.

(e.g., plural partitive, singular illative), resulting in 1400 total samples. Gold standard forms were generated using the FiloSoft Estonian morphological synthesizer (Kaalep and Vaino, 2001). An example question from this benchmark is shown in Table 3.

Word Meaning benchmark tests lexical understanding by prompting the model with a word definition and asking it to generate the correct term. The underlying dataset (https://huggingface.co/datasets/TalTechNLP/word_meanings_et) is derived from Ekilex, Estonian word and term base system of the Institute of the Estonian Language (Tavast et al., 2023). The full set contains over 54,000 entries, from which a random sample of 1000 items was used in the benchmark. Synonyms obtained via Estonian WordNet (Orav et al., 2018)

Prompt	Answer
Sa oled eesti keele ekspert. Sulle antakse sõnapaar ainsuse nimetavas käändes ja sinu ülesandeks on käänata see etteantud käändesse. Jälgi, kas käänat on küsitud ainsuses või mitmuses. Anna vastus kujul: "Vastus: {käänatud sõnapaar}". Sõnapaar ainsuse nimetavas käändes on: kollane päevalill. Kääna see käändesse: ainsuse omastav	kollase päevalille
You are an expert in the Estonian language. You are given a word pair in the singular nominative case, and your task is to inflect it into the given case. Pay attention to whether the requested case is singular or plural. Provide the answer in the format: "Answer: {inflected word pair}". The word pair in the singular nominative case is: kollane päevalill. Inflect it into the case: singular genitive	kollase päevalille

Table 3: Question-answer pair from the Declension benchmark, together with translation.

Prompt	Answer
Sa oled sõnade teadmise ekspert. Sulle antakse sõna definitsioon ning sina pead tagastama sõna, mida see definitsioon kirjeldab. Anna vastus kujul: "Vastus: {definitsioonile vastav sõna}". Vasta ainult vastusega. Sõna definitsioon(id): [suuri kulusid, väljaminekuid nõudev, 'esemete kohta:'] hinnalt kallis, kalli, kõrge hinnaga]	kulukas, kallis, kallivõitu, kõrgehinnaline, soolane, krõbe, kallis, kallihinnaline, hinnaline
You are an expert in knowledge about words. You are given a word definition, and you must return the word that the definition describes. Provide the answer in the format: "Answer: word corresponding to the definition". Respond with the answer only. Word definition(s): [requiring large expenses or costs, '(of objects:) with a high price']	expensive, high-priced, premium, pricey

Table 4: Question-answer pair from the Word Meaning benchmark, together with translation.

Prompt	Answer
Sa oled eesti keele ekspert. Sulle antakse ette eestikeelsed laused, kus võivad esineda grammatikavead. Tagasta korrektne lause, kus vead on parandatud. Juhul kui etteantud lause on juba õige, siis tagastagi see sama lause. Vasta ainult lausega! Originaalne: Kellena küll saaksin, kui ei oskaks lugeda? Mitte täiuslikku inimesena.	Kelleks küll saaksin, kui ei oskaks lugeda? Mitte täiuslikuks inimeseks.
You are an expert in the Estonian language. You are given Estonian sentences that may contain grammatical errors. Return the correct sentence with the errors fixed. If the given sentence is already correct, return it as is. Respond with the sentence only! Original: Whom could I possibly become if I couldn't read? Not as a perfect person.	Who could I possibly become if I couldn't read? Not a perfect person.

Table 5: Question-answer pair from the Grammar correction benchmark, together with translation.

were accepted as correct answers. A sample item from this benchmark is shown in Table 4.

Grammar Correction benchmark is based on Estonian grammar correction datasets (https://huggingface.co/datasets/TalTechNLP/grammar_et and https://huggingface.co/datasets/TalTechNLP/grammar2_et) which in turn are simplified versions of the EstGEC corpus. They contain original and corrected sentences written by native and non-native Estonian students (446 and 3285 sentences respectively). The model is tasked with correcting grammatical errors if present. Evaluation uses both exact-match (binary) accuracy and Levenshtein-based similarity scores. A sample question-answer pair from this benchmark is shown in Table 5.

News Summarization benchmark uses the ERRNews dataset (<https://huggingface.co/datasets/TalTechNLP/ERRnews>) that contains transcriptions of Estonian public broadcaster (ERR) radio news stories paired with human-written short summaries (Härm and Alumäe, 2022). Transcripts are generated using an automatic speech recognition (ASR) system.

Prompt	Answer
Sa oled tehisintellekti ekspert eesti keele alal. Su ülesanne on analüüsida uudislugusid ja genereerida ühe või kahe lausega lühike kokkuvõte, mis kirjeldab uudisloo sisu. Ära unusta alati vastata eesti keeles. Vasta ainult kokkuvõttega! Uudislugu: Eesti Energia hinnangul oli auvere elektrijaama ehitus keerukas pikk ja pingeline protsess. Ehituslepingu sõlmimisest kuni jaama käivitamiseni jõuti nelja aastaga ...	Eesti Energia võttis ehitajalt üle kauaoodatud Auvere elektrijaama. Eesti moodsaim põlevkivijaam suudab katta umbes neljandiku riigi elektrivajadusest.
You are an artificial intelligence expert in the Estonian language. Your task is to analyze news stories and generate a short one- or two-sentence summary that describes the content of the news story. Always remember to respond in Estonian. Respond with the summary only! News story: According to Eesti Energia, the construction of the Auvere power plant was a complex, long, and demanding process. It took four years from signing the construction contract to the commissioning of the plant...	Eesti Energia took over the long-awaited Auvere power plant from the builder. Estonia's most modern oil shale power plant can cover about a quarter of the country's electricity needs.

Table 6: Sample prompt and expected answer from the News Summarization benchmark, together with translation.

From over 10,000 records, a test split of 523 items was used. Evaluation uses ROUGE-L metrics. A sample response-answer pair from this benchmark is shown in Table 6.

Speaker Name Extraction benchmark is based on a dataset consisting of speaker-attributed transcripts from the main evening radio news program of the Estonian Public Broadcasting (*Päevakaja*), along with the corresponding list of speakers (https://huggingface.co/datasets/TalTechNLP/paevakaja_speakers). The speaker list includes both news anchors and journalists, who typically introduce themselves, as well as interviewees introduced by the moderators. The reference lists of speakers were manually compiled by human editors.

The transcripts were generated using an ASR system, and each speaking turn is prefixed with a speaker code obtained from a speaker diarization system. The LLM's task is to identify all speakers occurring in the show. This benchmark evaluates the model's ability to follow instructions, understand text, and extract structured information. Performance is measured using precision, recall, and F1-score (the main metric). A sample item from this benchmark is shown in Table 7.

4. Results

We evaluated a total of 6 base models and 26 chat models on the seven benchmark tasks described above. For all benchmarks, we used metrics scaled between 0 and 1, where higher values indicate better performance. This design allows for meaningful averaging of results across benchmarks.

For the grammar correction benchmark, we employed a custom metric based on the Levenshtein distance, defined as $\frac{1}{1 + \text{Levenshtein distance}}$, which also

Prompt	Answer
Sa oled ekspert Eesti avaliku elu tegelaste alal. Sulle antakse uudiste- või vestlussaate automaatne transkriptsioon, koos kõneleja koodidega. Proovi arvata, millised isikud saates kõnelevad, nii reporterid kui ka intervjuueeritavad. Veendu et sa vastad inimestega, kes kõnelevad, mitte inimestega, kellest räägitakse. Väljasta tulemus JSONi abil. JSON väljundi formaadi näide: ["nimi1", "nimi2"]. Ära liialt riski, meile on täpsus olulisem kui saagis. Nimed võivad olla valesti transkribeeritud, kasuta oma taustateadmisi, et neid vajadusel korrigeerida. Kindlasti anna vastuseks ainult üks JSON list ja mitte midagi muud! S1: Tere õhtust, kell on kuus ning Päevakaja võtab kokku tänasest olulisema meil ja maailmas, mina olen toimetaja Margitta otsmaa...	"Margitta Otsmaa", "Jaak Aab", "Alvar Soesoo", ...
You are an expert on Estonian public figures. You are given an automatic transcript of a news or talk show, along with speaker codes. Try to identify which individuals are speaking in the program, including both reporters and interviewees. Make sure you name the people who are speaking, not those who are being talked about. Output the result in JSON format. Example of JSON output format: ["name1", "name2"]. Don't take unnecessary risks — accuracy is more important than recall for us. The names may be misspelled in the transcript; use your background knowledge to correct them if necessary. Be sure to respond with only one JSON list and nothing else! S1: Good evening, it's six o'clock and Päevakaja brings you a summary of the most important news from Estonia and the world; I'm editor Margitta otsmaa...	"Margitta Otsmaa", "Jaak Aab", "Alvar Soesoo", ...

Table 7: Sample prompt and expected answer from the Speaker Name Extraction benchmark, together with translation. Note how the name of the editor *Margitta Otsmaa* has a casing error in the automatic transcript.

produces values in the range $[0, 1]$ and is therefore compatible with the overall evaluation framework.

4.1. Base Models

Table 8 shows the results of base models across the benchmarks, together with the average score. We observe that the ranking of models varies considerably across the tasks. Overall, Gemma-2 27B and Llama-3 70B consistently achieved the best results, with the exception of the trivia benchmark, where Llama-3 70B significantly outperformed other models.

4.2. Instruction-Tuned Models

To allow easier analysis, instruction-tuned models were divided into four categories: proprietary models from Google (Gemini), OpenAI (GPT) and Anthropic (Claude), and open models divided into small (up to 12B parameters), medium (up to 235B parameters) and large sizes. We acknowledge that this division is somewhat arbitrary.

Most recent proprietary models are *reasoning models*, meaning they generate internal “thinking

tokens” that are typically hidden from the end user, before producing the final response. In our experiments, these models were configured for minimal reasoning by setting the model parameter `reasoning_effort=minimal`.

Table 9 depicts the normalized results of the instruction-tuned models across the tasks. Note that the large and medium sized open models were evaluated using different API providers available through OpenRouter², while the small models were evaluated on our own server.

Among the tested models, Gemini 2.5 Pro achieved the highest average performance and the best scores in four individual tasks. GPT-5, Claude Sonnet 3.7 and GPT-4o each obtained highest ranked result in one benchmark.

While the large open models lag behind the highest-ranked proprietary models in all categories, the difference is surprisingly small. For example, the best open model – Kimi-K2 from Moonshot AI – equals Claude Sonnet 4 in the average score (albeit it must be acknowledged that Claude Sonnet 4 resulted in noticeable regression in its Estonian abilities, compared to Claude Sonnet 3.7). As expected, the results are highly correlated with models’ size, especially in the Exams and Trivia benchmarks that largely rely on factual knowledge. For tasks that expect more linguistic abilities (Declension, Words, Grammar), there is higher variety among similarly-sized models, probably resulting from different pre-training and instruction-tuning data.

The highest-ranked small model is Llama-3.1 finetuned by the EstLLM project³, with the specific goal of improving its Estonian abilities. This indicates that finetuning using language-specific datasets is a highly useful method for this goal, as the 8B EstLLM model outperforms all other small models and all but one medium-sized models.

4.3. Correlations between tasks

To explore the relationships between task types, we computed Pearson correlation coefficients across all benchmark results (Figure 2). No negative correlations were observed, indicating general consistency in model performance rankings. The strongest correlation ($r = 0.98$) was found between the inflection and word definition datasets, suggesting they assess highly overlapping linguistic capabilities. These tasks, along with trivia, exams, and

²During the benchmarking runs, we noticed that the quality of large open models varied dramatically, based on the actual API provider that was used by OpenRouter to fulfill the request. In order to mitigate this issue, we configured OpenRouter not to use those API providers: NovitaAI, Baseten, Together, Fireworks, SambaNova, AtlasCloud, Hyperbolic.

³<https://huggingface.co/tartuNLP/llama-estllm-prototype-0825>

Model	Mean Score	Exams Accuracy	Trivia Accuracy	Declension Accuracy	Words Accuracy	Grammar Levenshtein*	News ROUGE-L	Speakers F1-score
Llama 3.1 70B	0.438	0.667	0.433	0.586	0.379	0.291	0.160	0.550
Gemma 2 27B	0.389	0.693	0.311	0.505	0.354	0.306	0.164	n/a**
Qwen2 72B	0.352	0.659	0.328	0.306	0.188	0.257	0.134	0.591
Mistral Nemo Base (12B)	0.328	0.487	0.303	0.416	0.237	0.245	0.130	0.476
Llama 3.1 8B	0.295	0.405	0.318	0.268	0.182	0.242	0.126	0.522
Gemma 2 9B	0.331	0.616	0.295	0.396	0.259	0.263	0.157	n/a**

Table 8: Base LLM performances across 8 Estonian-language tasks with the metrics used for each task ordered by the average score in descending order. By the average score, the mean of numeric scores (excluding “n/a”) is meant. **Gemma models failed on the speaker recognition task due to too small context length.

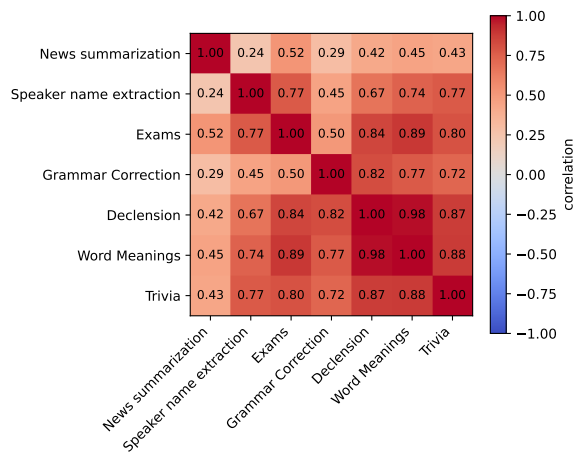


Figure 2: Pearson correlation matrix between the results of different datasets.

grammar, form a tightly interrelated group, with consistently strong inter-correlations, reflecting shared demands on syntactic precision and lexical knowledge. In contrast, the news summarization dataset exhibited the weakest correlations with other tasks. This divergence implies that ROUGE-based evaluation may not reliably reflect actual model quality in text generation tasks, especially when applied to noisy or semantically variable outputs such as news content. Overall, the analysis suggests that while most benchmarks target related competencies, summarization behaves anomalously and may require improved evaluation methodologies.

4.4. Impact of model scale

Figure 3 illustrates the relationship between model size and performance, based on the total parameter count rather than the number of active parameters, which may also influence deployment considerations. As expected, a clear positive correlation can be observed between model size and quality. Notably, the small Llama 3.1 model fine-tuned

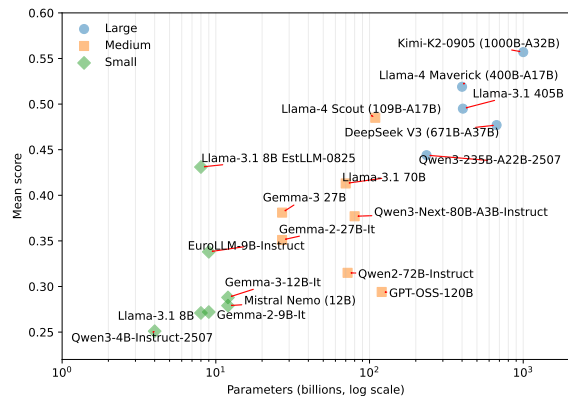


Figure 3: Model size vs. mean score (open models only). Absolute parameter size was used.

on Estonian data performs significantly above the general trendline, while some models (e.g. GPT-OSS-120B) perform below what their size would suggest.

5. Validation

To evaluate whether the developed Estonian-language benchmark reflects real-world usefulness, we conducted a validation study involving both human judgments and LLM-based evaluation. The goal was to assess whether model performance on benchmark tasks correlates with how helpful their outputs are perceived to be by users. This dual-validation approach provides insight into both the practical reliability of the benchmarks and the feasibility of using LLMs as automated evaluators.

5.1. Human Evaluation

A custom evaluation set was created using 47 questions translated from the MT-Bench dataset, spanning categories such as reasoning, writing, mathematics, general knowledge, and humanities (Zheng et al., 2023). Culturally irrelevant or non-Estonian

Model	Mean Score	Exams Accuracy	Trivia Accuracy	Decl. Accuracy	Words Accuracy	Grammar Leven- shtein*	News ROUGE- L	Speakers F1-score
<i>Proprietary models</i>								
Gemini 2.5 Pro	0.649	0.897	0.781	0.955	0.631	0.321	0.138	0.820
GPT-5	0.628	0.875	0.823	0.934	0.529	0.313	0.114	0.806
Claude Sonnet 3.7	0.605	0.905	0.511	0.937	0.570	0.391	0.135	0.785
Claude Sonnet 4	0.557	0.876	0.610	0.822	0.450	0.236	0.131	0.776
GPT-4o	0.595	0.900	0.513	0.948	0.555	0.395	0.127	0.724
Gemini 2.0 Flash	0.565	0.900	0.429	0.932	0.540	0.282	0.124	0.748
GPT-4o Mini	0.447	0.784	0.370	0.554	0.331	0.321	0.115	0.651
<i>Large open models</i>								
Kimi-K2-0905 (1000B-A32B)	0.557	0.880	0.513	0.876	0.435	0.270	0.140	0.783
Llama-4 Maverick (400B-A17B)	0.519	0.841	0.460	0.828	0.453	0.261	0.131	0.656
Llama-3.1 405B Instruct	0.495	0.817	0.440	0.680	0.410	0.286	0.129	0.702
Deepseek V3 (671B-A37B)	0.477	0.820	0.398	0.631	0.452	0.256	0.113	0.668
Qwen3-235B-A22B-2507	0.444	0.829	0.374	0.555	0.324	0.207	0.108	0.713
<i>Medium-sized open models</i>								
Llama-4 Scout (109B-A17B)	0.485	0.824	0.383	0.766	0.407	0.270	0.125	0.622
Llama-3.1 70B Instruct	0.413	0.760	0.359	0.464	0.306	0.199	0.129	0.674
Gemma-3 27B Instruct	0.381	0.756	0.336	0.404	0.356	0.226	0.118	0.471
Qwen3-Next-80B-A3B-Instruct	0.377	0.776	0.369	0.362	0.246	0.195	0.109	0.585
Gemma-2-27B-It	0.351	0.760	0.325	0.406	0.266	0.223	0.128	n/a**
Qwen2-72B-Instruct	0.315	0.739	0.376	0.177	0.152	0.060	0.112	0.591
GPT-OSS-120B	0.294	0.777	0.364	0.356	0.290	0.194	0.074	0.000
<i>Small open models</i>								
Llama-3.1 8B EstLLM-0825	0.431	0.575	0.425	0.811	0.327	0.275	0.152	0.452
EuroLLM-9B-Instruct	0.338	0.610	0.381	0.644	0.338	0.248	0.104	0.038
Gemma-3-12b-It	0.288	0.736	0.320	0.356	0.289	0.197	0.119	0.000
Mistral Nemo Instruct (12B)	0.279	0.570	0.299	0.274	0.177	0.187	0.034	0.410
Gemma-2-9B-It	0.272	0.556	0.278	0.308	0.178	0.182	0.128	n/a**
Llama-3.1 8B Instruct	0.271	0.542	0.309	0.099	0.132	0.189	0.119	0.505
Qwen3-4B-Instruct-2507	0.251	0.553	0.276	0.056	0.090	0.136	0.105	0.540

Table 9: Instruction-tuned LLM performances across the tasks. The scores between base LLMs and instruction-tuned LLMs are not directly comparable, as the tasks use different instructions and different few-shot setups.

content was excluded. Additionally, 10 Estonian-specific questions were authored to better reflect local contexts. All questions were open-ended and designed to simulate practical LLM usage.

Responses were generated using a subset of evaluated instruction-tuned models that were available in March 2025. Human annotators were presented pairs of anonymized model responses via a Django-based web interface and were asked choose which one was better, or indicate if they were of equal quality. Results were scored using both win rates and ELO ratings. The win rates are shown in Table 10. Among the models, Claude 3.7 Sonnet received the highest average score (0.75), followed closely by Deepseek V3 (0.75) and Gemini 2.0 Flash (0.70). Surprisingly, Gemma 2 27B outperformed both GPT-4o and GPT-4o Mini in human evaluations. At the lower end, Gemma 2 9B consistently received the weakest ratings.

To assess alignment between human preferences and benchmark task performance, we computed Pearson and Spearman correlation coefficients between normalized benchmark scores and human rating scores (Table 11). The correlations were generally strong or moderate, depending on the dataset. The highest correlations were observed for the Exams (Pearson $r = 0.86$, Spearman $\rho = 0.85$) and Word Definitions benchmarks ($r = 0.80$, $\rho = 0.79$), indicating that benchmark scores in these domains are highly predictive of perceived model quality. Lower correlations appeared in tasks such as Grammar Correction ($r = 0.48$) and especially News Summarization ($r = 0.44$), suggesting that automatic metrics like ROUGE may not capture meaningful differences in generation quality in all cases.

Model	Score
Claude 3.7 Sonnet	0.754
Deepseek Chat V3	0.746
Gemini 2.0 Flash 001	0.699
Gemma 2 27B It	0.642
Llama 3.1 70B Instruct	0.552
GPT-4o	0.545
GPT-4o Mini	0.533
Llama 3.1 405B Instruct	0.500
Qwen2 72B Instruct	0.442
Llama 3.1 8B Instruct	0.389
Mistral Nemo Instruct	0.327
Gemma 2 9B It	0.211

Table 10: Model win scores based on human evaluation.

Dataset	r
Exams	0.86
Word meanings	0.80
Declension	0.72
Word meanings	0.66
Speaker name extraction	0.63
Grammar	0.48
News summarization	0.44

Table 11: Pearson correlation coefficients (r) between benchmark test and human evaluation results.

5.2. LLM-as-a-Judge

To complement human evaluation and explore scalability, we also tested the effectiveness of using a strong LLM as an automatic evaluator. Claude 3.7 Sonnet was selected for this role due to its strong performance in both the benchmark and human evaluations⁴. The model was provided with instructions to select the better of two responses (or indicate a tie), following the same protocol used for human raters. Its judgments were then compared to those of human annotators.

The agreement between Claude’s ratings and human judgments was high, with a Pearson correlation of 0.94 and Spearman correlation of 0.92 ($p < 0.001$), indicating a strong linear and ordinal relationship. Table 12 shows the results of the LLM-as-a-judge evaluation using Claude 3.7 Sonnet. The model rankings closely matched those from human annotators (Table 10), with the same top three models (Deepseek Chat V3, Claude 3.7 Sonnet, and Gemini 2.0 Flash) though in a slightly different order. Interestingly, both evaluation methods ranked Llama 3.1 70B higher than the larger 405B variant. At the lower end, Llama 8B was ranked lowest

⁴This evaluation was done in March 2025 and thus excludes newer models.

Model	Score
Deepseek V3	0.772
Gemini 2.0 Flash 001	0.770
Claude 3.7 Sonnet	0.759
GPT-4o	0.677
Gemma 2 27B It	0.654
GPT-4o Mini	0.620
Llama 3.1 70B Instruct	0.473
Llama 3.1 405B Instruct	0.428
Qwen2 72B Instruct	0.403
Mistral Nemo Instruct	0.226
Gemma 2 9B It	0.214
Llama 3.1 8B Instruct	0.205

Table 12: Model win scores based on LLM evaluation using Claude Sonnet as a judge.

by Claude, while humans rated Gemma 2 9B as the weakest. Across both evaluation methods and scoring schemes (win rate and ELO), model rankings varied by no more than one position, indicating strong agreement overall.

6. Conclusion

We presented a comprehensive benchmark for evaluating LLMs on Estonian-language tasks. The benchmark covers seven diverse tasks addressing grammatical inflection, lexical semantics, summarization, factual knowledge, and information extraction from long texts. Our evaluation of six base models and 26 instruction-tuned models reveals substantial variation in performance across both tasks and model types. While flagship commercial models consistently outperform open-source models, recent large open models such as Kimi-K2 achieve competitive results in several tasks. Our findings also confirm that language-specific fine-tuning of open models leads to significant improvements across all Estonian-specific capabilities.

Our validation study shows that benchmark performance correlates strongly with human judgments of practical usefulness. In particular, the benchmark based on Estonian state exams showed the highest correlation with user-rated LLM quality in open-ended question answering. Furthermore, we demonstrate that a strong LLM – Claude 3.7 Sonnet in our case – can serve as a reliable automatic evaluator, closely aligning with human preferences. These results suggest that robust LLM benchmarking is feasible even for low-resource languages, and that top-tier LLMs can meaningfully support scalable evaluation.

Benchmark code and datasets are publicly released to encourage further research in Estonian NLP and to support benchmarking efforts for other underrepresented languages.

Ethics Statement

All datasets used in this work were collected and processed in accordance with Estonian and European Union data protection and copyright regulations.

Human evaluation was conducted with informed consent from voluntary participants. No personal or identifying information was collected during the evaluation.

We acknowledge that LLMs may reproduce social, cultural, or gender biases present in training data. While the benchmark aims to measure general and language-specific capabilities of LLMs, not their social alignment, we encourage responsible interpretation of results and further study of fairness and bias in Estonian-language models.

Bibliographical References

- Jwala Dhamala, Tony Sun, Varun Kumar, Satyapriya Krishna, Yada Pruksachatkun, Kai-Wei Chang, and Rahul Gupta. 2021. [BOLD: Dataset and metrics for measuring biases in open-ended language generation](#). In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21. ACM.
- Julen Etxaniz, Oscar Sainz, Naiara Perez, Itziar Aldabe, German Rigau, Eneko Agirre, Aitor Ormazabal, Mikel Artetxe, and Aitor Soroa. 2024. [Latxa: An open language model and evaluation suite for Basque](#). *arXiv preprint arXiv:2403.20266*.
- Martin Fajcik, Martin Docekal, Jan Dolezal, Karel Ondrej, Karel Beneš, Jan Kapsa, Pavel Smrz, Alexander Polok, Michal Hradis, Zuzana Neverilova, Ales Horak, Radoslav Sabol, Michal Stefanik, Adam Jirkovsky, David Adamczyk, Petr Hyner, Jan Hula, and Hynek Kydlíček. 2024. [BenCzechMark : A Czech-centric multitask and multimetric benchmark for large language models with duel scoring mechanism](#). *arXiv preprint arXiv:2412.17933*.
- Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac'h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. 2024. [The Language Model Evaluation Harness](#).
- Henry Härm and Tanel Alumäe. 2022. Abstractive summarization of broadcast news stories for Estonian. *Baltic Journal of Modern Computing*, 10(3).
- Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. [Measuring massive multitask language understanding](#). In *Proc. of ICLR*. OpenReview.net.
- Heiki-Jaan Kaalep and Tarmo Vaino. 2001. Complete morphological analysis in the linguist's toolbox. In *Congressus Nonus Internationalis Fenno-Ugristarum, Pars V*, pages 9–16.
- Fajri Koto, Haonan Li, Sara Shatnawi, Jad Doughman, Abdelrahman Boda Sadallah, Aisha Alraeesi, Khalid Almubarak, Zaid Alyafeai, Neha Sengupta, Shady Shehata, Nizar Habash, Preslav Nakov, and Timothy Baldwin. 2024. [ArabicMMLU: Assessing massive multitask language understanding in Arabic](#). *arXiv preprint arXiv:2402.12840*.
- Guokun Lai, Qizhe Xie, Hanxiao Liu, Yiming Yang, and Eduard Hovy. 2017. [RACE: Large-scale ReAding comprehension dataset from examinations](#). In *Proc. of EMNLP*, pages 785–794, Copenhagen, Denmark. Association for Computational Linguistics.
- Haonan Li, Yixuan Zhang, Fajri Koto, Yifei Yang, Hai Zhao, Yeyun Gong, Nan Duan, and Timothy Baldwin. 2024. [CMMLU: Measuring massive multitask language understanding in Chinese](#). In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 11260–11285, Bangkok, Thailand. Association for Computational Linguistics.
- Heili Orav, Kadri Vare, and Sirli Zupping. 2018. [Estonian Wordnet: Current state and future prospects](#). In *Proceedings of the 9th Global Wordnet Conference*, pages 347–351, Nanyang Technological University (NTU), Singapore. Global Wordnet Association.
- Qiwei Peng, Yekun Chai, and Xuhong Li. 2024. [HumanEval-XL: A multilingual code generation benchmark for cross-lingual natural language generalization](#). In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 8383–8394, Torino, Italia. ELRA and ICCL.
- Ruchir Puri, David S. Kung, Geert Janssen, Wei Zhang, Giacomo Domeniconi, Vladimir Zolotov, Julian Dolby, Jie Chen, Mihir Choudhury, Lindsey Decker, Veronika Thost, Luca Buratti,

- Saurabh Pujar, Shyam Ramji, Ulrich Finkler, Susan Malaika, and Frederick Reiss. 2021. [CoNet: A large-scale AI for code dataset for learning a diversity of coding tasks](#). *arXiv preprint arXiv:2105.12655*.
- Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. [SQuAD: 100,000+ questions for machine comprehension of text](#). In *Proc. of EMNLP*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.
- Guijin Son, Hanwool Lee, Sungdong Kim, Seung-gone Kim, Niklas Muennighoff, Taekyoon Choi, Cheonbok Park, Kang Min Yoo, and Stella Biderman. 2025. [KMMLU: Measuring massive multitask language understanding in Korean](#). In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 4076–4104, Albuquerque, New Mexico. Association for Computational Linguistics.
- Arvi Tavast, Indrek Hein, Jelena Kallas, Kristina Koppel, Margit Langemets, Tõnis Nurk, Esta Prangel, Signe Saar, Katrin Tsepelina, Mari Vaus, Ülle Viks, Marja Vaba, Jako Bucht, Martin Laubre, and Yogesh Sharma. 2023. Ekilex 2023. EKI sõnastiku- ja terminibaasisüsteem. Accessed 2023.
- Klaudia Thellmann, Bernhard Stadler, Michael Fromm, Jasper Schulze Buschhoff, Alex Jude, Fabio Barth, Johannes Leveling, Nicolas Flores-Herr, Joachim Köhler, René Jäkel, et al. 2024. [Towards multilingual LLM evaluation for European languages](#). *arXiv preprint arXiv:2410.08928*.
- Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019a. [Superglue: A stickier benchmark for general-purpose language understanding systems](#). In *Proc. of NeurIPS*, pages 3261–3275.
- Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019b. [GLUE: A multi-task benchmark and analysis platform for natural language understanding](#). In *Proc. of ICLR*. OpenReview.net.
- Arda Yüksel, Abdullatif Köksal, Lütfi Kerem Şenel, Anna Korhonen, and Hinrich Schütze. 2024. [TurkishMMLU: Measuring massive multitask language understanding in Turkish](#). *arXiv preprint arXiv:2407.12402*.
- Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. [HellaSwag: Can a machine really finish your sentence?](#) In *Proc. of ACL*, pages 4791–4800, Florence, Italy. Association for Computational Linguistics.
- Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. [Judging LLM-as-a-judge with MT-Bench and Chatbot Arena](#). In *Proc. of NeurIPS*.