

HELE-ANDRA KUULMETS

Cross-Lingual Transfer Learning and
Evaluation in Low-Resource Settings



HELE-ANDRA KUULMETS

Cross-Lingual Transfer Learning and
Evaluation in Low-Resource Settings



UNIVERSITY OF TARTU

Press

Institute of Computer Science, Faculty of Science and Technology, University of Tartu, Estonia.

Dissertation has been accepted for the commencement of the degree of Doctor of Philosophy (PhD) in Computer Science on February 3, 2026 by the Council of the Institute of Computer Science, University of Tartu.

Supervisor

Prof. Mark Fišel
University of Tartu
Estonia

Opponents

Prof. Barbara Plank
Ludwig-Maximilians-Universität München
Germany

Dr. Jindřich Helcl
University of Oslo
Norway

The public defense will take place on April 9, 2026 at 11:00 in Narva Rd. 18-1021.

The publication of this dissertation was financed by the Institute of Computer Science, University of Tartu.

ISSN 2613-5906 (print)

ISSN 2806-2345 (pdf)

ISBN 978-9908-57-151-5 (print)

ISBN 978-9908-57-152-2 (pdf)

Copyright © 2026 by Hele-Andra Kuulmets

University of Tartu Press

<http://www.tyk.ee/>

ABSTRACT

Modern AI applications rely on language models trained on massive datasets, leaving many languages with limited resources underrepresented in these systems. This gap reduces the usability of AI tools in local, native languages, discouraging their everyday use and reinforcing technological inequality across linguistic communities.

A common strategy to address this issue is cross-lingual transfer learning, where a model is trained on multiple languages, enabling knowledge learned from high-resource languages such as English to improve performance in lower-resource ones. Specifically, the transfer occurs because multilingual training encourages the model to align internal representations across languages. It has shown to be especially beneficial for improving performance of low-resource languages included in the training set.

This thesis investigates methods to strengthen cross-lingual knowledge transfer from English to Estonian by leveraging various resources, including textual data in Estonian, existing multilingual base models, large language models, and machine translation. Estonian is a Finnic language belonging to the Finno-Ugric branch of the Uralic language family and is most closely related to Finnish. It is natively spoken by around 1.1 million people and is the official language of Estonia. With 14 grammatical cases for nouns, an extensive verb inflection system, and relatively free word order, Estonian is considered a morphologically rich language, contrasting with many typologically different, high-resource languages (e.g., English) that have been the primary focus of NLP research.

The work is presented in two parts. The first focuses on pre-LLM settings, optimising Estonian performance on a commonsense reasoning task through multilingual training and synthetic data, and analysing the impact of translation quality. The second part focuses on adapting an English-centric large language model to Estonian by combining continued pretraining with multilingual instruction tuning and leveraging synthetic Estonian and English instructions alongside high-quality human-written English instructions.

This work also contributes insights into the evaluation of cross-lingual transfer learning methods. We show that translated test datasets can overestimate model performance compared to native test sets, and introduce a multilingual, multi-turn conversational benchmark for Estonian and other Finno-Ugric languages to enable more reliable assessment of conversational language models.

Across the experiments presented in this work, incorporating multilingual data consistently improves Estonian results, even when using only synthetic resources, suggesting that cost-effective cross-lingual strategies can significantly expand the reach of AI to low-resource languages. Based on these findings, we believe that approaches leveraging cross-lingual transfer offer a viable path to closing the performance gap between high- and low-resource languages, making advanced language technologies available to a wider range of linguistic communities.

CONTENTS

List of original publications	11
1. Introduction	13
2. Background	18
2.1. Fundamentals of NLP	18
2.1.1. The Transformer Architecture	18
2.1.2. Transformer Encoders and Decoders	20
2.1.3. Pre-Training	20
2.1.4. Fine-Tuning	21
2.2. Large Language Models	22
2.2.1. Model Scaling	22
2.2.2. Instruction Tuning	23
2.3. Cross-Lingual Transfer Learning	24
2.3.1. Multilingual Pre-Training	25
2.3.2. Zero-Shot Model Transfer	25
2.3.3. Multilingual Fine-Tuning	26
2.3.4. English as a Pivot	27
2.4. Evaluation	28
2.4.1. Creating Evaluation Datasets	28
2.4.2. Evaluating Instruction-Following Language Models	29
2.4.3. <i>LLM-as-a-Judge</i>	30
3. Cross-Lingual Transfer Learning for Estonian NLU	31
3.1. EstCOPA: An Estonian Commonsense Reasoning Benchmark	31
3.1.1. The COPA Dataset	32
3.1.2. Building the Estonian Dataset	32
3.1.3. Comparison with the XCOPA Dataset	33
3.2. Cross-Lingual Transfer Learning Experiments with EstCOPA	34
3.2.1. Experimental Setup	34
3.2.2. English Data Drives Estonian NLU	35
3.2.3. Supplementary Study: The Impact of Translation Quality	37
3.3. The Problem with Translated Test Datasets	38
3.3.1. Experiment with a Reading Comprehension Task	39
3.4. Conclusion	40
4. Cross-Lingual Transfer Learning with LLMs	41
4.1. Experiments on Eliciting Knowledge Transfer to Estonian	41
4.1.1. Alpaca-est Dataset	42
4.1.2. Experimental Setup	42
4.1.3. Knowledge Transfer with Continued Pre-Training	43

4.1.4. Knowledge Transfer with Translation Task Instructions . . .	45
4.1.5. Knowledge Transfer with High-Quality English Instructions	46
4.1.6. Will the Model Still Understand English?	47
4.2. State of Open LLMs in Understanding Finno-Ugric Languages . .	49
4.2.1. Experimental Setup	49
4.2.2. Knowledge Transfer from Pre-Training	50
4.2.3. Supplementary Study. Will English as a Pivot Improve Mul- tilingual Capabilities of LLMs?	52
5. Conversational Benchmark for Finno-Ugric Languages	55
5.1. Desiderata	56
5.2. Dataset Construction	57
5.3. The Benchmark in Practice. Human Evaluation of LLMs	58
6. Conclusion	62
Bibliography	65
Acknowledgements	82
Sisukokkuvõte (Summary in Estonian)	84
Publications	87
Curriculum Vitae	172
Elulookirjeldus (Curriculum Vitae in Estonian)	173

LIST OF FIGURES

1. Two examples from the COPA dataset.	32
2. An example from the COPA dataset and its translations to Estonian, illustrating differences in translation quality. The EstCOPA _{MT} version contains a lexical error, translating <i>pin</i> as <i>pinal</i> (‘pencil case’), which makes the premise lose its meaning. Both EstCOPA _{PE} and XCOPA translations preserve the intended meaning, although translating <i>pin</i> as <i>ora</i> (‘thorn’) introduces a minor semantic deviation.	33
3. Effect of continued pre-training dataset size on Estonian downstream tasks. The colors of the bars indicate the size of the pretraining dataset. The models are fine-tuned on cross-lingual general task instruction dataset Alpacas.	44
4. Effect of sequential fine-tuning, first on translation task instructions (TrTASK) and then on general task instructions (Alpacas), compared to fine-tuning without the translation task stage (Figure 3). Results are shown across different amounts of continued pre-training tokens; bar colors indicate the size of the pre-training dataset.	45
5. Comparison of few-shot prompting and few-shot chain-of-thought (CoT) prompting on Belebele (left, 1-shot), QA (middle, 1-shot), and XCOPA (right, 1-, 3- and 5-shot) tasks for five Finno-Ugric languages. CoT prompting generally provides little benefit for Belebele, reduces performance on QA, and yields mostly positive gains on XCOPA, with stronger models benefiting less. The bars show the scores with few-shot prompting. The tiny horizontal line (–) indicates the score with few-shot CoT prompting with the same number of shots.	53
6. Pairwise comparison of human and machine translations for Võro, Komi, and Livonian, showing the percentage of annotator preferences for each method. Results indicate a strong preference for human translations in Komi and Livonian, while Võro shows comparatively higher acceptance of machine-translated data.	58
7. Screenshot of the survey used to collect human annotations for responses generated by different models using instructions from SMUGRI-MT-BENCH. For readability, the survey shown here is translated into English (except the assistant’s answer, which is in Võro). The original instructions were presented to annotators in Estonian (for Estonian, Võro and Livonian) or in Russian (for Komi).	59
8. Human evaluation results for conversational LLMs on low-resource Finno-Ugric languages, comparing helpfulness and naturalness scores across models.	60
9. Category-level breakdown of helpfulness scores in human evaluation of conversational LLMs for low-resource Finno-Ugric languages.	61

LIST OF TABLES

1. Overview of training setups used in our experiments for encoder-only models. The first three rows show multilingual training configurations for improving Estonian natural language understanding and the last three rows shows the baselines. PT stands for pre-training and FT for fine-tuning.	35
2. Performance of encoder-only models on Estonian commonsense reasoning (XCOPA _{ET}). Best results per column are shown in bold. Scores from Ponti et al. (2020) are denoted with symbol †.	36
3. TRANSLATE-TEST performance on XCOPA _{ET} . Both results are from Ponti et al. (2020). Best result is shown in bold.	36
4. Performance of decoder-only models on XCOPA _{ET} under zero-shot and few-shot prompting. The top block reports our experiments with few-shot examples only in Estonian (et), only in English (en), or both in English and Estonian (mixed). The bottom block shows results reported by Lin et al. (2022b). TRTE stands for TRANSLATE-TEST.	37
5. Comparison of TRANSLATE-TRAIN performance using EstCOPA _{MT} and EstCOPA _{PE} training datasets.	37
6. Comparison of model accuracies on the XCOPA and EstCOPA _{PE} test datasets when trained on EstCOPA _{MT} +COPA.	38
7. Performance of the TRANSLATE-TEST method on native EstQA and translated XQuAD _{ET} test datasets.	39
8. Examples from Alpaca-est. These examples illustrate unstable quality of the synthetic data generated with gpt-3.5-turbo-0613. While the first two examples contain well-formed instructions and correct outputs, last two exhibit odd or problematic content.	42
9. Performance of various instruction-tuning strategies on Estonian NLP tasks. Adding high-quality English task instructions (HQI) and translation task instructions (HQTRTASK) improves zero-shot performance. We call (4) LLAMMAS and (5) LLAMMAS-MT.	47
10. Comparison of English QA and commonsense reasoning (COPA) performance across all instruction-tuned models.	48
11. Performance comparison of Llama 2 7B and LLAMMAS-BASE on five English benchmarks. Results show a minor average performance drop (-1.2%) for LLAMMAS-BASE, primarily due to MMLU.	48
12. Summary of evaluation tasks, datasets, and their language coverage for the five target Finno-Ugric languages.	50

13. Performance of Llama 2 and Llama 3.1 models across all tasks for five Finno-Ugric languages. Results show that Llama 3.1 models generally outperform Llama 2 models, with the exception of Finnish, where Llama 2 performs comparatively better. The results are obtained with 5-shot prompting. Accuracy is reported for SIB, Belebele, and XCOPA; BLEU is reported for FLORES; and BERTScore F1 is calculated using bert-base-multilingual-cased.	51
14. Comparison of zero-shot results of Llama 2 7B (L2), LLAMMAS-BASE (LB) and Llama 3.1 8B (L3.1).	52
15. Examples from MT-bench.	55
16. BLEU scores from translating Estonian subset of the SMUGRI-MT-bench into Võro, Komi, and Livonian. The results show that state-of-the-art LLMs do not yet achieve adequate performance for these extremely low-resource target languages, whereas dedicated machine translation systems can potentially produce high-quality translations.	58
17. Data collection statistics for human evaluation, including the number of participants and responses for each target language. Fractional values for <i>grades per question</i> indicate that multiple survey versions were used for each target language and they received uneven numbers of responses, resulting in a non-integer average ratings-per-question value.	60

LIST OF ORIGINAL PUBLICATIONS

Publications included in the thesis

- I **Kuulmets, Hele-Andra**, Tättar, Andre, and Fishel, Mark (Oct. 2022). “Estonian Language Understanding: a Case Study on the COPA Task”. In: *Proceedings of Baltic HLT 2022*. Vol. 10. Riga, Latvia: Baltic Journal of Modern Computing, pp. 470–480. DOI: 10.22364/bjmc.2022.10.3.19.
- II **Kuulmets, Hele-Andra** and Fishel, Mark (May 2023). “Translated Benchmarks Can Be Misleading: the Case of Estonian Question Answering”. In: *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*. Tórshavn, Faroe Islands: University of Tartu Library, pp. 710–716. URL: <https://aclanthology.org/2023.nodalida-1.71/>.
- III **Kuulmets, Hele-Andra***, Purason, Taido*, Luhtaru, Agnes, and Fishel, Mark (June 2024). “Teaching Llama a New Language Through Cross-Lingual Knowledge Transfer”. In: *Findings of the Association for Computational Linguistics: NAACL 2024*. Mexico City, Mexico: Association for Computational Linguistics, pp. 3309–3325. DOI: 10.18653/v1/2024.findings-naacl.210.
- IV **Kuulmets, Hele-Andra**, Purason, Taido, and Fishel, Mark (Mar. 2025). “How Well do LLMs know Finno-Ugric Languages? A Systematic Assessment”. In: *Proceedings of the Joint 25th Nordic Conference on Computational Linguistics and 11th Baltic Conference on Human Language Technologies (NoDaLiDa/Baltic-HLT 2025)*. Vol. 57. Tallinn, Estonia: University of Tartu Library, pp. 340–353. URL: <https://aclanthology.org/2025.nodalida-1.37.pdf>.
- V Purason, Taido*, **Kuulmets, Hele-Andra***, and Fishel, Mark (Apr. 2025b). “LLMs for Extremely Low-Resource Finno-Ugric Languages”. In: *Findings of the Association for Computational Linguistics: NAACL 2025*. Albuquerque, New Mexico: Association for Computational Linguistics, pp. 6692–6712. DOI: 10.18653/v1/2025.findings-naacl.373.

Publications not included in the thesis

- VI Bergmanis, Toms, Pinnis, Marcis, Rozis, Roberts, Šlapiņš, Jānis, Šics, Valters, Bernāne, Berta, Pužulis, Guntars, Titomers, Endijs, Tättar, Andre, Purason, Taido, **Kuulmets, Hele-Andra**, Luhtaru, Agnes, Rätsep, Liisa, Tars, Maali, Laumets-Tättar, Annika, and Fishel, Mark (June 2022). “MTee: Open Machine Translation Platform for Estonian Government”. In: *Proceedings of the 23rd Annual Conference of the European Association for Machine Translation*. Ghent, Belgium: European Association for Machine Translation, pp. 309–310. URL: <https://aclanthology.org/2022.eamt-1.44/>.

- VII Tättar, Andre, Purason, Taido, **Kuulmets, Hele-Andra**, Luhtaru, Agnes, Rätsep, Liisa, Tars, Maali, Pinnis, Mārcis, Bergmanis, Toms, and Fishel, Mark (2022). “Open and Competitive Multilingual Neural Machine Translation in Production”. In: *Baltic Journal of Modern Computing* Vol. 10, No. 3, pp. 422–434.

Other published work of the author

- VIII Šuvalov, Hendrik, Lepson, Mihkel, Kukk, Veronika, Malk, Maria, Ilves, Neeme, **Kuulmets, Hele-Andra**, and Kolde, Raivo (Mar. 2025). “Using Synthetic Health Care Data to Leverage Large Language Models for Named Entity Recognition: Development and Validation Study”. In: *Journal of Medical Internet Research* Vol. 27, e66279. DOI: 10.2196/66279.

Author’s contribution to the publications

The author’s contribution to Publications I, II, and IV was as the main author, that is, coming up with the idea, curating the data, designing and conducting the experiments, evaluating the models, analysing and interpreting the results, and writing the manuscript.

The author shares first authorship for Publications III and V. For Publication III, the author contributed by participating in coming up with the idea and designing experiments, conducting most evaluation, analysing and interpreting the results, and writing most of the manuscript. The other main author prepared the data and conducted all the experiments, and also participated in other parts of the work, including experimental design, some evaluation, result interpretation, and writing certain sections of the manuscript.

For Publication V, the author created the multi-turn benchmark and recruited the translators. The author was responsible for all aspects of the human evaluation. Other than that, the author’s contribution to the manuscript was minor. The other main author came up with the idea, designed and carried out all experiments and the automatic evaluation, and wrote most of the manuscript.

1. INTRODUCTION

Recent times have brought rapid progress in natural language processing (NLP), driven by advances in language modeling. These models have evolved from simple generative models that extend input text, to large language models capable of assisting with everyday tasks, and even to agents with the potential to automate complex workflows. Central to this progress is the Transformer architecture (Vaswani et al., 2017)—a type of neural network that has led to major improvements in capturing the meaning of natural language. While model architecture has played a crucial role in the progress, the availability of large datasets that are needed to teach models both human language and the knowledge encoded in written texts has been equally crucial. Yet, it is the scarcity of high-quality training data that has become one of the main obstacles to build even larger (and potentially smarter) language models.

The vast majority of available training data is in English. Consequently, most NLP research and development efforts have focused on English, resulting in much stronger models for English than for other languages. Even if there is interest in replicating the same level of progress in other languages, it is much harder to reach that as many of those languages suffer from significant resource scarcity which limits the achievable performance.

Fortunately, low-resource languages are not without options. Specifically, these languages can benefit from **cross-lingual knowledge transfer**, which happens when a model is trained on multiple languages at once. As a result, the model develops universal inner representations that allows knowledge to be transferred from high-resource languages (like English) to lower-resource ones (like Estonian). Thanks to this effect, a model that is trained partially on English can potentially improve its performance on other languages without requiring vast amounts of data for each language individually.

Estonian is one of such language that is considered low-resource in the context of language modeling. That means that it lacks sufficient data to train a language model from scratch, but it has enough resources to support in-depth study of cross-lingual knowledge transfer. The largest Estonian text corpus, the Estonian National Corpus (Koppel et al., 2023), prior to any cleaning or filtering, contains around 3.8 billion words. While this is only a fraction of what is used to train modern large-scale models, it can still be included in a larger multilingual dataset to help improve a model’s understanding of Estonian. In addition, several machine translation systems can translate between English and Estonian with reasonable quality, including Neurotõlge (Tättar et al., 2022; Yankovskaya et al., 2023), a neural machine translation system developed at the University of Tartu that currently supports 26 Finno-Ugric languages, many of which are low- or extremely low-resource. Such systems enable the creation of synthetic Estonian training data via machine translation at substantially lower cost than human annotation, facilitating the creation of fine-tuning datasets.

Research Questions

Advances in language modeling have not only made language models better in reasoning and richer in world knowledge, but have also significantly improved their ability to understand and generate non-English languages. The fact that these gains are evident even in low-resource languages such as Estonian, Latvian, and Icelandic (Holtermann et al., 2024; OpenAI et al., 2024) suggests that at least *some* of this progress can be attributed to improved cross-lingual alignment within the models. However, the understanding of how cross-lingual alignment emerges and manifests across diverse data settings remain insufficiently studied for many languages.

This thesis contributes to research on cross-lingual transfer learning with a particular focus on Estonian, while the methods explored are language-agnostic and applicable to other languages with similar resource profiles. Specifically, we investigate how to best leverage the modest amount of raw Estonian textual data available, together with different types of synthetic Estonian datasets, to enhance knowledge transfer from English.

We study cross-lingual transfer in both the pre-training and fine-tuning stages of language model training. First, we examine how cross-lingual knowledge transfer can be strengthened during the computationally less expensive fine-tuning phase. Our experimental setup assumes a constrained scenario in which no native, task-specific fine-tuning data in Estonian is available. We therefore rely on synthetic Estonian data, obtained either through machine translation or generated by stronger language models. These types of data are attractive because they are extremely inexpensive to produce, which helps to address data scarcity in many cases. However, their quality may be lower, potentially limiting model performance. Therefore, whenever possible, we compare synthetic-data-based approaches with alternatives that rely on higher-quality data to better understand the limitations of fine-tuning on synthetic data.

Second, we investigate cross-lingual transfer during the pre-training phase. Specifically, we study whether knowledge transfer can be improved through continued pre-training, that is, by taking an existing language model trained on other language(s) and further pre-training it on data with an increased proportion of Estonian. We are particularly interested in whether a small amount of additional data would already be sufficient to improve cross-lingual alignment to Estonian from other training languages, especially from English. In addition, we study the role of scale, considering both the scale of the original pre-training and the scale of continued pre-training.

Finally, we compare the cross-lingual transfer learning methods, that essentially rely on multilingual training, to monolingual approaches in understanding Estonian. We compare these approaches at different stages of training to identify the most advantageous setup for every stage.

Our experiments cannot be carried out without reliable model evaluation; how-

ever, Estonian lacks high-quality evaluation datasets. As collecting and annotating native datasets is expensive, synthetic translated datasets are often used as a substitute. As part of this work, we examine the potential implications of using synthetic data for evaluating language models.

Another challenge is evaluating conversational language models. Since evaluating natural language generation with automatic metrics is difficult, human evaluation is necessary. In low-resource languages, however, the limited availability of domain-expert human annotators makes reliable evaluation of conversational models particularly difficult. As part of this work, we identify key challenges in evaluating conversational models in low-resource settings and use these insights to guide the design of a conversational evaluation dataset, which we then use to conduct a human evaluation of language models.

Our research questions are as follows:

- To what extent can synthetic data be used in adapting language models to Estonian?
- How does multilingual training compare to monolingual training for Estonian language understanding across training stages?
- How does (continued) pre-training contribute to the establishment of cross-lingual knowledge transfer?
- How should native datasets and human evaluation be positioned relative to synthetic data and automatic metrics in the evaluation of language models for low-resource languages?

Contributions

This thesis consists of five main contributions, each corresponding to a publication included in the thesis. The contributions are summarized as follows.

Contribution I. We compare different fine-tuning and prompting strategies for strengthening cross-lingual knowledge transfer to Estonian in encoder-only and decoder-only models. By using an Estonian commonsense reasoning task as the evaluation setting, we show that multilingual fine-tuning and multilingual prompting of multilingual base models are superior to monolingual methods.

Contribution II. We compare translated and native test datasets in TRANSLATE-TEST settings, where Estonian inputs are translated into English at inference time and the outputs are translated back into Estonian. The results suggest that evaluation on translated test datasets may overestimate model performance.

Contribution III. We conduct a series of continued pre-training experiments on the English-centric decoder-only Llama 2 model to strengthen its ability to understand and generate Estonian. The results show that even a relatively small amount of continued pre-training data can substantially improve performance in the target language. In addition, by leveraging stronger language models for synthetic Estonian data generation, we introduce the first instruction-following model for

Estonian.

Contribution IV. Using available benchmark datasets, we evaluate a set of open language models on five Finno-Ugric languages, including Estonian, and provide an overview of their performance. We observe that massively scaling pre-training data improves performance in these languages, outperforming language-specific adaptation of models trained with smaller data scales.

Contribution V. We formulate a set of requirements for conversational benchmarks in low-resource settings. Based on these requirements, we introduce a new multi-turn conversational benchmark covering four Finno-Ugric languages and use it for human evaluation of language models adapted to these languages.

Structure of the Thesis

This thesis is organized into six chapters. Chapter 1, the chapter that you are currently reading, is an introduction to this thesis. It describes the motivation, formulates the research questions, and briefly described the contributions and structure of the work. Chapter 2 provides background and context for the work, while Chapters 3–5 present the core contributions and Chapter 6 concludes the thesis. During the course of the PhD research, cross-lingual transfer learning underwent a paradigm shift driven by the emergence of large language models (LLMs), which by default exhibit substantially improved performance in non-English language understanding. Chapter 3 focuses on methods and experiments conducted prior to the introduction of LLMs and bases on contributions **I** and **II**. Chapter 4 examines how to strengthen Estonian language understanding and generation capabilities in LLMs and investigates the role of scale in this process. It is based on contributions **III** and **IV**. Chapter 5 introduces a conversational benchmark for four low- or extremely low-resource Finno-Ugric languages and demonstrates its practical value. It is based on contribution **V**. Chapter 6 concludes the dissertation by summarizing the main findings.

The detailed structure of Chapters 2-6 is as follows:

- **Chapter 2.** Sections 2.1 and 2.2 provide a brief overview of how the development of the Transformer architecture enabled the rise of large language models. Section 2.3 introduces cross-lingual transfer learning, outlines common experimental setups, and discusses typical baselines. Section 2.4 focuses on evaluation, covering methods for creating synthetic evaluation datasets and approaches for evaluating instruction-following models.
- **Chapter 3.** Section 3.1 introduces a new Estonian commonsense dataset that will be used in subsequent transfer learning experiments. Section 3.2 presents the experiments, which compare different types of synthetic training data under various training setups, followed by a comparison of synthetic test datasets. Section 3.3 addresses a common but often overlooked problem in evaluating cross-lingual transfer learning. The chapter concludes with a short summary (Section 3.4).

- **Chapter 4.** Section 4.1 presents experiments aimed at eliciting knowledge transfer from English to Estonian in large language models. These experiments separately evaluate the effects of continued pretraining, synthetic Estonian instruction-tuning data, and parallel data on Estonian natural language understanding and generation. Section 4.2 provides a broader overview of the state of open LLMs in understanding Finno-Ugric languages, contextualizing the results presented in Section 4.1.
- **Chapter 5.** Section 5.1 formulates requirements for conversational benchmarks in low-resource languages. Section 5.2 describes the development of the novel benchmark, and Section 5.3 presents the design and results of a human evaluation of language models.
- **Chapter 6.** The final chapter synthesizes the key findings and outlines future research directions.

2. BACKGROUND

2.1. Fundamentals of NLP

The first step of any application processing text is to convert the text into a format that can be processed by a mathematical model. This is done by splitting the text into a predefined set of chunks, i.e subwords or tokens. Several different algorithms have been introduced to perform subword tokenization, the most common being Byte-Pair Encoding (Sennrich et al., 2016), WordPiece (Wu et al., 2016), and SentencePiece (Kudo and Richardson, 2018).

The tokenization process is followed by a data transformation step, during which the tokens are processed by a deep neural network to produce the desired output. This begins with mapping the tokens to learned continuous vectors, followed by their transformation through various layers of the deep neural network. The specific architecture of this neural network can vary greatly based on the particular task at hand. However, the field of natural language processing (NLP) nowadays generally recognizes two primary model architectures: **encoders** and **decoders**.

In current practice, both encoders and decoders are usually implemented using the transformer model architecture (Vaswani et al., 2017). It has become by far the most dominant neural network type used in NLP because it utilizes the computational hardware more effectively than the other types of neural networks. This efficiency enables the training of larger **language models** on bigger datasets, resulting in improved natural language understanding abilities of the models (Liu et al., 2019; Brown et al., 2020; Shoybi et al., 2020; Hoffmann et al., 2022).

Transformer-based language models are typically trained in two stages. First, to make use of the vast amounts of unlabeled text, the models are **pre-trained** on massive text corpora using an unsupervised training objective, i.e. predicting a masked token. This helps the model to learn general language patterns, e.g. syntactic structure and semantics of the words. However, the model still would not know how to solve a specific task. Thus, the models are then **fine-tuned** on the desired end task, typically using a small labelled dataset.

2.1.1. The Transformer Architecture

The original Transformer was designed to solve a machine translation task, which involves converting an input sequence of tokens into an output sequence of tokens. Following the typical architecture of a translation model, it consists of an encoder and a decoder, where the encoder's output serves as the input to the decoder. However, the Transformer encoder soon became popular as a standalone model, as it was better suited for sentence-level tasks. The decoder, initially offering no advantage over the encoder-decoder architecture when used alone, gained research attention after it was discovered that, with sufficient scaling, such models could solve tasks on which they had not been explicitly trained (Brown et al., 2020).

Both the Transformer encoder and decoder consist of a stack of N identical Transformer blocks. Each block takes as input the output of the previous block. For the first block, the input is the sum of token embeddings (W_e) and positional encodings (W_p):

$$h_0 = UW_e + W_p$$

$$h_i = \text{transformer_block}(h_{i-1}) \forall i \in [1, N]$$

In decoder-based models, the input $U = (u_{m-k}, \dots, u_m)$ is a sequence of tokens. The output of the final Transformer block is used to predict the probability distribution of the next token u_{m+1} :

$$P(u_{m+1}) = \text{softmax}(h_N^m W_e^T)$$

In encoder-based models, the input $U = (u_1, \dots, u_m)$ is a sequence in which some token u_j is replaced with a special [MASK] token. The output of the final Transformer block is then used to predict the probability distribution of the masked token:

$$P(u_j) = \text{softmax}(h_N^j W_e^T)$$

The novelty of the Transformer architecture—and a key reason for its improved performance in natural language understanding—is its exclusive reliance on the self-attention mechanism, eliminating the need for RNNs or convolutions. While the idea of self-attention was not new (e.g., Lin et al. (2017)), Vaswani et al. (2017) was the first to implement a sequence-to-sequence neural model without RNNs or convolutions, relying entirely on self-attention.

Each `transformer_block` hence includes a self-attention function, defined as:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

where Q , K , and V are linear projections of the output from the previous layer.

Essentially, the self-attention function is a mechanism that relates each token to all other tokens in a sequence to compute a representation for each token. This means the token representations are contextualized, depending on the other tokens in a sequence, helping to resolve semantic ambiguity.

In practice, each `transformer_block` proposed by Vaswani et al. (2017) uses `MultiHeadAttention`, which computes multiple independent sequence representations in parallel. Compared to RNNs, this allows the model to capture richer token-level information with significantly lower computation time through parallel execution. Additionally, each `transformer_block` employs residual connections to prevent vanishing gradients.

2.1.2. Transformer Encoders and Decoders

Encoders are models that map an input sequence of vectors (token embeddings) to a single fixed-length vector or to a sequence of vectors of the same length as the input. The latter is, for example, the output format of pre-trained transformer-based encoders, first introduced by Devlin et al. (2019), who referred to this model architecture as BERT, short for Bidirectional Encoder Representations from Transformers.

As the name suggests, these models are bidirectional, meaning that in order to create the final embedding the model is allowed to use information on both the left and right sides of the input text. During pre-training, the token to be predicted is masked using a special [MASK] token to prevent the model from seeing the correct output token.

Encoders are typically used for sentence and sequence classification tasks such as named entity recognition or textual entailment.

Decoders, by contrast, map an input sequence of tokens of arbitrary length to an output sequence of tokens of arbitrary length. Transformer-based decoders were first introduced by Vaswani et al. (2017).

These models are unidirectional, meaning that the model is only allowed to see tokens to the left of the current position when predicting the next token. Decoders are autoregressive, meaning that they produce one token at a time, using the previously generated tokens as context. Thus, the output text of arbitrary length is achieved by performing the next token prediction several times, each time incorporating the previously predicted tokens.

Decoders are used for open-ended text generation tasks such as machine translation or text summarization where the output sequence length is not fixed in advance.

Note that models that follow the architecture of a decoder or an encoder are commonly referred to as decoder-only and encoder-only models, respectively. In contrast, the original Transformer architecture proposed by Vaswani et al. (2017) combines both components into an encoder–decoder model.

2.1.3. Pre-Training

The advancements that the Transformer architecture has brought to NLP largely come from the fact that it is better at leveraging raw, unlabeled data. As task-specific labeled datasets are typically small and costly to create, models that can effectively learn from unlabeled data provide a cheaper alternative to the expensive manual creation of large labeled datasets.

Training the model on raw unlabeled data is called pre-training. Pre-training is the first phase of training language models during which the model learns the initial parameters using some unsupervised optimization objective (masked token prediction or next token prediction; see Section 2.1.2 for further details). Pre-training is typically done with whatever document-level data is available and does

not have to be from the same domain as the end task. On the contrary, diverse data is preferable as it prevents overfitting and leads to more generally applicable base model.

It has been shown (Brown et al., 2020; Kaplan et al., 2020; Hoffmann et al., 2022) that massively scaling model parameters and the number of pre-training tokens results in stronger base models. However, this makes this phase computationally very expensive (not to mention the efforts it takes to curate a high-quality pre-training dataset) and thus has become unavailable to many with more restricted budgets.

2.1.4. Fine-Tuning

The ability of language models to gain language understanding from raw, unlabeled data during the pre-training has significantly reduced the need to create large-scale annotated datasets for effective task-solving, while at the same increasing accuracy in solving the end task (Radford et al., 2019; Brown et al., 2020).

Training the model on an annotated dataset using a task-specific optimization objective is called fine-tuning. To adapt pre-trained encoders and decoders to a specific problem, a special layer is added on top of the base model. For example, to use an encoder-type model for sentence classification, a classification layer is simply added on top of a specific output token representation. Concretely, that specific output token representation is the embedding corresponding to a special input token [CLS], which is specifically added to the input for sentence classification tasks during pre-training. To solve a sentence classification task, this embedding is projected to a vector of length K where K is the number of classes, and where the parameters are optimized using the standard classification loss, log-softmax.

Another type of tasks that encoder-type models are well-suited for are token classification tasks such as POS tagging or NER, but also more unconventional tasks such as extractive question answering (QA), where the goal is to extract the excerpt from the given paragraph that answers the given question. For example, to solve an extractive QA task, one can fine-tune the model to predict, for each output embedding, whether the corresponding token starts the answer, ends the answer, or is neither (Devlin et al., 2019). The score for the i -th token being the beginning of an answer is computed as the dot product between the i -th embedding and a trainable vector \mathbf{S} . A similar calculation is performed to find the ending token of the span, using a trainable vector \mathbf{E} , and the predicted answer span is the one that maximizes

$$\mathbf{S} \cdot T_i + \mathbf{E} \cdot T_j, \quad \text{where } i \leq j.$$

During fine-tuning, the scores are softmaxed over all tokens, and the training objective is the sum of the log-likelihoods of the correct start and end tokens.

When adapting a model to a specific task, one can either update all parameters,

including those of the base model, or update only a subset, such as task-specific layers, while freezing the rest.

Decoders can also be fine-tuned, similarly to encoders, for example by projecting the last hidden embedding of the decoder to a vector of class scores. However, in practice, decoders are rarely fine-tuned on sentence-level tasks, as bidirectional encoder representations are generally more effective in such cases. A more typical use case for fine-tuning decoders would be tasks that require generating text of arbitrary length, such as machine translation or text summarization. However, such single-task setups have traditionally been addressed with encoder-decoder models. Fine-tuning decoder-only models, on the other hand, has been more focused on improving generalization to unseen tasks (see Section 2.2.2).

2.2. Large Language Models

Experiments with scaling both the number of parameters and the amount of pre-training data led to the observation that sufficiently large language models can learn to perform various downstream tasks without explicit supervision (Radford et al., 2019). In particular, it was found that by conditioning a pre-trained decoder-type model on natural language instructions, such as concatenating a task description (e.g., "translate to French") with a source sentence in English, the model is able to produce the correct output, such as generating the corresponding translation in French. This phenomenon, later termed **in-context learning** (Brown et al., 2020) was shown to continue improving as the model scale increased, with performance continuing to rise even when models became 10 times larger than any previously trained systems (Brown et al., 2020). It was at this stage of language modeling research that the term **large language model (LLM)** was introduced, to highlight capabilities that had not been observed in weaker language models.

2.2.1. Model Scaling

The ability to solve downstream tasks without explicit supervision had already been demonstrated by the 1.5 billion-parameter GPT-2 model (Radford et al., 2019), which was trained on 8 million documents totaling 40GB of text. For example, GPT-2 achieved non-trivial performance in French-to-English translation (BLEU 11.5) despite having only 10MB of French text in its training data.¹ Its successor, the 175B-parameter GPT-3 model (Brown et al., 2020), trained on 300 billion tokens, notably achieved zero-shot BLEU scores of 25.2 for English-to-French and 21.2 for French-to-English on the same WMT'14 dataset, despite only 7% of its training data consisting of non-English text.

Based on the scaling law analysis by Kaplan et al. (2020), which suggested a power-law relationship between model size and performance, most subsequent ef-

¹That said, it could not translate in the reverse direction, English-to-French, scoring a much lower BLEU of 5.

forts to improve language models focused on increasing model size when scaling computing power. This led to the development of extremely large models, such as Gopher (280B) (Rae et al., 2022), PaLM (540B) (Chowdhery et al., 2022), LaMDA (137B) (Thoppilan et al., 2022), and MT-NLG (530B) (Smith et al., 2022), while the number of training tokens remained constant at around 300 billion tokens (Hoffmann et al., 2022).

The question of training a compute-optimal model was revisited by Hoffmann et al. (2022) whose analysis led to a different conclusion from what was reported by Kaplan et al. (2020). Specifically the analysis by Hoffmann et al. (2022) predicted that when increasing the compute budget, training data size should also scale, and specifically it should scale at roughly the same rate as model size. They validated this hypothesis by training Chinchilla, a 70B-parameter model using four times as much data than previous models, which turned out significantly outperform much larger models across a wide range of downstream evaluation tasks.

Following these findings, new releases of language models were smaller in terms of parameters. More effort now put into scaling the pre-training data, and, as a side effect of reducing parameter size and thus hardware requirements, they became accessible to a larger audience for inference, increasing the interest in such models. Some of the early examples of models following the new training principles are Falcon models (with 7B to 180B parameters) trained on 3.5T tokens (Almazrouei et al., 2023), Qwen models (7B to 14B) trained on up to 3T tokens (Bai et al., 2023), LLaMA models (from 7B to 65B) trained on 1.4T tokens, (Touvron et al., 2023a) and their successors, Llama 2 models (7B to 70B) trained on 2T tokens (Touvron et al., 2023b).

2.2.2. Instruction Tuning

Although sufficiently pre-trained language models such as GPT-3 (Brown et al., 2020) showed strong performance on a wide range of tasks given only a few input-output examples, their results still did not match those of the best task-specific models. Moreover, in the zero-shot setting, performance remained fragile, with output quality varying considerably across tasks. Thus, in parallel with exploring model scaling, substantial research attention shifted towards improving the robustness of language models on diverse NLP tasks under both zero-shot and few-shot conditions.

To this end, several studies demonstrated that further training of base models on a mixture of NLP tasks—each accompanied by natural language instructions specifying the desired output—substantially improves generalization to unseen tasks (Mishra et al., 2022; Sanh et al., 2022; Wei et al., 2022). Further, Ouyang et al. (2022) fine-tuned GPT-3 on human-written queries and answer demonstrations, further aligning the model with user preferences through reinforcement learning from human feedback (RLHF). The authors observe that this model was significantly preferred over versions fine-tuned solely on instruction-augmented NLP

datasets.

This line of fine-tuning methodologies, referred to as **instruction tuning**, has since played a crucial role in enhancing language models’ ability to understand and follow user prompts, ultimately transforming them into conversational assistants capable of addressing a various user needs.

One of the key requirements for aligning model outputs with user expectations is the availability of high-quality instruction-tuning data that reflects real-life usage. To reduce the cost of creating such data, which traditionally requires human annotators, research has explored cheaper alternatives. Wang et al. (2023) demonstrated that generating instructions, inputs, and outputs from a language model and using this synthetic data to instruction-tune a base model can achieve performance on par with InstructGPT (Ouyang et al., 2022), but at a significantly lower cost. Building on this idea, Taori et al. (2023) generated 52K instruction-following demonstrations using `text-davinci-003` (a version of InstructGPT) and used this data to fine-tune LLaMA 7B. According to their preliminary evaluations, the resulting Alpaca model performed comparably to `text-davinci-003` despite being substantially smaller.

2.3. Cross-Lingual Transfer Learning

Cross-lingual transfer learning refers to a type of machine learning that leverages large datasets available in high-resource languages to improve language understanding in lower-resource languages. Since several very different methods that fit under this definition (see more at Section 2.3.2), in this work, we narrow the definition to methods that establish cross-lingual transfer through universal language-independent representations (Lample and Conneau, 2019). These representations are learned by training a (transformer-based) language model on multiple languages. This way, the model has to learn to project inputs in different languages into a shared embedding space, forcing it to create abstractions for language structures and semantics that are similar across languages.

Before the emergence of large language models, the primary objective in cross-lingual transfer learning research was to achieve **zero-shot model transfer**, that is, to develop a model capable of performing a task in a target language without relying on task-specific annotated data in that language, but only on shared multilingual representations (Pires et al., 2019). This line of research attracted broad interest (Conneau et al., 2020; Liu et al., 2020; Xue et al., 2021) and led to the development of numerous cross-lingual evaluation datasets (Conneau et al., 2018; Artetxe et al., 2020c; Ponti et al., 2020), in which training data is available only in English, while test data covers various target languages.

With the emergence of LLMs, both the goals and methods of cross-lingual transfer have changed. The focus is no longer on transferring a single task but rather on transferring the entire set of capabilities acquired during pretraining. However, training strong LLMs from scratch remains inaccessible to most re-

searchers. As a result, much of the current work in cross-lingual transfer learning centers on enhancing the multilingual capabilities of existing LLMs, which are predominantly trained on English data. This is typically pursued by continued pre-training on a corpus with increased representation of the target language, incorporating the target language during instruction tuning, or combining both strategies (Csaki et al., 2024; Cui et al., 2024; Etxaniz et al., 2024a; Rodríguez et al., 2025). Consequently, zero-shot model transfer is no longer considered a goal on its own. Instead, the focus has shifted to more flexible methods that may still rely on (potentially limited) target language data.

2.3.1. Multilingual Pre-Training

The foundation of cross-lingual knowledge transfer lies in pre-training on multiple languages. Prior to the dominance of large language models, multilingual pre-training deliberately included a high proportion of non-English data to induce multilingual capabilities. However, as noted by Conneau et al. (2020), increasing the number of languages for a fixed-size model reduces the per-language capacity, a phenomenon they refer to as the *curse of multilinguality*. This trade-off lowers the overall performance, though low-resource languages still benefit from positive transfer from typologically related languages. As a result, multilingual models tended to deliver state-of-the-art performance for medium- and low-resource languages, while high-resource languages were better served by language-specific models.

Model scaling (see Section 2.2.1) has since shifted this dynamic. Larger models demonstrate improved performance on non-English tasks despite being trained on only a small fraction of non-English data. This suggests that scale enhances a model’s ability to learn useful representations even from limited data in a given language (Shi et al., 2023; Holtermann et al., 2024). This insight has redefined the requirements for supporting low-resource languages — while more data remains beneficial, significant improvements can be achieved even with relatively small datasets (e.g., Etxaniz et al., 2024b).

2.3.2. Zero-Shot Model Transfer

More formally, in zero-shot model transfer, a model is fine-tuned on task data in a subset of languages and is expected to perform the same task in a target language that was not included in the training subset. This approach requires extensive multilingual pre-training to enable the model to learn shared representations across languages. The multilingual model is then fine-tuned on task-specific annotations in one or more source languages (typically English) and ideally, the learned skill is expected to transfer to the target language through the shared multilingual embedding space established during pre-training.

The motivation for zero-shot model transfer is the extremely high cost of creating task-specific datasets for every target language. Additionally, many languages

lack sufficient pre-training data to train a separate base model for that language. Finally, in multilingual scenarios, using a single model for all languages helps reduce both training and inference costs.

In addition to zero-shot transfer, there exist two other methods to tackle a task in the target language without requiring task-specific annotations in that language: TRANSLATE-TEST and TRANSLATE-TRAIN. Both rely on the availability of a machine translation system capable of translating to and from the target language. In the TRANSLATE-TEST approach, the test data is translated during inference from the target language into English, after which a model fine-tuned on English task-specific data is used to generate predictions that, depending on the task, might need to be translated back into the target language. In contrast, the TRANSLATE-TRAIN method involves translating the English training data into the target language and then fine-tuning a multilingual or monolingual base model on this translated data. These two approaches are often employed as baselines for ZERO-SHOT transfer, with TRANSLATE-TEST in particular being a difficult baseline to outperform, as it allows the reuse of well-performing English models.

Note that at the introduction of this section, we stated that cross-lingual transfer methods leverage large datasets in one language to improve performance in a lower-resourced target language. By this definition, both TRANSLATE-TEST and TRANSLATE-TRAIN can be considered and are sometimes referred to as cross-lingual transfer methods. However, if we define cross-lingual transfer learning more narrowly as learning a shared representation space within the model, then these methods do not fall under that category.

That said, both TRANSLATE-TEST and TRANSLATE-TRAIN can still benefit from shared representations. For example, TRANSLATE-TRAIN is typically more effective when used with multilingual language models that can leverage these shared spaces. Likewise, TRANSLATE-TEST can also be used with multilingual models, although it is less likely to yield improvements, since English language understanding tends not to benefit significantly from multilingual representations (Conneau et al., 2020).

2.3.3. Multilingual Fine-Tuning

Achieving zero-shot model transfer is desirable in scenarios where fine-tuning data in the target language is expensive to obtain. However, this motivation is weakened in the era of LLMs, as creating fine-tuning data has become substantially cheaper for many languages. As a result, zero-shot model transfer is often no longer a goal in itself.

One specific reason for the reduced interest in zero-shot model transfer is that LLMs can be used to generate synthetic fine-tuning data across a wide range of languages, increasing the likelihood that at least *some* fine-tuning data in the target language is available. In addition, when instruction tuning is desired — as it often is — existing task-specific NLP datasets can be reformulated as instruction-tuning

data (see Section 2.2.2). Together, these factors make multilingual fine-tuning more readily available than before the emergence of LLMs.

A further consideration is that most pre-trained LLMs achieving state-of-the-art performance are predominantly trained on English. Consequently, fine-tuning on target data is often necessary to introduce or strengthen capabilities in the target language. Moreover, it is computationally significantly cheaper than pre-training the model from scratch.

2.3.4. English as a Pivot

Many lesser-resourced languages lack sufficient data to train strong models. Yet, it has been shown that LLMs can still exhibit understanding of such languages (Holtermann et al., 2024), suggesting effective knowledge transfer from high-resource languages during pre-training. Although the exact compositions of pre-training data for these models is often undisclosed, prior work suggests that improved multilingual performance may arise from a more effective integration of lesser-resourced languages into the training corpus.

For example, Briakou et al. (2023) argue that the surprisingly strong translation performance of the PaLM model can be attributed to the unintentional inclusion of bilingual signals, including translation examples, in its pre-training data. Their analysis shows that 1.4% of PaLM’s training instances are bilingual, while 0.34% contain at least one translated sentence pair, amounting to approximately 30 million unintended translation examples. To support this claim, they train several scaled-down versions of PaLM using different training data configurations. They observe that ablating bilingual and translation instances from the training data has a clear negative effect to the translation capabilities.

In contrast, PaLM 2 (Anil et al., 2023) intentionally includes parallel data during pretraining. While its impact is not directly measured, the model achieves 77.4% accuracy on an Estonian subset of the commonsense reasoning benchmark XCOPA in a few-shot setting, and 91.0% when using chain-of-thought prompting where intermediate steps involve *translating* the input into English to perform reasoning in English (Shi et al., 2023).

Cross-lingual transfer specifically *from English* has received growing attention. Zhang et al. (2023) observe that ChatGPT exhibits a mixture of coordinate and subordinate multilingualism, suggesting that it either maintains concept representations separately across languages or internally translates the input into English before connecting it to the shared space. Wendler et al. (2024) apply mechanistic interpretability methods to Llama 2 (Touvron et al., 2023b) and find that token embeddings in the middle layers are closer to English representations, only shifting toward the target language in the final layers, which they interpret as evidence that the model’s “concept space” is anchored in English.

2.4. Evaluation

Evaluating and comparing language models has become increasingly challenging as their capabilities now extend beyond the tasks they were explicitly trained on (Mishra et al., 2022; Sanh et al., 2022). This challenge is even more pronounced for lesser-resourced languages, which have significantly benefited from stronger cross-lingual knowledge transfer arising from scaling model parameters and training data. However, as evaluation still depends heavily on human input, whether for curating, validating, or annotating data, it has not been able to keep pace with the rapid progress in model performance in these languages. Moreover, as the manual effort required for evaluation is very costly, it may not even be available for certain low-resource languages.

Another challenge lies in evaluating instruction-tuned models, which must generate adequate response to a wide range of user inputs. Beyond simply following user instructions, these models are expected to solve diverse problems, reason about the world, and maintain factual accuracy to be genuinely useful. There are all capabilities that are difficult to evaluate with automatic metrics. As a result, human annotators remain unavoidable.

A recent trend in model evaluation is to use stronger conversational language models to assess the output of weaker ones. This approach, known as *LLM-as-a-Judge*, offers a cost-effective alternative to human annotation, but it also has its limitations (further discussed in Section 2.4.3).

2.4.1. Creating Evaluation Datasets

Several methods can be used to create evaluation datasets for new languages, which is often necessary to assess cross-lingual transfer to lesser-resourced languages that have fewer evaluation datasets available. A popular approach is to translate datasets originally created for English, either fully or partially, into the target language. Unlike training datasets, evaluation datasets are usually translated manually by professional translators to ensure high data quality. Alternatively, data can be translated using a machine translation (MT) followed by optional post-editing. However, these approaches are less common for lesser-resourced languages, where MT systems lack sufficient quality. Moreover, MT systems often struggle with multi-sentence and document-level data, resulting in inconsistent translations.

A well-known issue with all translated datasets, including those translated by humans, is the presence of *translationese* (Koppel and Ornan, 2011; Artetxe et al., 2020a), that is, language artifacts that differ from natural, real-life data. Another problem is that translated datasets often contain source-language-specific content unfamiliar to native speakers of the target language, further contributing to the difference between the dataset and real-world data. Despite these drawbacks, translated datasets provide a straightforward way to compare model performance across different languages.

An alternative to translation is to create datasets from scratch in the target language. While this approach is more expensive, it eliminates problems related to *translationese* and unfamiliar content. However, building datasets from scratch also makes comparing models across languages more difficult—unlike translated benchmarks, newly created datasets usually do not come with parallel instances in other languages. Another problem of creating datasets from scratch arises for low-resource languages. For such languages, it is more difficult to find human annotators. Moreover, as the capabilities of language models get better, more challenging evaluation benchmarks must be created. However, constructing such benchmarks requires often expert-level knowledge, raising the bar for human annotators. For many low-resource languages, finding annotators with the required expertise can be prohibitively challenging.

2.4.2. Evaluating Instruction-Following Language Models

Instruction tuning allows language models to handle arbitrary tasks and problems through natural language input and output, effectively turning them into conversational assistants that users can interact with. However, evaluating these models with traditional task-specific NLP benchmarks gives only a limited overview of their capabilities and does not capture their conversational and instruction-following skills. Although benchmarks exist for automatically evaluating instruction-following capabilities (Zhou et al., 2023b), ultimately, instruction-tuned models must provide useful and accurate answers to user queries, and evaluation should reflect this as well.

In practice, the usefulness and accuracy of answers are measured by *helpfulness* (e.g., Touvron et al., 2023b). To do this, human evaluators are presented with a pre-written prompt and responses from two different anonymous models, and must choose which answer is more helpful. While pairwise comparison provides valuable insights into how models compare according to actual human preferences, it has some disadvantages. For example, the number of pairs to be evaluated increases rapidly with every new model introduced. Another issue is that a model preferred over another might not be objectively good, although this risk can be mitigated by choosing a strong baseline model for comparison.

A sufficiently large set of pairwise comparisons allows the creation of leaderboards where models are ranked according to some rating system. One notable example is LMArena (Chiang et al., 2024), where votes are crowdsourced via an online platform that invites users to submit their own prompts to randomly paired models. The platform generates traffic by regularly updating the list of models and displaying live leaderboards based on user votes and model performance on various task-specific prompts such as math, coding, and creative writing.

2.4.3. *LLM-as-a-Judge*

A recent trend in language model evaluation is to use stronger models to assess the outputs of weaker ones. This method, known as *LLM-as-a-Judge* (Zheng et al., 2023), has gained popularity as a cost-effective alternative to human annotators, who are needed to evaluate conversational models.

However, LLM-based judges have been shown to exhibit various biases that make them fragile evaluators. Such biases include, for example, position bias, where the model favors answers at certain position over others (Zheng et al., 2023); verbosity bias, where the model favors longer answers even when they are less accurate (Zheng et al., 2023); self-enhancement bias where the model favors the answer generated by itself (Zheng et al., 2023; Panickssery et al., 2024); authority bias where the models assign greater credibility to statements associated with authority (Chen et al., 2024); fallacy-oversight bias, where the model ignores logical errors in reasoning steps (Chen et al., 2024); compassion-fade bias where the model prefers well-known model’s name over the other (Koo et al., 2024); bandwagon-effect bias, where the model gives stronger preference to the majority’s belief (Koo et al., 2024) and distraction-bias where irrelevant details influence judgements (Koo et al., 2024).

Beyond biases, it is necessary to verify that *LLM-as-a-Judge* correlates well with human evaluations, before it can be reliably used in practice. While the strongest models have demonstrated high correlation with human judgments in English pairwise comparisons (Zheng et al., 2023), they show substantial variability in correlating with human evaluators when studied across a broader range of tasks (Bavaresco et al., 2025).

For non-English languages, *LLMs-as-Judges* have not yet been extensively studied. The few existing works in this area suggest that LLM-based judges struggle with consistency in its judgements across languages with particularly poor performance in low-resource settings (Fu and Liu, 2025). Moreover, even the strongest LLMs require calibration with human annotators before they can be reliably used as judges for non-English languages (Hada et al., 2024).

3. CROSS-LINGUAL TRANSFER LEARNING FOR ESTONIAN NLU

In this chapter we examine the strategies and challenges associated with cross-lingual knowledge transfer, with a focus on advancing Estonian natural language understanding (NLU). It is further divided into three sections.

In Section 3.1, we introduce the motivation behind creating EstCOPA, an Estonian commonsense reasoning benchmark. The dataset is constructed by translating the commonsense reasoning dataset COPA from English to Estonian using two different translation methods: machine translation and post-editing. The section concludes with a comparison between EstCOPA and the closely related dataset XCOPA.

In Section 3.2, we conduct a series of experiments using the training sets of the newly created EstCOPA dataset. We approach the problem by examining cross-lingual transfer over different phases of *training*. This is different from a typical comparison of ZERO-SHOT, TRANSLATE-TRAIN and TRANSLATE-TEST, which do not necessarily rely on multilingual training. Specifically, we compare a setup in which transfer occurs only during pretraining to one where it also occurs during fine-tuning. In other words, we investigate whether multilingual fine-tuning on top of multilingual pretraining has any additional effect on knowledge transfer. We find that it can significantly improve target task performance, but only if the underlying base model is sufficiently strong.

In Section 3.3, we focus on the evaluation of cross-lingual transfer methods. Specifically, we investigate the common practice of using translated test datasets for evaluating transfer learning strategies, especially the TRANSLATE-TEST strategy. We find that using such datasets might lead to an overestimation of the model’s performance on native datasets, exemplified by the results on the Estonian question-answering task.

3.1. EstCOPA: An Estonian Commonsense Reasoning Benchmark

Transformer-based language models have redefined the expected capabilities of systems performing natural language understanding. As a result, the need for stronger benchmarks to measure understanding of complex and nuanced aspects of language has become evident. Truly understanding natural language requires more than surface-level processing—it involves the ability to reason about various aspects of the world. One frequent everyday reasoning task is determining a causal relation between two events. However, to establish a causal relation between two everyday events, one usually needs to possess commonsense knowledge about the world. For example, to recognize the causal relation between the event *my body cast a shadow over the grass* and the event *the sun was rising*, one needs to know

<p><i>Premise:</i> The woman repaired her faucet. <i>Question:</i> What was the cause of this? (1) The faucet was leaky. (2) The faucet was turned off.</p> <p><i>Premise:</i> My favorite song came on the radio. <i>Question:</i> What happened as a result? (1) I covered my ears. (2) I sang along to it.</p>

Figure 1. Two examples from the COPA dataset.

that the rising sun creates shadows. Because the knowledge needed to connect the events is not stated directly, such reasoning problems are excellent for testing whether models can infer unstated, commonsense information.

3.1.1. The COPA Dataset

The Choice of Plausible Alternatives (COPA) dataset (Roemmele et al., 2011) is designed to evaluate English commonsense causal reasoning and contains 1,000 instances in total. Each instance in the dataset (see Figure 1 for example instances) consists of a premise and two choices, each of which could plausibly be causally related to the premise. The choices represent either plausible effects or plausible causes of the premise. The correct answer is the alternative that is more plausible. The dataset is split into 400 training, 500 testing, and 100 validation examples. With the rise of transformer-based models, COPA has become widely used in the NLP community and, for instance, is included in the general-purpose natural language understanding benchmark SuperGLUE (Wang et al., 2019).

3.1.2. Building the Estonian Dataset

We call our Estonian version of the dataset EstCOPA as it is based on the English COPA dataset. Specifically, we created two translated versions of the original COPA: a machine-translated version (EstCOPA_{MT}) generated using the MTee machine translation system (Bergmanis et al., 2022), and a post-edited version (EstCOPA_{PE}) where the machine translated output was further refined by a human. While post-editing is more costly, it results in translations that are more natural, take into account the context of other sentences (since MTee operates at the sentence level), and are more consistent in grammatical tense and gender agreement. Having these two versions allows us to investigate how translation quality affects performance in cross-lingual transfer learning setups that utilize a translation proxy (e.g. TRANSLATE-TRAIN).

To get a better understanding of the quality of the translations, we calculated the translation error rate (or TER; Snover et al., 2006) between machine-translated and post-edited data. TER is defined as the minimum number of edits needed to change a translation so that it exactly matches one of the reference translations, normalized by the length of the reference. We obtained a TER of 0.26, which indicates that the initial translations were already acceptable but required moderate edits when compared to the post-edited data. This is not surprising, as COPA

COPA

Premise: I pricked myself with the pin. *Question:* What happened as a result?
(1) A bead of sweat dripped from my face. (2) A drop of blood formed on my finger.

EstCOPAMT

Premise: Ma torkasin end pinaliga. *Question:* Mis oli selle tagajärg?
(1) Mu näolt tilkus higihelmes. (2) Tilk verd moodustus mu sõrmes.

EstCOPA_{PE}

Premise: Ma torkasin end nõelaga. *Question:* Mis oli selle tagajärg?
(1) Mu näolt tilkus higi. (2) Mu sõrmele tekkis veretilk.

XCOPA

Premise: Ma torkasin ennast oraga. *Question:* Mis oli selle tagajärg?
(1) Mu näolt tilkus higipiisk. (2) Mu sõrmel moodustus veretilk.

Figure 2. An example from the COPA dataset and its translations to Estonian, illustrating differences in translation quality. The EstCOPAMT version contains a lexical error, translating *pin* as *pinal* (‘pencil case’), which makes the premise lose its meaning. Both EstCOPA_{PE} and XCOPA translations preserve the intended meaning, although translating *pin* as *ora* (‘thorn’) introduces a minor semantic deviation.

consists of very short sentences.

During post-editing, editors were shown the question but not the label. To verify that the original labeling was preserved, we asked an editor to re-annotate the translated validation set. The re-annotations agreed with the original labels in 99% of cases which is comparable to prior work (Ponti et al., 2020). We assume a similar agreement rate for the training and test sets.

3.1.3. Comparison with the XCOPA Dataset

A related line of work is XCOPA (Ponti et al., 2020), the multilingual extension of COPA. XCOPA extends the COPA dataset to 11 typologically diverse languages, including Estonian, to advance research on cross-lingual generalization of language models. It consists of professionally translated validation and test sets from COPA. The training dataset of COPA is left untranslated as the goal is to encourage researchers to focus on methods that can generalize to new languages without requiring language-specific fine-tuning data (i.e., ZERO-SHOT model transfer).

EstCOPA, on the other hand, additionally includes translations of the full training set, promoting research on cross-lingual transfer learning strategies that utilize synthetic target data during fine-tuning. Moreover, the machine translated and post-edited translations in EstCOPA are less costly to create than the professionally translated XCOPA dataset, allowing researchers to explore cost-efficient alternatives to human translations for both training and testing. An example translation from each dataset is shown in Figure 2.

3.2. Cross-Lingual Transfer Learning Experiments with EstCOPA

Our experimental setup departs from conventional approaches as we distinguish between cross-lingual abstractions learned solely during multilingual pre-training and those learned during both multilingual pre-training *and* fine-tuning. This separation differs from the typical approach, in which comparisons are limited to ZERO-SHOT, TRANSLATE-TEST, and TRANSLATE-TRAIN, none of which make use of multilingual fine-tuning.

3.2.1. Experimental Setup

Our experimental setup studies cross-lingual transfer learning across two dimensions. First, we conduct experiments using both multilingual and monolingual (that is, trained only on Estonian) language models. Secondly, we compare both encoder-only and decoder-only models in our experiments. Finally, note that whenever possible, we will not re-run experiments for models and setups that has already been done by others (mostly ZERO-SHOT and TRANSLATE-TEST experiments) but will instead use the existing results.

Encoder-only models. The experiments conducted with encoder-only models are summarized in Table 1. The first three rows in the table present training setups that leverage multilingual training (during pre-training, fine-tuning or both) to solve a commonsense reasoning task in Estonian. The last three rows list the baseline configurations against which our methods are compared. Note that the baselines either do not employ multilingual training or do not directly process Estonian (e.g., TRANSLATE-TEST) and therefore do not constitute methods for Estonian NLU. Nevertheless, they serve as an upper bound on performance, indicating what can be achieved with the given base models when fine-tuning is performed in a language that the models understand best. The multilingual encoder-only model that we use in our experiments is XLM-R (Conneau et al., 2020). We conduct experiments with both the base (XLM-R) and large (XLM-R-L) variants of the model. The Estonian encoder-only model used in our experiments is EstBERT (Tanvir et al., 2021), which is available in two versions that differ in their maximum sequence length (128 or 512 tokens).

Our fine-tuning is conducted by formulating the task as a multiple-choice classification problem, where the classification token is transformed into a probability vector $\hat{\mathbf{y}}_i = [\hat{y}_1, \hat{y}_2]$, with each element representing the predicted probability of a class. The model is trained by minimizing the cross-entropy loss of that vector. To account for potential instability in results due to random initializations, we fine-tune the models using 10 random initializations, and the reported results are based on an ensemble of the 5 best-performing models based on validation accuracy.

Decoder-only models. Decoder-only models are typically not fine-tuned via gradient updates. Instead, they are prompted with an input text and the continuation that has the higher summed log-probability is treated as the model’s pre-

Pre-training language(s)	Fine-tuning language(s)	Model	Method	Knowledge transfer	Measures Estonian NLU
multilingual	Estonian	XML-R	TRANSLATE-TRAIN	during PT	yes
multilingual	English	XML-R	ZERO-SHOT	during PT	yes
multilingual	multilingual	XML-R	multilingual	during PT and FT	yes
Baselines					
multilingual	English	XML-R	TRANSLATE-TEST	during PT	no
Estonian	Estonian	EstBERT	TRANSLATE-TRAIN	-	yes
English	English	RoBERTa	TRANSLATE-TEST	-	no

Table 1. Overview of training setups used in our experiments for encoder-only models. The first three rows show multilingual training configurations for improving Estonian natural language understanding and the last three rows shows the baselines. PT stands for pre-training and FT for fine-tuning.

diction. To better guide the model toward the desired output, the prompt may include a small number of question–answer examples. We follow this practice and prompt decoder-only models with zero, one, or two question–answer pairs, which may be provided in English, Estonian, or combine both languages. The multilingual decoder-only model that we use in our experiments is XGLM (Lin et al., 2022b) that is trained on 30 languages, including Estonian, and comes in five different sizes. The Estonian decoder-only model we use is gpt-for-est¹ that bases on GPT-2 architecture and comes in two different sizes (119M and 724M parameters) and is trained on 2.2 billion words.

3.2.2. English Data Drives Estonian NLU

The results of fine-tuning encoder models with different setups are presented in Table 2. We observe that in the majority of the experiments, the task accuracy remains below 60%, often remaining close to the random baseline of 50%. However, there are two notable outliers to this pattern. The first is a ZERO-SHOT result reported by Ponti et al. (2020), in which a large auxiliary dataset, SIQA (Sap et al., 2019), containing 33k training instances, was used during fine-tuning. The second is our TRANSLATE-TRAIN result, where EstCOPA_{MT} was augmented with the COPA training split. This setup slightly outperforms the SIQA + COPA approach while using over 40 times less training data.

These results suggest that with a sufficiently strong multilingual base model, fine-tuning solely on English can yield representations that are generalizable across languages. However, a 40 times smaller combination of Estonian and English data slightly outperforms this method. Notably, the Estonian data in this setup is fully synthetic (i.e., machine translated from English), making the approach more cost-effective in terms of data collection (the comparison of machine translated and post-edited training data is presented in Section 3.2.2). In contrast, the

¹<https://huggingface.co/tartuNLP/gpt-for-est-base> and <https://huggingface.co/tartuNLP/gpt-for-est-large>

Method	Dataset	Size	XLM-R	XLM-R-L	EstBERT ₁₂₈	EstBERT ₅₁₂
ZERO-SHOT	COPA	400	57.8 [†]	49.4 [†]	–	–
ZERO-SHOT	SIQA+COPA	33,810	59.4 [†]	71.4 [†]	–	–
TRANSLATE-TRAIN	EstCOPA _{MT}	400	53.4	57.4	52.2	55.8
multilingual	EstCOPA _{MT} + COPA	800	56.6	73.8	–	–

Table 2. Performance of encoder-only models on Estonian commonsense reasoning (XCOPA_{ET}). Best results per column are shown in bold. Scores from Ponti et al. (2020) are denoted with symbol [†].

ZERO-SHOT method requires significantly more training data and engineering, such as identifying suitable auxiliary datasets.

Interestingly, in both cases, the model being fine-tuned was XLM-R large. This suggests that the base model possesses superior multilingual capabilities compared to the other models evaluated. On the other hand, it also highlights the importance of selecting the right base model for the downstream task, as task-specific data engineering seems to have less impact on the task performance than the choice of base model.

Method	Dataset	Size	XLM-R-L	RoBERTa-L
TRANSLATE-TEST	COPA	400	76.8	81.0

Table 3. TRANSLATE-TEST performance on XCOPA_{ET}. Both results are from Ponti et al. (2020). Best result is shown in bold.

Although our best result (73.8%) is slightly below the 76.8% achieved with XLM-R large and TRANSLATE-TEST strategy (see Table 3), it remains close, suggesting that our approach is close to the upper performance bound achievable with this model and dataset on English.

Experiments with decoder-only language models are presented in Table 4. The results indicate that for the multilingual XGLM, combining Estonian and English examples is more effective for solving the Estonian commonsense reasoning task than using examples only in one language. However, prompting the models with Estonian-only examples outperforms prompting with English-only examples. Yet, when comparing our best results with TRANSLATE-TEST reported by Lin et al. (2022b) we can see that there remains a notable gap between state-of-the-art autoregressive models on English and Estonian.

In all our experiments with decoder-only models, we achieved the best results with the largest 7.5B-parameter XGLM model. However, it also appears that a larger model does not always guarantee better performance. For instance, XGLM-4.5B is almost always outperformed by the smaller XGLM-1.7B and XGLM-2.9B models. Among the gpt-for-est models, the larger version is slightly stronger, outperforming both XGLM-564M and the XGLM-1.7B models, despite the latter being over twice its size. These Estonian-only models do not benefit from mixing Estonian and English examples, which is expected given their monolingual training.

Setting	gpt-for-est 119M	gpt-for-est 724M	xglm 564M	xglm 1.7B	xglm 2.9B	xglm 4.5B	xglm 7.5B	XGLM-6.7 en-only
0-shot	54.2	59.2	55.4	57.4	60.0	52.0	60.6	–
1-shot (et)	57.2	57.4	55.0	56.4	59.4	55.8	61.6	–
2-shot (et)	56.0	57.6	55.0	56.8	59.6	56.8	62.6	–
2-shot (mixed)	55.4	58.0	56.2	56.2	60.6	57.0	63.8	–
1-shot (en)	56.0	57.6	53.2	54.2	55.8	51.8	59.2	–
2-shot (en)	53.6	57.6	53.2	55.2	54.8	51.2	59.6	–

Results from Lin et al. (2022)								
0-shot	–	–	–	–	–	–	61.6	–
4-shot (et)	–	–	–	–	–	–	65.9	–
0-shot (TRTE)	–	–	–	–	–	–	–	72.4
4-shot (TRTE)	–	–	–	–	–	–	–	73.6

Table 4. Performance of decoder-only models on XCOPA_{ET} under zero-shot and few-shot prompting. The top block reports our experiments with few-shot examples only in Estonian (et), only in English (en), or both in English and Estonian (mixed). The bottom block shows results reported by Lin et al. (2022b). TRTE stands for TRANSLATE-TEST.

3.2.3. Supplementary Study: The Impact of Translation Quality

Having multiple translation versions of the same dataset provides an opportunity to investigate the impact of training data translation quality on model performance. Table 5 compares the accuracies of models trained on machine-translated EstCOPA_{MT} and post-edited EstCOPA_{PE}. Notably, higher quality training data does not lead to improved accuracy on the human-translated test dataset. This insight, combined with the previous finding that improvements in accuracy stem mostly from enabling knowledge transfer from high-resource languages, suggests reconsidering the need for expensive human-created training datasets.

Method	Dataset	Size	XLM-R	XLM-R-L	EstBERT ₁₂₈	EstBERT ₅₁₂
TRANSLATE-TRAIN	EstCOPA _{MT}	400	53.4	57.4	52.2	55.8
TRANSLATE-TRAIN	EstCOPA _{PE}	400	53.6	55.4	57.2	55.0

Table 5. Comparison of TRANSLATE-TRAIN performance using EstCOPA_{MT} and EstCOPA_{PE} training datasets.

Using the XCOPA test data allows us to assess whether expensive human-translated test sets can be replaced by more cost-efficient alternatives. Specifically, we evaluate whether the human-translated XCOPA test set could be substituted with the post-edited EstCOPA_{PE} test set by comparing model accuracies on both datasets. If the results are comparable, this would suggest that EstCOPA_{PE} is a viable, lower-cost alternative to XCOPA for evaluation purposes.

Table 6 reports the accuracies of models trained on EstCOPA_{MT}+COPA on both XCOPA and EstCOPA_{PE} test sets. The accuracy on the EstCOPA_{PE} test dataset is consistently higher, though by a small margin, than on the XCOPA test set, indicating that the lower-quality test set is easier for the models. One possible

Test dataset	EstBERT ₁₂₈	EstBERT ₅₁₂	XLM-R	XLM-R-L
XCOPA	58.6	57.0	56.6	73.8
EstCOPA _{PE}	59.0	59.0	57.2	75.6

Table 6. Comparison of model accuracies on the XCOPA and EstCOPA_{PE} test datasets when trained on EstCOPA_{MT}+COPA.

explanation is that EstCOPA_{MT} and EstCOPA_{PE} may contain more *translationese* than human-translated XCOPA. Since the models were trained on machine-translated data likely containing *translationese*, they may have learned to exploit these patterns, potentially at the expense of performance on higher-quality Estonian text.

Ideally, this comparison should also include native Estonian data to provide a more reliable assessment of the suitability of translated test datasets. However, no such benchmark currently exists for Estonian.

3.3. The Problem with Translated Test Datasets

Cross-lingual transfer learning methods are designed to address the data scarcity that prevents low-resource languages from benefiting from data-hungry language technology tools. The *goal* of these methods is to solve NLU problems in a target language *without* requiring language-specific training data. And indeed, such training data is typically unavailable.

However, test data in the target language is still needed. Usually, the test set is created specifically for evaluating cross-lingual transfer methods. If it were not created for that purpose, it means that it would have been a byproduct of a larger dataset that also includes training data in the target language but in that scenario, cross-lingual transfer learning methods would not be necessary.

Unfortunately, creating datasets from scratch, even test datasets, is expensive, as it involves hiring annotators and conducting extensive verification of data quality. By contrast, translating existing English datasets into other languages is much cheaper, even when done manually. Additionally, using widely recognized English datasets allows for better comparability with prior work. Another benefit of translated datasets is that they enable direct comparisons across languages, since the underlying content is identical.

For the reasons described above, evaluations of cross-lingual transfer learning methods are often conducted on test sets translated from English into the target language (Artetxe et al., 2020b; Ponti et al., 2020). However, these test sets are often part of a larger dataset that also includes training and validation sets, as they were originally created to advance English NLU. As a result, a common and convenient approach is to use the same training set, either the original English version (as in ZERO-SHOT and TRANSLATE-TEST), or a machine translated version (TRANSLATE-TRAIN) establish cross-lingual transfer. But this also means that the test and training sets come from the same source.

We argue that it is an issue, because during training, models may learn to exploit linguistic patterns or references to cultural context that are specific to English. As these patterns would persist in the translated test data, it might lead to inflated performance on the translated test sets compared to the native test sets, which may contain language- and culture-specific nuances absent from the English source.

3.3.1. Experiment with a Reading Comprehension Task

To compare native and translated test datasets, both must exist in the target language, which is an uncommon situation. Therefore, we select a task with an existing native Estonian dataset and create the translated dataset ourselves, as this is less expensive to produce. We opt for EstQA (Käver, 2021a), a reading comprehension dataset created to be an Estonian equivalent for English SQuAD (Rajpurkar et al., 2016).

English SQuAD itself is a popular benchmark for evaluating encoder-only models. Each instance in the dataset consists of a paragraph of text and a question based on this paragraph. To solve the task, the model has to correctly detect the span that contains the answer to the given question. With encoders, this is approached as a token classification task where the model has to predict the start and end tokens of the span. Artetxe et al. (2020b) translate 1190 instances from SQuAD into ten languages (not including Estonian), resulting in a cross-lingual version called XQuAD.

We create a machine-translated alternative to EstQA using XQuAD. The data from XQuAD are translated to Estonian with the Google Translation API. We first align the spans with SimAlign (Jalili Sabet et al., 2020) and then correct them manually. We call this dataset XQuAD_{ET}.

We train a QA model by fine-tuning XLM-Roberta-base on the SQuAD training dataset. We decide not to use existing QA models (although reusability is one of the main benefits of TRANSLATE-TEST) because XQuAD is a subset of the validation set of SQuAD and would give an unfair advantage to XQuAD in our experiments.

We use TRANSLATE-TEST to obtain predictions both on EstQA and XQuAD_{ET}. This means that the input is first machine-translated to English and then fed to the model. The predicted span in English is then automatically aligned with the Estonian input to get the final prediction in Estonian.

Train data	Test data	EM	F1
SQuAD	XQuAD _{ET}	58.74	72.26
	EstQA	57.04	70.35

Table 7. Performance of the TRANSLATE-TEST method on native EstQA and translated XQuAD_{ET} test datasets.

The results in Table 7 support the hypothesis that evaluating TRANSLATE-

TEST with a translated test dataset might lead to an overestimation of performance on the native target data. Note that in order to obtain the predictions for XQuA_{ET} the data was machine translated twice (first into Estonian and then back into English during inference) but is still easier for the model to solve, despite the potentially accumulating translation errors that can distort the meaning of the text.

3.4. Conclusion

In this chapter, we examined methods that leverage multilingual data to enable language models to perform commonsense reasoning in Estonian. For both encoder-only and decoder-only models, the most effective approaches consistently relied on multilingual data during pre-training, fine-tuning, and few-shot prompting. Notably, we found that even when the target-language data used in multilingual fine-tuning is fully synthetic, it can still yield substantial performance gains, whereas the impact of improving the quality of this data remained unclear in our experiments. However, we noted that performance was strongly dependent on the base model, with larger models consistently achieving higher accuracy.

We also investigated the use of translated test datasets for model evaluation and the effect of translation quality. Models fine-tuned on mixed data, including only synthetic Estonian, achieved slightly higher accuracy on post-edited test dataset than on professionally translated test dataset. We hypothesize that this may be due to the models learning to exploit *translationese*. Similarly, we compared machine translated test dataset with native test dataset in TRANSLATE-TEST setting where the training split for fine-tuning comes from the same dataset as the machine-translated test split. We observed slightly higher accuracy on machine-translated test sets when compared to native test set, despite multiple additional automatic processing steps in translating the test data, that may introduce noise. Together, these findings highlight the sensitivity of evaluation results to translation choices.

4. CROSS-LINGUAL TRANSFER LEARNING WITH LLMs

With LLMs, the goal of cross-lingual transfer has shifted from improving task-specific performance in the target language to teach models to utilize arbitrary knowledge and skills it has learned in the primary language (i.e., English) during the pre-training, also in the target language. In turn, that enables general-purpose conversational assistants also for languages with very limited training data available. In this chapter, we aim to use this method build a conversational assistant for Estonian. The chapter is further divided into three sections.

In Section 4.1, we systematically study various cross-lingual transfer learning strategies for transferring the *arbitrary knowledge* that the English-centric base model has learned to Estonian. For that purpose, we conduct an exhaustive series of experiments on the 7B-parameter Llama 2 model, combining continued pre-training with various instruction tuning setups. To make instruction tuning in Estonian possible at all, we also create a novel Estonian general task instruction dataset, Alpaca-est, as Estonian lacks instruction tuning datasets.

We evaluate our strategies on a set of standard NLP benchmarks and find that even a relatively small amount of monolingual Estonian pretraining data, complemented with cross-lingual instruction-tuning data, remarkably improves the performance of the Llama 2 model on various Estonian NLP tasks. We further showcase knowledge transfer from an additional set of high-quality English instructions and conversations, reflected in improved commonsense reasoning and the ability to engage in multi-turn conversations.

In Section 4.2, we place our findings in a broader context by conducting a comprehensive evaluation of six state-of-the-art open-source base language models across five Finno-Ugric languages, including Estonian. We observe substantial improvements in understanding Finno-Ugric languages, including extremely low-resource languages, such as Livonian, when compared to earlier ones. This indicates that newer base models are much better at transferring knowledge from high-resource languages to low-resource languages without further language-specific training, making us re-evaluate some of the previously held assumptions.

4.1. Experiments on Eliciting Knowledge Transfer to Estonian

Our goal is to enable an English-centric language model to apply its knowledge when responding to prompts in Estonian. We use the Llama 2 7B as the base model for our experiments. In general, our strategy of converting it into a conversational Estonian language model follows a two-step process. The first step is continued pre-training of the base model on Estonian data to improve the model’s familiarity with the target language. The second step is instruction-tuning to elicit instruction-following and conversational behavior. While continued pre-

training is relatively straightforward, instruction-tuning can be approached in various ways, especially under data constraints, where high-quality instructions in the target language are scarce.

4.1.1. Alpaca-est Dataset

The motivation to create Alpaca-est is the lack of instruction-tuning datasets in Estonian. Alpaca-est is generated following the methodology of Taori et al. (2023) and is purely synthetic. To create the dataset, we first create a set of Estonian seed instructions that consists of 90 translated examples from the Alpaca seed set and 17 new instructions written by the authors. We then randomly sample from a set of seed instructions and use an `gpt-3.5-turbo-0613` to generate new instructions based on the examples. We generate a total of 52,006 instructions for Estonian.

examples	
instruction	Nimeta kolm kuulsat Eesti muusikut. (<i>Name three famous Estonian musicians.</i>)
input	-
output	Mõned kuulsad Eesti muusikud on Tanel Padar, Kerli ja Ewert and The Two Dragons. (<i>Some of the famous Estonian musicians are Tanel Padar, Kerli and Ewert and The Two Dragons.</i>)
instruction	Antud loendis tuvastage kõik numbrid, mis on suuremad kui 10. (<i>From the given list, identify all numbers that are greater than 10.</i>)
input	Loend: [5, 12, 8, 15, 3, 9, 10] (<i>List: [5, 12, 8, 15, 3, 9, 10]</i>)
output	Suuremad kui 10: [12, 15] (<i>Greater than 10: [12, 15]</i>)
instruction	Looge lõbus ja kutsuv ürituse pealkiri. (<i>Create a fun and inviting event title.</i>)
input	-
output	Tantsiöö tähtede all: hullamiseks ja liugu laskmiseks! (<i>Dance night under the stars: for fooling around and sliding!</i>)
instruction	Leia tekstist viis nimisõna. (<i>Find five nouns from the text.</i>)
input	Kohtumine oli äärmiselt meeldejääv. (<i>The meeting was extremely memorable.</i>)
output	kohtumine, äärmiselt, meeldejääv. (<i>meeting, extremely, memorable.</i>)

Table 8. Examples from Alpaca-est. These examples illustrate unstable quality of the synthetic data generated with `gpt-3.5-turbo-0613`. While the first two examples contain well-formed instructions and correct outputs, last two exhibit odd or problematic content.

We observe that the resulting dataset exhibits uneven quality. While some examples are of high quality, others guide the model toward incorrect outputs, contain odd or problematic content, or display unnatural Estonian. Examples of the generated instruction are shown in Table 8.

4.1.2. Experimental Setup

In addition to the newly introduced Alpaca-est dataset, the data mixtures used in our experiments incorporate several existing datasets. This section provides a detailed overview of the data sources used for training and evaluation.

Continued pre-training data. For continued pre-training, we use a subset of English and Estonian data from CulturaX (Nguyen et al., 2024) that contains 8.8B tokens of cleaned and deduplicated Estonian data. After manually observing the

data, we decided to further limit the Estonian subset to documents that come from .ee, .org or .net domains to reduce the proportion of machine translated and other lower-quality documents.

Synthetic general task instructions. The synthetic general task instruction dataset is combined from the Stanford Alpaca dataset (Taori et al., 2023) and our own Alpaca-est dataset (see Section 4.1.1). We call the combination of these two datasets **Alpacas**.

High-quality general task instructions. We use human-generated OpenAssistant 1 (Köpf et al., 2023) multi-turn conversations, taking the top-scoring English-only path from each conversation tree. We also take 10,000 examples of both Chain-of-Thought and FLAN-2 mixtures (Longpre et al., 2023; Chung et al., 2024). We call this high-quality mixture of data, consisting of 22,363 examples in short as **HQI**.

Translation task instructions. This set of instructions is created from relatively low-quality bitexts: CCMatrix (Schwenk et al., 2019), WikiMatrix (Schwenk et al., 2021), OpenSubtitles (Lison and Tiedemann, 2016), and Europarl (Tiedemann, 2012). We filter the data with OpusFilter (Aulamo et al., 2020) using long word, sentence length, source-target length-ratio, character score, language-ID, terminal punctuation, and non-zero numerals filters. The final dataset consists of 1 million instructions among which 75% prompt the model to translate from English to Estonian and 25% the opposite direction. We call this relatively low-quality translation task instructions dataset **TRTASK**. Additionally, we create a high-quality translation task instruction dataset, called **HQTRTASK** from WMT18 dev set (Bojar et al., 2018) consisting of 245 documents and MTee (Tätär et al., 2022) held-out validation dataset consisting of 1,528 sentences.

Evaluation. We evaluate the models on five downstream tasks which are extractive question answering (QA), commonsense reasoning (CSR), grammatical error correction (GEC) and machine translation between English and Estonian (MT). Performance on question answering is measured using the EstQA dataset (Käver, 2021b), while commonsense reasoning is evaluated with EstCOPA using post-edited translations (see Chapter 3). Grammatical error correction is evaluated using the EstGEC-L2 dataset¹. Translation performance is reported on the FLORES-200 devtest (Team et al., 2022).

We report accuracy for commonsense reasoning, accuracy and F1 for question answering, BLEU (Papineni et al., 2002) for machine translation, and F0.5 (Bryant et al., 2017) for grammatical error correction.

4.1.3. Knowledge Transfer with Continued Pre-Training

The goal of continued pretraining is to make the model more familiar with the target language. Although it can be done with arbitrary amounts of data, in practice, due to data constraints, it is often done with relatively small datasets when

¹<https://github.com/tlu-dt-nlp/EstGEC-L2-Corpus>

compared to data that is used for full pretraining. For example, we estimate that there exists only around 12B tokens of raw monolingual data for Estonian.² This is approximately 167 times smaller than the size of the entire Llama 2 pretraining dataset (2 trillion tokens).³

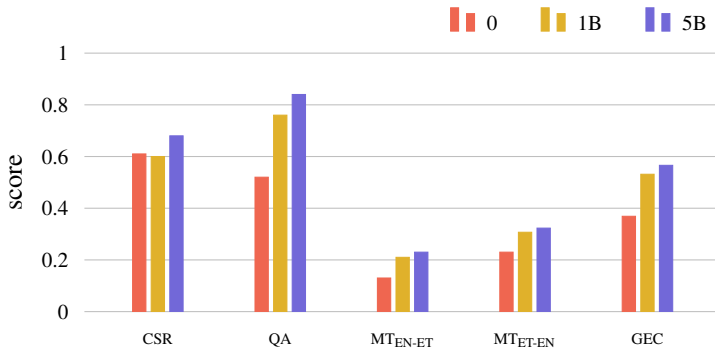


Figure 3. Effect of continued pre-training dataset size on Estonian downstream tasks. The colors of the bars indicate the size of the pretraining dataset. The models are fine-tuned on cross-lingual general task instruction dataset Alpacas.

We therefore design our pretraining experiments to follow a data-constrained scenario in which little or no monolingual data is available. Specifically, we perform continued pretraining of the Llama 2 7B model on 0B, 1B, and 5B tokens of data, 75% of which is in Estonian and 25% in English. The models are trained using bf16 precision with DeepSpeed (Rasley et al., 2020), with a context length of 1024 and a batch size of 256, using 4 AMD MI250x GPUs (acting as 8 GPUs) on the LUMI supercomputer. The pre-training on 5B tokens took 1184 GPU-hours.

To evaluate the impact of continued pre-training on Estonian capabilities, we instruction-tune all three models on the same cross-lingual general task instruction-tuning dataset, Alpacas. This allows us to quantify how much continued pre-training additionally contributes to knowledge transfer elicited with cross-lingual instruction-tuning.

Figure 3 shows the effect of pretraining dataset size on model performance for Estonian tasks. The results reveal a positive correlation between the continued pretraining dataset size and downstream performance on most of the tasks (the only exception is the accuracy on commonsense reasoning after continued pre-training on 1B tokens). Notably, even 1B tokens yield a substantial improvement, with the exception of commonsense reasoning. Using 5B tokens leads to further gains, although the improvement is smaller than that achieved with the initial 1B. We call the best performing model, the one with continued pre-training on 5B tokens, LLAMMAS-BASE.

²This approximation is calculated by multiplying the size of Estonian National Corpus 2023 (Koppel et al., 2023) in words (3.8B) with the average number of tokens per Estonian word (3.28) according to Llama 2 tokenizer.

³Note that in this scale, most of the languages would be considered low-resource.

4.1.4. Knowledge Transfer with Translation Task Instructions

While general task instructions are typically difficult to obtain for the target language, creating translation task instructions requires minimal effort, given that some parallel data is available for the target language. On one hand, teaching the model to translate between languages helps to strengthen cross-lingual alignment inside the model (Chowdhery et al., 2022) and is computationally less expensive than continued pretraining. On the other hand, the significantly larger volume of translation task instructions may adversely affect the model’s ability to generalize to new tasks. Moreover, it remains unclear how additional monolingual pretraining combines with translation task instruction tuning.

To answer these questions, we conduct a follow-up experiment where, after continued pre-training on 0B, 1B, or 5B tokens, the model is sequentially fine-tuned first on translation task instructions and then on cross-lingual general task instructions.

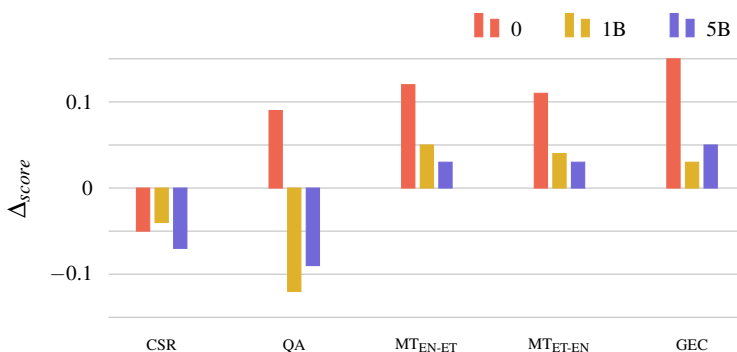


Figure 4. Effect of sequential fine-tuning, first on translation task instructions (TRTASK) and then on general task instructions (Alpacas), compared to fine-tuning without the translation task stage (Figure 3). Results are shown across different amounts of continued pre-training tokens; bar colors indicate the size of the pre-training dataset.

Figure 4 shows the performance gained or lost for each base model and task when sequential fine-tuning is applied, compared to using only general task instruction tuning. We observe that translation task instructions significantly improve results for QA, machine translation, and GEC when no additional pre-training is done. However, these benefits diminish considerably when a pre-training step is introduced. In fact, pretraining followed by translation task fine-tuning negatively affects performance on QA and commonsense reasoning tasks, suggesting that this setup reduces the model’s ability to generalize to out-of-distribution tasks. (It should be noted that since the model is both trained and evaluated on the machine translation task, and GEC is, in a sense, a closely related task — they may not be entirely appropriate tasks for measuring generalization.) Therefore, while sequential fine-tuning with translation tasks offers advantages in the complete absence of monolingual data, it does not replace continued pretraining and may negatively affect effective cross-lingual knowledge transfer.

4.1.5. Knowledge Transfer with High-Quality English Instructions

In the previous sections, we compared models trained on Alpacas with those trained sequentially on translation task instructions (TRTASK) and general task instructions (Alpacas). In this section, we use the best-performing base model from the previous experiments (LLAMMAS-BASE which was pre-trained on additional 5B tokens of data) to further investigate instruction-tuning strategies beyond synthetic instruction-tuning data.

There are two main motivations for this. On one hand, Alpacas is a synthetic dataset generated with Self-Instruct (Wang et al., 2023), which may suffer from various issues that lower the overall quality of the dataset. On the other hand, prior work has shown that it is possible to achieve remarkably strong instruction-following capabilities with just 1,000 high-quality training examples (Zhou et al., 2023a). We therefore hypothesize that supplementing Alpacas with a set of high-quality instructions could improve the model’s performance. However, as no high-quality instruction datasets exist for Estonian, we instead use high-quality English instructions.

Specifically, we include two additional instruction sets in our experiments: high-quality English general task instructions (HQI) and a high-quality translation task instructions (HQTRTASK)) (see Section 4.1.1 for the composition details). We the effect of combining HQI and HQTRTASK in several ways with Alpacas and TRTASK), as specified below:

- Alpacas + HQI
- Alpacas + HQI + HQTRTASK
- 1) TRTASK 2) Alpacas + HQI + HQTRTASK

Specifically, we build on the setups from the previous experiments: fine-tuning on Alpacas and sequentially fine-tuning on TRTASK and Alpacas. We now augment Alpacas with high-quality general task instructions (HQI) and high-quality translation task instructions (HQTRTASK) creating a composite instruction-tuning dataset. Using this dataset, we evaluate two strategies: (1) fine-tuning only on the composite dataset, and (2) sequential fine-tuning—first on TRTASK, then on the composite dataset.

We also train the model on Alpacas + HQI to isolate the effect of high-quality translation task instructions on model performance. Note that all these strategies include Estonian and English instruction, ensuring that the model benefits from English knowledge and maintains its English capabilities. We use Alpacas and sequential training setups from earlier experiments as baselines, as well as the proprietary GPT-3.5-Turbo and GPT-4 models.

The results are summarized in Table 9. Compared to the baseline model fine-tuned solely on Alpacas (model 1), we observe consistent performance improvements across all tasks when high-quality English general instructions (HQI) and translation task instructions (HQTRTASK) are added (model 4). This indicates positive cross-lingual knowledge transfer from the additional English datasets to

Model	CSR	QA		MT _{EN-ET}	MT _{ET-EN}	GEC
	acc.	F1	acc.	BLEU	BLEU	F0.5
LLAMMAS-BASE fine-tuned						
(1) Alpacas	63.6	46.5	81	22.5	32.3	56.6
(2) 1) TRTASK 2) Alpacas	59.2	46.1	73	25.0	34.5	59.4
(3) Alpacas + HQI	66.4	52.9	82	23.1	32.4	59.4
(4) Alpacas + HQI + HQTRTASK	66.4	54.8	84	22.6	34.6	60.3
(5) 1) TRTASK 2) (4)	62.2	43.5	76	26.9	36.9	61.2
Commercial baselines						
GPT-3.5-Turbo	86.0	34.2	93	26.0	37.5	63.4
GPT-4	98.4	35.1	97	28.5	37.7	67.4

Table 9. Performance of various instruction-tuning strategies on Estonian NLP tasks. Adding high-quality English task instructions (HQI) and translation task instructions (HQTRTASK) improves zero-shot performance. We call (4) LLAMMAS and (5) LLAMMAS-MT.

Estonian. We refer to this model (model 4) as LLAMMAS, as it achieves the best overall performance on zero-shot tasks.

Although one might attribute the improvements over the baseline primarily to HQTRTASK, which teaches the model to translate between English and Estonian, our results suggest that this explanation is incomplete. Notably, model 3 (Alpacas + HQI) also outperforms the baseline, indicating that high-quality English instructions alone contribute to cross-lingual knowledge transfer.

Finally, the best results on EN→ET, ET→EN, and GEC are obtained with the sequentially fine-tuned model 5, which we therefore refer to as LLAMMAS-MT. This model is first fine-tuned on TRTASK, followed by fine-tuning on the combined Alpacas + HQI + HQTRTASK dataset. The positive effect of sequential fine-tuning on translation and grammatical error correction tasks aligns with our earlier findings that sequential fine-tuning can be more effective for translation-related tasks (see Section 4.1.3). However, we observe that, even with the addition of high-quality instructions, sequential fine-tuning degrades performance on commonsense reasoning and question answering tasks.

4.1.6. Will the Model Still Understand English?

In the previous sections, we demonstrated that continued pre-training on the target language, combined with the right instruction-tuning setup, can teach an English-centric language model to understand and generate a new language—Estonian, in our case. However, we are not only interested in teaching the model Estonian but also maintaining its skills and capabilities in English. If the model was no longer able to perform well on English tasks, it would suggest that our training has led to catastrophic forgetting of its previously learned skills. This would mean that the model is not leveraging its English knowledge when responding to Estonian prompts.

To verify that our models still perform well in English, we evaluate our best-

performing models on the English equivalents of the datasets that we used for Estonian evaluation. Specifically, we use COPA and the English subset of XQuAD.

The results on English question answering and commonsense reasoning are shown in Table 10. On the QA task, LLAMMAS reaches a comparable level of accuracy in both languages (83% in English vs. 84% in Estonian). However, it tends to be more verbose in English, producing longer answers and leading to a lower F1 score compared to Estonian (41% vs. 55%). On the commonsense reasoning task, LLAMMAS performs significantly better in English (80.6%) than in Estonian (66.4%), indicating that it has not learned to utilize all the reasoning capabilities it has in English when the input is given in Estonian.

Model	CSR	QA	
	acc.	F1	acc.
Alpacas	63.4	30.4	85
1) TRTASK 2) Alpacas	70.2	29.5	81
Alpacas + HQI	78.6	33.3	87
LLAMMAS	80.6	41.0	83
LLAMMAS-MT	73.6	31.4	82
GPT-3.5	95.2	30.7	95
GPT-4	99.8	33.2	96

Table 10. Comparison of English QA and commonsense reasoning (COPA) performance across all instruction-tuned models.

Additionally, we compare Llama 2 7B and LLAMMAS-BASE on 5 popular English benchmarks: MMLU (Hendrycks et al., 2021), which covers 57 tasks with different levels of difficulty; TruthfulQA (Lin et al., 2022a) which measures truthfulness with questions designed to cause imitative falsehoods; WinoGrande (Sakaguchi et al., 2021), a pronoun resolution challenge; TriviaQA (Joshi et al., 2017), a question answering dataset; and HellaSwag (Zellers et al., 2019), a commonsense reasoning task.

Model	MMLU	TruthfulQA	WinoGrande	TriviaQA	HellaSwag
Llama 2 7B	45.8	32.1	68.8	52.4	76.0
LLAMMAS-BASE	42.6	31.9	70.0	49.8	74.8
difference	-3.1	-0.2	1.2	-2.6	-1.2

Table 11. Performance comparison of Llama 2 7B and LLAMMAS-BASE on five English benchmarks. Results show a minor average performance drop (-1.2%) for LLAMMAS-BASE, primarily due to MMLU.

The results in Table 11 show that LLAMMAS-BASE shows only a slight drop in performance on 4 out of 5 English benchmarks. On average, the difference is -1.2%, with MMLU contributing the most to this drop. These results show that the continued pre-training on 5B tokens of data, which is still hundreds of times less than the total seen during full pre-training, has only slightly affected English capabilities. Nevertheless, further continued pre-training would likely affect English results more.

4.2. State of Open LLMs in Understanding Finno-Ugric Languages

Ever since LLMs were first introduced, they have continued to improve rapidly with each iteration, both in terms of reasoning and complex problem solving skills but also in a multilingual setting. For example, while the early versions of ChatGPT produced awkward and translation-like Estonian, the best proprietary models are currently able to generate near fluent Estonian. However, for some time, strong multilingual capabilities were limited to proprietary models, while open-weight models such as the Llama 2 model family were struggling with most non-English languages. Fortunately, this gap is narrowing with each new iteration of open LLMs although the list of officially supported languages of these models remains limited. Additionally, the primary focus is on those with significantly more data available than for Finno-Ugric (FiU) languages.

On the other hand, it has been observed that even models optimized solely for English, such as the Llama 2 model family (Touvron et al., 2023b), demonstrate some understanding of a wide range of languages beyond their intended use (Holtermann et al., 2024). Yet, the exact level of proficiency of these models remains unclear for many languages. Moreover, comparative performance across models remains largely unexplored.

Motivated by these observations, we conduct a systematic evaluation of a set of open LLMs on five FiU languages: Finnish, Estonian, Livonian, Võro, and Komi.

4.2.1. Experimental Setup

Models. We include all Llama 2 and Llama 3.1 family models (except Llama 3.1 405B) in our experiments as they have demonstrated state-of-the-art performance and competitiveness with proprietary models. Additionally, they have been widely used in non-English adaptation. For further insights, we compare Llama models with Mistral NeMo (Jiang et al., 2024), another natively multilingual open model shown to be competitive with the Llama 3.1 model of the same size. Finally, we evaluate LLAMMAS and compare it to Llama 2 7B and Llama 3.1 8B.

Tasks and Datasets. Since we want results to be comparable across models and languages, we aim to use multilingual datasets that contain our target languages. Unfortunately, the availability of datasets in these languages is limited, especially for Komi, Võro and, Livonian. In total, we evaluate the models on five tasks using nine datasets. A summary of the tasks, datasets, and their language coverage is provided in Table 12.

Evaluation. We report 5-shot results on all tasks using log likelihoods of answer choices whenever possible and greedy decoding otherwise (machine translation and reading comprehension tasks). We use `lm-eval-harness` framework (Gao et al., 2024) to implement the evaluation.

Task	Datasets	est	fin	vro	kpv	liv
machine translation	FLORES-200 (Team et al., 2022), SMUGRI-FLORES (Yankovskaya et al., 2023)	✓	✓	✓	✓	✓
multiple choice QA	Belebele (Bandarkar et al., 2024), Belebele-smugri (Purason et al., 2025b)	✓	✓	✓	✓	✓
text classification	SIB-200 (Adelani et al., 2024), SIB-smugri (Purason et al., 2025b)	✓	✓	✓	✓	✓
extractive QA	EstQA (Käver, 2021a), TyDiQA Clark et al., 2020	✓	✓			
commonsense reasoning	XCOPA (Ponti et al., 2020)	✓				

Table 12. Summary of evaluation tasks, datasets, and their language coverage for the five target Finno-Ugric languages.

4.2.2. Knowledge Transfer from Pre-Training

The results across all tasks and Llama models are shown in Table 13. In general, Llama 2 7B and Llama 2 13B perform significantly worse on the FiU languages in our evaluation than the Llama 3.1 family models. The exception is Finnish, where the Llama 2 models perform notably better than on the other FiU languages. This may be due to the greater amount of Finnish data in the Llama 2 training dataset (Touvron et al., 2023b) when compared to data in other FiU languages. However, both Llama 2 7B and Llama 2 13B still appear weak on Finnish when compared to other models.

Llama 2 70B shows notable improvements over the smaller Llama 2 model on Estonian and Finnish across all tasks. The results for Belebele and SIB also indicate an improvement for Võro, though the improvement in machine translation (FLORES) is less pronounced. Additionally, SIB appears to be generally too easy benchmark for the models, as Llama 2 7B already achieves 86% accuracy for Finnish. For other languages, the benchmark saturates with Llama 2 70B.

We observe that on Estonian and Finnish, LLaMA 2 70B performs competitively with LLaMA 3.1 8B, despite the latter being nearly nine times smaller. However, LLaMA 3.1 8B slightly underperforms on Finnish, as indicated by the results on Belebele and FLORES. This is somewhat surprising, given that Finnish is one of the most well-resourced Finno-Ugric languages. Without access to detailed information about the pre-training data distributions, it is difficult to pinpoint the cause of this behavior. Nevertheless, we note that Finnish was among the most represented non-English languages in the LLaMA 2 training data, accounting for approximately 0.03% of the training corpus.

When comparing Llama 3.1 8B to Llama 3.1 70B, the larger model clearly outperforms the smaller one on Belebele, FLORES, and XCOPA. For Estonian and Finnish, the Llama 3.1 70B achieves nearly 90% accuracy on Belebele and XCOPA, along with very strong BLEU scores on the FLORES dataset. The improvements are also notable for extremely low-resource languages Võro, Komi, and Livonian.

The exact amount of Finno-Ugric data in the pretraining corpora of Llama

	L2-7B	L2-13B	L2-70B	L3.1-8B	L3.1-70B
SIB					
liv	64.8	61.6	83.2	74.4	77.6
kpv	68.0	59.2	83.2	77.6	87.2
vro	64.8	59.2	85.6	86.4	86.4
est	69.6	68.0	88.8	89.6	89.6
fin	85.6	81.6	91.2	87.2	89.6
Belebele					
liv	26.23	35.25	36.89	37.70	42.62
kpv	27.87	31.15	34.43	52.46	73.77
vro	27.05	32.79	44.26	50.82	73.77
est	28.69	36.07	66.39	68.03	88.52
fin	44.26	54.92	86.89	74.59	91.80
XCOPA					
est	49.2	51.8	67.6	69.2	92.6
FLORES (FiU → En)					
liv	6.8	9.3	12.0	10.5	16.1
kpv	5.4	6.0	7.3	10.3	21.9
vro	7.8	9.1	12.9	16.7	30.3
est	12.6	17.8	26.9	35.3	41.0
fin	29.6	31.9	34.6	32.0	37.1
Extractive QA					
<i>exact match</i>					
est	21.89	34.33	49.25	50.75	52.74
fin	51.66	48.34	53.45	58.31	47.06
<i>F1</i>					
est	35.35	51.39	66.72	70.87	73.76
fin	70.63	70.36	74.65	75.44	72.98
<i>BERTScore F1 (Zhang* et al., 2020)</i>					
est	76.88	82.95	88.86	91.76	93.02
fin	88.50	87.95	89.60	90.63	88.67

Table 13. Performance of Llama 2 and Llama 3.1 models across all tasks for five Finno-Ugric languages. Results show that Llama 3.1 models generally outperform Llama 2 models, with the exception of Finnish, where Llama 2 performs comparatively better. The results are obtained with 5-shot prompting. Accuracy is reported for SIB, Belebele, and XCOPA; BLEU is reported for FLORES; and BERTScore F1 is calculated using bert-base-multilingual-cased.

models is unknown, but is likely very limited. For example, Purason et al. (2025b) presented experiments on adapting LLMs to FiU languages, but gathered only 2.6 million characters of pretraining data for Livonian, 14 million for Võro, and 579 million for Komi. Thus, the extremely low-resource nature of Võro, Komi, and Livonian and the consistent improvements of Llama models on these languages indicate that the cross-lingual knowledge transfer capabilities of the pretrained models are improving steadily. However, the contrast in results between Livonian and other extremely low-resource languages Võro and Komi shows that the models struggle more in transferring knowledge to an almost completely unseen language while even limited exposure to Võro and Komi may help the models solve tasks in these languages. However, this is just speculation and further investigation is needed to confirm that.

On the other hand, it might be that better understanding of Võro and Komi comes from their high similarity to related higher-resource languages, Estonian and Russian, which are more exposed to the models, but the counterargument for this is that Livonian is heavily influenced by Latvian which is also a higher-resource language.

	Belebele			FLORES			XCOPA		
	L2	LB	L3.1	L2	LB	L3.1	L2	LB	L3.1
liv	24.59	38.52	33.61	4.74	4.62	4.91	-	-	-
vro	23.77	33.61	48.36	4.61	9.92	12.19	-	-	-
kpj	26.23	29.51	38.52	2.88	1.44	8.18	-	-	-
est	22.95	39.34	62.30	8.53	28.90	31.0	48.80	56.60	56.60
fin	32.79	34.43	68.03	27.16	11.57	28.54	-	-	-
avg	26.07	35.08	50.16	9.59	11.29	16.96	48.80	56.60	56.60

Table 14. Comparison of zero-shot results of Llama 2 7B (L2), LLAMMAS-BASE (LB) and Llama 3.1 8B (L3.1).

The comparison of Llama 2 7B, Llama 3.1 8B, and LLAMMAS shows that LLAMMAS outperforms Llama 2 7B on Estonian and Võro by a significant margin. However, its performance on Finnish decreases, indicating that continued pretraining on Estonian has negatively impacted Finnish language understanding. However, Llama 3.1 8B surpasses both models by a large margin, except on the Belebele task in Livonian. Notably, Llama 3.1 8B outperforms LLAMMAS even on Estonian, demonstrating that language-specific adaptation of a weaker base model cannot match a stronger, unadapted base model. In other words, Llama 3.1 8B has obtained significantly stronger cross-lingual knowledge transfer capabilities through pre-training alone than continued pre-training and careful fine-tuning of a weaker base model on the target language could achieve.

4.2.3. Supplementary Study. Will English as a Pivot Improve Multilingual Capabilities of LLMs?

Some studies have found that using English as a pivot language improves the multilingual capabilities of LLMs (Huang et al., 2023; Shi et al., 2023; Zhang et al.,

2024). More specifically, directing the model to produce intermediate reasoning steps that involve translating the source-language input into English has proven to be beneficial for solving tasks in non-English languages.

We conduct an experiment similar to these existing studies. Specifically, we guide models to translate the prompt to English, produce an answer in English, and then translate the answer back to the source language. Since we guide the model by showing it example input-output pairs, we call this method few-shot chain-of-thought (CoT) (Kojima et al., 2022) prompting. Our intuition is that solving tasks in English might be easier for the model than in some low-resource language. This is supported by the results in Table 13 which show a correlation between a model’s performance in a language and the amount of resources available for it.

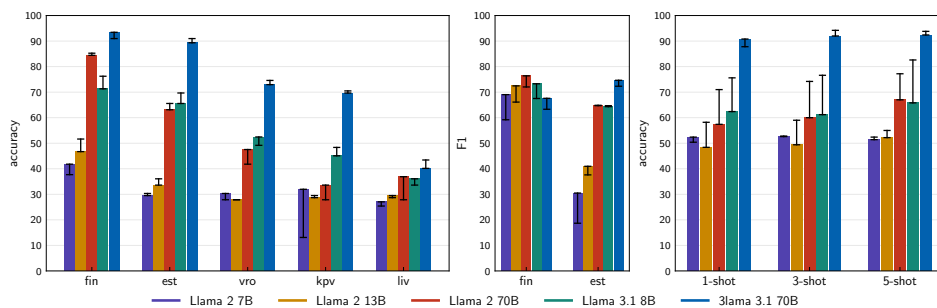


Figure 5. Comparison of few-shot prompting and few-shot chain-of-thought (CoT) prompting on Belebele (left, 1-shot), QA (middle, 1-shot), and XCOPA (right, 1-, 3- and 5-shot) tasks for five Finno-Ugric languages. CoT prompting generally provides little benefit for Belebele, reduces performance on QA, and yields mostly positive gains on XCOPA, with stronger models benefiting less. The bars shows the scores with few-shot prompting. The tiny horizontal line (-) indicates the score with few-shot CoT prompting with the same number of shots.

We test this method on Belebele, QA, and XCOPA tasks as they remained challenging even for the strongest models. Due to the significant increase in the input length with additional CoT examples, we use only one-shot CoT prompting for Belebele and QA. For XCOPA we consider 1-, 3-, and 5-shot scenarios.

Figure 5 shows the comparison of few-shot and CoT few-shot prompting. In general, we see that CoT prompting brings rather few benefits for Belebele and negatively affects the QA task. The leftmost plot shows that most of the time CoT is not beneficial for extremely low-resource languages, which might be caused by the low translation quality. However, good translation abilities also do not necessarily mean benefits from CoT prompting, as exemplified by the results on Finnish. Moreover, Llama 3.1 70B, which has the best translation skills, also benefits less from CoT prompting than weaker models, suggesting that the model has strong cross-lingual capabilities that mitigate the need for CoT prompting.

For the QA task, CoT prompting consistently results in lower performance.

This could be attributed to the nature of the extractive QA task, which requires the output to precisely match the correct text snippet. The intermediate translation steps involved in CoT prompting may lead to slight alterations in the morphological form of the answer, causing a mismatch with the expected output.

In XCOPA, we see mostly positive improvements from CoT prompting, with even Llama 2 13B benefiting, while Llama 2 7B does not. However, the benefit of CoT prompting decreases significantly for Llama 3.1 70B, following the trend observed in the Belebele task.

The discrepancy in our results aligns with Sprague et al. (2024), whose experiments and extensive meta-analysis of existing studies show that CoT provides significant benefits on tasks involving math and logic but yields much smaller gains on other task types.

5. CONVERSATIONAL BENCHMARK FOR FINNO-UGRIC LANGUAGES

As instruction-tuned LLMs can be used to solve arbitrary problems—as long as the problem can be described in a natural language—evaluating them only on standard NLP benchmarks that focus on specific tasks is not sufficient to fully understand their usability as a conversational assistant. However, this is the primary use case of such models. In fact, it is extremely difficult to automate the evaluation of conversational models, as use cases and preferences differ from user to user. Because of that, the performance of conversational models is usually reported only in relation to other models by asking annotators to compare the outputs of two models and pick the preferred one. The winning model in such a pairwise comparison is considered to be more aligned with human preferences than the other model.

To compare models, a set of predefined questions can be used to generate outputs. One test instance can consist of a question and a follow-up question, thus forming a template for a multi-turn conversation. MT-bench¹ (Zheng et al., 2023) is one such dataset. It consists of 80 manually created multi-turn questions across 8 common prompt categories, with a focus on challenging examples to better differentiate models.

category	sample questions	
STEM	1st turn	You have been tasked with designing a solar-powered water heating system for a residential building. Describe the key components and considerations you would include in your design. Design a five-step workflow.
	2nd turn	If the system is intended for a building with a capacity of 100 individuals, what would be the estimated budget for implementing this system?
Humanities	1st turn	How do the stages of life shape our understanding of time and mortality?
	2nd turn	Write an allegorical poem that illustrates the above.

Table 15. Examples from MT-bench.

In the following chapters we introduce a multi-turn conversational benchmark for 4 Finno-Ugric languages, including Estonian. Although the benchmark is inspired by MT-bench, its creation process is different, as we need to account for a low-resource scenario (Section 5.1 and Section 5.2). We then use the benchmark in practice by conducting a human evaluation on a set of Finno-Ugric language models (Section 5.3). In terms of performance, our human evaluation results agree with the conclusions of automatic evaluation. However, we gain additional insights into the generation quality of the models.

¹Here, MT stands for multi-turn.

5.1. Desiderata

Our benchmark aims to cover 4 Finno-Ugric languages, specifically Estonian, Võro, Komi and Livonian, among which Võro, Komi and Livonian are extremely low-resource languages, both in terms of speakers and data resources. Considering these constraints and their potential implications, we formulate the following desiderata for the multi-turn conversational benchmark.

1. Questions should cover real-life scenarios to reflect practical usefulness. The easiest and most likely way for speakers of low-resource Finno-Ugric languages to benefit from LLMs is through interaction via a chat-like interface. Since we want our benchmark to cover real-life use cases, it should consist of prompts similar to natural user queries. Another benefit of using real-life data is that it helps quickly reveal the model’s usefulness and potential weaknesses in practical scenarios, which standard NLP benchmarks typically do not cover.

2. Questions should be challenging enough for LLMs to differentiate the models accurately. Zheng et al. (2024) show that challenging prompts from real-life conversations reveal larger performance gaps between models than manually designed benchmarks, even if the latter contain high-quality challenging questions

3. Answering questions should not require expert knowledge. A key requirement for the benchmark is that it should comprise questions that are challenging for language models. However, such questions are often challenging for humans as well, requiring expert-level knowledge in various domains. Unfortunately, this would significantly shrink the potential evaluator pool, making it nearly impossible to find such evaluators within the communities of extremely low-resource language speakers.

4. Translating the benchmark into a new language should be feasible in terms of time and content (i.e., should not require expert knowledge). Since no data on human interactions with chat LLMs exists for extremely low-resource (XLR) languages, we collect the data in English and translate it. Given the limited availability of professional translators, we carefully select examples that are straightforward to translate even by non-professional translators.

One possible way to create a multi-turn benchmark for Finno-Ugric languages is to translate MT-bench to the target languages. However, it would not meet out benchmark requirements. Firstly, MT-bench is designed to be a challenging benchmark and requires expert-level human labelers (Zheng et al., 2023). It would be almost impossible to find experts in certain domains for some of our target languages. Secondly, our extremely low-resource languages Võro, Livonian, and Komi are mostly used in everyday conversations and might lack vocabulary for domains beyond their typical local usage. This would make domain-specific questions in these languages look unnatural and artificial, drifting too much away from real-life use cases.

5.2. Dataset Construction

Ideally, we would like our dataset to consist of interactions with chat LLMs written by the native speakers of the target languages. However, there exists no such data. We therefore opt for the LMSYS-Chat-1M dataset, which consists of real-world user interactions with LLMs submitted to the Chatbot Arena website (Zheng et al., 2024). We use this dataset to first create the English benchmark that would meet the criteria specified in Section 5.1. The dataset will then be manually translated to the target languages.

The process for creating the initial benchmark is as follows. First, we extract all two-turn English conversations that have not been redacted or flagged by OpenAI moderation API. We only allow conversations with user prompts no longer than 50 tokens to ease the translation process. We then use sentence-transformers to compute the sentence embeddings. We use the embeddings to cluster the data, examine the clusters, and pick user prompts that fill the criteria specified in Section 5.1. By repeating the process multiple times, we end up collecting 248 frequent conversations that meet our requirements. We organize these conversations into four categories: math, reasoning, writing, and general. As we aim for a smaller dataset, we further manually remove near duplicate conversations and conversations where judging the answer still seemed to require overly specific knowledge. The final dataset consists of 80 questions, 20 per category and 42 with a follow-up question.

The dataset is then translated into Estonian, Võro, Komi, and Livonian by fluent speakers with a linguistic background or previous experience in translation. The translators are asked to preserve any informality of the text in the translations, e.g. missing capitalization and punctuation. We make the dataset publicly available.²

One could argue, that instead of human translators, a machine translation system or a proprietary LLM could have been used for translating or generating the test data. To assess the necessity for human translations, we explore both of these options. The BLEU scores of translating the benchmark from Estonian to Võro, Komi and Livonian is shown in Table 16. We can see that the best proprietary LLMs have not yet learned to translate to our extremely low-resource target languages. The model that is unable to translate text to the target language would also not be able to generate synthetic test data of sufficient quality.

The BLEU score in Table 16 suggest that it might be possible to obtain high-quality translations to our low-resource languages with some machine translation systems. To verify that, we take translations that got the best BLEU score for each target language and compare these with human translations using pairwise comparison. For Võro and Livonian the translations are produced by Neurotõlge (Tättar et al., 2022) and for Komi by Llama-SMUGRI-translate (Purason et al., 2025b). The human annotators (who are different from the translators) are

²<https://huggingface.co/datasets/tartuNLP/smugri-mt-bench>

	EST→VRO	EST→LIV	ENG→KPV
gpt-4o-mini-2024-07-18	9.3	5	4.6
gpt-4o-2024-08-06	4.5	5.6	4.2
gpt-4-turbo-2024-04-09	18.9	5.9	3.6
Neurotõlge (Tättar et al., 2022)	24.7	21.4	31.7
Llama-SMUGRI-translate (Purason et al., 2025b)	26.4	25.3	19.1

Table 16. BLEU scores from translating Estonian subset of the SMUGRI-MT-bench into Võro, Komi, and Livonian. The results show that state-of-the-art LLMs do not yet achieve adequate performance for these extremely low-resource target languages, whereas dedicated machine translation systems can potentially produce high-quality translations.

asked to choose a better translation from the two alternatives (ties allowed). We gather 3 sets of annotations for Livonian, 2 sets for Võro and 1 for Komi. Figure 6 shows that Komi and Livonian speakers mostly prefer human translations over machine-translated data, however, Võro speakers surprisingly often prefer machine-translated data suggesting a good quality of Võro machine translation. The average agreement between the pairs of Livonian annotations was 67.5%, while between Võro annotations 42.5%.

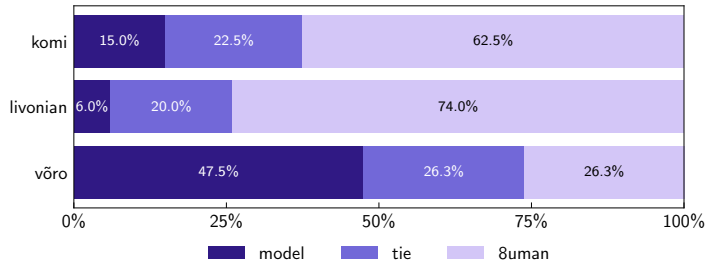


Figure 6. Pairwise comparison of human and machine translations for Võro, Komi, and Livonian, showing the percentage of annotator preferences for each method. Results indicate a strong preference for human translations in Komi and Livonian, while Võro shows comparatively higher acceptance of machine-translated data.

5.3. The Benchmark in Practice. Human Evaluation of LLMs

We apply our newly constructed dataset, SMUGRI-MT-BENCH, in practice to evaluate a set of conversational LLMs optimized for low-resource Finno-Ugric languages (Purason et al., 2025a).³ The goal of our evaluation is to shed more light on the performance differences among these models as automatic evaluations based on a small number of benchmarks did not lead to conclusive results.

The SMUGRI-MT-BENCH is used to conduct human evaluation of these models. For this purpose, we generate responses to each question using every LLM under evaluation. When a question includes a follow-up, it is presented to the

³Both the models and SMUGRI-MT-BENCH were published by Purason et al. (2025a); however, the models themselves are not part of this thesis.

model together with the preceding interaction, thereby simulating a multi-turn conversation.

progress

Your task is to evaluate the quality of the AI assistant's response from two different aspects. First, please assess how **useful** the AI assistant is in answering the user's question. A response is useful if it addresses the user's question accurately, clearly, and in a friendly manner. A response is not useful if it does not answer the question, answers incorrectly, or is unclear. Second, please assess how **natural** the assistant's response sounds in Võro. For example, the use of incorrect case endings can make a response sound unnatural. **Note that a response may sound natural while still not answering the question.** If a **follow-up question** is also displayed, only that response should be evaluated.

Most questions do not have a single correct answer, and you should provide your evaluation based on your best judgment. However, to simplify the evaluation, a **reference answer** is provided for calculation questions and some logic tasks; this reference may be considered during evaluation but does not have to be followed.

User
Write me a heartwarming and reassuring birthday congratulation text for my friend

Assistant
Õnnitõm! Soovi sullõ imelist päivä, miä om täüs armastust, rõõmu ja naardmist. Olkõ tuu kinkalgi, kiä umma sündüst väärt tähüstämä - olõssi tenolik su perrele, armsilõ inemiisile vai hindäle esihinnäst. Looda, et saat jaka ütenukun latsõga rõõmu ja luuta, et egäl sammul juhtus midägi ilosat.

The assistant's response is useful.

strongly disagree somewhat disagree neutral somewhat agree strongly agree

The assistant's response sounds natural in Võro.

strongly disagree somewhat disagree neutral somewhat agree strongly agree

Next [space bar]

Figure 7. Screenshot of the survey used to collect human annotations for responses generated by different models using instructions from SMUGRI-MT-BENCH. **For readability, the survey shown here is translated into English** (except the assistant’s answer, which is in Võro). The original instructions were presented to annotators in Estonian (for Estonian, Võro and Livonian) or in Russian (for Komi).

We use the resulting conversations to carry out a survey for each of the target language. The participants of this survey are shown the generated conversation and asked to rate the helpfulness and naturalness of each response on a 5-point Likert scale. We chose to measure helpfulness as it is a commonly used indicator for assessing practical usefulness of the model (Touvron et al., 2023b). The motivation to ask participants to rate how natural the response sounds in the target language comes from observing the output of the LLAMMAS model, which often sounds unnatural and translation-like. The surveys, which were anonymous, were distributed within target language communities via social media and direct outreach to speakers. The data collection statistics are shown in Table 17.

We use the resulting conversations to conduct a survey for each target language. Survey participants are shown the generated conversations and asked to rate the helpfulness and naturalness of each response on a 5-point Likert scale. We measure helpfulness because it is a commonly used indicator of a model’s

practical usefulness (Touvron et al., 2023b). The decision to additionally assess naturalness is motivated by observing that the output the LLAMMAS model often sounds unnatural and translation-like in Estonian. The surveys were anonymous and distributed within target-language communities via social media and direct outreach to speakers. The screenshot of the survey is shown in 7 and the data collection statistics are reported in Table 17.

	ET	VRO	LIV	KPV
surveys submitted	45	17	6	27
answers graded	1708	836	279	1306
grades per question	2.8	1.74	0.58	2.7

Table 17. Data collection statistics for human evaluation, including the number of participants and responses for each target language. Fractional values for *grades per question* indicate that multiple survey versions were used for each target language and they received uneven numbers of responses, resulting in a non-integer average ratings-per-question value.

The results, together with the baseline GPT-3.5-Turbo, are shown in Figures 8 and 9. Figure 8 shows that all the models by (Purason et al., 2025a)—TrAlpaca, LLMTrAlpaca+TrInst, TrAlpaca+TrInst, for simplicity, we refer them here as SMUGRI-MODELS—perform similarly on all of the languages and therefore do not help in ranking the models, as opposed to our hopes and dreams. Nonetheless, these results are still valuable as they reinforce the findings from automatic benchmarks which also were not able to rank the models.

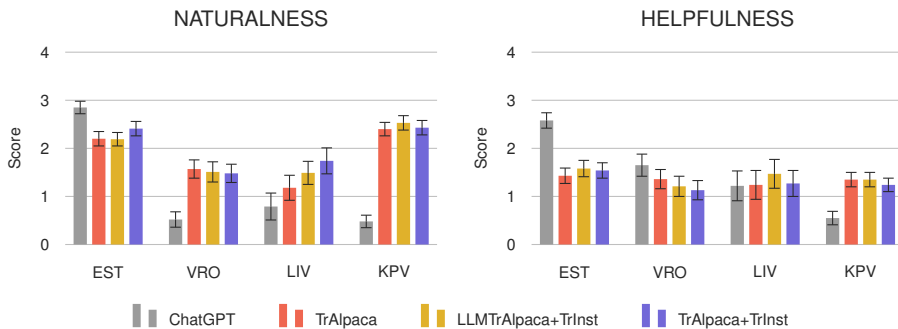


Figure 8. Human evaluation results for conversational LLMs on low-resource Finno-Ugric languages, comparing helpfulness and naturalness scores across models.

However, the results offer valuable insights when compared with GPT-3.5-Turbo. Specifically, it appears that GPT-3.5-Turbo answers in Estonian are significantly more helpful than then answers by SMUGRI-MODELS, which makes sense as Estonian is the most well-resourced. For Võro and Livonian, the helpfulness is on par with GPT-3.5-Turbo and in case of Komi, the SMUGRI-MODELS outperform GPT-3.5-Turbo. While variations in annotator expectations may influence results across different languages, it is noteworthy that our models consistently achieve similar helpfulness scores across various languages.

In terms of naturalness GPT-3.5-Turbo outperforms SMUGRI-MODELS in Estonian, however, is rated significantly lower for Võro, Livonian and Komi. We hypothesize that the reason for this is the extremely limited amount of training resources for these languages that are not sufficient for the model to learn fluency in these target languages.

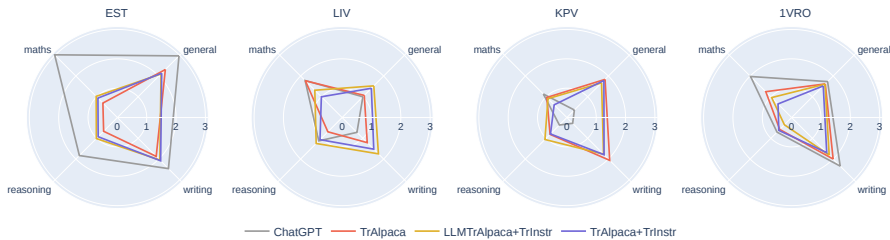


Figure 9. Category-level breakdown of helpfulness scores in human evaluation of conversational LLMs for low-resource Finno-Ugric languages.

Comparisons by category reveals that the scores for GPT-3.5-Turbo are inflated by examples in the maths and reasoning categories, where SMUGRI-MODELS models show less helpfulness. In contrast, SMUGRI-MODELS models perform comparably in the general and writing categories. Notably, in Komi, SMUGRI-MODELS models surpass GPT-3.5-Turbo in both general and writing tasks, while achieving similar scores in the maths and reasoning tasks.

6. CONCLUSION

The goal of this thesis was to investigate how available Estonian resources can be most effectively leveraged to enable cross-lingual knowledge transfer from other languages to Estonian, thereby mitigating the limitations caused by Estonian’s data scarcity. The contributions of this thesis are threefold, addressing this goal from the perspectives of data, language modeling, and evaluation.

Data

From a data perspective, this thesis examined the suitability of synthetic training data for strengthening cross-lingual knowledge transfer to Estonian. While synthetic datasets are considerably less expensive to produce than datasets created with human annotators, they often contain noise, translation artifacts, and other quality issues that may affect model learning.

To study this trade-off, we introduced two new synthetic datasets for Estonian. The first, EstCOPA, targets commonsense reasoning by translating an existing benchmark to Estonian. The second, Alpaca-est, is a general instruction-following dataset generated using a large language model, designed to improve the model’s ability to follow user instructions.

Our experimental results consistently show that training on synthetic data improves models’ ability to process and understand Estonian. However, within the scope of our experiments, replacing synthetic data with higher-quality alternatives did not lead to clear or consistent additional gains.

Finding 1. *Synthetic data can provide substantial benefits during fine-tuning, enabling existing models to better understand a low-resource target language.*

Language Modeling

From a modeling perspective, we conducted an extensive set of experiments with both encoder-only and decoder-only models to understand when and how multilingual data should be incorporated during training.

The first group of experiments focused on identifying the most effective training stages for multilingual data. We found that leveraging multilingual pre-training followed by multilingual fine-tuning consistently outperforms monolingual training approaches. However, the effectiveness of multilingual fine-tuning strongly depends on the quality and capacity of the underlying base model — weaker base models benefit less from additional multilingual supervision.

The second group of experiments explored continued pre-training on Estonian. We show that continued pre-training improves models’ Estonian language understanding, but that increasing the size of the continued pre-training corpus does not yield proportional performance gains, suggesting diminishing returns. Crucially,

stronger general-purpose base models can outperform weaker language-adapted models, highlighting the importance of initial cross-lingual alignment.

Finding 2. *Multilingual training is beneficial for low-resource languages across all stages of model training.*

Finding 3. *Cross-lingual alignment is largely established during pre-training, and its strength strongly determines the effectiveness of downstream adaptation.*

Evaluation

Finally, this thesis addressed challenges related to evaluating models for low-resource languages. We described the limitations of existing evaluation practices and introduced SMUGRI-MT-BENCH, a benchmark designed to assess conversational capabilities in Estonian, Võro, Komi, and Livonian.

Using this benchmark, we conducted a human evaluation measuring both helpfulness and naturalness of various language models. The results indicate that most models—including GPT-3.5-Turbo—struggle to provide helpful responses in these languages (with Estonian being an exception). In contrast, smaller Llama-2 7B-based models were consistently judged to produce more natural-sounding outputs than GPT-3.5-Turbo for Võro, Komi, and Livonian.

Finding 4. *Manual evaluation remains essential for gaining meaningful insights into model behavior.*

Summary

Overall, this thesis demonstrates that the careful integration of multilingual resources remains the most effective strategy for building NLP systems for low-resource languages such as Estonian. Even modest amounts of target-language data, when combined with strong multilingual foundations, can yield capable models. At the same time, robust and carefully designed evaluation is critical to avoid overestimating progress and to ensure that improvements translate into meaningful benefits for speakers of lesser-resourced languages in real-world settings.

Future Work

There are several promising directions for future research.

First, our experiments with LLAMMAS indicate that not all capabilities transfer equally across languages. In particular, reasoning performance for Llama 2 7B degraded after fine-tuning on Estonian. The underlying cause remains unclear. Possible explanations include limitations in the quality of Estonian training data or contamination in English benchmarks that artificially inflate performance. Without detailed knowledge of the original training corpora, drawing final conclusions is difficult.

Second, low-resource languages continue to lack robust benchmarks. However, in the era of large language models, it is increasingly unclear what constitutes a “good” evaluation dataset. Future work could investigate the limitations of current evaluation paradigms and develop benchmarks that better reflect real-world language use and model capabilities.

Finally, our human evaluation shows that even strong models struggle to generate natural-sounding text in extremely low-resource Finno-Ugric languages such as Võro, Livonian, and Komi. However, measuring naturalness remains challenging, and reliable automatic metrics are largely absent. Developing methods to assess and improve output naturalness in low-resource languages is therefore an important direction for future research.

BIBLIOGRAPHY

- Papineni, Kishore, Roukos, Salim, Ward, Todd, and Zhu, Wei-Jing (July 2002). “Bleu: a Method for Automatic Evaluation of Machine Translation”. In: *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*. Philadelphia, Pennsylvania, USA: Association for Computational Linguistics, pp. 311–318. DOI: 10.3115/1073083.1073135. URL: <https://aclanthology.org/P02-1040/>.
- Snover, Matthew, Dorr, Bonnie, Schwartz, Rich, Micciulla, Linnea, and Makhoul, John (Aug. 2006). “A Study of Translation Edit Rate with Targeted Human Annotation”. In: *Proceedings of the 7th Conference of the Association for Machine Translation in the Americas: Technical Papers*. Cambridge, Massachusetts, USA: Association for Machine Translation in the Americas, pp. 223–231. URL: <https://aclanthology.org/2006.amta-papers.25/>.
- Koppel, Moshe and Ordan, Noam (June 2011). “Translationese and Its Dialects”. In: *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*. Portland, Oregon, USA: Association for Computational Linguistics, pp. 1318–1326. URL: <https://aclanthology.org/P11-1132/>.
- Roemmele, Melissa, Bejan, Cosmin Adrian, and Gordon, Andrew S. (2011). “Choice of plausible alternatives: An evaluation of commonsense causal reasoning”. In: *2011 AAI Spring Symposium Series*.
- Tiedemann, Jörg (May 2012). “Parallel Data, Tools and Interfaces in OPUS”. In: *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC’12)*. Istanbul, Turkey: European Language Resources Association (ELRA), pp. 2214–2218. URL: <https://aclanthology.org/L12-1246/>.
- Lison, Pierre and Tiedemann, Jörg (May 2016). “OpenSubtitles2016: Extracting Large Parallel Corpora from Movie and TV Subtitles”. In: *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC’16)*. Portorož, Slovenia: European Language Resources Association (ELRA), pp. 923–929. URL: <https://aclanthology.org/L16-1147/>.
- Rajpurkar, Pranav, Zhang, Jian, Lopyrev, Konstantin, and Liang, Percy (Nov. 2016). “SQuAD: 100,000+ Questions for Machine Comprehension of Text”. In: *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Austin, Texas: Association for Computational Linguistics, pp. 2383–2392. DOI: 10.18653/v1/D16-1264. URL: <https://aclanthology.org/D16-1264>.
- Sennrich, Rico, Haddow, Barry, and Birch, Alexandra (Aug. 2016). “Neural Machine Translation of Rare Words with Subword Units”. In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Berlin, Germany: Association for Computational Linguistics,

- pp. 1715–1725. DOI: 10.18653/v1/P16-1162. URL: <https://aclanthology.org/P16-1162/>.
- Wu, Yonghui, Schuster, Mike, Chen, Zhifeng, Le, Quoc V., Norouzi, Mohammad, Macherey, Wolfgang, Krikun, Maxim, Cao, Yuan, Gao, Qin, Macherey, Klaus, Klingner, Jeff, Shah, Apurva, Johnson, Melvin, Liu, Xiaobing, Kaiser, Łukasz, Gouws, Stephan, Kato, Yoshikiyo, Kudo, Taku, Kazawa, Hideto, Stevens, Keith, Kurian, George, Patil, Nishant, Wang, Wei, Young, Cliff, Smith, Jason, Riesa, Jason, Rudnick, Alex, Vinyals, Oriol, Corrado, Greg, Hughes, Macduff, and Dean, Jeffrey (2016). *Google’s Neural Machine Translation System: Bridging the Gap between Human and Machine Translation*. arXiv: 1609.08144 [cs.CL]. URL: <https://arxiv.org/abs/1609.08144>.
- Bryant, Christopher, Felice, Mariano, and Briscoe, Ted (July 2017). “Automatic Annotation and Evaluation of Error Types for Grammatical Error Correction”. In: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Vancouver, Canada: Association for Computational Linguistics, pp. 793–805. DOI: 10.18653/v1/P17-1074. URL: <https://aclanthology.org/P17-1074/>.
- Joshi, Mandar, Choi, Eunsol, Weld, Daniel, and Zettlemoyer, Luke (July 2017). “TriviaQA: A Large Scale Distantly Supervised Challenge Dataset for Reading Comprehension”. In: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Vancouver, Canada: Association for Computational Linguistics, pp. 1601–1611. DOI: 10.18653/v1/P17-1147. URL: <https://aclanthology.org/P17-1147/>.
- Lin, Zhouhan, Feng, Minwei, Santos, Cícero Nogueira dos, Yu, Mo, Xiang, Bing, Zhou, Bowen, and Bengio, Yoshua (2017). “A Structured Self-Attentive Sentence Embedding”. In: *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*. OpenReview.net. URL: https://openreview.net/forum?id=BJC%5C_jUqxe.
- Vaswani, Ashish, Shazeer, Noam, Parmar, Niki, Uszkoreit, Jakob, Jones, Llion, Gomez, Aidan N, Kaiser, Łukasz, and Polosukhin, Illia (2017). “Attention is All you Need”. In: *Advances in Neural Information Processing Systems*. Vol. 30. Curran Associates, Inc. URL: https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.
- Bojar, Ondřej, Federmann, Christian, Fishel, Mark, Graham, Yvette, Haddow, Barry, Huck, Matthias, Koehn, Philipp, and Monz, Christof (Oct. 2018). “Findings of the 2018 Conference on Machine Translation (WMT18)”. In: *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*. Belgium, Brussels: Association for Computational Linguistics, pp. 272–303. DOI: 10.18653/v1/W18-6401. URL: <https://aclanthology.org/W18-6401/>.

- Conneau, Alexis, Rinott, Ruty, Lample, Guillaume, Williams, Adina, Bowman, Samuel R., Schwenk, Holger, and Stoyanov, Veselin (2018). “XNLI: Evaluating Cross-lingual Sentence Representations”. In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Brussels, Belgium: Association for Computational Linguistics.
- Kudo, Taku and Richardson, John (Nov. 2018). “SentencePiece: A simple and language independent subword tokenizer and detokenizer for Neural Text Processing”. In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Brussels, Belgium: Association for Computational Linguistics, pp. 66–71. DOI: 10.18653/v1/D18-2012. URL: <https://aclanthology.org/D18-2012/>.
- Radford, Alec, Narasimhan, Karthik, Salimans, Tim, and Sutskever, Ilya (2018). *Improving Language Understanding by Generative Pre-Training*. https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf. OpenAI preprint.
- Devlin, Jacob, Chang, Ming-Wei, Lee, Kenton, and Toutanova, Kristina (June 2019). “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding”. In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, pp. 4171–4186. DOI: 10.18653/v1/N19-1423. URL: <https://aclanthology.org/N19-1423/>.
- Lample, Guillaume and Conneau, Alexis (2019). *Cross-lingual Language Model Pretraining*. arXiv: 1901.07291 [cs.CL]. URL: <https://arxiv.org/abs/1901.07291>.
- Liu, Yinhan, Ott, Myle, Goyal, Naman, Du, Jingfei, Joshi, Mandar, Chen, Danqi, Levy, Omer, Lewis, Mike, Zettlemoyer, Luke, and Stoyanov, Veselin (2019). *RoBERTa: A Robustly Optimized BERT Pretraining Approach*. arXiv: 1907.11692 [cs.CL]. URL: <https://arxiv.org/abs/1907.11692>.
- Pires, Telmo, Schlinger, Eva, and Garrette, Dan (2019). *How multilingual is Multilingual BERT?* arXiv: 1906.01502 [cs.CL]. URL: <https://arxiv.org/abs/1906.01502>.
- Radford, Alec, Wu, Jeffrey, Child, Rewon, Luan, David, Amodei, Dario, and Sutskever, Ilya (2019). *Language models are unsupervised multitask learners*.
- Sap, Maarten, Rashkin, Hannah, Chen, Derek, Le Bras, Ronan, and Choi, Yejin (Nov. 2019). “Social IQa: Commonsense Reasoning about Social Interactions”. 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)*. Hong Kong, China: Association for Computational Linguistics, pp. 4463–4473. DOI: 10.18653/v1/D19-1454. URL: <https://aclanthology.org/D19-1454/>.

- Schwenk, Holger, Wenzek, Guillaume, Edunov, Sergey, Grave, Edouard, and Joulin, Armand (2019). “Ccmatrix: Mining billions of high-quality parallel sentences on the web”. In: *arXiv preprint arXiv:1911.04944*.
- Wang, Alex, Pruksachatkun, Yada, Nangia, Nikita, Singh, Amanpreet, Michael, Julian, Hill, Felix, Levy, Omer, and Bowman, Samuel (2019). “SuperGLUE: A Stickier Benchmark for General-Purpose Language Understanding Systems”. In: *Advances in Neural Information Processing Systems*. Vol. 32. Curran Associates, Inc. URL: https://proceedings.neurips.cc/paper_files/paper/2019/file/4496bf24afe7fab6f046bf4923da8de6-Paper.pdf.
- Zellers, Rowan, Holtzman, Ari, Bisk, Yonatan, Farhadi, Ali, and Choi, Yejin (July 2019). “HellaSwag: Can a Machine Really Finish Your Sentence?” In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics, pp. 4791–4800. DOI: 10.18653/v1/P19-1472. URL: <https://aclanthology.org/P19-1472/>.
- Artetxe, Mikel, Labaka, Gorka, and Agirre, Eneko (Nov. 2020a). “Translation Artifacts in Cross-lingual Transfer Learning”. In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Online: Association for Computational Linguistics, pp. 7674–7684. DOI: 10.18653/v1/2020.emnlp-main.618. URL: <https://aclanthology.org/2020.emnlp-main.618/>.
- Artetxe, Mikel, Ruder, Sebastian, and Yogatama, Dani (July 2020b). “On the Cross-lingual Transferability of Monolingual Representations”. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, pp. 4623–4637. DOI: 10.18653/v1/2020.acl-main.421. URL: <https://aclanthology.org/2020.acl-main.421/>.
- Artetxe, Mikel, Ruder, Sebastian, and Yogatama, Dani (2020c). “On the Cross-lingual Transferability of Monolingual Representations”. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics. DOI: 10.18653/v1/2020.acl-main.421. URL: <http://dx.doi.org/10.18653/v1/2020.acl-main.421>.
- Aulamo, Mikko, Virpioja, Sami, and Tiedemann, Jörg (July 2020). “OpusFilter: A Configurable Parallel Corpus Filtering Toolbox”. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*. Online: Association for Computational Linguistics, pp. 150–156. DOI: 10.18653/v1/2020.acl-demos.20. URL: <https://aclanthology.org/2020.acl-demos.20/>.
- Brown, Tom, Mann, Benjamin, Ryder, Nick, Subbiah, Melanie, Kaplan, Jared D, Dhariwal, Prafulla, Neelakantan, Arvind, Shyam, Pranav, Sastry, Girish, Askell, Amanda, Agarwal, Sandhini, Herbert-Voss, Ariel, Krueger, Gretchen, Henighan, Tom, Child, Rewon, Ramesh, Aditya, Ziegler, Daniel, Wu, Jeffrey, Winter, Clemens, Hesse, Chris, Chen, Mark, Sigler, Eric, Litwin, Mateusz,

- Gray, Scott, Chess, Benjamin, Clark, Jack, Berner, Christopher, McCandlish, Sam, Radford, Alec, Sutskever, Ilya, and Amodei, Dario (2020). “Language Models are Few-Shot Learners”. In: *Advances in Neural Information Processing Systems*. Vol. 33. Curran Associates, Inc., pp. 1877–1901. URL: https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfc4967418bfb8ac142f64a-Paper.pdf.
- Clark, Jonathan H., Choi, Eunsol, Collins, Michael, Garrette, Dan, Kwiatkowski, Tom, Nikolaev, Vitaly, and Palomaki, Jennimaria (2020). “TyDi QA: A Benchmark for Information-Seeking Question Answering in Typologically Diverse Languages”. In: *Transactions of the Association for Computational Linguistics*.
- Conneau, Alexis, Khandelwal, Kartikay, Goyal, Naman, Chaudhary, Vishrav, Wenzek, Guillaume, Guzmán, Francisco, Grave, Edouard, Ott, Myle, Zettlemoyer, Luke, and Stoyanov, Veselin (July 2020). “Unsupervised Cross-lingual Representation Learning at Scale”. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, pp. 8440–8451. DOI: 10.18653/v1/2020.acl-main.747. URL: <https://aclanthology.org/2020.acl-main.747/>.
- Jalili Sabet, Masoud, Dufter, Philipp, Yvon, François, and Schütze, Hinrich (Nov. 2020). “SimAlign: High Quality Word Alignments Without Parallel Training Data Using Static and Contextualized Embeddings”. In: *Findings of the Association for Computational Linguistics: EMNLP 2020*. Online: Association for Computational Linguistics, pp. 1627–1643. DOI: 10.18653/v1/2020.findings-emnlp.147. URL: <https://aclanthology.org/2020.findings-emnlp.147/>.
- Kaplan, Jared, McCandlish, Sam, Henighan, Tom, Brown, Tom B, Chess, Benjamin, Child, Rewon, Gray, Scott, Radford, Alec, Wu, Jeffrey, and Amodei, Dario (2020). “Scaling laws for neural language models”. In: *arXiv preprint arXiv:2001.08361*.
- Liu, Yinhan, Gu, Jiatao, Goyal, Naman, Li, Xian, Edunov, Sergey, Ghazvininejad, Marjan, Lewis, Mike, and Zettlemoyer, Luke (2020). *Multilingual Denoising Pre-training for Neural Machine Translation*. arXiv: 2001.08210 [cs.CL]. URL: <https://arxiv.org/abs/2001.08210>.
- Ponti, Edoardo Maria, Glavaš, Goran, Majewska, Olga, Liu, Qianchu, Vulić, Ivan, and Korhonen, Anna (Nov. 2020). “XCOPA: A Multilingual Dataset for Causal Commonsense Reasoning”. In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Online: Association for Computational Linguistics, pp. 2362–2376. DOI: 10.18653/v1/2020.emnlp-main.185. URL: <https://aclanthology.org/2020.emnlp-main.185/>.
- Rasley, Jeff, Rajbhandari, Samyam, Ruwase, Olatunji, and He, Yuxiong (2020). “DeepSpeed: System Optimizations Enable Training Deep Learning Models with Over 100 Billion Parameters”. In: *Proceedings of the 26th ACM SIGKDD*

- International Conference on Knowledge Discovery & Data Mining*. KDD '20. Virtual Event, CA, USA: Association for Computing Machinery, pp. 3505–3506. ISBN: 9781450379984. DOI: 10.1145/3394486.3406703. URL: <https://doi.org/10.1145/3394486.3406703>.
- Shoeybi, Mohammad, Patwary, Mostofa, Puri, Raul, LeGresley, Patrick, Casper, Jared, and Catanzaro, Bryan (2020). *Megatron-LM: Training Multi-Billion Parameter Language Models Using Model Parallelism*. arXiv: 1909.08053 [cs.CL]. URL: <https://arxiv.org/abs/1909.08053>.
- Zhang*, Tianyi, Kishore*, Varsha, Wu*, Felix, Weinberger, Kilian Q., and Artzi, Yoav (2020). “BERTScore: Evaluating Text Generation with BERT”. In: *International Conference on Learning Representations*. URL: <https://openreview.net/forum?id=SkeHuCVFDr>.
- Hendrycks, Dan, Burns, Collin, Basart, Steven, Zou, Andy, Mazeika, Mantas, Song, Dawn, and Steinhardt, Jacob (2021). “Measuring Massive Multitask Language Understanding”. In: *Proceedings of the International Conference on Learning Representations (ICLR)*.
- Käver, Anu (2021a). “Extractive Question Answering for Estonian Language”. MA thesis. Tallinn University of Technology (TalTech).
- Käver, Anu (2021b). “Extractive Question Answering for Estonian Language”. MA thesis. Tallinn University of Technology.
- Sakaguchi, Keisuke, Bras, Ronan Le, Bhagavatula, Chandra, and Choi, Yejin (Aug. 2021). “WinoGrande: an adversarial winograd schema challenge at scale”. In: *Commun. ACM* 64.9, pp. 99–106. ISSN: 0001-0782. DOI: 10.1145/3474381. URL: <https://doi.org/10.1145/3474381>.
- Schwenk, Holger, Chaudhary, Vishrav, Sun, Shuo, Gong, Hongyu, and Guzmán, Francisco (Apr. 2021). “WikiMatrix: Mining 135M Parallel Sentences in 1620 Language Pairs from Wikipedia”. In: *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*. Online: Association for Computational Linguistics, pp. 1351–1361. DOI: 10.18653/v1/2021.eacl-main.115. URL: <https://aclanthology.org/2021.eacl-main.115/>.
- Tanvir, Hasan, Kittask, Claudia, Eiche, Sandra, and Sirts, Kairit (May 2021). “EstBERT: A Pretrained Language-Specific BERT for Estonian”. In: *Proceedings of the 23rd Nordic Conference on Computational Linguistics (NoDaLiDa)*. Reykjavik, Iceland (Online): Linköping University Electronic Press, Sweden, pp. 11–19. URL: <https://aclanthology.org/2021.nodalida-main.2/>.
- Xue, Linting, Constant, Noah, Roberts, Adam, Kale, Mihir, Al-Rfou, Rami, Siddhant, Aditya, Barua, Aditya, and Raffel, Colin (June 2021). “mT5: A Massively Multilingual Pre-trained Text-to-Text Transformer”. In: *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Online: Association for Computational Linguistics, pp. 483–498. DOI: 10.18653/v1/2021.n

- aacl-main.41. URL: <https://aclanthology.org/2021.naacl-main.41/>.
- Bergmanis, Toms, Pinnis, Marcis, Rozis, Roberts, Šlapiņš, Jānis, Šics, Valters, Bernāne, Berta, Pužulis, Guntars, Titomers, Endijs, Tättar, Andre, Purason, Taido, **Kuulmets, Hele-Andra**, Luhtaru, Agnes, Rätsep, Liisa, Tars, Maali, Laumets-Tättar, Annika, and Fishel, Mark (June 2022). “MTee: Open Machine Translation Platform for Estonian Government”. In: *Proceedings of the 23rd Annual Conference of the European Association for Machine Translation*. Ghent, Belgium: European Association for Machine Translation, pp. 309–310. URL: <https://aclanthology.org/2022.eamt-1.44/>.
- Chowdhery, Aakanksha, Narang, Sharan, Devlin, Jacob, Bosma, Maarten, Mishra, Gaurav, Roberts, Adam, Barham, Paul, Chung, Hyung Won, Sutton, Charles, Gehrmann, Sebastian, Schuh, Parker, Shi, Kensen, Tsvyashchenko, Sasha, Maynez, Joshua, Rao, Abhishek, Barnes, Parker, Tay, Yi, Shazeer, Noam, Prabhakaran, Vinodkumar, Reif, Emily, Du, Nan, Hutchinson, Ben, Pope, Reiner, Bradbury, James, Austin, Jacob, Isard, Michael, Gur-Ari, Guy, Yin, Pengcheng, Duke, Toju, Levskaya, Anselm, Ghemawat, Sanjay, Dev, Sunipa, Michalewski, Henryk, Garcia, Xavier, Misra, Vedant, Robinson, Kevin, Fedus, Liam, Zhou, Denny, Ippolito, Daphne, Luan, David, Lim, Hyeontaek, Zoph, Barret, Spiridonov, Alexander, Sepassi, Ryan, Dohan, David, Agrawal, Shivani, Omernick, Mark, Dai, Andrew M., Pillai, Thanumalayan Sankaranarayana, Pellat, Marie, Lewkowycz, Aitor, Moreira, Erica, Child, Rewon, Polozov, Oleksandr, Lee, Katherine, Zhou, Zongwei, Wang, Xuezhi, Saeta, Brennan, Diaz, Mark, Firat, Orhan, Catasta, Michele, Wei, Jason, Meier-Hellstern, Kathy, Eck, Douglas, Dean, Jeff, Petrov, Slav, and Fiedel, Noah (2022). *PaLM: Scaling Language Modeling with Pathways*. arXiv: 2204.02311 [cs.CL]. URL: <https://arxiv.org/abs/2204.02311>.
- Hoffmann, Jordan, Borgeaud, Sebastian, Mensch, Arthur, Buchatskaya, Elena, Cai, Trevor, Rutherford, Eliza, Las Casas, Diego de, Hendricks, Lisa Anne, Welbl, Johannes, Clark, Aidan, Hennigan, Thomas, Noland, Eric, Millican, Katherine, Driessche, George van den, Damoc, Bogdan, Guy, Aurelia, Osindero, Simon, Simonyan, Karén, Elsen, Erich, Vinyals, Oriol, Rae, Jack, and Sifre, Laurent (2022). “An empirical analysis of compute-optimal large language model training”. In: *Advances in Neural Information Processing Systems*. Vol. 35. Curran Associates, Inc., pp. 30016–30030. URL: https://proceedings.neurips.cc/paper_files/paper/2022/file/c1e2faff6f588870935f114ebe04a3e5-Paper-Conference.pdf.
- Kojima, Takeshi, Gu, Shixiang (Shane), Reid, Machel, Matsuo, Yutaka, and Iwasawa, Yusuke (2022). “Large Language Models are Zero-Shot Reasoners”. In: *Advances in Neural Information Processing Systems*. Vol. 35. Curran Associates, Inc., pp. 22199–22213. URL: https://proceedings.neurips.cc/paper_files/paper/2022/file/8bb0d291acd4acf06ef112099c16f326-Paper-Conference.pdf.

- Kuulmets, Hele-Andra**, Tättar, Andre, and Fishel, Mark (Oct. 2022). “Estonian Language Understanding: a Case Study on the COPA Task”. In: *Proceedings of Baltic HLT 2022*. Vol. 10. Riga, Latvia: Baltic Journal of Modern Computing, pp. 470–480. DOI: 10.22364/bjmc.2022.10.3.19.
- Lin, Stephanie, Hilton, Jacob, and Evans, Owain (May 2022a). “TruthfulQA: Measuring How Models Mimic Human Falsehoods”. In: *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Dublin, Ireland: Association for Computational Linguistics, pp. 3214–3252. DOI: 10.18653/v1/2022.acl-long.229. URL: <https://aclanthology.org/2022.acl-long.229/>.
- Lin, Xi Victoria, Mihaylov, Todor, Artetxe, Mikel, Wang, Tianlu, Chen, Shuo-hui, Simig, Daniel, Ott, Myle, Goyal, Naman, Bhosale, Shruti, Du, Jingfei, Pasunuru, Ramakanth, Shleifer, Sam, Chaudhary, Vishrav, O’Horo, Brian, Wang, Jeff, Zettlemoyer, Luke, Kozareva, Zornitsa, Diab, Mona, Stoyanov, Veselin, and Li, Xian (Dec. 2022b). “Few-shot Learning with Multilingual Generative Language Models”. In: *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. Abu Dhabi, United Arab Emirates: Association for Computational Linguistics, pp. 9019–9052. DOI: 10.18653/v1/2022.emnlp-main.616. URL: <https://aclanthology.org/2022.emnlp-main.616/>.
- Mishra, Swaroop, Khashabi, Daniel, Baral, Chitta, and Hajishirzi, Hannaneh (2022). *Cross-Task Generalization via Natural Language Crowdsourcing Instructions*. arXiv: 2104.08773 [cs.CL]. URL: <https://arxiv.org/abs/2104.08773>.
- Ouyang, Long, Wu, Jeffrey, Jiang, Xu, Almeida, Diogo, Wainwright, Carroll, Mishkin, Pamela, Zhang, Chong, Agarwal, Sandhini, Slama, Katarina, Ray, Alex, Schulman, John, Hilton, Jacob, Kelton, Fraser, Miller, Luke, Simens, Maddie, Askell, Amanda, Welinder, Peter, Christiano, Paul F, Leike, Jan, and Lowe, Ryan (2022). “Training language models to follow instructions with human feedback”. In: *Advances in Neural Information Processing Systems*. Vol. 35. Curran Associates, Inc., pp. 27730–27744. URL: https://proceedings.neurips.cc/paper_files/paper/2022/file/b1efde53be364a73914f58805a001731-Paper-Conference.pdf.
- Rae, Jack W., Borgeaud, Sebastian, Cai, Trevor, Millican, Katie, Hoffmann, Jordan, Song, Francis, Aslanides, John, Henderson, Sarah, Ring, Roman, Young, Susannah, Rutherford, Eliza, Hennigan, Tom, Menick, Jacob, Cassirer, Albin, Powell, Richard, Driessche, George van den, Hendricks, Lisa Anne, Rauh, Maribeth, Huang, Po-Sen, Glaese, Amelia, Welbl, Johannes, Dathathri, Sumanth, Huang, Saffron, Uesato, Jonathan, Mellor, John, Higgins, Irina, Creswell, Antonia, McAleese, Nat, Wu, Amy, Elsen, Erich, Jayakumar, Siddhant, Buchatskaya, Elena, Budden, David, Sutherland, Esme, Simonyan, Karen, Paganini, Michela, Sifre, Laurent, Martens, Lena, Li, Xiang Lorraine, Kuncoro, Adhiguna, Nematzadeh, Aida, Gribovskaya, Elena, Donato, Domenic, Lazaridou, Angeliki, Mensch, Arthur, Lespiau, Jean-Baptiste, Tsimpoukelli, Maria, Grigorev, Niko-

- lai, Fritz, Doug, Sottiaux, Thibault, Pajarskas, Mantas, Pohlen, Toby, Gong, Zhitao, Toyama, Daniel, Masson d’Autume, Cyprien de, Li, Yujia, Terzi, Tayfun, Mikulik, Vladimir, Babuschkin, Igor, Clark, Aidan, Las Casas, Diego de, Guy, Aurelia, Jones, Chris, Bradbury, James, Johnson, Matthew, Hechtman, Blake, Weidinger, Laura, Gabriel, Iason, Isaac, William, Lockhart, Ed, Osindero, Simon, Rimell, Laura, Dyer, Chris, Vinyals, Oriol, Ayoub, Kareem, Stanway, Jeff, Bennett, Lorraine, Hassabis, Demis, Kavukcuoglu, Koray, and Irving, Geoffrey (2022). *Scaling Language Models: Methods, Analysis & Insights from Training Gopher*. arXiv: 2112.11446 [cs.CL]. URL: <https://arxiv.org/abs/2112.11446>.
- Sanh, Victor, Webson, Albert, Raffel, Colin, Bach, Stephen, Sutawika, Lintang, Alyafeai, Zaid, Chaffin, Antoine, Stiegler, Arnaud, Raja, Arun, Dey, Manan, Bari, M Saiful, Xu, Canwen, Thakker, Urmish, Sharma, Shanya Sharma, Szczechla, Eliza, Kim, Taewoon, Chhablani, Gunjan, Nayak, Nihal, Datta, Debajyoti, Chang, Jonathan, Jiang, Mike Tian-Jian, Wang, Han, Manica, Matteo, Shen, Sheng, Yong, Zheng Xin, Pandey, Harshit, Bawden, Rachel, Wang, Thomas, Neeraj, Trishala, Rozen, Jos, Sharma, Abheesht, Santilli, Andrea, Fevry, Thibault, Fries, Jason Alan, Teehan, Ryan, Scao, Teven Le, Biderman, Stella, Gao, Leo, Wolf, Thomas, and Rush, Alexander M (2022). “Multitask Prompted Training Enables Zero-Shot Task Generalization”. In: *International Conference on Learning Representations*. URL: <https://openreview.net/forum?id=9Vrb9DOWI4>.
- Smith, Shaden, Patwary, Mostofa, Norick, Brandon, LeGresley, Patrick, Rajbhandari, Samyam, Casper, Jared, Liu, Zhun, Prabhunoye, Shrimai, Zerveas, George, Korthikanti, Vijay, Zhang, Elton, Child, Rewon, Aminabadi, Reza Yazdani, Bernauer, Julie, Song, Xia, Shoeybi, Mohammad, He, Yuxiong, Houston, Michael, Tiwary, Saurabh, and Catanzaro, Bryan (2022). *Using DeepSpeed and Megatron to Train Megatron-Turing NLG 530B, A Large-Scale Generative Language Model*. arXiv: 2201.11990 [cs.CL]. URL: <https://arxiv.org/abs/2201.11990>.
- Tättar, Andre, Purason, Taido, **Kuulmets, Hele-Andra**, Luhtaru, Agnes, Rätsep, Liisa, Tars, Maali, Pinnis, Mārcis, Bergmanis, Toms, and Fishel, Mark (2022). “Open and Competitive Multilingual Neural Machine Translation in Production”. In: *Baltic Journal of Modern Computing* Vol. 10, No. 3, pp. 422–434.
- Team, NLLB, Costa-jussà, Marta R., Cross, James, Çelebi, Onur, Elbayad, Maha, Heafield, Kenneth, Heffernan, Kevin, Kalbassi, Elahe, Lam, Janice, Licht, Daniel, Maillard, Jean, Sun, Anna, Wang, Skyler, Wenzek, Guillaume, Youngblood, Al, Akula, Bapi, Barrault, Loic, Gonzalez, Gabriel Mejia, Hansanti, Prangthip, Hoffman, John, Jarrett, Semarley, Sadagopan, Kaushik Ram, Rowe, Dirk, Spruit, Shannon, Tran, Chau, Andrews, Pierre, Ayan, Necip Fazil, Bhosale, Shruti, Edunov, Sergey, Fan, Angela, Gao, Cynthia, Goswami, Vedanuj, Guzmán, Francisco, Koehn, Philipp, Mourachko, Alexandre, Ropers, Christophe, Saleem, Safiyyah, Schwenk, Holger, and Wang, Jeff (2022). *No Language Left Behind:*

- Scaling Human-Centered Machine Translation*. arXiv: 2207.04672 [cs.CL]. URL: <https://arxiv.org/abs/2207.04672>.
- Thoppilan, Romal, Freitas, Daniel De, Hall, Jamie, Shazeer, Noam, Kulshreshtha, Apoorv, Cheng, Heng-Tze, Jin, Alicia, Bos, Taylor, Baker, Leslie, Du, Yu, Li, YaGuang, Lee, Hongrae, Zheng, Huaixiu Steven, Ghafouri, Amin, Menegali, Marcelo, Huang, Yanping, Krikun, Maxim, Lepikhin, Dmitry, Qin, James, Chen, Dehao, Xu, Yuanzhong, Chen, Zhifeng, Roberts, Adam, Bosma, Maarten, Zhao, Vincent, Zhou, Yanqi, Chang, Chung-Ching, Krivokon, Igor, Rusch, Will, Pickett, Marc, Srinivasan, Pranesh, Man, Laichee, Meier-Hellstern, Kathleen, Morris, Meredith Ringel, Doshi, Tulse, Santos, Renelito Delos, Duke, Toju, Soraker, Johnny, Zevenbergen, Ben, Prabhakaran, Vinodkumar, Diaz, Mark, Hutchinson, Ben, Olson, Kristen, Molina, Alejandra, Hoffman-John, Erin, Lee, Josh, Aroyo, Lora, Rajakumar, Ravi, Butryna, Alena, Lamm, Matthew, Kuzmina, Viktoriya, Fenton, Joe, Cohen, Aaron, Bernstein, Rachel, Kurzweil, Ray, Aguera-Arcas, Blaise, Cui, Claire, Croak, Marian, Chi, Ed, and Le, Quoc (2022). *LaMDA: Language Models for Dialog Applications*. arXiv: 2201.08239 [cs.CL]. URL: <https://arxiv.org/abs/2201.08239>.
- Wei, Jason, Bosma, Maarten, Zhao, Vincent, Guu, Kelvin, Yu, Adams Wei, Lester, Brian, Du, Nan, Dai, Andrew M., and Le, Quoc V (2022). “Finetuned Language Models are Zero-Shot Learners”. In: *International Conference on Learning Representations*. URL: <https://openreview.net/forum?id=gEZrGCozdqR>.
- Almazrouei, Ebtesam, Alobeidli, Hamza, Alshamsi, Abdulaziz, Cappelli, Alessandro, Cojocar, Ruxandra, Debbah, M erouane, Goffinet,  tienne, Hesslow, Daniel, Launay, Julien, Malartic, Quentin, Mazzotta, Daniele, Noun, Badreddine, Pannier, Baptiste, and Penedo, Guilherme (2023). “The Falcon Series of Open Language Models”. In: *CoRR* abs/2311.16867. URL: <https://doi.org/10.48550/arXiv.2311.16867>.
- Anil, Rohan et al. (2023). *PaLM 2 Technical Report*. arXiv: 2305.10403 [cs.CL]. URL: <https://arxiv.org/abs/2305.10403>.
- Bai, Jinze, Bai, Shuai, Chu, Yunfei, Cui, Zeyu, Dang, Kai, Deng, Xiaodong, Fan, Yang, Ge, Wenbin, Han, Yu, Huang, Fei, Hui, Binyuan, Ji, Luo, Li, Mei, Lin, Junyang, Lin, Runji, Liu, Dayiheng, Liu, Gao, Lu, Chengqiang, Lu, Keming, Ma, Jianxin, Men, Rui, Ren, Xingzhang, Ren, Xuancheng, Tan, Chuanqi, Tan, Sinan, Tu, Jianhong, Wang, Peng, Wang, Shijie, Wang, Wei, Wu, Shengguang, Xu, Benfeng, Xu, Jin, Yang, An, Yang, Hao, Yang, Jian, Yang, Shusheng, Yao, Yang, Yu, Bowen, Yuan, Hongyi, Yuan, Zheng, Zhang, Jianwei, Zhang, Xingxuan, Zhang, Yichang, Zhang, Zhenru, Zhou, Chang, Zhou, Jingren, Zhou, Xiaohuan, and Zhu, Tianhang (2023). *Qwen Technical Report*. arXiv: 2309.16609 [cs.CL]. URL: <https://arxiv.org/abs/2309.16609>.
- Briakou, Eleftheria, Cherry, Colin, and Foster, George (July 2023). “Searching for Needles in a Haystack: On the Role of Incidental Bilingualism in PaLM’s Translation Capability”. In: *Proceedings of the 61st Annual Meeting of the*

- Association for Computational Linguistics (Volume 1: Long Papers)*. Toronto, Canada: Association for Computational Linguistics, pp. 9432–9452. DOI: 10.18653/v1/2023.acl-long.524. URL: <https://aclanthology.org/2023.acl-long.524/>.
- Huang, Haoyang, Tang, Tianyi, Zhang, Dongdong, Zhao, Xin, Song, Ting, Xia, Yan, and Wei, Furu (Dec. 2023). “Not All Languages Are Created Equal in LLMs: Improving Multilingual Capability by Cross-Lingual-Thought Prompting”. In: *Findings of the Association for Computational Linguistics: EMNLP 2023*. Singapore: Association for Computational Linguistics, pp. 12365–12394. DOI: 10.18653/v1/2023.findings-emnlp.826. URL: <https://aclanthology.org/2023.findings-emnlp.826/>.
- Köpf, Andreas, Kilcher, Yannic, Rütte, Dimitri von, Anagnostidis, Sotiris, Tam, Zhi Rui, Stevens, Keith, Barhoum, Abdullah, Nguyen, Duc, Stanley, Oliver, Nagyfi, Richárd, ES, Shahul, Suri, Sameer, Glushkov, David, Dantuluri, Arnav, Maguire, Andrew, Schuhmann, Christoph, Nguyen, Huu, and Mattick, Alexander (2023). “OpenAssistant Conversations - Democratizing Large Language Model Alignment”. In: *Advances in Neural Information Processing Systems*. Vol. 36. Curran Associates, Inc., pp. 47669–47681. URL: https://proceedings.neurips.cc/paper_files/paper/2023/file/949f0f8f32267d297c2d4e3ee10a2e7e-Paper-Datasets_and_Benchmarks.pdf.
- Koppel, Kristina, Kallas, Jelena, Jürviste, Madis, and Kaljumäe, Helen (2023). *Eesti keele ühendkorpus 2023*. In Estonian.
- Kuulmets, Hele-Andra** and Fishel, Mark (May 2023). “Translated Benchmarks Can Be Misleading: the Case of Estonian Question Answering”. In: *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*. Tórshavn, Faroe Islands: University of Tartu Library, pp. 710–716. URL: <https://aclanthology.org/2023.nodalida-1.71/>.
- Longpre, Shayne, Hou, Le, Vu, Tu, Webson, Albert, Chung, Hyung Won, Tay, Yi, Zhou, Denny, Le, Quoc V, Zoph, Barret, Wei, Jason, and Roberts, Adam (July 2023). “The Flan Collection: Designing Data and Methods for Effective Instruction Tuning”. In: *Proceedings of the 40th International Conference on Machine Learning*. Vol. 202. Proceedings of Machine Learning Research. PMLR, pp. 22631–22648. URL: <https://proceedings.mlr.press/v202/longpre23a.html>.
- Shi, Freda, Suzgun, Mirac, Freitag, Markus, Wang, Xuezhi, Srivats, Suraj, Vosoughi, Soroush, Chung, Hyung Won, Tay, Yi, Ruder, Sebastian, Zhou, Denny, Das, Dipanjan, and Wei, Jason (2023). “Language models are multilingual chain-of-thought reasoners”. In: *The Eleventh International Conference on Learning Representations*. URL: <https://openreview.net/forum?id=fR3wGck-IXp>.
- Taori, Rohan, Gulrajani, Ishaan, Zhang, Tianyi, Dubois, Yann, Li, Xuechen, Guestrin, Carlos, Liang, Percy, and Hashimoto, Tatsunori B. (2023). *Stanford Alpaca*:

- An Instruction-following LLaMA model*. https://github.com/tatsu-lab/stanford_alpaca.
- Touvron, Hugo, Lavril, Thibaut, Izacard, Gautier, Martinet, Xavier, Lachaux, Marie-Anne, Lacroix, Timothée, Rozière, Baptiste, Goyal, Naman, Hambro, Eric, Azhar, Faisal, Rodriguez, Aurelien, Joulin, Armand, Grave, Edouard, and Lample, Guillaume (2023a). *LLaMA: Open and Efficient Foundation Language Models*. arXiv: 2302.13971 [cs.CL]. URL: <https://arxiv.org/abs/2302.13971>.
- Touvron, Hugo, Martin, Louis, Stone, Kevin, Albert, Peter, Almahairi, Amjad, Babaei, Yasmine, Bashlykov, Nikolay, Batra, Soumya, Bhargava, Prajjwal, Bhosale, Shruti, Bikel, Dan, Blecher, Lukas, Ferrer, Cristian Canton, Chen, Moya, Cucurull, Guillem, Esiobu, David, Fernandes, Jude, Fu, Jeremy, Fu, Wenyin, Fuller, Brian, Gao, Cynthia, Goswami, Vedanuj, Goyal, Naman, Hartshorn, Anthony, Hosseini, Saghar, Hou, Rui, Inan, Hakan, Kardas, Marcin, Kerkez, Viktor, Khabsa, Madian, Kloumann, Isabel, Korenev, Artem, Koura, Punit Singh, Lachaux, Marie-Anne, Lavril, Thibaut, Lee, Jenya, Liskovich, Diana, Lu, Yinghai, Mao, Yuning, Martinet, Xavier, Mihaylov, Todor, Mishra, Pushkar, Molybog, Igor, Nie, Yixin, Poulton, Andrew, Reizenstein, Jeremy, Rungta, Rashi, Saladi, Kalyan, Schelten, Alan, Silva, Ruan, Smith, Eric Michael, Subramanian, Ranjan, Tan, Xiaoqing Ellen, Tang, Binh, Taylor, Ross, Williams, Adina, Kuan, Jian Xiang, Xu, Puxin, Yan, Zheng, Zarov, Iliyan, Zhang, Yuchen, Fan, Angela, Kambadur, Melanie, Narang, Sharan, Rodriguez, Aurelien, Stojnic, Robert, Edunov, Sergey, and Scialom, Thomas (2023b). *Llama 2: Open Foundation and Fine-Tuned Chat Models*. arXiv: 2307.09288 [cs.CL]. URL: <https://arxiv.org/abs/2307.09288>.
- Wang, Yizhong, Kordi, Yeganeh, Mishra, Swaroop, Liu, Alisa, Smith, Noah A., Khashabi, Daniel, and Hajishirzi, Hannaneh (2023). *Self-Instruct: Aligning Language Models with Self-Generated Instructions*. arXiv: 2212.10560 [cs.CL]. URL: <https://arxiv.org/abs/2212.10560>.
- Yankovskaya, Lisa, Tars, Maali, Tättar, Andre, and Fishel, Mark (2023). “Machine Translation for Low-resource Finno-Ugric Languages”. In: *The 24rd Nordic Conference on Computational Linguistics*.
- Zhang, Xiang, Li, Senyu, Hauer, Bradley, Shi, Ning, and Kondrak, Grzegorz (Dec. 2023). “Don’t Trust ChatGPT when your Question is not in English: A Study of Multilingual Abilities and Types of LLMs”. In: *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. Singapore: Association for Computational Linguistics, pp. 7915–7927. DOI: 10.18653/v1/2023.emnlp-main.491. URL: <https://aclanthology.org/2023.emnlp-main.491/>.
- Zheng, Lianmin, Chiang, Wei-Lin, Sheng, Ying, Zhuang, Siyuan, Wu, Zhanghao, Zhuang, Yonghao, Lin, Zi, Li, Zhuohan, Li, Dacheng, Xing, Eric, Zhang, Hao, Gonzalez, Joseph E, and Stoica, Ion (2023). “Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena”. In: *Advances in Neural Information Process-*

- ing Systems*. Vol. 36. Curran Associates, Inc., pp. 46595–46623. URL: https://proceedings.neurips.cc/paper_files/paper/2023/file/91f18a1287b398d378ef22505bf41832-Paper-Datasets_and_Benchmarks.pdf.
- Zhou, Chunting, Liu, Pengfei, Xu, Puxin, Iyer, Srinivasan, Sun, Jiao, Mao, Yun-
ing, Ma, Xuezhe, Efrat, Avia, Yu, Ping, Yu, Lili, Zhang, Susan, Ghosh, Gargi,
Lewis, Mike, Zettlemoyer, Luke, and Levy, Omer (2023a). “LIMA: Less Is
More for Alignment”. In: *NeurIPS*. URL: http://papers.nips.cc/paper_files/paper/2023/hash/ac662d74829e4407ce1d126477f4a03a-Abstract-Conference.html.
- Zhou, Jeffrey, Lu, Tianjian, Mishra, Swaroop, Brahma, Siddhartha, Basu, Sujoy,
Luan, Yi, Zhou, Denny, and Hou, Le (2023b). “Instruction-Following Evalu-
ation for Large Language Models”. In: *CoRR* abs/2311.07911. URL: <https://doi.org/10.48550/arXiv.2311.07911>.
- Adelani, David, Liu, Hannah, Shen, Xiaoyu, Vassilyev, Nikita, Alabi, Jesujoba,
Mao, Yanke, Gao, Haonan, and Lee, En-Shiun (Mar. 2024). “SIB-200: A Sim-
ple, Inclusive, and Big Evaluation Dataset for Topic Classification in 200+
Languages and Dialects”. In: *Proceedings of the 18th Conference of the Eu-
ropean Chapter of the Association for Computational Linguistics (Volume 1:
Long Papers)*. St. Julian’s, Malta: Association for Computational Linguistics,
pp. 226–245. URL: <https://aclanthology.org/2024.eacl-long.14>.
- Bandarkar, Lucas, Liang, Davis, Muller, Benjamin, Artetxe, Mikel, Shukla, Satya
Narayan, Husa, Donald, Goyal, Naman, Krishnan, Abhinandan, Zettlemoyer,
Luke, and Khabsa, Madian (Aug. 2024). “The Belebele Benchmark: a Parallel
Reading Comprehension Dataset in 122 Language Variants”. In: *Proceedings
of the 62nd Annual Meeting of the Association for Computational Linguistics
(Volume 1: Long Papers)*. Bangkok, Thailand: Association for Computational
Linguistics, pp. 749–775. DOI: 10.18653/v1/2024.acl-long.44. URL:
<https://aclanthology.org/2024.acl-long.44>.
- Chen, Guiming Hardy, Chen, Shunian, Liu, Ziche, Jiang, Feng, and Wang, Benyou
(Nov. 2024). “Humans or LLMs as the Judge? A Study on Judgement Bias”.
In: *Proceedings of the 2024 Conference on Empirical Methods in Natural Lan-
guage Processing*. Miami, Florida, USA: Association for Computational Lin-
guistics, pp. 8301–8327. DOI: 10.18653/v1/2024.emnlp-main.474. URL:
<https://aclanthology.org/2024.emnlp-main.474/>.
- Chiang, Wei-Lin, Zheng, Lianmin, Sheng, Ying, Angelopoulos, Anastasios Niko-
las, Li, Tianle, Li, Dacheng, Zhu, Banghua, Zhang, Hao, Jordan, Michael,
Gonzalez, Joseph E., and Stoica, Ion (2024). “Chatbot Arena: An Open Plat-
form for Evaluating LLMs by Human Preference”. In: *Forty-first International
Conference on Machine Learning*. URL: <https://openreview.net/forum?id=3MW8GKNyzI>.
- Chung, Hyung Won, Hou, Le, Longpre, Shayne, Zoph, Barret, Tay, Yi, Fedus,
William, Li, Yunxuan, Wang, Xuezhi, Dehghani, Mostafa, Brahma, Siddhartha,
Webson, Albert, Gu, Shixiang Shane, Dai, Zhuyun, Suzgun, Mirac, Chen,

- Xinyun, Chowdhery, Aakanksha, Castro-Ros, Alex, Pellat, Marie, Robinson, Kevin, Valter, Dasha, Narang, Sharan, Mishra, Gaurav, Yu, Adams, Zhao, Vincent, Huang, Yanping, Dai, Andrew, Yu, Hongkun, Petrov, Slav, Chi, Ed H., Dean, Jeff, Devlin, Jacob, Roberts, Adam, Zhou, Denny, Le, Quoc V., and Wei, Jason (2024). “Scaling Instruction-Finetuned Language Models”. In: *Journal of Machine Learning Research* 25.70, pp. 1–53. URL: <http://jmlr.org/papers/v25/23-0870.html>.
- Csaki, Zoltan, Li, Bo, Li, Jonathan Lingjie, Xu, Qiantong, Pawakapan, Pian, Zhang, Leon, Du, Yun, Zhao, Hengyu, Hu, Changran, and Thakker, Urmish (Nov. 2024). “SambaLingo: Teaching Large Language Models New Languages”. In: *Proceedings of the Fourth Workshop on Multilingual Representation Learning (MRL 2024)*. Miami, Florida, USA: Association for Computational Linguistics, pp. 1–21. DOI: 10.18653/v1/2024.mrl-1.1. URL: <https://aclanthology.org/2024.mrl-1.1/>.
- Cui, Yiming, Yang, Ziqing, and Yao, Xin (2024). *Efficient and Effective Text Encoding for Chinese LLaMA and Alpaca*. arXiv: 2304.08177 [cs.CL].
- Etxaniz, Julen, Sainz, Oscar, Perez, Naiara, Aldabe, Itziar, Rigau, German, Agirre, Eneko, Ormazabal, Aitor, Artetxe, Mikel, and Soroa, Aitor (Aug. 2024a). “Latxa: An Open Language Model and Evaluation Suite for Basque”. In: *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Bangkok, Thailand: Association for Computational Linguistics, pp. 14952–14972. DOI: 10.18653/v1/2024.acl-long.799. URL: <https://aclanthology.org/2024.acl-long.799/>.
- Etxaniz, Julen, Sainz, Oscar, Perez, Naiara, Aldabe, Itziar, Rigau, German, Agirre, Eneko, Ormazabal, Aitor, Artetxe, Mikel, and Soroa, Aitor (2024b). *Latxa: An Open Language Model and Evaluation Suite for Basque*. arXiv: 2403.20266 [cs.CL]. URL: <https://arxiv.org/abs/2403.20266>.
- Gao, Leo, Tow, Jonathan, Abbasi, Baber, Biderman, Stella, Black, Sid, DiPofi, Anthony, Foster, Charles, Golding, Laurence, Hsu, Jeffrey, Le Noac’h, Alain, Li, Haonan, McDonell, Kyle, Muennighoff, Niklas, Ociepa, Chris, Phang, Jason, Reynolds, Laria, Schoelkopf, Hailey, Skowron, Aviya, Sutawika, Lintang, Tang, Eric, Thite, Anish, Wang, Ben, Wang, Kevin, and Zou, Andy (July 2024). *A framework for few-shot language model evaluation*. Version v0.4.3. DOI: 10.5281/zenodo.12608602. URL: <https://zenodo.org/records/12608602>.
- Hada, Rishav, Gumma, Varun, Wynter, Adrian de, Diddee, Harshita, Ahmed, Mohamed, Choudhury, Monojit, Bali, Kalika, and Sitaram, Sunayana (2024). *Are Large Language Model-based Evaluators the Solution to Scaling Up Multilingual Evaluation?* arXiv: 2309.07462 [cs.CL]. URL: <https://arxiv.org/abs/2309.07462>.
- Holtermann, Carolin, Röttger, Paul, Dill, Timm, and Lauscher, Anne (Aug. 2024). “Evaluating the Elementary Multilingual Capabilities of Large Language Models with MultiQ”. In: *Findings of the Association for Computational Linguistics*.

- tics ACL 2024*. Bangkok, Thailand and virtual meeting: Association for Computational Linguistics, pp. 4476–4494. DOI: 10.18653/v1/2024.findings-acl.265. URL: <https://aclanthology.org/2024.findings-acl.265>.
- Jiang, Albert, Sablayrolles, Alexandre, Tacnet, Alexis, Kothari, Alok, Roux, Antoine, Mensch, Arthur, Herblin-Stoop, Audrey, Garreau, Augustin, Birky, Austin, Bam4d, Bout, Baptiste, Monicault, Baudouin de, Savary, Blanche, Rambaud, Carole, Feldman, Caroline, Chaplot, Devendra Singh, Casas, Diego de las, Arcelin, Eleonore, Hanna, Emma Bou, Metzger, Etienne, Blanchet, Gaspard, Lengyel, Gianna, Bour, Guillaume, Lample, Guillaume, Rajaona, Harizo, Roussez, Henri, Sattouf, Hichem, Mack, Ian, Delignon, Jean-Malo, Chudnovsky, Jessica, Murke, Justus, Khandelwal, Kartik, Stewart, Lawrence, Martin, Louis, TERNON, Louis, Saulnier, Lucile, Lavaud, L elio Renard, Jennings, Margaret, Pellat, Marie, Torelli, Marie, Lachaux, Marie-Anne, Janiewicz, Marjorie, Seznec, Micka el, Schuhl, Nicolas, Muhs, Niklas, Garrigues, Olivier de, Platen, Patrick von, Jacob, Paul, Buche, Pauline, Reddy, Pavan Kumar, Savas, Perry, Stock, Pierre, Sauvestre, Romain, Vaze, Sagar, Subramanian, Sandeep, Garg, Saurabh, Yang, Sophia, Antoniak, Szymon, Scao, Teven Le, Schueller, Thibault, Lavril, Thibaut, Wang, Thomas, Gervet, Th eophile, Lacroix, Timoth ee, Nemychnikova, Valera, Shang, Wendy, Sayed, William El, and Marshall, William (2024). *Mistral-Nemo-Base-2407*. URL: <https://huggingface.co/mistralai/Mistral-Nemo-Base-2407>.
- Koo, Ryan, Lee, Minhwa, Raheja, Vipul, Park, Jong Inn, Kim, Zae Myung, and Kang, Dongyeop (Aug. 2024). “Benchmarking Cognitive Biases in Large Language Models as Evaluators”. In: *Findings of the Association for Computational Linguistics: ACL 2024*. Bangkok, Thailand: Association for Computational Linguistics, pp. 517–545. DOI: 10.18653/v1/2024.findings-acl.29. URL: <https://aclanthology.org/2024.findings-acl.29/>.
- Kuulmets, Hele-Andra***, Purason, Taido*, Luhtaru, Agnes, and Fishel, Mark (June 2024). “Teaching Llama a New Language Through Cross-Lingual Knowledge Transfer”. In: *Findings of the Association for Computational Linguistics: NAACL 2024*. Mexico City, Mexico: Association for Computational Linguistics, pp. 3309–3325. DOI: 10.18653/v1/2024.findings-naacl.210.
- Nguyen, Thuat, Nguyen, Chien Van, Lai, Viet Dac, Man, Hieu, Ngo, Nghia Trung, DERNONCOURT, Franck, Rossi, Ryan A., and Nguyen, Thien Huu (May 2024). “CulturaX: A Cleaned, Enormous, and Multilingual Dataset for Large Language Models in 167 Languages”. In: *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*. Torino, Italia: ELRA and ICCL, pp. 4226–4237. URL: <https://aclanthology.org/2024.lrec-main.377/>.
- OpenAI et al. (2024). *GPT-4 Technical Report*. arXiv: 2303.08774 [cs.CL]. URL: <https://arxiv.org/abs/2303.08774>.
- Panickssery, Arjun, Bowman, Samuel R., and Feng, Shi (2024). “LLM Evaluators Recognize and Favor Their Own Generations”. In: *Advances in Neural*

- Information Processing Systems*. Vol. 37. Curran Associates, Inc., pp. 68772–68802. DOI: 10.52202/079017-2197. URL: https://proceedings.neurips.cc/paper_files/paper/2024/file/7f1f0218e45f5414c79c0679633e47bc-Paper-Conference.pdf.
- Sprague, Zayne, Yin, Fangcong, Rodriguez, Juan Diego, Jiang, Dongwei, Wadhwa, Manya, Singhal, Prasann, Zhao, Xinyu, Ye, Xi, Mahowald, Kyle, and Durrett, Greg (2024). *To CoT or not to CoT? Chain-of-thought helps mainly on math and symbolic reasoning*. arXiv: 2409.12183 [cs.CL]. URL: <https://arxiv.org/abs/2409.12183>.
- Wendler, Chris, Veselovsky, Veniamin, Monea, Giovanni, and West, Robert (Aug. 2024). “Do Llamas Work in English? On the Latent Language of Multilingual Transformers”. In: *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Bangkok, Thailand: Association for Computational Linguistics, pp. 15366–15394. DOI: 10.18653/v1/2024.acl-long.820. URL: <https://aclanthology.org/2024.acl-long.820/>.
- Zhang, Zhihan, Lee, Dong-Ho, Fang, Yuwei, Yu, Wenhao, Jia, Mengzhao, Jiang, Meng, and Barbieri, Francesco (Aug. 2024). “PLUG: Leveraging Pivot Language in Cross-Lingual Instruction Tuning”. In: *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Bangkok, Thailand: Association for Computational Linguistics, pp. 7025–7046. DOI: 10.18653/v1/2024.acl-long.379. URL: <https://aclanthology.org/2024.acl-long.379>.
- Zheng, Lianmin, Chiang, Wei-Lin, Sheng, Ying, Li, Tianle, Zhuang, Siyuan, Wu, Zhanghao, Zhuang, Yonghao, Li, Zhuohan, Lin, Zi, Xing, Eric, Gonzalez, Joseph E., Stoica, Ion, and Zhang, Hao (2024). “LMSYS-Chat-1M: A Large-Scale Real-World LLM Conversation Dataset”. In: *The Twelfth International Conference on Learning Representations*. URL: <https://openreview.net/forum?id=B0fDKxftw0>.
- Bavaresco, Anna, Bernardi, Raffaella, Bertolazzi, Leonardo, Elliott, Desmond, Fernández, Raquel, Gatt, Albert, Ghaleb, Esam, Giulianelli, Mario, Hanna, Michael, Koller, Alexander, Martins, Andre, Mondorf, Philipp, Neplenbroek, Vera, Pezzelle, Sandro, Plank, Barbara, Schlangen, David, Suglia, Alessandro, Surikuchi, Aditya K, Takmaz, Ece, and Testoni, Alberto (July 2025). “LLMs instead of Human Judges? A Large Scale Empirical Study across 20 NLP Evaluation Tasks”. In: *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Vienna, Austria: Association for Computational Linguistics, pp. 238–255. ISBN: 979-8-89176-252-7. DOI: 10.18653/v1/2025.acl-short.20. URL: <https://aclanthology.org/2025.acl-short.20/>.
- Fu, Xiyan and Liu, Wei (Nov. 2025). “How Reliable is Multilingual LLM-as-a-Judge?” In: *Findings of the Association for Computational Linguistics: EMNLP 2025*. Suzhou, China: Association for Computational Linguistics, pp. 11040–

11053. ISBN: 979-8-89176-335-7. DOI: 10.18653/v1/2025.findings-emnlp.587. URL: <https://aclanthology.org/2025.findings-emnlp.587/>.
- Kuulmets, Hele-Andra**, Purason, Taido, and Fishel, Mark (Mar. 2025). “How Well do LLMs know Finno-Ugric Languages? A Systematic Assessment”. In: *Proceedings of the Joint 25th Nordic Conference on Computational Linguistics and 11th Baltic Conference on Human Language Technologies (NoDaLiDa/Baltic-HLT 2025)*. Vol. 57. Tallinn, Estonia: University of Tartu Library, pp. 340–353. URL: <https://aclanthology.org/2025.nodalida-1.37.pdf>.
- Purason, Taido, Kuulmets, Hele-Andra, and Fishel, Mark (Apr. 2025a). “LLMs for Extremely Low-Resource Finno-Ugric Languages”. In: *Findings of the Association for Computational Linguistics: NAACL 2025*. Albuquerque, New Mexico: Association for Computational Linguistics, pp. 6677–6697. ISBN: 979-8-89176-195-7. DOI: 10.18653/v1/2025.findings-naacl.373. URL: <https://aclanthology.org/2025.findings-naacl.373/>.
- Purason, Taido^{*}, **Kuulmets, Hele-Andra**^{*}, and Fishel, Mark (Apr. 2025b). “LLMs for Extremely Low-Resource Finno-Ugric Languages”. In: *Findings of the Association for Computational Linguistics: NAACL 2025*. Albuquerque, New Mexico: Association for Computational Linguistics, pp. 6677–6697. DOI: 10.18653/v1/2025.findings-naacl.373.
- Rodríguez, Pablo, Suárez, Silvia Paniagua, Gamallo, Pablo, and Docio, Susana Sotelo (July 2025). “Continued Pretraining and Interpretability-Based Evaluation for Low-Resource Languages: A Galician Case Study”. In: *Findings of the Association for Computational Linguistics: ACL 2025*. Vienna, Austria: Association for Computational Linguistics, pp. 4622–4637. ISBN: 979-8-89176-256-5. DOI: 10.18653/v1/2025.findings-acl.240. URL: <https://aclanthology.org/2025.findings-acl.240/>.
- Šuvalov, Hendrik, Lepson, Mihkel, Kukk, Veronika, Malk, Maria, Ilves, Neeme, **Kuulmets, Hele-Andra**, and Kolde, Raivo (Mar. 2025). “Using Synthetic Health Care Data to Leverage Large Language Models for Named Entity Recognition: Development and Validation Study”. In: *Journal of Medical Internet Research* Vol. 27, e66279. DOI: 10.2196/66279.

ACKNOWLEDGEMENTS

I would like to thank my supervisor Prof. Mark Fishel for trusting me with a PhD position when I showed up at his office five years ago and said that I wanted to do a PhD. Whenever we had a meeting, I always left feeling more encouraged and optimistic about my research. Mark has a way of making things feel manageable, even when they do not feel that way at all.

I am deeply grateful to my fellow PhD student, Taido Purason. During the summer of my third year, after having worked alone for some time, I found myself feeling stuck and unsure about what to do next. With a new generation of LLMs emerging, things felt hopeless for low-resource languages. When I mentioned my low point to Mark, he put Taido and me to work together so I would feel less alone. It quickly became clear that Taido and I had very similar ideas about what we wanted to work on. We went on to publish two papers together, both of which I am very proud of.

I would also like to thank my friend and former fellow PhD student Andre Tättar. Andre was the person who originally encouraged me to pursue a PhD. His guidance and mentorship while writing my first paper helped me ease my self-doubt and made me realize that perhaps my self-criticism is, at times, my worst enemy.

I am thankful to my internal reviewer, Dr. Ahmed Sabir, and to my external reviewers, Prof. Barbara Plank and Dr. Jindřich Helcl. Their constructive and detailed feedback helped me make this thesis much better. I should also mention that deciphering Barbara's handwritten comments was unexpectedly a lot of fun.

I have been lucky to be surrounded by so many smart, talented and kind PhD students at the Institute of Computer Science at the University of Tartu. Being able to share the struggles and celebrate the wins with people who understand what they mean has made this journey much less lonely than I had feared. I would especially like to thank the organizers of the PhD community events, which created space for social gatherings and discussions of important topics.

I am thankful for the wonderful people of TartuNLP. Our cake Fridays were something I always looked forward to. Thank you to Agnes, Giacomo, Liisa, Lisa K., Lisa Y., Maksym, Taido, Maali, Rasmus, Rauno, Andre, Annika, Martin, Dmytro, Aleksei I., and Britt-Kathleen.

Many thanks as well to the Estonian Chatbot Arena team—Aleksei Dorkin, Krister Kruusmaa, and Kairit Sirts.

This thesis would not have been possible without all the collaborators from the Finno-Ugric communities who have, over the years, helped us create and gather data for these languages. I would like to thank the translators: Janek Vaab for Võro, Marili Tomingas for Livonian, and Aleksei Ivanov for Komi translations. I am especially grateful to Aleksei for gathering annotations from Komi speakers, and to my mother, Heli, for organizing Võro speakers to participate in the surveys.

I also thank all anonymous participants who took the time to respond, as well as Valts Ernštreits and Sulev Iva for their expertise.

* * *

I wouldn't be here today without Martin. Thank you for carrying me through every low point over the years. Thank you for your unconditional support.

This thesis was mostly written in the spring and summer of 2025, and keeping the writing going was not easy. Thank you, Olavi, for your support and patience during that time. Thank you for all the adventures.

Finally, thank you, Mom.

SISUKOKKUVÕTE

Keeltevaheline siirdeõpe ja selle hindamine väheste ressurssidega oludes

Doktoritöös uuritakse, kuidas parandada ressursivaestes oludes eesti keele mõistmist keelemudelites. Tekstiressursside vähesuse tõttu ei saa eestikeelsete mudelite treenimisel üks ühele rakendada ingliskeelsete mudelite treenimiseks kasutatud lähenemist. Üks võimalus olukorda lahendada on kasutada ära keeltevahelist teadmussiiret, nähtust, mis käib kaasas keelemudelite mitmel keelel korraga treenimisega. Nimelt, kui treenida keelemudeleid mitmel keelel korraga, muutuvad eri keeltes sisendite vektoretsitused mudeli sees üksteisega sarnasemaks. Seega, kui mingis treenimiseks kasutatavas keeles näeb mudel piisavalt tekste selleks, et neist midagi kasulikku õppida, lõikavad sellest mingil määral kasu ka teised treenimisel kasutatud keeled.

Sellise keeltevahelise teadmussiirde tõhusus sõltub paljudest teguritest, näiteks mudeli parameetrite arvust, treeningandmestiku suurusest ja keelte osakaalu jaotusest, mudeli treeningdetailidest ja paljust muust, mille kohta tehakse järeldusi peamiselt empiiriliste eksperimentide põhjal. Seni domineerivad kirjanduses aga uurimused, mis üritavad saavutada võimalikult head teadmussiiret inglise keelest paljudesse teistesse keeltesse korraga, optimeerimata lahendust konkreetselt mitte üheleegi keelele.

Siinne doktoritöö uurib keeltevahelist siirdeõpet eesti keele kontekstis. Täpsemalt uuritakse töös, kui palju on võimalik keelemudelite eesti keele oskust parandada, kui treenimisel kasutatakse maksimaalselt ära olemasolevaid tekstiressursse ja tehnoloogiaid, mis võimaldavad sünteetilisi eestikeelseid andmeid juurde luua (masintõlkesüsteemid ja suured keelemudelid).

Autori originaalpanused on esitatud kolmes (3., 4. ja 5.) peatükis. Esimene neist kirjeldab keeltevahelise siirdeõppe eksperimente väikeste, konkreetse ülesande lahendamiseks optimeeritud keelemudelitega. Teises uuritakse, kuidas kõige paremini kasutada ära suuri keelemudeleid sama eesmärgi saavutamiseks. Viimases peatükis tutvustatakse uut võrdlusalust suurte keelemudelite hindamiseks nelja erineva soome-ugri keele mõistmisel.

Kolmanda peatüki alguses tutvustatakse esmalt võrdlusalust, mille peal mudeleid hindama hakatakse. Iga ülesanne võrdlusaluses koosneb kahest argisest lausest ning mudeli eesmärk on tuvastada, kas nende vahel võib olla põhjuslik seos. Ülesanded on koostatud selliselt, et inimesele piisab selle lahendamiseks argimõistuslikust arutluskäigust. Näiteks suudab inimene sellele tuginedes näha põhjuslikkust sündmuste *keha heitis murule varju* ja *päike tõusis* vahel, sest ta teab lisaks, et tõusev päike tekitab varje. Põhjuslikkuse tuvastamine on üks loomuliku keele mõistmiseks vajalik oskus ning kuigi see on inimesele lihtne, eksivad selles aeg-ajalt ka parimad keelemudelid.

Pärast võrdlusaluse tutvustamist peenhäälestatakse nii eesti- kui ka mitmekeel-

seid eeltreenitud keelemudeleid kirjeldatud ülesannet lahendama eesmärgiga hinnata mitmekeelse eeltreenimise ja peenhäälestamise mõju ülesande lahendamise oskusele. Tulemuste põhjal tehakse järgmised järeldused:

- parim tulemus selle ülesande lahendamisel saadakse mudeliga, mis on nii eeltreenitud kui ka peenhäälestatud mitmekeelsetel andmetel. See mudel on parem ka nendest mudelitest, mis on kas ühes või mõlemas etapis treenitud ainult eestikeelsetel andmetel;
- kvaliteetsete treeningandmete asendamine masintõlgitud vähemkvaliteetsete treeningandmetega ei muuda tulemusi olulisel määral, viidates sellele, et sihtkeelsete andmete kvaliteet ei ole siirdeõppe kontekstis nii oluline kui muud aspektid;
- alusmudeli valik määrab suures osas ära teadmussiirde tõhususe: teadmussiire eesti keelde avaldub ülesande lahendamiseks vajalikul määral ainult piisavalt suure parameetrite arvu ja treeningandmete mahu korral.

Neljas peatükk uurib erinevaid meetodeid suurtele, peamiselt ingliskeelsetel tekstidel treenitud keelemudelitele eesti keele õpetamiseks. Kuigi esmapilgul võib asjaolu, et need mudelid on õppinud põhiliselt ainult inglise keelsetest tekstidest, mitte kuigi julgustavalt mõjuda, tasub arvesse võtta, et suured keelemudelid on kümneid kordi suuremad ja treenitud sadu kordi suurematel andmekogudel kui väiksemad mudelid. See asjaolu muudab need mudelid erinevate ülesannete lahendamisel oluliselt võimekamaks, kui olid väiksemad mudelid, ja annab ühtlasi lootust selleks, et need on paremini suutelised ühes keeles õpitud teadmisi teises keeles vastamisel kasutama.

Võttes aluseks 7 miljardist parameetrist koosneva Llama 2 mudeli, tehakse selle peatükis rida eksperimente katsetamaks nii jätkatud eeltreenimisega kui ka juhistega peenhäälestamisega. Jätkatud eeltreenimiseks kasutatakse andmestikku, mis suures osas sisaldab eesti keelt, kuid vähesel määral ka inglise keelt. Juhistega peenhäälestamiseks kasutatakse eesti- ja ingliskeelsete sünteetiliste ja pärisandmete kombinatsioone. Eksperimentide põhjal tehakse järgmised, eelmistega sarnanevad järeldused:

- mudeli jätkatud eeltreenimine peamiselt eestikeelsetel andmetel parandab oluliselt mudeli eesti keele mõistmise oskust isegi nii vähese kui 1 miljardi tokeni korral;
- mudelit mitmekeelsete juhistega peenhäälestades on võimalik õpetada see eestikeelset sisendit mõistma ja sellele vastama isegi siis, kui treeningandmetena kasutada madalakvaliteedilisi sünteetilisi juhiseid;
- mudeli eesti keele mõistmise oskus paraneb veelgi, kui treeningandmeid täiendada kvaliteetsete ingliskeelsete juhistega, viidates täiendavale teadmussiirdele nendest andmetest.

Et keelemudelite valdkond areneb kiiresti ning pidevalt antakse välja uusi ja paremaid mudeleid, võrdles autor peatüki lõpetuseks selle töö raames loodud parimat spetsiaalselt eesti keelele kohandatud mudelit (4. peatükk) uuemate ja

suuremate, kuid eesti keelele mitte kohandatud mudelitega. Võrdlusest selgus, et uuemad mudelid on eesti keele mõistmises oluliselt paremad kui vanem ja nõrgem spetsiaalselt eesti keelele kohandatud mudel. See tulemus näitab sarnaselt eespool tooduga, et tugeva keeltevahelise teadmussiirde saavutamiseks on eeltreenimisel märksa olulisem roll kui teistel treeningetappidel, sealhulgas jätkatud eeltreenimisel piiratud suurusel eestikeelsele andmehulgal.

Viies peatükk tutvustab uut soome-ugri keeltele mõeldud võrdlusalust vestlusmudelite hindamiseks. Loodud võrdlusalus sisaldab ülesandeid nii eesti kui ka võru, liivi ja komi keeles. Testandmestik on koostatud, võttes arvesse samas peatükis kirjeldatud piiranguid, mida vähete ressurssidega keeled mudelite hindamisele seavad. Seejärel kasutatakse võrdlusalust selleks, et koguda nimetatud keelte kõnelejatelt tagasisidet Tartu Ülikooli keeletehnoloogia uurimisrühma loodud uutele soome-ugri keelemudelitele ja täheldatakse, et võrdlusaluse abil kogutud tagasiside täiendab oluliselt mudelite automaatset hindamist.

Kokkuvõttes uuritakse selles doktoritöös keeltevahelise siirdeõppe meetodeid, et parandada keelemudelite eesti keele mõistmise oskust ressursivaestes oludes. Töö tulemused näitavad siirdeõppemeetodite kasulikkust ja üleolekut mudelitele eesti keele õpetamisel isegi ainult sünteetiliste eestikeelsete andmete kasutamise korral. Töö tulemusena avaldatakse ka uus andmestik vestlusmudelite hindamiseks. Töö autor loodab, et doktoritöö tulemustest on kasu paremate eestikeelsete keelemudelite loomisel, ja julgustab selleks rohkem ära kasutama siirdeõppe võimalusi.

PUBLICATIONS

CURRICULUM VITAE

Personal data

Name: Hele-Andra Kuulmets
Birth: 11.05.1995
Citizenship: Estonia
Contact: kuulmets@gmail.com

Education

2021 – 2026 PhD in Computer Science, University of Tartu
2017 – 2019 MSc in Computer Science, University of Tartu
2014 – 2017 BSc in Computer Science, University of Tartu
2011 – 2014 Tartu Secondary School of Business
2002 – 2011 Kuldre Basic School

Employment

2021 – 2025 Junior Research Fellow in Natural Language Processing,
Institute of Computer Science, University of Tartu
2019 – 2021 Data Scientist, Texta OÜ
2017 – 2018 Software Developer, Nortal AS

Teaching

Spring 2023 Introduction to Data Science
Spring 2023 Object Oriented Programming
Spring 2022 Introduction to Data Science
Fall 2025 Seminar on Natural Language Processing

Scientific work

Main fields of interest:

- Artificial Intelligence
- Natural Language Processing
- Large Language Models
- Cross-lingual Transfer Learning
- Benchmarking and Evaluation

ELULOOKIRJELDUS

Isikuandmed

Nimi: Hele-Andra Kuulmets
Sünniaeg: 11.05.1995
Kodakondsus: Eesti
E-post: kuulmets@gmail.com

Haridus

2021 – 2026 Doktoriõpe informaatika erialal, Tartu Ülikool
2017 – 2019 Magistriõpe informaatika erialal, Tartu Ülikool
2014 – 2017 Bakalaureuseõpe informaatika erialal, Tartu Ülikool
2011 – 2014 Tartu Kommertsgümnaasium
2002 – 2011 Kuldre Kool

Teenistuskäik

2021 – 2025 Nooremteadur loomuliku keele töötamise alal,
Arvutiteaduse instituut, Tartu Ülikool
2019 – 2021 Andmeteadlane, Texta OÜ
2017 – 2018 Tarkvaraarendaja, Nortal AS

Õpetamine

Kevad 2023 Sissejuhatus andmeteadusesse
Kevad 2023 Objektorienteeritud programmeerimine
Kevad 2022 Sissejuhatus andmeteadusesse
Sügis 2025 Keeletehnoloogia seminar

Teadustegevus

Peamised uurimisvaldkonnad:

- Tehisintellekt
- Loomuliku keele töötlus
- Suured keelemudelid
- Keeltevaheline siirdeõpe
- Võrdlusalused ja keelemudelite hindamine

**DISSERTATIONES INFORMATICAЕ
PREVIOUSLY PUBLISHED IN
DISSERTATIONES MATHEMATICAE
UNIVERSITATIS TARTUENSIS**

19. **Helger Lipmaa.** Secure and efficient time-stamping systems. Tartu, 1999, 56 p.
22. **Kaili Müürisep.** Eesti keele arvutigrammatika: süntaks. Tartu, 2000, 107 lk.
23. **Varmo Vene.** Categorical programming with inductive and coinductive types. Tartu, 2000, 116 p.
24. **Olga Sokratova.** Ω -rings, their flat and projective acts with some applications. Tartu, 2000, 120 p.
27. **Tiina Puolakainen.** Eesti keele arvutigrammatika: morfoloogiline ühestamine. Tartu, 2001, 138 lk.
29. **Jan Villemson.** Size-efficient interval time stamps. Tartu, 2002, 82 p.
45. **Kristo Heero.** Path planning and learning strategies for mobile robots in dynamic partially unknown environments. Tartu 2006, 123 p.
49. **Härmel Nestra.** Iteratively defined transfinite trace semantics and program slicing with respect to them. Tartu 2006, 116 p.
53. **Marina Issakova.** Solving of linear equations, linear inequalities and systems of linear equations in interactive learning environment. Tartu 2007, 170 p.
55. **Kaarel Kaljurand.** Attempto controlled English as a Semantic Web language. Tartu 2007, 162 p.
56. **Mart Anton.** Mechanical modeling of IPMC actuators at large deformations. Tartu 2008, 123 p.
59. **Reimo Palm.** Numerical Comparison of Regularization Algorithms for Solving Ill-Posed Problems. Tartu 2010, 105 p.
61. **Jüri Reimand.** Functional analysis of gene lists, networks and regulatory systems. Tartu 2010, 153 p.
62. **Ahti Peder.** Superpositional Graphs and Finding the Description of Structure by Counting Method. Tartu 2010, 87 p.
64. **Vesal Vojdani.** Static Data Race Analysis of Heap-Manipulating C Programs. Tartu 2010, 137 p.
66. **Mark Fišel.** Optimizing Statistical Machine Translation via Input Modification. Tartu 2011, 104 p.
67. **Margus Niitsoo.** Black-box Oracle Separation Techniques with Applications in Time-stamping. Tartu 2011, 174 p.
71. **Siim Karus.** Maintainability of XML Transformations. Tartu 2011, 142 p.
72. **Margus Treumuth.** A Framework for Asynchronous Dialogue Systems: Concepts, Issues and Design Aspects. Tartu 2011, 95 p.
73. **Dmitri Lepp.** Solving simplification problems in the domain of exponents, monomials and polynomials in interactive learning environment T-algebra. Tartu 2011, 202 p.

74. **Meelis Kull.** Statistical enrichment analysis in algorithms for studying gene regulation. Tartu 2011, 151 p.
77. **Bingsheng Zhang.** Efficient cryptographic protocols for secure and private remote databases. Tartu 2011, 206 p.
78. **Reina Uba.** Merging business process models. Tartu 2011, 166 p.
79. **Uuno Puus.** Structural performance as a success factor in software development projects – Estonian experience. Tartu 2012, 106 p.
81. **Georg Singer.** Web search engines and complex information needs. Tartu 2012, 218 p.
83. **Dan Bogdanov.** Sharemind: programmable secure computations with practical applications. Tartu 2013, 191 p.
84. **Jevgeni Kabanov.** Towards a more productive Java EE ecosystem. Tartu 2013, 151 p.
87. **Margus Freudenthal.** Simpl: A toolkit for Domain-Specific Language development in enterprise information systems. Tartu, 2013, 151 p.
90. **Raivo Kolde.** Methods for re-using public gene expression data. Tartu, 2014, 121 p.
91. **Vladimir Sor.** Statistical Approach for Memory Leak Detection in Java Applications. Tartu, 2014, 155 p.
92. **Naved Ahmed.** Deriving Security Requirements from Business Process Models. Tartu, 2014, 171 p.
94. **Liina Kamm.** Privacy-preserving statistical analysis using secure multi-party computation. Tartu, 2015, 201 p.
100. **Abel Armas Cervantes.** Diagnosing Behavioral Differences between Business Process Models. Tartu, 2015, 193 p.
101. **Fredrik Milani.** On Sub-Processes, Process Variation and their Interplay: An Integrated Divide-and-Conquer Method for Modeling Business Processes with Variation. Tartu, 2015, 164 p.
102. **Huber Raul Flores Macario.** Service-Oriented and Evidence-aware Mobile Cloud Computing. Tartu, 2015, 163 p.
103. **Tauno Metsalu.** Statistical analysis of multivariate data in bioinformatics. Tartu, 2016, 197 p.
104. **Riivo Talviste.** Applying Secure Multi-party Computation in Practice. Tartu, 2016, 144 p.
108. **Siim Orasmaa.** Explorations of the Problem of Broad-coverage and General Domain Event Analysis: The Estonian Experience. Tartu, 2016, 186 p.
109. **Prastudy Mungkas Fauzi.** Efficient Non-interactive Zero-knowledge Protocols in the CRS Model. Tartu, 2017, 193 p.
110. **Pelle Jakovits.** Adapting Scientific Computing Algorithms to Distributed Computing Frameworks. Tartu, 2017, 168 p.
111. **Anna Leontjeva.** Using Generative Models to Combine Static and Sequential Features for Classification. Tartu, 2017, 167 p.
112. **Mozhgan Pourmoradnasseri.** Some Problems Related to Extensions of Polytopes. Tartu, 2017, 168 p.

113. **Jaak Randmets.** Programming Languages for Secure Multi-party Computation Application Development. Tartu, 2017, 172 p.
114. **Alisa Pankova.** Efficient Multiparty Computation Secure against Covert and Active Adversaries. Tartu, 2017, 316 p.
116. **Toomas Saarsen.** On the Structure and Use of Process Models and Their Interplay. Tartu, 2017, 123 p.
121. **Kristjan Korjus.** Analyzing EEG Data and Improving Data Partitioning for Machine Learning Algorithms. Tartu, 2017, 106 p.
122. **Eno Tõnisson.** Differences between Expected Answers and the Answers Offered by Computer Algebra Systems to School Mathematics Equations. Tartu, 2017, 195 p.

DISSERTATIONES INFORMATICAЕ UNIVERSITATIS TARTUENSIS

1. **Abdullah Makkeh.** Applications of Optimization in Some Complex Systems. Tartu 2018, 179 p.
2. **Riivo Kikas.** Analysis of Issue and Dependency Management in Open-Source Software Projects. Tartu 2018, 115 p.
3. **Ehsan Ebrahimi.** Post-Quantum Security in the Presence of Superposition Queries. Tartu 2018, 200 p.
4. **Ilya Verenich.** Explainable Predictive Monitoring of Temporal Measures of Business Processes. Tartu 2019, 151 p.
5. **Yauhen Yakimenka.** Failure Structures of Message-Passing Algorithms in Erasure Decoding and Compressed Sensing. Tartu 2019, 134 p.
6. **Irene Teinmaa.** Predictive and Prescriptive Monitoring of Business Process Outcomes. Tartu 2019, 196 p.
7. **Mohan Liyanage.** A Framework for Mobile Web of Things. Tartu 2019, 131 p.
8. **Toomas Krips.** Improving performance of secure real-number operations. Tartu 2019, 146 p.
9. **Vijayachitra Modhukur.** Profiling of DNA methylation patterns as biomarkers of human disease. Tartu 2019, 134 p.
10. **Elena Sügis.** Integration Methods for Heterogeneous Biological Data. Tartu 2019, 250 p.
11. **Tõnis Tasa.** Bioinformatics Approaches in Personalised Pharmacotherapy. Tartu 2019, 150 p.
12. **Sulev Reisberg.** Developing Computational Solutions for Personalized Medicine. Tartu 2019, 126 p.
13. **Huishi Yin.** Using a Kano-like Model to Facilitate Open Innovation in Requirements Engineering. Tartu 2019, 129 p.
14. **Faiz Ali Shah.** Extracting Information from App Reviews to Facilitate Software Development Activities. Tartu 2020, 149 p.
15. **Adriano Augusto.** Accurate and Efficient Discovery of Process Models from Event Logs. Tartu 2020, 194 p.
16. **Karim Baghery.** Reducing Trust and Improving Security in zk-SNARKs and Commitments. Tartu 2020, 245 p.
17. **Behzad Abdolmaleki.** On Succinct Non-Interactive Zero-Knowledge Protocols Under Weaker Trust Assumptions. Tartu 2020, 209 p.
18. **Janno Siim.** Non-Interactive Shuffle Arguments. Tartu 2020, 154 p.
19. **Ilya Kuzovkin.** Understanding Information Processing in Human Brain by Interpreting Machine Learning Models. Tartu 2020, 149 p.
20. **Orlenys López Pintado.** Collaborative Business Process Execution on the Blockchain: The Caterpillar System. Tartu 2020, 170 p.
21. **Ardi Tampuu.** Neural Networks for Analyzing Biological Data. Tartu 2020, 152 p.

22. **Madis Vasser.** Testing a Computational Theory of Brain Functioning with Virtual Reality. Tartu 2020, 106 p.
23. **Ljubov Jaanuska.** Haar Wavelet Method for Vibration Analysis of Beams and Parameter Quantification. Tartu 2021, 192 p.
24. **Arnis Parsovs.** Estonian Electronic Identity Card and its Security Challenges. Tartu 2021, 214 p.
25. **Kaido Lepik.** Inferring causality between transcriptome and complex traits. Tartu 2021, 224 p.
26. **Tauno Palts.** A Model for Assessing Computational Thinking Skills. Tartu 2021, 134 p.
27. **Liis Kolberg.** Developing and applying bioinformatics tools for gene expression data interpretation. Tartu 2021, 195 p.
28. **Dmytro Fishman.** Developing a data analysis pipeline for automated protein profiling in immunology. Tartu 2021, 155 p.
29. **Ivo Kubjas.** Algebraic Approaches to Problems Arising in Decentralized Systems. Tartu 2021, 120 p.
30. **Hina Anwar.** Towards Greener Software Engineering Using Software Analytics. Tartu 2021, 186 p.
31. **Veronika Plotnikova.** FIN-DM: A Data Mining Process for the Financial Services. Tartu 2021, 197 p.
32. **Manuel Camargo.** Automated Discovery of Business Process Simulation Models From Event Logs: A Hybrid Process Mining and Deep Learning Approach. Tartu 2021, 130 p.
33. **Volodymyr Leno.** Robotic Process Mining: Accelerating the Adoption of Robotic Process Automation. Tartu 2021, 119 p.
34. **Kristjan Krips.** Privacy and Coercion-Resistance in Voting. Tartu 2022, 173 p.
35. **Elizaveta Yankovskaya.** Quality Estimation through Attention. Tartu 2022, 115 p.
36. **Mubashar Iqbal.** Reference Framework for Managing Security Risks Using Blockchain. Tartu 2022, 203 p.
37. **Jakob Mass.** Process Management for Internet of Mobile Things. Tartu 2022, 151 p.
38. **Gamal Elkoumy.** Privacy-Enhancing Technologies for Business Process Mining. Tartu 2022, 135 p.
39. **Lidia Feklistova.** Learners of an Introductory Programming MOOC: Background Variables, Engagement Patterns and Performance. Tartu 2022, 151 p.
40. **Mohamed Ragab.** Bench-Ranking: A Prescriptive Analysis Approach for Large Knowledge Graphs Query Workloads. Tartu 2022, 158 p.
41. **Mohammad Anagreh.** Privacy-Preserving Parallel Computations for Graph Problems. Tartu 2023, 181 p.
42. **Rahul Goel.** Mining Social Well-being Using Mobile Data. Tartu 2023, 104 p.

43. **Anti Ingel.** Algorithms using information theory: classification in brain-computer interfaces and characterising reinforcement-learning agents. Tartu 2023, 142 p.
44. **Shakshi Sharma.** Fighting Misinformation in the Digital Age: A Comprehensive Strategy for Characterizing, Identifying, and Mitigating Misinformation on Online Social Media Platforms. Tartu 2023, 158 p.
45. **Kristiina Rahkema.** Quality Analysis of iOS Applications with Focus on Maintainability and Security Aspects. Tartu 2023, 182 p.
46. **Ivan Slobozhan.** Studying Online Social Media Engagement in CIS Countries during Protests, Mass Demonstrations and War. Tartu 2023, 81 p.
47. **Nurlan Kerimov.** Building a catalogue of molecular quantitative trait loci to interpret complex trait associations. Tartu 2023, 248 p.
48. **Pavlo Tertychnyi.** Machine Learning Methods for Anti-Money Laundering Monitoring. Tartu 2023, 117 p.
49. **Abasi-amefon Obot Affia.** A Framework and Teaching Approach for IoT Security Risk Management. Tartu 2023, 180 p.
50. **Raimond-Hendrik Tunnel.** Video Game Design and Development Bachelor's Curriculum for Estonia. Tartu 2024, 137 p.
51. **Ahto Salumets.** Bioinformatics analysis of various aspects in immunology. Tartu 2024, 198 p.
52. **Mohammed Abdulhameed Shaif Ali.** Deep Learning Methods for Cell Microscopy Image Analysis. Tartu 2024, 143 p.
53. **Pille Pullonen-Raudvere.** Foundations of Efficient and Secure Algorithm Development for Secure Multiparty Computation. Tartu 2024, 265 p.
54. **Marili Rõõm.** Multiple approaches to learners' success and factors affecting it in computer programming MOOCs. Tartu 2024, 170 p.
55. **Shivananda Rangappa Poojara.** Design and Orchestration of Scalable, Event-Driven Serverless Data Pipelines for Internet of Things (IoT) Applications. Tartu 2024, 172 p.
56. **Hassan Abdulgaleel Hassan Salim Eldeeb.** Empowering Machine Learning Pipelines with Automated Feature Engineering. Tartu 2024, 121 p.
57. **Muhammad Uzair.** Soft decision making for agri-food 4.0. Tartu 2024, 158 p.
58. **Kirill Milintsevich.** Estimation of Depression Level from Text: Symptom-Based Approach, External Knowledge, Dataset Validity. Tartu 2024, 130 p.
59. **Maksym Del.** Multilingual and Multi-Domain Representational Patterns Across Trpansformer-Based Models. Tartu 2024, 131 p.
60. **Kristo Raun.** Adaptive Out-of-order Handling in Streaming Conformance Checking. Tartu 2024, 118 p.
61. **Toivo Vajakas.** Towards integration of mobile network data into analyzing human mobility. Tartu 2024, 103 p.
62. **Katsiaryna Lashkevich.** Data-Driven Analysis and Optimization of Waiting Times in Business Processes. Tartu 2024, 169 p.
63. **Alejandra Duque-Torres.** Classifying, Constraining and Ranking Metamorphic Relations. Tartu 2025, 159 p.

64. **Mariia Bakhtina.** A Method for Information Security and Privacy Management in Smart Solutions. Tartu 2025, 199 p.
65. **Andre Tättar.** Multilingual Machine Translation for Under-Resourced Languages. Tartu 2025, 170 p.
66. **Mahmoud Shoush.** Prescriptive Process Monitoring Under Uncertainty and Resource Constraints. Tartu 2025, 178 p.
67. **Alireza Akhavi Zadegan.** A Multimodal approach for refining Mapping and Localization by Integrating Generative AI and Pedestrian-Centric Data. Tartu 2025, 147 p.
68. **Eerik Muuli.** Automating the assessment and feedback processes in IT teaching – improving creation and maintenance from the teaching staff perspective. Tartu 2025, 196 p.
69. **Kateryna Kubrak.** Towards User-Centered Prescriptive Process Monitoring Systems. Tartu 2025, 151 p.
70. **Zhigang Yin.** Computing and Sensing in a Smart Ring. Tartu 2025, 251 p.
71. **Abdul-Rasheed Olatunji Ottun.** Practical Trustworthy Artificial Intelligence with Human Oversight. Tartu 2025, 239 p.
72. **Sander Mikelsaar.** Analysis and Optimization of Iteratively Decodable Codes. Tartu 2025, 146 p.
73. **Marharyta Domnich.** Advancing Human-Centric Counterfactual Explanations in Explainable AI. Tartu 2025, 210 p.
74. **Viacheslav Komisarenko.** Aligning Training Loss to Evaluation Metrics in Deep Learning. Tartu 2026, 165 p.
75. **Heidi Taveter.** Using Programming-Process Data of Introductory Programming Courses: Finding Solver Types, Giving Feedback, and Detecting Plagiarism. Tartu 2026, 184 p.
76. **Daniel Majoral Lopez.** Deep neural networks for microscopy images. Tartu 2026, 81 p.
77. **Mahir Gulzar.** Addressing Real-world Scenarios via Motion Prediction in Autonomous Driving. Tartu 2026, 141 p.