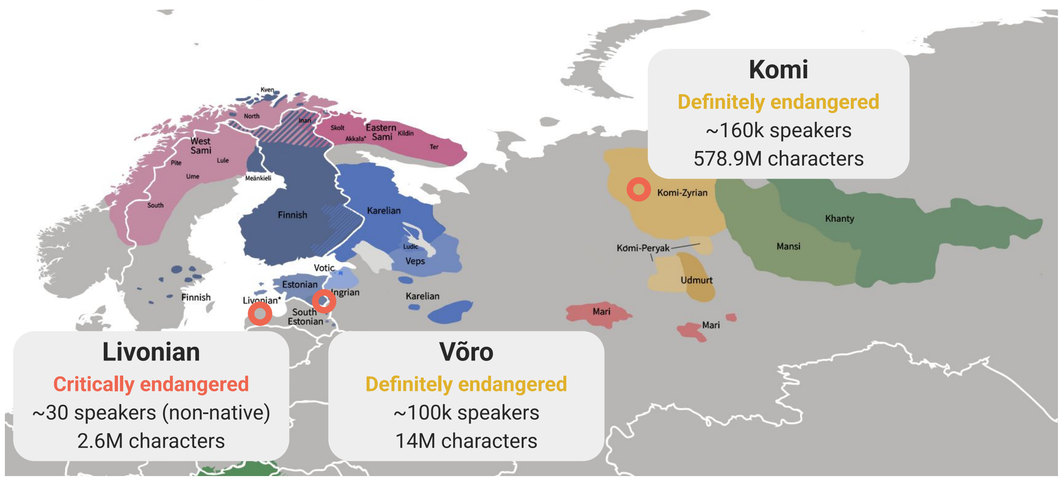
LLMs for Extremely Low-Resource Finno-Ugric Languages

Taido Purason*, Hele-Andra Kuulmets*, Mark Fishel

This work develops multilingual LLMs for Võro, Livonian, and Komi—three extremely low-resource (XLR) Finno-Ugric languages. Key contributions include the creation of Llama-SMUGRI models, new evaluation benchmarks, and extensive human evaluation. Our models achieve competitive performance with GPT-3.5-turbo in helpfulness and surpass it in naturalness.

Languages



GalaxMaps, CC BY-SA 4.0, via Wikimedia Commons

Benchmarks

SIB-200 → **SIB-SMUGRI** (topic classification)

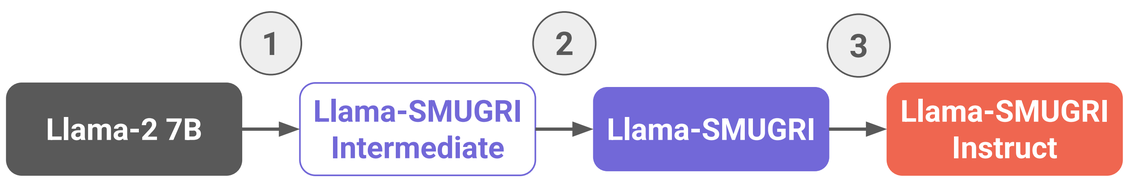
Belebele → **Belebele-SMUGRI** (MCQA)

We partially **extend existing benchmarks** to cover Livonian, Võro, and Komi.

SMUGRI-MT-Bench [NEW]

We **create a novel multi-turn conversational benchmark** for evaluating LLMs in Võro, Livonian, and Komi using 80 real-world prompts covering *math*, *reasoning*, *writing*, and *general* topics. It enables human assessment of model performance in these low-resource languages.

Training methodology



1. Continued pre-training on **high-resource supporting languages** (*ET*, *FI*, *EN*, *RU*, *LV*) for 10B tokens.
2. Continued pre-training on **XLR Finno-Ugric languages**, supporting languages, and parallel data for 3B characters, repeating the XLR data at most 4 times.
3. **Instruction-tuning** on machine translated and supporting language instructions.

Results

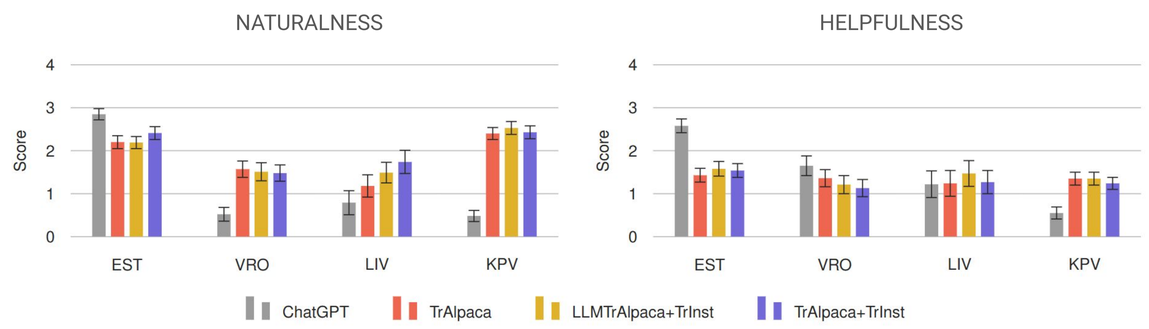
Model	SIB-SMUGRI 5-shot, acc			BELEBELE-SMUGRI 3-shot, acc			FLORES-SMUGRI 5-shot, BLEU					
	VRO	LIV	KPV	VRO	LIV	KPV	ET-VRO	ET-LIV	RU-KPV	VRO-ET	LIV-ET	KPV-RU
Llammas-base	78.4 (3.7)	69.6 (4.1)	64.0 (4.3)	30.7 (4.1)	28.4 (4.0)	32.3 (4.2)	11.5 (0.9)	4.3 (0.5)	1.7 (0.4)	28.7 (1.5)	8.0 (0.8)	2.2 (0.3)
Llama-2-7B	57.6 (4.4)	60.0 (4.4)	58.4 (4.4)	29.1 (4.1)	**29.9** (4.1)	**36.2** (4.3)	11.1 (1.0)	4.6 (0.6)	1.5 (0.3)	11.3 (0.9)	4.4 (0.6)	2.4 (0.3)
Llama-SMUGRI (ours)												
Stage 1	80.8 (3.5)	**75.2** (3.9)	65.6 (4.3)	32.3 (4.2)	26.8 (3.9)	26.0 (3.9)	11.5 (1.0)	4.2 (0.5)	2.6 (0.6)	29.6 (1.4)	7.2 (0.7)	4.1 (0.7)
Stage 2	78.4 (3.7)	65.6 (4.3)	74.4 (3.9)	31.5 (4.1)	26.0 (3.9)	28.4 (4.0)	26.5 (1.1)	3.4 (0.4)	15.7 (1.0)	45.3 (1.5)	10.6 (0.9)	18.6 (0.9)
Stage 2 + parallel	**84.0** (3.3)	66.4 (4.2)	**76.8** (3.8)	**35.4** (4.3)	27.6 (4.0)	29.1 (4.1)	**29.1** (1.2)	**4.3** (0.5)	**16.0** (1.0)	**48.7** (1.4)	**17.6** (1.0)	**22.1** (1.3)

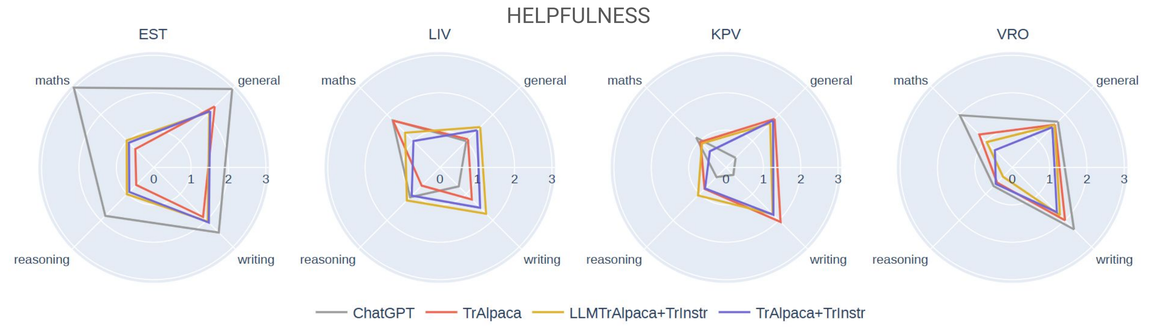
Continued pre-training stage 1 positively affects some discriminative benchmark scores while in **stage 2** the model learns to generate text in those languages.

Model	BELEBELE-SMUGRI 0-shot, acc			SIB-SMUGRI 5-shot, acc		
	VRO	LIV	KPV	VRO	LIV	KPV
GPT-3.5-turbo	45.7 (4.4)	37.8 (4.3)	34.6 (4.2)	81.6 (3.5)	73.6 (4.0)	68.8 (4.2)
GPT-4-turbo	**70.1** (4.1)	40.2 (4.3)	**44.1** (4.4)	**92.0** (2.5)	72.0 (4.0)	67.2 (4.2)
Llammas (Kuulmets et al., 2024)	36.2 (4.3)	32.3 (4.2)	27.6 (4.0)	80.8 (3.5)	78.4 (3.7)	63.2 (4.3)
Llama-SMUGRI-Instruct						
SupInst	42.5 (4.4)	30.7 (4.1)	**44.1** (4.4)	86.4 (3.1)	79.2 (3.6)	**88.8** (2.8)
SupInst+LLMTrAlpaca	39.4 (4.3)	35.4 (4.3)	42.5 (4.4)	85.6 (3.1)	**81.6** (3.5)	84.8 (3.2)
SupInst+TrAlpaca	35.4 (4.2)	32.3 (4.2)	40.2 (4.3)	85.6 (3.1)	79.2 (3.6)	85.6 (3.1)
SupInst+LLMTrAlpaca+TrInst	44.9 (4.4)	**40.9** (4.4)	**44.1** (4.4)	86.4 (3.1)	76.0 (3.8)	78.4 (3.7)
SupInst+TrAlpaca+TrInst	45.7 (4.4)	32.3 (4.2)	**44.1** (4.4)	86.4 (3.1)	78.4 (3.7)	78.4 (3.7)

Automatic evaluation on the instruction-tuned models reveals that *GPT-4-turbo* outperforms our models on Võro, while our models are competitive on the other languages.

Human evaluation





Our models achieve **higher naturalness than GPT-3.5-turbo** with competitive helpfulness — often outperforming GPT-3.5-turbo in *general* and *writing* tasks.

www.tartunlp.ai taido.purason@ut.ee hele-andra.kuulmets@ut.ee

UNIVERSITY OF TARTU
Institute of Computer Science