How Well do LLMs know Finno-Ugric Languages? A Systematic Assessment

Hele-Andra Kuulmets, Taido Purason, Mark Fishel

Institute of Computer Science, University of Tartu

March 3, 2025

Introduction

- Multilingual capabilities of LLMs are getting better and better

Introduction

- Multilingual capabilities of LLMs are getting better and better
- Proprietary models are better than open models

Introduction

- Multilingual capabilities of LLMs are getting better and better
- Proprietary models are better than open models
    - GPT-4 achieves 80.9% for Latvian and 76.5% for Icelandic on the 3-shot MMLU benchmark [1]

---

[1] OpenAI. 2024. GPT-4 technical report.

Introduction

- Multilingual capabilities of LLMs are getting better and better
- Proprietary models are better than open models
    - GPT-4 achieves 80.9% for Latvian and 76.5% for Icelandic on the 3-shot MMLU benchmark [1]
- Open models have started to catch-up but officially supported languages remain limited

[1] OpenAI. 2024. GPT-4 technical report.

Introduction

- Multilingual capabilities of LLMs are getting better and better
- Proprietary models are better than open models
    - GPT-4 achieves 80.9% for Latvian and 76.5% for Icelandic on the 3-shot MMLU benchmark [1]
- Open models have started to catch-up but officially supported languages remain limited
    - Yet, Llama English-centric Llama 2 7B correctly answered 14% and 40% of basic open-ended questions in Estonian and Finnish even though only 0.03% of the Llama 2 training data was in Finnish and less than 0.005% in Estonian. [2] [3]

---

[1] OpenAI. 2024. GPT-4 technical report.

[2] Carolin Holtermann, Paul Röttger, Timm Dill, and Anne Lauscher. 2024. Evaluating the elementary multilingual capabilities of large language models with MultiQ. In *Findings of the Association for Computational Linguistics ACL 2024.*

[3] Hugo Touvron, et al. 2023. Llama 2: Open foundation and fine-tuned chat models.

Introduction

- Multilingual capabilities of LLMs are getting better and better
- Proprietary models are better than open models
    - GPT-4 achieves 80.9% for Latvian and 76.5% for Icelandic on the 3-shot MMLU benchmark [1]
- Open models have started to catch-up but officially supported languages remain limited
    - Yet, Llama English-centric Llama 2 7B correctly answered 14% and 40% of basic open-ended questions in Estonian and Finnish even though only 0.03% of the Llama 2 training data was in Finnish and less than 0.005% in Estonian. [2] [3]
- Proficiency and comparative performance across models remain largely unexplored

---

[1] OpenAI. 2024. GPT-4 technical report.

[2] Carolin Holtermann, Paul Röttger, Timm Dill, and Anne Lauscher. 2024. Evaluating the elementary multilingual capabilities of large language models with MultiQ. In *Findings of the Association for Computational Linguistics ACL 2024.*

[3] Hugo Touvron, et al. 2023. Llama 2: Open foundation and fine-tuned chat models.

Introduction

- Multilingual capabilities of LLMs are getting better and better
- Proprietary models are better than open models
  - GPT-4 achieves 80.9% for Latvian and 76.5% for Icelandic on the 3-shot MMLU benchmark [1]
- Open models have started to catch-up but officially supported languages remain limited
  - Yet, Llama English-centric Llama 2 7B correctly answered 14% and 40% of basic open-ended questions in Estonian and Finnish even though only 0.03% of the Llama 2 training data was in Finnish and less than 0.005% in Estonian. [2] [3]
- Proficiency and comparative performance across models remain largely unexplored
- We benchmark 7 open LLMs on 5 Finno-Ugric languages across 5 tasks and 9 datasets

---

[1] OpenAI. 2024. GPT-4 technical report.

[2] Carolin Holtermann, Paul Röttger, Timm Dill, and Anne Lauscher. 2024. Evaluating the elementary multilingual capabilities of large language models with MultiQ. In *Findings of the Association for Computational Linguistics ACL 2024.*

[3] Hugo Touvron, et al. 2023. Llama 2: Open foundation and fine-tuned chat models.

Languages and models

- 5 Finno-Ugric languages: Finnish, Estonian, Livonian, Võro and Komi

Languages and models

- 5 Finno-Ugric languages: Finnish, Estonian, Livonian, Võro and Komi
    - Finnish and Estonian most well-resourced

Languages and models

- 5 Finno-Ugric languages: Finnish, Estonian, Livonian, Võro and Komi
    - Finnish and Estonian most well-resourced
    - Others are extremely low-resource languages

Languages and models

- 5 Finno-Ugric languages: Finnish, Estonian, Livonian, Võro and Komi
    - Finnish and Estonian most well-resourced
    - Others are extremely low-resource languages

- 7 models

Languages and models

- 5 Finno-Ugric languages: Finnish, Estonian, Livonian, Võro and Komi
    - Finnish and Estonian most well-resourced
    - Others are extremely low-resource languages

- 7 models
    - 5 models from Llama 2 and 3.1 families (7B-70B) (Meta AI)

Languages and models

- 5 Finno-Ugric languages: Finnish, Estonian, Livonian, Võro and Komi
    - Finnish and Estonian most well-resourced
    - Others are extremely low-resource languages

- 7 models
    - 5 models from Llama 2 and 3.1 families (7B-70B) (Meta AI)
    - Mistral NeMo (12B) (Mistral AI)

Languages and models

- 5 Finno-Ugric languages: Finnish, Estonian, Livonian, Võro and Komi
    - Finnish and Estonian most well-resourced
    - Others are extremely low-resource languages

- 7 models
    - 5 models from Llama 2 and 3.1 families (7B-70B) (Meta AI)
    - Mistral NeMo (12B) (Mistral AI)
    - Llammas (TartuNLP)

Research questions

Research questions

- How well can open LLMs solve tasks in Finno-Ugric languages?

Research questions

- How well can open LLMs solve tasks in Finno-Ugric languages?
- What is the expected improvement from few-shot prompting over zero-shot prompting?

Research questions

- How well can open LLMs solve tasks in Finno-Ugric languages?
- What is the expected improvement from few-shot prompting over zero-shot prompting?
- Can chain-of-thought prompting improve the performance of open LLMs on Finno-Ugric languages?

Prompting strategies

Zero-shot prompting

model input

Given a passage and a question, select the correct answer from the given choices.

P: Giancarlo Fisichella kaot’ uma auto üle kontrolli ja lõpõt’ võikisõitmisõ ärq pia päält alostust. Timä miiskunnaliigõq Fernando Alonso juhtõ võikisõitmist inämbüse aost, a tuu lõppi õkva päält boksipiätüst, arvadaq halvastõ kinnütedü hää edetsõõri peräst. Michael Schumacher lõpõt’ võistlusõ pia päält Alonsot palľodõst müüdasõitmiisist tulnuidõ vedrostusõ häti peräst.
Q: Mille Schumacher võikisõidu lõpõt'?
A.Tä kaot' uma auto üle kontrolli; **B**. Ka pand' Alonsoga kokko; **C**. Fisichella lahk' tsõõri ärq; **D**. Auto vedrostusõ man oll' hädä
Answer:

model output

D

Few-shot prompting

model input

Given a passage and a question, select the correct answer from the given choices.

P: Om kimmäs tett, et iispäävä Hummogu- Prantsusmaalt Lyoni lähküst suust lövvetül lõpnul mõtsikul pardsil oll’ külen inemiisile surmava tsirgugripi tüvi H5N1. Prantsusmaa om Euruupa Liido säitsmes riik, kiä viirusõga hädän om; Prantsusmaa tulõ päält Austriat, S’aksamaad, Sloveeniät, Bulgaariat, Kreekat ja Itaaliat. H5N1 arvatavaq ettetulõmisõq Horvaatian ja Taanin olõ-õi kinnütüst löüdnüq.
Q: Mitmõst Õuruupa Liido riigist H5N1 viirust om lövvet?
A. Viiest; **B.** Kuvvõst; **C.** Säitsmest; **D.** Katsast
Answer: C

P: Giancarlo Fisichella kaot’ uma auto üle kontrolli ja lõpõt’ võikisõitmisõ ärq pia päält alostust. Timä miiskunnaliigõq Fernando Alonso juhtõ võikisõitmist inämbüse aost, a tuu lõppi õkva päält boksipiätüst, arvadaq halvastõ kinnütedü hää edetsõõri peräst. Michael Schumacher lõpõt’ võistlusõ pia päält Alonsot palľodõst müüdasõitmiisist tulnuidõ vedrostusõ häti peräst.
Q: Mille Schumacher võikisõidu lõpõt'?
A.Tä kaot' uma auto üle kontrolli; **B**. Ka pand' Alonsoga kokko; **C**. Fisichella lahk' tsõõri ärq; **D**. Auto vedrostusõ man oll' hädä
Answer:

model output

D

Chain-of-thought (CoT) prompting

model input (CoT prompting)

Given a passage and a question, select the correct answer from the given choices.

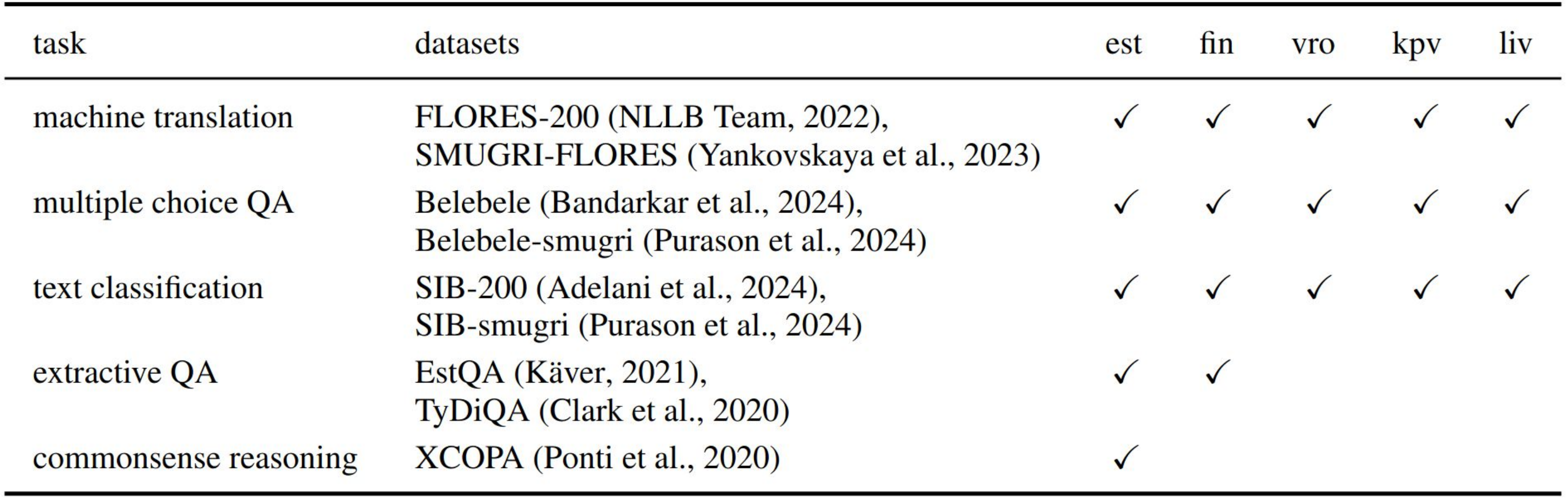
P: Om kimmäs tett, et iispäävä Hummogu-Prantsusmaalt Lyoni lähküst suust lövvetül lõpnul mõtsikul pardsil oll' külen inemiisile surmava tsirgugripi tüvi H5N1. Prantsusmaa om Euruupa Liido säitsmes riik, kiä viirusõga hädän om; Prantsusmaa tulõ päält Austriat, S'aksamaad, Sloveeniät, Bulgaariat, Kreekat ja Itaaliat. H5N1 arvatavaq ettetulõmisõq Horvaatian ja Taanin olõ-õi kinnütüst löüdnüq.
Q: Mitmõst Õuruupa Liido riigist H5N1 viirust om lövvet?
A. Viiest; **B**. Kuvvõst; **C**. Säitsmest; **D**. Katsast
Answer: Let's think step by step.
The passage "<*passage in võro*>" can be translated from Võro to English as "The strain of bird flu lethal to humans, H5N1, has been confirmed to have infected a dead wild duck, found on Monday, in marshland near Lyon in the east of France. France is the seventh country in the European Union to suffer this virus; following Austria, Germany, Slovenia, Bulgaria, Greece and Italy. Suspected cases of H5N1 in Croatia and Denmark remain unconfirmed."
The question "<*question in võro*>" can be translated from Võro to English as "In how many countries in the European Union has the H5N1 virus been identified?" and the options translate as follows: "A. Five", "B. Six", "C. Seven", "D. Eight". According to text the correct answer in English is "C. Seven". Therefore, the answer in Võro is "**C. Säitsmest**".

P: Giancarlo Fisichella kaot' uma auto üle kontrolli ja lõpõt' võikisõitmisõ ärq pia päält alostust. Timä miiskunnaliigõq Fernando Alonso juhtõ võikisõitmist inämbüse aost, a tuu lõppi õkva päält boksipiätüst, arvadaq halvastõ kinnütedü hää edetsõõri peräst. Michael Schumacher lõpõt' võistlusõ pia päält Alonsot pallodõst müüdasõitmiisist tulnuidõ vedrostusõ häti peräst.
Q: Mille Schumacher võikisõidu lõpõt'?
A. Tä kaot' uma auto üle kontrolli; **B**. Ka pand' Alonsoga kokko; **C.** Fisichella lahk' tsõõri ärq; **D**. Auto vedrostusõ man oll' hädä
Answer:

model output

Let's think step by step.
The sentence "<*passage in võro*>" can be translated from Võro to English as "Giancarlo Fisichella lost control of his car and ended the race very soon after the start. His teammate Fernando Alonso was in the lead for most of the race, but ended it right after his pit-stop, probably because a badly tucked right front wheel. Michael Schumacher ended his race not long after Alonso, because of the suspension damage in the numerous battles during the race.".
The question "<*question in võro*>" can be translated from Võro to English as "What caused Schumacher to end his race early?" and the options translate as follows: "A. He lost control of his car", "B. A collision with Alonso", "C. Wheel damage caused by Fisichella", "D. Damage to the car's suspension". According to text the correct answer in English is "D. Damage to the car's suspension". Therefore, the answer in Võro is "**D. Auto vedrostusõ man oll' hädä**".

Datasets



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

Machine translation

Sentence in Finnish: Tämän teorian mukaan suurin osa galaksin pimeästä aineesta ympäröi sitä eräänlaisena sädekehänä ja koostuu suuresta määrästä pieniä hiukkasia.

Sentence in English: This theory says that most dark matter around a galaxy is located around a galaxy in a kind of halo, and is made of lots of small particles.

Multiple choice QA

Passage: Giancarlo Fisichella lost control of his car and ended the race very soon after the start. His teammate Fernando Alonso was in the lead for most of the race, but ended it right after his pit-stop, probably because a badly tucked right front wheel. Michael Schumacher ended his race not long after Alonso, because of the suspension damage in the numerous battles during the race.

Question: What caused Schumacher to end his race early?

A. He lost control of his car; **B**. A collision with Alonso; **C**. Wheel damage caused by Fisichella; **D**. Damage to the car's suspension

Answer: D

Text classification

Text: Visitors with limited time would be best to spend their time elsewhere.

Categories: science/technology, travel, politics, sports, health, entertainment, geography

Answer: travel

Extractive QA (reading comprehension)

Text: Throughout its existence, Warsaw has been a multi-cultural city. According to the 1901 census, out of 711,988 inhabitants 56.2% were Catholics, 35.7% Jews, 5% Greek orthodox Christians and 2.8% Protestants. Eight years later, in 1909, there were 281,754 Jews (36.9%), 18,189 Protestants (2.4%) and 2,818 Mariavites (0.4%). This led to construction of hundreds of places of religious worship in all parts of the town. Most of them were destroyed in the aftermath of the Warsaw Uprising of 1944. After the war, the new communist authorities of Poland discouraged church construction and only a small number were rebuilt.

Question: What was Warsaw's population in 1901?

Answer: 711,988

Commonsense reasoning

Premise: My favorite song came on the radio.

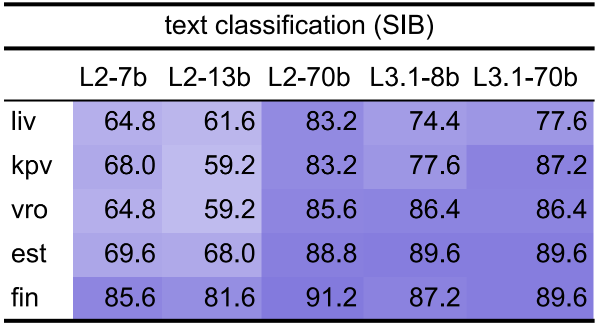
Question: What happened as a result?

Choice1: I covered my ears. **Choice2:** I sang along to it.

Answer: I sang along to it.

How well can open LLMs solve tasks in Finno-Ugric languages?

How well can open LLMs solve tasks in Finno-Ugric languages?



	text classification (SIB)				
	L2-7b	L2-13b	L2-70b	L3.1-8b	L3.1-70b
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

How well can open LLMs solve tasks in Finno-Ugric languages?

	text classification (SIB)				
	L2-7b	L2-13b	L2-70b	L3.1-8b	L3.1-70b
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

	multiple choice QA (Belebele)				
	L2-7b	L2-13b	L2-70b	L3.1-8b	L3.1-70b
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

How well can open LLMs solve tasks in Finno-Ugric languages?

text classification (SIB)

	L2-7b	L2-13b	L2-70b	L3.1-8b	L3.1-70b
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

multiple choice QA (Belebele)

	L2-7b	L2-13b	L2-70b	L3.1-8b	L3.1-70b
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

machine translation (FLORES) (Fiu->En)

	L2-7b	L2-13b	L2-70b	L3.1-8b	L3.1-70b
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

How well can open LLMs solve tasks in Finno-Ugric languages?

text classification (SIB)

	L2-7b	L2-13b	L2-70b	L3.1-8b	L3.1-70b
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

multiple choice QA (Belebele)

	L2-7b	L2-13b	L2-70b	L3.1-8b	L3.1-70b
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

machine translation (FLORES) (Fiu->En)

	L2-7b	L2-13b	L2-70b	L3.1-8b	L3.1-70b
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

commonsense reasoning (XCOPA)

	L2-7b	L2-13b	L2-70b	L3.1-8b	L3.1-70b
est	49.2	51.8	67.6	69.2	92.6

How well can open LLMs solve tasks in Finno-Ugric languages?

text classification (SIB)

	L2-7b	L2-13b	L2-70b	L3.1-8b	L3.1-70b
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

multiple choice QA (Belebele)

	L2-7b	L2-13b	L2-70b	L3.1-8b	L3.1-70b
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

machine translation (FLORES) (Fiu->En)

	L2-7b	L2-13b	L2-70b	L3.1-8b	L3.1-70b
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

commonsense reasoning (XCOPA)

	L2-7b	L2-13b	L2-70b	L3.1-8b	L3.1-70b
est	49.2	51.8	67.6	69.2	92.6

extractive QA (EstQA, TiDyQA) *(exact match)*

	L2-7b	L2-13b	L2-70b	L3.1-8b	L3.1-70b
est	21.89	34.33	49.25	50.75	52.74
fin	51.66	48.34	53.45	58.31	47.06

What is the expected improvement from few-shot prompting over zero-shot prompting?

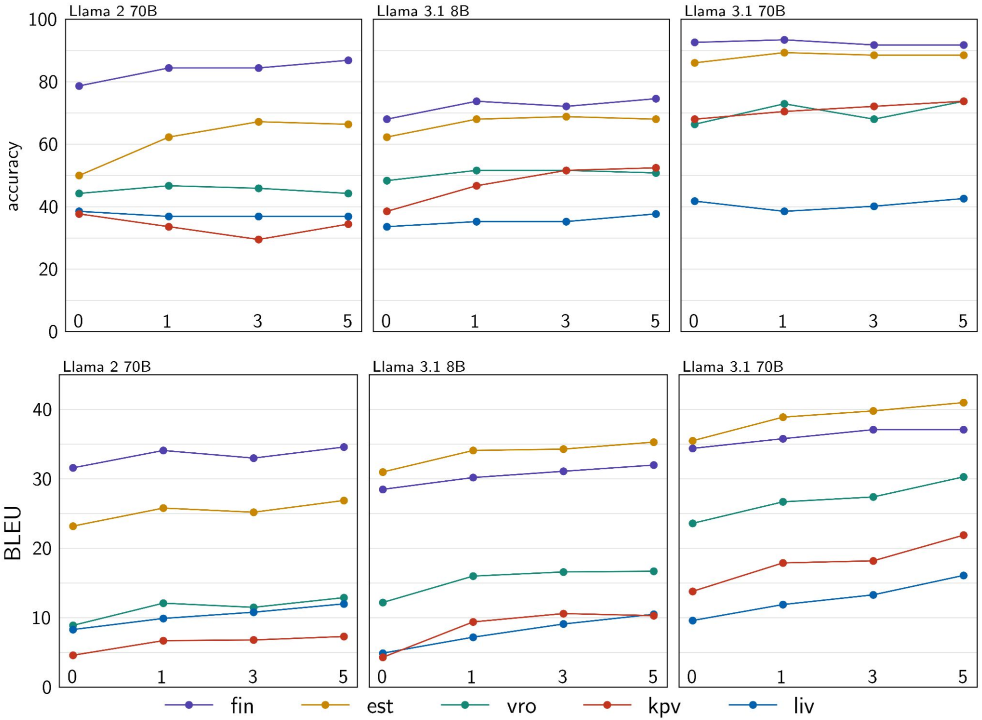
What is the expected improvement from few-shot prompting over zero-shot prompting?

Multiple choice QA

(Belebele)

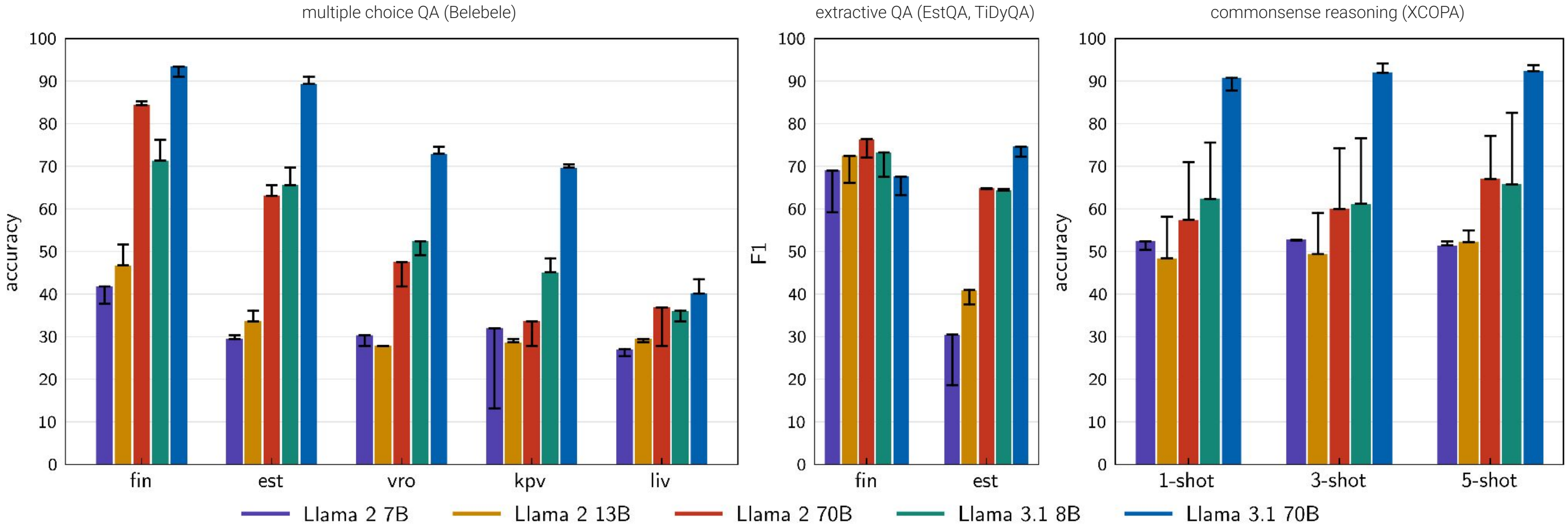
Machine translation

(FLORES)



Can CoT prompting improve the performance of open LLMs on Finno-Ugric languages?

Can CoT prompting improve the performance of open LLMs on Finno-Ugric languages?



Llama 3.1 8B vs Mistral NeMo (12B)

Llama 3.1 8B vs Mistral NeMo (12B)

	Belebele		FLORES		XCOPA	
	L3.1	MN	L3.1	MN	L3.1	MN
liv	33.61	35.25	4.91	5.85	-	-
vro	48.36	50.82	12.19	8.18	-	-
kpv	38.52	36.89	8.18	3.45	-	-
est	62.30	74.59	31.00	33.04	56.80	56.40
fin	68.03	74.59	28.54	30.39	-	-
avg	50.16	**54.43**	**16.96**	16.18	**56.80**	56.40

zero-shot

	Belebele		FLORES		XCOPA		QA	
	L3.1	MN	L3.1	MN	L3.1	MN	L3.1	MN
liv	37.70	37.70	10.50	10.10	-	-	-	-
vro	50.82	50.00	16.72	12.55	-	-	-	-
kpv	52.46	34.43	10.32	6.01	-	-	-	-
est	68.03	83.61	35.28	32.28	69.20	71.60	70.87	71.86
fin	74.59	78.69	31.97	33.24	-	-	75.44	77.39
avg	56.72	**56.89**	**20.96**	18.83	69.20	**71.60**	73.16	**74.63**

five-shot

Llama 2 7B vs Llammas-base (7B) vs Llama 3.1 8B

Llama 2 7B vs Llammas-base (7B) vs Llama 3.1 8B

	Belebele			FLORES			XCOPA		
	L2	Lam	L3.1	L2	Lam	L3.1	L2	Lam	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	-	-	-
kpv	26.23	29.51	38.52	2.88	1.44	8.18	-	-	-
est	22.95	39.34	62.30	8.53	28.90	31.00	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**

zero-shot

	Belebele			FLORES			XCOPA			QA		
	L2	Lam	L3.1	L2	Lam	L3.1	L2	Lam	L3.1	L2	Lam	L3.1
liv	26.23	23.77	37.70	6.76	7.70	10.50	-	-	-	-	-	-
vro	27.05	31.97	50.82	7.83	16.23	16.72	-	-	-	-	-	-
kpv	27.87	24.59	52.46	5.36	3.64	10.32	-	-	-	-	-	-
est	28.69	36.89	68.03	12.65	34.29	35.28	49.20	68.20	69.00	35.35	63.76	70.87
fin	44.26	27.87	74.59	29.63	18.36	31.97	-	-	-	70.63	56.32	75.44
avg	30.82	29.02	**56.72**	12.44	16.04	**20.96**	49.20	68.20	**69.00**	52.99	60.04	**73.16**

five-shot

How Well do LLMs know Finno-Ugric Languages?

- Llama 3.1 > Llama 2 models
- Few-shot prompting and CoT prompting – beneficial to some level
- Near-perfect results with Llama 3.1 70B => we need stronger benchmarks
- Mistral NeMo > Llama 3.1 8B for Estonian and Finnish
- Stronger base model > weaker language adapted base model

https://github.com/TartuNLP/smugri-lm-eval-configs

TARTUNLP

How Well do LLMs know Finno-Ugric Languages?

- Llama 3.1 > Llama 2 models
- Few-shot prompting and CoT prompting – beneficial to some level
- Near-perfect results with Llama 3.1 70B => we need stronger benchmarks
- Mistral NeMo > Llama 3.1 8B for Estonian and Finnish
- Stronger base model > weaker language adapted base model

https://github.com/TartuNLP/smugri-lm-eval-configs

Questions?