Prompting Large Language Models for Church Slavic Translation

Edoardo Signoroni and Pavel Rychlý

e.signoroni@mail.muni.cz, pary@fi.muni.cz

RASLAN 2025 – December 5, 2025

Introduction

MT has made great progress for living, high-resource languages

Most of the world's languages are, however, **underresourced** and **underrepresented**

Historical languages are often in this situation

Church Slavic lacks both data and human experts

Introduction

LLMs have show promise for **zero-shot LRs MT** and other tasks

Can **off-the-shelf LLMs** aid scholars in the translation process?

How do LLMs fare on **historical LRs MT**?

Church Slavic

A **Slavic language** created in the 9th cen. by Cyril and Methodius to Christianize the Moravian Slavs

Based on the South Slavic Macedonian dialects around Thessalonica

The **first Slavic literary language**, written in Glagolitic and **Cyrillic**

Adopted in other Slavic regions and evolved

Current NLP work divide it chronologically in Early (9-11th cc.), Middle (12-16th cc.), Late (17-18th cc.) and geographically in South and East (based on available data)

NLP for Church Slavic

Church Slavic has **little presence in NLP** research.

Two main variants: Old Church Slavic and Old East Slavic (also in Shared Tasks and Workshops)

Recent work include **datasets**, **tools**, and wider **interdisciplinary studies**

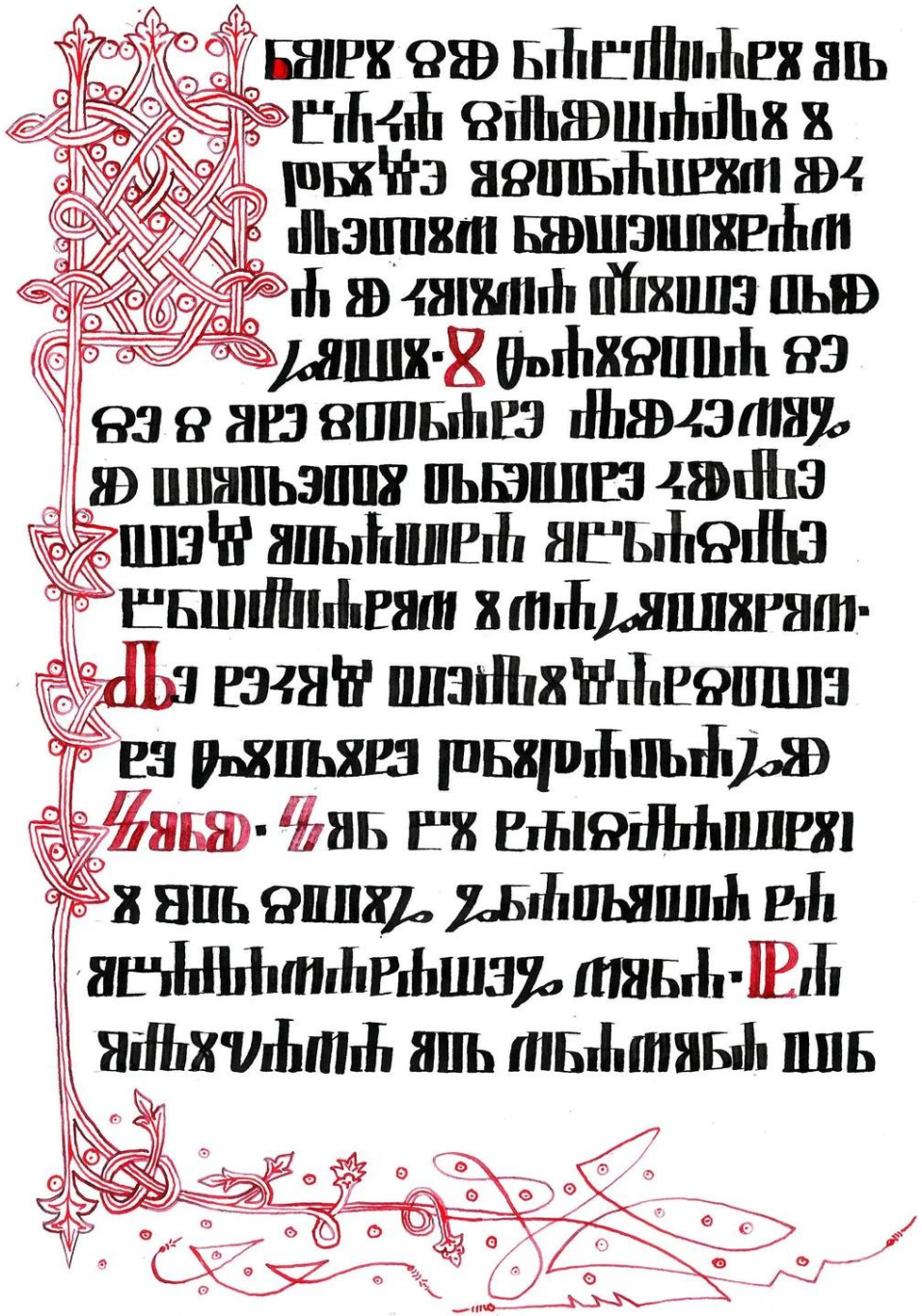
Notably, some work employs LLMs to classify **diachronic** and **geographical** (Lendvai et al., 2025 RANLP) variants, and in **translation to aid automated alignment** (Janssen et al., RANLP 2025)

Dataset

The data is a section of the corpus created by Jenssen et al. (RANLP 2025) from five different versions of ***De Lepra ad Sistelium*** **by Methodius Olympius**, an allegorical interpretation of the norms on leprosy from the Old Testament.

The earliest manuscripts are preserved in a **10th cen. Old Church Slavic translation**

We take the **Slavic** and **German** from a recent edition (Jouravel et al. 2024) and manually align the **English** from a public domain translation (Cleminson and Eastbourne, 2015)



Dataset

We have **two translation directions:**

Church Slavic - English

Church Slavic - German

with 262 lines each.

Example of Data

"src":	"Откуду, о Евъвулие?",
"eng":	"“Whence have you come, O Eubulius?”",
"deu":	"Woher kommst du, Eubulius?“

Experiments

We carry out **two experiments**:

I. **Zero-shot and few shot** prompting

0,1,3 shots

three small LMs (7-9B)

one large (120B)

II. LLM **Size**

same LM

five different sizes

(0.5-14B)

Experiments

Few-shots prompt

```
"You will be asked to translate a sentence from Old Church Slavonic
to {target_language}. You will output only the translation in
{target_language}, without any additional text. Here are some
examples:
Old Church Slavonic: {example}
{target_language}: {example}
...
```

Zero-shot prompt

```
Translate the following text from Old Church Slavonic to
{target_language}:
{example}
Write only the translation without any additional text."
```

Zero and Few-shot Prompting

We prompt **four models**:

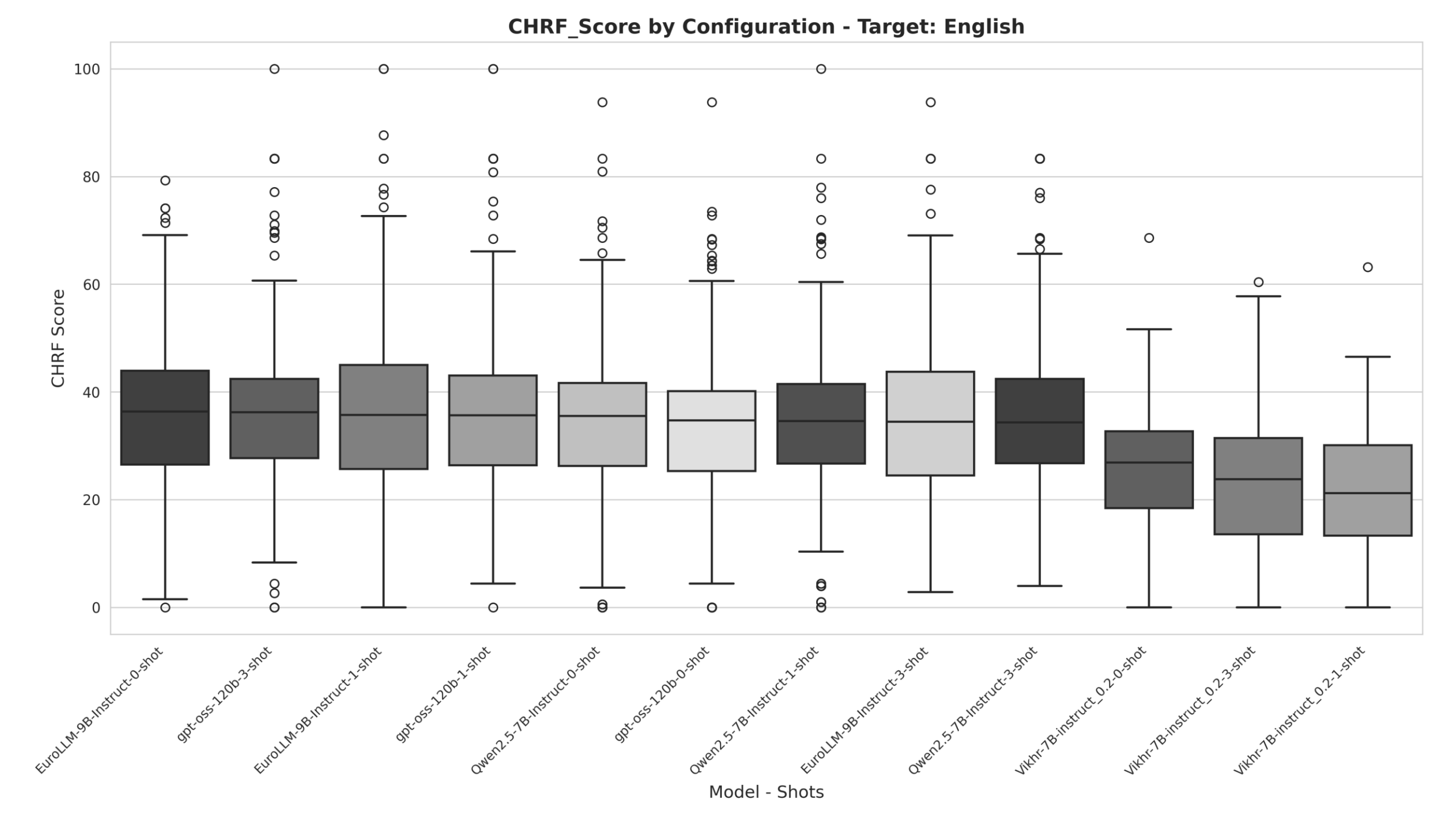
- EuroLLM-9B-Instruct (Martins et al. 2024)
- Qwen2.5-7B-Instruct (Yang et al. 2024)
- Vikhr-7B-Instruct (Nikolich et al. 2024)
- gpt-oss-120b (OpenAI, 2025, local instance)

In **three settings**:

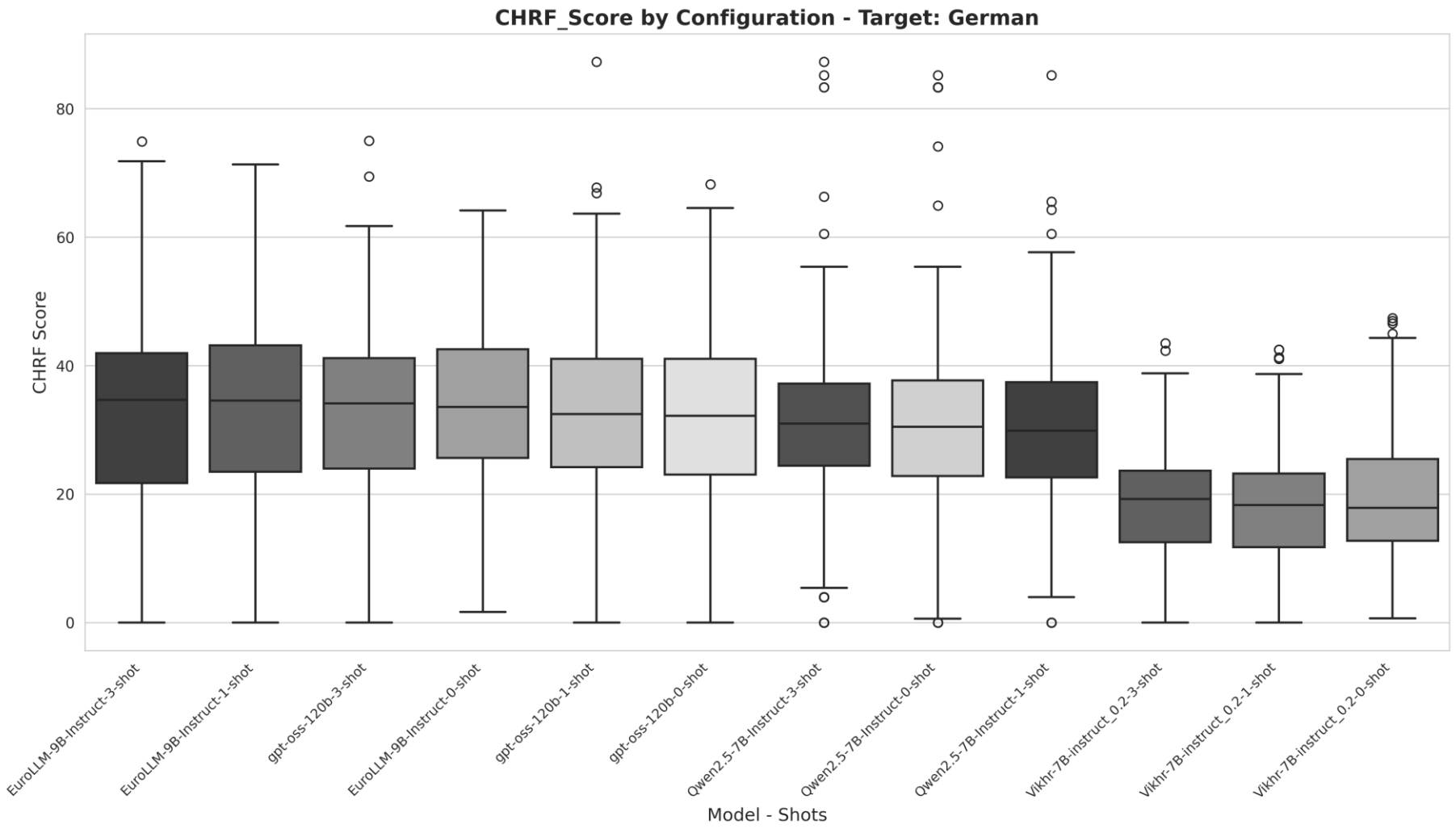
- 0,1,3-shot

Zero and Few-shot Prompting

(a) Church Slavic to English					(b) Church Slavic to German				
Model	**Shots**	**chrF**	**COMET**	**BLEU**	**Model**	**Shots**	**chrF**	**COMET**	**BLEU**
EuroLLM-9B	0-shot	35.6	0.614	0.065	EuroLLM-9B	0-shot	**33.5**	0.578	0.055
EuroLLM-9B	1-shot	35.8	0.598	0.077	EuroLLM-9B	1-shot	33.0	0.564	0.051
EuroLLM-9B	3-shot	35.1	0.610	0.069	EuroLLM-9B	3-shot	32.2	0.555	0.051
gpt-oss-120b	0-shot	33.7	0.608	0.061	gpt-oss-120b	0-shot	31.7	0.560	0.038
gpt-oss-120b	1-shot	**36.3**	0.620	0.066	gpt-oss-120b	1-shot	32.0	0.568	0.038
gpt-oss-120b	3-shot	**36.3**	0.617	0.064	gpt-oss-120b	3-shot	32.3	0.569	0.044
Qwen2.5-7B	0-shot	34.1	0.596	0.050	Qwen2.5-7B	0-shot	30.8	0.535	0.025
Qwen2.5-7B	1-shot	34.1	0.597	0.046	Qwen2.5-7B	1-shot	29.6	0.537	0.025
Qwen2.5-7B	3-shot	35.6	0.617	0.043	Qwen2.5-7B	3-shot	30.6	0.523	0.032
Vikhr-7B	0-shot	25.3	0.496	0.015	Vikhr-7B	0-shot	19.4	0.406	0.004
Vikhr-7B	1-shot	21.4	0.435	0.008	Vikhr-7B	1-shot	17.9	0.338	0.003
Vikhr-7B	3-shot	22.8	0.449	0.010	Vikhr-7B	3-shot	18.1	0.351	0.001


CHRF_Score by Configuration - Target: English
CHRF Score
100
80
60
40
20
0
EuroLLM-9B-Instruct-0-shot
gpt-oss-120b-3-shot
EuroLLM-9B-Instruct-1-shot
gpt-oss-120b-1-shot
Qwen2.5-7B-Instruct-0-shot
gpt-oss-120b-0-shot
Qwen2.5-7B-Instruct-1-shot
EuroLLM-9B-Instruct-3-shot
Qwen2.5-7B-Instruct-3-shot
Vikhr-7B-instruct_0.2-0-shot
Vikhr-7B-instruct_0.2-3-shot
Vikhr-7B-instruct_0.2-1-shot
Model - Shots

Zero and Few-shot Prompting



Weirdly, Vikhr is the worst model

All the other models are comparable, regardless of size (even gpt-oss-120b)

Minimal differences in zero and few-shot configurations

Higher scores for English translation

LLM Size

We prompt **five models** of the same type:

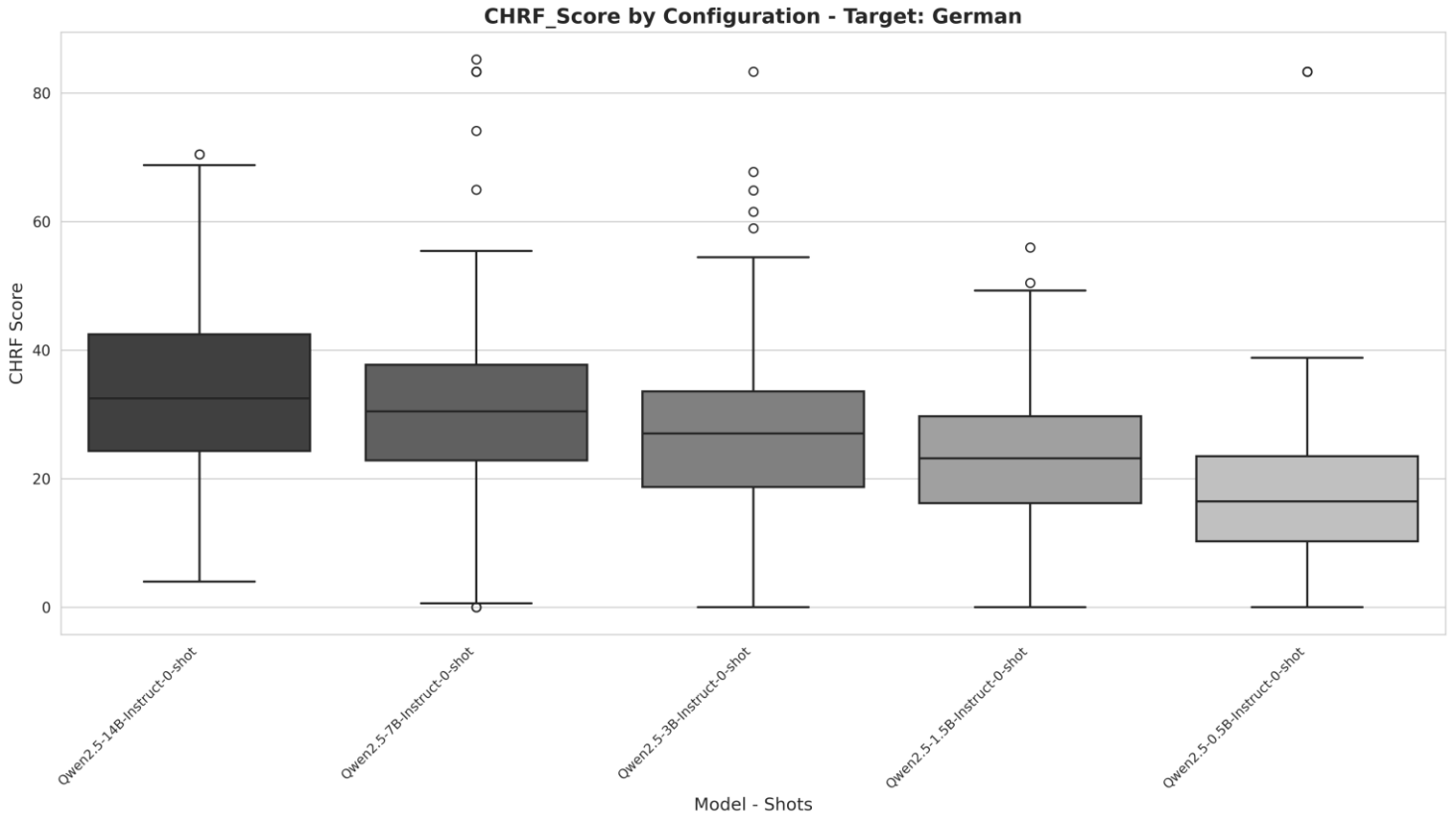
- Qwen2.5-0.5B-Instruct
- Qwen2.5-1.5B-Instruct
- Qwen2.5-3B-Instruct
- Qwen2.5-7B-Instruct
- Qwen2.5-14B-Instruct

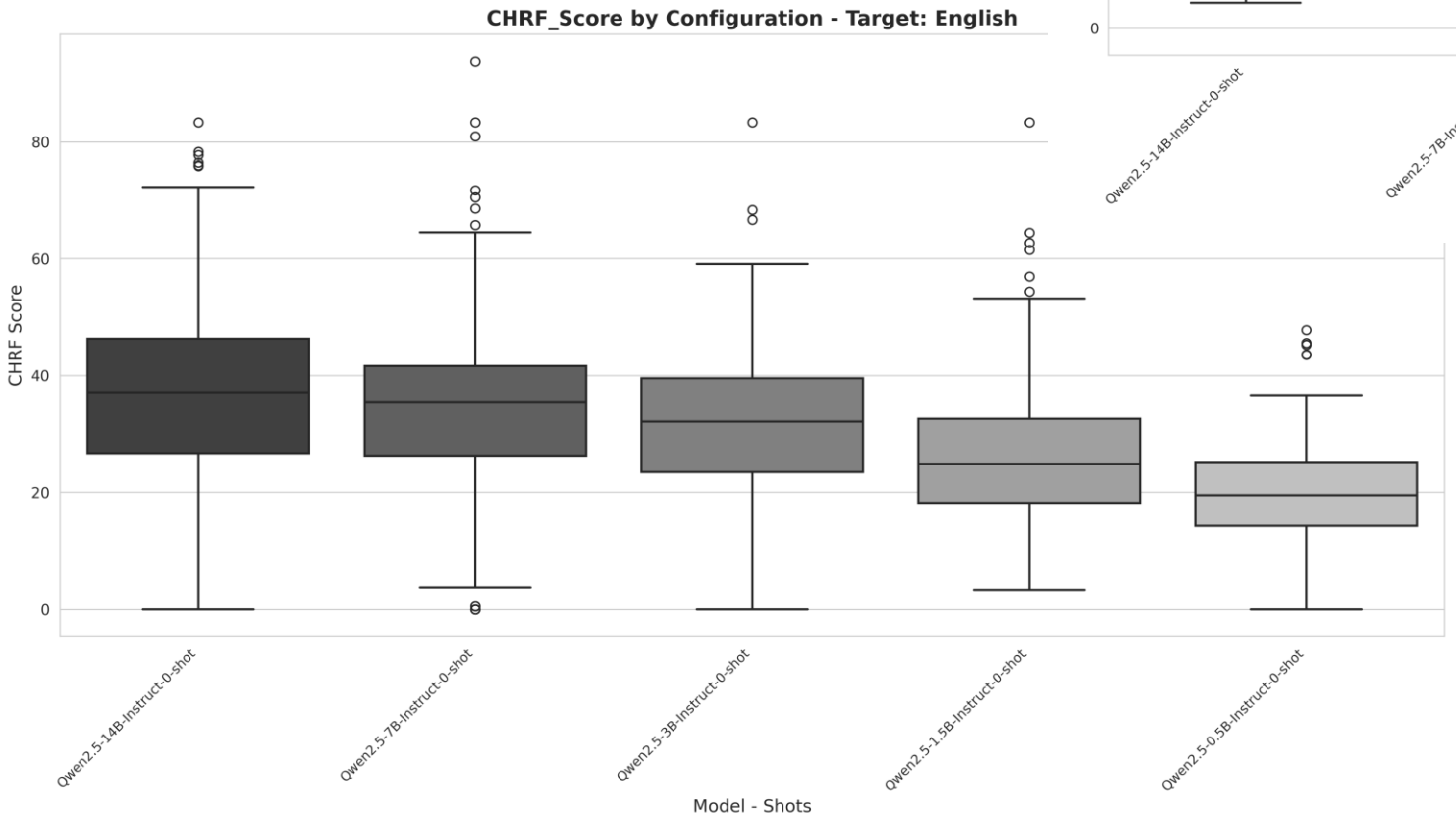
In a **zero-shot** scenario

LLM Size

(a) Church Slavic to English				(b) Church Slavic to German			
Model	**chrF**	**COMET**	**BLEU**	**Model**	**chrF**	**COMET**	**BLEU**
Qwen2.5-14B	**36.4**	0.629	0.075	Qwen2.5-14B	**32.7**	0.566	0.044
Qwen2.5-7B	34.1	0.596	0.050	Qwen2.5-7B	30.8	0.535	0.025
Qwen2.5-3B	31.2	0.578	0.037	Qwen2.5-3B	26.7	0.487	0.014
Qwen2.5-1.5B	26.1	0.550	0.016	Qwen2.5-1.5B	23.0	0.455	0.008
Qwen2.5-0.5B	19.5	0.479	0.004	Qwen2.5-0.5B	18.4	0.387	0.001

Larger models perform better, but with diminishing returns





Again, English fares better than German

Limitations

Small dataset from a single work and genre > Expand the corpus

Work with human annotators < **Automated metrics** for evaluation

Limited **coverage of model** families > consider other approaches

Fine-tune models < Only zero and few-shot **prompting**

Conclusions

LLMs have the potential to **aid human experts** in translating and studying historical texts

Few-shots and size may have a lower impact than **model and training data selection**

Performance varies by language pair