Universität Zürich^UZH

nccr evolving language NATIONAL CENTRE OF COMPETENCE IN RESEARCH

30th ANNIVERSARY EMNLP 2025 Suzhou, China | 中国苏州

Kirill Semenov, Rico Sennrich
University of Zurich
kirill.semenov@uzh.ch

Paper: Github:

Measuring the Effect of Disfluency in Multilingual Knowledge Probing Benchmarks

Template-Based Benchmarks Underestimate Multilingual Factual Retrieval

Problem

English factual benchmarks like LAMA: (Petroni et al., 2019)

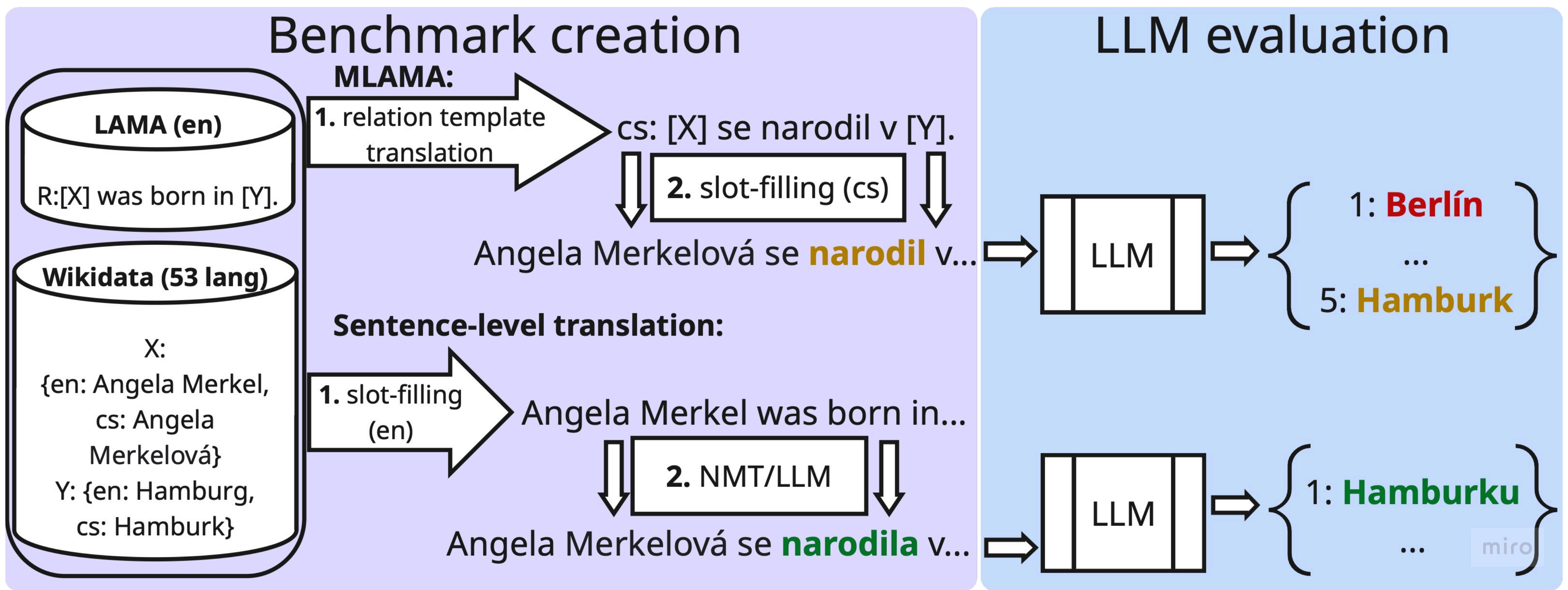
- prompt LLM with template filled with NEs

multilingual benchmarks like MLAMA: (Kassner et al., 2021)

- translate templates and NEs separately

→ ungrammatical or poorly worded prompts

How much do we lose with disfluencies?



Benchmark creation

LLM evaluation

LAMA (en)

R:[X] was born in [Y].

Wikidata (53 lang)

X: {en: Angela Merkel, cs: Angela Merkelová}
Y: {en: Hamburg, cs: Hamburk}

MLAMA:

1. relation template translation

cs: [X] se narodil v [Y].

2. slot-filling (cs)

Angela Merkelová se **narodil** v...

Sentence-level translation:

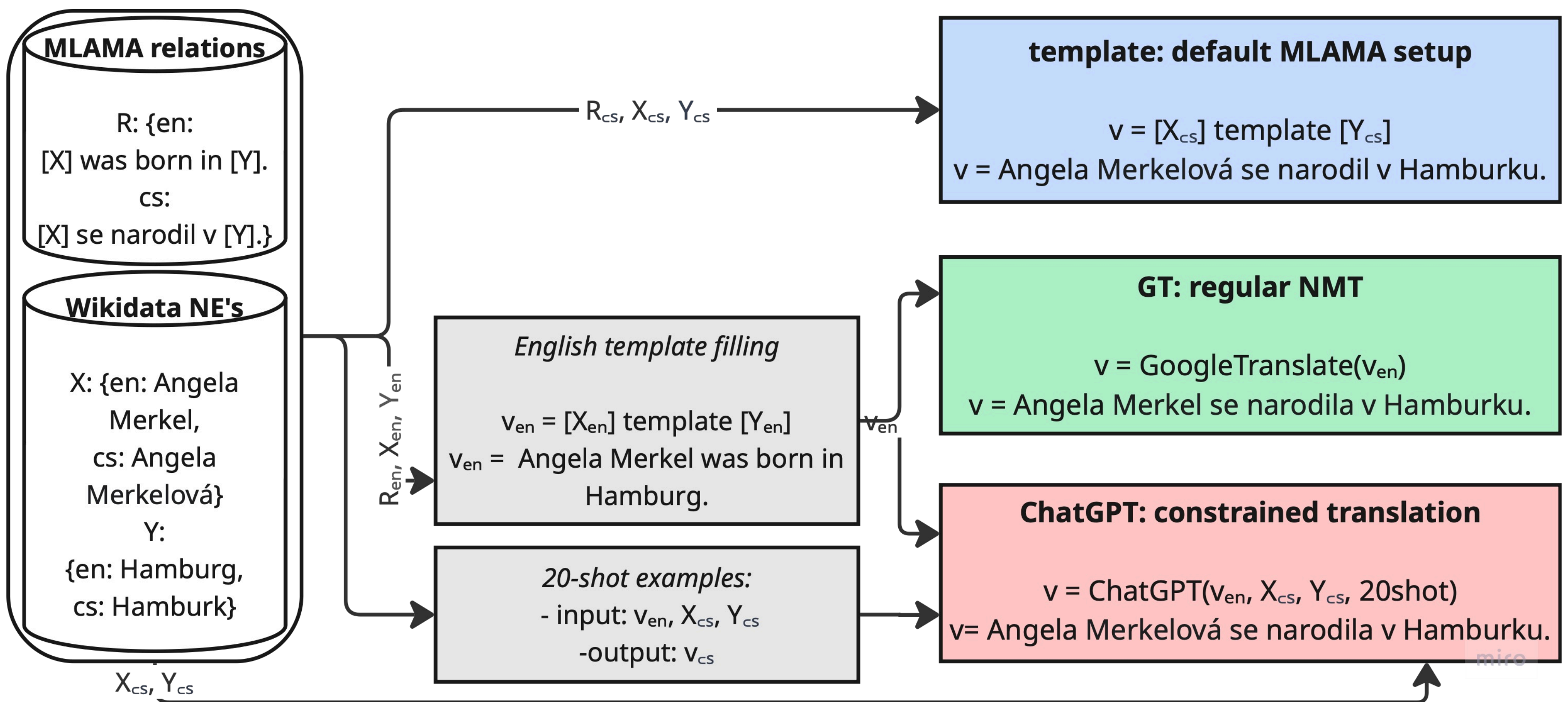
1. slot-filling (en)

Angela Merkel was born in...

2. NMT/LLM

Angela Merkelová se **narodila** v...

Method



MLAMA relations

R: {en: [X] was born in [Y]. cs: [X] se narodil v [Y].}

Wikidata NE's

X: {en: Angela Merkel, cs: Angela Merkelová}
Y: {en: Hamburg, cs: Hamburk}

English template filling

v_{en} = [X_{en}] template [Y_{en}]
v_{en} = Angela Merkel was born in Hamburg.

20-shot examples:
- input: v_{en}, X_{cs}, Y_{cs}
-output: v_{cs}

template: default MLAMA setup

v = [X_{cs}] template [Y_{cs}]
v = Angela Merkelová se narodil v Hamburku.

GT: regular NMT

v = GoogleTranslate(v_{en})
v = Angela Merkel se narodila v Hamburku.

ChatGPT: constrained translation

v = ChatGPT(v_{en}, X_{cs}, Y_{cs}, 20shot)
v= Angela Merkelová se narodila v Hamburku.

Languages	4 Slavic (ru, cs, uk, hr)	5 Non-Slavic (es, zh, vi, id, da)
Verbalizations	Template, GT, ChatGPT	Template, GT
Relations	15	8
Model	Llama 2-7b base (Touvron et al., 2023)	
Metrics	$R@n$ ↑: correct object in top n ranks Correct object ranks distribution ↓	

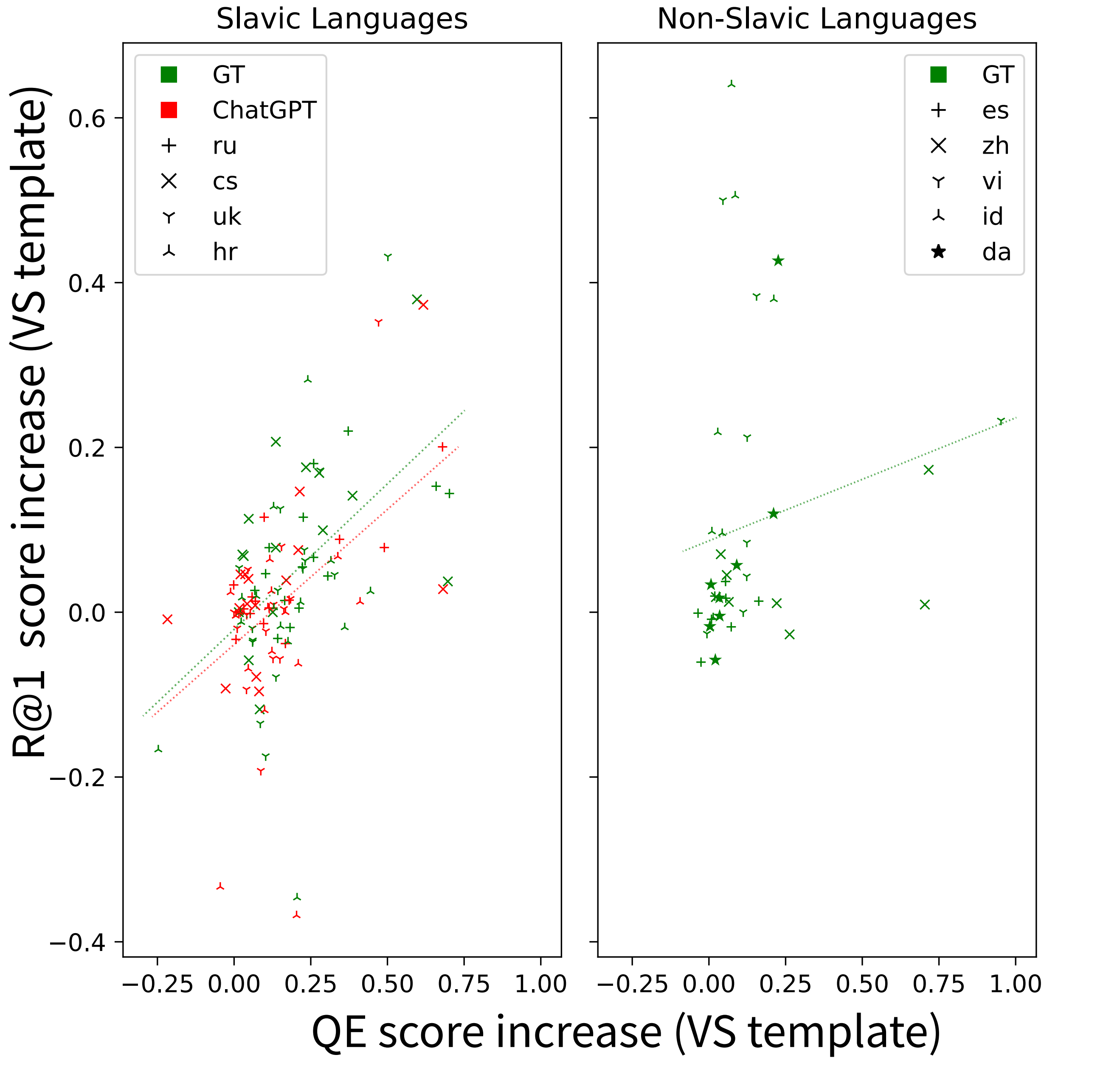
Results

Sentence-level translation improves retrieval in 7 of 9 languages

R@1 scores

Verbalization	Slavic				Non-Slavic				
	ru	cs	uk	hr	es	zh	vi	id	da
Template	.512	.579	.488	**.707**	**.725**	.020	.587	.547	.568
GT	**.586**	**.670**	**.525**	.704	**.725**	**.059**	**.765**	**.786**	**.640**
ChatGPT	.545	.615	.492	.653	–	–	–	–	–

Fluency ∝ Factual Retrieval: Predictors



Discussion

Why does GT perform better than ChatGPT?

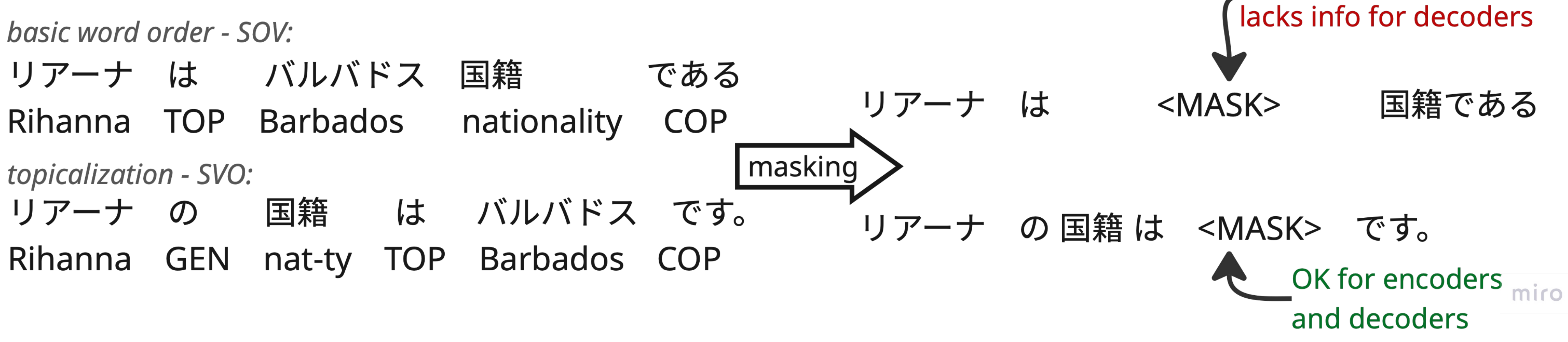
One of the factors - **explicitation**: adding descriptions to NEs

Template:		Times	было	написано	в	английский	язык.
		Times	was-N	written-N	in	English-NOM	language.NOM
GT:	**Журнал**	Times	издавался	на	английском	языке.
	Journal	*Times*	was.issued.M	in	English-INS	lang-INS

Perspectives

How to prompt languages multilingually?

Rihanna was born in Barbados: en -> ja



basic word order - SOV:

リアーナ	は	バルバドス	国籍	である
Rihanna	TOP	Barbados	nationality	COP

topicalization - SVO:

リアーナ	の	国籍	は	バルバドス	です。
Rihanna	GEN	nat-ty	TOP	Barbados	COP

masking

リアーナ　は　<MASK>　国籍である

OK for encoders
lacks info for decoders

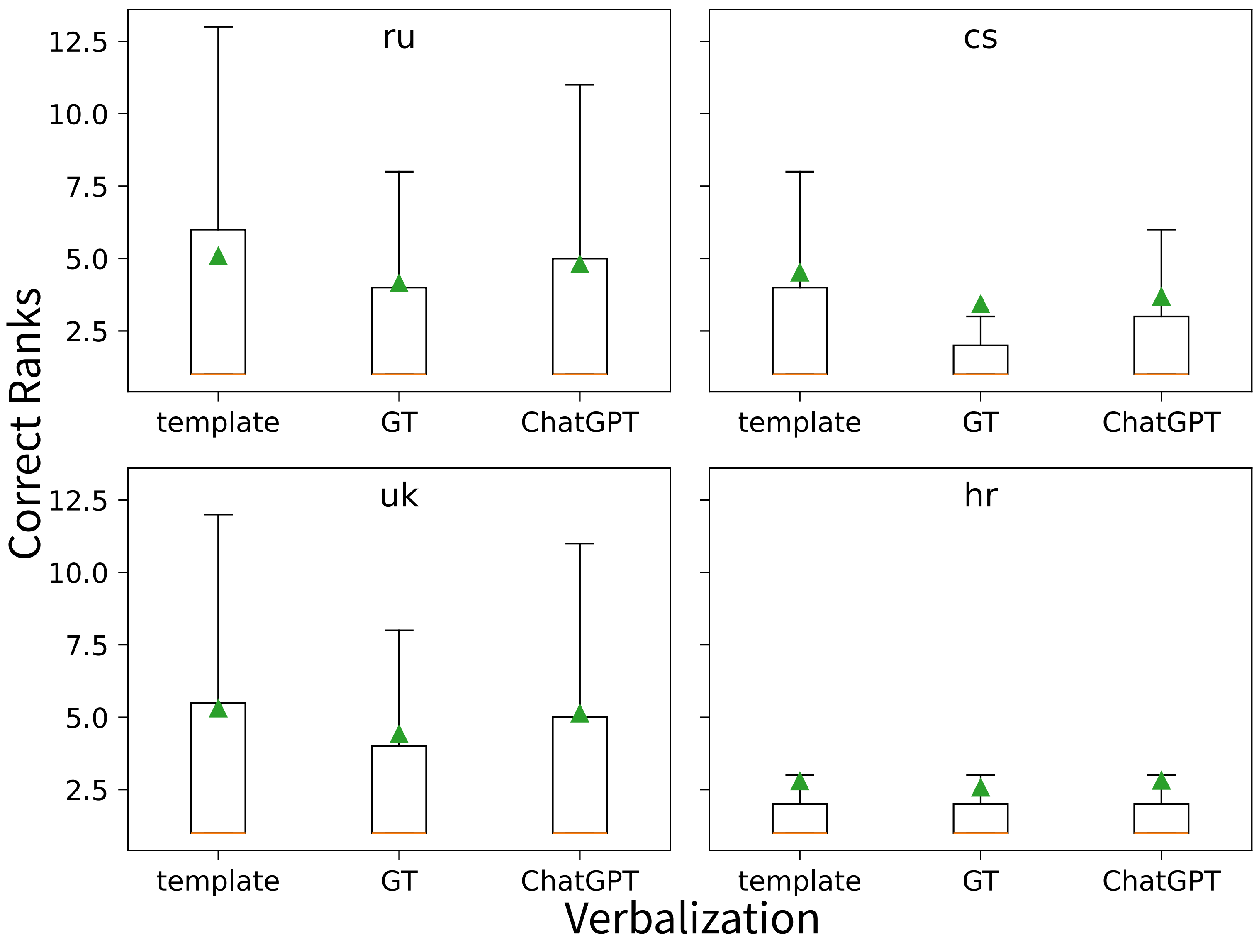
リアーナ　の 国籍 は　<MASK>　です。

OK for encoders and decoders

Conclusions

- for most languages, templates significantly decrease factual retrieval
- simple translation of full sentences with NMT helps
- we need to think more how to:
  - evaluate grammaticality multilingually
  - adapt the behavioral datasets of the encoder "era" to decoder models

Correct NE ranks move towards 1



This research was funded by NCCR Evolving Language, Swiss National Science Foundation Agreement 51NF40_180888.