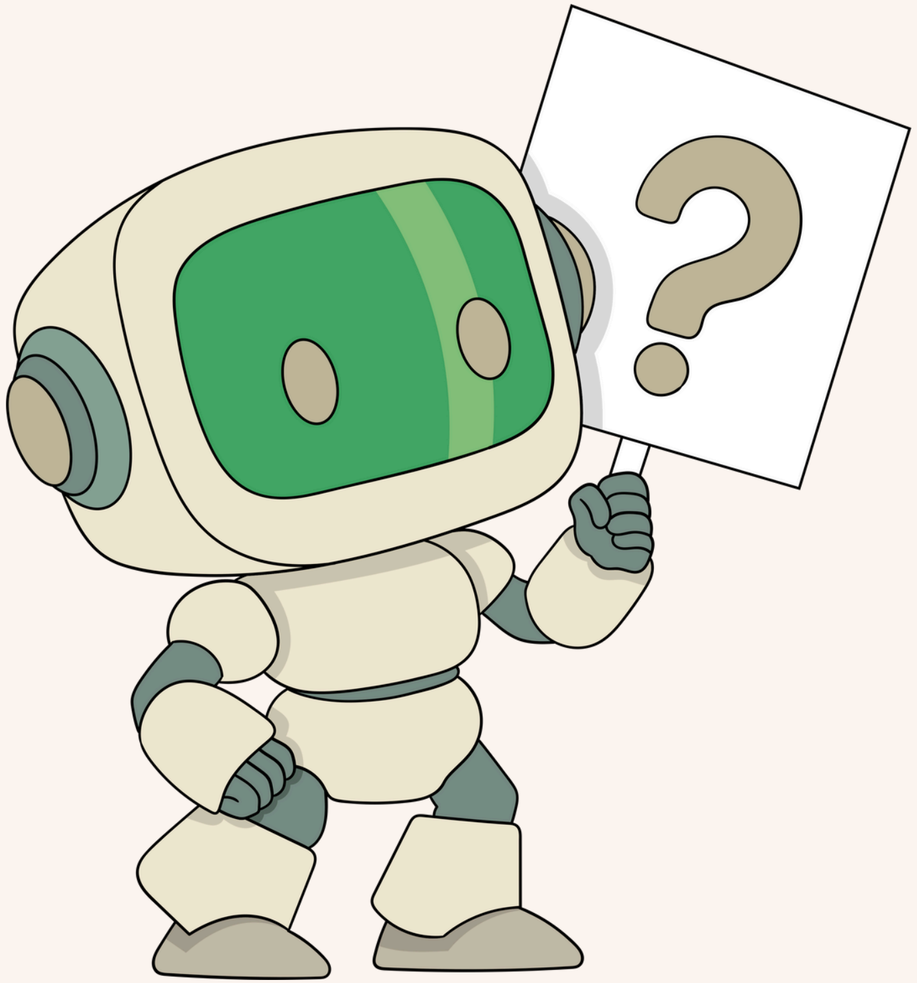


Iván Martínez Murillo

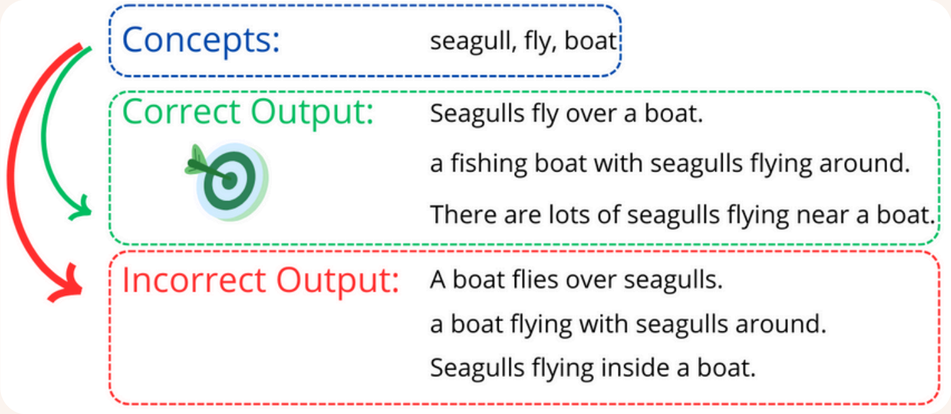
Do LLMs exhibit the same commonsense capabilities across languages?

Universitat d'Alacant
Universidad de Alicante

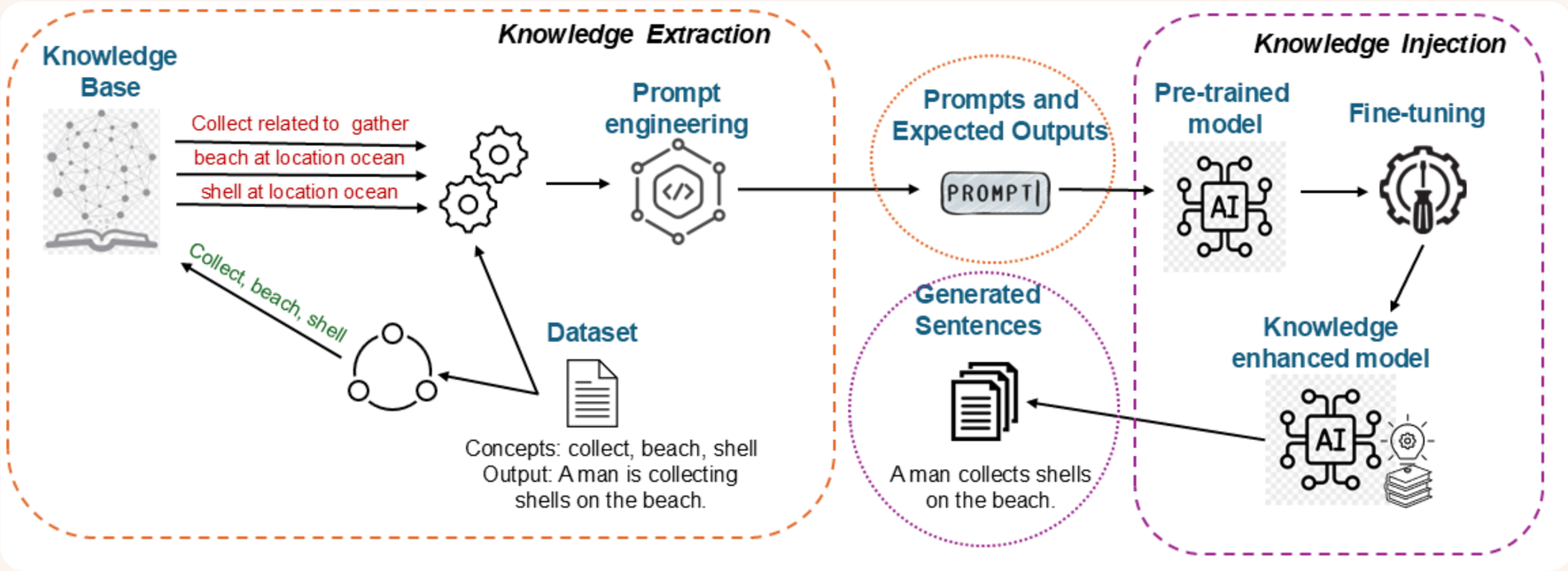
Index

1. Background

CommonGen



Knowledge-Enhanced T5 & BART



1. Background

Limitations:

- Lack of a faithful evaluation method for open-ended commonsense generation
- Research focuses solely on English
- Benchmarks use outdated models

2. Hypothesis

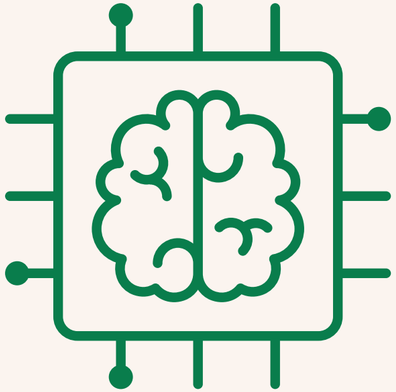
LLMs do not exhibit the same reasoning and commonsense capabilities in other languages as they do in English.

3. Objectives

1

Extend and propose a multilingual dataset to evaluate the commonsense reasoning of LLMs.

2



Test the commonsense capabilities of LLMs in a multilingual setting.

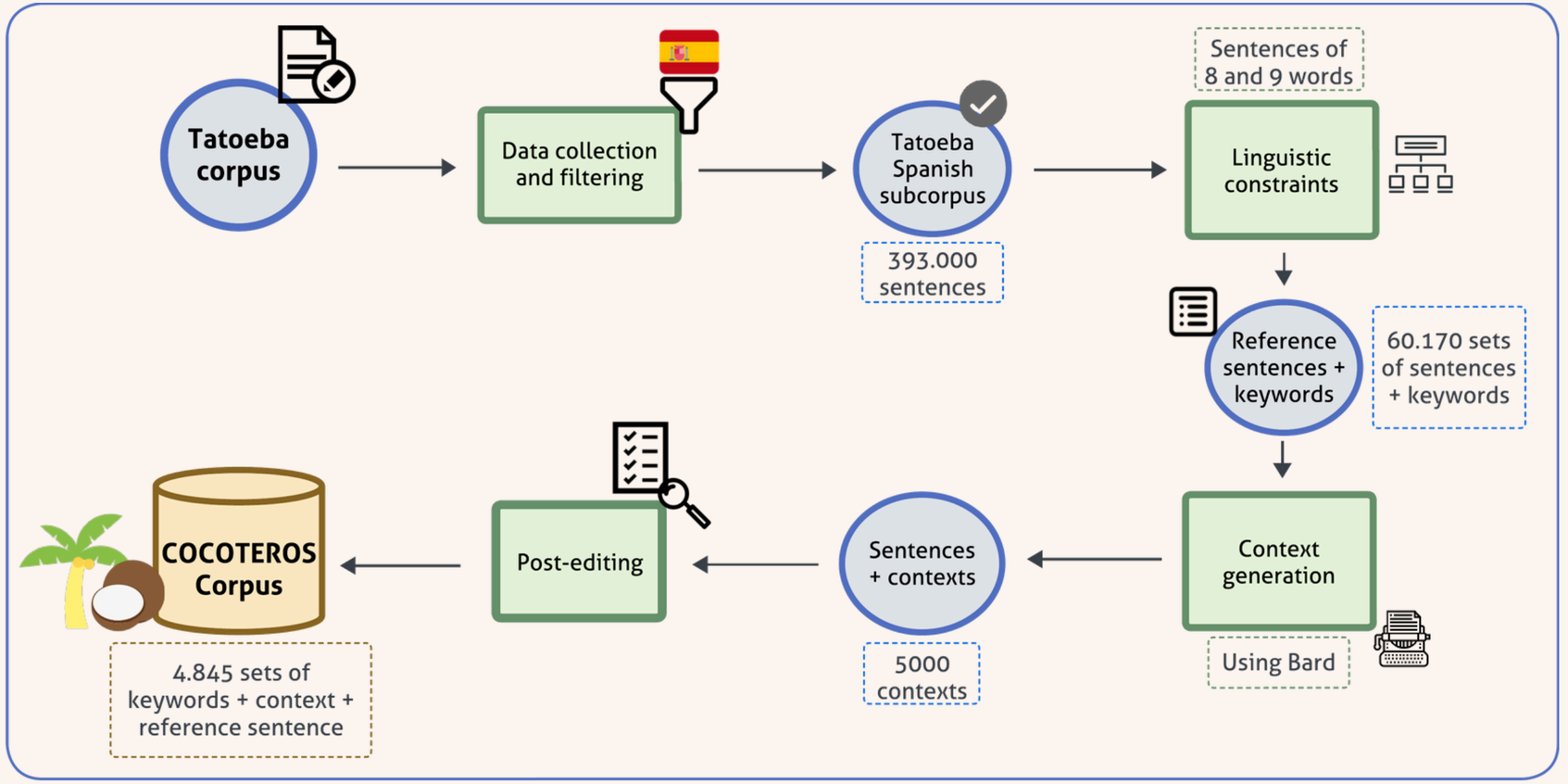
3

Find the optimal way to evaluate this open-ended, multilingual commonsense generation task.

4. Initial Dataset

COCOTEROS dataset:

https://huggingface.co/datasets/gplsi/cocoteros



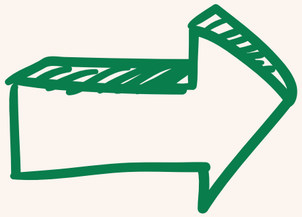
4. Initial Dataset

COCOTEROS dataset:
https://huggingface.co/datasets/gplsi/cocoteros

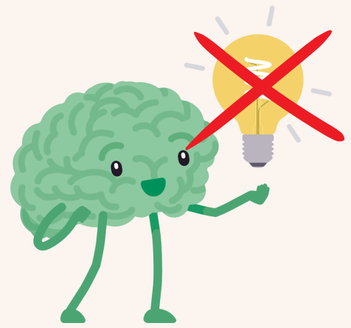
Reference sentence	Keywords	Generated context
Debes dejar de actuar de forma tan tonta. --- *You should stop acting in such a dumb way.*	['dejar', 'actuar', 'forma'] --- *['stop', 'act', 'way']*	Tu comportamiento es inapropiado y está perjudicando tu reputación. Por favor, sé más maduro y responsable. --- *Your behaviour is inappropriate and it's damaging your reputation. Please, be more mature and responsible.*
El vecindario entero se sorprendió por la noticia. --- *The whole neighbourhood was surprised by the news.*	['sorprender', 'vecindario', 'noticia'] --- *['surprise', 'neighbourhood', 'news']*	La noticia del cierre de la escuela fue como un jarro de agua fría para los vecinos. --- *The news of the school's closure came as a complete shock to the neighbours.*
Te prometo que las cosas van a cambiar. --- *I promise you things are going to change.*	['prometer', 'cambiar', 'cosa'] --- *['promise', 'change', 'things']*	Te aseguro que todo va a mejorar. Estoy trabajando duro para que eso suceda. --- *I assure you that everything will get better. I'm working hard to make it happen.*

5. Dataset Expansion

Cocoteros



Llama-3.2-3B-Instruct



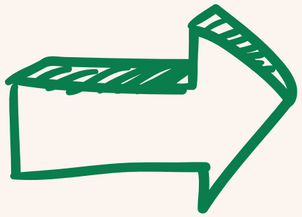
Conterfactual sentence

Ensuring:

1. *Uses all the keywords*
2. *It's fluent in Spanish (not containing grammatical errors)*
3. *Is non sensical*

5. Dataset Expansion

Cocoteros



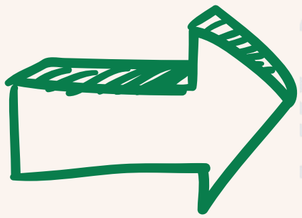
From the reference sentences

Non related sentence

Ensuring:

1. *It does not relate to the same input*
2. *Does not contain any keyword*

5. Dataset Expansion

Cocoteros dataset	Translating the whole dataset to:	Fixing keywords	Translated Version
1 	OPUS mt		
	English		
2  	Salt.usu	Grok 	
	Valencian		
3 	OPUS mt		
	Dutch		

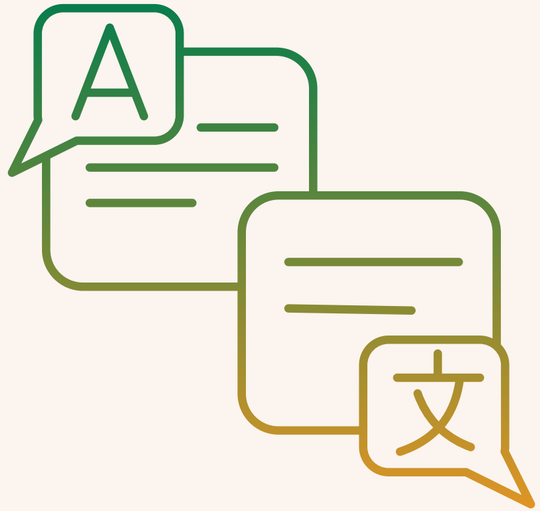
6. Benchmark

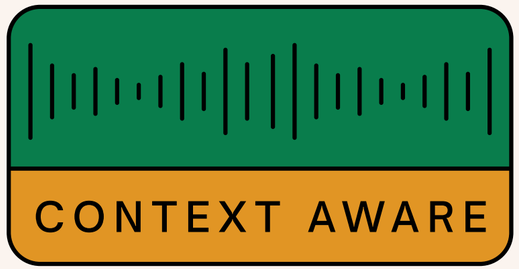
Two focus:

1. *Multilingual capacity*
2. *Context helpfulness*

Five families:

1. *LLaMA*
2. *Gemma*
3. *EuroLLM*
4. *Qwen*
5. *Salamandra*

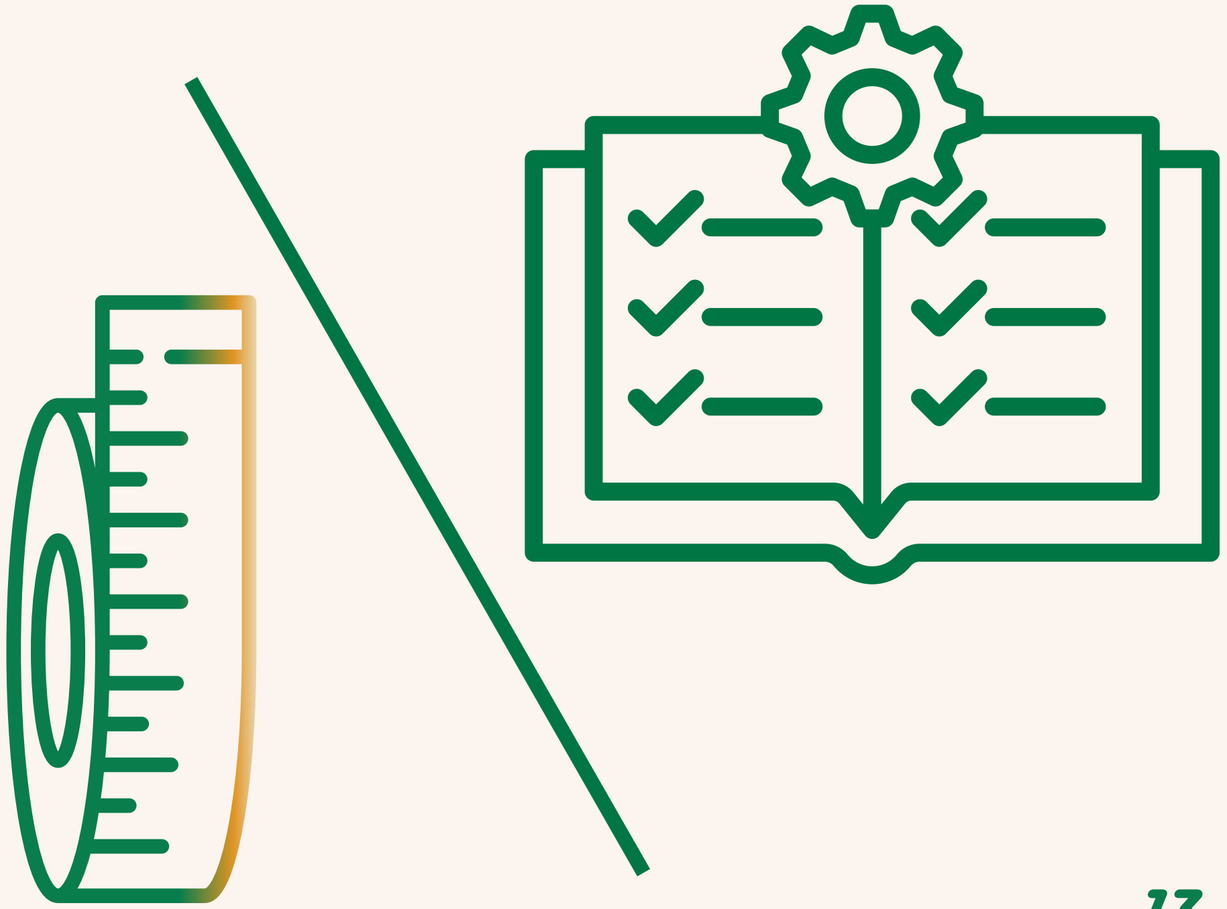




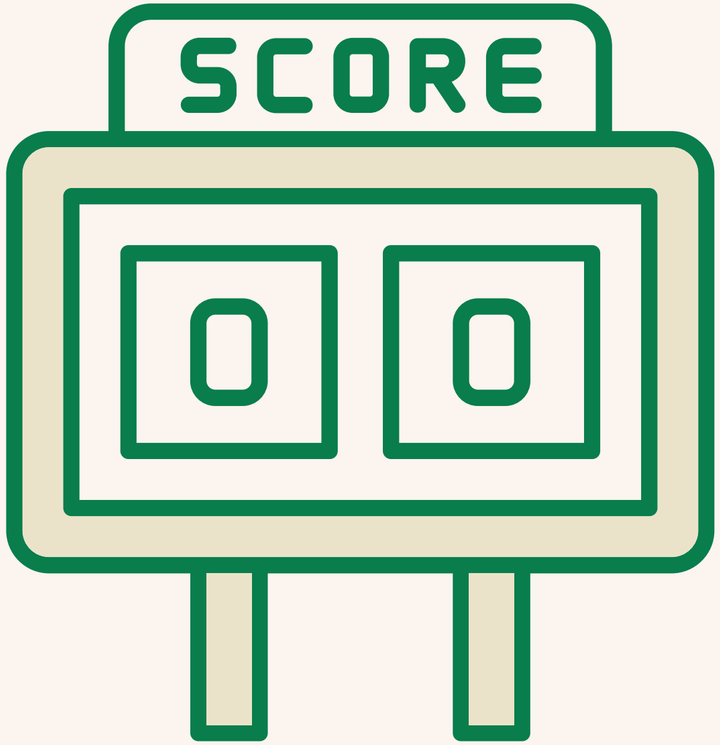
7. LLM Selection

We present a benchmark of 20 LLMs

1. *Llama-3.2-(1/3)B(-Instruct)*
2. *Gemma-2-(2/9)b(-it)*
3. *EuroLLM(1.7/9)B(-Instruct)*
4. *Qwen3-(4/8)B(-Base)*
5. *Salamandra-(2/7)b(-Instruct)*



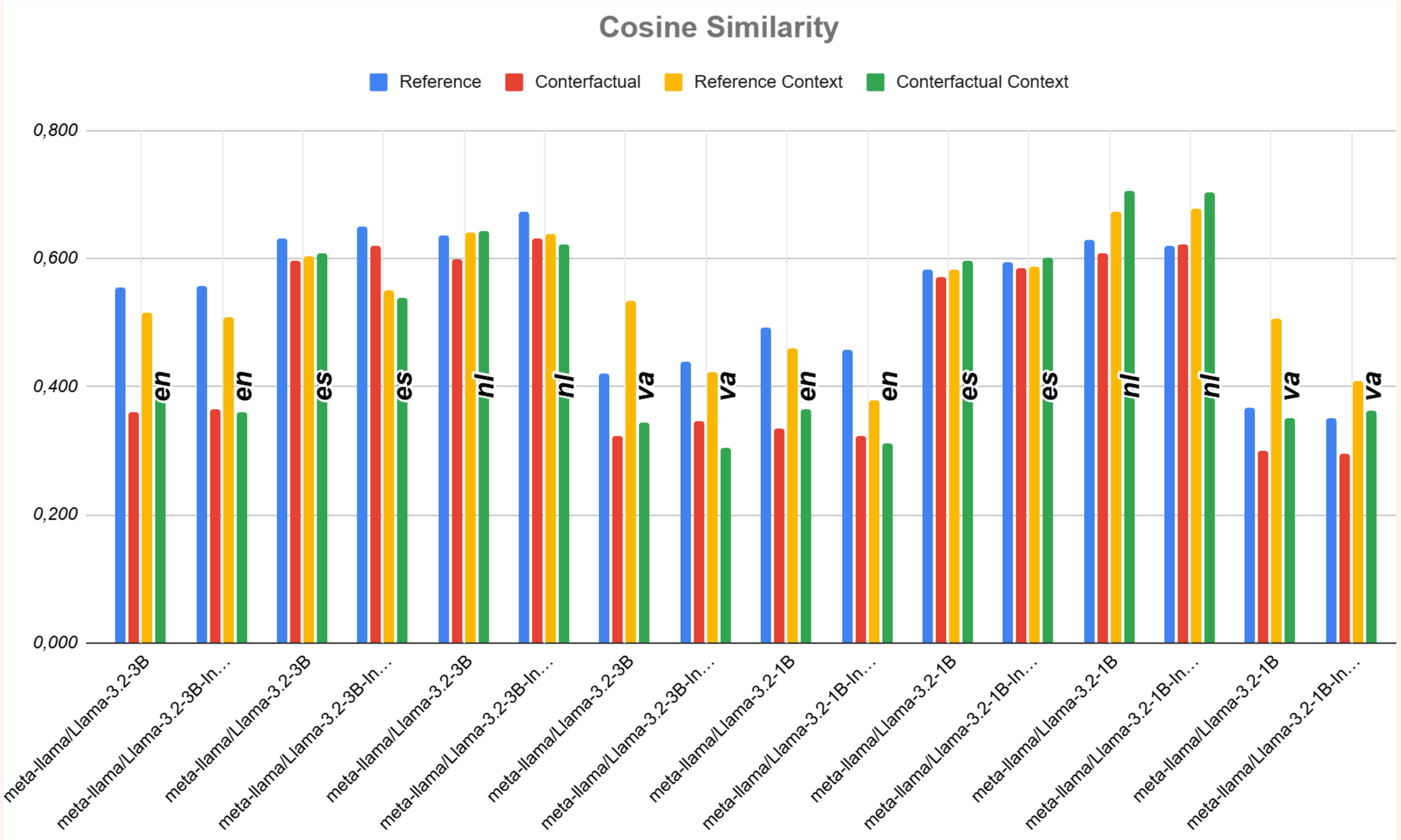
8. Evaluation



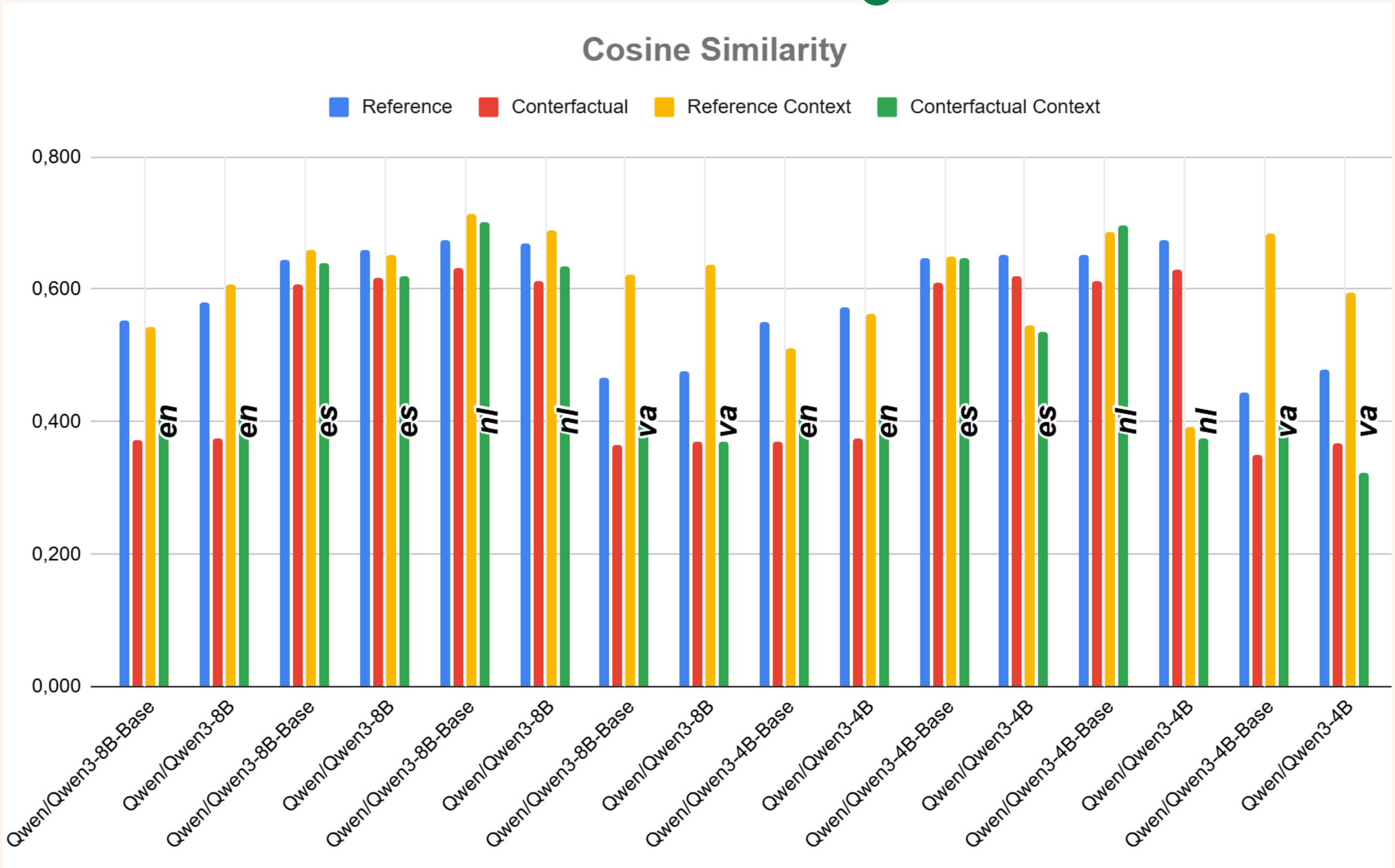
Methods

1. *Automatic metrics:*
    a. *BertScore*
    b. *NLI based evaluation*
    c. *Dependency tree parsing evaluation*
    d. *Cosine similarity with Universal sentence encoders*
2. *LLM-as-a-judge:*
    a. *Prometheus: prometheus-7b-v2.0*
    b. *JudgeLM (ongoing)*

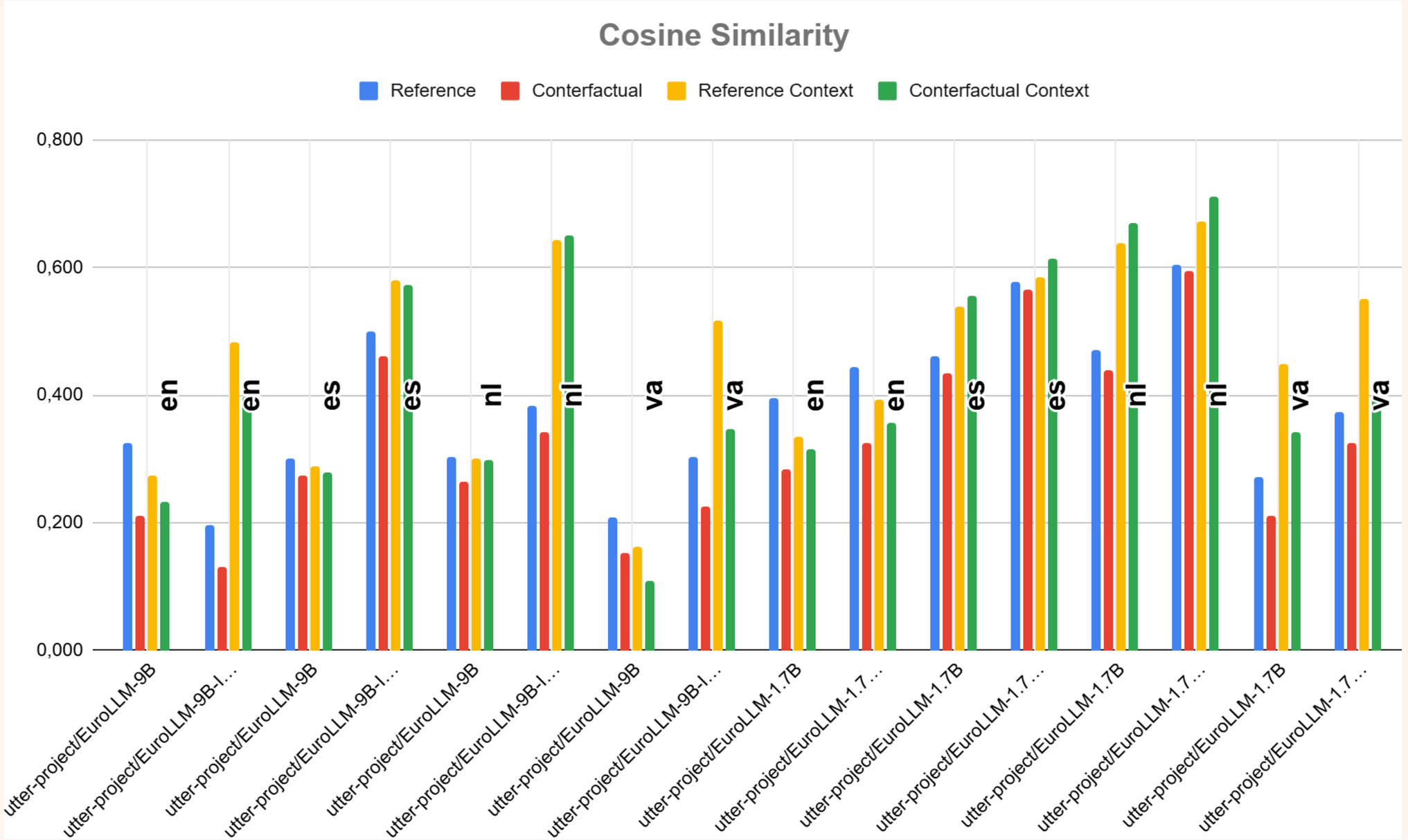
9. Results: LLaMA



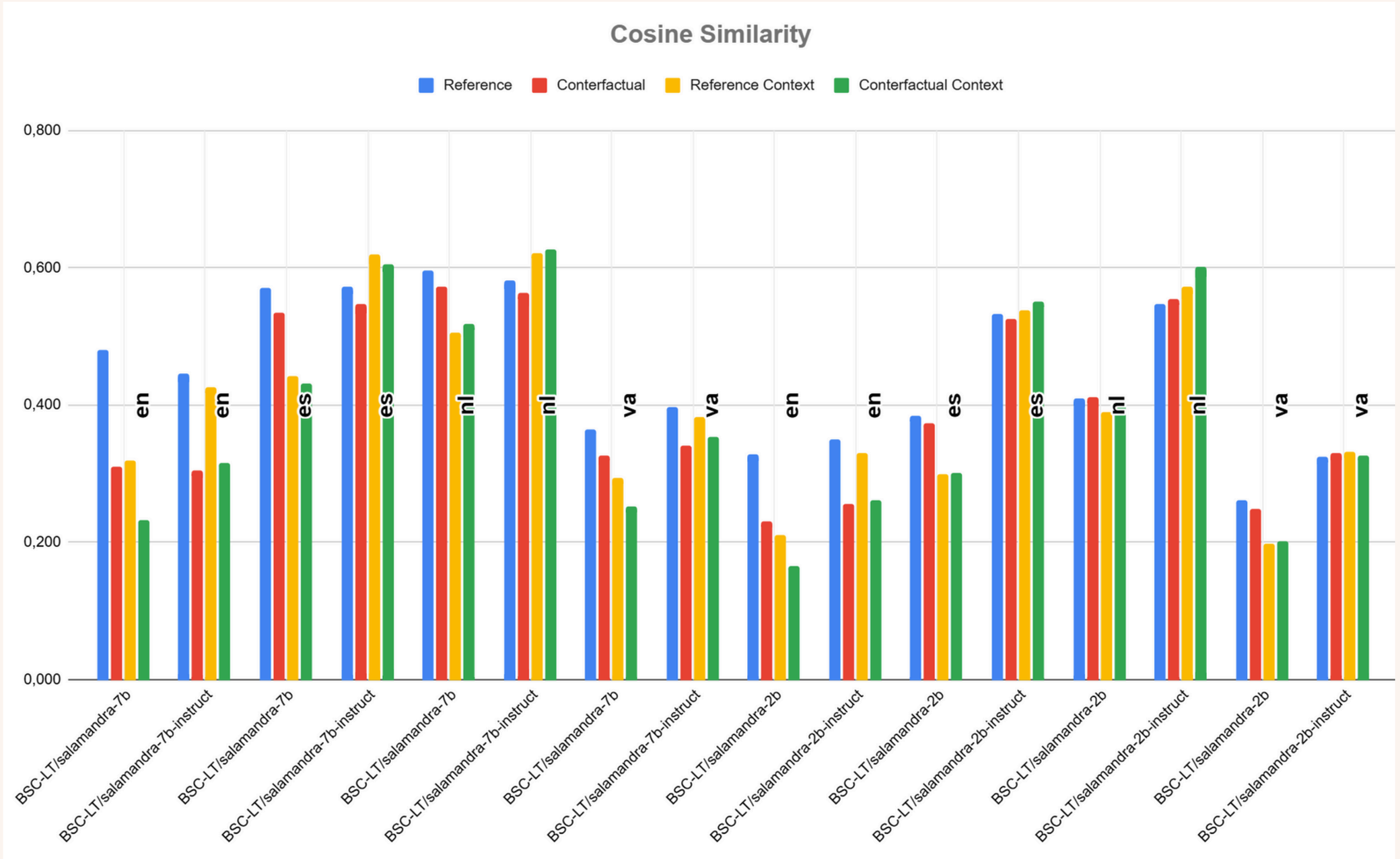
9. Results: QWEN



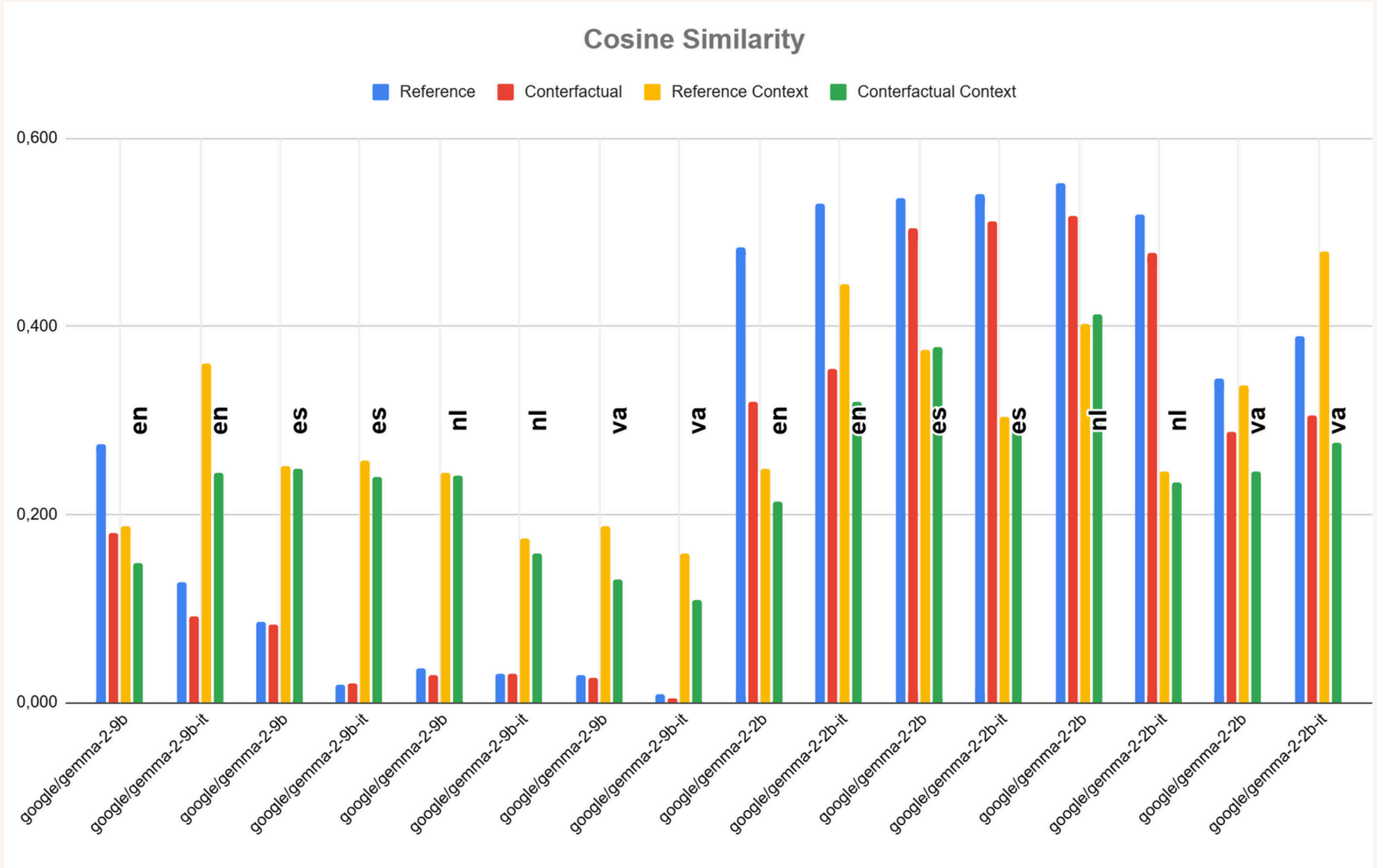
9. Results: EuroLLM



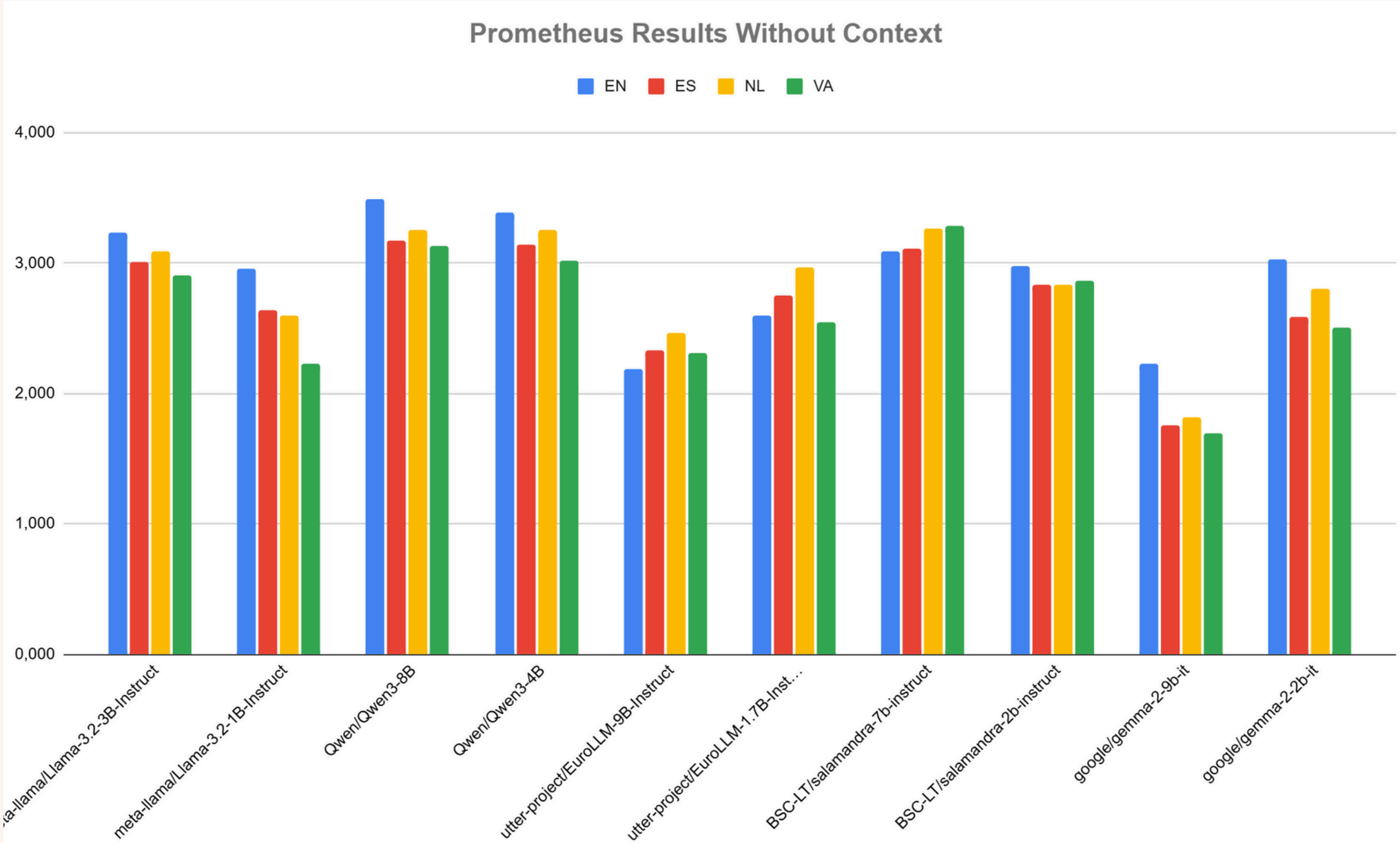
9. Results: Salamandra



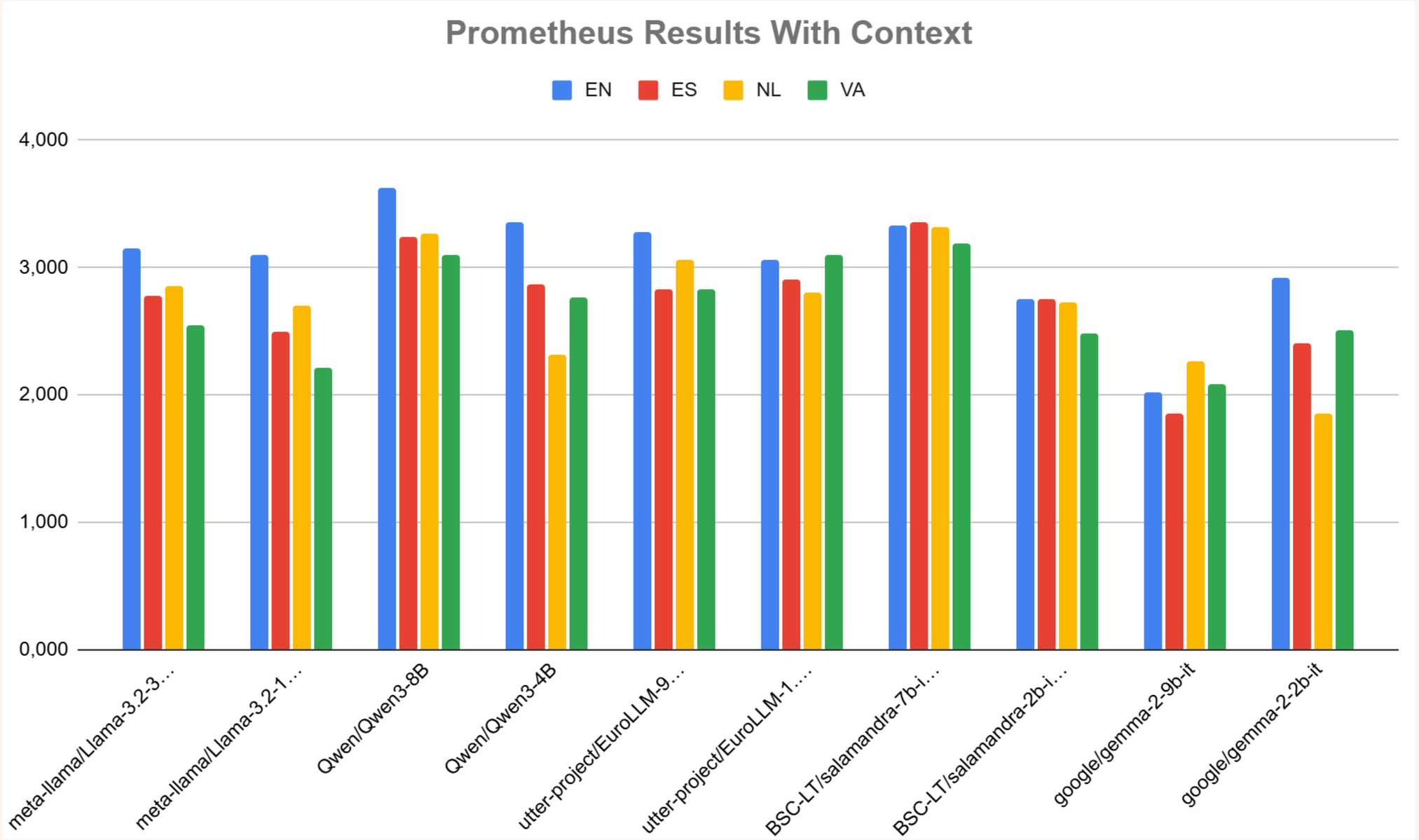
9. Results: Gemma

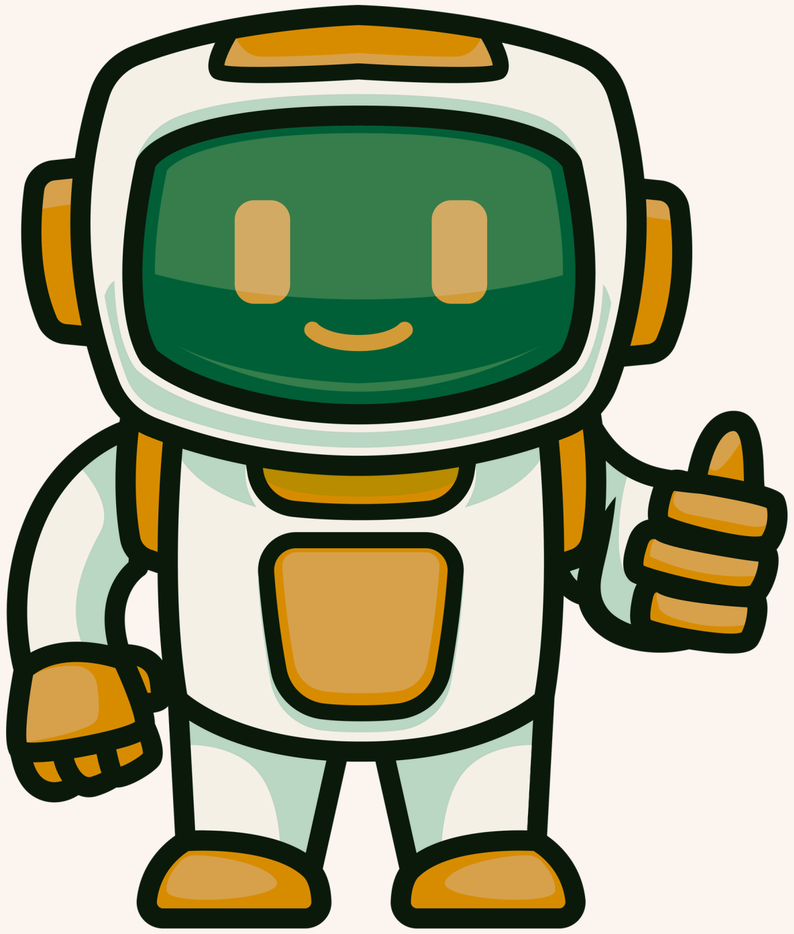


9. Results: Prometheus



9. Results: Prometheus





10. Findings

- *Not every model performs the commonsense generation task equally well across languages.*
- *Context appears to be partially helpful for generating commonsense sentences.*
- *LLMs, when used as judges, can capture multilingual nuances.*

12. Next Steps

1

JUDGELM: FINE-TUNED LARGE LANGUAGE MODELS ARE SCALABLE JUDGES

Lianghui Zhu[1,2] * Xinggang Wang[1†] Xinlong Wang[2†]

[1] School of EIC, Huazhong University of Science & Technology
[2] Beijing Academy of Artificial Intelligence

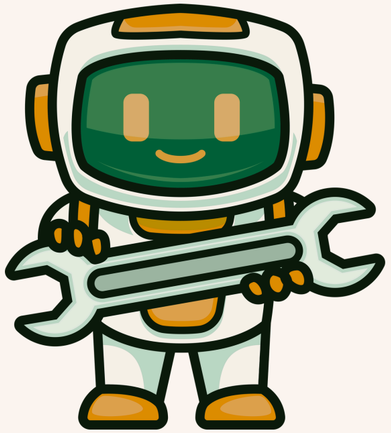
Code & Models: https://github.com/baaivision/JudgeLM

JudgeLM

2

Human Evaluation

3



Train a LLM as a judge

Any questions?

ivan.martinezmurillo@ua.es

University of Alicante, Spain

Universitat d'Alacant
Universidad de Alicante

MINISTERIO DE CIENCIA E INNOVACIÓN
Financiado por la Unión Europea NextGenerationEU
Plan de Recuperación, Transformación y Resiliencia
AGENCIA ESTATAL DE INVESTIGACIÓN

This research work is part of the R&D project CORTEX: Conscious Natural Text Generation (PID2021-123956OB-I00), funded by MCIN/ AEI/10.13039/501100011033/ and by ERDF A way of making Europe