dfkinit2b at CheckThat! 2025: Leveraging LLMs and Ensemble of Methods for Multilingual Claim Normalization

Tatiana Anikina, Ivan Vykopal, Sebastian Kula, Ravi Kiran Chikkala, Natalia Skachkova, Jing Yang, Veronika Solopova, Vera Schmitt, Simon Ostermann

Ivan Vykopal

CLEF CheckThat! 2025, 11.09.2025

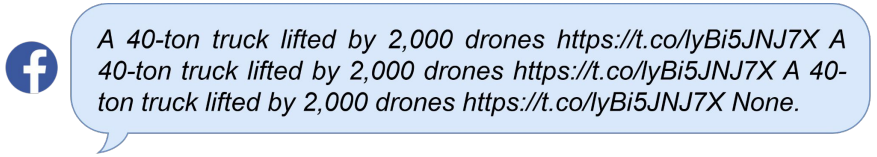
ivan.vykopal@kinit.sk

Kempelen Institute of Intelligent Technologies

Motivation

- Misinformation is spreading rapidly on social media

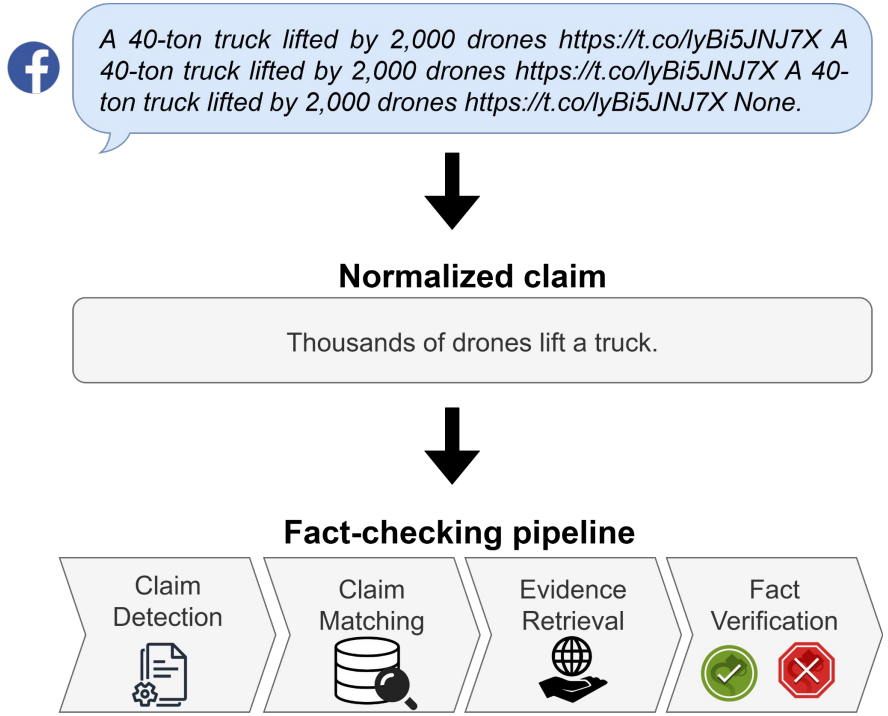
Motivation



- Misinformation is spreading rapidly on social media
- **Posts are noisy, informal and often ambiguous**

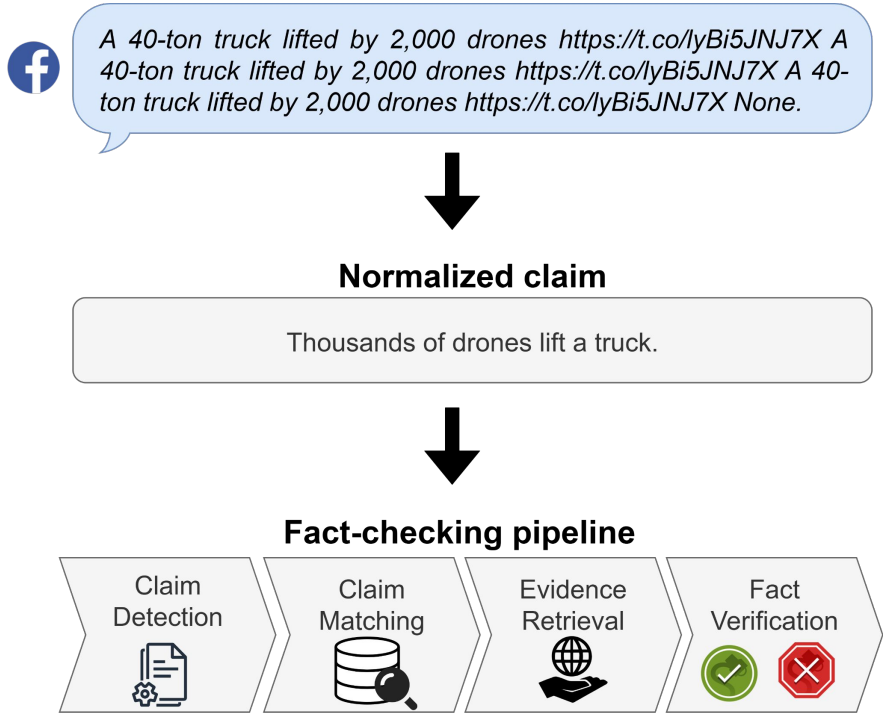
Motivation

- Misinformation is spreading rapidly on social media
- **Posts are noisy, informal and often ambiguous**
- Need for ***Claim Normalization*** to support automated fact-checking



Motivation

- Misinformation is spreading rapidly on social media
- **Posts are noisy, informal and often ambiguous**
- Need for ***Claim Normalization*** to support automated fact-checking
- Multilingual dimension adds complexity

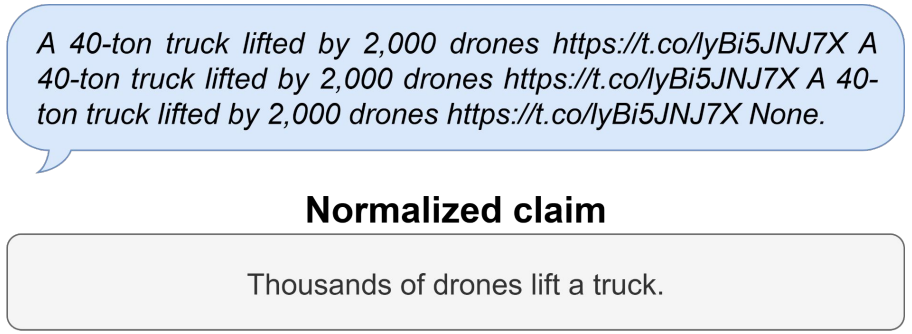


CheckThat! 2025 Task 2 Overview

Task: *“Given a noisy, unstructured social media post, the task is to simplify it into a concise form.”*

CheckThat! 2025 Task 2 Overview

Task: *"Given a noisy, unstructured social media post, the task is to simplify it into a concise form."*



CheckThat! 2025 Task 2 Overview

Task: *"Given a noisy, unstructured social media post, the task is to simplify it into a concise form."*

A 40-ton truck lifted by 2,000 drones https://t.co/lyBi5JNJ7X A 40-ton truck lifted by 2,000 drones https://t.co/lyBi5JNJ7X A 40-ton truck lifted by 2,000 drones https://t.co/lyBi5JNJ7X None.

Normalized claim

Thousands of drones lift a truck.

Two settings:

CheckThat! 2025 Task 2 Overview

Task: *"Given a noisy, unstructured social media post, the task is to simplify it into a concise form."*

A 40-ton truck lifted by 2,000 drones https://t.co/lyBi5JNJ7X A 40-ton truck lifted by 2,000 drones https://t.co/lyBi5JNJ7X A 40-ton truck lifted by 2,000 drones https://t.co/lyBi5JNJ7X None.

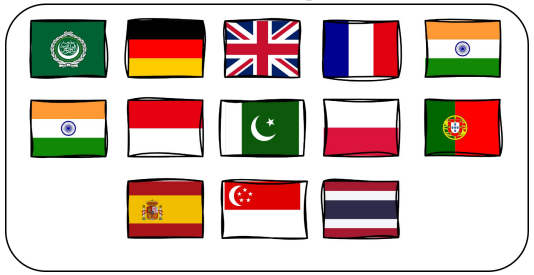
Normalized claim

Thousands of drones lift a truck.

Two settings:

- Monolingual: 13 languages with full datasets (train, dev, test)

Monolingual



CheckThat! 2025 Task 2 Overview

Task: *"Given a noisy, unstructured social media post, the task is to simplify it into a concise form."*

A 40-ton truck lifted by 2,000 drones https://t.co/lyBi5JNJ7X A 40-ton truck lifted by 2,000 drones https://t.co/lyBi5JNJ7X A 40-ton truck lifted by 2,000 drones https://t.co/lyBi5JNJ7X None.

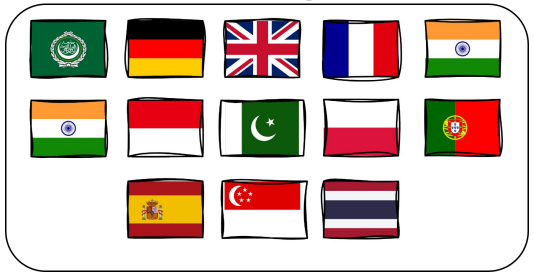
Normalized claim

Thousands of drones lift a truck.

Two settings:

- Monolingual: 13 languages with full datasets (train, dev, test)
- Zero-shot: 7 unseen languages, test set only

Monolingual



Zero-shot

Data & Pre-Processing

- Data are sourced from Google Fact-Check Explorer via ClaimReview

Data & Pre-Processing

- Data are sourced from Google Fact-Check Explorer via ClaimReview

Challenges

Data & Pre-Processing

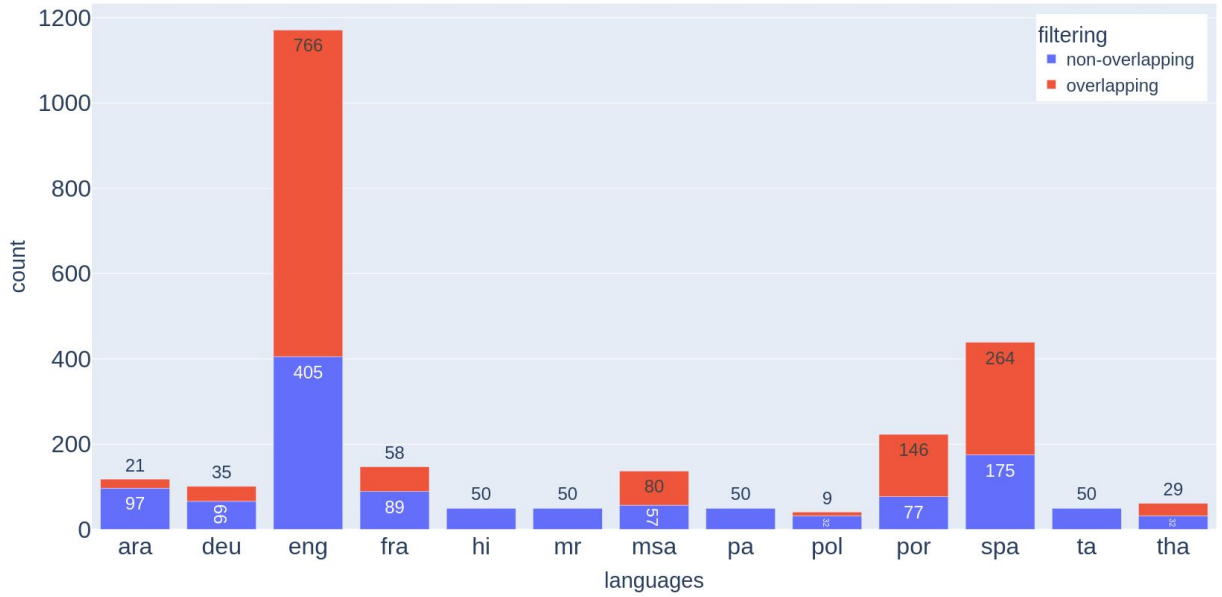
- Data are sourced from Google Fact-Check Explorer via ClaimReview

Challenges

- Mixed languages in post/claim

Data & Pre-Processing

- Data are sourced from Google Fact-Check Explorer via ClaimReview



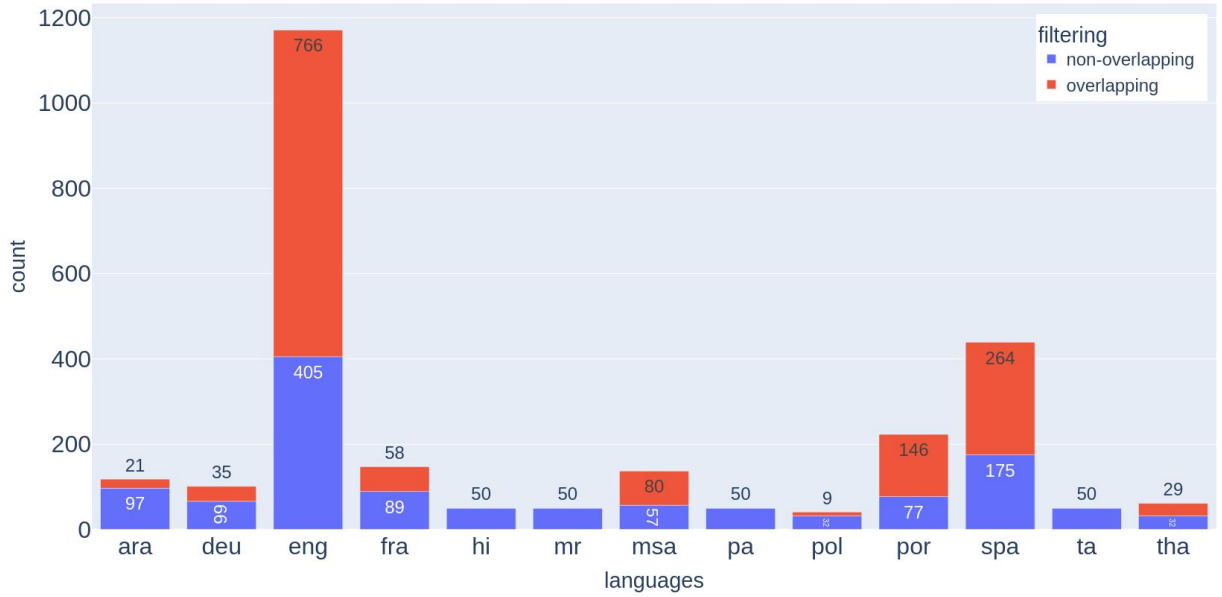
Claim overlap between the gold train and development data.

Challenges

- Mixed languages in post/claim
- Duplicates across train/dev

Data & Pre-Processing

- Data are sourced from Google Fact-Check Explorer via ClaimReview



Claim overlap between the gold train and development data.

Challenges

- Mixed languages in post/claim
- Duplicates across train/dev

Pre-processing: de-duplication & similarity filtering (discarded if cos. sim < 0.05)

Models & Approaches

- 9 LLMs

Model	# Params	# Langs	Citation	LoRA	Prompting
Llama3.1 Instruct	405 B	8	Grattafiori et al. [27]		✓
Llama3.1 Nemotron Ultra	253 B	8	Bercovich et al. [28]		✓
Qwen2.5 Instruct	72 B	29	Yang et al. [29]		✓
Llama3.3 Instruct	70 B	8	Grattafiori et al. [27]		✓
Qwen3	32 B	100+	Yang et al. [30]		✓
Gemma3 IT	27 B	140+	Team et al. [31]	✓	✓
Qwen3	14 B	100+	Yang et al. [30]	✓	
Bielik Instuct v2.3	11 B	1	Ociepa et al. [32]		✓
Qwen3	8 B	100+	Yang et al. [30]		✓

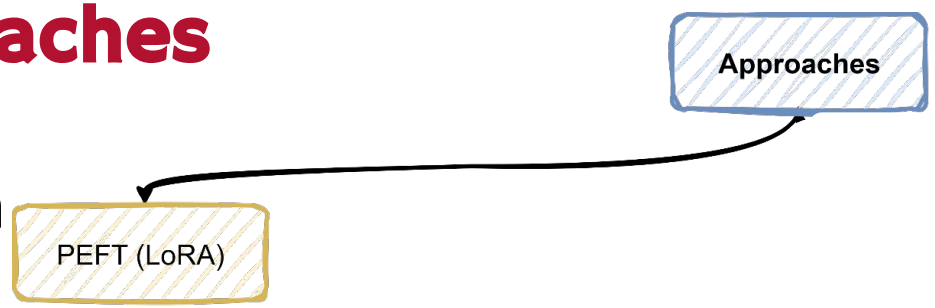
Models & Approaches

- 9 LLMs and 3 main approaches

Model	# Params	# Langs	Citation	LoRA	Prompting
Llama3.1 Instruct	405 B	8	Grattafiori et al. [27]		✓
Llama3.1 Nemotron Ultra	253 B	8	Bercovich et al. [28]		✓
Qwen2.5 Instruct	72 B	29	Yang et al. [29]		✓
Llama3.3 Instruct	70 B	8	Grattafiori et al. [27]		✓
Qwen3	32 B	100+	Yang et al. [30]		✓
Gemma3 IT	27 B	140+	Team et al. [31]	✓	✓
Qwen3	14 B	100+	Yang et al. [30]	✓	
Bielik Instuct v2.3	11 B	1	Ociepa et al. [32]		✓
Qwen3	8 B	100+	Yang et al. [30]		✓

Models & Approaches

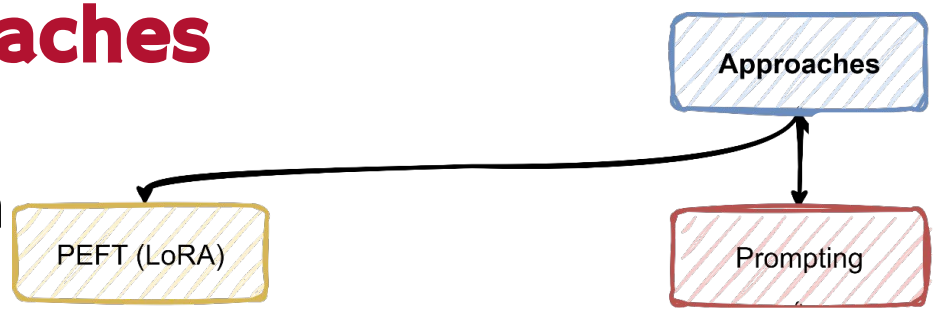
- 9 LLMs and 3 main approaches



Model	# Params	# Langs	Citation	LoRA	Prompting
Llama3.1 Instruct	405 B	8	Grattafiori et al. [27]		✓
Llama3.1 Nemotron Ultra	253 B	8	Bercovich et al. [28]		✓
Qwen2.5 Instruct	72 B	29	Yang et al. [29]		✓
Llama3.3 Instruct	70 B	8	Grattafiori et al. [27]		✓
Qwen3	32 B	100+	Yang et al. [30]		✓
Gemma3 IT	27 B	140+	Team et al. [31]	✓	✓
Qwen3	14 B	100+	Yang et al. [30]	✓	
Bielik Instuct v2.3	11 B	1	Ociepa et al. [32]		✓
Qwen3	8 B	100+	Yang et al. [30]		✓

Models & Approaches

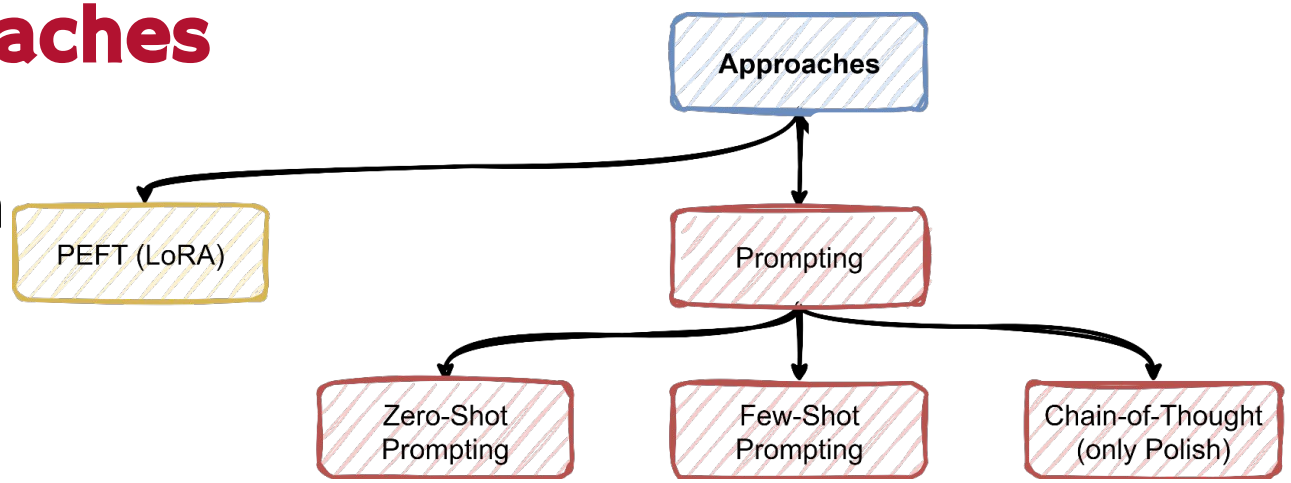
- 9 LLMs and 3 main approaches



Model	# Params	# Langs	Citation	LoRA	Prompting
Llama3.1 Instruct	405 B	8	Grattafiori et al. [27]		✓
Llama3.1 Nemotron Ultra	253 B	8	Bercovich et al. [28]		✓
Qwen2.5 Instruct	72 B	29	Yang et al. [29]		✓
Llama3.3 Instruct	70 B	8	Grattafiori et al. [27]		✓
Qwen3	32 B	100+	Yang et al. [30]		✓
Gemma3 IT	27 B	140+	Team et al. [31]	✓	✓
Qwen3	14 B	100+	Yang et al. [30]	✓	
Bielik Instuct v2.3	11 B	1	Ociepa et al. [32]		✓
Qwen3	8 B	100+	Yang et al. [30]		✓

Models & Approaches

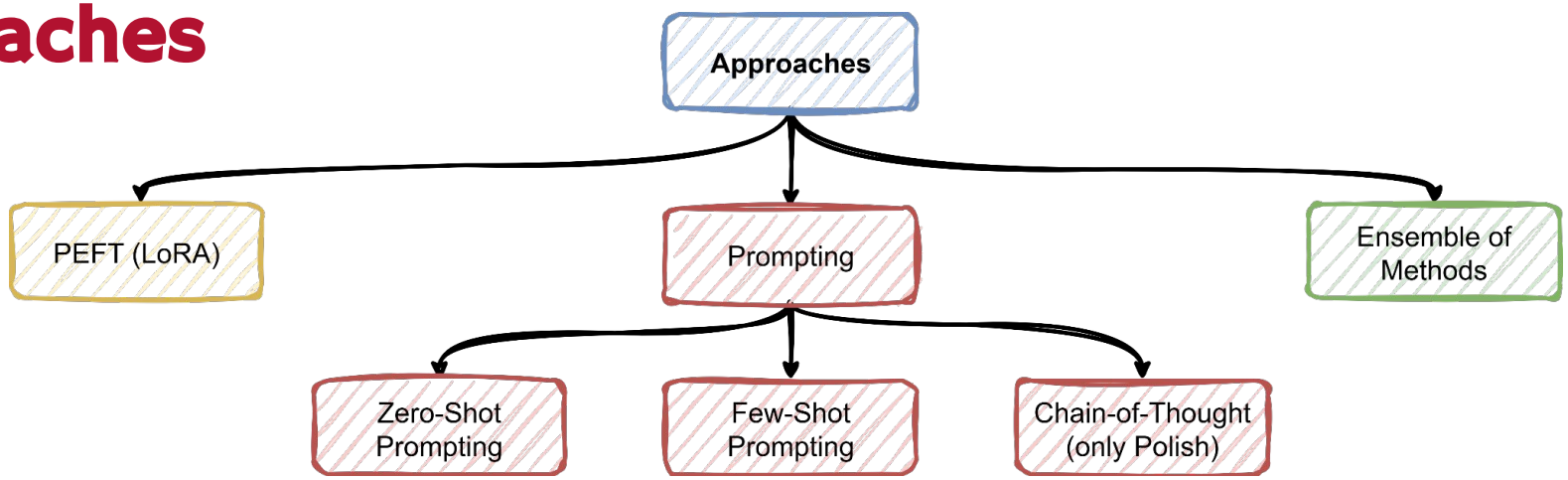
- 9 LLMs and 3 main approaches



Model	# Params	# Langs	Citation	LoRA	Prompting
Llama3.1 Instruct	405 B	8	Grattafiori et al. [27]		✓
Llama3.1 Nemotron Ultra	253 B	8	Bercovich et al. [28]		✓
Qwen2.5 Instruct	72 B	29	Yang et al. [29]		✓
Llama3.3 Instruct	70 B	8	Grattafiori et al. [27]		✓
Qwen3	32 B	100+	Yang et al. [30]		✓
Gemma3 IT	27 B	140+	Team et al. [31]	✓	✓
Qwen3	14 B	100+	Yang et al. [30]	✓	
Bielik Instuct v2.3	11 B	1	Ociepa et al. [32]		✓
Qwen3	8 B	100+	Yang et al. [30]		✓

Models & Approaches

- 9 LLMs and 3 main approaches



Model	# Params	# Langs	Citation	LoRA	Prompting
Llama3.1 Instruct	405 B	8	Grattafiori et al. [27]		✓
Llama3.1 Nemotron Ultra	253 B	8	Bercovich et al. [28]		✓
Qwen2.5 Instruct	72 B	29	Yang et al. [29]		✓
Llama3.3 Instruct	70 B	8	Grattafiori et al. [27]		✓
Qwen3	32 B	100+	Yang et al. [30]		✓
Gemma3 IT	27 B	140+	Team et al. [31]	✓	✓
Qwen3	14 B	100+	Yang et al. [30]	✓	
Bielik Instuct v2.3	11 B	1	Ociepa et al. [32]		✓
Qwen3	8 B	100+	Yang et al. [30]		✓

Fine-Tuning with LoRA

Fine-Tuning with LoRA

Fine-Tuning with LoRA

- Only for **monolingual** setting

Fine-Tuning with LoRA

- Only for **monolingual** setting
- Experimented with Gemma3 27B and **Qwen3 14B**

Fine-Tuning with LoRA

- Only for **monolingual** setting
- Experimented with Gemma3 27B and **Qwen3 14B**
- 3 approaches:

Fine-Tuning with LoRA

- Only for **monolingual** setting
- Experimented with Gemma3 27B and **Qwen3 14B**
- 3 approaches:
  - Single LoRA adapter for each language

Fine-Tuning with LoRA

- Only for **monolingual** setting
- Experimented with Gemma3 27B and **Qwen3 14B**
- 3 approaches:
  - Single LoRA adapter for each language
  - Mixed-language fine-tuning (single LoRA for all languages)

Fine-Tuning with LoRA

- Only for **monolingual** setting
- Experimented with Gemma3 27B and **Qwen3 14B**
- 3 approaches:
  - Single LoRA adapter for each language
  - Mixed-language fine-tuning (single LoRA for all languages)
  - Translation-based augmentation (translation from English data)

Prompting Experiments

Zero-Shot Prompting

- Task description + characteristics of normalized claims

Prompting Experiments

Zero-Shot Prompting

- Task description + characteristics of normalized claims

Few-Shot Prompting

- Demonstration selection via similarity (mGTE-base)
- Experimented with 1, 2, 5 and 10 examples

Prompting Experiments

Zero-Shot Prompting

- Task description + characteristics of normalized claims

Few-Shot Prompting

- Demonstration selection via similarity (mGTE-base)
- Experimented with 1, 2, 5 and 10 examples

Translated Prompts (Zero- & Few-Shot)

- Instructions translated to the target language via Google Translate API

Prompting Experiments

Zero-Shot Prompting

- Task description + characteristics of normalized claims

Few-Shot Prompting

- Demonstration selection via similarity (mGTE-base)
- Experimented with 1, 2, 5 and 10 examples

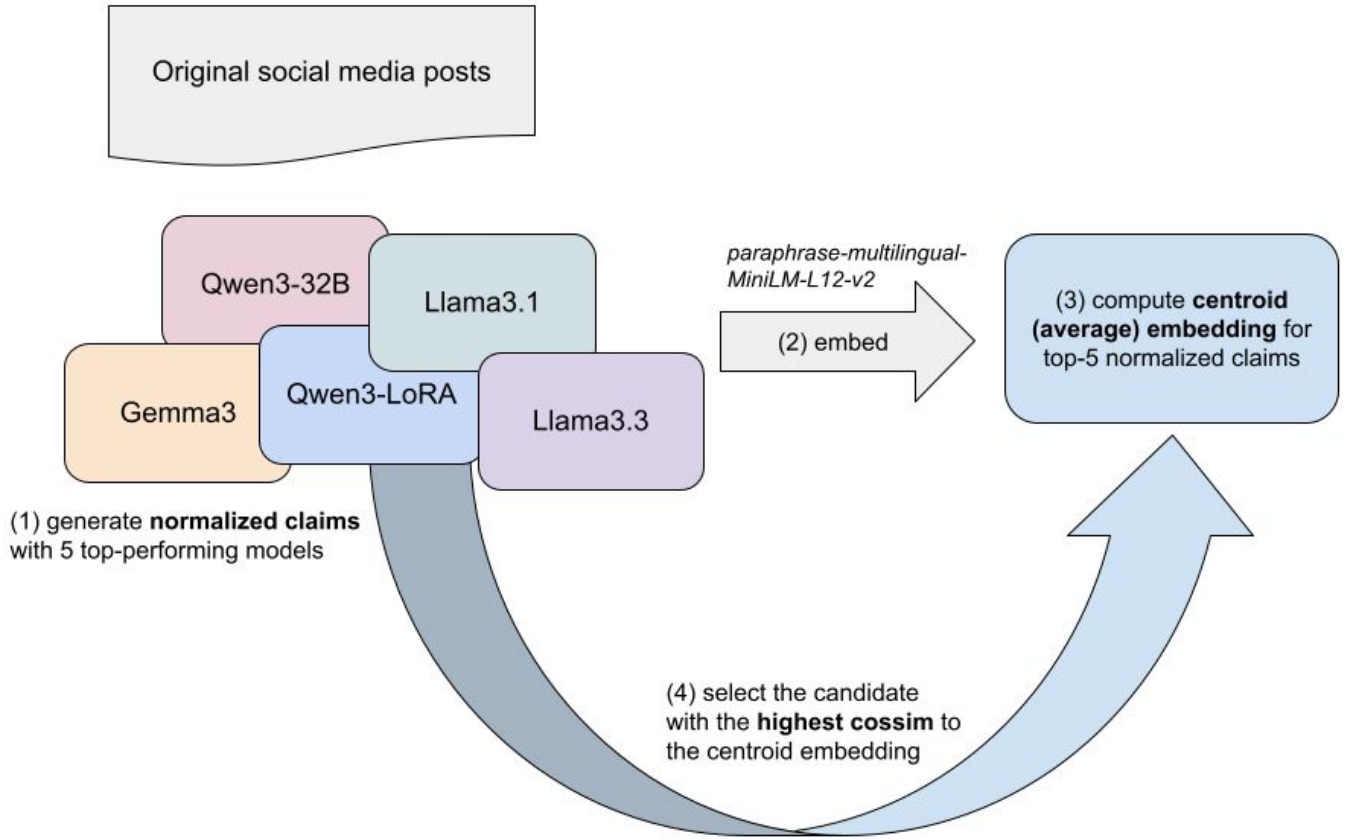
Translated Prompts (Zero- & Few-Shot)

- Instructions translated to the target language via Google Translate API

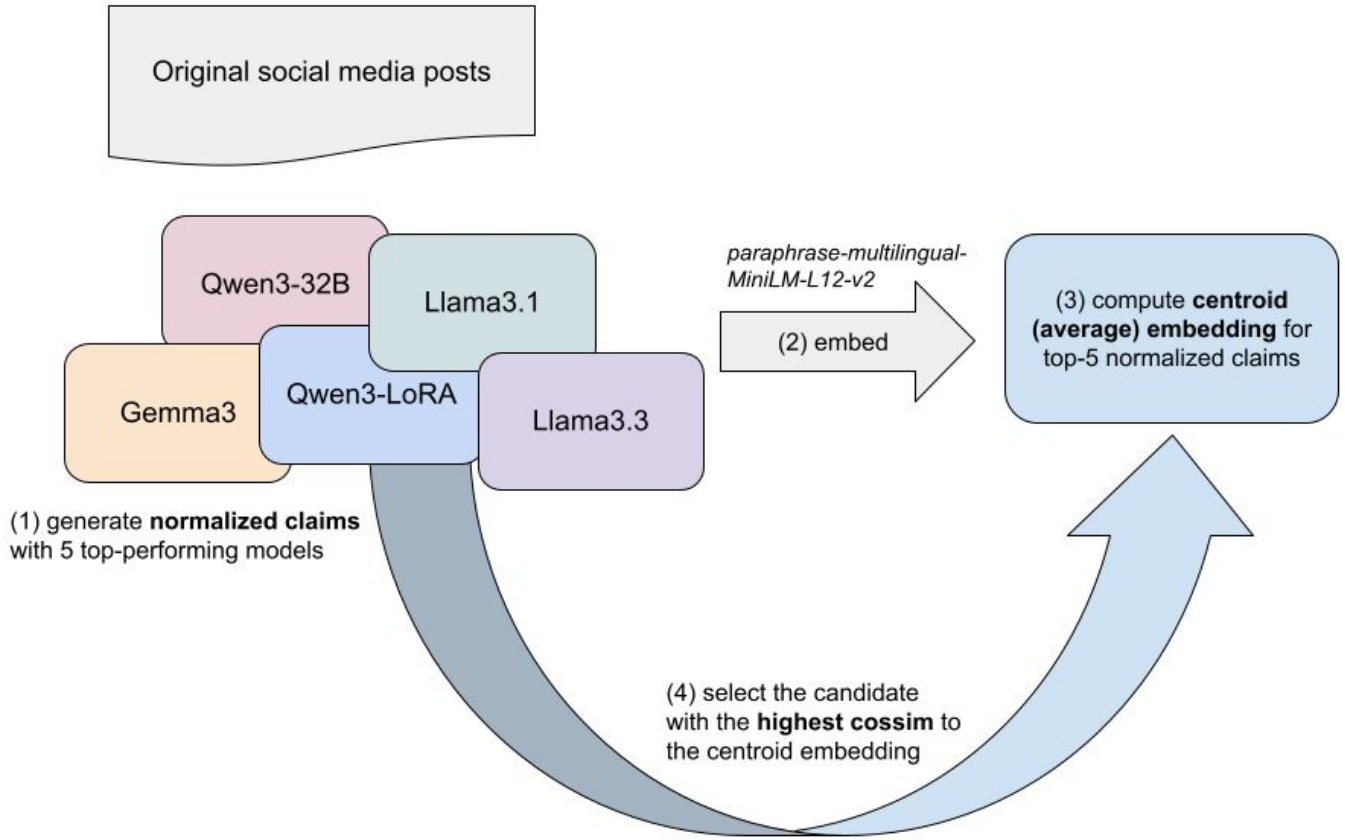
Experiments with Polish

- *Polish-CoT* vs. *Few-shot prompting*
- Experiments also with the Polish LLM

Ensemble of Methods



Ensemble of Methods



Monolingual Setting

LoRA tuned per target language achieves optimal results

LoRA tuned per target language achieves optimal results

- Preliminary results with Gemma3 27B, final submission with Qwen3 14B

LoRA tuned per target language achieves optimal results

- Preliminary results with Gemma3 27B, final submission with Qwen3 14B

Fine-tuning results for `Gemma3 27B` evaluated on the official development set. The first row shows the zero-shot prompting results for comparison.

Approach	ara	deu	eng	fra	hi	mr	msa	pa	pol	por	spa	ta	tha
Zero-shot	0.305	0.161	0.244	0.265	0.224	0.275	0.219	0.311	0.194	0.294	0.268	0.340	0.054
LoRA-target	0.361	0.298	**0.658**	0.439	**0.290**	**0.311**	0.599	**0.352**	0.267	0.509	0.518	0.450	**0.217**
LoRA-all-balanced	**0.390**	0.293	N/A	**0.454**	0.285	0.287	0.570	0.309	0.265	**0.510**	0.531	0.438	0.213
LoRA-translated	0.379	**0.302**	N/A	0.430	0.254	0.297	0.551	0.290	0.236	0.497	0.509	0.332	0.175
LoRA-translated-v2	0.369	0.280	N/A	0.420	0.283	0.286	**0.623**	0.313	**0.315**	0.504	**0.535**	**0.457**	0.198

LoRA tuned per target language achieves optimal results

- Preliminary results with Gemma3 27B, final submission with Qwen3 14B

Fine-tuning results for `Gemma3 27B` evaluated on the official development set. The first row shows the zero-shot prompting results for comparison.

Approach	ara	deu	eng	fra	hi	mr	msa	pa	pol	por	spa	ta	tha
Zero-shot	0.305	0.161	0.244	0.265	0.224	0.275	0.219	0.311	0.194	0.294	0.268	0.340	0.054
LoRA-target	0.361	0.298	**0.658**	0.439	**0.290**	**0.311**	0.599	**0.352**	0.267	0.509	0.518	0.450	**0.217**
LoRA-all-balanced	**0.390**	0.293	N/A	**0.454**	0.285	0.287	0.570	0.309	0.265	**0.510**	0.531	0.438	0.213
LoRA-translated	0.379	**0.302**	N/A	0.430	0.254	0.297	0.551	0.290	0.236	0.497	0.509	0.332	0.175
LoRA-translated-v2	0.369	0.280	N/A	0.420	0.283	0.286	**0.623**	0.313	**0.315**	0.504	**0.535**	**0.457**	0.198

LoRA tuned per target language achieves optimal results

- Preliminary results with Gemma3 27B, final submission with Qwen3 14B

Fine-tuning results for `Gemma3 27B` evaluated on the official development set. The first row shows the zero-shot prompting results for comparison.

Approach	ara	deu	eng	fra	hi	mr	msa	pa	pol	por	spa	ta	tha
Zero-shot	0.305	0.161	0.244	0.265	0.224	0.275	0.219	0.311	0.194	0.294	0.268	0.340	0.054
LoRA-target	0.361	0.298	**0.658**	0.439	**0.290**	**0.311**	0.599	**0.352**	0.267	0.509	0.518	0.450	**0.217**
LoRA-all-balanced	**0.390**	0.293	N/A	**0.454**	0.285	0.287	0.570	0.309	0.265	**0.510**	0.531	0.438	0.213
LoRA-translated	0.379	**0.302**	N/A	0.430	0.254	0.297	0.551	0.290	0.236	0.497	0.509	0.332	0.175
LoRA-translated-v2	0.369	0.280	N/A	0.420	0.283	0.286	**0.623**	0.313	**0.315**	0.504	**0.535**	**0.457**	0.198

LoRA tuned per target language achieves optimal results

- Preliminary results with Gemma3 27B, final submission with Qwen3 14B

Fine-tuning results for `Gemma3 27B` evaluated on the official development set. The first row shows the zero-shot prompting results for comparison.

Approach	ara	deu	eng	fra	hi	mr	msa	pa	pol	por	spa	ta	tha
Zero-shot	0.305	0.161	0.244	0.265	0.224	0.275	0.219	0.311	0.194	0.294	0.268	0.340	0.054
LoRA-target	0.361	0.298	**0.658**	0.439	**0.290**	**0.311**	0.599	**0.352**	0.267	0.509	0.518	0.450	**0.217**
LoRA-all-balanced	**0.390**	0.293	N/A	**0.454**	0.285	0.287	0.570	0.309	0.265	**0.510**	0.531	0.438	0.213
LoRA-translated	0.379	**0.302**	N/A	0.430	0.254	0.297	0.551	0.290	0.236	0.497	0.509	0.332	0.175
LoRA-translated-v2	0.369	0.280	N/A	0.420	0.283	0.286	**0.623**	0.313	**0.315**	0.504	**0.535**	**0.457**	0.198

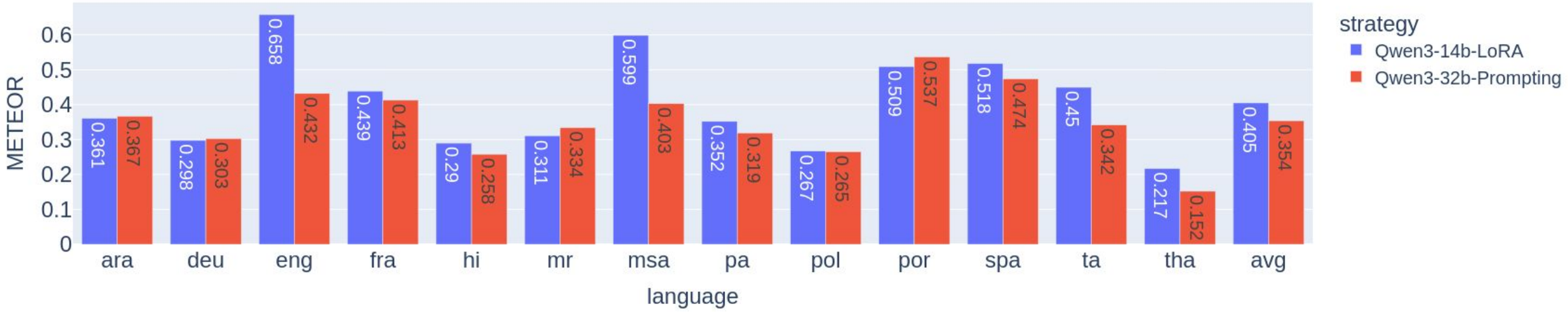
LoRA tuned per target language achieves optimal results

- Preliminary results with Gemma3 27B, final submission with Qwen3 14B

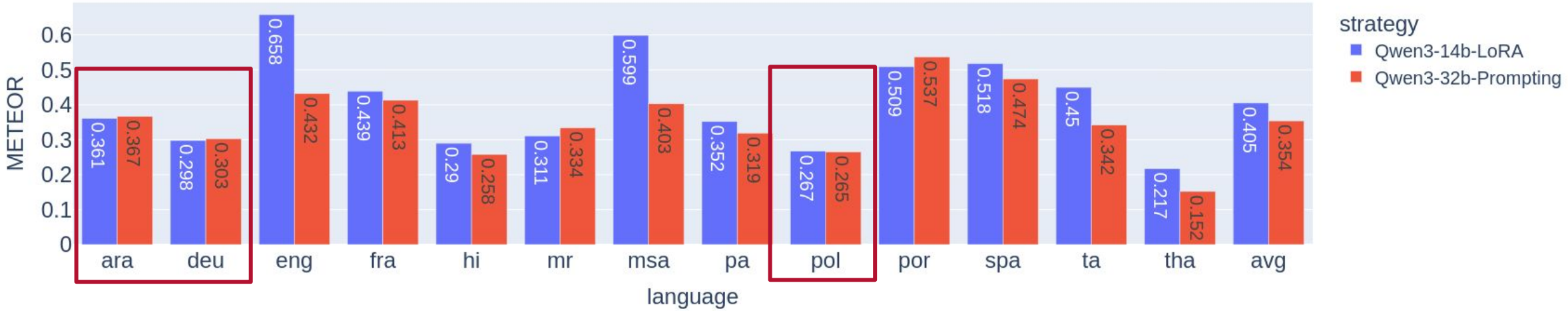
Fine-tuning results for `Gemma3 27B` evaluated on the official development set. The first row shows the zero-shot prompting results for comparison.

Approach	ara	deu	eng	fra	hi	mr	msa	pa	pol	por	spa	ta	tha
Zero-shot	0.305	0.161	0.244	0.265	0.224	0.275	0.219	0.311	0.194	0.294	0.268	0.340	0.054
LoRA-target	0.361	0.298	**0.658**	0.439	**0.290**	**0.311**	0.599	**0.352**	0.267	0.509	0.518	0.450	**0.217**
LoRA-all-balanced	**0.390**	0.293	N/A	**0.454**	0.285	0.287	0.570	0.309	0.265	**0.510**	0.531	0.438	0.213
LoRA-translated	0.379	**0.302**	N/A	0.430	0.254	0.297	0.551	0.290	0.236	0.497	0.509	0.332	0.175
LoRA-translated-v2	0.369	0.280	N/A	0.420	0.283	0.286	**0.623**	0.313	**0.315**	0.504	**0.535**	**0.457**	0.198

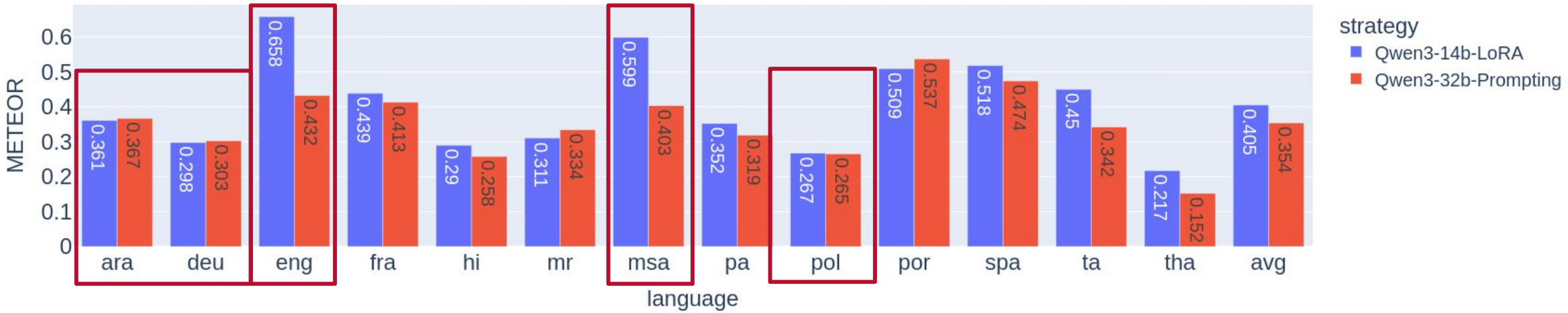
LoRA offers negligible gains for some languages



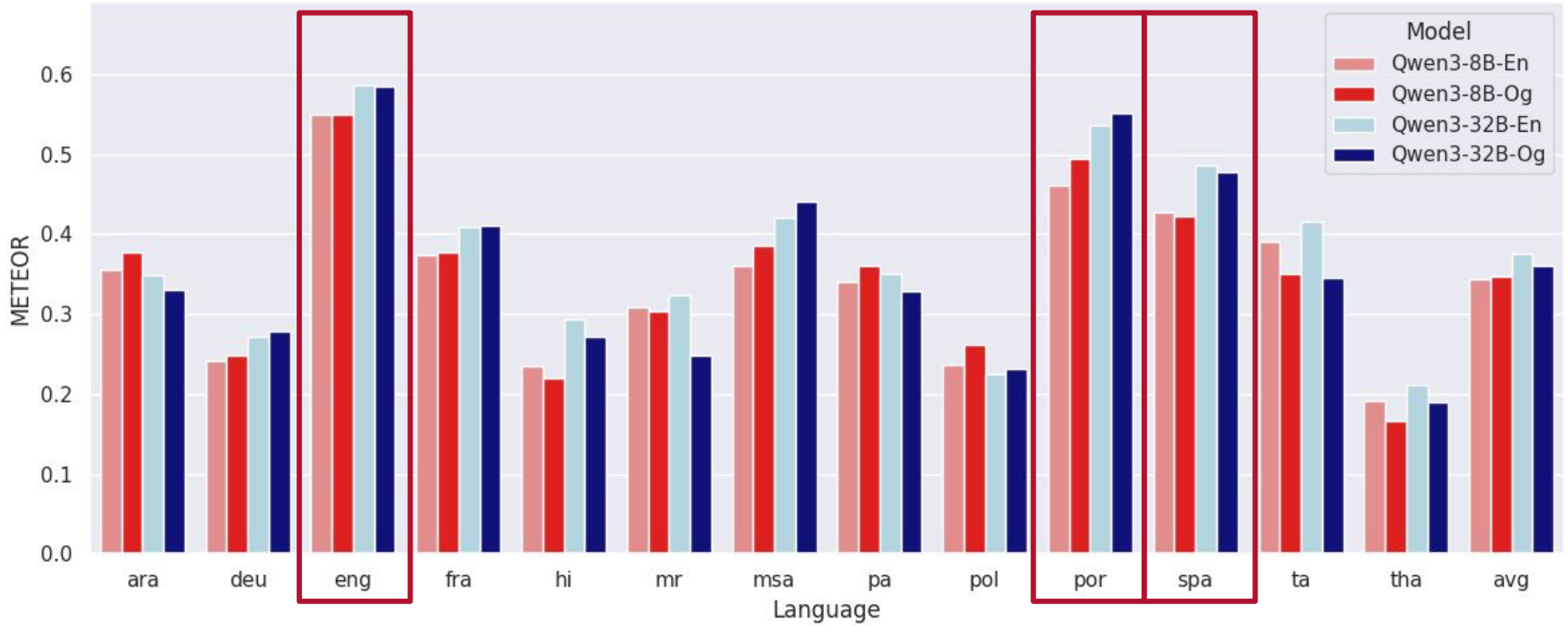
LoRA offers negligible gains for some languages



LoRA offers negligible gains for some languages but larger improvements for others



High-resource European languages demonstrated consistently strong performance



Performance of Qwen3 in the monolingual setting on the dev set using 10-shot prompting with unfiltered data for sample selection.

Multilingual models outperform the Polish LLM

Model	Prompt Type	Dev Set	Test Set
Bielik Instuct v2.3	Polish-CoT	0.198	N/A
	3-shot	0.282	0.297
Llama3.1 Nemotron Ultra	3-shot	0.254	N/A
	10-shot	**0.296**	0.347
Llama3.1 405B	10-shot	0.271	0.393
	20-shot	N/A	**0.396**

Overall Results

We ranked first on all 7 zero-shot languages and on 6 out of 13 languages in the monolingual setting

We ranked first on all 7 zero-shot languages and on 6 out of 13 languages in the monolingual setting

Language	Arabic (ara)	German (deu)	English (eng)	French (fra)	Hindi (hi)	Marathi (mr)	Indonesian (msa)	Punjabi (pa)	Polish (pol)	Portuguese (por)
Best Score	0.504	0.386	0.457	0.527	0.328	0.389	0.565	0.331	0.407	0.577
Our Score	0.504	0.347	0.457	0.470	0.328	0.389	0.502	0.331	0.396	0.574
Δ (Ours vs Best)	0	-0.039	0	-0.057	0	0	-0.063	0	-0.011	-0.003
Our Strategy	Ensemble	Qwen3-32b	Ensemble	Qwen3$_{LoRA}$	Ensemble	Qwen3$_{LoRA}$	Qwen3$_{LoRA}$	Qwen3-8b	Llama3.1	Ensemble
Our Rank	1	2	1	2	1	1	2	1	2	2
Language	**Spanish (spa)**	**Tamil (ta)**	**Thai (tha)**	**Bengali$_{zero}$ (ben)**	**Czech$_{zero}$ (ces)**	**Greek$_{zero}$ (ell)**	**Korean$_{zero}$ (kor)**	**Dutch$_{zero}$ (ndl)**	**Romanian$_{zero}$ (ron)**	**Telugu$_{zero}$ (te)**
Best Score	0.608	0.632	0.586	0.378	0.252	0.262	0.134	0.200	0.295	0.526
Our Score	0.554	0.632	0.300	0.378	0.252	0.262	0.134	0.200	0.295	0.526
Δ (Ours vs Best)	-0.054	0	-0.286	0	0	0	0	0	0	0
Our Strategy	Ensemble	Qwen3$_{LoRA}$	Ensemble	Ensemble	Gemma3	Ensemble	Gemma3	Ensemble	Ensemble	Ensemble
Our Rank	2	1	3	1	1	1	1	1	1	1

We ranked first on all 7 zero-shot languages and on 6 out of 13 languages in the monolingual setting

Language	Arabic (ara)	German (deu)	English (eng)	French (fra)	Hindi (hi)	Marathi (mr)	Indonesian (msa)	Punjabi (pa)	Polish (pol)	Portuguese (por)
Best Score	0.504	0.386	0.457	0.527	0.328	0.389	0.565	0.331	0.407	0.577
Our Score	0.504	0.347	0.457	0.470	0.328	0.389	0.502	0.331	0.396	0.574
Δ (Ours vs Best)	0	-0.039	0	-0.057	0	0	-0.063	0	-0.011	-0.003
Our Strategy	Ensemble	Qwen3-32b	Ensemble	Qwen3$_{LoRA}$	Ensemble	Qwen3$_{LoRA}$	Qwen3$_{LoRA}$	Qwen3-8b	Llama3.1	Ensemble
Our Rank	1	2	1	2	1	1	2	1	2	2

Language	Spanish (spa)	Tamil (ta)	Thai (tha)	Bengali$_{zero}$ (ben)	Czech$_{zero}$ (ces)	Greek$_{zero}$ (ell)	Korean$_{zero}$ (kor)	Dutch$_{zero}$ (ndl)	Romanian$_{zero}$ (ron)	Telugu$_{zero}$ (te)
Best Score	0.608	0.632	0.586	0.378	0.252	0.262	0.134	0.200	0.295	0.526
Our Score	0.554	0.632	0.300	0.378	0.252	0.262	0.134	0.200	0.295	0.526
Δ (Ours vs Best)	-0.054	0	-0.286	0	0	0	0	0	0	0
Our Strategy	Ensemble	Qwen3$_{LoRA}$	Ensemble	Ensemble	Gemma3	Ensemble	Gemma3	Ensemble	Ensemble	Ensemble
Our Rank	2	1	3	1	1	1	1	1	1	1

We ranked first on all 7 zero-shot languages and on 6 out of 13 languages in the monolingual setting

Language	Arabic (ara)	German (deu)	English (eng)	French (fra)	Hindi (hi)	Marathi (mr)	Indonesian (msa)	Punjabi (pa)	Polish (pol)	Portuguese (por)
Best Score	0.504	0.386	0.457	0.527	0.328	0.389	0.565	0.331	0.407	0.577
Our Score	0.504	0.347	0.457	0.470	0.328	0.389	0.502	0.331	0.396	0.574
Δ (Ours vs Best)	0	-0.039	0	-0.057	0	0	-0.063	0	-0.011	-0.003
Our Strategy	Ensemble	Qwen3-32b	Ensemble	Qwen3_{LoRA}	Ensemble	Qwen3_{LoRA}	Qwen3_{LoRA}	Qwen3-8b	Llama3.1	Ensemble
Our Rank	1	2	1	2	1	1	2	1	2	2

Language	Spanish (spa)	Tamil (ta)	Thai (tha)	Bengali_{zero} (ben)	Czech_{zero} (ces)	Greek_{zero} (ell)	Korean_{zero} (kor)	Dutch_{zero} (ndl)	Romanian_{zero} (ron)	Telugu_{zero} (te)
Best Score	0.608	0.632	0.586	0.378	0.252	0.262	0.134	0.200	0.295	0.526
Our Score	0.554	0.632	0.300	0.378	0.252	0.262	0.134	0.200	0.295	0.526
Δ (Ours vs Best)	-0.054	0	-0.286	0	0	0	0	0	0	0
Our Strategy	Ensemble	Qwen3_{LoRA}	Ensemble	Ensemble	Gemma3	Ensemble	Gemma3	Ensemble	Ensemble	Ensemble
Our Rank	2	1	3	1	1	1	1	1	1	1

Best strategies – Gemma3 prompting & Ensemble method

Key Findings & Conclusion

- LoRA adapters are efficient for low-resource languages

Key Findings & Conclusion

- LoRA adapters are efficient for low-resource languages
- Prompting works well with high-resource languages

Key Findings & Conclusion

- LoRA adapters are efficient for low-resource languages
- Prompting works well with high-resource languages
- English prompts outperform target language in zero-shot

Key Findings & Conclusion

- LoRA adapters are efficient for low-resource languages
- Prompting works well with high-resource languages
- English prompts outperform target language in zero-shot
- Ensemble smooths out model variance and achieves the best performance in 11 out of 20 languages *(5 out 7 cases in the zero-shot setting)*

Current Challenges

- Dataset imbalance and overlap between splits

Current Challenges

- Dataset imbalance and overlap between splits
- Missing context (especially visual, e.g., images or videos)

Future Directions

- Use LLMs for data augmentation

Future Directions

- Use LLMs for data augmentation
- Evaluate the impact of normalized claims in fact-checking tasks

Thank you!

Funded by
the European Union

Federal Ministry
of Research, Technology
and Space