TECHNISCHE UNIVERSITÄT DARMSTADT

isep
École d'ingénieurs du numérique

CULTURE, LANGUAGE AND COMPLEXITY IN AI REASONING

Does the language you ask in change how well an AI answers?

Saïd Chaoui · Orlando Kodjo · Esther Mangani
Weimar Triangle 2026 · Pilot experiment

TRIANGLE DE WEIMAR
TRÓJKĄT WEIMARSKI
WEIMARER DREIECK

Simple questions: 100 % right in every language.
Harder questions: 85 % Polish, 80 % French, 75 % English, 70 % German.

4	16	5×	197
languages: EN · FR · DE · PL	questions: 8 science, 8 political	fresh chats per question and language	answers logged of 320 planned (ChatGPT)

01 · WHY

One model, many Europes

LLM quality depends on model size, data and compute [1], [2], and many multilingual benchmarks are translated from English [3], [4]. A citizen asking in Polish may not get the answer a citizen asking in English gets.

Cross-language tests covering the Weimar languages are rare: MMLU-ProX, for instance, has no Polish [5].

METHOD

Complexity: 8 official EU-MMLU questions [6] (biology, chemistry), the same text in 4 languages from EU translators. 4 *direct* (one fact), 4 *multistep* (2+ steps), labelled before any run.

Culture: 8 euandi 2024 policy statements [7]. No right answer: does the position move with the language?

02 · RESULT

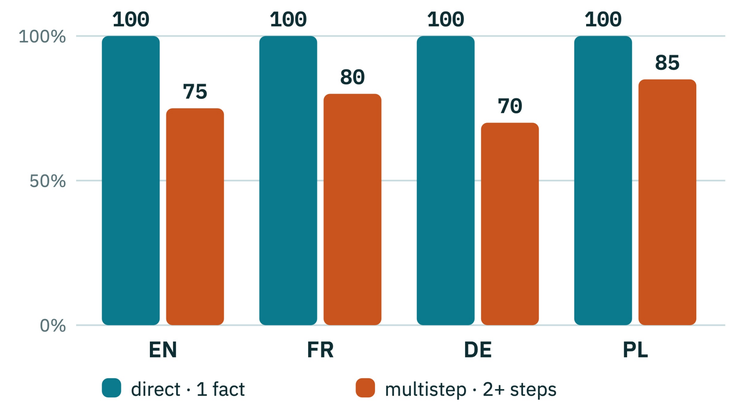
Only one question really breaks the model

		EN	FR	DE	PL
X01	direct	5/5	5/5	5/5	5/5
X02	direct	4/4	5/5	5/5	5/5
X03	direct	5/5	5/5	5/5	5/5
X04	direct	5/5	5/5	5/5	5/5
X05	multistep	5/5	5/5	5/5	5/5
X06	multistep	5/5	5/5	4/5	5/5
X07	multistep	5/5	5/5	5/5	4/5
X08	multistep	0/5	1/5	0/5	3/5

Correct answers out of 5 runs per question and language. 7 of 8 questions score 4/5 or better everywhere; X08 fails in 3 of 4 languages. X02-EN: only 4 runs logged, all correct.

03 · RESULT

The gap appears only when it gets hard



Accuracy per level. Direct: 100 % everywhere, gap 0. Multistep: 70–85 %, a 15-point gap between Polish (best) and German (worst).

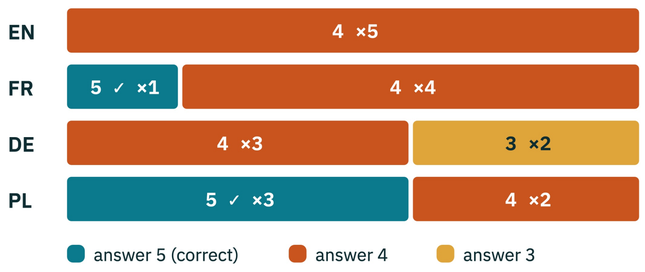
04 · ZOOM

Same question, different answers

The one question the AI failed, **X08** (chemistry): *"What is the maximum number of phases that can be at equilibrium with each other in a three component mixture?"*

F = C − P + 2 Gibbs rule. C = 3 components. Most phases when nothing can vary (F = 0), so **P = 3 + 2 = 5**.

How to read: one row per language, one block per chat (5 chats each). Colour = the answer the AI gave.



The model mostly answers "4". Polish finds 5 three times out of five; English and German never do.

Why it matters: same question, same official EU translation, same AI, and yet the answer changes with the language. One question only: an example, not a proof.

05 · RESULT

Consistent is not correct

Runs giving the most frequent answer	EN	FR	DE	PL
direct	100 %	100 %	100 %	100 %
multistep	100 %	95 %	85 %	85 %

On hard items the model changes its mind between chats, more in German and Polish. English is the most stable, yet stably wrong on X08.

38/38

CULTURE (FR ONLY)

French runs on the 8 political statements all answered **"no opinion"**. The EN, DE and PL statements were not entered in time, so no comparison yet.

06 · TAKEAWAYS

What we learned, in 4 points

100 % — **Simple question? Language doesn't matter.**
When the answer is a single fact, the AI got it right in English, French, German and Polish alike. No gap at all.

15 pts — **Harder question? Language starts to matter.**
With 2 or more reasoning steps, accuracy drops, and more in some languages: 85 % in Polish vs 70 % in German.

0/5 — **English is not a safe default.**
On our hardest question, asking in English or German never gave the right answer; asking in Polish did, 3 times out of 5. The language you ask in can change the answer.

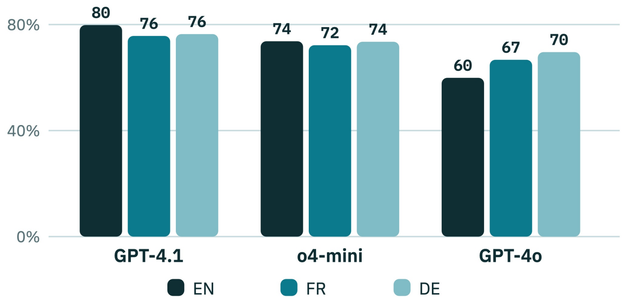
38/38 — **On politics, the AI stays neutral.**
Asked for its position on 8 EU policy statements in French, it always answered "no opinion". We still need EN, DE and PL to compare.

In one sentence: the harder the question, the more the language you ask in can change what the AI answers.

07 · CONTEXT

PUBLISHED NUMBERS · NOT OUR DATA

Big benchmarks: English is not always first

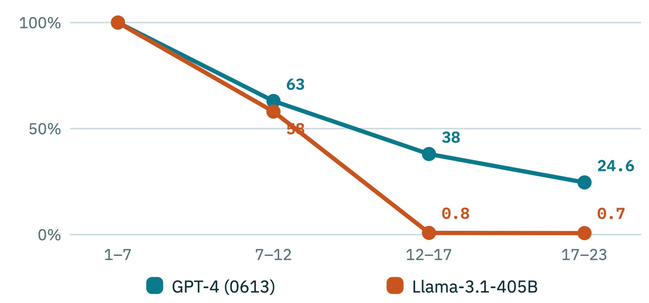


MMLU-ProX accuracy [5], Tables 3 & 6. 5-shot chain-of-thought. Gaps of 1–10 points; GPT-4o scores higher in German than in English. Polish is not tested.

08 · CONTEXT

PUBLISHED NUMBERS · NOT OUR DATA

Longer tasks: accuracy collapses



MLissard object counting [8], Table 3: mean of six languages (no French or Polish), bins = sequence length 1–7, 7–12, 12–17, 17–23. Length and complexity co-vary.

09 · RECOMMENDATIONS

- Test AI in **all EU languages**, especially for complex tasks: health, law, science.
- Measure **consistency** across languages [9] and across repeated chats, not only accuracy.
- Use **official EU translations** such as EU-MMLU to compare languages fairly.
- Add **Polish** and other Central-European languages to benchmarks.

LIMITS

Pilot: one model (exact version not recorded), 16 items, 197 of 320 runs. The multistep gap rests mostly on one item: a signal, not a proof. Culture in French only. Official texts kept as published. No causal claims.

REFERENCES **AI disclosure:** this poster's layout, charts and text were generated with AI (Claude, Anthropic) from the authors' data, then reviewed by the authors. Data: our runs, Weimar Triangle 2026.

[1] J. Kaplan *et al.*, "Scaling laws for neural language models," arXiv:2001.08361, 2020.

[2] J. Hoffmann *et al.*, "Training compute-optimal large language models," in *Proc. NeurIPS*, 2022.

[3] S. Singh *et al.*, "Global MMLU: Understanding and addressing cultural and linguistic biases in multilingual evaluation," in *Proc. ACL*, 2025. [Online]. Available: aclanthology.org/2025.acl-long.919

[4] S. Rajaee *et al.*, "An empirical analysis of machine translation for expanding multilingual benchmarks," in *Proc. WMT*, 2025. [Online]. Available: aclanthology.org/2025.wmt-1.1

[5] W. Xuan *et al.*, "MMLU-ProX: A multilingual benchmark for advanced large language model evaluation," in *Proc. EMNLP*, 2025. [Online]. Available: aclanthology.org/2025.emnlp-main.79

[6] European Commission, DG Translation, "EU-MMLU," dataset, 2026. [Online]. Available: huggingface.co/datasets/EC-DGT-AI/EU-MMLU

[7] European University Institute, "euandi 2024," voting advice application, 2024. [Online]. Available: euandi.eu

[8] M. C. Bueno *et al.*, "MLissard: Multilingual long and simple sequential reasoning benchmarks," in *Proc. GenBench Workshop*, 2024. [Online]. Available: aclanthology.org/2024.genbench-1.6

[9] H. Lai and M. Nissim, "mCoT: Multilingual instruction tuning for reasoning consistency in language models," in *Proc. ACL*, 2024. [Online]. Available: aclanthology.org/2024.acl-long.649