TU Wien Methodology 4 at ELOQUENT 2026: A Factorial Study of Retrieval and Reasoning for Cultural Robustness

Seifeldin Negm[1,*], Zeyad Zohny[1,*], Kora Sulo[1] and Ana Kormaku[1]

[1]TU Wien, Faculty of Informatics, Vienna, Austria

Abstract

We report the TU Wien Methodology 4 submission to the ELOQUENT 2026 evaluation of cultural robustness. The task asks whether a multilingual model gives culturally varied answers to everyday questions and whether that variation responds to an explicit country context. We treat the task as a question about model behavior rather than a score to maximize, and we test two common interventions: retrieval of a single cultural fact, and chain-of-thought reasoning. We run a 2×4 factorial study over Qwen3-32B across six languages, crossing retrieval against four reasoning modes, for eight conditions. The organizers score each condition with Cultural Signal Purity (CSP), which measures how far a model's language defaults vary independently of its response to country anchors. The results contradict the intuition that more scaffolding helps. A single retrieved fact with no added reasoning scores highest (CSP 0.640 against a plain-prompt baseline of 0.492), while the same fact combined with explicit chain-of-thought scores lowest and falls below baseline (0.403). Every isolated intervention improves on the baseline, but stacking them does not. A small qualitative audit suggests a cautious reading: retrieval helps when the inserted fact is on target, while weak or off-topic facts make the RAG variants fragile. We describe the system, the design, and this interpretation.

Keywords

Cultural robustness, Multilingual LLMs, Retrieval-augmented generation, Chain-of-thought, Cultural Signal Purity, ELOQUENT 2026

1. Introduction

ELOQUENT [1] is a set of shared tasks on the quality and robustness of generative language models. One task studies cultural robustness. It asks a model everyday questions, for example what people eat for breakfast or how a school day is structured, and checks two properties. First, do answers vary across languages in a way that reflects local custom rather than a single global default? Second, when the prompt names a specific country, does the answer move toward that country?

These two properties pull in different directions, and the tension is the point of the task. A model can vary its answers by language yet ignore an explicit country anchor, or it can follow the anchor while giving identical language defaults. The task rewards systems whose language-level variation and anchor-level variation are independent of each other.

We approached the task as a chance to learn how standard tooling shapes cultural signal, not as a leaderboard exercise. Two techniques dominate current practice: retrieval-augmented generation [2], which grounds an answer in an external fact, and chain-of-thought prompting [3], which asks the model to reason before it answers. Both are usually presented as improvements. We wanted to know whether they help here, and our guiding question was direct:

> Does grounding reasoning in retrieved cultural facts reduce culturally generic answers?

To answer it we built a factorial pipeline that turns retrieval and reasoning on and off in every combination, then submitted all eight conditions for official scoring. The headline finding is a negative result we did not expect. The lightest intervention wins, and the heaviest combination of retrieval and explicit reasoning is the only condition that scores below the plain baseline.

CLEF 2026 Working Notes, 21 – 24 September 2026, Jena, Germany

*Corresponding author.

EMAIL: seifznegm@gmail.com (S. Negm); zeyadalaaeldin20@gmail.com (Z. Zohny); korasulo26@gmail.com (K. Sulo); kormaku.ana@gmail.com (A. Kormaku)

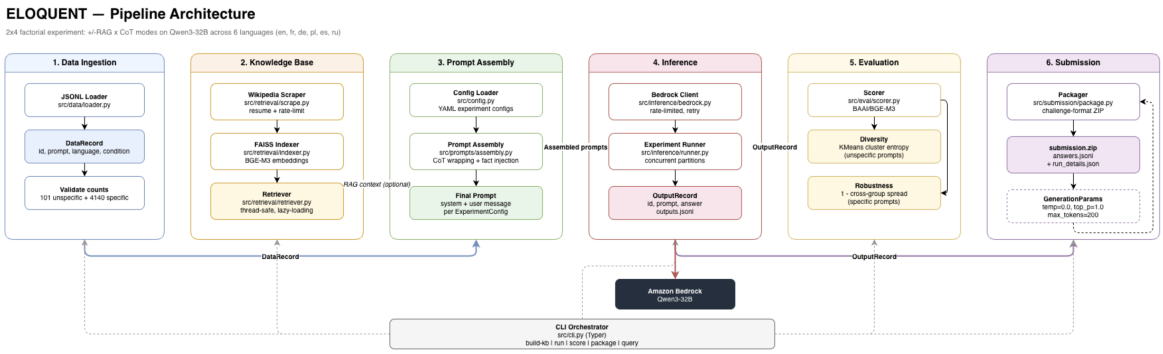

Figure 1: System pipeline. A question is optionally grounded with one retrieved cultural fact, assembled into a prompt under the active reasoning mode, answered by Qwen3-32B on Amazon Bedrock, and packaged for official scoring. Retrieval and reasoning are the only components that change across conditions.

2. Related work

Work on cultural alignment shows that prompting language does not by itself move a model toward a target culture. Durmus et al. [4] find that asking a model to answer in a language does not shift its opinions toward speakers of that language, while naming the country explicitly does. That split maps onto the task's distinction between language defaults and country anchors, and it frames our retrieval and reasoning conditions as attempts to strengthen one side without flattening the other.

Retrieval quality depends on the source. Naous et al. [5] show that English-centric resources, including Wikipedia, carry weak or skewed cultural signal, which limits what a retrieval step can supply. Targeted methods such as CultureLLM [6] instead inject culture through training data. We keep retrieval at inference time so that we can isolate its effect, and we read our retrieval results in light of the source limits these papers describe.

Chain-of-thought reasoning [3] improves performance on arithmetic and symbolic tasks by making intermediate steps explicit. Whether that benefit transfers to open cultural questions is unclear, since a longer reasoning trace can also converge on a single encyclopedic answer. Our design lets us measure that risk directly. We run inference on Qwen3-32B [7], which exposes a native reasoning mode that we treat as a separate factor from explicit chain-of-thought.

3. Task and data

The task provides everyday-life questions in two forms. Unspecific questions carry no location, so the only cultural cue is the language of the prompt. Specific questions name a country, which acts as an explicit anchor. The benchmark pairs the two forms so that the scorer can compare a model's language default against its anchored response for the same underlying question.

We worked in six languages: English, French, German, Polish, Spanish, and Russian. Each language partition contains 101 unspecific prompts and 4140 specific prompts, for 4241 prompts per condition. With eight conditions across six languages, the submission covers 48 partitions. The questions span everyday domains such as food, education, work, and social norms.

4. System description

Figure 1 shows the pipeline. A question enters, an optional retrieval step attaches one cultural fact, a prompt assembler builds the final input under the active reasoning mode, the model generates an answer, and the answer is packaged for scoring. The retrieval and reasoning switches are the only moving parts between conditions; everything else stays fixed.

Table 1
The eight conditions in the 2×4 design.

Condition	Retrieval	Reasoning mode
baseline	off	none
cot_only	off	explicit CoT
reasoning_only	off	native reasoning
cot_reasoning	off	CoT + reasoning
rag_only	on	none
rag_cot	on	explicit CoT
rag_reasoning	on	native reasoning
rag_cot_reasoning	on	CoT + reasoning

4.1. Model and inference

We use Qwen3-32B through Amazon Bedrock with the Converse API in the `eu-central-1` region. Generation is deterministic: temperature 0.0, top-p 1.0, and a cap of 200 answer tokens. When a reasoning mode is active, the model receives a separate reasoning budget of 2000 tokens that does not appear in the final answer. Each prompt runs in an independent session with greedy decoding, so a condition's result depends only on its prompt, not on order or sampling.

4.2. Knowledge base and retrieval

The retrieval index draws on Wikipedia content for 35 European countries across 14 topic patterns. We embed passages with BGE-M3 [8], a multilingual model that produces 1024-dimensional vectors, and index them with FAISS [9]. At inference time a retrieval-enabled condition fetches the single best-matching passage for the question and inserts it under a `Cultural fact:` label. We deliberately retrieve one fact rather than many, so that grounding stays light and its effect is easy to read.

4.3. Prompt assembly

Every condition shares a base instruction that tells the model to answer the question as written, directly, and at the requested length, and an answer rule that asks for the answer alone, in the language of the question. On top of that base, the reasoning modes differ. The explicit chain-of-thought mode adds a five-step silent wrapper that asks the model to classify the question, identify the cultural signal, decide what not to repeat, reason about the locally typical answer, and then write the answer without revealing the reasoning. The native reasoning mode instead uses the model's own thinking budget. Appendix A lists the wrapper in full.

4.4. Experimental design

We cross two factors. The first is retrieval, either off or on. The second is reasoning mode, with four settings: none, explicit chain-of-thought, native reasoning, and both together. The cross gives the 2×4 design in Figure 2 and the eight conditions in Table 1. The design isolates each intervention and every combination, which is what lets us separate a main effect from an interaction.

5. Evaluation methodology

The organizers score each condition with Cultural Signal Purity (CSP). CSP needs no ground-truth answers. It embeds the model's responses, here with Qwen3-8B embeddings, and asks whether a model's language defaults vary independently of its response to country anchors.

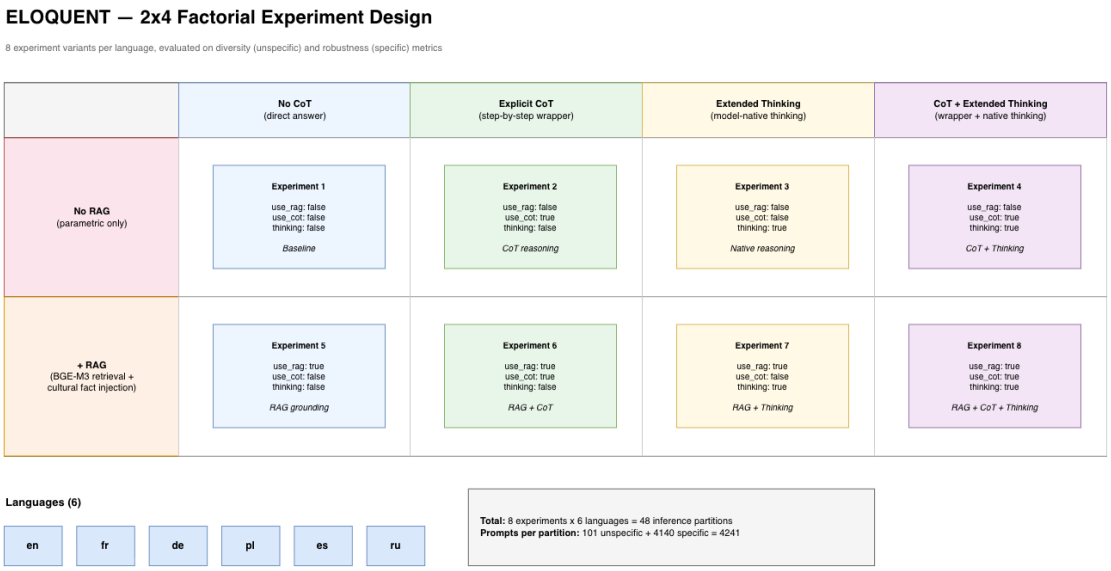

Figure 2: The 2×4 factorial design. Retrieval (off or on) is crossed with four reasoning modes (none, explicit chain-of-thought, native reasoning, and both), giving eight conditions that each run across all six languages.

For a question q and a language pair (ℓ_1, ℓ_2), the unanchored distance compares the two language defaults, that is, the answers given with no country named:

$$U_q(\ell_1, \ell_2) = d_{\cos}\left(\mathbf{r}_q^{\ell_1},\, \mathbf{r}_q^{\ell_2}\right), \tag{1}$$

where $\mathbf{r}$ is the embedding of an answer and $d_{\cos}$ is cosine distance. The anchored distance averages the same comparison over a set C of country contexts:

$$A_q(\ell_1, \ell_2) = \frac{1}{|C|} \sum_{c \in C} d_{\cos}\left(\mathbf{r}_{q,c}^{\ell_1},\, \mathbf{r}_{q,c}^{\ell_2}\right). \tag{2}$$

For each language pair the scorer takes the Spearman rank correlation ρ between the vector of unanchored distances and the vector of anchored distances across questions. The final score subtracts the squared mean correlation from one:

$$\mathrm{CSP} = 1 - \bar{\rho}^2, \qquad \bar{\rho} = \frac{1}{|P|} \sum_{(\ell_1,\ell_2) \in P} \rho\left(\mathbf{U}^{(\ell_1,\ell_2)},\, \mathbf{A}^{(\ell_1,\ell_2)}\right), \tag{3}$$

where P is the set of language pairs.

A high CSP means the two kinds of variation are independent: how a model differs by language tells you little about how it responds to an anchor. A low CSP means anchoring only rescales the pattern the model already shows by language, or that answers are so generic that both signals collapse together. Our submission was scored on six languages, which gives 15 language pairs. Because the organizers' reference baselines run on 23 languages, our absolute values are not comparable to theirs, so we read our conditions against each other and against our own baseline.

Table 2
Official Cultural Signal Purity by condition, sorted high to low. Δ is the change against the baseline condition.

Condition	CSP	Δ
rag_only	0.640	$+0.148$
reasoning_only	0.592	$+0.100$
cot_only	0.560	$+0.068$
rag_reasoning	0.558	$+0.066$
cot_reasoning	0.546	$+0.054$
rag_cot_reasoning	0.525	$+0.033$
baseline	0.492	0.000
rag_cot	0.403	-0.089

6. Results

Table 2 lists the official CSP for each condition, sorted high to low, with the change against our baseline. Three facts stand out. Every single intervention beats the plain baseline. The best condition is the lightest one, a single retrieved fact with no added reasoning, at 0.640. The worst condition is `rag_cot`, which combines retrieval with explicit chain-of-thought and is the only condition below baseline, at 0.403.

During development we also used a local diversity–robustness proxy for sanity checks. We do not use that proxy for conclusions, because the organizer evaluation uses its own embeddings and aggregation. All rankings in Table 2 therefore refer to official CSP.

The interactions matter more than the main effects. Retrieval on its own is the strongest move in the study, but adding reasoning to it coincides with lower CSP. Retrieval plus native reasoning drops to 0.558, and retrieval plus explicit chain-of-thought falls to 0.403. To check the pattern qualitatively, we manually audited 10 matched English cases across `baseline`, `cot_only`, `rag_only`, and `rag_cot`. We judged the single retrieved fact in each RAG prompt for top-1 relevance. Five facts were directly relevant, two were partially relevant, and three were irrelevant or tied to the wrong cultural setting. The same sample showed six locally grounded answers, three generic or no-effect answers, and one case where a partial demographic fact narrowed the answer too strongly. For example, a breakfast fact led the RAG answers to name local foods such as yoghurt, muesli, porridge, and open sandwiches; an unrelated legal fragment for a friend-making prompt produced only generic advice about clubs and events.

7. Discussion

The pattern suggests a mechanism, but does not prove one. CSP rewards independence between language defaults and anchored answers. A short answer built on one clean fact can keep that independence, because the fact acts as a distinct anchor that the model has not already folded into its language-default behavior. Retrieval with no reasoning matches that shape, and it scores highest.

The lower reasoning scores are consistent with a different failure mode: extra reasoning can turn a sparse cue into a safer, more generic answer. Our qualitative audit gives mixed but useful evidence. When retrieval is on target, it often adds local detail. When the fact is weak, off-topic, or tied to the wrong setting, the answer either ignores it or becomes fragile. The source itself may contribute, since the cultural signal in Wikipedia-derived passages is known to be thin [5]. This makes the central result less a rejection of reasoning in general than a warning that reasoning does not automatically repair weak retrieval.

We expected the opposite. The intuition behind the submission was that the full pipeline, retrieval feeding explicit reasoning, would combine a grounded fact with careful thought and win. It came last among the retrieval conditions and below the baseline. The cautious reading is that the interventions did not compose in this setup. Each one nudged the model in a useful direction on its own, but their

combination was more sensitive to weak retrieval than we expected.

8. Limitations

Several limits bound these conclusions. We scored on 6 of the 23 languages in the full benchmark, so our absolute numbers are not comparable to the organizers' baselines and our relative effects may not hold across the wider set. Our retrieval anchors cover European countries only, which narrows the cultural range and may interact with the language set. The knowledge base is Wikipedia-derived, and prior work shows that source carries weak cultural signal [5], so a stronger source could change the retrieval results. We tested a single model, Qwen3-32B, and a single deterministic run per condition, so we cannot separate a stable effect from run-to-run variance. The manual audit covers only 10 English cases, so it is explanatory rather than representative. Finally we report CSP at the condition level and do not break it down by question category, which would show whether the interaction we describe is uniform or concentrated in domains such as food or education.

9. Conclusion

We submitted a 2×4 factorial study of retrieval and reasoning to ELOQUENT 2026 and measured each condition with Cultural Signal Purity. The lightest intervention won. A single retrieved fact with no added reasoning scored highest, well above the plain baseline, while the same fact combined with explicit chain-of-thought was the only condition to score below baseline. Every isolated intervention helped, but stacking retrieval and reasoning did not in this setup. Our manual audit suggests why: light grounding can add concrete local detail when retrieval is relevant, while weak facts make RAG outputs fragile. For cultural robustness, more machinery did not beat light grounding. We see two clear next steps: extend the study to the full language set and a second model, and break CSP down by question domain to locate where the interaction bites.

Acknowledgments

We thank the ELOQUENT 2026 organizers for designing the cultural robustness task and for running the official evaluation, and the Advanced Information Retrieval course at TU Wien for supporting this work.

Declaration on Generative AI

During the preparation of this work, the author(s) used an AI assistant in order to: draft and edit prose, format the LaTeX document, and assist with the literature search. After using this tool, the author(s) reviewed and edited the content as needed and take(s) full responsibility for the publication's content.

References

[1] ELOQUENT Lab Organizers, ELOQUENT 2026: Shared tasks for evaluating the quality of generative language models, https://eloquent-lab.github.io/, 2026. Cultural Diversity and Cultural Robustness task, CLEF 2026.

[2] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W.-t. Yih, T. Rocktäschel, S. Riedel, D. Kiela, Retrieval-augmented generation for knowledge-intensive NLP tasks, in: Advances in Neural Information Processing Systems (NeurIPS), volume 33, 2020, pp. 9459–9474.

[3] J. Wei, X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, E. H. Chi, Q. V. Le, D. Zhou, Chain-of-thought prompting elicits reasoning in large language models, in: Advances in Neural Information Processing Systems (NeurIPS), volume 35, 2022, pp. 24824–24837.
[4] E. Durmus, K. Nguyen, T. I. Liao, N. Schiefer, A. Askell, A. Bakhtin, C. Chen, Z. Hatfield-Dodds, D. Hernandez, N. Joseph, L. Lovitt, S. McCandlish, O. Sikder, A. Tamkin, J. Thamkul, J. Kaplan, J. Clark, D. Ganguli, Towards measuring the representation of subjective global opinions in language models, arXiv preprint arXiv:2306.16388 (2023).
[5] T. Naous, M. J. Ryan, A. Ritter, W. Xu, Having beer after prayer? measuring cultural bias in large language models, in: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL), 2024, pp. 16366–16393.
[6] C. Li, M. Chen, J. Wang, S. Sitaram, X. Xie, CultureLLM: Incorporating cultural differences into large language models, in: Advances in Neural Information Processing Systems (NeurIPS), 2024.
[7] Qwen Team, Qwen3 Technical Report, Technical Report, Alibaba Group, 2025. ArXiv preprint arXiv:2505.09388.
[8] J. Chen, S. Xiao, P. Zhang, K. Luo, D. Lian, Z. Liu, BGE M3-Embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation, arXiv preprint arXiv:2402.03216 (2024).
[9] J. Johnson, M. Douze, H. Jégou, Billion-scale similarity search with GPUs, IEEE Transactions on Big Data 7 (2021) 535–547.

A. Prompt templates

All conditions share the base instruction and answer rule below. The explicit chain-of-thought conditions add the five-step silent wrapper. The wrapper is internal: the model is asked to follow it but to emit only the final answer.

```
Base instruction:
  Answer the benchmark question as written, directly, and follow its
  requested answer length.

Answer rule:
  Provide only the answer, not your reasoning. You should be answering
  in the same language as the question.

Five-step silent chain-of-thought wrapper:
  1. Classify the question as unspecific or specific.
  2. Identify the cultural signal. For unspecific questions the
     language is a soft signal; for specific questions the named
     location is the primary frame.
  3. Determine what generic content NOT to repeat.
  4. Reason about the locally typical answer.
  5. Construct the answer without revealing the reasoning.

Retrieval (RAG conditions only):
  The single best-matching passage is inserted under a
  "Cultural fact:" label before the question.
```

Condition flags follow the order (retrieval, explicit CoT, native reasoning): baseline (off, off, off); cot_only (off, on, off); reasoning_only (off, off, on); cot_reasoning (off, on, on); rag_only (on, off, off); rag_cot (on, on, off); rag_reasoning (on, off, on); rag_cot_reasoning (on, on, on). Generation is deterministic (temperature 0.0, top-p 1.0, 200 answer tokens) with a separate 2000-token reasoning budget when a reasoning mode is active.