University of Amsterdam at the CLEF 2026 ELOQUENT Cultural Robustness Task

Influence of Conversational Context, Personalization, and Prompt Framing on Cultural Signal Purity

Bruno N. Sotic[1], Liza Borman[1], Bram Koorstra[1] and Jaap Kamps[1]

[1]*University of Amsterdam, Amsterdam, The Netherlands*

Abstract

This paper describes the University of Amsterdam's participation in the ELOQUENT Cultural Robustness task at CLEF 2026. We study how three aspects of modern LLM interaction, namely conversational context, personalization through retrieved user memory, and system-prompt framing, affect a model's cultural behaviour. Using the task's open-ended, culturally variable questions in five languages (German, English, French, Italian, Swedish), we evaluate four open instruction-tuned models and one closed model under a context-free baseline and a set of injected-context conditions that vary whether the injected context conflicts with or is neutral to the prompt's culture, and whether it is delivered as prior conversation turns or as retrieved memory; for the closed model we instead vary the system-prompt style. We assess the submissions with the organizers' Cultural Signal Purity (CSP) metric and with a qualitative analysis of the generated answers. We find that injecting any conversational context lowers CSP for the capable models, and that a culture-neutral context is at least as damaging as a conflicting one. The qualitative analysis shows why the scalar alone is insufficient: a conflicting context homogenizes answers toward the injected culture, at times overriding a country named explicitly in the prompt, whereas a neutral context flattens answers toward generic, culture-free content. These are opposite failures that a single CSP score cannot tell apart.

Keywords

Large Language Models, Cultural Robustness, Cross-Lingual Evaluation, Conversational Context, Prompt Framing, Personalization

1. Introduction

Large language models are now deployed across many languages and cultural settings, and users reasonably expect culturally appropriate answers. This involves two distinct demands. When no cultural context is given, a model should reflect plausible cultural diversity rather than collapse onto a single default. When a region is specified, it should respect that region consistently, regardless of the language it is asked in. The ELOQUENT lab[1] runs a series of CLEF shared tasks for evaluating the quality of generative language models [1, 2]. Its Cultural Robustness and Diversity task at CLEF 2026 [3] operationalizes both demands with a set of open-ended, culturally variable questions posed in many languages.

Real deployments, however, rarely present a question in isolation. Assistants carry conversational history, retrieve personalized memory about the user, and run under system prompts that shape their tone and behaviour. These interaction-design choices may interfere with a model's cultural behaviour in ways a single-turn benchmark does not capture. Our participation is built around this observation: rather than evaluating the bare questions, we deliberately inject conversational context and personalization and measure the effect on cultural signal.

Our submissions form a small two-by-two over injected context. The injected material either conflicts with or is neutral to the prompt's culture, and it is delivered either as prior conversation turns or as retrieved user memory; we also include context-free baselines and, for a closed model, a set of

CLEF 2026: Conference and Labs of the Evaluation Forum, September 21–24, 2026, Jena, Germany

nadalic.sotic@uva.nl (B. N. Sotic); kamps@uva.nl (J. Kamps)

0009-0003-2122-2235 (B. N. Sotic); 0000-0002-6614-0087 (J. Kamps)

[1] https://eloquent-lab.github.io/

system-prompt styles. We evaluate the submissions with the organizers' Cultural Signal Purity (CSP) metric [3] and a qualitative reading of the outputs, in five languages. Our main findings are twofold. Quantitatively, injecting any context degrades CSP for the capable models, with a culture-neutral context as harmful as a conflicting one. Qualitatively, the two conditions fail in opposite directions, a distinction the scalar metric cannot reveal on its own.

Section 2 describes the experimental setup, Section 3 the evaluation methodology, Section 4 reports and analyses the results, and Section 5 concludes.

2. Experimental Setup

We compare a context-free baseline against a set of conditions that inject material into the prompt: prior conversation turns that either conflict with or are neutral to the prompt's culture, retrieved user-memory personalization, and, for the closed model, several system-prompt styles.

2.1. Data and prompts

The task provides open-ended, culturally variable questions covering everyday domains: food, work, education, and social and family norms, and each question comes in two forms. In the *unspecific* form the question is asked on its own, with no country mentioned, so any cultural grounding has to be inferred from the language it is written in. In the *specific* form the same question is accompanied by a short grounding that names a country, for example "We live in Austria and we want to eat like locals," together with an instruction not to mention that country in the answer; each question is paired with many such country groundings. Because the difference between the two is simply whether a cultural anchor is present in the prompt, we refer to them interchangeably as culturally *anchored* (specific) and *unanchored* (unspecific). We work with five of the languages the task offers: German, English, French, Italian, and Swedish.

2.2. Baseline

We evaluate four open-weight instruction-tuned models, Mistral-7B-Instruct-v0.2, Llama-3.1-8B-Instruct, Qwen2.5-7B-Instruct, and Aya-Expanse-8B, together with GPT-4.1-mini as a closed reference point. The baseline is a single-turn setup: each question is passed on its own, with no added conversational context, and it serves as the reference against which the injected-context conditions are measured. We decode with a sampling temperature of 0.7 rather than greedily, since this reflects how these models are typically run in deployment. To confirm that this sampling does not make the results unstable, we first generated each prompt 15 times per model and found that, within a model, the outputs were largely the same from run to run. Given this low variation, and following the submission guidelines, the baseline scores we report are computed from a single one of these runs.

2.3. Prior Conversational Turns (History)

In these conditions we prepend a short, fixed conversation before the target question and then ask the question as the final turn. The prepended conversation has two turns, each a user message and an assistant reply (four messages in total, U_1, A_1, U_2, A_2), and is written in the same language as the question. We use two versions of this history, a neutral one and a conflicting one. The question itself is left unchanged, and we generate one response per question, language, model, and condition.

2.3.1. Neutral culture

The neutral history is culturally generic and names no country: the user asks how to behave appropriately across different situations and cultures, and the assistant replies about the importance of context and norms. The two turns are:

- *User:* "I'm trying to understand how to behave appropriately in different situations and cultures." *Assistant:* "It depends a lot on the situation and the local norms, so the best approach is to stay observant, be respectful, and adapt to the context you are in."
- *User:* "I'm worried about misreading what is expected of me and accidentally doing the wrong thing." *Assistant:* "That is understandable; paying attention to how others behave and asking when you are unsure usually goes a long way."

2.3.2. Conflict culture

The conflicting history places the user in a specific country that differs from the one named in the question. The user explains that they have recently moved there and want to fit in with the local culture, and the assistant replies about adjusting. The two turns are:

- *User:* "I recently moved to [country] and I want to embrace the local culture." *Assistant:* "That is wonderful; immersing yourself in the local food, customs, and everyday habits is the best way to settle in and feel at home."
- *User:* "I'm worried about misreading what is expected here and accidentally doing the wrong thing." *Assistant:* "That is natural when adjusting to a new place; observing how locals do things and following their lead will help you fit in."

The injected country is selected per question and is different from the country in the target prompt: we seed a random number generator with the question's identifier and draw a country from a fixed list, which makes the selection reproducible and specific to each question. Because this country differs from the one named in an anchored prompt, the history pulls toward one culture while the question grounds another.

2.4. RAG (Retrieval-Based Personalization)

Instead of conversation turns, this condition prepends retrieved user-memory statements to the question. For each question we retrieve from a fixed bank of 15 short user-profile statements by dense vector similarity: the question and the statements are encoded with OpenAI's `text-embedding-3-large`, scored by cosine similarity to the question, and the three highest-scoring statements are selected and prepended to the unchanged question, each marked explicitly as a memory item. We use the same models as in the baseline and generate one response per question, language, and model. We run two versions, differing only in the contents of the memory bank.

2.4.1. Neutral

The 15 statements describe general, culture-free preferences and habits, such as an interest in local customs, a preference for authentic regional food, and a tendency to ask locals for recommendations. For a question like "What is a typical breakfast?", food- and custom-related statements rank above unrelated ones; the prepended block looks like:

> [Memory: I prefer homemade meals using fresh, locally sourced ingredients.]
> [Memory: I value authentic regional food over generic international options.]
> [Memory: I try to follow local customs and traditions when in a new place.]

2.4.2. Dutch persona

This version is identical to the neutral one in retrieval and generation; only the bank changes, with the 15 statements now describing norms, preferences, and habits typical of the Netherlands, for example:

> [Memory: For breakfast I usually have bread with cheese or hagelslag.]

Injecting these Dutch preferences regardless of the question's language lets us probe cross-lingual cultural transfer: whether an injected cultural profile shifts responses even in languages other than the culture's own.

2.5. System Prompt Modifications

In addition, we investigate whether changes to the system prompt can influence cultural robustness without changing the underlying model or the prompt content. We keep the user prompts unchanged and vary only the system instructions given to the model, which makes the responses styles the main differences between conditions.

All experiments in this section use GPT-4.1-mini. The prompts remain the same for all conditions, which means that the system instruction is the only factor that changes between experiments. We define the following three conditions:

Structured, asks the model to produce clear and well-organised responses. This was included to examine whether a more structured response style affects cross-lingual consistency.

Evidence, instructs the model to answer based on the supported information from the prompt and to avoid making unsupported assumptions. In this case, we analyze whether relying more closely on the information provided in the prompt affects consistency across languages.

Concise, encourages short and direct answers. The condition focuses on the relationship between response length and variation in the outputs.

All conditions were generated using GPT-4.1-mini with the same generation parameters. The only difference between runs was the system instruction. This design allows us to examine how changes to the system prompt affect cultural robustness while keeping the model unchanged.

3. Evaluation

We assess the submissions primarily with the organizers' Cultural Signal Purity (CSP) metric, complemented by a qualitative analysis of the generated answers.

3.1. Cultural Signal Purity

CSP asks whether a model's cultural diversity and its regional robustness are independent capabilities rather than two faces of a single language-default pattern, in the spirit of correlation-based probes of model behaviour. For each question q and pair of languages (L_1, L_2), the responses are embedded with a multilingual sentence encoder (the organizers use Qwen3-Embedding). The diversity signal U_q is the cosine distance between the two unanchored answers, which should be high if the model varies appropriately across languages. The robustness signal A_q is the mean, over groundings, of the cosine distance between the two anchored answers, which should be low if the model converges on the specified region regardless of language. For each language pair one computes the Spearman correlation ρ between $\{U_q\}$ and $\{A_q\}$; writing $\bar{\rho}$ for the mean over pairs,

$$\mathrm{CSP} = \frac{1 - \bar{\rho}}{2}. \tag{1}$$

A high CSP requires the two signals to be decoupled, that is a negative $\bar{\rho}$: the model can be diverse and robust independently rather than having both driven by one language default. CSP therefore measures the separability of these capabilities, not factual correctness. We report the organizers' official CSP, computed with Qwen3-Embedding. As a cross-check we also re-implemented CSP with LaBSE sentence embeddings; this reproduced the broad ordering of models but not the finer ordering among conditions, which we attribute to the choice of encoder and discuss in Appendix A. We do not rely on it in the main text.

3.2. Qualitative protocol

Because CSP is a single scalar, it cannot show how an affected model's answers change, so we also inspect the outputs directly, in two parts. We begin with the history conditions, conflict and neutral, because they isolate the effect most cleanly: a single, controlled prefix is the only thing that varies. For

Table 1
Cultural Signal Purity (CSP) scores for all UAmsterdam submissions. Higher values indicate stronger cultural signal purity.

Model	Baseline	Conflict	Neutral	RAG-N	RAG-DP
Aya-Expanse-8B	0.865	0.669	0.474	0.527	0.571
Llama-3.1-8B-Instruct	**0.875**	**0.691**	0.464	**0.666**	**0.732**
Mistral-7B-Instruct-v0.2	0.731	0.555	0.405	0.425	0.611
Qwen2.5-7B-Instruct	0.432	0.388	**0.487**	0.581	0.428
GPT-4.1-mini prompt variants					
Baseline	0.651	–	–	–	
Concise	0.616	–	–	–	
Evidence	**0.753**	–	–	–	
Structured	0.659	–	–	–	

these we focus on the two highest-baseline-CSP open models, Aya-Expanse-8B and Llama-3.1-8B, in German and English, on both anchored and unanchored prompts, comparing the two conditions across a handful of questions spanning the task's categories. We then examine the GPT-4.1-mini system-prompt conditions, inspecting responses across the prompt styles to understand what drives the differences in their scores. Both analyses are interpretive rather than statistical, and German outputs are translated into English for presentation.

4. Results

Table 1 reports CSP for the four open models across the baseline and the four conditions that inject context, together with a set of GPT-4.1-mini runs that vary the system prompt instead of injecting context (discussed below). For the three open models with a strong baseline (Aya, Llama, and Mistral) the pattern is consistent: putting any context before the question lowers CSP, and substantially. The conflict condition costs roughly 0.18 to 0.20 relative to baseline, about a fifth of the score; the neutral condition costs roughly 0.33 to 0.41, close to half; The two retrieval (RAG) conditions split the same way as the history pair: the neutral memory bank (RAG-N) is more damaging than the Dutch-persona bank (RAG-DP) for every strong model (Aya 0.527 vs. 0.571, Llama 0.666 vs. 0.732, Mistral 0.425 vs. 0.611), and for Llama and Mistral the Dutch-persona condition is in fact the gentlest of the four, edging out even conflict. What this means is clearest in the two quantities CSP is built from. For a given question, one measures how much a model's answers vary across languages when no country is named (which should be high), the other how much they vary when a country is named (which should be low), and CSP rewards models in which these two move independently. A fall in CSP therefore signals that this independence is eroding: once a prior conversation precedes the question, the model's answers across languages converge whether or not a country was specified, so the variation that should track language and the consistency that should track the named country are both overwritten by the injected context.

Two features of these results are counterintuitive, and a single score makes them easy to overlook. First, for all three strong models the neutral condition is more damaging than the conflict condition, even though neutral names no country and conflict injects one that usually clashes with the prompt. Were a clash between the injected country and the prompt's own grounding the main cause of the degradation, conflict should score below neutral; the reverse holds. The same ordering recurs in the retrieval channel, where the neutral memory bank (RAG-N) again scores below the Dutch-persona bank (RAG-DP) for all three strong models, so the effect is not an artifact of the conversation-history format. This suggests the damage has less to do with cultural conflict than with the mere presence of a dominating prefix that pulls the answers together, though the score cannot establish this on its own. Second, Qwen behaves in the opposite direction, improving under both neutral (0.487) and RAG (0.581) from a much weaker baseline (0.432), with conflict its only damaging condition and the Dutch-persona

RAG (0.428) leaves it essentially unchanged. We are cautious in interpreting this. Because CSP averages rank correlations, a model whose baseline is already this low has diversity and robustness signals that were barely separated to begin with, so a correlation this close to zero can move in either direction for reasons unrelated to culture; we read Qwen as a single-model anomaly rather than evidence that context benefits weaker models.

The GPT-4.1-mini rows probe a different lever: rather than injecting history or memory, they vary the system prompt, asking whether the framing of the instruction alone moves the cultural signal. It does, but within a far narrower band than context injection moves the open models. The variants span 0.614 to 0.753, with the default at 0.651, and only the evidence prompt stands out, raising CSP by about 0.10 above the default. Two points temper even this. The two prompts intended as neutral defaults (0.651 and 0.614) themselves differ by about 0.04, so part of the spread reflects sensitivity to phrasing rather than to the named styles, and we do not read the smaller gaps as meaningful. And as before, the score registers that the evidence prompt shifted the balance between diversity and robustness without telling us why: whether it drew out genuinely more region-specific content or merely changed the wording and format of the answers. Finally, GPT-4.1-mini's baseline (0.651) sits below the three strong open models, but this is not a judgment on quality. CSP rewards keeping diversity and robustness separate, so a lower value here indicates that these two signals are more entangled for this model, not that its answers are worse or less culturally aware.

These observations expose the limit of the metric for our purpose. CSP is a single number, an average over language pairs of the correlation between two distance signals, and by construction it reports that a model's behaviour across languages changed without revealing how. It cannot distinguish a model that homogenizes its answers toward an injected culture from one that strips out cultural detail and returns the same generic answer in every language: both reduce the separation between diversity and robustness, both lower CSP, yet they are opposite failures. Nor can it tell us, on the prompts that name a country, whether a model has lost robustness by drifting toward the country supplied in prior turns or simply by becoming inconsistent across languages. The numbers in Table 1 are therefore best read as evidence that conversational context disrupts how these models handle culture, while the nature of that disruption, and the reason the conditions fall in such a surprising order, is left open by the score. Because several of these results are counterintuitive and a single figure aggregates the mechanism away, they warrant a closer look at the responses themselves.

We therefore turn to a qualitative analysis, focusing on the two history conditions, conflict and neutral, where the manipulation is cleanest and the contrast between supplying a specific culture and supplying none is most directly visible in the answers.

4.1. Qualitative Analysis

The scores tell us that context changes a model's cross-language behaviour but not what changes in the answers; we read the outputs to recover that, first for the history conditions and then for the GPT-4.1-mini prompt styles

4.1.1. History conditions

Reading the responses (Table 2), the two history conditions turn out to fail in opposite directions. Under conflict, the country introduced through the prior conversation becomes the operative cultural anchor: asked what to feed a child both models describe a Turkish breakfast, asked about a gift a Bosnian one, asked about name days a Luxembourgish custom, and the injected country often surfaces by name despite the explicit instruction not to mention it. Under neutral, the same questions collapse onto generic, culture-free advice ("a balanced breakfast", "a gift matched to their hobbies"), and occasionally onto the culture tied to the language rather than none at all, as when Llama answers the German name-day question with German custom. The two conditions that looked alike through CSP are therefore doing opposite things: conflict homogenises answers toward an injected culture, neutral flattens them toward the generic. This is exactly the distinction a single correlation cannot draw, since both compress

Table 2
Representative responses from Aya-Expanse-8B and Llama-3.1-8B under the two history conditions. Glosses summarise the cultural content of each answer (German answers translated), with both models agreeing unless a split is noted. In "prompt / history", the first entry is the country named in an anchored prompt ("–" if none) and the second is the country injected through the prior conversation turns.

Question (cat.)	Lang	Prompt / history	Conflict	Neutral
Breakfast (food)	en	– / Turkey	Turkish: menemen, sucuk, poğaça	Generic "balanced breakfast"
Breakfast (food)	en	Austria / Hungary	*Hungarian*: goulash, paprikás csirke (overrides named country)	Austrian/European: Müsli, rolls
Breakfast (food)	de	Austria / Hungary	*Austrian*: Kaiserschmarrn (named country wins)	Austrian: bread, jam
Gift (social norms)	en	– / Bosnia	Bosnian: rakija, local crafts	Generic: wine, gift to hobbies
Gift (social norms)	de	Austria / Iceland	*Icelandic* (Llama, overrides) vs. Austrian (Aya)	Generic regional gift
Name days (soc. norms)	en	– / Luxembourg	Luxembourg patron-saint day	Generic; Llama leans German/Eur.

cross-language variation and both lower the score.

On the anchored prompts the effect sharpens into outright override of the country the prompt names. With Austria named and Hungary injected, the English answers swap the Austrian breakfast for Hungarian dishes (goulash, paprikás csirke) while the German answers stay Austrian (Kaiserschmarrn); for the gift question the split falls by model rather than language, with German Llama drifting to an Icelandic gift while Aya keeps it Austrian. Override happens, then, but not systematically: whether the named or the injected country wins depends on the model, the language, and the item. A model that "loses robustness" here is sometimes drifting toward the injected country and sometimes not, which is just the heterogeneity a per-pair correlation averages into one number.

This also suggests why neutral, counterintuitively, scored below conflict. Under conflict the injected culture still supplies differentiated, language-specific content, so the answers keep some cross-language distance; under neutral they fall back on near-identical generic templates that read almost as translations of each other, compressing that distance more completely. We offer this only as a reading of a few cases, not a measurement, especially since the ordering did not survive our change of encoder (Appendix A). And the whole effect is confined to culturally loaded questions: for an item with no real cultural dimension, such as the drawbacks of a teenage job, conflict and neutral return essentially the same generic answer, with no injected country anywhere in it.

4.1.2. System-prompt conditions

To complement the quantitative scores, we manually inspected several responses produced by the different system-prompt conditions. The examples suggest that the difference in scores reflects different response patterns rather than differences in writing style. A recurring pattern in the Structured and Concise conditions was filling in missing contextual information. For example, when asked about the best way to get around "the capital", the Structured response assumed that a metro system existed and recommended it as the best option, whereas the Evidence condition stated that the prompt did not say which capital was being referred to. A similar pattern appeared in the homework example: asked how many hours of homework high-school students should expect per week, the Concise and Structured conditions gave an estimate of about 5 to 10 hours even though no country or educational system was specified, while the Evidence condition instead stated that the prompt did not provide enough information about the school system. The same happened for workplace norms: asked whether it is common to send Slack messages to colleagues on weekends or late at night, the Concise and Structured conditions gave general advice, whereas the Evidence condition again declined for lack of context. These examples indicate that the Evidence condition improves its score mainly by reducing

assumptions about missing context: the model avoids guessing and stays with the information in the prompt, which is consistent with its having the highest CSP among the GPT-4.1-mini variants. The Concise condition behaves differently. Its responses are shorter and more direct, but in many cases the underlying assumptions remain unchanged (it still estimates homework hours and still makes workplace assumptions), so it affects response length more than content. This likely explains why Concise produces only a small change while Evidence moves the score much more.

5. Conclusion

We examined how conversational context, retrieved personalization, and system-prompt framing affect the cultural behaviour of LLMs on the ELOQUENT Cultural Robustness task. Across the capable models, injecting any context lowered Cultural Signal Purity, and a culture-neutral context was at least as damaging as a conflicting one; varying only the system prompt moved the closed model far less. The qualitative analysis showed that the scalar conflates two opposite effects: a conflicting context steers answers toward the injected culture and can override an explicitly named country, while a neutral context flattens them toward generic content. For deployed assistants this indicates that conversational memory and personalization are not culturally neutral by default, and that evaluating cultural robustness benefits from measures sensitive to the direction of change, not only its magnitude. A natural next step is to turn the difference between two minimally different context branches into a controllable signal for steering a model's cultural behaviour rather than merely diagnosing its drift.

Acknowledgments

This research was conducted as part of the final research projects of the Bachelor's in Artificial Intelligence at the University of Amsterdam, We thank the coordinator, Dr. Sander van Splunter, for his support and flexibility to work around the CLEF deadlines. We thank the track and task organizers for their outstanding service and effort in making realistic benchmarks available to evaluate the quality of generative language models.

Bruno Sotic is partly funded by the Netherlands Organization for Scientific Research (NWO NWA # 1518.22.105). Jaap Kamps is partly funded by the Netherlands Organization for Scientific Research (NWO CI # CISC.CC.016, NWO NWA # 1518.22.105), the University of Amsterdam (AI4FinTech program), and ICAI (AI for Open Government Lab). Views expressed in this paper are not necessarily shared or endorsed by those funding the research.

Declaration on Generative AI

During the preparation of this work, the authors used and Grammarly in order to: Grammar and spelling check and Paraphrase and reword. After using these tools/services, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] J. Karlgren, L. Dürlich, E. Gogoulou, L. Guillou, J. Nivre, M. Sahlgren, A. Talman, ELOQUENT CLEF shared tasks for evaluation of generative language model quality, in: Advances in Information Retrieval: 46th European Conference on Information Retrieval, ECIR 2024, Glasgow, UK, March 24–28, 2024, Proceedings, Part V, Springer-Verlag, Berlin, Heidelberg, 2024, pp. 459–465. URL: https://doi.org/10.1007/978-3-031-56069-9_63. doi:10.1007/978-3-031-56069-9_63.

[2] J. Karlgren, M. Barrett, O. Bojar, M. I. Engels, D. Fabre, S. Ettejjari, L. Goeuriot, R. R. Gunti, J. Mothe, P. Mulhem, M. Piacentini, L. F. V. Madriz, D. Schwab, P. Šindelář, G. Stampoulidis, K. Thomas, M. Vartampetian, Overview of ELOQUENT 2026: shared tasks for evaluating generative language model quality, in: M. Hagen, M. Potthast, B. Stein, P. Schaer, E. Zangerle, S. MacAvaney, J. M. Struß, E. S. Salido, A. Barrón-Cedeño, A. G. S. de Herrera (Eds.), Experimental IR Meets Multilinguality,

Multimodality, and Interaction. Proceedings of the Seventeenth International Conference of the CLEF Association (CLEF 2026), Springer, 2026.

[3] M. Barrett, J. Karlgren, J. Mothe, G. Stampoulidis, Overview of the Cultural Robustness and Diversity Task at the ELOQUENT 2026 lab for evaluating generative language model quality, in: E. S. Salido, A. Barrón-Cedeño, A. G. S. de Herrera, S. MacAvaney, J. M. Struß (Eds.), Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2026), CEUR-WS, 2026.

A. Appendix

All data, code, and material necessary to reproduce this study are publicly available at: https://huggingface.co/datasets/StagerLee/eloquent-2026

Table 3
LaBSE-based Cultural Signal Purity (CSP) scores by model and condition. CSP is computed as $(1 - \bar{\rho})/2$, where $\bar{\rho}$ is the mean Spearman correlation between unanchored and anchored cross-language distances across language pairs.

Model	Baseline	Conflict	Neutral	RAG-N	RAG-DP	GPT:Concise	GPT:Evidence	GPT:Structured
Aya	0.346	0.184	0.223	0.207	0.253	–	–	–
GPT-4.1-mini	0.297	–	–	–	–	0.258	0.307	0.284
Llama	0.378	0.242	0.254	0.274	0.301	–	–	–
Mistral	0.290	0.216	0.237	0.269	0.300	–	–	–
Qwen	0.243	0.201	0.270	0.239	0.248	–	–	–

Table 4
LaBSE-based mean unanchored cross-language distance ($\mathrm{mean\_U}$) by model and condition. Higher values indicate larger cross-language distance for unspecific answers.

Model	Baseline	Conflict	Neutral	RAG-N	RAG-DP	GPT:Concise	GPT:Evidence	GPT:Structured
Aya	0.379	0.291	0.291	0.325	0.331	–	–	–
GPT-4.1-mini	0.253	–	–	–	–	0.277	0.280	0.231
Llama	0.335	0.302	0.311	0.341	0.347	–	–	–
Mistral	0.464	0.391	0.413	0.454	0.434	–	–	–
Qwen	0.368	0.321	0.334	0.347	0.351	–	–	–

Table 5
LaBSE-based mean anchored cross-language distance ($\mathrm{mean\_A}$) by model and condition. Higher values indicate larger cross-language distance for specific grounded answers averaged over shared groundings.

Model	Baseline	Conflict	Neutral	RAG-N	RAG-DP	GPT:Concise	GPT:Evidence	GPT:Structured
Aya	0.415	0.329	0.318	0.327	0.333	–	–	–
GPT-4.1-mini	0.284	–	–	–	–	0.292	0.278	0.249
Llama	0.359	0.341	0.327	0.350	0.394	–	–	–
Mistral	0.495	0.441	0.439	0.462	0.489	–	–	–
Qwen	0.387	0.357	0.351	0.357	0.371	–	–	–

Table 6
LaBSE-based CSP values compared with the organizers' reported Qwen3 CSP scores for the overlapping model–condition settings.

Model	Condition	LaBSE CSP	Organizers' Qwen3 CSP
Aya	Baseline	0.346	0.865
Aya	Conflict	0.184	0.669
Aya	Neutral	0.223	0.474
Aya	RAG-N	0.207	0.527
Aya	RAG-DP	0.253	0.571
Llama	Baseline	0.378	0.875
Llama	Conflict	0.242	0.691
Llama	Neutral	0.254	0.464
Llama	RAG-N	0.274	0.666
Llama	RAG-DP	0.301	0.732
Mistral	Baseline	0.290	0.731
Mistral	Conflict	0.216	0.555
Mistral	Neutral	0.237	0.405
Mistral	RAG-N	0.269	0.425
Mistral	RAG-DP	0.300	0.611
Qwen	Baseline	0.243	0.432
Qwen	Conflict	0.201	0.388
Qwen	Neutral	0.270	0.487
Qwen	RAG-N	0.239	0.581
Qwen	RAG-DP	0.248	0.428
GPT-4.1-mini	Baseline	0.297	0.651
GPT-4.1-mini	GPT:Concise	0.258	0.616
GPT-4.1-mini	GPT:Evidence	0.307	0.753
GPT-4.1-mini	GPT:Structured	0.284	0.659

Table 7
Correlation between LaBSE-based CSP scores and the organizers' reported Qwen3 CSP scores for the overlapping model–condition pairs.

Statistic	Value
Spearman ρ	0.627
Pearson r	0.692