Overview of the Cultural Robustness and Diversity Task of the ELOQUENT 2026 Lab for Evaluating Generative Language Model Quality

Notebook for the ELOQUENT Lab at CLEF 2026

Maria Barrett[1], Jussi Karlgren[2], Josiane Mothe[3] and Georgios Stampoulidis[2]

[1]AMD Silo AI, Denmark
[2]AMD Silo AI, Sweden
[3]University of Toulouse

Abstract

Generative language models are intended to be creative and responsive to the style of the conversation they engage in. The experimental Cultural Robustness and Diversity task is designed to explore how language models vary in question answering tasks with respect to cultural factors, conditioned on the language used to pose the question and on the presence of a location anchor in the prompt. Participants were encouraged to experiment with approaches that either infer cultural grounding from language or utilize explicit cultural information in prompts. Rather than optimizing solely for performance, the task emphasized experimental design to understand factors affecting cultural robustness. We received 107 valid submissions and six system reports. Teams explored diverse approaches including prompt prefixing and retrieval-augmented generation to inject cultural context. Results demonstrate that cultural robustness is strongest when models are directly anchored to relevant facts in the appropriate language, and weakest when culture must be inferred through multiple layers of context or reasoning.

1. Evaluating Generative Language Model Quality

The shared task experiment described in this report is part of the ELOQUENT lab for evaluation of generative language model quality at CLEF, the Conference and Labs of the Evaluation Forum. A full report from the lab can be found in the CLEF Conference proceedings [1]. This task is the successor of two earlier tasks probing how generative language models are affected by variation in input.

Our general hypothesis is that training data will carry value systems from the culture they are taken from and that instruction training and other tuning procedures will systematically modify the responses in some direction which indeed is the entire purpose of such training. We wish to demonstrate what sort of variation can be traced to cultural background of models and to the data they are trained on.

In its first year, the task experimented with stylistic and dialectal variation in a set of English-language questions to explore how type of question might influence the response content. The results varied over topic, and were largely influenced by specific lexical items [2]. In its second year, the experiment consisted of a set of questions about values and habits given in a selection of languages, to explore how cultural variation is predicated on cross-linguistic variation, by differential prompting, and by different variants of models, as shown by differences between systems trained in different languages [3].

This year, we explored model diversity and robustness to culturally sensitive prompts when the cultural context needs to be inferred from the prompt language (unanchored) and across languages when the cultural context is explicitly defined in the prompt (anchored).

The key metric, is cultural signal purity (CSP) which is an evaluation method that measures cultural capabilities of multilingual LLMs without ground-truth labels. It compares model outputs across languages for the anchored and unancored conditon. Responses are embedded into a shared vector space, and for each language pair the method computes distances between outputs in both conditions; CSP is derived from the correlation between these distance patterns, rewarding models that both

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

Figure 1: Illustration of example model output with and without cultural grounding in the prompts. For clarity and cross-linguistic accessibility, actual model outputs are represented using emojis in this illustrative example.

(a) Illustration of the unanchored condition where the cultural context is inferred from the language.

(b) Illustration of the anchored condition where the cultural context is specified in the prompt across languages.

produce culturally diverse outputs when context is implicit and converge appropriately when context is explicitly specified. The method is presented in Barrett et al. [4].

2. Participation

The task had 72 registered teams. By the deadline, 22 teams participated, with 141 submitted experimental conditions using a variety of models and prompt variants.

Using the metadata file, we were able to combine multiple submissions from the same team into fewer valid submissions. Each submission need as a minimum four complete languages. We took out 31 submissions due to incompleteness or inablity to merge into complete submissions. We end up with 107 valid submissions from 13 teams. Six teams submitted approved reports.

3. Data

The test set consists of 101 unspecific prompts and 4140 specific prompts in each language, exemplified in Figure 1. For culturally anchoring the prompts, we used one region where each language is an official language as well as all EU countries.[1] A few languages also have anchoring in the form of three ages (25, 53, 77) in a few cases where we believed age would be an interaction effect, e.g., `I would like to start dating. I am [age] years old and live in [region]`. The test items were authored in English and manually translated and localised by organisers, participants, and volunteers into 23 languages, at time of writing, shown in Table 1.

Our key metric *Cultural Signal Purity* (CSP), is a combination of two measures, one for unanchored variation and one for anchored variation. We compute for each question q an unanchored score, U_q, as

[1]Austria, Belgium, Bosnia-Herzegovina, Bulgaria, Catalonia, Croatia, the Republic of Cyprus, the Czech Republic, Denmark, Estonia, the Faroe Islands, Finland, France, Georgia, Germany, Greece, Hungary, Iceland, India, Ireland, Israel, Italy, Karnataka, Latvia, Liechtenstein, Lithuania, Luxembourg, Maharashtra, Malta, the Netherlands, North Macedonia, Norway, Poland, Portugal, the Republic of Cyprus, Romania, Russia, Slovakia, Slovenia, Spain, Sweden, Tamil Nadu, Telangana, Turkey, Ukraine

a measure of how much the responses to the unanchored questions vary across languages, to see how much the system under consideration infers from the language used.

If U_q is high, we assume that the system infers information from the language that the question is posed in; if U_q is low, the variation across languages is low. We also compute A_{qc}, anchored scores, as measures of how much the system responses varies across questions where cultural context c is explicitly specified. If an A_{qc} score for some given c is high across languages we assume that the system pays more attention to the language used than the location anchor c given; if it is low, we assume that the system uses the location for shaping the response. We combine the A_{qc} scores computed over all given location anchors c into an averaged A_q score.

This procedure is performed over each possible pair of languages (L_1, L_2): from the set of languages to yield sets of $U_q(L_1, L_2)$ and $A_q(L_1, L_2)$ scores for each system. For each language pair (L_1, L_2) we compute a Spearman correlation ρ_{L_1,L_2} between the $U_q(L_1, L_2)$ and $A_q(L_1, L_2)$ scores to examine how U and A covary over the set of questions q. We expect a system that pays attention to the location anchor to show lower rather than higher correlation and consider this to be a desirable behaviour with respect to cultural sensitivity. We average the ρ_{L_1,L_2}s to $\bar{\rho}$ and compute the final score for the system as $CSP = 1 - \bar{\rho}^2$. We require a system submission to cover a minimum of 4 languages for the results to be reliable. More detail on the procedure is given in a separate publication on the test collection [4].

The CSP computation relies on representing the multilingual responses generated by a language model using a scheme which can assess the semantic similarity of response sentences quantitatively, reliably, and deterministically and which performs reasonably well for both high-resource and low-resource languages. For these experiments we have for this purpose selected `Qwen3-8B` embeddings [5], a vectorisation scheme which represents sentences as vectors in a multilingual semantic space, with semantically similar sentences located near each other, across languages. We use cosine between the representation vectors as the basis for our similarity score. This scheme can be replaced with any other vectorisation procedure or indeed any other quantifiable multilingual semantic similarity measure.

The CSP metric however does not assess the topical quality of system responses. Random or nonsensical responses will presumably result in a high correlation, but this metric is not intended to measure truthfulness or factual correctness.

The Cultural Robustness and Diversity task is not a classical competitive shared task where participants aim solely for the highest score. Instead, participants were encouraged to try various strategies to improve the cultural diversity and robustness and make multiple submissions with different strategies. Minimal participation was a baseline submission with no changes to the prompts. A submission consisted of minimum four languages with complete output for the anchored and unanchored condition, where the same strategy should be used for the entire submission.

4. Submissions and results

4.1. Language coverage

Table 1 show an overview of the languages covered by the valid submissions. We observe higher frequency among higher-resource Western languages such as English, German, French, Italian, Spanish. Some languages are not used in any submissions. One team added Arabic to the data set.

4.2. Best CSP scores

Table 2 show an overview of all teams with at least one valid submission, their best CSP score, but most importantly the δ between their highest and lowest CSP score. This table shows that even one-model submissions have substantial δ between their best and worst system, meaning that CSP can be moved quite a lot with a diverse range of methods. Figure 2 show an overview of the CSP for all submission by teams with accepted lab reports by model. We observe quite wide CSP sc ore variability within each model.

Table 1
Language distribution of the 107 valid submissions.

Language		in n valid submissions
Arabic [2]	ar	1
Bengali	bn	0
Catalan	ca	0
Czech	cs	0
Danish	da	5
English	en	108
Faroese	fo	2
Finnish	fi	3
French	fr	94
German	de	101
Greek	el	15
Hebrew	he	1
Hindi	hi	4
Italian	it	92
Kannada	kn	0
Marathi	mr	0
Polish	pl	23
Russian	ru	23
Slovak	sk	0
Spanish	es	79
Swedish	sv	30
Tamil	ta	5
Telugu	te	0

Table 2
Result summary of the 107 valid Cultural Robustness and Diversity task submissions at ELOQUENT 2026 showing best CSP, n submissions, n models, and δ between the highest and lowest score per team. Teams are ranked by δ. Boldfaced team names denote an accepted lab note report.

Ref	Team Name	n subm.	n models	Best score	δ
[7]	**UAmsterdam**	25	9	0.875	0.488
	The Lamas	9	1	0.825	0.477
[8]	**MiPV AI-IR**	12	2	0.783	0.405
	Master MIAGE Toulouse | DreamBear	4	2	0.647	0.275
	fbe4bb208901f	36	1	0.636	0.261
[9]	**TU Wien - Methodology 4 - Group 1**	8	1	0.640	0.237
[10]	**SRH ReliableAI**	3	1	0.718	0.236
	Fatma Nurefsan Ates	3	1	0.581	0.205
	annika-simonsen-uoi	2	2	0.646	0.117
[11]	**MIAGE_M2**	2	2	0.995	0.114
[6]	**L3_TOULOUSE**	1	1	0.744	0.000
	Master MIAGE Toulouse - DECAP SOUTRIC ZAIT AMBLARD	1	1	0.414	0.000
	Nicolas Rieuvernet	1	1	0.959	0.000

4.3. Detailed analysis

Table 3 reports for each team the languages they consider, the models they used to generate the responses, the evaluation method they used as well as the main strategies they tried. Many teams complement the official CSP score report with a qualitative analysis [7, 8, 10, 6]. A few also add different visualizations [11, 6].

Table 3
Overview team submissions with approved reports.

Ref.	Evaluation / Analysis	Languages	Models	Strategies
[7] **UAmsterdam**	CSP, qualitative analysis of conflicting vs. neutral context, language-default homogenization, and culture override effects	English, French, German, Italian, Swedish	Mistral-7B-Instruct-v0.2, Llama-3.1-8B-Instruct, Qwen2.5-7B-Instruct, Aya-Expanse-8B, GPT-4.1-mini	Context-free baseline; conflicting vs. neutral prior conversation; retrieved user-memory personalization; system-prompt styles for the closed model
[8] **MiPV**	CSP, qualitative analysis of Western defaults, stereotypical associations, and weak grounding	English, German, Greek, Italian, Polish, Russian, Spanish, Turkish	Llama 3.1 8B Instruct, Aya Expanse 8B	Baseline, chain-of-thought, persona, uncertainty-aware prompting, combined prompting, key word injection
[9] **TU Wien**	CSP, comparison against baseline, condition-by-condition analysis, qualitative mechanism interpretation	English, French, German, Polish, Russian, Spanish	Qwen3-32B	2×4 factorial over retrieval (on/off) and four reasoning modes (none, explicit chain-of-thought, native reasoning, both)
[10] **SRH**	CSP; response examples; qualitative inspection of cultural grounding, language stability, and failure cases	Danish, English, Finnish, French, German, Greek, Hindi, Italian, Polish, Russian, Swedish, Spanish	Qwen2.5-32B-Instruct-AWQ	Baseline vs. RAG with English-language cultural facts vs. RAG with target language cultural facts; keyword matching; retrieval of top cultural facts; translated knowledge-base injection
[11] **MIAGE**$_{M2}$	Word count, type-token ratio, sentence embeddings, K-means clustering, PCA, Kruskal-Wallis, Mann-Whitney U, qualitative anomaly analysis	English, French, German, Italian, Spanish	Llama 3 8B, Llama 3 70B, Mistral Small, Mistral Large	Parameter sweep over Top-p and presence penalty; compare unspecific vs. specific prompts
[6] **L3**	CSP, Silhouette Score, cluster inspection, representative-response analysis, response-length analysis, Arabic parenthetical-expression analysis	Arabic, English, French, Italian, Spanish	openai/gpt-oss-120b	Specific vs. unspecific prompts; Arabic prompt translation and manual curation

The University of Amsterdam team [7] used the largest number of different models. They considered five languages and five models: Mistral-7B-Instruct-v0.2, Llama-3.1-8B-Instruct, Qwen2.5-7B-Instruct, Aya-Expanse-8B, and GPT-4.1-mini, the latter being the only model that was not an open-weight model. This paper studied how conversational context, retrieved user memory, and system-prompt framing affected outputs. It found that injecting context generally lowered CSP, that neutral context could be as damaging as conflicting context, and concluded the scalar metric alone missed important failure modes. More specifically, the team tested several injected-context settings, varying both whether

the added context conflicted with the prompt's culture or remained culturally neutral, and whether it appeared as earlier dialogue or retrieved memory. The results showed that adding conversational context of any kind reduced CSP for the stronger models, with culture-neutral context proving at least as harmful as conflicting context. The qualitative analysis clarified why the scalar score was not enough: conflicting context tended to pull responses toward the injected culture, sometimes even overriding a country explicitly named in the prompt, while neutral context pushed responses toward generic, culture-agnostic content. Their highest CSP score (0.875) was obtained by Llama-3.1-8B-Instruct with the initial prompts.

The University of Pavia team [8] tried the highest number of strategies (six). They used eight languages (English, German, Greek, Italian, Polish, Russian, Spanish, and Turkish) and two models (Llama 3.1 8B Instruct and Aya Expanse 8B). In addition to the baseline strategy, they used chain-of-thought (identify the cultural context, determine the relevant norms, and infer the culture the user is referring to). They also employed the following strategies: *Persona* strategy where the model was asked to endorse the role of someone living in a specific country and familiar with its customs, traditions and everyday norms; *Uncertainty* where the model was asked first to evaluate what could and could not be inferred from the question and use this uncertainty; *Combined* that combined the previous strategies; and *Keyword injection*. The authors found that CSP was very similar for Aya whatever the strategy used while it differed quite a lot for Llama. Their qualitative analysis revealed a tendency toward superficial cultural adaptation: the model recognized the prompt required adaptation but their responses remained weakly grounded in the target culture. They also detected cultural rigidity, generic Western default outputs, stereotypical cultural associations, invented cultural content, ungrounded cultural references, asymmetric cultural representation, and gender norm reproduction. Their highest CSP score (0.783) was obtained by Aya using the *Persona* strategy.

The TU Wien team [9] considered six languages (English, French, German, Polish, Russian, Spanish) and a single model: Qwen3-32B. The team investigated the use of retrieval-augmented generation (RAG) (either off or on) on Wikipedia content after it had been split into passages with BGE-M3 [12] and indexed with FAISS [13]; when on, the best retrieved passage was added to the prompt as a cultural fact. The team also used two types of chain-of-thought reasoning: (a) native reasoning mode that used the model's own thinking and (b) non-native that asked 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 study did not include any evaluation or analysis beyond the official score. The team found that a single retrieved fact without reasoning scored best, while combining retrieval with chain-of-thought scored worst. Their highest CSP score (0.640) was obtained when RAG only was used.

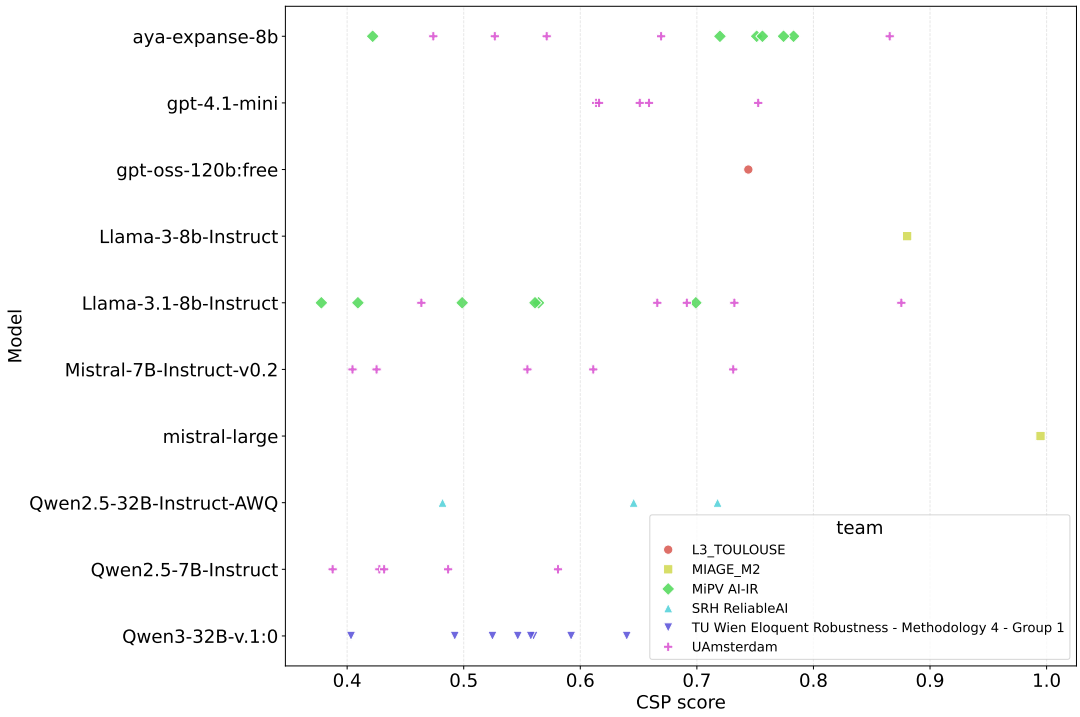
The SHR University of Heidelberg team [10] used the highest number of languages (12) with a large variety (Romance and Indo-European). They used a single model (Qwen2.5-32B-Instruct-AWQ) with three types of contexts, two of which used simulated RAG mechanisms to provide additional context. They used (a) prompts as they were (b) prompts augmented with three English facts (c) prompts augmented with three facts translated into the prompt language. The facts were chosen by a pipeline that first detected the topic of the prompt (e.g., food, education, etc.), then the culture (either inferred by the language if there was no geographic anchor or by extracting the geographic anchor). From their qualitative analysis, the authors concluded that incorrect cultural grounding could be worse than no grounding at all. The authors concluded that the Qwen2.5-32B-Instruct-AWQ benefited from being told what it already knew and that the language in which the facts were written mattered more than initially thought. Their highest CSP score (0.718) was obtained using translated facts.

The University of Toulouse (MIAGE) [11] used four models from two families: two open source (Llama 3 8B and Llama 3 70B) and two proprietary models via their APIs (Mistral Small and Mistral

Large). The team considered five languages (English, French, German, Italian, Spanish). Here the team analysed the models' parameter influence. First the top-p parameter (Nucleus Sampling) that controlled the cumulative probability threshold for token sampling, managing the model's creativity or risk-taking window; it ranged from 0.0 (highly predictable) to 1.0 (higher risk of hallucination). Second the Presence Penalty parameter that controlled the model's likelihood to repeat tokens that had already appeared in the generated text; ranging from -2.0 (encouraged to reuse identical words) to 2.0 (penalized for repetitions). The team tested three hypotheses: (H1) A high Top-p value (0.9) would generate significantly longer responses than a constrained Top-p value (0.1) [H1 was wrong, differences were not statistically significant]; (H2) The Presence Penalty parameter would directly influence vocabulary richness, as measured by the Type-Token Ratio (TTR) [H2 was wrong; the differences were more linked to the model than to the parameter value]; (H3) The target language dictated the syntax structure and naturally reflected on the baseline lexical diversity score [while length did not vary when changing Top-p value, it was dependent on the syntax diversity of the target language]. The team also proposed other visualizations. Their highest CSP score was (0.995).

The University of Toulouse (L3) [6] used a single model gpt-oss-120b on five languages (English, French, Italian, Spanish and Arabic that the team added to the collection). The team did not try different strategies but rather provided different analyses like word count, type-token ratio, cosine-similarity embeddings, K-means clustering, PCA, and non-parametric tests. The analyses showed that specific prompts—those mentioning cultural context such as country or age—often produced more diverse responses than unspecific prompts. This was especially clear for Arabic, French, and Italian. The paper also found that specific prompts usually improved Silhouette Scores, meaning the responses formed more coherent and better-separated semantic clusters, especially in English and Italian. Qualitatively, the authors concluded that unspecific prompts tended to generate generic everyday advice, while culturally specific prompts encouraged references to local foods, traditions, family heritage, memories, and culturally grounded values. For example, breakfast prompts led to culturally specific items such as burek, halloumi, waffles, Turkish tea, or regional breads in Arabic. Their CSP score was (0.744).

Figure 2: CSP score by model and team for all teams with accepted lab notes report.



5. Main takeaway

The teams have taken very different approaches. Some have chosen to incorporate a wide range of languages, others a wide range of models, and still others a wide range of analytical methods. They made some hypotheses, but the intuition was rarely met, showing that we are far from understanding model behaviour.

From these studies, cultural robustness seems to be strongest when the model is lightly anchored to a relevant fact in the right language, and weakest when it is asked to infer culture through layers of context or reasoning.

6. Plans for Next Year

After this shared task, we will release the code for scoring and the task material in LM Eval Harness. We will not repeat the task in this form at CLEF, but the task will be available under a permissive license for the community. Rather, we will explore different aspects of dialectal and geographic variations. We encourage the community to refine the methodology as well as translate the dataset to other languages.

Acknowledgments

The ELOQUENT lab is partially supported by the European Commission through the OpenEuroLLM (101195233-DIGITAL-2024-AI-06) and the DeployAI (101146490-DIGITAL-2022-CLOUD-AI-B-03) projects as part of their activities on building, evaluating, and disseminating generative language models. Views and opinions expressed are those of the authors and do not necessarily reflect those of the European Union or the Digital Europe programme. Neither the European Union nor the programme can be held responsible for them. It is also partially supported by the ANR (French National Research Agency) for its financial support of the GUIDANCE project n°ANR-23-IAS1-0003.

References

[1] 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.

[2] M. Sahlgren, J. Karlgren, L. Dürlich, E. Gogoulou, A. Talman, S. Zahra, ELOQUENT 2024 — Robustness Task, in: G. Faggioli, N. Ferro, M. Vlachos, P. Galuščáková, A. G. S. de Herrera (Eds.), Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2024), volume 3740, CEUR-WS.org, 2024.

[3] J. Karlgren, M. I. Engels, M. Barrett, R. R. Gunti, M. Hoveyda, B. Nadalic Sotic, J. Kamps, M. Koistinen, E. Zosa, Overview and Joint Report of the Robustness and Consistency Task at the ELOQUENT 2025 lab for evaluating generative language model quality, in: G. Faggioli, N. Ferro, P. Rosso, D. Spina (Eds.), Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2025), CEUR-WS, 2025.

[4] M. Barrett, G. Stampoulidis, D. Zautner, J. Burdge, J. Karlgren, Measuring the cultural capabilities of large language models, in: TBD - Under review, Under review, 2026.

[5] Y. Zhang, M. Li, D. Long, X. Zhang, H. Lin, B. Yang, P. Xie, A. Yang, D. Liu, J. Lin, F. Huang, J. Zhou, Qwen3 embedding: Advancing text embedding and reranking through foundation models, arXiv preprint arXiv:2506.05176 (2025).

[6] H. Mekkaoui, J. Mothe, Univ. Toulouse (IRIT Lab.) at ELOQUENT Lab at CLEF 2026, 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.
[7] B. N. Sotic, B. Koorstra, L. Borman, J. Kamps, University of Amsterdam at ELOQUENT 2026: Probing Cultural Signal in Multilingual LLMs through Conversation History and Cultural Anchoring, 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.
[8] P. Topaloglu, T. Vaccari, A. Avdeiko, G. Peikos, Investigating the Effect of Prompting Strategies on the Cultural Awareness of Large Language Models, 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.
[9] S. Negm, Z. Zohny, K. Sulo, A. Kormaku, TU Wien Methodology 4 at ELOQUENT 2026: A Factorial Study of Retrieval and Reasoning for Cultural Robustness, 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.
[10] P. Priyanshu, S. Chandna, SRH-ReliableAI at the 2026 ELOQUENT lab for evaluating generative language model quality: CultuRAG for Cultural Robustness and Diversity, 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.
[11] E. Minguez, L. Bonenfant, T. Renauld, T. Schimdt, Challenge ELOQUENT @ CLEF 2026 – Robustesse Culturelle & Diversité - m2 MIAGE, 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.
[12] 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 4 (2024).
[13] J. Johnson, M. Douze, H. Jégou, Billion-scale similarity search with gpus, IEEE transactions on big data 7 (2019) 535–547.