Investigating the Effect of Prompting Strategies on the Cultural Awareness of Large Language Models*

MiPV AI-IR: ELOQUENT Lab at CLEF 2026

Pelinsu Topaloglu[1,2], Teresa Vaccari[1,2], Artem Avdeiko[1,2] and Georgios Peikos[2,*]

[1]University of Pavia, 27100 Pavia, Italy
[2]University of Milano-Bicocca (DISCo), 20126 Milan, Italy

Abstract

This paper describes the systems submitted by our team, MiPV AI-IR, to the cultural robustness and diversity task in the ELOQUENT Lab at CLEF 2026. Large Language Models are increasingly used for advice-seeking in socio-cultural contexts, including etiquette, social norms, and cultural practices. However, prior work has shown that these models may exhibit cultural biases and favor certain cultural perspectives over others. We investigate prompting as a non-invasive strategy for steering culturally adaptive generation, without additional training or technical expertise. We evaluate six prompting strategies, namely baseline, chain-of-thought, persona, uncertainty-aware prompting, combined prompting, and keyword injection. These strategies are tested across Llama 3.1 8B and Aya Expanse 8B, allowing us to compare a general-purpose instruction-tuned model with a model optimized for multilingual generation. Experiments are conducted in eight languages: English, Turkish, Italian, Greek, Russian, Spanish, German, and Polish. Quantitative results show that prompting strategies have model-specific effects, with chain-of-thought prompting performing best for Llama and persona prompting achieving the strongest result for Aya. Our qualitative analysis suggests that both models often exhibit superficial cultural adaptation, where responses appear culturally aware but remain weakly grounded in the target culture. This behavior appears through Western defaults, limited cultural variation, stereotypical associations, ungrounded cultural references, and asymmetric representations of non-Western cultures across languages. Overall, our findings show that prompting can improve cultural signal behavior, but that qualitative analysis remains important for understanding how cultural adaptation is expressed in model outputs.

Keywords

Large Language Models, Multilingual Robustness, Cultural Alignment, Prompt Engineering, Cross-lingual Evaluation, Conversational AI

1. Introduction

Large Language Models (LLMs) are increasingly deployed in socially situated settings, where users communicate with conversational systems through language patterns, expressions, and interaction styles shaped by their cultural and social context [1, 2, 3, 4, 5]. In fact, language cannot be perceived as a culturally neutral means of personal expression, as it reflects socially grounded norms, values, communicative conventions, cognitive patterns, and shared contextual knowledge [6, 7, 8]. LLM-based conversational systems are now commonly used for information access, recommendations, education, and everyday assistance, including requests related to social norms, etiquette, and personal advice [9, 10, 11]. This use also extends to morally, culturally, and socially sensitive situations, where users may seek guidance on interpersonal behavior, ethical judgment, etiquette, and social norms [12, 13]. Given that LLM-generated content can achieve human-level persuasiveness [14], failures in cultural adaptation may directly affect the reliability and appropriateness of responses generated by conversational systems. This concern has motivated recent work on whether LLMs reflect culturally specific values,

CLEF 2026 Working Notes, 21 – 24 September 2026, Jena, Germany
*Corresponding author.
pelinsu.topaloglu01@universitadipavia.it (P. Topaloglu); teresa.vaccari01@universitadipavia.it (T. Vaccari); artem.avdeiko01@universitadipavia.it (A. Avdeiko); georgios.peikos@unimib.it (G. Peikos)
0009-0002-3690-2463 (P. Topaloglu); 0009-0000-6696-2374 (T. Vaccari); 0009-0004-7448-9272 (A. Avdeiko); 0000-0002-2862-8209 (G. Peikos)

represent low-resource cultures accurately, and respond consistently to culturally grounded emotional contexts [15, 16, 17].

A key difficulty in evaluating cultural robustness is that language and culture are closely intertwined, yet they may signal different cultural contexts. In such interaction, prompts are not only semantic inputs to a model, but also carriers of cultural cues that may influence how an LLM interprets a request and generates a response. Consequently, semantically related prompts expressed in different languages may reasonably lead to culturally different outputs. For example, asking for a "typical student lunch" may evoke different foods, habits, and everyday practices depending on the language of interaction. However, when the target culture is explicitly specified, such as in a prompt asking for a "typical Italian student lunch," responses should remain semantically consistent across languages. This distinction creates a central evaluation challenge for LLMs, particularly for models trained or deployed in multilingual settings. Such models should adapt to culturally grounded contexts implicitly conveyed through language, while remaining robust when the intended cultural setting is explicitly controlled.

In this context, cultural diversity captures the extent to which an LLM adapts its responses when cultural context is implicitly inferred from the language of the prompt. Cultural robustness, on the other hand, captures the extent to which an LLM preserves the intended cultural context across languages and prompt variations when that context is explicitly specified. These two dimensions are complementary, i.e., a model that produces nearly identical outputs across all languages may fail to capture culturally situated expectations, while excessive variation may indicate unstable or inconsistent multilingual behavior. Balancing them is important because multilingual systems should neither ignore culturally grounded expectations nor exhibit uncontrolled behavioral variation across languages.

We investigate cultural diversity and cultural robustness within the Cultural Robustness and Diversity Track of the ELOQUENT 2026 lab [18, 19]. Previous studies suggest that multilingual LLMs may exhibit uneven cultural adaptation due to imbalances in training data distributions and the predominance of English-centric resources [20]. Existing approaches for improving multilingual cultural behavior mainly rely on fine-tuning and prompting strategies[21, 22, 23]. While fine-tuning attempts to encode culturally relevant knowledge directly into model parameters, prompting approaches aim to guide model behavior through natural language instructions and contextual steering.

In this work, we focus on prompting strategies because they provide a lightweight and non-invasive mechanism for studying multilingual cultural behavior without modifying the underlying model parameters. Furthermore, prompting strategies can be easily transferred across languages and deployment scenarios, and can be communicated to end users as practical guidelines for improving culturally aware interactions with LLMs. For this reason, prompting strategies offer a useful bridge between controlled multilingual evaluation and practical guidance for real-world LLM use.

We conduct experiments with multiple prompting strategies across two instruction-tuned LLMs and eight languages, namely **English**, **Turkish**, **Italian**, **Greek**, **Russian**, **Spanish**, **German**, and **Polish**. Specifically, we compare two similarly sized instruction-tuned models, Llama 3.1 8B Instruct [24] and Aya Expanse 8B [25]. Llama 3.1 8B Instruct is a general-purpose model with multilingual capabilities, while Aya Expanse 8B is explicitly optimized for multilingual instruction following. This comparison allows us to examine whether multilingual specialization affects culturally grounded response generation, particularly in balancing cultural adaptation with cross-lingual behavioral robustness.

2. Related Work

Recent research has increasingly investigated the cultural behavior and robustness of multilingual LLMs. Previous studies suggest that such LLMs may exhibit uneven cultural adaptation due to imbalances in training data distributions and the predominance of English-centric resources [20]. Other work has explored whether LLMs reflect culturally specific values, preserve culturally grounded emotional representations, and accurately represent low-resource cultures across languages [15, 16, 17].

Prior work has mainly approached cultural alignment in LLMs from two directions, namely, post-training methods and inference-time methods. Post-training approaches typically rely on culturally

informed fine-tuning, synthetic cultural data generation, or specialized training objectives. For example, Li et al. [21] use the World Values Survey (WVS) to generate synthetic culturally aligned training data for specific countries, while Wu et al. [22] apply a similar WVS-based approach for Chinese cultural adaptation. Both studies show improvements in cultural awareness after fine-tuning. Similarly, Liu et al. [26] leverage assemblies of LLM agents to simulate social interactions and capture implicit cultural norms and values, whereas Li et al. [27] use multi-agent interactions to generate culturally grounded datasets for subsequent fine-tuning. Other studies focus on modifying the training objective itself. For instance, Elazar et al. [28] introduce a consistency loss during pretraining to improve response consistency, while Feng et al. [29] propose a multi-stage training paradigm with culturally aware reward modeling based on explicit metrics such as cultural recall, precision, and F1.

Although post-training methods have shown promising results, they are computationally expensive and require curated datasets, repeated training, and access to model parameters. As an alternative, several studies explored inference-time prompting strategies, which guide model behavior without modifying the underlying model. Wei et al. [23] show that chain-of-thought (CoT) prompting improves reasoning performance by encouraging intermediate reasoning steps before producing a final answer. Persona prompting approaches instead guide model behavior through predefined characteristics or social roles [30]. Within the previous edition of the ELOQUENT lab, Sotic and Kamps [31] evaluate persona and CoT prompting strategies for multilingual robustness and observe higher consistency under persona prompting settings. However, the effectiveness of persona prompting remains debated. Zheng et al. [32] show that selecting effective personas is challenging and may negatively affect model performance when inappropriate personas are used. Other inference-time methods focused on uncertainty-aware reasoning. For example, Hu et al. [33] introduce uncertainty-based reward mechanisms for information-seeking tasks, while Yin et al. [34] propose adaptive guidance methods that identify and reduce uncertainty sources during inference.

Overall, previous work suggests that cultural alignment can be improved either through post-training adaptation or inference-time guidance. In this work, we focus on inference-time prompting strategies because they provide a lightweight and non-invasive mechanism for studying multilingual cultural behavior without modifying model parameters. Furthermore, prompting strategies can be transferred across languages and deployment scenarios, and can also be communicated to end users as practical guidelines for culturally aware interaction with LLMs.

3. Methodology

We introduce an inference-time methodology for studying cultural diversity and robustness in multilingual LLMs. Rather than adapting the models through additional training, we focus on how cultural information can be provided through prompts. This allows us to evaluate the role of prompt design in guiding models toward culturally appropriate response generation.

Our pipeline is organized into three stages, as shown in Figure 1. Given a user input, the system first infers the cultural context that should guide the response before constructing the final prompt. In our methodology this step is handled by KeyBERT [35], using the paraphrase-multilingual-MiniLM-L12-v2 backbone [36], a lightweight multilingual model well-suited for keyword extraction across the eight target languages. The model takes the user's input and extracts two types of information, namely the *country*, which is explicitly stated by the user in their query, and a set of culturally relevant keywords which serve as *contextual cues*. For instance, given the input "What do people in Japan wear to weddings?," the system extracts the country "Japan" and keywords such as *wear* and *weddings*. The extracted country and keywords are then inserted into predefined prompt templates to construct the final prompt according to one of six prompting strategies we developed in our study. The constructed prompt is finally passed to an instruction-tuned LLM, which generates an answer. Instruction-tuned models are used because they are optimized to follow structured natural language instructions, allowing us to isolate the effect of prompt design on culturally appropriate generation while keeping the model parameters fixed.

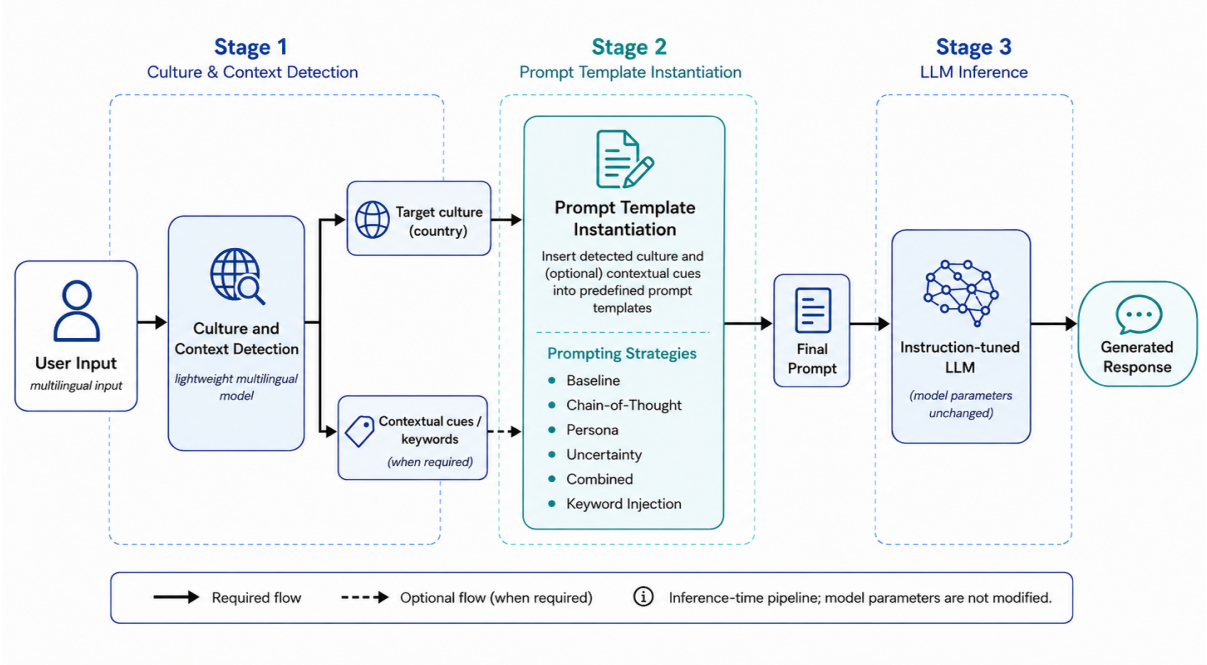

Figure 1: Overview of the proposed inference-time prompting pipeline for culturally grounded response generation.

Table 1 summarizes the six prompting strategies used in our experiments and their intended goals. We define these strategies to examine how different forms of cultural guidance affect response generation at inference time.

The Baseline strategy uses the original user input without any additional prompting instructions, providing a reference point for evaluating the other strategies. CoT prompting instructs the model to reason about the cultural aspects of the input before producing an answer, encouraging deliberate cultural reasoning rather than surface-level pattern matching. This makes it particularly relevant for culturally grounded generation, and similar reasoning-based prompting strategies have been shown to be effective in prior ELOQUENT submissions [31]. Persona prompting encourages the model to answer from the perspective of someone familiar with the target cultural background. Despite mixed evidence on its effectiveness, persona prompting remains a relevant strategy for culturally grounded generation, as prior work suggests that culturally situated personas can influence model responses [37]. Uncertainty prompting is intended to reduce unsupported assumptions by asking the model to consider what can and cannot be inferred from the input. Prior work suggests that explicitly prompting models to recognize the limits of their knowledge during reasoning can support more reliable LLM behavior [38]. The Combined strategy integrates CoT reasoning, persona-based framing, and uncertainty awareness into a single prompt. This aims to examine whether stacking complementary strategies compounds their individual benefits. This allows us to examine whether combining prompting strategies leads to stronger performance than using each strategy independently. Finally, Keyword Injection provides the model with the contextual cues extracted during the detection stage, together with the target culture represented as a country name. This strategy offers a lightweight alternative to CoT reasoning by making selected cultural cues explicit without requiring the model to generate an intermediate reasoning process.

Overall, this methodology enables a controlled comparison of how different prompting strategies encode and expose cultural information to the model at inference time.

Table 1
Prompting strategies for culturally grounded response generation and their intended goals.

Strategy	Description	Goal
Baseline	The model is prompted using only the original user input.	Provides a reference point for comparing the effect of the other prompting strategies.
CoT	The model is instructed to identify the relevant cultural aspects before generating the response.	Encourages explicit reasoning about cultural norms and contextual factors.
Persona	The model is instructed to answer from the perspective of a person familiar with the inferred cultural background.	Aims to improve cultural alignment within the target context.
Uncertainty	The model is instructed to consider what can and cannot be inferred from the input before answering.	Aims to reduce overconfident or culturally unsupported responses.
Combined	The model is provided with a prompt that combines reasoning, persona-based framing, and uncertainty awareness.	Evaluates the joint effect of multiple cultural guidance strategies.
Keyword Injection	The model is provided with the extracted contextual cues and the target culture, expressed as a country name.	Provides explicit cultural information that should guide the generated response.

4. Experimental Setup

This section describes the experimental setting used to evaluate the proposed prompting methodology. We first present the datasets used for cultural diversity and cultural robustness evaluation. We then describe the prompt templates used to instantiate the prompting strategies introduced in Table 1. Finally, we report the instruction-tuned models used for response generation.

4.1. Dataset

For our empirical evaluation, we use the official datasets released by the organizers of the Cultural Diversity and Cultural Robustness task at the ELOQUENT Lab of CLEF 2026 [18]. The task evaluates how LLMs respond to culturally grounded questions across 24 languages and different cultural cues. The dataset includes two complementary evaluation settings, with a culture-unspecific setup used to assess cultural diversity when the cultural context must be inferred from the input language, and a culture-specific setup used to assess cultural robustness when the target country is explicitly provided.

The cultural diversity setting consists of 101 questions per language in which the cultural context is not explicitly specified. The questions cover a broad range of culturally dependent situations, including social, workplace, family, and societal norms, as well as socially sensitive topics such as gender roles and areas requiring explicit cultural knowledge, such as food preferences, holiday practices, and home decoration. Most inputs are open-ended and require the model to identify the relevant cultural background and generate a complete answer. A smaller subset follows a more closed formulation, for example questions of the form *"Is it accepted to ...?"*.

The cultural robustness setting contains culture-specific variants of the same questions. In this case, the target cultural context is explicitly provided through a country reference, by a direct addition to the prompt. For each language, this setting includes 4,140 questions. This setup makes the intended cultural context explicit, but it also allows us to examine whether the model follows the specified country context or is influenced by the language in which the query is written.

4.2. Models

For response generation, we evaluate two open-source instruction-tuned language models with the same parameter scale, Llama 3.1 8B Instruct [24] and Aya Expanse 8B [25]. Using models of comparable size allows us to reduce the influence of model scale when comparing their behavior. The two models differ mainly in their multilingual specialization. Llama 3.1 8B Instruct is a general-purpose instruction-tuned model with multilingual support, while Aya Expanse 8B is specifically designed for multilingual alignment and covers all eight languages considered in our evaluation. This comparison allows us to examine whether multilingual alignment affects culturally grounded response generation beyond the effect of model size.

4.3. Prompts & Generation

The prompt templates used to implement the prompting strategies are presented in Table 2. Each template modifies the original user input by adding a specific form of cultural guidance, while the Baseline condition keeps the input unchanged. For strategies that require additional cultural information, the detected target culture and the extracted contextual cues are inserted into the corresponding template before generation. In the templates, **{country}** and **{culture}** refer to the detected target culture expressed as a country name, while **{context}** refers to the contextual cues extracted during the detection stage. For prompting strategies requiring an explicit cultural anchor, **{country}** and **{culture}** were populated with the provided country in the culture-specific setting. In the culture-unspecific setting, these fields were set to an *unknown* placeholder, so the model had to infer the relevant cultural context from the input.

Table 2
Prompt templates used to instantiate the prompting strategies.

Strategy	Template
Baseline	No additional instructions are added, and the original user input is passed to the model as given.
CoT	Think through the question step by step before answering. Identify the cultural concepts being asked about, determine which kinds of norms are relevant, for example social, workplace, family, or religious norms, and infer the specific culture or cultural setting the user is referring to. Then interpret the question within that context and provide a culturally appropriate, accurate, and context-sensitive answer.
Persona	Answer as someone who lives in **{country}** and is familiar with its customs, traditions, and everyday norms. Provide responses that reflect how people in that culture would typically understand and handle the situation.
Uncertainty	Before answering, carefully evaluate what can be inferred from the question and what cannot. Avoid making unsupported assumptions about cultural specifics. When the cultural context is ambiguous or your confidence is limited, explicitly acknowledge the uncertainty and answer cautiously.
Combined	Reason step by step before answering. Identify the cultural concepts and relevant norms, for example social, workplace, and family norms, and infer the cultural context. Answer as a local from **{country}**, reflecting realistic customs, traditions, and expectations. Avoid unsupported assumptions. If cultural details are uncertain, acknowledge this and respond cautiously. Give a clear and culturally appropriate answer.
Keyword Injection	Given the extracted context **{context}** and the target culture **{culture}**, use this information to provide a culturally appropriate answer.

During generation, each question is processed in a separate session to avoid context carryover across questions, and generation is limited by setting `max_new_tokens=200`.

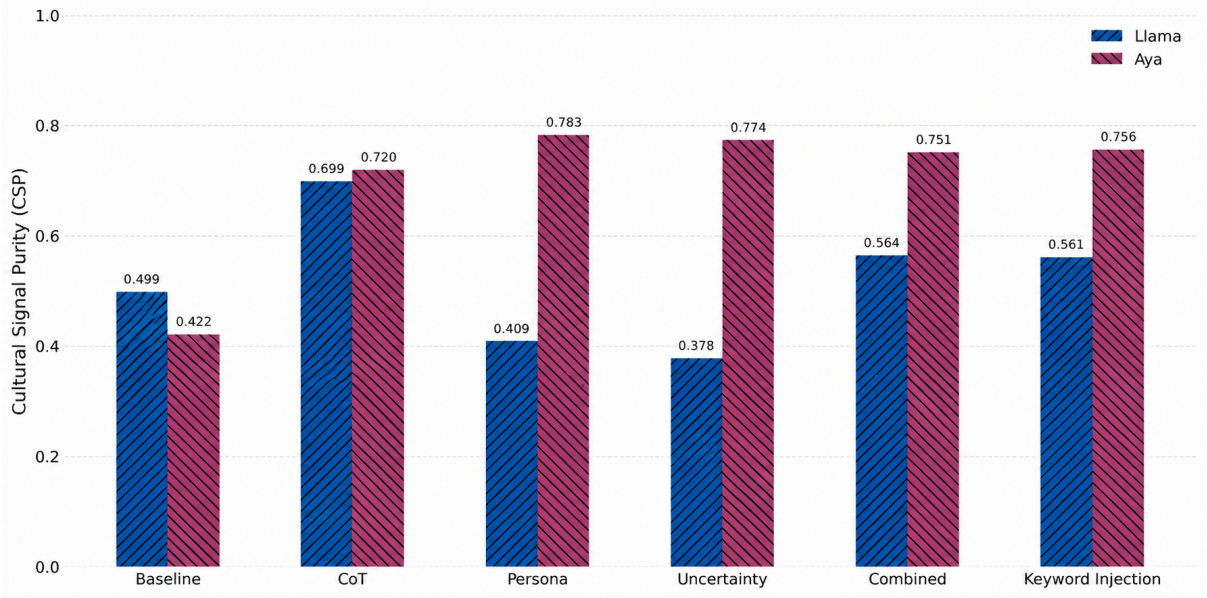

Figure 2: Cultural Signal Purity (CSP) scores of the MiPV AI-IR systems across Llama 3.1 8B Instruct and Aya Expanse 8B.

5. Results & Discussion

This section presents the results of our evaluation. We first report the quantitative robustness results based on the official scores shared by the task organizers. We then present a qualitative analysis of sampled model responses across the eight languages considered in our experiments. This analysis was conducted by the authors together with colleagues who are either native speakers of, or proficient in, the evaluated languages. Finally, we discuss the main findings by relating the quantitative results to the qualitative observations.

5.1. Quantitative Results

The official evaluation is based on the Cultural Signal Purity (CSP) score, which measures whether multilingual LLMs can both produce culturally diverse responses when the cultural context is inferred from the prompt language and adapt to explicit cultural grounding regardless of the prompt language. The score is computed by comparing embedding-based distances between model responses in two conditions, using Qwen3-8B [39] to encode the responses. The first condition measures cross-language response variation when no cultural grounding is provided, while the second measures cross-language variation when the same cultural context is explicitly anchored in the prompt. A higher CSP score indicates that the model can produce culturally diverse responses when the cultural context is inferred from language, while also overriding language-based defaults when explicit cultural grounding is provided. Figure 2 reports the Cultural Signal Purity (CSP) scores obtained by the MiPV AI-IR systems across the two evaluated models. Overall, Aya obtains higher CSP scores than Llama in almost all experimental settings. The only exception is the base setting, where Llama slightly outperforms Aya. Prompting strategies produce notably different effects depending on the model.

For Llama, chain-of-thought prompting is the most effective strategy, substantially increasing the CSP score from 0.499 in the base setting to 0.699. This is the highest score achieved by Llama across all experiments. The combined and keyword injection settings also improve over the base score, reaching 0.564 and 0.561, respectively, although the gains are more moderate. By contrast, the persona and uncertainty-aware prompts reduce performance, with scores of 0.409 and 0.378, respectively. This suggests that Llama benefits most from reasoning-oriented prompting, while persona-based and uncertainty-aware instructions may introduce ambiguity or interfere with the cultural signal captured by the evaluation.

For Aya, all prompting strategies improve over the base setting. The strongest result is obtained with the persona prompt, which increases the CSP score from 0.422 to 0.783. The uncertainty-aware, keyword injection, and combined prompts also perform strongly, with scores of 0.774, 0.756, and 0.751, respectively. Chain-of-thought prompting also improves Aya's performance, reaching 0.720, although it is not the best configuration for this model. These results suggest that Aya is particularly responsive to explicit cultural or role-based prompting, and that it can use these instructions more effectively than Llama in this setting.

A clear model-specific pattern emerges from these results. Chain-of-thought prompting is the most effective configuration for Llama, whereas persona prompting is the strongest configuration for Aya. This indicates that prompting strategies have model-specific effects. Although the base setting is relatively weak for both models, Aya shows a larger improvement after prompting, moving from the lowest base score to the strongest overall scores. The combined prompt does not outperform the best individual strategies, suggesting that combining multiple instructions does not necessarily produce additive gains. Similarly, keyword injection improves both models over their base settings, but it is not the best configuration for either model. This indicates that explicit cultural cues are useful, although their effectiveness depends on how they are integrated into the prompt and on the model being used.

Overall, the obtained results indicate that prompt design substantially affects the model's ability to produce culturally diverse responses when culture is inferred from language, while also adapting to explicit cultural grounding when it is provided. However, this effect is model dependent. Aya shows consistent gains across all prompting strategies and achieves the strongest scores among our submitted configurations, while Llama benefits mainly from chain-of-thought prompting. The combined prompt and keyword injection results suggest that adding more instructional or cultural cues does not necessarily produce additive gains, since their effectiveness depends on the model and on how they are integrated into the prompt.

This trend is further supported by the overall task results shared by the organizers. The best Aya configuration in our experiments achieves the second-highest CSP score among the reported systems, following only another approach that also uses Aya Expanse 8B. This suggests that Aya Expanse 8B may provide a strong basis for culturally grounded response generation in this evaluation setting.

5.2. Qualitative Results

The qualitative analysis of model outputs across languages, countries, and prompting strategies reveals a recurring tendency toward **superficial cultural adaptation**. By this, we mean that models often recognize when a prompt requires cultural adaptation, but their responses remain weakly grounded in the target culture. This behavior appears in different forms, including generic Western defaults, limited cultural variation across countries, stereotypical associations, vague cultural references, and asymmetric representations of non-Western cultures across languages. The analysis was conducted by the authors together with colleagues who are native speakers of, or proficient in, the evaluated languages. Table 3 summarizes the main qualitative patterns observed in the model outputs.

Cultural rigidity and generic Western defaults. We use cultural rigidity to describe cases where the model produces highly similar answers across countries or languages, even when cultural variation would be expected. This pattern is especially visible for Llama, where the target country often has little effect on the content of the response. For example, when asked about graduation attire in Italy and Turkey, the model gives near-identical recommendations, typically suggesting a suit and tie for men and a dress or formal skirt for women. In these cases, the model appears to perform occasion-based reasoning rather than country-specific cultural reasoning. That is, it recognizes the general event described in the prompt, but does not meaningfully adapt the answer to the cultural context. This is also visible when the model explicitly acknowledges the country in the response, but changes little beyond the country name. For instance, responses about wearing vibrant, patterned clothing at work in Germany and Greece are almost identical, with both answers referring to mixed reactions and expectations of a more conservative dress code.

Table 3
Qualitative patterns of superficial cultural adaptation observed in model outputs.

Pattern	Observed in	Description
Cultural rigidity	Both models	The model produces highly similar answers across countries, languages, or settings, even when cultural variation would be expected.
Generic Western default	Both models	The model defaults to Western or North American norms when no country is specified, even when the input language could imply a different cultural context.
Stereotypical cultural association	Mainly Aya	The model adapts to a target culture through outdated or stereotypical associations rather than contemporary cultural knowledge.
Invented cultural content	Both models	The model produces fabricated or non-existing cultural terms when trying to provide culturally specific advice.
Vague cultural references	Mainly Aya	The model uses culturally marked phrases without providing concrete, verifiable, or actionable cultural information.
Asymmetric cultural representation	Both models	The model describes a culture differently depending on the language used in the prompt, suggesting reliance on external representations of that culture.
Gender-norm reproduction	Both models	The model gives gender-binary recommendations and associates modesty concerns mainly with women's clothing.

A related pattern appears in the culture-unspecific setup, where no country is stated and the cultural context must be inferred from the input language. In this setting, both models frequently default to Western or North American norms. In several CoT outputs, this assumption is made explicit, with models stating that they assume a Western European cultural context, such as "westlichen, insbesondere europäischen Kultur", or framing Western norms as a universal default, as in "uniwersalny kontekst kulturowy". This pattern is particularly evident in the food task, where both models converge on finger foods, sandwiches, and salads, for example "σαλάτες, σάντουιτς" and "ensalada de queso y frutas", rather than culturally specific dinner options. These items are more characteristic of informal North American or Northern European entertaining than of a multi-course dinner. Critically, this behavior is not limited to English or other Western-language inputs. It also appears in Greek and Turkish prompts, suggesting that the models often rely on a generic Western frame rather than inferring culturally grounded expectations from the prompt language.

The same limited adaptation also appears across social variables such as age. For example, when asked what to bring as a gift when invited to someone's home for the first time, model outputs remain largely uniform across cultural contexts and age groups, consistently recommending wine, chocolate, or flowers. Although these suggestions may be reasonable in many Western settings, they do not capture culturally specific norms that may vary across countries and social contexts. For instance, locally specific gifts, such as baklava for Turkey, are not mentioned. Similarly, changing the age of the host does not substantially affect the recommendations, despite the fact that appropriate house gifts may depend on the social context of the host.

Stereotypical cultural associations and invented cultural content. A second pattern concerns cases where cultural adaptation becomes stereotypical or fabricated. This is especially visible for Aya, which sometimes adapts strongly to the target culture but relies on outdated or stereotypical associations rather than grounded contemporary knowledge. For example, when asked about graduation attire in Turkey, the model repeatedly recommends Ottoman-era clothing. This pattern appears across multiple languages, with outputs referring to turbans, such as "тюрбан", and kaftans, such as "traje tradicional turco (kaftán)" and "κεφτάν". Rather than providing culturally appropriate contemporary

advice, the model substitutes period-inappropriate cultural symbols for genuine cultural reasoning.

In some cases, this behavior extends to invented cultural content. For example, Russian and Polish prompts elicited food terms such as “нагретки” and “serowe tabliczki” , which do not appear to correspond to real dishes. These cases suggest that, when cultural grounding is weak, the model may not only rely on stereotypes but also generate arbitrary cultural associations. This poses a risk beyond cultural insensitivity, since fabricated cultural content can mislead users who are seeking genuine cultural advice.

Ungrounded cultural references. A further recurring pattern concerns ungrounded cultural references. By this, we refer to cases where models use culturally marked expressions, but do not provide concrete, verifiable, or actionable cultural information. For example, a German-language output suggests a graduation outfit that incorporates elements of a Turkish national dress, described as a “Variante des türkischen Nationalkleides”, without specifying what this dress looks like or whether it is appropriate for the occasion. Similarly, Italian graduation suggestions are described as having an “Italian flair” or involving “local craftsmanship”, without further explanation.

These phrases create the appearance of cultural awareness, while providing little concrete guidance. This pattern differs from generic advice, which makes no cultural reference, and from stereotyping, which makes a culturally specific but inaccurate reference. Ungrounded cultural references therefore occupy a middle ground: the culture is named or signaled, but the response does not explain what the cultural reference means or why it is relevant.

Asymmetric cultural representation. Another pattern concerns asymmetric cultural representation, where the same target culture is described differently depending on the language used in the prompt. This pattern is particularly visible in responses about Turkish graduation attire. When prompted in Turkish, both models produce largely Western-style suggestions, such as formal suits and dresses, with limited culturally distinctive content. However, when the same question is asked in German, Italian, Greek, or Spanish, the outputs shift noticeably. Responses more frequently refer to modesty, Islamic tradition, or the need to show respect and reverence, using expressions such as “nicht zu auffällig”, “ισλαμική παράδοση”, “με ευλάβεια”, and “respeto a las normas sociales”.

This asymmetry suggests that the models may represent the target culture through the cultural associations encoded in the prompt language. For instance, the emphasis on modesty and religious respect appears mainly when Turkey is discussed through non-Turkish prompts, rather than when the prompt is written in Turkish. To put it simple, when the question is asked in Greek about Turkey, the model may rely on Greek-language representations of Turkish culture rather than on Turkish cultural norms themselves. This indicates that the models may rely on external representations of Turkish culture in Western-language discourse, instead of grounded knowledge of Turkish cultural practices. Together with the stereotypical references to Ottoman-era clothing discussed above, this pattern shows how attempts at cultural adaptation can reproduce culturally biased representations when they are not supported by grounded cultural knowledge.

Gender-norm reproduction. A further pattern concerns the reproduction of conventional gender norms. Across languages and prompting strategies, both models frequently organize clothing recommendations along binary gender lines, assigning suits to men and dresses or skirts to women. This pattern is visible in graduation attire suggestions, where responses in multiple languages explicitly separate recommendations by gender, for example “un abito elegante per le donne” and “элегантное платье для женщин”, rather than providing more neutral or flexible advice.

When modesty is mentioned, it follows the same tendency, since it is directed primarily at women’s clothing. These outputs suggest that the models reproduce gendered social expectations encoded in their training data, alongside the broader tendency to reproduce external representations of cultures rather than grounded cultural knowledge.

5.3. Overall Findings

Taken together, the quantitative and qualitative results suggest that CSP is useful for comparing cultural signal behavior across systems, while qualitative analysis helps clarify how this adaptation is expressed in the generated outputs. This is particularly visible for Aya Expanse 8B. Quantitatively, Aya benefits strongly from all prompting strategies, with the persona prompt achieving the highest score among our submitted configurations. This suggests that Aya is highly responsive to explicit culture-related instructions, and that such prompts help the model better separate language-based cultural signals from explicitly provided cultural grounding.

However, the qualitative analysis provides a more nuanced interpretation of this result. Although the persona prompt improves Aya's CSP score, Aya also shows several forms of superficial cultural adaptation. In particular, the model sometimes adapts strongly to the target culture, but does so through stereotypical, vague, or weakly grounded cultural associations. Examples include Ottoman-era clothing suggestions for Turkish graduation attire, ungrounded references such as "Italian flair", and cases where culturally specific content becomes fabricated. This suggests that the persona prompt may increase cultural responsiveness, but not necessarily guarantee culturally accurate or contextually grounded responses.

Overall, our findings suggest that prompt design can substantially improve culturally adaptive behavior, especially for multilingual models such as Aya. At the same time, qualitative analysis remains necessary to identify whether this adaptation is grounded in reliable cultural knowledge or whether it reflects superficial cultural cues, stereotypes, or language-specific cultural biases. Future evaluations should therefore combine quantitative measures such as CSP with human-centered qualitative analysis, especially for culturally sensitive generation tasks.

6. Conclusions

This work shows that culturally adaptive generation remains a challenging problem for multilingual LLMs. Across our experiments, models often default to Western cultural norms when no explicit cultural context is provided, even when the input language could suggest a different cultural background. This suggests that cultural inference from language alone remains limited, and that language-based cultural signals are not always sufficient for producing genuinely culture-specific responses. Our qualitative analysis also shows that explicit cultural prompting does not always lead to grounded cultural adaptation. In culturally specific settings, models may produce historically inaccurate, stereotypical, or fabricated suggestions, especially when they lack reliable cultural knowledge. This creates a risk of misinformation, since users seeking cultural advice may receive responses that appear culturally specific but are not accurate or appropriate. Finally, our findings suggest that models may not represent cultures directly, but rather reproduce how cultures are described across different language contexts, especially the one used to create the prompt. As a result, the same target culture can be represented differently depending on the language of the prompt. This has important implications for cultural fairness, since attempts to make models more culturally inclusive may still reinforce the stereotypes and assumptions encoded in their training data.

Overall, our results highlight the need to evaluate culturally adaptive generation not only through aggregate scores, but also through qualitative analysis of the actual cultural content produced by the models. Future work should further investigate culturally grounded prompting, multilingual evaluation protocols, and mechanisms for reducing stereotypical or weakly grounded cultural adaptation.

Acknowledgments

We acknowledge the CINECA award under the ISCRA initiative, for the availability of high-performance computing resources and support. This paper was supported by the National Plan for NRRP Complementary Investments (PNC, established with the decree-law 6 May 2021, n. 59, con-

verted by law n. 101 of 2021) in the call for the funding of research initiatives for technologies and innovative trajectories in the health and care sectors (Directorial Decree n. 931 of 06-06-2022) - project n. PNC0000003 - AdvaNced Technologies for Human-centrEd Medicine (project acronym: ANTHEM). This work reflects only the authors' views and opinions, neither the Ministry for University and Research nor the European Commission can be considered responsible for them.

Declaration on Generative AI

During the preparation of this work, the authors used ChatGPT, Claude, and Grammarly in order to: paraphrase and reword, grammar and spelling check. Further, the authors used GPT-5.5 for Figure 1 in order to: Generate a new image based on a draft image we created using draw.io, with the goal of improving content alignment and overall presentation. After using these tools, the authors reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] T. Larina, Culture-specific communicative styles as a framework for interpreting linguistic and cultural idiosyncrasies, International Review of Pragmatics 7 (2015) 195–215.

[2] M. Liu, Verbal communication styles and culture, in: M. Powers (Ed.), Oxford Research Encyclopedia of Communication, Oxford University Press, 2016. doi:10.1093/acrefore/9780190228613.013.162.

[3] UNESCO, Culture, 2026. URL: https://www.unesco.org/en/query-list/c/423986#:~:text=Culture%20refers%20to%20the%20collective,another20and%20perceive%20the%20worl.

[4] B. AlKhamissi, M. N. ElNokrashy, M. Alkhamissi, M. T. Diab, Investigating cultural alignment of large language models, in: L. Ku, A. Martins, V. Srikumar (Eds.), Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024, Association for Computational Linguistics, 2024, pp. 12404–12422. doi:10.18653/V1/2024.ACL-LONG.671.

[5] R. K. McBain, R. Bozick, M. Diliberti, L. A. Zhang, F. Zhang, A. Burnett, A. Kofner, B. Rader, J. Breslau, B. D. Stein, A. Mehrotra, L. U. Pines, J. Cantor, H. Yu, Use of generative ai for mental health advice among us adolescents and young adults, JAMA Network Open 8 (2025) e2542281–e2542281. doi:10.1001/jamanetworkopen.2025.42281.

[6] K. Dokic, B. Pisker, B. Radisic, Mirroring cultural dominance: Disclosing large language models social values, attitudes and stereotypes, Societies 15 (2025). doi:10.3390/soc15050142.

[7] E. Sokli, G. Peikos, P. Kasela, G. Pasi, Leveraging cognitive complexity of texts for contextualization in dense retrieval, in: Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Suzhou, China, 2025, pp. 27083–27096. URL: https://aclanthology.org/2025.emnlp-main.1377/. doi:10.18653/v1/2025.emnlp-main.1377.

[8] T. Safavi, D. Koutra, Relational world knowledge representation in contextual language models: A review, in: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, 2021, pp. 1053–1067.

[9] W. Sleegers, J. Elsey, D. Moss, Adoption and Uses of LLMs among U.S. Tech Workers, Technical Report, Rethink Priorities, 2025. URL: https://rethinkpriorities.org/research-area/adoption-llms-tech-workers/.

[10] Z. Liao, M. Antoniak, I. Cheong, E. Y.-Y. Cheng, A.-H. Lee, K. Lo, J. C. Chang, A. X. Zhang, Llms as research tools: A large scale survey of researchers' usage and perceptions, arXiv preprint arXiv:2411.05025 (2024).

[11] J. Wang, W. Ma, P. Sun, M. Zhang, J. Nie, Understanding user experience in large language model interactions, CoRR abs/2401.08329 (2024). doi:10.48550/ARXIV.2401.08329. arXiv:2401.08329.

[12] S. Krügel, A. Ostermaier, M. Uhl, Chatgpt's inconsistent moral advice influences users' judgment, Scientific Reports 13 (2023) 4569.

[13] S. Ikeda, Inconsistent advice by chatgpt influences decision making in various areas, Scientific Reports 14 (2024) 15876.

[14] S. Noels, A. Rogiers, M. Buyl, T. D. Bie, Persuasion with large language models: A survey of empirical evidence, study methodologies, and ethical implications, 2026. URL: https://arxiv.org/abs/2411.06837. arXiv:2411.06837.

[15] R. Masoud, Z. Liu, M. Ferianc, P. C. Treleaven, M. R. Rodrigues, Cultural alignment in large language models: An explanatory analysis based on hofstede's cultural dimensions, in: Proceedings of the 31st International Conference on Computational Linguistics, 2025, pp. 8474–8503.

[16] I. Ahmad, S. Dudy, R. Ramachandranpillai, K. Church, Are generative language models multicultural? a study on hausa culture and emotions using chatgpt, in: Proceedings of the 2nd Workshop on Cross-Cultural Considerations in NLP, 2024, pp. 98–106.

[17] S. Dudy, I. S. Ahmad, R. Kitajima, A. Lapedriza, Analyzing cultural representations of emotions in llms through mixed emotion survey, in: 2024 12th International Conference on Affective Computing and Intelligent Interaction (ACII), IEEE, 2024, pp. 346–354.

[18] 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.

[19] 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.

[20] Y. Tao, O. Viberg, R. S. Baker, R. F. Kizilcec, Cultural bias and cultural alignment of large language models, PNAS Nexus 3 (2024) pgae346. doi:10.1093/pnasnexus/pgae346.

[21] 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 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024, 2024.

[22] J. Wu, S. Chen, P. Chen, W. Du, H. Gao, L. Wang, P. Hansen, A resource-efficient framework for cultural alignment in large language models (llms): The chinese context, Design and Artificial Intelligence 2 (2026) 100063. doi:https://doi.org/10.1016/j.daai.2026.100063.

[23] 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: S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, A. Oh (Eds.), Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022, 2022.

[24] Meta AI, Llama-3.1-8b-instruct, https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct, 2024. Hugging Face model repository, accessed 2026-05-25.

[25] J. Dang, S. Singh, D. D'souza, A. Ahmadian, A. Salamanca, M. Smith, A. Peppin, S. Hong, M. Govindassamy, T. Zhao, S. Kublik, M. Amer, V. Aryabumi, J. A. Campos, Y.-C. Tan, T. Kocmi, F. Strub, N. Grinsztajn, Y. Flet-Berliac, A. Locatelli, H. Lin, D. Talupuru, B. Venkitesh, D. Cairuz, B. Yang, T. Chung, W.-Y. Ko, S. S. Shi, A. Shukayev, S. Bae, A. Piktus, R. Castagné, F. Cruz-Salinas, E. Kim, L. Crawhall-Stein, A. Morisot, S. Roy, P. Blunsom, I. Zhang, A. Gomez, N. Frosst, M. Fadaee, B. Ermis, A. Üstün, S. Hooker, Aya expanse: Combining research breakthroughs for a new multilingual frontier, 2024. URL: https://arxiv.org/abs/2412.04261. arXiv:2412.04261.

[26] C. C. Liu, A. Korhonen, I. Gurevych, Cultural learning-based culture adaptation of language models, in: W. Che, J. Nabende, E. Shutova, M. T. Pilehvar (Eds.), Proceedings of the 63rd Annual

Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2025, Vienna, Austria, July 27 - August 1, 2025, Association for Computational Linguistics, 2025, pp. 3114–3134.

[27] C. Li, D. Teney, L. Yang, Q. Wen, X. Xie, J. Wang, Culturepark: Boosting cross-cultural understanding in large language models, in: A. Globersons, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. M. Tomczak, C. Zhang (Eds.), Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024, 2024.

[28] Y. Elazar, N. Kassner, S. Ravfogel, A. Ravichander, E. H. Hovy, H. Schütze, Y. Goldberg, Measuring and improving consistency in pretrained language models, CoRR abs/2102.01017 (2021). arXiv:2102.01017.

[29] R. Feng, S. Gao, X. Chen, L. Chen, S. Shang, Culfit: A fine-grained cultural-aware LLM training paradigm via multilingual critique data synthesis, in: Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2025, Vienna, Austria, July 27 - August 1, 2025, Association for Computational Linguistics, 2025, pp. 22413–22430.

[30] C. Olea, H. Tucker, J. Phelan, C. Pattison, S. Zhang, M. Lieb, D. Schmidt, J. White, Evaluating persona prompting for question answering tasks, in: Proceedings of th e 10th international conference on artificial intelligence and soft computing, Sydney, Australia, 2024.

[31] B. N. Sotic, J. Kamps, University of amsterdam at the CLEF 2025 eloquent track, in: G. Faggioli, N. Ferro, P. Rosso, D. Spina (Eds.), Working Notes of the Conference and Labs of the Evaluation Forum, CLEF 2025, Madrid, Spain, 9-12 September 2025, CEUR Workshop Proceedings, CEUR-WS.org, 2025, pp. 1435–1442.

[32] M. Zheng, J. Pei, L. Logeswaran, M. Lee, D. Jurgens, When "a helpful assistant" is not really helpful: Personas in system prompts do not improve performances of large language models, in: Findings of the Association for Computational Linguistics: EMNLP 2024, Miami, Florida, USA, November 12-16, 2024, Findings of ACL, Association for Computational Linguistics, 2024, pp. 15126–15154. doi:10.18653/V1/2024.FINDINGS-EMNLP.888.

[33] Z. Hu, C. Liu, X. Feng, Y. Zhao, S. Ng, A. T. Luu, J. He, P. W. Koh, B. Hooi, Uncertainty of thoughts: Uncertainty-aware planning enhances information seeking in large language models, CoRR abs/2402.03271 (2024). doi:10.48550/ARXIV.2402.03271. arXiv:2402.03271.

[34] Z. Yin, Q. Sun, Q. Guo, Z. Zeng, X. Li, J. Dai, Q. Cheng, X. Huang, X. Qiu, Reasoning in flux: Enhancing large language models reasoning through uncertainty-aware adaptive guidance, in: Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024, Association for Computational Linguistics, 2024, pp. 2401–2416. doi:10.18653/V1/2024.ACL-LONG.131.

[35] M. Grootendorst, Keybert: Minimal keyword extraction with bert., 2020. URL: https://doi.org/10.5281/zenodo.4461265. doi:10.5281/zenodo.4461265.

[36] N. Reimers, I. Gurevych, Sentence-bert: Sentence embeddings using siamese bert-networks, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, 2019. URL: http://arxiv.org/abs/1908.10084.

[37] P. Dey, A. Bothra, Y. Khanter, J. Zhao, E. Ferrara, Can llms express personality across cultures? introducing culturalpersonas for evaluating trait alignment, in: Findings of the Association for Computational Linguistics: EMNLP 2025, Suzhou, China, November 4-9, 2025, Association for Computational Linguistics, 2025, pp. 20241–20262.

[38] Z. Peng, X. Liu, G. Yang, J. Liu, X. Peng, Y. Long, The uncertainty advantage: Enhancing large language models' reliability through chain of uncertainty reasoning, Pattern Recognit. Lett. 200 (2026) 30–36. URL: https://doi.org/10.1016/j.patrec.2025.11.040. doi:10.1016/J.PATREC.2025.11.040.

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