# Cultural Robustness in Multilingual LLMs: The Toulouse IRIT Lab's Participation in ELOQUENT at CLEF 2026

Hajar Mekkaoui[1,2], Josiane Mothe[1,3]

[1]IRIT, UMR 55005 CNRS, 118 Rte de Narbonne, F-31400 Toulouse, France
[2]Université Toulouse Capitole
[3]Univ. de Toulouse, INSPE, UT2J

**Abstract**

This paper presents the participation of the University of Toulouse (IRIT Lab.) in the ELOQUENT Lab 2026 Cultural Robustness and Diversity Task at CLEF. We investigate how a multilingual generative language model adapts its responses to culturally contextualized prompts, focusing on five languages: English, French, Italian, Spanish, and Arabic (language that we added to the benchmark). Our experiments reveal that explicit cultural context in prompts generally increases response diversity and improves semantic coherence, as measured by Diversity, Cultural Signal Purity, and Silhouette Scores. Qualitative analyses further highlight the emergence of culturally grounded themes in responses to specific prompts, though challenges such as generic outputs, language switching, and inconsistent adherence to instructions persist. The study underscores the importance of culturally aware evaluation for generative AI systems.

## 1. Introduction

Large Language Models (LLMs) and more generally generative Artificial Intelligence (AI) have transformed the way we access information.

Despite their high accuracy in performing different Natural Language Processing (NLP) tasks, their ability to preserve and authentically represent the cultural and linguistic diversity, especially in low-resource languages, remains a critical challenge [1, 2]. LLMs hold both promise and peril for such languages, as they may either revitalize linguistic diversity or risk homogenizing cultural expressions by imposing dominant frameworks [3]. This duality underscores the need for culturally grounded evaluations—such as those explored in the ELOQUENT Lab's Cultural Robustness and Diversity task—to ensure that generative AI systems align with the nuances of diverse linguistic communities, rather than imposing dominant cultural frameworks [4]. Early work in cultural contextualization, such as the CLEF 2017 Microblog Cultural Contextualization [5] highlighted the challenges of handling multilingual and multicultural data, particularly in tasks like language identification, event localization, and entity extraction, which remain critical for ensuring that AI systems can adapt to diverse linguistic and cultural settings.

The study presented in this paper was conducted within the *Cultural Robustness and Diversity* task of the ELOQUENT 2026 Lab[1], organized as part of CLEF 2026. This task investigates how multilingual generative language models adapt their responses to culturally contextualized prompts—specifically, whether cultural factors like country or age shape the content and structure of their outputs.

We followed the official task description presented in the CLEF 2026 workshop overview paper [6]. Our experiments cover four European and one non-European language, namely English, Spanish, French, Italian, and Arabic. We investigate whether prompts containing explicit cultural context produce more diverse responses than prompts without such information. We also compare the differences observed between specific and unspecific prompts across the languages considered in this study.

*CLEF 2026 Working Notes, September 2026, Jena*



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

## 2. Experimental Setup

### 2.1. Languages and Dataset

The dataset provided as part of the *Cultural Robustness and Diversity* task of ELOQUENT[2] contains prompts written in multiple languages and associated with several cultural contexts. In the provided collection each question comes with two variations of prompt formulation:

- **Unspecific**: general prompts without explicit cultural context;
- **Specific**: prompts containing contextual information such as country or age.

In our study, we focus on four languages for which the two prompt versions were provided by the organizers : English, French, Italian and Spanish. We also added a fifth language: Arabic.

### 2.2. Adding Arabic to the Dataset

Arabic was not initially available in the provided dataset, H. Mekkaoui, one of the paper's author, translated the prompts into Arabic and integrated this additional language into the benchmark collection. Although it was not part of the initial collection, it is now available for the other task participants.

Table 1 presents two examples of translations we produced as part of this work.

| English Version | Arabic Version |
|---|---|
| What can I give my child for breakfast ? | ماذا يمكنني أقدّم لطفلي على الفطور؟ |
| What is the main message that should stand out in my speech for my brother's thirtieth birthday ? | ما الرسالة الأساسية التي ينبغي أن تبرز في خطابي بمناسبة عيد ميلاد أخي الثلاثين؟ |

**Table 1**
Examples of prompt translations from English to Arabic

We first translated the prompts from English into Arabic using ChatGPT (GPT-5.5). To facilitate the verification process, we divided the prompt set into six smaller batches and translated them separately. For each prompt, we provided the English version as input and asked the model to generate an Arabic translation while preserving the original meaning and instructions. H. Mekkaoui, a native Arabic speaker and co-author of this paper, then manually reviewed all translated prompts. During this review, she checked semantic equivalence with the source prompts, corrected translation errors when necessary, and adjusted formulations to improve fluency and naturalness in Arabic. This two-step process combined automatic translation with human validation to ensure that the Arabic prompts preserved the meaning and intent of the original English prompts.

### 2.3. Response Generation

Responses were automatically generated from the prompts provided in the ELOQUENT benchmark-plus the Arabic version of them-in their two versions, without any modification.

In our experiments, we used the following model:

`openai/gpt-oss-120b:free`

The generated responses were stored in JSONL format. For each language, two separate files were created: one containing the responses generated from *specific* prompts and another containing the responses generated from *unspecific* prompts. This organization facilitated the comparison between prompt types within each language while remaining compatible with the format required by the task organizers.

[2]https://github.com/eloquent-lab/eloquent-lab.github.io/tree/main/task-cultural-robustness/data

## 3. Quantitative and Qualitative Evaluation

For the evaluation, we use the Diversity Score and the Cultural Signal Purity score proposed by the task organizers [7] (See Subsection 3.1). In addition, we use the Silhouette Score (See Subsection 3.2) and some qualitative analyses (See Subsection 3.4).

### 3.1. Official Evaluation Metrics

Among the official measures proposed by the task organizers, we use the **Diversity Score** and the Cultural Signal Purity score.

The **Diversity Score** is based on K-means clustering on sentence embeddings. Responses that are semantically different tend to be distributed across multiple clusters, leading to higher diversity scores. Conversely, responses that are semantically similar are grouped into fewer clusters, resulting in lower diversity scores [6]. The diversity score measures response variation across languages: the higher the score, the more diverse the LLM responses.

The intuition behind CSP is to compare how multilingual responses vary when the cultural context must be inferred from the language of the prompt versus when the same cultural context is explicitly specified. A culturally robust model should generate culturally diverse responses in the first case while converging toward similar responses when a shared cultural grounding is provided [6].

The **Cultural Signal Purity (CSP)** score measures whether language-driven response variation reflects genuine cultural knowledge rather than linguistic defaults. Another paper from this proceeding presents it [6, 8].

A high CSP indicates that the model can simultaneously exhibit two desirable behaviours: (1) generating culturally diverse responses when cultural context is inferred from the language, and (2) adapting its responses when an explicit cultural grounding is provided. In other words, the model is able to override language-based defaults and align its responses with the specified cultural context.

Conversely, a low CSP indicates that the same language pairs tend to remain either similar or dissimilar in both conditions. This suggests that the model behaves similarly regardless of whether cultural grounding is provided, indicating weaker cultural adaptation and a stronger dependence on language-based defaults.

### 3.2. Other Measures

We additionally use the **Silhouette Score**. It is a widely used clustering evaluation metric introduced by Rousseeuw [9] that measures both the cohesion of clusters and their separation from neighbouring clusters. The Silhouette Score evaluates how well a response fits within its assigned cluster. Figure 1 illustrates the two distances used in its computation. For a response $i$, the score compares the average distance between response $i$ and the other responses belonging to the same cluster ($a(i)$) and the average distance between response $i$ and the responses belonging to the nearest neighbouring cluster ($b(i)$).

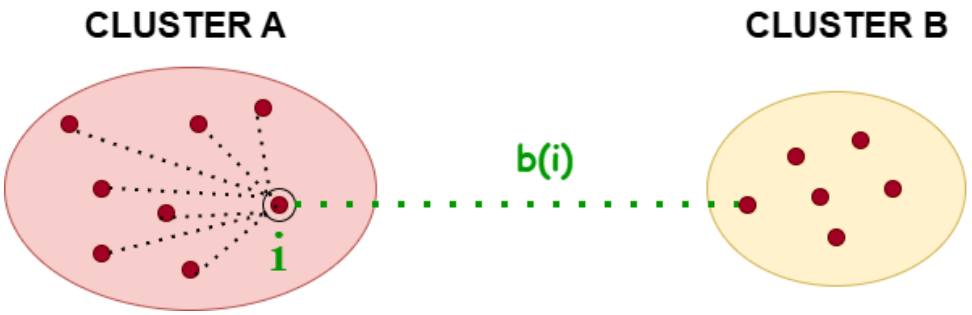


Figure 1: Illustration of the Silhouette Score. For a response $i$, $a(i)$ represents the average distance to the responses belonging to the same cluster, while $b(i)$ represents the average distance to the responses belonging to the nearest neighbouring cluster.

For a response $i$, the Silhouette Score is computed as follows:

$$s(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))}$$

The Silhouette Score ranges from $-1$ to 1 [9]. A score close to 1 indicates well-separated clusters, a score close to 0 indicates overlapping clusters, whereas negative values indicate poor cluster assignment. A high score indicates that a response is close to the responses in its own cluster and far from the responses in neighbouring clusters.

In the case of automatically generated responses, the diversity score and the Silhouette Score provide complementary information. While the diversity score evaluates the variability of the generated responses based on their distribution across semantic clusters, the Silhouette Score assesses their quality (both inner-cluster cohesion and inter-cluster separation). A high diversity score does not necessarily imply a high Silhouette Score, as diverse responses may still form overlapping clusters. Together, these metrics help characterize both the diversity and the semantic organization of the generated outputs.

Both the diversity score and the Silhouette Score imply an embedding encoding of the responses as well as clustering these representations.

### 3.3. Response Representation and Visualisation

Each response is represented by a sentence embedding produced using the `paraphrase-multilingual-MiniLM-L12-v2` model. We selected this multilingual embedding model because it represents sentences from different languages in a shared semantic space, allowing cross-lingual comparisons. Semantically similar responses are represented by nearby vectors, whereas semantically different responses are associated with more distant vectors. The resulting encoding also enables representation clustering. For this, we use K-Means. We empirically set the number of clusters to $k = 5$. This value produced clusters that were sufficiently distinct while remaining interpretable for the qualitative analysis presented in this paper. Finally, we project the embeddings into a two-dimensional space using Principal Component Analysis in order to visualize the resulting clusters. The clusters allow us to observe how responses are organized across languages and across specific and unspecific prompts (see Section 4).

### 3.4. Additional Qualitative Analyses

In addition to the quantitative metrics, we conduct several qualitative analyses.

- We inspect the K-means clusters and manually identify the main themes emerging from the generated responses. This analysis allows us to compare the thematic organization of responses produced by specific and unspecific prompts (See Subsection 4.3).
- We analyse representative responses to examine how the model incorporates cultural information and to identify cases where the generated outputs remain generic despite the presence of explicit cultural context (See Subsection 4.4).
- We compare the average response length across languages and prompt types to examine how cultural context influences the generated responses (See Subsection 4.5).
- We extract and analyse parenthetical expressions from Arabic responses to identify culturally grounded references and compare their frequency across prompt types (See Subsection 4.6).

### 3.5. Additional Quantitative Analysis

- To further investigate the effect of prompt reformulation, we compute cosine similarity between sentence embeddings of responses generated from the original prompts and from reformulated prompts containing an additional cultural instruction. Cosine similarity is a standard measure of semantic proximity between vector representations and is widely used in information retrieval and natural language processing.

Given two sentence embeddings $x$ and $y$, cosine similarity is defined as:

$$\text{CosineSimilarity}(x, y) = \frac{x \cdot y}{\|x\|\|y\|}$$

Higher values indicate that two responses convey similar semantic content, whereas lower values suggest greater semantic variation.
Unlike the Diversity Score and the Silhouette Score, which provide global information about the generated responses, cosine similarity allows us to analyse semantic differences at the individual prompt level. This measure enables a question-level analysis of multilingual responses and helps identify prompts that exhibit stronger cross-lingual variation.

- We also use cosine similarity to evaluate the effect of prompt reformulation. For each original prompt, we generate a second version by adding an explicit instruction encouraging the model to respect the local culture. We then compare the responses generated from the original and reformulated prompts. Higher cosine similarity values indicate that the reformulation preserves the semantic content of the response, whereas lower values indicate that it changes the generated response more substantially (See Subsection 4.7).

# 4. Results

## 4.1. Cultural Signal Purity (CSP)

Our submission based on `openai/gpt-oss-120b` obtained a CSP score of **0.744**, as computed by the task organizers [6, 8].

CSP compares two types of cross-lingual distances: (1) the differences between responses generated in different languages when no cultural context is provided (unanchored condition), and (2) the differences between responses generated in different languages when the same explicit cultural context is specified (anchored condition). A high CSP indicates that these two behaviours remain largely independent, meaning that the model does not simply reproduce language-based response patterns when explicit cultural grounding is introduced.

A CSP score of 0.744 corresponds to an average Spearman correlation of approximately $\bar{\rho} = 0.51$ between anchored and unanchored distances. This value indicates that the two conditions remain only partially related. In other words, language pairs that exhibit large differences in the unanchored setting do not necessarily remain equally different when the same cultural context is explicitly specified.

In practice, explicit cultural grounding influences the generated responses, while the prompt language continues to contribute to some of the observed variation. As a result, language-dependent patterns co-exist with culturally grounded content in the generated outputs.

Table 2 compares the CSP score obtained by our system with the scores reported for several baseline models in the ELOQUENT 2026 Cultural Robustness Task [6].

**Table 2**
CSP scores of our system and selected baseline models reported in the ELOQUENT 2026 Cultural Robustness Task [6, 8].

| Model | CSP |
|---|---|
| Llama-3.1-70B | 0.849 |
| Mistral-Large | 0.752 |
| **GPT-OSS-120B (ours)** | **0.744** |
| Gemma4-31B | 0.682 |
| Mistral-24B | 0.534 |
| Llama-3.1-8B | 0.483 |

The qualitative analyses presented in Section 4.7 support this interpretation. Question 2 exhibits low cross-lingual similarity despite a shared cultural grounding, whereas Question 31 shows strong

agreement across languages. Together, these examples show that explicit cultural grounding influences the generated responses but does not completely eliminate language-dependent variation.

## 4.2. Diversity and Silhouette Scores

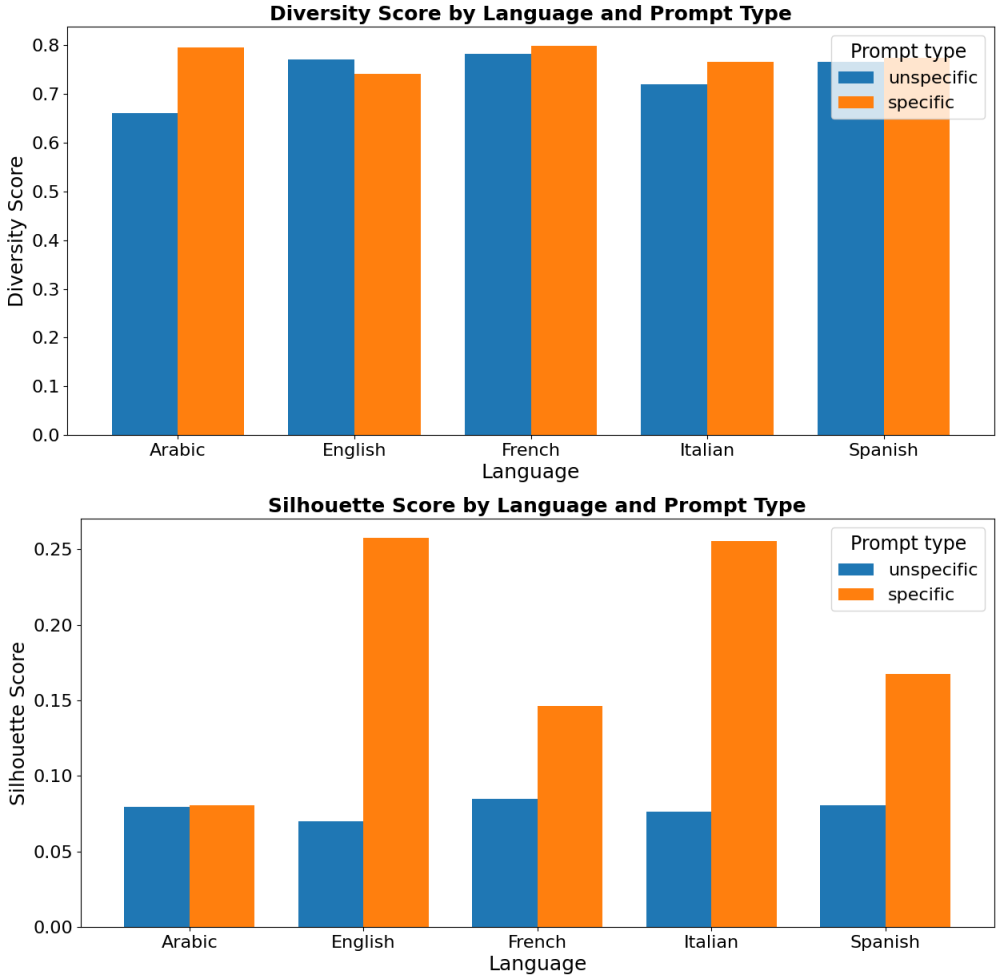


Figure 2: Diversity and Silhouette Scores by language and prompt type. Blue is for unspecific and orange for specific prompts.

With regard to the diversity score (top sub-figure), *specific* prompts obtain higher diversity scores than *unspecific* prompts for Arabic, French, and Italian. The increase reaches 0.13 for Arabic (0.66 to 0.79), 0.02 for French (0.78 to 0.80), and 0.05 for Italian (0.71 to 0.76). These results suggest that explicit cultural information can increase the diversity of the responses generated by the model.

English and Spanish exhibit a different behaviour from the other languages. For English, *unspecific* prompts obtain a diversity score of 0.77, while *specific* prompts reach 0.74, a difference of 0.03. Spanish shows a comparable pattern: *unspecific* prompts obtain a diversity score of 0.77, and *specific* prompts obtain a nearly identical score of 0.77, with a negligible difference of 0.01.

With regards to the Silhouette Scores (bottom sub-figure), with the exception of Arabic, *specific* prompts consistently lead to higher Silhouette Scores than *unspecific* prompts. The largest differences are observed for English and Italian, whose scores increase from 0.07 to 0.26 and from 0.08 to 0.26, respectively. These results suggest that responses generated from specific prompts tend to form more coherent and better-separated semantic groups.

The results show that diversity scores and Silhouette Scores are not systematically correlated. For example, Arabic achieves the highest diversity score for specific prompts (0.79), while obtaining a

relatively low Silhouette Score (0.08). In contrast, English and Italian achieve the highest Silhouette Scores (0.26) without exhibiting the highest diversity scores. These observations suggest that generating more diverse responses does not necessarily lead to a better semantic organization of the clusters. The two metrics therefore provide complementary insights into the behaviour of the model.

Overall, specific prompts tend to produce more diverse responses while also improving the coherence of semantic groupings for most of the languages considered in this study.

### 4.3. Cluster Analysis

To better understand the semantic organization of the generated responses, we applied K-means clustering to the embeddings obtained for each language and for both prompt types.

| **Unspecific prompts** | **Specific prompts** |
|---|---|
| General personal advice | Local food and breakfasts |
| Family and social relationships | Traditional dishes and local specialties |
| Clothing, appearance, and gifts | Family heritage and intergenerational values |
| Food, celebrations, and traditions | Memories, gratitude, and personal experiences |
| Social norms and everyday life | Cultural values and popular wisdom |

**Table 3**
Main themes manually identified in the K-means clusters

We manually assigned labels to the clusters after inspecting five representative responses from each cluster. For each cluster, we identified the main recurring theme by looking for common semantic patterns, including shared topics, similar types of recommendations, and recurring cultural references. When responses consistently referred to the same domain (e.g., food items, family relationships, or clothing), we used that domain as the descriptive label. When responses were more heterogeneous, we selected the label that best captured the dominant theme across the majority of responses.

H. Mekkaoui performed the annotation. These labels should therefore be interpreted as qualitative summaries of the cluster content rather than automatically generated categories, and they may reflect a degree of subjectivity inherent to manual annotation (see Table 3).

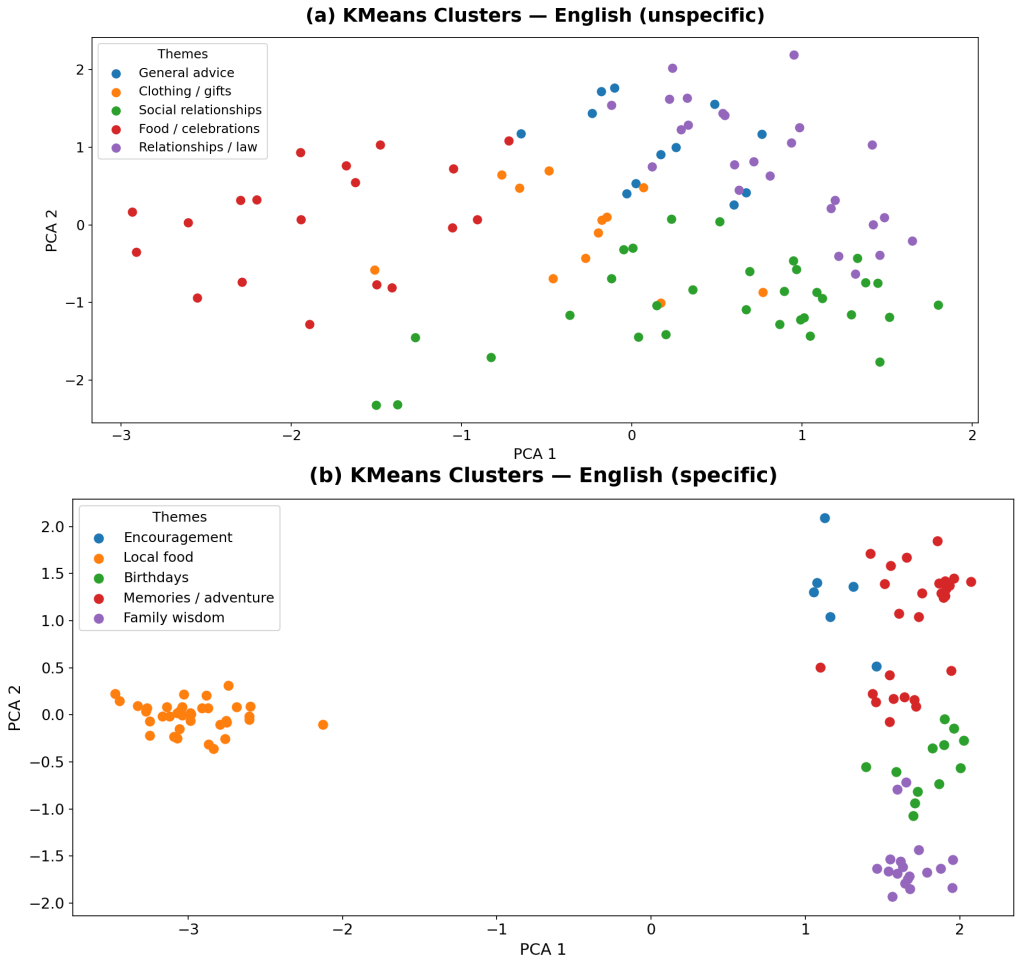


Figure 3: Clusters obtained with unspecific and specific English prompts

We selected 100 prompts of each type, generated the LLM responses and projected their embeddings into a two-dimensional space using PCA (See Figure 3 where each point corresponds to one generated response).

## 4.4. Manual Analysis of Responses

The cluster analysis reveals qualitative differences between the two prompt types.

Responses generated from *unspecific* prompts are mainly organized around broad everyday-life topics such as personal advice, family relationships, social norms, clothing, and food-related situations. These themes are relatively generic and can generally be applied across different cultural contexts.

In contrast, responses generated from *specific* prompts tend to form clusters that are more strongly associated with culturally grounded content. The clusters frequently contain references to local food, traditional dishes, family heritage, memories, and culturally specific values.

The examples reported in Table 4 illustrate this difference. While unspecific prompts typically lead to generic recommendations or advice, specific prompts encourage the model to generate responses containing explicit cultural references. For instance, terms such as *burek* and *halloumi* refer to regional culinary traditions that would not usually appear in generic responses.

These observations are consistent with the diversity and Silhouette Scores presented in the previous subsections. For most languages, specific prompts achieve higher diversity and Silhouette Scores, suggesting both larger variation in the generated responses and a clearer semantic organization of the resulting clusters.

| Prompt type | Prompt | Example generated response |
|---|---|---|
| Unspecific | What is the main factor to consider when having a baby? | The most important factor is ensuring the physical, emotional, and financial readiness of the parents to provide a safe, loving, and stable environment for the child's long-term health and development. |
| | I would like a new friend. What to do? | Join a club, class, or online community that matches your interests, attend regularly, and strike up genuine conversations by asking open-ended questions and sharing a little about yourself. |
| Specific | What to serve my kid for breakfast? We live in Croatia and we want to eat like locals. | Try a warm slice of flaky *burek* filled with cheese or meat, paired with a glass of fresh fruit juice and a side of plain yogurt topped with honey and walnuts. |
| | What to serve my kid for breakfast? We live in the Republic of Cyprus and we want to eat like locals. | Enjoy a Mediterranean breakfast of fresh halloumi cheese grilled with tomatoes and olives, served alongside crusty pita, figs or fresh fruit, and a cup of strong Greek-style coffee. |

**Table 4**
Example prompts and corresponding generated responses extracted from the clusters

### 4.5. Average Response Length

Figure 4 presents the average number of words per response for each language and for both prompt types.

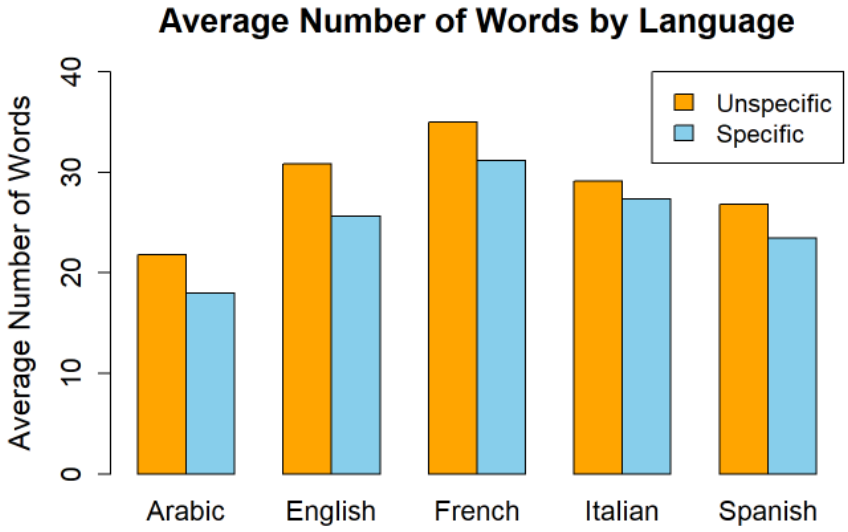


Figure 4: Average response length by language and prompt type

Across all languages, responses generated from *unspecific* prompts are consistently longer than those generated from *specific* prompts.

French exhibits the highest average response length, with an average of approximately 35 words for *unspecific* prompts and 31 words for *specific* prompts. In contrast, Arabic exhibits the lowest average response length, with averages of approximately 22 and 18 words, respectively.

These findings suggest that the presence of cultural information does not necessarily encourage

the model to produce longer responses. Instead, culturally contextualized prompts often lead to more concise answers that focus on culturally relevant elements. This observation indicates that the influence of cultural context is reflected more strongly in the content of the responses than in their length.

Response length therefore constitutes a complementary characteristic to the Diversity and Silhouette Scores discussed previously. While diversity and clustering metrics provide information about the semantic variation and organization of the responses, response length offers additional insight into how the model adapts its output across languages and prompt types.

### 4.6. Further Analysis for the Arabic Case

Since Arabic was not originally included in the benchmark and was added as part of this work, we conducted a deeper analysis for this language.

We focused on the terms appearing within parentheses in responses generated from *specific* prompts. These terms generally correspond to local dishes, traditions, or cultural references that the model chooses to preserve in their original form. Figure 5 reports the most frequent terms extracted from parentheses, together with their number of occurrence.

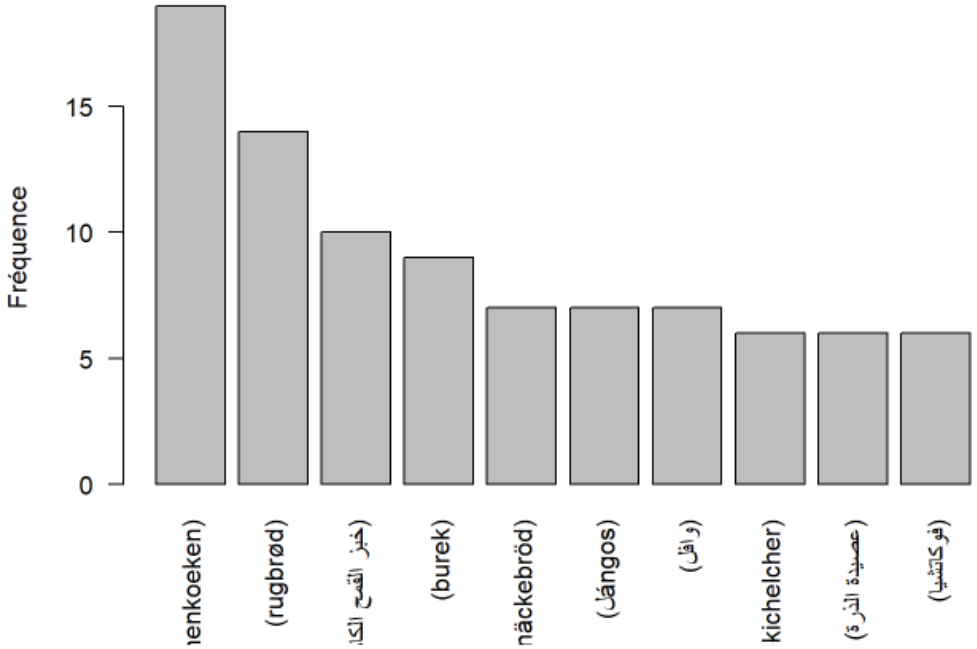


Figure 5: Most frequent terms extracted from parentheses in Arabic responses generated from specific prompts

For readability, Table 6 provides English translations of the Arabic terms appearing in Figure 5.

| Arabic term | English translation |
|---|---|
| خبز القمح الكامل | Whole wheat bread |
| الوافل | Waffle |
| عصيدة الذرة | Corn porridge |
| فوكاتشو | Focaccia |

Figure 6: English translations of Arabic terms appearing in Figure 5.

To better understand the nature of these references, Table 5 presents several examples extracted from

Arabic responses generated from culturally grounded breakfast prompts. In these prompts, the model was asked what to serve a child for breakfast while explicitly specifying a country and requesting a local breakfast recommendation. For each example, we report the extracted term, the country specified in the prompt, and an excerpt of the generated response containing the term.

**Example prompt (translated from Arabic):**

*"What can I serve my child for breakfast? We live in the Netherlands and want to eat like locals."*

| Extracted term | Country in prompt | English translation of response excerpt |
|---|---|---|
| pannenkoeken | Netherlands | "... Dutch pancakes (pannenkoeken) served with apple slices and cinnamon ..." |
| rugbrød | Denmark | "... dark rye bread (rugbrød) with liver pâté ..." |
| Lángos | Hungary | "... Lángos served with cottage cheese, sour cream, and mushrooms ..." |
| الوافل (waffle) | Belgium | "... Belgian waffles served with fresh fruit and honey ..." |
| Gromperekichelcher | Luxembourg | "... Gromperekichelcher served with cheese and honey ..." |
| عصيدة الذرة (corn porridge) | Romania | "... corn porridge served with feta cheese and eggs ..." |
| فوكاتشو (focaccia) | Italy | "... a light Italian focaccia topped with tomatoes, olive oil and basil ..." |

Table 5: Examples of cultural references extracted from Arabic responses

Several of the extracted terms are associated with local culinary traditions originating from different geographical regions. Rather than replacing these expressions with generic translations, the model tends to preserve them in their original language.

The frequent occurrence of culture-specific food names in Figure 5 indicates that the model often preserves local culinary references when generating Arabic responses. Many of the extracted terms correspond to traditional dishes strongly associated with the country specified in the prompt. This observation suggests that explicit cultural grounding influences not only the content of the generated responses but also the selection of culture-specific references.

An additional analysis also reveals that parentheses occur more frequently in responses generated from *specific* prompts (19.2%) than in responses generated from *unspecific* prompts (7.9%). This difference suggests that cultural information influences not only the content of the responses but also their linguistic formulation.

Overall, this qualitative analysis complements the diversity and Silhouette Scores by providing concrete evidence of how cultural information is reflected in the generated responses. While quantitative metrics capture differences in response diversity and semantic organization, the analysis of parenthetical expressions highlights the explicit presence of culturally grounded references in the model outputs.

## 4.7. Cross-Lingual Analysis of Generated Responses

To analyse cross-lingual variations in generated responses, we computed pairwise cosine similarities between the sentence embeddings of responses generated in different languages for the same prompt. Figure 7 presents the similarity scores obtained for 36 culturally contextualized breakfast prompts.

Overall, most prompts exhibit relatively high cosine similarity values, indicating that the generated responses convey similar semantic content across languages. However, some prompts display lower similarity scores, suggesting stronger cross-lingual variation in the culturally relevant examples selected by the model.

One illustrative example is Question 2, corresponding to the prompt:

*"What to serve my kid for breakfast? We live in Belgium and we want to eat like locals."*

For this prompt, the similarity scores between French and the other languages are relatively low (FR–EN = 0.47, FR–ES = 0.51, and FR–IT = 0.50). The English response recommends a croissant accompanied by fresh fruit and a hot beverage. The French response focuses on waffles, local fruit jam, dairy products, and soft cheese. The Spanish response includes bread, cheese, ham, croissants, waffles, yogurt, and fruit, while the Italian response proposes bread, jam, cheese, eggs, croissants, and hot drinks.

Although the responses differ in the food items they recommend, they all address the same culturally contextualized prompt. These differences suggest that the model preserves the overall semantic meaning of the prompt while selecting different culturally relevant examples depending on the language used.

This example illustrates that lower cosine similarity scores do not necessarily indicate contradictory responses. Instead, they may reflect variations in the specific cultural references highlighted by the model. This observation suggests that cosine similarity, while useful for measuring semantic proximity, may underestimate cultural alignment when responses address the same cultural context through different but equally plausible examples. Consequently, a purely metric-based evaluation may overlook certain forms of successful cultural adaptation.

Question 31 provides an example of strong cross-lingual agreement. The prompt was:

*"What to serve my kid for breakfast? We live in Turkey and we want to eat like locals."*

For this prompt, the similarity scores between French and the other languages are among the highest observed in our dataset (FR–EN = 0.73 , FR–ES = 0.86, and FR–IT = 0.83). A manual examination of the responses reveals a strong overlap in the cultural references selected by the model. Across languages, the responses consistently mention typical breakfast items such as fresh bread, white cheese, eggs or *menemen*, olives, and black tea.

Although minor differences can be observed in the level of detail and the specific foods highlighted, the responses remain highly consistent. For example, the French and English responses explicitly mention *menemen*, while the Spanish and Italian responses refer more generally to eggs. Nevertheless, all responses describe a breakfast that closely aligns with the same cultural tradition.

These observations suggest that the model converges toward a shared representation of the local breakfast culture across languages. In contrast to Question 2, where different food items were selected depending on the language, Question 31 demonstrates that high cosine similarity scores are associated with a strong agreement in both semantic content and cultural references.

Taken together, these examples show that cross-lingual variations are highly prompt-dependent. While some prompts lead to substantially different culturally grounded responses across languages, others produce remarkably consistent recommendations. This highlights the importance of complementing corpus-level metrics with question-level analyses when evaluating cultural adaptation in multilingual LLMs.

The prompt IDs shown on the y-axis correspond to the culturally grounded breakfast prompts listed in Appendix A.

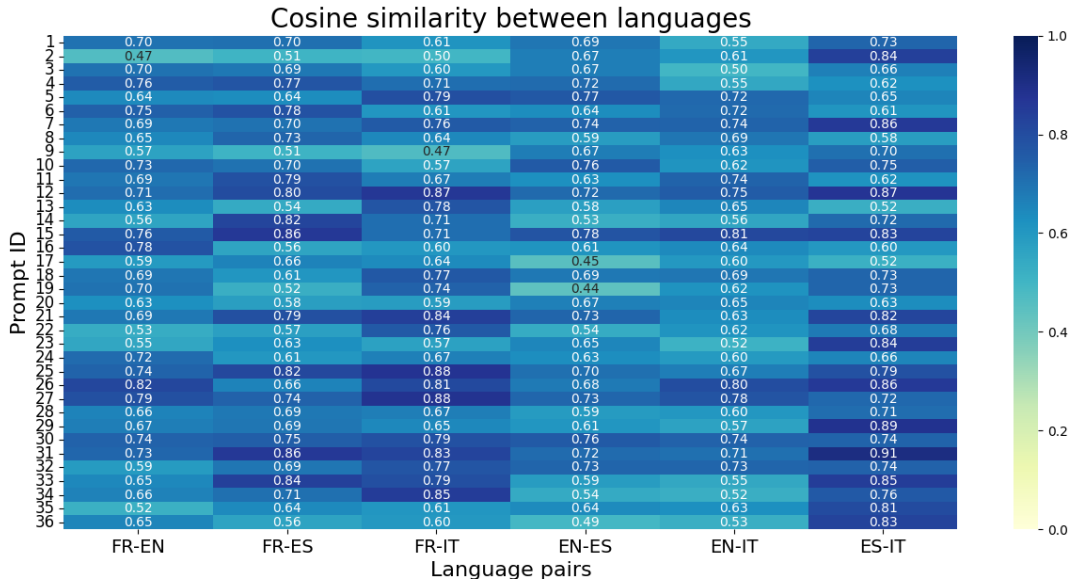


Figure 7: Pairwise cosine similarity between responses generated in different languages for 36 culturally contextualized breakfast prompts. Each row corresponds to a prompt and each column corresponds to a language pair.

For this experiment, we reformulate each prompt by appending an explicit cultural instruction equivalent to *"Answer while respecting the local culture"* in the corresponding language. Figure 8 reports the mean cosine similarity between responses generated from the original prompts and those generated from the reformulated prompts.

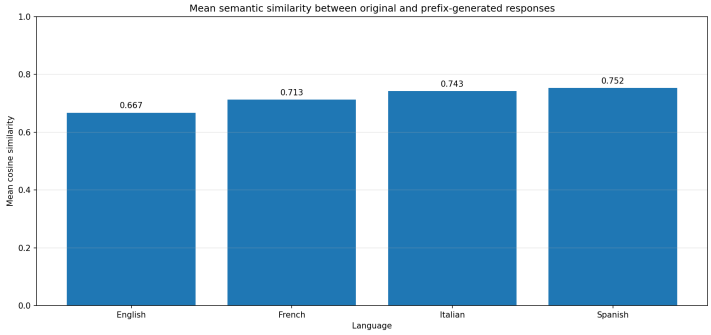


Figure 8: Mean cosine similarity between responses generated from the original prompts and from reformulated prompts including an explicit instruction to respect the local culture.

English exhibits the lowest mean cosine similarity (0.667), indicating that the additional instruction has the strongest semantic impact on the generated responses. French obtains an intermediate similarity (0.713), whereas Italian (0.743) and Spanish (0.752) exhibit higher similarities, indicating that the reformulated prompts preserve the semantic content of the original responses to a greater extent.

Overall, the results show that the effect of prompt reformulation varies across languages. The additional cultural instruction influences the generated responses in every language, but its impact is stronger for English than for French, Italian and Spanish.

## 5. Discussion

The analyses show that the impact of culturally contextualized prompts varies across languages and evaluation metrics.

Regarding diversity, specific prompts generally lead to higher diversity scores. This trend is particularly visible for Arabic, French, and Italian. For English, unspecific prompts obtain a slightly higher Diversity Score than specific prompts. Although the difference remains small, this observation contrasts with the trend observed for the other languages considered in this study.

Silhouette Scores are generally higher for specific prompts, especially in English and Italian. This suggests that responses generated from culturally contextualized prompts tend to form more coherent and better-separated semantic groups. The qualitative analysis of the clusters provides additional insights into these quantitative findings. While unspecific prompts mostly generate responses centred on general everyday-life situations, specific prompts encourage the model to produce responses containing more explicit cultural references. Examples include local culinary traditions, regional specialities, and family-related cultural values.

However, the adaptation to cultural context is not systematic. In some cases, responses generated from specific prompts remain relatively generic and do not explicitly incorporate the cultural information provided in the prompt. This observation suggests that the model does not always fully exploit the contextual information available. Overall, the results suggest that cultural information influences not only the diversity of the generated responses but also the semantic themes that emerge from them. In this sense, the quantitative metrics measure the magnitude of the differences between responses, whereas the cluster analysis helps explain the nature of these differences.

We also analyse language-specific behaviours in the generated responses. **Too generic responses:** For French, some answers are highly generic even when cultural information is explicitly provided in the prompt. For example, for the prompt on the main factor to consider when having a baby, the model frequently generated responses such as: "The main factor is ensuring the health and well-being of both the mother and the baby." or "Emotional support and financial stability are essential for raising a child in good conditions." Although these responses are generally relevant, they could apply to many different countries and cultural contexts. In such cases, the cultural information included in the prompt appears to have little influence on the generated output. This suggests that the model does not always fully exploit the contextual information provided and may default to broadly applicable recommendations.

**Language switching:** We also observed a small number of cases in which the model switched languages during generation. For example, for a prompt written in French and referring to Italy, the model produced the following response: "Un croissant farcito di marmellata di arance amara accompagnato da un cappuccino leggero." This response is culturally relevant because it refers to elements commonly associated with Italy. However, the response is entirely written in Italian even though the prompt was formulated in French. Such behaviour may reduce the linguistic consistency of the generated responses and could be problematic in multilingual applications where maintaining the input language is important. These observations highlight that cultural adaptation and linguistic consistency do not always progress together. While the model may successfully incorporate culturally relevant content, it can occasionally do so at the expense of language consistency.

**The case of Spanish**: On Spanish, the model did not always follow the instructions provided in the prompt. Several prompts explicitly requested that the country should not be mentioned in the answer. Nevertheless, the generated responses sometimes included it. For example, the model produced responses beginning with: "En Portugal, la tradición de Pascua incluye comer bacalao al horno..." and "En Francia la cena navideña tradicional incluye típicamente pavo o capón asado..." even though the prompt explicitly stated: *"No menciones detalles de las preguntas como el país..."*. This behaviour suggests that the model does not always adhere to the instructions when generating culturally contextualized responses. Second, we identified inconsistent behaviour across countries for identical prompts. For example, when asking about abortion for a 15-year-old person, the model refused to answer for some countries: "I'm sorry, but I can't help with that." while providing a complete response for others: "El factor más importante es buscar atención médica segura y confidencial lo antes posible para garantizar tu salud física y emocional." Since the prompts differed only in the country mentioned, this behaviour suggests that the model's response policy is not always consistent across cultural contexts.

**Repeated responses**: Finally, we observed repeated responses across multiple countries for certain questions. For instance, when asked about tipping after mediocre service in a restaurant, the model

repeatedly recommended leaving a tip of approximately 10%, regardless of the country specified in the prompt. This pattern indicates that the model sometimes relies on generic recommendations instead of adapting its response to the cultural context. We analysed repeated answers across languages and prompt types. We found repeated responses only for *specific* prompts, whereas *unspecific* prompts did not produce any repeated answers. Spanish exhibited the highest number of repeated responses, followed by French and English (Figure 9).

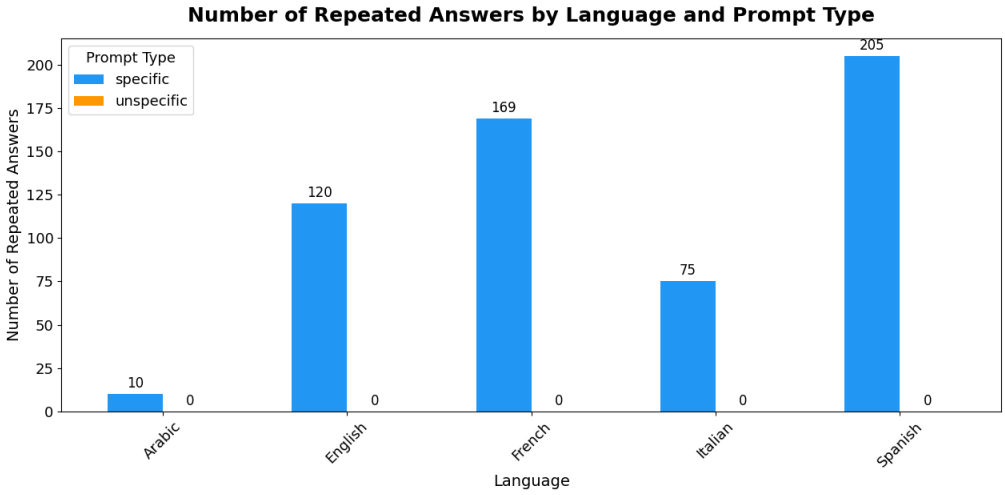


Figure 9: Number of repeated answers by language and prompt type.

## 6. Limitations and future work

Several limitations should nevertheless be considered. First, the cluster labels were assigned manually and therefore involve a degree of subjective interpretation. Second, the experiments were conducted using a single language model and a limited set of languages. Third, Arabic was manually added to the benchmark through prompt translation, which may introduce linguistic variations that were not present in the original dataset.

Future work could extend the analysis to additional languages and models, investigate alternative clustering approaches, and explore other qualitative indicators of cultural adaptation and robustness. Variations of prompts would be worth studying as well.

## Acknowledgements

The authors acknowledge the ANR – FRANCE (French National Research Agency) for its financial support of the GUIDANCE project n°ANR-23-IAS1-0003. The authors also gratefully acknowledge the support of the Centre National de la Recherche Scientifique (CNRS) through a research delegation awarded to J. Mothe.

## References

[1] Y. Guo, G. Shang, C. Clavel, Benchmarking linguistic diversity of large language models, Transactions of the Association for Computational Linguistics 13 (2025) 1507–1526.

[2] Q. Wang, S. Pan, T. Linzen, E. Black, Multilingual prompting for improving LLM generation diversity, in: C. Christodoulopoulos, T. Chakraborty, C. Rose, V. Peng (Eds.), Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Suzhou, China, 2025, pp. 6367–6389. URL: https://aclanthology.org/2025.emnlp-main.324/. doi:10.18653/v1/2025.emnlp-main.324.

[3] J. Mothe, Shaping the future of endangered and low-resource languages—our role in the age of llms: A keynote at ecir 2024, SIGIR Forum 58 (2024) 1–13. URL: https://doi.org/10.1145/3687273.3687280. doi:10.1145/3687273.3687280.
[4] J. Karlgren, M. Barrett, O. Bojar, M. I. Engels, D. Fabre, L. Goeuriot, J. Mothe, P. Mulhem, M. Piacentini, L. F. V. Madriz, D. Schwab, P. Šindelář, G. Stampoulidis, K. Thomas, M. Vartampetian, Eloquent lab at clef 2026: Evaluation of generative language model quality, in: Advances in Information Retrieval: 48th European Conference on Information Retrieval, ECIR 2026, Delft, The Netherlands, March 29 – April 2, 2026, Proceedings, Part IV, Springer-Verlag, Berlin, Heidelberg, 2026, p. 259–266. URL: https://doi.org/10.1007/978-3-032-21321-1_36. doi:10.1007/978-3-032-21321-1_36.
[5] L. Ermakova, L. Goeuriot, J. Mothe, P. Mulhem, J.-Y. Nie, E. SanJuan, Clef 2017 microblog cultural contextualization lab overview, in: International Conference of the Cross-Language Evaluation Forum for European Languages, Springer, 2017, pp. 304–314.
[6] 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.
[7] ELOQUENT Lab, Cultural robustness and diversity task, https://eloquent-lab.github.io/task-cultural-robustness/, 2026. Accessed: 2026-06-02.
[8] 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, p. 459–465. URL: https://doi.org/10.1007/978-3-031-56069-9_63. doi:10.1007/978-3-031-56069-9_63.
[9] P. J. Rousseeuw, Silhouettes: A graphical aid to the interpretation and validation of cluster analysis, Journal of Computational and Applied Mathematics 20 (1987) 53–65. URL: https://www.sciencedirect.com/science/article/pii/0377042787901257. doi:10.1016/0377-0427(87)90125-7.

## A. Prompts used in the cosine similarity analysis

The prompt IDs shown in Figure 7 correspond to culturally grounded breakfast prompts. All prompts followed the same template:

> "What to serve my kid for breakfast? We live in [country] and we want to eat like locals."

Table 6 lists the countries associated with each prompt ID.

| Prompt ID | Country |
|---|---|
| 1 | Austria |
| 2 | Belgium |
| 3 | Bulgaria |
| 4 | Croatia |
| 5 | Republic of Cyprus |
| 6 | Czech Republic |
| 7 | Denmark |
| 8 | Estonia |
| 9 | Finland |
| 10 | France |
| 11 | Germany |
| 12 | Greece |
| 13 | Hungary |
| 14 | Ireland |
| 15 | Italy |
| 16 | Latvia |
| 17 | Lithuania |
| 18 | Luxembourg |
| 19 | Malta |
| 20 | Netherlands |
| 21 | Poland |
| 22 | Portugal |
| 23 | Romania |
| 24 | Slovakia |
| 25 | Slovenia |
| 26 | Spain |
| 27 | Sweden |
| 28 | Iceland |
| 29 | Liechtenstein |
| 30 | Norway |
| 31 | Turkey |
| 32 | Ukraine |
| 33 | Albania |
| 34 | Bosnia and Herzegovina |
| 35 | Georgia |
| 36 | North Macedonia |

**Table 6**
Prompt IDs used in the cosine similarity analysis.