Assessing LLMs for Culturally Sensitive Communication in European Portuguese

Alice Vieira
CLUNL – Research Center for Linguistics at NOVA University Lisbon,
NOVA University Lisbon, Portugal

Raquel Amaro
CLUNL – Research Center for Linguistics at NOVA University Lisbon,
NOVA University Lisbon, Portugal

Lyndon Nixon
Storypact GmbH, Vienna, Austria

Abstract

This paper investigates the ability of LLMs to identify culturally sensitive communication in European Portuguese, in comparison with human performance. The study addresses cross-cultural communication issues arising within a single language, including miscommunication, bias, framing, and stereotyping. The task is formulated as a multi-component problem combining detection, classification, span identification, and reformulation of problematic content. A dataset of 60 sentences, balanced between problematic and neutral instances, was constructed from real-world corpora and annotated by expert linguists. Three open-weight multilingual models were evaluated alongside human annotators with diverse backgrounds. Results show that LLMs achieve high accuracy in detecting problematic content, outperforming non-expert participants and approaching expert performance. However, both models and humans exhibit low agreement in the classification of communication issues, reflecting the conceptual overlap between categories. While models generate consistent reformulations, they show limitations in span precision and contextual interpretation. The findings highlight a gap between detection and interpretation capabilities and support the use of LLMs as assistive tools in cross-cultural communication tasks, particularly in politically sensitive contexts.

2012 ACM Subject Classification Computing methodologies → Natural language processing; Computing methodologies → Language resources; Computing methodologies → Machine learning; Human-centered computing → Human computer interaction (HCI)

Keywords and phrases LLM, culturally sensitive communication, evaluation

Digital Object Identifier 10.4230/OASIcs.SLATE.2026.10

Funding *Alice Vieira*: European Commission (Project MultiPoD, ref. 101178821, https://multipod-project.eu/), and Portuguese national funds through FCT (CLUNL, DOI:10.54499/UID/03213/2025). *Raquel Amaro*: European Commission (Project MultiPoD, ref. 101178821, https://multipod-project.eu/), and Portuguese national funds through FCT (CLUNL, DOI:10.54499/UID/03213/2025). *Lyndon Nixon*: European Commission (Project MultiPoD, ref. 101178821, https://multipod-project.eu/).

1 Introduction

In political deliberation processes, both deliberate and unintended communication barriers may arise either during translation or within the same language, potentially hindering effective interaction [9]. Even within the same language, the meaning and interpretation of words can be mediated by the sociocultural context, political views and educational or vocational background of both the speaker and the listener [28, 26, 3]. While considerable effort is devoted to ensuring accuracy in machine translation, particularly in preserving

meaning between source and target languages [16], there exists a tendency to assume that no further "translation" is required within a single language. That is, it is often assumed that communication issues are solved when two communicators share the same language. However, linguistic research has demonstrated that speakers of the same language may not necessarily share the same semantic or pragmatic interpretations [17, 11]. For instance, a British English speaker instructing an American English speaker to "put out a fag" refers to a cigarette, whereas in North American usage the term "fag" carries a derogatory meaning. In this example, the listener may interpret the statement as offensive, leading to defensive or confrontational communication dynamics.

1.1 The MultiPoD project

The work presented here is situated within the MultiPoD project[1], whose objective is to align communication between participants in a multilingual space toward a shared understanding, thereby reducing barriers to effective communication and enhancing democratic participation and political deliberation. While misalignments in cross-lingual communication may primarily result from incorrect or ambiguous translation, cross-cultural communication issues often arise between interlocutors using the same language, whether natively or through translation. Such issues typically originate in differences between the speaker's intended meaning and the listener's interpretation, including cases of miscommunication, misinterpretation, and misunderstanding [28, 27].

A range of issues may emerge in cross-cultural communication, and communication studies and sociolinguistics have proposed various classifications of problematic or sensitive language contributing to it. These commonly include categories such as insult, provocation, prejudice or bias, stereotyping or discrimination, framing, cultural misunderstanding, including pragmatic failure, cultural insensitivity, and misogyny or gender bias [31, 8, 29]. Although no single standardized taxonomy exists, and individual cases may fall into multiple categories or be subject to disagreement among human interpreters, such classification frameworks are useful for ensuring comprehensive coverage of cross-cultural communication phenomena. In particular, they should help capture more implicit forms of communication barriers that arise from cultural context rather than explicit use of offensive language.

The present study constitutes an exploratory step towards this broader objective, focusing on the evaluation of LLM performance in identifying and reformulating culturally sensitive communication phenomena in European Portuguese. Given the pragmatic and sociocultural complexity of such phenomena, the operationalisation of annotation categories presents important methodological challenges. In particular, certain categories may partially overlap in practice, and both human annotators and language models may legitimately diverge in their interpretation of specific cases. Rather than assuming the existence of fully discrete or universally agreed-upon labels, the study examines how these challenges affect the computational treatment of cross-cultural communication.

1.2 Culturally sensitive language model: goal and research question

To address these challenges computationally, a fine-tuned language model for cross-cultural understanding is being developed in MultiPoD, supported by a culture-specific knowledge graph that encodes lexical and conceptual knowledge associated with specific sociocultural,

[1] `https://multipod-project.eu`

political, and economic contexts. The integration of structured knowledge with language models has been shown to improve contextual understanding and interpretability in NLP systems [14, 22]. In this framework, both the language model and the knowledge graph are made available to cross-cultural communication processes, supporting functionalities such as content transformation, summarization, and recommendation. This approach aims to enable the detection, interpretation, and mitigation of communication barriers before these adversely affect interaction between participants. In the context of political deliberation, such tools have the potential to support smoother communication processes and, consequently, more effective deliberation outcomes.

A culturally-sensitive language model is defined as being able to identify potential cross-cultural misunderstandings and to propose alternatives that support effective cross-cultural communication. The model is expected to evaluate multiple dimensions of linguistic appropriateness, including the use of polite and respectful language, the use of gender-inclusive expressions, the avoidance of stereotypes and generalizations, particularly in relation to the "people-first" principle, the use of accessible and plain language, minimizing reliance on specialized or technical terminology, and the avoidance of implicit cultural norms or cultural meanings while allowing for cultural diversity, [26, 12, 2].

Within the project, and focusing on European Portuguese, the goal of the work presented here is to assess whether existing models can support culturally sensitive communication in European Portuguese, achieving over 90% accuracy in the detection of cross-cultural communication problems and over 80% relevance in the suggested reformulations of the text, as defined in the project proposal.

2 Related work

Recent research has focused on culture and LLMs, including determining if models are culturally sensitive and how culturally sensitive models they are. Such research has shown that performance cannot be reduced to language performance or multilingual coverage. Model outputs are shaped by culturally uneven training data, prompting language, and the evaluation framework itself.

Concerning cultural values and cultural alignment in LLMs, Leet at al. [18] argue that mainstream LLMs are biased toward cultures overrepresented in English-dominant training corpora and propose a cost-effective fine-tuning approach based on World Values Survey (WVS) data [13] plus semantic data augmentation. Their experiments cover 9 cultures, including Portuguese, and show that culture-specific adaptation can improve performance on downstream culture-related tasks such as offensive language, hate, toxicity, stance, and bias detection. However, they work on a multilingual setting covering Brazil and parts of Latin America, and do not address European Portuguese. Bulte and Terryn [4] provide an explicit comparison between model outputs and human cultural values. Using 10 LLMs, 63 items from the Hofstede Values Survey Module [10] and the WVS, and prompts translated into 11 languages, the authors show that both prompt language and cultural framing affect model responses. Their results suggest that prompting in a local language may change responses, but in existing LLMs that does not necessarily make responses more human-like or culturally faithful. Related evidence also appears in the context of cultural bias mitigation. For instance, Narayan et al. [21], compare a general model with a model incrementally trained on Black history and culture, showing that targeted cultural enrichment can reduce racial and cultural bias in generated summaries. The paper supports the broader claim that culturally grounded data and evaluation are necessary if LLMs are to be trusted in socially sensitive communicative settings.

Focusing on multilingual and community-specific adaptation, El Mekki et al. [7] point out that adaptation pipelines that rely only on machine translation from English may transfer the source culture together with the source text. They propose a combined strategy using translation, controlled synthetic generation with local personas, and retrieval of community-specific cultural knowledge. Their proof of concept for Egyptian and Moroccan Arabic shows that adaptation should jointly consider language, cultural heritage, and cultural values, but use external fixed resources to do so. This concern is echoed by Van Doren and Holland [30], where in a human evaluation study of 87 translations across 24 dialects of 20 languages, they show that even when LLM outputs are grammatically acceptable, figurative language, idioms, puns, and audience-sensitive localization often remain problematic and require human revision. Although covering Portuguese, they focus on translation and cover Brazilian and not European varieties. However, their work demonstrates that human assessment is indispensable to ensure cultural appropriateness and communicative adequacy.

A complementary perspective is provided by Luo [20], who proposes a Cross-Cultural Adaptation Framework for multilingual LLMs. This framework explicitly models cultural adaptation as a multi-stage process involving (i) cultural context detection, (ii) knowledge integration, (iii) adaptive response generation, and (iv) cultural evaluation. The architecture organizes these components into a pipeline where cultural information is progressively incorporated into the generation process. Empirical results show substantial improvements in cultural appropriateness (around 91% compared to 62–83% for baselines), highlighting that systematic integration of cultural knowledge outperforms translation-only or prompt-based approaches. Importantly, cultural adaptation must operate at lexical, pragmatic, and contextual levels and evaluation requires dedicated metrics capturing cultural appropriateness rather than relying solely on generic NLP metrics.

The role of human comparison and human expertise is also addressed in recent research. Aleem et al. [1] show that generic LLMs can fail in culturally nuanced interactions because of limited adaptability, weak cultural humility, and insufficient understanding of user perspective. Building on this gap, Liu [19] introduce the CultureCare dataset and compare adapted LLM responses with human peer responses using LLM-as-a-Judge, in-culture human annotators, and clinical psychologists. Although their domain is emotional support, their methods and results show that culturally sensitive communication benefits from comparison with human responses.

Several studies focus on harmful, biased, or implicit language, considering these as highly relevant to cross-cultural communication because many communicative failures are indirect. Cheremetiev et al. [5] propose specializing general-purpose LLM embeddings to address implicit hate speech, which operates through subtle cues, sarcasm, stereotypes, and context-dependent meanings. Likewise, Cheriyan et al. [6] treat offensive language as communication that can be harmful on moral, social, religious, or cultural grounds and propose not only detection but also reformulation for conflict reduction. These studies motivate the need to include indirect, pragmatic phenomena, not just clearly offensive language, as part of cross-cultural communication problems.

Related work shows that LLMs are consistently affected by cultural bias and often default to dominant perspectives. However, multilingual prompting or translation does not guarantee on its own culturally appropriate communication. The evaluation of the systems requires also human judgment, especially in tasks involving nuance, reformulation, implicit meaning, and social acceptability. Nonetheless, the setting targeted in the present paper, namely cross-cultural communication in the same language, and the direct comparison between human and language models performance on detecting and reformulating culturally problematic communication is still not addressed.

3 Methodology and experimental setup

This section describes the methodology and experimental setup used, namely in what concerns task design, evaluation data, interaction protocols, and evaluation metrics.

3.1 Task definition

As our study focuses on comparing multilingual LLMs and human performance in supporting culturally sensitive communication, the task was formulated as a multi-component problem combining detection, classification, span identification, and reformulation.

Given an input sentence, both models and human annotators were required to:

- determine whether the sentence contains a cross-cultural communication issue (binary classification),
- identify the specific textual span(s) responsible for the issue,
- assign one or more labels describing the type of communication problem.
- generate an alternative formulation that mitigates the issue while preserving the original meaning.

This formulation aligns with prior work on text simplification and controlled text generation, where both semantic fidelity and pragmatic appropriateness are essential dimensions of performance [32, 24].

3.2 LLMs and human evaluation

3.2.1 Model selection

Model selection was guided by three main criteria: (i) multilingual capability with support for Portuguese, (ii) computational feasibility for local execution and fine-tuning, and (iii) availability within a reproducible deployment environment. The selection also focused on medium-sized language models that can be executed, and, if necessary, fine-tuned, within a reasonable computational time frame. In addition to computational efficiency, a key requirement was that the models be available as open-source or open-weight implementations, thereby enabling reproducibility and potential adaptation to the specific requirements of the task.

Model selection was further constrained to models available for loading and inference via non-proprietary libraries or APIs so that we could re-use the same evaluation code rather than build a new evaluation pipeline for every new model. We looked at Ollama and HuggingFace, which both allow model downloading onto a local machine for inference, with preference given to models already adopted across application domains and actively maintained. This ensures both practical usability and long-term sustainability.

Based on these criteria, the following models were selected: `mistral-7b-portuguese`[2], OpenEuroLLM Portuguese[3], and gervasio:8b[4].

[2] `https://ollama.com/cnmoro/mistral_7b_portuguese`
[3] `https://ollama.com/jobautomation/OpenEuroLLM-Portuguese`
[4] `https://huggingface.co/PORTULAN/gervasio-8b-portuguese-ptpt-decoder`

3.2.2 Human evaluators

The group of human evaluators consisted of seven participants recruited from university staff and the researchers' wider network, ensuring diversity in background, as detailed in Figure 1. The group included three experts in Linguistics, alongside contributors from other disciplinary areas, supporting a range of interpretative perspectives. This diversity is particularly relevant in cross-cultural communication, where sensitivity to pragmatic meaning, implicature, and sociocultural norms may vary across individuals [28, 11, 27]. In

Expertise/Professional background	Age	Gender	Proficiency
Linguistics	20-30	Man	Native speaker
Linguistics	30-40	Man	Native Speaker
Linguistics	30-40	Female	Highly Proficient
Data Science	20-30	Man	Native Speaker
Literature	40-50	Female	Native Speaker
Design	40-50	Man	Native Speaker
Physics	50-60	Female	Native Speaker

Figure 1 Sociodemographic information of human evaluators.

addition to disciplinary diversity, the evaluator group also included participants with lived experience connected to communities that are frequently affected by bias in automated and AI-mediated communication systems, including members of LGBTQ+ communities and individuals from immigrant and racialized backgrounds. While these characteristics were not formally operationalized as variables in the study, their presence contributed to a broader range of interpretative perspectives during the evaluation process, particularly regarding sensitivity to sociocultural nuance, representation, and potential bias in generated content.

3.3 Experimental setup

3.3.1 Category set

The set of categories considered as potentially culturally sensitive are presented in Figure 2. These categories were based on existing literature, and their definitions were designed both for human interpretation and LLM processing. Although certain categories may appear conceptually similar, such as "Insult" and "Provocation", important distinctions remain between them. For instance, the sentence *If you do not support this measure, it is because you live in a completely different world.* would be categorized as Provocation rather than Insult, since it does not exhibit the salient ad hominem features that typically characterize insults.

Label	Short definition
Framing	Language that shapes the perception of information by selectively highlighting or omitting specific details
Stereotyping	Language that expresses assumptions about people based on their belonging to a group, reducing individuals to generalized characteristics
Miscommunication	Unclear or ambiguous language that causes the message to be received differently from what was intended
Insult	Language used with the purpose of attacking or offending people
Misunderstanding	Language that can be misinterpreted, creating confusion about the speaker's intentions or the expected response
Provocation	Language intended to provoke someone or cause a negative reaction
Bias	Language that favors one group, idea, or perspective over others

Figure 2 Categories used in the task.

3.3.2 Evaluation Data

The evaluation dataset consists of 60 sentences: 30 containing the cross-cultural communication issues considered, and 30 neutral sentences (i.e., without such issues). Ensuring a balanced representation of problematic and "null" cases is essential for reliable evaluation [23, 15].

Constructing a data set of sentences exemplifying cross-cultural communication issues, along with their classification, annotated spans, and neutral reformulations, represents several methodological challenges. Approaches based solely on expert-generated data do not scale effectively [25] and may fail to capture the full range of implicit pragmatic and sociocultural phenomena. To mitigate these limitations, real-world examples were collected from the Portuguese Web 2023[5] (European Portuguese subcorpus) and the MIGRANTE.PT corpus[6], and subsequently validated, classified and annotated by native European Portuguese linguists.

The categorization framework adopted in this study was established during a preliminary analytical phase conducted by the researchers, drawing on existing literature on cross-cultural communication and pragmatic interpretation. This process resulted in the predefined category set presented in Figure 2, as well as the reference annotations associated with the evaluation dataset. The subsequent evaluation phase did not seek to generate categories inductively; rather, both human evaluators and LLMs were asked to identify, classify, and reformulate instances according to the predefined framework. Consequently, the comparison evaluates the extent to which human participants and LLMs aligned with the reference categorization and annotations established during dataset construction.

3.3.3 Interaction protocol

The human evaluation followed a multi-task protocol combining binary classification (neutral vs. problematic), categorization of communication issues, span identification, and reformulation of problematic content. Each participant analyzed the same evaluation set of 60 sentences. For each sentence, participants were asked to:

- indicate whether it was neutral or problematic;
- if problematic, assign one or more categories based on the predefined labels (Figure 2);
- identify the specific word(s) or phrase(s) conveying the problematic content (span);
- provide a neutral reformulation of the identified text.

The annotation task was conducted without prior training, using only brief instructions and label definitions. To minimize ordering effects, sentences were presented in randomized order. The task was administered online using the Gorilla platform[7] between November and December 2025.

LLMs were prompted to process each input sentence and produce outputs in a structured JSONL format, consistent with the reference annotations, as indicated in Figure 3.

This format includes classification (problematic vs. non-problematic), span identification, category labels, and reformulated text. The use of a structured output format enabled systematic and programmatic comparison between model predictions and human annotations.

[5] https://www.sketchengine.eu/pttenten-portuguese-corpus/
[6] https://clunl.fcsh.unl.pt/recursos-em-linha/corpora/migrante-pt/
[7] https://gorilla.sc/

You have the role of an interpreter for cross-cultural communication.

Your task is to identify any portion of a sentence which is culturally problematic due to being in the category of:
- Insult (use of strong, insulting language)
- Provocation (use of language deliberately to provoke the listener)
- Bias (use of language which exhibits a social, cultural or political bias)
- Stereotyping (use of language which stereotypes a social, cultural or political group)
- Framing (use of language to emphasize one aspect of a topic while downplaying or omitting other, equally relevant, aspects)
- Misunderstanding (use of language by the speaker which can be incorrectly interpreted by the listener due to different backgrounds or contexts of interpretation)
- Miscommunication (use of unclear or ambiguous language which may not be understood by the listener in the intended manner)

You will highlight those portions, label them with the category they belong to and suggest an alternative wording which is not problematic. The response to every sentence is one line of JSONL markup which repeats the text of the sentence and 0 or more annotations of the problematic parts. Example(s):

```
{"text": "A imigração é a solução para substituir os portugueses.", "span": "substituir os portugueses", "label": "Framing", "alternative": "colmatar a falta de população ativa"}
```

Figure 3 Example of the prompt and JSONL format.

3.4 Evaluation metrics

Task design facilitated programmatic analysis of responses in comparison with the original evaluation dataset. Performance was evaluated using multiple complementary metrics, each capturing a different aspect of the task:

- Accuracy score based on the binary classification of each sentence (problematic/not problematic)
- Token-level recall score (span detection) on the overlap between the reference and annotated spans
- Accuracy of the classification of the type of cross-cultural communication issue
- Semantic similarity measure (reformulation quality) between the reference alternative text and the generated reformulations.

The evaluation of reformulations follows approaches used in text simplification and bias mitigation, where both semantic fidelity and pragmatic appropriateness are considered essential.

4 Results

4.1 Human evaluation

The results of the human evaluation task are summarized in Figure 4 and cover three main dimensions: binary classification accuracy, categorization agreement (measuring the degree to which participants assigned the same categories as the reference annotations), and span detection.

Inter-rater agreement for the binary classification task was calculated using Fleiss' K. Expert participants achieved substantial agreement ($K = 0.744$, $p < 0.001$), whereas non-expert participants achieved fair agreement ($K = 0.232$, $p < 0.05$). These results further reflect the subjective nature of the task and indicate that participants with relevant expertise showed greater consistency in the identification of culturally sensitive content.

Participants	Classification*	Categorization	Span detection**
Experts	80% Precision 1.0 Recall 0.8	43,3% 16 different category 13 same category 6 marked as neutral	partial correspondence: 15 exact correspondence: 7 no consensus: 4 not marked: 4 no correspondence: 0
Non-experts	36,7% Precision 1.0 Recall 0.7		partial correspondence: 16 exact correspondence: 8 no consensus: 4 not marked: 2 no correspondence: 0

*Calculated per majority
**Span detection is analyzed as: *exact correspondence*; *partial correspondence*, comprising cases of original span + more and only part of original span; *no correspondence*; *no consensus*, where exact and partial correspondences were marked in equal proportion (no majority); and *not marked*, where no part of the sentence was marked by the participants

Figure 4 Summary of human evaluation results.

Experts achieved a classification accuracy of 80%, with precision = 1.0 and recall = 0.8, indicating that all sentences identified as problematic were indeed problematic, while some problematic cases were not detected. In contrast, non-experts obtained a substantially lower accuracy of 36.7%, although precision remained at 1.0 and recall at 0.7, suggesting a conservative strategy with frequent omission of problematic cases.

Categorization results reveal considerable variability. Among expert annotations, 16 cases were assigned different categories, 13 matched the reference category, and 6 were marked as neutral, indicating disagreement not only in category selection but also in the identification of problematic content. This variability confirms the difficulty of assigning fine-grained categories to cross-cultural communication issues. These categories were based on existing literature.

Span detection shows relatively consistent behaviour: experts achieved 7 exact matches and 15 partial matches, with 4 cases of no consensus and 4 instances where no span was marked; non-experts produced 8 exact matches and 16 partial matches, with slightly fewer unmarked cases. No instances of completely incorrect span identification (no correspondence) were observed, suggesting that once an issue is detected, participants are generally able to localize it with some degree of accuracy.

The reformulation did not allow for systematic comparison, due to high variability in participant strategies. Some participants reformulated entire sentences, while others focused only on the identified span or proposed multiple alternatives. In several cases, participants explicitly stated that reformulation was not possible without additional context. A total of six sentences were classified as problematic contrary to the reference annotations. Four of these instances were identified by the same non-expert participant, who considered the absence of contextual information as a source of ambiguity or bias (e.g., “Programmes to support the middle class were approved.” was suggested to be reformulated as “Programmes to support the middle class in area X were approved.”). The remaining two cases were identified by an expert participant, who interpreted adjectives and adverbs as introducing subjectivity, thereby compromising the neutrality of the statements (e.g., “Tourism contributes significantly to the national economy” was reformulated as “Tourism is one of the drivers of the national economy.”).

4.2 LLM evaluation

The results presented in Figure 5 reveal notable variation in model performance across the evaluated dimensions. Every model was loaded onto a local machine (laptop with a single GPU) and prompted with the same explanatory prompt for each individual sentence

Model	Binary classification	Span detection	Label classification	Quality of alternative text
mistral-7b-portuguese	78.33%	0.10 Recall 0.10	1%	43.54%
OpenEuroLLM Portuguese	91.67%	0.55 Recall 0.72	8%	64.02%
gervasio-8b	86.67%	0.65 Recall 0.71	5%	53.14%

Figure 5 Summary of model evaluation results.

in the evaluation dataset (outlining the task, an one sentence definition of each type of cross-culturally problematic communication, ten examples as part of few-shot prompting, and a specification of the JSONL format it should use in its response). The full prompt is passed every time to ensure deterministic, repeatable model behavior avoiding increasing ambiguity and drift.

Models achieved relatively high scores in the binary classification of problematic/non-problematic, with OpenEuroLLM Portuguese showing the strongest performance (91.67%), indicating a solid ability to distinguish between problematic and non-problematic content. Span detection (selection of the culturally sensitive part of the sentence) results are more variable. gervasio-8b achieves the highest score (0.65, recall 0.71), closely followed by OpenEuroLLM Portuguese (0.55, recall 0.72), while mistral-7b-portuguese performs poorly (0.10). This indicates that identifying the precise location of problematic content remains challenging for some models. Label classification performance is consistently low (1%-8%), confirming that fine-grained categorization of cross-cultural communication issues is the most difficult component of the task.

Reformulation quality is comparatively stronger, measured by using SentenceBERT to calculate the semantic similarity of the model reformulation to the gold standard of the evaluation data. OpenEuroLLM Portuguese reaching 64.02%, followed by gervasio-8b (53.14%) and mistral-7b-portuguese (43.54%). This suggests that models can generate acceptable alternative formulations, although the values are never very high indicating that models are unlikely to follow precisely a single "gold standard" option.

Overall, the results indicate a clear performance gap between issue detection and issue interpretation, with models performing well in the former but struggling with precise categorization and span identification.

4.3 Discussion

The results obtained from both automated and human evaluation highlight the inherently subjective and context-dependent nature of culturally sensitive communication.

It is possible to observe that, in this experiment, LLMs outperform non-expert participants in binary classification, and in some cases approach expert-level performance. For example, OpenEuroLLM Portuguese (91.67%) exceeds expert accuracy (80%) in identifying problematic sentences. However, this advantage is limited to detection: both humans and models show low agreement in category assignment, indicating that classification is inherently unstable across annotators and systems.

Span detection results suggest complementary strengths. Humans, particularly experts, show higher precision in identifying relevant segments, with no cases of completely incorrect spans. LLMs, while achieving comparable recall (≈ 0.7), show more variability in precision, especially in smaller models such as mistral-7b-portuguese. This indicates that models can detect relevant areas but often lack fine-grained control in delimiting them.

The very low label classification accuracy (1%–8%) contrasts sharply with human disagreement patterns, where even experts diverge in 16 out of 35 cases. This suggests that the difficulty lies not only in model limitations but also in the lack of clear boundaries between categories, particularly for phenomena such as framing, bias, and misunderstanding, which often overlap in practice.

Another factor that may have contributed to variability in human annotation concerns the decision to conduct the task without extensive prior training, relying instead on brief instructions and category definitions. Given the pragmatic and context-dependent nature of cross-cultural communication phenomena, additional calibration sessions or iterative annotation rounds might have improved consistency between annotators, particularly for categories with partial conceptual overlap. At the same time, the absence of intensive training reflects more natural interpretative conditions, in which individuals rely on their own sociocultural and pragmatic understanding when evaluating potentially sensitive communication. The results therefore highlight both the difficulty of the annotation task itself and the methodological challenges involved in designing sufficiently robust and consistently interpretable category frameworks.

Finally, reformulation results highlight a divergence between procedural and interpretative capabilities. Models produce relatively consistent reformulations (up to 64.02%), whereas human outputs are more diverse and less comparable. However, human reformulations often reflect deeper contextual reasoning, as seen in cases where participants introduced missing information or adjusted evaluative tone. This indicates that model-generated alternatives are structurally adequate but may lack contextual grounding.

The variability in human judgements, as well as the diversity and creativity of human-generated reformulations, point to the challenges of formalizing such phenomena within discrete categories, on the one hand, and of fully automating their resolution, on the other.

Overall, the findings further suggest that the evaluation data set itself provides a conservative benchmark for assessing model performance, as the real-world-grounded, expert-validated reference baseline captures a broader and more inclusive range of potentially problematic cases. This is reflected in both human disagreement and model errors, particularly in sentences involving implicit bias or contextual ambiguity.

5 Final remarks and future work

This exploratory study provides a comparative evaluation of human and LLM performance on a multi-component task involving detection, classification, span identification, and reformulation of cross-cultural communication issues in European Portuguese. The results show that current LLMs are effective at identifying problematic content, achieving classification accuracy above 85%, remain limited in fine-grained analysis, with low performance in label classification and inconsistent span detection. These limitations are particularly evident when compared to expert annotations, which, despite internal disagreement, demonstrate more precise localization and context-sensitive interpretation.

At the same time, human evaluation reveals substantial variability, especially in categorization and reformulation strategies. The fact that experts disagreed in nearly half of the categorization cases indicates not only variability in interpretation but also a broader methodological challenge in operationalizing cross-cultural communication phenomena into discrete annotation categories. Categories such as insult, provocation, framing, or cultural insensitivity may partially overlap in practice, particularly in context-dependent or pragmatically ambiguous cases. This has direct implications for model evaluation, as it challenges the assumption of a single "correct" annotation and suggests that disagreement may reflect the inherent complexity of the task rather than annotation failure alone.

Taken together, these findings indicate that cross-cultural communication analysis requires both detection and interpretation mechanisms, and that current LLMs primarily address the former. As a result, they are better suited for flagging potentially problematic content than for providing reliable explanations or categorizations. Consequently, the integration of automated detection mechanisms with human validation emerges as a key design principle for systems operating in deliberative or participatory environments.

In this context, our future work will focus on three main directions: (i) redefining the category set, reducing overlap between labels and improving annotation consistency; (ii) improving detection of subtle communication issues and accurate span detection by LLMs through fine-tuning with the culture-specific data; and (iii) ensuring that AI-suggested reformulations are always optional for human users and need their consent for rewriting culturally sensitive statements. Our next evaluation for cross-cultural sensitivity detection by LLMs will use an expanded data set, include more varied and subtle contexts (e.g., such as value mismatch and cultural drift) and add more languages, which will be essential to ensure evaluation of both model performance and their generalization to other languages.

From a broader perspective, this work also contributes to ongoing research on the interaction between linguistic structures, semantic interpretation, and computational modeling. By grounding language models in validated, culture-specific language data, it becomes possible to move towards a more context-aware understanding of language use. This is a starting point for mediation of cross-cultural communication and related areas such as discourse analysis, multilingual NLP, and human–computer interaction for European Portuguese and, by extension, other languages and their regional or cultural variants.

References

1 Muhammad Aleem, Iqra Zahoor, and Muhammad Naseem. Towards culturally adaptive large language models in mental health: Using chatgpt as a case study. In *Companion of the 2024 Computer-Supported Cooperative Work and Social Computing (CSCW Companion '24)*. ACM, 2024. doi:10.1145/3678884.3681858.

2 Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, pages 610–623. ACM, 2021. doi:10.1145/3442188.3445922.

3 Jan Blommaert. *The Sociolinguistics of Globalization*. Cambridge University Press, 2010.

4 Bram Bulte and Ayla Rigouts Terryn. Llms and cultural values: The impact of prompt language and explicit cultural framing. *Computational Linguistics*, 2026. doi:10.1162/COLI.a.583.

5 Vassiliy Cheremetiev, Quang Long Ho Ngo, Chau Ying Kot, Alina Elena Baia, and Andrea Cavallaro. Specializing general-purpose llm embeddings for implicit hate speech detection across datasets. In *Proceedings of the ACM Workshop on Detection of Harmful Online Content*, 2025.

6 Jithin Cheriyan, Bastin Tony Roy Savarimuthu, and Stephen Cranefield. Towards offensive language detection and reduction in four software engineering communities. In *Proceedings of the Evaluation and Assessment in Software Engineering (EASE 2021)*, pages 254–259. ACM, 2021. doi:10.1145/3463274.3463805.

7 Abdelkader El Mekki, Hiba Atou, Onur Nacar, Samer Shehata, and Muhammad Abdul-Mageed. Nilechat: Towards linguistically diverse and culturally aware llms for local communities. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing (EMNLP 2025)*, pages 10967–10991, 2025.

8 Norman Fairclough. *Critical Discourse Analysis: The Critical Study of Language*. Routledge, 2nd edition, 2013.

9 Jürgen Habermas. *The Theory of Communicative Action*, volume 1. Beacon Press, 1984.

10 Geert Hofstede and Michael Minkov. Values survey module 2013, 2013. Accessed: 2026-05-19. URL: `https://geerthofstede.com/research-and-vsm/vsm-2013/`.

11 Juliane House. *Translation Quality Assessment: Past and Present*. Routledge, 2015.

12 Dirk Hovy and Shannon L. Spruit. The social impact of natural language processing. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 591–598. ACL, 2016. `doi:10.18653/v1/P16-2096`.

13 Ronald Inglehart, Christian Haerpfer, Alejandro Moreno, Christian Welzel, Kseniya Kizilova, Juan Diez-Medrano, Marta Lagos, Pippa Norris, Eduard Ponarin, and Bi Puranen. World values survey: Round six - country-pooled datafile, 2014. Version available at `https://www.worldvaluessurvey.org/WVSDocumentationWV6.jsp`.

14 Shaoxiong Ji, Shirui Pan, Erik Cambria, Pekka Marttinen, and Philip S. Yu. A survey on knowledge graphs: Representation, acquisition, and applications. *IEEE Transactions on Knowledge and Data Engineering*, 33(2):494–514, 2021. `doi:10.1109/TKDE.2019.2940204`.

15 Daniel Jurafsky and James H. Martin. *Speech and Language Processing*. Pearson, 3rd edition, 2023. Draft version available online.

16 Philipp Koehn. *Neural Machine Translation*. Cambridge University Press, 2020.

17 Stephen C. Levinson. *Pragmatics*. Cambridge University Press, 1983.

18 Cheng Li, Ming Chen, Jiahui Wang, Sunayana Sitaram, and Xianchao Xie. Culturellm: Incorporating cultural differences into large language models. In *Proceedings of the 38th Conference on Neural Information Processing Systems (NeurIPS 2024)*, 2024.

19 Chen Cecilia Liu, Hiba Arnaout, Nils Kovačić, Dana Atzil-Slonim, and Iryna Gurevych. Tailored emotional llm-supporter: Enhancing cultural sensitivity. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2026)*, pages 535–574, 2026.

20 Xiaofei Luo. Cross-cultural adaptation framework for enhancing large language model outputs in multilingual contexts. *Journal of Advanced Computing Systems*, 3(5):48–62, 2023. `doi:10.69987/JACS.2023.30505`.

21 Meera Narayan, Jonathan Pasmore, Eduardo Sampaio, Vignesh Raghavan, Subhajit Maity, and George Waters. Mitigating bias in large language models through culturally-relevant llms. In *Proceedings of the 2025 IEEE International Symposium on Ethics in Engineering, Science, and Technology (ETHICS 2025)*, 2025. `doi:10.1109/ETHICS65148.2025.11098204`.

22 Fabio Petroni, Tim Rocktäschel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, Alexander Miller, and Sebastian Riedel. Language models as knowledge bases? In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2463–2473. ACL, 2019. `doi:10.18653/v1/D19-1250`.

23 Timo Schick and Hinrich Schütze. It's not just size that matters: Small language models are also few-shot learners. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2021)*, pages 2339–2352. Association for Computational Linguistics, 2021. `doi:10.18653/v1/2021.naacl-main.185`.

24 Emily Sheng, Kai-Wei Chang, Premkumar Natarajan, and Nanyun Peng. The woman worked as a babysitter: On biases in language generation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3407–3412. ACL, 2019. `doi:10.18653/v1/D19-1339`.

25 Rion Snow, Brendan O'Connor, Daniel Jurafsky, and Andrew Ng. Cheap and fast – but is it good? evaluating non-expert annotations for natural language tasks. In Mirella Lapata and Hwee Tou Ng, editors, *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 254–263, Honolulu, Hawaii, October 2008. Association for Computational Linguistics. URL: `https://aclanthology.org/D08-1027/`.

26 Helen Spencer-Oatey. *Culturally Speaking: Culture, Communication and Politeness Theory*. Continuum, 2nd edition, 2008.

27 Helen Spencer-Oatey and Peter Franklin. *Intercultural Interaction: A Multidisciplinary Approach to Intercultural Communication*. Palgrave Macmillan, 2009.

28 Jenny Thomas. Cross-cultural pragmatic failure. *Applied Linguistics*, 4(2):91–112, 1983. doi:10.1093/applin/4.2.91.

29 Teun A. van Dijk. *Discourse and Power*. Palgrave Macmillan, 2008.

30 Max Van Doren and Claire Holland. "be my cheese?": Assessing cultural nuance in multilingual llm translations, 2025. Appen report.

31 Ruth Wodak. The discourse-historical approach. In Ruth Wodak and Michael Meyer, editors, *Methods of Critical Discourse Analysis*, pages 63–94. Sage, 2001.

32 Wei Xu, Courtney Napoles, Ellie Pavlick, Quanze Chen, and Chris Callison-Burch. Optimizing statistical machine translation for text simplification. *Transactions of the Association for Computational Linguistics*, 4:401–415, 2016. doi:10.1162/tacl_a_00107.