Found in Translation: Measuring Multilingual LLM Consistency as Simple as Translate then Evaluate

Ashim Gupta **Maitrey Mehta** **Zhichao Xu** **Vivek Srikumar**
Kahlert School of Computing
University of Utah
svivek@cs.utah.edu

Abstract

Large language models (LLMs) provide detailed and impressive responses to queries in English. However, *are they really consistent at responding to the same query in other languages?* The popular way of evaluating for multilingual performance of LLMs requires expensive-to-collect annotated datasets. Further, evaluating for tasks like open-ended generation, where multiple correct answers may exist, is nontrivial. Instead, we propose to evaluate the predictability of model responses across different languages. In this work, we propose a framework to evaluate LLM's cross-lingual consistency based on a simple *Translate then Evaluate* strategy. We instantiate this evaluation framework along two dimensions of consistency: information and empathy. Our results reveal pronounced inconsistencies in popular LLM responses across thirty languages, with severe performance deficits in certain language families and scripts, underscoring critical weaknesses in their multilingual capabilities. These findings necessitate cross-lingual evaluations that are consistent along multiple dimensions. We invite practitioners to use our framework for future multilingual LLM benchmarking.

1 Introduction

Large language models (LLMs) have demonstrated impressive capabilities on English-language tasks. Most popular LLMs like GPT-4o, Llama-3, Qwen-2.5, etc. achieve performances north of 85% on popular benchmarks like MMLU, HellaSwag, WinoGrande, etc. for English (Yang et al., 2024). Increasingly, these models are now deployed in multilingual, multicultural contexts, and are marketed as capable multilingual systems (Meta, 2024). However, understanding and characterizing their performance profiles in different languages remains a critical and open challenge. This difficulty is partly due to the high cost of scaling annotation efforts across many languages. Moreover, even with ample resources to support annotation, evaluation remains difficult for open-ended tasks like long-form question answering or dialogue generation, where a single correct answer may not exist.

In this paper, we propose a framework to study multilingual performance profiles of LLMs by assessing their *cross-lingual consistency* — comparing model outputs in various languages against their corresponding English responses. Our framework *sidesteps the need for expensive annotated data and the formulation of close-form tasks*. Cross-lingual consistency matters for two key reasons. First, consistency is inherently valuable in model behavior, as noted in prior work (Gardner et al., 2020; Ribeiro et al., 2020). A system that behaves consistently across languages is more predictable, which is especially critical in high-stakes, user-facing domains such as healthcare or legal services (Schlicht et al., 2025). Second, consistency ensures equitable access to information. Consider the example in Fig. 1; a scientific question posed in Hindi elicits a response that is both shorter and less informative than its English counterpart.

It is challenging to evaluate consistency on open-ended tasks like open-domain question answering. For close-form tasks like classification or multi-choice question answering, consistency can be quantified through agreement rates or accuracy. For open-ended tasks, ground-truth data is usually absent, and when present, does not encompass all the correct possibilities. In such cases, model-based evaluation metrics are often adopted — such as FActScore (Min et al., 2023) — which require accurate identification of claims and verification of factual entailment. Crucially, such evaluators are optimized for English and a few high-resource languages, making them less reliable when applied to responses in low-resource languages such as Gujarati or Vietnamese.

To tackle this challenge, we adopt a simple yet effective strategy: *Translate then Evaluate*. We trans-

Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics, pages 3477–3496
December 20-24, 2025 ©2025 Association for Computational Linguistics

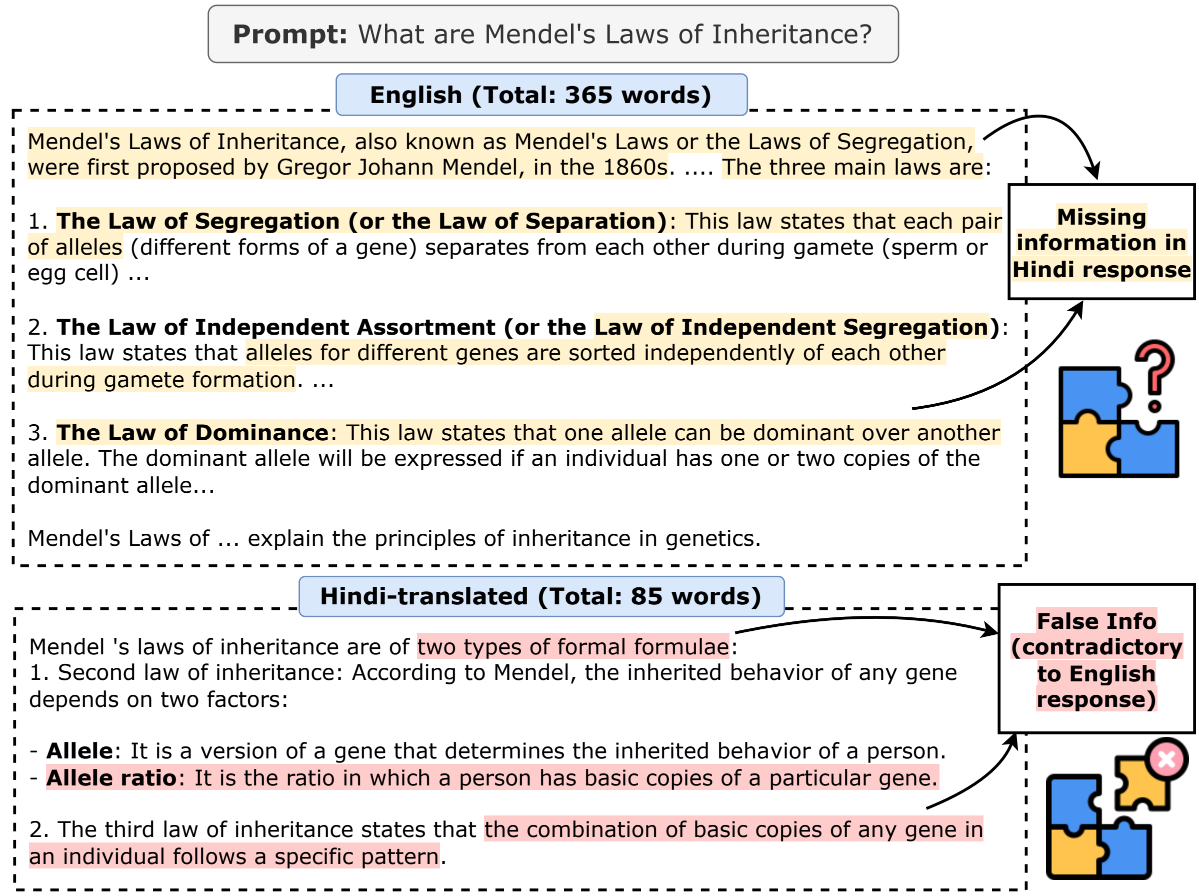

Figure 1: Llama-3 8B (🦙) response to an English (en) prompt versus its response to the translation of the same prompt in a different language, like Hindi (hi). We observe a clear gap in the information generated between the two languages. The response in Hindi has been translated for the reader. The English response is truncated for brevity.

late LLM responses to non-English queries into English using a high-fidelity translation system, allowing for direct comparison between this translated response and the response to its corresponding English query. We use a contemporary translator to enable a scalable evaluation of cross-lingual consistency, making it feasible to include a much broader spectrum of languages in our subsequent analyses.

We instantiate our framework across two disparate consistency dimensions: *information consistency*, to study the information overlap in the responses for the same information-seeking prompt across multiple languages, and *empathy consistency*, to study the persistence of mechanisms of empathy for the same therapeutic prompts across different languages. We adopt off-the-shelf English-based classifiers to act as our consistency evaluators for both these dimensions, with intermediate translation systems facilitating language compatibility.

Our findings show low rates of consistency along both dimensions when tested on mainstream LLMs for 30 languages. We observe that models are generally less consistent for certain language families, scripts, and, generally, for languages not explicitly considered in the training data. Further, stronger proprietary models are more consistent compared to open-weight models. These findings underscore weaknesses in mainstream LLMs' multilingual capabilities, necessitating a more rigorous, comprehensive multi-dimensional consistency evaluation.[1]

2 A Framework for Cross-Lingual Consistency

Evaluating cross-lingual consistency of LLMs is uniquely challenging, especially for open-ended or nuanced tasks where standard accuracy metrics are inapplicable. A key obstacle is that most modern evaluators — be it for factuality, empathy, or even coherence — are monolingual and validated primarily on English data. Multilingual evaluation is either unreliable or infeasible in the absence of

[1]Our code and data is available at https://github.com/ashim95/discrepancy

high-quality resources for other languages.

We address this challenge via a general and scalable framework for measuring cross-lingual consistency, centered around a simple yet effective principle: *Translate then Evaluate*.[2] This paradigm allows us to exploit well-established English-language evaluation pipelines, extending their utility to a much broader set of languages.

Let us formalize this idea. We begin by defining the notion of an evaluator. An evaluator $\mathcal{E}$ is a function that takes two text inputs and returns a compatibility score. Evaluators may be based on traditional heuristic metrics, such as BLEU (Papineni et al., 2002) or ROUGE (Lin, 2004), which compute n-gram overlap, or on more recent LLM-based evaluation metrics like FActScore (Min et al., 2023) and VeriScore (Song et al., 2024).

Let $\mathcal{X}$ denote a set of abstract prompts, where each prompt $x \in \mathcal{X}$ represents a language-agnostic intent or input (e.g., a question or instruction). For language $\ell \in \mathcal{L}$, we denote the realization of x in language ℓ by x_ℓ. In particular, x_{en} refers to the English version of x. We assume that all prompts in $\mathcal{X}$ have valid and semantically equivalent translations in each language $\ell \in \mathcal{L}$. Additionally, for a given multilingual LLM $\mathcal{M}$, we define its generated response to prompt x in language ℓ as:

$$r_\ell(x) = \mathcal{M}(x_\ell).$$

We denote consistency of the model $\mathcal{M}$ for language ℓ with respect to an evaluator $\mathcal{E}$, as $\mathcal{C}_{\mathcal{M},\mathcal{E}}(\ell)$. This consistency is defined as the expected compatibility score computed by an evaluator $\mathcal{E}$ between the English response and non-English response, $r_{\text{en}}(x)$ and $r_\ell(x)$. The expectation is computed over prompts x, which gives us:

$$\mathcal{C}_{\mathcal{M},\mathcal{E}}(\ell) := \mathbb{E}_x \left[\mathcal{E}\left(r_{\text{en}}(x), r_\ell(x)\right)\right].$$

Directly comparing $r_{\text{en}}(x)$ and $r_\ell(x)$ is non-trivial; most model-based evaluators are monolingual and optimized for English. Therefore, we use a translation model $\mathcal{T}_{\ell\to\text{en}}$ that maps responses from the target language ℓ to English:

$$\hat{r}_\ell(x) := \hat{r}_{\ell\to\text{en}}(x) = \mathcal{T}_{\ell\to\text{en}}(r_\ell(x))$$

This translation step brings both responses into the same language, allowing the use of English-only evaluators. Assuming $\mathcal{E}$ produces similar scores before and after translation, we can write the consistency as:

$$\mathcal{C}_{\mathcal{M},\mathcal{E}}(\ell) \approx \mathbb{E}_x \left[\mathcal{E}\left(r_{\text{en}}(x), \hat{r}_\ell(x)\right)\right]$$

Practically, we cannot compute the expectation over all possible inputs x, therefore we approximate the consistency score as the empirical mean over a set of prompts $x \in \mathcal{X}$:

$$\mathcal{C}_{\mathcal{M},\mathcal{E}}(\ell) \approx \frac{1}{|\mathcal{X}|} \sum_{x\in\mathcal{X}} \mathcal{E}\left(r_{\text{en}}(x), \hat{r}_\ell(x)\right) \quad (1)$$

The above framework is fully defined by defining the translation model and an (English) evaluator. We now offer some discussion about how each of these might be chosen.

Desideratum for the translation model. As noted above, a critical requirement of this framework is that the translation model $\mathcal{T}$ preserves the semantic and pragmatic properties that the evaluator $\mathcal{E}$ is designed to measure. In our experiments, we employ the contemporary NLLB 54B-parameter Mixture-of-Experts (MoE) model (Costa-Jussà et al., 2022), which has demonstrated superior performance over conventional systems such as Google Translate for English-targeted translation across a wide range of languages.

Evaluators. To assess the consistency of multilingual LLMs, it is important to use a variety of evaluators. Our proposed framework is evaluator-agnostic and can be instantiated with any that support English. Ideally, a broad suite of evaluators — each targeting a specific linguistic or functional property — will help characterize the multilingual performance profiles of LLMs. The evaluator itself may operate in one of the two modes. In the first, it can directly compare a pair of texts to produce a similarity or compatibility score similar to FActScore. In the second mode, it can target a specific property (e.g., sentiment), compute this property independently for each text, and then compare the resulting attributes to derive a score. In this work, we instantiate our framework using two readily available, off-the-shelf evaluators operating in these two modes, to be detailed in the next section.

3 Instantiations of Cross-Lingual Consistency

As noted earlier (§ 2), the proposed framework does not depend on the specific choice of an evaluator.

[2]This approach is similar to the widely used and highly effective baseline of translating the input before answering, as noted in Shi et al. (2023).

We discuss two instantiations of this framework for scenarios where high-quality model-based evaluators are not available for languages other than English. Other instantiations of the framework could be similarly applied (e.g., toxicity, sentiment, etc.).

3.1 Information Consistency

Motivation. For applications involving scientific or factual content, it is critical that multilingual language models produce consistent information regardless of the language in which a question is posed. Inconsistent outputs can lead to unequal access to reliable knowledge across languages, undermining trust in the model. As illustrated in Fig. 1, the same scientific question asked in English and Hindi yields responses that differ not only in length but also in factual accuracy — the Hindi response is both incomplete and scientifically incorrect. In this subsection, we formalize the notion of cross-lingual information consistency to capture and quantify such discrepancies.

Notation and setup. Given that most state-of-the-art LLMs exhibit stronger performance in English compared to other languages (Grattafiori et al., 2024), it is natural to use the English response as a reference when measuring information consistency. To this end, we leverage FActScore (Min et al., 2023), an off-the-shelf model-based evaluator designed to assess the factual accuracy of generated text against a reference. FActScore fits seamlessly into our framework, as it enables us to evaluate the factual consistency of non-English responses with respect to the information conveyed in the English output. Specifically, FActScore represents the factual precision of a generated response r with respect to a knowledge source s:

$$\mathcal{E}_{\text{FAct}}(r, s) = \frac{1}{|F(r)|} \sum_{f \in F(r)} \mathbf{1}_{\text{supported}(f,s)} \quad (2)$$

where $F(r)$ denotes the set of claims or pieces of information extracted from the response r, and $\text{supported}(f, s)$ indicates whether claim f is supported by the reference text s. In our setup, we use the English response $r_{\mathsf{en}}(x)$ as the reference, and evaluate the translated non-English response $\hat{r}_\ell(x)$ for factual consistency (substituting into Eqn. (1)):

$$\mathcal{C}^{\text{Precision}}_{\mathcal{M},\mathcal{E}_{\text{FAct}}}(\ell) \approx \frac{1}{|\mathcal{X}|} \sum_{x \in \mathcal{X}} \mathcal{E}^{\text{Precision}}_{\text{FAct}}\left(\hat{r}_\ell(x), r_{\mathsf{en}}(x)\right)$$

While FActScore is originally designed to measure factual precision, relying solely on precision is insufficient in our cross-lingual consistency setting. A model that abstains from answering in another language, or responds with vague, easily verifiable statements, may achieve high precision despite omitting key information present in the English response (see the example in Fig. 1, where the Hindi response omits large amount of information). To properly evaluate consistency, we should also compute recall–i.e., whether the non-English response captures the full set of relevant information conveyed in English output. As FActScore is asymmetric — extracting claims from the first argument and verifying them against the second — we obtain a recall-oriented variant of consistency by simply swapping the response and reference positions:

$$\mathcal{C}^{\text{Recall}}_{\mathcal{M},\mathcal{E}_{\text{FAct}}}(\ell) = \frac{1}{|\mathcal{X}|} \sum_{x \in \mathcal{X}} \mathcal{E}^{\text{Recall}}_{\text{FAct}}\left(r_{\mathsf{en}}(x),\ \hat{r}_\ell(x)\right)$$

In our experiments, we report the harmonic mean of these two measures to get our final information consistency metric in the form of an F-score.

3.2 Empathy Consistency

Motivation. LLMs are increasingly being used in mental health contexts — from providing therapeutic feedback (Sharma et al., 2023; Hsu et al., 2025) and conducting diagnostic assessments (Raihan et al., 2024; Mohammadi et al., 2024) to offering direct support in mental health therapy (Gilbert, 2024; McLennan, 2025). In such high-stakes applications, it is essential that these systems behave consistently and predictably for users across diverse linguistic and cultural backgrounds. A particularly important property in this regard is the ability to express comparable levels of empathy across languages — both for ensuring equitable user experience and for upholding ethical standards in emotionally sensitive interactions. In this subsection, we introduce the notion of cross-lingual empathy consistency by leveraging the EPITOME framework (Sharma et al., 2020) for measuring expressed empathy in LLM-human conversations.

The EPITOME framework. The EPITOME framework characterizes expressed empathy through three mechanisms: a) Emotional Reactions (`ER`): an expression of emotions (e.g., warmth, compassion, etc.), b) Interpretations (`IP`): an expression of acknowledgment of the user's predicament, and c) Explorations (`EX`): an expression of intent to explore user's feelings in further detail.[3] Originally, each LLM response is evaluated for each of these three mechanisms on a three-point Likert scale. We follow Gabriel et al. (2024) and condense this three-point scale (no, weak, and strong communication) into a binary categorization by merging the 'weak' and 'strong' communication categories. This means that we are only concerned with the absence or presence of empathetic communication in terms of the three mechanisms.

Notation and setup. Unlike information consistency, where the FActScore-based evaluator directly compares two texts to produce a consistency score, the empathy evaluator first independently computes the empathy profile of each input, and then compares them to produce an empathy consistency score. Consider the three binary empathy classifiers f_m for $m \in \{\texttt{ER}, \texttt{IP}, \texttt{EX}\}$, each trained to detect a specific empathy mechanism in a given response. For any response r, we define its *empathy profile* as a binary vector:

$$\mathbf{e}(r) = [f_{\texttt{ER}}(r),\ f_{\texttt{IP}}(r),\ f_{\texttt{EX}}(r)] \in \{0, 1\}^3.$$

We define the empathy evaluator, $\mathcal{E}_{\text{Emp}}$ as:

$$\mathcal{E}_{\text{Emp}}(r_{\text{en}}(x), \hat{r}_\ell(x)) = \begin{cases} 1, & \text{if } \mathbf{e}(r_{\text{en}}(x)) = \mathbf{e}(\hat{r}_\ell(x)) \\ 0, & \text{otherwise} \end{cases}$$

We then use Eqn. (1) to compute the empathy consistency score.

4 Experimental Setup

In this section, first we will describe experimental details like the LLMs evaluated, languages considered, etc. Following that, we will discuss the specific details for each of the two cross-lingual consistency metrics we use.

LLMs Evaluated. We evaluate several open-weight as well as proprietary LLMs at different parameter scales. Specifically, among the open-weight modes, we evaluate Qwen-2.5 (Yang et al., 2024) at 14B, 72B parameter scales, Llama-3 at 8B and 70B scales (Grattafiori et al., 2024) [4], Gemma-2 (Gemma, 2024) at 9B and 27B, and Aya-expanse at 8B and 32B (Dang et al., 2024). For proprietary models, we consider GPT-4o (Achiam et al., 2023) –both mini and standard size [5]–along with Gemini-2.0-Flash and Flash-Lite (Gemini, 2023).

Language	Consistency	Language	Consistency
Arabic	0.90	Malayalam	0.90
Czech	0.91	Persian	0.89
French	0.93	Russian	0.94
Hindi	0.91	Spanish	0.94
Japanese	0.86	Tamil	0.86

Table 1: Self-consistency and Translationese. We measure information consistency between English and back-translated responses from each language. See § 4.1 for more details.

Languages. For both the consistency evaluations (*information* and *empathy*), we evaluate models on 30 languages with different scripts and covering several language families. We perform evaluation on: Arabic (`ar`), Bengali (`bn`), Czech (`cs`), German (`de`), Greek (`el`), Spanish (`es`), Persian (`fa`), French (`fr`), Gujarati (`gu`), Hindi (`hi`), Indonesian (`id`), Italian (`it`), Hebrew (`iw`), Japanese (`ja`), Kannada (`kn`), Korean (`ko`), Malayalam (`ml`), Dutch (`nl`), Punjabi (`pa`), Polish (`pl`), Portuguese (`pt`), Romanian (`ro`), Russian (`ru`), Tamil (`ta`), Telugu (`te`), Thai (`th`), Turkish (`tr`), Urdu (`ur`), Vietnamese (`vi`), Chinese (`zh`).

4.1 Information Consistency

Prompts. The authors of this paper manually curate a list of 150 information-seeking prompts from eight different categories from MMLU (Hendrycks et al., 2020), including categories such as college chemistry, African history, etc. We specifically focus on questions where answers are long-form and should elicit factual information, excluding non-information seeking prompts like "How are you?", those based on person opinion like "What cities have the best nightlife?", etc. We show an example question and its translations in Tables 8 to 10.

Evaluator. The FActScore evaluator operates in two stages: it first extracts atomic pieces of information from the generated response, and then verifies each claim against the knowledge source (English response in our case). While the original implementation employs GPT-4 for both extraction and verification, relying on a proprietary model is costly at the scale of our experiments. Therefore, we use

[3] We point the reader to Sharma et al. (2020) for a detailed description of the mechanisms and examples.

[4] We use Llama-3.1 for the 8B model, and Llama-3.3 for the 70B model. Llama-3.3 does not have a corresponding 8B model and its 70B-variant is known to be much superior to the Llama-3.1 counterpart.

[5] We use gpt-4o-mini-2024-07-18 and gpt-4o-2024-08-06.

Mechanism	**F1**	**Accuracy**
Emo. Reactions (ER)	86.1	87.2
Explorations (EX)	92.0	95.8
Interpretations (IP)	87.3	87.4

Table 2: Performance (in %) of the evaluator for detecting the three empathy mechanisms.

the open-weight Qwen-2 model at the 14B scale, which we find performs both steps effectively while supporting scalable evaluation.

Translation impact on the evaluator. In our information consistency setup, both inputs to the evaluator are in English: the original English response $r_{\text{en}}(x)$ and the translated version of a non-English response, $\hat{r}_\ell(x)$. To assess the potential impact of translation artifacts — referred to as translationese in literature (Graham et al., 2019) — on the evaluator, we conduct an automatic validation of the evaluation pipeline. For a subset of languages, we take English responses generated by the Llama-3 8B model, translate them into each target language using the NLLB model, and then back into English using the same system. We then compute information consistency between the original and back-translated responses, effectively treating this as a self-consistency check. Results in Table 1 suggest that the Qwen-based evaluator maintains high consistency scores across languages, indicating robustness to translation effects.

4.2 Empathy Consistency

Evaluator and prompts. We use the publicly-available Reddit subset of the EPITOME dataset (∼3k samples) to train three binary classifiers: f_{ER}, f_{IP}, f_{EX}. We sample 100 user utterances from the test split to serve as the evaluation set for computing empathy consistency. These are originally in English, so we translate each of these into the 30 languages. The performance of these classifiers in shown in Table 2, their comparison with few-shot GPT-4o, and other details are mentioned in Appendix C.

5 Results and Analysis

Recall that one critical assumption in our framework is the fidelity of the translation model we use. As such, we first discuss translation system's fidelity (§ 5.1) before analyzing the cross-lingual consistency evaluation results (§ 5.2).

Language	Fidelity	Language	Fidelity
Arabic	4.23	Malayalam	4.12
Czech	4.13	Persian	4.10
French	4.59	Russian	4.28
Hindi	4.38	Spanish	4.48
Japanese	4.14	Tamil	4.16

Table 3: Information Fidelity scores for ten languages as judged by native speakers.

5.1 Fidelity of the Translation System

Consider the case of information consistency, where we rely on translating non-English responses into English to enable evaluation with an English-centric metric. As previously noted (§ 2), we assume that the translation process preserves the underlying property of interest–namely, the factual information of the original response. To assess the validity of this assumption, we conduct a human evaluation of translation fidelity.[6] Specifically, for a generated response in a given language and its corresponding English translation, annotators rate how well the translated text preserves the original information content, using a 5-point Likert scale (with 5 indicating highest fidelity). This evaluation is performed across a subset of ten languages, and the resulting fidelity scores are reported in Table 3. We observe an average rating of 4.26, indicating *consistently high scores across all languages*.

While a similar fidelity assessment for empathy consistency is desirable, specifically, to determine whether the translation preserves the empathy profile of the original response, such evaluations require domain expertise in therapeutic communication. Due to the difficulty of sourcing multilingual clinical annotators, we leave it to future work.

5.2 Consistency Results and Analysis

We present the evaluation results in Table 4, and extend the detailed discussion below.

Stronger proprietary models are generally more consistent. We observe that almost all the open-weight model (first eight rows) are less consistent than proprietary models. The best-performing proprietary model (`gpt-4o-mini`) is significantly more consistent than its best performing open-weight counterpart (`gemma-2-27b`) in open-ended generation (0.68 vs 0.59). The

[6]Given page limit, we report details about our human evaluation setup in § E.3

Model	Avg.	Script				Language Family					Train Data	
		Latin	Indic	PArab	Oth1	Rom	Slav	IA	Drav	Oth2	In	Out
Information Consistency												
`aya-exp-8b`	0.42	0.59	0.11	0.43	0.48	0.61	0.57	0.17	0.06	0.52	0.56	0.09
`aya-exp-32b`	0.49	0.65	0.18	0.51	0.55	0.66	0.64	0.26	0.13	0.59	0.63	0.16
`gemma-2-9b`	0.52	0.60	0.41	0.49	0.53	0.61	0.58	0.44	0.39	0.54	–	–
`gemma-2-27b`	0.59	0.65	0.53	0.56	0.58	0.66	0.63	0.55	0.50	0.60	–	–
`Llama-3.1-8B`	0.32	0.45	0.14	0.25	0.34	0.48	0.42	0.17	0.10	0.35	0.44	0.28
`Llama-3.3-70B`	0.49	0.59	0.33	0.45	0.53	0.61	0.58	0.39	0.29	0.53	0.58	0.47
`Qwen-2.5-14B`	0.39	0.54	0.14	0.34	0.44	0.58	0.50	0.19	0.09	0.46	0.53	0.29
`Qwen-2.5-72B`	0.49	0.62	0.27	0.49	0.54	0.64	0.60	0.35	0.19	0.56	0.61	0.42
`gemini-2.0-lite`	0.65	0.69	0.61	0.64	0.63	0.70	0.68	0.63	0.59	0.65	0.66	0.62
`gemini-2.0`	0.64	0.67	0.60	0.63	0.62	0.68	0.66	0.61	0.59	0.64	0.65	0.61
`gpt-4o-mini`	0.62	0.69	0.53	0.60	0.62	0.70	0.67	0.56	0.51	0.63	–	–
`gpt-4o`	0.68	0.73	0.62	0.67	0.66	0.74	0.72	0.65	0.60	0.68	–	–
Empathy Consistency												
`aya-exp-8b`	0.65	0.74	0.49	0.76	0.65	0.73	0.75	0.59	0.43	0.70	0.72	0.50
`aya-exp-32b`	0.67	0.76	0.52	0.76	0.64	0.75	0.79	0.55	0.51	0.70	0.73	0.52
`gemma-2-9b`	0.59	0.66	0.55	0.53	0.53	0.65	0.67	0.57	0.53	0.57	–	–
`gemma-2-27b`	0.60	0.64	0.58	0.57	0.56	0.66	0.62	0.56	0.58	0.60	–	–
`Llama-3.1-8B`	0.45	0.53	0.36	0.38	0.44	0.55	0.51	0.40	0.29	0.46	0.55	0.42
`Llama-3.3-70B`	0.48	0.55	0.41	0.40	0.49	0.57	0.58	0.41	0.39	0.48	0.54	0.47
`Qwen-2.5-14B`	0.60	0.67	0.49	0.62	0.59	0.69	0.66	0.50	0.45	0.63	0.66	0.56
`Qwen-2.5-72B`	0.61	0.67	0.54	0.57	0.59	0.72	0.64	0.59	0.47	0.61	0.65	0.58
`gemini-2.0-lite`	0.65	0.71	0.63	0.64	0.60	0.68	0.72	0.60	0.63	0.66	0.66	0.63
`gemini-2.0`	0.60	0.64	0.54	0.61	0.59	0.62	0.62	0.55	0.53	0.62	0.61	0.57
`gpt-4o-mini`	0.76	0.81	0.72	0.73	0.73	0.83	0.80	0.71	0.72	0.76	–	–
`gpt-4o`	0.82	0.86	0.77	0.82	0.81	0.88	0.86	0.77	0.79	0.82	–	–

Table 4: Averaged consistency scores for *information* (top) and *empathy* (bottom) consistency. We provide details for language and script groupings in § E.1 and § E.2. For each model, we also group languages by if they were explicitly included in the training data or not. We refer to each of the model documentations for this.

consistency gap is even more acute when observed in the empathy scenario with best close-weight model (`gpt-4o`) outperforming the best open-weight model (`aya-exp-32b`) (0.82 vs 0.67). We hypothesize that the consistency of language models has a positive correlation with their performance, as a similar observation also applies for models within the same family, e.g., `gpt-4o` is more consistent than `gpt-4o-mini`.

LLMs are less consistent for lower resource scripts like Indic. In Table 4, we present results grouped by language scripts. We divide the languages we consider into four script families: {`Latin`, `Indic`, `Perso-Arabic`, `Other`} (details on the grouping of the languages by script in § E.1). Notice that models are the most consistent on languages using the Latin script, compared to lower resource scripts like Indic. We hypothesize that this result is a consequence of our choice to use English as the reference language. Further, we find that for `aya-exp-8b`, consistency substantially drops for languages using the Indic scripts compared to the corresponding Latin script languages (0.11 vs 0.59). This finding emphasizes the need for attention towards certain non-Latin scripts.

LLM responses in low-resource language families are less consistent with English. Consistency results along language family groups follow the related findings observed across script families. We observe that the consistency gap is particularly acute for Dravidian languages, for example, `aya-expanse-8b` only scores 0.06 compared to Romance languages' 0.61 in the information consistency setting. For empathetic consistency, similar degradations are also observed, though are not as significant. Indo-Aryan languages also show significant degradation. For example, `Llama-3.1-8B` gets 0.48→0.17, 0.55→0.40 from Romance to Indo-

Aryan for information and empathy consistency, respectively. Details about language grouping by language family is in Appendix E.2.

LLMs are more consistent with in-training languages. We find that LLMs are generally more consistent in their responses for languages that are seen in the training data. The differences in consistency are substantial in the `Aya` family of models, while the `Gemini` models generalize the most to languages not explicitly considered for training.

Scaling mostly helps. We consider two model sizes per language family in our experiment setup. An increase in model size almost always translates into better task performance (e.g., Grattafiori et al., 2024). We observe that in most cases, larger models are more consistent, but it is not always the case. For example, information and empathy consistency improve for open-weight models like `gemma-2` and `Qwen`, but show minor degradation for `gemini-2.0`. We intend to more rigorously evaluate the scaling trends of multilingual consistency in the future.

The need for multi-dimensional consistency evaluation. We recommend the proposed consistency framework be used in a multi-dimensional manner. In this paper, we instantiated our framework with two such dimensions in information and empathy. We note a highly consistent model along a certain dimension does not guarantee high consistency across all dimensions. For example, `aya-exp-8b` displays poor information consistency when compared to other open-weight models. However, it is more consistent than certain proprietary models like `gemini-2.0` in empathy consistency. A comprehensive, multi-dimensional evaluation of consistency should provide further insight into the specific aspects where models may behave differently across different languages.

6 Related Works

Leveraging output consistency in NLP. The application of consistency to enhance and evaluate NLP systems has a rich history. For instance, Alberti et al. (2019) use round-trip consistency between a question generator and an answer extractor to generate synthetic question-answer pairs. Li et al. (2019) use softened logical constraints as regularizers to improve performance and consistency of NLI models. In a similar vein, Elazar et al. (2021) propose a loss-based method to make models robust to relational paraphrases.

Recently, Chen et al. (2024) evaluated LLMs for self-consistency when posited with hypotheticals and compositional queries in a few-shot setting. In the cross-lingual setting, Qi et al. (2023) evaluate for cross-lingual consistency of LLMs on factual knowledge. Xing et al. (2024) use an embedding similarity-based metric to evaluate cross-lingual semantic consistency for factual question-answers. Our work differs from these works in that we specifically evaluate for open-ended generation tasks, which may or may not have a correct answer. We point the reader to Novikova et al. (2025) for a recent survey regarding consistency in language models.

Cross-lingual evaluation of LLMs. The most popular way of evaluating and comparing cross-lingual performance of LLMs is to evaluate on multilingual benchmarks consisting of parallel data. Such benchmarks focus on commonsense reasoning (Ponti et al., 2020; Lin et al., 2021), natural language inference (Conneau et al., 2018), world knowledge (Lai et al., 2023; Singh et al., 2025, *inter alia*), which are close-form tasks where cross-lingual performance can be simply evaluated in terms of their respective metrics (e.g., accuracy). Instead, our framework aims to evaluate consistency of responses between languages without the assumption of a ground truth.

7 Conclusions and Future Works

Cross-lingual evaluation of LLMs is generally nontrivial. It is especially challenging for open-ended generation tasks which do not have one correct answer. In this work, we propose a cross-lingual consistency framework that evaluates an LLM's ability to respond similarly for the same inputs across different languages. We instantiate our framework using two consistency evaluators: *information* and *empathy*, and tested 12 popular models in 30 diverse languages. Our findings indicate that models can be highly inconsistent across languages. The inconsistency is higher in open-weight models, especially for less-represented language families and scripts. We leave investigations on the disparity of low-resource language response and its mitigation strategies to future work. We invite future LLM creators to include our framework in addition to task performance as an essential barometer for multilingual performance.

Limitations

The *Translate then Evaluate* approach is central to our framework. Hence, our framework relies heavily on the availability of high-quality translation systems that can translate from a source language l to English, and vice-versa. Of the 7000 languages in the world, most are not supported by commercial translation systems. Google Translate supports 249 languages while, NLLB 54B MoE (Costa-Jussà et al., 2022) supports 204 languages. The development of translation systems for such languages is its own research area (Haddow et al., 2022). Currently, our framework is not useful for evaluating for such languages without translation support.

We perform a human study for translation fidelity test in the information consistency, as shown in table 3. A similar study would have been ideal to test translation fidelity for the empathy consistency setting. However, such a study requires annotators with domain expertise which is not always straightforward (Mehta and Srikumar, 2023).

Acknowledgements

We thank the members of the UtahNLP group for their valuable insights. We especially thank Ana Marasović for her help in clarifying the motivation and the experimental setup used in this work. We thank the anonymous reviewers for their insightful feedback. Ashim Gupta is supported by the Bloomberg Data Science Ph.D. Fellowship. This material is based in part upon work supported by the National Science Foundation under Grant #2217154. The support and resources from the Center for High Performance Computing at the University of Utah are gratefully acknowledged. This research is also supported by National Artificial Intelligence Research Resource (NAIRR) Pilot (award NAIRR240144) and the Delta advanced computing and data resource which is supported by the National Science Foundation (award NSF-OAC 2005572).

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. GPT-4 Technical Report. *arXiv preprint arXiv:2303.08774*.

Chris Alberti, Daniel Andor, Emily Pitler, Jacob Devlin, and Michael Collins. 2019. Synthetic QA Corpora Generation with Roundtrip Consistency. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6168–6173, Florence, Italy. Association for Computational Linguistics.

Angelica Chen, Jason Phang, Alicia Parrish, Vishakh Padmakumar, Chen Zhao, Samuel R. Bowman, and Kyunghyun Cho. 2024. Two Failures of Self-Consistency in the Multi-Step Reasoning of LLMs. *Trans. Mach. Learn. Res.*, 2024.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel R. Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. XNLI: Evaluating Cross-lingual Sentence Representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Marta R Costa-Jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, and 1 others. 2022. No Language Left Behind: Scaling Human-Centered Machine Translation. *arXiv preprint arXiv:2207.04672*.

John Dang, Shivalika Singh, Daniel D'souza, Arash Ahmadian, Alejandro Salamanca, Madeline Smith, Aidan Peppin, Sungjin Hong, Manoj Govindassamy, Terrence Zhao, and 1 others. 2024. Aya Expanse: Combining Research Breakthroughs for a New Multilingual Frontier. *arXiv preprint arXiv:2412.04261*.

Yanai Elazar, Nora Kassner, Shauli Ravfogel, Abhilasha Ravichander, Eduard Hovy, Hinrich Schütze, and Yoav Goldberg. 2021. Measuring and Improving Consistency in Pretrained Language Models. *Transactions of the Association for Computational Linguistics*, 9:1012–1031.

Saadia Gabriel, Isha Puri, Xuhai Xu, Matteo Malgaroli, and Marzyeh Ghassemi. 2024. Can AI Relate: Testing Large Language Model Response for Mental Health Support. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 2206–2221, Miami, Florida, USA. Association for Computational Linguistics.

Matt Gardner, Yoav Artzi, Victoria Basmov, Jonathan Berant, Ben Bogin, Sihao Chen, Pradeep Dasigi, Dheeru Dua, Yanai Elazar, Ananth Gottumukkala, Nitish Gupta, Hannaneh Hajishirzi, Gabriel Ilharco, Daniel Khashabi, Kevin Lin, Jiangming Liu, Nelson F. Liu, Phoebe Mulcaire, Qiang Ning, and 7 others. 2020. Evaluating Models' Local Decision Boundaries via Contrast Sets. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1307–1323, Online. Association for Computational Linguistics.

Gemini. 2023. Gemini: A Family of Highly Capable Multimodal Models. *arXiv preprint arXiv:2312.11805*.

Gemma. 2024. Gemma.

Daniel Gilbert. 2024. Despite uncertain risks, many turn to ai like chatgpt for mental health.

Yvette Graham, Barry Haddow, and Philipp Koehn. 2019. Translationese in Machine Translation Evaluation. *arXiv preprint arXiv:1906.09833*.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The Llama 3 Herd of Models. *arXiv preprint arXiv:2407.21783*.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A. Smith. 2018. Annotation Artifacts in Natural Language Inference Data. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 107–112, New Orleans, Louisiana. Association for Computational Linguistics.

Barry Haddow, Rachel Bawden, Antonio Valerio Miceli Barone, Jindřich Helcl, and Alexandra Birch. 2022. Survey of Low-Resource Machine Translation. *Computational Linguistics*, 48(3):673–732.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring Massive Multitask Language Understanding. *arXiv preprint arXiv:2009.03300*.

Shang-Ling Hsu, Raj Sanjay Shah, Prathik Senthil, Zahra Ashktorab, Casey Dugan, Werner Geyer, and Diyi Yang. 2025. Helping the Helper: Supporting Peer Counselors via AI-Empowered Practice and Feedback. *Proceedings of the ACM on Human-Computer Interaction*, 9(2):1–45.

Viet Lai, Chien Nguyen, Nghia Ngo, Thuat Nguyen, Franck Dernoncourt, Ryan Rossi, and Thien Nguyen. 2023. Okapi: Instruction-tuned Large Language Models in Multiple Languages with Reinforcement Learning from Human Feedback. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 318–327, Singapore. Association for Computational Linguistics.

Tao Li, Vivek Gupta, Maitrey Mehta, and Vivek Srikumar. 2019. A Logic-Driven Framework for Consistency of Neural Models. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3924–3935, Hong Kong, China. Association for Computational Linguistics.

Chin-Yew Lin. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Xi Victoria Lin, Todor Mihaylov, Mikel Artetxe, Tianlu Wang, Shuohui Chen, Daniel Simig, Myle Ott, Naman Goyal, Shruti Bhosale, Jingfei Du, Ramakanth Pasunuru, Sam Shleifer, Punit Singh Koura, Vishrav Chaudhary, Brian O'Horo, Jeff Wang, Luke Zettlemoyer, Zornitsa Kozareva, Mona T. Diab, and 2 others. 2021. Few-shot Learning with Multilingual Language Models. *CoRR*, abs/2112.10668.

April McLennan. 2025. Young Australians using AI bots for therapy.

Maitrey Mehta and Vivek Srikumar. 2023. Verifying Annotation Agreement without Multiple Experts: A Case Study with Gujarati SNACS. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 10941–10958, Toronto, Canada. Association for Computational Linguistics.

Meta. 2024. Meta AI is Now Multilingual, More Creative and Smarter. Accessed: 2025-04-16.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. FActScore: Fine-grained Atomic Evaluation of Factual Precision in Long Form Text Generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12076–12100.

Seyedali Mohammadi, Edward Raff, Jinendra Malekar, Vedant Palit, Francis Ferraro, and Manas Gaur. 2024. WellDunn: On the Robustness and Explainability of Language Models and Large Language Models in Identifying Wellness Dimensions. In *Proceedings of the 7th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pages 364–388, Miami, Florida, US. Association for Computational Linguistics.

Jekaterina Novikova, Carol Anderson, Borhane Blili-Hamelin, and Subhabrata Majumdar. 2025. Consistency in Language Models: Current Landscape, Challenges, and Future Directions. *arXiv preprint arXiv:2505.00268*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: A Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Edoardo Maria Ponti, Goran Glavaš, Olga Majewska, Qianchu Liu, Ivan Vulić, and Anna Korhonen. 2020. XCOPA: A Multilingual Dataset for Causal Commonsense Reasoning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2362–2376, Online. Association for Computational Linguistics.

Jirui Qi, Raquel Fernández, and Arianna Bisazza. 2023. Cross-Lingual Consistency of Factual Knowledge in Multilingual Language Models. In *Proceedings of the 2023 Conference on Empirical Methods in*

Natural Language Processing, pages 10650–10666, Singapore. Association for Computational Linguistics.

Nishat Raihan, Sadiya Sayara Chowdhury Puspo, Shafkat Farabi, Ana-Maria Bucur, Tharindu Ranasinghe, and Marcos Zampieri. 2024. MentalHelp: A Multi-Task Dataset for Mental Health in Social Media. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 11196–11203, Torino, Italia. ELRA and ICCL.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 2020. Beyond Accuracy: Behavioral Testing of NLP Models with CheckList. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4902–4912, Online. Association for Computational Linguistics.

Ipek Baris Schlicht, Zhixue Zhao, Burcu Sayin, Lucie Flek, and Paolo Rosso. 2025. Do LLMs Provide Consistent Answers to Health-related Questions across Languages? In *European Conference on Information Retrieval*, pages 314–322. Springer.

Ashish Sharma, Inna W Lin, Adam S Miner, David C Atkins, and Tim Althoff. 2023. Human–AI Collaboration Enables More Empathic Conversations in Text-Based Peer-to-Peer Mental Health Support. *Nature Machine Intelligence*, 5(1):46–57.

Ashish Sharma, Adam Miner, David Atkins, and Tim Althoff. 2020. A Computational Approach to Understanding Empathy Expressed in Text-Based Mental Health Support. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5263–5276, Online. Association for Computational Linguistics.

Freda Shi, Mirac Suzgun, Markus Freitag, Xuezhi Wang, Suraj Srivats, Soroush Vosoughi, Hyung Won Chung, Yi Tay, Sebastian Ruder, Denny Zhou, Dipanjan Das, and Jason Wei. 2023. Language Models are Multilingual Chain-of-Thought Reasoners. In *The Eleventh International Conference on Learning Representations*.

Shivalika Singh, Angelika Romanou, Clémentine Fourrier, David Ifeoluwa Adelani, Jian Gang Ngui, Daniel Vila-Suero, Peerat Limkonchotiwat, Kelly Marchisio, Wei Qi Leong, Yosephine Susanto, Raymond Ng, Shayne Longpre, Sebastian Ruder, Wei-Yin Ko, Antoine Bosselut, Alice Oh, Andre Martins, Leshem Choshen, Daphne Ippolito, and 4 others. 2025. Global MMLU: Understanding and Addressing Cultural and Linguistic Biases in Multilingual Evaluation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 18761–18799, Vienna, Austria. Association for Computational Linguistics.

Yixiao Song, Yekyung Kim, and Mohit Iyyer. 2024. VeriScore: Evaluating the Factuality of Verifiable Claims in Long-Form Text Generation. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 9447–9474.

Xiaolin Xing, Zhiwei He, Haoyu Xu, Xing Wang, Rui Wang, and Yu Hong. 2024. Evaluating Knowledge-based Cross-lingual Inconsistency in Large Language Models. *arXiv preprint arXiv:2407.01358*.

An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, and 1 others. 2024. Qwen2.5 Technical Report. *arXiv preprint arXiv:2412.15115*.

A Experimental Details and Observations

When running Gemini models, we sometimes receive a `RECITATION` error from the Google's API. We found there to be no workaround and therefore use a default string of "`No answer`" in those cases.[7]

B Experimental Details for Measuring Information Consistency

B.1 Translation Related

1. As mentioned in the NLLB paper (Costa-Jussà et al., 2022), we observe that translation quality with NLLB models significantly degrades with an increase in input length. Therefore, we translate sentence-by-sentence instead of translating the entire generated text as one input. We find that for splitting sentences properly across languages, we need to use language specific sentence splitters.

2. We use the largest MoE model instead of the smaller dense 3.3B model because anecdotally, it produced fewer degenerate translations (although the reported numbers are comparable – only 1.1% drop in chrf++).

3. We set the translation model to generate greedy generations as the original code base uses greedy decoding[8]. We also experimented with using sampling instead of greedy decoding, and found that it had little to no affect on degeneration.

4. We also handle degeneration wherever we can detect it. Given a translation, we see if there are any repeated phrases or repeated patterns.

[7] This is an open bug, and there is no existing solution as documented here: `https://github.com/google/generative-ai-docs/issues/257`

[8] `https://github.com/facebookresearch/fairseq/blob/nllb/examples/nllb/modeling/evaluation/generate_multi.py`

If they do, we generate a translation by sampling for that particular example with temperature as 0.7, and top_p as 0.95.

C Experimental Details for Measuring Discrepancy in Empathy

This section deals with the experimental details regarding the psychotherapy (empathy) experiments. In this section, we will describe the selection criteria for the empathy evaluator and the evaluation prompts to measure the discrepancy.

C.1 Training and Choosing the Empathy Evaluator

We rely on prior works (Sharma et al., 2020; Gabriel et al., 2024) to motivate our experimental decisions for training the empathy evaluator. Our evaluator detects the presence/absence of the use of an empathy communication mechanism by a respondent for a certain seeker post. Following Sharma et al. (2020)'s EPITOME framework, we evaluate for three communication mechanisms: emotional reaction (ER), interpretation (IP), and exploration (EX).

Data splits. Since the authors do not publicly share the original splits, we generated our own train-dev-test split using the ratios (75-5-20) mentioned in Sharma et al. (2020). These splits are released with the project code base for reproducibility. Note that we create three train-dev-test splits, one each for a kind of empathy communication mechanism. However, the splits are generated at the seeker prompt level; that is, if a certain seeker prompt occurs in the train split for ER empathy communication detection, it will remain in the train split for IP and EX empathy communication detection. We take this decision to avoid test-time leakage. As a result, stratified sampling to maintain the exact distribution between splits becomes a difficult problem. However, our random splits show a similar distribution across the splits (Table 5). The drastic differences in ratios across the empathy communication mechanisms are not relevant since we train different models for each of these mechanisms.

Models selection and hyper-parameter tuning. We adapt Sharma et al. (2020)'s codebase for our evaluator training with the above-mentioned splits. (Sharma et al., 2020) use a bi-encoder model to detect empathetic communication. For the encoder, we consider the following models: `RoBERTa-base`, `RoBERTa-large`, `ModernBERT-base`, and `ModernBERT-large`. The original work uses `RoBERTa-base`. We consider the `ModernBERT` model family for its faster training and larger context window (8192 vs RoBERTa's 512). We tune the learning rate for the following values: {1e-5, 2.5e-5, 5e-5, 7.5e-5, 1e-4} on the dev split. Similarly, we tune the epochs for the following values: {5, 10, 15, 20}. We find that `ModernBERT-large` for the learning rate 2.5e-5 and ten epochs for ER and IP to be the best configuration. For EX, we find that the same model with 5.5e-5 and five epochs to work the best. We train the final classifier with this configuration on the combined train-dev set. The performance of the final evaluator is mentioned in Table 6.

Comm. Mechanism	Train	Dev	Test
ER	33.7	31.2	35.5
IP	47.2	48.1	47.5
EX	15.8	13.0	15.4

Table 5: Proportion of datapoints (in %) where a certain empathy communication mechanism was used across the splits for the three empathy mechanisms: emotional reaction (ER), interpretation (IP), and exploration (EX)

Comm. Mechanism	F1	Accuracy
ER	86.1	87.2
IP	87.3	87.4
EX	92.0	95.8

Table 6: Performance (in %) of the evaluator for detecting the use of the three communication mechanisms: emotional reaction (ER), interpretation (IP), and exploration (EX)

Comparison with GPT-4o. We compare this bi-encoder based evaluator against a 9-shot inference with GPT-4o. We find that GPT-4o significantly underperforms when compared to the bi-encoder model. The results are mentioned in Table 7. The prompt template will be released with the codebase.

C.2 Choice of Evaluation Prompts

We choose 100 examples from the test split to serve as the evaluation set for computing discrepancy. We carefully select the prompts so that they are stratified across the two labels. Note that it does not matter if we stratify as we are only interested in the seeker utterances and labeling is

	F1		Accuracy	
Comm. Mechanism	**Bi-Encoder**	**GPT-4o**	**Bi-Encoder**	**GPT-4o**
ER		66.8		66.8
IP		59.8		61.1
EX		70.1		78.0

Table 7: Performance (in %) comparison between GPT-4o and the best-performing bi-encoder model.

performed in conjunction with the human response to the prompt. However, we choose to stratify to reduce any chance of spurious correlations (Gururangan et al., 2018). As explained in the previous subsection, since the stratification is performed at the seeker post level, perfect stratification is not possible.

D Discrepancy Examples

In this section, we have compiled examples where the model generates disparate information for the same prompt in different languages.

D.1 Open-ended Generation Example

An example of generations from GPT-4o for the prompt, "*What were the roles of women in political leadership across pre-colonial African societies?*", in English, Indonesian, and Korean are shown in Tables 8, 9, and 10, respectively. We observe that responses to the prompts in Indonesian and Korean to be much more concise than their English counterpart. Our metric aims to capture this discrepancy of information.

E Other Experimental Details

We define other useful details related to our experiments and results in this section.

E.1 Language Grouping by Scripts

The language grouping by scripts as used in Table 4 are provided in Table 11.

E.2 Language Grouping by Scripts

The language grouping by scripts as used in Table 4 are provided in table Table 12.

E.3 Details about Human Study

For each language, we recruit 10 annotators through prolific and ask each of them to label ten examples each (a total of 100 examples per language). During our pilot study, we found that comparing translations of full long-form responses is a cognitively challenging task. Therefore, we create small snippets of text (usually 1-4 sentences) and present annotators with text in a given language and its corresponding translation. The study was run through qualtrics. Each HIT (10 responses) was expected to take a 3-4 minutes, and we paid annotators at the rate of 12$ per hour (roughly 0.80$ per HIT). There are some languages for which we finding ten unique annotators would have been hard (the prolific platform provides an estimate of the number of people available for a language). For these languages, we asked annotators to perform multiple HITs with a maximum of 4 per annotator. See Fig. 2 for an example annotation screen.

E.4 Sentence Tokenizers Used.

As mentioned before, we translate the multilingual generations one sentence at a time. This necessitates using good, robust sentence tokenizers for each of the 30 languages we considered. For English, German, Spanish, French, Italian, Dutch, Portuguese, Russian, Chinese, Japanese, Korean, Greek, Romanian, Polish, and Indonesian, we used the tokenizers provide by the spacy toolkit. For English and Chinese, we use the `core_web_sm` tokenizers, and for all others we used the `core_news_sm` tokenizers.

For Indian languages like Bengali, Gujarati, Hindi, Kannada, Malayalam, Punjabi, Tamil, and Telugu, we used the tokenizers provided by the IndicNLP toolkit. For Urdu, we used the tokenizer provided by urduhack [9]. For Vietnamese, we sed the tokenizer from the `underthesea` library.

For all the other languages, we used the Custom tokenizers provided by spacy for each language and use the `sentencizer` for further rule-based sentence tokenizer.

We manually verified that each of these sentence tokenizers worked as expected.

[9] `https://github.com/urduhack/urduhack/blob/master/urduhack/tokenization/eos.py`

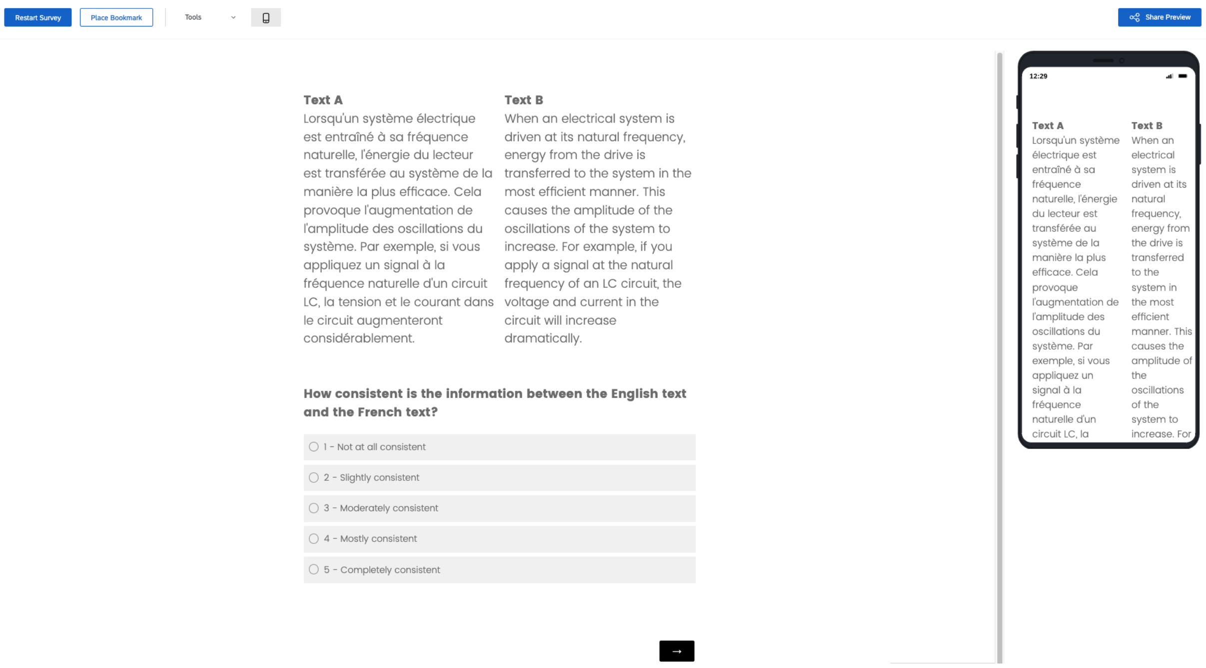

Figure 2: Annotation screen used in the user study. The annotator is asked to rate the translation quality on a Likert scale of 1-5, and higher number indicates better consistency.

Prompt	What were the roles of women in political leadership across pre-colonial African societies?
Generation	In pre-colonial African societies, women played diverse and often significant roles in political leadership, though the extent and nature of their influence varied by region, culture, and historical period. Key roles included: 1. **Queens and Queen Mothers (Matriarchs):** * In several African societies, women held formal authority as sovereigns or co-rulers. For example: * **Nzinga Mbande** of Ndongo and Matamba (Angola) was a warrior queen who led resistance against Portuguese colonization in the 17th century. * The **Queen Mothers of the Asante** (modern-day Ghana), such as Yaa Asantewaa, were powerful political and military leaders. * Among the **Lovedu of South Africa**, the Rain Queen (Modjadji) held spiritual and political significance, believed to control rainfall. 2. **Advisors and Regents:** * Women often acted as advisors to kings or served as regents during the minority of a male heir. In many cases, Queen Mothers (e.g., in the Oyo and Benin Kingdoms) had substantial influence in court affairs and succession decisions. 3. **Leaders in Matrilineal Societies:** * In matrilineal systems, such as among the Akan and Bemba, lineage and inheritance passed through the mother's line, giving women and maternal uncles significant political influence in succession and property decisions. 4. **Spiritual and Religious Leaders:** * Women often held high-ranking spiritual positions that conferred political influence. For example, female priestesses and diviners (e.g., the Igbo *dibia*) mediated between the spiritual and political realms. 5. **Military Leadership:** * In some societies, women served as military commanders or warriors. The **Dahomey Amazons** (Fon people of present-day Benin) were an all-female military regiment under the king's command, demonstrating institutionalized female power in warfare. 6. **Village and Clan Leadership:** * At the local level, women frequently served as heads of lineages, clan leaders, or representatives in councils, particularly in societies with decentralized political structures like the Igbo and Yoruba. In sum, pre-colonial African societies often recognized women's authority in both formal and informal governance structures, contradicting colonial-era and Western portrayals of uniformly patriarchal systems. Would you like examples from a specific region or society?

Table 8: GPT-4o generation for the given English prompt.

Prompt	Apa peran wanita dalam kepemimpinan politik di masyarakat Afrika pra-kolonial?
Generation	Dalam masyarakat Afrika pra-kolonial, wanita memainkan peran penting dalam kepemimpinan politik, meskipun peran tersebut berbeda-beda antar wilayah dan suku. Beberapa peran utama mereka meliputi: 1. **Pemimpin politik dan militer** – Beberapa wanita memegang kekuasaan tertinggi, seperti *Yaa Asantewaa* dari Kekaisaran Asante (Ghana) yang memimpin perang melawan kolonial Inggris, atau *Amina dari Zazzau* (Nigeria) yang dikenal sebagai ratu pejuang dan penakluk. 2. **Ratu ibu (Queen Mothers)** – Di banyak kerajaan seperti Ashanti dan Buganda, ratu ibu memiliki otoritas signifikan, termasuk pengaruh dalam penunjukan raja, nasihat politik, dan pengelolaan wilayah atau istana kerajaan. 3. **Kepala klan atau suku** – Dalam masyarakat matrilineal seperti suku Akan, wanita tua dalam klan (biasanya nenek atau ibu kepala) sering menjadi penentu garis keturunan, warisan, dan keputusan penting komunitas. 4. **Mediator dan penasihat** – Wanita sering memainkan peran sebagai penasihat raja atau dewan, serta mediator dalam konflik antar-kelompok atau keluarga. Peran ini menunjukkan bahwa kepemimpinan perempuan bukanlah hal yang asing atau sekunder, melainkan integral dalam struktur sosial dan politik pra-kolonial Afrika. Apakah kamu tertarik pada wilayah atau kerajaan tertentu di Afrika?
Generation^{trans}	In pre-colonial African societies, women played important roles in political leadership, although these roles varied across regions and tribes. Some of their key roles included: 1. **Political and military leaders** – Some women held supreme power, such as *Yaa Asantewaa* of the Asante Empire (Ghana) who led the war against British colonial rule, or *Amina of Zazzau* (Nigeria) who was known as a warrior queen and conqueror. 2. **Queen Mothers** – In many kingdoms such as Ashanti and Buganda, queen mothers wielded significant authority, including influence over king appointments, political advice, and the management of the royal court or territory. 3. **Clan or tribal chiefs** – In matrilineal societies such as the Akan, the elder women in the clan (usually the grandmother or matriarch) often determined lineages, inheritance, and important community decisions. 4. **Mediators and advisors** – Women often played roles as advisors to the king or council, as well as mediators in conflicts between groups or families. This role shows that female leadership was not an extraneous or secondary thing, but rather integral to the social and political structures of pre-colonial Africa. Are you interested in a particular region or kingdom in Africa?

Table 9: GPT-4o generation for the Indonesian translation of the English prompt: "*What were the roles of women in political leadership across pre-colonial African societies?*"

Prompt	식민지 이전 아프리카 사회에서 여성의 정치적 리더십 역할은 무엇이었나요?
Generation	식민지 이전 아프리카 사회에서 여성은 다양한 방식으로 정치적 리더십을 발휘했습니다. 이들은 단순한 조력자에 그치지 않고, 왕비, 여왕, 여사제, 장로, 또는 군사 지도자로 활약하며 중요한 권력을 행사했습니다. 여성 리더들은 종종 영적 권위와 정치 권력을 겸비하여 지역 공동체의 의사결정, 분쟁 조정, 상징적 통치에 깊이 관여했습니다. 예를 들어: * 다호메이 왕국(현 베냉)에서는 여성 전사 집단인 '아마존'이 존재했고, 이들은 왕을 호위하고 전쟁에 참여하는 핵심 전투 부대였습니다. * 아샨티 왕국(현 가나)에서는 여성 지도자인 '퀸 마더(Queen Mother)'가 국왕을 조언하고 후계자 선정에 중요한 역할을 했습니다. * 이보족 사회(현 나이지리아)에서는 여성들이 '오무아다' 또는 '움우나'와 같은 여성 단체를 조직해 공동체 내 정의 실현과 여성 권익을 위한 자율적 정치 기구를 운영했습니다. 이처럼 여성은 공동체의 생존과 통치를 위한 핵심 축이었으며, 식민주의 이전의 아프리카에서는 정치 권력과 젠더가 더 유연하게 구성되는 경우도 많았습니다. 궁금한 특정 사회나 인물에 대해 더 알아보고 싶으신가요?
Generation^{trans}	In pre-colonial African societies, women exercised political leadership in a variety of ways. They were not just helpers; they were also queens, priestesses, elders, and military leaders, exercising significant power. Often, women leaders held both spiritual and political authority, and were deeply involved in community decision-making, conflict resolution, and symbolic governance. For example: * In the Kingdom of Dahomey (now Benin) there was a group of female warriors called the Amazons, who were a key fighting force that guarded the king and fought in wars. * In the Kingdom of Ashanti (now Ghana) a female leader, the Queen Mother, advised the king and played a key role in selecting a successor. * In the Igbo community (now Nigeria) women organized women's groups such as the Omuada or Um'una to operate autonomous political institutions for justice and women's rights within the community. In this way, women were the linchpin of the survival and governance of the community, and in pre-colonial Africa, political power and gender were often more flexibly structured. Interested in learning more about a particular society or person?

Table 10: GPT-4o generation for the Korean translation of the English prompt: "*What were the roles of women in political leadership across pre-colonial African societies?*"

Script Group	Languages
Latin	`cs,de,es,fr,id,it,nl,pl,pt,ro,tr,vi`
Indic	`bn,gu,hi,kn,ml,pa,ta,te`
PArb	`ar,fa,ur`
Oth1	`el,iw,ja,ko,ru,th,zh`

Table 11: Language Groups by script. PArb refers to the Perso-Arabic scripts; Oth1 groups all other scripts.

Language Family	Languages
Rom	de,es,fr,it,nl,pt,ro
Slav	cs,pl,ru
IA	bn,fa,gu,hi,pa,ur
Drav	kn,ml,ta,te
Oth2	ar,el,id,iw,ja,ko,th,tr,vi,zh

Table 12: Language Groups by language families. Rom: Romance; Slav:Slavic, IA:Indo-Aryan, Drav:Dravidian, Oth2:Other language families

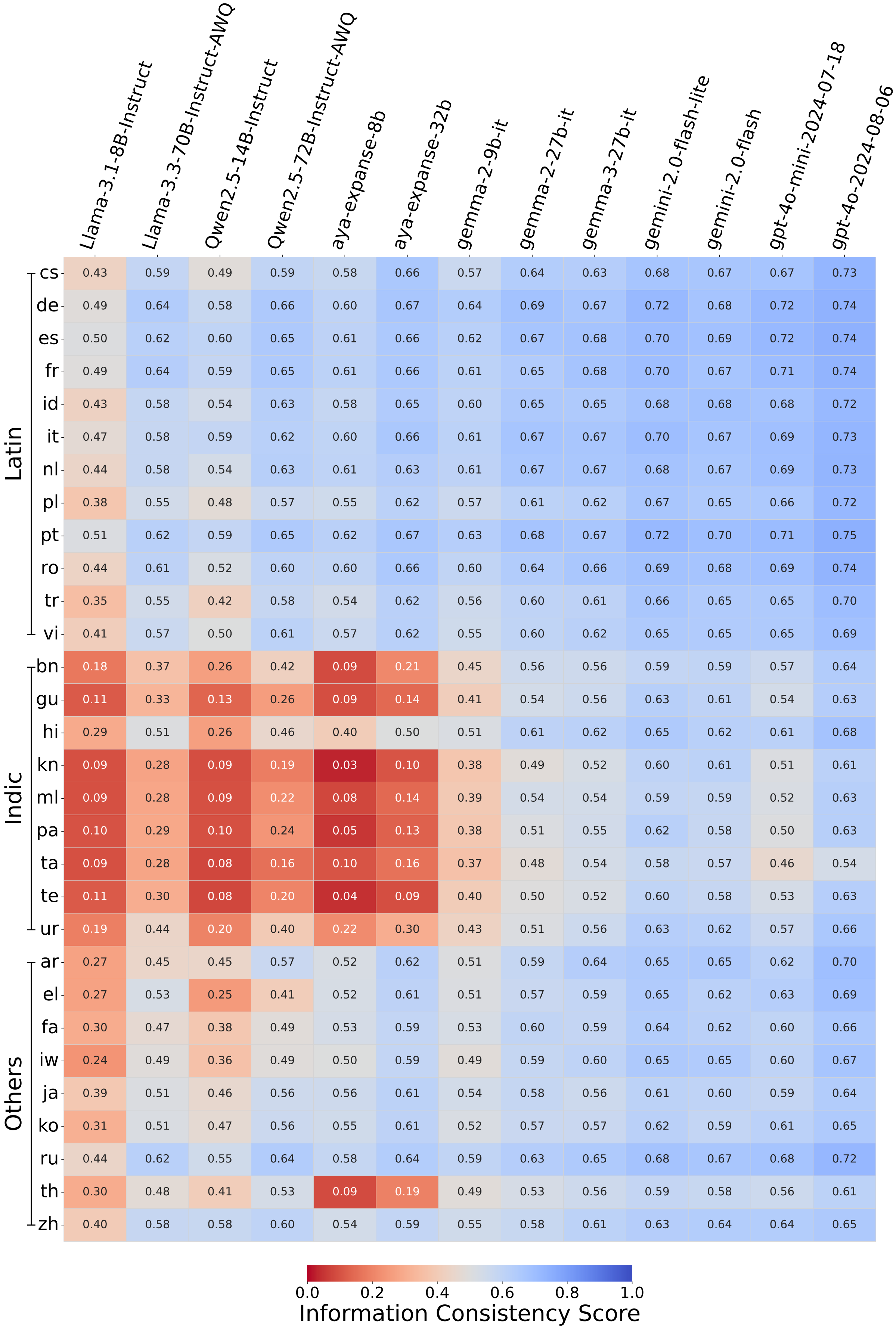

Figure 3: Cross-lingual information consistency scores. The languages are grouped roughly by their respective script family (Latin, Indic or others).

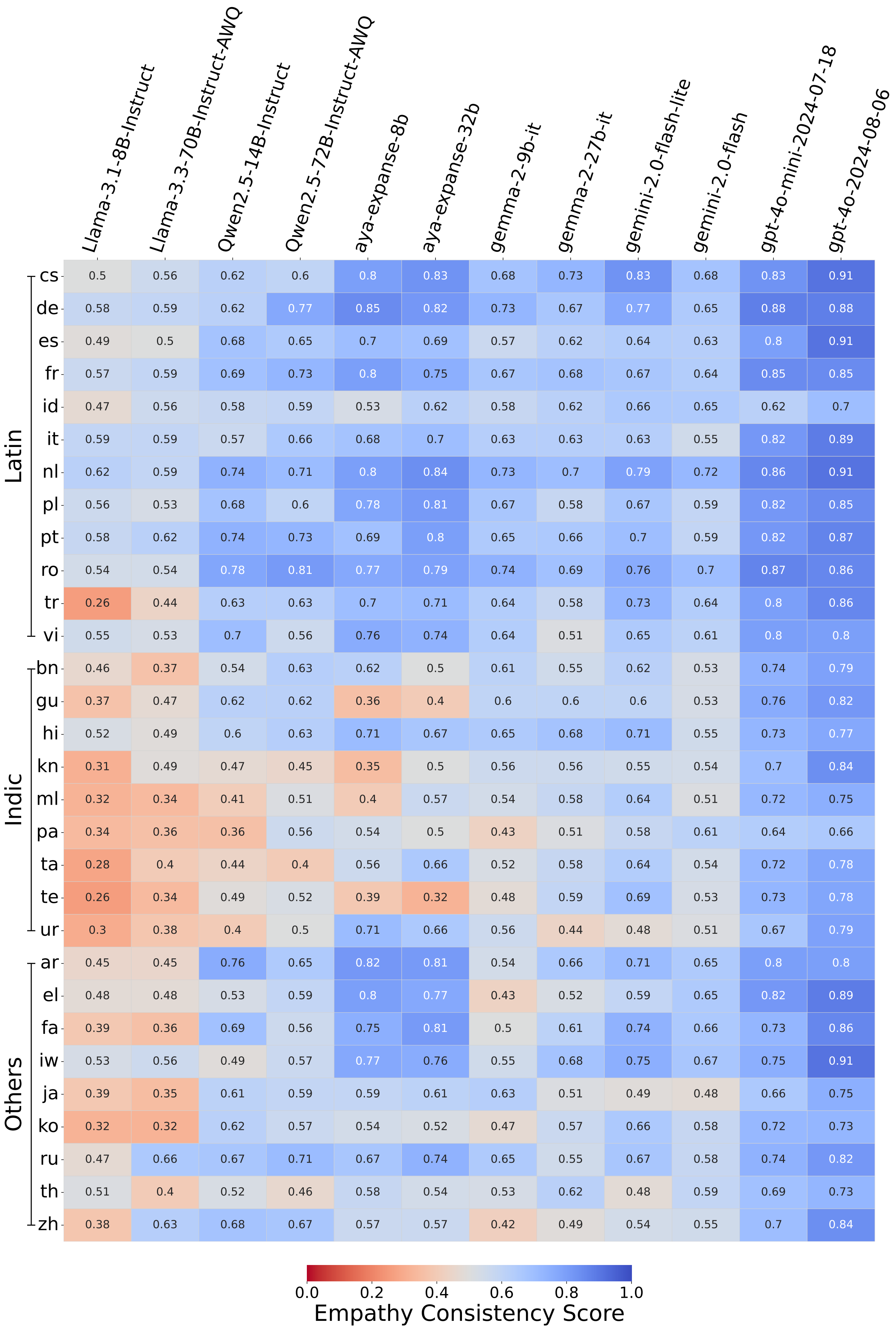

Figure 4: Cross-lingual empathy consistency scores. The languages are grouped roughly by their respective script family (Latin, Indic or others).