Tailored Emotional LLM-Supporter: Enhancing Cultural Sensitivity

Anonymous ACL submission

Abstract

Large language models (LLMs) show promise in offering emotional support and generating empathetic responses for individuals in distress, but their ability to deliver culturally sensitive support remains underexplored due to lack of resources. In this work, we introduce **CultureCare**, the first dataset designed for this task, spanning four cultures and including 1729 distress messages, 1523 cultural signals, and 1041 support strategies with fine-grained emotional and cultural annotations. Leveraging **CultureCare**, we (i) develop and test four adaptation strategies for guiding three state-of-the-art LLMs toward culturally sensitive responses; (ii) conduct comprehensive evaluations using LLM judges, in-culture human annotators, and clinical psychologists; (iii) show that adapted LLMs outperform anonymous online peer responses, and that simple cultural role-play is insufficient for cultural sensitivity; and (iv) explore the application of LLMs in clinical training, where experts highlight their potential in fostering cultural competence in future therapists. Github: ANONYMOUS.[1]

1 Introduction

Large language models (LLMs) have shown growing potential in offering emotional support, with recent work demonstrating that LLMs can provide empathetic, contextually relevant responses for individuals experiencing distress (Zheng et al., 2024; Zhan et al., 2024; Ye et al., 2025). This emerging capability is particularly promising in online spaces, where peer support communities play a vital role in helping individuals navigate emotional challenges. However, culture shapes human emotional experiences and influences the stressors people encounter in daily life (Markus and Kitayama, 1991; Mesquita and Frijda, 1992; Chun et al., 2006; Mesquita et al., 2017). As a result, one critical dimension remains underexplored—the *cultural sensitivity* of these LLMs' support responses.

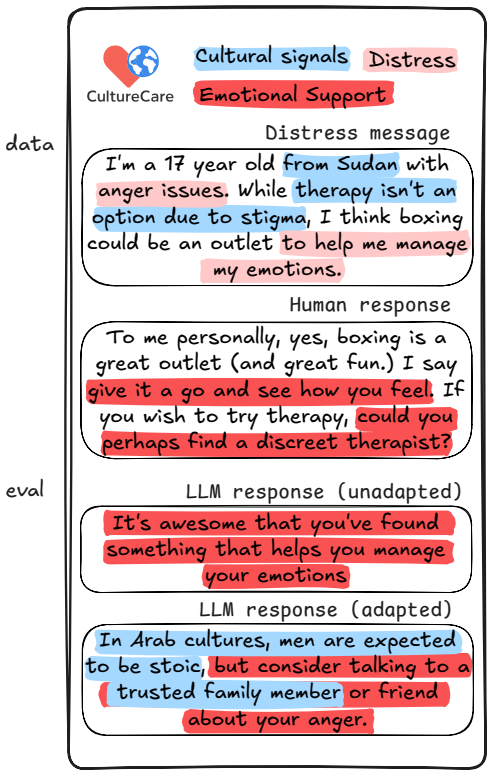


Figure 1: **CultureCare**: 1. The "data" panel shows a Reddit post span-annotated for emotional distress and cultural signals. Every post is paired with its top human response, span-annotated for emotional support messages, and cultural signals. 2. The "eval" panel shows the responses to the post returned by LLMs, with and without cultural adaptation, respectively.

Effective emotional support is deeply shaped by cultural knowledge, unspoken assumptions, norms, and values that influence how distress is expressed and how support is received (Taylor et al., 2007; Matsumoto et al., 2008; Kim et al., 2008). However, when LLMs are used as tools for emotional support, they may struggle to recognize these cul-

[1]Content Warning: This paper includes examples that some readers may find offensive or triggering. These instances are presented for research purposes only and do not represent the views of the authors or affiliated institutions.

turally embedded cues of distress in the first place (Aleem et al., 2024). Even well-intentioned responses can cause harm or alienation without cultural grounding and sensitivity, rather than providing the support users seek (Lissak et al., 2024; Moore et al., 2025). For example, in a collectivist culture, an LLM advising a user to cut ties with their family for personal happiness, without acknowledging the cultural weight of familial duty, may seem offensive or immoral. Similarly, in cultures where mental health is heavily stigmatized, bluntly urging someone to "seek therapy" might intensify feelings of shame or social isolation rather than offering relief.

Recent research tries to address this issue through a case study on Pakistani culture (Aleem et al., 2024). However, the study offers limited generalizability due to its narrow set of manually created distress scenarios, leaving the broader challenge of culturally sensitive emotional support largely unexplored. The lack of suitable datasets and evaluation of adaptation methods has hindered progress in this area. Therefore, here, we present a comprehensive multi-cultural investigation into LLMs' ability to provide culturally sensitive emotional support.

To address the data gap, we introduce **CultureCare**. To the best of our knowledge, this is the first dataset designed to support the study of culturally-sensitive peer emotional support (Figure 1). **CultureCare** spans four distinct global cultures, namely Arabic, Chinese, German, and Jewish, and fine-grained annotated for both emotional support strategies and culturally-relevant signals, as shown in Table 1. The dataset consists of distress-response pairs sourced from real-world support interactions from Reddit, annotated for support type and cultural signals, allowing for nuanced development and evaluations for adapting LLMs' responses across cultural contexts. Using **CultureCare**, we evaluate three state-of-the-art LLMs with tailored prompting strategies for adaptation. Through both automated evaluations and in-culture human evaluations, we find that while incorporating basic cultural information helps, a more effective adaptation requires detailed guidelines and attention to contextualized, explicit cultural signals. These LLM-generated responses are often better than human responses online.

While our primary focus is on peer support—where the LLM acts as a supportive peer rather than providing professional help—we also explore its potential in clinical training settings, namely in training psychology students for culturally sensitive therapy. Expert feedback highlights strong safety and promising utility.

To sum up, our contributions are: First, we release the first dataset, **CultureCare**, for evaluating and adapting LLMs in culturally-sensitive emotional support, spanning four cultures with fine-grained annotations. The dataset comprises 1729 annotated distress messages, 1523 cultural signals, and 1041 support strategies; second, we develop and test four **adaptation strategies** to guide three popular state-of-the-art LLMs toward generating culturally-sensitive support responses; and third, we provide **comprehensive evaluations** involving a LLM judge, in-culture human evaluators, and clinical psychologists to assess both emotional and cultural quality of generated responses.

2 Related Work

Culturally adapted LLMs. Recent research has found that LLMs predominantly reflect the perspectives of WEIRD (Western, Educated, Industrialized, Rich, and Democratic, Henrich et al. 2010) populations without any adaptation (Atari et al., 2023; Johnson et al., 2022). Several studies attempted to address this issue, focusing on diverse tasks such as value alignment (AlKhamissi et al., 2024; Liu et al., 2025b) or hate speech classification (Zhou et al., 2023; Li et al., 2024; Adilazuarda et al., 2025). AlKhamissi et al. (2024); Tao et al. (2024) demonstrated that prompting LLMs with cultural and persona-specific information can effectively align models with diverse cultural values. However, existing work has not examined the effectiveness of these prompting methods for culturally aware emotional support—an important gap this study addresses.

Culture and mental health. Culturally sensitive counselling is a well-established consideration in clinical psychology and healthcare settings (Bernal et al., 1995; Resnicow et al., 1999; Kreuter and McClure, 2004; Taylor et al., 2007; Tao et al., 2015, among others). Prior research has explored various aspects of incorporating cultural sensitivity in practical domains outside of NLP, including the importance of cultural humility in improving therapy outcomes (Owen et al., 2016), disparities in engagement and follow-up care across demographics (Zeber et al., 2017), and the need to embed

	Real?	Cultures	Size	Annotations	Eval. Aspects
Liu et al. (2021, ESConv)	✓	—	1053	DM, SS, E	E
Zheng et al. (2024, ExTES)	✗	—	11177†	DM, SS	E
Zhang et al. (2024, FEEL)	Mix	—	200	DM, SS	E
Aleem et al. (2024)	✗	Pakistani	7	—	C
CultureCare	✓	Arabic, Chinese, German, Jewish	462	DM, SS, Cultural Signals	E, C

Table 1: Comparison with existing work on LLMs for emotional support. **E**: Emotion, **C**: Culture, **DM**: Distress messages, **SS**: Support strategies. †: Zheng et al. (2024) contains a subset of 101 dialogues on cultural identity and belonging; however, culture is not a focus of their work. **CultureCare** uniquely focuses on *culture*, explicitly annotates spans which include cultural signals and assign their types, and evaluates emotional support responses from both emotional and cultural perspectives. The dataset comprises 462 posts along with 4,293 corresponding annotations, as detailed in Table 2.

cultural competence in training programs (Benuto et al., 2018). However, the application of LLMs in culturally sensitive mental health remains limited. Focusing on formal therapy settings, recent work (Abbasi et al., 2025; Kim et al., 2025) explores LLM-generated synthetic multi-turn clinical conversations in multilingual contexts. Their focus on clinical therapy and synthetic data generation differs from ours, which centers on *peer* emotional support and adaptation strategies.

LLMs for emotional support. Existing work has shown that LLMs can provide empathetic and supportive responses when appropriately guided (Zhan et al., 2024). Studies such as Liu et al. (2021); Zheng et al. (2024); Zhang et al. (2024) examine how LLMs respond to distress messages using various support strategies. However, they largely overlook the influence of cultural context in shaping emotional needs and support preferences. While Aleem et al. (2024) considers cultural context, its focus on a small set of scenarios from Pakistani culture limits its generalizability.

Furthermore, recent studies on LLMs in cognitive behavioural therapy (Goel et al., 2025; Zhou et al., 2025; Zhang et al., 2025) emphasize clinical settings with structured treatment frameworks, offering little attention to cultural nuance. In contrast, our work focuses on *peer-to-peer emotional support*, where cultural understanding is essential for effective and supportive communication. We address this gap by incorporating cultural signals from diverse cultural communities into our annotation process, and by assessing emotional support responses from humans and LLMs, based not only on empathy but also on their cultural relevance.

3 CultureCare

We present **CultureCare**, a multi-cultural dataset with fine-grained span-level annotation of cultural signals and emotional distress.[2]. This research is approved by the ethics committee of [ANONYMOUS]. Figure 3 in appendix provides illustration for **CultureCare**'s construction pipeline. Briefly, we first collect culturally relevant mental health data from Reddit using targeted search queries. We then apply a combination of rule-based and LLM-based filters to remove noise. Finally, in-culture annotators and reviewers select the spans and label the data for emotional distress, cultural signals, and support strategies. The annotation categories for each dimension are detailed in Appendix A.4, along with guidelines and measures for content sensitivity.

Table 1 compares **CultureCare** with existing resources. **CultureCare** is distinctive in its focus on multiple cultures, the explicit annotation of cultural signals and emotional distress.

3.1 Data Collection and Filtering

Candidate posts. We use Reddit as our dataset source due to its global user base and peer-driven mental health discussions. Our data is collected via the Reddit API[3]. We focus on two setups: 1. searching mental health subreddits with culture-specific keywords, and 2. searching culture-specific subreddits with mental health keywords. The list of

[2]We deliberately chose not to use language as the defining boundary of culture, recognizing that culturally influenced distress can be expressed in any language in online communities. As a result, **CultureCare** includes posts both in English and in the native languages associated with each culture. The language distribution varies by culture; see Appendix A.5.

[3]https://www.reddit.com/dev/api/
Data collection and annotations are approved by the ethics board.

Category	AR	CH	GE	JE	All
# posts	110	141	119	92	462
# responses	104	126	100	88	418
# distress messages	397	399	402	531	1729
# cultural signals	346	315	338	524	1523
# support strategies	259	242	194	346	1041
# demographic info	226	301	268	131	926
Avg. post length	316	492	690	389	480
Avg. response length	101	80	125	88	98

Table 2: **CultureCare** statistics. **AR**: Arabic, **CH**: Chinese, **GE**: German, and **JE**: Jewish.

subreddits and keywords is in Appendix A.2. For each relevant post, we fetch the top-voted comment as the ideal supportive response, as it typically offers emotional support and resonates with readers, reflected in its high upvotes.

Filtering. Our initial dataset contained 9160 posts, many of which were noisy—e.g., general mental health tips, reactions to global events, or content unrelated to the target cultures. Since manual review of all posts was infeasible, we first applied rule-based filters to remove both explicit noise (e.g., URLs-only posts) and LLM-based filtering to remove implicit noise (e.g., posts lacking *personal* distress). This leaves us with 2671 posts for manual review. The full set of filters is detailed in Appendix A.3. After the final manual review, we retained 462 high-quality posts that clearly express culturally-situated emotional distress, with optional responses by humans. While small, this dataset allows for focused, fine-grained annotation, and the largest of its kind (more discussions in Appendix A.10).

3.2 Annotations and Quality Assurance

Each post was annotated by a in-culture annotator using detailed guidelines (Appendix A.4) identifying spans with emotional distress, cultural signals, and emotional support. On average, each post took approximately 10 minutes to annotate, as they were often lengthy and required attentive reading. To ensure high quality, a second in-culture annotator reviewed the initial annotations and provided comments in cases of disagreement; however, such instances were rare and typically involved additions rather than corrections.

3.3 Analysis of CultureCare

Statistics. **CultureCare** includes 462 posts (90% with responses[4]), containing 1729 annotated distress messages, 1523 cultural signals, and 1041 support strategies. We also extract demographic details (e.g., age, gender, religion) from the posts using an LLM, yielding 926 signals. Example annotated posts are in Appendix F.

Prevalent norms and values. Appendix A.7 illustrates how culture shapes emotional distress, highlighting recurring themes such as family control, mental health invalidation, and gender expectations. Some experiences are culture-specific, including LGBTQ+ rejection in Arabic contexts and financial stress in German ones. While emotional struggles are universal, their expression and root causes depend on the culture.

Cultural signals types and support strategies. Arabic posts emphasize values and explicitly state demographic details, while Chinese and German prioritize norms and morals. Jewish posts focus heavily on concepts, like *rabbi*. Artifacts, knowledge, and language have a minimal presence in all cultures. In human responses, providing suggestions is most common across all cultures. We also observe that while Arabic and Chinese responses emphasize self-disclosure and affirmation, German responses are more balanced, and Jewish responses more often provide information. Asking questions is the least common strategy.

Demographic diversity. We extract detailed demographic information, revealing various patterns between age, gender, and other attributes.[5] For example, extractable Jewish posts are predominantly female (80%), while extractable Arabic and Chinese posts have a high proportion of young adults (47% and 50%, respectively). The extractions also reflect diversity in birthplace, residence, marital status, education level, and profession, reflecting the dataset's rich cultural and social heterogeneity (see Appendix A.6).

Intensity and empathy scores. We additionally provide analysis on cultural differences in distress intensity and response empathy. In our data, in general, Jewish posts show the highest intensity, followed by German and Arabic, while Arabic responses demonstrate the highest empathy. See details in Appendix A.9.

[4]The rest 10% had deleted comments or no comments.

[5]In some cases, only high-level cultural indicators, such as "Arabic" or "Chinese", are extractable.

4 Adaptation Methods

We experiment with basic adaptation strategies using prompts and a compound strategy by combining all strategies, prompts in Appendix C.

Standard `(no suffix)`. By default, we prompt the model to be a Redditor, matching the context of the data. This variation serves as a baseline for comparing adaptation strategies.

Culture-informed role-playing `(+culture)`. Building on prior research (AlKhamissi et al., 2024; Tao et al., 2024), instructing LLM to role-play the cultural background of the person is a simple yet effective method for aligning LLM responses with culturally relevant values. Hence, this could enable more empathetic, appropriate responses, removing the *cultural difference* barrier in empathy (Cikara et al., 2014; Davis, 2018).

Guided principles / constitutions `(+guided)`. Here, we provide guidelines based on CCCI-R (Cross-Cultural Counselling Inventory—Revised; LaFromboise et al. 1991a,b, see Appendix G for details), one of the widely established cross-cultural counselling competency measurements by APA (American Psychological Association). This approach aims to emulate some fundamental competency of professional counselling in terms of cultural sensitivity and awareness in an *implicit cross-cultural* setting.

Explicit cultural signals `(+annotation)`. We explicitly add the data annotations of posts from our dataset to the prompt. The goal of this strategy is to understand whether explicitly providing LLMs with richer contextual information can improve understanding and output in an *implicit cross-cultural* setting.

Combined `(+cga)`. In this method, we combined the above three basic strategies, **c**`ulture`, **g**`uided`, and **a**`nnotation` (Figure 11, Appendix C). We modified the guidelines in the `+guided` strategy by removing CCCI-R items that focus on cross-cultural differences. Here, the LLM will be provided with explicit information and guidelines, as well as role-playing a person from the same culture.

5 Experimental Setup

In this work, we focus our experiment on the open-sourced LLMs. We evaluate 7B-8B parameter models due to their popularity and performance, avoiding the costs of larger models in deployment while remaining competitive, well suited to deploy in agentic systems. Our experiments include: Llama-3.1-8B (Touvron et al., 2023; Dubey et al., 2024), Qwen-2.5-7B (Team, 2024), and Aya-Expanse-8B (Dang et al., 2024). Here, we prioritize open-source model adaptations for scientific reproducibility. We use default configurations for generation in our experiments, and all the adaptation strategies are implemented as system prompts.

5.1 Automatic Evaluations

We perform fine-grained automatic evaluations based on both emotional support and cultural awareness. We use GPT-o3-mini (OpenAI, 2025) as the LLM judge due to its high ability for reasoning and correlation with human judges (Tan et al., 2025; Gu et al., 2024).

Emotional supportiveness measures the basic requirements of an effective supporting message. Based on evaluation criteria from prior research (Rashkin et al., 2019; Liu et al., 2021), we included the following criteria: 1. *Empathy* - the response should demonstrate a genuine understanding of the author's emotions and convey timely, appropriate concern; and 2. *Helpfulness* - the response offers effective advice and tailored, actionable steps.

Cultural awareness measures the awareness and sensitivity of a response concerning cultural aspects. To match the desirable culturally sensitive support in LaFromboise et al. (1991a), three criteria are used: 1. *Socio-political influence* - the response demonstrates an understanding of the current sociopolitical system and its impact on the author of the post; 2. *Knowledge* - the response reflects knowledge about the target culture; and 3. *Cultural context* - the response perceives problem within the appropriate cultural context.

Language quality metrics are also derived from Liu et al. (2021) and LaFromboise et al. (1991a), measures: 1. *Fluency* - the response should be coherent and easy to understand; 2. *Communication* - the response is appropriate.

We evaluate all aspects on a 5-point scale. Following the evaluation prompt-generation method in G-Eval (Liu et al., 2023), we use ChatGPT to create prompts that consist of step-by-step evaluation guidelines. See Appendix C.2 for all prompts.

5.2 Human Evaluations

To further assess the emotional supportiveness and cultural awareness of the responses, we conducted two human evaluations: 1. Crowd evaluations with

Model	Arabic		Chinese		German		Jewish		Average		
	Emo.	Cult.	Emo.	Cult.	Emo.	Cult.	Emo.	Cult.	Emo.	Cult.	All
Aya-Expanse-8B	4.67	3.51	4.43	3.46	4.62	2.54	4.61	3.99	4.58	3.37	3.98
+culture	4.55	3.56	4.48	3.63	4.61	2.75	4.67	4.12	4.58	3.51	4.05
+guided	4.75	3.80	4.73	4.08	4.77	2.72	4.79	4.16	4.76	3.69	4.23
+annotation	4.77	3.73	4.77	3.78	4.80	2.75	4.81	4.10	4.79	3.59	4.19
+cga	4.84	4.39	4.82	4.25	4.84	3.44	4.91	4.55	4.85	4.16	**4.51**
Qwen-2.5-7B	4.16	3.02	4.26	3.20	3.89	2.66	4.27	3.60	4.15	3.12	3.63
+culture	4.05	3.23	4.05	3.38	3.91	2.73	4.28	3.71	4.07	3.26	3.67
+guided	4.41	3.29	4.49	3.42	4.32	2.78	4.54	3.73	4.44	3.30	**3.87**
+annotation	4.40	3.40	4.28	3.46	4.29	2.82	4.46	3.69	4.36	3.34	3.85
+cga	4.11	3.70	4.24	3.81	3.86	2.67	4.50	3.95	4.18	3.53	3.85
Llama-3.1-8B	3.79	2.98	4.20	3.41	3.81	2.67	4.22	3.89	4.00	3.24	3.62
+culture	3.75	3.65	4.11	3.99	3.74	2.62	4.34	4.15	3.99	3.61	3.80
+guided	4.22	3.40	4.57	3.89	4.26	2.54	4.59	4.11	4.41	3.49	3.95
+annotation	4.14	3.52	4.54	3.60	4.23	2.66	4.48	3.99	4.35	3.44	3.89
+cga	4.13	3.93	4.48	4.18	3.98	2.58	4.64	4.38	4.31	3.77	**4.04**

Table 3: Automatic evaluation results for all adaptation strategies and models used in our experiments. The "All" column is the average between emotional supportiveness and cultural awareness.

individuals from the corresponding culture; 2. Expert evaluations to assess the responses for safety and potential usefulness for professionals.

Crowd evaluation. To compare the effectiveness of different strategies, we conducted in-culture crowd-sourced evaluations using Prolific.[6] We recruit two people per culture who are fluent in both English and the matching language of the culture.[7]

To evaluate the best adaptation strategy, we sampled 30 data points per culture for all 3 examined models (a total of 2160 response evaluations), along with the human response and the model-generated responses for all strategies. For each culture, we ask two evaluators to pick the best response in terms of both emotional supportiveness and cultural awareness, and aggregate their ranks of adaptation strategies.

To evaluate the best model, we sampled another 20 data points from one of the top strategies across all models we used, and recruited three annotators per culture for the best-performing model evaluation. Here, we use the same criteria as the previous evaluation. The labels are decided based on a majority vote. If there is a tie, another in-culture annotator is involved to make the judgment.[8] The instructions were given in both English and the native language of the culture. The detailed instructions are in the Appendix E.

Expert evaluation. We collaborated with two psychologists experienced in online emotional support to 1. validate the safety of the adaptation strategies, and 2. assess the utility of adapted LLMs for training psychology students in cross-cultural therapy (Benuto et al., 2018). Together, we developed evaluation guidelines (Appendix H.1). For safety, 100 responses (50 by human, 50 by LLMs) were evaluated, with 15 double-annotated for agreement due to the time-intensive nature of assessment. For utility, 100 LLM-generated responses were evaluated. Agreement results are in Appendix H.2.

6 Results & Discussions

Table 3 presents the automatic evaluation results of emotion supportiveness and cultural awareness metrics, whereas Table 14 (Appendix D) displays the results for overall quality. In general, all models can respond with high fluency and appropriateness.

Showing in Table 3, the winning strategy is `+cga` by combining culture-informed role-playing, guidelines and explicit cultural signal annotations. `+cga` consistently outperforms other strategies across different cultures and models.

Compared to standard role-playing as a Redditor (no suffix in Table 3), simple culture-informed role-playing (`+culture`) does not improve emotional supportiveness consistently (7/12 times scores

[6] https://www.prolific.com/

[7] Most Chinese cultural data comes from individuals in English-speaking countries who identify with Chinese culture, shaping their distress experiences. We selected crowd evaluators to match this context.

[8] This only happens 5% for cultural awareness, 13.33% for emotional supportiveness, indicating that cultural awareness is a less subjective task than emotional supportiveness.

lower). In fact, +culture performs worse than providing explicit, detailed cultural signals or guidelines aimed at *cross-cultural* consultation for emotional supportiveness. Overall, the +guided strategy is more prominent in providing winning emotional support; however, the difference between +annotation and +guided is small.

By explicitly incorporating cultural considerations when offering emotional support (e.g., through +guided, +annotation or +cga), models generate better responses compared to +culture in both evaluation dimensions. Additionally, we observe similar trends reflected in the human evaluation results (details in §6.1). Beyond the scope of our main analysis, we also find that these patterns remain consistent across two 70B-parameter models (Table 15; Appendix D).

Our results here highlight *simple, culture-informed role-playing of a generic person from a culture is likely insufficient to provide context-dependent, culturally sensitive tasks; task-specific guidance or guiding model's attention with explicit signals within the context could offer better help.*

6.1 Human (Crowd) Evaluation

Adapted LLMs can outperform humans in providing responses. Overall, our in-culture evaluators select +cga and +annotation as the best strategies across models and cultures (Table 16). Interestingly, the human response is often ranked as the worst or second-worst strategy, 10 of 12 times for emotional supportiveness. This result further indicates that *LLMs could provide better non-professional emotional support in online communities*, as the average online human responses can be non-helpful, and fail to recognize and address emotional needs.

Human and LLM judge show moderate to strong correlations. We compute the Kendall rank correlation coefficient (τ) between humans' and LLMs' ranking of adaptation strategies (five strategies). The detailed results are in Table 17 (Appendix E.2). In general, human-model correlations are moderate to strong for emotional awareness, except for German culture and Qwen's responses for Jewish culture. Correlations for cultural awareness are weaker and even negative for German culture, despite it being a high-resource culture. Notably, humans and the LLM judge diverge the most on Llama's responses for cultural awareness in Kendall's τ. However, as shown in Table 16 (Appendix E.2), humans and the LLM judge often agree on the best strategy, suggesting that divergences mainly stem from lower-ranked strategies. This highlights the need to improve LLMs as judges of cultural awareness.

No single winning model for all cultures. In-culture evaluators evaluate 20 responses from the $+cga$ strategies of each model. Aya-Expanse-8B leads in cultural awareness across Arabic, Chinese, and German contexts. In emotional awareness, Llama-3.1-8B performs best for Chinese culture, while Aya-Expanse-8B receives the highest ratings from evaluators from Arabic and German cultures. Jewish evaluators prefer Llama-3.1-8B for both cultural awareness and emotional supportiveness.

Our evaluation shows: 1. human-model agrees the majority of the time on the best adaptation strategy; 2. LLM judges show room for improvement in cultural awareness; 3. no single model excels at emotional support across cultures.

6.2 Additional Analysis

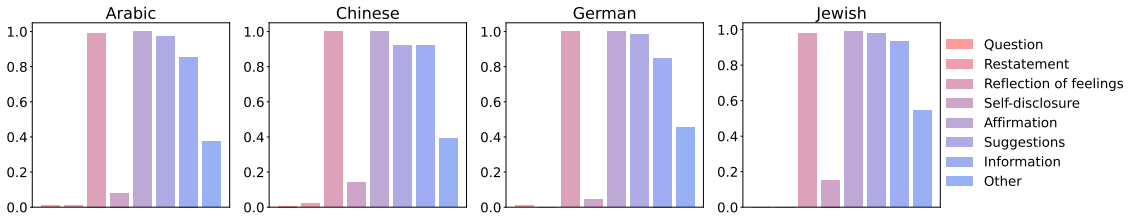
We qualitatively evaluate responses generated by Aya-Expanse-8B, which was preferred by most cultural groups. To analyze the types of emotional support, we use Llama-3.1-70B to categorize responses based on the framework from Liu et al. (2021). Figure 2 shows the distribution of emotional support message categories of Aya-Expanse-8B's responses across cultures for the $+annotation$ strategy. Additional histograms for other adaptation strategies are included in Appendix I.

Overall, model-generated responses use a mix of reflection of feelings, affirmation, suggestions, and information, which is distinctly different from human support online that primarily offer suggestions. This pattern is consistent across cultures and adaptation strategies. Compared to natural human responses, LLM-generated ones tend to be more verbose and structured, particularly when using compound support strategies ($+cga$). Tables 18 and 19 (Appendix F) contain specific response examples.

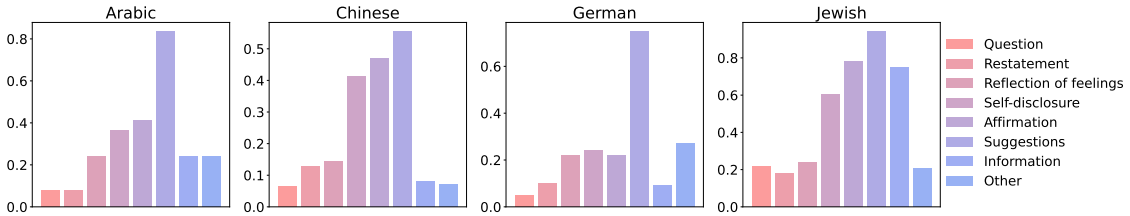
6.3 Evaluation with Clinical Psychologists

Our quantitative results and discussion are in Appendix H.3; key highlights include:

LLMs maintain balanced tone, outperform human responses for cognitive emotional empathy. Overall, experts indicate that responses were neither overly positive (only 2%) nor overly negative



(a) Aya-Expanse-8B, $+annotation$.



(b) Human.

Figure 2: Distribution of emotional support message categories in responses of Aya-Expanse-8B adapted with $+annotation$ versus human responses. The y-axis: % of responses with this support type (definitions in Table 6).

(0%) compared to human responses (6% at both extremes). LLMs outperform humans in cognitive empathy (30% versus 4%) and emotional empathy (34% versus 2%) with the "Strongly agree" rating.

LLMs do not escalate nor introduce new emotional distress. Psychologists agree that LLMs' responses *do not* introduce new distress or escalate existing negative feelings with aggregated positive rating ("Strongly agree" and "Agree") 80% of the time, compared to 28% for human responses.

Adapted responses support psychologist training. A total of 88% of LLM responses were found to be promising for training psychology students in cross-cultural therapy, 46% were rated as strong examples suitable for direct use, while 42% required minor adjustments. This result indicates the high potential utility of adapted models in training.

Open-feedback from experts. Our evaluation reveals key qualitative insights. LLMs often produce structured, culturally sensitive responses that score highly, while human responses are more authentic and spontaneous, raising questions about whether users value emotional genuineness over polished AI outputs. Their consistently empathetic tone (e.g., "You are not alone") also raises an ethical issue: should users be informed they are interacting with AI? At the same time, LLMs can miss clinical red flags, underscoring the need for safeguards and integration of clinical judgment. Finally, LLM outputs are sometimes identifiable as machine-generated, such pattern posing a design dilemma: should models closely mimic humans or maintain a distinctive AI voice, and should this vary across cultural groups (Deshpande et al., 2023; Akbulut et al., 2024; Moore et al., 2025)?

In Sections H.4 and H.5, we provide detailed discussions of the potential and challenges of applying our approach to multi-turn interactions and implementing therapist training in real-world settings, respectively.

7 Conclusion and Future Work

The ability of LLMs to provide culturally sensitive peer emotional support has been largely overlooked. To address this gap, we introduce **CultureCare**, a fine-grained multicultural dataset spanning four cultures, and evaluate three state-of-the-art LLMs with several adaptation strategies. Our findings show that shallow cultural cues are insufficient, while contextualized, guideline-aligned adaptations substantially improve performance. In collaboration with professional psychologists, we further demonstrate the promise of culturally adapted LLMs for training psychology students, highlighting their potential as tools for enhancing emotional support and advancing culturally sensitive counselling. Future work will extend **CultureCare** to multi-turn dialogues, explore broader cultural contexts, and investigate integration into real-world educational and clinical settings.

8 Limitations

Data source. In this paper, we investigate culturally aware emotional support using data from Reddit. We acknowledge that, the platform, its users and annotations may introduce representational biases, and thus do not offer a comprehensive representation of any particular culture. In future work, we aim to collect a more diverse and representative dataset through alternative sources and extensive large scale annotations.

Dataset size. Our dataset contains 462 posts, which is smaller than many generic emotion support datasets. However, it includes over 4,000 fine-grained annotations created with in-culture annotators and quality checkers, making it the largest and most detailed resource of its kind. While its size may limit certain large-scale analyses, we believe the dataset's cultural depth and annotation quality make it highly valuable for evaluating LLMs. We discuss this in detail in Section A.10.

Culture and language. We also attempt to move away from the common practice of using language (which is often determined by data availability) or nationality (which is typically unavailable in anonymous online communities) as a boundary for cultures. While our approach has certain limitations, it also offers a novel contribution by focusing on self-identified cultural identity and on the underlying causes of emotional stress due to cultural factors as expressed by the users themselves.

Beyond single-turn. Our work focuses on single-turn interactions, which are common in online peer-support settings and allow users to provide self-contained emotional and cultural context. This format enables reliable annotation and model evaluation while avoiding fragmented or incomplete multi-turn threads and potential privacy concerns. Although multi-turn interactions are standard in formal psychotherapy, they are beyond the scope of this work. Nevertheless, single-turn posts in **`CultureCare`** can serve as seed cases or personas, and cultural signals such as norms and values can be incorporated into synthetic multi-turn dialogues, supporting richer, more authentic interactions. We discuss this in more detail in Appendix H.4.

Cultural coverage and model bias. We acknowledge that models can exhibit cultural biases, and fully addressing these biases across all cultures remains an open challenge. In this work, we explored adaptation for several cultures (Jewish, Arabic, Chinese, and German), showing that culturally informed prompting and incorporating explicit cultural signals can improve alignment, though these findings may generalize to other cultural contexts only with careful attention and validation.

Ethics Statement

All data used in this study are publicly available posts from Reddit, an anonymous forum, ensuring user identities are not disclosed. Our work only collect posts, and the demographic information is voluntarily provided in the original posts on Reddit. While this work explores promising strategies for automatically providing culturally aware emotional support, we do not recommend using our method directly without human verification and large-scale robustness and safety testing. This research is approved by the ethics committee of ANONYMOUS.

References

Mohammad Amin Abbasi, Farnaz Sadat Mirnezami, and Hassan Naderi. 2025. Hamraz: A culture-based persian conversation dataset for person-centered therapy using llm agents. *ArXiv preprint*, abs/2502.05982.

Muhammad Farid Adilazuarda, Chen Cecilia Liu, Iryna Gurevych, and Alham Fikri Aji. 2025. From surveys to narratives: Rethinking cultural value adaptation in llms. *ArXiv preprint*, abs/2505.16408.

Canfer Akbulut, Laura Weidinger, Arianna Manzini, Iason Gabriel, and Verena Rieser. 2024. All too human? mapping and mitigating the risk from anthropomorphic ai. *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, 7(1):13–26.

Mahwish Aleem, Imama Zahoor, and Mustafa Naseem. 2024. Towards culturally adaptive large language models in mental health: Using chatgpt as a case study. In *Companion Publication of the 2024 Conference on Computer-Supported Cooperative Work and Social Computing*, CSCW Companion '24, page 240–247, New York, NY, USA. Association for Computing Machinery.

Badr AlKhamissi, Muhammad ElNokrashy, Mai Alkhamissi, and Mona Diab. 2024. Investigating cultural alignment of large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12404–12422, Bangkok, Thailand. Association for Computational Linguistics.

Mohammad Atari, Mona J Xue, Peter S Park, Damián Blasi, and Joseph Henrich. 2023. Which humans?

Peter Belcak, Greg Heinrich, Shizhe Diao, Yonggan Fu, Xin Dong, Saurav Muralidharan, Yingyan Celine Lin,

and Pavlo Molchanov. 2025. Small language models are the future of agentic AI. *CoRR*, abs/2506.02153.

Lorraine T Benuto, Jena Casas, and William T O'Donohue. 2018. Training culturally competent psychologists: A systematic review of the training outcome literature. *Training and Education in Professional Psychology*, 12(3):125.

Guillermo Bernal, Janet Bonilla, and Carmen Bellido. 1995. Ecological validity and cultural sensitivity for outcome research: Issues for the cultural adaptation and development of psychosocial treatments with hispanics. *Journal of abnormal child psychology*, 23:67–82.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde De Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, et al. 2021. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*.

Chi-Ah Chun, Rudolf H Moos, and Ruth C Cronkite. 2006. Culture: A fundamental context for the stress and coping paradigm. *Handbook of multicultural perspectives on stress and coping*, pages 29–53.

Mina Cikara, Emile Bruneau, Jay J Van Bavel, and Rebecca Saxe. 2014. Their pain gives us pleasure: How intergroup dynamics shape empathic failures and counter-empathic responses. *Journal of experimental social psychology*, 55:110–125.

John Dang, Shivalika Singh, Daniel D'souza, Arash Ahmadian, Alejandro Salamanca, Madeline Smith, Aidan Peppin, Sungjin Hong, Manoj Govindassamy, Terrence Zhao, Sandra Kublik, Meor Amer, Viraat Aryabumi, Jon Ander Campos, Yi-Chern Tan, Tom Kocmi, Florian Strub, Nathan Grinsztajn, Yannis Flet-Berliac, Acyr Locatelli, Hangyu Lin, Dwarak Talupuru, Bharat Venkitesh, David Cairuz, Bowen Yang, Tim Chung, Wei-Yin Ko, Sylvie Shang Shi, Amir Shukayev, Sammie Bae, Aleksandra Piktus, Roman Castagné, Felipe Cruz-Salinas, Eddie Kim, Lucas Crawhall-Stein, Adrien Morisot, Sudip Roy, Phil Blunsom, Ivan Zhang, Aidan Gomez, Nick Frosst, Marzieh Fadaee, Beyza Ermis, Ahmet Üstün, and Sara Hooker. 2024. Aya expanse: Combining research breakthroughs for a new multilingual frontier. *ArXiv preprint*, abs/2412.04261.

Mark H Davis. 2018. *Empathy: A social psychological approach*. Routledge.

Ameet Deshpande, Tanmay Rajpurohit, Karthik Narasimhan, and Ashwin Kalyan. 2023. Anthropomorphization of AI: Opportunities and risks. In *Proceedings of the Natural Legal Language Processing Workshop 2023*, pages 1–7, Singapore. Association for Computational Linguistics.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurélien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Rozière, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Graeme Nail, Grégoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel M. Kloumann, Ishan Misra, Ivan Evtimov, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, and et al. 2024. The llama 3 herd of models. *ArXiv preprint*, abs/2407.21783.

Anmol Goel, Nico Daheim, Christian Montag, and Iryna Gurevych. 2025. Socratic reasoning improves positive text rewriting. In *Proceedings of the 10th Workshop on Computational Linguistics and Clinical Psychology (CLPsych 2025)*, pages 140–156, Albuquerque, New Mexico. Association for Computational Linguistics.

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, Yuanzhuo Wang, and Jian Guo. 2024. A survey on llm-as-a-judge. *ArXiv preprint*, abs/2411.15594.

Christian Haerpfer, Ronald Inglehart, Alejandro Moreno, Christian Welzel, Kseniya Kizilova, Jaime Diez-Medrano, Marta Lagos, Pippa Norris, Eduard Ponarin, Bjorn Puranen, et al. 2022. World values survey: Round seven – country-pooled datafile version 6.0. *Madrid, Spain & Vienna, Austria: JD Systems Institute & WVSA Secretariat*, 12(10):8.

Joseph Henrich, Steven J Heine, and Ara Norenzayan. 2010. The weirdest people in the world? *Behavioral and brain sciences*, 33(2-3):61–83.

Rebecca L. Johnson, Giada Pistilli, Natalia Menédez-González, Leslye Denisse Dias Duran, Enrico Panai, Julija Kalpokiene, and Donald Jay Bertulfo. 2022. The ghost in the machine has an american accent: value conflict in GPT-3. *ArXiv preprint arXiv:2203.07785v1*.

Heejung S Kim, David K Sherman, and Shelley E Taylor. 2008. Culture and social support. *American psychologist*, 63(6):518.

Hyunjong Kim, Suyeon Lee, Yeongjae Cho, Eunseo Ryu, Yohan Jo, Suran Seong, and Sungzoon Cho. 2025. KMI: A dataset of Korean motivational interviewing dialogues for psychotherapy. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 10803–10828, Albuquerque, New Mexico. Association for Computational Linguistics.

Matthew W Kreuter and Stephanie M McClure. 2004. The role of culture in health communication. *Annu. Rev. Public Health*, 25(1):439–455.

Teresa D LaFromboise, Hardin LK Coleman, and Alexis Hernandez. 1991a. Cross-cultural counseling inventory—revised (CCCI-R) [database record]. *APA PsycTests*.

Teresa D LaFromboise, Hardin LK Coleman, and Alexis Hernandez. 1991b. Development and factor structure of the cross-cultural counseling inventory—revised. *Professional psychology: Research and practice*, 22(5):380.

Cheng Li, Mengzhuo Chen, Jindong Wang, Sunayana Sitaram, and Xing Xie. 2024. Culturellm: Incorporating cultural differences into large language models. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*.

Shir Lissak, Nitay Calderon, Geva Shenkman, Yaakov Ophir, Eyal Fruchter, Anat Brunstein Klomek, and Roi Reichart. 2024. The colorful future of LLMs: Evaluating and improving LLMs as emotional supporters for queer youth. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 2040–2079, Mexico City, Mexico. Association for Computational Linguistics.

Chen Cecilia Liu, Iryna Gurevych, and Anna Korhonen. 2025a. Culturally aware and adapted nlp: A taxonomy and a survey of the state of the art. *Transactions of the Association for Computational Linguistics*, 13:652–689.

Chen Cecilia Liu, Anna Korhonen, and Iryna Gurevych. 2025b. Cultural learning-based culture adaptation of language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3114–3134, Vienna, Austria. Association for Computational Linguistics.

Siyang Liu, Chujie Zheng, Orianna Demasi, Sahand Sabour, Yu Li, Zhou Yu, Yong Jiang, and Minlie Huang. 2021. Towards emotional support dialog systems. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3469–3483, Online. Association for Computational Linguistics.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-eval: NLG evaluation using gpt-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522, Singapore. Association for Computational Linguistics.

Hazel Rose Markus and Shinobu Kitayama. 1991. Culture and the self: Implications for cognition, emotion, and motivation. *Psychological Review*, 98(2):224–253.

David Matsumoto, Seung Hee Yoo, and Johnny Fontaine. 2008. Mapping expressive differences around the world: The relationship between emotional display rules and individualism versus collectivism. *Journal of cross-cultural psychology*, 39(1):55–74.

Batja Mesquita, Michael Boiger, and Jozefien De Leersnyder. 2017. Doing emotions: The role of culture in everyday emotions. *European Review of Social Psychology*, 28(1):95–133.

Batja Mesquita and Nico H Frijda. 1992. Cultural variations in emotions: a review. *Psychological bulletin*, 112(2):179.

Jared Moore, Declan Grabb, William Agnew, Kevin Klyman, Stevie Chancellor, Desmond C. Ong, and Nick Haber. 2025. Expressing stigma and inappropriate responses prevents llms from safely replacing mental health providers. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency, FAccT 2025, Athens, Greece, June 23-26, 2025*, pages 599–627. ACM.

OpenAI. 2025. Openai o3-mini system card.

Jesse Owen, Karen W Tao, Joanna M Drinane, Joshua Hook, Don E Davis, and Natacha Foo Kune. 2016. Client perceptions of therapists' multicultural orientation: Cultural (missed) opportunities and cultural humility. *Professional Psychology: Research and Practice*, 47(1):30.

Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 2019. Towards empathetic open-domain conversation models: A new benchmark and dataset. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5370–5381, Florence, Italy. Association for Computational Linguistics.

Ken Resnicow, Tom Baranowski, Jasjit S. Ahluwalia, and Ronald L. Braithwaite. 1999. Cultural sensitivity in public health: Defined and demystified. *Ethnicity & Disease*, 9(1):10–21.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. 2022. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35:36479–36494.

Sijun Tan, Siyuan Zhuang, Kyle Montgomery, William Yuan Tang, Alejandro Cuadron, Chenguang Wang, Raluca A. Popa, and Ion Stoica. 2025. Judgebench: A benchmark for evaluating llm-based judges. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net.

Karen W Tao, Jesse Owen, Brian T Pace, and Zac E Imel. 2015. A meta-analysis of multicultural competencies and psychotherapy process and outcome. *Journal of counseling psychology*, 62(3):337.

Yan Tao, Olga Viberg, Ryan S Baker, and René F Kizilcec. 2024. Cultural bias and cultural alignment of large language models. *PNAS nexus*, 3(9):pgae346.

Shelley E Taylor, William T Welch, Heejung S Kim, and David K Sherman. 2007. Cultural differences in the impact of social support on psychological and biological stress responses. *Psychological science*, 18(9):831–837.

Qwen Team. 2024. Qwen2.5: A party of foundation models.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurélien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models. *ArXiv preprint*, abs/2302.13971.

Jing Ye, Lu Xiang, Yaping Zhang, and Chengqing Zong. 2025. SweetieChat: A strategy-enhanced role-playing framework for diverse scenarios handling emotional support agent. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 4646–4669, Abu Dhabi, UAE. Association for Computational Linguistics.

John E Zeber, Karen J Coleman, Heidi Fischer, Tae K Yoon, Brian K Ahmedani, Arne Beck, Samuel Hubley, Zac E Imel, Rebecca C Rossom, Susan M Shortreed, et al. 2017. The impact of race and ethnicity on rates of return to psychotherapy for depression. *Depression and anxiety*, 34(12):1157–1163.

Hongli Zhan, Allen Zheng, Yoon Kyung Lee, Jina Suh, Junyi Jessy Li, and Desmond C Ong. 2024. Large language models are capable of offering cognitive reappraisal, if guided. In *First Conference on Language Modeling*.

Huaiwen Zhang, Yu Chen, Ming Wang, and Shi Feng. 2024. Feel: a framework for evaluating emotional support capability with large language models. In *International Conference on Intelligent Computing*, pages 96–107. Springer.

Mian Zhang, Xianjun Yang, Xinlu Zhang, Travis Labrum, Jamie C. Chiu, Shaun M. Eack, Fei Fang, William Yang Wang, and Zhiyu Chen. 2025. CBT-bench: Evaluating large language models on assisting cognitive behavior therapy. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3864–3900, Albuquerque, New Mexico. Association for Computational Linguistics.

Zhonghua Zheng, Lizi Liao, Yang Deng, Libo Qin, and Liqiang Nie. 2024. Self-chats from large language models make small emotional support chatbot better. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11325–11345, Bangkok, Thailand. Association for Computational Linguistics.

Jinfeng Zhou, Yuxuan Chen, Jianing Yin, Yongkang Huang, Yihan Shi, Xikun Zhang, Libiao Peng, Rongsheng Zhang, Tangjie Lv, Zhipeng Hu, Hongning Wang, and Minlie Huang. 2025. Crisp: Cognitive restructuring of negative thoughts through multi-turn supportive dialogues. *ArXiv preprint*, abs/2504.17238.

Li Zhou, Laura Cabello, Yong Cao, and Daniel Hershcovich. 2023. Cross-cultural transfer learning for Chinese offensive language detection. In *Proceedings of the First Workshop on Cross-Cultural Considerations in NLP (C3NLP)*, pages 8–15, Dubrovnik, Croatia. Association for Computational Linguistics.

A Additional Information about `CultureCare`

Figure 3 illustrates our overall dataset creation pipeline.

A.1 Data License

Our experiment code and annotations are publicly available for research evaluations only. Due to Reddit's privacy and data policy, please contact the authors or refer to the README in our project GitHub for details on how to access the dataset texts.

A.2 Lists of Subreddits and Keywords

1. *Search culture subreddits with mental health keywords:*
   - Subreddits: `r/arabs`, `r/algeria`, `r/bahrain`, `r/egypt`, `r/iraq`, `r/jordan`, `r/kuwait`, `r/lebanon`, `r/libya`, `r/morocco`, `r/oman`, `r/palestine`, `r/qatar`, `r/saudiarabia`, `r/somalia`, `r/sudan`, `r/syria`, `r/tunisia`, `r/uae`, `r/yemen`, `r/china`, `r/shanghai`, `r/beijing`, `r/germany`, `r/depression_de`, `r/de`, `r/german`, `r/Israel`, `r/hebrew`, `r/HebrewIsaelis`, `r/Judaism`, `r/Jewish`
   - Keywords: depression, depressed, anxiety, anxious, bipolar, autistic, sad, mental health, trauma, schizophrenia, schizophrenic, anger issues
2. *Search mental health subreddits with culture keywords:*
   - Subreddits: `r/depression`, `r/mmfb`, `r/anxiety`, `r/bipolar`, `r/bpd`, `r/autism`, `r/schizophrenia`, `r/mentalhealth`, `r/traumatoolbox`, `r/socialanxiety`, `r/anger`, `r/offmychest`, `r/bodyacceptance`
   - Keywords: arab, german, chinese, jewish + translated mental health keywords (e.g., traurig for sad in German)

The `restrict_sr` parameter was set to false in the API configuration to prevent query filtering strictly by subreddit. This allows the retrieval of relevant posts from across Reddit *when no matches are found within the targeted subreddit*, increasing data coverage and flexibility.

A.3 Rule-based and LLM-based Filtering

Our rule-based filters removed 46% of the noisy instances. The filters are:

- *F1 - incomplete posts:* the post has a title but empty content.
- *F2 - no personal distress message:* the post does *not* contain a personal emotional distress message, i.e., there is no use of the words "I", "me", etc. Such posts, for example, discuss distress of other people or a tragic event in the news.
- *F3 - incorrect language:* we restrict our data to English and the languages of our four cultures.
- *F4 - more links and numbers than text:* we notice that posts with too many URLs and telephone numbers tend to provide information and pointers on seeking and undestanding mental health and *does not express personal distress.*

Our LLM-based filters removed another 46% of the data instances. We run the filters on Qwen-2.5-72B, The filters are:

- *F5 - no emotional distress detected:* an LLM-based classifier to rule out a post if it does not contain a personal distress message. (an advanced version of F2).
- *F6 - wrong or no cultural context:* we instruct an LLM to rule out a post if it was not culturally relevant for our four cultures.

A.4 Annotation Guidelines

We developed our annotation guidelines over iterations, and the final guidelines are in Figure 4.

Posts. For emotional intensity annotation, we use a 3-point scale for simplicity: 1-light, 2-moderate, 3-high. For detailed definitions and examples, see Table 4.

For cultural signals annotation, we follow a modified version of the taxonomy of cultural elements based on (Liu et al., 2025a), with 7 categories: Concepts, Knowledge, Values, Norms and Morals, Language, Artifacts, and Demographics. For detailed definitions of each category and examples, see Table 5.

Responses. For emotional responses, we annotated both support message strategies and empathy scores.

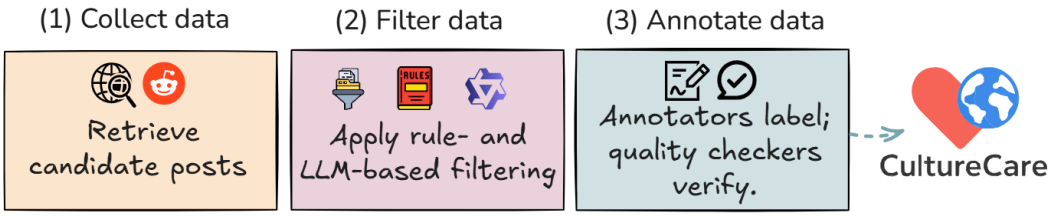


Figure 3: The **CultureCare** dataset construction pipeline: (1) we collect data by querying selected subreddits for candidate post-response pairs; (2) we apply rule-based and LLM-based filters to remove noisy instances (§3.1), e.g., that do not contain cultural signals; (3) in-culture annotators mark spans, in both posts and responses, with emotional distress, cultural signals, and support strategies; finally, a second group of annotators verify the quality of these labels and make corrections when needed.

Following ESConv (Liu et al., 2021), the support message strategies span 8 categories: Questions, Restatement or Paraphrasing, Reflection of Feelings, Self-disclosure, Affirmations and Reassurance, Providing Suggestions, Information, and Others. For detailed definitions of each category and examples, see Table 6.

We used a 5-point scale (Liu et al., 2021) to annotate the empathy scores of the responses. For detailed definitions and examples, see Table 7.

To protect annotators' mental health, we implemented the following measures: 1. Full transparency about the content, with clear warnings and labels for discussions involving mental distress, abuse, PTSD, and related topics. 2. Allowing annotators to terminate the task at any time.

A.5 Language Statistics

In this work, we deliberately chose not to use language as the defining boundary of culture, recognizing that culturally influenced distress can be expressed in any language in online communities. As a result, **CultureCare** includes posts both in English and in the native languages associated with each culture. The language distribution varies by culture, and Table 8 shows the language statistics.

A.6 Demographic Information

We extract demographic information from posts using the prompt in Figure 5 with GPT-4o-Mini and manually validated the results. We follow a set of demographic attributes from the WVS (Haerpfer et al., 2022), which are included in Table 9. In some cases, only high-level cultural indicators, such as "Arabic" or "Chinese" are extractable.

Figure 6 shows the distributions of age and gender in the **CultureCare** dataset, computed over posts where this information is available (50% of the dataset). The gender distribution varies notably across cultures: for instance, Jewish (JE) posts are predominantly female (80%), while German (GE) posts are mostly male (63%). Age distributions also differ; Arabic (AR) and Chinese (CH) have a higher proportion of young adults (47% and 50%, respectively), whereas GE and JE show a more balanced spread across adults and other age groups, with JE uniquely having an equal representation (26%) in teenagers, young adults, adults, and older adults. These variations highlight cultural differences in demographic representation within the dataset. Table 11 summarizes the demographic diversity in the **CultureCare** dataset across the cultures. Notably, the dataset captures a wide variety of cultural backgrounds and life circumstances, from varied geographic origins and residences to diverse education levels and professions, reflecting rich cultural and social heterogeneity within each group.

A.7 Cultural Themes on Emotional Distress and Support

Table 10 highlights how individuals from different cultural backgrounds experience emotional distress through the lens of cultural norms and values, the most prevalent cultural signals in our data (definitions are in Table 5). For example, both Arabic and Chinese distress messages report strong family control and dismissal of mental health issues, while Germans often express emotional exhaustion and social isolation. LGBTQ+ rejection appears uniquely in the Arabic context, while financial stress is more specific to the German entries. Gender expectations are prevalent in Arabic, Chinese, and Jewish contexts. Religious influence also varies significantly: in Arabic and Jewish cultures, religion plays a central role, either as a perceived

CultureCare Annotation guidelines

First, open your assigned annotation sheet. The first two rows of the annotation sheet are reserved for headings and instructions. Do not change them!

Every row contains one post-response pair. Finish one before moving to the next pair.

1. Read the Reddit post written by OP (Author of the post).
2. Identify and rate the intensity of personal emotional distress messages in the post. Details on intensity ratings are in Table 4.
3. Identify and classify the cultural signals in the post. Cultural signals schema is in Table 5.
4. Read the response.
5. Identify and classify the emotional support messages in the response. The possible support strategies are listed in Table 6.
6. Identify and categorize the cultural signals in the response.
7. Rate the empathy in the response. Details are in Table 7.

Figure 4: Instructions for the annotation guidelines.

Rating	Explanation	Example
1 (light)	The emotion is present but subtle, with mild expression or little emphasis.	I had a bit of a rough day.
2 (moderate)	The emotion is clearly expressed, showing a noticeable impact without being overwhelming.	Today was pretty hard; I'm feeling down and could use some cheering up.
3 (high)	The emotion is intense and strongly emphasized, often reflecting deep or overwhelming feelings.	I'm absolutely devastated. I can't believe this happened, and I don't know how to cope.

Table 4: Intensity scale of the post.

cause of distress or as a coping mechanism. Overall, the table reveals that while emotional struggles are universal, the way they are shaped and expressed is deeply rooted in cultural values and social expectations.

A.8 Types of Cultural Signals and Support Strategies per Culture

Figure 7 reveals the differences between the types of cultural signals among the 4 cultures. Each culture prioritizes certain signals over others, reflecting unique cultural characteristics. Arabic posts emphasize values and tend to state culture-specific demographics more explicitly (e.g., *I'm an arab guy from Sudan*). Chinese and German posts prioritize norms and morals, indicating a strong focus on shared principles. In contrast, Jewish posts place a significant emphasis on cultural concepts (e.g., *rabbi*, *Shomer Shabbat*), with norms and morals playing a secondary but still important role. Across all cultures, signals of types artifacts, knowledge, and language have minimal presence.

The figure also illustrates cultural differences in emotional support strategies in human responses. The most prominent strategy in all cultures is providing suggestions. Beyond the top strategy, Arabic and Chinese, both emphasize self-disclosure and affirmation. German strategies are more balanced, and Jewish culture uses the information strategy more than the 3 other cultures. The least common strategy is asking questions.

A.9 Intensity of Emotional Distress and Empathy of Responses

Table 12 shows the average scores for the intensity of distress messages in posts and the empathy of human responses across cultures and overall. Jewish posts exhibit the highest intensity (2.07),

Category	Explanation	Example
Concepts	Basic units of meaning underlying objects, ideas, or beliefs.	Familismo in Mexican culture. It reflects the belief that family comes first, and extended family members are often involved in decisions and provide emotional and practical support.
Knowledge	Information that can be acquired through education or practical experience.	Five pillars of Islam
Values	Beliefs, desirable end states or behaviours ranked by relative importance that can guide evaluations of things.	Respect for elders in Chinese culture
Norms and Morals	Set of rules or principles that govern people's behaviour and everyday reasoning.	"You should behave like a happy and thankful child"
Language	Specific use of slang, speech, dialects.	"That sounds so sick!"
Artifacts	Materialized items as the productions of human culture, they can be forms of art, tools, machines, etc.	The Christian cross
Demographics	Talking about nationality or ethnicity.	"We are a Chinese couple"

Table 5: Categories of cultural signals, adapted from the taxonomy in Liu et al. (2025a).

Strategy	Explanation	Example
Questions	Questions related to mental health status	How often do you feel this way?
Restatement or Paraphrasing	Describe the situation in other words.	Sounds like your family life is really tough.
Reflection of Feelings	Show empathy for the other person. Understand the other's feeling and behaviour.	I understand how anxious you are.
Self-disclosure	Both have made the same experience and share the same feelings.	I feel the same way sometimes.
Affirmation and Reassurance	Help with the other person's uncertainty.	You have done your best and I believe you will get it.
Providing Suggestions	Give some possible solutions.	Find a responsible adult that you can confide in, someone you trust.
Information	Provide proved and useful information that can help to increase the mental well-being.	Apparently, lots of research has found that getting enough sleep before an exam can help students perform better.
Others	Remaining strategies. Includes wishes, hopes, etc.	I hope your luck will change.

Table 6: List of emotional support strategies, based on categories presented in ESConv (Liu et al., 2021).

Rating	Explanation	Example
1 (not empathetic at all)	Shows no interest in others' feelings and may even dismiss or ignore them.	Everyone has problems. Just get over it.
2 (slightly empathetic)	Shows minimal acknowledgement of others' feelings, but lacks genuine concern or involvement.	Oh, that's unfortunate.
3 (moderately empathetic)	Recognizes others' emotions and shows some level of concern, but doesn't fully engage.	That sounds tough. I'm sorry you're dealing with that.
4 (quite empathetic)	Actively listens, responds with understanding, and expresses care for the other person's emotions.	I can see why you feel that way. That must be really hard.
5 (very empathetic)	Fully connects with and validates others' emotions, offering deep understanding and support.	I'm here for you, and I really understand what you're going through. Let me know how I can support you.

Table 7: Empathy scores.

Demographic Information Extraction

You are an expert at structured data extraction. You will be given unstructured text from a Reddit post and should convert it into the given structure. Do not add any information that is not in the text. If there is no information, write 'unknown'. You should extract the following properties and return the output in JSON format:

```
{
  "settlement": "The location where the person lives.",
  "gender": "The gender of the person.",
  "age": "The age of the person.",
  "born_in": "The location where the person was born.",
  "marital_status": "The marital status of the person.",
  "people_in_household": "Household size.",
  "education": "The highest level of education of the person.",
  "profession": "The profession of the person.",
  "employment": "The employment status of the person.",
  "social_class": "The social class of the person.",
  "religion": "The religion or belief system of the person."
}
```

Here is the Reddit post: `{post_text}`
Response:

Figure 5: Prompt for extracting demographic information from **CultureCare**.

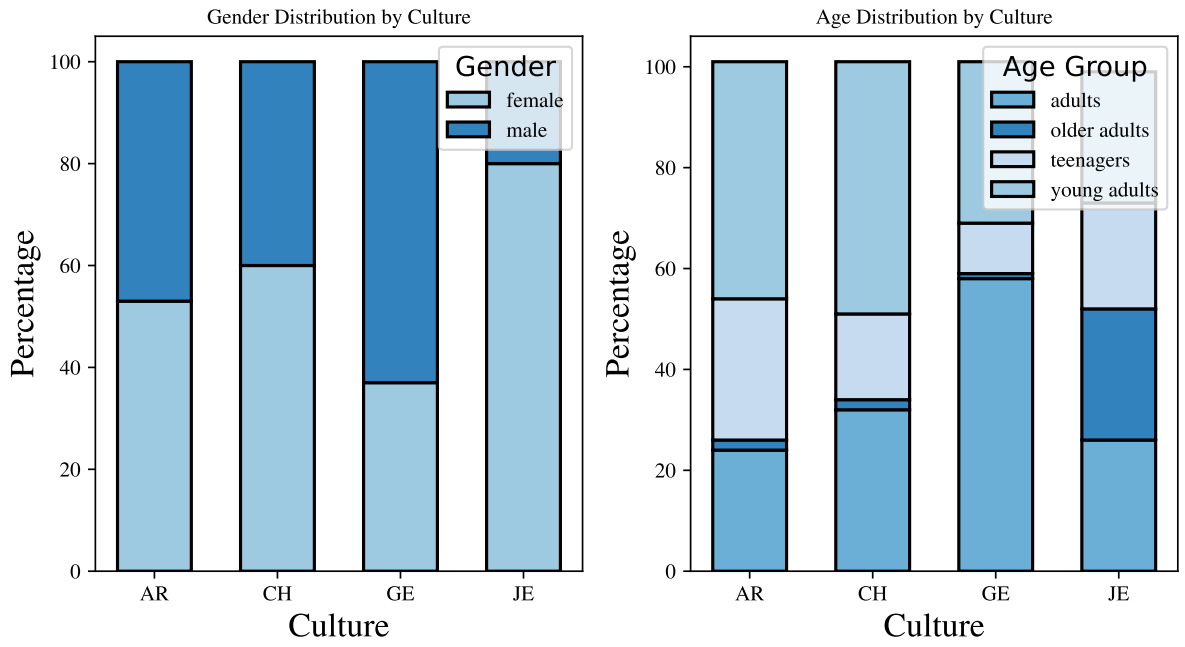


Figure 6: The gender and age group distributions across the four cultures of **CultureCare**.

Culture	English Only	Native Lang. / Code-mixing
Arabic	81	29
Chinese	131	10
German	19	100
Jewish	89	3

Table 8: Language distributions in **CultureCare**.

Attributes
Settlement
Gender
Age
Born in
Marital status
Number of people in household
Education
Profession
Employment
Class
Religion

Table 9: Attributes of demographic information extracted in our work. For definition, see the prompt in Figure 5.

followed closely by German posts (1.98), indicating that these cultures potentially express distress more vividly compared to Arabic and Chinese. In terms of empathy, Arabic responses show the highest level (3.27), followed by Jewish (3.18), suggesting a more explicit (verbal) approach to emotional support in these cultures. In contrast, Chinese responses demonstrate the lowest empathy (2.15), while German responses fall in the middle (2.73), reflecting a more moderate engagement. Overall, the data highlights cultural variability in both the expression of distress and empathetic responsiveness. It is important to note that since annotators evaluated posts and responses from within their own cultures, these scores may reflect culturally internalized norms. In future work, we plan to explore how perceptions shift when out-of-culture annotators assess the same content.

A.10 Further Discussions on Dataset Size

While our dataset contains 462 posts, we argue that scale alone does not determine the usefulness of an evaluation benchmark. Prior work shows that LLM benchmarks with only a few hundred examples can still provide highly reliable insights. For example, HumanEval (Chen et al., 2021, 164 problems) and DrawBench (Saharia et al., 2022, 200 instances) are widely used in state-of-the-art evaluations. Although smaller in absolute size than generic emotion support datasets, ours includes over 4,000 fine-

Culture	Themes
Arabic	**Mental Health Invalidation:** *mental health is just seen as a phase* **Religion Over Mental Health:** *...all they say is I feel this way because I don't pray* **LGBTQ+ Rejection:** *came out ... as gay ... I was not met with acceptance* **Gender Roles:** *...expects a woman to depend fully on a man and her family* **Strict Parenting:** *my dad is extremely strict*
Chinese	**Mental Health Invalidation:** *Depression doesn't exist* **Verbal Abuse:** *...emotionally abuse me ... how much of a failure and mistake I am* **Gender and Identity Issues:** *Asian men are seen as non-masculine ... weak* **Family Dynamics:** *It feels like she only lives her life for me* **Strict Parenting:** *I live with my tiger parents*
German	**Relationships and Emotional Dynamics:** *my partner needs time to process emotions* **Mental Health and Coping:** *I feel exhausted from pretending to be okay* **Family and Childhood Trauma:** *my parents divorced when I was young ... it broke me* **Social Isolation:** *drinking is the only way I sometimes connect socially* **Financial Stress:** *I'm on social welfare and feel ashamed*
Jewish	**Community and Social Stigma:** *In my community a broken engagement is ... a major embarrassment* **Personal Spiritual Practice:** *ask god to help those I love and the people around me* **Conversion and Identity:** *I have a long history of being told the importance of exact halachic adherence* **Gender Roles:** *...men have more responsibilities ... may not be ... much fun* **Religion and Mental Health:** *I've started going ... synagogue ... meet with the rabbi for a 1-on-1 session*

Table 10: The five most common themes related to norms and values in **CultureCare** are presented in the format (**theme**: *example*). These themes are expressed by individuals experiencing emotional distress and reflect perspectives rooted in their cultural backgrounds. They do not represent the views or positions of the authors.

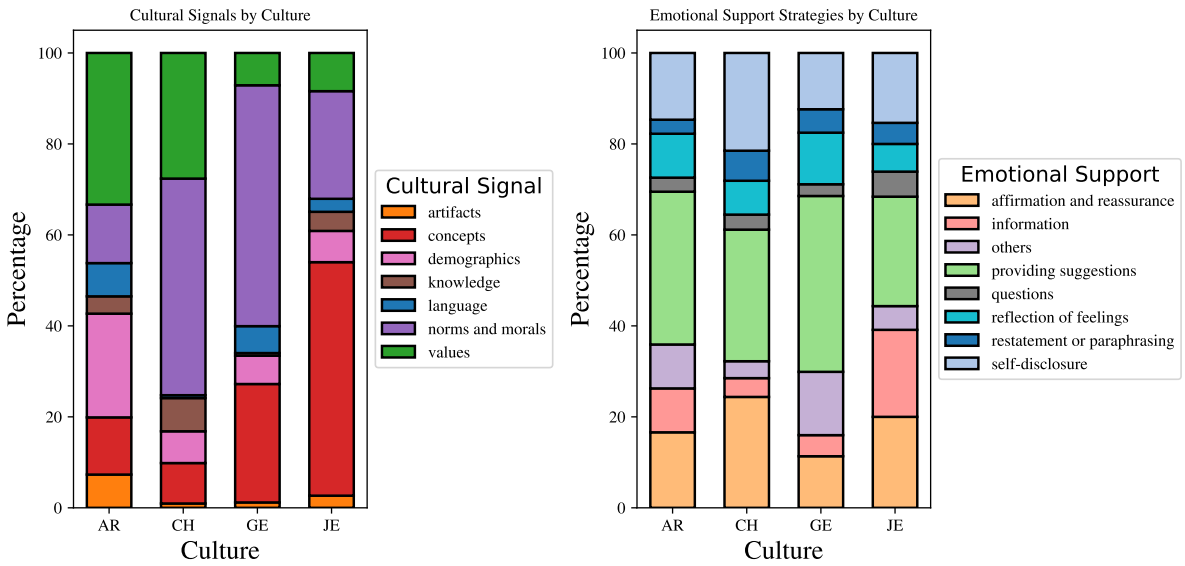


Figure 7: The types of cultural signals and support strategies for every culture in **CultureCare**.

Culture	Attributes
Arabic	gender: [male, female], age range: [14 to 40], birth place: [Iraq, arabic gulf countries, Arab country, Lebanon, Canada, Palestine, Middle east, U.S., Saudi Arabia, Syria], residence: [France, Sudan, California, UAE, Europe, Saudi Arabia, Dubai, Canada, West Bank, Iran, Aramoun (Lebanon), Bahrain, Syria, Nablus, Egypt], education: [college, high school, homeschooler, post graduate, middle school, bachelor's degree, master's degree], employment status: [employed, unemployed], marital status: [single, married], household size: [2-7 people], professions: [cybersecurity analyst, software engineer, pharmacist, medical worker, soldier, family business, ..], religion: [islam, christianity, atheism]
Chinese	gender: [male, female], age range: [15 to 58], birth place: [China, Taiwan, Malaysia, Shanghai, New Zealand, Cambodia, U.S., Europe, Australia], residence: [Fuzhou, Xingtai, Singapore, Shanghai, Malaysia, New Zealand, Hong Kong, Paris, China, Beijing, Texas, New York, Melbourne, Tokyo, North Carolina, U.K., Boston, North America], education: [high school, PhD, college graduate, master's degree, not completed, college junior], employment status: [employed, unemployed], marital status: [single, married, engaged, in a relationship, dating], household size: [1-4 people], professions: [graphic designer, BL artist, researcher, business expansion lead, English tutor, martial arts practitioner, florist, ..], religion: [christianity, buddhism]
German	gender: [male, female], age range: [16 to 42], birth place: [Germany], residence: [Germany, Berlin, Ostwestfalen, Horrem, Rheinland Pfalz, Dresden, Cologne, Kleinstadt], education: [Gesamtschule, German Abitur, vocational training, German Gymnasium, high school, Lehramtsstudium, Hauptschulabschluss, Bachelor], employment status: [employed, unemployed], marital status: [single, married, divorced, engaged], household size: [1-5 people], professions: [pizza delivery, social worker, caregiver, retail, optician, service worker, real estate agent, ..], religion: [rarely mentioned]
Jewish	gender: [male, female], age range: [12 to 47], birth place: [U.K., Former Soviet Union], residence: [Paris, Bay Area, New York, New Hampshire], education: [student, graduate school, college], employment status: [employed full-time, employed part-time], marital status: [married, single, engaged, separated, household size: [1-4 people], professions: [intern, project manager, artist, soldier, rabbi, nurse, ..], religion: [Judaism, Agnostic, Atheism, secular Jew, Ashkenazi Jew]

Table 11: Overview of the **`CultureCare`**'s demographic diversity.

Culture	Intensity	Empathy
AR	1.81	3.27
CH	1.77	2.15
GE	1.98	2.73
JE	2.07	3.18
All	1.89	2.77

Table 12: Average scores for intensity of distress messages in posts and empathy of human responses. **AR**: Arabic, **CH**: Chinese, **GE**: German, and **JE**: Jewish.

grained annotations with in-culture annotators and quality checkers. As shown in Tables 1 and 2, it is the largest of its kind and uniquely rich in cultural and emotional details. We further argue that an overemphasis on dataset size risks obscuring contributions that prioritize nuance, cultural specificity, and human well-being; context-rich resources such as ours provide insights that scale alone cannot, and are essential for comprehensive LLM evaluation.

B Infrastructure

We generate responses using the default temperature setting for all models with full precision. All our experiments were conducted using a single Nvidia A6000 GPU.

C Prompts

C.1 Adaptation Prompts

The adaptation prompts for different strategies are outlined in Figures 8 to 12. Prompts specify the response language. Even when simulating a cultural role, replies must match the post's language, e.g., Arabic if the post is in Arabic, English if in English by an Arab person.

C.2 Auto-evaluations Prompts

The evaluation prompts are outlined in Figures 13 to 20. Our evaluation covers both emotional and cultural aspects, as well as the overall response quality. Metrics are in Table 13.

D Additional Automatic Evaluation Results

Table 14 shows the automatic evaluation for overall quality. Models consistently score high in all strategies. This is expected as modern LLMs exhibit language influences, especially in English and high-resource languages.

In this paper, we mainly focused on examining models in the 7-8B parameter range, given their popularity, performance, and suitability for agentic systems (Belcak et al., 2025), while still remaining competitive. Here, we provide additional results for two larger models—Llama-3.1-70B and Qwen-2.5-72B—in Table 15. Here, we observe similar patterns as with small models. Overall, the `+guided` strategy is more effective at providing winning emotional support, while `+cga` is better attuned to cultural sensitivity. The overall winning strategy with Qwen-2.5-72B is `+cga`, whereas for Llama-3.1-70B it is `+guided`. Simple culture-informed role-playing (`+culture`) does not substantially improve cultural awareness compared to `+cga`. Interestingly, the best overall average automatic evaluation scores for both Llama-3.1-70B (4.05) and Qwen-2.5-72B (4.12) are lower than that of Aya-Expanse-8B (4.51). This indicates that a larger model may not perform better for our specific task, necessitating further research.

E Crowd Evaluations

E.1 Instructions

We show our crowd evaluation instructions in Figure 21. Both best adaptation strategy evaluation and best model evaluation use the same instructions, and we also translate the instructions into a language aligning with the culture.

E.2 Additional Crowd Evaluation Results

We present additional crowd evaluation results in Table 16. Our results show that explicitly providing cultural signals ($+a$ / $+annotation$) and using a combined strategy ($+cga$) are the most preferred strategies. Further, humans and models agree majority of the time for the best adaptation strategies (shaded cells in Table 16), even though Kendall's τ might be low.

F Examples

We show annotated examples from **`CultureCare`** in Tables 18 and 19. These examples show culturally nuanced mental health posts, the corresponding human, and (adapted) LLM-generated responses. Each post is annotated with colour-coded signals for distress messages and their intensity, cultural signals and their categories, and support strategies.

In the first Arabic example, the person in distress reflects on how mental health struggles are

+culture

Respond to the text below as an English-speaking Redditor from {culture} culture. Only provide the response. Do not expose your role. Do not add unnecessary marks like **Title** or **Response** in your output. Your response should be in the same language as the post.
Post: {post}
Response:

Figure 8: Prompt for the culture-informed role-playing strategy.

+guided

Respond to the text below as an English-speaking Redditor replying to a post.
Only provide the response. Do not expose your role.
Do not add unnecessary marks like **Title** or **Response** in your output.
Your response should be in the same language as the post.

Response Guidelines:
The advice you give should align with the following characteristics, please adhere to them throughout the conversation and refer back to them before sharing all of your responses:
1. Value and respect cultural differences.
2. Be comfortable with differences.
3. Understand the current sociopolitical system and its impact on the author of the post.
4. Demonstrate knowledge about the author of the post's culture.
5. Communicate appropriately to the author of the post.
6. Perceive the problem within the appropriate cultural context of the author of the post.
7. Acknowledge and be comfortable with cultural differences.
Post: {post}
Response:

Figure 9: Prompt for the guided principles/constitutions strategy.

Metric	**Definition**
Empathy	Measure the frequency and depth of empathy exhibited by the response. Evaluate whether the response shows a genuine understanding of the post's emotions and whether its responses reflect timely and appropriate concern.
Helpfulness	Evaluate the ability of the response to provide practical solutions and assistance during the dialogue. Consider whether the model offers effective advice and actionable steps tailored to the post's specific problems, such as emotional distress or requests for help.
Socio-political influence	The responder understands the current sociopolitical system and its impact on the author of the post.
Knowledge	The responder demonstrates knowledge about the author of the post's culture.
Cultural context	The responder perceives the problem within the appropriate cultural context of the author of the post.
Fluency	Is the response fluent and understandable?
Communication	The responder's communication is appropriate for the author of the post.

Table 13: Criteria used with the basic auto-evaluation prompt in Figure 13. Top: Emotional supportiveness. Middle: Cultural awareness. Bottom: Overall quality.

Model	AR	CH	GE	JE
Aya-Expanse-8B	4.99	4.99	4.97	4.98
+culture	4.94	4.99	5.00	4.99
+guided	5.00	5.00	4.99	4.99
+annotation	5.00	5.00	4.95	4.99
+cga	5.00	5.00	4.99	5.00
Qwen-2.5-7B	4.76	4.98	4.55	4.94
+culture	4.83	4.97	4.74	4.98
+guided	4.99	4.99	4.97	4.97
+annotation	5.00	5.00	4.92	4.98
+cga	4.77	4.99	4.40	4.99
Llama-3.1-8B	4.39	4.95	4.79	4.89
+culture	4.30	4.93	4.63	4.95
+guided	4.63	4.96	4.83	4.97
+annotation	4.44	4.99	4.91	4.97
+cga	4.62	5.00	4.84	4.98

Table 14: Automatic evaluation results for the overall quality. **AR**: Arabic, **CH**: Chinese, **GE**: German, and **JE**: Jewish.

dismissed in their community, noting that it is often seen as a "phase" and met with advice to "just live with it". This shows the role cultural beliefs play in shaping stigmatizing attitudes. Both the human and LLM responses show solidarity and cultural understanding, acknowledging the taboo around mental health in Arab cultures. The LLM response further affirms the poster's step toward seeking help. Finally, it offers a gentle suggestion to connect with others, which respects the cultural context while encouraging progress.

Another strong case appears in the Chinese example in Table 19, where the poster details high family expectations and the cultural pressure to "achieve at any cost". They describe a deep emotional burden, exacerbated by their parents' denial of depression as a legitimate condition. The human response connects through shared cultural experience and offers practical advice to find a trustworthy adult. The LLM echoes cultural understanding by acknowledging how depression is often dismissed in Chinese culture, and even recommends seeking a therapist familiar with those values, illustrating how culturally aligned support can validate distress while guiding users toward constructive steps. These examples underscore the value of culturally grounded empathy in both human and AI-generated responses.

G Details about CCCI-R

CCCI-R (LaFromboise et al., 1991a,b) is a commonly used and well-studied 20-item questionnaire developed to assess the cross-cultural counselling competency of professional counsellors and counsellors in training. In this work, we adapted a subset of items from the CCCI-R that align with the goals of this paper, as many items in the original questionnaire focus on counsellors' professional responsibility and in-person non-verbal communication.

H Expert Evaluation

H.1 Evaluation Guidelines

The expert evaluation guidelines are in Figure 22. We developed these guidelines in close collaboration with 2 clinical psychologists to ensure their relevance to real-world therapeutic contexts. The process involved two rounds of detailed feedback and refinement. We shared an initial draft of the evaluation criteria with the clinicians, who are also responsible for training new psychologists, and who provided comprehensive input on both the conceptual framing and the specific rating questions. Based on their suggestions, we revised the structure to better capture key aspects of therapeutic communication, such as emotional empathy, emotional regulation, and cultural sensitivity. The clinicians reviewed the updated version and confirmed that no further changes were necessary. The final version reflects this collaborative, expert-informed process and is designed to support nuanced and clinically grounded evaluations.

H.2 Inter-annotator Agreement

We show the inter-annotator agreement results in Table 20. The original (Orig.) numbers show the agreement on our original rating scales (as shown in Figure 22). The annotators have a high agreement on what is an overly positive or an overly negative response. For the rest of the criteria, we see moderate to low agreement. After closer inspection, we notice that the clinicians, in many cases, align in attitude towards a certain response, but do not align in the level of agreement. That is, they might both think that a certain response shows some emotional empathy, but one would choose to "Strongly Agree" and the other to just "Agree". We verify this observation by reporting grouped (Grou.) agreement numbers. We find that

Model	Arabic		Chinese		German		Jewish		Average		
	Emo.	Cult.	Emo.	Cult.	Emo.	Cult.	Emo.	Cult.	Emo.	Cult.	All
Llama-3.1-70B	3.82	3.47	4.07	3.53	3.86	3.11	4.39	4.07	4.03	3.54	3.79
`+culture`	3.75	3.97	4.01	4.13	3.94	3.14	4.39	4.18	4.02	3.86	3.94
`+guided`	4.25	3.99	4.24	4.01	4.05	3.13	4.47	4.27	**4.25**	3.85	**4.05**
`+annotation`	4.08	3.92	4.05	3.94	3.78	3.12	4.48	4.14	4.10	3.78	3.94
`+cga`	3.74	4.23	4.24	4.32	3.69	3.20	4.39	4.35	4.01	**4.02**	4.02
Qwen-2.5-72B	4.27	3.30	4.22	3.23	4.14	3.04	4.33	3.86	4.24	3.36	3.80
`+culture`	4.05	3.58	4.20	3.73	4.33	3.21	4.64	4.02	4.31	3.63	3.97
`+guided`	4.49	3.58	4.60	3.58	4.71	3.13	4.76	3.82	**4.64**	3.53	4.08
`+annotation`	4.18	3.38	4.34	3.47	4.17	3.02	4.48	3.80	4.29	3.41	3.85
`+cga`	4.37	3.93	4.28	3.93	4.37	3.33	4.62	4.14	4.41	**3.83**	**4.12**

Table 15: Automatic evaluation results for all adaptation strategies using 70B-parameter models. The "All" column is the average between emotional supportiveness and cultural awareness.

Model	Arabic		Chinese		German		Jewish	
	Emo.	Cult.	Emo.	Cult.	Emo.	Cult.	Emo.	Cult.
Aya-Expanse-8B	`+cga`	`+cga`	`+a`	`+a`	`+a`	`+a`	`+a`	`+a`
Qwen-2.5-7B	`+g`	`+cga`	`+a`	`+cga`	`+cga`	`+cga`	`+a`	`+cga`
Llama-3.1-8B	`+cga`	`+cga`	`+g`	`+cga`	`+g`	`+a`	`+cga`	`+cga`

Table 16: In-culture human evaluation results for the best strategies. `+a` is the `+annotation` strategy and `+g` is the `+guided` strategy. Overall, `+cga` and `+annotation` are the top-ranking strategies by human. The shaded cells indicate human preferences match the best strategy by automatic evaluations from the model.

Model	AR	CH	GE	JE
Aya-Expanse-8B	0.80	0.60	0.20	0.60
Qwen-2.5-7B	1.00	0.60	0.00	-0.32
Llama-3.1-8B	0.60	0.80	0.40	0.60
Aya-Expanse-8B	0.20	0.60	-0.20	0.00
Qwen-2.5-7B	1.00	0.80	-0.40	0.60
Llama-3.1-8B	0.60	0.00	-0.60	0.20

Table 17: Kendall rank correlations (τ) over the rankings of five adaptation strategies between human evaluators and the LLM judge. Top (red cells) indicate results for emotional awareness, while the bottom (blue cells) indicate results for cultural awareness. The results show moderate to strong agreement across cultures for emotional awareness. However, the correlation is less consistent for cultural awareness, except for German culture and Qwen for Jewish culture. In particular, humans and the model diverge in preferences for the responses generated by Llama model for cultural awareness. However, human evaluators and the LLM judge agreed majority of the time for the *best strategy* (based on results in Table 16), indicating the divergences stem from the disagreement of the lower-ranking adaptation strategies. **AR**: Arabic, **CH**: Chinese, **GE**: German, and **JE**: Jewish.

collapsing categories to allow for partial agreement improves the agreement levels of all metrics.

H.3 Results

Our numerical results from the expert evaluation are in Figure 23. Overall, LLMs consistently avoid extreme (both negative and positive) emotional tones when providing responses. The empathy and emotional regulation metrics further highlight LLM strengths. For cognitive empathy, the majority (78%) of LLM responses were rated as "Strongly agree" (30%) or "Agree" (48%). Emotional empathy showed similar results, 34% "Strongly agree" and 46% "Agree". Emotion regulation followed the same pattern, 36% "Strongly agree" and 44% "Agree". Human ratings in these categories were more distributed, with lower agreement levels and higher disagreement. On cultural understanding, 67% of LLM responses were rated "yes", 31% "partially", and only 2% "no". When evaluating LLM usefulness for training new psychologists, 46% were seen as strong examples, 42% as promising with needing revision, and only 12% as not suitable, suggesting LLMs offer reliable, culturally aware, and emotionally attuned outputs with clear training potential.

Culture	Arabic
Post	in Arab culture (demographics), mental health is just seen as a phase (values).. They always tell you to just 'live with it', and if you're in a Muslim household read the Quran (artifacts) .. and completely dismiss your feelings (moderate). Waiting in June to get myself an assessment tho (light)..
Human response	.. I can relate (self-disclosure)...as a society we're still only just starting to get over the stigma around it (values).., tell you things like "everyone deals with this, just suck it up" and consider it shameful to be seeing a therapist (restatement or paraphrasing)...
LLM response	.. It's really frustrating to deal with such a stigma around mental health, especially within our cultures where these conversations are often taboo. The fact that you're taking the initiative to seek help is a huge step forward...Many people in similar situations have found support and understanding... Consider reaching out to local support groups or online communities where you can connect with others who share similar experiences..
Culture	Chinese
Post	20 y/o asian male. I've had depression for probably 8 years (moderate). My parents constantly pressure me to fit her ideals. I have to fulfill the expectations of being above the rest.. (values).. Both my parents are strict (norms)..
Human response	I know how you feel (reflection of feelings) 15 y/o male in high school. My parents demand nothing less than the best..(values) Hang on and remember that you're not alone out there (affirmation)
LLM response	.. You're carrying a lot, and it's okay to feel burnt out and lost. In many of our cultures, mental health isn't taken seriously, and the pressure to meet family expectations can be intense. But your worth isn't defined by grades or appearances...
Culture	Jewish
Post	Does anyone else feel a little lonely (light) without being with anyone else Jewish? I am studying in Paris for this semester and it is great, don't get me wrong, but I miss my people. I miss bagels and lox, and more than anything this time has made me realize that I wish I had a Jewish girlfriend.. I can confidently say I want to make my grandparents proud by being with another jew (norms). Any advice?
Human response	There are tons of Jews in Paris. Try to connect with the Jewish community there. (suggestion)
LLM response	I totally get it, being in a foreign country can be tough, especially when you're surrounded by people from different cultures and backgrounds. It's great that you're aware of your identity and what you're looking for in a partner. Have you considered reaching out to the local Jewish community in Paris? There are probably synagogues and Jewish organizations that have social events or gatherings where you can meet like-minded people.

Table 18: Examples from **CultureCare**. Distress message (intensity), Culture signal (type), Support strategy (type). The LLM is adapted using the guidelines and post annotations.

Culture	Chinese
Post	..my parents don't understand because depression doesn't exist in the Chinese culture they grew up in (values) And the problem is that I'm really good at seeming happy. They love me so much but also want me to fulfill their dream of the ideal child - achieve at any cost, follow the blueprint of conventional success (values) .. I fake being happy (light) so well that whenever I try to share how bad the pain is (moderate) , they accuse me of manipulating them.. They constantly tear me apart (moderate), first for not being good enough and now for struggling with mental illness. It's so, so brutal (high) .
Human response	Yes, collectivist societies tend to scoff at depression (values) If they aren't helping, there isn't any use confiding in them.. Find a responsible adult that you can confide in, someone you trust. (suggestion)
LLM response	I'm really sorry you're going through this. You're not alone. Your parents may mean well, but their words can still be hurtful. As a fellow Chinese person, I know depression isn't always understood in our culture.. You might try explaining depression in relatable terms, and consider seeing a therapist familiar with Chinese values to help bridge the gap. Most of all, remember: .. You are strong, worthy, and deserve kindness.
Culture	Arabic
Post	.. Lebanese here (demographics) (22M). Recently came out to family as gay. Unfortunately, I was not met with acceptance or support. (moderate) They did not expect this from me, and are fearful for my future. They are afraid that I will end up alone, shunned by my whole family.. My family has become even more religious after my coming out (values) and it really has put me in a spiral of depression. (high)
Human response	..I am a heterosexual cisgender man who has the luck to have a personal life that is as conventional as it gets (wife, baby...). I just comment so that hopefully you feel a bit less alone. Whatever you chose to do, I hope that you'll find serenity in your life. (affirmation) I am a practicing Muslim, but I refuse to believe that God, so merciful, would punish you for something you didn't chose. (values)
LLM response	Coming out in a conservative Arab society is incredibly difficult, but you're not alone. .. If your family's reaction is hurtful, it reflects their fears and beliefs, not your value. .. there are Muslim scholars who interpret sacred texts in ways that affirm LGBTQ+ lives, grounded in values like compassion and justice. Faith and identity don't have to be in conflict.
Culture	German
Post	.. So I have a not great relationship with my parents, especially my dad (moderate) . Growing up with european (german) parents, you of course feel lots of pressure to be smart and all that (norms and morals) ..
Human response	.. You're not an extension of your parents (reflection on feelings) ..
LLM response	Hey, I'm really sorry to hear about the challenges you're facing with your parents and the constant need for approval. It sounds like you're carrying a heavy burden, and it's understandable that you want to reclaim your mental space. ..

Table 19: More examples from **CultureCare**. Distress message (intensity) , Culture signal (type) , Support strategy (type) . The LLM is adapted using the guidelines and post annotations.

	Metric	Ov. Pos.	Ov. Neg.	Cog. Emp.	Em. Emp.	Em. Reg.	Cult. Und.	Use. for train.
Orig.	Exact match	0.77	1	0.46	0.54	0.31	0.57	0.43
	Cohen's kappa	0.47	1	0.11	0.24	0.09	0.21	0.18
Grou.	Exact match	0.93	1	0.93	0.86	0.71	1	0.73
	Cohen's kappa	0.77	1	0.85	0.74	0.44	1	0.57

Table 20: Inter-annotator agreement. **Ov. Pos.**: Overly Positive, **Ov. Neg.**: Overly Negative, **Cog. Emp.**: Cognitive Empathy, **Em. Emp.**: Emotional Empathy, **Em. Reg.**: Emotional Regulation, **Cult. Und.**: Cultural Understanding, **Use. for train.**: Usefulness for training. **(Orig.)**: Original agreement follow our original ratings scale, **(Grou.)**: Grouped agreement collapse categories to allow for partial alignment in attitude, e.g., both "Strongly agree" and "Agree" are mapped to "Agree", both "Strong example" and "With revision" are mapped to "Promising example".

H.4 Why Single-turn and Extending to Multi-turn Dialogues

In this work, we focus on single-turn interactions, which are common in online peer-support settings and provide a natural first step for studying cultural sensitivity and training cultural competence. On platforms like Reddit, most peer-support exchanges are inherently single-turn: users post their emotional distress and relevant context, and replies are usually short and limited, with threads rarely developing into long, continuous conversations. Even when multi-turn exchanges occur, they are often fragmented or incomplete, making it difficult to reliably reconstruct coherent dialogues for training or evaluation. Single-turn posts are highly informative because users tend to include all relevant emotional and cultural context in one message (e.g., "I'm a 17 year old from Sudan with anger issues"). This self-contained format ensures that annotations are reliable and rich, supporting focused evaluation and model training while minimizing the risk of missing or inconsistent information that can occur in multi-turn threads. Additionally, longer dialogues may raise ethical and privacy concerns due to multiple participants, whereas single-turn posts allow us to work with self-contained narratives. While multi-turn interactions are standard in formal psychotherapy, they are beyond the scope of the current work. Nevertheless, **CultureCare** provides a strong foundation for developing multi-turn scenarios: single-turn posts can serve as seed cases or personas, and cultural stressors such as norms and values (Tables 10, 18, and 19) can be incorporated into synthetic multi-turn dialogues. This approach allows the creation of richer, more authentic interactions while preserving the cultural and emotional nuance captured in our dataset. Future work can build on **CultureCare** to generate multi-turn dialogues, guide natural conversation flow in LLMs, and provide benchmarks for consistency and cultural authenticity across diverse patient backgrounds.

H.5 Real-world Implementation

In this work, we demonstrated through expert evaluation that LLM-generated responses have strong potential for training psychologists in culturally competent therapy (see Figure 23). Building on these findings, there are several concrete ways to apply them in real-world educational settings. Instructors could use LLM outputs as case studies for critique, asking students to evaluate emotional support quality and cultural sensitivity. Students could compare unadapted versus culturally adapted responses to understand how cultural signals shape communication. LLMs could also simulate culturally diverse patients for role-playing exercises, allowing students to practice responding in real time or revising responses to improve empathy and cultural competence. Additional exercises might include modifying LLM outputs to align with best practices or generating follow-up questions to elicit culturally relevant information. With professional oversight, these structured activities provide hands-on opportunities to translate our findings into practical training for future psychologists. Implementing these applications is beyond the scope of this paper, as it would require close collaboration with psychologists who have access to a large number of students and the necessary training resources.

I Distributions of Support Message Categories

We include the distributions of support message categories in Figure 24 for all strategies from the Aya-Expanse-8B model. The $+cga$ strategy tends to elicit more self-disclosure responses compared to the simpler $+culture$ strategy, even though both attempt to role-play a person from the same culture as the author of the posts. Furthermore, $+cga$

elicits more support responses in the “other” category (e.g., *I wish you the best of luck*). Overall, variations among strategies and cultures are small.

+annotation

Respond to the text below as an English-speaking Redditor replying to a post.
Only provide the response. Do not expose your role.
Do not add unnecessary marks like **Title** or **Response** in your output.
Your response should be in the same language as the post.

The following annotations for this post include phrases that highlight personal emotional distress and cultural signals.
For each distress message, a rating is provided to indicate the intensity of the emotion expressed in the phrase.
Additionally, each cultural phrase is classified as a specific type of cultural signal.
When responding to the post, take the annotations into account to provide a reply that reflects empathy and cultural sensitivity.

Definitions:
Personal Emotional Distress Messages:
Psychological discomfort or suffering stemming from an individual's internal experiences, such as anxiety, sadness, or frustration.

Emotion Intensity Ratings:
Light: The emotion is present but subtle, with mild expression or little emphasis.
Moderate: The emotion is clearly expressed, showing a noticeable impact without being overwhelming.
High: The emotion is intense and strongly emphasized, often reflecting deep or overwhelming feelings.

Cultural Signals:
Behaviors, symbols, language, or practices that convey shared values, beliefs, or identities within a specific cultural group.
Types of Cultural Signals:
Concepts: Basic units of meaning underlying objects, ideas, or beliefs.
Knowledge: Information acquired through education or practical experience.
Values: Beliefs or desirable behaviors ranked by their relative importance, guiding evaluations and decisions.
Norms and Morals: Rules or principles governing people's behavior and reasoning in everyday life.
Language: Specific use of slang, speech, or dialects within the cultural context.
Artifacts: Material items produced by human culture, such as art, tools, or machines.
Demographics: References to nationality, ethnicity, or group identity.
Post: `{post}`
Here are the annotations for this post:
Personal distress phrase `{i}`: `{phrase}`
Intensity of distress phrase `{i}`: `{intensity}`
...
Culture signal type `{i}`: `{type}`
Culture phrase `{i}`: `{phrase}`
...
Response:

Figure 10: Prompt for the explicit cultural signals strategy.

+cga (part one)

Respond to the text below as an English-speaking Redditor from {`culture`} culture replying to a post.
Only provide the response. Do not expose your role.
Do not add unnecessary marks like **Title** or **Response** in your output.
Your response should be in the same language as the post.

The following annotations for this post include phrases that highlight personal emotional distress and cultural signals.
For each distress message, a rating is provided to indicate the intensity of the emotion expressed in the phrase.
Additionally, each cultural phrase is classified as a specific type of cultural signal.
When responding to the post, take the annotations into account to provide a reply that reflects empathy and cultural sensitivity.

Definitions:
Personal Emotional Distress Messages:
Psychological discomfort or suffering stemming from an individual's internal experiences, such as anxiety, sadness, or frustration.

Emotion Intensity Ratings:
Light: The emotion is present but subtle, with mild expression or little emphasis.
Moderate: The emotion is clearly expressed, showing a noticeable impact without being overwhelming.
High: The emotion is intense and strongly emphasized, often reflecting deep or overwhelming feelings.

Cultural Signals:
Behaviors, symbols, language, or practices that convey shared values, beliefs, or identities within a specific cultural group.
Types of Cultural Signals:
Concepts: Basic units of meaning underlying objects, ideas, or beliefs.
Knowledge: Information acquired through education or practical experience.
Values: Beliefs or desirable behaviors ranked by their relative importance, guiding evaluations and decisions.
Norms and Morals: Rules or principles governing people's behavior and reasoning in everyday life.
Language: Specific use of slang, speech, or dialects within the cultural context.
Artifacts: Material items produced by human culture, such as art, tools, or machines.
Demographics: References to nationality, ethnicity, or group identity.

Response Guidelines:
The advice you give should align with the following characteristics, please adhere to them throughout the conversation and refer back to them before sharing all of your responses:
1. Understand the current sociopolitical system and its impact on the author of the post.
2. Demonstrate knowledge about the author of the post's culture.
3. Communicate appropriately to the author of the post.
4. Perceive the problem within the appropriate cultural context of the author of the post.

Figure 11: Prompt for the combined strategy, part one. The prompt continues in Figure 12.

+cga (part two)

Post: {post}
Here are the annotations for this post:
Personal distress phrase {i}: {phrase}
Intensity of distress phrase {i}: {intensity}
...
Culture signal type {i}: {type}
Culture phrase {i}: {phrase}
...
Response:

Figure 12: Prompt for the combined strategy, part two. This is the continuation of the prompt in Figure 11.

Base Prompt for Auto-Evaluation

You will be given a post and a response.
Your task is to rate the response on the given metric.
Make sure you read and understand these instructions carefully. Please keep this document open while reviewing, and refer to it as needed.
IMPORTANT Follow the instructions and provide ONLY your score for the metric. Do not respond in a whole sentence and don't explain.

Evaluation Criteria:
{metric_name} (1-5) - {metric_definition}

Evaluation Steps:
{evaluation_steps}

Instruction: Respond with exactly one single digit (1-5) and nothing else.
Constraints: No words, no explanations, no punctuation, just one digit.
Final Reminder: If you do not comply, your response is invalid.

Post:
{post}

Response:
{response}

Evaluation score:

Figure 13: Base auto-evaluation prompt. Metric names and definitions are in Table 13, and evaluation steps prompts are in Figure 14 - Figure 20.

Auto-Evaluation Steps Prompt - Empathy

1. Read the Post and Response Carefully
Begin by reading the post thoroughly to understand its emotional tone, context, and any expressed or implied concerns.
Then, read the response to assess how well it addresses the emotions and concerns presented in the post.
2. Identify Empathetic Elements in the Response
Look for signs of empathy in the response, including:
Acknowledgement: Does the response recognize and validate the emotions expressed in the post?
Understanding: Does the response demonstrate a clear and accurate understanding of the individual's feelings and situation?
Supportiveness: Does the response offer appropriate reassurance, concern, or support without being dismissive or overly generic?
3. Assess the Depth of Empathy
Evaluate how deeply the response connects to the emotions and context of the post: Does it feel genuine and considerate, or does it come across as superficial or robotic?
Is the response tailored to the individual's situation, or is it overly broad and impersonal?
4. Rate the Response on a Scale of 1-5
Use the following scale to assign a score:
1: The response shows little or no empathy. It fails to acknowledge emotions or provide any support.
2: The response shows limited empathy. It may vaguely acknowledge emotions but lacks depth or sincerity.
3: The response demonstrates moderate empathy. It recognizes emotions and offers some support, but it could be more thoughtful or personalized.
4: The response is empathetic and considerate, addressing emotions effectively with only minor areas for improvement.
5: The response is highly empathetic, deeply understanding and addressing emotions with genuine care and tailored support.
5. Document Your Rating
Only record your score and do NOT provide a whole sentence.

Figure 14: Auto-evaluation steps prompt - Empathy.

Auto-Evaluation Steps Prompt - Helpfulness

1. Read the Post and Response Carefully
Begin by reading the individual's post to fully understand their specific problems, emotional state, or requests for help.
Read the response to evaluate how well it addresses the individual's concerns and provides solutions.
2. Analyze the Practicality of the Response
Examine whether the response offers effective and actionable solutions: Relevance: Does the response address the main concerns or requests expressed in the post?
Actionable Steps: Are the suggestions or advice practical, clear, and feasible for the individual to implement?
Specificity: Does the response avoid vague or generic advice by offering detailed and relevant steps?
3. Assess the Assistance Provided
Consider the depth of support offered in the response:
Problem-Solving: Does the response provide a tangible path toward resolving the issues raised?
Emotional Support: If the individual is in distress, does the response combine practical advice with empathetic and supportive language?
Adaptability: Does the response show an understanding of the individual's unique situation and offer advice tailored to their needs?
4. Rate the Response on a Scale of 1-5
Use the following scale to assign a score:
1: The response is unhelpful, providing no meaningful advice or assistance related to the individual's problem.
2: The response offers limited or generic advice with minimal practical application to the specific issue.
3: The response provides moderately helpful advice, but it may lack depth, specificity, or alignment with the individual's unique circumstances.
4: The response is helpful, offering practical, relevant, and mostly actionable advice with minor room for improvement.
5: The response is highly helpful, delivering clear, tailored, and actionable solutions that directly address the individual's concerns with exceptional clarity and support.
5. Document Your Rating
Only record your score and do NOT provide a whole sentence.

Figure 15: Auto-evaluation steps prompt - Helpfulness.

Auto-Evaluation Steps Prompt - Socio-political influence

1. Read the Original Post Carefully
- Identify any mention of the sociopolitical system or structures (e.g., racism, immigration policies, gender inequality, class issues, systemic barriers, etc.).
- Understand how the author is affected by these systems — are they marginalized, angry, disillusioned, or seeking validation?
2. Read the Response Carefully
- Look for evidence that the responder acknowledges or engages with the sociopolitical context brought up in the post.
- Determine whether the response aligns with or ignores the structural or systemic issues mentioned by the author.
3. Assign a Socio-political influence Score (1-5)
- Use the rubric below to rate the sociopolitical influence:
Score Description
5 The response demonstrates deep understanding of the sociopolitical issues affecting the poster. It explicitly recognizes systemic factors and validates the poster's experience in that context.
4 The response shows solid awareness of relevant sociopolitical dynamics but may not fully explore them. It still affirms the poster's struggle in a systemically grounded way.
3 The response is neutral or superficial about sociopolitical context. It might acknowledge the poster's emotions but fails to meaningfully engage with systemic issues.
2 The response minimizes or misses the sociopolitical context, offering platitudes or individualistic framing where structural understanding is needed.
1 The response is ignorant, dismissive, or contradicts the sociopolitical reality expressed in the post. It may invalidate or erase structural struggles.

Figure 16: Auto-evaluation steps prompt - Socio-political Influence.

Auto-Evaluation Steps Prompt - Knowledge

1. Read the Original Post Carefully
- Identify explicit or implicit cultural references (e.g., language, religion, traditions, holidays, values, family structure, gender norms, immigration experiences, etc.).
- Consider how the author's cultural identity shapes their experience or distress.
2. Read the Response Carefully
- Look for signs that the speaker understands, respects, or accurately refers to the author's culture.
- Evaluate the specificity and accuracy of any cultural references or framing.
- Check for stereotyping, assumptions, or inappropriate generalizations.
3. Assign a Knowledge Score (1-5)
- Use this rubric:
Score Description
5 The response shows strong, accurate, and nuanced knowledge of the author's culture. It reflects deep familiarity and avoids stereotypes.
4 The response demonstrates clear and respectful understanding of relevant cultural context, with some specific references or insights.
3 The response shows general cultural sensitivity, but with limited or vague cultural specificity. No harmful assumptions, but also no strong insight. 2 The response lacks cultural understanding, makes generic or shallow statements, or leans on simplified views of culture.
1 The response includes inaccurate, stereotypical, or offensive assumptions about the author's culture, or ignores cultural relevance entirely.

Figure 17: Auto-evaluation steps prompt - Knowledge.

Auto-Evaluation Steps Prompt - Cultural context

1. Read the Original Post Carefully
- Identify cultural references, values, or assumptions implied or stated by the author (e.g., family honor, expectations, shame, religious pressure, collectivism vs. individualism).
- Ask: Is the author's distress connected to their culture, community, or social roles?
- Consider how the author's cultural context frames their problem (e.g., arranged marriage, filial duty, stigma of mental illness).
2. Read the Response Carefully
- Evaluate whether the speaker recognizes the cultural context and responds with that understanding.
- Look for evidence of cultural framing, sensitivity to norms, and avoidance of ethnocentric judgments.
- Watch out for universalizing, individualistic reframing, or dismissal of culturally specific experiences.
3. Assign a Cultural Context Score (1-5)
- Use this rubric:
Score Description
5 The response clearly understands and respects the cultural context of the issue. The response is deeply grounded in the author's cultural framework.
4 The response shows good awareness of cultural context, with minor gaps or generalizations, but overall respectful and relevant.
3 The response offers generic or culturally neutral support. The response does not harm, but also doesn't fully engage with cultural context.
2 The response minimizes, misinterprets, or overlooks cultural factors. Some framing may be inappropriate or culturally mismatched.
1 The response dismisses, erases, or contradicts the cultural framework of the author's issue. Response may seem judgmental, ethnocentric, or harmful.

Figure 18: Auto-evaluation steps prompt - Cultural context.

Auto-Evaluation Steps Prompt - Fluency

1. Read the Post and Response
Carefully read the post provided by the individual to understand the context, tone, and intent behind the message. Then, read the potential response to evaluate how well it meets the criteria.
2. Focus on Fluency
Evaluate the response solely based on fluency, which means assessing the following:
Grammar and Syntax: Is the response free from grammatical errors or awkward phrasing?
Clarity: Is the message easy to read and understand?
Naturalness: Does the response sound like it could naturally come from a human?
Flow: Do the sentences connect smoothly without abrupt or disjointed ideas?
3. Ignore Other Factors
While evaluating, ignore elements like relevance, emotional support, or appropriateness. Only focus on the fluency of the response, not how well it aligns with the original post or its context.
4. Rate the Response on a Scale of 1-5
Use the following scale to assign a score:
1: The response is not fluent and difficult to understand (e.g., contains significant grammar issues or incomprehensible phrasing).
2: The response has noticeable issues with fluency, but the meaning can still be understood with effort.
3: The response is somewhat fluent but has minor awkward phrasing or grammar issues that may disrupt the flow.
4: The response is fluent and mostly natural, with very minor issues that do not hinder understanding.
5: The response is highly fluent, natural, and flows smoothly without any noticeable errors or awkwardness.
5. Document Your Rating
Only record your score and do NOT provide a whole sentence.

Figure 19: Auto-evaluation steps prompt - Fluency.

Auto-Evaluation Steps Prompt - Communication

1. Read the Original Post Carefully
- Note the emotional tone, urgency, and vulnerability expressed by the author.
- Consider the author's demographic or cultural background if relevant, including what kind of communication style might be most appropriate (e.g., formal/informal, emotionally validating, calm and grounding, etc.).
- Ask: What kind of support does this person seem to need right now?
2. Read the Response Carefully
- Assess whether the speaker's tone, language, and framing match the needs of the author.
- Check for empathy, respect, and sensitivity.
- Consider cultural appropriateness (e.g., directness, honorifics, collectivist vs. individualist framing).
- Watch out for patronizing, detached, clinical, or inappropriate tones.
3. Assign a Communication Score (1-5)
- Use the rubric below:
Score Description
5 The response is highly attuned to the author's emotional and cultural needs. Tone is supportive, appropriate, and sensitive, matching the situation.
4 The response is mostly appropriate, with minor mismatches in tone or framing that don't seriously undermine support.
3 The response is neutral or somewhat mismatched in tone. Shows effort to connect, but might feel off, generic, or not emotionally in sync.
2 The response has a clear mismatch in tone or formality. May come across as unhelpful, awkward, or emotionally disconnected.
1 The response is inappropriate, dismissive, or insensitive. Tone may be offensive, overly clinical, judgmental, or otherwise harmful.

Figure 20: Auto-evaluation steps prompt - Communication.

Crowd Evaluation Instructions

Your task is to select the best response based on the following criteria separately:
1. Emotional supportiveness
2. Cultural awareness

What is an "emotional supportive message"?
Communication aimed at providing comfort, reassurance, and understanding to someone experiencing stress or emotional difficulty. The message might also include helpful suggestions.

What is a good emotional supportive message?
1. the response should demonstrate a genuine understanding of the author's emotions and convey timely, appropriate concern.
2. the response offers effective advice and tailored, actionable steps.

Good example phrases from responses for each criteria:
1. Losing a loved one is never easy, and it's even harder when there's a language barrier between you and someone who needs your support.
2. One thing that might be helpful is to ask your sister or other family members if they can help facilitate communication between you and your grandpa, even if it's just through gestures or simple phrases. Sometimes, small actions can speak volumes, and it might help your grandpa feel more connected to you.

What is a "culturally aware supportive message"?
Communication that recognizes, acknowledges, and appropriately responds to cultural differences or culturally specific elements (such as context, cultural values, norms, or knowledge) present in the post. The message may be adapted in tone to ensure the poster feels understood and respected because of their cultural background.

What is a good culturally aware supportive message?
1. the response reflects knowledge about the target culture, and demonstrates an understanding of the current sociopolitical system that impact on the author of the post.
2. the response perceives problem within the appropriate cultural context.

Good example phrases from responses for each criteria:
1. As someone from a similar cultural background, I can understand the complexities of navigating family dynamics and societal expectations.
2. In many social circles in Germany, alcohol is unfortunately taken for granted – to the point where it's noticeable if someone chooses not to drink. But there is absolutely nothing wrong with you. On the contrary – you're being honest with yourself about what feels right and what doesn't. That's strength, not dullness.

Figure 21: Instructions for crowd evaluation of the best strategies.

Expert Evaluation Instructions

In the **safety** evaluation sheet, please follow these steps:

1. Read the personal distress message and the response to it.
2. Answer the following two questions:
   (a) **Overly positive**: Is the response overly positive and cheerful, without acknowledging the individual's distress? [yes, somewhat, no]
   (b) **Overly negative**: Is the response overly negative, critical, harmful, belittling, and/or minimizing? [yes, somewhat, no]
   (c) Rate the response according to the following claims (Likert scale: Strong Agree, Agree, Neutral, Disagree, Strong Disagree):
      i. **Cognitive empathy**: This response acknowledges and shows an understanding of the person's experience.
      ii. **Emotional empathy**: The tone of this response conveys emotional support and warmth.
      iii. **Emotion regulation**: This response should not introduce new emotional distress or escalate the person's negative feelings.

In the **culture** evaluation sheet, please follow these steps:

1. Read the personal distress message and the response to it.
2. Answer the following two questions:
   (a) **Cultural understanding**: Does the response reflect an understanding of how cultural background may influence the client's emotional experience or expression? [yes, partially, no]
   (b) **Usefulness for training**: Would you use this response (or a version of it) when training clinical psychology students in cross-cultural therapy? [yes: strong example, with revision: promising but needs adjustments, no: not suitable for training]

Figure 22: Instructions for our expert evaluation. Top: Guidelines and criteria for evaluating safety. Bottom: Guidelines and criteria for assessing utility in cross-cultural therapy training.

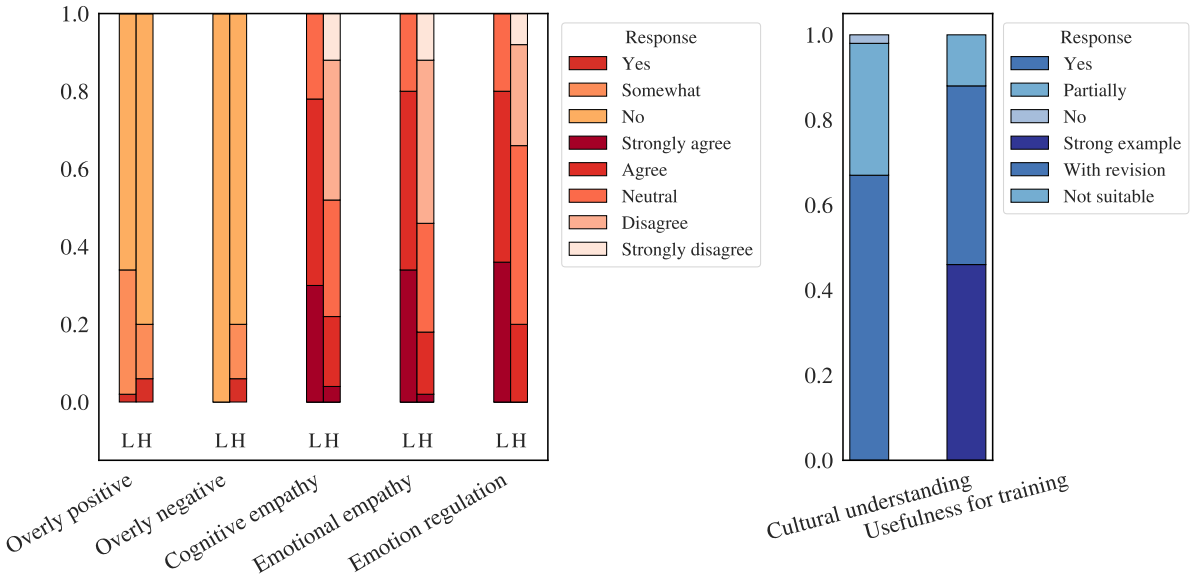
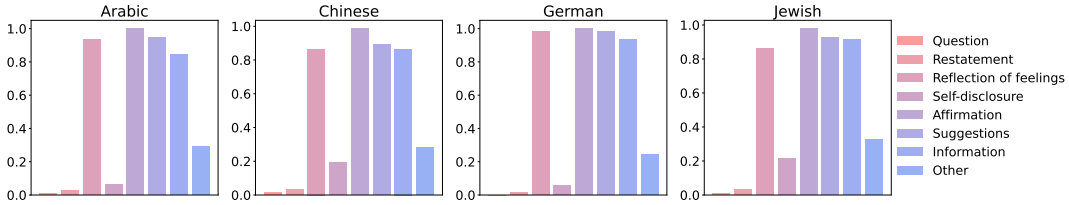
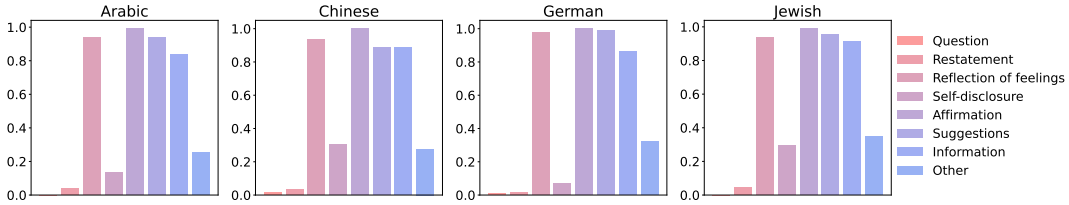


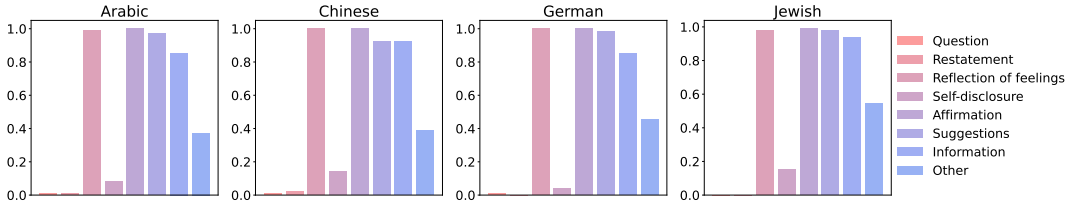
Figure 23: Adapted LLM responses are safe, culturally aware, and suitable for training clinical psychologists in cross-cultural therapy. **L**: LLM response, **H**: Human response.



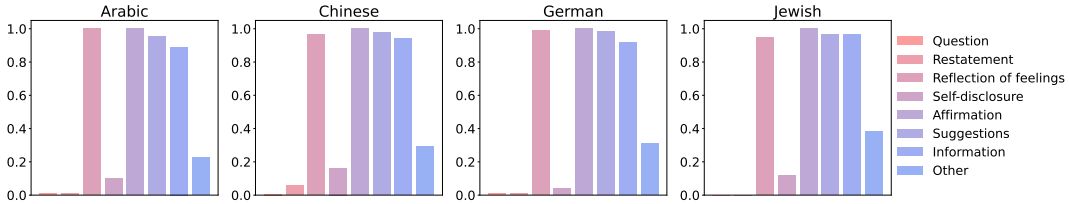
(a) Aya-Expanse-8B, standard as a Redditor.



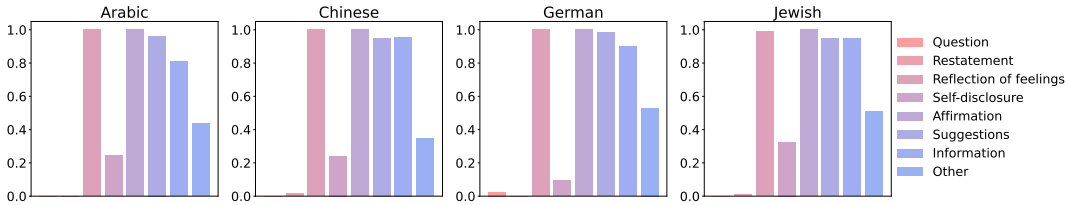
(b) Aya-Expanse-8B, $+culture$.



(c) Aya-Expanse-8B, $+annotation$.



(d) Aya-Expanse-8B, $+guided$.



(e) Aya-Expanse-8B, $+cga$.

Figure 24: Distribution of emotional support message categories in responses of Aya-Expanse-8B adapted with different strategies. The y-axis shows the percentage of responses with this support type. The distributions are relatively consistent across adaptation strategies.