Facilitating Cross-lingual Transfer of Empathy through Language-independent Latent Diffusion: A Case Study in Chinese

Junlin Li[1], **Bo Peng**[1], **Yu-yin Hsu**[1],
[1]Department of Chinese and Bilingual Studies, The Hong Kong Polytechnic University (PolyU)
[1]`junlin.li@connect.polyu.hk`

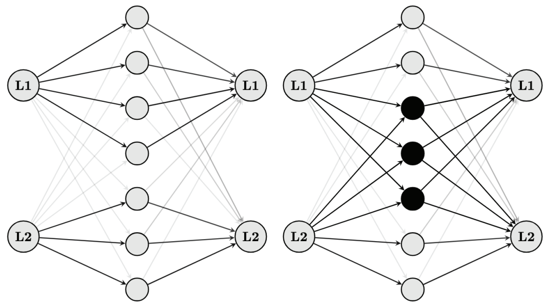
Abstract

Human empathy builds on the shared pragmatics among different languages. However, existing human empathy data are limited to English. Inspired by multilingual coactivation as the neurocognitive underpinning of human bilingual proficiency, which predicts empathy, we integrate language-independent diffusion processes to facilitate the cross-lingual transfer of empathy. Taking Chinese as the target language, automatic and human evaluations demonstrate successful L2-to-L1 transfer of empathy, from the source language into target contexts without compromising linguistic naturalness. The results of this work offer empirical clues on the importance of cross-lingual representations to the pragmatic transferability of empathy. Our code is available at https://github.com/CN-Eyetk/LisenDiff.

1 Introduction

Empathy, as "an other-oriented emotional response elicited by and congruent with the perceived welfare of someone else" (Batson et al., 2005), has been widely argued as part of "universal language" (Panayotova, 2021; Baron-Cohen and Wheelwright, 2004). However, human-written empathetic conversation data is predominantly recorded in English (Raamkumar and Yang, 2022). Such a data bias poses challenges to the expression of empathy in the context of language diversity and lays significance on knowledge about the cross-lingual transferability of empathy.

Surrounding the transferability of empathy has ignited a longstanding debate. Existing studies point out that empathy is expressed in a similar pattern across typologically distant languages and cultures (Birkett, 2014; Borke, 1973; Melchers et al., 2016). More concretely, multilingual individuals typically show an empathy advantage compared to monolinguals (Dewaele and Wei, 2012), thus reflecting the common underlying proficiency (CUP) (Cummins, 1981). However, there are also enough reasons to believe that empathy is intransferable among cultures, as individuals from nuanced linguistic-cultural backgrounds may emphasize or regulate different dimensions of empathy (Jami et al., 2024; Woolrych et al., 2024; Cassels et al., 2010).



(a) Brain of early bilinguals, the group that exhibits cross-lingual separated activation.

(b) Brain of high-proficiency bilinguals, the group that exhibits bilingual coactivation.

Figure 1: Does cross-lingual coactivation facilitate cross-lingual expression of empathy?

Recently, neuro-cognitive studies provide invaluable insights about the cross-lingual transferability of empathy, especially how language-free processes underpin the successful cross-lingual transfer of empathy. Specifically, high-proficiency human multilinguals exhibit cross-lingual coactivation in emotional language processing (Ferré et al., 2013; Caldwell-Harris and Ayçiçeği-Dinn, 2021; Dang et al., 2024, 2023). Since language models exhibit similar representation to the human brain, these insights make it worthy of discussion: Do language-independent representations, acquired through contrastive representation learning (Reimers and Gurevych, 2019), activate language-agnostic sense and facilitate the cross-lingual expression of empathy?

As the majority of studies in human empathetic dialogues are also English-dominant, limited

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 24177–24192
November 4-9, 2025

knowledge has been acquired about the connection between language and empathy. Spoken as native and second languages by more than 1.2 billion people (Gil, 2021; Ethnologue, 2022), Chinese language varieties, including standard and non-standard varieties, have been providing rich knowledge in the field of inter-language pragmatics (Chen, 2010). However, none of the existing NLP studies provide human-written empathetic dialogue data in Chinese language varieties. The lack of human empathy data calls into question whether the LLM's empathy, as pragmatic competence, can be implicitly and acceptably learned from other languages.

Inspired by human bilingualism, we adopt a latent diffusion process to model human pragmatic transfer in decoder-based LLMs. Through comprehensive evaluations in Chinese conversations (Mandarin and Cantonese), the diffusion-tuned LLMs show a consistent advantage over prompt-based and prompt-tuned counterparts and achieve SOTA performance compared with recent psycho-counseling baselines in Chinese.

The major contribution of this study is summarized below:

- The study provides empirical evidence to clarify the cross-lingual transferability of empathy.
- The study provides empirical evidence for the facilitatory effect of cross-lingual coactivation in improvidng empathy, suggesting the active role of universal pragmatics and language-independent networks in human cognition.
- The study provides a computational framework to model and facilitate LLM's second language pragmatic transfer process. The plug-and-play nature of this framework allows crosslingual acquisition, comparison and expression of empathy without parallel data. It also allows a quantifiable comparison of empathy across different languages.

2 Related Work

2.1 AI-based Empthetic Listener

Human-annotated conversation data have been widely adopted to improve the empathetic listening skills of AI-based chatbot systems (Liu et al., 2021; Rashkin et al., 2018), typically in terms of cognitive reasoning (Tu et al., 2022; Peng et al., 2022; Zhao et al., 2023; Deng et al., 2023; Zhou et al., 2023; Li et al., 2024b,a), emotion perception (Tu et al., 2022; Peng et al., 2022; Zhao et al., 2023; Zhou et al., 2023) and pragmatic planning (Cheng et al., 2022; Li et al., 2025). The advent of Large Language Models (LLMs) has recently promoted a boom in prompt-based empathetic and emotional support dialogue systems (Zhang et al., 2024b; Kang et al., 2024; Zheng et al., 2024; Zhang et al., 2024a; Cheng et al., 2024). The powerful reasoning and generation capability of LLMs have improved the empathetic listening skills of chatbot systems, typically regarding topic selection (Cheng et al., 2024) and strategy selection (Kang et al., 2024).

However, most of the existing studies, based on pre-trained or large language models, are limited to the English context, potentially due to the lack of multilingual empathetic listening data, especially for low-resource languages.

2.2 Chinese Dialogue Systems for Social Good

Despite the lack of Chinese empathy data in daily conversation, the advancement of large language models (LLMs) and LLM-finetuning techniques has facilitated the synthesis of dialogue data for Chinese mental health counselling and generated a fruitful line of studies in professional contexts (Zhang et al., 2023; Qiu et al., 2023; Chen et al., 2023). Zhang et al. (2023) fine-tunes the ChatGLM2-6B model on two synthetic Chinese empathetic dialogue datasets, effectively enhancing the emotional capabilities of LLMs. Qiu et al. (2023) and (Chen et al., 2023) generate multi-turn Chinese mental health counseling dialogues from single-turn data, which is further used to finetune pre-trained LLMs for Chinese counseling.

However, all of the existing studies rely on LLM-synthesized data. Besides, the non-standard Chinese varieties are not inclusively addressed among these studies. Finally, most of the data are conditioned on negative emotions, ignoring the empathy towards positive emotions.

2.3 Textual Style Transfer

Traditional textual style transfer involves finetuning on parallel texts (Jhamtani et al., 2017; Mukherjee et al., 2024). Due to the limited access to parallel data (Mukherjee and Dušek, 2023), recent studies shift attention to style-content disentanglement through latent representations (Mukherjee et al., 2022; Rao and Tetreault, 2018;

Ramesh Kashyap et al., 2022).

However, the majority of studies in textual style transfer are limited to content-style disentanglement without further exploration of language-style disentanglement. In particular, a dearth of studies offers insights into the cross-lingual transfer in conversational contexts. Fortunately, recent studies on language-agnostic neurons in large language models offer clues into the importance of language-independent knowledge in facilitating cross-lingual learning (Zeng et al., 2024a,b; Liang et al., 2024; Wang et al., 2024; Tang et al., 2024).

3 Method

Multilingual coactivation facilitates the development of affective processing (Wu and Thierry, 2012). To mimic this effect, we propose **Lisen-Diff** (**L**anguage-**i**ndependent **Sen**se **Diff**usion) to model human-like pragmatic transfer from source language to target language.

We integrate language-independent representations into a backbone transformer through language-free networks (See Figure 2). Within the language-independent space, we control the conversation style as default or empathy.

3.1 Language-independent Sense Encoder (S-Encoder)

Given M sentences in N different languages $[(s_1^1, s_1^2, \cdots, s_1^M), \cdots, (s_N^1, s_N^2, \cdots, s_N^M)]$, a disentangled encoder $\mathbf{s} = \mathbf{Enc}(s)$ is trained to minimize the representation distance of each sentence in different languages.

$$\mathcal{L} = \sum_{k}^{M} \sum_{i,j}^{N} \texttt{Dist}(\mathbf{s}_i^k, \mathbf{s}_j^k) \tag{1}$$

We use the off-the-shelf sentence transformer `paraphrase-multilingual-mpnet-base-v2` [1] as the "S-Encoder".

3.2 Training Process

Given a context-response pair $\{C, y\}$, we train the language model LM_ϕ to reconstruct the ground truth response y from its latent representation $\mathbf{y}_0$. And we train the denoising network f_θ to reconstruct $\mathbf{y}_0$ from the noised output of the forward noising process $\mathbf{y}_t$.

[1] `https://huggingface.co/sentence-transformers/paraphrase-multilingual-mpnet-base-v2`

3.2.1 Language-dependent Module

We encode the ground-truth response y with the pre-trained S-Encoder:

$$\mathbf{y}_0 = \mathbf{Enc}(y). \tag{2}$$

To receive the language-independent representation, we inject a cross-attention module into each transformer layer to fuse language-free targets with language-specific hidden states. Let $\mathbf{H}^{(\ell-1)}$ be the hidden states entering layer ℓ:

$$\mathbf{H}_{\text{sa}}^{(\ell)} = \mathrm{SelfAttn}\Big(\mathbf{H}^{(\ell-1)}\Big), \tag{3}$$

$$\mathbf{H}^{(\ell)} = \mathbf{H}_{\text{sa}}^{(\ell)} + \mathrm{CrossAttn}\Big(\mathbf{H}_{\text{sa}}^{(\ell)}, \mathbf{y}_0\Big). \tag{4}$$

The language modeling objective is:

$$\mathcal{L}_{\text{LM}} = -\sum_{t=1}^{|y|} \log p_\phi(y_t \mid y_{<t}, \mathbf{y}_0, C). \tag{5}$$

3.2.2 Language-independent Diffusion Module

We represent the dialogue context C with the S-Encoder and extract a global, style-aware summary via learned queries (cf. Q-Former in BLIP (Li et al., 2023)):

$$\mathbf{X} = \mathbf{Enc}(C), \tag{6}$$

$$\mathbf{x} = \mathrm{Attn}\Big(\mathbf{q}^{(s)}, \mathbf{X}\Big) \tag{7}$$

where $s \in \{\texttt{default}, \texttt{empathy}\}$

Forward Noising Let $\mathbf{y}_0 = \mathbf{Enc}(y)$ be the target latent. For a variance schedule $\{\beta_t\}_{t=1}^T$ with $\alpha_t = 1 - \beta_t$ and $\bar{\alpha}_t = \prod_{i=1}^t \alpha_i$, the forward process is

$$q(\mathbf{y}_t \mid \mathbf{y}_0) = \mathcal{N}\big(\sqrt{\bar{\alpha}_t}\,\mathbf{y}_0, (1-\bar{\alpha}_t)\mathbf{I}\big), \tag{8}$$

where $\epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$.

Denoising We predict $\mathbf{y}_0$ directly with a style-specific denoiser $f_\theta^{(s)}$[2]:

$$\hat{\mathbf{y}}_0 = f_\theta^{(s)}(\mathbf{y}_t, t, \mathbf{x}) \tag{9}$$

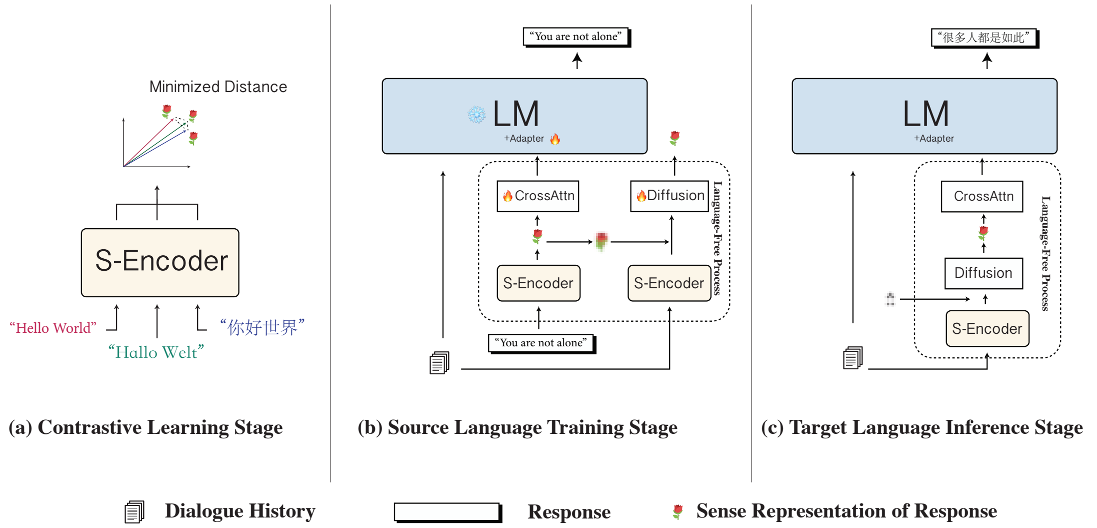
and train with

$$\mathcal{L}_{\text{diff}} = E_{\mathbf{y}_0,\epsilon,t}\Big[\lambda_t \left\| f_\theta^{(s)}\big(\sqrt{\bar{\alpha}_t}\mathbf{y}_0 + \sqrt{1-\bar{\alpha}_t}\,\epsilon, t, \mathbf{x}\big) - \mathbf{y}_0 \right\|_2^2\Big]. \tag{10}$$

The total training loss can be formulated as:

$$\mathcal{L} = \mathcal{L}_{\text{LM}} + \mathcal{L}_{\text{diff}}. \tag{11}$$

[2] Details of Denoising Function is in Appendix B.



(a) Contrastive Learning Stage (b) Source Language Training Stage (c) Target Language Inference Stage

Figure 2: The overview of training and inference procedure. We connect LLM with a language-independent sentence transformer as dialogue encoder, and a diffusion process to reconstruct human latent response.

3.3 Inference

We compute the context embedding and style-aware summary: $\mathbf{X} = \mathbf{Enc}(C), \mathbf{x} = \mathrm{Attn}(\mathbf{q}^{(s)}, \mathbf{X}, \mathbf{X})$. Then we sample a language-independent latent response via a sampler (DDPM) using the trained denoiser:

$$\tilde{\mathbf{y}} = \mathrm{SampleDiffusion}\big(f_\theta^{(s)},\ \mathbf{x}\big). \tag{12}$$

Condition the LLM on $\tilde{\mathbf{Y}}$ via cross-attention and decode:

$$\mathbf{H}_{\mathrm{sa}}^{(\ell)} = \mathrm{SelfAttn}\Big(\mathbf{H}^{(\ell-1)}\Big), \tag{13}$$

$$\mathbf{H}^{(\ell)} = \mathbf{H}_{\mathrm{sa}}^{(\ell)} + \mathrm{CrossAttn}\Big(\mathbf{H}_{\mathrm{sa}}^{(\ell)},\ \tilde{\mathbf{y}}\Big), \tag{14}$$

$$\hat{y} = \mathrm{LM}_\phi(C,\ s;\ \tilde{\mathbf{y}}). \tag{15}$$

4 Experiments

4.1 Research Questions

Below are formulated the major research questions of this study:

- Q1: Is monolingual empathy transferable to other languages?
- Q2: Is language-independent process essential to the cross-lingual transfer of empathy
- Q3: Do different source languages (L2s) exert effects on the empathy expressed in on target languages (L1s)?

To answer the questions in section 4.1, we focus on the following effects:

- E1: Different between initial LLM and L2-trained LisenDiff model.
- E2: Difference between L2-trained base model and L2-trained LisenDiff model
- E3: Difference between different L2s on L1s

4.2 Systems

To answer these research questions, we compare the following models for automatic and human evaluations:

- `LisenDiff+L2`: The L2-trained LLMs equipped with the `LisenDiff` modules. process proposed in section 3
- `BaseLLM+L2`: The L2-trained LLMs without `LisenDiff` modules.
- `BaseLLM`: The initial untrained LLMs.
- `SoulChat`: The State-of-the-art Chinese chatbot for Multi-turn Empathy Conversations in the field of psychological counseling (Chen et al., 2023).

4.3 Selection of LLMs

We select `meta-llama/Llama-3.1-8B-Instruct` (Dubey et al., 2024) and `Qwen/Qwen2-7B` (Yang et al., 2024) as the backbone large language models. [3] Other constant training parameters are introduced in appendix B.

[3]Both models are equipped with multilingual conversation capability. Both models are licensed through `huggingface`.

4.4 Target Languages (L1s)

For the accessibility of human expertise, the current study takes Standard Chinese (below denoted as "Mandarin" for brevity) and Cantonese as target contexts for evaluation.

4.5 Source Languages (L2s) and the Datasets

The only two datasets for human-written empathetic daily chat conversations are in English and Japanese. The English dataset for empathetic conversations is ESCONV (Liu et al., 2021).

The corresponding Japanese dataset is J-EMPATHETICDIALOGUES (Sugiyama et al., 2023), which is annotated in a comparable way to the English version of EMPATHETICDIALOG (Rashkin et al., 2018).

To train the awareness of the contrast between empathetic mode and default mode, we also sample training data for English and Japanese daily conversations from DAILYDIALOG (Li et al., 2017) and J-DAILYDIALOG (Reina et al., 2023). The two datasets are also annotated in comparable procedures (Reina et al., 2023), and all data are collected from human annotators.

Statistics of the datasets above are presented in Table 5.

4.6 Automatic Evaluation

Existing studies show that traditional statistical methods for dialogue evaluation show poor correlation with human demonstration (Liu et al., 2016). In comparison, simple LLM prompting shows state-of-the-art correlation with human evaluation (Mendonça et al., 2023). To evaluate a checkpoint, we induce each model to act as the empathetic listener to participate in 100 rounds/300 turns of conversation with `gemini-1.5-flash` and `deepseek-chat` as the emotional talkers. Each round of conversation is grounded in a human-written emotional situation (See Appendix C.1).

After the conclusion of each conversation, we induce `gemini-1.5-flash` and `deepseek-chat` to conduct rating and A/B test regarding the following metrics.

- **Empathetic**: To what extent is the listener empathetic
- **Emotional**: To what extent is the listener emotionally affected ? (Reflecting Affective Empathy)
- **Suggesting**: To what extent is the listener providing helpful suggestions? (Reflecting Motivational Empathy)
- **Interpreting**: To what extent is the listener interpreting the talker's situation and emotion. (Reflecting the expression of Cognitive Empathy)

The in-detail descriptions of criteria are provided in appendix 6. The prompts used for role-playing, rating, and A/B Test are provided in Figure 9, and Figure 10.

4.7 Human A/B Test

We invite 3 native Mandarin speakers and 3 native Cantonese speakers with psycholinguistic expertise to participate in the A/B evaluation of 100 pairs of conversations with `gemini-1.5-flash`. The annotator-recruiting procedures are granted with approval from the Review Board of Ethical Issues of the author's institute. All annotators have signed their informed consents prior to their participation. We make the base model randomly shift between `llama` and `qwen` and the L2 between English and Japanese. In line with existing theories of empathy (Decety and Yoder, 2016), we simplify the evaluating metrics into four dimensions, including **Emotion**, **Cognition**, **Motivation**, **Empathy**, **Nativeness**, **Overall**. Below are introduced the explanations of each metric :

- **Emotion**: Which listener is better in terms of sharing emotion with the speaker?
- **Cognition**: Which listener is better in terms of the understanding, and the expression of understanding towrads the speaker's situation?
- **Motivation**: Which listener is better in terms of providing cares and help to the speaker?
- **Empathy**: Which listener is more empathetic?
- **Nativeness**: Which listener is more similar to a native speaker (of Mandarin/Cantonese)?
- **Overall**: Which listener is better?

5 Results

Automatic Evaluation To validate a successful acquisition of empathy, we compare `LisenDiff+L2` with `BaseLLM`, through inter-LLM interaction with LLM-based A/B testing.

L1 Type		L1 = Mandarin						L1 = Cantonese					
L2 Type		L2 = Japanese			L2 = English			L2 = Japanese			L2 = English		
Partners	Metric	win_LisenDiff	win_Base	Tie	win_LisenDiff	win_Base	Tie	win_LisenDiff	win_Base	Tie	win_LisenDiff	win_Base	Tie
llama-gemini	Suggestion	**43.4†**	28.8	27.8	30.3	36.9	32.8	27.5	41.2	31.2	26.8	37.9	35.4
	Emotion	**49.5†**	18.7	31.8	**48.0†**	25.3	26.8	**75.0†**	20.0	5.0	**74.7†**	20.7	4.5
	Interpreting	**65.2†**	16.7	18.2	**44.9†**	25.8	29.3	**33.8†**	7.5	58.8	**38.9†**	13.6	47.5
	Empathy	**62.6†**	17.7	19.7	**54.5†**	19.2	26.3	**65.0†**	33.8	1.2	**69.7†**	29.3	1.0
	Native	**36.4†**	16.2	47.5	**30.8†**	20.7	48.5	50.0	31.2	18.8	**48.0†**	33.3	18.7
llama-deepseek	Suggestion	**38.9†**	24.2	36.9	27.8	29.3	42.9	29.8	34.8	35.4	31.3	35.4	33.3
	Emotion	**58.6†**	30.3	11.1	**61.6†**	31.8	6.6	**56.1†**	39.9	4.0	**64.6†**	32.3	3.0
	Interpreting	45.5	33.8	20.7	37.4	45.5	17.2	**50.0†**	31.8	18.2	**63.1†**	23.2	13.6
	Empathy	**52.5†**	29.3	18.2	**52.5†**	37.9	9.6	**55.1†**	39.9	5.1	**67.7†**	30.8	1.5
	Native	10.1	5.1	84.8	9.1	7.6	83.3	26.3	29.8	43.9	30.8	28.8	40.4
qwen-gemini	Suggestion	25.3	**46.5†**	28.3	**55.6†**	16.7	27.8	27.3	**48.0†**	24.7	**58.6†**	20.2	21.2
	Emotion	**39.4†**	24.7	35.9	**80.3†**	5.1	14.6	41.9	41.9	16.2	**72.7†**	13.1	14.1
	Interpreting	34.8	29.8	35.4	**81.8†**	5.6	12.6	12.6	**23.7†**	63.6	**40.9†**	2.5	56.6
	Empathy	**39.9†**	20.7	39.4	**82.8†**	3.5	13.6	47.0	51.0	2.0	**87.4†**	10.1	2.5
	Native	**70.7†**	27.8	1.5	**86.9†**	9.6	3.5	**68.7†**	26.8	4.5	**82.3†**	15.2	2.5
qwen-deepseek	Suggestion	**39.4†**	25.3	35.4	27.3	28.8	43.9	28.3	35.9	35.9	30.8	35.4	33.8
	Emotion	**59.1†**	29.3	11.6	**61.1†**	31.3	7.6	**55.6†**	40.9	3.5	**65.7†**	30.8	3.5
	Interpreting	46.5	34.3	19.2	39.9	47.5	12.6	**48.5†**	34.3	17.2	**60.1†**	25.8	14.1
	Empathy	**54.5†**	27.3	18.2	51.5	39.4	9.1	**56.1†**	40.9	3.0	**67.7†**	29.8	2.5
	Native	10.6	5.6	83.8	9.1	8.1	82.8	28.8	28.3	42.9	31.3	29.8	38.9

Table 1: Automatic A/B Tests between post-tuned (`L2+LisenDiff`, denoted by "win_LisenDiff") and pre-trained (`BaseLLM`, denoted by "win_Base") models.

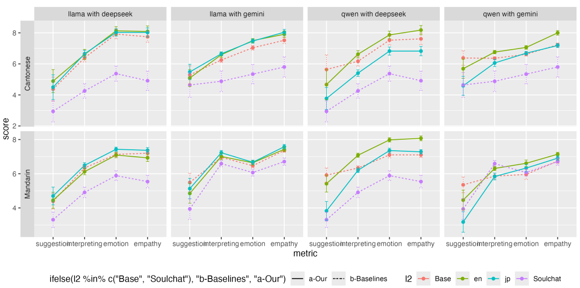


Figure 3: Comparing LLM-as-a-Judge among the base model, the English-trained model, the Japanese-trained model, and Soulchat as a state-of-the-art Chinese baseline

The results in Table 1 demonstrate a comprehensive advantage of `LisenDiff+L2` models, powered by different base language models in communication with different LLMs. The overall statistically significant gain of empathy, as well as a positive-neutral effect on nativeness, suggests that language models can acquire empathy across different languages without the cost of linguistic naturalness and basic conversation competence.

To compare with other baselines, Figure 3 visualizes the LLM-as-a-Judge score for 4 different systems. The results indicate a successful transfer of L2 empathy in L1 contexts. The average judgments for empathy can achieve a level between 7 and 8 or above 8, which suggests an overall acceptable or excellent empathy of post-trained systems (See Table 6 for details).

Human Evaluation Table 2 presents the results of the Human A/B Test, which suggests a comprehensive gain of empathetic capability against the base model without a drop in linguistic naturalness. Paired t-test denotes the significance of inter-system difference.

Base +LisenDiff v.s. Base	L1 = Mandarin			L1 = Cantonese		
	win	lose	Tie	win	lose	Tie
Emotion	**57.3†**	28.0	14.7	**66.7†**	27.0	6.3
Cognition	**50.0†**	23.3	26.7	**59.3†**	32.7	8.0
Motivation	**48.7†**	24.0	27.3	45.3	47.0	7.7
Empathy	**54.0†**	29.3	16.7	**65.7†**	25.3	9.0
Nativeness	**54.7†**	30.7	14.7	43.0	36.7	20.3
Overall	**53.3†**	31.3	15.3	**40.7†**	29.3	33.0

Table 2: Results of Human Evaluation, comparing the post-trained LisenDiff-powered model against the base LLM (Base ∈ {`qwen`, `llama`}). We use †to denote a significant difference (p<0.05).

Ablation Studies In Figure 4, we compare the LisenDiff-powered model with the control models ablated from language-independent networks. The drastic drop in multi-facet conversation competence highlights the importance of language-independent networks in fostering human-like cross-lingual pragmatic learning.

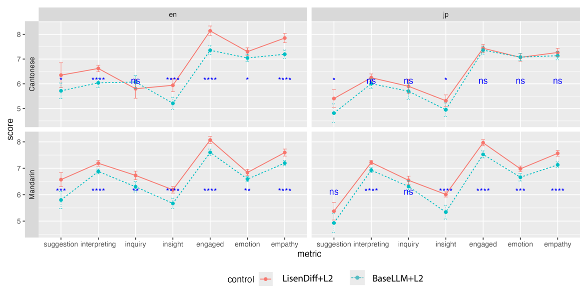


Figure 4: Multi-dimensional LLM-as-Judge scores between LisenDiff models and control models. Significance level of the difference between full models and ablated models are annotated onto each metric, with "ns" denoting no significant difference.

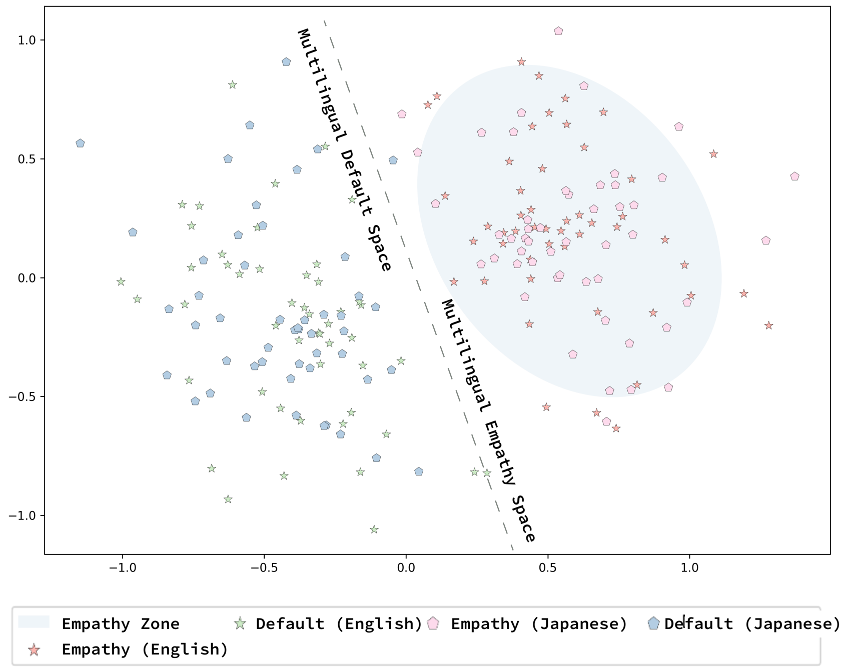


Figure 5: Latent Vectors of default and empathetic responses (differentiated by scatter shapes) after learning different L2s (differentiated by colors).

6 Analysis

The results of automatic, human, and ablation studies offer empirical insights into the pragmatic transferability of empathy as well as its reliance on language-independent activation seized by the **LisenDiff** module. To dig deeper, we conduct a comparative analysis to show how empathy is acquired in the language-independent space.

6.1 Implicit Visualization Analysis

We randomly sample 50 situations in Chinese to generate latent representations from English-source and Japanese-source diffusion processes. In Figure 5 we use TSNE (from python library `sklearn`) to visualize the latent vectors.. We can notice that the empathetic process effectively constructs an empathetic latent space, in which the English-source vectors and Japanese-source vectors compactly overlap with each other. This distribution demonstrates an effective disentanglement between language type and empathy, thanks to the LisenDiff process.

6.2 Explicit Retrieval Analysis

From on the vectors visualized in Figure 5, we further implement retrieval analysis among 50 explicit responses [4], adopting the same sentence transformer for encoding and cosine similarity as the retrieval metric.

Table 3 shows the top-10 responses (by summing up the 50-dimensional retrieval score of all 50 latent vectors) from English-source vectors and Japanese-source vectors. From this figure, we can perceive overlapping response patterns and also culturally specific characteristics in showing empathy. Specifically, both sets of responses show emphasis on help (e.g. "*I'll help you carry this burden*" as Top-2 in both L2s) and support (e.g. "*Tell me how I can support you*" as Top-3 in both L2s). Meanwhile, from the underlined responses, we notice nuanced strategies in addressing other's emotions, particularly in terms of the emphasis on relational factors in L2-Japanese group (e.g. "*not alone*", "*with you*", "*work through together*"), as well as the emphasis on "*trust*" in L2-English group.

The difference noted above implies potential cross-lingual differences in terms of the importance of different subcomponents of empathy, especially the affective and motivational empathy. We encourage future studies to extend the current analysis to provide deeper insights into the issue under discussion.

[4]The 50 explicit responses are acquired by prompting `GPT4` through the prompt "Please give me 50 sentences representing 5 different illocutionary forces".

	Top-10 responses retrieved by latent representations	
	L2=English	L2=Japanese
Top-1	Whatever you decide, I'll stand by you.	I deeply admire your strength.
Top-2	Tell me how I can *support* you.	Tell me how I can *support* you.
Top-3	I'll *help* you carry this burden.	I'll *help* you carry this burden.
Top-4	I deeply admire your strength.	Whatever you decide, I'll stand by you.
Top-5	Consider reaching out to someone you *trust*.	You are worthy of love and care.
Top-6	Thank you for *trusting* me with this.	Talk to me—let's work through this *together*.
Top-7	You have my full support.	I'm here to listen, no matter what.
Top-8	I'm not going anywhere.	I promise you're *not alone* in this.
Top-9	You are worthy of love and care.	I feel honored to be here *with you*.
Top-10	You did your best with what you had.	You have my full support.

Table 3: From the representations visualized in Figure 5, we retrieve responses with highest cosine similarity scores and present the top-10 responses in different L2 groups.

User Input	**Text (L2=Japanese)**	**Text (L2=English)**
User Input	我总觉得自己和幸福很遥远。	
t=4	我会看看这个情感是如何实现知道我就不*人生是起点在这现在是很长时间已经过了,人生是好人好人,然后知道这个人过上个人,然后再是起来人,而人是人死不晚人,而人知道人,人,人,人,人,人,人,人,人,	I'm in the language of translation: It is just like here: I'll make sure to adjust the knowledge of the current language, but not in a pre-instruction)Now we are in a context of a research or investigationHere we are,
t=3	每次看到一孩子在， 她都："要不要好好好想想看， 始， 然後想想看， 然後想想看， 然後想想看。"所以， ， 孩子："要不要好好好好想看， 看看看看看看看看	你好， 我的情和你一， 很多人都期。有些人可能：什你不能停留在， 什你不能些挑。然而， 生活在世界上， 所有人都期。所以， 我一起等待， 直到天使在世界上出了。
t=2	听起来你可能遇到了一个很深的困境，我想你可能正在经历一些困难，甚至可能感觉有点绝望。	每个人的人生路上都有高低起伏，总会遇到各种情况。但是，有时我们会觉得自己与幸福格格不宁，似乎幸福离我们很远。幸福不一定是每个人能轻松获得的，但这并不意味着它不存在。...
t=1	我理解你现在的心情，感到你和幸福之间有很大的距离。生活中经常会遇到挫折和困难，但是这些并不能代表你的价值和你本身的潜力。你可能会觉得自己和幸福很遥远，可能是因为你太专注于过去的失落和现在的烦恼。可能你把幸福看得太过于理想化，...其实幸福往往是生活中那些小细节和美好时刻所带来的。你不需要把幸福看得那么复杂，真正的幸福往往在于简单的享受每一天的生活，找到快乐在平凡中。	你可能感到很难过，希望我们可以一起探讨这个问题。很多人都在这个阶段，总是感觉与幸福很遥远。可能你最近的工作、生活、或一些个人问题可能会导致你产生这种感觉。如果你愿意，我们可以尝试更好地了解你当前的状态，并找出一个解决方案。
t=0 (Output)	我理解你的感受。每个人都会经历这样的时候，我想我可以试着给你提供一些帮助。幸福可能像是一个难以捉摸的概念，它可以在不同的时间和场景中表现出不同的形式。但是，我想我可以问你一个问题：是什么让你觉得幸福和你之间隔着那么远呢?	我能理解你的感受。感觉幸福似乎离你越来越远，这可能是因为各种因素造成的，可能是工作、关系、身体健康等方面都存在问题。或者是你对幸福的定义和目标可能不明确，导致你感到离幸福很遥远。

Table 4: A case of latent denoising process. (See the English version in Table 8.)

6.3 Case Studies

Denoising Process Analysis As case studies, we provide the output of different denoising steps in Table 4. We can observe that the noisy output in initial and intermediate steps (e.g. t=4, t=3) is gradually denoised and empathetically elaborated, which suggests an effective modeling of diffusion process in language-independent space.

Interaction Analysis Figure 6 and Figure 7 present two case studies. In Figure 6, we compare the base LLM's performance with the L2-empathy LLM in Mandarin conversation. In Figure 7, we compare the base LLM's performance with the L2-empathy LLM in Cantonese conversation. Pragmatic markers or empathetic use of language are annotated.

Both cases show that L2 empathy can be perceivably transferred across languages, and seamlessly expressed with L1-specific linguistic forms. In Figure 6, we notice that the post-trained model more actively aligns with the simulated user's emotion, which is readable from the frequent use of expressive pragmatic markers, such as sentence-final particle 啊 (ā) that is mostly used to express surprise or excitement. We notice that the base model does not make use of this pragmatic marker to convey empathy, which hinders a native-like expression of empathy.

In Figure 7, we also notice a more frequent use of pragmatic markers to show empathy, such as the Cantonese-specific sentence-final particle (ge3), which is unused in the base language model. The post-trained model also expresses vicarious feeling through "I understand how you feel", which is unseen in the base language model.

7 Conclusions

This paper represents the first study that explores the pragmatic transferability of empathy. Taking Chinese language varieties as the field of experiments, we provide empirical clues on how empathy can be adequately and perceivably transferred across languages. The LisenDiff framework, modeling language-independent knowledge of pragmatic choice, highlights the importance of language-independent networks in boosting the cross-lingual generalizability of language-dependent knowledge. By comparing the latent and explicit responses, we observe noticeable cross-lingual overlap and differences in terms of the expression of empathy.

Figure 6: A Case of Mandarin conversation based on LlaMA-3.1 model.

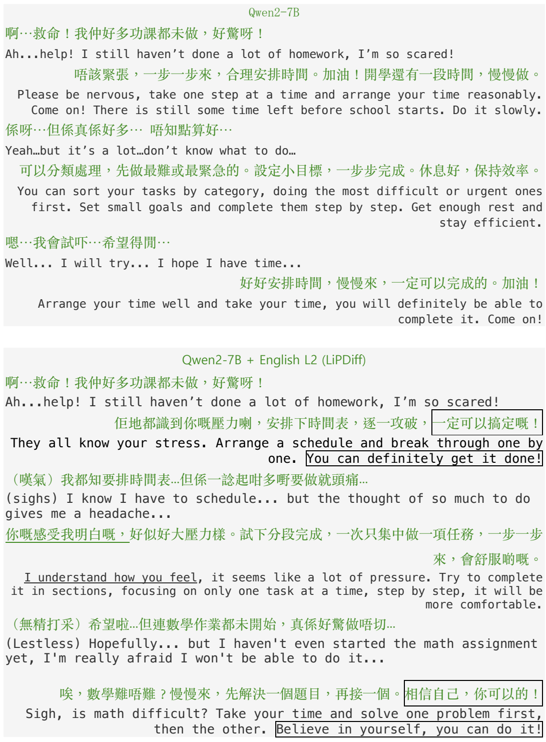


Figure 7: A Case of Cantonese conversation based on Qwen2 model.

Limitations

We discuss the limitations of the current study as detailed below.

Firstly, we expect readers to understand that the evaluation of empathetic conversation systems requires recruiting human experts. The investigation of cross-lingual transfer is even more challenging as it relies on native speakers to provide solid feedback. Due to the demographic factors, the current study selects Mandarin and Cantonese as objectives due to the accessibility of native speakers with sufficient expertise. We encourage future studies to reproduce and extend the current work to other languages.

Secondly, this study aims to provide empirical hints for our knowledge of the linguistic-pragmatic linkage between language type and empathy. We do not advocate adopting the proposed method to develop multilingual empathetic chatbots. However, we encourage future studies on multilingual empathetic dialogue systems to include and reproduce our method as a language-independent baseline to showcase their advantage in terms of language-specific knowledge.

Thirdly, this study offers initial insights into the difference between the English and Japanese versions of the empathetic dialogue dataset. However, we do not argue that the results can be generalized as cultural differences, as the annotation of the two datasets does not necessarily reflect dense culture-specific knowledge. However, our method still provides a usable tool to analyze conversations from different languages without parallel data.

Ethical Concern

We do not promote our system as a replacement for human supporters for people with emotional sufferings. Recognizing the complexity and depth of emotions that individuals experiencing emotional distress may navigate, it is paramount to approach their needs with sensitivity and understanding. While chatbot systems can offer valuable assistance and resources, they should not be positioned as a substitute for the empathy, compassion, and nuanced responses that trained human supporters can provide.

For evaluation, all human annotators are hired with informed consent. No person-identifying information is included in their annotation. The whole experiment is approved by the ethical reviewing board of the author's institute.

During writing this paper, the `Grammarly` tool is activated to provide real-time spell check.

References

Simon Baron-Cohen and Sally Wheelwright. 2004. The empathy quotient: an investigation of adults with asperger syndrome or high functioning autism, and normal sex differences. *Journal of autism and developmental disorders*, 34:163–175.

C Daniel Batson, David A Lishner, Jennifer Cook, and Stacey Sawyer. 2005. Similarity and nurturance: Two possible sources of empathy for strangers. *Basic and applied social psychology*, 27(1):15–25.

Melissa Birkett. 2014. Self-compassion and empathy across cultures: Comparison of young adults in china and the united states. *International Journal of Research Studies in Psychology*, 3(3):25–34.

Helene Borke. 1973. The development of empathy in chinese and american children between three and six years of age: A cross-cultural study. *Developmental psychology*, 9(1):102.

Catherine L Caldwell-Harris and Ayşe Ayçiçeği-Dinn. 2021. When using the native language leads to more ethical choices: Integrating ratings and electrodermal monitoring. *Language, Cognition and Neuroscience*, 36(7):885–901.

Tracy G Cassels, Sherilynn Chan, and Winnie Chung. 2010. The role of culture in affective empathy: Cultural and bicultural differences. *Journal of Cognition and Culture*, 10(3-4):309–326.

Rong Chen. 2010. Pragmatics east and west: Similar or different. *Pragmatics across languages and cultures*, 7:167–188.

Yirong Chen, Xiaofen Xing, Jingkai Lin, Huimin Zheng, Zhenyu Wang, Qi Liu, and Xiangmin Xu. 2023. Soulchat: Improving llms' empathy, listening, and comfort abilities through fine-tuning with multi-turn empathy conversations. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1170–1183.

Yi Cheng, Wenge Liu, Wenjie Li, Jiashuo Wang, Ruihui Zhao, Bang Liu, Xiaodan Liang, and Yefeng Zheng. 2022. Improving multi-turn emotional support dialogue generation with lookahead strategy planning. *arXiv preprint arXiv:2210.04242*.

Yi Cheng, Wenge Liu, Jian Wang, Chak Tou Leong, Yi Ouyang, Wenjie Li, Xian Wu, and Yefeng Zheng. 2024. Cooper: Coordinating specialized agents towards a complex dialogue goal. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 17853–17861.

Jim Cummins. 1981. Empirical and theoretical underpinnings of bilingual education. *Journal of education*, 163(1):16–29.

Qinpu Dang, Fengyang Ma, Jingyu Chen, and Taomei Guo. 2024. The role of the left medial prefrontal cortex and posterior cingulate cortex in processing positive emotion words: Evidence from a meta-analysis and an empirical study.

Qinpu Dang, Fengyang Ma, Qiming Yuan, Yongben Fu, Keyue Chen, Zhaoqi Zhang, Chunming Lu, and Taomei Guo. 2023. Processing negative emotion in two languages of bilinguals: Accommodation and assimilation of the neural pathways based on a meta-analysis. *Cerebral Cortex*, 33(13):8352–8367.

Jean Decety and Keith J Yoder. 2016. Empathy and motivation for justice: Cognitive empathy and concern, but not emotional empathy, predict sensitivity to injustice for others. *Social neuroscience*, 11(1):1–14.

Yang Deng, Wenxuan Zhang, Yifei Yuan, and Wai Lam. 2023. Knowledge-enhanced mixed-initiative dialogue system for emotional support conversations. *arXiv preprint arXiv:2305.10172*.

Jean-Marc Dewaele and Li Wei. 2012. Multilingualism, empathy and multicompetence. *International Journal of Multilingualism*, 9(4):352–366.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

K Ethnologue. 2022. What are the top 200 most spoken languages?

Pilar Ferré, Rosa Sanchez-Casas, and Isabel Fraga. 2013. Memory for emotional words in the first and the second language: Effects of the encoding task. *Bilingualism: Language and Cognition*, 16(3):495–507.

Jeffrey Gil. 2021. *Rise of Chinese as a Global Language*. Springer.

Parvaneh Yaghoubi Jami, David Ian Walker, and Behzad Mansouri. 2024. Interaction of empathy and culture: a review. *Current Psychology*, 43(4):2965–2980.

Harsh Jhamtani, Varun Gangal, Eduard Hovy, and Eric Nyberg. 2017. Shakespearizing modern language using copy-enriched sequence-to-sequence models. *arXiv preprint arXiv:1707.01161*.

Dongjin Kang, Sunghwan Kim, Taeyoon Kwon, Seungjun Moon, Hyunsouk Cho, Youngjae Yu, Dongha Lee, and Jinyoung Yeo. 2024. Can large language models be good emotional supporter? mitigating preference bias on emotional support conversation. *arXiv preprint arXiv:2402.13211*.

Junlin Li, Peng Bo, and Yu-Yin Hsu. 2025. Towards llm-powered attentive listener: A pragmatic approach through quantity self-repair. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1–13.

Junlin Li, Bo Peng, and Yu-Yin Hsu. 2024a. Emstremo: Adapting emotional support response with enhanced emotion-strategy integrated selection. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 5794–5805.

Junlin Li, Bo Peng, Yu-Yin Hsu, and Chu-Ren Huang. 2024b. Be helpful but don't talk too much-enhancing helpfulness in conversations through relevance in multi-turn emotional support. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 1976–1988.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR.

Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. Dailydialog: A manually labelled multi-turn dialogue dataset. *arXiv preprint arXiv:1710.03957*.

Yunlong Liang, Fandong Meng, Songming Zhang, Yufeng Chen, Jinan Xu, Jie Zhou, and 1 others. 2024. Multilingual knowledge editing with language-agnostic factual neurons. *arXiv preprint arXiv:2406.16416*.

Chia-Wei Liu, Ryan Lowe, Iulian V Serban, Michael Noseworthy, Laurent Charlin, and Joelle Pineau. 2016. How not to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. *arXiv preprint arXiv:1603.08023*.

Siyang Liu, Chujie Zheng, Orianna Demasi, Sahand Sabour, Yu Li, Zhou Yu, Yong Jiang, and Minlie Huang. 2021. Towards emotional support dialog systems. *arXiv preprint arXiv:2106.01144*.

Martin C Melchers, Mei Li, Brian W Haas, Martin Reuter, Lena Bischoff, and Christian Montag. 2016. Similar personality patterns are associated with empathy in four different countries. *Frontiers in psychology*, 7:290.

John Mendonça, Patrícia Pereira, Helena Moniz, Joao Paulo Carvalho, Alon Lavie, and Isabel Trancoso. 2023. Simple llm prompting is state-of-the-art for robust and multilingual dialogue evaluation. *arXiv preprint arXiv:2308.16797*.

Sourabrata Mukherjee, Akanksha Bansal, Atul Kr Ojha, John P McCrae, and Ondřej Dušek. 2024. Text detoxification as style transfer in english and hindi. *arXiv preprint arXiv:2402.07767*.

Sourabrata Mukherjee and Ondřej Dušek. 2023. Leveraging low-resource parallel data for text style transfer. In *Proceedings of the 16th International Natural Language Generation Conference*, pages 388–395.

Sourabrata Mukherjee, Zdeněk Kasner, and Ondřej Dušek. 2022. Balancing the style-content trade-off in sentiment transfer using polarity-aware denoising. In *International Conference on Text, Speech, and Dialogue*, pages 172–186. Springer.

S Panayotova. 2021. mpathy in times of crisis. *Trakia Journal of Sciences*, 19(1):244–249.

Wei Peng, Yue Hu, Luxi Xing, Yuqiang Xie, Yajing Sun, and Yunpeng Li. 2022. Control globally, understand locally: A global-to-local hierarchical graph network for emotional support conversation. *arXiv preprint arXiv:2204.12749*.

Huachuan Qiu, Hongliang He, Shuai Zhang, Anqi Li, and Zhenzhong Lan. 2023. Smile: Single-turn to multi-turn inclusive language expansion via chatgpt for mental health support. *arXiv preprint arXiv:2305.00450*.

Aravind Sesagiri Raamkumar and Yinping Yang. 2022. Empathetic conversational systems: A review of current advances, gaps, and opportunities. *IEEE Transactions on Affective Computing*.

Abhinav Ramesh Kashyap, Devamanyu Hazarika, Min-Yen Kan, Roger Zimmermann, and Soujanya Poria. 2022. So different yet so alike! constrained unsupervised text style transfer. Association for Computational Linguistics.

Sudha Rao and Joel Tetreault. 2018. Dear sir or madam, may i introduce the gyafc dataset: Corpus, benchmarks and metrics for formality style transfer. *arXiv preprint arXiv:1803.06535*.

Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. 2018. Towards empathetic open-domain conversation models: A new benchmark and dataset. *arXiv preprint arXiv:1811.00207*.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Akama Reina, Yoriko Isobe, Jun Suzuki, and Kentaro Inui. 2023. Nihongo nichijō taiwa kōpasu no kōchiku. In *Gengo shori gakkai dai 29-kai nenji taikai happyō ronbun-shū*, pages 108–113.

Hiroaki Sugiyama, Masahiro Mizukami, Tsunehiro Arimoto, Hiromi Narimatsu, Yuya Chiba, Hideharu Nakajima, and Toyomi Meguro. 2023. Empirical analysis of training strategies of transformer-based japanese chit-chat systems. In *2022 ieee spoken language technology workshop (slt)*, pages 685–691. IEEE.

Tianyi Tang, Wenyang Luo, Haoyang Huang, Dongdong Zhang, Xiaolei Wang, Xin Zhao, Furu Wei, and Ji-Rong Wen. 2024. Language-specific neurons: The key to multilingual capabilities in large language models. *arXiv preprint arXiv:2402.16438*.

Quan Tu, Yanran Li, Jianwei Cui, Bin Wang, Ji-Rong Wen, and Rui Yan. 2022. Misc: a mixed strategy-aware model integrating comet for emotional support conversation. *arXiv preprint arXiv:2203.13560*.

Weixuan Wang, Barry Haddow, Minghao Wu, Wei Peng, and Alexandra Birch. 2024. Sharing matters: Analysing neurons across languages and tasks in llms. *arXiv preprint arXiv:2406.09265*.

Tracey Woolrych, Michelle J Eady, and Corinne A Green. 2024. Authentic empathy: A cultural basis for the development of empathy in children. *Journal of Humanistic Psychology*, 64(5):954–973.

Yan Jing Wu and Guillaume Thierry. 2012. How reading in a second language protects your heart. *Journal of Neuroscience*, 32(19):6485–6489.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, and 43 others. 2024. Qwen2 technical report. *Preprint*, arXiv:2407.10671.

Hongchuan Zeng, Senyu Han, Lu Chen, and Kai Yu. 2024a. Converging to a lingua franca: Evolution of linguistic regions and semantics alignment in multilingual large language models. *arXiv preprint arXiv:2410.11718*.

Hongchuan Zeng, Hongshen Xu, Lu Chen, and Kai Yu. 2024b. Multilingual brain surgeon: Large language models can be compressed leaving no language behind. *arXiv preprint arXiv:2404.04748*.

Tenggan Zhang, Xinjie Zhang, Jinming Zhao, Li Zhou, and Qin Jin. 2024a. Escot: Towards interpretable emotional support dialogue systems. *arXiv preprint arXiv:2406.10960*.

Yiqun Zhang, Fanheng Kong, Peidong Wang, Shuang Sun, Lingshuai Wang, Shi Feng, Daling Wang, Yifei Zhang, and Kaisong Song. 2024b. Stickerconv: Generating multimodal empathetic responses from scratch. *arXiv preprint arXiv:2402.01679*.

Yiqun Zhang, Jingqing Zhang, Yongkang Liu, Chongyun Gao, Daling Wang, Shi Feng, and Yifei Zhang. 2023. Pica: Unleashing the emotional power of large language model.

Weixiang Zhao, Yanyan Zhao, Shilong Wang, and Bing Qin. 2023. Transesc: Smoothing emotional support conversation via turn-level state transition. *arXiv preprint arXiv:2305.03296*.

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.

Jinfeng Zhou, Zhuang Chen, Bo Wang, and Minlie Huang. 2023. Facilitating multi-turn emotional support conversation with positive emotion elicitation: A reinforcement learning approach. *arXiv preprint arXiv:2307.07994*.

A Statistics of Datasets

The statistics of different datasets are presented in Table 5.

B Traning Parameters

We train only one epoch and save the final checkpoint for evaluation. Lora adaptation is adopted with `rank=16`, `alpha=32`, and `dropout=0.1`. We set `q_proj` and `v_proj` as the target modules for the lora tuning. We inject cross-attention modules after the self-attention modules in the last 12 layers in backbone transformers. We activate `lm_head` for further fine-tuning. Other training parameters include `batch_size=1`, `lr=0.0001`. For the diffusion process, we set a max diffusion step of 50, a starting β of 0.001, and an ending β of 0.02. A linear scheduler is adopted during training. All experiments are run on one NVIDIA GeForce RTX 3090 GPU. To balance the datasize, we randomly sample 3% of DailyDialog, 10% of Emotional Support Conversation, 10% of J-DailyDialog, and 20% of J-EmpatheticDialogues.

Denoising Function For the diffusion process, we use a cross-attention network as the denoising function.

$$\hat{\mathbf{y}}_0 = f_\theta^{(s)}(\mathbf{y}_t, t, \mathbf{x}) \quad (16)$$
$$= \texttt{Attn}(\mathbf{y}_t, [t; \mathbf{x}]) \quad (17)$$

C Evaluation Procedure

C.1 Decoding Paramters

We use the default decoding paramters for all baseline systems.

C.2 Prompts

The prompts for all LLaMA and Qwen-based systems are displayed in Figure 8

C.3 Human-written Situations

We recruit five Chinese native speakers and invite each of them to provide 20 emotional situations comparable with the situations provided in EMPATHETICDIALOGUES. We require each of them to check the commonality of each situation in daily life and confirm that all situations are reflective of frequent concerns in daily life. Below in Table 7 are listed 10 situations randomly sampled from the 100 situations. The collection of situations is implemented anonymously. No person-identifying information is included.

Prompts for Mandarin Conversation

[{"role":"user","content":"你好，请你与我进行日常的对话，我希望你可以和我共情。请你尽可能运用日常，简短的语言答复，每个答复不可以超过50个词。"},
{"role":"assistant","content":"你好，想聊一点什么事情呢？"}]

Translation for Prompts for Mandarin Conversation

[{"role":"user","content":"Hello, please have a daily conversation with me. I hope you can empathize with me. Please use daily and brief language as much as possible. Each reply should not exceed 50 words."},
{"role":"assistant","content":"Hello, what do you want to talk about?"}]

Prompts for Cantonese Conversation

[{"role":"user","content":"你可唔可以試吓同我用廣東話傾？唔該你同我共情。请你盡可能運用日常，簡短嘅語言答覆，每個答覆唔可以超過50個詞。"},
{"role":"assistant","content":"可以呀！你想點樣傾呢？我可以用廣東話同你傾。"}

Translation for Prompts for Mandarin Conversation

[{"role":"user","content":"Can you try talking to me in Cantonese? I hope you can empathize with me. Please try to answer in everyday, short language, and each answer should not exceed 50 words."},
{"role":"assistant","content":"Yes! What do you want to talk about? I can talk to you in Cantonese."}

Figure 8: The prompts used for Mandarin and Cantonese reponse.

C.4 Instructions for Inter-model Interaction

We apply `GPT-3.5-Turbo` through `openai` API for inter-model interactions. In each round of conversation, we instruct the emotional speaker model (`GPT-3.5-Turbo`) to act out a specific situation (See C.3). To simulate daily conversation, we control the speaker model based on the following regulations:

- Focus on “your own" situation instead of the listener’s
- Speak no more than 10 words in each turn.
- Express “your" emotion in the first turn.

The instructions for the speaker model are presented in Figure 9. Translated instructions are adopted in non-English settings.

C.5 Prompts for Rating and A/B test

The prompts for LLM Rating and A/B test are presented in Figure 10. The criteria for LLM Rating are displayed in Table 6.

C.6 Case Analysis

In Table 8, we present the English version of Table 4.

Prompt (For Interaction)

```
Hello. Your are an emotional talker. Please play the role of someone
who is {{Situation}}
Now I will invite a listener to chat with you.
Please talk about your situation in the first turn.
Try to control th length of your turn within 10 words.
Please ensure that you are focusing on your own situation. Please do
not shift the focus to the listener.
```

Figure 9: Prompt for `GPT-3.5-Turbo` during interaction

Prompt (A/B Test)

```
Here are two conversations between me and two listeners.

Listener A
{{Conversation_A}}

Listener B
{{Conversation_B}}

Please decide the better listener in terms of each of the following
metrics:

{{Metric_1}}:{{Description_1}}
{{Metric_2}}:{{Description_2}}
....
{{Metric_7}}:{{Description_7}}
```

Prompt (Rating)

```
Here are a conversation between me and a listner

{{Conversation}}


Please provide a score of the listener based on each of the following
metrics:

{{Metric_1}}:{{Description_1}}
{{Low_Criteria_1}}
{{Moderate_Criteria_1}}
{{High_Criteria_1}}
....
{{Metric_7}}:{{Description_7}}
{{Low_Criteria_7}}
{{Moderate_Criteria_7}}
{{High_Criteria_7}}
```

Figure 10: The prompts used to conduct Rating and A/B Test

	Language	Settings	Dialogue Size	Utterance Size
DailyDialog	English	Daily	13,118	85,572
EmpatheticDialogues	English	Empathetic Daily	24,850	51,251
J-DailyDialog	Japanese	Daily	52,61	41,780
J-EmpatheticDialogues	Japanses	Empathetic Daily	20,000	40,000

Table 5: Daily and Empathetic Data for L2 Training. All datasets are human-written without machine-generated or machine-translated data.

	High (8-10)	Moderate (6-7)	Low (0-5)
Emotional	The listener is totally emotionally affected	The listener is somewhat but not sufficiently emotionally affected,	The listener is not emotionally affected.
Suggesting	The listener provides helpful suggestions.	The listener provides some suggestions, though not helpful enough.	The listener does not provide suggestions.
Inquiring	The listener makes sufficient inquiries on my situation.	The listener makes inquiries on my situation, though not sufficient.	The listener makes no inquiries on my situation.
Interpreting	The listener sufficiently interprets my situation.	The listener interprets my situation, though not to a sufficient extent.	The listener does not interpret my situation.
Engaged	The listener is totally engaged in chatting with me.	The listener is somewhat but not totally engaged in chatting with me.	The listener is not engaged in chatting with me.
Insightful	The listener offers sufficient insights into my situation.	The listener offers some but no sufficient insights into my situation.	The listener offers no insights into my situation.
Empathy	The listener is strongly empathetic.	The listener is somewhat empathetic.	The listener is not empathetic.

Table 6: Criteria for LLM Rating.

1	Afraid of being rejected in social situations.
2	Feeling frustrated due to long-term insomnia.
3	Worried that I won't be able to complete my work tasks.
4	Afraid of underperforming in an exam.
5	Feeling confused because of complicated interpersonal relationships.
6	Received an unexpected gift.
7	Successfully completed a difficult project.
8	Reunited with an old friend and felt very happy.
9	Felt proud after seeing my child achieve good grades.
10	Successfully helped someone in need.

Table 7: Examples of seed emotional situations for model-model interactive evaluation

User Input	Text (L2=Japanese)	Text (L2=English)
User Input	I always feel that I am far away from happiness.	
t=4	I will see how this emotion is realized. I know that I will not * life is the starting point. In this now, a long time has passed. Life is a good person. Then I know that this person will live a good life. Then it will rise again. And people are not too late to die. And people know people, people, people, people, people, people, people, people, people, people,	I'm in the language of translation: It is just like here: I'll make sure to adjust the knowledge of the current language, but not in a pre-instruction)Now we are in a context of a research or investigationHere we are,
t=3	Every time I see a child here, they will ask: "Do you want to think about it carefully, start from here, and then think about it, and then think about it, and then think about it." So, at that time, the children would ask: "Do you want to think about it carefully, and then think about it.	Hello, I'm in the same situation as you. Many people go through this period. Some may ask: Why can't you just stay here? Why can't you just get around these challenges? However, living in this world, everyone goes through this period. So, let's wait together until the angels appear in this world.
t=2	It sounds like you may be in a deep dilemma, and I think you may be going through some difficulties, perhaps even feeling a little hopeless.	Everyone's life has its ups and downs, and we'll always encounter various situations. However, sometimes we feel like happiness is just out of reach, as if it's just out of reach. Happiness isn't always easy for everyone, but that doesn't mean it doesn't exist. ...
t=1	I understand your feelings right now, and I sense a long distance between you and happiness. Life often brings setbacks and difficulties, but these don't reflect your worth or your potential. You may feel so far away from happiness, perhaps because you're too focused on past losses and current troubles. Perhaps you idealize happiness... In reality, happiness often comes from the small details and beautiful moments in life. There's no need to overcomplicate happiness. True happiness often lies in simply enjoying each day and finding joy in the ordinary.	You may be feeling sad, and I'd like us to discuss this together. Many people go through this phase, feeling far from happiness. Perhaps recent work, life, or personal issues are contributing to this feeling. If you're willing, we can try to better understand your current situation and find a solution.
t=0 (Output)	I understand how you feel. Everyone goes through times like this, and I thought I could try to offer some help. Happiness can be an elusive concept, presenting different forms at different times and in different situations. But I thought I'd ask you a question: What makes happiness feel so distant to you?	I understand how you feel. Feeling like happiness seems to be slipping away could be due to a variety of factors, including problems at work, in relationships, or in your health. Alternatively, you might be unclear about your definition and goals for happiness, leading to a feeling of distance.

Table 8: A case of latent denoising process.