

# Multilingual KokoroChat: A Multi-LLM Ensemble Translation Method for Creating a Multilingual Counseling Dialogue Dataset

Ryoma Suzuki, Zhiyang Qi, Michimasa Inaba

The University of Electro-Communications  
1-5-1, Chofugaoka, Chofu, Tokyo, Japan  
{r-suzuki, qizhiyang, m-inaba}@uec.ac.jp

## Abstract

To address the critical scarcity of high-quality, publicly available counseling dialogue datasets, we created Multilingual KokoroChat by translating KokoroChat, a large-scale manually authored Japanese counseling corpus, into both English and Chinese. A key challenge in this process is that the optimal model for translation varies by input, making it impossible for any single model to consistently guarantee the highest quality. In a sensitive domain like counseling, where the highest possible translation fidelity is essential, relying on a single LLM is therefore insufficient. To overcome this challenge, we developed and employed a novel multi-LLM ensemble method. Our approach first generates diverse hypotheses from multiple distinct LLMs. A single LLM then produces a high-quality translation based on an analysis of the respective strengths and weaknesses of all presented hypotheses. The quality of “Multilingual KokoroChat” was rigorously validated through human preference studies. These evaluations confirmed that the translations produced by our ensemble method were preferred from any individual state-of-the-art LLM. This strong preference confirms the superior quality of our method’s outputs. The Multilingual KokoroChat is available at <https://github.com/UEC-InabaLab/MultilingualKokoroChat>.

**Keywords:** LLM Ensemble, Counseling Dialogue, Multilingual Dataset

## 1. Introduction

The COVID-19 pandemic has exacerbated the global mental health crisis, worsening numerous factors that contribute to psychological distress (Santomauro et al., 2021). Despite the growing need, access to professional mental healthcare remains a significant challenge for many, largely due to a shortage of skilled counselors. To bridge this gap, mental health support systems leveraging Large Language Models (LLMs) have emerged as a promising solution, showing great potential in alleviating distress and helping individuals navigate personal challenges (Sharma et al., 2023). However, the development of such systems is severely hampered by the scarcity of publicly available, high-quality, human-authored counseling dialogue datasets. To address this limitation, our work aims to extend the large-scale Japanese counseling dialogue corpus, KokoroChat (Qi et al., 2025), into a multilingual resource through translation.

For this multilingual extension, machine translation using LLMs is a powerful tool (Enis and Hopkins, 2024; He et al., 2024; Xu et al.). However, the optimal LLM for generating the best translation is inconsistent across different inputs, as models exhibit varying strengths in handling specific contexts and expressions (Jiang et al., 2023; Wang et al., 2024). Consequently, even a single state-of-the-art model cannot always guarantee the highest translation quality. This issue becomes particularly critical in sensitive domains like counseling, where superior quality and stability are paramount to avoid the

risk of misleading translations. To overcome this challenge, we propose a novel multi-LLM ensemble method. Our approach feeds the source text, along with translations from multiple LLMs, into a final LLM that synthesizes their strengths and mitigates their weaknesses to generate a more refined output. Using this method, we have developed the first multilingual psychological counseling dataset, named **Multilingual KokoroChat**, which provides a new foundational resource for cross-lingual counseling research and the development of multilingual dialogue systems.

## 2. Related Work

### 2.1. Psychological Counseling Datasets

In recent years, psychological counseling has attracted increasing attention in the field of Natural Language Processing (NLP). Early studies mainly focused on empathic response generation, where systems aim to produce emotionally appropriate responses by recognizing the user’s affective state (Rashkin et al., 2019; Sharma et al., 2020; Zheng et al., 2021). With the rapid advancement of LLMs, research interest has gradually shifted from simple empathy generation to more complex modeling of counseling dialogues. Consequently, a number of LLM-based counseling chat systems have been proposed, such as ChatCounselor (Liu et al., 2023), MeChat (Qiu et al., 2024), and SoulChat (Chen et al., 2023).

The performance of these systems largely de-

Dataset	Method	Open-source	Language	Dialogues	Avg. utterances
HealMe (Xiao et al., 2024)	LLM	✓	En	1,300	6.0
ESD-CoT (Zhang et al., 2024)	LLM	✓	En	1,708	23.4
CACTUS (Lee et al., 2024)	LLM	✓	En	31,577	31.5
SMILECHAT (Qiu et al., 2024)	LLM	✓	Zh	55,165	33.2
AUGESC (Zheng et al., 2023)	LLM	✓	En	65,000	26.7
<hr/>					
Anno-MI (Wu et al., 2022)	Human	✓	En	133	72.9
ESConv (Liu et al., 2021)	Human	✓	En	1,300	29.5
Client-Reactions (Li et al., 2023)	Human	✗	Zh	2,382	78.5
<b>KokoroChat</b> (Qi et al., 2025)	Human	✓	Ja	<b>6,589</b>	<b>91.2</b>
<hr/>					
PsyDial (Qiu and Lan, 2025)	Hybrid	✓	Zh	2,382	75.6
<b>Multilingual KokoroChat</b>	Hybrid	✓	<b>Zh/Ja</b>	<b>6,565</b>	<b>91.2</b>
			<b>En/Ja</b>	<b>6,582</b>	<b>91.2</b>

Table 1: Comparison of psychological counseling datasets categorized by construction method: LLM-based (top), human-collected (middle), and hybrid (bottom). We further build and release a multilingual version based on the high-quality human-collected dataset *KokoroChat*. Note that the slight reduction in the number of dialogues for the multilingual versions is due to the exclusion of content that triggered the LLM’s safety filters.

depends on the availability of high-quality counseling datasets, making data construction a central research issue. Existing efforts for building such datasets can be broadly categorized into two approaches: **human-created** and **LLM-augmented**.

Human-created datasets are typically collected and curated from real human dialogues. For instance, Anno-MI (Wu et al., 2022) was constructed from motivational interviewing conversations extracted from online videos; ESConv (Liu et al., 2021) collected emotional support dialogues from crowdworkers who had undergone professional skill training; ClientReactions (Li et al., 2023) was derived from real counselor–client interactions on online counseling platforms; and KokoroChat (Qi et al., 2025) was systematically collected through role-playing by trained counselors.

On the other hand, many studies adopt LLM-based automatic generation, where the model simultaneously plays both counselor and client to create data efficiently and at scale. Representative examples include HealMe (Xiao et al., 2024) and CACTUS (Lee et al., 2024), which generate multi-turn dialogues using prompts inspired by Cognitive Behavioral Therapy (CBT); ESD-CoT (Zhang et al., 2024), which generates full conversations based on scenarios extracted from existing datasets; SMILECHAT (Qiu et al., 2024), which expands single-turn Q&A pairs into multi-turn dialogues; and AUGESC (Zheng et al., 2023), which formulates data augmentation as a dialogue completion task. In addition, PsyDial (Qiu and Lan, 2025) adopts a **hybrid construction approach** that regenerates client utterances using LLMs based on ClientReactions, producing a semi-authentic dataset that

preserves realism while protecting privacy.

As shown in Table 1, although automatically generated data are advantageous in scalability and efficiency, they remain limited in authenticity—often with fewer dialogue turns—and in content diversity (Zheng et al., 2024). Therefore, this study focuses on human-collected dialogues and adopts KokoroChat (Qi et al., 2025), the largest and most professionally curated open counseling dataset, as the foundation.

## 2.2. LLM Ensemble

Ensemble learning, a method that combines multiple models to surpass the performance of any single model, has long been recognized as an effective technique in machine learning (Littlestone and Warmuth, 1994; Ganaie et al., 2022) and has been shown to improve system performance in machine translation (Zhou et al., 2017).

LLM ensembles initially emerged through selection-based approaches, such as the Pair-Ranker module in LLM-Blender (Jiang et al., 2023) and MBR (Minimum Bayes Risk) decoding (Fernandes et al., 2022; Farinhas et al., 2023), which choose the best output from multiple candidate outputs generated by multiple LLMs. However, this approach has a fundamental limitation: it cannot produce output quality that surpasses that of the best initial candidate. Furthermore, for our task of translating entire dialogues, selection-based approaches are inadequate. Because the dialogue is assembled from outputs of multiple models, terminology and stylistic tone can become inconsistent across turns.

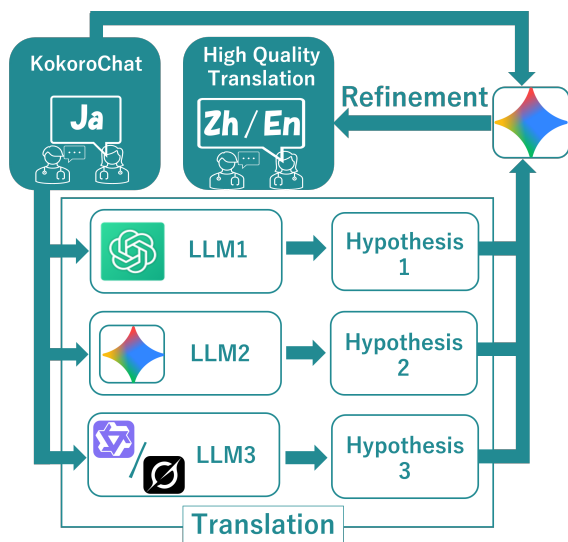


Figure 1: Proposed Multi-LLM Ensemble Translation Method

To overcome these limitations, research has progressed toward fusion-based approaches that integrate the strengths of multiple candidates to synthesize a new output superior to any individual candidate. For example, the GenFuser module in LLM-Blender fuses top-ranked candidates to generate a refined output (Jiang et al., 2023), and a layered Mixture-of-Agents (MoA), in which agents use prior-layer outputs as auxiliary context, enables hierarchical collaboration (Wang et al., 2024); both approaches consistently outperform single models. EVA (Ensemble via Vocabulary Alignment), which addresses lexical inconsistencies between different LLMs by fusing probability distributions at the token level, has demonstrated superior performance over selection-based ensemble methods like MBR and PairRanker, demonstrating the advantages of fusion-based approaches (Xu et al., 2024).

Furthermore, in Medical QA tasks, LLM-Synergy surpassed the performance of individual models by applying a boosting-based weighted majority vote (Yang et al., 2025). The vote aggregated outputs from LLMs that were dynamically selected based on question context through clustering. These successes in specialized domains strongly support applying a tailored ensemble method to counseling dialogue translation.

Our method is a fusion-based approach, but it is distinguished by its direct, analysis-driven refinement process. Specifically, the refiner LLM is instructed to analyze and articulate both the strengths and weaknesses of all presented translation candidates. Based on this generated analysis, the final translation is synthesized. Our method is distinguished from existing research by incorporating refiner LLMs’ advanced reasoning capabilities directly into the fusion process itself.

### # Data Description

- This is Japanese text counseling data from role-playing sessions where counselors acted as both counselor and client.
- Each line is separated by a colon (':'). The left side indicates the role name, and the right side is the utterance.

### # Translation Instructions

- As a professional translator, translate this data into **{target language}**.
- For the translation, please use polite and natural expressions that are appropriate for a counseling context.

Table 2: Prompt for Hypothesis Generation

## 3. KokoroChat

The foundation of our work is KokoroChat (Qi et al., 2025), a large-scale, publicly available collection of manually authored Japanese text-based counseling dialogues created through role-playing. The participants are professional counselors and trainee counsellors. The corpus comprises 6,589 dialogue sessions, each lasting 60 minutes, totalling approximately 600,000 utterances. The corpus covers a wide range of counseling topics, including issues related to school, work, family relationships, romance, and financial problems. A key feature of KokoroChat is that every dialogue is annotated with 20-item feedback scores provided by the participant playing the client role. These scores enable researchers to develop counseling dialogue systems using only high-quality data or automatic evaluation models for counselling dialogues. Given its status as the world’s largest manually curated counseling corpus and its unique integration of feedback scores, KokoroChat is a critical resource. We argue that extending this corpus into a multilingual format will provide significant value to the international research community.

## 4. Multi-LLM Ensemble Translation

To construct a high-quality multilingual counseling dialogue dataset, this study aims to address the challenge that the optimal LLM varies depending on the input, meaning no single model can consistently guarantee the best possible quality. Even high-performance models have distinct strengths and weaknesses, leading to inconsistent output quality. This instability is particularly unacceptable in sensitive domains like counseling, where the risk of mistranslation must be minimized.

To overcome this challenge, we propose a novel multi-LLM ensemble method, as shown in Figure 1. The approach consists of two key stages: generating diverse translation candidates (hypotheses)

---

### # Data Description

- This data includes Japanese text counseling data and its {target language} translation candidates.
- Japanese text counseling data was collected through role-playing between counselors acting as counselor and client.

### # Input Data Format

- The input is a list of dictionary objects.

### ## Keys of each dictionary

- 'role': Role of the speaker ('Counselor' or 'Client')
- 'source': Japanese original text
- 'hypothesis1', 'hypothesis2', 'hypothesis3': {target language} translation candidates

### # Evaluation Instructions

- You are a professional translator, evaluate the {target language} translation candidates.
  - For each utterance, follow these steps for evaluation:
    - 1. Analysis of Each Translation Candidate**
      - Compare each translation candidate and describe specifically which parts are superior.
      - Describe specifically which parts need improvement.
    - 2. Construction of an Improved Translation**
      - Based on your analysis, synthesize a revised translation by combining the strengths of both candidates.
      - Make corrections based on the areas for improvement you identified.
      - Ensure consistent terminology to maintain consistency throughout the translation.
- 

Table 3: Prompt for Integration and Refinement

and refining them through integration.

## 4.1. Diverse Hypothesis Generation

The objective of this stage is not merely to obtain multiple translations, but to construct a diverse and high-quality set of translation hypotheses. To achieve this, we instruct three distinct LLMs, using the system prompt detailed in Table 2, to generate three independent translation hypotheses per language.

Crucially, we process each entire dialogue as a single input rather than translating utterance by utterance. This approach is essential for preserving dialogue-level phenomena often lost in sentence-by-sentence translation. Providing the complete conversational history enables the LLMs to generate contextually coherent translations, maintain consistent terminology, and preserve the distinct linguistic styles of each speaker. The result is a translation that functions as a single, cohesive, and natural-flowing conversation, not merely a collection of disjointed sentences.

## 4.2. Integration and Refinement

In the second stage, a specialized “refiner” LLM is tasked with integration and refinement. The objective of this stage is not simply to select the best among the three candidates, but to create a new translation that surpasses any of the original hypotheses by synthesizing their respective strengths.

To achieve this, the refiner LLM is provided with both the original Japanese dialogue and the three translation hypotheses generated in the first stage. Crucially, these inputs are processed at the dialogue level to ensure the refiner LLM fully comprehends the overall context, including the flow of conversation and the consistency of speaker roles.

While maintaining this global context, the model is prompted to perform its analysis and synthesis on an utterance-by-utterance basis. Specifically, following the system prompt detailed in Table 3, we instruct the model to first analyze and articulate the strengths and weaknesses of each of the three candidate translations for a given utterance. Based on this analysis, it is then directed to synthesize a final, polished translation for that utterance, integrating the identified strengths while mitigating the noted shortcomings.

## 5. Experiments

To validate proposed multi-LLM ensemble method and assess the quality of the resulting Multilingual KokoroChat, we conducted experiments on translation into English and Chinese. We compared the output of our method with that of single LLMs using both automatic and human evaluation. The automatic evaluation provides a large-scale, objective measure of quality, while the human evaluation captures crucial nuances such as naturalness and contextual appropriateness, which automated metrics often fail to capture.

### 5.1. Experimental Settings

For the English translation, we selected three models representing the state-of-the-art at the time of manuscript preparation: GPT-5 (gpt-5-2025-0907) <sup>1</sup>, Gemini 2.5 Pro (gemini-2.5-pro) <sup>2</sup> and Grok-4 (grok-4-0709) <sup>3</sup>. For Chinese translation, we replaced Grok-4 with Qwen-Plus (qwen-plus-2025-07-28) <sup>4</sup>, which demonstrates superior performance and produces more stable outputs for Chinese-language tasks. This approach allows us

---

<sup>1</sup> <https://platform.openai.com/docs/models/gpt-5>

<sup>2</sup> <https://ai.google.dev/gemini-api/docs/models>

<sup>3</sup> <https://docs.x.ai/docs/models>

<sup>4</sup> <https://www.alibabacloud.com/help/ja/model-studio/what-is-qwen-llm>

System	MetricX↓	XCOMET↑
GPT	3.409*	0.921*
Gemini	3.356*	<b>0.927*</b>
Grok	3.369*	<b>0.927</b>
Proposed	<b>3.345</b>	0.926

Table 4: Automatic Evaluation Results for Japanese-to-English Translation. \*indicates a statistically significant performance difference  $p < 0.01$  compared to the proposed method.

System	MetricX↓	XCOMET↑
GPT	2.547*	0.859*
Gemini	2.483*	0.870*
Qwen	2.476*	<b>0.871</b>
Proposed	<b>2.445</b>	<b>0.871</b>

Table 5: Automatic Evaluation Results for Japanese-to-Chinese Translation. \*indicates a statistically significant performance difference  $p < 0.01$  compared to the proposed method.

to collect a set of hypotheses for each language that ensures both quality and diversity.

We selected Gemini 2.5 Pro as the refiner LLM, tasked with integrating and polishing the candidate translations. This decision was based on a preliminary analysis where Gemini 2.5 Pro demonstrated a superior ability over other models in comparing multiple texts, identifying specific areas for improvement, and synthesizing these insights into a cohesive and improved output. All outputs were generated using the default parameter settings for each respective model.

## 5.2. Automated Evaluation

The source KokoroChat dataset lacks human translations. Consequently, conventional reference-based metrics such as BLEURT (Sellam et al., 2020) or BARTScore (Yuan et al., 2021) were not applicable. Therefore, we employed two reference-free evaluation metrics: XCOMET-QE<sup>5</sup> (Guerreiro et al., 2024) and MetricX-QE<sup>6</sup> (Juraska et al., 2024). Both were chosen for their reported high correlation with human preferences in the WMT24 Metrics Shared Task (Freitag et al., 2024). XCOMET scores range from 0 to 1, where higher is better, while MetricX scores range from 0 to 25, where lower is better.

For this evaluation, we used 200 randomly sampled dialogues from KokoroChat. To remove the

<sup>5</sup><https://huggingface.co/Unbabel/XCOMET-XL>

<sup>6</sup><https://huggingface.co/google/metricx-24-hybrid-xl-v2p6>

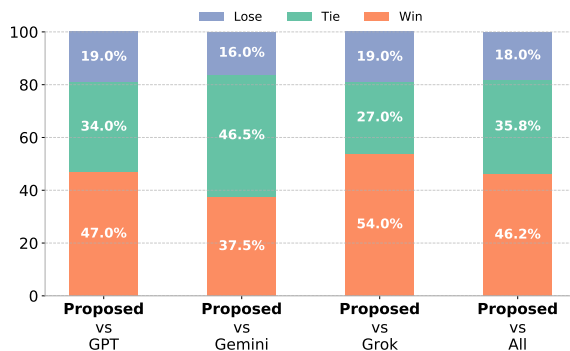


Figure 2: Human Evaluation Results for Japanese-to-English Translation

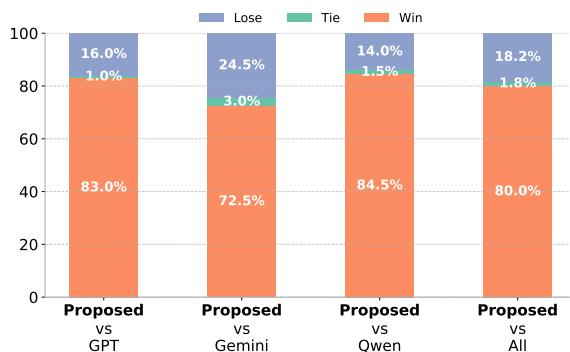


Figure 3: Human Evaluation Results for Japanese-to-Chinese Translation

trivial samples that would not yield meaningful differences in evaluation scores, we filtered this set to include only utterances for which at least three of the four systems (proposed method and the three single LLMs) generated unique translations. The final evaluation datasets consisted of 13,106 English utterances and 14,544 Chinese utterances.

## 5.3. Results of Automated Evaluation

The results are presented in Tables 4 and 5. To ensure statistical rigor, we performed a pairwise significance test using the Wilcoxon Signed-Rank Test (Wilcoxon, 1945) between our method and each single-LLM baseline. The Bonferroni correction (Bonferroni, 1936) was applied to counteract the increased risk of Type I errors from multiple comparisons; specifically, all p-values were multiplied by three.

Based on the MetricX, proposed method achieved the best (lowest) scores for both English and Chinese. For the Chinese translation, this superiority was statistically significant over all single-LLM baselines.

In the XCOMET metrics, our method tied with Qwen for the top performance in Chinese, achieving statistically significant improvements over both

Part of Dialogue History	
Grok (Presented to Evaluators)	Proposed (Not Presented to Evaluators)
<b>Counselor:</b> I see, so when you're talking with someone, you're concerned about whether you're getting attention there? But you wouldn't want others to notice that, right?	<b>Counselor:</b> I see. So when you're talking with someone, you're concerned about whether or not you're the focus of their attention? But you also don't want anyone to notice you feel that way, right?
<b>Client:</b> Yes. I don't want to show my <b>uncool</b> side.	<b>Client:</b> That's right. I don't want to show a <b>pathetic</b> side of myself.
Comparison	
Grok	Proposed
<b>Counselor:</b> Yes, you feel it's <b>uncool</b> if others notice that you're concerned things, right?	<b>Counselor:</b> Mm-hmm, so you feel that it would be <b>pathetic</b> if people noticed you were worried about such things, is that right?

Figure 4: Comparison of Japanese-to-English translation where the proposed method was judged inferior to Grok. This judge resulted from a contextual mismatch inherent in the randomized experimental design: the use of "pathetic" for internal consistency conflicted with the term "uncool" in the context randomly assigned to the evaluators.

GPT and Gemini. For English, while marginally behind Gemini and Grok, our method still demonstrated a statistically significant improvement over GPT.

Collectively, these results provide strong quantitative evidence that the translations generated by proposed method are consistently superior to, or at worst comparable with, those from single-LLM approaches.

#### 5.4. Human Evaluation

To assess qualitative aspects that automated metrics cannot capture, such as naturalness and contextual coherence, we conducted a human evaluation using a pairwise comparison framework. Evaluators were first presented with four turns of the preceding dialogue history. To prevent any particular system's translation style from biasing the context, the dialogue history for each evaluation was constructed using the output of a randomly selected system from the four candidates (proposed method and the three single LLMs). Each evaluator was then presented with a pair of translations: one from proposed method and one from a single LLM. Without access to the original Japanese text, evaluators were instructed to select the superior translation based on its contextual appropriateness and naturalness within the dialogue history.

The evaluation materials were drawn from 10 randomly selected dialogues from KokoroChat. From these, we sampled 100 unique utterances (10 per dialogue) where the translation from our method was distinct from those of all three single-LLM baselines. Each of these 100 utterances was then systematically paired against the three baseline translations, yielding a total of 300 evaluation pairs per language. This approach ensured that identical translations were never compared.

The assessments were performed by five native Chinese speakers for the Chinese translations. For the English translations, we recruited workers living

in English-speaking countries<sup>7</sup> through the Amazon Mechanical Turk (MTurk)<sup>8</sup>.

#### 5.5. Results of Human Evaluation

The results of the pairwise human evaluation, shown in Figures 2 and 3, were clear and strongly favored the proposed method. For both English and Chinese, the proposed method was strongly and consistently preferred by human evaluators over every single-LLM baseline. In every head-to-head comparison, the percentage of times the proposed method was judged superior ("Win") far surpassed the percentage of times it was judged inferior ("Lose"). This powerful result confirms the central hypothesis of the research: the process of integrating and refining outputs from multiple LLMs produces translations that are perceptibly and significantly more natural and contextually appropriate to human readers.

However, it is important to note that some instances where the proposed method was judged inferior were not due to a flaw in its refinement logic, but were a side effect of the randomized experimental design. Figure 4 illustrates an example where human evaluators judged the proposed method to be inferior to Grok. From the evaluator's perspective, the counselor appears to substitute the user's negative emotion ("uncool") with a more negative expression ("pathetic"). This apparent error occurs because, to ensure fairness, the dialogue history presented to evaluators was constructed from the output of a randomly selected system. This occasionally created a critical mismatch between the context provided to the evaluator and the internal context upon which our method relied.

<sup>7</sup>Specifically, the United States, Canada, the United Kingdom, Australia, and New Zealand.

<sup>8</sup><https://www.mturk.com/>

Source and Three Hypotheses		
(Source)	それはお辛いですね...わかる気がします。私も23歳で実家を離れましたが、居場所がないように感じていました。	That must be difficult... I can relate. I left my parents' home at 23, and I felt like I didn't have a place where I belonged.
(GPT)	↓ 咨询师: 那确实很辛苦.....我能理解。我在23岁离开老家时, 也觉得自己没有归属感。	
(Gemini)	咨询师: 那一定很难受吧.....我有点理解。我23岁时也离开了老家, 当时也感觉没有自己的容身之处。	
(Qwen)	咨询师: 这一定很痛苦.....我能理解。我23岁时也离开了家, 当时感觉没有归属感。	
Analysis of each Hypotheses and Overall Assessment		
	Original Output	Translated by Authors
(GPT)	全体的に良い翻訳です。「归属感」は「居場所」の訳として適切です。	A strong translation overall. The term "归属感" is an appropriate translation for the "居場所" (a place where one belongs).
(Gemini)	「那一定很难受吧」という共感の言葉が非常に良いです。「我有点理解」は「わかる気がします」の謙虚なニュアンスをよく捉えています。「容身之处」も「居場所」の良い訳です。	The empathetic phrase "那一定很难受吧" is a notable strength. The phrase "我有点理解" effectively captures the humble nuance of the original Japanese. "容身之处" is also an excellent translation.
(Qwen)	「这一定很痛苦」も良い共感の表現です。「归属感」も適切です。	"这一定很痛苦" is also an effective expression of empathy. The use of "归属感" is appropriate as well.
(Overall)	h2の出だし「那一定很难受吧」が最も共感的で、カウンセラーの言葉として適しています。また、「我有点理解」も原文のニュアンスに近いです。「居場所」の訳は「归属感」と「容身之处」のどちらも良いですが、後の会話との一貫性を考慮して選択する必要があります。ここでは「归属感」で統一します。	The opening from Hypothesis 2 (Gemini), "那一定很难受吧", is the most empathetic and suitable for a counselor's response. Its use of "我有点理解" also closely captures the original's nuance. While both "归属感" and "容身之处" are effective translations for the "居場所", consistency with the subsequent dialogue is a key consideration. Therefore, "归属感" is selected for the final synthesis to maintain uniformity.
Translation generated based on the Analysis		
(Proposed)	咨询师: 那一定很难受吧.....我能理解您的感受。我23岁离开家时, 也曾感觉没有归属感。	

Figure 5: Japanese-to-Chinese translation example 1. This demonstrates the analysis and synthesis process that produced a final translation highly preferred by human evaluators over any of the three single-LLM hypotheses.

## 5.6. Case Study

While the human evaluation results demonstrated the superiority of proposed method, some cases were consistently judged to be inferior. We analyze both successful and unsuccessful cases to explore avenues for future improvement.

Figure 5 illustrates a case where the proposed method functioned as intended, resulting in a translation that was highly preferred by evaluators. The refiner LLM correctly identified the opening of Gemini's translation as the most empathetic among the hypotheses. As noted in the model's internal analysis, this phrase's use of a modal particle creates a gentle, validating tone, which is more suitable for a counselor than a direct, assertive statement. Furthermore, the final synthesized phrase is more natural and appropriately empathetic than the other candidates. This case demonstrates that the method can successfully analyze the outputs of single LLMs and synthesize a superior utterance suitable for a counselor's role based on that analysis.

Conversely, Figure 6 presents another instance

where the proposed method was judged inferior to a single LLM (Qwen) in Chinese translation. The refiner LLM correctly identifies that Qwen's translation deviates semantically from the source text, shifting the question's focus from "what" to "when." Despite this semantic deviation, Qwen's translation was perceived by some human evaluators as a more natural conversational phrase in Chinese. This suggests that the refiner LLM, in this instance, over-prioritized strict semantic fidelity to the source text at the expense of conversational naturalness in the target language. By focusing too heavily on preserving the original question's exact focus, it produced a translation that, while technically more faithful, was ultimately judged inferior to Qwen's more natural-sounding translation.

## 6. Discussion

An analysis of our experimental results reveals a discrepancy between the automated and human evaluations, as well as performance variations

Source and Three Hypotheses		
(Source)	そうでしたか。相談者さんの悪口を言っていると 思うようになったきっかけは、どのような事だっ たんでしょうか。	I understand. What was it that made you start thinking they were talking badly about you?
(GPT)	咨询师: 是这样啊。您觉得她们在说您坏话, 最初是因为什么契机呢?	
(Gemini)	咨询师: 这样啊。那让您觉得她们在说您坏话的契机是什么呢?	
(Qwen)	咨询师: 这样啊。您是从什么时候开始觉得她们在说您坏话的呢?	
Analysis of each Hypotheses and Overall Assessment		
	Original Output	Translated by Authors
(GPT)	"...最初是因为什么契机呢?", "契机" 一词完美 对应了原文的 "きっかけ"。	"...最初是因为什么契机呢?" — The term "契机" (trigger) corresponds perfectly to the original Japanese, "きっかけ".
(Gemini)	"...的契机是什么呢?", 同样使用了 "契机", 翻译非常准确。	"...的契机是什么呢?" — This translation is also highly accurate, using the term "契机"(trigger).
(Qwen)	"您是从什么时候开始觉得..." , 把提问的重点从 "きっかけ" (契机/what) 错误地理解成了 "いつから" (从何时起/when)。	"您是从什么时候开始觉得..." — This <b>incorrectly</b> interprets the focus of the question, shifting it from "the trigger" (what) to "since when" (when).
(Overall)	候选项1和2都准确地翻译了 "きっかけ"。候选项 1的句子结构更流畅一些。候选项3的翻译有误。	Both Candidate 1 (GPT) and Candidate 2 (Gemini) accurately translate "きっかけ". However, Candidate 1's sentence structure is more fluid. <b>Candidate 3 (Qwen) is incorrect.</b>
Translation generated based on the Analysis		
(Proposed)	咨询师: 是这样啊。让您觉得她们在说您坏话的契机是什么呢?	

Figure 6: Japanese-to-Chinese translation example 2. This demonstrates the analysis and synthesis process that produced a translation judged inferior by human evaluators to Qwen’s hypothesis due to being slightly unnatural as a conversational text. The refiner LLM rejected Qwen’s hypothesis for semantic deviation.

across languages. This section examines these findings to discuss the effectiveness of our method.

### 6.1. The Limitations of Reference-Free Automated Metrics

While the proposed method was overwhelmingly preferred in human evaluations, some single-LLM systems achieved comparable or superior scores in the automated evaluation, particularly with the English XCOMET metric. This divergence strongly suggests that current reference-free metrics, despite their sophistication, are not yet capable of capturing the full range of qualities that humans perceive as high-quality dialogue. Specifically, the automatic evaluation metrics used in this study evaluate each utterance in isolation. Consequently, they are ill-equipped to assess factors such as contextual coherence across utterances, appropriateness for speaker roles (counselor or client). Specific examples of these cases are provided in Appendix A.1. While human evaluation incorporates these aspects by presenting the dialogue history, the automatic approach neglects them entirely. This difference in consideration likely contributes to the divergence between automatic and human evaluation results.

### 6.2. The Trade-off Between Semantic Fidelity and Conversational Naturalness

The translation example in Figure 6 shows that discrepancies can arise between human evaluators’ criteria and the optimization objective of the refiner LLM. Human evaluators were not shown the source text and based their judgments solely on conversational fluency, so semantic fidelity to the source was not considered. This explains why our proposed method, which is more semantically faithful to the source, was judged inferior in this particular case.

Conversely, the refiner LLM is provided with the source and prompted to act as a “professional translator,” which implicitly biases it toward prioritizing semantic fidelity. However, the primary purpose of the Multilingual KokoroChat dataset is to serve as a high-quality training corpus for AI counselors in the target languages. For such applications, the dialogue quality in the target language is more important than strict semantic fidelity for our counseling dataset. The example in Figure 6 suggests that in some cases, sacrificing a degree of fidelity can produce translations that are substantially more natural. A key direction for future work is to develop methods that can better balance this trade-off.

### 6.3. Language-Specific Effects

The results from both automated and human evaluations indicated that the performance uplift from our ensemble method was more pronounced for Chinese than for English. We hypothesize that this is because the baseline performance of the single LLMs for Japanese-to-English translation is already exceptionally high, leaving less room for improvement. In contrast, for Japanese-to-Chinese, where single-model performance may be slightly less mature, the ensemble’s ability to combine strengths and mitigate weaknesses provides a more significant benefit.

This suggests a promising direction for future research: the Multi-LLM Ensemble method may prove to be even more impactful when applied to low-resource language pairs, where the performance of any single translation model is often far from perfect.

## 7. Conclusion

This study introduced a novel, two-stage refinement framework to address the challenge that the optimal LLM varies depending on the input, meaning no single model can consistently guarantee the best possible quality. Our method first generates a diverse set of translation hypotheses using multiple distinct LLMs. Subsequently, a single refiner LLM analyzes the respective strengths and weaknesses of these candidates to synthesize a final, superior translation. By applying this framework to KokoroChat, the world’s largest Japanese counseling dialogue corpus, we have created Multilingual KokoroChat, a high-quality, large-scale dataset for English and Chinese.

Our approach’s effectiveness was confirmed through a thorough two-part evaluation. First, automated metrics provided strong quantitative validation, demonstrating that our method was statistically superior or highly competitive. Second, the pairwise human evaluation produced decisive results. Human evaluators overwhelmingly preferred the translations generated by our ensemble method over those from any single LLM, citing their superior naturalness and contextual coherence.

The resulting Multilingual KokoroChat dataset, which will be made publicly available, is expected to become a valuable resource for the global research community, accelerating advancements in AI-driven mental health support.

## 8. Acknowledgments

I would like to thank Associate Professor Michimasa Inaba of the Department of Informatics, Graduate School of Informatics and Engineering, The Univer-

sity of Electro-Communications, for his continuous guidance and support throughout this research. I am also deeply grateful to all the members of the Inaba Laboratory for their thoughtful suggestions and generous cooperation during the progression of this study. This work was supported by JSPS KAKENHI Grant Number 25H01382.

## 9. Bibliographical References

- Carlo E. Bonferroni. 1936. *Teoria statistica delle classi e calcolo delle probabilità*. Pubblicazioni del R. Istituto superiore di scienze economiche e commerciali di Firenze. Seeber.
- Maxim Enis and Mark Hopkins. 2024. [From LLM to NMT: Advancing Low-Resource Machine Translation with Claude](#).
- António Farinhas, José de Souza, and Andre Martins. 2023. [An empirical study of translation hypothesis ensembling with large language models](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 11956–11970. Association for Computational Linguistics.
- Patrick Fernandes, António Farinhas, Ricardo Rei, José G. C. de Souza, Perez Ogayo, Graham Neubig, and Andre Martins. 2022. [Quality-aware decoding for neural machine translation](#). In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1396–1412. Association for Computational Linguistics.
- Markus Freitag, Nitika Mathur, Daniel Deutsch, Chikiu Lo, Eleftherios Avramidis, Ricardo Rei, Brian Thompson, Frederic Blain, Tom Kocmi, Jiayi Wang, David Ifeoluwa Adelani, Marianna Buchichio, Chrysoula Zerva, and Alon Lavie. 2024. [Are LLMs breaking MT metrics? results of the WMT24 metrics shared task](#). In *Proceedings of the Ninth Conference on Machine Translation*, pages 47–81. the Association for Computational Linguistics.
- M.A. Ganaie, Minghui Hu, A.K. Malik, M. Tanveer, and P.N. Suganthan. 2022. [Ensemble deep learning: A review](#). *Engineering Applications of Artificial Intelligence*, 115:105151.
- Nuno M. Guerreiro, Ricardo Rei, Daan van Stigt, Luisa Coheur, Pierre Colombo, and André F. T. Martins. 2024. [xcomet: Transparent machine translation evaluation through fine-grained error detection](#). *Transactions of the Association for Computational Linguistics*, 12:979–995.

- Zhiwei He, Tian Liang, Wenxiang Jiao, Zhuosheng Zhang, Yujiu Yang, Rui Wang, Zhaopeng Tu, Shuming Shi, and Xing Wang. 2024. Exploring human-like translation strategy with large language models. *Transactions of the Association for Computational Linguistics*, 12:229–246.
- Dongfu Jiang, Xiang Ren, and Bill Yuchen Lin. 2023. [LLM-blender: Ensembling large language models with pairwise ranking and generative fusion](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics*, pages 14165–14178.
- Juraj Juraska, Daniel Deutsch, Mara Finkelstein, and Markus Freitag. 2024. [MetricX-24: The Google submission to the WMT 2024 metrics shared task](#). In *Proceedings of the Ninth Conference on Machine Translation*, pages 492–504. the Association for Computational Linguistics.
- N. Littlestone and M.K. Warmuth. 1994. [The weighted majority algorithm](#). *Information and Computation*, 108(2):212–261.
- June M. Liu, Donghao Li, He Cao, Tianhe Ren, Zeyi Liao, and Jiamin Wu. 2023. [Chatcounselor: A Large Language Models for Mental Health Support](#).
- 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.
- Damian F Santomauro, Ana M Mantilla Herrera, Jamileh Shadid, Peng Zheng, Charlie Ashbaugh, David M Pigott, Cristiana Abbafati, Christopher Adolph, Joanne O Amlag, Aleksandr Y Aravkin, et al. 2021. Global prevalence and burden of depressive and anxiety disorders in 204 countries and territories in 2020 due to the covid-19 pandemic. *The Lancet*, 398(10312):1700–1712.
- Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. [BLEURT: Learning robust metrics for text generation](#). In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892.
- Ashish Sharma, Inna W Lin, Adam S Miner, David C Atkins, and Tim Althoff. 2023. Human–AI collaboration enables more empathic conversations in text-based peer-to-peer mental health support. *Nature Machine Intelligence*, 5(1):46–57.
- Ashish Sharma, Adam Miner, David Atkins, and Tim Althoff. 2020. [A computational approach to understanding empathy expressed in text-based mental health support](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5263–5276, Online. Association for Computational Linguistics.
- Junlin Wang, Jue Wang, Ben Athiwaratkun, Ce Zhang, and James Zou. 2024. [Mixture-of-agents enhances large language model capabilities](#).
- Frank Wilcoxon. 1945. [Individual comparisons by ranking methods](#). *Biometrics Bulletin*, 1(6):80–83.
- Haoran Xu, Young Jin Kim, Amr Sharaf, and Hany Hassan Awadalla. A paradigm shift in machine translation: Boosting translation performance of large language models. In *The Twelfth International Conference on Learning Representations*.
- Yangyifan Xu, Jinliang Lu, and Jiajun Zhang. 2024. [Bridging the gap between different vocabularies for LLM ensemble](#). 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 7140–7152. Association for Computational Linguistics.
- Han Yang, Mingchen Li, Huixue Zhou, Yongkang Xiao, Qian Fang, Shuang Zhou, and Rui Zhang. 2025. [Large language model synergy for ensemble learning in medical question answering: Design and evaluation study](#). *J Med Internet Res*, 27:e70080.
- Weizhe Yuan, Graham Neubig, and Pengfei Liu. 2021. [BartScore: Evaluating generated text as text generation](#). In *Advances in Neural Information Processing Systems*, volume 34, pages 27263–27277. Curran Associates, Inc.
- Chujie Zheng, Yong Liu, Wei Chen, Yongcai Leng, and Minlie Huang. 2021. [CoMAE: A multi-factor hierarchical framework for empathetic response generation](#). In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 813–824, Online. 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.

Long Zhou, Wenpeng Hu, Jiajun Zhang, and Chengqing Zong. 2017. [Neural system combination for machine translation](#). In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 378–384. Association for Computational Linguistics.

## 10. Language Resource References

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, Singapore. Association for Computational Linguistics.

Suyeon Lee, Sunghwan Kim, Minju Kim, Dongjin Kang, Dongil Yang, Harim Kim, Minseok Kang, Dayi Jung, Min Hee Kim, Seungbeen Lee, Kyong-Mee Chung, Youngjae Yu, Dongha Lee, and Jinyoung Yeo. 2024. [Cactus: Towards psychological counseling conversations using cognitive behavioral theory](#). In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 14245–14274, Miami, Florida, USA. Association for Computational Linguistics.

Anqi Li, Lizhi Ma, Yaling Mei, Hongliang He, Shuai Zhang, Huachuan Qiu, and Zhenzhong Lan. 2023. [Understanding client reactions in online mental health counseling](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10358–10376, Toronto, Canada. 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.

Zhiyang Qi, Takumasa Kaneko, Keiko Takamizo, Mariko Ukiyo, and Michimasa Inaba. 2025. [KokoroChat: A Japanese psychological counseling dialogue dataset collected via role-playing by trained counselors](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*, pages 12424–12443.

Huachuan Qiu, Hongliang He, Shuai Zhang, Anqi Li, and Zhenzhong Lan. 2024. [SMILE: Single-turn to multi-turn inclusive language expansion via ChatGPT for mental health support](#). In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 615–636, Miami, Florida, USA. Association for Computational Linguistics.

Huachuan Qiu and Zhenzhong Lan. 2025. [PsyDial: A large-scale long-term conversational dataset for mental health support](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 21624–21655, Vienna, Austria. Association for Computational Linguistics.

Zixiu Wu, Simone Balloccu, Vivek Kumar, Rim Helaoui, Ehud Reiter, Diego Reforgiato Recupero, and Daniele Riboni. 2022. [Anno-mi: A dataset of expert-annotated counselling dialogues](#). In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6177–6181.

Mengxi Xiao, Qianqian Xie, Ziyang Kuang, Zhicheng Liu, Kailai Yang, Min Peng, Weiguang Han, and Jimin Huang. 2024. [HealMe: Harnessing cognitive reframing in large language models for psychotherapy](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1707–1725, Bangkok, Thailand. Association for Computational Linguistics.

Tenggan Zhang, Xinjie Zhang, Jinming Zhao, Li Zhou, and Qin Jin. 2024. [ESCoT: Towards interpretable emotional support dialogue systems](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13395–13412, Bangkok, Thailand. Association for Computational Linguistics.

Chujie Zheng, Sahand Sabour, Jiaxin Wen, Zheng Zhang, and Minlie Huang. 2023. [AugESC: Dialogue augmentation with large language models for emotional support conversation](#). In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1552–1568, Toronto, Canada. Association for Computational Linguistics.

## A. Appendix

Section A.1 presents examples where translations that are contextually appropriate within a dialogue were assigned low scores by automated metrics because they were evaluated at an utterance-by-utterance level.

Speaker	Original Utterance	Translated by Gemini
Counselor	夫さんがイライラされているんですね お困りですね	I see your husband is feeling irritable. That must be difficult for you.
Client	困っています (I am having a hard time.)	It is.

Figure 7: A case where contextually natural translation is penalized by utterance-level evaluation. The translation "It is" captures the conversational flow but is judged as semantically incomplete by reference-free metrics that do not account for the preceding context.

Speaker	Original Utterance	Translated by GPT
Client	いろいろ勉強して資格を取ろうかと思っています	I'm thinking of studying to earn a certification,
Client	その時間が無くなった (That time has run out.)	and I felt like
Client	感じでした (It was a feeling.)	I lost that time.

Figure 8: A case where contextually natural translation is penalized by utterance-level evaluation. While the translation elegantly resolves a sentence split, utterance-by-utterance evaluation incorrectly penalizes them.

### A.1. Discrepancy between dialogue-level naturalness and utterance-level automatic evaluation.

Figure 7 and Figure 8 illustrate a limitation of reference-free automated metrics that evaluate utterances in isolation. In the dialogue shown in Figure 7, the Client responds to the Counselor's empathetic remark, "That must be difficult for you." While the original Japanese utterance literally translates to "I am having a hard time," Gemini translates this as "It is." This is a highly natural and contextually appropriate English expression. However, automated metrics evaluating translations on an utterance-by-utterance basis penalize this output for the lack of explicit semantic overlap with the source text. Similarly, the dialogue shown in Figure 8 demonstrates how speakers naturally divided their utterances into three parts to maintain proper pacing in this text-based counseling format. GPT faithfully reproduces this speaker's characteristic utterance segmentation by dividing the entire dialogue into three parts while maintaining the natural speech flow of the conversation. However, the second utterance and the third utterance were penalized by automated metrics evaluating in isolation because the translation's meanings no longer matched the individual original Japanese utterance.

These cases highlight a fundamental discrepancy between dialogue-level naturalness and utterance-level automatic evaluation. This discrepancy contributes to the gap between automated evaluations and human evaluation results.