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

The University of Electro-Communications

Ryoma Suzuki, Zhiyang Qi, Michimasa Inaba

Inaba Lab. The University of Electro-Communications

LREC 2026 Palma

Background

- **Mental health is a global challenge.**
  - The COVID-19 pandemic has worsened the global mental health crisis [Santomauro+ Lancet21]
- **Growing demand for mental health**
  - Shortage of skilled counselors

  → **Solution: Mental health support systems**
- **Challenges in Developing Mental Health Support Systems**
  - **Essentials: High-quality, human-authored counseling dialogue datasets**
  - **Challenges**: **Scarcity** of such **publicly available** datasets
  - Manual data collection requires significant time and costs
- **Objective of this Study**
  - Extend **KokoroChat** [Qi+ ACL25] into a multilingual dataset in **English** and **Chinese**

KokoroChat

- **A large-scale Japanese counseling dialogue dataset** collected via role-playing by trained counselors
- **6,589 text-based dialogues (average 91.2 utterances)**

  → **High Quality | Large Scale | Long Context**

Our Dataset: Multilingual KokoroChat

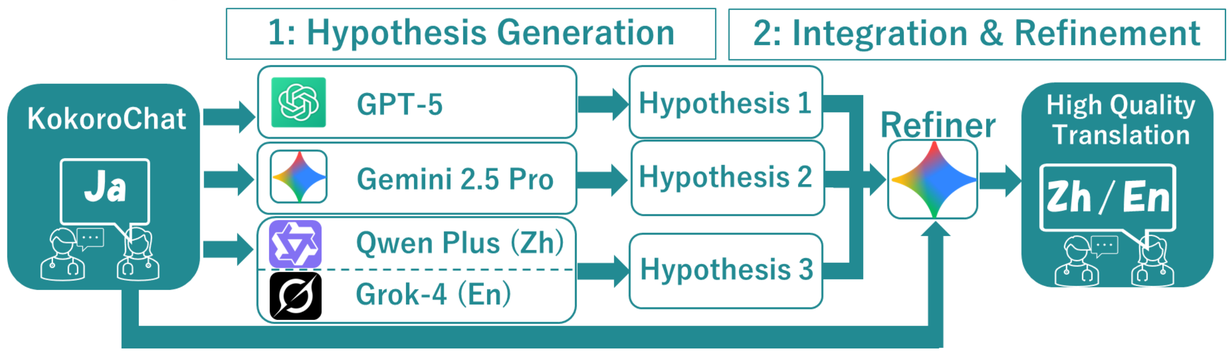
- **High-quality English and Chinese** translations of **KokoroChat**
- **Employed a novel Multi-LLM Ensemble Translation Method**

GitHub

Multi-LLM Ensemble Translation

- Powerful tool in machine translation**: LLM**
- **Challenge in translation using LLMs**
  - The optimal model for translation varies by input [Jiang+ ACL23, Wang+ ICLR25]
- **Counseling: Sensitive domain**
  - The highest possible translation fidelity is essential
- Relying on a single LLM is **insufficient**

→ **Developed a novel Multi-LLM Ensemble Translation Method**



- **STEP1: Hypothesis Generation**
  - Generate **three distinct translation hypotheses** per language using **different LLMs**
  - Process **the entire dialogue** at once to preserve dialogue-level phenomena
- **STEP2: Integration and Refinement**
  - Feed the source text and three hypotheses into a refiner LLM
  - **Analyze** strengths & weaknesses of each hypothesis using dialogue-level context
  - Generate a **high-quality** translation utterance-by-utterance by integrating the strengths of each hypothesis

Example of the Translation Process

Japanese Source Translated by Authors

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.

Hypothesis Generation

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**。

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**。

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 (Translated from Japanese by authors)

A strong translation overall. The term "**归属感**" is an appropriate translation for the "居場所" (a place where one belongs).

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.

"**这一定很痛苦**" is also an effective expression of empathy. The use of "**归属感**" is appropriate as well.

Overall Assessment (Translated from Japanese by authors)

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.

Final Refined Translation

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。

Automated Evaluation

- **Evaluation Data**
  - Randomly sampled **200 dialogues** from KokoroChat
  - Evaluation utterance size (**En: 13,106, Zh: 14,544**)
- **KokoroChat has no human translation**
  - Employed two reference-free evaluation metrics
  - **Correlate highly** with human preference even without reference translations [Freitag+ WMT24]
- **Reference-Free Evaluation Metrics**
  - **XCOMET-QE** [Guerreiro+ TACL24]
    - **Score: 0～1**
    - **Higher** is better
  - **MetricX-QE** [Juraska+ WMT24]
    - **Score: 0～25**
    - **Lower** is better
- Scores are the average rating per utterance
  - **Cannot** assess dialogue-level naturalness
- **Proposed Method Achieved:**
  - **Best MetricX** in **En** & **Zh**
  - **Best XCOMET** in **Zh**
  - Slightly **lower XCOMET** than Gemini & Grok in **En**

English Translation Score

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

Chinese Translation Score

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

Human Evaluation

- **Evaluation Data**
  - Randomly sampled **10 dialogues** from KokoroChat
  - 10 utterances per dialogue (Total: **100 utterances**)
  - With four turns of the preceding dialogue history
- **Participants**
  - **English:** Workers living in English-speaking countries recruited via Amazon Mechanical Turk
  - **Chinese**: **5** native speakers
- **Criteria**
  - **Contextual coherence, Dialogue-level naturalness**
  - Without Japanese source → **Faithfulness not evaluated**
- **Proposed Method Achieved:**
  - Strongly **preferred** in both languages
  - **Stronger** preference in **Chinese** than in **English**