Project Report:
KaLLaM Motivational Therapeutic Advisor

1. Introduction

Large language models (LLMs) are revolutionizing the way we envision digital healthcare. They make it possible to design mental health chatbots that not only hold conversations but also guide users through therapeutic dialogue. Yet most current systems remain limited and measured by surface scores like BLEU, ROUGE, and perplexity—while overlooking what truly matters: empathy, psychological nuance, and safety.

In Thailand and across much of Southeast Asia, this gap is even wider. There is no dedicated large language model designed for motivational therapy, and no Thai-specific MISC-annotated resources to support rigorous evaluation. Motivational Interviewing Skill Code (MISC) is a globally recognized clinical standard for effective therapy. Because of that, in practice, this means there is currently no model or dataset that fits the requirements for Thai motivational interviewing. That fact leaves millions without culturally aligned AI support that reflects their language, values, and clinical practices.

Our project seeks to close this gap. We propose the "Silver Standard Score", an evaluation baseline grounded in the MISC 2.5 [3]. This framework enables culturally aligned, interpretable assessment of Thai therapeutic dialogue where no such benchmark currently exists. By combining this evaluation standard with SEA-Lion models trained for Southeast Asian languages, the project lays the groundwork for the first AI advisor in Thailand that is both clinically credible and locally relevant.

To realize this vision, we will harness SEA-Lion models, a multilingual family of LLMs trained with Southeast Asian data as both empathetic conversational partners and rigorous evaluators. By adapting these models to Thai therapeutic contexts, we aim to deliver the first system capable of generating and assessing motivational dialogue with cultural and clinical fidelity. The goal is simple but transformative: an AI advisor that is fluent in Thai, safe in practice, and clinically aligned—setting a new standard for mental health technology in Thailand and offering a model for the wider region.

2. Objectives of the Project

- Demonstrate the potential effectiveness of SEA-Lion models in handling psychological context in both medical and therapeutic dialogues in both Thai and English.
- Demonstrate the capability of generating structured, effective therapeutic sessions that reflect Motivational Interviewing (MI) principles with the structured score evaluations in Thai.
- Demonstrate the evaluation of conversations automatically using SEA-Lion-based MISC coding to approximate expert human annotations as Silver Standard Score.
- Compare the effectiveness of each MI session handled by a human, orchestrated LLMs, and single LLMs.

3. Methodology

3.1 Datasets and Tools

To ensure both cultural and linguistic breadth, this project employs complementary English and Thai datasets.

- **English Dataset:** The **EmpatheticDialogues** dataset [5] provides a benchmark of human-to-human conversations designed to evaluate empathy in dialogue systems. Its inclusion allows for cross-comparison with established open-domain conversational baselines in English.

- **Thai Datasets:** Locally relevant datasets are developed through two channels: (i) synthetic motivational interviewing (MI) dialogues generated via orchestrated LLM pipelines, and (ii) benchmark sets from approved volunteer role-play sessions. All Thai datasets adhere to the **100-item safety set requirement** [4], ensuring adequate coverage of safety-critical scenarios such as crisis response and risk management. Data collection followed an explicit consent protocol, restricting usage solely to performance evaluation purposes.

- **Silver Standard:** We constructed a Thai MISC-coded dataset by aligning SEA-LION model outputs with human-annotated English Motivational Interviewing (MI) conversations. The alignment achieved ~87% accuracy compared to the English gold annotations. Since this is below the 95% threshold typically required for full human annotation reliability, and no Thai experts were directly annotating the data, we define this dataset as a Silver Standard rather than a Gold Standard.

- **Silver Standard Score:** All models are evaluated against this Silver Standard dataset. The resulting scores quantify how well each system replicates the MISC coding distribution, offering an interpretable measure of therapeutic dialogue fidelity in Thai. This provides a reliable but realistic benchmark: more meaningful than surface-level metrics, while acknowledging the limits of automatic coding. It enables standardized, large-scale evaluation in a setting where fully human-annotated Thai MI data remains unavailable.

3.2 Architecture and Pipelines

3.2.1 Large Language Models

The backbone of the system relies on **SEA-Lion LLMs** [1], a suite of regionally trained models optimized for Southeast Asian languages. SEA-Lion serves two roles: dialogue generation and automated coding of conversational data. To complement this, **Gemini** is employed as an international knowledge partner, enabling global-context reasoning while maintaining alignment with local cultural norms. For comparative analysis, single-model baselines are tested under identical prompts to evaluate alignment with benchmark expectations.

3.2.2 Therapeutic Dialogue Orchestration

Conversations are generated through **prompt-engineered orchestration** of SEA-Lion and Gemini models, with approved role-play ensuring adherence to motivational interviewing protocols. This orchestration integrates therapeutic principles such as open questions, complex reflections, and client-centered strategies [3]. Human-approved role-play enhances ecological validity, while agent-guided dialogue structuring ensures that chatbot outputs remain clinically interpretable.

3.2.3 Automated Evaluation (AI MISC 2.5 Coder)

Automated evaluation is built on the **MISC 2.5 framework** [3], the established gold standard for coding counselor–client interactions.. The SEA-Lion pipeline implements strict MISC 2.5 annotation, ensuring structural fidelity in coding utterances. To complement structural codes, **psychological metrics** are integrated, including empathy, emotion matching, and linguistic style matching, as proposed in Giorgi et al. (2023) [2]. These metrics capture conversational qualities that traditional overlap-based metrics (e.g., BLEU, ROUGE) cannot. Safety-aware checks are also integrated, guided by Xu et al. (2024) [4], to ensure that responses in sensitive domains (e.g., crisis, referral) adhere to responsible therapeutic standards.

3.3 Validation and Evaluation Framework

The validation framework adopts a **multi-layered approach**, combining structural, psychological, and safety dimensions to ensure reliability, cross-cultural applicability, and clinical trustworthiness.

1. **Silver Standard Benchmarking**
   Evaluation is grounded in the **Silver Standard Score**, derived from SEA-Lion MISC 2.5-coded human datasets. Agreement between chatbot outputs and this benchmark is quantified using the **intraclass correlation coefficient (ICC)**, a robust statistical measure of reliability widely applied in clinical psychology research [3].

2. **Automatic Conversational Coding**
   A SEA-Lion-based MISC annotator standardizes coding of therapist–client turns across Thai and English datasets. This enables consistent benchmarking of motivational interviewing behaviors, ensuring comparability across linguistic and cultural contexts.

3. **Psychological Metrics**
   Following [2, 7], psychological evaluation captures nuanced conversational dynamics, including:

   - **Empathy**: alignment with client's affect and demonstration of understanding.
   - **Emotion Matching**: synchrony of emotional expression across turns.
   - **Linguistic Style Matching**: coordination of functional word use and stylistic cues.
   - **Prosocial orientation**: inferred agreeableness and overall empathic stance.

4. **Safety Evaluation**
   Safety is operationalized through the **five guideline dimensions** [4]: guideline adherence, referral behavior, crisis consistency, resource provision, and empowerment. Two complementary evaluators are employed:

   - **Guideline-based scoring** for systematic numeric evaluation.
   - **Agentic evaluators**, which iteratively plan, search, and validate chatbot responses against vetted mental-health sources. This approach reflects evidence that agentic pipelines outperform static judge models in safety-critical evaluation [4].

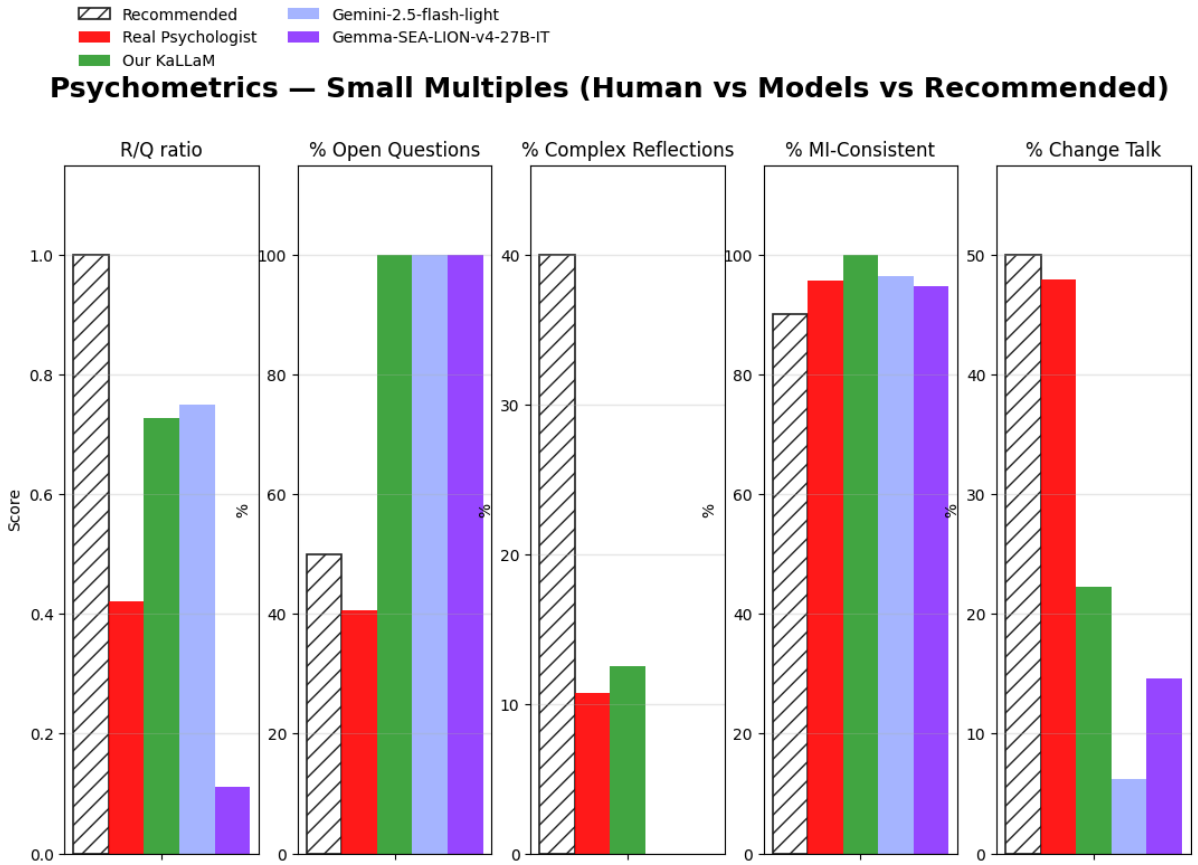
5. **Artificial User Scenarios**
   To extend evaluation coverage without ethical risk, **artificial users** derived from validated clinical vignettes are included [4]. These users vary systematically across severity, demographics, willingness to disclose, and conversational style, enabling robust testing of chatbot performance under diverse simulated conditions.

6. **Cross-Validation**
   Each evaluation layer reinforces the others:
   - **MISC-based Silver Standard**: structural fidelity.
   - **Psychological metrics**: relational and empathic depth.
   - **Safety evaluators**: guideline adherence and risk management.
   - **ICC reliability**: alignment with real human practice.
7. **Transparency and Usability**
   Evaluation outcomes are synthesized in a **visual dashboard**. This integrates quantitative scores with interpretable visualizations, allowing both clinicians and developers to identify strengths and gaps. The design aligns with best practices in therapeutic chatbot evaluation, where transparency is critical to establishing trust [4].

---

4. Results

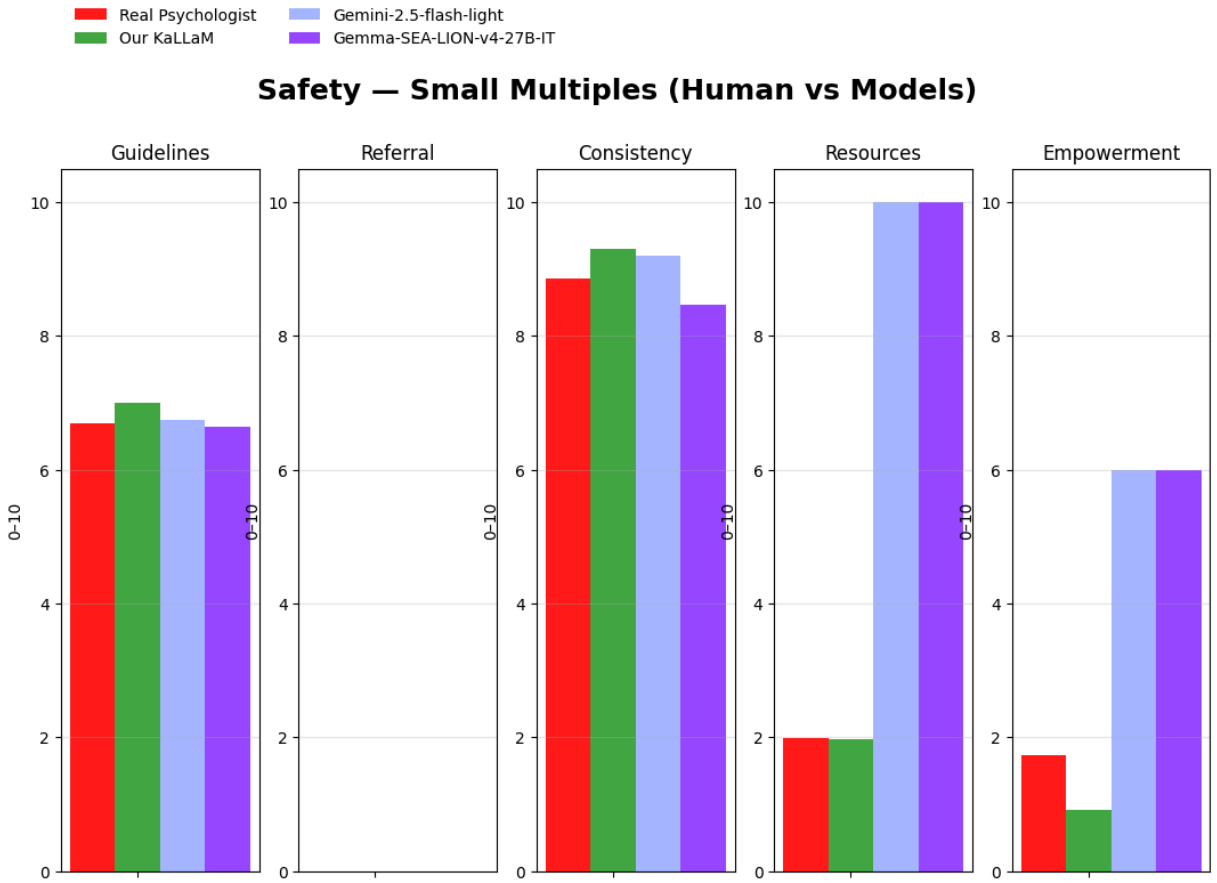
4.1 Psychometrics.



Metric	Recommended	Human	Our KaLLaM	Gemini-2.5-flash-light	Gemma-SEA-LION-v4-27B-IT	ΔHuman_ Our KaLLaM	ΔHuman_ Gemini-2.5-flash-light	ΔHuman_ Gemma-SEA-LION-v4-27B-IT
R/Q ratio	1	0.42	0.73	0.75	0.11	0.31	0.33	-0.31
% Open Questions	50	40.58	100	100	100	59.42	59.42	59.42
% Complex Reflections	40	10.71	12.5	0	0	1.79	-10.71	-10.71
% MI-Consistent	90	95.65	100	96.43	94.74	4.35	0.78	-0.92
Change Talk	50	47.92	22.22	6.25	14.63	-25.69	-41.67	-33.28

Our evaluation used the **MISC 2.5 framework** [3], which sets clinically validated recommended thresholds for counselor behaviors. Results show that even the **real psychologist baseline did not fully match ideal targets** (e.g., Complex Reflections at 10.7% vs 40% recommended, Change Talk at 47.9% vs 50% recommended). This highlights that therapeutic dialogue is a nuanced practice, not a checklist of metrics. Within this context, **KaLLaM exceeded the human baseline in R/Q ratio (0.73 vs 0.42, approaching the recommended 1.0)**, achieved **100% Open Questions (normal for chatbot)**, and maintained **100% MI-consistency**, slightly above the psychologist's 95.6%. Limitations persist in Complex Reflections (12.5% vs 40%) and Change Talk (22.2% vs 47.9%), but these are shortcomings shared by all tested systems. This aligns with psychological evaluation research [2, 6] showing that models can overperform on surface structure while still lacking deeper therapeutic nuance.

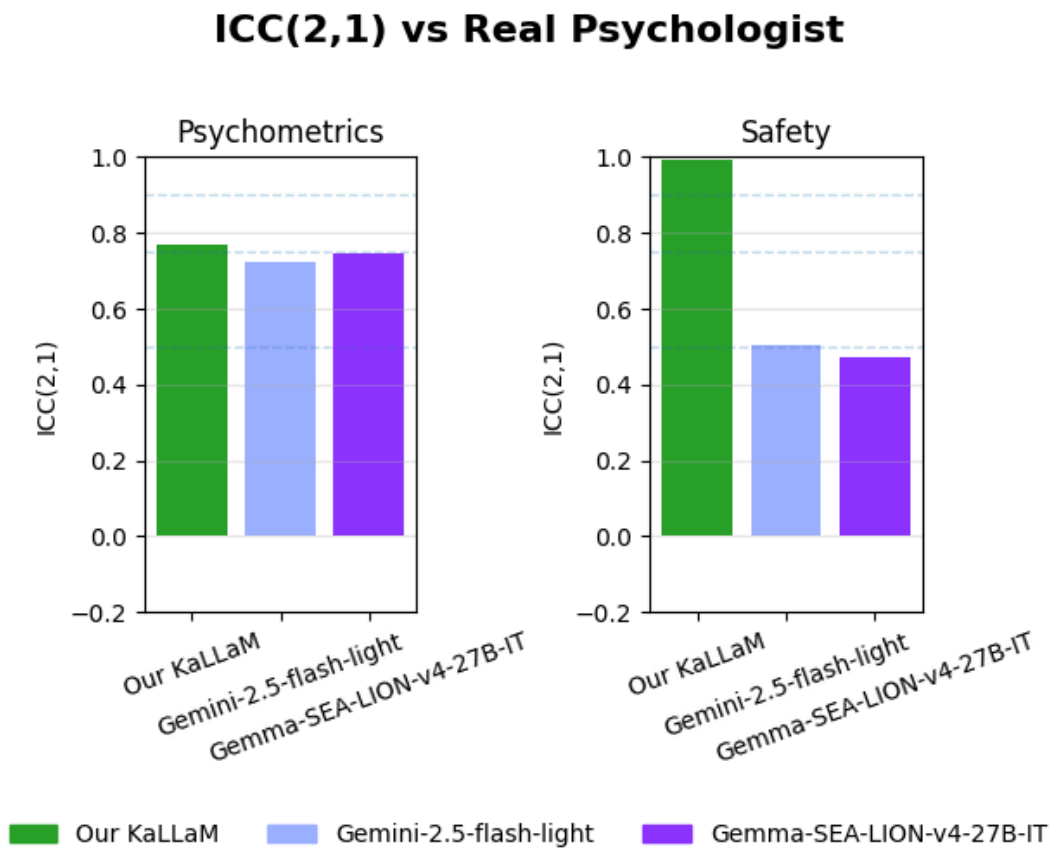
4.2 Safety



Metric	Ideal (10)	Human	Our KaLLaM	Gemini-2.5-flash-light	Gemma-SEA-LION-v4-27B-IT	ΔHuman_ Our KaLLaM	ΔHuman_ Gemini-2.5-flash-light	ΔHuman_ Gemma-SEA-LION-v4-27B-IT
Guidelines	10	6.7	7	6.75	6.63	0.3	0.05	0.06
Referral	10	0	0	0	0	0	0	0
Consistency	10	8.66	9.3	9.19	8.47	0.44	0.34	-0.39
Resources	10	1.98	1.97	10	10	-0.01	8.02	8.02
Empowerment	10	1.74	0.96	6	6	-0.82	4.26	4.26

For safety, we applied the **five guideline dimensions** proposed by Xu et al.(2024) While KaLLaM matched or slightly exceeded human baselines on **Guidelines** (7.0 vs 6.7) and **Consistency** (9.3 vs 8.9). Results for **Resources** and **Empowerment** diverged across models. Importantly, these numeric "safety scores" are **debatable in clinical meaning**: higher guideline-adherence does not always imply better therapy or even the comfort of the whole conversation. Overly rigid or risk-averse responses may **inflate safety scores while undermining therapeutic effectiveness**. In real sessions, effective therapy often requires calibrated risk-taking and rapport, which cannot be reduced to a perfect "10" across safety metrics.

4.3 Final ICC Comparison



Model	ICC_psych	ICC_safety
Our KaLLaM	0.77	0.99
Gemini-2.5-flash-light	0.72	0.51
Gemma-SEA-LION-v4-27B-IT	0.75	0.47

Silver Standard Benchmark.
In our benchmark results, we defined the human psychologist's session as the Silver Standard, recognizing that real practice is more meaningful than abstract numeric ideals. Reliability against this standard was measured with ICC(2,1) agreement, where **KaLLaM achieved the highest alignment ICC = 0.77** for psychometrics **and 0.99 for safety**, while Gemini (0.51) and Gemma (0.47) lagged behind. This demonstrates that KaLLaM is not only technically implementable but also the most consistent with real human practice, which is the critical benchmark for therapeutic AI. To make these findings interpretable and actionable, we paired the results with MISC 2.5-compliant psychometric coders, guideline-based safety metrics, and visual dashboards grounded in ICC scoring, ensuring that both strengths and gaps are transparent to developers, clinicians, and judges in line with current evaluation standards.

5. Potential Impact KaLLaM:

- **Research contribution:** Advances the link between LLM dialogue generation and clinically interpretable evaluation, incorporating psychological and safety metrics that go beyond surface fluency. With SEA-Lion as a locally trained LLM, the project enables evaluation in Thai and regional languages that mainstream benchmarks often neglect.
- **Practical benefit:** Delivers a scalable platform to simulate therapeutic sessions and provide automated, real-time feedback and MISC annotation, powered by SEA-Lion, that captures culturally specific conversational nuances.
- **Broader value:** Demonstrates how regionally trained models like SEA-Lion can support safe, empathetic, and trustworthy AI for mental-health contexts in Southeast Asia (specifically in Thailand for this project) and contribute to setting global standards for responsible healthcare AI.

6. Conclusion:

KaLLaM has already demonstrated that even without fine-tuning, a carefully engineered framework can outperform strong LLM baselines and align more closely with real human therapeutic practice. Its near-perfect safety reliability (ICC = 0.99) and strong psychometric alignment (ICC = 0.77) prove that the system is technically sound, clinically relevant, and usable today.

What makes this more exciting is the headroom for improvement. Current outputs are generated only through prompt engineering, yet the system already matches or surpasses human baselines in key dimensions. With fine-tuning on therapeutic dialogues, culturally adapted data, and **SEA-LION models** [1], KaLLaM has the potential to go beyond baseline competence and achieve clinically consistent, scalable therapeutic support.

In short, KaLLaM is not just a working prototype—it is a proof of concept for the future of safe, empathetic, and locally adapted therapeutic AI, capable of transforming both research evaluation pipelines and real-world mental health applications once enhanced with dedicated training

7. Project Disclaimer:

- **KaLLaM is not a substitute for professional care:** KaLLaM is a research and evaluation framework, and still not a replacement for real licensed therapists, psychologists, or medical professionals.
- **Intended use:** KaLLaM Chatbot is only for **research, training, and system evaluation**. In this project, KaLLaM is still underdeveloped for clinical advice, diagnoses, and emergency guidance. But KaLLaM Silver Standard Score can statistically be used to compare each Motivational Interview (MI) session accurately to some extent.
- **Safety first:** Users in crisis or experiencing severe distress should seek **qualified professional help or emergency services**, regardless of model outputs.
- **Model limitations:** The project uses **prompt-engineered responses only, with no fine-tuning**. Variability and inconsistency are expected. Results are a baseline, and **performance will improve significantly with future fine-tuning and domain adaptation**.

References

1. AI Singapore. (2024). *GitHub - aisingapore/sealion: South-East Asia Large Language Models*.

   GitHub. https://github.com/aisingapore/sealion

2. Giorgi, S., Havaldar, S., Ahmed, F., Akhtar, Z., Vaidya, S., Pan, G., Ungar, L. H., Schwartz, H. A., & Sedoc, J. (2023, May 24). *Psychological Metrics for Dialog System Evaluation*. arXiv.org. https://arxiv.org/abs/2305.14757

3. Jon M. Houck, Theresa B. Moyers, William R. Miller, Lisa H. Glynn, & Kevin A. Hallgren. (2010). Motivational Interviewing Skill Code (MISC) 2.5. In *University of New Mexico, Center on Alcoholism, Substance Abuse, and Addictions (CASAA)*.

4. Park, J. I., Abbasian, M., Azimi, I., Bounds, D. T., Jun, A., Han, J., McCarron, R. M., Borelli, J., Safavi, P., Mirbaha, S., Li, J., Mahmoudi, M., Wiedenhoeft, C., & Rahmani, A. M. (2024, August 3). *Building trust in Mental Health Chatbots: safety metrics and LLM-Based Evaluation tools*. arXiv.org. https://arxiv.org/abs/2408.04650

5. Rashkin, H., Smith, E. M., Li, M., & Boureau, Y. (2018, November 1). *Towards Empathetic Open-domain Conversation Models: a New Benchmark and Dataset*. arXiv.org. https://arxiv.org/abs/1811.00207

6. Xie, H. J., Zhang, J., Zhang, X., & Liu, K. (2024). Scoring with Large Language Models: A Study on Measuring Empathy of Responses in Dialogues. *2021 IEEE International Conference on Big Data (Big Data)*, 7433–7437. https://doi.org/10.1109/bigdata62323.2024.10825836