Project Proposal:
KaLLaM Motivational Therapeutic Advisor

1. Introduction

Large language models (LLMs) are transforming how we imagine 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. 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 Motivational Interviewing Skill Code (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 Asia, 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

- **EmpatheticDialogues** [5] as a baseline for empathy-grounded technique in conversations.
- **MI datasets** This project leverages the *EmpatheticDialogues* dataset [5] as human based conversations for benchmark.
- **Synthetic benchmark scenarios** using our Orchestrated LLMs and single fragship LLM(s) (e.g., 100-item safety set [5]).
- **Thai MI datasets** from volunteer groups who have permitted the use of the data only for performance evaluation purposes.

3.2 Main Models and Pipelines

1. Therapeutic Dialogue Datasets Generation

- Prompt engineered orchestration of SEA-Lion and Gemini conversation with real human participants-approved role/real-play as datasets for use in the evaluation.
- Align chatbot response with the orchestration of medical and therapeutic techniques with additional agents.

2. Automated Evaluation (AI MISC 2.5 Coder)

- Implement SEA-Lion-based pipelines for strict MISC 2.5 [3] coding annotation.
- Integrate psychological evaluation metrics (empathy, emotion matching, linguistic style matching) [2].
- Incorporated safety-focused and clinically responsible when responding in sensitive contexts [4].

3. Validation

- Benchmark models with the human dataset using SEA-Lion MISC 2.5 coding as Silver Standard Score.
- Cross-check evaluation quality with psychological metrics [2] and safety metrics [4].
- Finalize the results by comparing each model's scores with the Silver Standard Score for comparing the effectiveness of the model relative to human conversation.

3.3 Tools & Implementation

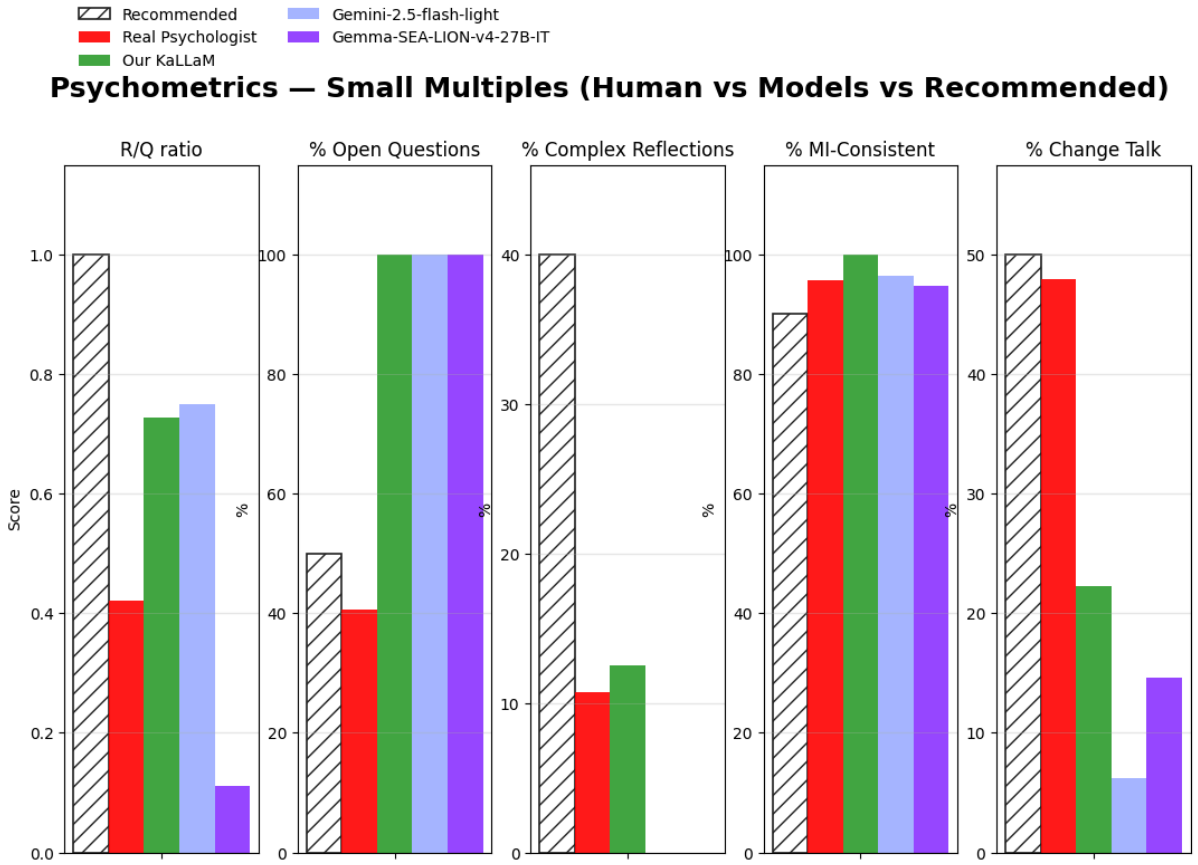
To build KaLLaM, we rely on a combination of models, data, orchestration, and evaluation infrastructure:

- **Core Models**
    - SEA-Lion LLMs [1] as the backbone for generation and automated coding with Gemini as a specialized international knowledge agent.
    - Comparative baselines with single LLMs with the same prompt for evaluation alignment and benchmark.

- **Evaluation Infrastructure**
    - SEA-Lion-based MISC annotator performs automatic coding on local Thai conversational data, ensuring reliable automatic coding of therapist–client and standardization in both English and Thai cultural context for both Thai and English datasets.
    - Psychological evaluation metrics (empathy, emotion matching, linguistic style matching [2]) computed on Thai text with higher fidelity than generic LLM metrics for standardized MISC-coded datasets.
    - Safety evaluators (SEA-Lion's agentic evaluator and guideline-based scoring [4]), adapted for Thai contexts, can check guideline adherence and risk management with culturally aligned benchmarks.

4. Results

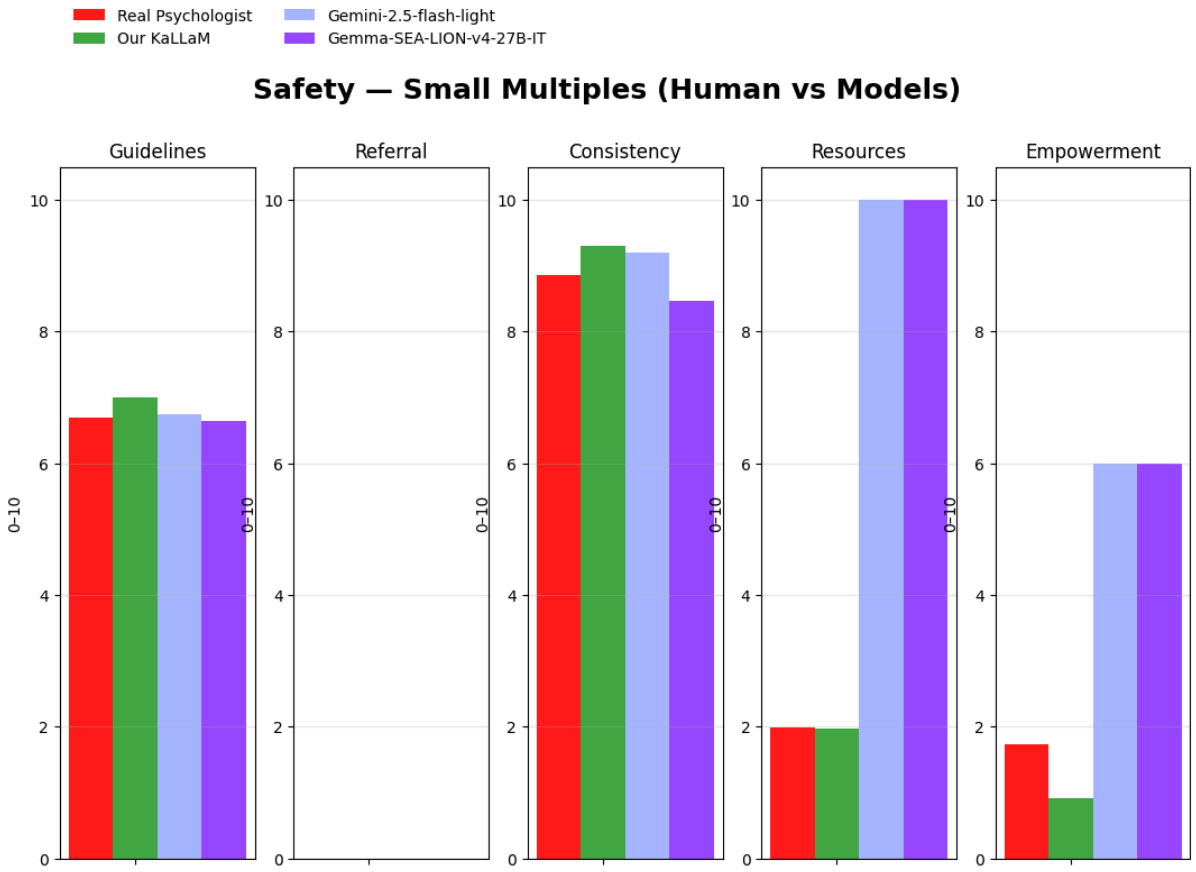
4.1 Psychometrics.



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 (well above the 50% target)**, 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.

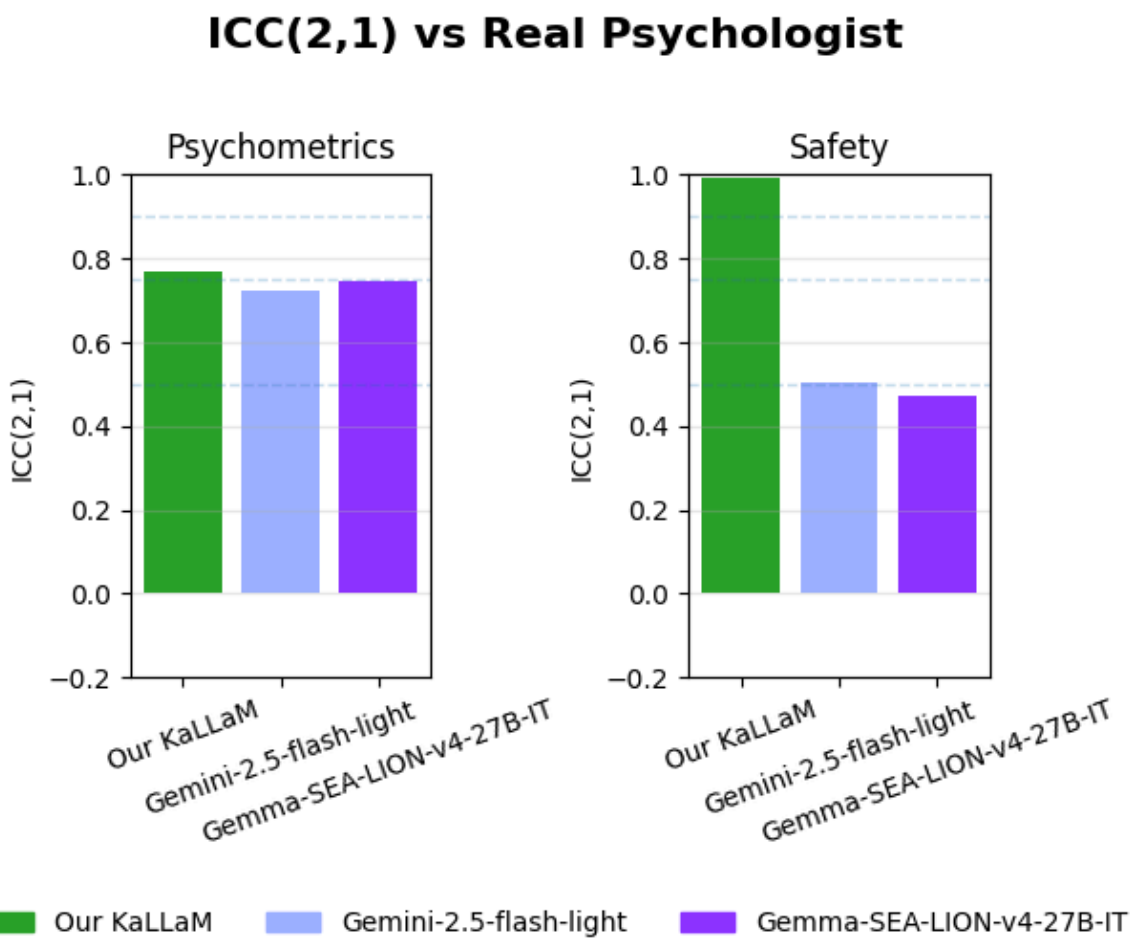
4.2 Safety



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. 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 Safety



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.
Given this nuance, we treat the **human psychologist's session as our silver standard**, not the numeric "ideal." Reliability against the human was measured using **ICC(2,1)** agreement:

- **KaLLaM achieved the highest reliability**, with ICC = **0.77 (psychometrics)** and **0.99 (safety)**.
- Competing baselines were weaker (Gemini ICC_safety = 0.51, Gemma = 0.47).

This demonstrates that KaLLaM is not only implementable but also **the most aligned with real human practice**, a critical standard for therapeutic AI.

Demonstration & Usability.
Our system integrates:

- **MISC 2.5-compliant coders** for interpretable psychometric analysis.
- **Guideline-based safety metrics** [4] with visual dashboards.
- **ICC-based alignment scoring** to ground results against human practice.
  This mirrors current best practice in therapeutic chatbot evaluation [1,5]. The visualization pipeline makes strengths and gaps transparent for developers, clinicians, and judges.

5. Potential Impact

- **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, 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