Large Language Models for Scalable Mental Health Support

Shaoxiong Ji

Principal Investigator, ELLIS Institute Finland
Assistant Professor, University of Turku
shaoxiong.ji@utu.fi

Planetary Health Symposium

10 August 2026

https://olaresearch.org

Mental Health & Planetary Well-Being

Ecological & Psychological Interconnections

- **Stressors:** Climate change drives eco-anxiety.
- **Resilience:** Vital for addressing global challenges.
- **Gap:** Traditional care systems cannot scale.

One Health Perspective

Interconnection:

- Environmental health
- Physical health
- **Mental well-being**

Digital tools scale public health support.

The Mental Health Crisis & AI Opportunity

- **Global Health Crisis:** Surging mental disorders vs. care bottlenecks.
- **Digital Footprints:** Emotional states & distress shared on social media.
- **AI Paradigm Shift:** Large Language Models (LLMs) for scalable digital support.

The Mental Health Crisis & AI Opportunity

- **Global Health Crisis:** Surging mental disorders vs. care bottlenecks.
- **Digital Footprints:** Emotional states & distress shared on social media.
- **AI Paradigm Shift:** Large Language Models (LLMs) for scalable digital support.

Three Pillars of this Talk

1. **Mental Health Analysis:** Evaluating interpretable reasoning.
2. **Conversational AI:** Grounding dialogue in clinical profiles.
3. **Psychological Defense:** Identifying protective mechanisms.

Mental Health in Social Content

- **Global Burden:** Depression & anxiety are leading causes of disability.
- **Escalation Risks:** Untreated distress can lead to self-harm.
- **Early Detection:** NLP on digital footprints triggers timely prevention.

Figure: Warning signs of mental illness[a]

[a] Source
https://www.nami.org/Blogs/NAMI-Blog/May-2022/Understanding-The-Early-Warning-Signs-of-Mental-Illness

Significance in Mental Health

- **Linguistic Expression:** Reveals emotional states.
- **Diagnostics:** Language is central to understanding.
- **AI Opportunity:** NLP/ML innovate therapeutics.

Large Language Models: A New Era

The LLM Revolution:

- **Redefined:** Language understanding & generation.
- **Scalable Support:** Psychoeducation guidance.
- **Data Insights:** Sentiment analysis in social data.
- **Clinical Augmentation:** Dialogue systems.

In-context learning & advanced reasoning.

Mental Health Analysis: Motivation & Research Questions[1]

The Challenge:

- Rapid evolution from PLMs to LLMs.
- **Gap:** Explainability.

[1] K. Yang, S. Ji, T. Zhang, Q. Xie, Z. Kuang, and S. Ananiadou. "Towards Interpretable Mental Health Analysis with Large Language Models". In: *Proceedings of EMNLP*. 2023. URL: https://arxiv.org/abs/2304.03347.

Mental Health Analysis: Motivation & Research Questions[1]

The Challenge:

- Rapid evolution from PLMs to LLMs.
- **Gap:** Explainability.

Research Questions

- **RQ1:** Mental health task performance?
- **RQ2:** Impact of emotional prompts?
- **RQ3:** Human-like reasoning ability?

[1] K. Yang, S. Ji, T. Zhang, Q. Xie, Z. Kuang, and S. Ananiadou. "Towards Interpretable Mental Health Analysis with Large Language Models". In: *Proceedings of EMNLP*. 2023. URL: https://arxiv.org/abs/2304.03347.

Evaluation Pipeline: Datasets & Setup

Evaluation Setup:

- **Models:** InstructGPT-3, ChatGPT.
- **Datasets:** 11 datasets spanning depression, stress, and suicidality (e.g., DR, Dreaddit, CAMS, SAD).
- **Tasks:** Binary detection, multi-class, and cause detection.
- **Paradigm:** Zero-shot and emotion-enhanced CoT prompting.

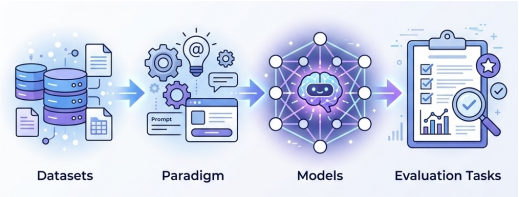


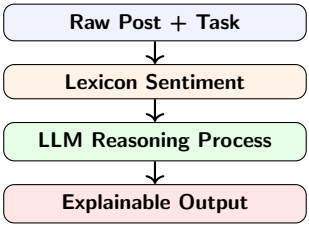
Figure: Evaluation Pipeline

Emotion-Enhanced Prompting

Prompting Strategies:

- **Zero-Shot CoT:** "Explain step-by-step".
- **Lexicon Fusion:** VADER & NRC EmoLex.
- **Few-Shot:** Expert annotations.

Task Settings: Binary, Multi-Class, Cause.



Explainable Detection with ChatGPT

Explainable AI:

- **Reasoning:** Step-by-step classification.
- **Fusion:** Emotion labels + CoT logic.

Objectives:

- **Transparency:** Natural language explanations.
- **Performance:** Explainable vs. pure prediction.

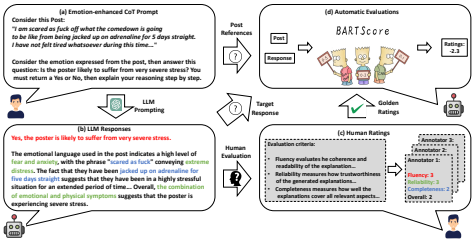


Figure: Prompting and Reasoner Pipeline

Explanations are evaluated on fluency, reliability, and completeness.

Human Evaluation of Explainability

Human-in-the-Loop Study:

- **Criteria (0–3 scale):** Fluency, Reliability, Completeness, and Overall performance.
- **Expert Assessment:** Scored by psychology PhD students.
- **Agreement:** High inter-annotator agreement (95.9% for correct predictions, $\geq$0.41 Fleiss' Kappa).

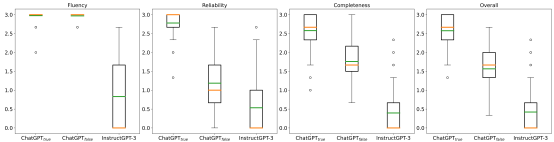


Figure: Human Evaluation Score Distributions

Experimental Results & Key Findings

F1 Correctness Results (Weighted F1 %)

Model	DR	CLPsych	Dreaddit	T-SID	SAD	CAMS
Supervised SOTA						
MentalRoB	**94.23**	**69.71**	**81.76**	**89.01**	**68.44**	**47.62**
Zero-Shot LLM-based						
LLaMA-13B	54.07	39.29	36.28	25.27	13.20	14.64
ChatGPT	82.41	56.31	71.79	33.30	54.05	33.85
Emotion-enhanced CoT (Ours)						
$\text{ChatGPT}_{\text{CoT\_emo}}$	83.10	58.24	74.83	27.71	56.68	42.29
$\text{ChatGPT}_{\text{CoT\_FS}}$	**84.22**	**61.63**	**75.38**	**43.95**	**63.56**	**45.99**

Key Insights:

- **LLM Gap:** Zero-shot LLMs lag behind MentalRoBERTa.
- **Emotion CoT:** Grounding in emotional reasoning improves performance.
- **Few-Shot Boost:** Calibration via expert examples bridges the gap (especially on CAMS).

Conversational AI in Mental Health

Therapeutic Chatbots[2]:

- **Personalization:** History tracking.
- **Empathetic Alignment:** Active listening.
- **Safety vs. Utility:** Harm mitigation.

Technical Challenge

Require clinical grounding (e.g., CBT) and structured states.

[2] S. **Ji**, T. Zhang, K. Yang, S. Ananiadou, and E. Cambria. "Rethinking Large Language Models in Mental Health Applications". In: *arXiv preprint arXiv:2311.11267* (2023).

Dialogue Context & Risk Assessment

Sequential Risk Assessment in Conversations[3]:

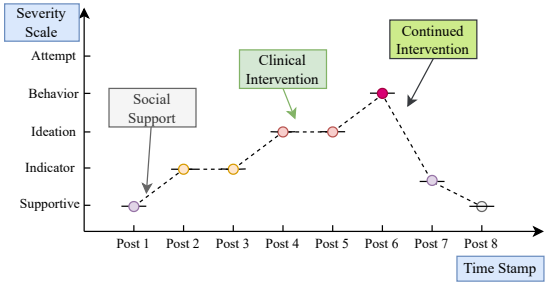


Figure: Sequential tracking of risk levels

Conversational Language Analysis:

- **Dynamics:** Evolving risk states.
- **Memory:** Long-term patient history.
- **Grounding:** In clinical bases.

[3] S. Ji. "Towards Intention Understanding in Suicidal Risk Assessment with Natural Language Processing". In: *Findings of EMNLP*. Association for Computational Linguistics, 2022, pp. 4028–4038. URL: https://aclanthology.org/2022.findings-emnlp.297.

SQPsych: Dialogue Generation from Questionnaires

Privacy Bottleneck:

- **Regulations:** GDPR/HIPAA block data sharing.
- **Scarcity:** Limits open-weight model training.

SQPsych Solution:

- Condition dialogues on structured patient questionnaires (HAM-D, HAM-A, BDI-II).
- Bypasses privacy barriers; preserves clinical diagnostic logic.

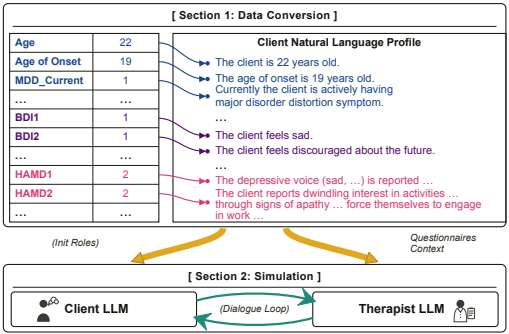


Figure: SQPsych Generation Pipeline

SQPsych: Roleplay Simulation & Expert Preference

Roleplay Framework:

- **Client Agent:** Emotion grounded in questionnaire profiles.
- **Therapist Agent:** Applies standard CBT techniques.

Expert Validation:

- Evaluated 7 models (**SQPsychLLM**) with trained therapists.
- SQPsych makes models significantly better at therapist roleplay.

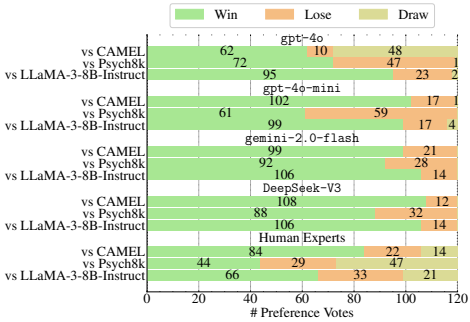


Figure: Pairwise preference by expert therapists (SQPsych vs Baselines)

Resources at
https://ai-mh.github.io/SQPsych.html

PsyDefConv: Psychological Defenses in Conversations[4]

Objective:

- Adapt DMRS to detect defenses in supportive dialogues.
- **Online Setting:** Real-time detection (no future context).

Problem Formulation

Input: Context $D = (u_1, \ldots, u_t)$, seeker turn u_t.
Output: Defense level $y \in \{0, \ldots, 8\}$.

DMRS Defense Levels (L0–L8)

- **L0:** No Defenses (phatic/functional)
- **L1:** Action (acting out, passive aggression)
- **L2:** Major Image-Distorting (splitting)
- **L3:** Disavowal (denial, projection)
- **L4:** Minor Image-Distorting (devaluation)
- **L5:** Neurotic (repression, displacement)
- **L6:** Obsessional (affect isolation)
- **L7:** High-Adaptive (affiliation, humor)
- **L8:** Needs More Information

[4] H. Na, Z. Wang, Z. Chen, P. Zhou, Y. Hua, G. Z. Zhou, H. Zhang, T. Shen, W. Wang, J. Torous, et al. "You never know a person, you only know defenses: Detecting levels of psychological defense mechanisms in supportive conversations". In: *Findings of the Association for Computational Linguistics: ACL 2026*. 2026, pp. 14428–14448.

PsyDefConv: Dataset Construction & Statistics

Construction:

- **Base:** ESConv (Emotional Support Conversations).
- **Size:** 200 dialogues, 4,709 utterances (2,336 seeker turns).
- **Annotation:** Double-blind (Cohen's $\kappa = 0.639$).

Skewed Distribution:

- **High-Adaptive (L7):** 51.8% (dominant)
- **Immature (L1-4):** 18.9%
- **Neurotic (L5-6):** 11.9%
- **Ambiguous/Non-def (L0, 8):** 17.4%

Evaluation Metric:

- **Macro F1 (classes 1–8)** to penalize majority-class bias.

DMRS Co-Pilot: AI-Assisted Annotation

Mitigating Clinician Burnout:

- Clinical DMRS annotation is highly intensive and exhausting.
- **DMRS Co-Pilot:** A 4-stage cascaded LLM pipeline (Gemini 2.5 Pro/Flash):
  1. **Stressor ID:** Highlights stressors.
  2. **Screening:** Screens 150 DMRS criteria.
  3. **Validation:** Gathers textual evidence.
  4. **Synthesis:** Ranks recommendations.
- **Impact:** Speeds up annotation by **24.0%**, reducing cognitive fatigue.

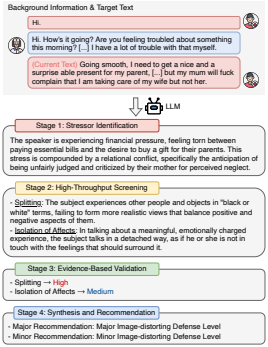


Figure: DMRS Co-Pilot Architecture

PsyDefDetect Shared Task @ BioNLP 2026[5]

Shared Task Overview:

- **Task:** Predict DMRS defense levels (1–8) from context.
- **Participation:** 172 participants, 21 teams.

Official Baselines:

Model	Type	Macro F1	Acc
Gemini 2.5 Pro	ZS	0.2599	0.5636
DeepSeek-V3.2	ZS	0.2617	0.5572
Ministral-8B	FT	**0.3148**	**0.6483**
InternLM3-8B	FT	0.3053	0.6398

Evaluation Metric

Macro F1-score over positive classes (1–8) to penalize majority L7 class bias.

[5] H. Na, Z. Wang, Z. Chen, Y. Hua, R. Gao, K. Yang, L. Chen, W. Wang, S. Ji, J. Torous, and S. Ananiadou. "Overview of the PsyDefDetect Shared Task at BioNLP 2026: Detecting Levels of Psychological Defense Mechanisms in Supportive Conversations". In: *Proceedings of Biomedical Natural Language Processing Workshop*. 2026, pp. 932–943.

PsyDefDetect: Top Systems & Key Insights

Leaderboard Results:

- **Rank 1: Nürnberg NLP (F1 = 0.4200)**
  9-voter ensemble (Ministral, Phi-4).
- **Rank 2: UTS (F1 = 0.4055)**
  Gemini agent council + Qwen QLoRA.
- **Rank 3: PerceptionLab (F1 = 0.3956)**
  DMRS-Q retrieval + Flash fine-tuning.

Key Insights:

- Top systems exceed baseline by **+10.5 F1**.
- Accuracy is misleading due to 52% L7 class skew.

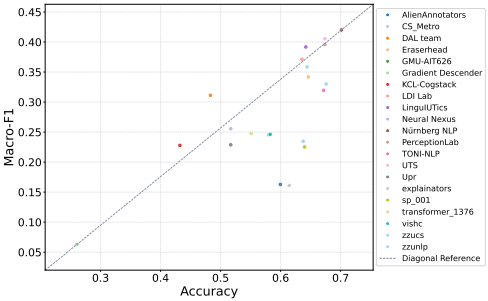


Figure: Accuracy vs. Macro F1 across systems

Ethical Considerations in Clinical AI

- **Privacy:** Local open-weight models (HIPAA/GDPR).
- **Safety:** Prototypes only; mitigate hallucinations.
- **Human-in-the-Loop:** AI as co-pilot, not replacement.

Future Directions

- **Pilot Studies:** Therapist-guided clinical trials.
- **Diversification:** ACT, DBT, and psychoanalysis.
- **Multimodal:** Fusing text, voice, and physiology.

Future Directions

- **Pilot Studies:** Therapist-guided clinical trials.
- **Diversification:** ACT, DBT, and psychoanalysis.
- **Multimodal:** Fusing text, voice, and physiology.

The Ultimate Goal

Secure, compliant, clinically aligned AI assisting human therapists globally.

Conclusion

- **Explainability:** Essential to bridge accuracy and clinical justification.
- **Therapeutic simulation:** Structured profile grounding and roleplay (SQPsych) enables therapeutic simulation.
- **Psychological defense:** Multi-stage pipelines automate complex clinical annotations (PsyDefConv).

Conclusion

- **Explainability:** Essential to bridge accuracy and clinical justification.
- **Therapeutic simulation:** Structured profile grounding and roleplay (SQPsych) enables therapeutic simulation.
- **Psychological defense:** Multi-stage pipelines automate complex clinical annotations (PsyDefConv).

https://olaresearch.org/

Join us on Discord

Thank You!

Collaborators & Partners:

ELLIS Institute Finland | University of Turku | University of Manchester | TU Darmstadt | University of Technology Sydney

Hongbin Na, Doan Nam Long Vu, Kailai Yang, Sophia Ananiadou, Ling Chen, John Torous, Hamidreza Jamalabadi, Tilo Kircher, Udo Dannlowski, Christiane Hermann, Simone Balloccu, Zimu Wang, Zhaoming Chen, Yining Hua, Erik Cambria, Tianlin Zhang, Annika Schoene, Rui Tan, Svenja Jule Francke, Lena Moench, Daniel Woiwod, Florian Thomas-Odenthal, Sanna Stroth, Peilin Zhou, Grace Ziqi Zhou, Haiyang Zhang, Tao Shen, Wei Wang, Rena Gao, Luna Ansari, Qian Chen, Qianqian Xie, Ziyan Kuang