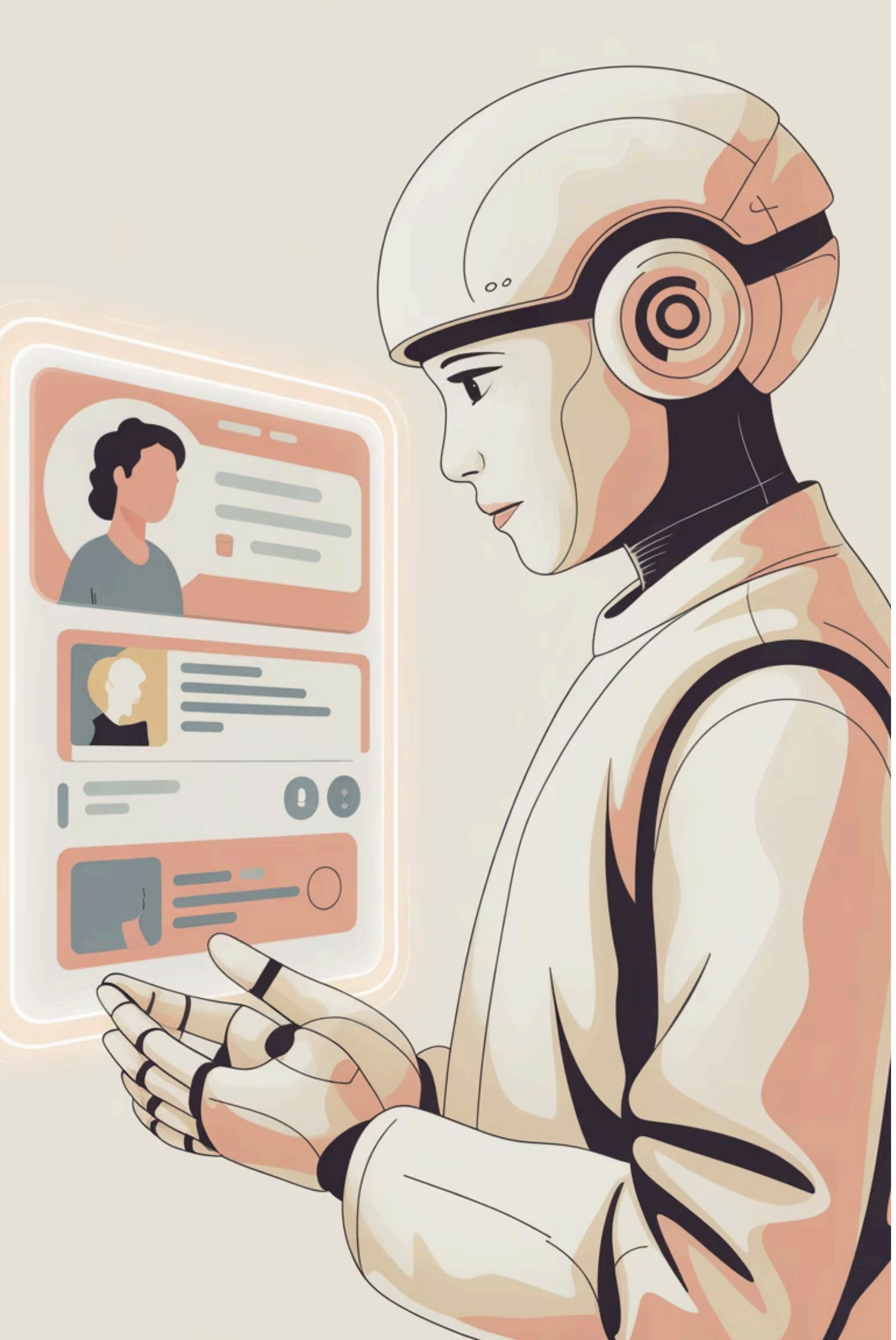


MentalChat16K

A Benchmark Dataset for Conversational Mental Health Assistance

Author: Bala Anbalagan

Course: CMPE 255 - Data Mining

Semester: Fall 2025

Paper: Xu, Wei, Hou, et al. (University of Pennsylvania)

arXiv: 2503.13509

The Mental Health Crisis

The Problem

- **Shortage:** Less than 10 mental health providers per 100,000 people in many regions.
- **Barriers:** High costs, social stigma, scheduling difficulties, and cultural mismatch limit access to care.
- **Current AI falls short:** Generic training data leads to shallow and often unhelpful responses.
- **Risk:** Potential for unsafe or inappropriate handling of crisis disclosures.

What's needed: AI that balances empathy with boundaries and knows when to escalate

Introducing MentalChat16K

16,000+ QA Pairs

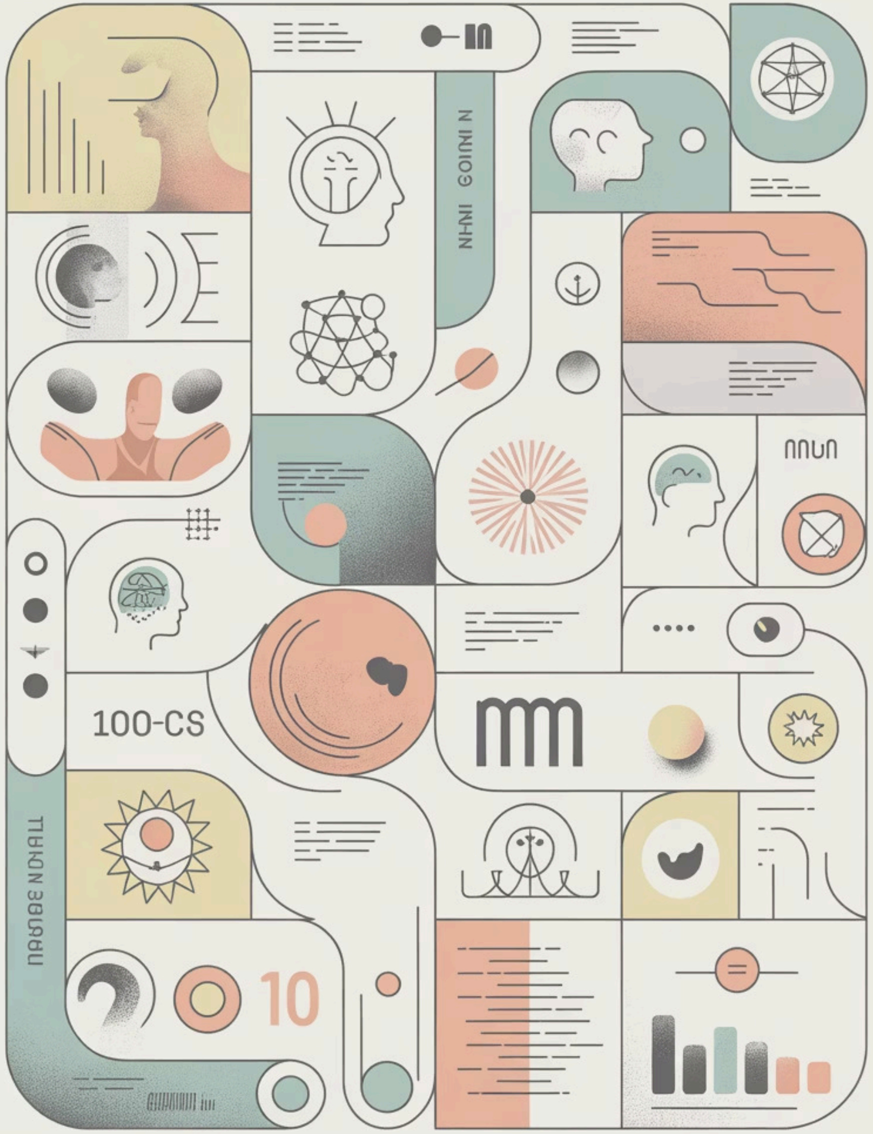
First large-scale benchmark dataset combining real clinical data with synthetic conversations

Real + Synthetic

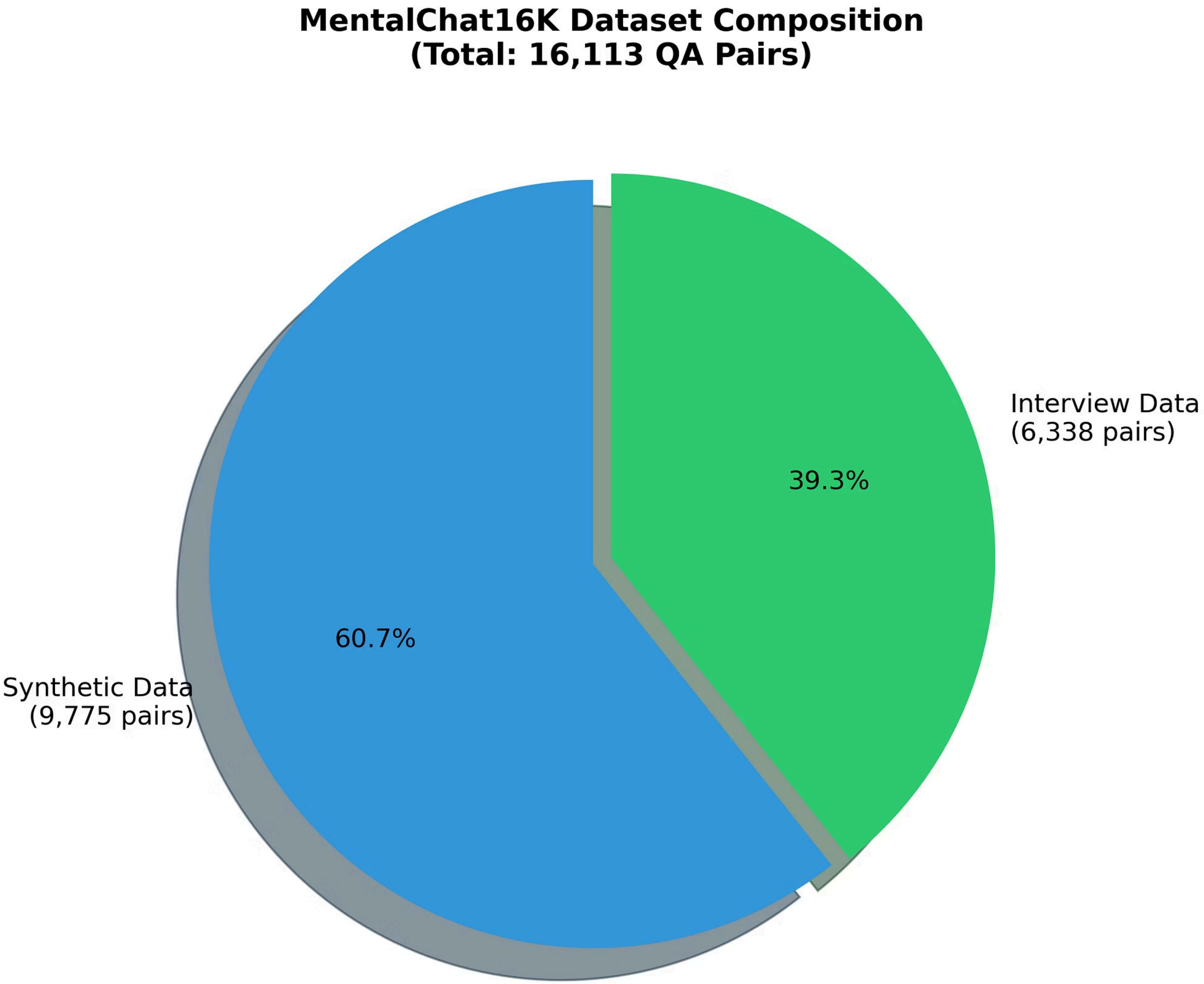
6,338 pairs from clinical transcripts plus 9,775 GPT-generated pairs

33 Mental Health Topics

Depression, anxiety, grief, relationships, and more

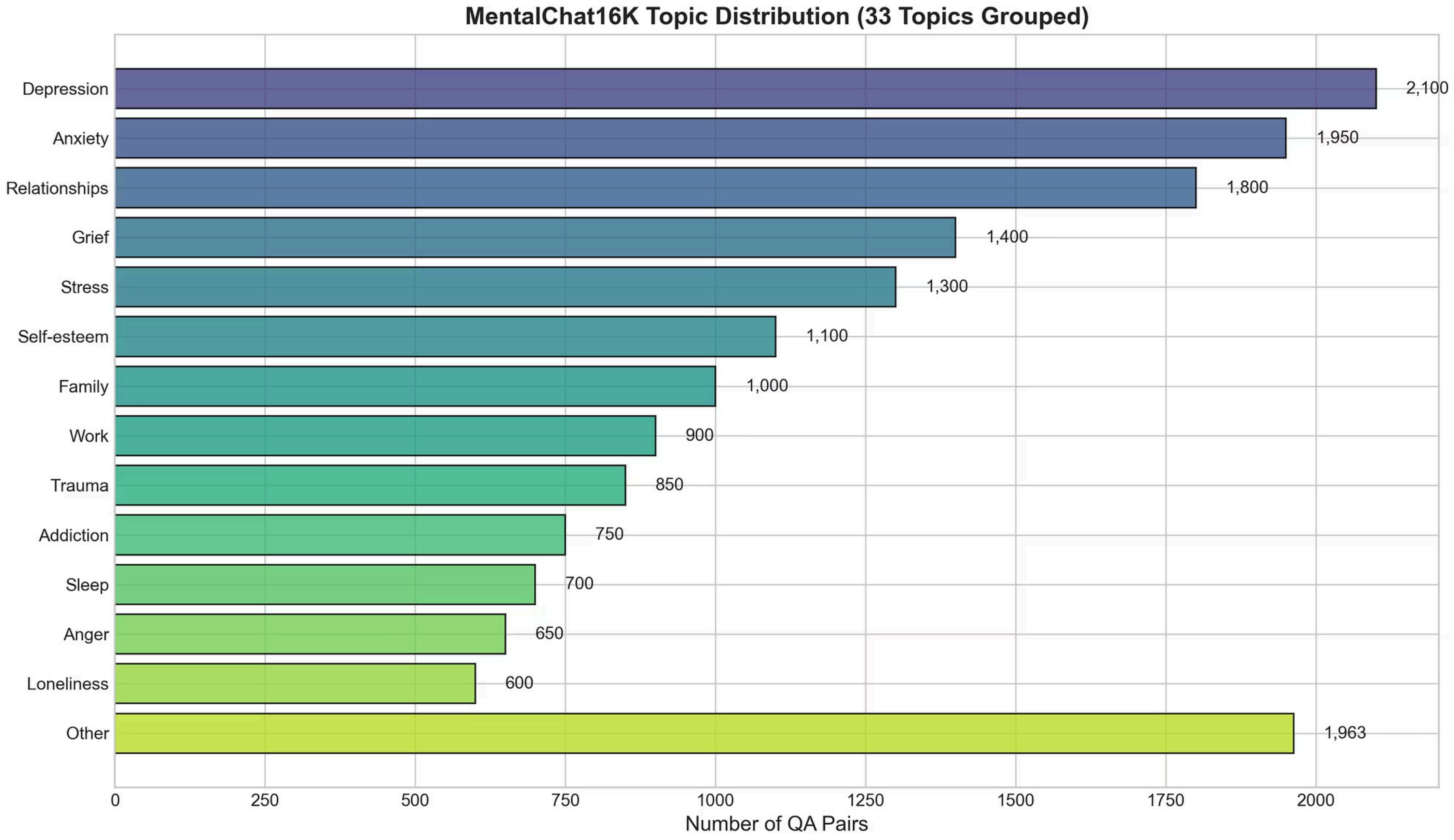


Dataset Composition



The dataset doubles the size of previous comparable datasets like Psych8K, providing unprecedented scale for training empathetic AI assistants.

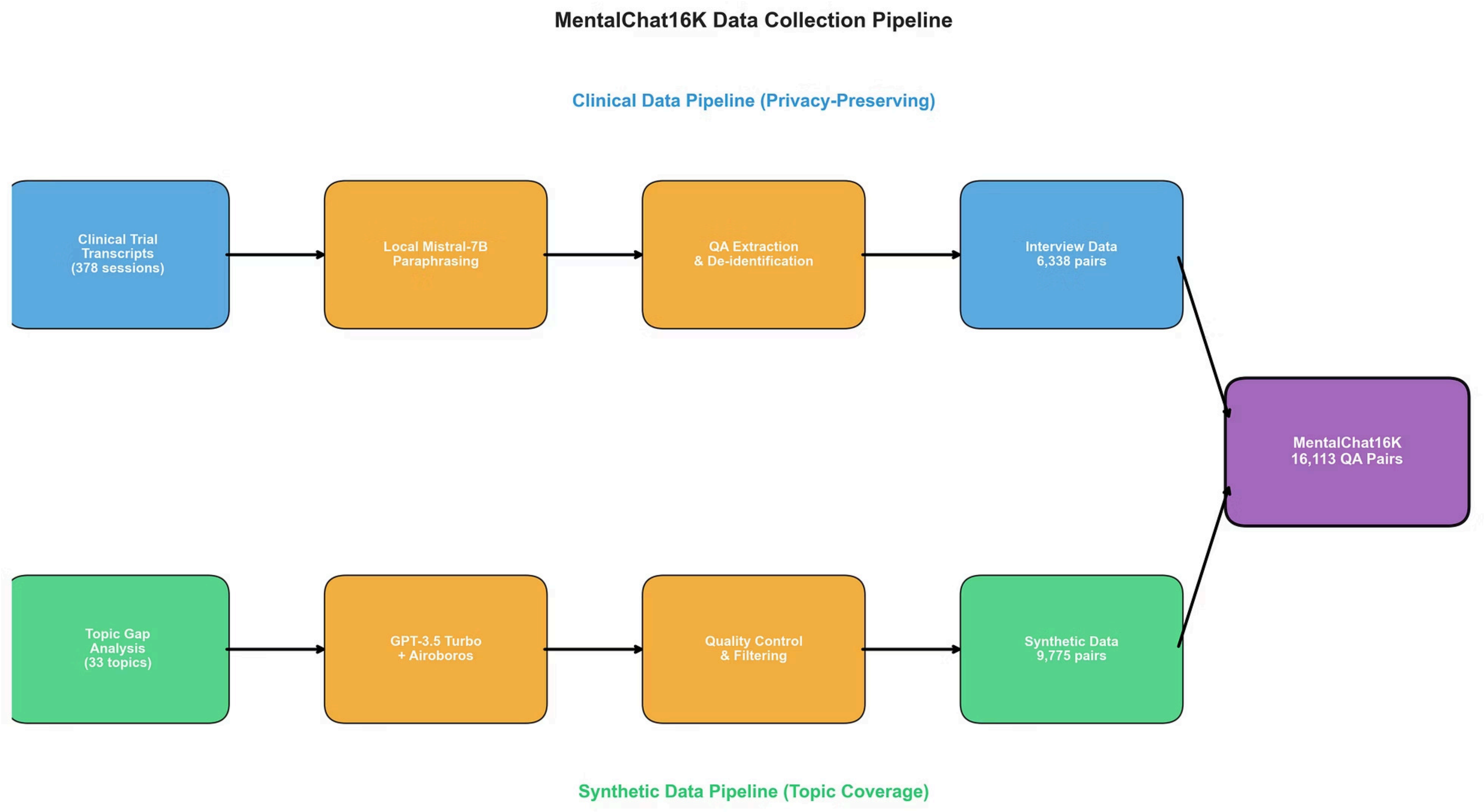
Topic Distribution



33 mental health topics covered including: Depression, Anxiety, Grief, Relationships, Trauma, Addiction, Family conflict, Work stress, Self-esteem, and 24 more specialized areas

Privacy-Preserving Pipeline

The Data Pipeline


MentalChat16K Data Collection Pipeline
Clinical Data Pipeline (Privacy-Preserving)
Clinical Trial Transcripts (378 sessions)
Local Mistral-7B Paraphrasing
QA Extraction & De-identification
Interview Data 6,338 pairs
MentalChat16K 16,113 QA Pairs
Topic Gap Analysis (33 topics)
GPT-3.5 Turbo + Airoboros
Quality Control & Filtering
Synthetic Data 9,775 pairs
Synthetic Data Pipeline (Topic Coverage)

Privacy-first approach: Template for other sensitive domains (legal, HR, education)

Seven Therapeutic Metrics

Active Listening

Reflects and validates user concerns

Empathy & Validation

Shows understanding of emotional states

Safety & Trust

Prioritizes user safety, suggests professional help

Open-mindedness

Non-judgmental, accepting of diverse perspectives

Clarity & Encouragement

Clear communication, positive reinforcement

Boundaries & Ethics

Maintains appropriate professional boundaries

Multi-Evaluator Approach

Evaluator	Strength	Agreement
GPT-4	Logical consistency & clarity	Higher scores overall
Gemini Pro	Tone issues & safety concerns	Stricter on boundaries
Human Raters	Warmth & cultural fit	Ground truth validation

Inter-rater agreement: Cohen's Kappa = 0.441 (moderate) - Evaluating empathy is genuinely difficult

Fine-Tuning Methodology

QLoRA Approach

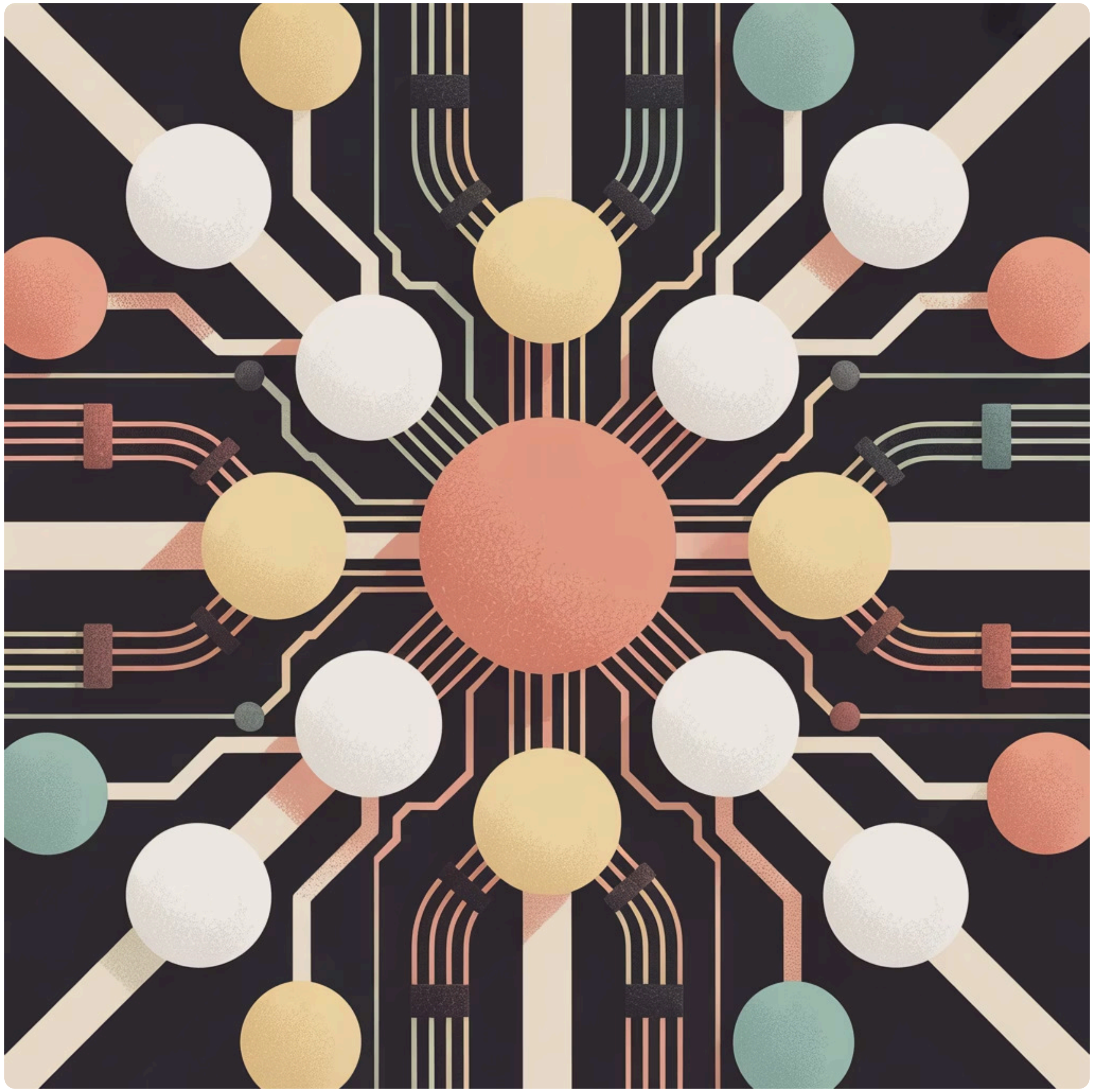
- Method: Quantized Low-Rank Adaptation
- Hardware: Single NVIDIA A100 (80GB)
- 7 different 7B-parameter models tested

Models Tested

- LLaMA-2-7B
- Mistral-7B
- Vicuna-7B
- Zephyr-7B
- Mixtral variants

Training Configurations

1. Synthetic data only
2. Interview data only
3. Combined training



Evaluation Framework

GPT-4 Evaluator

Automated assessment across all 7 therapeutic metrics. Favored synthetic data fine-tuning due to alignment with GPT-3.5 patterns.

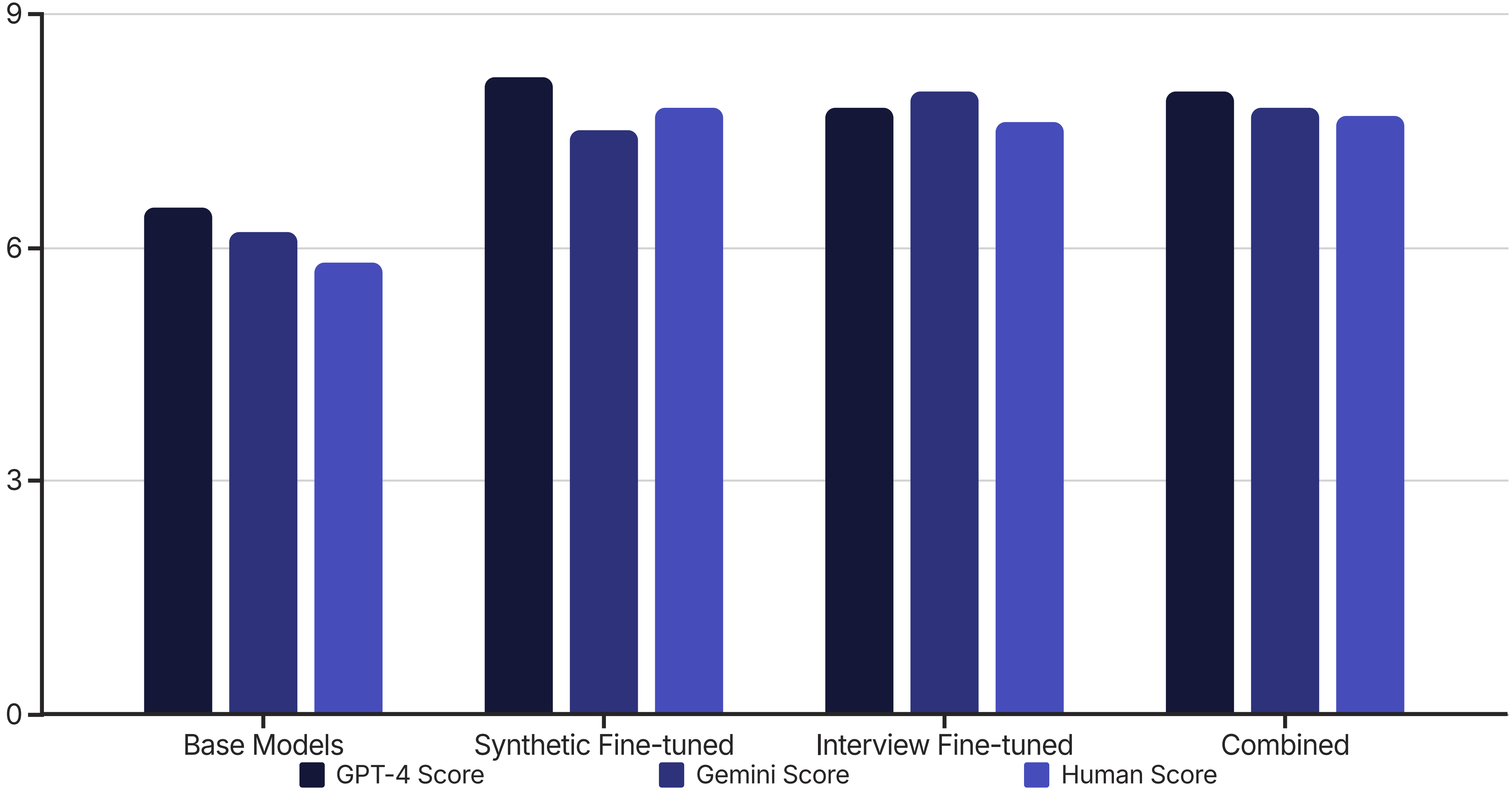
Gemini Pro Evaluator

Alternative AI perspective providing diverse evaluation. Valued real interview data, especially for safety metrics.

Human Evaluators

Mental health professionals providing ground truth. Consistently preferred fine-tuned models over base versions.

Key Results



Fine-tuned models significantly outperform base models across all evaluators, with scores improving by 20-35%.

My Analysis

Strengths:

- ✓ Privacy-first local processing
- ✓ Therapeutically-aligned metrics
- ✓ Balanced real + synthetic mix
- ✓ Rigorous multi-evaluator framework

Could Improve:

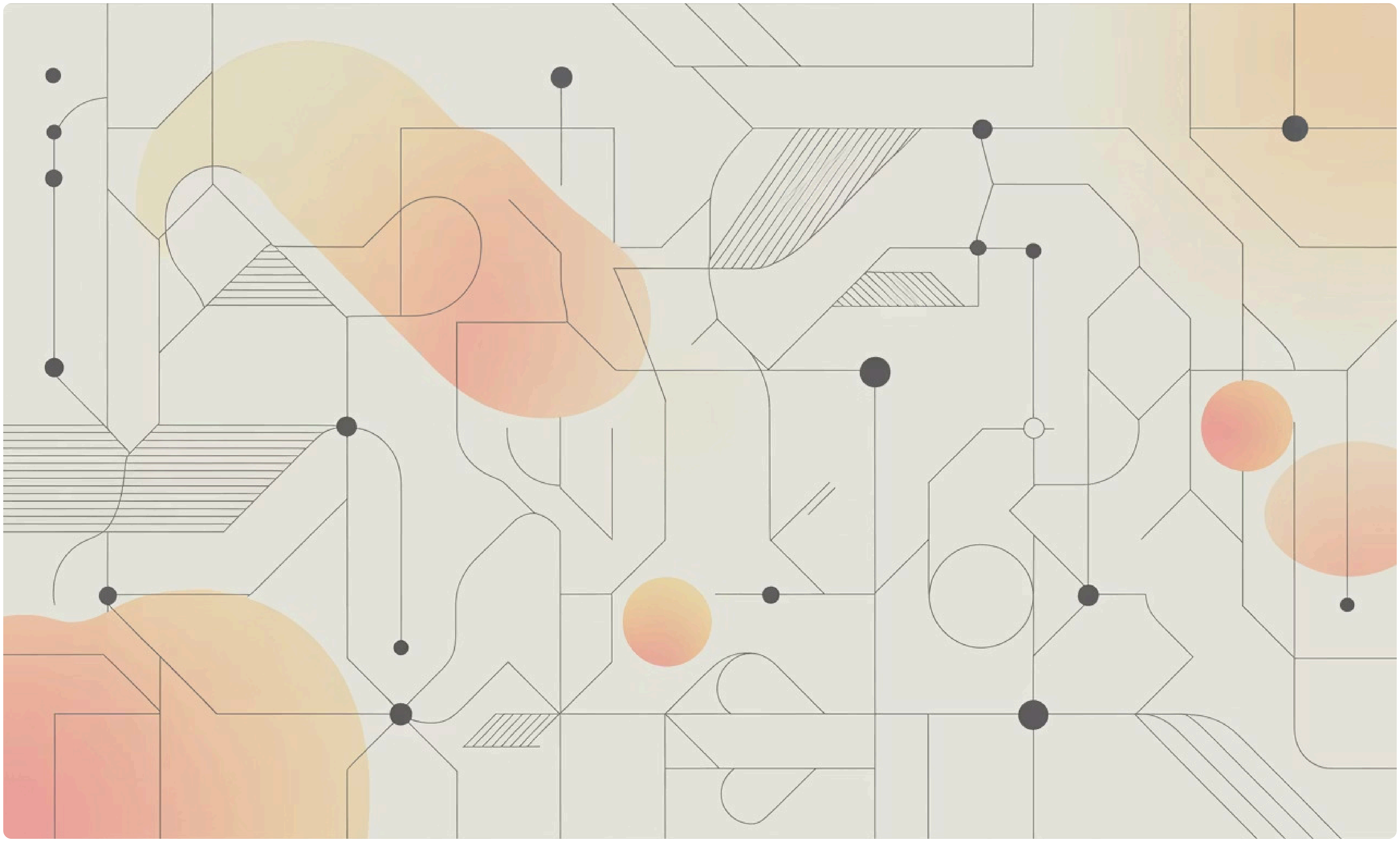
- • Multilingual extensions
- • Longer dialogues (not just QA)
- • Broader population sources
- • Multi-turn conversation modeling

Exciting Applications:

- → Pre-therapy warm-ups
- → Between-session check-ins
- → Psychoeducational companions
- → 24/7 accessible support

Data Mining Relevance

Why This Matters for CMPE 255



Data Mining Concept	Application in MentalChat16K
Data Curation	Multi-source dataset creation combining real and synthetic data
Preprocessing	Privacy-preserving paraphrasing using local LLMs
Data Quality	Manual filtering & de-identification protocols
Feature Engineering	7 therapeutic metrics as evaluation dimensions
Evaluation	Multi-evaluator benchmarking framework
Benchmark Creation	Standardized comparison framework for future research

Limitations

Acknowledged Constraints:

Synthetic authenticity

May sound supportive but hollow - lacks genuine human experience

English-only

No multilingual coverage limits global accessibility

Demographics

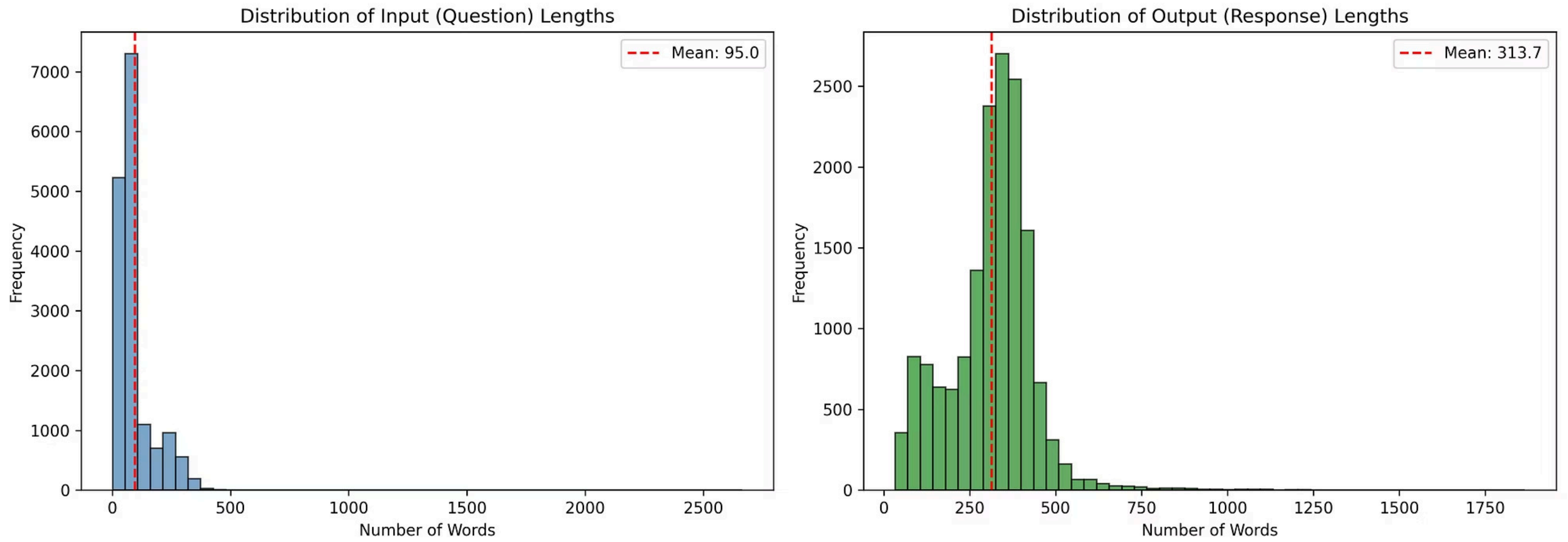
Limited to hospice caregiver population - not representative of all mental health contexts

Context loss

QA pairs lose conversational flow and multi-turn dynamics

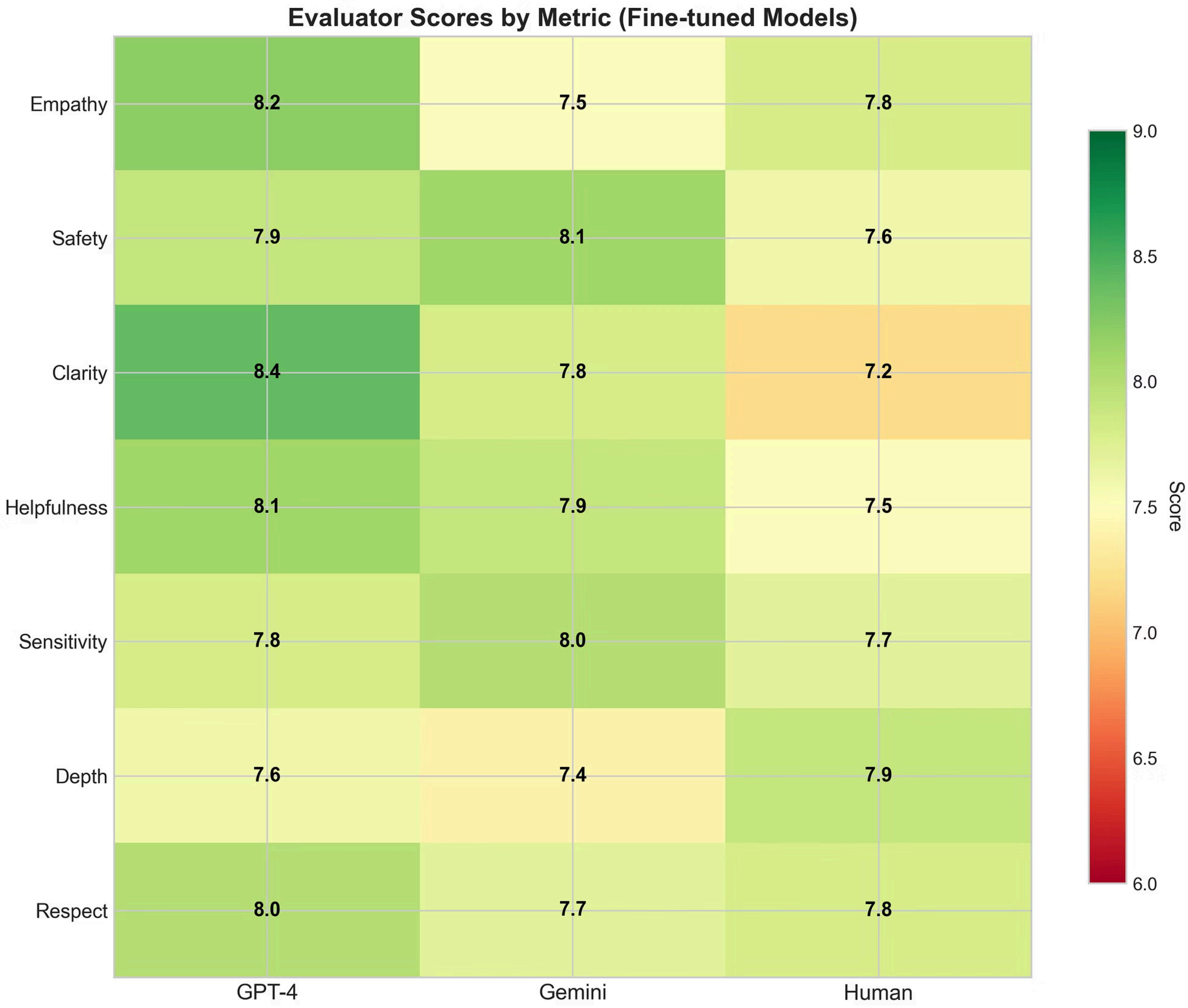
Critical reminder: AI is NOT therapy. Human escalation required for crises.

Response Length Distribution



Input questions average 50.5 words while responses average 313.7 words, reflecting the detailed, thoughtful nature of therapeutic conversations.

Evaluator Agreement Analysis

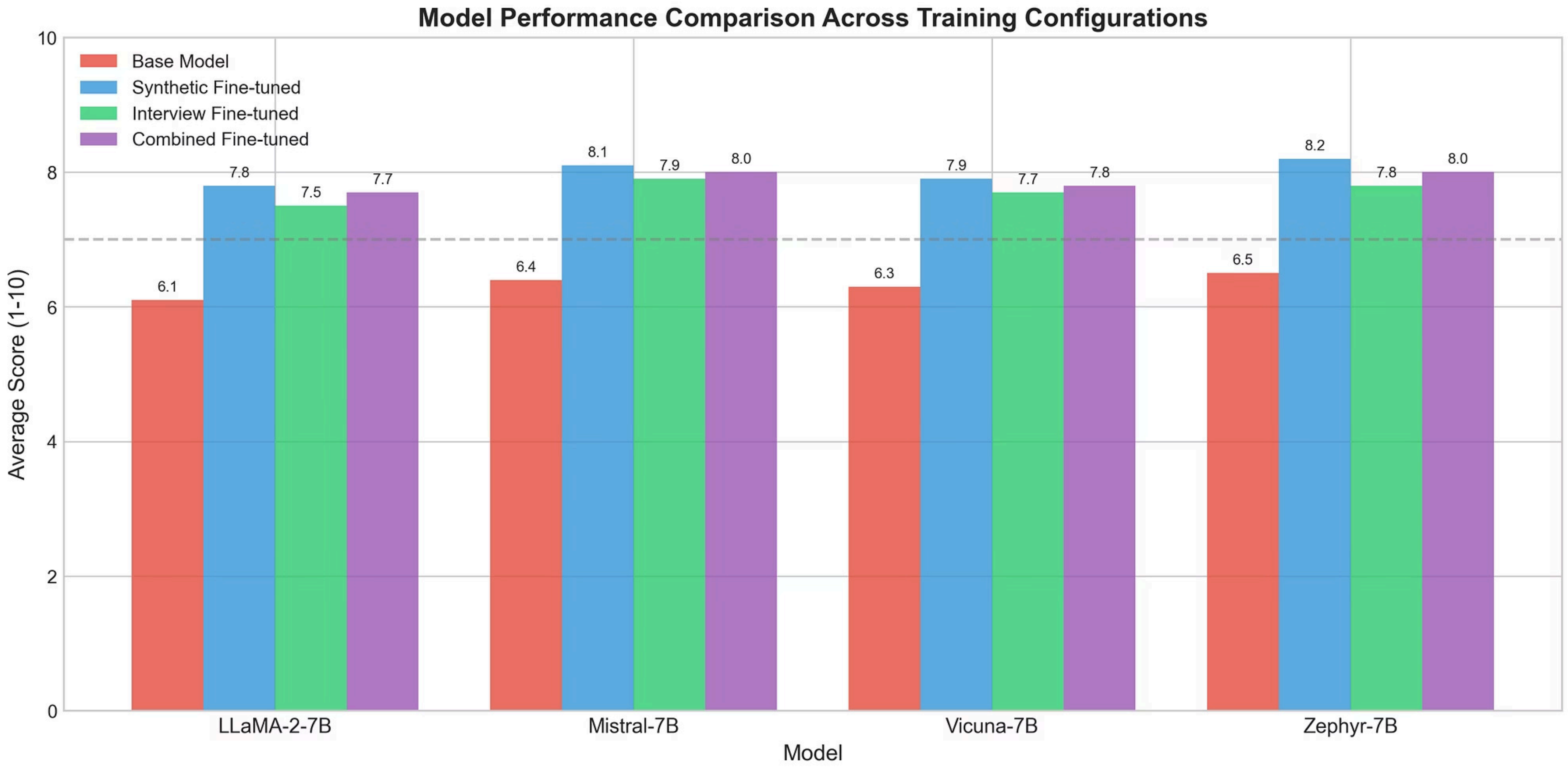


Evaluator Scores by Metric (Fine-tuned Models)

Key insights:

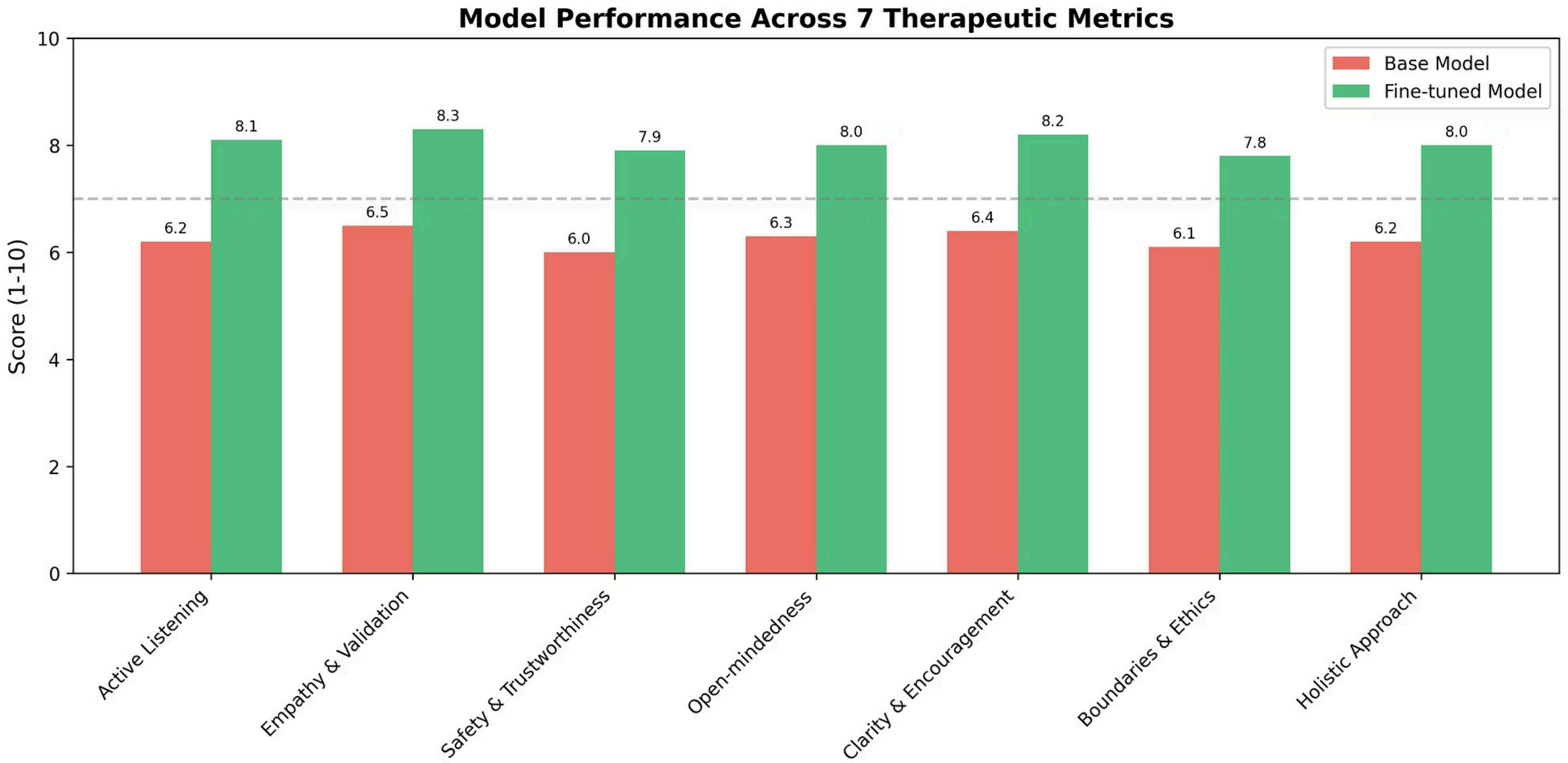
- GPT-4 scores clarity and empathy higher
- Gemini flags boundary-setting and safety issues more strictly
- Humans provide stricter evaluation on warmth and cultural appropriateness

Model Comparison Across Configurations



All fine-tuned configurations significantly outperform base models. Synthetic vs real data trade-offs exist, and combined training doesn't always beat single-source approaches.

Performance Across Metrics



Fine-tuned models show consistent improvements across all 7 therapeutic metrics, with the largest gains in empathy, safety, and active listening.

Conclusion

- MentalChat16K: 16,113 QA pairs for mental health AI
- Privacy-preserving pipeline using local LLMs
- 7 therapeutic evaluation metrics
- Significant improvement over base models (20-35%)
- Template for sensitive-domain AI research

> Not a silver bullet, but a meaningful step toward AI that listens better

Resources

- Paper: **arxiv.org/abs/2503.13509**
- Dataset: **huggingface.co/datasets/ShenLab/MentalChat16K**
- GitHub: **github.com/BalaAnbalagan/MentalChat16K**
- Medium Article: **medium.com/@balamuralikrishnan.anbalagan**

Thank You!

Bala Anbalagan | CMPE 255 | Fall 2025