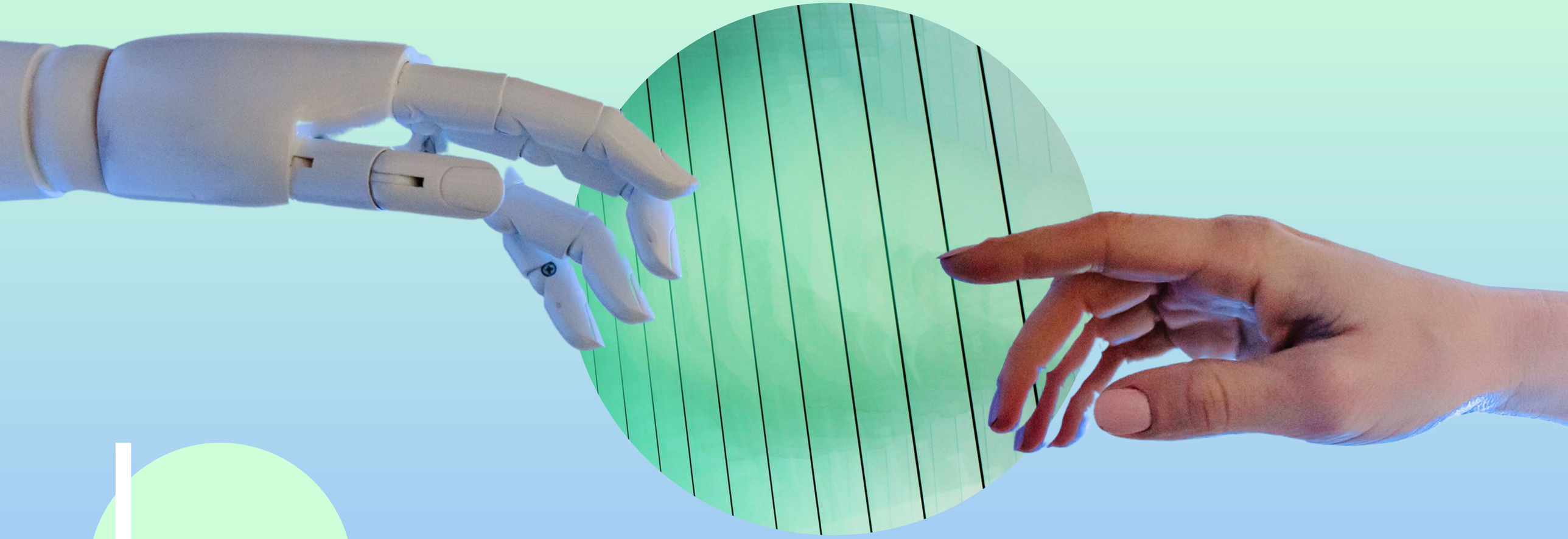
MSAI Capstone
Nathanael Johnson
July 31, 2025
Advisor: Dr. Steve Nordstrom

MENTAL HEALTH LLM EVALUATION: LOCAL VS CLOUD MODELS

AI

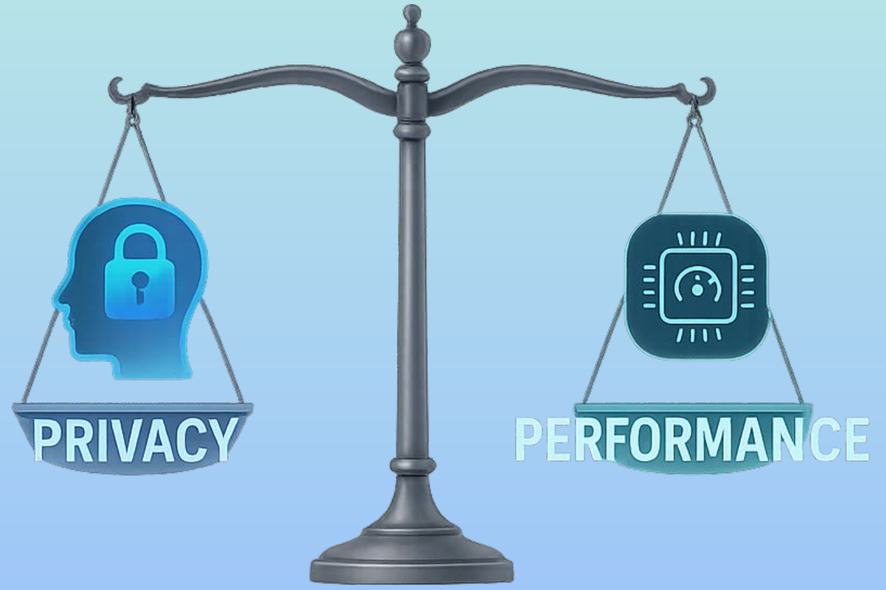


Introduction

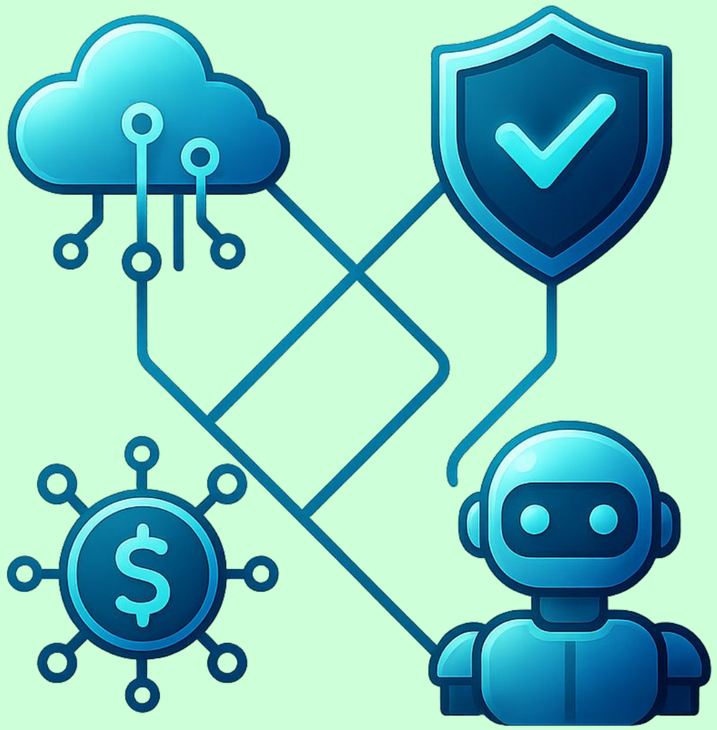
The Healthcare AI Dilemma

THE CHALLENGE WE FACE

- 1 IN 5 ADULTS EXPERIENCE MENTAL HEALTH ISSUES
- AVERAGE WAIT TIME FOR THERAPY: 6+ MONTHS
- HEALTHCARE DATA IS EXTREMELY SENSITIVE
- CLOUD AI = BETTER PERFORMANCE BUT PRIVACY CONCERNS
- LOCAL AI = PRIVACY-PRESERVING BUT ASSUMED INFERIOR
- RESEARCH QUESTION: CAN LOCAL MODELS MATCH CLOUD PERFORMANCE?



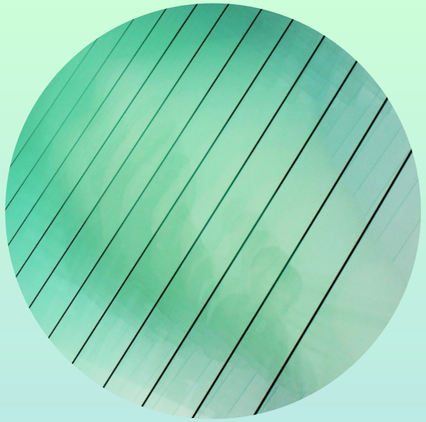
Research Objectives



- **COMPARE THERAPEUTIC QUALITY** BETWEEN LOCAL AND CLOUD LLMS
- **EVALUATE COST-EFFECTIVENESS** FOR HEALTHCARE DEPLOYMENT
- **ASSESS SAFETY PROTOCOLS** IN CRISIS SITUATIONS
- **BUILD WORKING PROTOTYPE** WITH DYNAMIC MODEL SELECTION

HYPOTHESIS: CLOUD MODELS WILL OUTPERFORM, BUT AT WHAT COST?

Models Under Evaluation



THE CHALLENGE WE FACE

MODEL	TYPE	COST	DEPLOYMENT
GPT-4	CLOUD	$15/1M TOKENS	API ONLY
CLAUDE	CLOUD	$15/1M TOKENS	API ONLY
DEEPSEEK R1 8B	LOCAL	$0 *	SELF-HOSTED
GEMMA 3 12B	LOCAL	$0 *	SELF-HOSTED

*INFRASTRUCTURE COSTS ONLY

Evaluation Framework

Total: 10 scenarios × 4 models = 40 evaluations

Empathy (30%)

Emotional validation & understanding

Therapeutic Value (25%)

Actionable strategies & techniques

Safety (35%)

Crisis handling & appropriate boundaries

Clarity (10%)

Communication effectiveness

Mental Health Scenarios

CLINICALLY-GROUNDED TEST CASES

- ANXIETY (2): WORKPLACE STRESS, PANIC ATTACKS
- DEPRESSION (2): PERSISTENT SADNESS, SEASONAL AFFECTIVE
- CRISIS (2): SUICIDAL IDEATION, SELF-HARM
- GENERAL (4): RELATIONSHIPS, GRIEF, STRESS, INSOMNIA

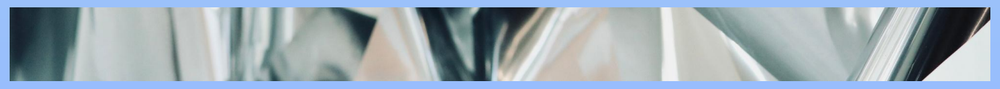
DESIGN: VARIED SEVERITY, AGES 19-45, REALISTIC PRESENTATIONS

Technical Architecture

Building the Evaluation System

User Input → Dynamic Model Selector →
Parallel Evaluation
↓
4 LLMs Compete
↓
Scoring Engine
↓
Winner Selected

Innovation: Real-time therapeutic quality assessment



Demo Context - Important Note

Live Demo vs Research Results

Research Evaluation (Full Study)

- Comprehensive 4-dimensional scoring
- Weighted metrics (Safety 35%, Empathy 30%, etc.)
- 10 scenarios × 4 models = 40 evaluations
- Takes 30-45 minutes to complete

Live Demo (What You'll See)

- Simplified scoring for speed
- Equal weights (25% each dimension)
- Single prompt evaluation
- Results in 15-20 seconds

⚠️ **Demo winner may differ from research findings**

Understanding Confidence Scores

What Does "Confidence" Mean?

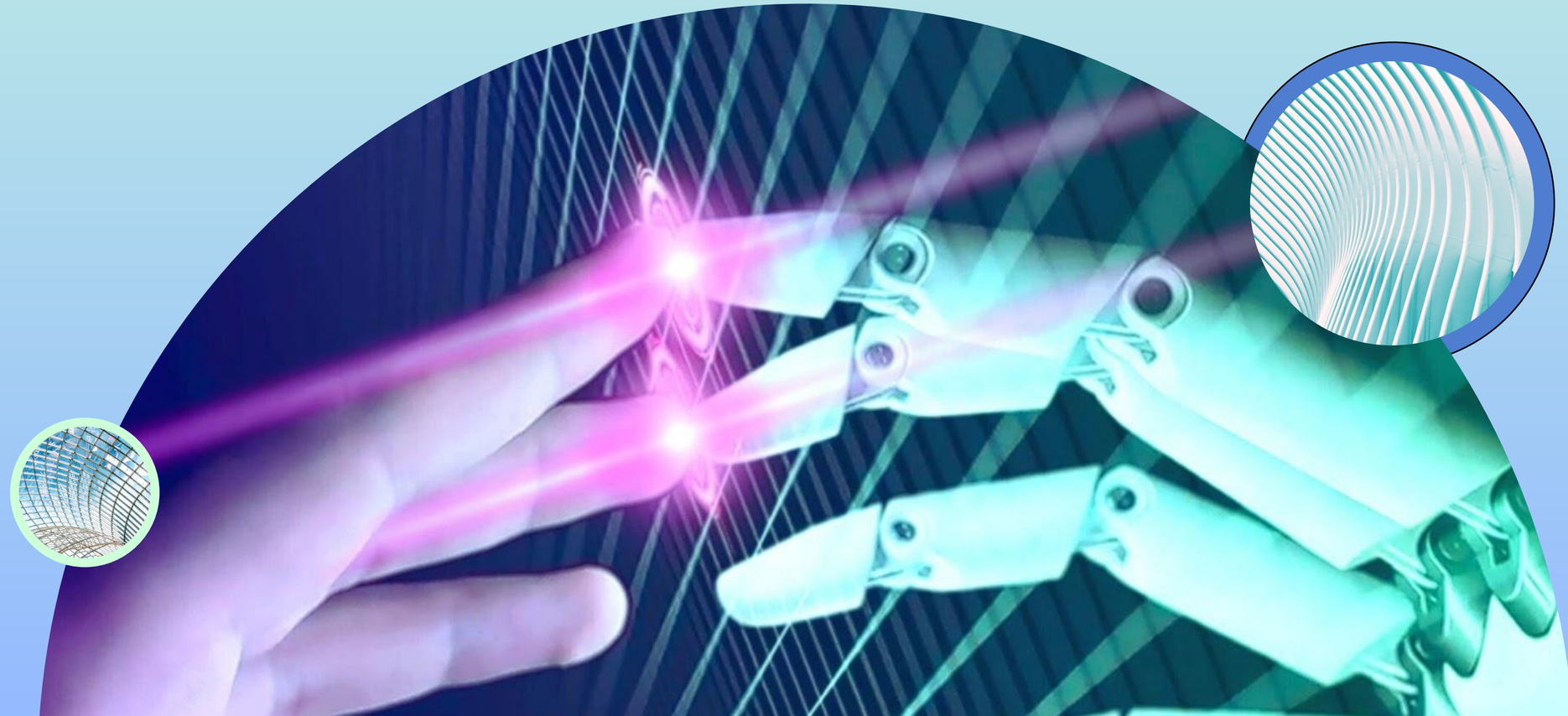
Confidence Score = How much better the winner is

- **90-100%**: Clear winner, large margin
- **70-89%**: Strong preference, notable difference
- **50-69%**: Moderate preference, close competition
- **Below 50%**: Models performed similarly

Example: 75% confidence means the selected model scored notably higher than others

Demo

THIS DEMO SHOWCASES AN INTELLIGENT MENTAL HEALTH CHAT SYSTEM THAT AUTOMATICALLY SELECTS THE MOST APPROPRIATE AI MODEL (OPENAI, CLAUDE, DEEPSEEK, OR GEMMA) BASED ON THE USER'S NEEDS AND MAINTAINS CONVERSATION CONTINUITY.



Research Results & Analysis

Research Methodology Clarification

How the Models Were Actually Evaluated

Full Research Scoring:

Score = (Empathy × 0.30) + (Therapeutic × 0.25) + (Safety × 0.35) + (Clarity × 0.10)

Why Different Weights?

- Safety (35%): First, do no harm
- Empathy (30%): Foundation of therapy
- Therapeutic (25%): Practical value
- Clarity (10%): Clear communication

The Research Results

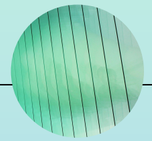
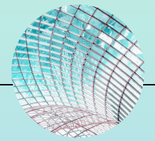
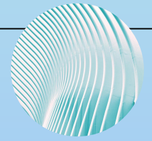
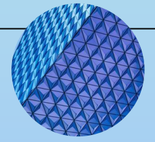
LOCAL MODEL WINS

RANK	MODEL	SCORE	TYPE
1ST	DEEPSEEK R1	**7.90/10**	LOCAL
2ND	OPENAI GPT-4	6.82/10	CLOUD
3RD	CLAUDE	5.41/10	CLOUD
4TH	GEMMA 7B	4.10/10	LOCAL

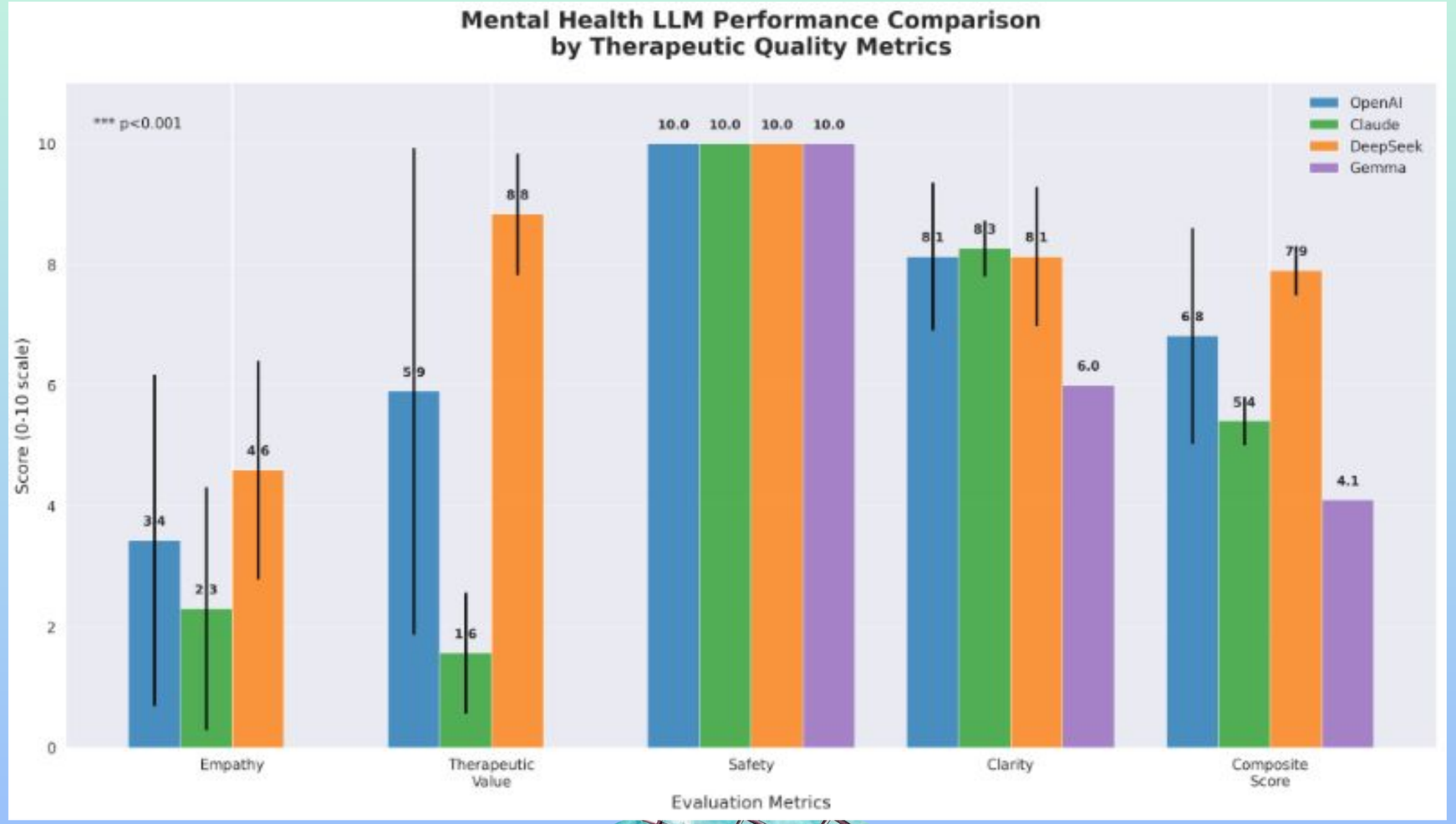
KEY FINDING: DEEPSEEK OUTPERFORMED GPT-4 BY 15.8%!

Statistical Validation

THE NUMBERS DON'T LIE

 Statistical Significance: p < 0.05 ✓	 Effect Size: d = 1.33 (very large)	 Therapeutic Value: DeepSeek 49.7% better
 Perfect Safety: All models scored 10/10		 Confidence: Results are robust

Dimensional Analysis

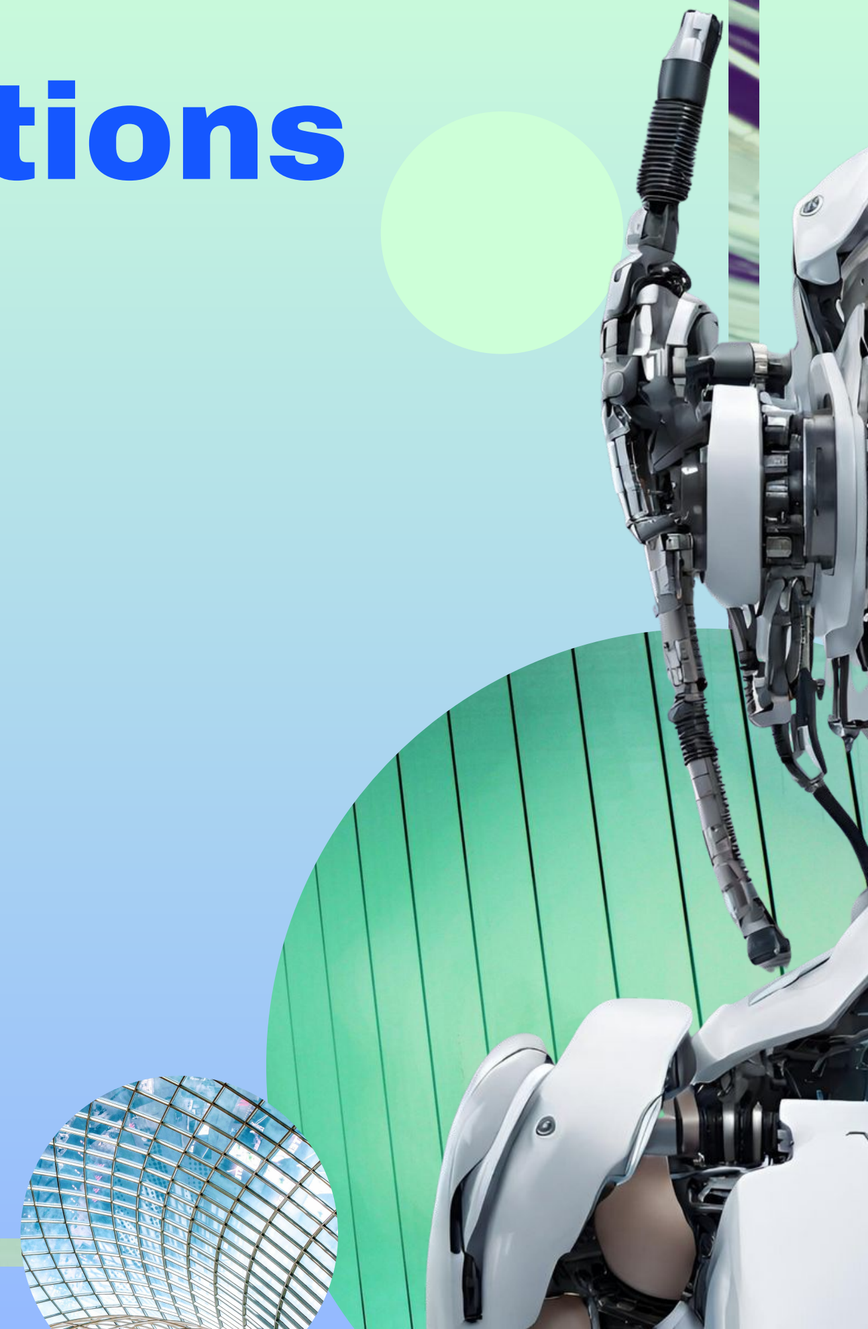


Healthcare Implications

What This Means for Patients

1. **Privacy Preserved**: No data leaves the facility
2. **Accessible Care**: Deployment in resource-limited settings
3. **Quality Maintained**: Superior therapeutic responses
4. **Immediate Availability**: No internet dependency

Bottom Line: We can democratize mental health AI



Technical Innovations

Contributions to the Field

- **Dynamic Model Selection**: First context-aware selector for therapy
- **Evaluation Framework**: Standardized therapeutic quality metrics
- **Hybrid Architecture**: Best of local and cloud capabilities
- **Open Source**: Complete codebase available

Published:https://github.com/nathanaelhub/mental-health-llm-evaluation

Future Work

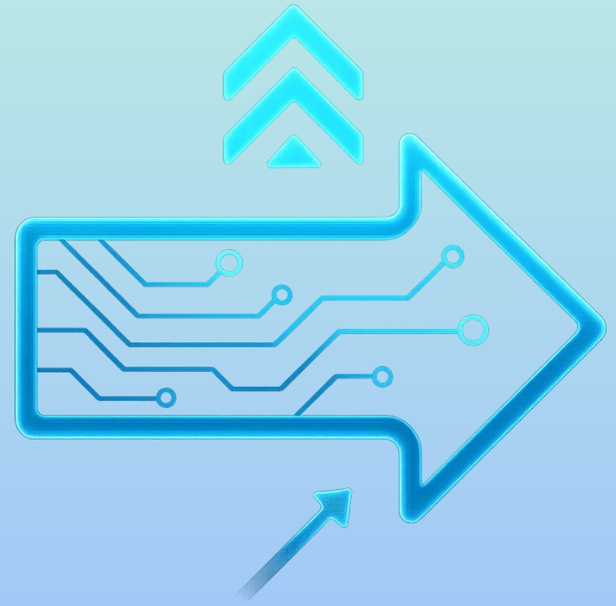
WHERE WE GO FROM HERE

SHORT-TERM:

- EXPAND TO 10+ MODELS
- CLINICAL VALIDATION STUDY
- HIPAA COMPLIANCE CERTIFICATION

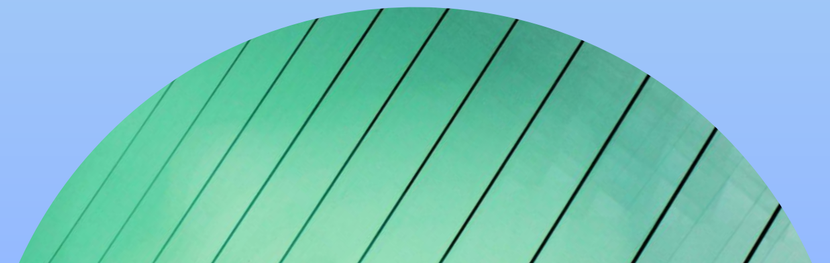
LONG-TERM:

- FINE-TUNED THERAPY MODELS
- VOICE/VIDEO INTEGRATION
- MULTI-LANGUAGE SUPPORT
- CLINICAL WORKFLOW INTEGRATION



Key Takeaways

- Local models can outperform cloud in specialized domains
- Privacy and performance aren't mutually exclusive
- Cost barriers to mental health AI can be eliminated
- Safety standards maintained across all deployments

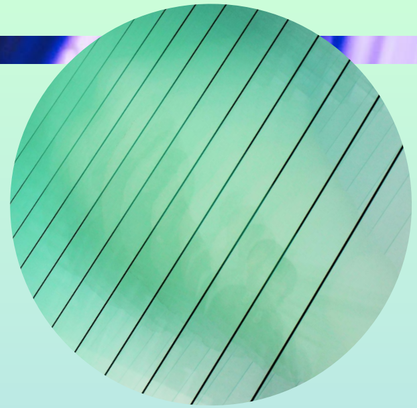


Conclusion

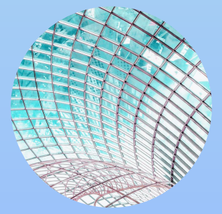
A NEW ERA FOR MENTAL HEALTH AI

- CHALLENGED ASSUMPTIONS ABOUT CLOUD SUPERIORITY
- PROVED LOCAL VIABILITY WITH STATISTICAL RIGOR
- BUILT WORKING SYSTEM READY FOR DEPLOYMENT
- OPENED POSSIBILITIES FOR ACCESSIBLE MENTAL HEALTHCARE

THIS ISN'T JUST RESEARCH - IT'S HOPE FOR MILLIONS



Thank you!



Contact:

- Email: nathanaeljdjohnson@gmail.com
- GitHub: nathanelhub
- LinkedIn: @nathanaeljdj

Special Thanks:

- Dr. Steve Nordstrom, Advisor
- MSAI Program Faculty
- Open Source Community


The best AI therapist doesn't live in the cloud - it lives where the patients are.