AI Engineer World's Fair 2024: A Summary

Generated by Bard

September 3, 2024

The End of the GPT-4 Barrier

- GPT-4's dominance challenged by new models (Gemini, Claude, LLaMA)
- Open-source models democratize access to high-performing AI
- Evaluation beyond benchmarks: "Vibes" matter (user preference)

The AI Trust Crisis

- Public concerns about data privacy and model training (Dropbox, Slack)
- Need for transparency and responsible AI development
- Anthropic's Claude 3.5 Sonnet: Training without user data

Democratizing AI with LLaMa.file

- Mozilla's open-source project: Run LLMs locally on any device
- Addresses environmental and cost concerns of GPU reliance
- Innovative techniques for CPU inference: Significant speed improvements

Building and Scaling AI Applications

- Convex: Simplifying backend engineering with reactive paradigm
- Huru: Connecting AI to live data with unified query language
- Addressing challenges: Prompt injection, "slop" (unreviewed AI content)

Code Generation and the Future of Programming

- GitHub Next: Next-edit suggestions, co-pilot workspaces
- Codium: M query for comprehensive context retrieval
- Cursor: AI-powered code editor for advanced code generation

AI Engineering: A New Frontier

- Respecting human limitations (reading speed vs. speaking speed)
- Acknowledging contingent facts and trends (cost of intelligence)
- AI engineers as utility maximizers, driving benefits for humanity