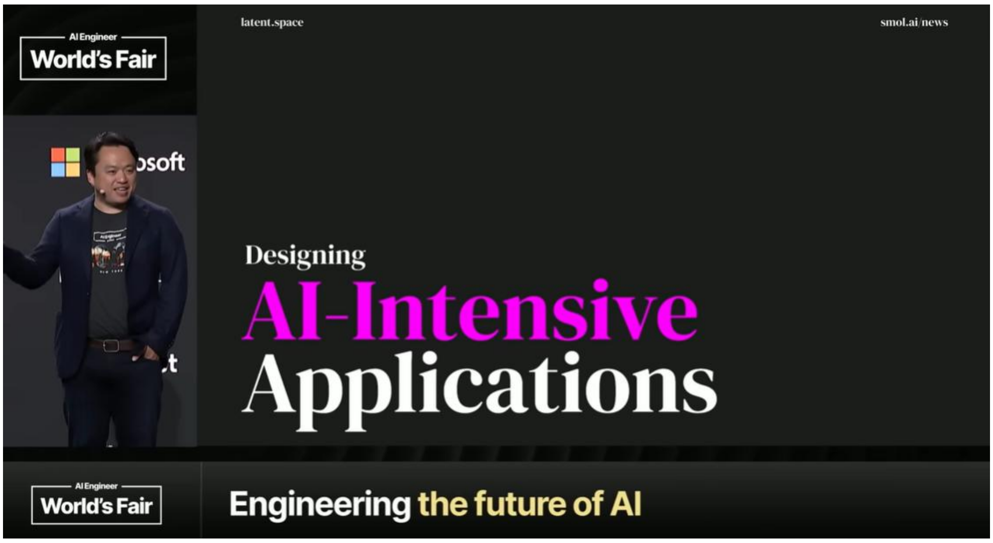
Designing AI-Intensive Applications - swyx

https://youtu.be/IHkyFhU6JEY?si=xQd98xHxAcjS0rEj

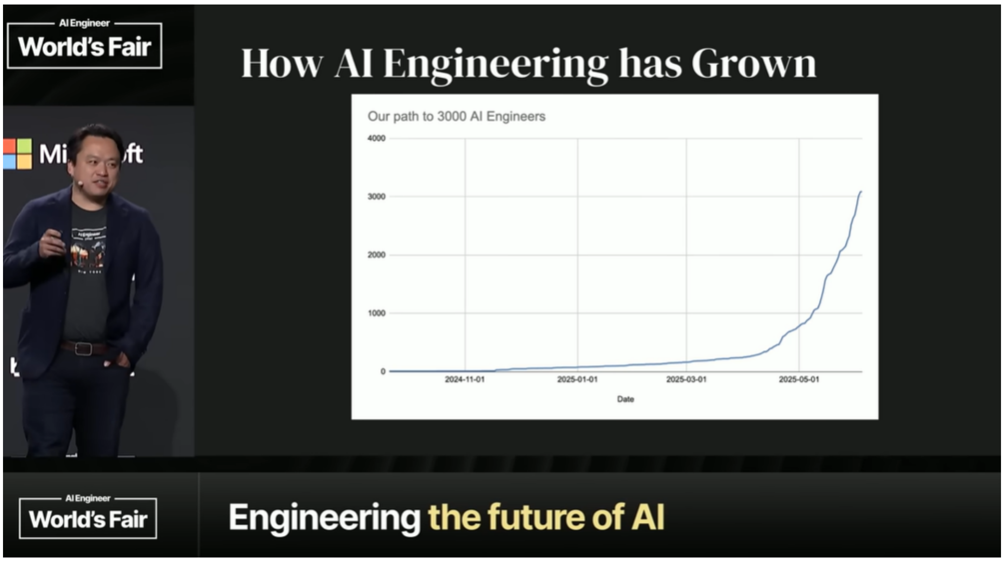
~ Prepared by Riya Kharade

➢ Introduction



The speaker welcomes everyone to the conference. He explains that this talk will cover the current state of AI engineering in 2025. He says the conference is growing and more topics are being added to help everyone learn more. The goal is to answer questions about AI, the conference, and how things are changing.

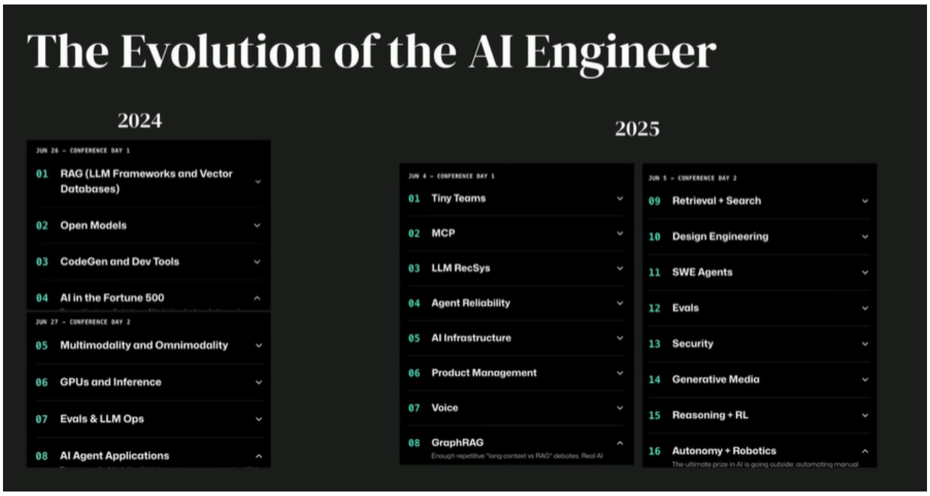
➢ Background



- Over 3,000 people registered for the conference, most at the last minute.
- The number of conference tracks has been doubled to give more value.
- The conference tries to cover all areas of AI engineering.
- Organizers use surveys and feedback to create content based on the community's needs.
- They are more technical than other events and respond quickly to attendees' requests.

➢ **All Points Discussed**

1. AI Engineering is Growing



The Evolution of the AI Engineer

2024

JUN 26 - CONFERENCE DAY 1

01 RAG (LLM Frameworks and Vector Databases)

02 Open Models

03 CodeGen and Dev Tools

04 AI in the Fortune 500

JUN 27 - CONFERENCE DAY 2

05 Multimodality and Omnimodality

06 GPUs and Inference

07 Evals & LLM Ops

08 AI Agent Applications

2025

JUN 4 - CONFERENCE DAY 1

01 Tiny Teams

02 MCP

03 LLM RecSys

04 Agent Reliability

05 AI Infrastructure

06 Product Management

07 Voice

08 GraphRAG

Enough repetitive "long context vs RAG" debates. Real AI

JUN 5 - CONFERENCE DAY 2

09 Retrieval + Search

10 Design Engineering

11 SWE Agents

12 Evals

13 Security

14 Generative Media

15 Reasoning + RL

16 Autonomy + Robotics

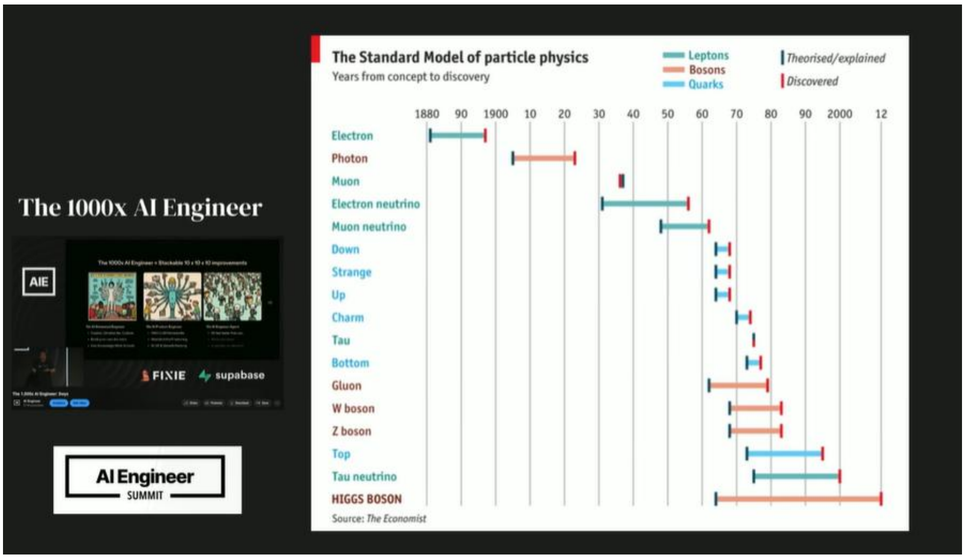
The ultimate prize in AI is going outside: automating manual

- AI engineering was once mocked and considered a small field.
- Now it is respected and seen as a high-paying career.
- It's compared to the growth of physics in the early 1900s when big discoveries were made.

2. The Need for Standard Models

- Engineering fields use standard frameworks like MVC, ETL, MapReduce.
- The speaker asks: What will be the standard model for AI engineering?
- Having models helps organize ideas and make sense of complex tasks.

2. Examples of AI Standard Models



The 1000x AI Engineer

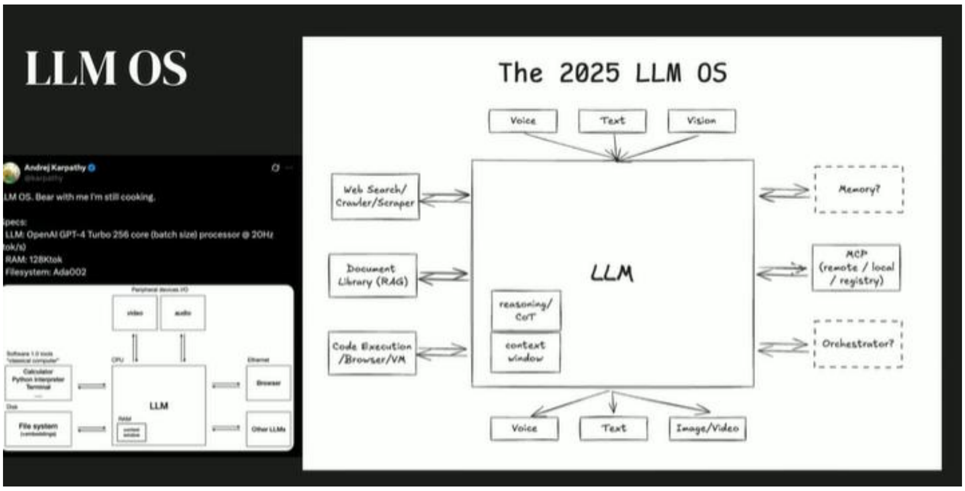
The 1000x AI Engineer = Stackable 10 x 10 x 10 improvements

The Standard Model of particle physics

Years from concept to discovery

Source: *The Economist*

- **LM OS (Language Model Operating System):** Helps in building systems that use language tools.



LLM OS

Andrej Karpathy

@karpathy

LLM OS. Bear with me I'm still cooking.

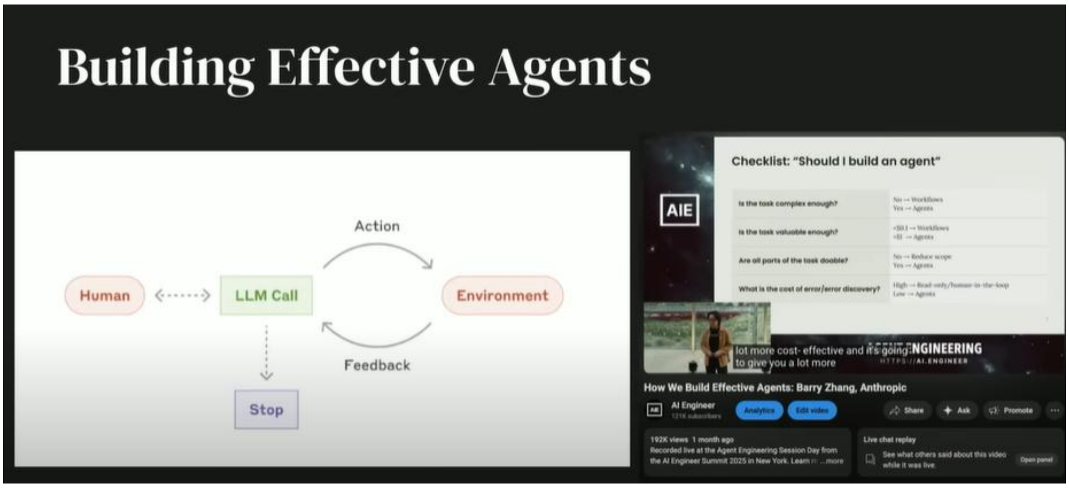
Specs:
- LLM: OpenAI GPT-4 Turbo 256 core (batch size) processor @ 20Hz (tok/s)
- RAM: 128Ktok
- Filesystem: Ada002

The 2025 LLM OS

- **MCP (Multipurpose Conversational Protocol):** Connects users to AI tools like chatbots and voice assistants.

- **AI SDLC (Software Development Life Cycle):** Guides how AI products are built, tested, and maintained.

3. Building Effective Agents



Building Effective Agents

Checklist: "Should I build an agent"

Is the task complex enough?	No → Workflows Yes → Agents
Is the task valuable enough?	<$0.1 → Workflows >$1 → Agents
Are all parts of the task doable?	No → Reduce scope Yes → Agents
What is the cost of error/error discovery?	High → Read-only/human-in-the-loop Low → Agents

lot more cost- effective and it's going to give you a lot more

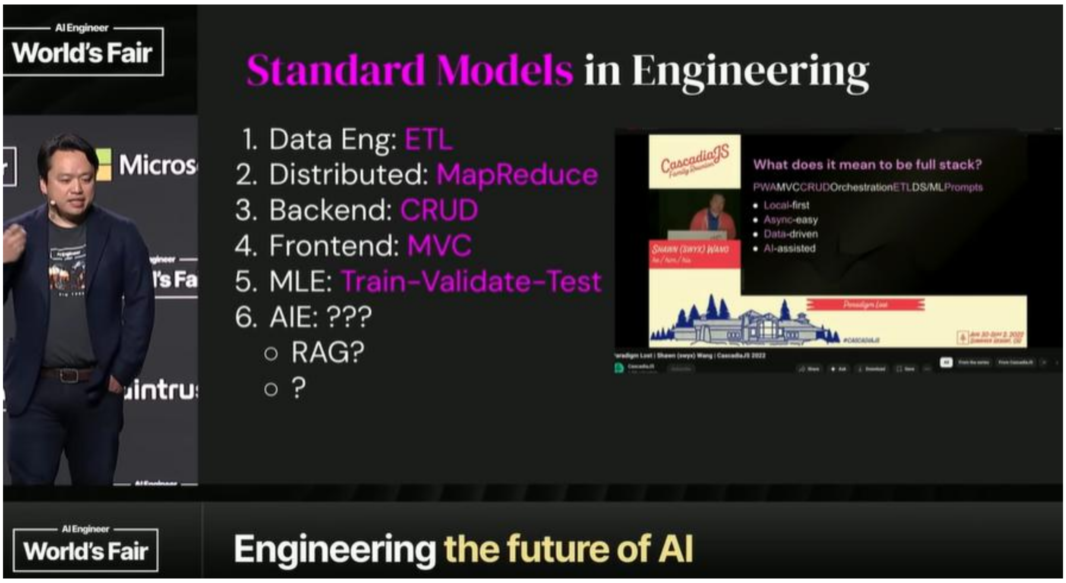
How We Build Effective Agents: Barry Zhang, Anthropic

AI Engineer
121K subscribers

192K views 1 month ago
Recorded live at the Agent Engineering Session Day from the AI Engineer Summit 2025 in New York. Learn ...more

Live chat replay
See what others said about this video while it was live.

4. Standard Models in Engineering



Standard Models in Engineering

1. Data Eng: ETL
2. Distributed: MapReduce
3. Backend: CRUD
4. Frontend: MVC
5. MLE: Train-Validate-Test
6. AIE: ???
   - RAG?
   - ?

What does it mean to be full stack?

PWAMVCCRUDOrchestrationETLDS/MLPrompts

- Local-first
- Async-easy
- Data-driven
- AI-assisted

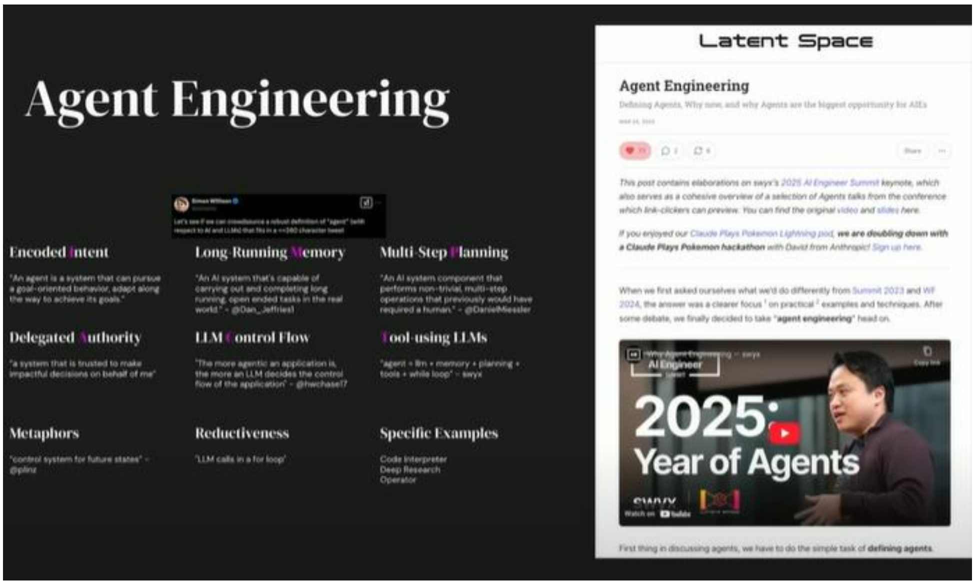
'aradigm Lost | Shawn (swyx) Wang | CascadiaJS 2022

Engineering the future of AI

5. Challenges in AI Engineering

- Early steps like training AI are becoming easier with free tools.
- Real challenges are in securing, evaluating, and scaling AI systems for businesses.

6. **Agent Engineering**



Agent Engineering

Simon Willison

Let's see if we can crowdsource a robust definition of "agent" (with respect to AI and LLMs) that fits in a <=280 character tweet

Encoded Intent

"An agent is a system that can pursue a goal-oriented behavior, adapt along the way to achieve its goals."

Long-Running Memory

"An AI system that's capable of carrying out and completing long running, open ended tasks in the real world." - @Dan_Jeffries1

Multi-Step Planning

"An AI system component that performs non-trivial, multi-step operations that previously would have required a human." - @DanielMiessler

Delegated Authority

"a system that is trusted to make impactful decisions on behalf of me"

LLM Control Flow

"The more agentic an application is, the more an LLM decides the control flow of the application" - @hwchase17

Tool-using LLMs

"agent = llm + memory + planning + tools + while loop" - swyx

Metaphors

"control system for future states" - @pliny

Reductiveness

"LLM calls in a for loop"

Specific Examples

Code Interpreter
Deep Research
Operator

Latent Space

Agent Engineering

Defining Agents, Why now, and why Agents are the biggest opportunity for AIEs

This post contains elaborations on swyx's 2025 AI Engineer Summit keynote, which also serves as a cohesive overview of a selection of Agents talks from the conference which link-clickers can preview. You can find the original video and slides here.

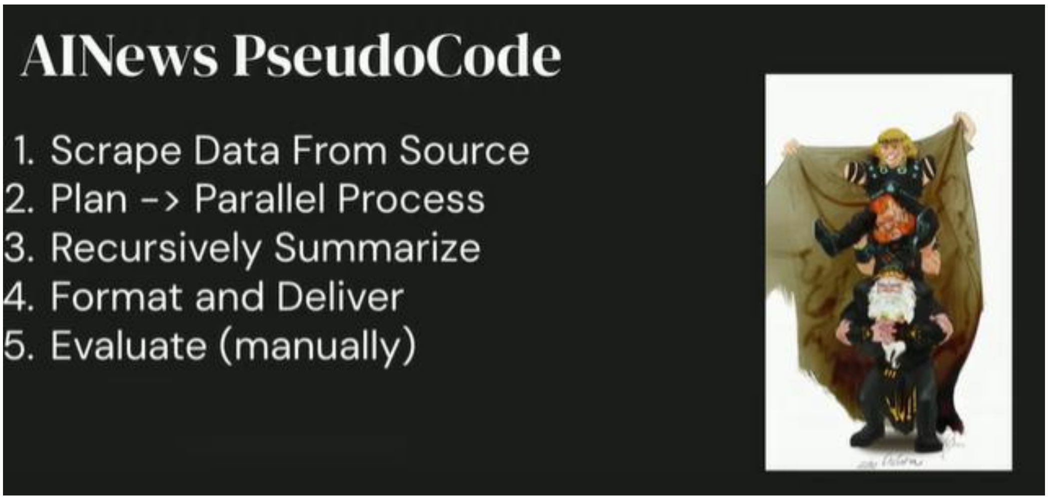
If you enjoyed our Claude Plays Pokemon Lightning pod, ***we are doubling down with a Claude Plays Pokemon hackathon*** *with David from Anthropic! Sign up here.*

When we first asked ourselves what we'd do differently from Summit 2023 and WF 2024, the answer was a clearer focus [1] on practical [2] examples and techniques. After some debate, we finally decided to take "**agent engineering**" head on.

First thing in discussing agents, we have to do the simple task of **defining agents**.

- An agent is a system that can perform tasks like planning and problem-solving.
- There is no single definition of what an agent is.
- Important parts include memory, planning, control flow, and tool use.

7. **Human Input vs AI Output**



AINews PseudoCode

1. Scrape Data From Source
2. Plan -> Parallel Process
3. Recursively Summarize
4. Format and Deliver
5. Evaluate (manually)

- Instead of arguing about definitions, it's better to think about how much input humans give and how much output AI provides.
- This helps in designing systems that are efficient and useful.

7. Mental Model of Input and Output

- Simple tools like autocomplete need small human input and give small output.
- Advanced tools need more planning and give bigger, smarter output.
- Some systems can even operate with very little human involvement.

8. AI News Example

- The speaker built an AI news tool for himself and others.
- It is not technically an "agent" but still helps people by organizing and summarizing information.
- The process it follows is called **SPAD**:
    1. **Scrape** – Collect data from websites.
    2. **Plan** – Organize how to process the data.
    3. **Analyze/Summarize** – Understand and create summaries.
    4. **Deliver** – Present the information to users.
    5. **Evaluate** – Check if the information is useful and improve it.

9. Scalable AI Systems

- Systems can make thousands of AI calls every day to serve users.
- Structured processes like SPAD make this easier and more efficient.

10. Beyond Text Outputs

- AI tools can now create structured outputs like graphs and computer code.
- This makes AI even more useful for developers and businesses.

11. Encouragement to Innovate

- Attendees are encouraged to think about new models and improve their work.
- The field is still in its early phase with many opportunities to create something new.

➢ Key Takeaways

- AI engineering is growing fast and is now respected.
- Standard models help in organizing and building better systems.
- LM OS, MCP, and AI SDLC are important frameworks to guide work.
- The ratio of human input to AI output is a better way to think about systems than debating names.
- SPAD is a helpful method for building scalable AI products.
- AI tools can now produce structured outputs like code and graphs, not just text.
- Innovation is needed, and the community must work together to build better models.

➢ Conclusion

The speaker ends with a positive message. He encourages everyone to explore and define new models for AI engineering. The goal is to create systems that are helpful, efficient, and widely used. AI is still in its early days, and there are many chances to contribute, learn, and build something meaningful.