

# Introduction

*CS 6501: Serverless AI*

*Fall 2025*

Lecture 1

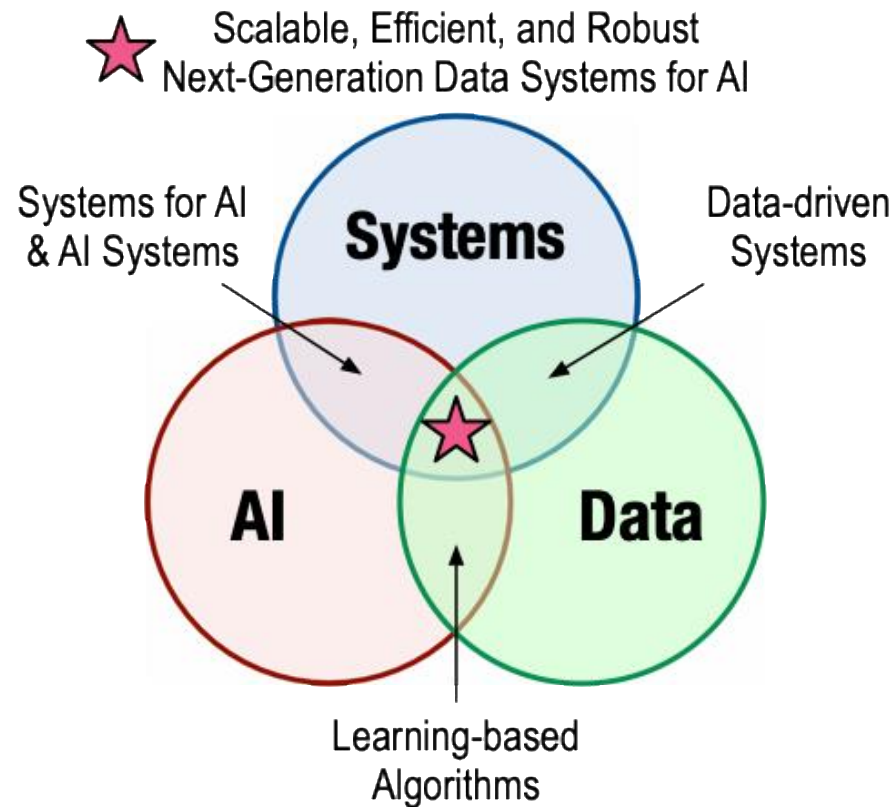
Yue Cheng



UNIVERSITY  
*of* VIRGINIA

# Introduction

- Yue Cheng
  - Associate professor of Data Science & Computer Science
  - Web: <https://tddg.github.io>
  - Email: [mrz7dp@virginia.edu](mailto:mrz7dp@virginia.edu)
- Current research:  
Designing better data systems for AI
  - (Storage) Systems for AI
  - (Affordable) Serverless AI



# Course staff and office hours

- Instructor: Yue Cheng
  - Office hours: M/W, 2:30pm-3:30pm on Zoom
- GTA:
  - TBD

# Discussion and getting help

- Discussion, questions: Ed
  - <https://edstem.org/us/dashboard>
  - Alternative place to ask questions, brainstorming, and discussing anything related to course topics
  - No anonymous posts or questions
  - Can use private posts to instructor/GTA
  - We are monitoring Ed several times a day
  - We will respond to questions in a batch manner

# Today's agenda


- Why are we studying Serverless + AI? What is this course about?
  - What is Serverless and FaaS?
  - What are serverless AI applications?
- What will you do in this course?

# What is serverless?

- Operationally
  - “No-ops” – (almost) no configuration
  - Autoscaling down to 0
  - Closer to pay-per-use (rather than pay-per-allocation)
  - Fine-grained billing
- Popular offering: serverless custom code
  - Function-as-a-Service (FaaS)



# What is FaaS?

- A **programming abstraction** that enables users to upload code, run them at (**virtually**) any scale, and **pay only for the resources used**
- First model of mostly general computing to have all those properties → **AWS Lambda** 
- Well-defined life-cycle: trigger, invocation
- Quite many limitations in **short** **small** duration, memory, communication, state  
**indirect** **stateless**

A marketing term

# Serverless industry trend



2006



2008



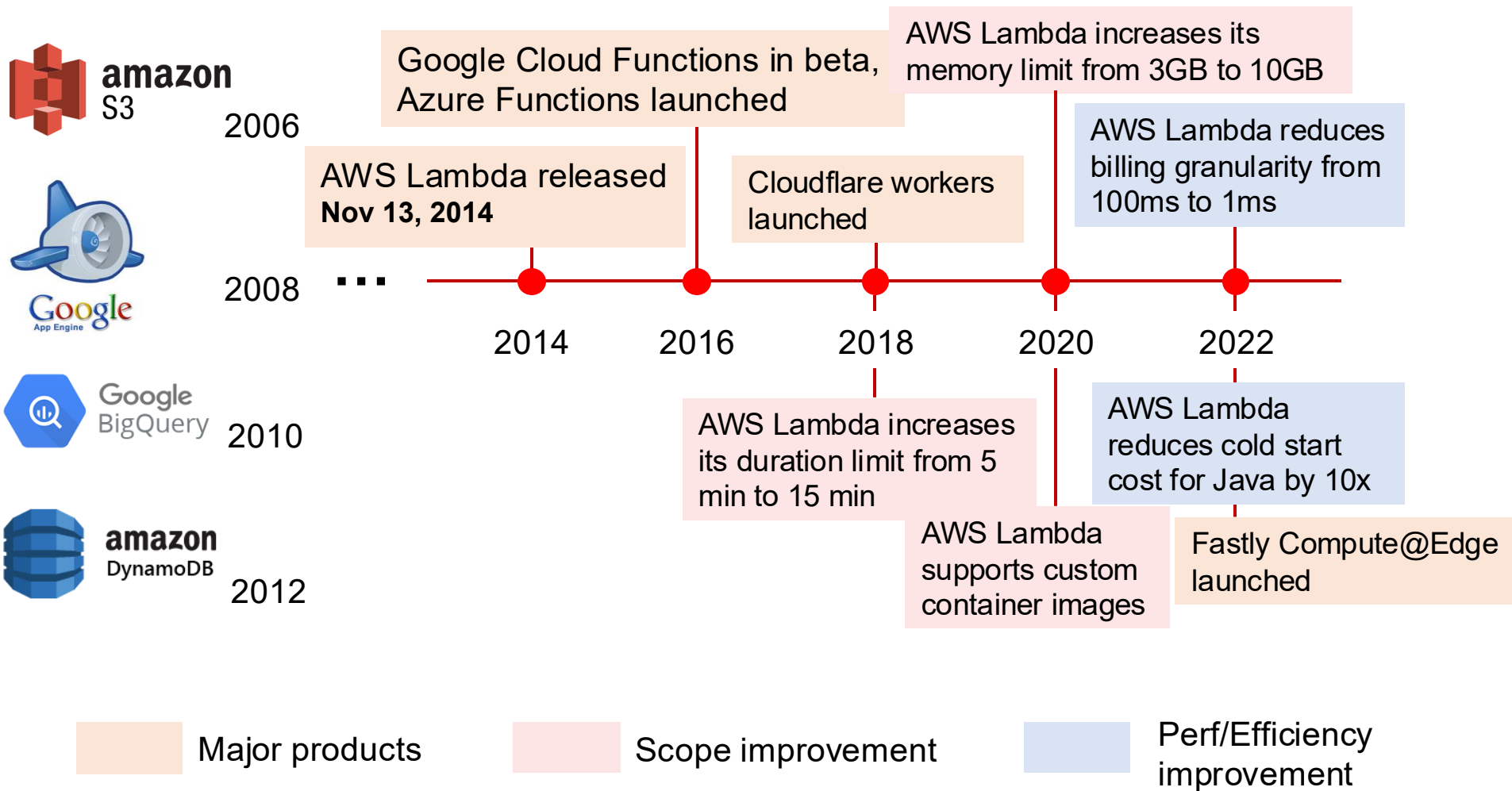
2010



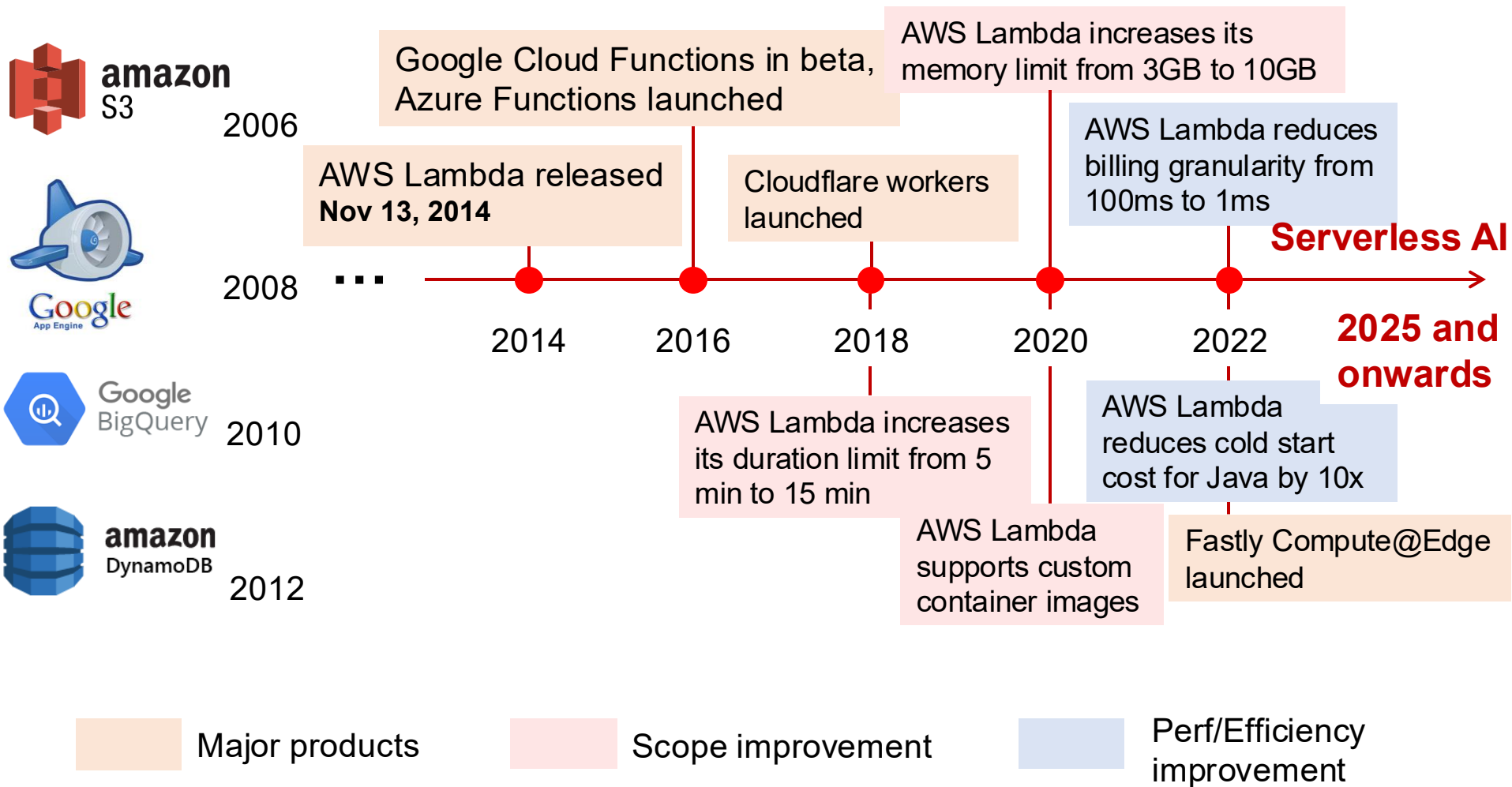
2012



# Serverless industry trend



# Serverless industry trend: 2025 and onwards ...



# When AI meets serverless

- Users don't need to manage AI hardware or infrastructure
  - but can run/use AI applications on demand (anytime)

SaaS



Gemini

MaaS



**Inference Endpoints**  
Transformers in production: solved

Workers AI



**CLOUDFLARE**



Serverless GPU FaaS  
(GPUaaS)



**Modal**



**RunPod**

# Serverless GPU demo



- Problem:
  - Jupyter/Colab Notebooks reserve GPUs even though GPUs are not utilized
  - Users need to pay for GPUs
- We built NotebookOS, a new Notebook with serverless GPUs
  - Notebooks run anywhere (your local laptop with no NVIDIA GPUs)
  - GPU cells transfer code to modal's serverless GPU for remote execution



# Course syllabus

# Big picture course goals

- Learn key works in serverless and modern AI systems
- Learn how to read, present, and evaluate a systems research paper
- Learn how to manage project and write code with AI coding tools
  - Sharpen your AI coding skill: an exciting time to do so
- Learn how to approach, discuss, and communicate about technical subject matter

# Schedule (tentative)

- Readings, project due dates, resources
- Less concrete further out

<https://tddg.github.io/cs6501-serverless-ai-fall25/>

CS6501, Fall'25

[Syllabus](#)  
[Staff](#)  
**[Schedule](#)**  
[Calendar](#)  
[Project](#)  
[Resources](#)

Q Search CS6501, Fall'25

Last updated: | [Permalink](#)

## Course Schedule

Being less concrete further out, the course scheduling is tentative and subject to changes.

### Introduction

Week 1	
08/27:	<a href="#">Course introduction</a>
Week 2	
09/01:	<a href="#">Cloud computing</a>
09/03:	<a href="#">Serverless computing</a>

# Schedule (tentative)

- Most lectures will have two paper presentations, each led by one of us (including me :-)
- Week 1-2: Cloud & serverless computing (Foundations)
- Week 3-5: FaaS platforms, workloads, cold starts
- Week 6: Stateful serverless computing (Serverless & FaaS)
- Week 7-9: Serverless applications
- Week 10-11: LLM serving and inference (Serverless AI)
- Week 12-14: Serverless AI
- Week 15-16: Final project presentation



# Readings

- Mostly papers, occasionally blog articles
  - As most topics are not directly covered by a text
- Slides/lecture notes, scribed summaries
- Optional textbooks (both are free)
  - **“Operating Systems: Three Easy Pieces (OSTEP)”** by Remzi H. Arpaci-Dusseau and Andrea C. Arpaci-Dusseau
  - **“Distributed Systems (3<sup>rd</sup> edition)”** by van Steen and Tenenbaum will supply optional alternate explanations

# Paper reviews and presentations

- Three roles: Audience, Scribes, Presenters
- **Audience:** Read both papers for that class
  - Ensure to understand key ideas, designs, findings
  - Fill the survey for that day (**5 wild cards**)
- **Scribes:** Capture in-class discussions
  - Sign up as scribe for at least 3 presentations
- **Presenters:** Prepare slides and present the paper of your choice in class
  - Presenter doesn't need to submit survey for that class
  - Send your slides to the instructor at least **3 days before the class**

# Projects

- A term-long, research / development project
  - I'll supply ideas (and codebase for some)
  - You'll build the idea into an **MVP** (minimum viable product) with **AI assisted coding**
- Week 2: Team signup (size up to two)
- Week 4: Proposal report due
- Week 8: Midterm project reflection
- Week 9: Checkpoint report due
- Week 15-16: Project presentation and everything due
- My experience of working with AI coding tools
  - Started earlier this summer, tried many tools
  - Actively driving three projects in parallel

# Me before AI coding is a thing

# Summer experiments supercharged by AI tools

Teaching

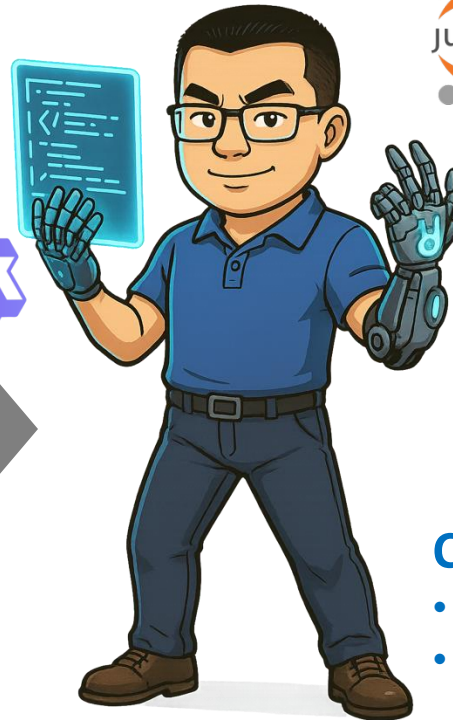
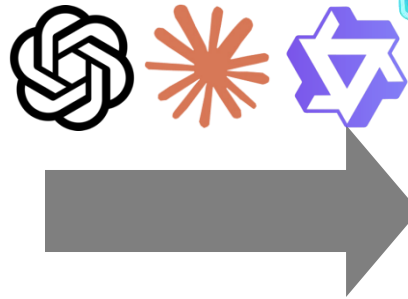
Writing  
papers

Managing  
students

Writing  
proposals

Endless  
meetings

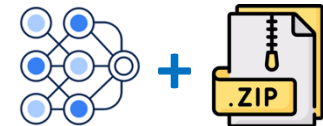
Not enough time for coding...



**Serverless GPU notebooks**  
derived from NotebookOS



**A better compressor**  
atop ZipLLM



**Customized Qwen CLI**

- Reflective learning
- Workload tracing



[Experience so far] AI coding:

- Excellent for rapid prototyping & hypothesis validation
- But gets confused easily – requires smart user management

# By the end of the semester...

Hopefully, you will have built some sophisticated, functional **MVP** that addresses an important **real-world painpoint**

# Grading

- Paper reviews (10% total)
- Class participation (10%)
- Presentation (30%)
- Project (50%)

# Resources and TODOs

- Sign-ups
  - Sign up for **Ed** (enrollment email sent)
  - Sign up for **paper presentations & scribing**
  - Redeem **Google Cloud edu credit** (\$50 each)
  - Install **Qwen Code CLI** customized for this class
    - Interfaces Qwen 3 Coder Plus (reasonably powerful LLM)
    - Offers 2,000 free requests per day for free-tier users
    - Provides dashboard visualization for historical prompts
- Next week
  - Fundamentals of cloud & serverless computing

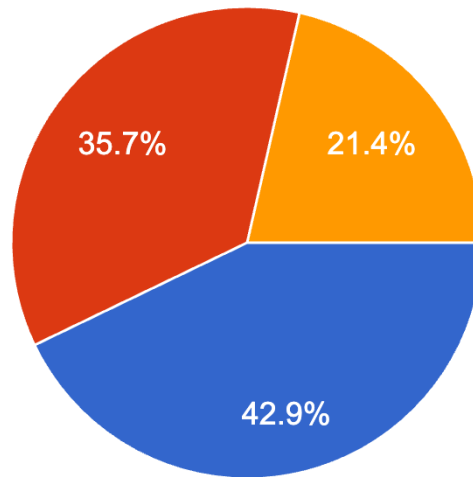
# Background survey



# Background survey

How much experience do you have in critically reading and evaluating systems research papers?

14 responses

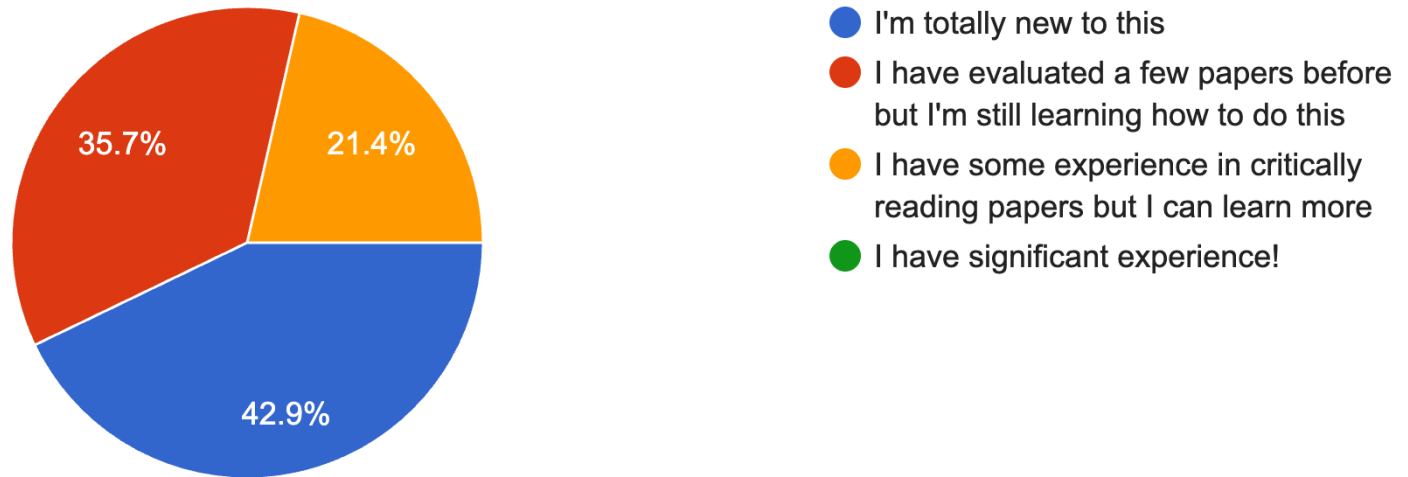


- I'm totally new to this
- I have evaluated a few papers before but I'm still learning how to do this
- I have some experience in critically reading papers but I can learn more
- I have significant experience!

# Background survey

How much experience do you have in critically reading and evaluating systems research papers?

14 responses

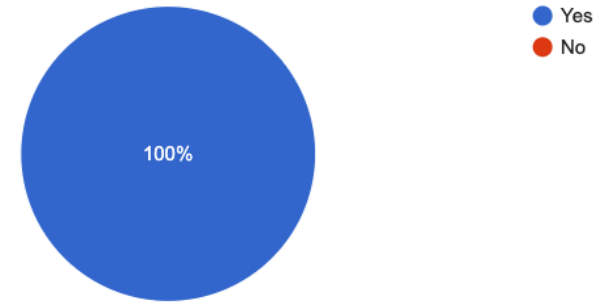


- Best way to test if you understand a technique is by **teaching** it
- Best way to understand a technique is by **doing** it
- **AI provides effective tools for smoothing the steep learning curve**
- **Bonus credit: Reproduce any result or make a demo from your presented paper and earn +2%**

# Background survey

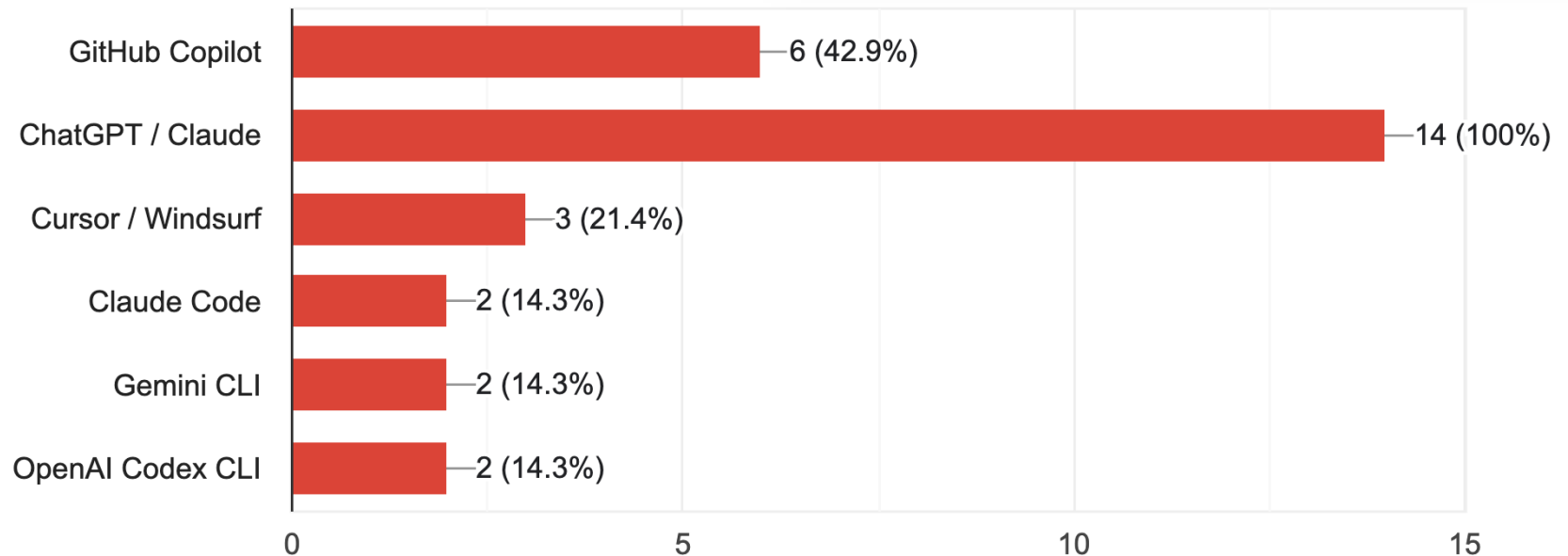
Have you ever used AI-assisted coding tools (e.g., GitHub Copilot, ChatGPT, Claude, Cursor, Windsurf, or CLIs like Claude Code, Gemini CLI, Codex)?

14 responses



If yes, which tools have you used?

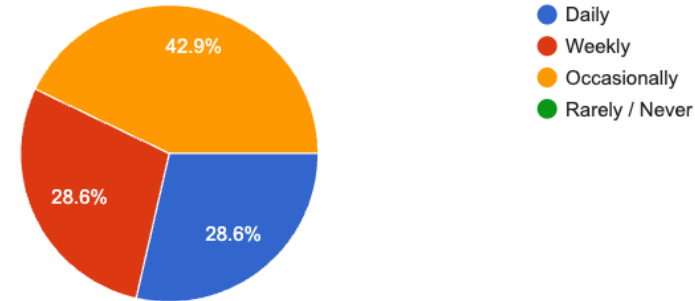
14 responses



# Background survey

How often do you currently use AI coding tools?

14 responses



For which tasks do you primarily use AI coding tools?

14 responses

