

September 18 - Outline

- 1 Announcements
- 2 TA Office Hours
- 3 Course Website Updates
- 4 Presentation Evaluation
- 5 Course Project
- 6 SWE-Marathon
- 7 The task: a Slack clone
- 8 Milestone 0

Announcements

- Please make sure you are enrolled in **Ed Discussion**. That is where all course announcements go out.
 - ▶ Every announcement is also sent to your email.
- Several announcements and reminders to sign up for presentations have been posted, with the Excel spreadsheet linked. Thank you to those who already volunteered!
- If you did not sign up, or joined the course late, you were manually and randomly assigned to one of the remaining presentation times.
- Presentation times will be **locked after the add/drop deadline**.
- If there is a day you know you cannot present, tell us **by today**. After that, if something comes up, you will have to coordinate a swap with other presenters yourselves.

TA Office Hours

- TA office hours have been decided.
- **Ira OH:**
 - ▶ Virtual, Google Meet slots
 - ▶ Mondays and Fridays, multiple slots throughout the day
- **Alex OH:**
 - ▶ Virtual, Google Meet link
 - ▶ Thursdays and Fridays, 6–8 pm
- Check the Ed Discussion announcement for exact links.

Course Website Updates

- Updated regularly. Each session lists the papers plus **supplementary materials**.
- **Sample Presenter Schedule:** how to split a paper between two presenters.
 - ▶ Student 1: intro, methods, evaluation. Student 2: related work, critique, discussion.
 - ▶ For papers with a non-standard structure (like today's survey paper) we may post a suggested split. Suggestions only.
- **Two sample presentations:** one for intro + methods + evaluation, one for related work + critique + discussion questions.
- Papers for the next session are posted **by Friday**, so everyone gets the same prep time.

Presentation Evaluation

Your presentation evaluation:

- **Instructor score (80%)**: considers your individual contribution
- **Group score (20%)**: scored by the class
- At the beginning of class, both the instructor and students will receive a form link.
 - ▶ Submit it during class, once the presenters and discussion have wrapped up. Forms submitted hours after class will not be counted.
- We will always aim to finish the presentations within the allotted time, so plan your timing accordingly.
- **Attendance is mandatory.**

Course Project

- As mentioned in the first class, the project runs in **two phases**:
 - ▶ **Phase 1: the building phase**
 - ▶ **Phase 2: the adversarial phase**
- Today: background on the task, and the motivation for Phase 1.

- Your task is based on a recent benchmark, **SWE-Marathon**: a benchmark of **long-horizon** software engineering tasks.
- 20 tasks that take humans a long time to develop, but which agents are expected to complete in a much shorter time.
 - ▶ A clone of Excel. A clone of Slack. A C compiler in Rust.
 - ▶ If it takes a human 60 hours, the agent gets 4.
- Why it exists: most benchmarks test short tasks. A single pull request, a small ticket, or at most something that takes a human 3 to 4 hours.

SWE-Marathon vs. existing benchmarks

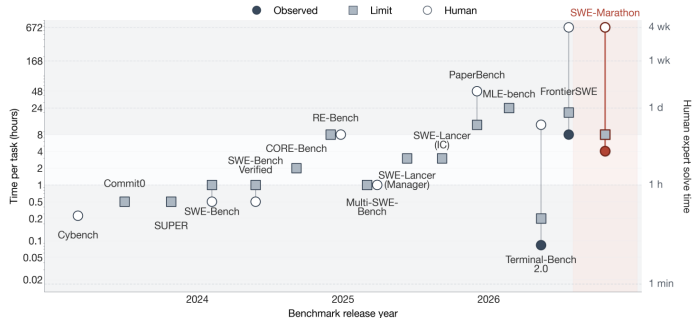
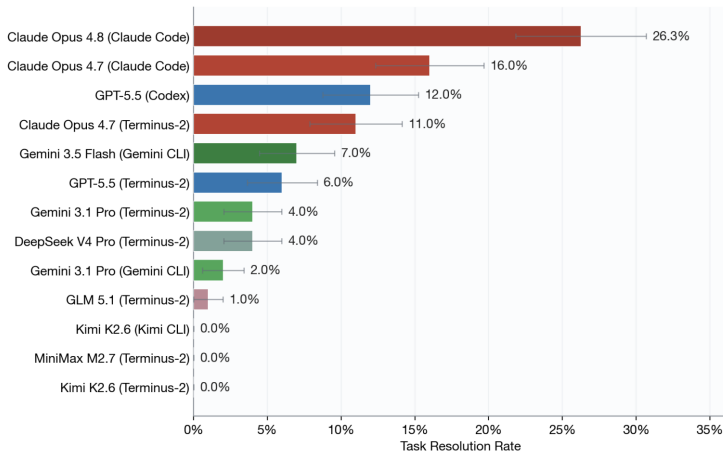


Figure 1: SWE-Marathon compared to existing software-engineering and agentic benchmarks. SWE-Marathon tasks average 27.2M total tokens per rollout with a right tail reaching 877M tokens.

- x-axis: benchmark release year
- y-axis: time per task, in hours
- **Grey square:** the time budget the benchmark gives an agent
- **White circle:** how long a human expert takes
- **Dark circle:** how long agents actually take
- A tall vertical line means a big gap between those numbers
- Left: older benchmarks like SWE-Bench. Right, in red: SWE-Marathon

How do agents do?



Frontier agents, even paired with frontier LLMs, solve **fewer than 30%** of the tasks.

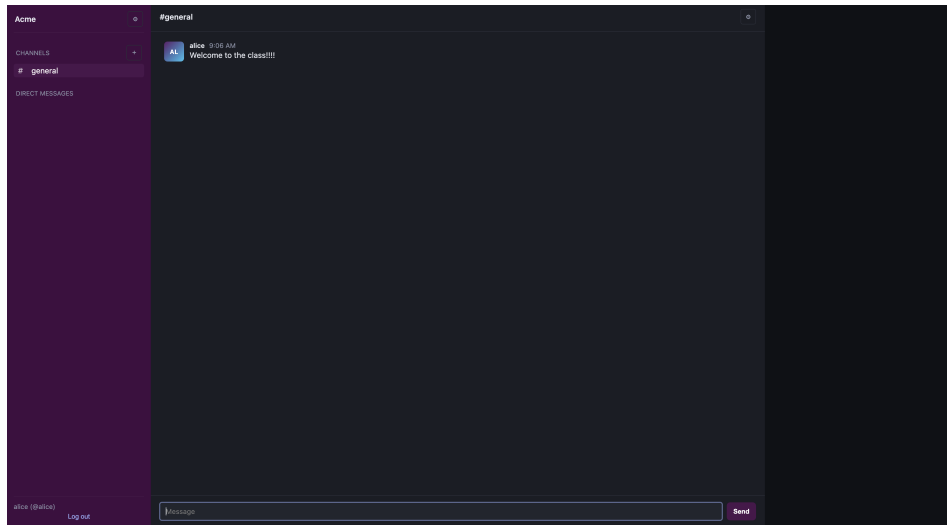
What is a SWE-Marathon task?

- A **unique executable environment**
- A **test suite**
- A **reference solution**, which, if assessed on that test suite, gets a perfect score (a reward of 1)

Our task

A **Slack clone**.

The Slack clone: reference solution



The benchmark's own reference solution, running locally.

Your role

- You will **not** build the Slack clone. The **agent** builds it.
- You design what the agent has access to, such as **specifications** and **tools**, so that the agent gets better at the task.
- Some conditions are predefined:
 - ▶ The agent that builds the Slack clone is fixed (e.g., mini-swe-agent).
 - ▶ The base LLM it uses is fixed (e.g., DeepSeek V4 Pro or DeepSeek Flash).

What the agent sees

- `instruction.md`: a very detailed description of what to build, about 5,000 words.
 - ▶ You will need to read through it and apply some domain knowledge.
- An **empty** `/app` directory, where the code will live.
- A **timer script** (`timer.sh`).
 - ▶ The agent runs for a set amount of time (e.g., 4 hours), then submits whatever it has.

Then grading starts

The grader checks, among other things:

- Does the app start?
- Did it copy someone else's code, or a publicly available solution?
- Does the API work?
- Does the web page work?

Example: grading the frontend

- The grader opens the agent's site in a **real browser**, driven by a script. No human involved.
- It clicks through: sign up, make a channel, post a message, edit it, reply in a thread, add an emoji reaction, reload, and check it is all still there.
- Every step works: this part of the test suite **passes**.
- The sign-up button does nothing: it **fails**.

Milestone 0

- **Your first task:** launch one **end-to-end run** with the setup just described (the fixed agent and model, `instruction.md`, empty `/app`), so the agent produces some app.
 - ▶ This is a **baseline**. You will not have the test suite; the app only has to satisfy a “healthy node” precondition, meaning it starts and responds.
- Milestone 0 is **individual**, to get everyone up to speed and familiar with the codebase.
- **Milestone 1 and onward:** iteratively developing a component of the system.
- **Credits:** we anticipate support from an API provider (e.g., Amazon Bedrock), with more than enough per student.
- We expect to have your credit access and GitHub access set up in the next week or so, so you are ready to tackle Milestone 0.