Collaborative Language Model Agents

Applications in AI-Assisted Software Development

Valerie Chen

Stanford University

valchen@stanford.edu

valeriechen.github.io

Models are not standalone







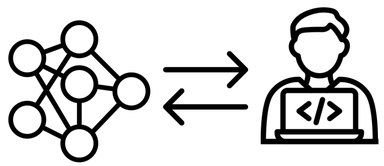
Models participate in collaborative work







How do we build LM agents for collaborative workflows?



Development



Base Model

Post-training

Evaluation

What should we do differently?

Move from a development-centered workflow

Development

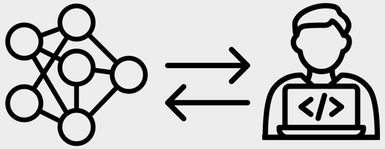


Base Model

Post-training

Evaluation

Deployment



Interactive Eval

Feedback

Interface

What should we do differently?

Move from a development-centered workflow

To one that considers deployment settings

Development

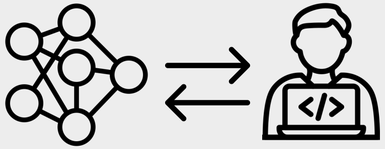


Base Model

Post-training

Evaluation

Deployment



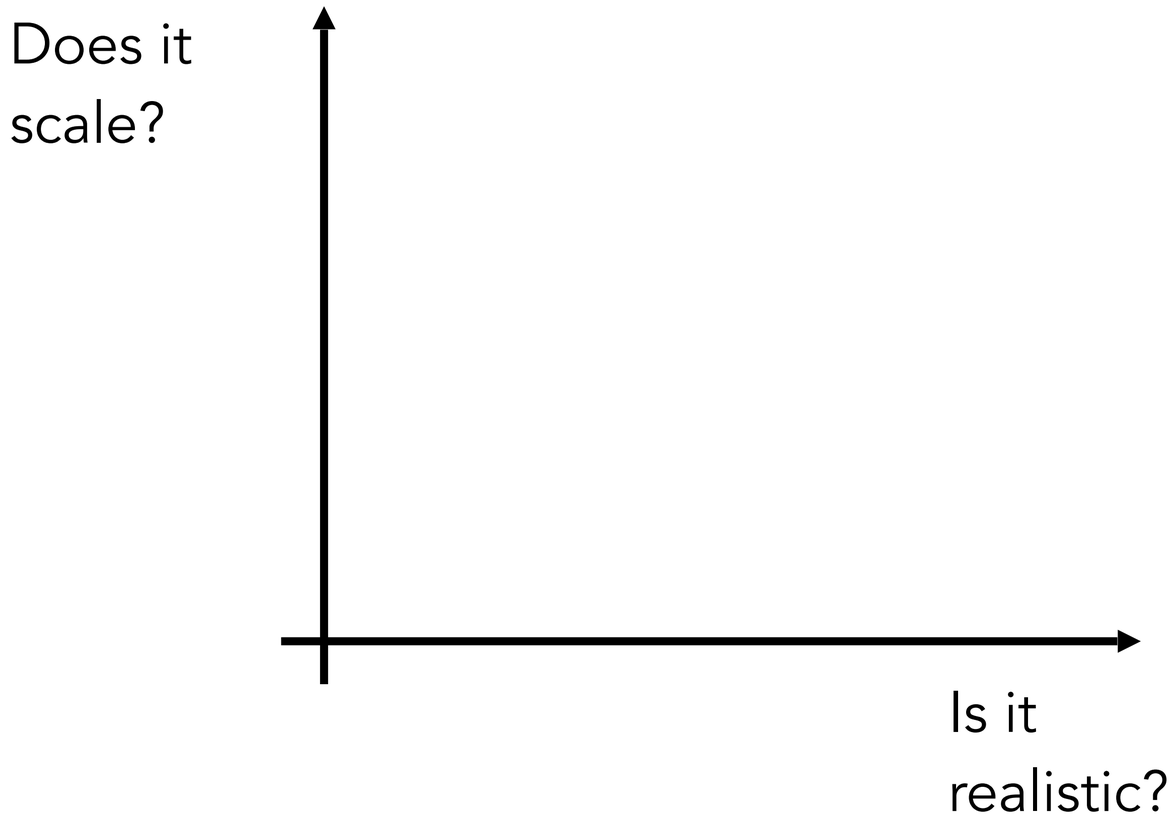
Interactive Eval

Feedback

Interface

Scientific Challenges

Measure *model performance in collaborative contexts*



Development

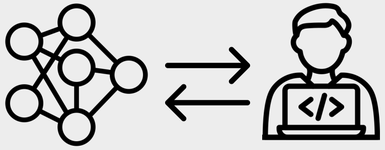


Base Model

Post-training

Static Evaluation

Deployment



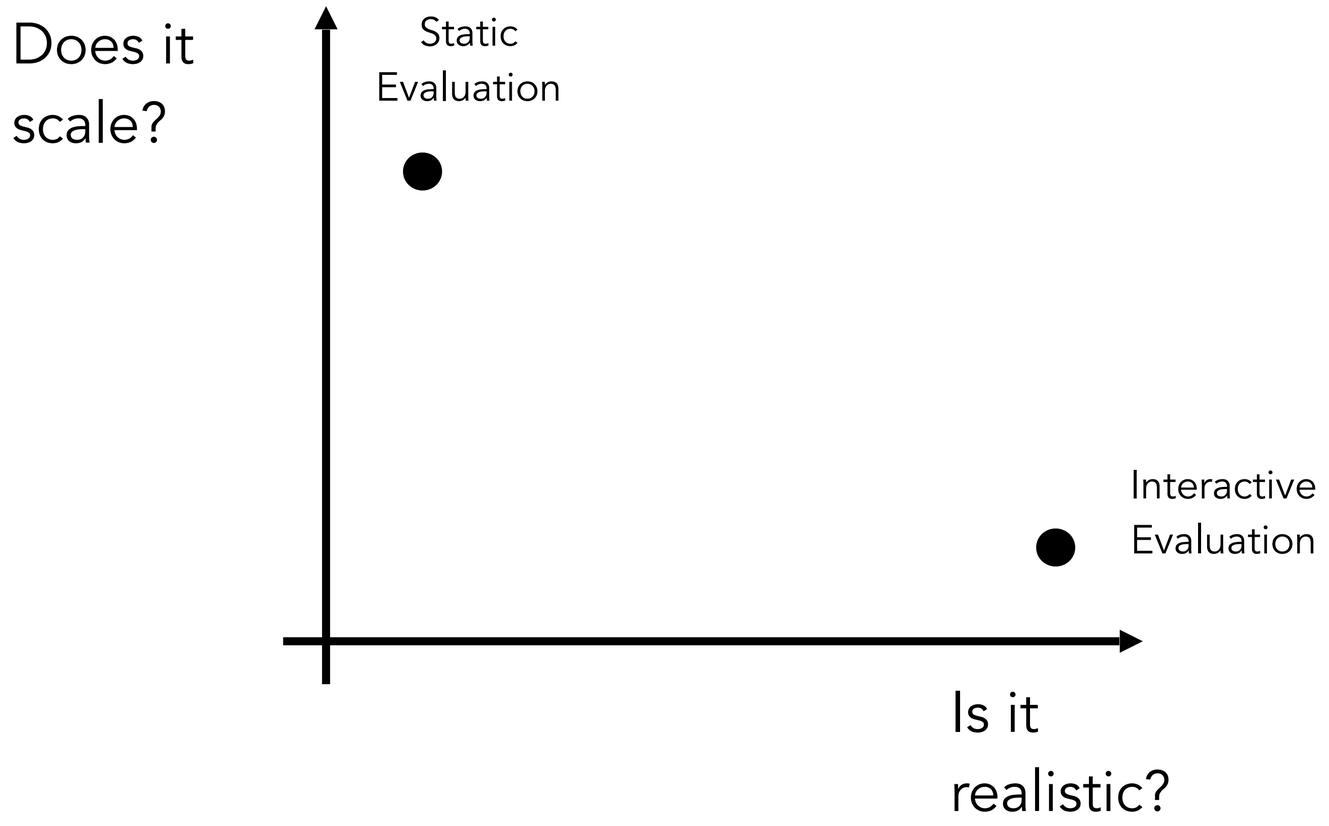
Interactive Eval

Feedback

Interface

Scientific Challenges

Measure *model performance in collaborative contexts*



Development

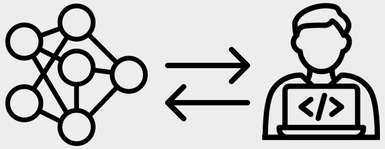


Base Model

Post-training

Static Evaluation

Deployment



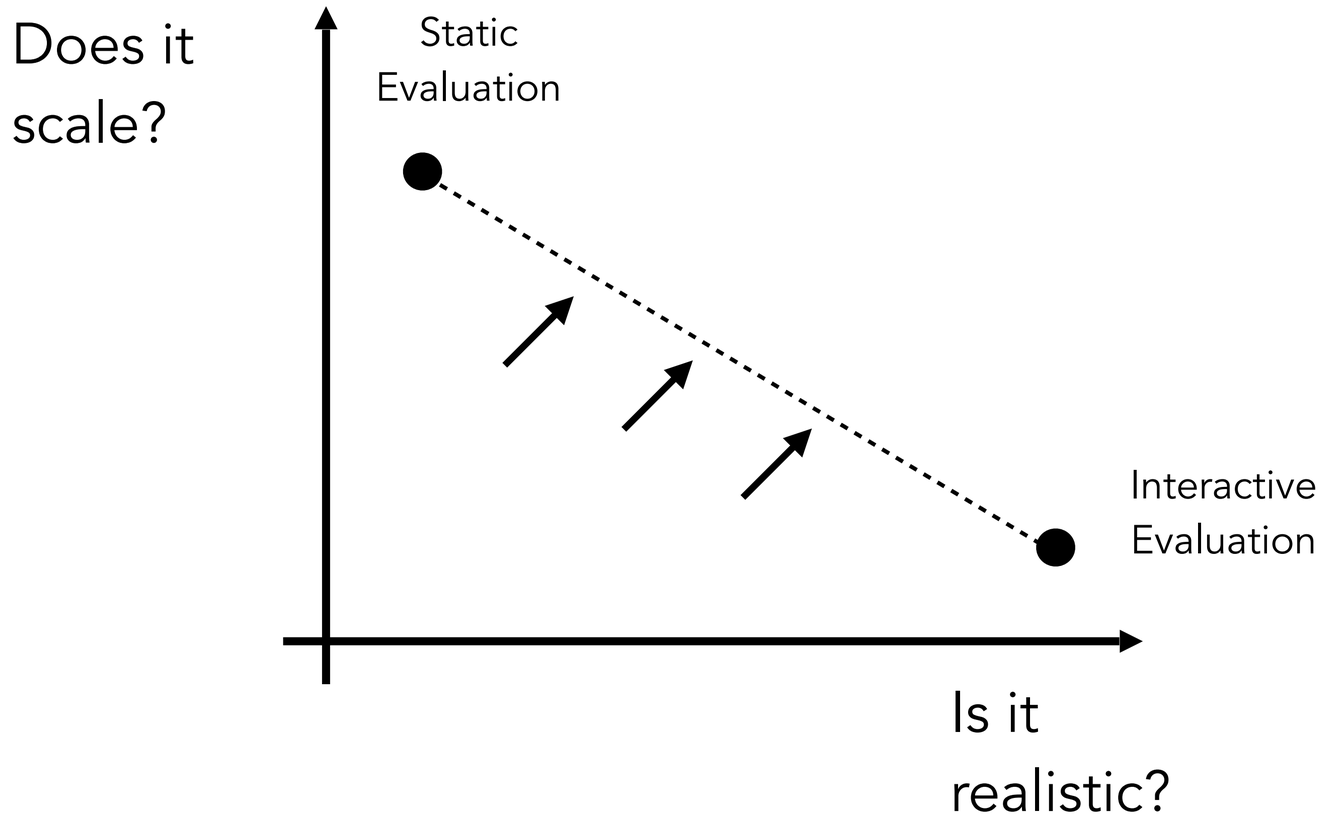
Interactive Eval

Feedback

Interface

Scientific Challenges

Measure *model performance in collaborative contexts*



Scientific Challenges

Learn from interaction in collaborative contexts

Development

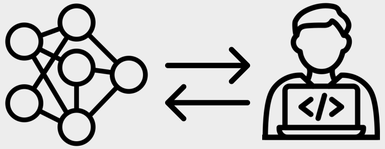


Base Model

Post-training

Evaluation

Deployment



Interactive Eval

Feedback

Interface

Development

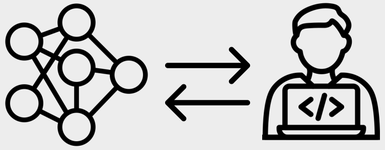


Base Model

Post-training

Evaluation

Deployment



Interactive Eval

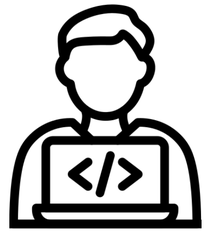
Feedback

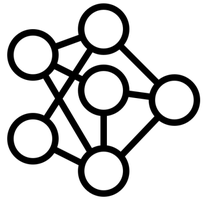
Interface

Scientific Challenges

Learn from interaction in collaborative contexts

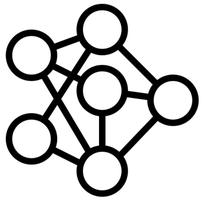
Build me a website to keep track of faculty I will meet…





Love this idea—this will make prep + follow-ups *so* much saner. Let's set..





That sounds like an exciting (and busy) journey ahead! Moving..



Ouyang et al. NeurIPS, 2022

Development

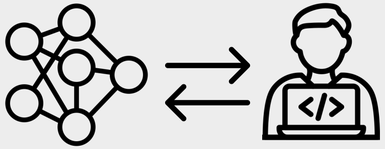


Base Model

Post-training

Evaluation

Deployment



Interactive Eval

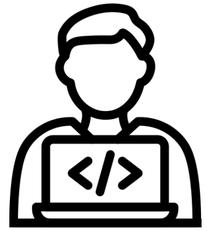
Feedback

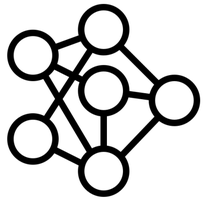
Interface

Scientific Challenges

Learn from interaction in collaborative contexts

Build me a website to keep track of faculty I will meet…

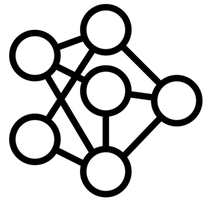




Read /workspace

Create task list

Edit index.html



Ask clarifying q

Is this the website what I envisioned?

…back and forth…

Development

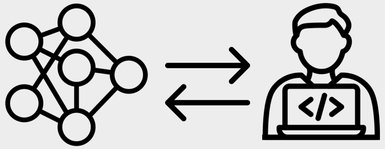


Base Model

Post-training

Static Evaluation

Deployment



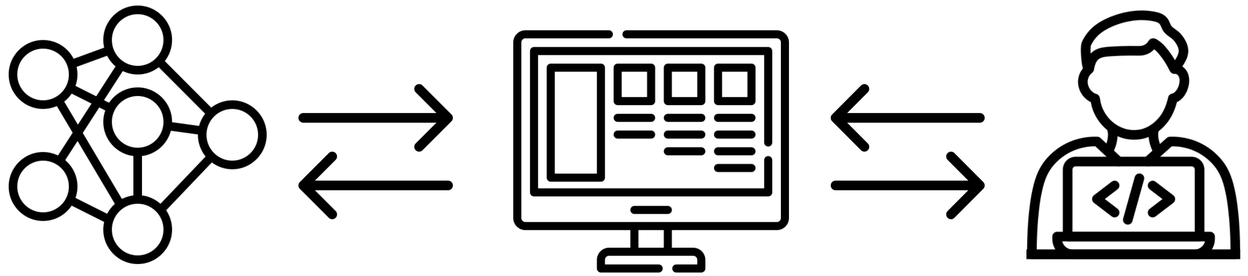
Interactive Eval

Feedback

Interface

Scientific Challenges

Optimize interfaces for sustainable interactions



↓ Prompting Effort

↓ Verification Effort

Development

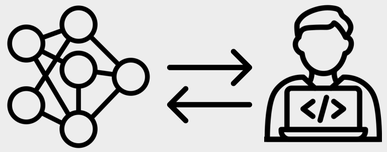


Base Model

Post-training

Static Evaluation

Deployment



Interactive Eval

Feedback

Interface

Building Collaborative LLM Agents

Measure model performance in collaborative contexts

Learn from interaction in collaborative contexts

Optimize interfaces for sustainable interactions

Development

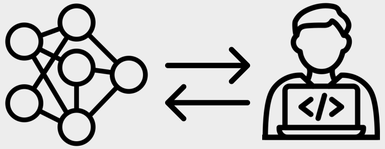


Base Model

Post-training

Static Evaluation

Deployment



Interactive Eval

Feedback

Interface

Building Collaborative LLM Agents

Measure model performance in collaborative contexts

[ICML'25] Chi* and **Chen*** et al. Copilot Arena
[ICLR'26] Chi*, **Chen***, Shar* et al. EDIT-Bench

Learn from interaction in collaborative contexts

[ICML'26] Chen et al. PULSE.
[COLM'26] Wang et al. [incl **Chen**]. Critic Rubrics.

Optimize interfaces for sustainable interactions

[CHI'25] Chen et al. Proactive AI.
[IUI'26] Kuo et al. [incl **Chen**]. ProAIDE.

Development

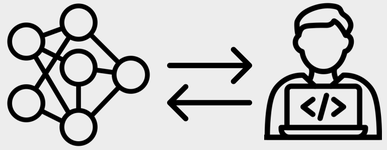


Base Model

Post-training

Static Evaluation

Deployment



Interactive Eval

Feedback

Interface

Building Collaborative LLM Agents

Measure model performance in collaborative contexts

[ICML'25] Chi* and **Chen*** et al. Copilot Arena
[ICLR'26] Chi*, **Chen***, Shar* et al. EDIT-Bench

Learn from interaction in collaborative contexts

[ICML'26] Chen et al. PULSE.
[COLM'26] Wang et al. [incl **Chen**]. Critic Rubrics.

OpenHands

Optimize interfaces for sustainable interactions

[CHI'25] Chen et al. Proactive AI.
[IUI'26] Kuo et al. [incl **Chen**]. ProAIDE.

JET BRAINS

Development

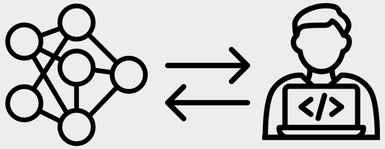


Base Model

Post-training

Static Evaluation

Deployment



Interactive Eval

Feedback

Interface

Building Collaborative LLM Agents

Measure model performance in collaborative contexts

Learn from interaction in collaborative contexts

Optimize interfaces for sustainable interactions

Measuring Performance ( )



GPT 6 Astra

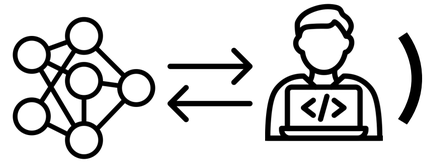


Scalability

= easily evaluate many models

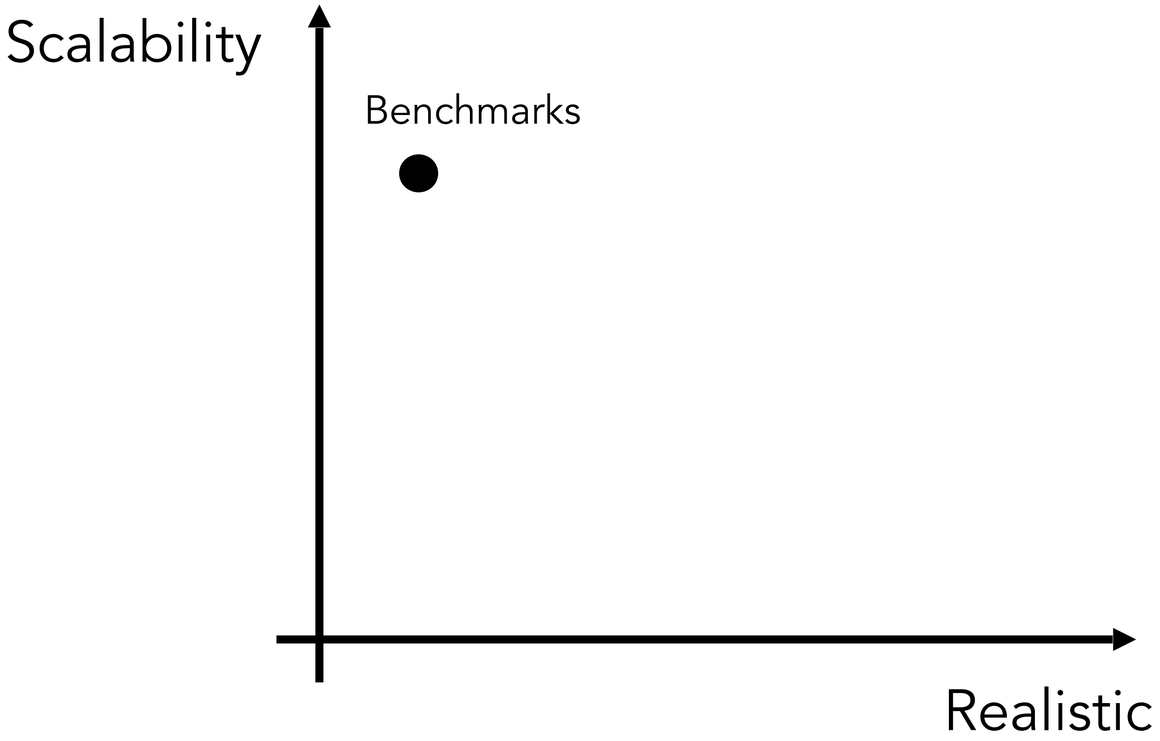
Realistic

= capture usage in actual user environments

Measuring Performance ()



Challenge: existing approaches are flawed



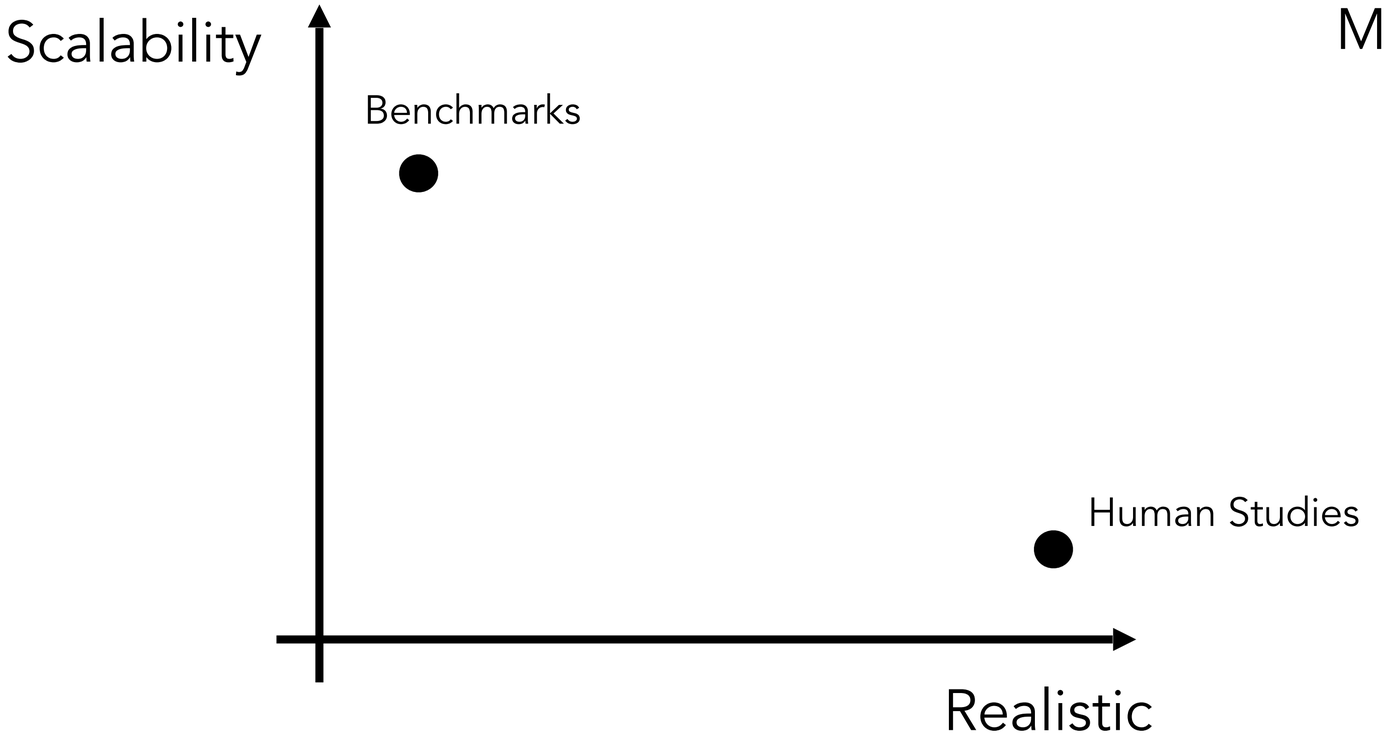
Model performance is measured by test cases

Unit Tests

Pre PR	Post PR	Tests
✗	✓	join_struct_col
✗	✓	vstack_struct_col
✗	✓	dstack_struct_col
✓	✓	matrix_transform
✓	✓	euclidean_diff

Jimenez*, Yang* et al. ICLR, 2024

Challenge: existing approaches are flawed



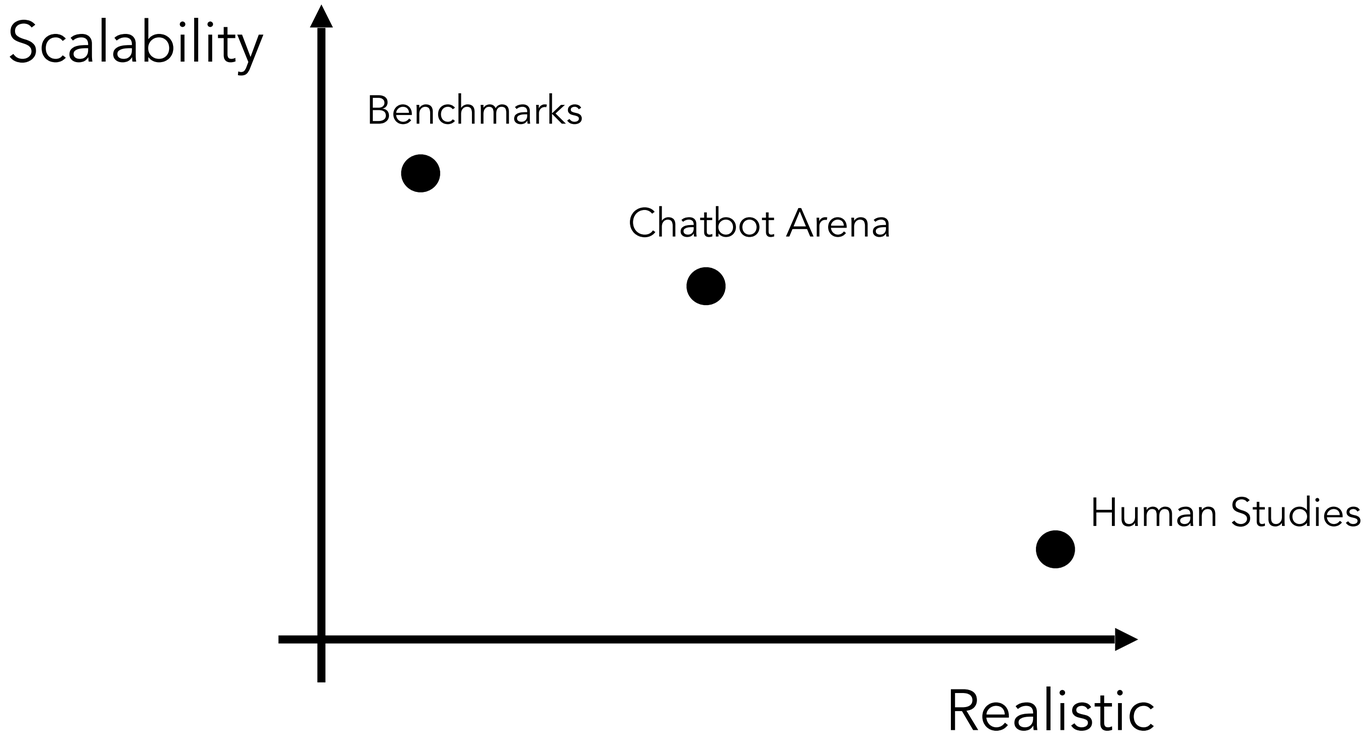
Model performance is measured by human task completion



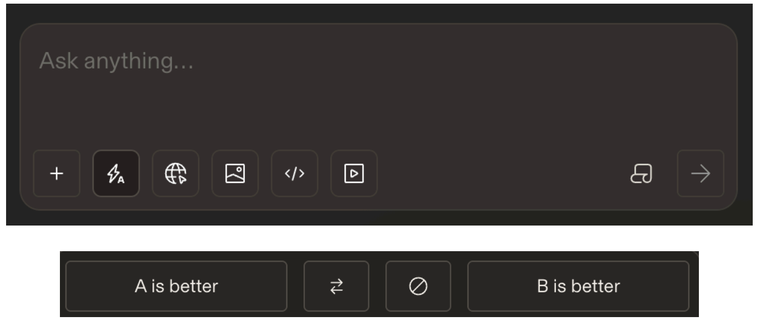


Chen et al., CHI 2026

Challenge: existing approaches are flawed

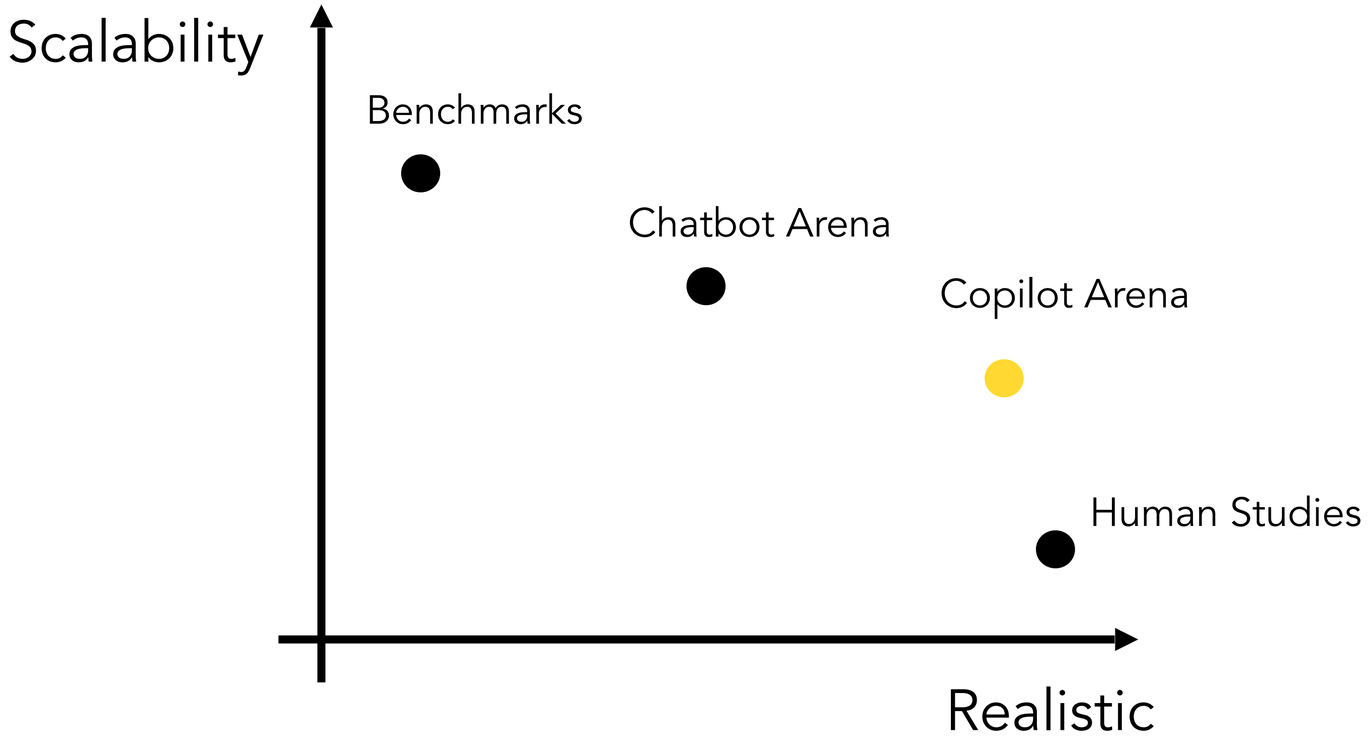


Model performance is measured by ELO scores based on human preference

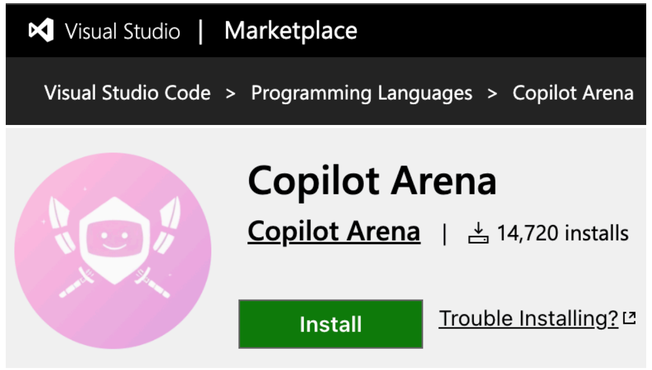


Chiang*, Zheng* et al. ICML, 2024

Copilot Arena pushes the Pareto frontier

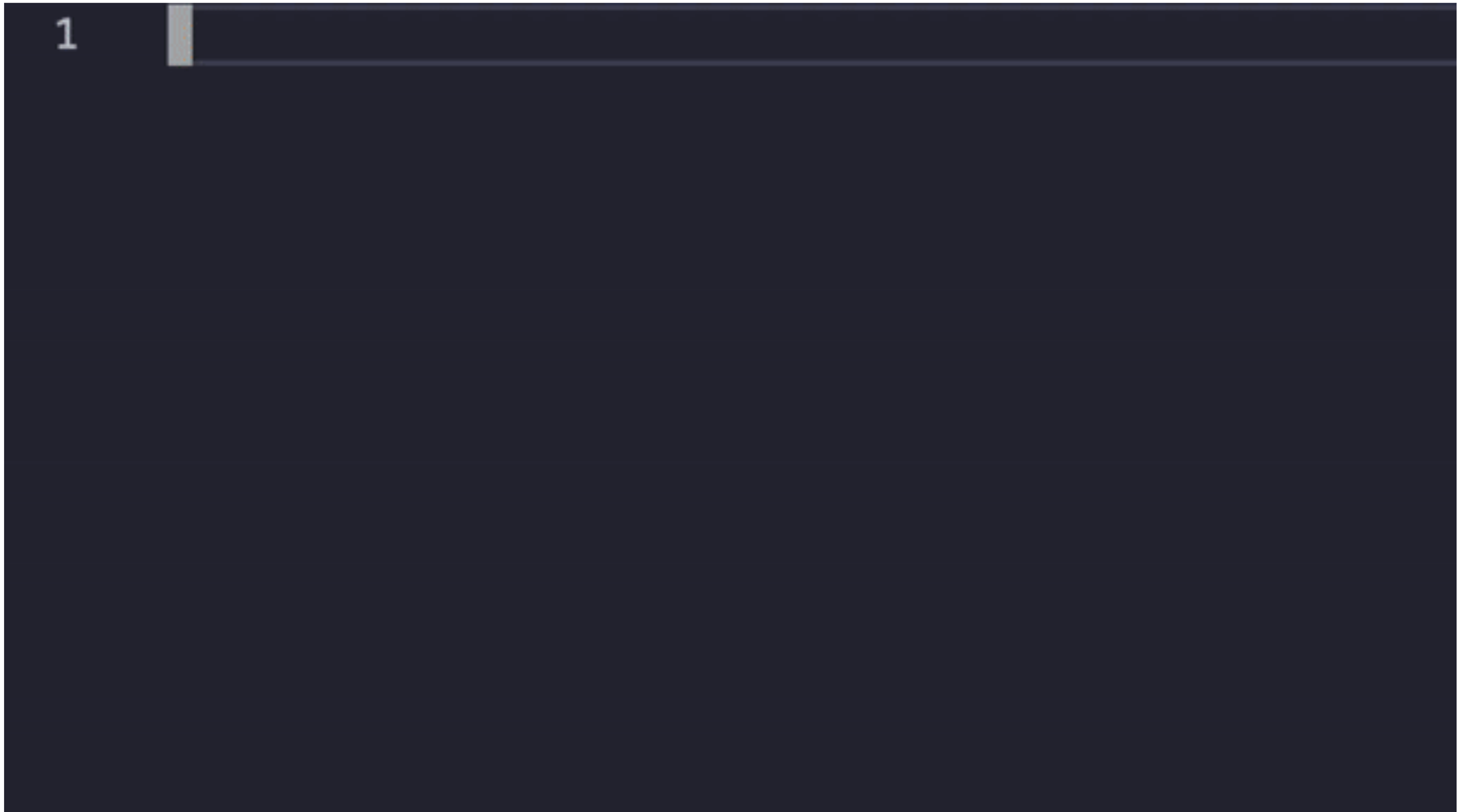


Model performance is measured by ELO scores based on human preference



Chi* and **Chen*** et al. Copilot Arena: A Platform for Code LLM Evaluation in the Wild. ICML, 2025.

Copilot Arena



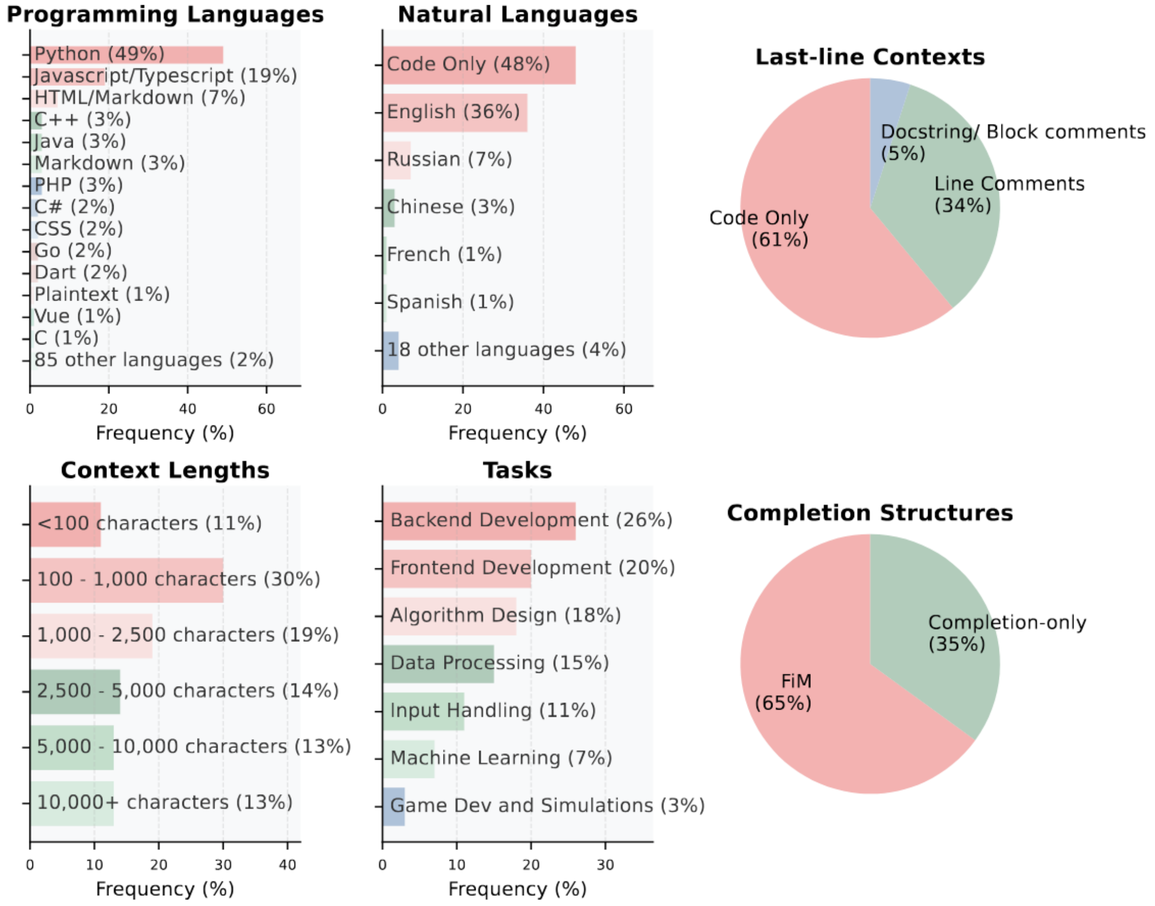
System Overview

Design tackles technical challenges of familiarity, latency, and broad applicability



Copilot Arena is the first interactive evaluation system in user environment

We collect over 30k pairwise comparisons from more than 4.5k users



Excitement about interactive evals from AI community

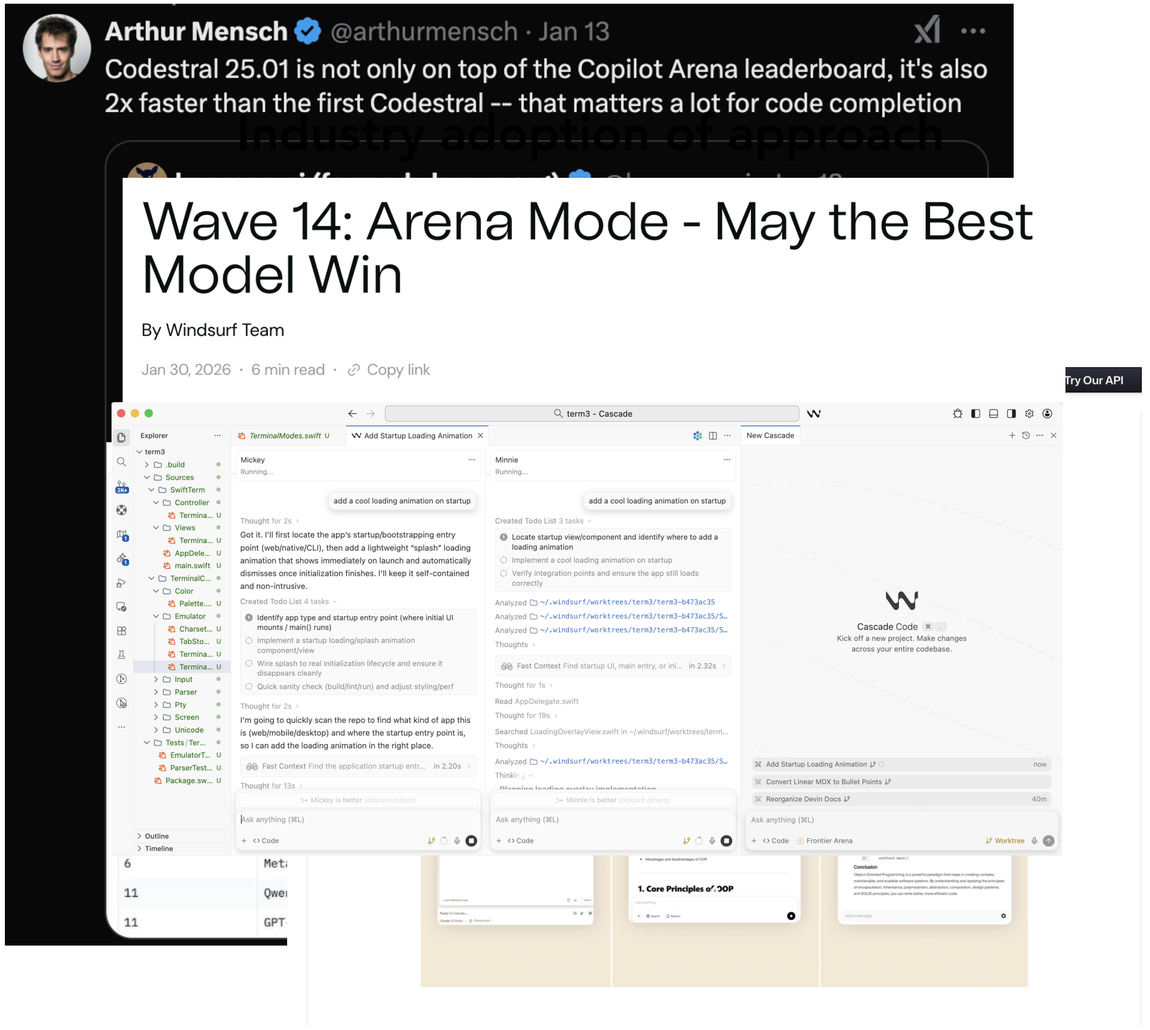
Results cited in new model releases



Covered in articles and multiple YouTube videos (>50k+ views)

THE WALL STREET JOURNAL
WSJ





Development

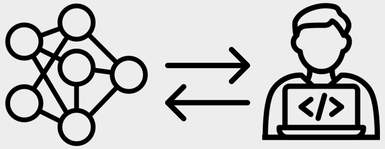


Base Model

Post-training

Static Evaluation

Deployment

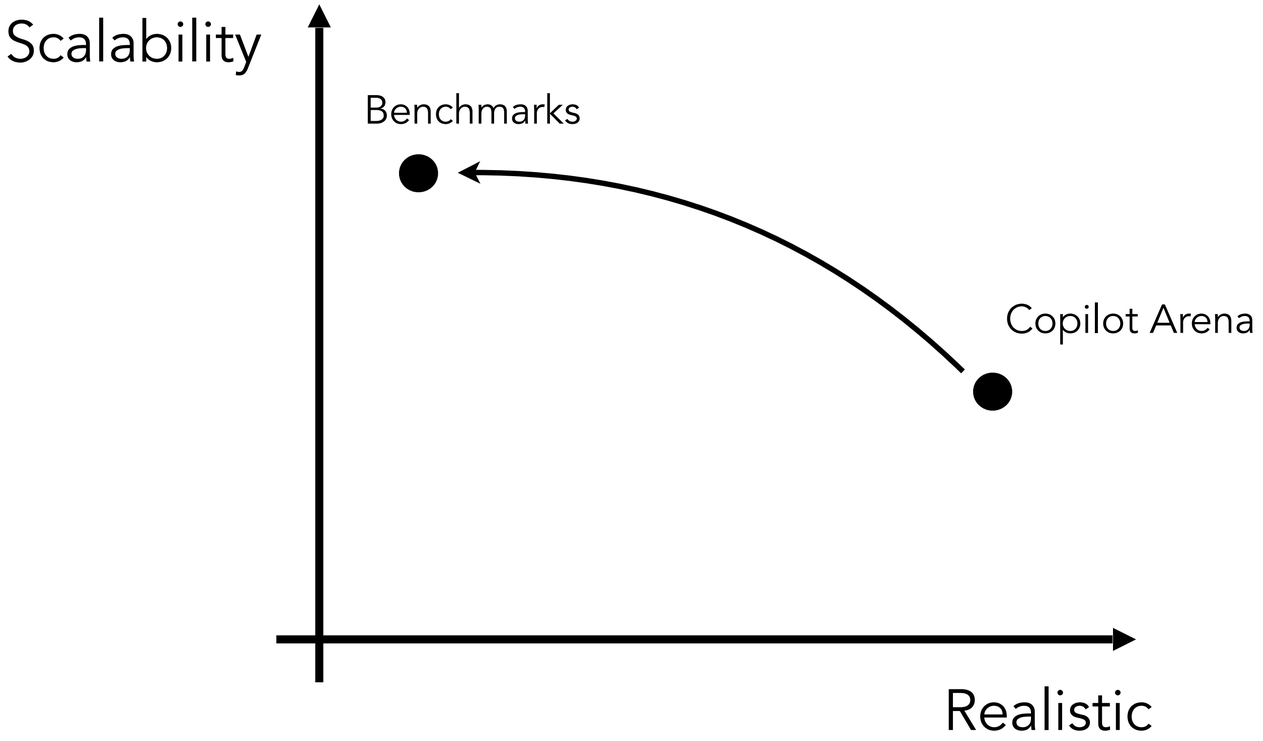


Interactive Eval

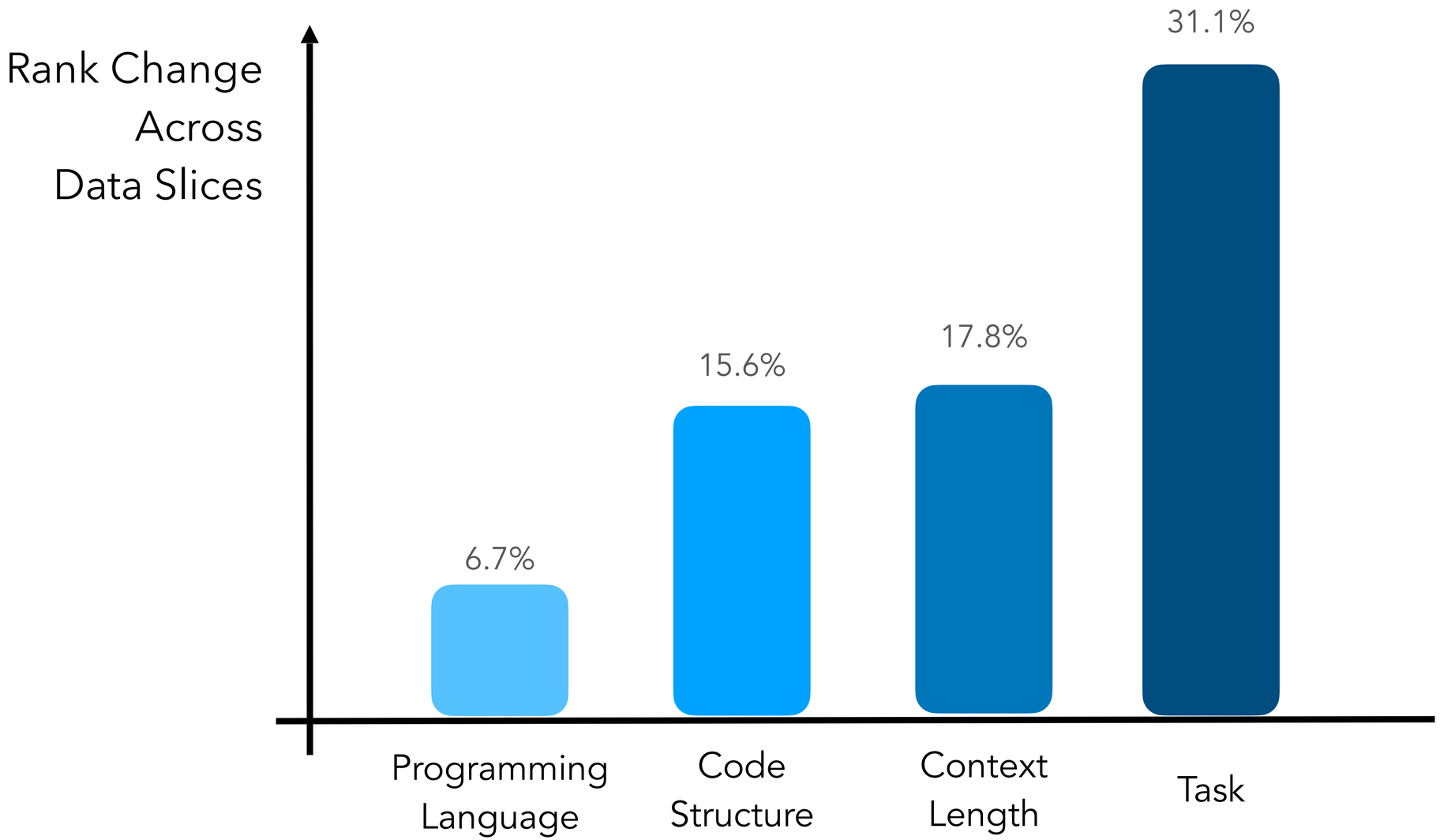
Feedback

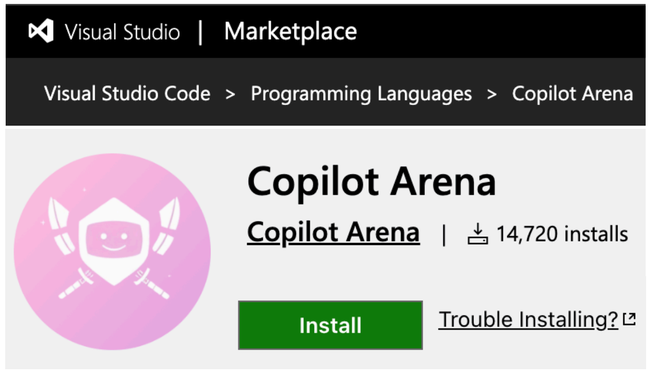
Interface

Can interactive evaluation enable new benchmarks?

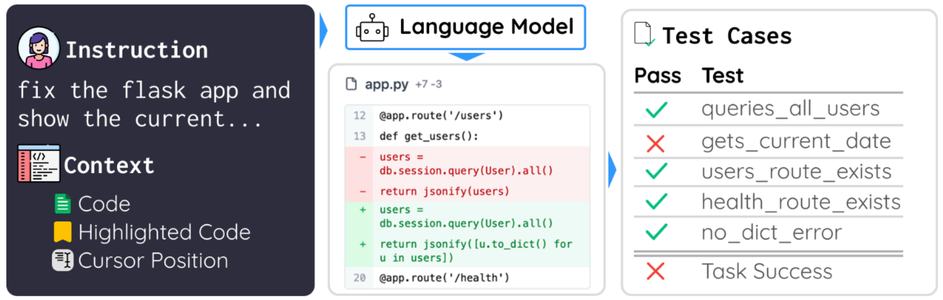


How should we create new benchmarks?



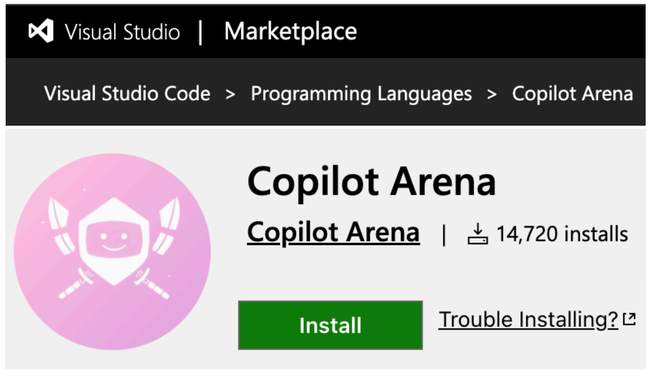






Creating benchmark problems from in-the-wild data captures more realistic usage

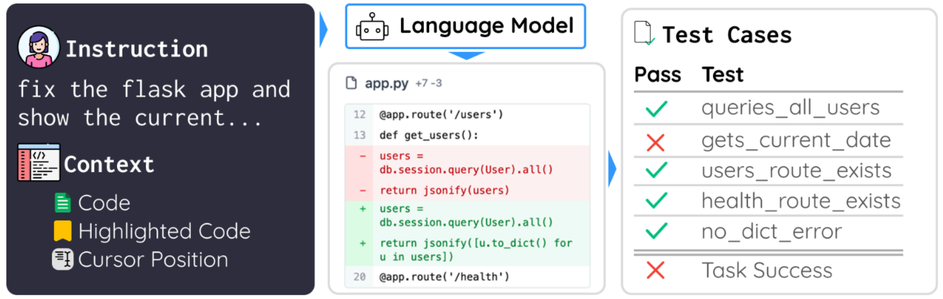
EditBench	CanItEdit (Cassano et al., 2023b)	EditEval (Hu et al., 2023)
RuntimeError: Cannot close a running event loop sys:1: RuntimeWarning: coroutine 'Application.shutdown' was never...	Fix combination_unlimited_rep() so that it returns the right result. The function combination_unlimited_rep should...	Fix the bug in 'sum_even_and_even_index' to make it return the sum of even numbers at even indices.

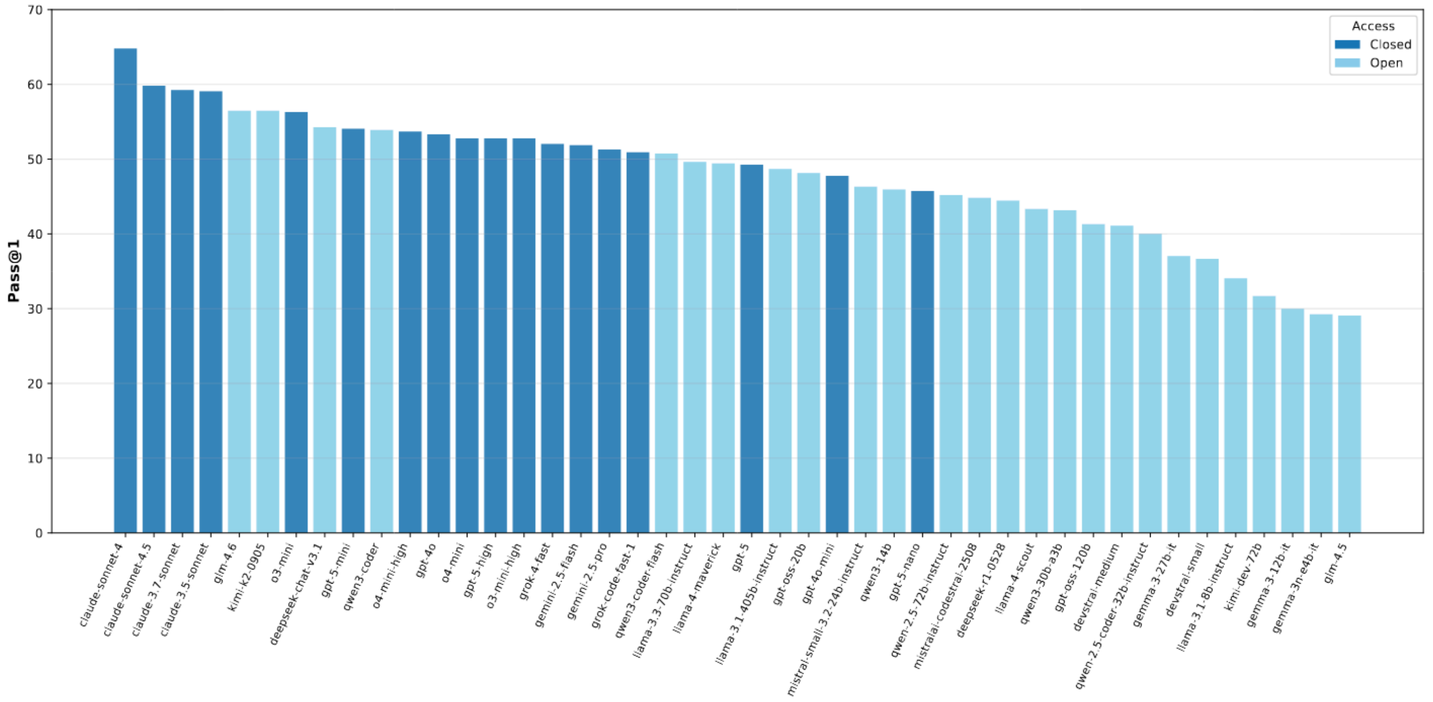


Insights

Data filtering

Benchmark





Current Failure Modes

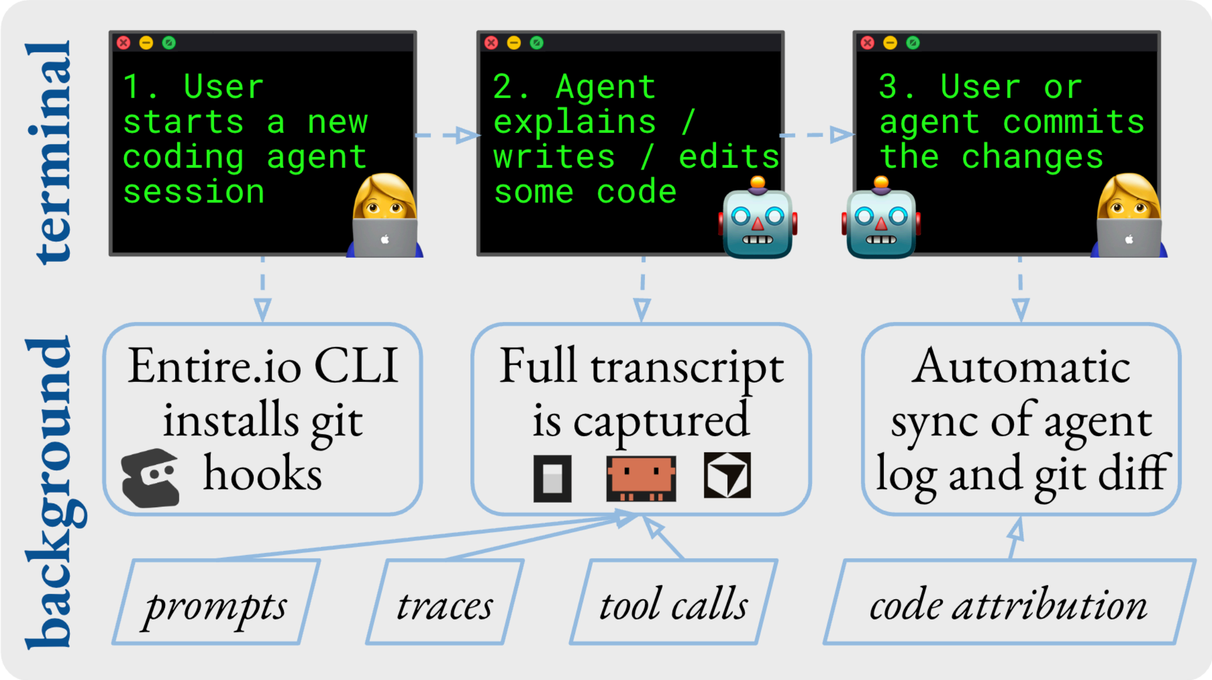
- EDIT-Bench allows us to easily evaluate 40 LLMs
- Only 1 scores over greater than 60% pass@1
- Problem type: 10-15% gap between bug fixing and optimization or feature addition problems
- Level of explicit instruction: 59% gap when instructions are short

Can interactive evaluation enable new benchmarks?

Baumann et al. SWE-chat: Coding Agent Interactions From Real Users in the Wild. COLM, 2026.

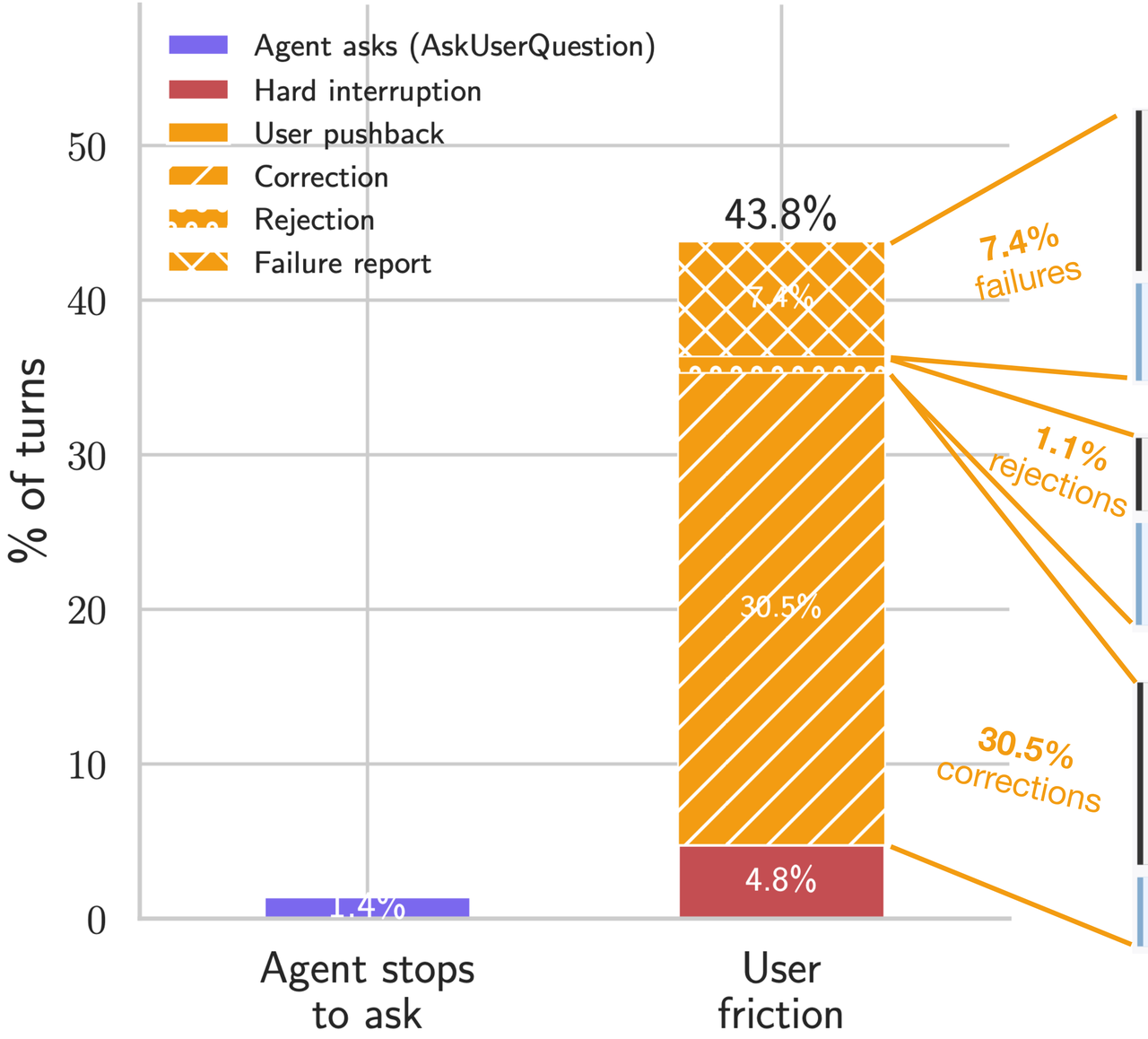
Can interactive evaluation enable new benchmarks?

SWE-chat: Data collection



Baumann et al. SWE-chat: Coding Agent Interactions From Real Users in the Wild. COLM, 2026.

How should we create new benchmarks?



7.4% failures

Agent: Now instead of adding padding to the element (which breaks the chart layout), it wraps the element in a temporary container div with 24px padding, captures that, then restores the DOM. The chart content stays untouched.
▷ `Edit` *ChartExport.tsx — wrapped capture target in padding container*

User: This is broken
[User sends screenshot]

1.1% rejections

Agent: Done. Issue #53 committed (`766901b`) and closed. Ready for the next one.

User: I undid the commit. We need to find a better approach from scratch. Activate the party mode.

30.5% corrections

Agent: The challenge is that `PrepareCommitMsg` doesn't have access to the transcript path — that's passed to the Stop hook by Claude Code.
What we *do* have in PrepareCommitMsg: session state (`FilesTouched`, `BaseCommit`, etc.) and staged changes from git. So we'd need to find another approach to access the transcript content.

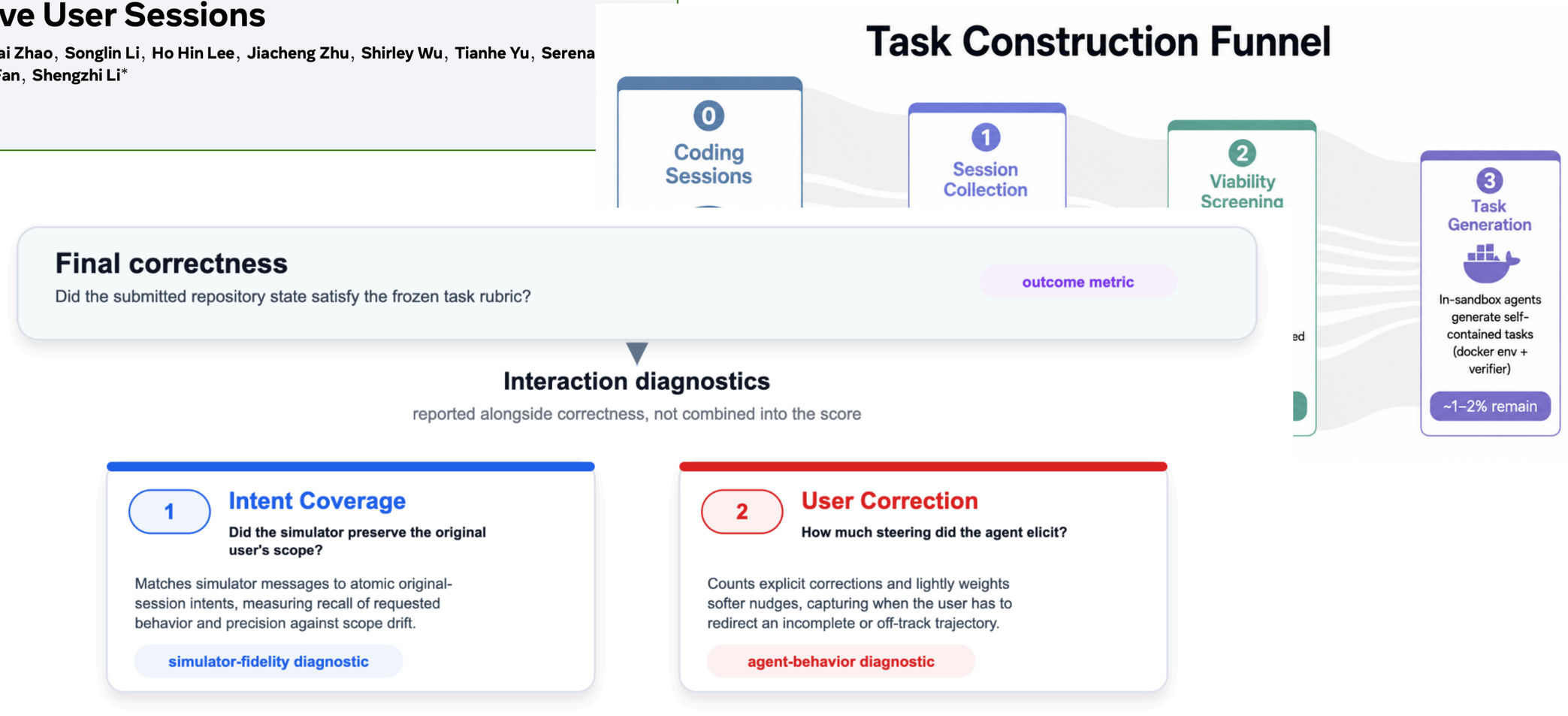
User: but we do have access to the `session.CondensedTranscriptLines` ...?

SWE-Together: Evaluating Coding Agents in Interactive User Sessions

Yifan Wu*, **Zhuokai Zhao**, **Songlin Li**, **Ho Hin Lee**, **Jiacheng Zhu**, **Shirley Wu**, **Tianhe Yu**, **Serena Zhang**, **Xiangjun Fan**, **Shengzhi Li***

Meta

*Project Co-Lead



Wu et al. SWE-Together: Evaluating Coding Agents in Interactive User Sessions. Preprint, 2026.

Development

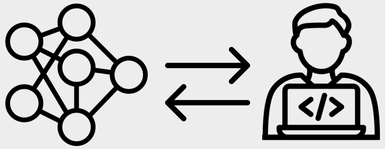


Base Model

Post-training

Static Evaluation

Deployment



Interactive Eval

Feedback

Interface

Building Collaborative LLM Agents

Measure model performance in collaborative contexts

Learn from interaction in collaborative contexts

Optimize interfaces for sustainable interactions

Development

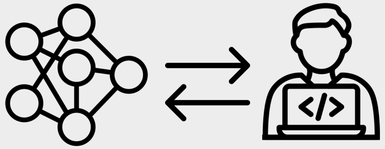


Base Model

Post-training

Static Evaluation

Deployment



Interactive Eval

Feedback

Interface

Building Collaborative LLM Agents

Measure model performance in collaborative contexts

- Reframing model evaluation as the scalability-realism frontier
- We need more work to bridge static and interactive evaluation

Learn from interaction in collaborative contexts

Optimize interfaces for sustainable interactions

Development

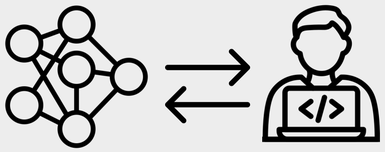


Base Model

Post-training

Static Evaluation

Deployment



Interactive Eval

Feedback

Interface

Building Collaborative LLM Agents

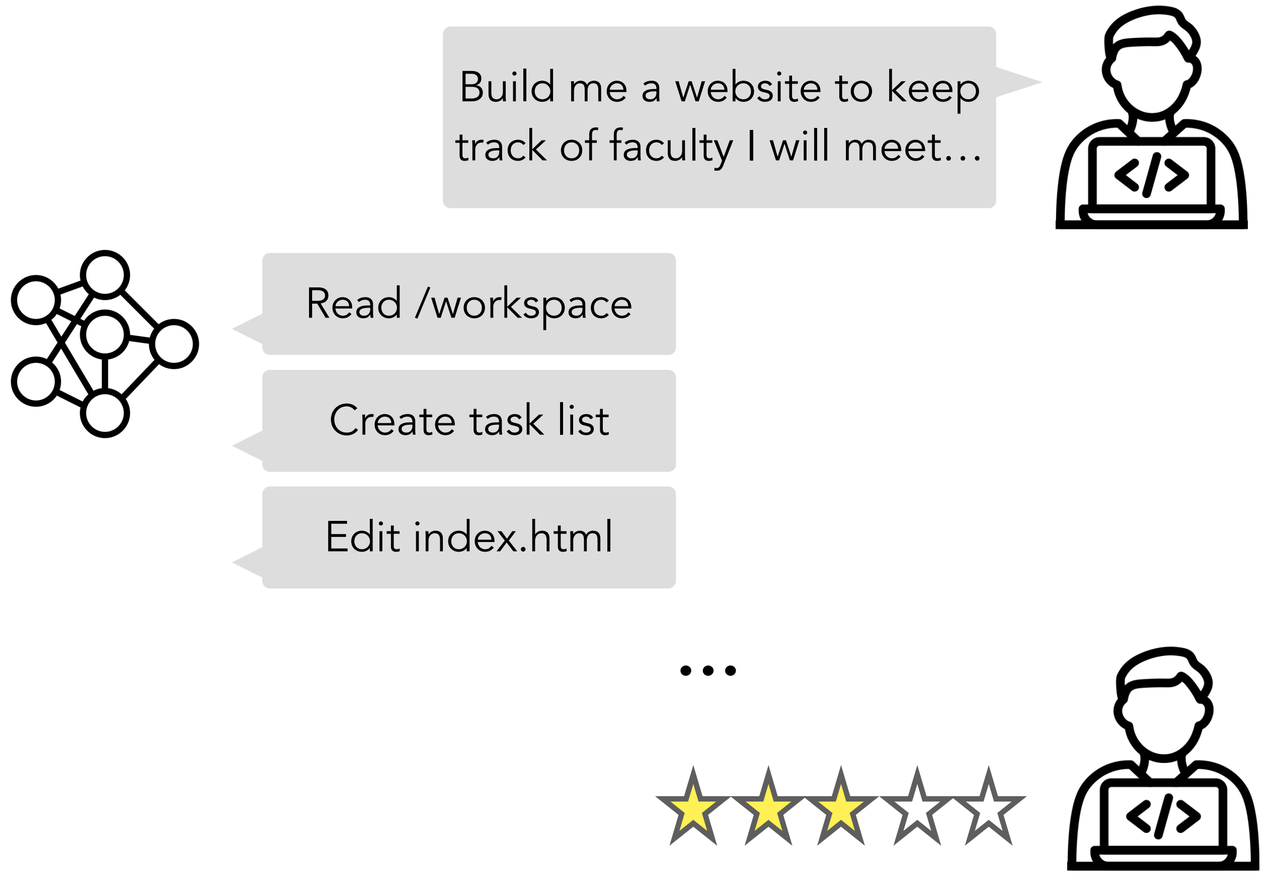
Measure model performance in collaborative contexts

- Reframing model evaluation as the scalability-realism frontier
- We need more work to bridge static and interactive evaluation

Learn from interaction in collaborative contexts

Optimize interfaces for sustainable interactions

Collaboration requires learning from multi-turn interactions



Most interactions do not receive explicit feedback



Prior approaches have looked for other proxies



Unit Tests

Pre PR	Post PR	Tests
✗	✓	join_struct_col
✗	✓	vstack_struct_col
✗	✓	dstack_struct_col
✓	✓	matrix_transform
✓	✓	euclidean_diff



Insight: can we make use of the unlabeled interactions?

Development

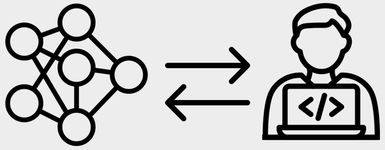


Base Model

Post-training

Static Evaluation

Deployment

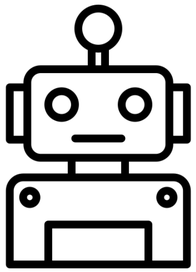


Interactive Eval

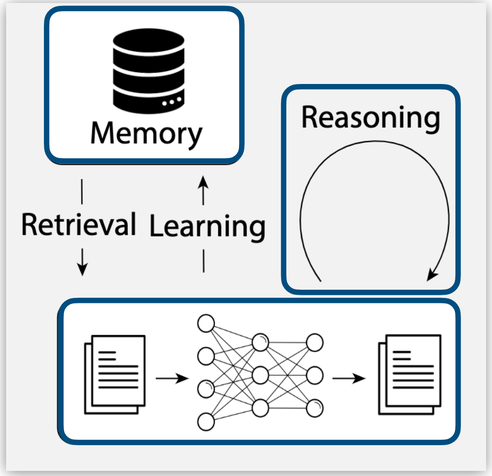
Feedback

Interface

Optimizing agent design in deployment



Coding Agent



Sumers et al. 2024

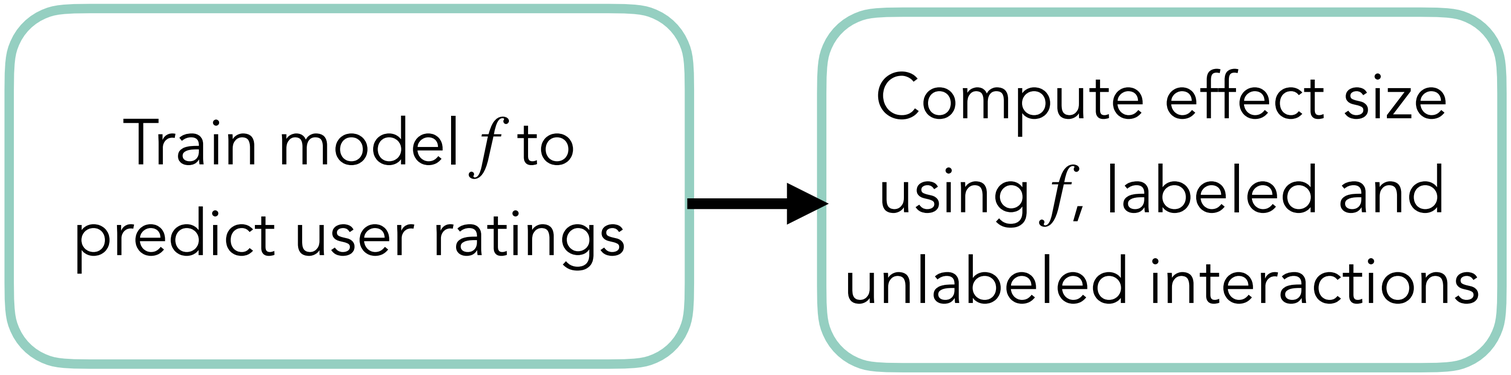
Agents comprise…

- Language model
- Memory
- Reasoning and Planning
- …and more!

Which changes have the biggest impact on user satisfaction?

PULSE efficiently compares agent designs

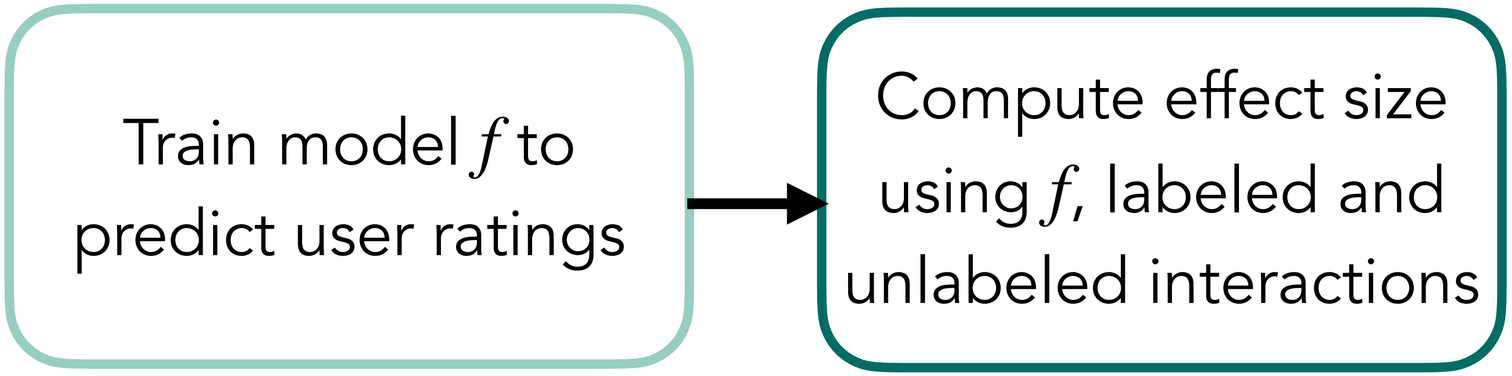
Suppose we want to decide whether users prefer design A or B



Chen et al. How can we assess human-agent interactions? Case studies in software agent design. ICML, 2026.

PULSE efficiently compares agent designs

Suppose we want to decide whether users prefer design A or B



Computing Effect Size

n_c = # labeled samples in group c $\quad$ N_c = # unlabeled samples in group c

$$(X_i, Y_i)_{i=1}^{n_c} \sim \mathcal{D} \qquad \tilde{X}_j \sim \mathcal{D},\ j = 1,\ldots,N_c$$

Prediction powered inference on design c given model $f : \mathcal{X} \to \mathbb{R}$

$$\hat{\mu}_c(\lambda_c) = \frac{1}{n_c} \sum_i Y_i$$

Combine both estimators

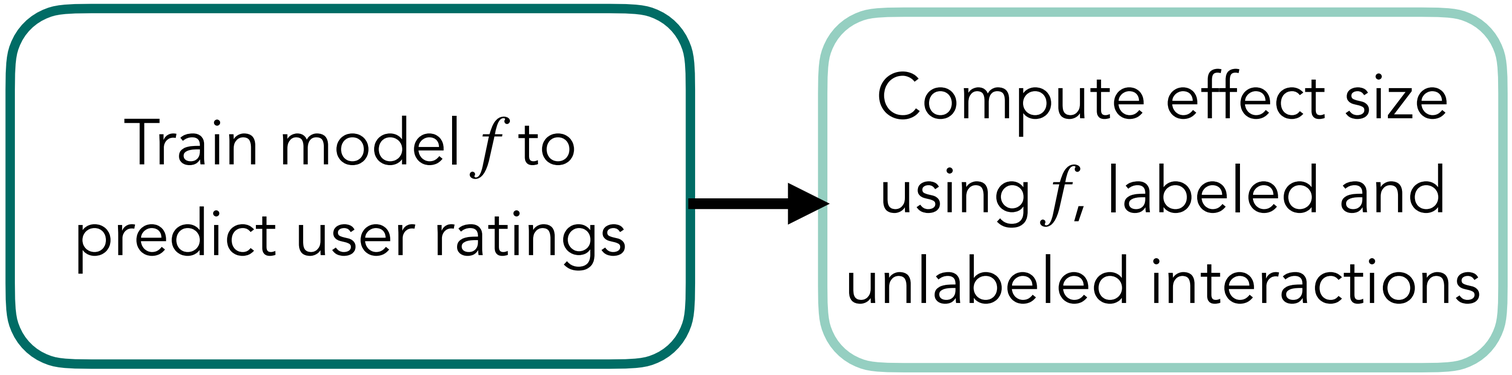
$$\widehat{\Delta}_{\text{augment}} = \hat{\mu}_A(\hat{\lambda}_A) - \hat{\mu}_B(\hat{\lambda}_B)$$

Comparison to baseline

$$\widehat{\Delta}_{\text{naive}} = \frac{1}{n_A} \sum_{i \in A} Y_i - \frac{1}{n_B} \sum_{i \in B} Y_i$$

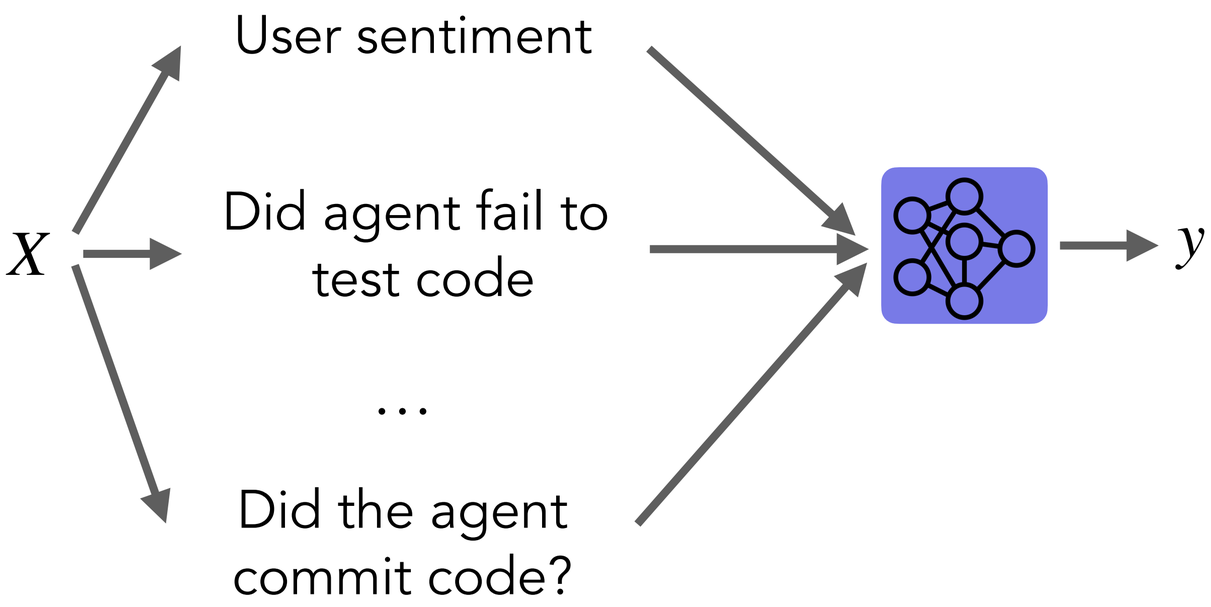
PULSE efficiently compares agent designs

Suppose we want to decide whether users prefer design A or B



How do we train f ?

Challenge: X is a complex, high-dimensional object



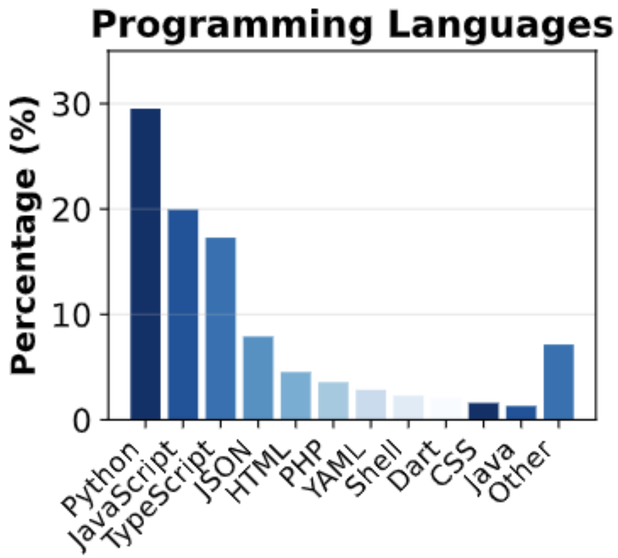
Results: Improves correlation by over 26% compared to baselines that are given X

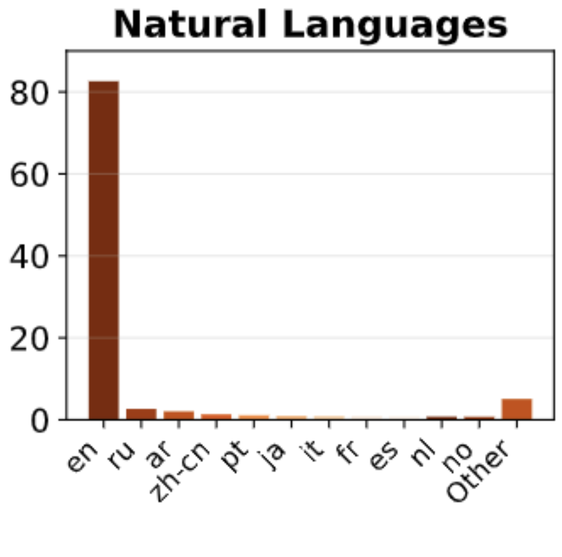
Deploying PULSE on a real platform

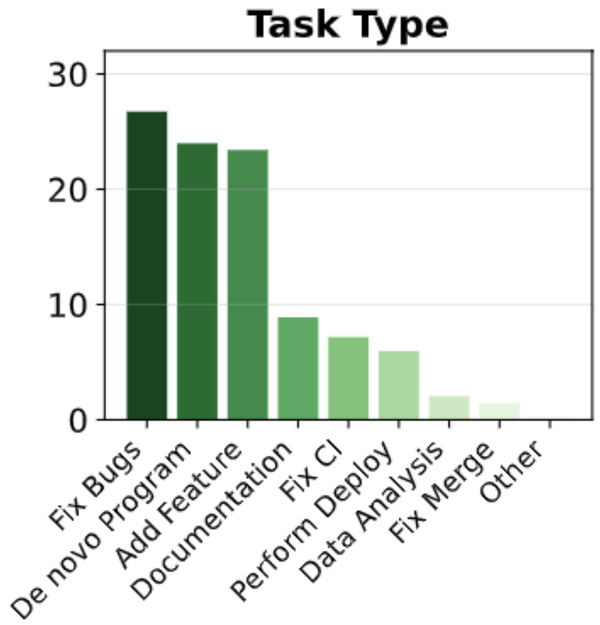


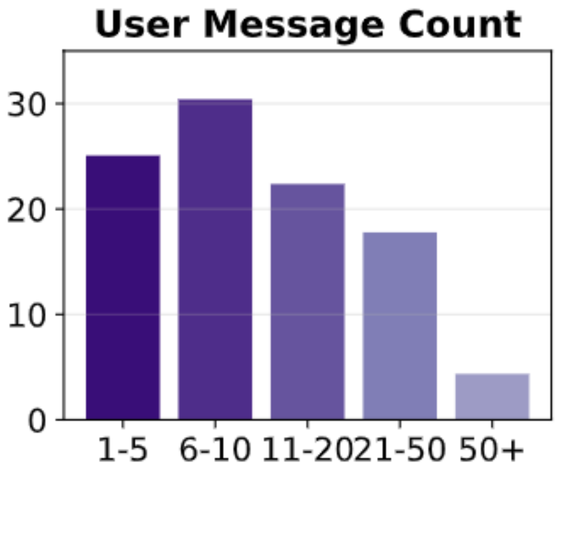
Overview of Users

Over the course of multiple months, we ran experiments on over 15k users:

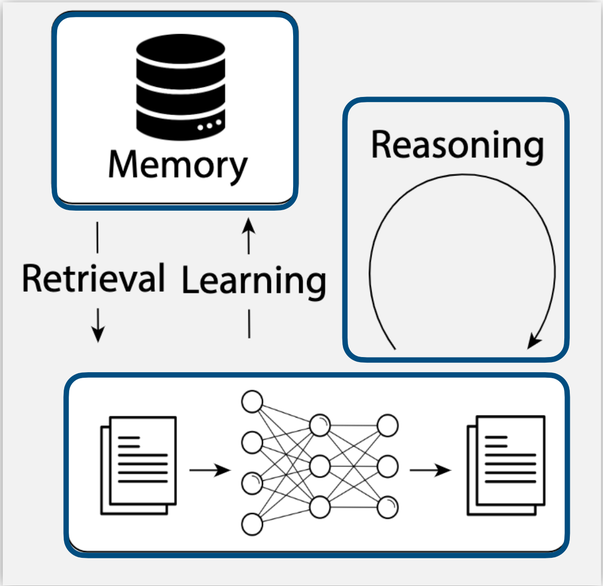






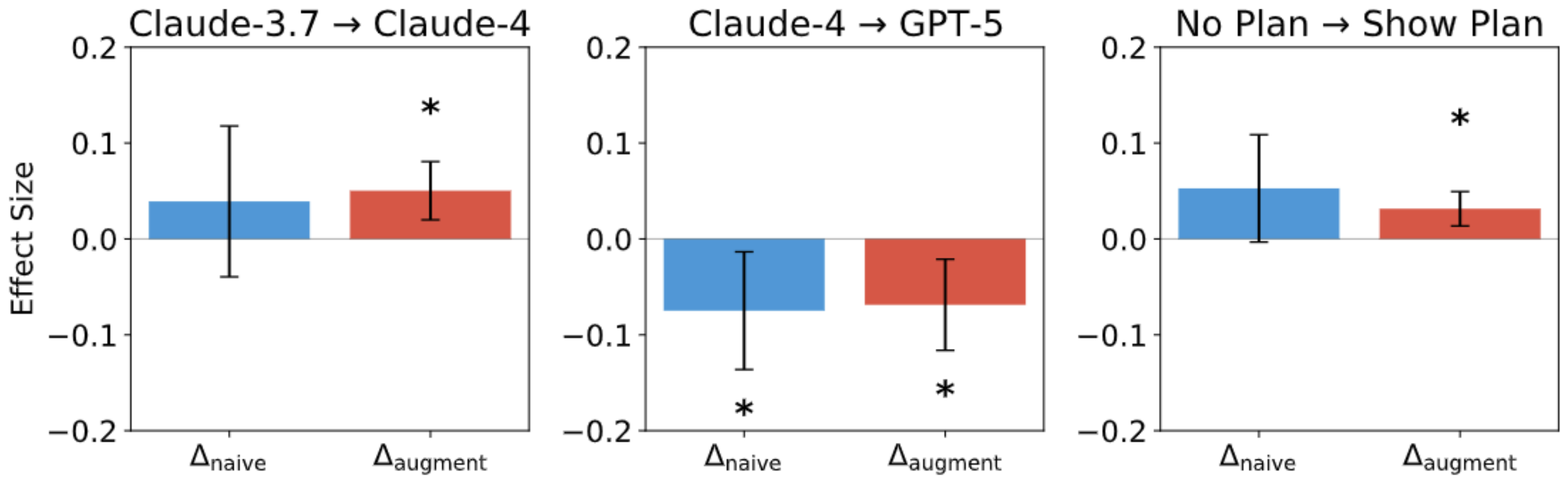


Agent design has significant impact on user experience



- Changing **LLM** had the largest effect (6-8% difference)
- Changing **reasoning** strategy had a smaller but significant effect (3% difference)
- Reducing **memory** had no significant difference

PULSE improves over naive baseline



PULSE reduces confidence intervals by up to 40%, while remaining faithful to naive baseline

Development

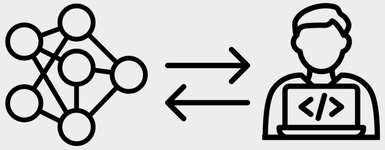


Base Model

Post-training

Static Evaluation

Deployment

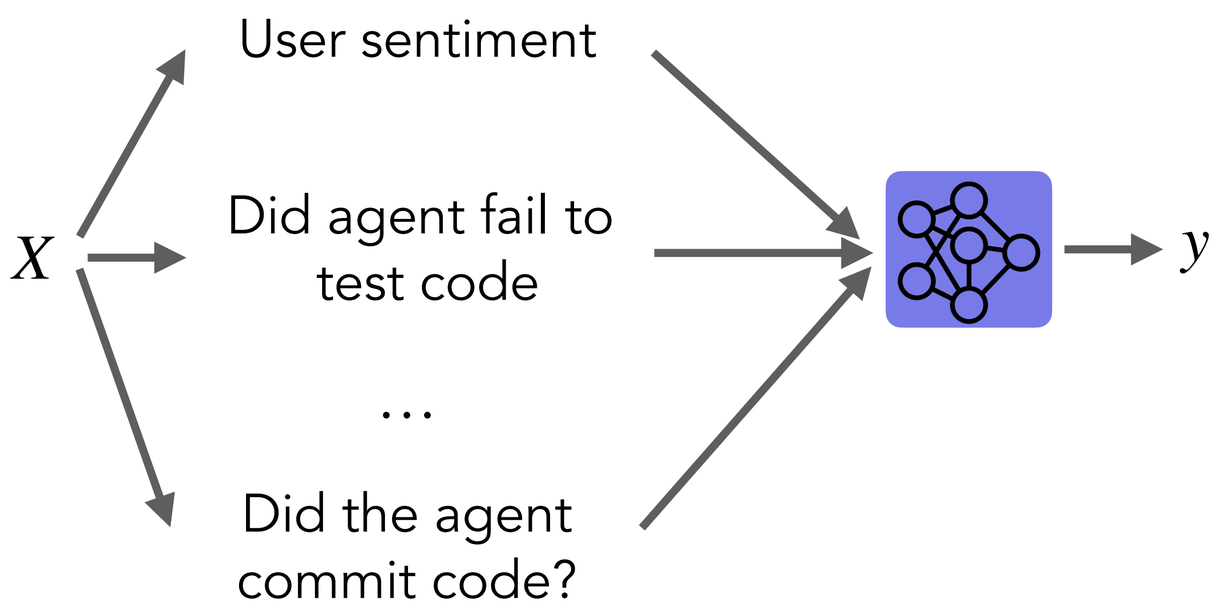


Interactive Eval

Feedback

Interface

Can we also use these implicit signals for model development?

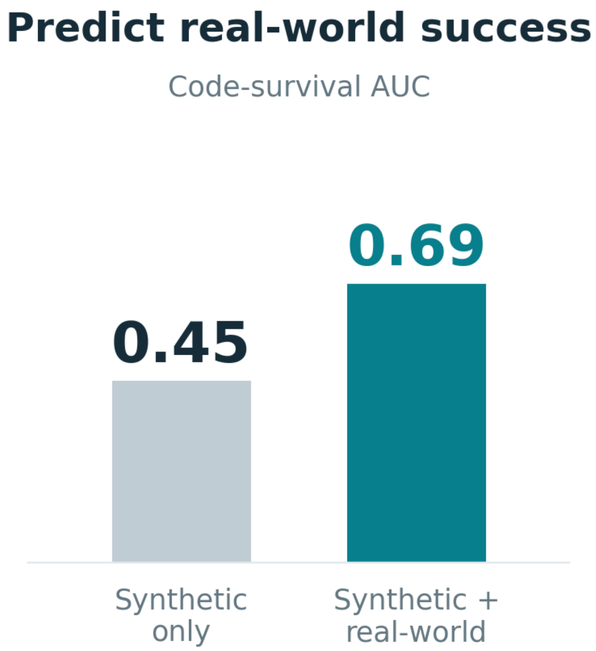


We train a critic model to predict both implicit signals (dense) and outcomes like code survival (sparse)



Wang et al. [incl **Chen**]. A Rubric-Supervised Critic from Sparse Real-World Outcomes. COLM 2026.

Demonstrating usefulness of critic model



Wang et al. [incl **Chen**]. A Rubric-Supervised Critic from Sparse Real-World Outcomes. COLM 2026.

CURSOR

Mar 26, 2026

Improving Composer through real-time RL



Each real-time RL cycle starts by collecting billions of tokens from user interactions with the current checkpoint and *distilling them into reward signals*.

Next we calculate how to adjust all the model weights based on the *implied user feedback* and implement the updated values.



Development

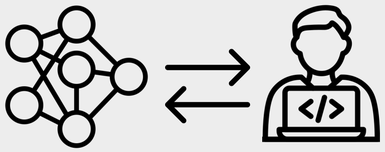


Base Model

Post-training

Static Evaluation

Deployment



Interactive Eval

Feedback

Interface

Building Collaborative LLM Agents

Measure model performance in collaborative contexts

- Reframing model evaluation as the scalability-realism frontier
- We need more work to bridge static and interactive evaluation

Learn from interaction in collaborative contexts

Optimize interfaces for sustainable interactions

Development

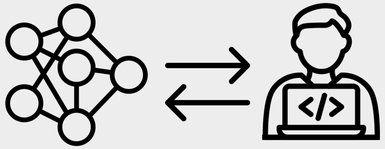


Base Model

Post-training

Static Evaluation

Deployment



Interactive Eval

Feedback

Interface

Building Collaborative LLM Agents

Measure model performance in collaborative contexts

- Reframing model evaluation as the scalability-realism frontier
- We need more work to bridge static and interactive evaluation

Learn from interaction in collaborative contexts

- Reframing learning from human feedback towards task completion
- Much value lies in learning from implicit human-AI interaction signal

Optimize interfaces for sustainable interactions

Development

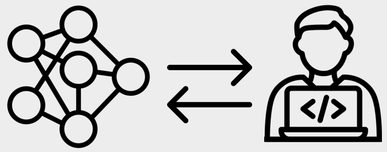


Base Model

Post-training

Static Evaluation

Deployment



Interactive Eval

Feedback

Interface

Building Collaborative LLM Agents

Measure model performance in collaborative contexts

- Reframing model evaluation as the scalability-realism frontier
- We need more work to bridge static and interactive evaluation

Learn from interaction in collaborative contexts

- Reframing learning from human feedback towards task completion
- Much value lies in learning from implicit human-AI interaction signal

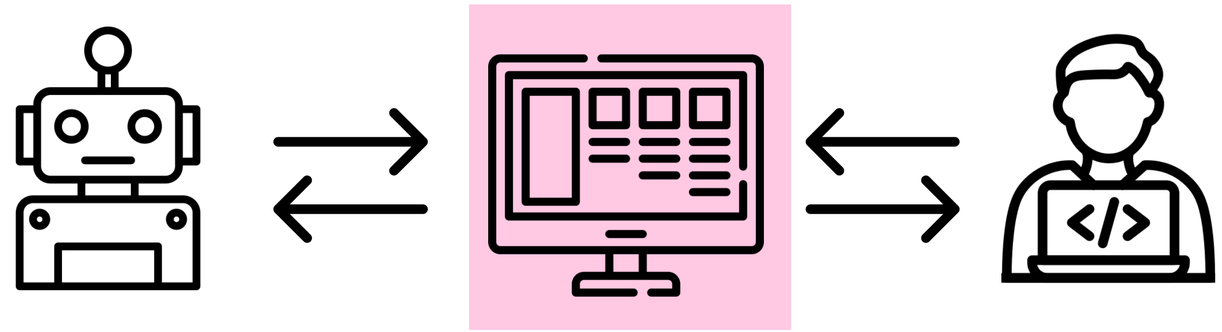
Optimize interfaces for sustainable interactions

Building collaborative agents is a multi-objective problem



Human-AI Performance = AI Quality

Building collaborative agents is a multi-objective problem

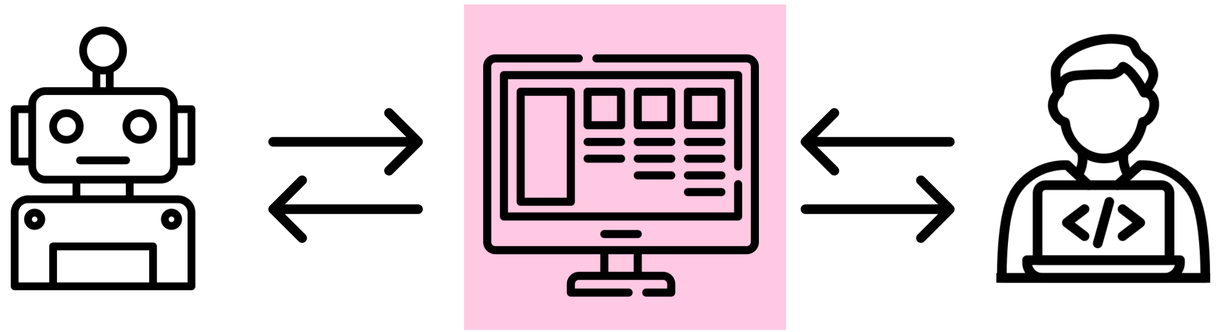


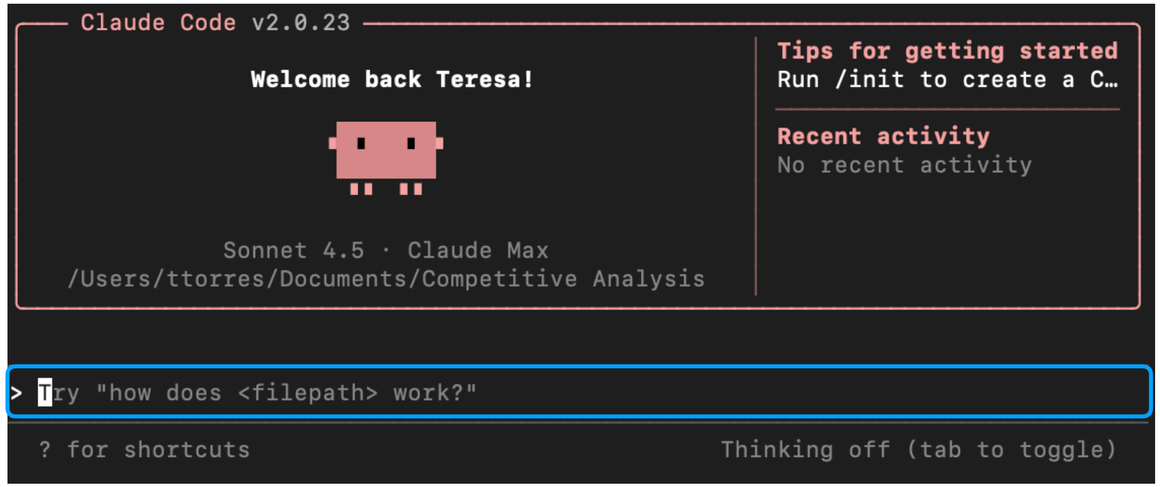
Human-AI Performance = AI Quality

– Prompting Effort

– Verification Effort

Challenge: significant user effort spent prompting AI agents





Insight: can we move from *reactive* to *proactive* AI systems?

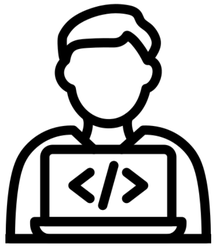
From reactive to proactive AI

Interactions with reactive AI system

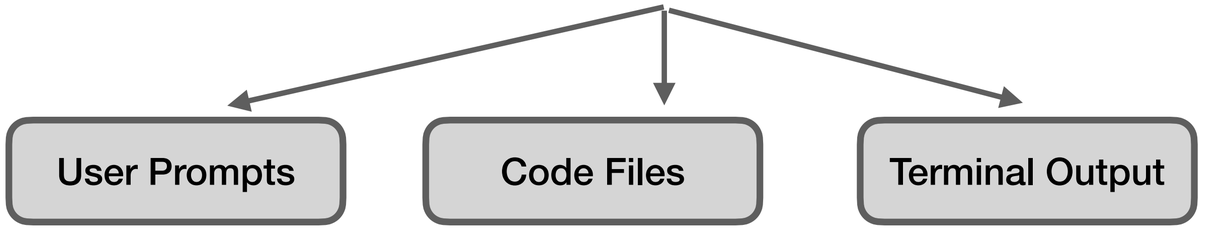
$$\text{output} = \text{LM}(\text{prompt})$$

Interactions with a proactive AI system

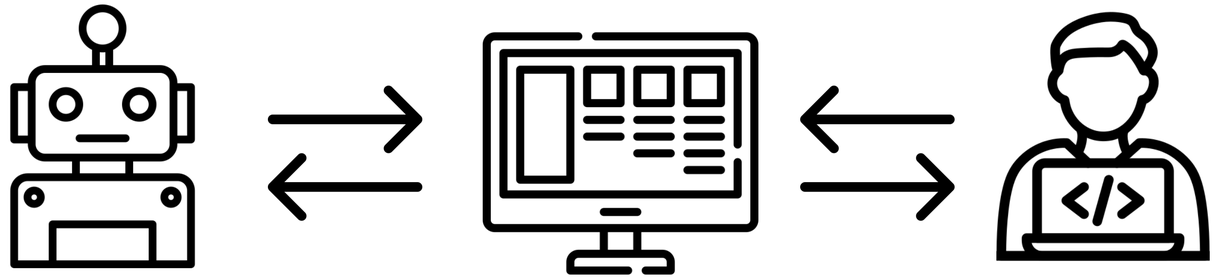
$$\text{when to show}, \text{what to show} = \text{LM}(\text{context})$$



User



Challenge: Proactive AI may add verification cost



Human-AI Performance = AI Quality

– Prompting Effort ↓

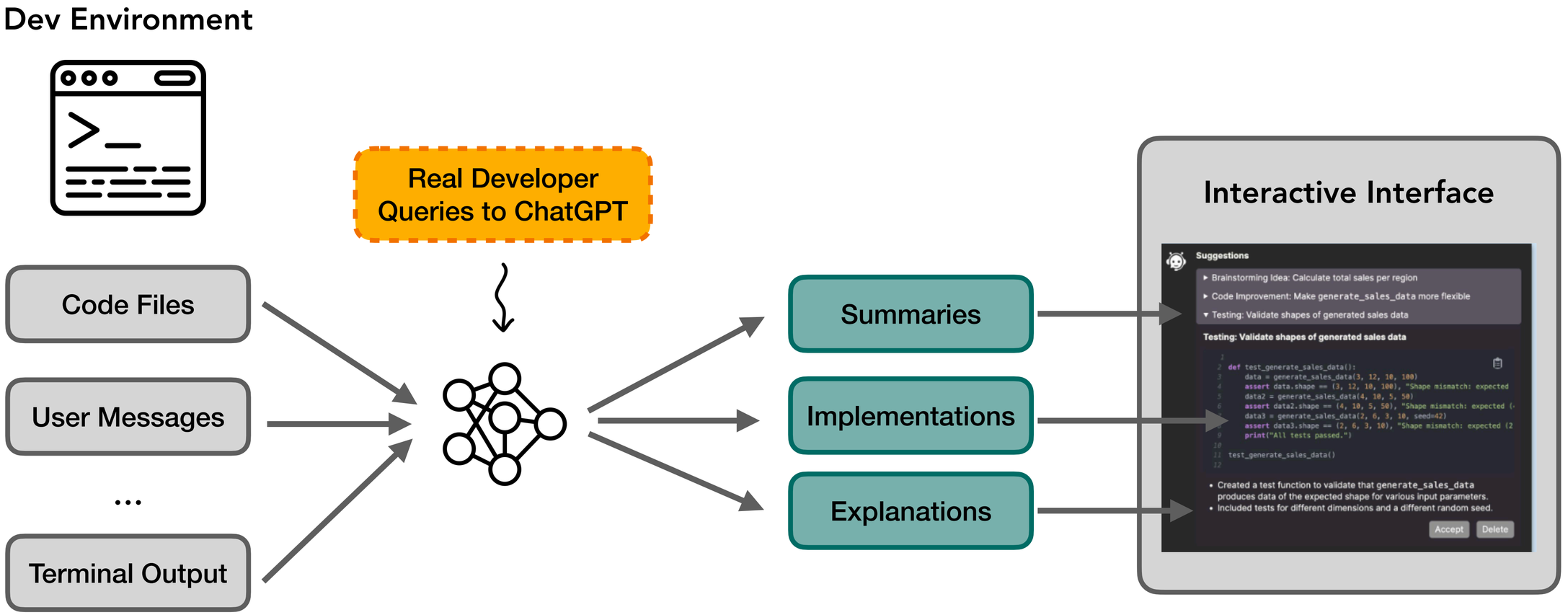
– Verification Effort ↑

Challenge: How do we balance a user's prompting and verification effort?

Initially	During interaction	When wrong	Over time
G1 Make clear what the system can do	G3 Time services based on context	G7 Support efficient invocation	G12 Remember recent interactions
G2 Make clear how well it can do it	G4 Show contextually relevant information	G8 Support efficient dismissal	G13 Learn from user behavior
	G5 Match relevant social norms	G9 Support efficient correction	G14 Update and adapt cautiously
	G6 Mitigate social biases	G10 Scope services when in doubt	G15 Encourage granular feedback
		G11 Make clear why the system did what it did	G16 Convey the consequences of user actions
			G17 Provide global controls
			G18 Notify users about changes

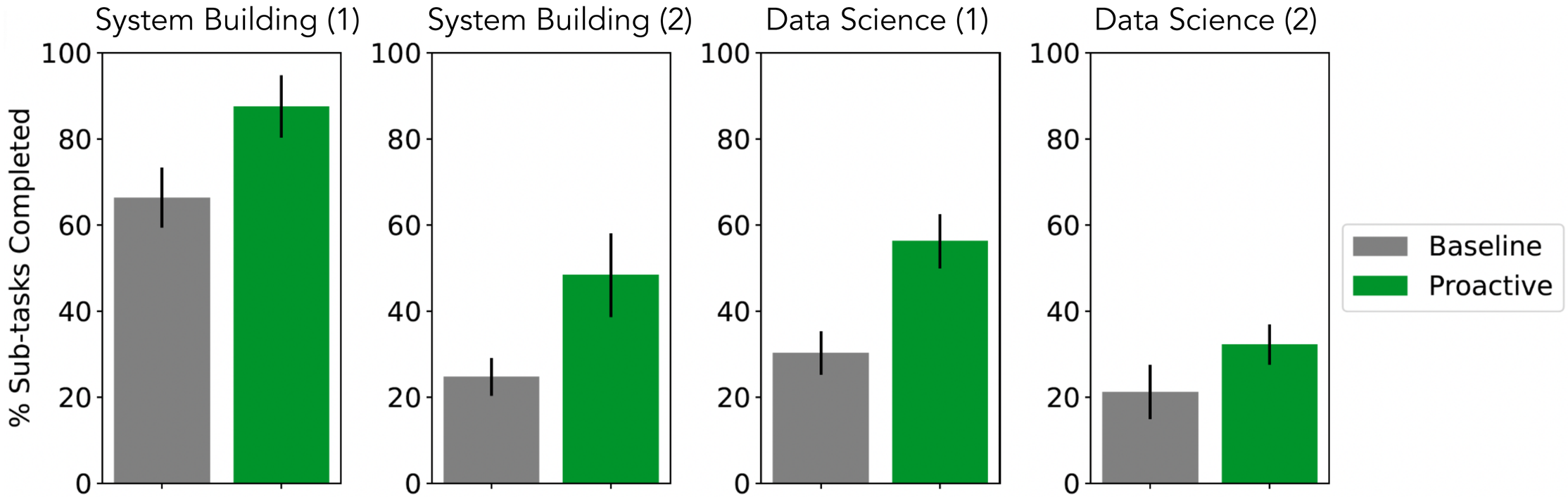
Amershi et al. Guidelines for Human-AI Interaction. CHI, 2019.

Challenge: How do we balance a user's prompting and verification effort?

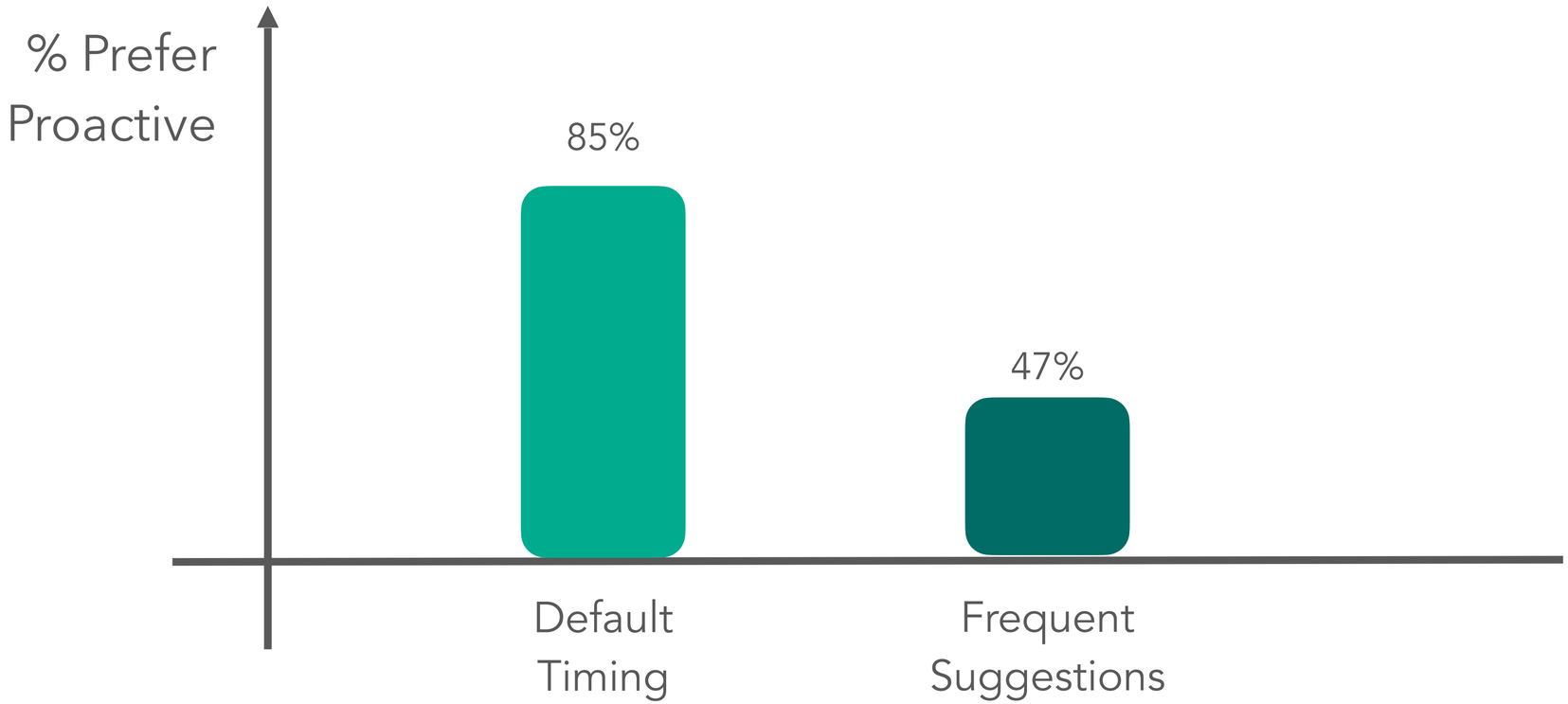


Chen et al. Need Help? Designing Proactive AI Assistants for Programming. CHI, 2025.

Proactive AI improves human-AI performance



Participants also prefer the proactive assistant



Only impacts user experience, not task completion!

Do findings translate to in-the-wild usage?

We collaborated with a company to study in-the-wild usage:

JetBrains IDEs are used by over 15 million developers worldwide

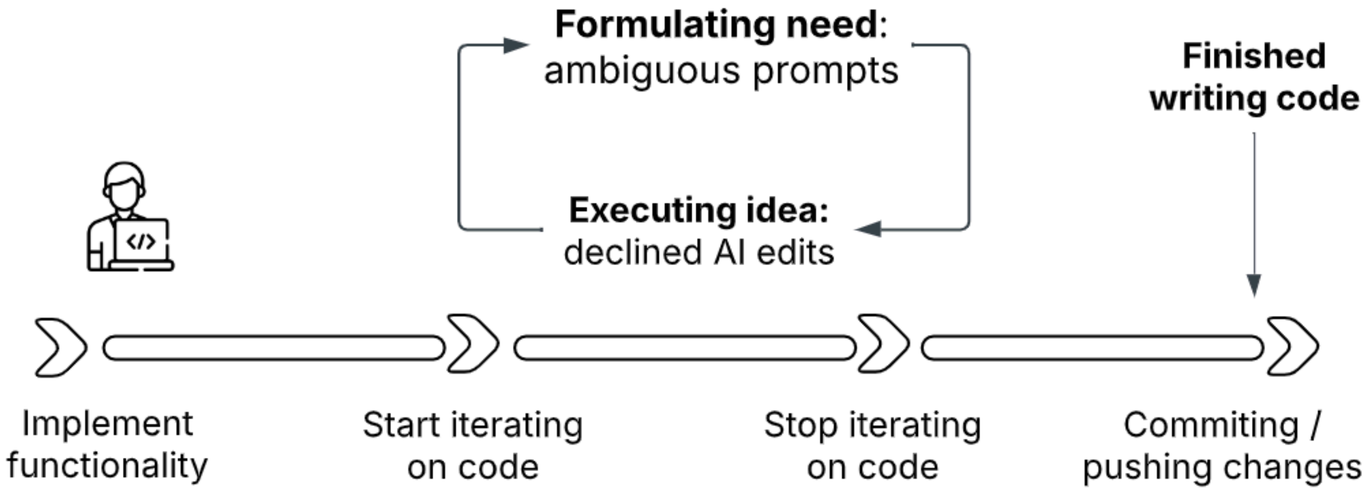


Kuo et al. [incl. **Chen**] Developer Interaction Patterns with Proactive AI: A Five-Day Field Study. IUI, 2026.

Do findings translate to in-the-wild usage?

We collaborated with a company to study in-the-wild usage:

JetBrains IDEs are used by over 15 million developers worldwide



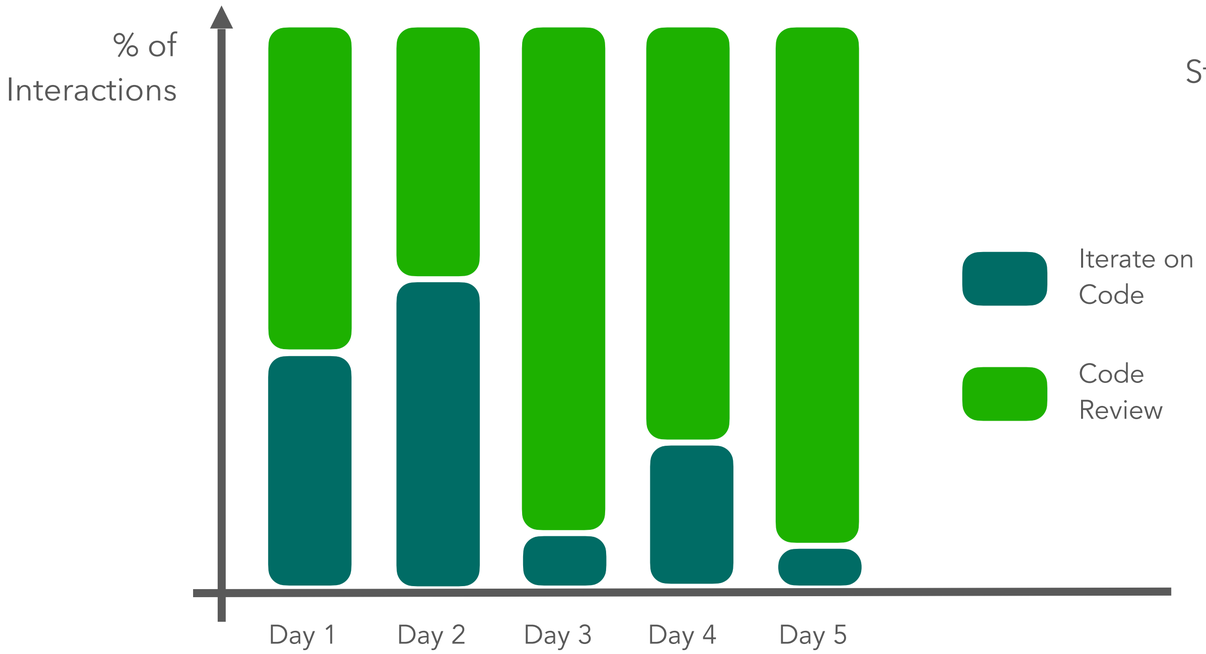
Kuo et al. [incl. **Chen**] Developer Interaction Patterns with Proactive AI: A Five-Day Field Study. IUI, 2026.

Strong system usability and insight into interaction timing

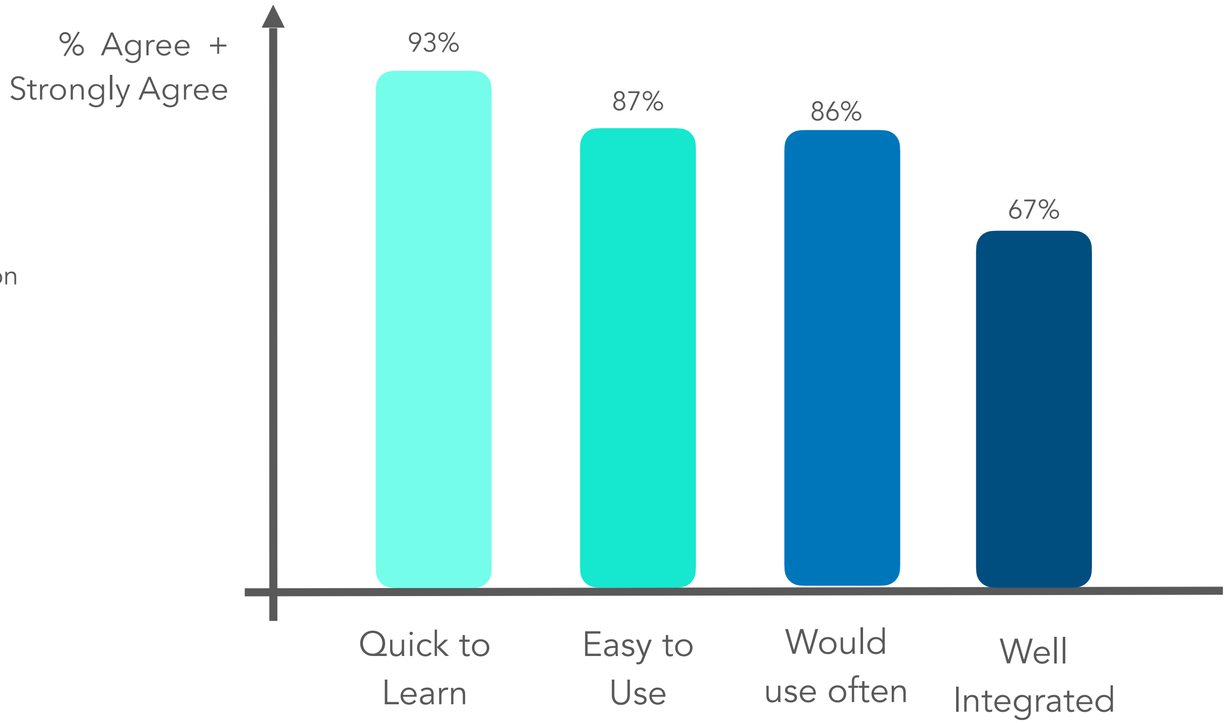
Through an in-the-wild, longitudinal study...

We collect 229 interactions across 5,732 intervention points

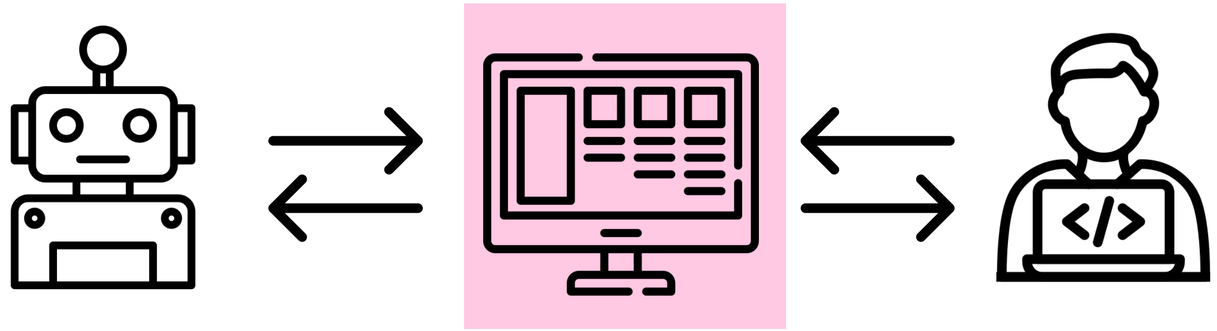
Engagement by Intervention



System Usability



Challenge: significant user effort spent prompting AI agents



Today, we're opening a research preview of Claude Duet, a new app where Claude works as a proactive agent. It comes to you with what needs your attention, completes tasks you ask of it, keeps working even when you step away, and checks with you before it acts on your behalf. We built Claude Duet for professionals owners with more on their plates than they

ClaudeAI-mod-bot App MOD • 4d ago • Stickied comment • Edited 4d ago

Wilson, lead ClaudeAI modbot

TL;DR of the discussion generated automatically after 50 comments.

The consensus in this thread is that you're right, OP—Claude Duet feels pretty underbaked and useless for most existing users.

Development

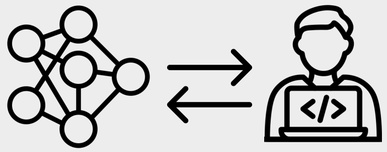


Base Model

Post-training

Static Evaluation

Deployment



Interactive Eval

Feedback

Interface

Building Collaborative LLM Agents

Measure model performance in collaborative contexts

- Reframing model evaluation as the scalability-realism frontier
- We need more work to bridge static and interactive evaluation

Learn from interaction in collaborative contexts

- Reframing learning from human feedback towards task completion
- Much value lies in learning from implicit human-AI interaction signal

Optimize interfaces for sustainable interactions

Development

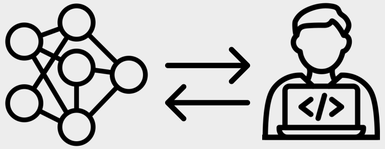


Base Model

Post-training

Static Evaluation

Deployment



Interactive Eval

Feedback

Interface

Building Collaborative LLM Agents

Measure model performance in collaborative contexts

- Reframing model evaluation as the scalability-realism frontier
- We need more work to bridge static and interactive evaluation

Learn from interaction in collaborative contexts

- Reframing learning from human feedback towards task completion
- Much value lies in learning from implicit human-AI interaction signal

Optimize interfaces for sustainable interactions

- Reframing interface design as reducing prompting and verifiability effort
- Proactive AI interfaces can improve interaction but needs careful design

Development

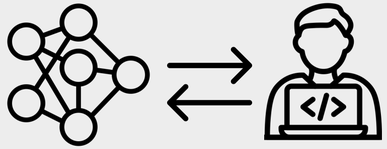


Base Model

Post-training

Static Evaluation

Deployment



Interactive Eval

Feedback

Interface

Building Collaborative LLM Agents

Measure model performance in collaborative contexts

- Reframing model evaluation as the scalability-realism frontier
- We need more work to bridge static and interactive evaluation

Learn from interaction in collaborative contexts

- Reframing learning from human feedback towards task completion
- Much value lies in learning from implicit human-AI interaction signal

Optimize interfaces for sustainable interactions

- Reframing interface design as reducing prompting and verifiability effort
- Proactive AI interfaces can improve interaction but needs careful design