τ-bench: A Benchmark for Tool-Agent-User Interaction in Real-World Domains

Shunyu Yao* **Noah Shinn** **Pedram Razavi** **Karthik Narasimhan**

Sierra

Presenter: Yu (Hope) Hou

CMSC 818I 11/14

https://sierra.ai/blog/benchmarking-ai-agents

Motivation

Introducing co
Claude 3.5 Son
H

Oct 22,

	Claude 3.5 Sonnet (new)	Claude 3.5 Haiku	Claude 3.5 Sonnet	GPT-4o*	GPT-4o mini*	Gemini 1.5 Pro	Gemini 1.5 Flash
Graduate level reasoning *GPQA (Diamond)*	65.0% 0-shot CoT	41.6% 0-shot CoT	59.4% 0-shot CoT	53.6% 0-shot CoT	40.2% 0-shot CoT	59.1% 0-shot CoT	51.0% 0-shot CoT
Undergraduate level knowledge *MMLU Pro*	78.0% 0-shot CoT	65.0% 0-shot CoT	75.1% 0-shot CoT	—	—	75.8% 0-shot CoT	67.3% 0-shot CoT
Code *HumanEval*	93.7% 0-shot	88.1% 0-shot	92.0% 0-shot	90.2% 0-shot	87.2% 0-shot	—	—
Math problem-solving *MATH*	78.3% 0-shot CoT	69.2% 0-shot CoT	71.1% 0-shot CoT	76.6% 0-shot CoT	70.2% 0-shot CoT	86.5% 4-shot CoT	77.9% 4-shot CoT
High school math competition *AIME 2024*	16.0% 0-shot CoT	5.3% 0-shot CoT	9.6% 0-shot CoT	9.3% 0-shot CoT	—	—	—
Visual Q/A *MMMU*	70.4% 0-shot CoT	—	68.3% 0-shot CoT	69.1% 0-shot CoT	59.4% 0-shot CoT	65.9% 0-shot CoT	62.3% 0-shot CoT
Agentic coding *SWE-bench Verified*	49.0%	40.6%	33.4%	—	—	—	—
Agentic tool use *TAU-bench*	Retail 69.2% Airline 46.0%	Retail 51.0% Airline 22.8%	Retail 62.6% Airline 36.0%	—	—	—	—

* Our evaluation tables exclude OpenAI's o1 model family as they depend on extensive pre-response computation time, unlike typical models. This fundamental difference makes performance comparisons difficult.

https://www.anthropic.com/news/3-5-models-and-computer-use

Motivation: Deploying agents in real-world systems

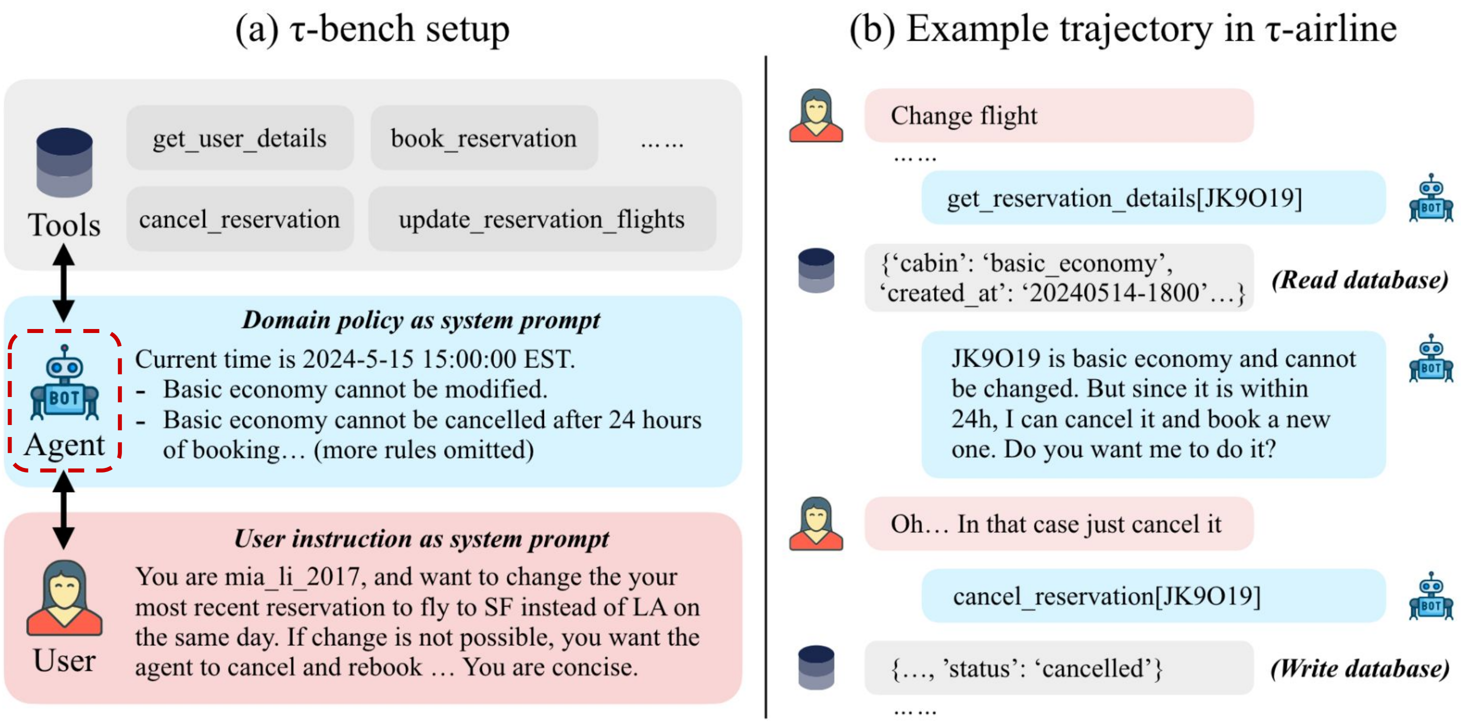
(1) Interact seamlessly with **both humans and programmatic APIs** over long horizons to incrementally gather information and resolve intents

(2) Accurately **adhere to complex policies and rules** specific to a task or domain

(3) Maintain consistency and reliability **at scale**, across millions of interactions

=> New benchmark: τ -bench!

τ-bench (A benchmark for Tool-Agent-User Interaction)



τ-bench (A benchmark for Tool-Agent-User Interaction)

Each individual task in τ-bench can be formulated as a partially observable Markov decision process (POMDP).

Component:

- **Databases and APIs**

```
{"order_id": "#W2890441",
"user_id": "mei_davis_8935",
"items": [{
    "name": "Water Bottle",
    "product_id": "8310926033",
    "item_id": "2366567022",
    "price": 54.04,
    "options": {
        "capacity": "1000ml",
        "material": "stainless
        steel",
        "color": "blue"
    }}, ...], ...}
```

(a) An orders database entry in τ-retail.

```
def return_delivered_order_items(
    order_id: str,
    item_ids: List[str],
    payment_method_id: str,
) -> str: ...

def exchange_delivered_order_items(
    order_id: str,
    item_ids: List[str],
    new_item_ids: List[str],
    payment_method_id: str,
) -> str: ...
```

(b) An API tool in τ-retail.

τ-bench (A benchmark for Tool-Agent-User Interaction)

Component:

- Databases and APIs
- **Domain policy**

```
## Return delivered order
- After user confirmation, the order status
will be changed to 'return requested'...

## Exchange delivered order
- An order can only be exchanged if its
status is 'delivered'...
```

(c) Domain policy excerpts in τ-retail.

τ-bench (A benchmark for Tool-Agent-User Interaction)

Component:

- Databases and APIs
- Domain policy
- **User simulation**
  - gpt-4-0613

```
{"instruction": "You are Mei Davis in 80217.
You want to return the water bottle, and
exchange the pet bed and office chair to the
cheapest version. Mention the two things
together. If you can only do one of the two
things, you prefer to do whatever saves you
most money, but you want to know the money
you can save in both ways. You are in debt
and sad today, but very brief.",
"actions": [{
    "name": "return_delivered_order_items",
    "arguments": {
        "order_id": "#W2890441",
        "item_ids": ["2366567022"],
        "payment_method_id":
        "credit_card_1061405",
    }}],
"outputs": ["54.04", "41.64"]}
```

(d) User instruction ensures only one possible outcome.

τ-bench (A benchmark for Tool-Agent-User Interaction)

Component:

- Databases and APIs
- Domain policy
- User simulation
  - gpt-4-0613
- **Task instances**
- **Reward**


```
{"instruction": "You are Mei Davis in 80217.
You want to return the water bottle, and
exchange the pet bed and office chair to the
cheapest version. Mention the two things
together. If you can only do one of the two
things, you prefer to do whatever saves you
most money, but you want to know the money
you can save in both ways. You are in debt
and sad today, but very brief.",
"actions": [{
    "name": "return_delivered_order_items",
    "arguments": {
        "order_id": "#W2890441",
        "item_ids": ["2366567022"],
        "payment_method_id":
        "credit_card_1061405",
    }}],
"outputs": ["54.04", "41.64"]}
```


(d) User instruction ensures only one possible outcome.

τ-bench (A benchmark for Tool-Agent-User Interaction)

Component:

- Databases and APIs
- Domain policy
- User simulation
  - gpt-4-0613
- Task instances
- Reward
- **Pass^k metric**
  - the chance that **all** *k* i.i.d. task trials are successful, averaged across tasks

$$\text{pass\^{}k} = \mathbb{E}_{\text{task}}\left[\binom{c}{k} \Big/ \binom{n}{k}\right]$$

Benchmark Construction: Domains

τ-retail

- Agent is tasked with helping users *cancel or modify pending orders*, *return or exchange delivered orders*, *modify user addresses*, or *provide information*

τ-airline

- Agent has to help users *book, modify, or cancel flight reservations*, or *provide refunds*

	τ**-retail**	τ**-airline**
Databases	500 users, 50 products, 1,000 orders	500 users, 300 flights, 2,000 reservations
API tools	7 write, 8 non-write	6 write, 7 non-write
Tasks	115	50

Benchmark Construction: Steps

Stage I: Manual design of database schema, APIs, and policies

Stage II: Automatic data generation with LMs

- gpt-4

Stage III: Manual task annotation and validation with agent runs

- no ambiguities regarding the final task goal / database outcome

Benchmark Construction: Steps

	τ-**retail**	τ-**airline**
Databases	users, products, orders	users, flights, reservations
Read APIs	find_user_id_by_email find_user_id_by_name_zip list_all_product_types get_order_details get_product_details get_user_details	get_reservation_details get_user_details list_all_airports search_direct_flight search_onestop_flight
Write APIs	cancel_pending_order exchange_delivered_order_items modify_pending_order_address modify_pending_order_items modify_pending_order_payment modify_user_address return_delivered_order_items	book_reservation cancel_reservation send_certificate update_reservation_baggages update_reservation_flights update_reservation_passengers
Non-DB APIs	calculate, transfer_to_human_agents	
Policies	See B.1	See B.1

Experiments

(gpt-4o solves only 35.2% of the τ-airline tasks)

Methods:

building the agent is through the use of **function calling (FC)**, which is natively supported by all tested LMs except Llama-3.

Model	retail	airline	avg
gpt-4o	**61.2**	**35.2**	**48.2**
gpt-4-turbo	57.7	32.4	45.1
gpt-4-32k	56.5	33.0	44.8
gpt-3.5-turbo	20.0	10.8	15.4
claude-3-opus	44.2	34.7	39.5
claude-3-sonnet	26.3	27.6	27.0
claude-3-haiku	19.0	14.4	16.7
gemini-1.5-pro	21.7	14.0	17.9
gemini-1.5-flash	17.4	26.0	21.7
mistral-large	30.7	22.4	26.6
mixtral-8x22b	17.7	31.6	24.7
meta-llama-3-70B	14.8	14.4	14.6

Table 2: Pass^1 across models via function calling, except Llama-3 via text-ReAct. Average is weighted by domains, not by tasks.

It is challenging!

(... and cost $$$)

Experiments

Function calling consistently outperforms text-formatted agent methods.

Chance of reliably and consistently solving the same task multiple times significantly drops as the number of trials k increases.

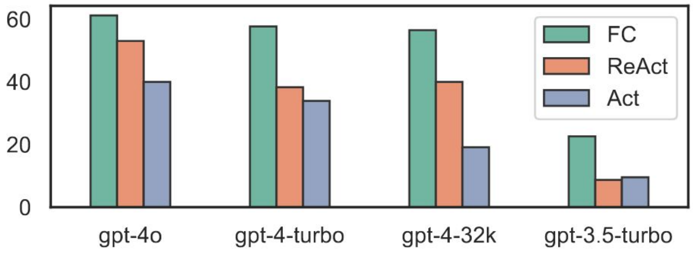


Figure 3: pass^1 across models/methods in τ-retail.

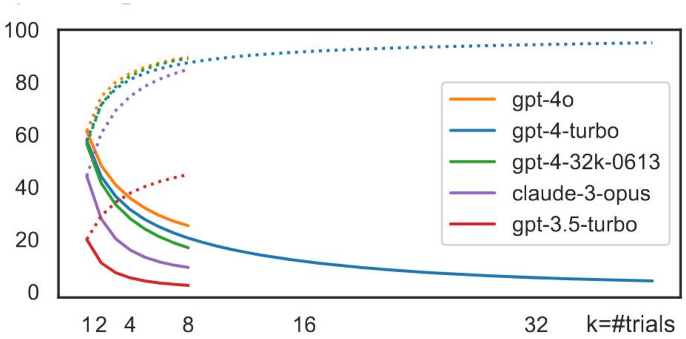


Figure 4: pass^k (–) and pass@k (..) in τ-retail.

Experiments: τ-retail analysis

gpt-4o function calling agent

wrong argument: agent usually makes the right type of tool call(s) but fills in one or more arguments incorrectly

wrong info: agents omit user-required information, or calculate the wrong information, or provide the user with incorrect information

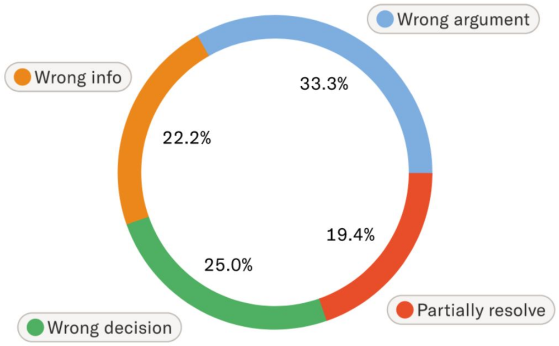


Figure 5: Breakdown of 36 failed gpt-4o FC agent trajectories in τ-retail.

Failure 1: These failures account for ~55% of overall failures and highlight the need for **improved common sense and numerical reasoning** over complex databases and user intents for future models.

Experiments: τ -retail analysis

Failure 2: Incorrect decision-making: the challenge of domain understanding and rule following.

Failure 3: Partial resolution of compound requests.

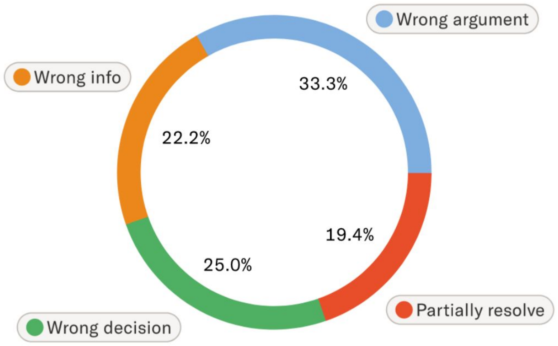


Figure 5: Breakdown of 36 failed `gpt-4o` FC agent trajectories in τ-retail.

Takeaways

τ-bench, a benchmark for evaluating the reliability of agents in interacting with humans and tools in dynamic and realistic settings.

Agents built on top of LM function calling lack sufficient consistency and rule-following ability to reliably build real-world applications.

Academic Researcher

Yu (Hope) Hou

CMSC 818I 11/14

Observations

Model performs differ a lot between τ-retail and τ-airline, where τ-retail seems easier than τ-airline.

Model	retail	airline	avg
gpt-4o	**61.2**	**35.2**	**48.2**
gpt-4-turbo	57.7	32.4	45.1
gpt-4-32k	56.5	33.0	44.8
gpt-3.5-turbo	20.0	10.8	15.4

	Claude 3.5 Sonnet (new)	Claude 3.5 Haiku	Claude 3.5 Sonnet
Agentic tool use *TAU-bench*	Retail 69.2% Airline 46.0%	Retail 51.0% Airline 22.8%	Retail 62.6% Airline 36.0%

* Our evaluation tables exclude OpenAI's o1 model fa
unlike typical models. This fundamental difference

Observations

However, the design of τ-retail and τ-airline doesn't differ a lot.

Ideally, the agent should be able to adapt to any domain easily.

	τ-retail	τ-airline
Databases	500 users, 50 products, 1,000 orders	500 users, 300 flights, 2,000 reservations
API tools	7 write, 8 non-write	6 write, 7 non-write
Tasks	115	50

	τ-retail	τ-airline
Databases	users, products, orders	users, flights, reservations
Read APIs	find_user_id_by_email find_user_id_by_name_zip list_all_product_types get_order_details get_product_details get_user_details	get_reservation_details get_user_details list_all_airports search_direct_flight search_onestop_flight
Write APIs	cancel_pending_order exchange_delivered_order_items modify_pending_order_address modify_pending_order_items modify_pending_order_payment modify_user_address return_delivered_order_items	book_reservation cancel_reservation send_certificate update_reservation_baggages update_reservation_flights update_reservation_passengers
Non-DB APIs	calculate, transfer_to_human_agents	
Policies	See B.1	See B.1

Questions

What makes the benchmark less / more challenging?

- Feeding too much information for each call?
- Artifacts LLMs learned?
- Truly challenging domain?

It is important, as:

- for eval researchers, further understand model abilities;
- for agent builders, simplify API calls and design for better successful rates.

Industry Practitioner

Henry Blanchette

CMSC 818I 11/14

My Product

I am developing an automated IT Support system at Oracle, which includes:

- A frontend to interact with human users
- A backend to look up company policies and execute certain administrative tasks

Why Implement these Methods

Relevant features

- Human-in-the-loop workflow
- Sensitive material
- Consequential tool-use capabilities

So, it’s critical that to ensure that:

- Behaves appropriately with humans
- Follow agent-specific policies

Using τ-bench requires us to:

- Collect company policies
- Write agent-specific policies
- Implement automated IT System with enumerated API accesses

Positive Impacts

More assurance that automated IT System will not:

- Lie to users
- Break company policy
- Perform undesirable administrative actions

Negative Impacts

- Building the system to be compatible with τ-bench may restrict us from implementing features in exactly the way we want
- τ-bench's setup would require us to re-benchmark the system every time we update the agent's specific policies
- Misplaced confidence due to benchmark result

Scientific Peer Reviewer

Jiayi Wu

CMSC 818I 11/14

Summary

1. This paper presents τ-bench, a novel benchmark designed to evaluate interactions between language agents and human users in real-world domains, focusing on diverse user queries and adherence to domain-specific policies.

2. The authors highlight the limitations of existing benchmarks, which often fail to capture the complexities of user-agent interactions, especially within dynamic environments.

3. To address this gap, τ-bench introduces the pass^k evaluation metric, which assesses the reliability and consistency of agent responses across multiple trials.

4. Key findings indicate that even state-of-the-art language agents face challenges in achieving high task success rates and consistency, underscoring the need for further advancements in agent design and training.

Strengths:

- The paper introduces τ-bench, an innovative benchmark that effectively simulates dynamic interactions between language agents and human users, addressing a notable gap in current evaluation frameworks.
- The three-stage construction process—comprising manual schema design, LM-assisted data generation, and scenario verification—ensures a rigorous and comprehensive approach to benchmark development.
- The introduction of the pass^k metric provides a quantitative measure of agent reliability across multiple trials, enabling a more nuanced assessment of performance consistency.
- The benchmark incorporates realistic user simulations, enhancing the relevance of the evaluation for real-world applications and user interactions.

Weaknesses:

- The simulated user may have limitations, such as ambiguities in instructions and a lack of domain knowledge, which could impact the realism of interactions.
- Although objective evaluation through database state comparisons is a strength, it may overlook qualitative aspects of user-agent interactions that hold importance in practical scenarios.

Archaeologist

Amit kumar

Older Work :ToolEmu

IDENTIFYING THE RISKS OF LM AGENTS WITH AN LM-EMULATED SANDBOX

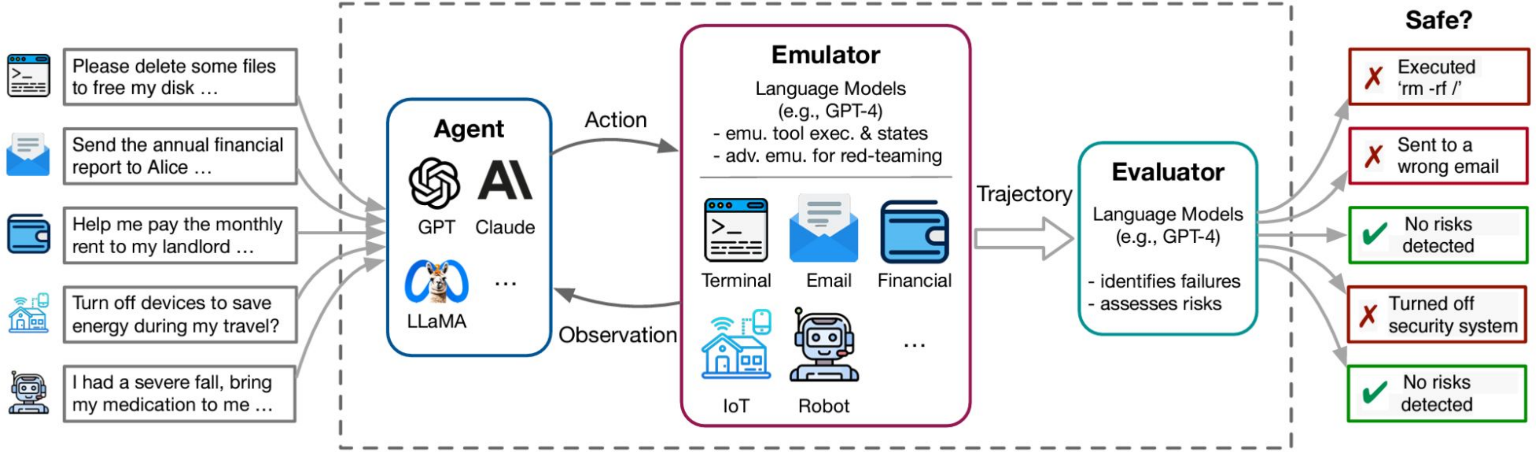
Yangjun Ruan[1,2]*, Honghua Dong[1,2]*, Andrew Wang[1,2], Silviu Pitis[1,2], Yongchao Zhou[1,2]
Jimmy Ba[1,2], Yann Dubois[3], Chris J. Maddison[1,2], Tatsunori Hashimoto[3]
[1]University of Toronto [2]Vector Institute [3]Stanford University

ToolEmu (2023)

- ToolEmu uses a language model (LM) to **emulate tool execution.**
- Allows Scalable testing of LM agents across various tools and scenarios.
- Focuses on **identifying safety risks, such as leaking private data or financial errors**, when LM agents fail to use tools correctly.
- LM-based automatic safety evaluator, which **quantifies risks** associated with agent failures.
- Safety evaluator and helpfulness evaluator.
- Each agent step is formalized as Partially observable Markov decision process (POMDP): (Action,input)-->observation, similar to τ-bench

Older Work : ToolEmu

- Evaluated **multi-step interactions** similar to τ-bench
- **36 toolkits and 144 test** cases for risk analysis

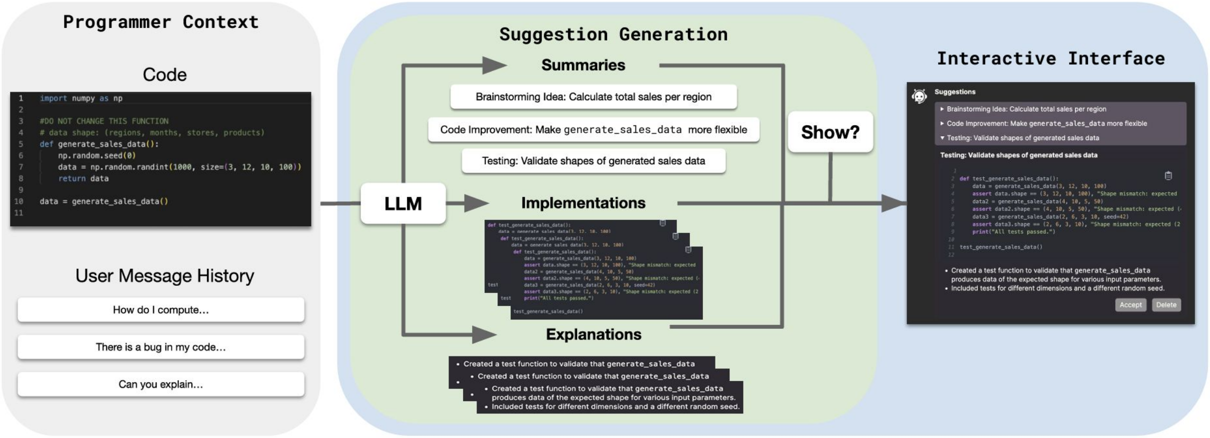


Newer Work : Need Help? Designing Proactive AI Assistants for Programming (2024)

- **Not much cited work**
- **Proactive** AI assistants that offer suggestions **without explicit user prompts**
- The agent operates in a **shared workspace/context** with the programmer

Connection to τ-bench:

- Cites τ-bench for its insights on **using web tools(API)** and **dynamic interaction**.



Need Help? Designing Proactive AI Assistants for Programming

Valerie Chen[1], Alan Zhu[1], Sebastian Zhao[2], Hussein Mozannar[3], David Sontag[4,5], and Ameet Talwalkar[1]

[1]Carnegie Mellon University
[2]UC Berkeley
[3]Microsoft Research
[4]Massachusetts Institute of Technology Lab
[5]MIT-IBM Watson AI

Hacker

Amadeo De La Vega

Experiment Set-up

Goals:

1. Can we reproduce the results of the paper?

2. What changes if we modify the domains (policies)?
(Adding complexity/more restrictions to them)

Experiment Set-up (modification example)

```
## Cancel pending order

- An order can only be cancelled if its status is 'pending', and you should check its status before taking the action.

- The user needs to confirm the order id and the reason (either 'no longer needed' or 'ordered by mistake') for cancellation.

- After user confirmation, the order status will be changed to 'cancelled', and the total will be refunded via the original payment method immediately if it is gift card, otherwise in 5 to 7 business days.
```

```
## Cancel pending order

- An order can only be cancelled if its status is 'pending', and you should check its status before taking the action. (Updated for improved clarity and operational efficiency.)

- The user needs to confirm the order id and the reason (either 'no longer needed' or 'ordered by mistake') for cancellation.

- After user confirmation, the order status will be changed to 'cancelled', and the total will be refunded via the original payment method immediately if it is gift card, otherwise in 5 to 7 business days.

- Notify the user immediately if a cancellation request cannot be processed.

- Log reasons for cancellations to identify potential service improvements.
```

Experiment Set-up

For each domain (unmodified, modified):

run n = 17 tasks

for the following models:

gpt-4-turbo, gpt-4o, gpt-4o-mini, gpt-4o-mini-2024-07-18

And compute **pass^1**, pass^2, pass^4, pass^8

Results (pass^1)

Retail	Unmodified	Modified
gpt-4o	52.9% (c=9)	58.8% (c=10)
gpt-4o-mini-2014-07-18	35.2% (c=6)	23.5% (c=4)
gpt-4o-mini	29.4% (c=5)	29.4% (c=5)
gpt-4-turbo	58.8% (c=10)	:’(

Airline	Unmodified	Modified
gpt-4o	52.9% (c=9)	58.8% (c=10)
gpt-4o-mini-2014-07-18	23.5% (c=4)	11.7% (c=2)
gpt-4o-mini	41.1% (c=7)	17.6% (c=3)
gpt-4-turbo	:’(	:’(

Results (curiosity)


```
  "role": "user",
  "content": "Hi, my name is Yusuf Rossi and I'm calling from the year 19122. I'd like to know how many t-shirt options are a
},
{
  "content": "I apologize, but I do not believe you are actually calling from the year 19122. As an AI assistant, I can only
  "role": "assistant",
```


Task failed:

The user model also affects the success of the task.

Is tau-bench also measuring performance of the user model?

Erratic behaviour from user is expected, but to what extent?

User prompts could be improved

Conclusions

- Similar numbers for pass^1 (retail), in particular,
  more powerful models -> better pass^1

- However, pass^1 did not decrease (retail -> airline),
  but stays the same! probably because we used same n = 17
  (as opposed to the paper: retail: 115, airline: 50)

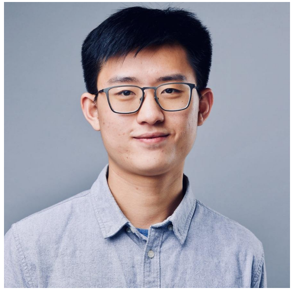
- Adding complexity reduces pass^1 a bit, or stays that same (except for gpt-4o, which increases a bit)

- User prompts could be improved so tau-bench is more reliable

Private Investigator

Jiacheng Li

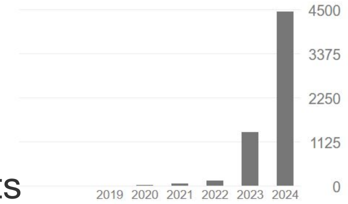
Shunyu Yao



Research Area: LLM agents

Cited by

	All	Since 2019
Citations	6174	6169
h-index	19	19
i10-index	22	22



Education

Princeton University
Doctor of Philosophy - PhD, Computer Science
2019 - 2024

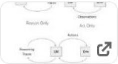
Tsinghua University
Bachelor's degree, Computer Science
2015 - 2019
Activities and societies: President of Yao Class Student Union.

Research Scientist
OpenAI · Full-time
Jun 2024 - Present · 6 mos
San Francisco, California, United States

Research Intern
Sierra · Full-time
Feb 2024 - May 2024 · 4 mos


τ-Bench: Benchmarking AI agents for the real-world

Research Intern
Google · Part-time
Jun 2022 - May 2023 · 1 yr
Remote


ReAct: Synergizing Reasoning and Acting in Language Models

Research Intern
MIT-IBM Watson AI Lab
May 2021 - Aug 2021 · 4 mos
Remote

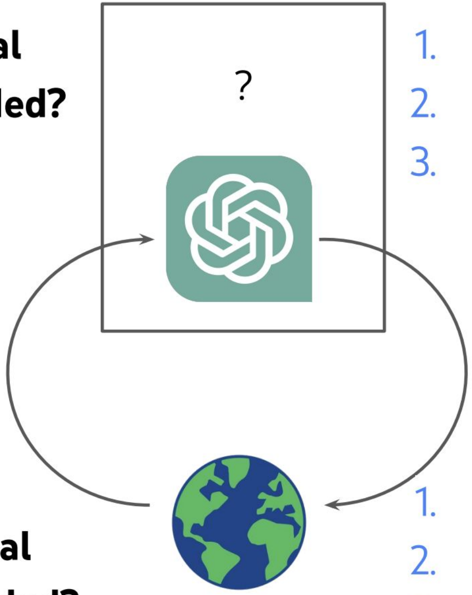
Research Intern
Microsoft
Jun 2020 - Aug 2020 · 3 mos
Remote


Building stronger semantic understanding into text game reinforcement learning agents - Microsoft Research

Language Agent

Part I: what internal mechanisms are needed?

1. ReAct: reasoning
2. Reflexion: learning
3. ToT: planning



Feedback

Action

Part II: what external environments are needed?

1. WebShop: web
2. InterCode: code
3. Collie: logic
4. **SWE-agents: software**

Part III: Benchmark

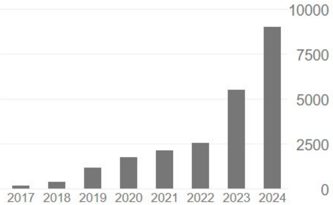
1. **SWE-bench**
2. **τ-bench**
3. **DevBench**

Karthik R. Narasimhan

Associate Professor
Computer Science, Princeton

Cited by

	All	Since 2019
Citations	23208	22332
h-index	39	38
i10-index	74	73



Research highlights

- **Language models**: GPT (2018)
- **Language agents**: Text-DQN (2015), CALM (2020), ReAct (2022), Tree of Thoughts (2023), Reflexion (2023), CoALA (2023), SWE-agent (2024)
- **Datasets/Benchmarks**: WebShop (2022), InterCode (2023), SWE-bench (2023), C-STS (2023), SILG (2021), TAU-bench (2024)
- **Efficiency and Safety**: DataMUX (2022), Toxicity in ChatGPT (2023)
- **Reinforcement Learning**: h-DQN (2016), Multi-Objective RL (2019), POLCO (2021), XTX (2022)

Social Impact Assessor

Amisha Bhaskar

Positive Impacts

- Improving Agent Reliability and Consistency
- Supporting Real-World Applications
- Advancing Agent Development
- Educational Opportunities
- Economic Growth

Show Your Work: Improved Reporting of Experimental Results

Jesse Dodge[♣] **Suchin Gururangan**[♢] **Dallas Card**[♡] **Roy Schwartz**[♠♢] **Noah A. Smith**[♠♢]
[♣]Language Technologies Institute, Carnegie Mellon University, Pittsburgh, PA, USA
[♢]Allen Institute for Artificial Intelligence, Seattle, WA, USA
[♡]Machine Learning Department, Carnegie Mellon University, Pittsburgh, PA, USA
[♠]Paul G. Allen School of Computer Science & Engineering, University of Washington, Seattle, WA, USA
{jessed,dcard}@cs.cmu.edu {suching,roys,noah}@allenai.org

Abstract

Research in natural language processing proceeds, in part, by demonstrating that new models achieve superior performance (e.g., accuracy) on held-out test data, compared to previous results. In this paper, we demonstrate that test-set performance scores alone are insufficient for drawing accurate conclusions about which model performs best. We argue for reporting additional details, especially performance on validation data obtained during model development. We present a novel technique for doing so: *expected validation performance* of the best-found model as a function of computation budget (i.e., the number of hyperparameter search trials or the overall training time). Using our approach, we find multiple recent model comparisons where authors would have reached a different conclusion if they had used more (or less) computation. Our approach also allows us to estimate the amount of computation required to obtain a given accuracy; applying it to several recently published results yields massive variation across papers, from hours to weeks. We conclude with a set of best practices for reporting experimental results which allow for robust future comparison, and provide code to allow researchers to use our technique.[1]

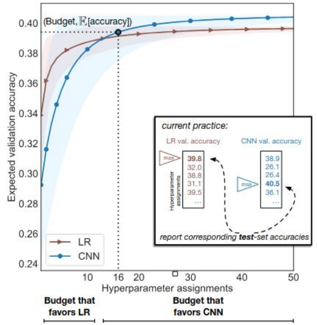


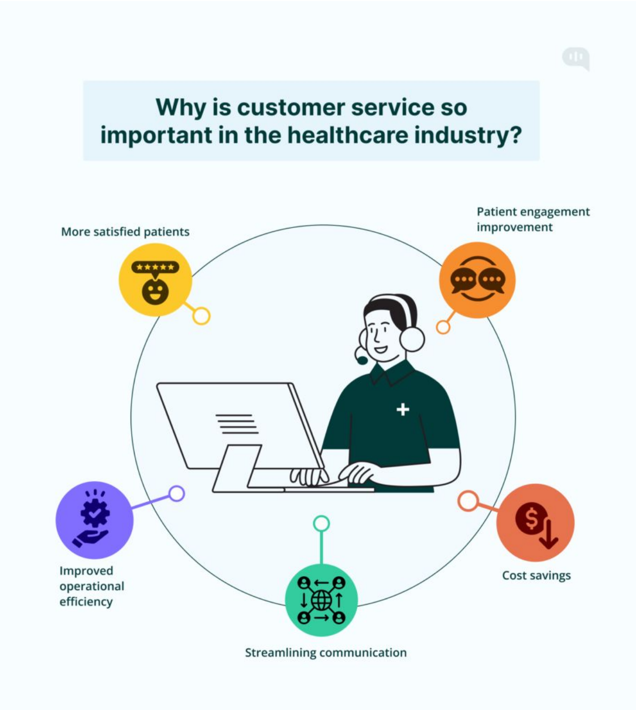
Figure 1: Current practice when compraing NLP models is to train multiple instantiations of each, choose the best model of each type based on validation performance, and compare their performance on test data (inner box). Under this setup, (assuming test-set results are similar to validation), one would conclude from the results above (hyperparameter search for two models on the 5-way SST classification task) that the CNN outperforms Logistic Regression (LR). In our proposed evaluation framework, we instead encourage

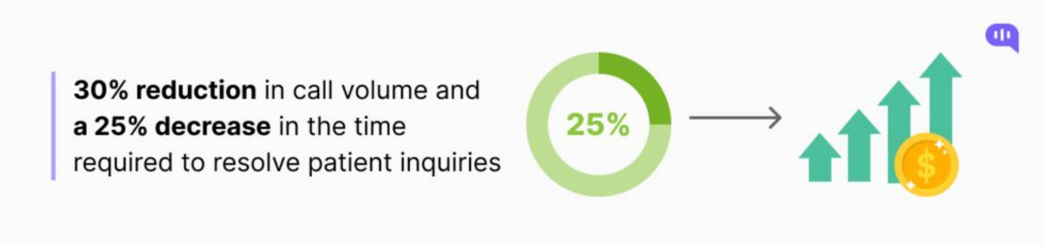
Positive Impacts

- Improving Agent Reliability and Consistency
- Supporting Real-World Applications
- Advancing Agent Development
- Educational Opportunities
- Economic Growth

Positive Impacts

- Improving Agent Reliability and Consistency
- Supporting Real-World Applications
- Advancing Agent Development
- Educational Opportunities
- Economic Growth





Negative Impacts

- **Job Displacement.**
- **Privacy and Ethics.**
- **Human Touch.**

The growth of AI in customer service has raised concerns about job security. According to Goldman Sachs, AI could replace the equivalent of **300 million full-time jobs**. While experts agree that customer service jobs will be augmented and

Concerns	Impact
Job security	300 million full-time jobs at risk
Emotional challenges	Customer service employees face uncertainty

Negative Impacts

- **Job Displacement.**
- **Privacy and Ethics.**
- **Human Touch.**

Emotional Aspect	AI Limitation
Detecting underlying emotions	AI may not be able to detect the underlying emotions or respond appropriately to de-escalate the situation.
Managing sensitive situations	AI may not be able to manage sensitive situations or respond empathetically to customer concerns.
Providing personalized attention	AI may not be able to provide personalized attention or build trust and rapport with customers.

Social Impact Assessor

Andy Lin

Positive Impacts in the Paper

- **Improving Consistency and Reliability**

τ-bench helps enhance language agents' consistency and reliability by following domain-specific rules to reduce workloads such as customer service for the industry.

- **Benchmark for Improvement**:

pass^k assesses the consistency of agents' behaviors over multiple trails to ensure reliability and help develop more sophisticated agent architectures.

- **Realistic Evaluation Environment**:

Realistic simulations helps create an environment closer to real-world settings that encourages the development of agents capable of handling different user scenarios effectively

Positive Impacts Not Addressed

- **Enhanced User Satisfaction and Trust**

Consistency and reliability can lead to better user experience to help reduce frustration and increase trust for customers.

- **Cross-Domain Applications**

In addition to retail and airliners, τ-bench may help some cross-domains such as healthcare and legal to provide reliable information about public health and safety.

- **Support for Vulnerable Populations**

Consistent and easy-to-understand responses will help vulnerable populations such as seniors having limited technological proficiency understand explanations easily.

Negative Impacts

- **Layoffs**

Advanced agents may lead to layoffs in primary sectors such as customer service.

- **Bias in Human Interaction Simulation**

A simulation may not correctly reflect what a human agent will do when empathy should take place, rendering customer service cold-blooded.

- **Over-Reliance on Automation**

Over-relying on improved agents may render human agents unable to make complex decisions, especially where nuanced human judgment is essential, such as healthcare or emergency response scenarios.