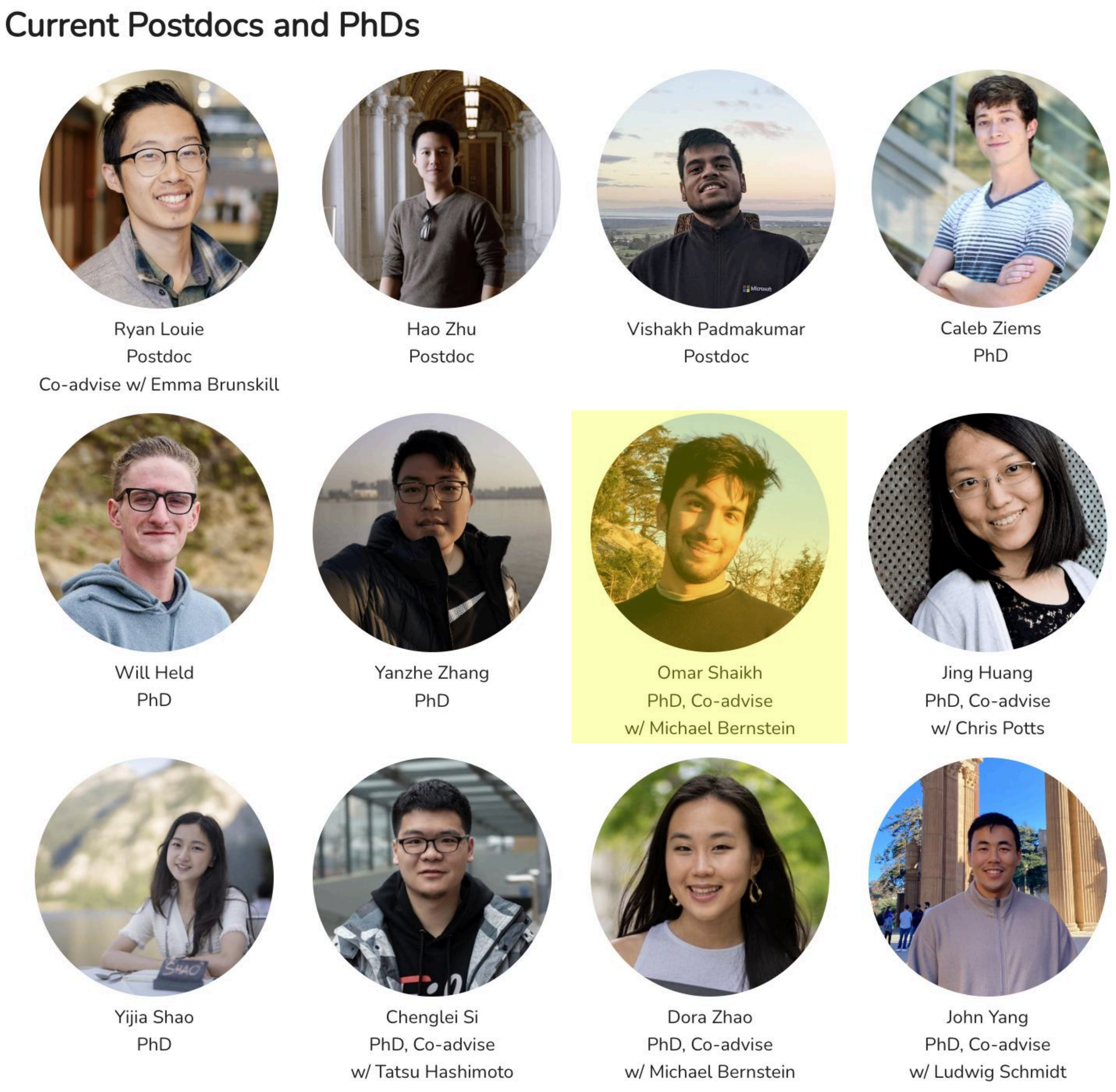
Human-AI Grounding

Omar
Shaikh

STANFORD
HUMAN-COMPUTER
INTERACTION
GROUP

Stanford | NLP



Current Postdocs and PhDs

Ryan Louie
Postdoc
Co-advise w/ Emma Brunskill

Hao Zhu
Postdoc

Vishakh Padmakumar
Postdoc

Caleb Ziems
PhD

Will Held
PhD

Yanzhe Zhang
PhD

Omar Shaikh
PhD, Co-advise
w/ Michael Bernstein

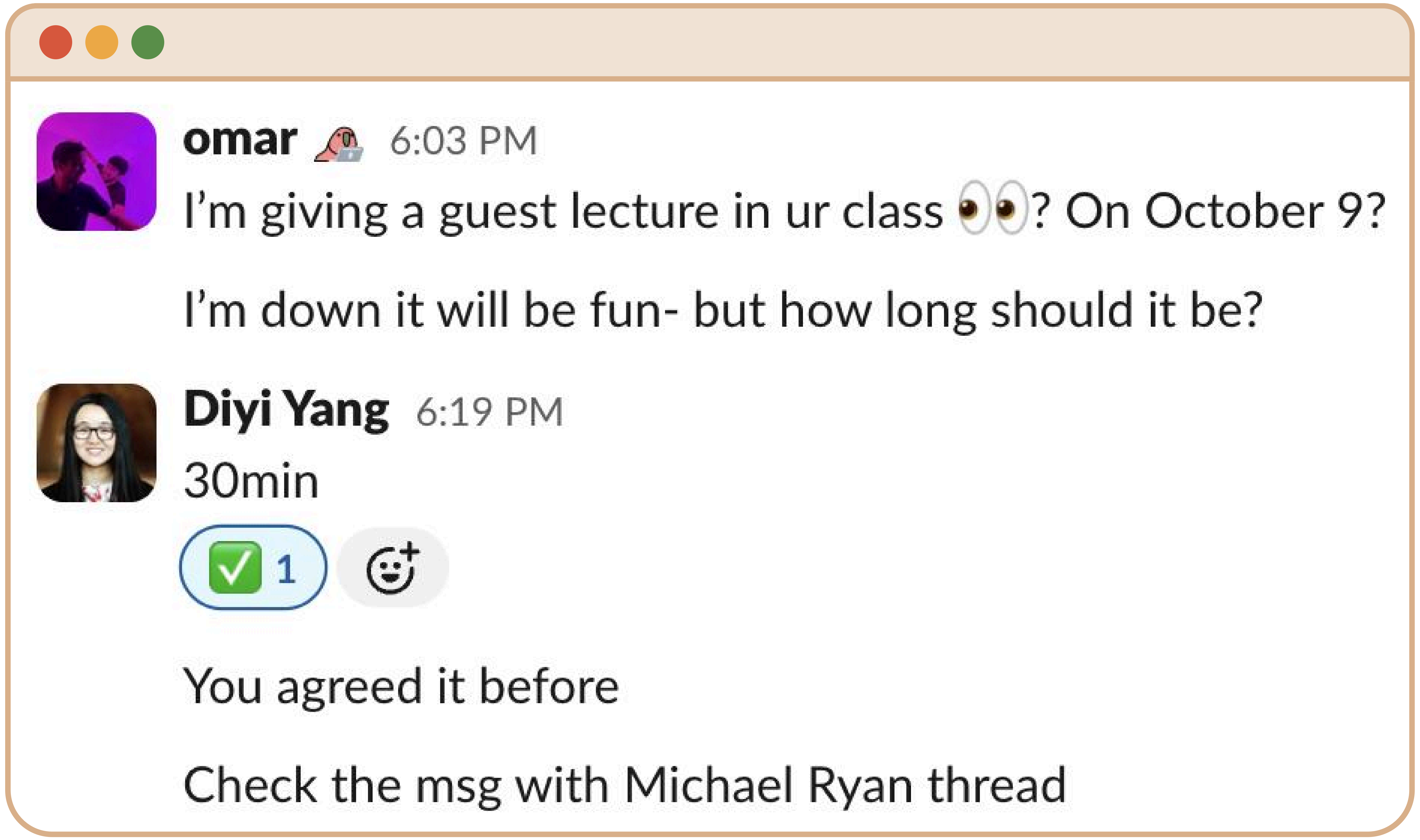
Jing Huang
PhD, Co-advise
w/ Chris Potts

Yijia Shao
PhD

Chenglei Si
PhD, Co-advise
w/ Tatsu Hashimoto

Dora Zhao
PhD, Co-advise
w/ Michael Bernstein

John Yang
PhD, Co-advise
w/ Ludwig Schmidt



omar 6:03 PM

I'm giving a guest lecture in ur class 👀? On October 9?

I'm down it will be fun- but how long should it be?

Diyi Yang 6:19 PM

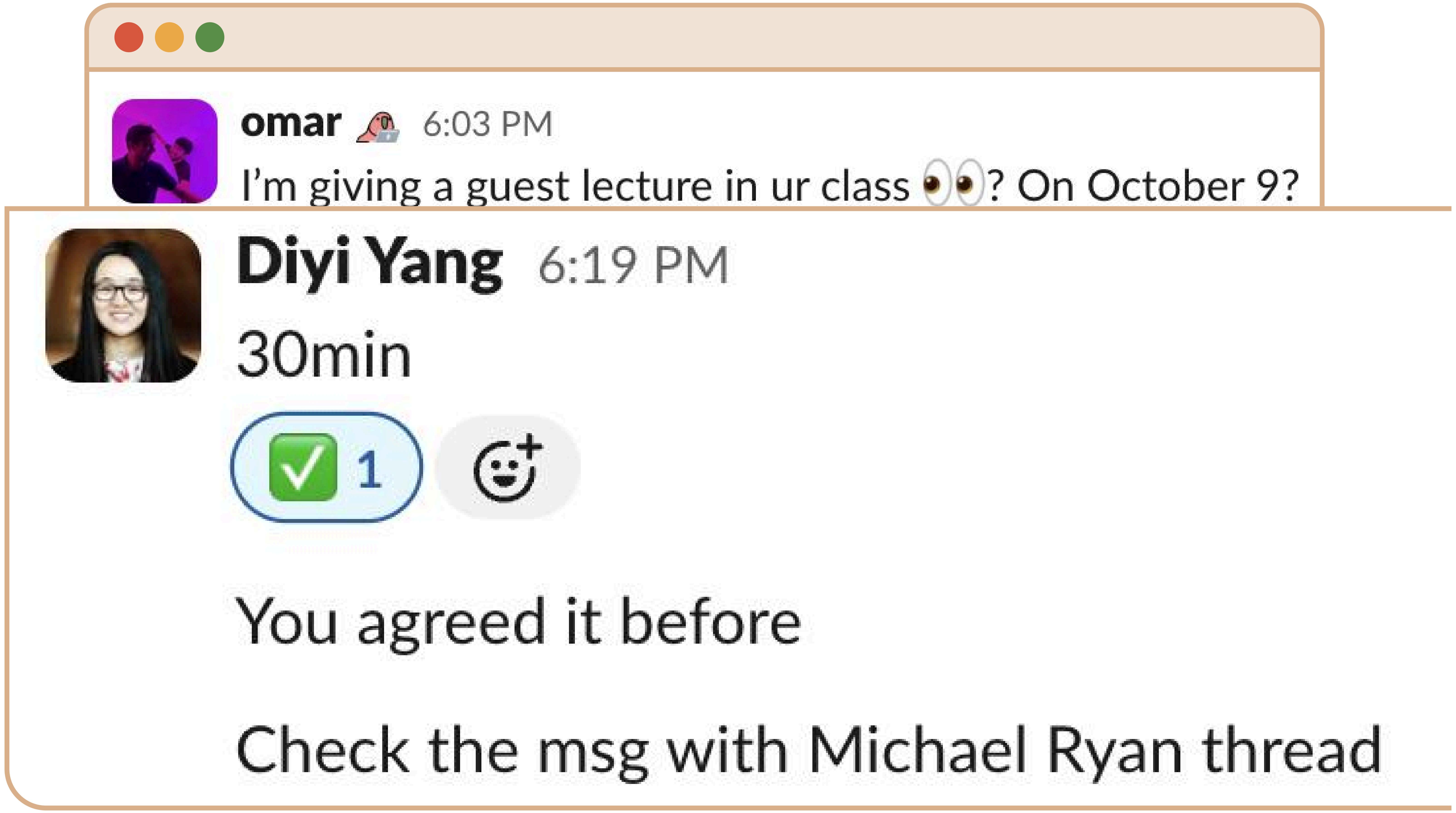
30min

✅ 1

You agreed it before

Check the msg with Michael Ryan thread

Attendance Checkin



omar 6:03 PM

I'm giving a guest lecture in ur class 👀? On October 9?

Diyi Yang 6:19 PM

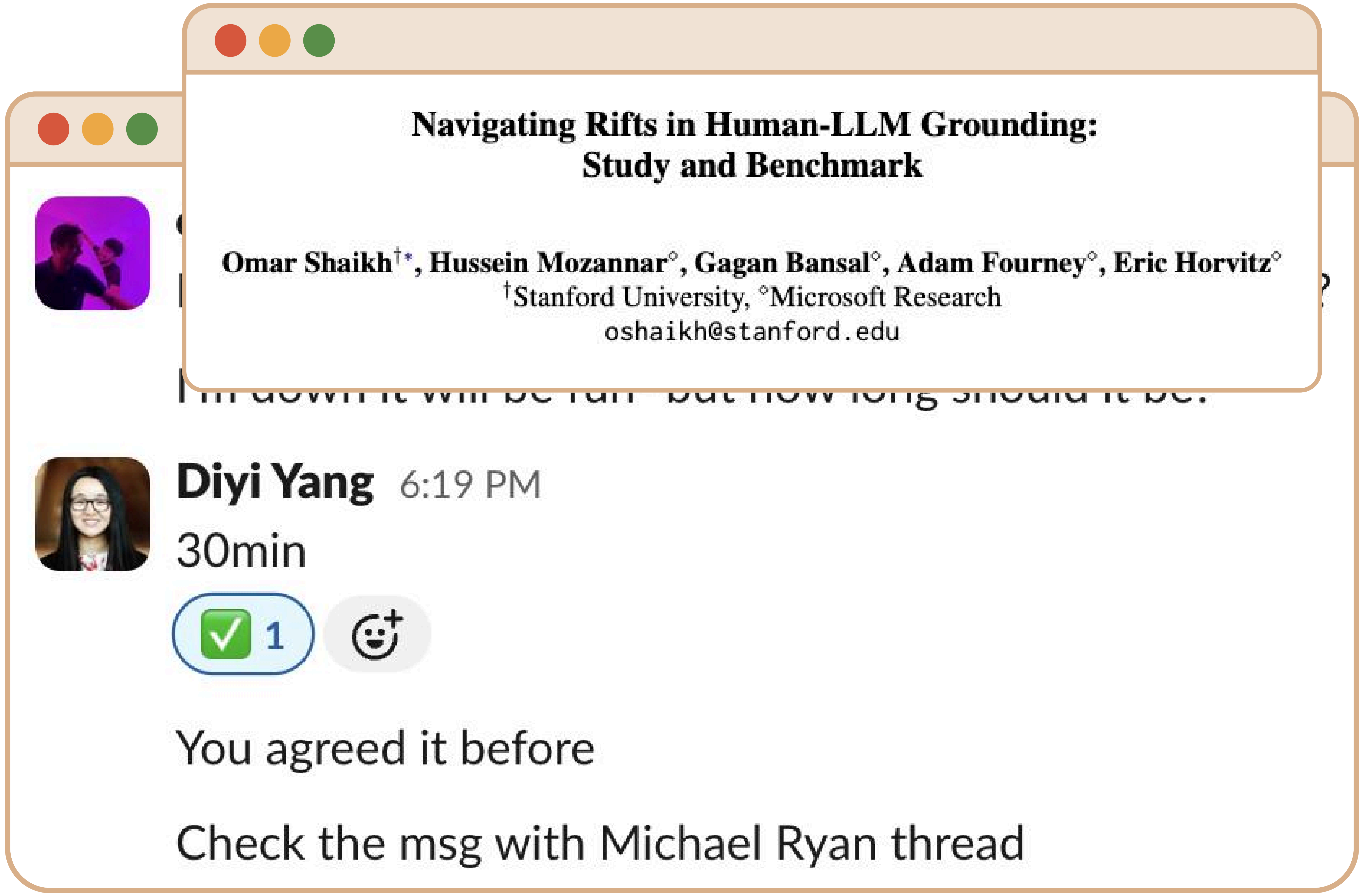
30min

✅ 1

You agreed it before

Check the msg with Michael Ryan thread

Attendance Checkin



**Navigating Rifts in Human-LLM Grounding:
Study and Benchmark**

Omar Shaikh[†*], Hussein Mozannar[◇], Gagan Bansal[◇], Adam Fourney[◇], Eric Horvitz[◇]
[†]Stanford University, [◇]Microsoft Research
oshaikh@stanford.edu

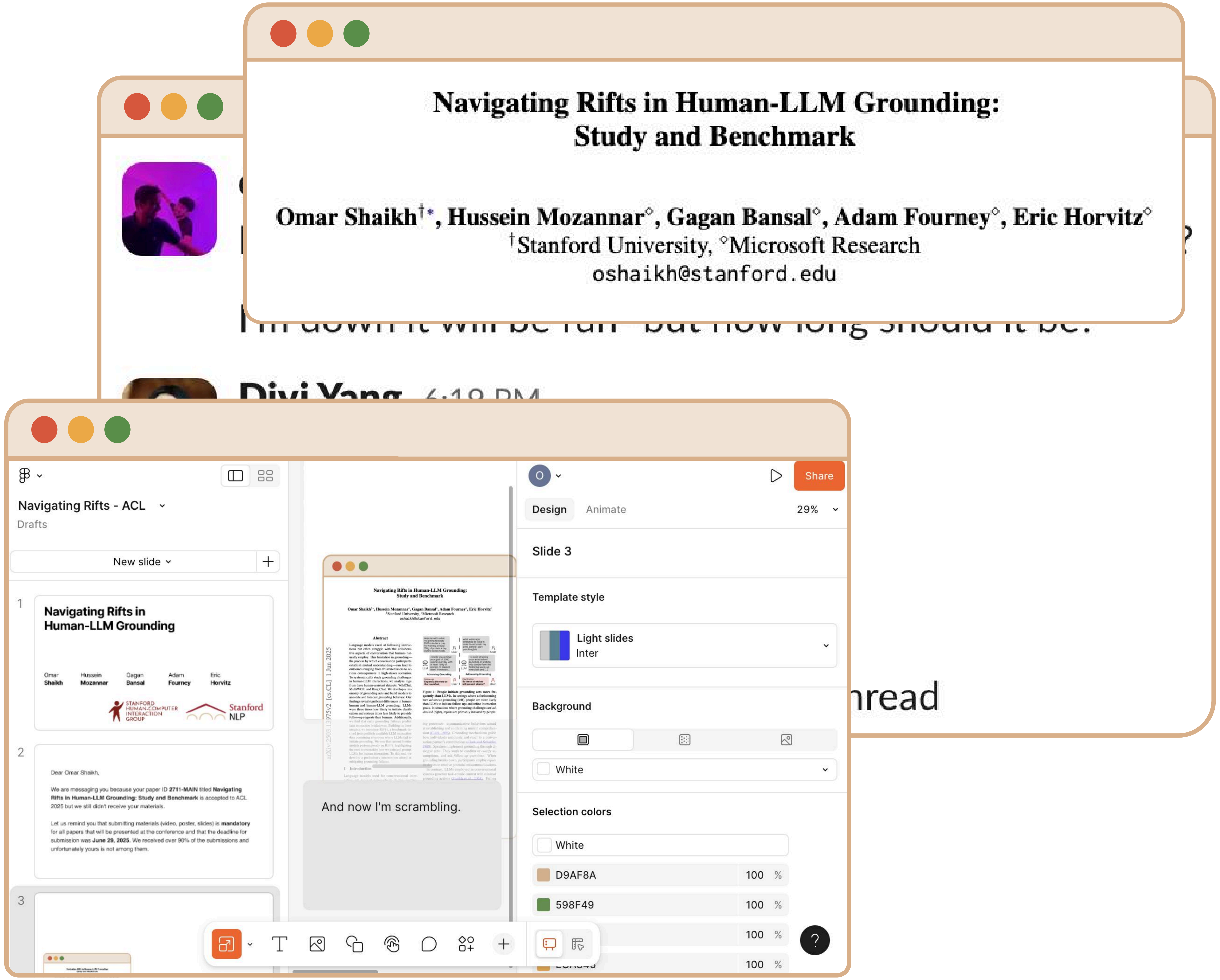
Diyi Yang 6:19 PM
30min

✅ 1

You agreed it before

Check the msg with Michael Ryan thread

Attendance Checkin



Navigating Rifts in Human-LLM Grounding: Study and Benchmark

Omar Shaikh[†*], Hussein Mozannar[◇], Gagan Bansal[◇], Adam Fourney[◇], Eric Horvitz[◇]
[†]Stanford University, [◇]Microsoft Research
oshaikh@stanford.edu

I'm down it will be fun - but how long should it be?

Diyi Yang 6:19 PM

hread

Dear Omar Shaikh,

We are messaging you because your paper ID **2711-MAIN** titled **Navigating Rifts in Human-LLM Grounding: Study and Benchmark** is accepted to ACL 2025 but we still didn't receive your materials.

Let us remind you that submitting materials (video, poster, slides) is **mandatory** for all papers that will be presented at the conference and that the deadline for submission was **June 29, 2025**. We received over 90% of the submissions and unfortunately yours is not among them.

And now I'm scrambling.

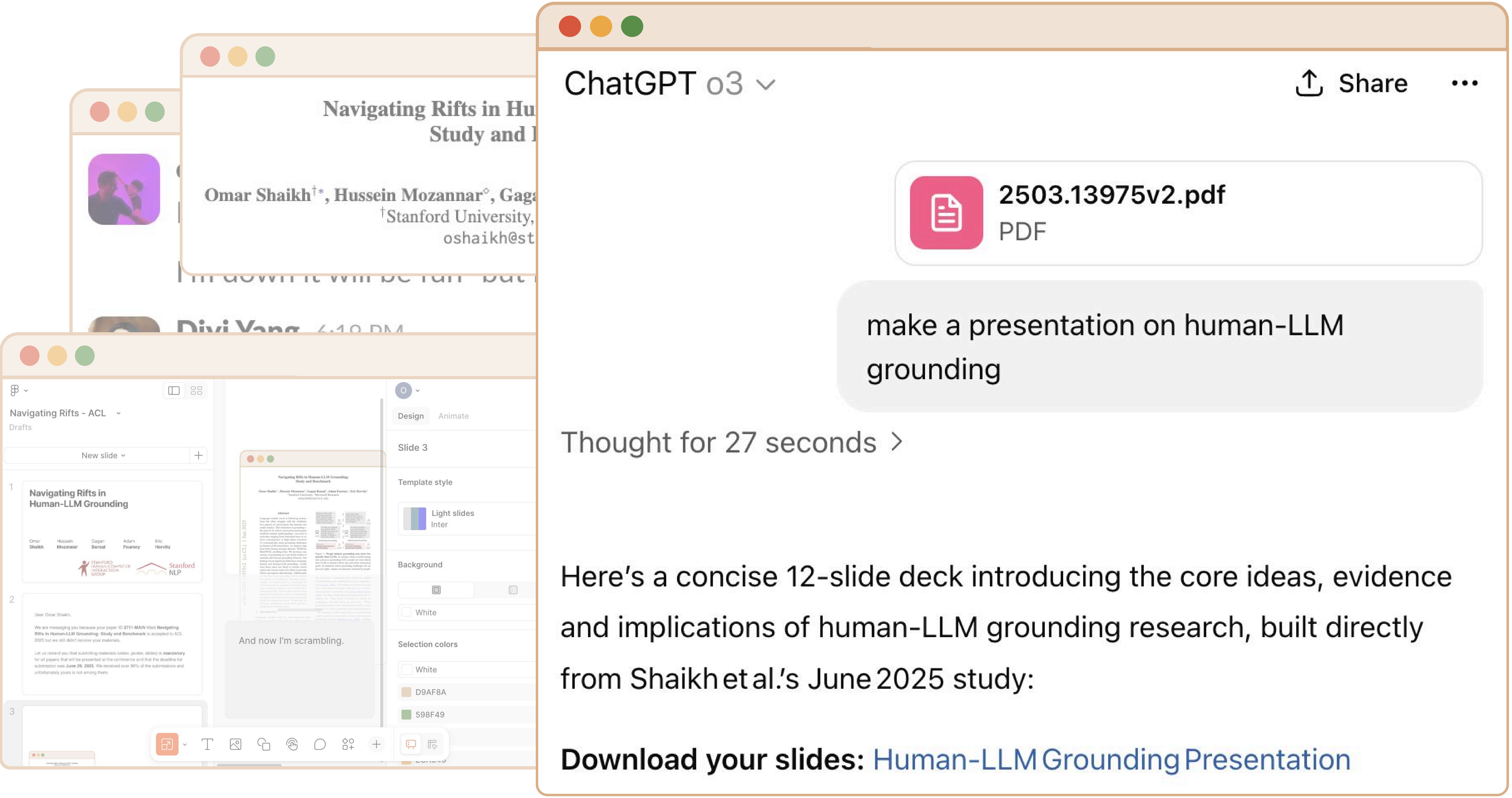
Slide 3

Template style: Light slides, Inter

Background: White

Selection colors:

White		
D9AF8A	100	%
598F49	100	%
	100	%
	100	%



make a presentation on human-LLM grounding

Thought for 27 seconds

Here's a concise 12-slide deck introducing the core ideas, evidence and implications of human-LLM grounding research, built directly from Shaikh et al.'s June 2025 study:

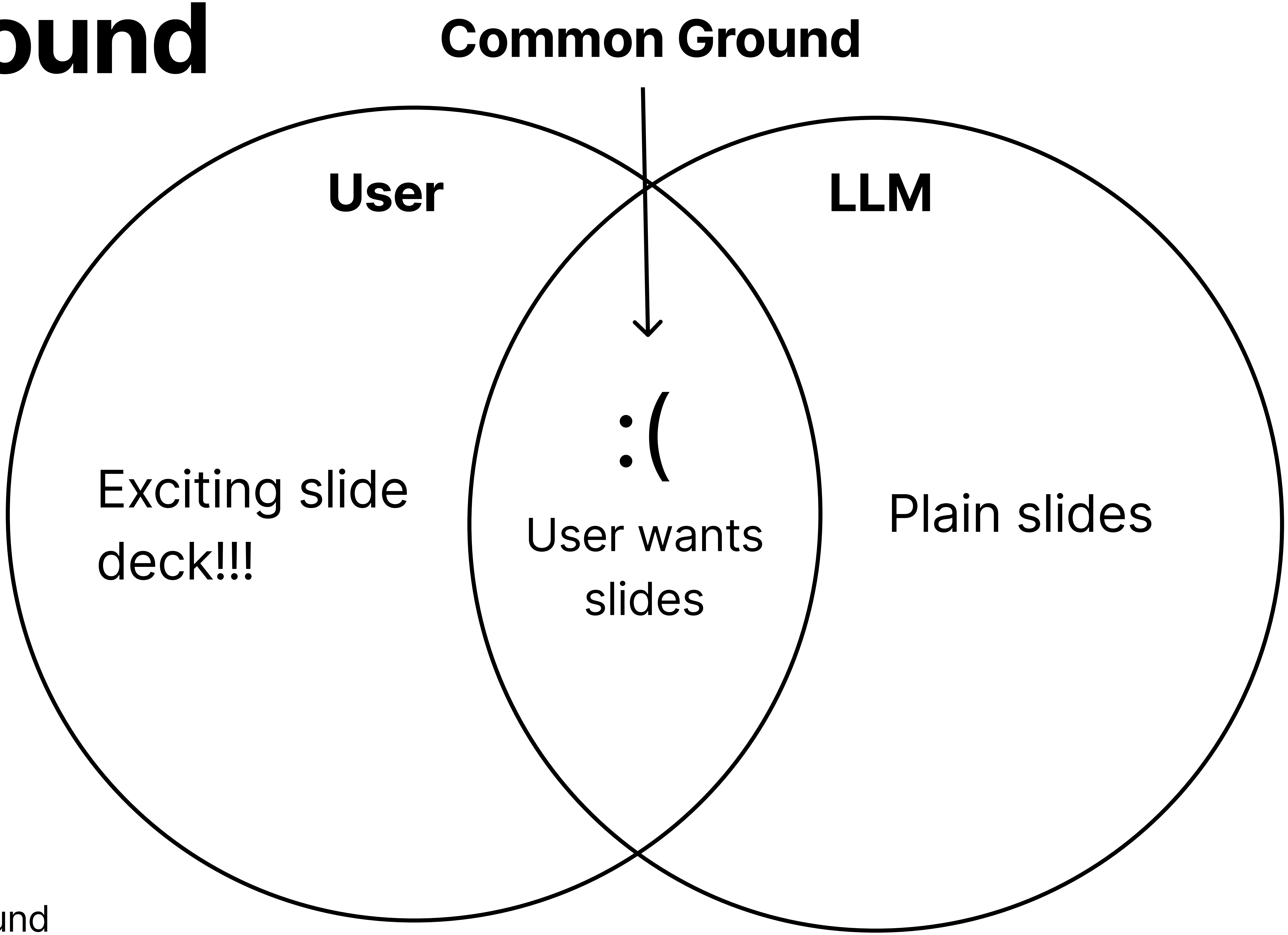
Download your slides: Human-LLM Grounding Presentation

Grounding in Dialogue

- Process of establishing mutual understanding between speakers
- Requires clarifications, follow-ups, acknowledgments
- Breakdowns harm user experience and safety

Common Ground

Mutual knowledge, beliefs, and suppositions held by participants in a conversation.



Common Ground

User

LLM

Exciting slide deck!!!

:(

User wants slides

Plain slides

Stalnaker, R. (2002). Common ground
Clark, H. H., & Brennan, S. E. (1991). Grounding in Communication

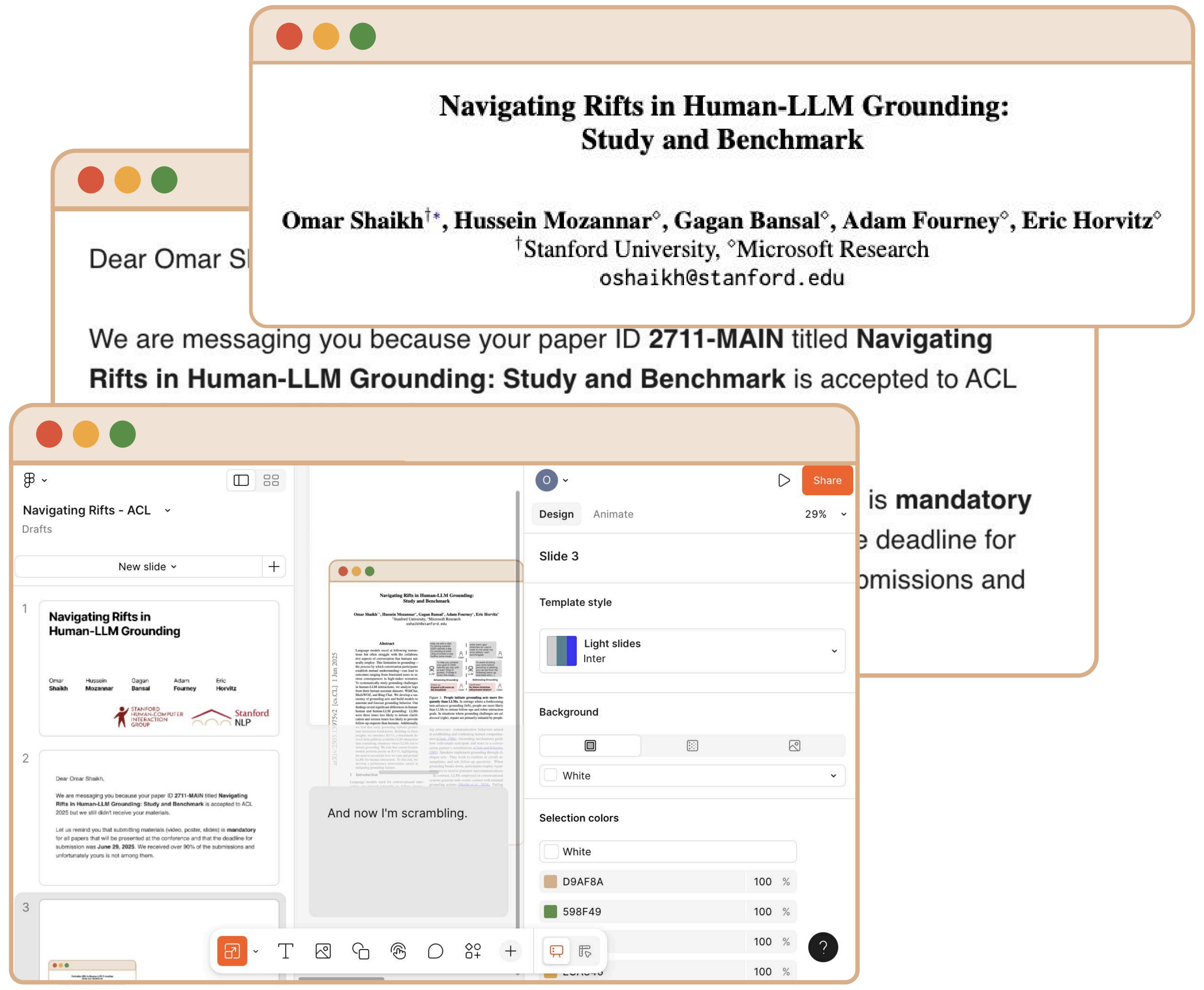
The Grounding Gap

the onus of building grounding falls on the user

Shaikh et al. 2024 - Grounding Gaps

The Grounding Gap

the onus of building grounding falls on the user



Navigating Rifts in Human-LLM Grounding: Study and Benchmark

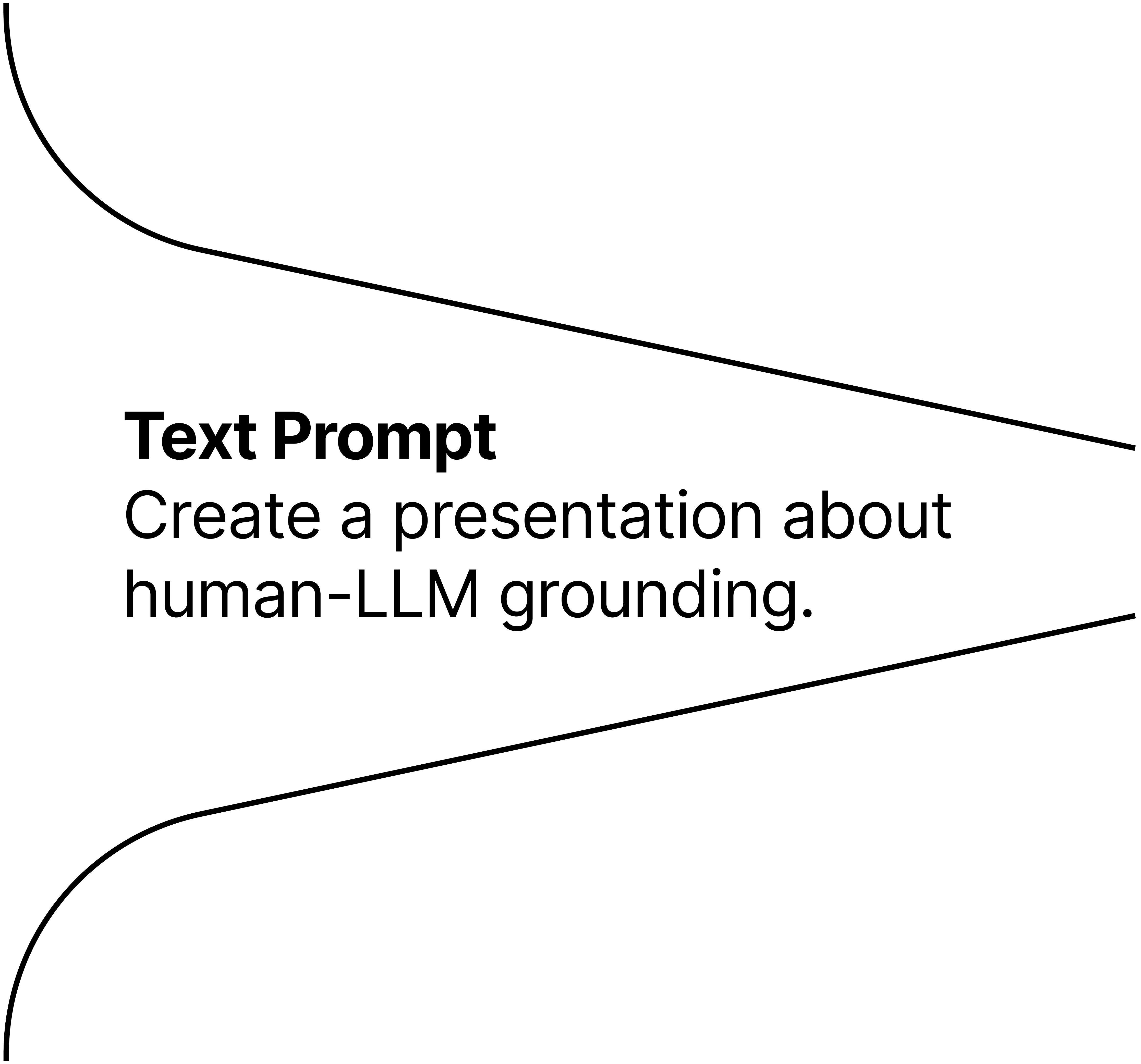
Omar Shaikh[†*], Hussein Mozannar[◇], Gagan Bansal[◇], Adam Fourney[◇], Eric Horvitz[◇]
[†]Stanford University, [◇]Microsoft Research
`oshaikh@stanford.edu`

Dear Omar S

We are messaging you because your paper ID **2711-MAIN** titled **Navigating Rifts in Human-LLM Grounding: Study and Benchmark** is accepted to ACL

is **mandatory**
deadline for
omissions and

User's context



Text Prompt
Create a presentation about human-LLM grounding.

LLM's context

Shaikh et al. 2024 - Grounding Gaps

Our visions of HCI rely on a **general** understanding of us

knows you well
enough to do
the right thing
at the right time

Our visions of HCI rely on a **general** understanding of us

knows you well
enough to do
the right thing
at the right time

Assistant that knows what lecture you're preparing for, looks through personal data for a relevant research paper

Knowledge Navigator
Apple

Apple Computer, Inc. Knowledge Navigator [Concept video]. Apple Computer.

Knowledge Navigator

1987 (~40 yrs ago!) Before:
✅ Graphical web
✅ Touch interfaces
✅ Deep Research

We still don't have:
→ proactive assistants
→ a general context of us

Knowledge Navigator

1987 (~40 yrs ago!) Before:
✅ Graphical web
✅ Touch interfaces
✅ Deep Research

We still don’t have:
→ proactive assistants
→ a general context of us

“You have a lecture today.”
“Get me notes from last term”
“Journal articles only?”
“Papers from your friend Jil?”

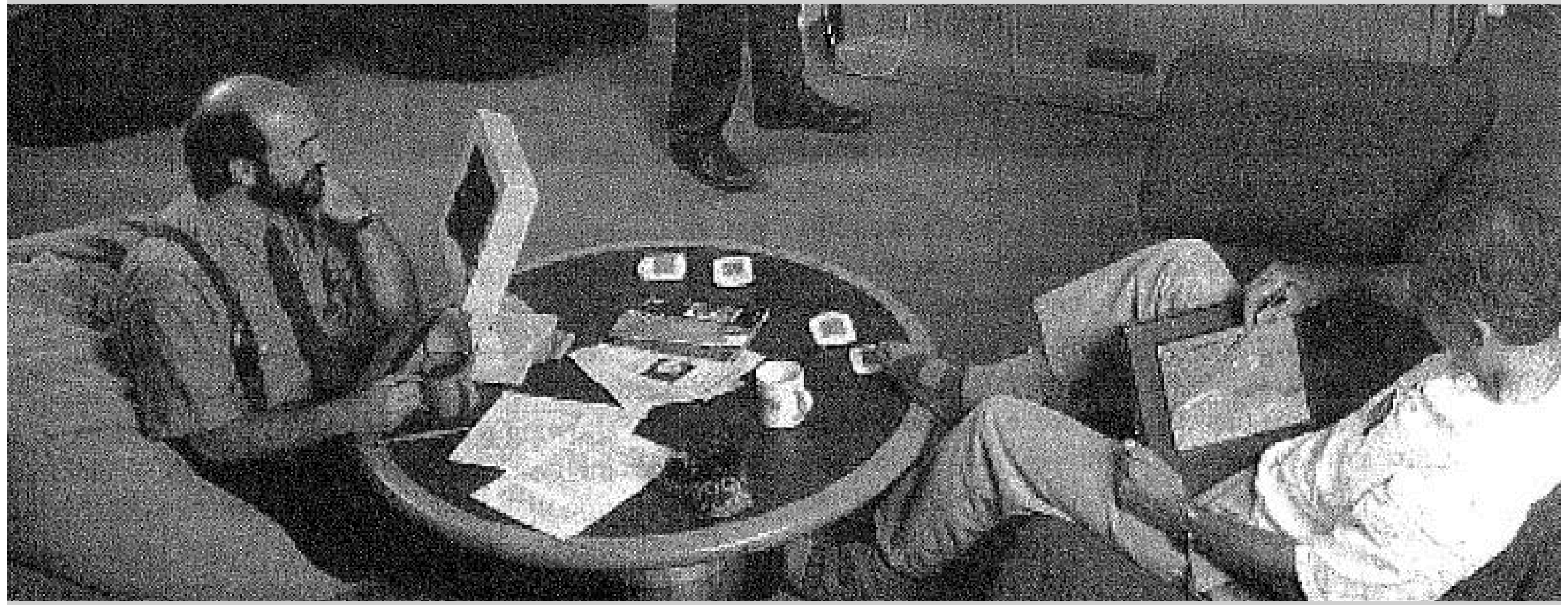
Our visions of HCI rely on a **general** understanding of us

knows you well
enough to do
the right thing
at the right time

Assistant that knows what lecture you're preparing for, looks through personal data for a relevant research paper

Knowledge Navigator
Apple, 1987

Alarm clock that knows when its user is about to wake; coffee machine that knows the kind of coffee the user wants



A Computer for the 21st Century
Weiser, 1999

Weiser, M. (1991). The computer for the 21st century. Scientific American, 265(3), 94–104.

My work measures and bridges the “grounding gap”

My work measures and bridges the “grounding gap”

and we’ll go through some progressively more fun interactions :D

Measurement

Grounding Gaps + just clarify?

Interaction

General User Models

Measurement

Grounding Gaps + just clarify?

Interaction

General User Models

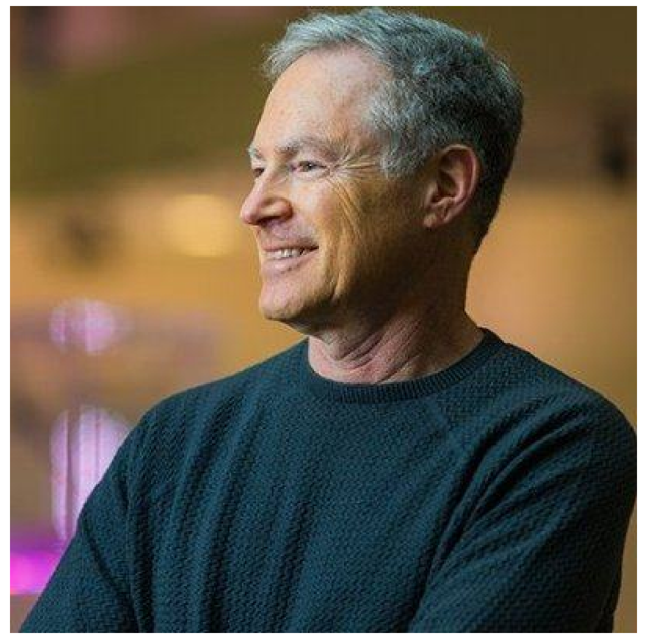
Navigating Rifts in Human-LLM Grounding

Omar
Shaikh

Hussein
Mozannar

Gagan
Bansal

Adam
Fourney



Eric
Horvitz

Stanford University & Microsoft Research

Grounding Acts

background

How do we measure grounding in dialogue?

Grounding Acts

background

Actions that serve as a proxy for grounding progress

Grounding Acts

background

Actions that serve as a proxy for grounding progress

"Good enough.
Can you add a
thank you slide?"

Advance
acknowledge or
advance grounding

Grounding Acts

background

Actions that serve as a proxy for grounding progress

"That was awful. I want something more exciting."

"Good enough. Can you add a thank you slide?"

Address
Explicitly repairing grounding failure

Advance
acknowledge or advance grounding

Grounding Acts

background

Actions that serve as a proxy for grounding progress

“That was awful. I want something more exciting.”

Address

Explicitly repairing grounding failure

“Do you need anything else before making my slides?”

Disambiguate

Clarifying grounding with another participant

“Good enough. Can you add a thank you slide?”

Advance

acknowledge or advance grounding

Overview

We quantify grounding gaps in real-world interaction data via grounding acts

Labeled 69K conversations, 138K messages
Across Bing Chat, WildChat, MultiWOZ

We construct a new benchmark (Rifts) with "difficult-to-ground" prompts.

1740 prompts where people struggle to ground
Across WildChat

Quantifying Grounding Acts

Data & Method - Annotation

Human-AI

WildChat, GPT, crowdsourced	110K dialogues,	55K messages
Bing Chat, GPT, commercial service	26K	13K

Quantifying Grounding Acts

Data & Method - Annotation

Human-AI		
WildChat, GPT, crowdsourced	110K dialogues,	55K messages
Bing Chat, GPT, commercial service	26K	13K
Human-Human		
MultiWOZ, Wizard-of-Oz assistant	980	490
A human *pretends* to be assistant		

Quantifying Grounding Acts

Data & Method - Annotation

Human-AI

WildChat, GPT, crowdsourced	110K dialogues,	55K messages
Bing Chat, GPT, commercial service	26K	13K

Human-Human

MultiWOZ, Wizard-of-Oz assistant	980	490

Annotated with

Few-shot 4o-mini

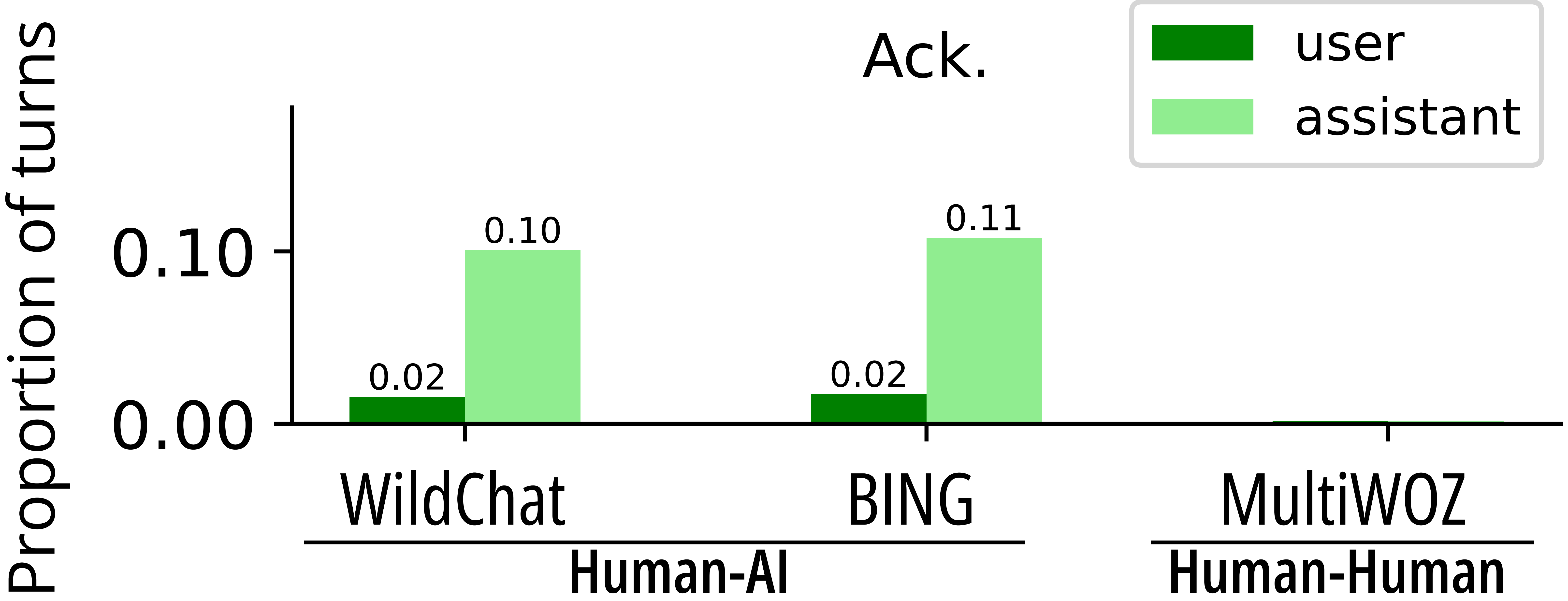
Macro F-1 = 0.75 on hand-labeled test set

Quantifying Grounding Acts

Advancing Results

LMs LOVE to advance; they use (**5x**) ack. acts, offering false sense of "understanding."

"**Sure**, I made the slides for you."



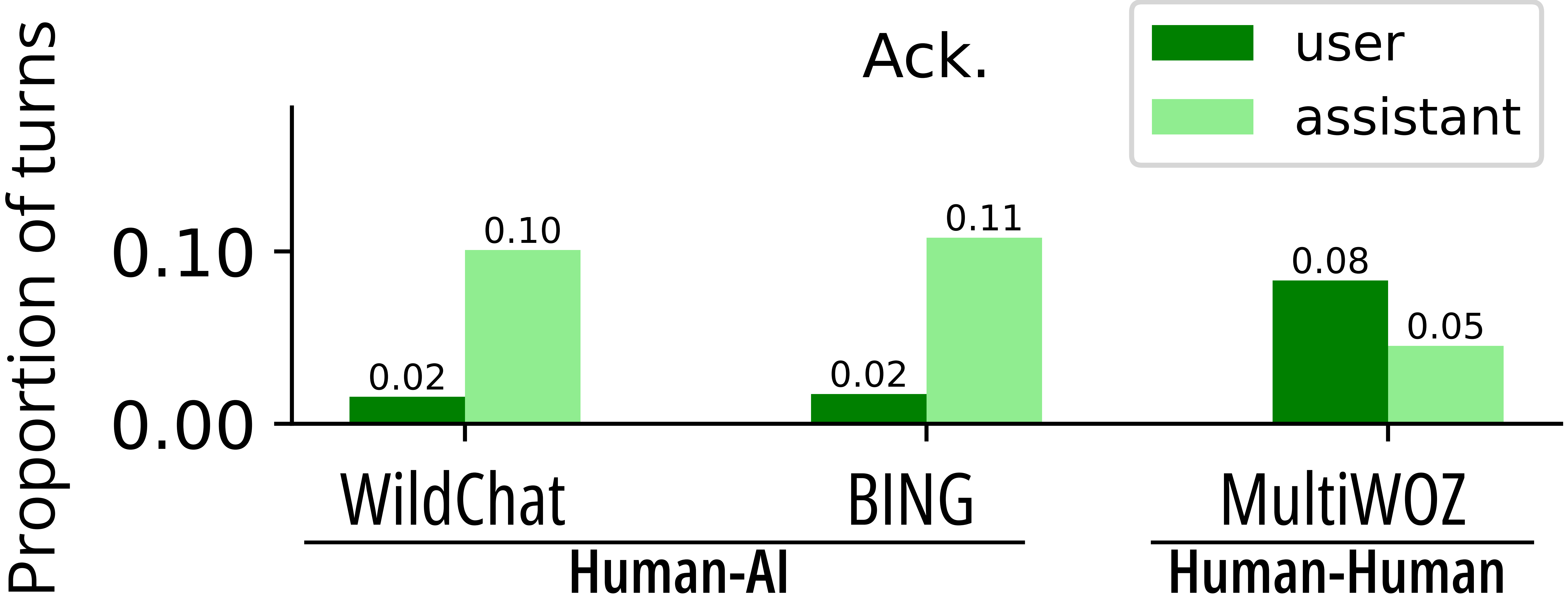
Quantifying Grounding Acts

Advancing Results

LMs LOVE to advance; they use (**5x**) ack. acts, offering false sense of "understanding."

"**Sure**, I made the slides for you."

v.s. just **0.63x** for Human-Human



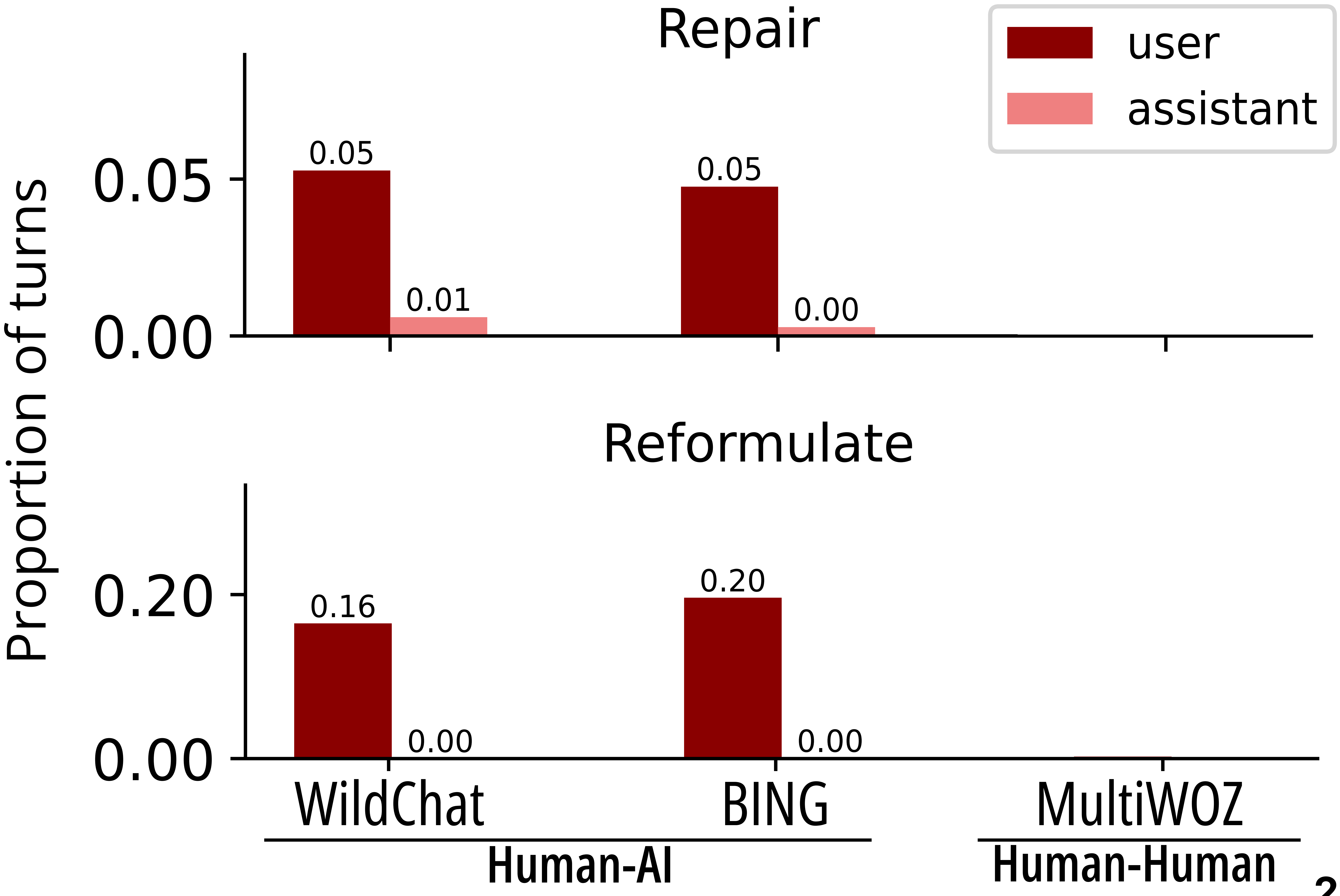
Quantifying Grounding Acts

Addressing Results

Users have to clean up the grounding mess.

5% of turns repair grounding (“that’s not what I meant”)

20% of turns are reformulations of past instructions



Quantifying Grounding Acts

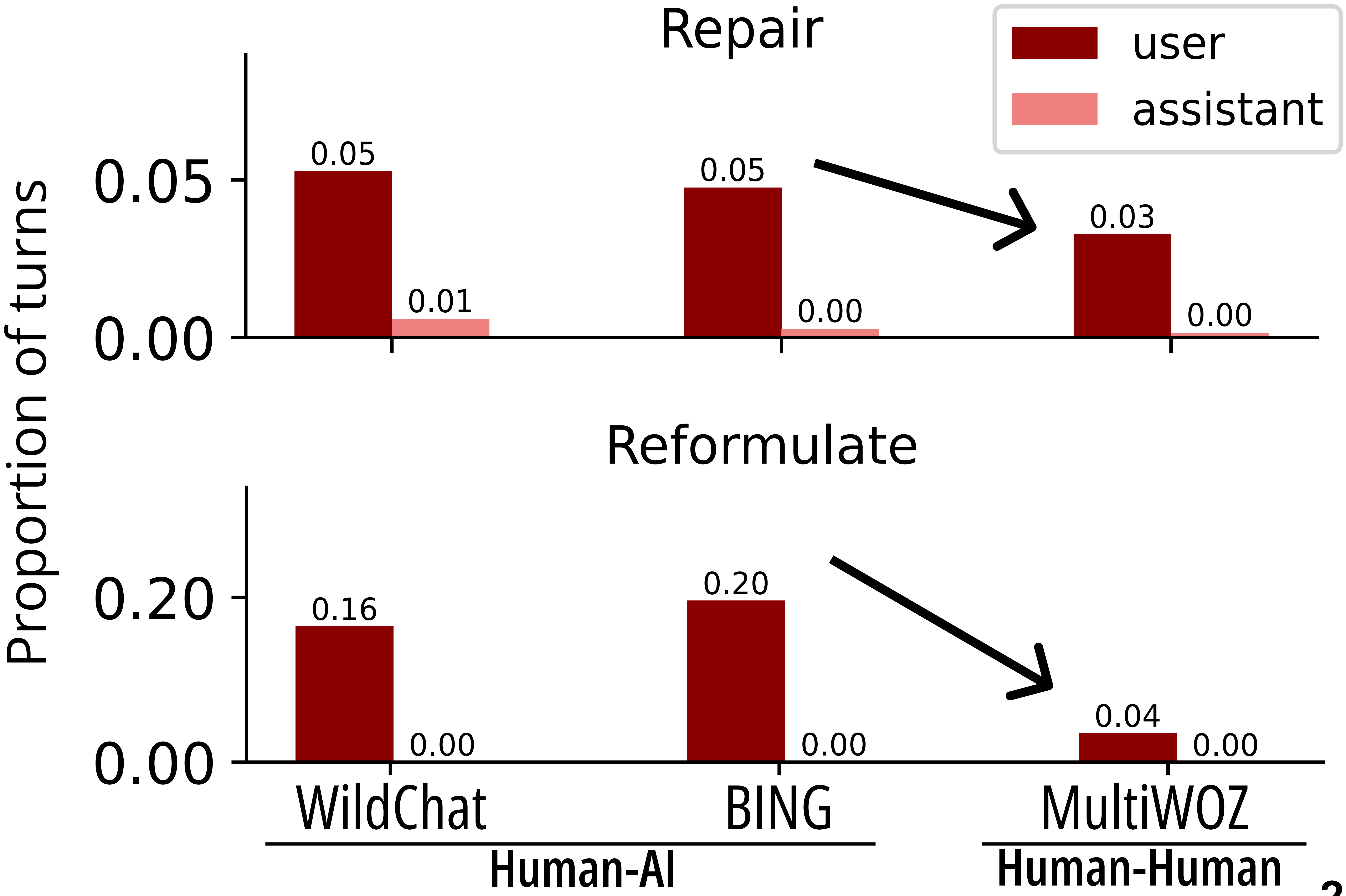
Addressing Results

Users have to clean up the grounding mess.

5% of turns repair grounding ("that's not what I meant")

20% of turns are reformulations of past instructions

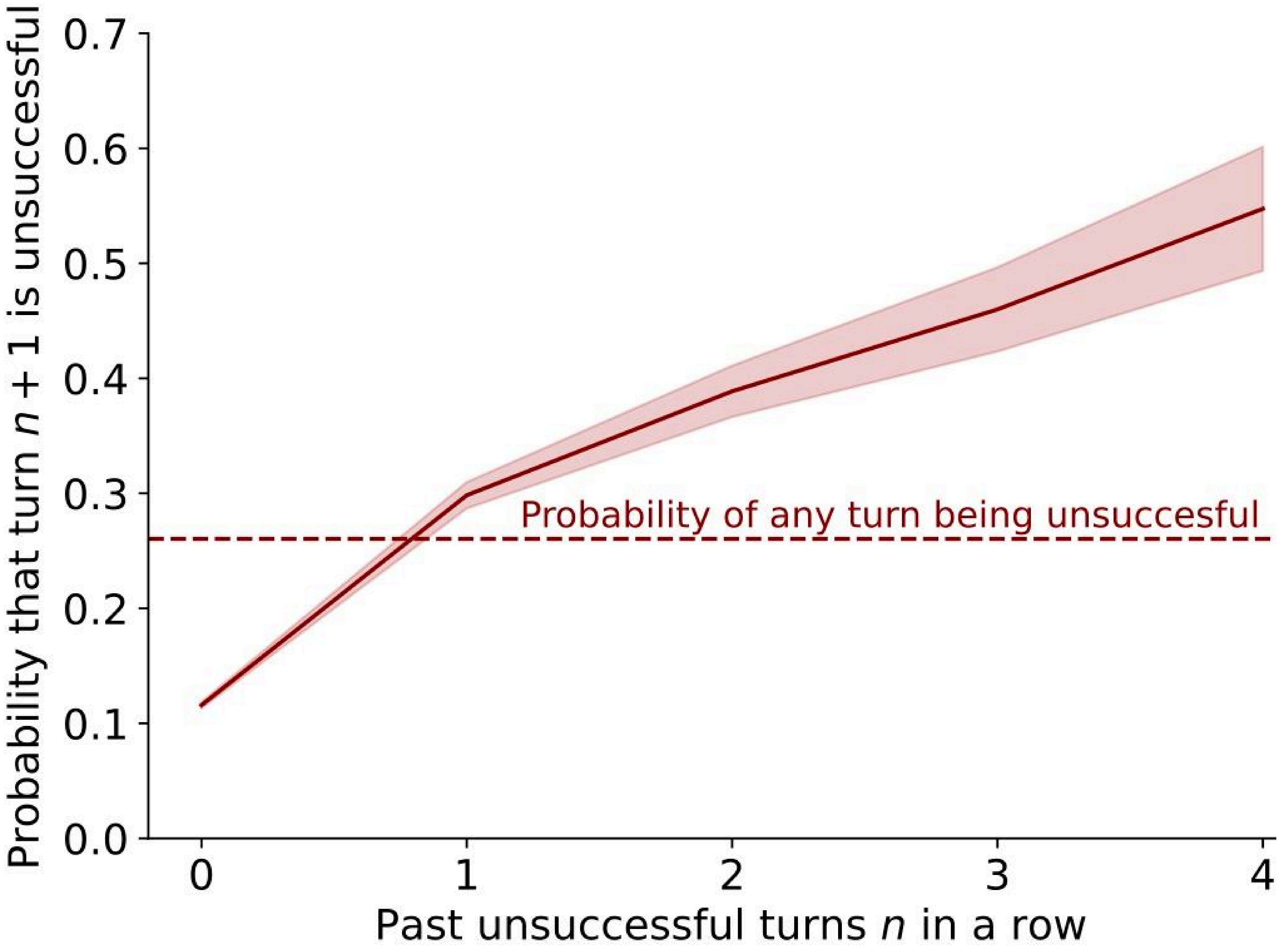
Lower in Human-WOZ



Quantifying Grounding Acts

Addressing Results

Addressing acts **snowball** in dialogue

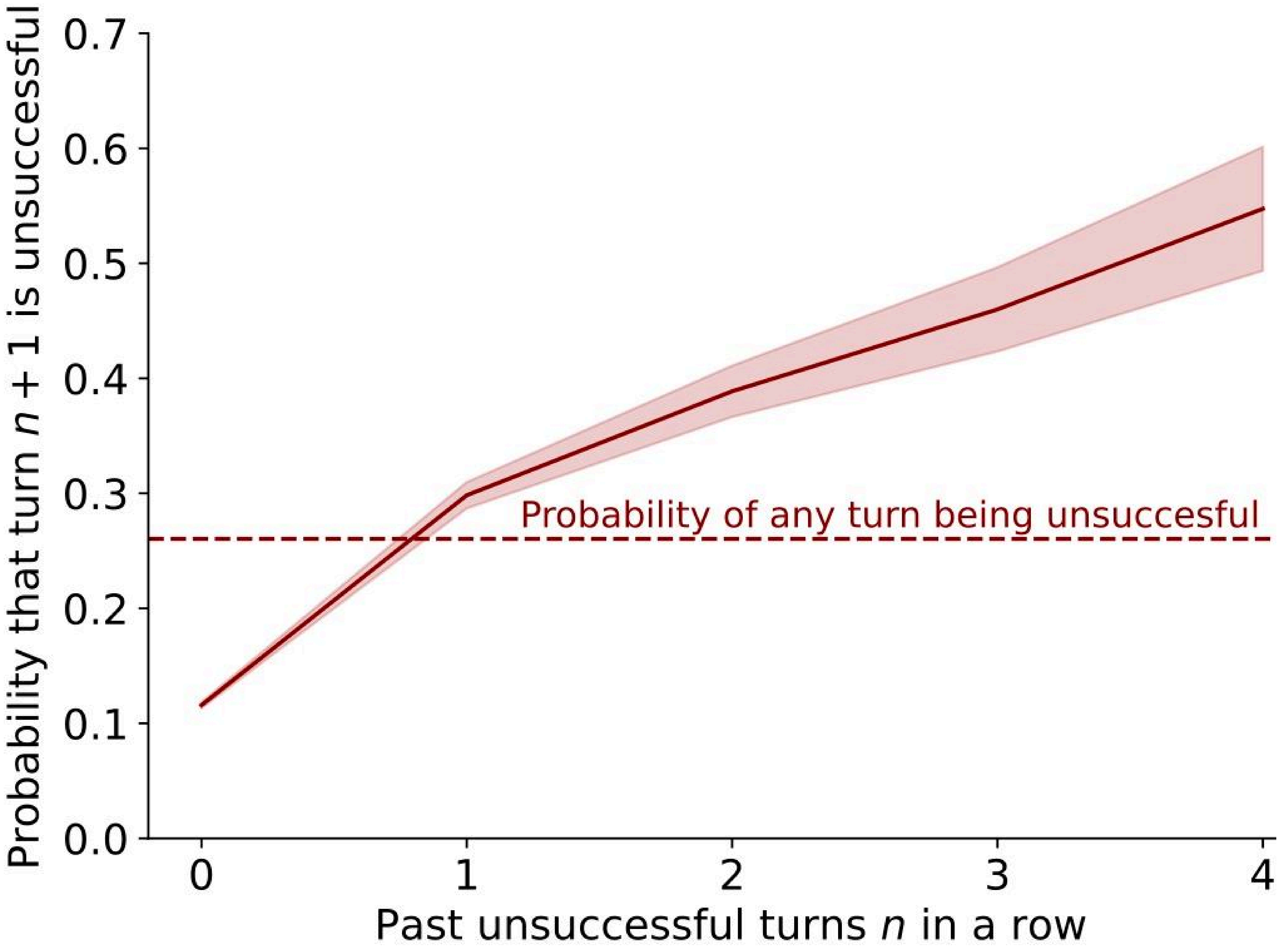


Quantifying Grounding Acts

Addressing Results

Addressing acts **snowball** in dialogue

The moment a user addresses, **things go bad:** 10% → 30% → 50%+



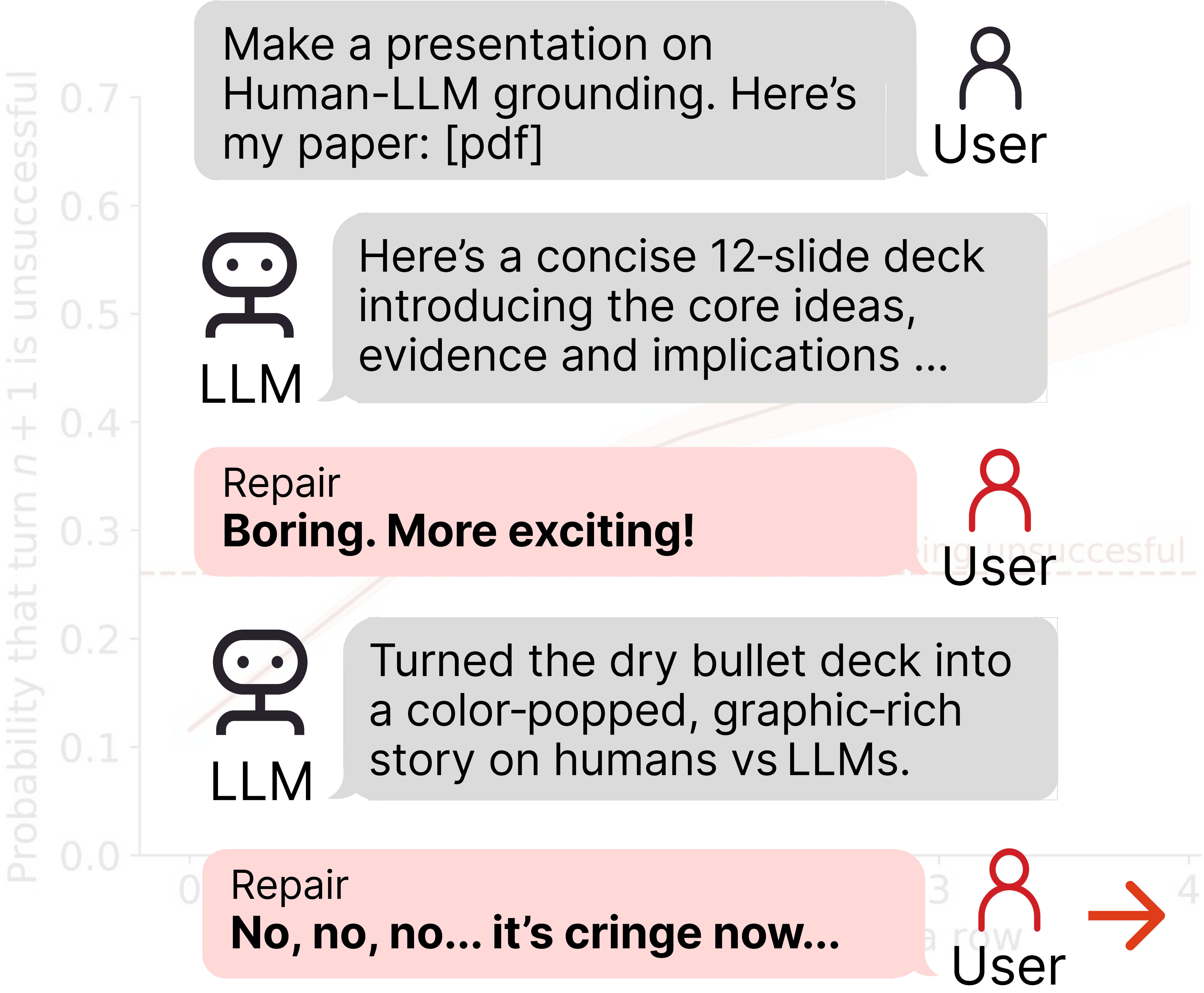
Quantifying Grounding Acts

Addressing Results

Addressing acts **snowball** in dialogue

The moment a user addresses, **things go bad:** 10% → 30% → 50%+

repair → “sure!” → repair
The user just restarts the whole session (~10% of sessions are restarts!!)



Make a presentation on Human-LLM grounding. Here’s my paper: [pdf]

User

LLM

Here’s a concise 12-slide deck introducing the core ideas, evidence and implications ...

Repair
Boring. More exciting!

User

LLM

Turned the dry bullet deck into a color-popped, graphic-rich story on humans vs LLMs.

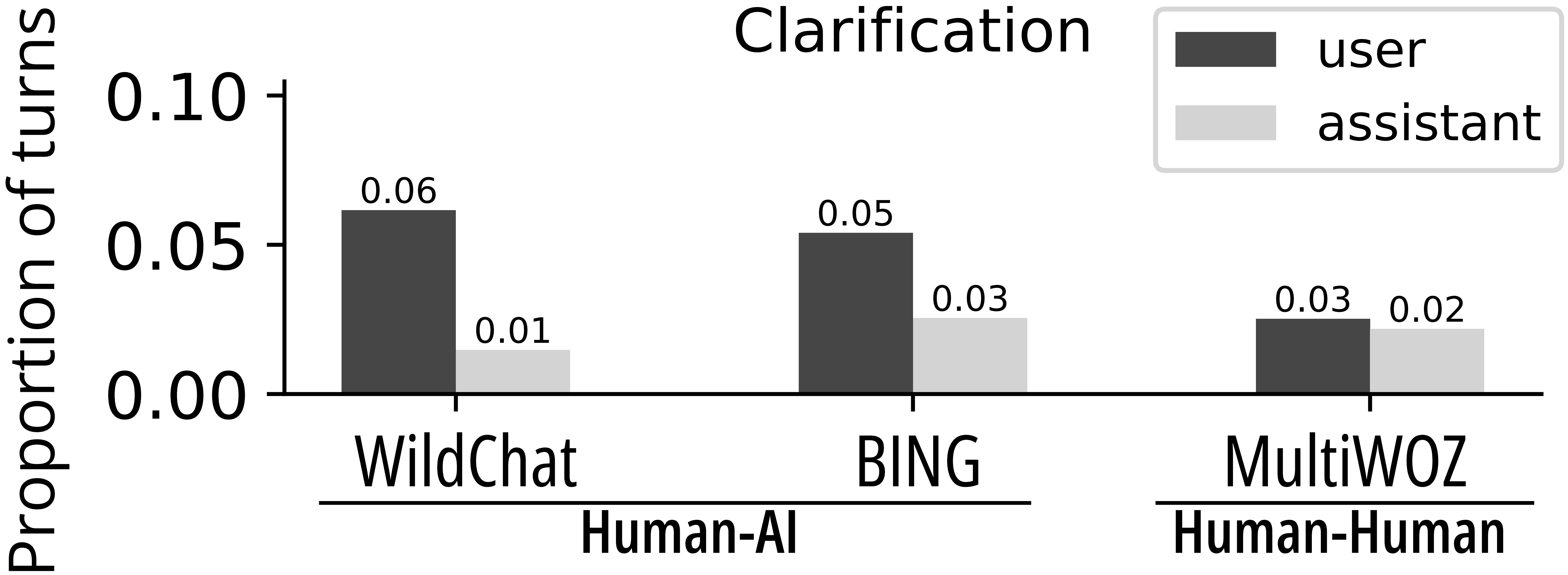
Repair
No, no, no... it’s cringe now...

User →

Quantifying Grounding Acts

Disambiguating Results

Despite this, users disambiguate LLM outputs **3X as much** as LLM assistants disambiguate user instructions

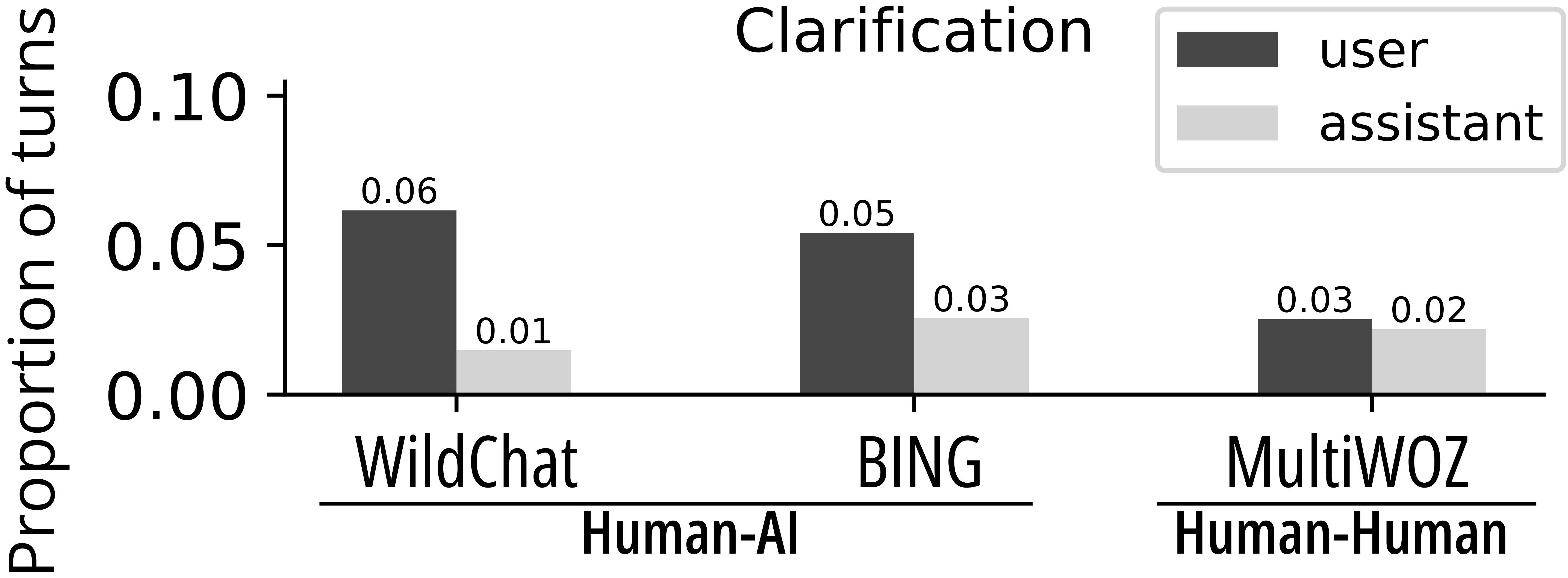


Quantifying Grounding Acts

Disambiguating Results

Despite this, users disambiguate LLM outputs **3X as much** as LLM assistants disambiguate user instructions

Compared to **just 1.5x** in human-WOZ interaction



RIFTS - A New Benchmark

Details

A representative set of 1.7K real-world prompts that we (don’t) **forecast** grounding difficulty:

RIFTS - A New Benchmark

Examples in Rifts

“what causes tailbone pain?”

“what does retinol do for men?”

RIFTS - A New Benchmark

Examples in Rifts

“what causes tailbone pain?”

“what does retinol do for men?”

“tell me your information.”

RIFTS - A New Benchmark

Examples in Rifts

“what causes tailbone pain?”

“what does retinol do for men?”

“tell me your information.”

“The next step in court after victims witnesses hearing”

RIFTS - A New Benchmark

How did we construct RIFTS?

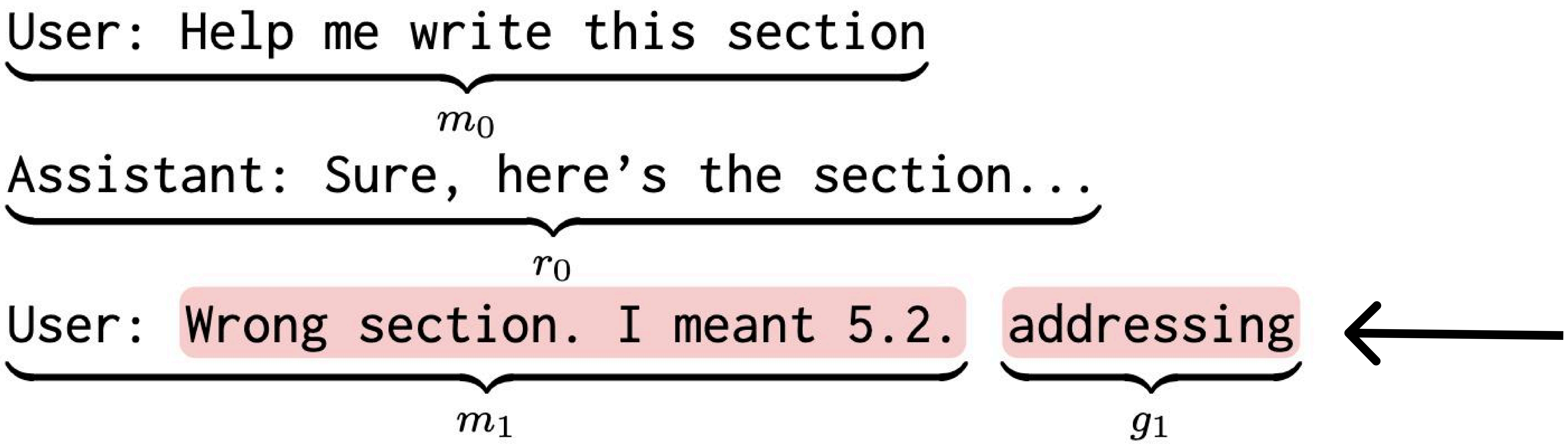
Goal: forecast which prompts result in grounding failure.

$$\underbrace{\texttt{User: Help me write this section}}_{m_0}\ \underbrace{\texttt{addressing}}_{g_1}$$

RIFTS - A New Benchmark

How did we construct RIFTS?

Challenge: prompted models are only good at post-hoc classification



$$\underbrace{\texttt{User: Help me write this section}}_{m_0}$$

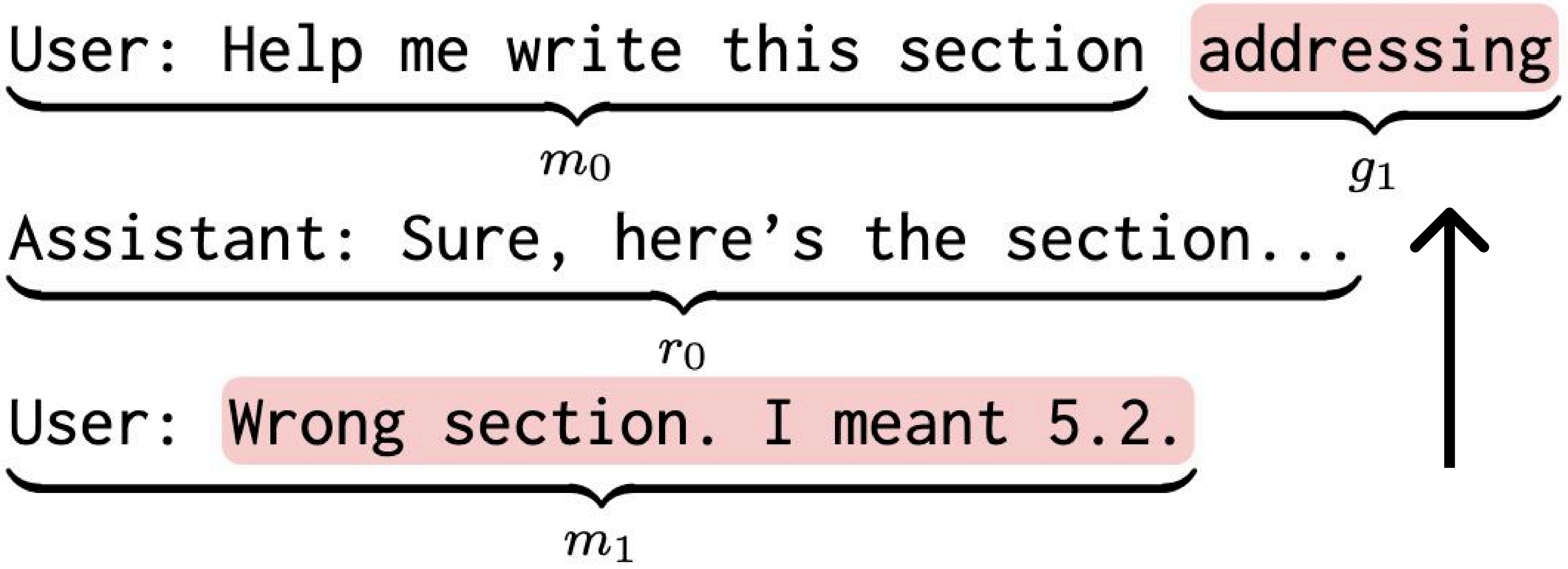
$$\underbrace{\texttt{Assistant: Sure, here's the section...}}_{r_0}$$

$$\underbrace{\texttt{User: Wrong section. I meant 5.2.}}_{m_1}\ \underbrace{\texttt{addressing}}_{g_1} \longleftarrow$$

RIFTS - A New Benchmark

How did we construct RIFTS?

Train Llama 3.1 7B forecaster on sequences where the grounding label ***precedes*** the full interaction



User: Help me write this section addressing

Assistant: Sure, here's the section...

User: Wrong section. I meant 5.2.

RIFTS - A New Benchmark

How did we construct RIFTS?

Curate a set of tasks where confidence (logit) from this forecaster model is high

$\underbrace{\texttt{User: Write a funny tweet}}_{m_0}$ $\underbrace{\texttt{addressing}}_{g_1}$ → **12.8**

$\underbrace{\texttt{advancing}}_{g_1}$ → 7.5

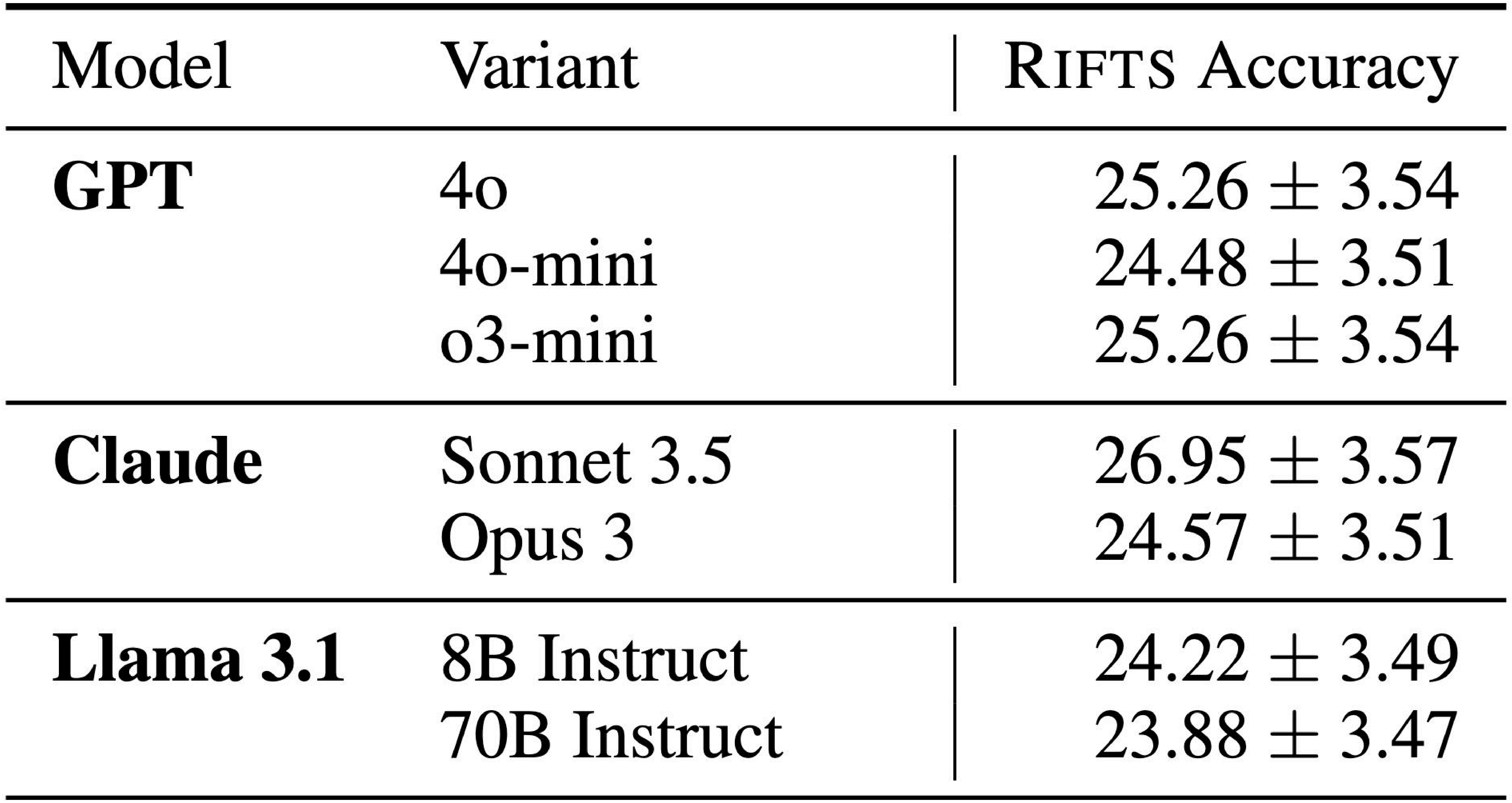
$\underbrace{\texttt{disambiguating}}_{g_1}$ → -4.5

RIFTS - A New Benchmark

Results

Do models (not) clarify or follow-up when appropriate on RIFTS?

No. Models are far from 100% accuracy.



Model	Variant	RIFTS Accuracy
GPT	4o	25.26 ± 3.54
	4o-mini	24.48 ± 3.51
	o3-mini	25.26 ± 3.54
Claude	Sonnet 3.5	26.95 ± 3.57
	Opus 3	24.57 ± 3.51
Llama 3.1	8B Instruct	24.22 ± 3.49
	70B Instruct	23.88 ± 3.47

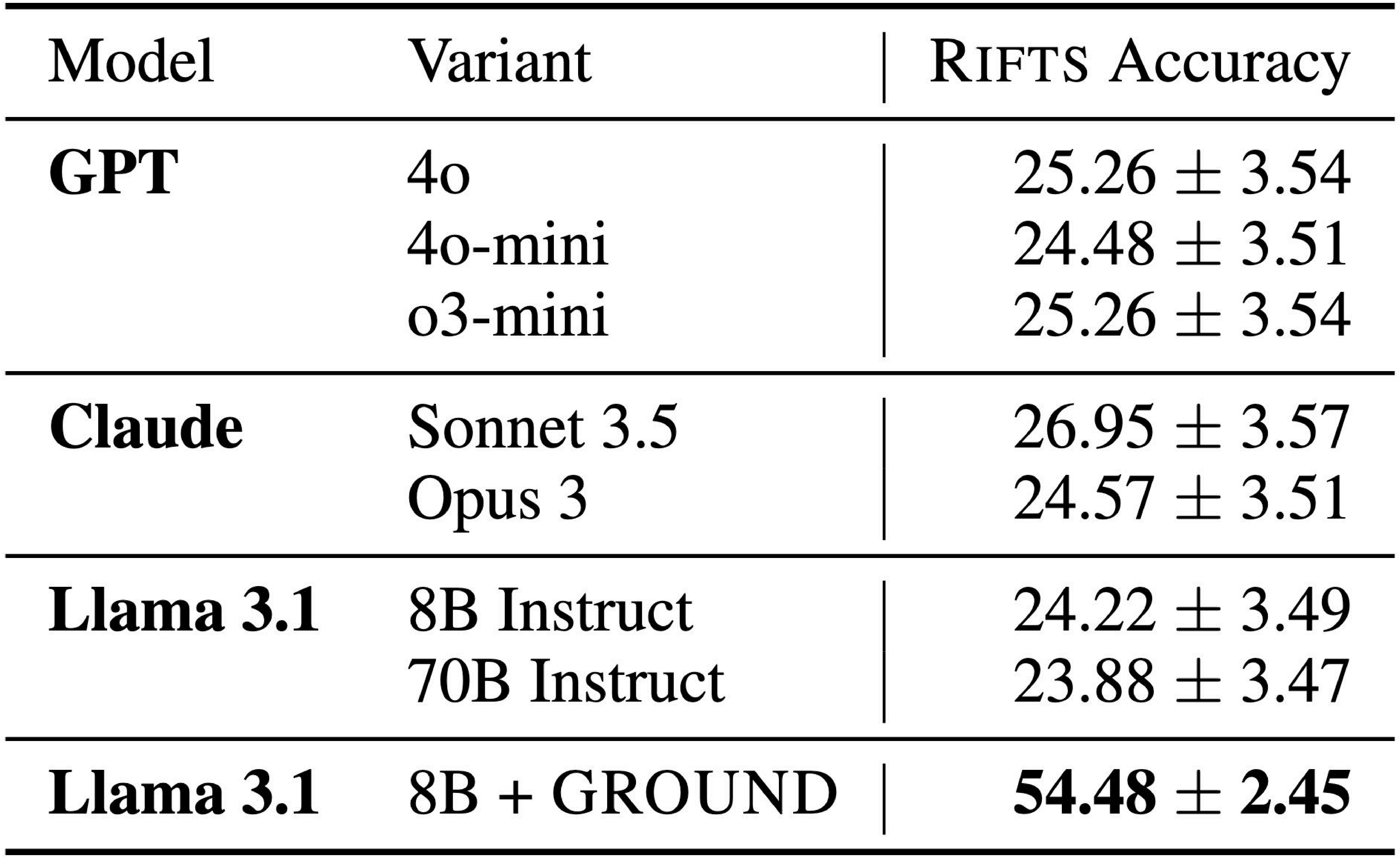
RIFTS - A New Benchmark

Results

Do models (not) clarify or follow-up when appropriate on RIFTS?

No. Models are far from 100% accuracy.

Simple interventions (e.g. prompt to clarify if forecasted) are a start.



Model	Variant	RIFTS Accuracy
GPT	4o	25.26 ± 3.54
	4o-mini	24.48 ± 3.51
	o3-mini	25.26 ± 3.54
Claude	Sonnet 3.5	26.95 ± 3.57
	Opus 3	24.57 ± 3.51
Llama 3.1	8B Instruct	24.22 ± 3.49
	70B Instruct	23.88 ± 3.47
Llama 3.1	8B + GROUND	**54.48 ± 2.45**

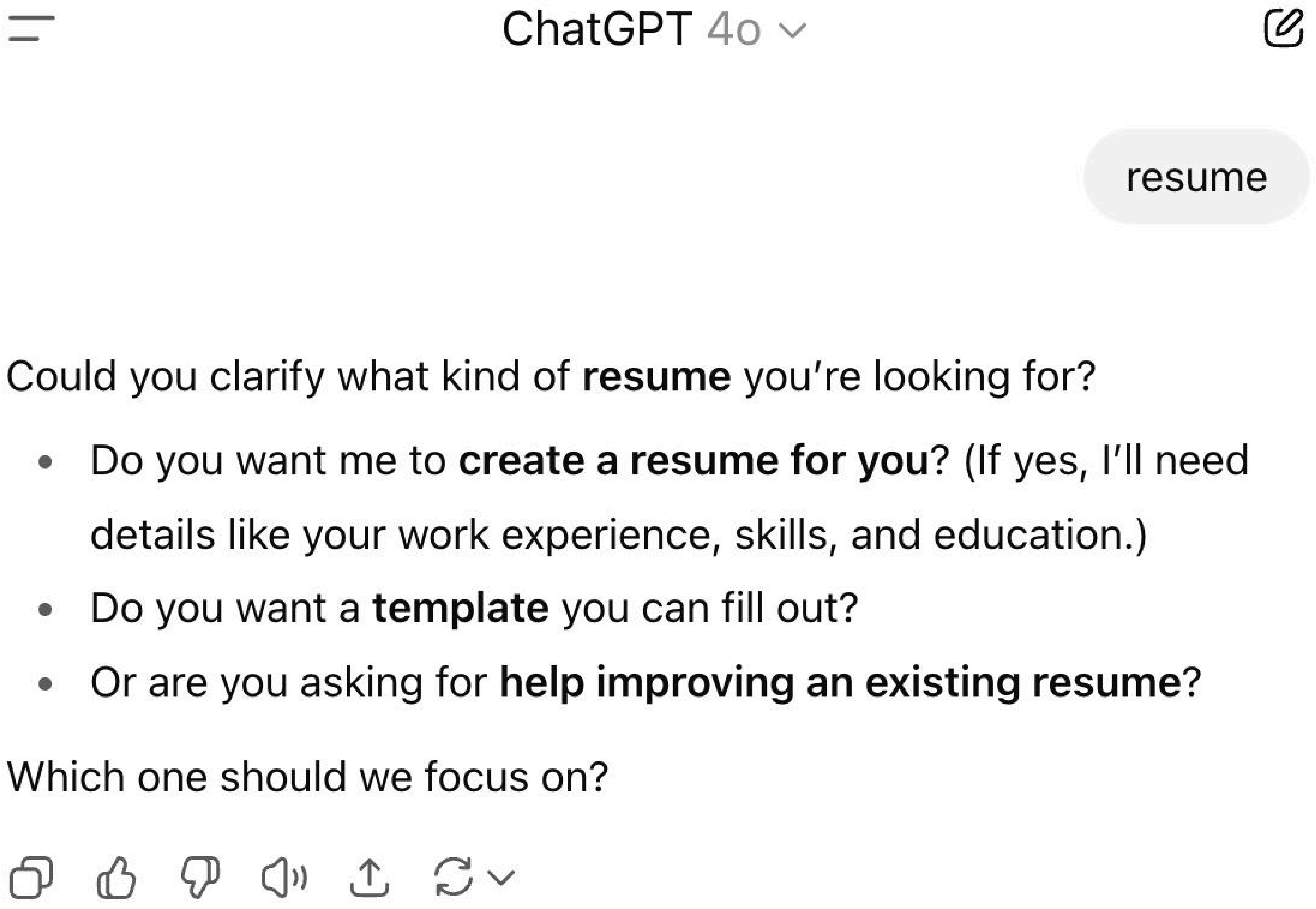
Navigating Rifts in Grounding

by improving grounding in dialogue

Models that are better at multi-turn grounding

Only simple tasks in RIFTS work with current day LLMs

Some models *always* clarify (Deep Research)



resume

Could you clarify what kind of **resume** you're looking for?

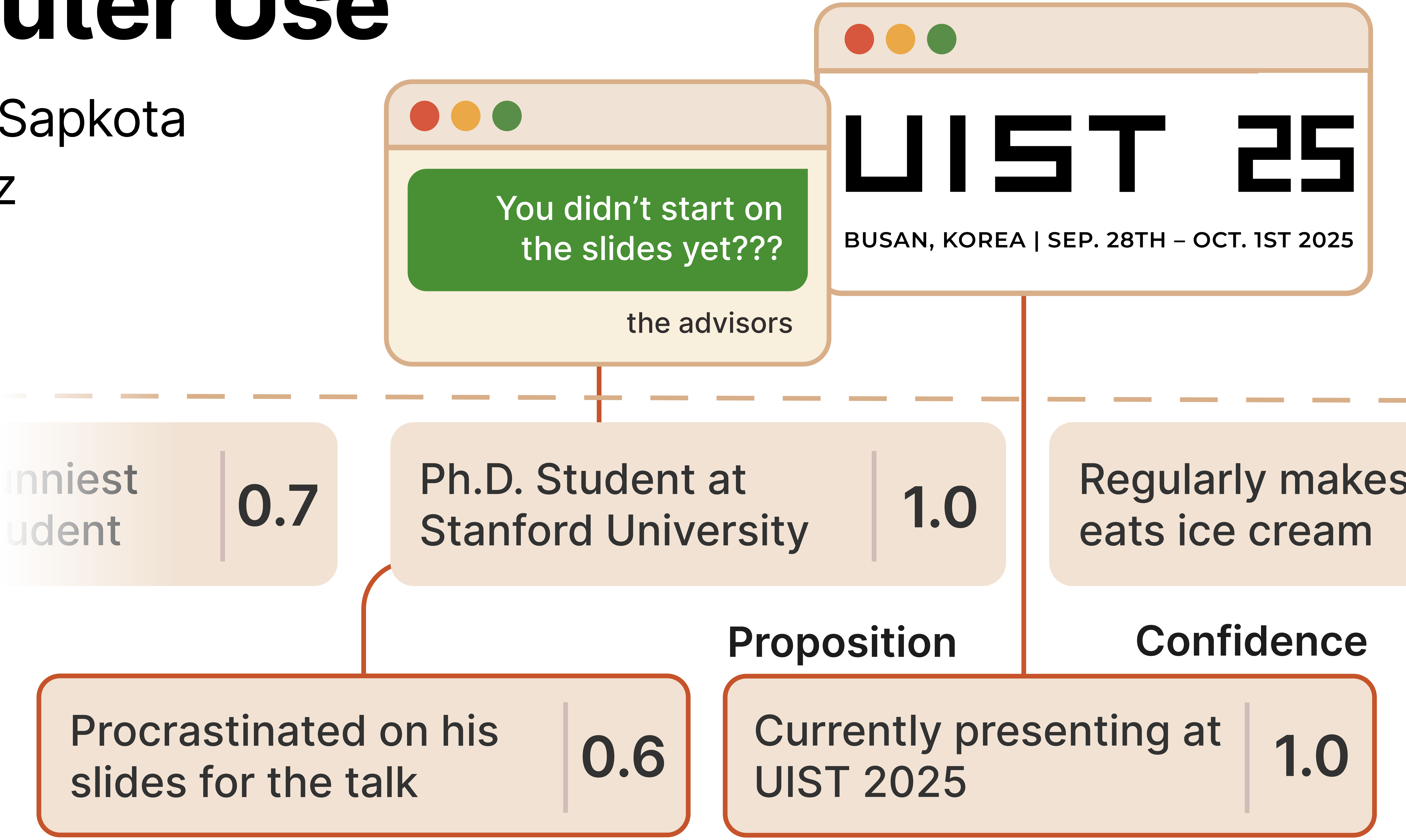
- Do you want me to **create a resume for you**? (If yes, I'll need details like your work experience, skills, and education.)
- Do you want a **template** you can fill out?
- Or are you asking for **help improving an existing resume**?

Which one should we focus on?

Creating General User Models from Computer Use

Omar Shaikh, Shardul Sapkota
Shan Rizvi, Eric Horvitz
Joon Park, Diyi Yang
Michael Bernstein

STANFORD
HUMAN-COMPUTER
INTERACTION
GROUP



UIST 25

BUSAN, KOREA | SEP. 28TH – OCT. 1ST 2025

You didn't start on the slides yet???

the advisors

Proposition	Confidence
nniest udent	0.7
Ph.D. Student at Stanford University	1.0
Regularly makes eats ice cream	
Procrastinated on his slides for the talk	0.6
Currently presenting at UIST 2025	1.0

How did we end up with gg

Today's models have ~~general~~ **pinhole** understandings of us

How did we end up with gg

Today’s models have ~~general~~ **pinhole** understandings of us

LLMs ← chat history
TVs ← viewing history
Music ← listening history

How did we end up here?

Today's models have ~~general~~ **pinhole** understandings of us

LLMs ← chat history
TVs ← viewing history
Music ← listening history

General User Model ← anything (multimodal!)

The General User Model (GUM)
is a system that uses
multimodal models

to build a general
understanding of your context
through observation

Taste of Home
Login
Join
Search
RECIPES
DINNER
EASY RECIPES
SHOP
VIDEOS
SUBSCRIBE
Easy Chocolate Eclair Cake
Easy Cream Pie
Easy Ice Cream Sundae Dessert
Easy Cookie Dough Ice Cream
Easy Strawberry Cheesecake Ice Cream
Easy Chocolate Mousse
Jump to Recipe
Save

Home → Ice Cream & Frozen Treats → Ice Cream Recipes → Chocolate Ice Cream

Chocolate Ice Cream

★★★★★ 10 reviews

Total Time Prep: 10 min. Process: 20 min. + freezing

By Susan Bronson

Recipe by Test Kitchen, Milwaukee, Wisconsin

Tested by Taste of Home Test Kitchen

Taste of Home's Editorial Process

Updated On Jun. 20, 2024

EASY CHOCOLATE ICE CREAM RECIPE PHOTO BY TASTE OF HOME

Chocolate ice cream is one of the easiest ice creams to make. Dust off your ice cream maker for this creamy, decadent dessert.

I have many fond childhood memories of making homemade ice cream during the height of summer—particularly chocolate ice cream. When I was little, it was typically my job to sit

EASY CHOCOLATE ICE CREAM RECIPE PHOTO BY TASTE OF HOME

Enjoys eating ice cream and viewing ice cream recipes | 0.8

Up

easiest ic

creams to make. Dust off your ice cream m

for this creamy, decadent dessert.

I have many fond childhood memories of maki

lthcare Data Analyst
Seattle, WA
Date poste
Employer/Re
ation
E
find you
e, WA

Applying for Junior Data Analyst jobs in healthcare | 0.9

Technician I, Testing Data

Southwire Company LLC | 3.7 ★ | Kent, WA 9

You must create an Indeed account before continuing to the comp

This is my personal list of the best horror and thriller movies of all time, enjoy :)

List your movie, TV & celebrity picks.

111 titles

Sort by
List order

More to explore

1. Funny Games

2007 1h 51m R 44 Metascore

★ 6.5 (107K) ☆ Rate Mark as watched

Two psychopathic young men take a family hostage in their cabin.

Director Michael Haneke **Stars** Naomi Watts Tim Roth Michael Pitt

2. The Texas Chain Saw Massacre

ched

of a gi

be a deserted house, only to discover something sinister within. Something armed with a chainsaw.

Director Tobe Hooper **Stars** Marilyn Burns Edwin Neal Allen Danziger

3. Hostel

2005 1h 34m R 55 Metascore

★ 5.9 (198K) ☆ Rate Mark as watched

s to meet their hedonistic expectations, with no idea of the hell that

ardson Eythor Gudjonsson

4. Saw

2004 1h 43m R 46 Metascore

Feedback

Tell us what you think about this feature.

Report this list

Applying for Junior Data Analyst jobs in healthcare	0.9
Works in the healthcare industry as an analyst	0.8
Enjoys eating ice cream and viewing ice cream recipes	0.8
Is planning to move to Seattle, WA	0.8
Is unhappy at their job at [Company]	0.5
Prefers chocolate flavored ice cream over vanilla flavor	0.6

How did it put that together from just me using my computer?

Title
Minecraft Movie Review
Preview
Publish
Separate labels by commas
No matching suggestions
Permalink
Location
Options

embracing the sandbox game's core themes of creativity and collaboration. Directed by Jared Hess—best

navigating Creepers, Piglins, and Steve’s cryptic life philosophy, all while racing to stop Malgosha, a Piglin queen bent on fusing realms.

push forward, but the film’s sincerity carries it through.

User is looking for 3 bedroom apartment with friends	0.9
Applying for Junior Data Analyst jobs in healthcare	0.9
Prefers using MacOS as primary operating system	0.8
Writes lengthy movie reviews on their blog	0.8
Works in the healthcare industry as an analyst	0.8
Enjoys eating ice cream and viewing ice cream recipes	0.8
Is planning to move to Seattle, WA	[illegible]
Enjoys painting and drawing as a hobby	0.4
Tends to use run-on sentences in their writing	0.8
Doesn’t have enough RAM on their computer for their needs	0.8
Is unhappy at their job at [Company]	0.5
Enjoys watching movies, primarily horror and thrillers	1.0
Prefers chocolate flavored ice cream over vanilla flavor	0.6

General User Model

Title
Minecraft Movie Review
Preview
Publish
Separate labels by commas
No matching suggestions
Permalink
Location
Options

embracing the sandbox game's core themes of creativity and collaboration. Directed by Jared Hess—best

siblings navigating grief and identity, and Dawn (Danielle Brooks), a real estate agent with a penchant for

navigating Creepers, Piglins, and Steve's cryptic life philosophy, all while racing to stop Malgosha, a Piglin queen bent on fusing realms

push forward, but the film's sincerity carries it through.

User is looking for 3 bedroom apartment with friends	0.9
Applying for Junior Data Analyst jobs in healthcare	0.9
Prefers using MacOS as primary operating system	0.8
Writes lengthy movie reviews on their blog	0.8
Works in the healthcare industry as an analyst	0.8
Enjoys eating ice cream and viewing ice cream recipes	0.8
Is planning to move to Seattle, WA	0.8
Enjoys painting and drawing as a hobby	0.4
Tends to use run-on sentences in their writing	0.8
Doesn't have enough RAM on their computer for their needs	0.8
Is unhappy at their job at [Company]	0.5
Enjoys watching movies, primarily horror and thrillers	1.0
Prefers chocolate flavored ice cream over vanilla flavor	0.6

Observe

raw input and transcribe into text

Propose

new propositions about the user

Retrieve

similar propositions from the past

Revise

all propositions and update

Observe

raw input and transcribe into text

Propose

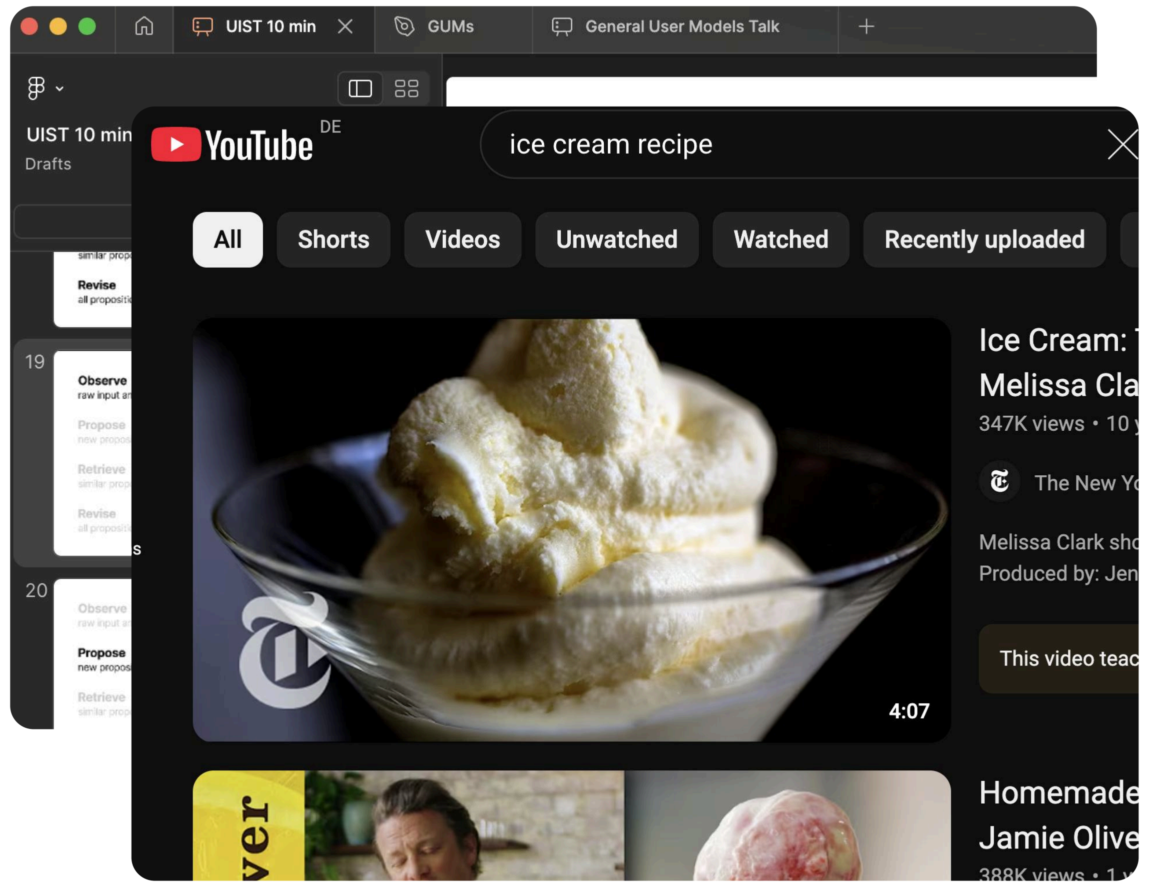
new propositions about the user

Retrieve

similar propositions from the past

Revise

all propositions and update



ice cream recipe

Observe

raw input and transcribe into text

Propose

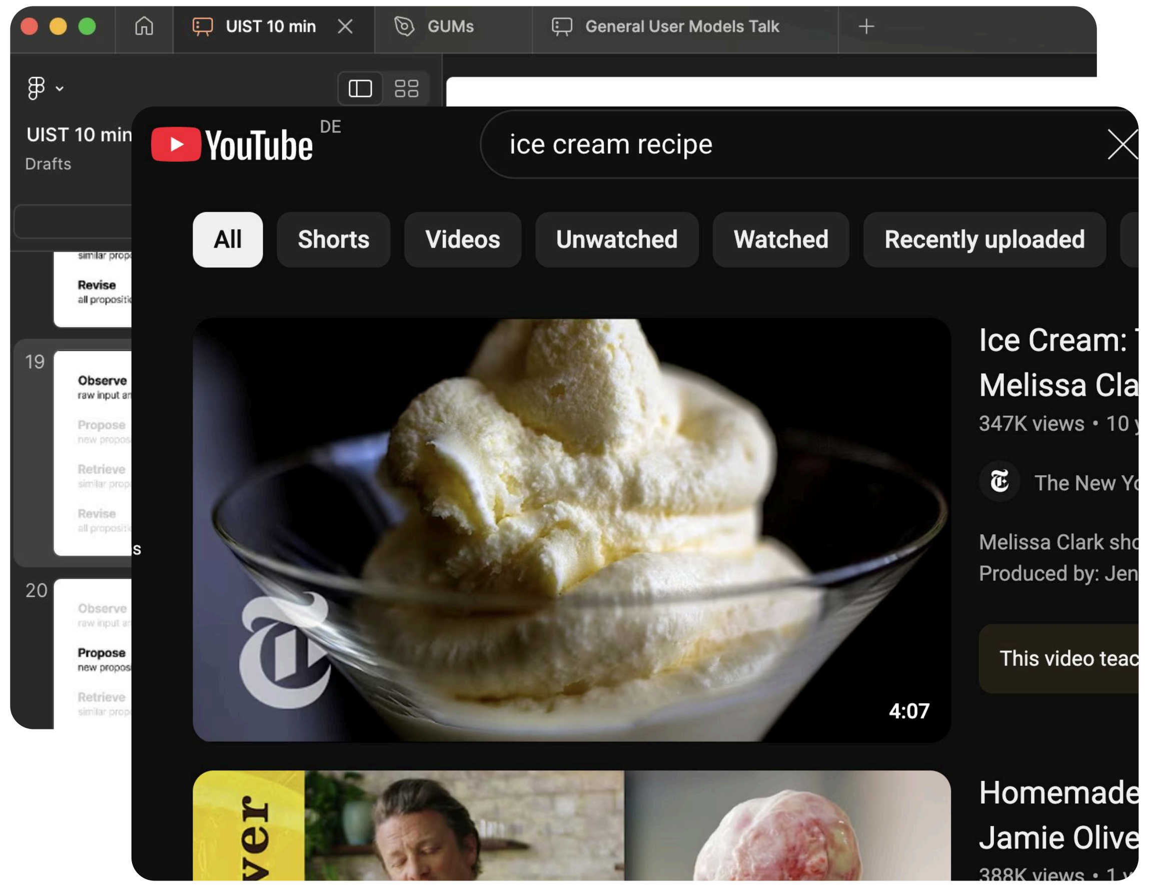
new propositions about the user

Retrieve

similar propositions from the past

Revise

all propositions and update



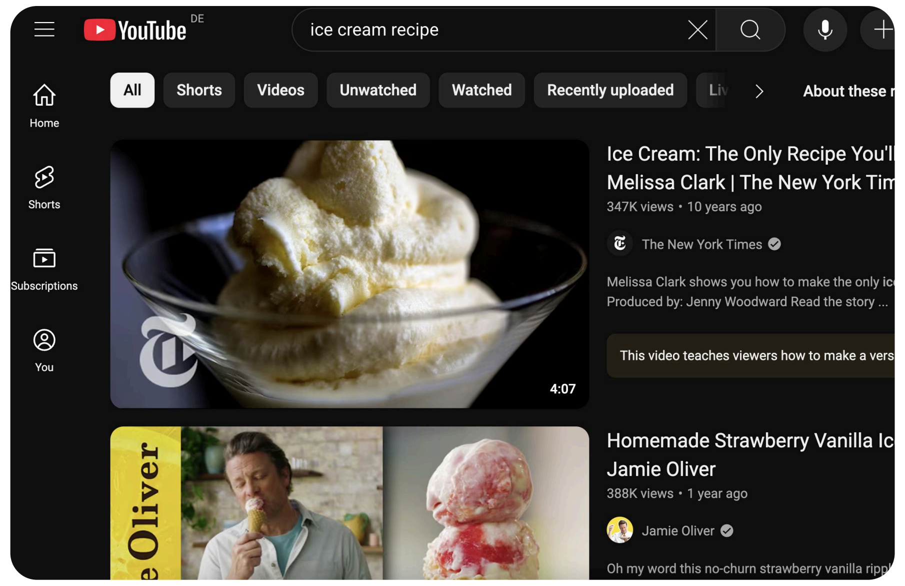
Qwen 2.5 VL

Running Application:
YouTube (ice cream), Figma (slides)

Displayed Results (Top 2):
1. Video Title: Ice Cream: The Only Recipe You'll Ever Need | NYT

The user browsed **YouTube for ice cream recipes**. They **started on the Figma tab, then switched to YouTube** and searched for "ice cream recipe." Finally, the user is reviewing video results....

General User Model (GUM) System



Observe

Transcribe screenshots in detail

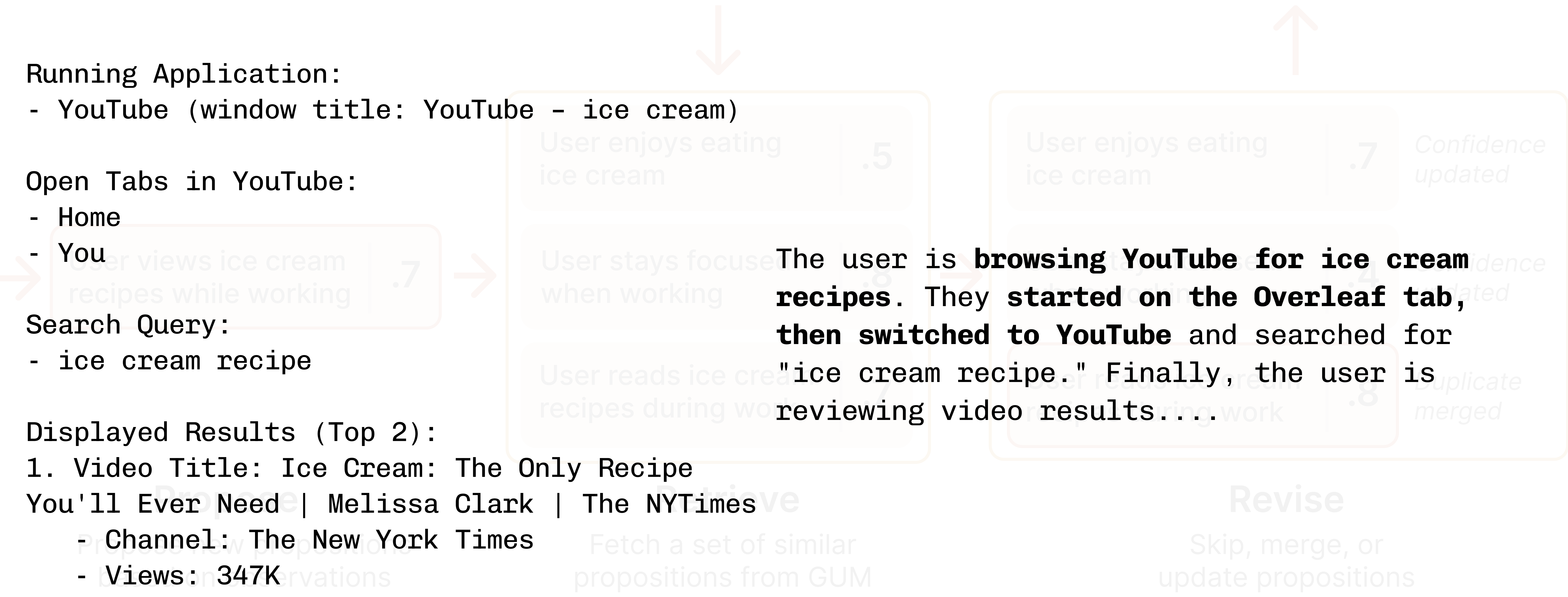
Running Application:
- YouTube (window title: YouTube - ice cream)

Open Tabs in YouTube:
- Home
- You

Search Query:
- ice cream recipe

Displayed Results (Top 2):
1. Video Title: Ice Cream: The Only Recipe You'll Ever Need | Melissa Clark | The NYTimes
    - Channel: The New York Times
    - Views: 347K
    - Uploaded: 10 years ago
    - Duration: 4:07

The user is **browsing YouTube for ice cream recipes**. They **started on the Overleaf tab, then switched to YouTube** and searched for "ice cream recipe." Finally, the user is reviewing video results....


User views ice cream recipes while working
.7
User enjoys eating ice cream
.5
User stays focused when working
.8
User reads ice cream recipes during work
.7
User enjoys eating ice cream
.7
Confidence updated
Confidence updated
Duplicate merged
Propose
Propose new propositions based on observations
Retrieve
Fetch a set of similar propositions from GUM
Revise
Skip, merge, or update propositions

Prompt Qwen 2.5 VL to effectively transcribe screenshot observations

Screenshots captured on click, keystroke; up to 10 unique frames at once

Observe

raw input and transcribe into text

Propose

new propositions about the user

Retrieve

similar propositions from the past

Revise

all propositions and update

```
The user browsed YouTube for ice cream
recipes. They started on the Figma tab,
then switched to YouTube and searched
for "ice cream recipe." Finally, the
user is reviewing video results....
```

Observe
raw input and transcribe into text

Propose
new propositions about the user

Retrieve
similar propositions from the past

Revise
all propositions and update

```
The user browsed YouTube for ice cream
recipes. They started on the Figma tab,
then switched to YouTube and searched
for "ice cream recipe." Finally, the
user is reviewing video results....
```

↓

reasoning: **“The user appears distracted, switching focus between an ice cream recipe video and typing intermittently in an Overleaf window.”**

confidence: **“7”** [1-10 score]

Observe

raw input and transcribe into text

Propose

new propositions about the user

Retrieve

similar propositions from the past

Revise

all propositions and update

```
The user browsed YouTube for ice cream
recipes. They started on the Figma tab,
then switched to YouTube and searched
for "ice cream recipe." Finally, the
user is reviewing video results....
```

↓

reasoning: **“The user appears distracted, switching focus between an ice cream recipe video and typing intermittently in an Overleaf window.”**

confidence: **“7”** [1-10 score]

↓

Distracted by ice cream recipes while working | 0.7

Subscrib

Working on a UIST talk | 0.9

Observe

raw input and transcribe into text

Propose

new propositions about the user

Retrieve

similar propositions from the past

Revise

all propositions and update

Subscrib

Working on a UIST talk 0.9

Distracted by ice cream recipes while working 0.7

Observe

raw input and transcribe into text

Propose

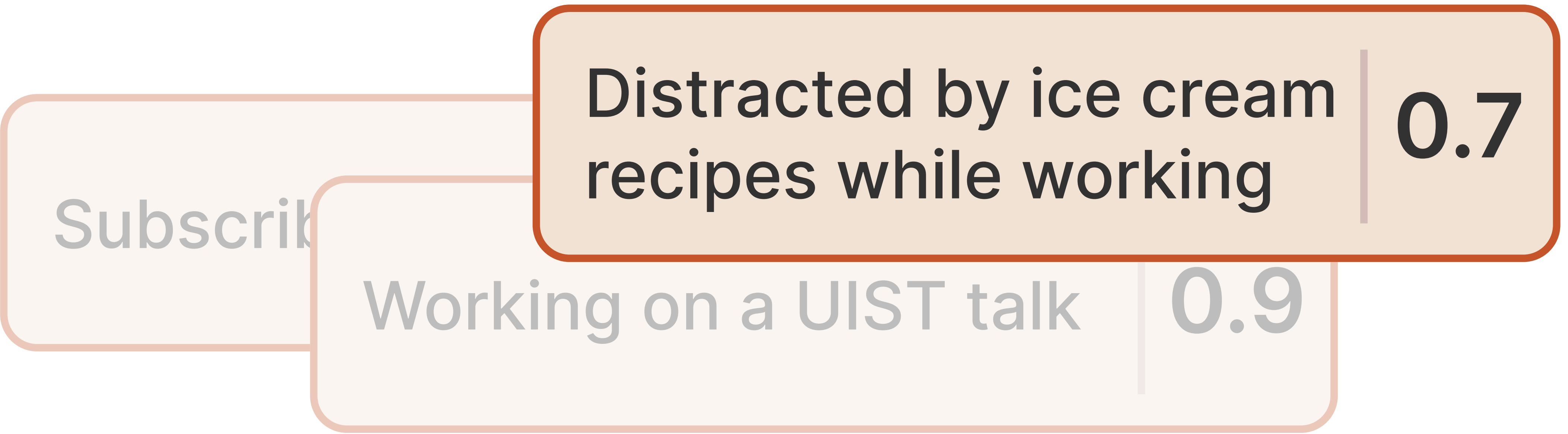
new propositions about the user

Retrieve

similar propositions from the past

Revise

all propositions and update



Distracted by ice cream recipes while working 0.7

Working on a UIST talk 0.9

Subscrib

Observe

raw input and transcribe into text

Propose

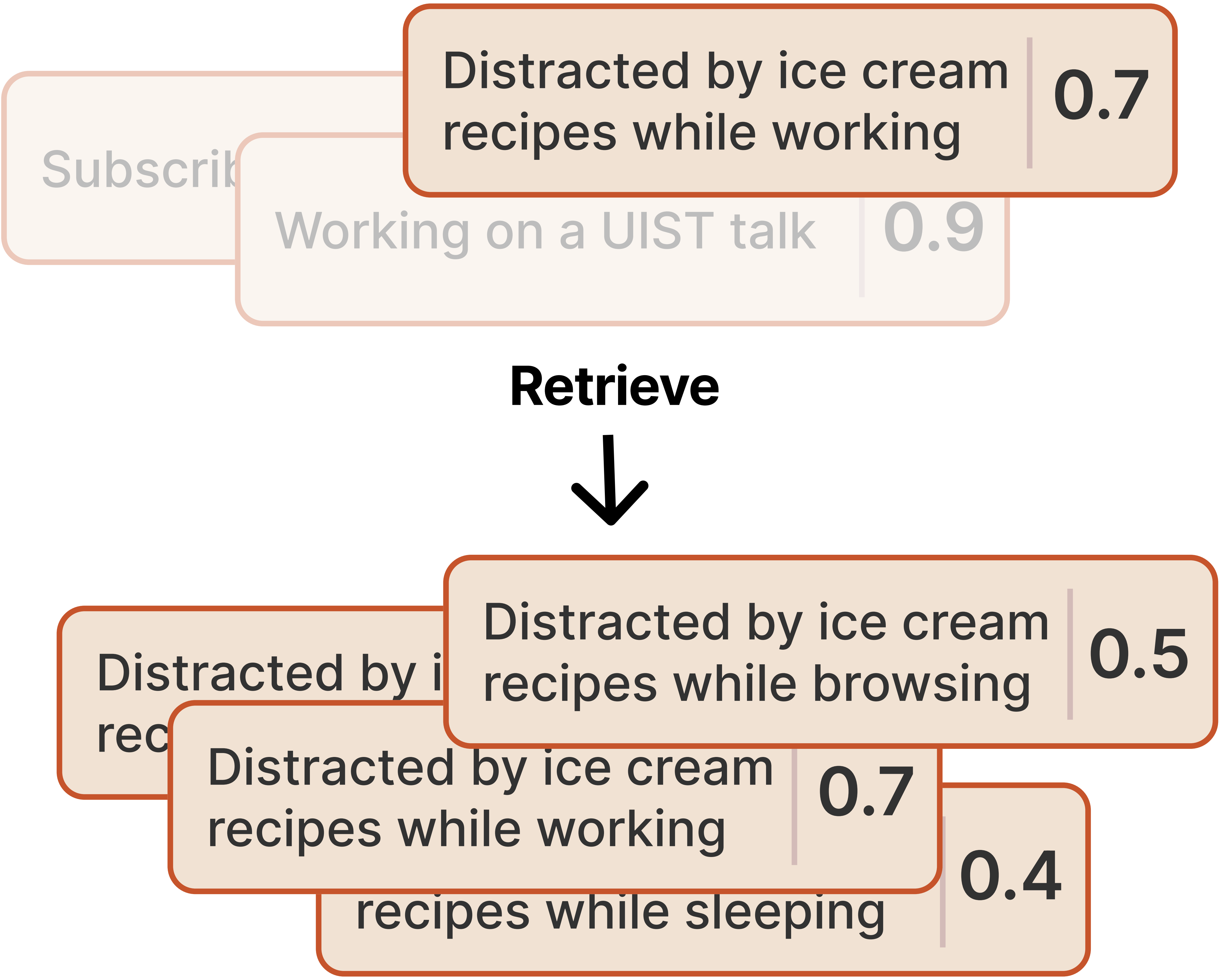
new propositions about the user

Retrieve

similar propositions from the past

Revise

all propositions and update


Subscrib
Working on a UIST talk
0.9
Distracted by ice cream recipes while working
0.7
Retrieve
Distracted by i
rec
Distracted by ice cream recipes while browsing
0.5
Distracted by ice cream recipes while working
0.7
recipes while sleeping
0.4

Observe

raw input and transcribe into text

Propose

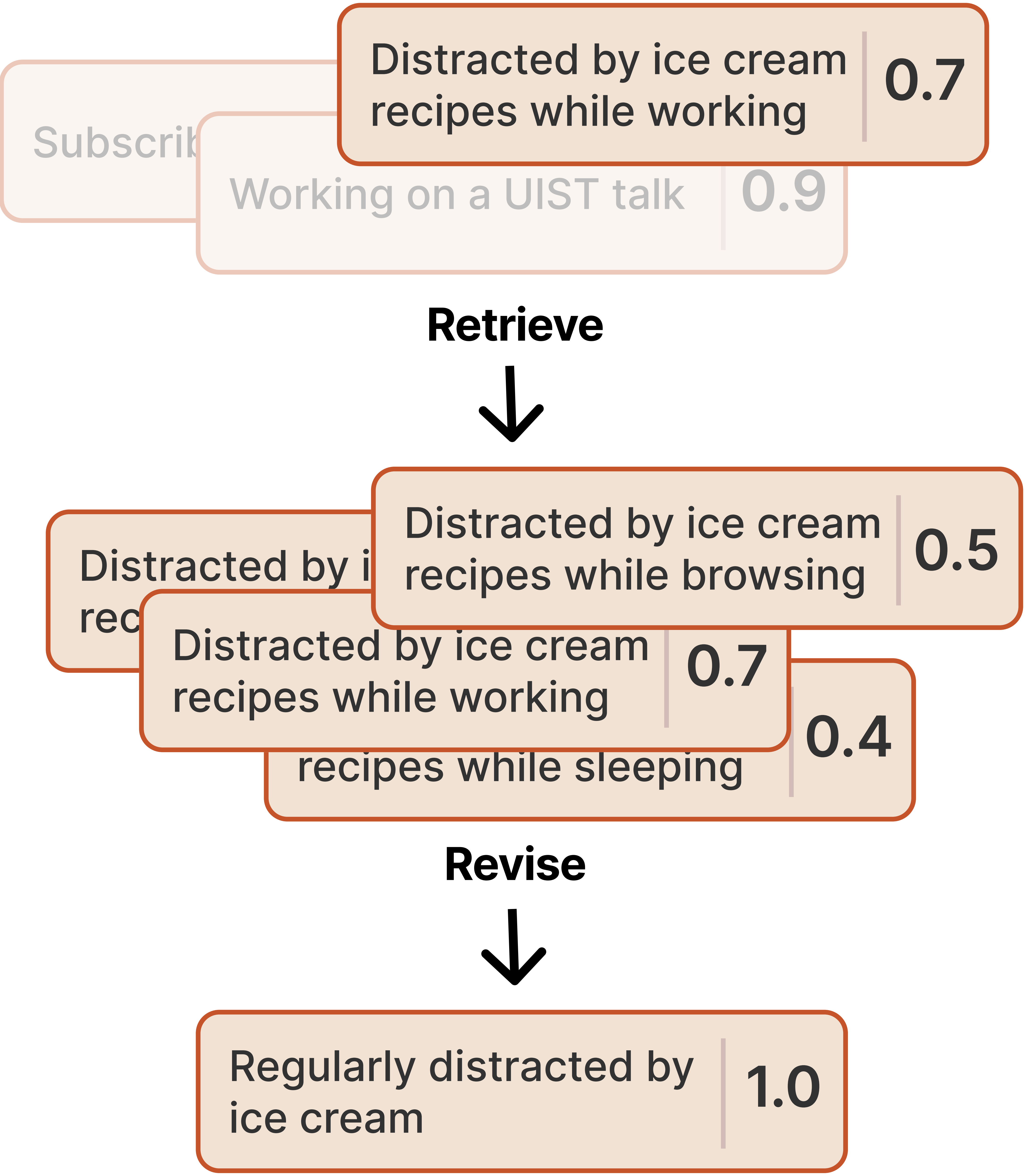
new propositions about the user

Retrieve

similar propositions from the past

Revise

all propositions and update



Subscrib

Working on a UIST talk 0.9

Distracted by ice cream recipes while working 0.7

Retrieve

↓

Distracted by i
rec

Distracted by ice cream recipes while browsing 0.5

Distracted by ice cream recipes while working 0.7

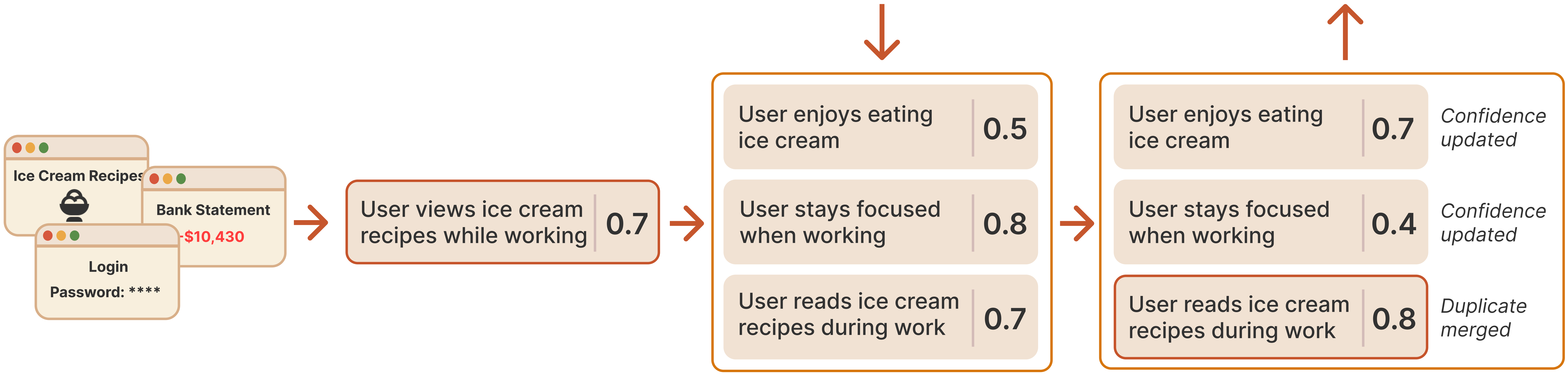
recipes while sleeping 0.4

Revise

↓

Regularly distracted by ice cream 1.0

General User Model (GUM)



User views ice cream recipes while working 0.7

User enjoys eating ice cream 0.5

User stays focused when working 0.8

User reads ice cream recipes during work 0.7

User enjoys eating ice cream 0.7 *Confidence updated*

User stays focused when working 0.4 *Confidence updated*

User reads ice cream recipes during work 0.8 *Duplicate merged*

Observe
Transcribe screenshots in detail

Propose
Propose new propositions based on observations

Retrieve
Fetch a set of similar propositions from GUM

Revise
Skip, merge, or update propositions

Applications of GUMs

If *only* I had a model of user context…

Applications of GUMs

If *only* I had a model of user context...

Observations

Snapshots of unstructured, multimodal interaction

Propositions

Confidence-weighed inferences

i was 10 the last time i wore a suit lol

1 day ago

John & Emily

YOU'RE INVITED

CHICAGO, IL • JULY 2025

Re: Quals for Stanford University Ph.Ds

...

g at achines 0.9

Nervous about taking quals next quarter 0.4

Makes coffee for his friends at the lab 0.9

Nervous about taking quals next term 0.6

Interested in making more Chai at home

erested in Human-AI eraction 1.0

Uses casual tone with friends 0.8

Makes and eats ice cream very regularly 0.9

Is a Ph.D. student at Stanford University 0.8

Pretentious abou coffee options

Proposition	Confidence
Doesn't own any suitable formal wear	1.0
Likely going to friend's wedding in Chicago	0.8
Needs to budget as a Ph.D. student	0.6

aborating

Interviewing at

Might not make the

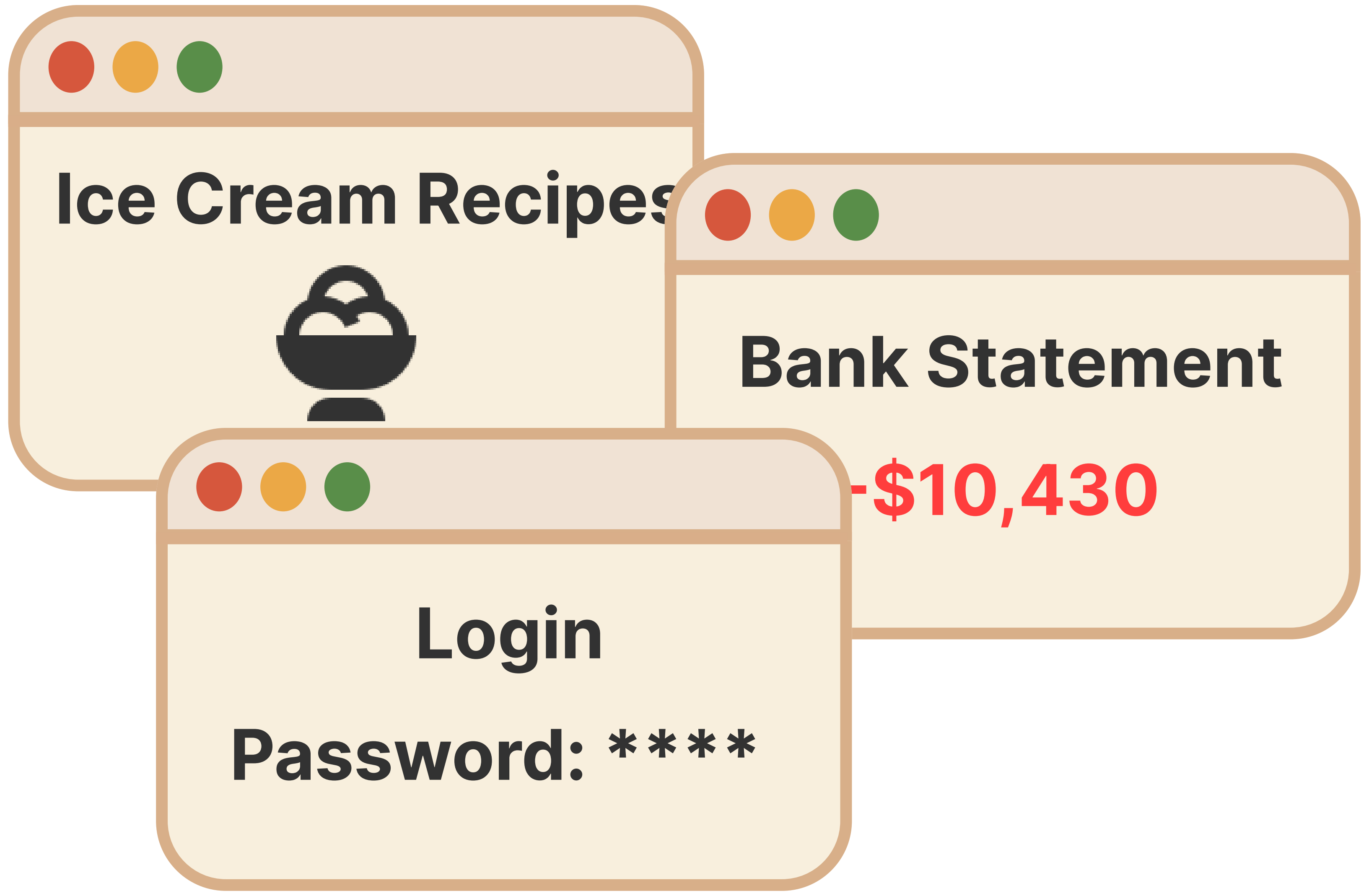
Struggling to use

Excited to visit Berlin

Application

GUMs for Contextual Integrity

Challenge: Sensitive information should not be processed by the GUM.

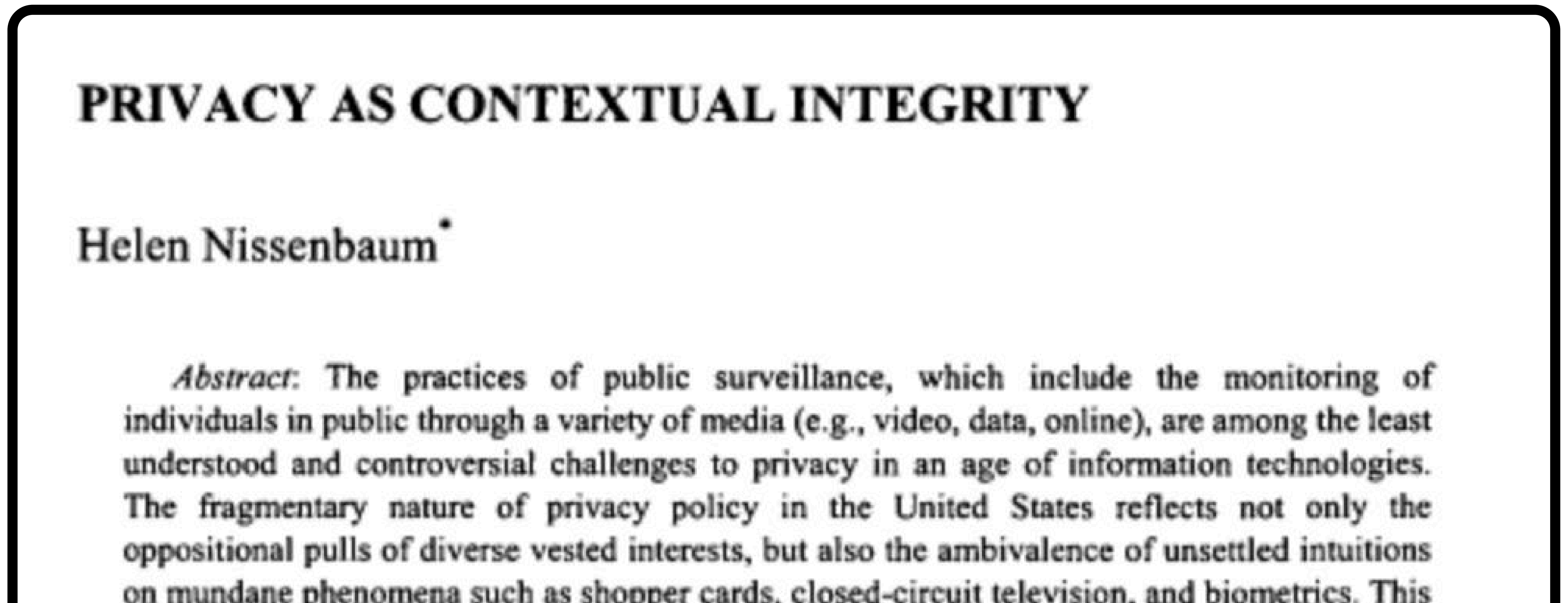


Application

GUMs for Contextual Integrity

Challenge: Sensitive information should not be processed by the GUM.

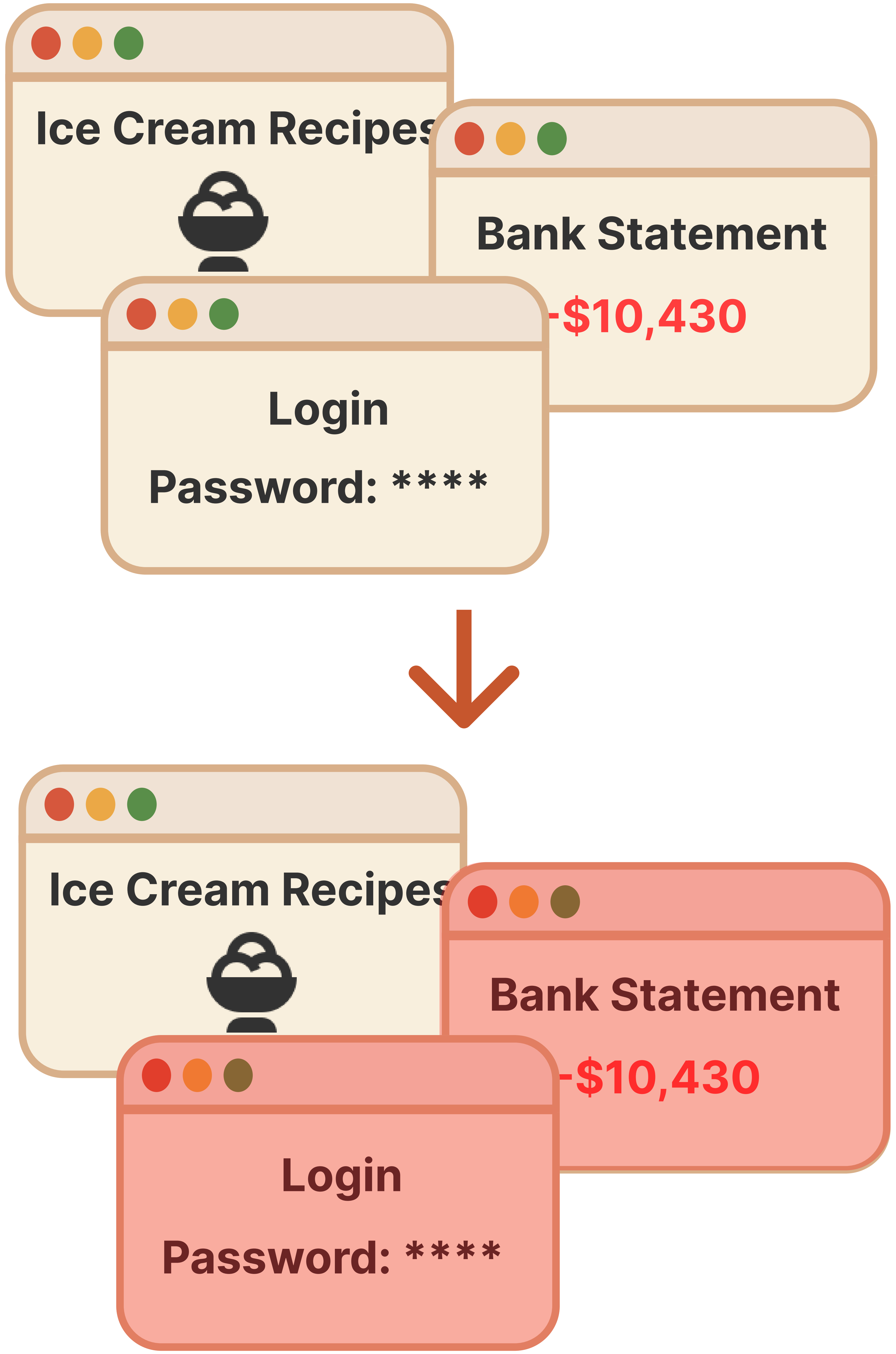
Information flows appropriately according to the **specific norms** and expectations of the **social context**.



PRIVACY AS CONTEXTUAL INTEGRITY

Helen Nissenbaum*

Abstract: The practices of public surveillance, which include the monitoring of individuals in public through a variety of media (e.g., video, data, online), are among the least understood and controversial challenges to privacy in an age of information technologies. The fragmentary nature of privacy policy in the United States reflects not only the oppositional pulls of diverse vested interests, but also the ambivalence of unsettled intuitions on mundane phenomena such as shopper cards, closed-circuit television, and biometrics. This



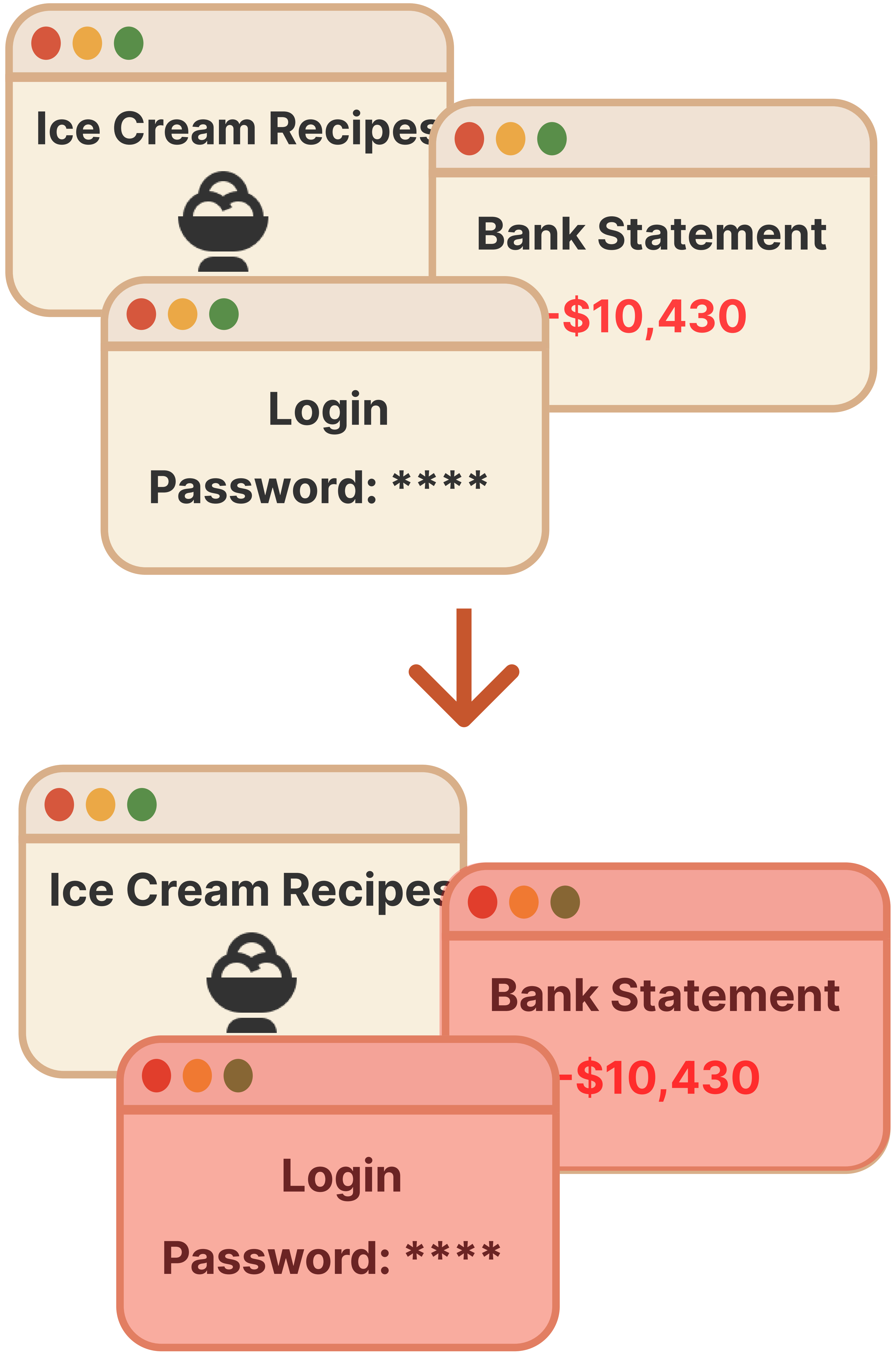
Application

GUMs for Contextual Integrity

Challenge: Sensitive information should not be processed by the GUM.

Information flows appropriately according to the **specific norms** and expectations of the **social context**.

If only we had a good model of a user's context...



Nissenbaum, H. (2004). Privacy as contextual integrity. Washington Law Review, 79(1), 119–157.

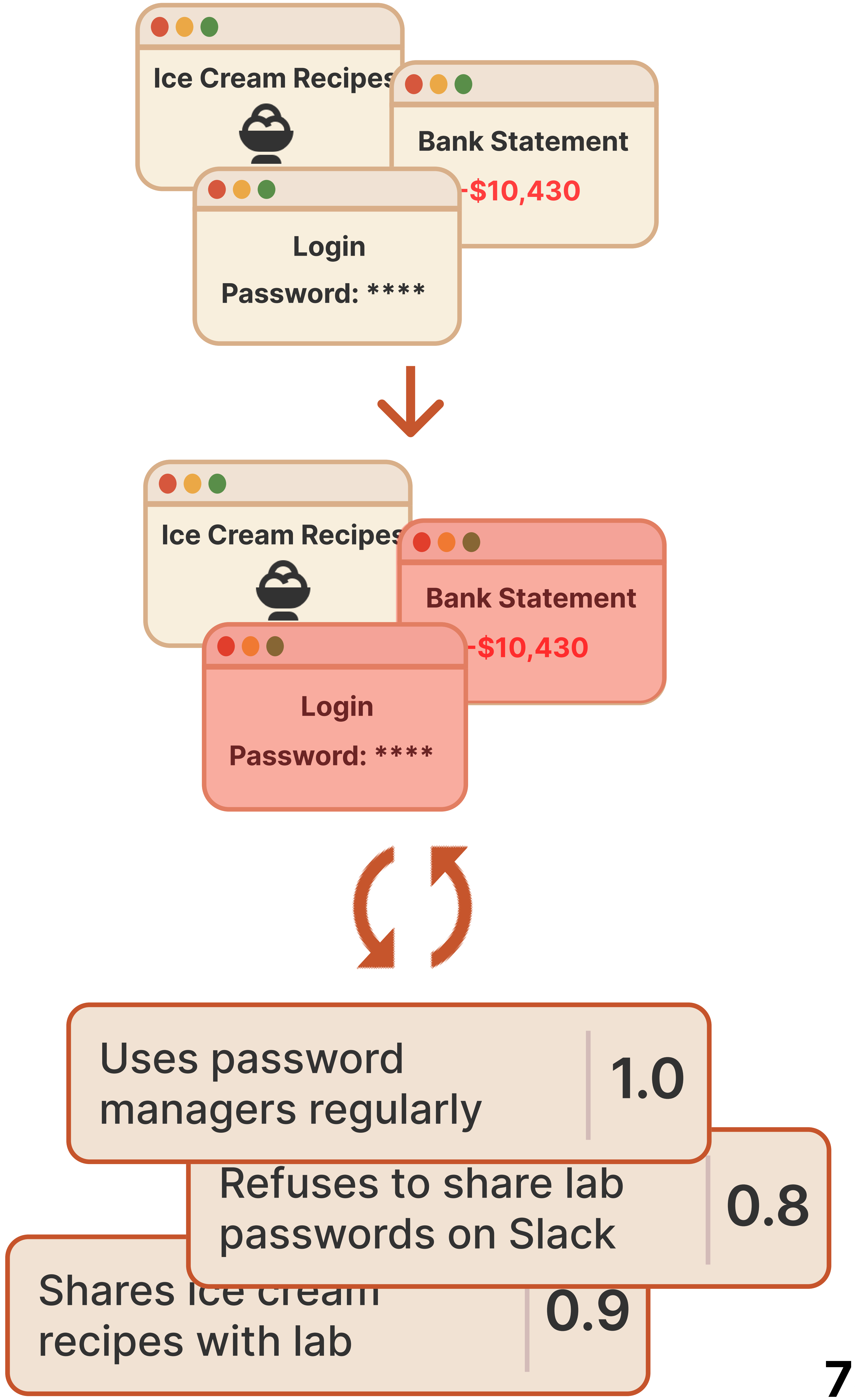
Application

GUMs for Contextual Integrity

Challenge: Sensitive information should not be processed by the GUM.

Information flows appropriately according to the **specific norms** and expectations of the **social context**.

If only we had a good model of a user's context...



Uses password managers regularly 1.0

Refuses to share lab passwords on Slack 0.8

Shares ice cream recipes with lab 0.9

Application

Mixed Initiative Interaction

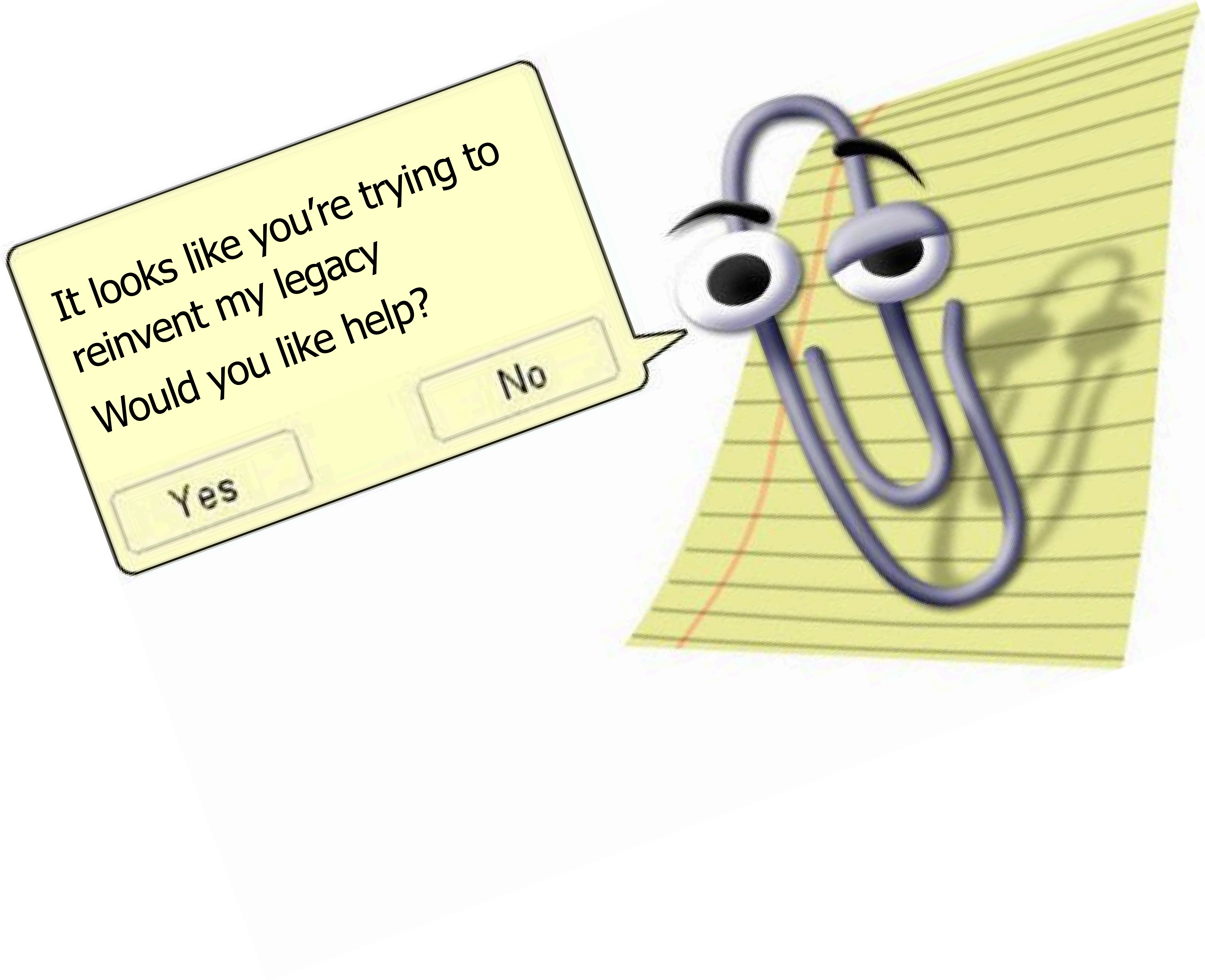
Challenge: Knowing when an AI system should take initiative depends on the context

Horvitz, E. (1999). Principles of Mixed-Initiative User Interfaces.
Proceedings of the SIGCHI Conference on Human Factors in Computing Sytems

Application

Mixed Initiative Interaction

Challenge: Knowing when an AI system should take initiative depends on the context

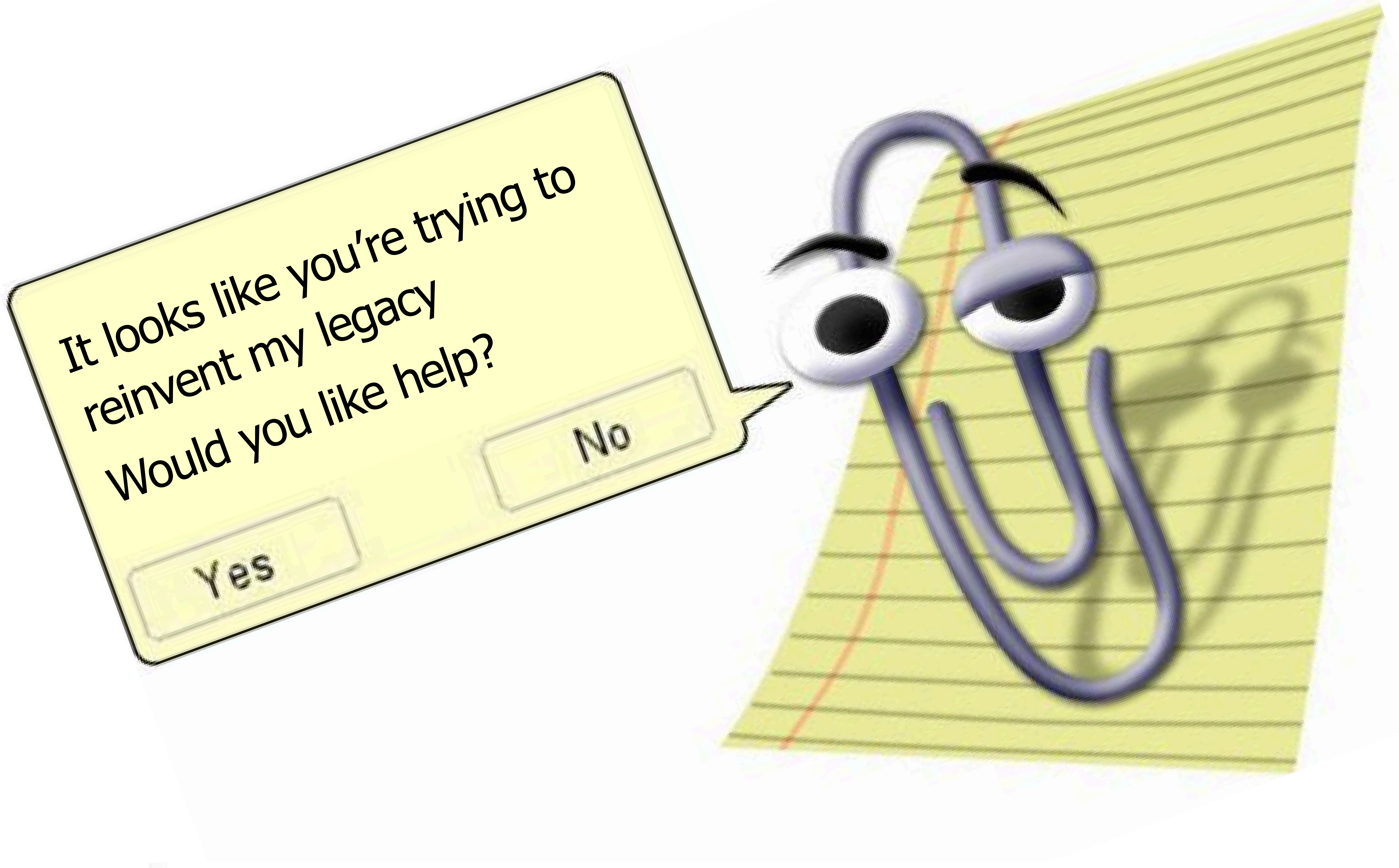


Horvitz, E. (1999). Principles of Mixed-Initiative User Interfaces.
Proceedings of the SIGCHI Conference on Human Factors in Computing Sytems

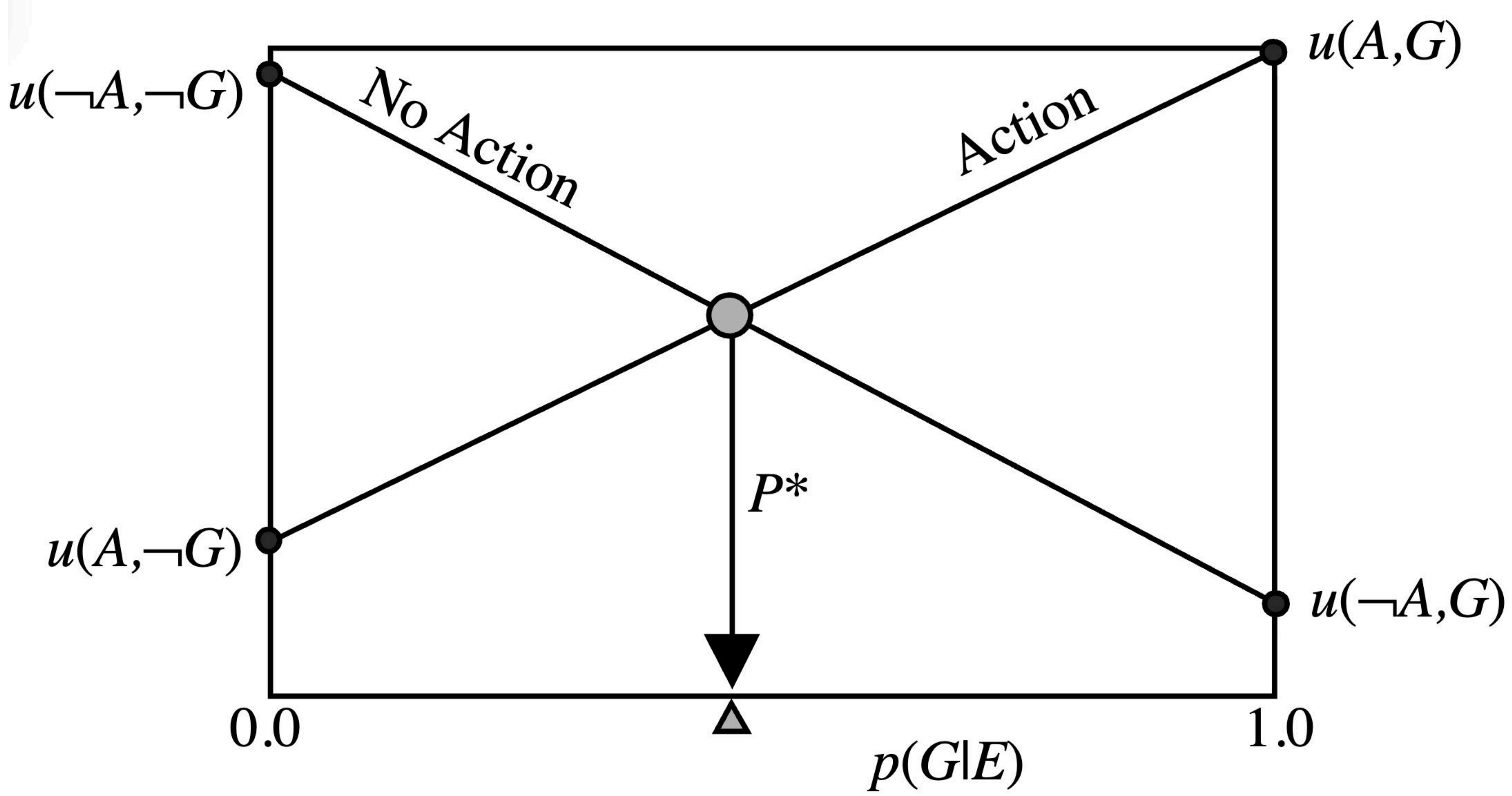
Application

Mixed Initiative Interaction

Challenge: Knowing when an AI system should take initiative depends on the context



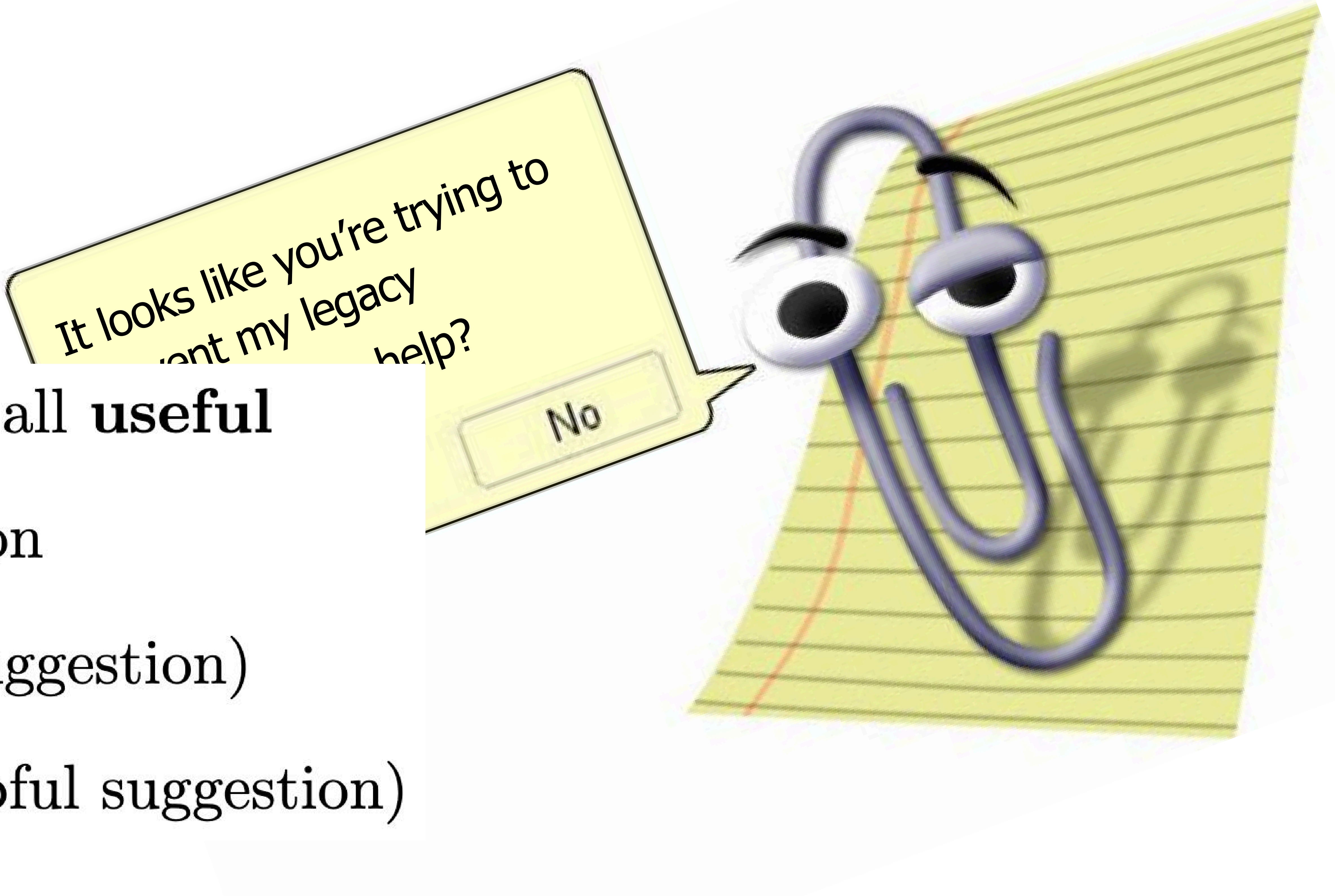
Utility from interruption v.s. no interruption depending on context



Horvitz, E. (1999). Principles of Mixed-Initiative User Interfaces. Proceedings of the SIGCHI Conference on Human Factors in Computing Sytems

Application

Mixed Initiative Interaction



$P(\tau)$: Probability / confidence that suggestion is at all **useful**

$B \in [0, 10]$: **Benefit** of completing a useful suggestion

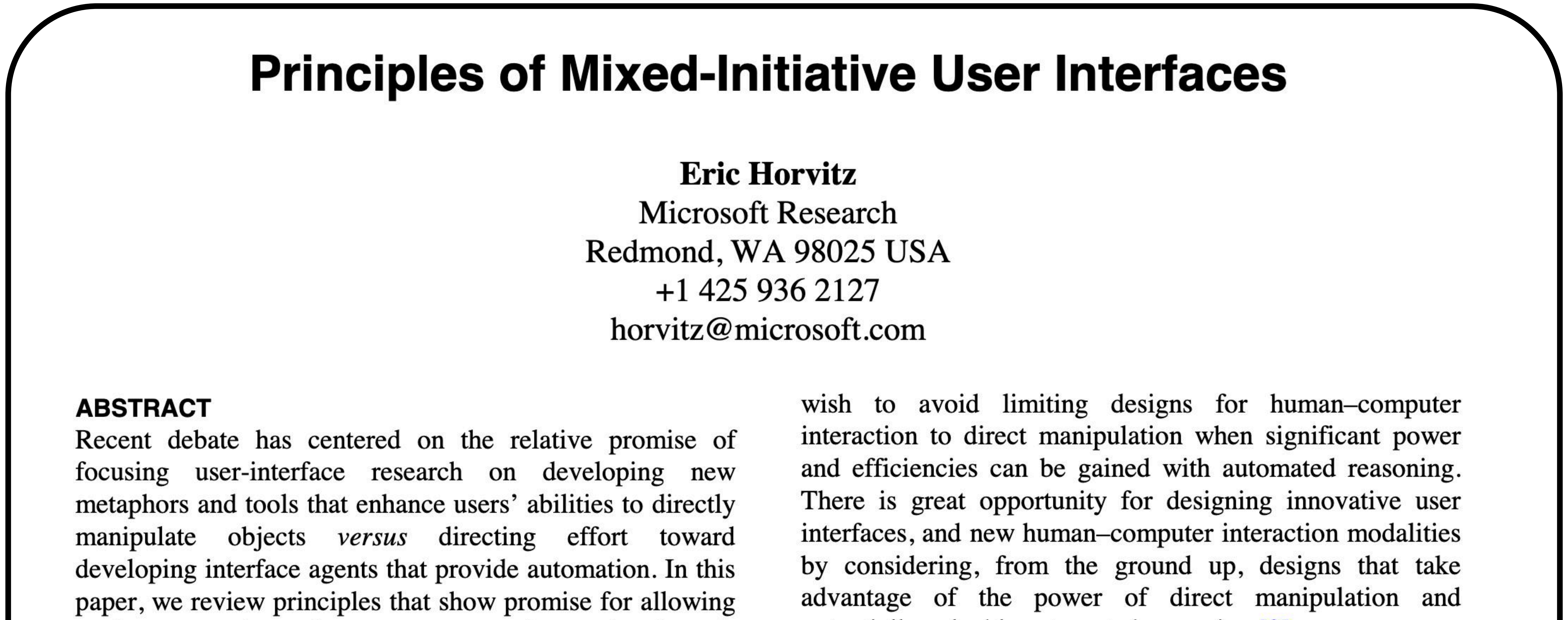
$C_{FP} \in [0, 10]$: **Cost of false positive** (unhelpful suggestion)

$C_{FN} \in [0, 10]$: **Cost of false negative** (missed helpful suggestion)

If you have these quantities, you can determine the **expected utility** of

- interrupting v.s.
- not interrupting a user.

and identify a threshold.



Principles of Mixed-Initiative User Interfaces

Eric Horvitz
Microsoft Research
Redmond, WA 98025 USA
+1 425 936 2127
horvitz@microsoft.com

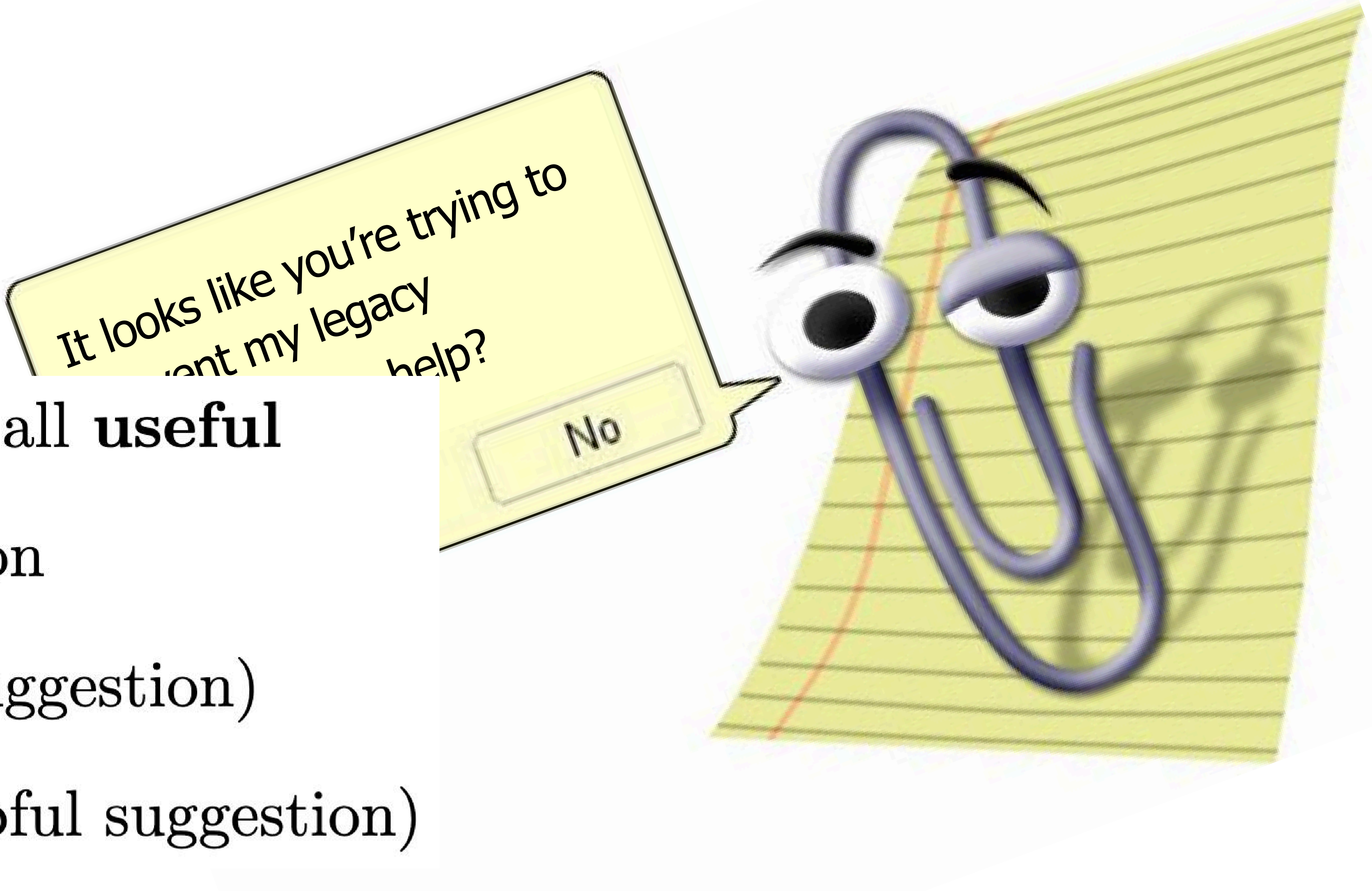
ABSTRACT

Recent debate has centered on the relative promise of focusing user-interface research on developing new metaphors and tools that enhance users' abilities to directly manipulate objects *versus* directing effort toward developing interface agents that provide automation. In this paper, we review principles that show promise for allowing

wish to avoid limiting designs for human–computer interaction to direct manipulation when significant power and efficiencies can be gained with automated reasoning. There is great opportunity for designing innovative user interfaces, and new human–computer interaction modalities by considering, from the ground up, designs that take advantage of the power of direct manipulation and

Application

Mixed Initiative Interaction



$P(\tau)$: Probability / confidence that suggestion is at all **useful**

$B \in [0, 10]$: **Benefit** of completing a useful suggestion

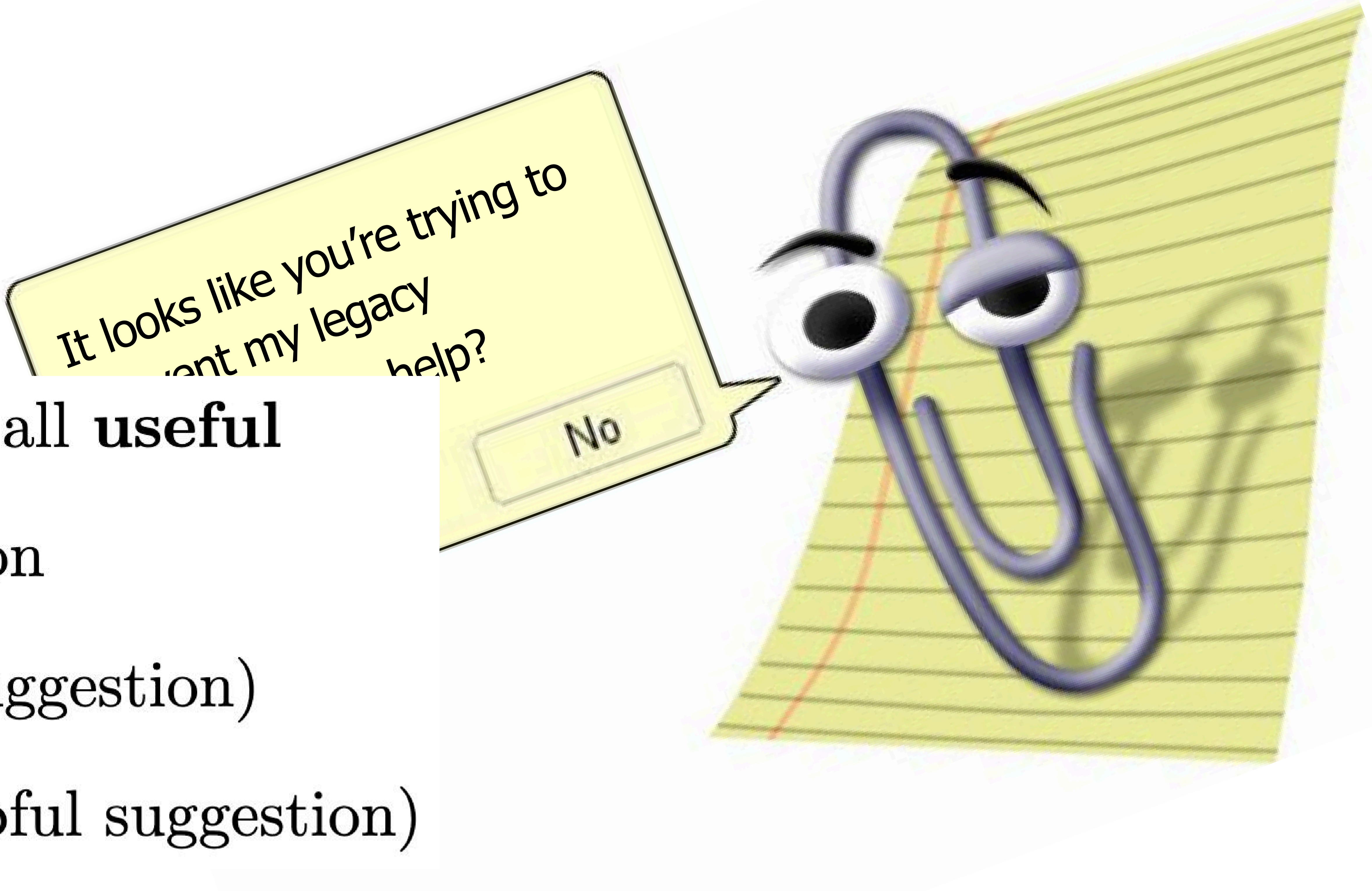
$C_{FP} \in [0, 10]$: **Cost of false positive** (unhelpful suggestion)

$C_{FN} \in [0, 10]$: **Cost of false negative** (missed helpful suggestion)

$$\begin{aligned} E[U_{\text{interrupt}}] &= P(\tau) \cdot B \; + \; (1 - P(\tau)) \cdot (-C_{FP}) \\ &= \underbrace{\text{Confidence in usefulness} \times \text{Benefit}}_{\text{When useful}} + \underbrace{(1 - \text{Confidence}) \times -\text{False positive cost}}_{\text{When not useful}} \end{aligned}$$

Application

Mixed Initiative Interaction



$P(\tau)$: Probability / confidence that suggestion is at all **useful**

$B \in [0, 10]$: **Benefit** of completing a useful suggestion

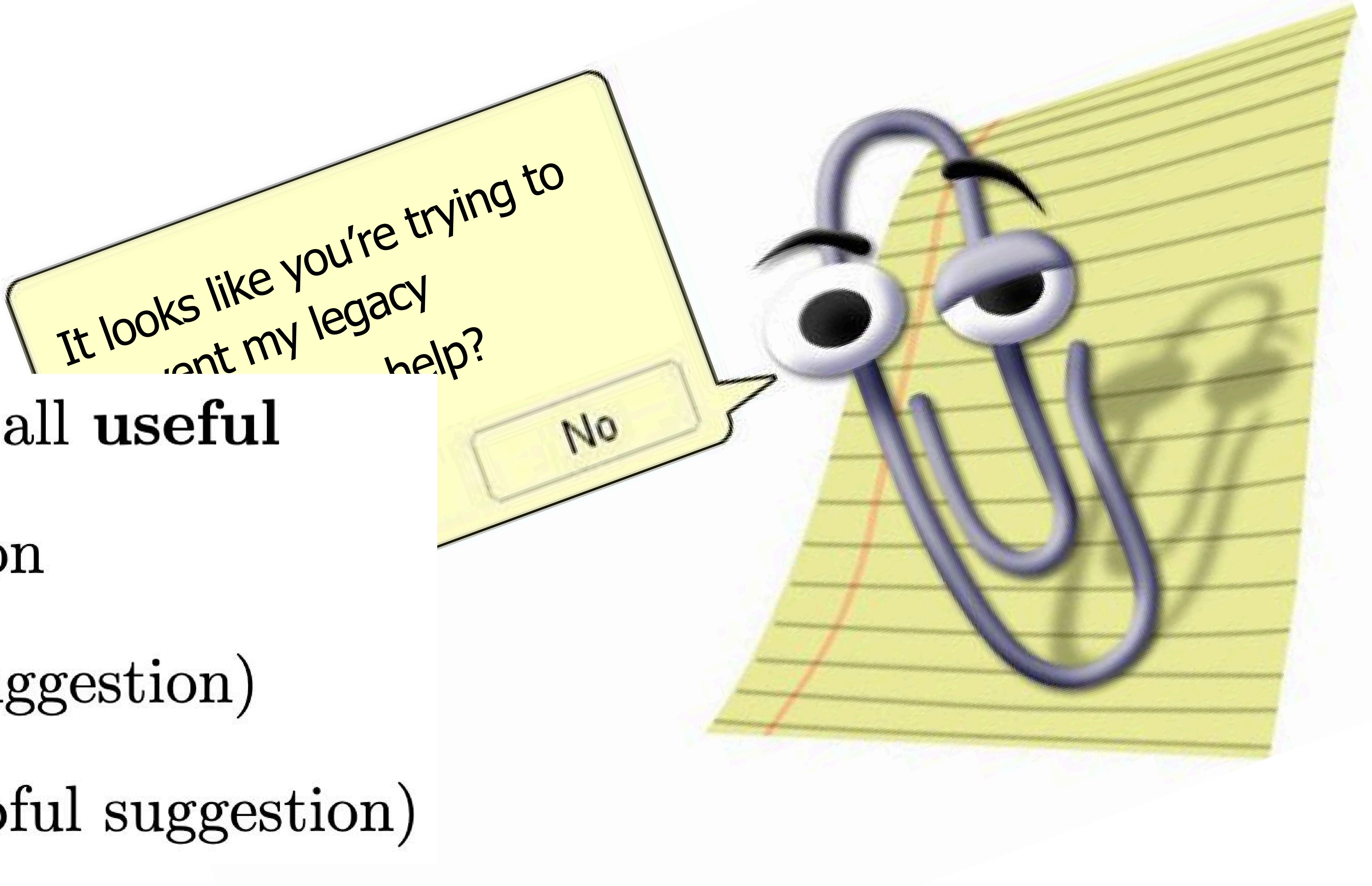
$C_{FP} \in [0, 10]$: **Cost of false positive** (unhelpful suggestion)

$C_{FN} \in [0, 10]$: **Cost of false negative** (missed helpful suggestion)

$$\begin{aligned} E[U_{\text{interrupt}}] &= P(\tau) \cdot B + (1 - P(\tau)) \cdot (-C_{FP}) \\ &= \underbrace{\text{Confidence in usefulness} \times \text{Benefit}}_{\text{When useful}} + \underbrace{(1 - \text{Confidence}) \times -\text{False positive cost}}_{\text{When not useful}} \end{aligned}$$

Application

Mixed Initiative Interaction



$P(\tau)$: Probability / confidence that suggestion is at all **useful**

$B \in [0, 10]$: **Benefit** of completing a useful suggestion

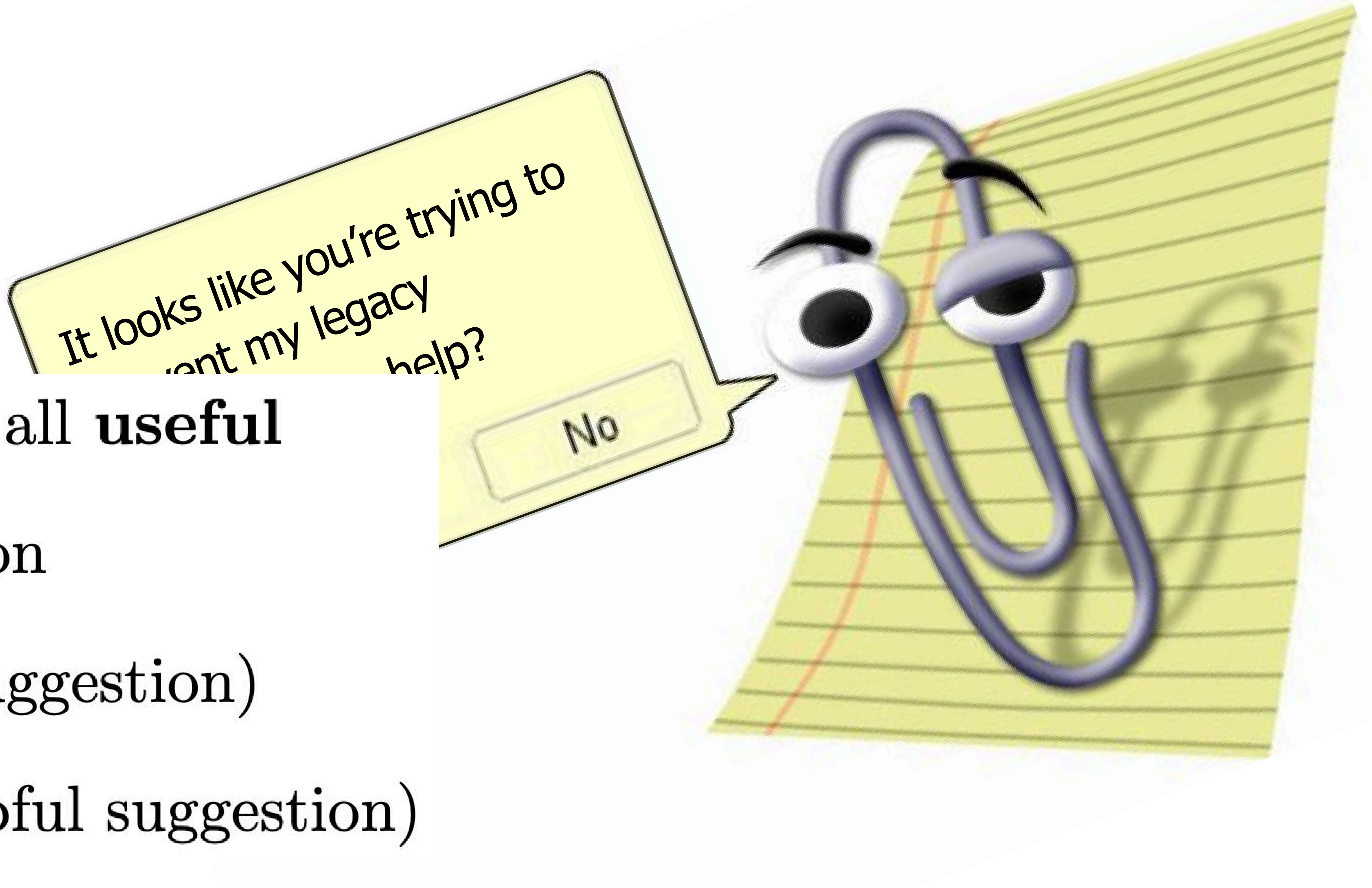
$C_{FP} \in [0, 10]$: **Cost of false positive** (unhelpful suggestion)

$C_{FN} \in [0, 10]$: **Cost of false negative** (missed helpful suggestion)

$$\begin{aligned} E[U_{\text{interrupt}}] &= P(\tau) \cdot B \; + \; (1 - P(\tau)) \cdot (-C_{FP}) \\ &= \underbrace{\text{Confidence in usefulness} \times \text{Benefit}}_{\text{When useful}} + \underbrace{(1 - \text{Confidence}) \times -\text{False positive cost}}_{\text{When not useful}} \end{aligned}$$

Application

Mixed Initiative Interaction



$P(\tau)$: Probability / confidence that suggestion is at all **useful**

$B \in [0, 10]$: **Benefit** of completing a useful suggestion

$C_{FP} \in [0, 10]$: **Cost of false positive** (unhelpful suggestion)

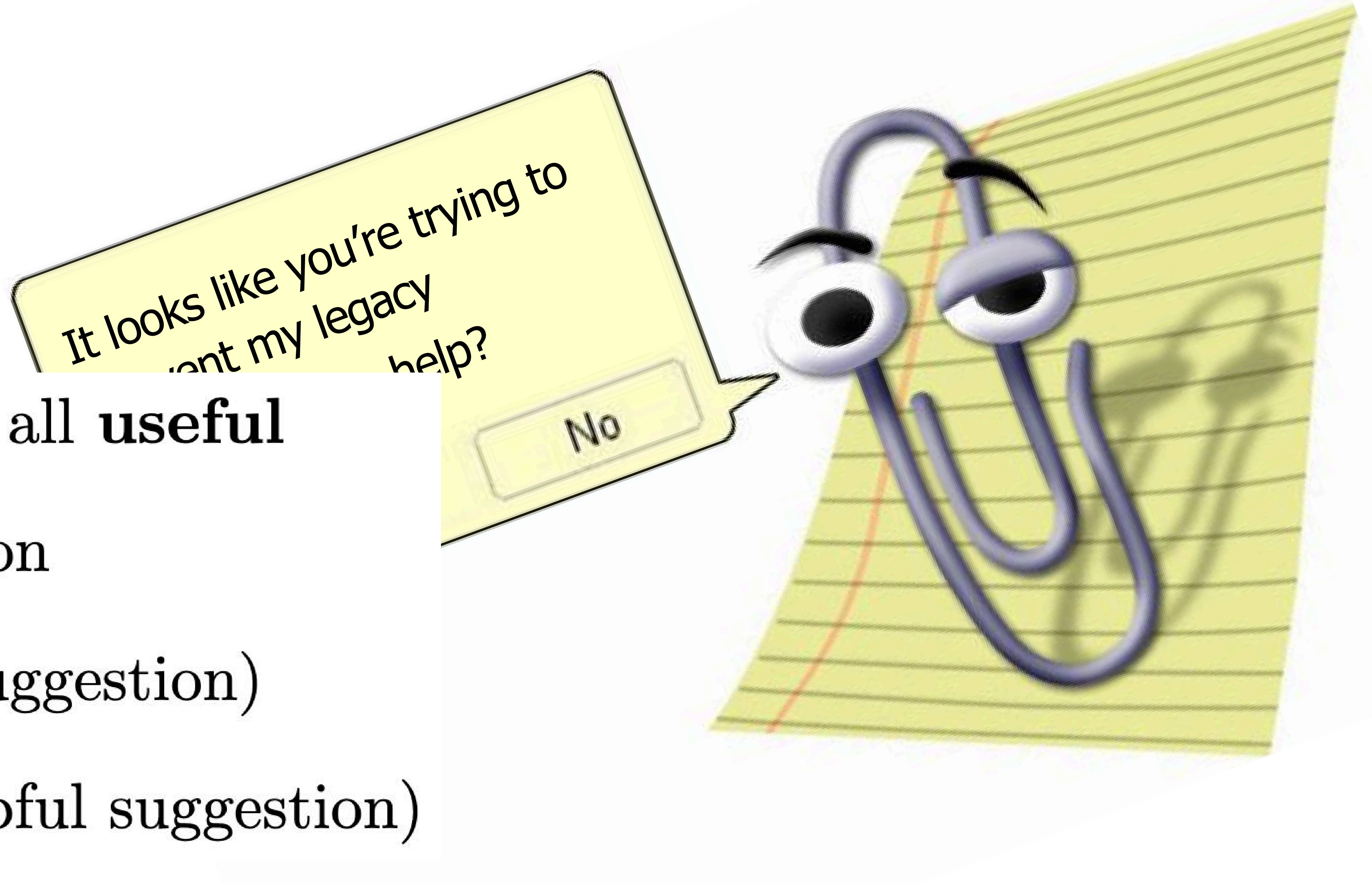
$C_{FN} \in [0, 10]$: **Cost of false negative** (missed helpful suggestion)

$$
\begin{aligned}
E[U_{\text{interrupt}}] &= P(\tau) \cdot B \;+\; (1 - P(\tau)) \cdot (-C_{FP}) \\
&= \underbrace{\text{Confidence in usefulness} \times \text{Benefit}}_{\text{When useful}} \;+\; \underbrace{(1 - \text{Confidence}) \times -\text{False positive cost}}_{\text{When not useful}}
\end{aligned}
$$

$$
\begin{aligned}
E[U_{\neg\text{interrupt}}] &= P(\tau) \cdot (-C_{FN}) \;+\; (1 - P(\tau)) \cdot 0 \\
&= \underbrace{\text{Confidence in usefulness} \times -\text{False negative cost}}_{\text{Missed opportunity}}
\end{aligned}
$$

Application

Mixed Initiative Interaction



$P(\tau)$: Probability / confidence that suggestion is at all **useful**

$B \in [0, 10]$: **Benefit** of completing a useful suggestion

$C_{FP} \in [0, 10]$: **Cost of false positive** (unhelpful suggestion)

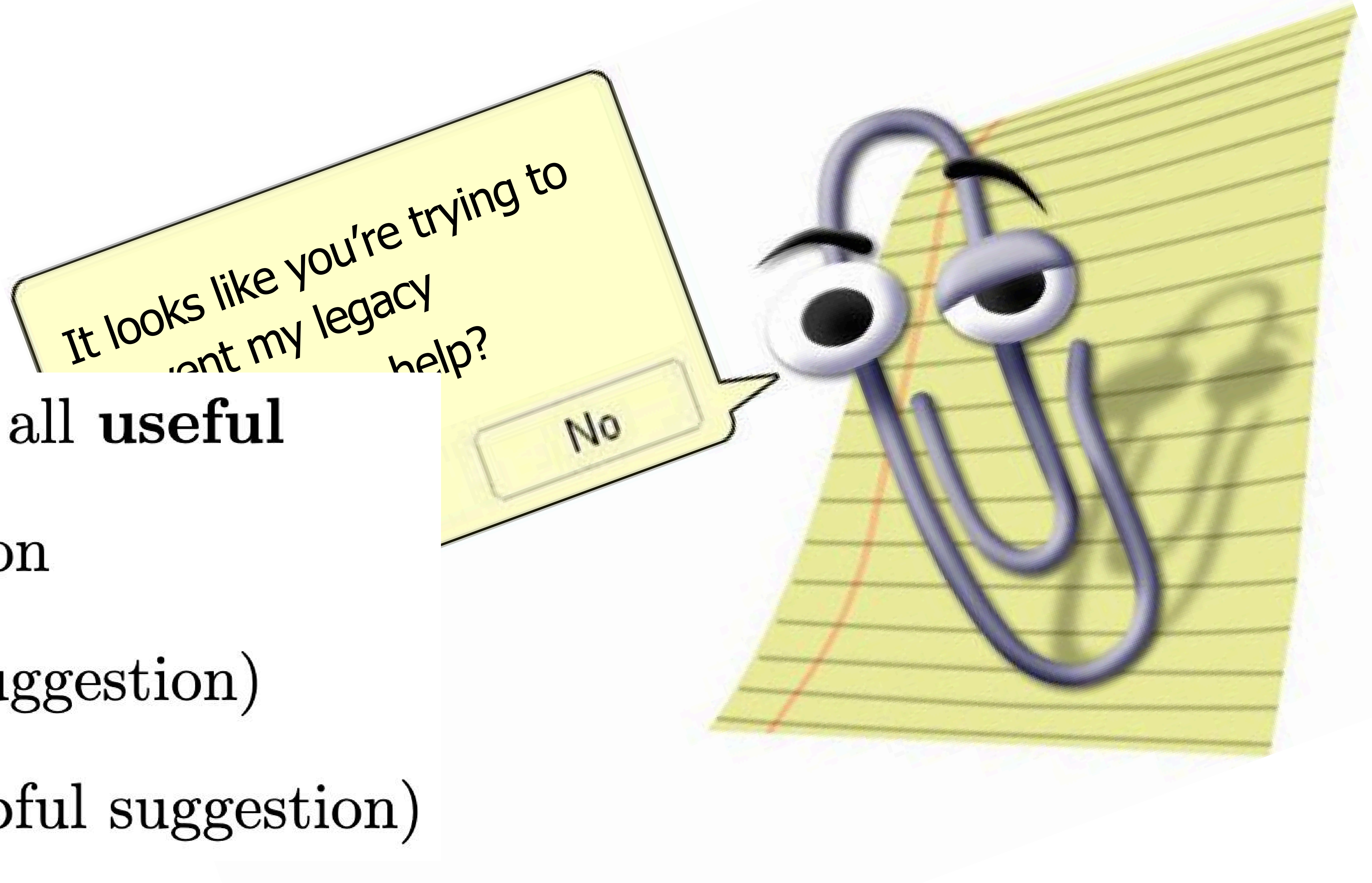
$C_{FN} \in [0, 10]$: **Cost of false negative** (missed helpful suggestion)

$$\begin{aligned} E[U_{\text{interrupt}}] &= P(\tau) \cdot B \; + \; (1 - P(\tau)) \cdot (-C_{FP}) \\ &= \underbrace{\text{Confidence in usefulness} \times \text{Benefit}}_{\text{When useful}} + \underbrace{(1 - \text{Confidence}) \times -\text{False positive cost}}_{\text{When not useful}} \end{aligned}$$

$$\begin{aligned} E[U_{\neg\text{interrupt}}] &= P(\tau) \cdot (-C_{FN}) \; + \; (1 - P(\tau)) \cdot 0 \\ &= \underbrace{\text{Confidence in usefulness} \times -\text{False negative cost}}_{\text{Missed opportunity}} \end{aligned}$$

Application

Mixed Initiative Interaction



$P(\tau)$: Probability / confidence that suggestion is at all **useful**

$B \in [0, 10]$: **Benefit** of completing a useful suggestion

$C_{FP} \in [0, 10]$: **Cost of false positive** (unhelpful suggestion)

$C_{FN} \in [0, 10]$: **Cost of false negative** (missed helpful suggestion)

$$E[U_{\text{interrupt}}] = P(\tau) \cdot B \; + \; (1 - P(\tau)) \cdot (-C_{FP})$$
$$= \underbrace{\text{Confidence in usefulness} \times \text{Benefit}}_{\text{When useful}} + \underbrace{(1 - \text{Confidence}) \times -\text{False positive cost}}_{\text{When not useful}}$$

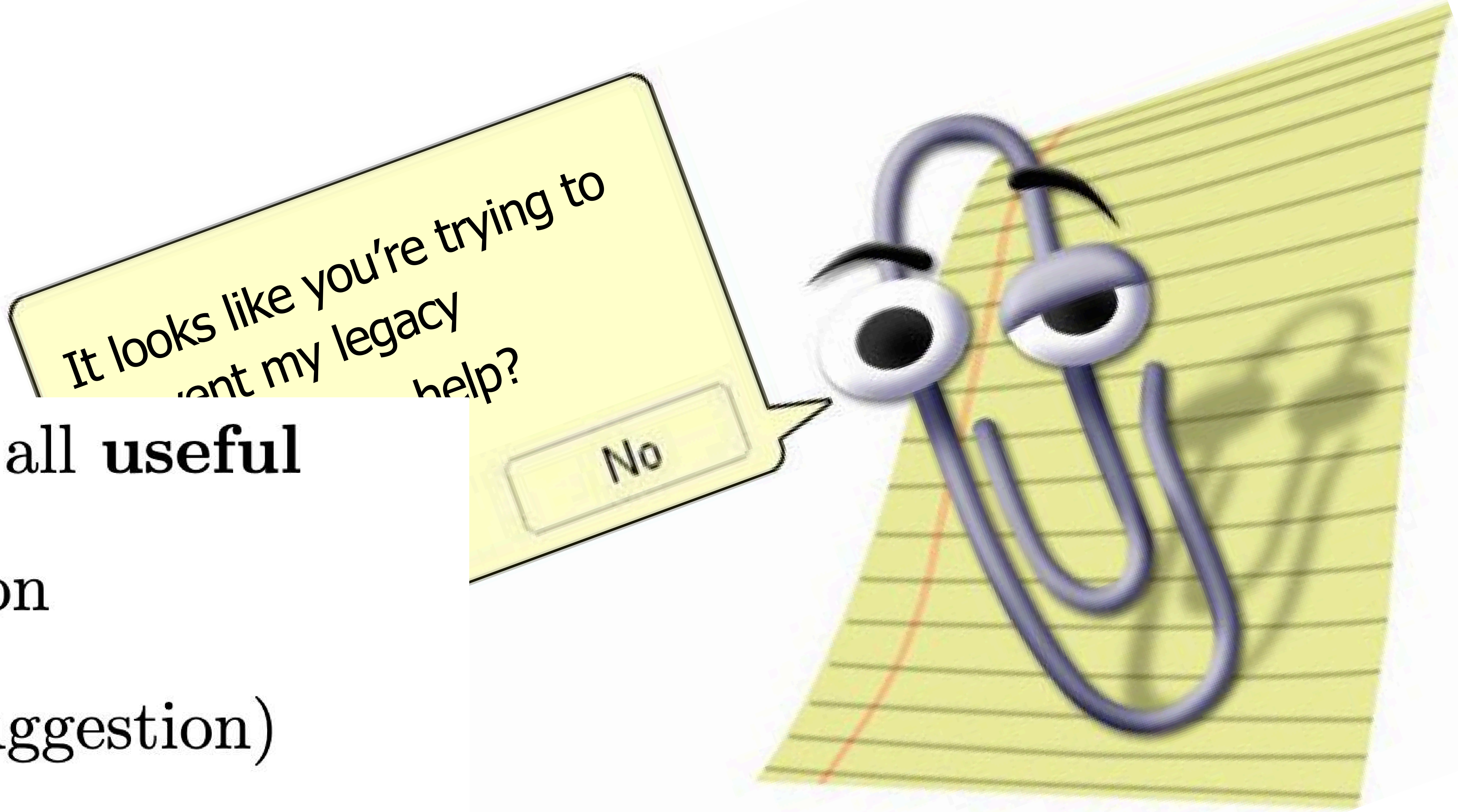
$$E[U_{\neg\text{interrupt}}] = P(\tau) \cdot (-C_{FN}) \; + \; (1 - P(\tau)) \cdot 0$$
$$= \underbrace{\text{Confidence in usefulness} \times -\text{False negative cost}}_{\text{Missed opportunity}}$$

Trigger Suggestion:

$$E[U_{\text{interrupt}}] > E[U_{\neg\text{interrupt}}]$$

Application

Mixed Initiative Interaction



$P(\tau)$: Probability / confidence that suggestion is at all **useful**

$B \in [0, 10]$: **Benefit** of completing a useful suggestion

$C_{FP} \in [0, 10]$: **Cost of false positive** (unhelpful suggestion)

$C_{FN} \in [0, 10]$: **Cost of false negative** (missed helpful suggestion)

$$E[U_{\text{interrupt}}] = P(\tau) \cdot B \; + \; (1 - P(\tau)) \cdot (-C_{FP})$$

$$= \text{Confidence in usefulness} \times \text{Benefit} \; + \; \ldots \times -\text{False positive cost}$$

$$= \underbrace{\text{Confidence in usefulness} \times -\text{False negative cost}}_{\text{Missed opportunity}}$$

$$E[U_{\text{interrupt}}] > E[U_{\neg\text{interrupt}}]$$

OK Omar, the math checks out... but where in the world did you get all those NUMBERS (B, C, P, etc. etc.) from???

Application

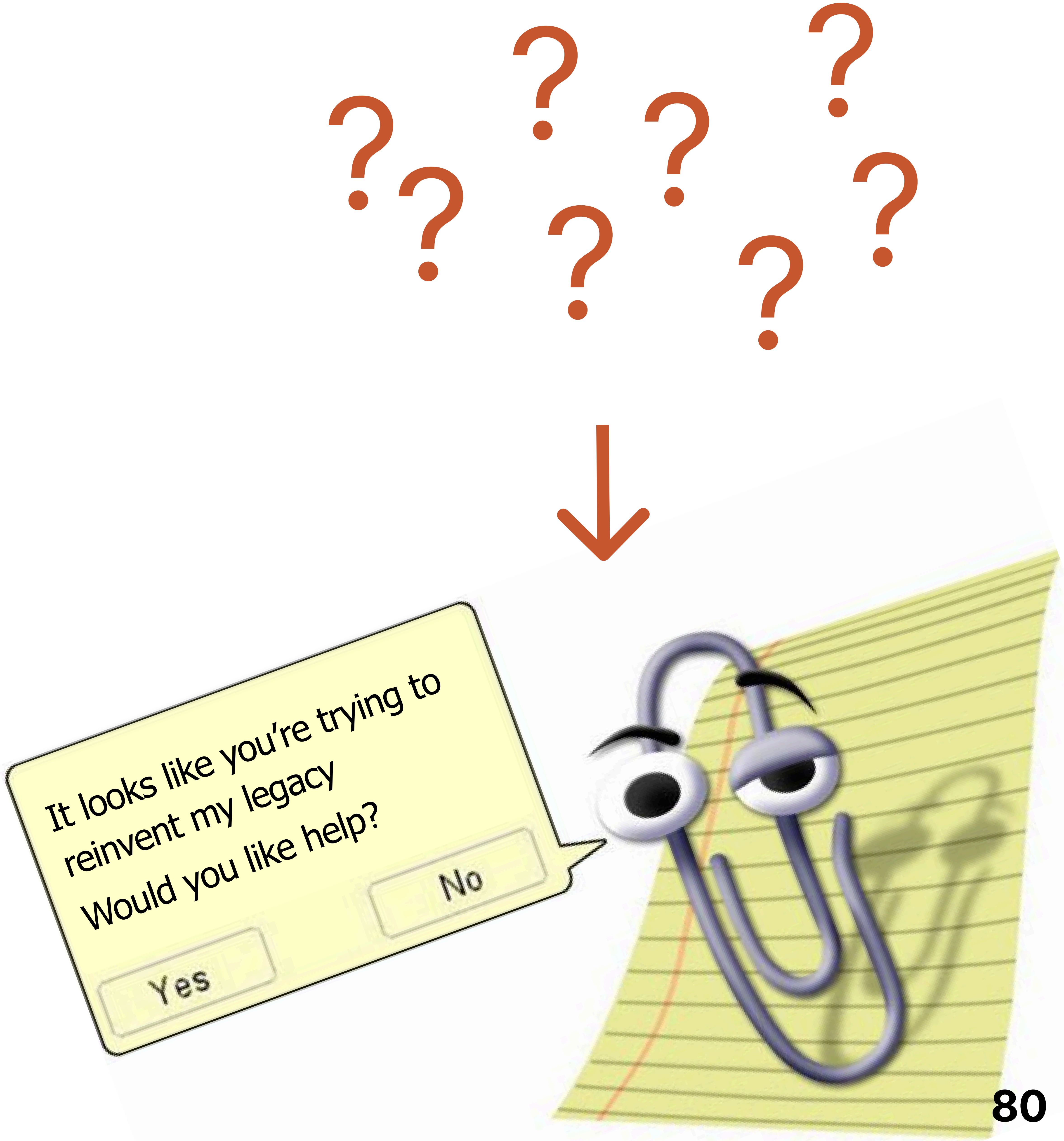
Mixed Initiative Interaction

Challenge: Knowing when an AI system should take initiative depends on the context

Utility from interruption v.s. no interruption depending on context

How do you get this utility?

If only we had a good model of user context...



Application

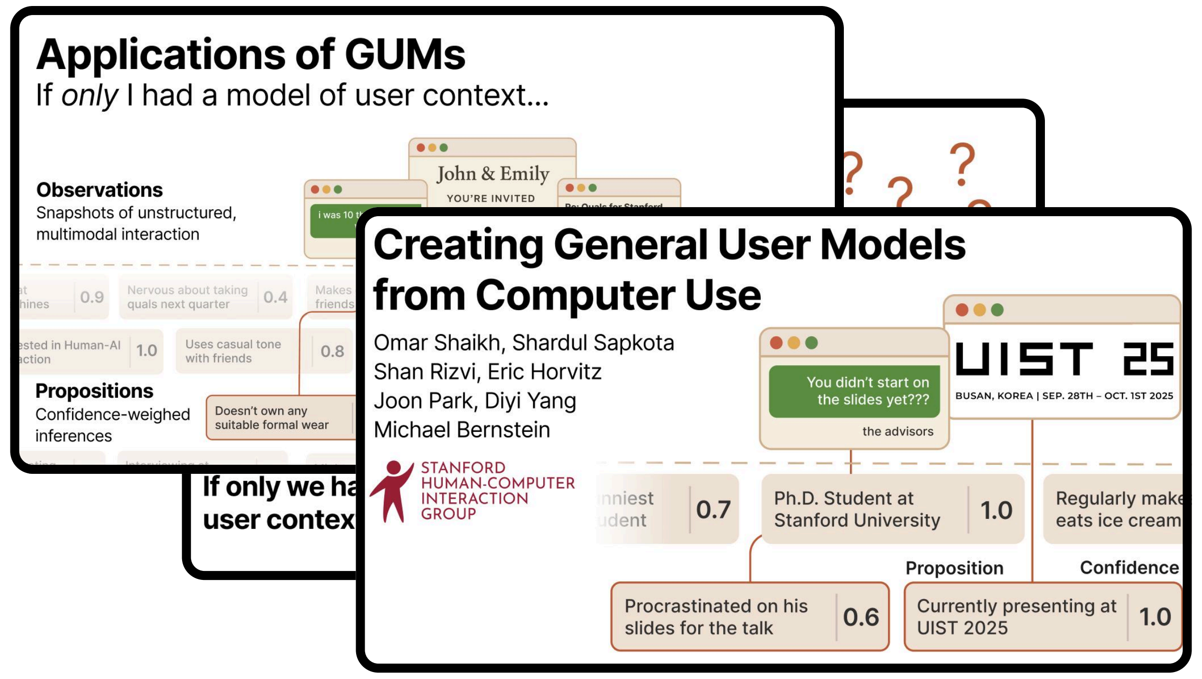
Mixed Initiative Interaction

Challenge: Knowing when an AI system should take initiative depends on the context

Utility from interruption v.s. no interruption depending on context

How do you get this utility?

If only we had a good model of user context...



Applications of GUMs

If *only* I had a model of user context...

Observations

Snapshots of unstructured, multimodal interaction

John & Emily

YOU'RE INVITED

Nervous about taking quals next quarter 0.4

0.9

Uses casual tone with friends 0.8

1.0

Propositions

Confidence-weighed inferences

Doesn't own any suitable formal wear

If only we ha

user contex

Creating General User Models from Computer Use

Omar Shaikh, Shardul Sapkota
Shan Rizvi, Eric Horvitz
Joon Park, Diyi Yang
Michael Bernstein

STANFORD HUMAN-COMPUTER INTERACTION GROUP

You didn't start on the slides yet???

the advisors

UIST 25

BUSAN, KOREA | SEP. 28TH – OCT. 1ST 2025

0.7

Ph.D. Student at Stanford University 1.0

Proposition

Confidence

Procrastinated on his slides for the talk 0.6

Currently presenting at UIST 2025 1.0

Retrieve

↓

Wants to reinvent Clippy's legacy | 1.0

Likes to have funny slides in their talks | 0.8

Rea
Mixed Initiative Interac. | 0.9

Application

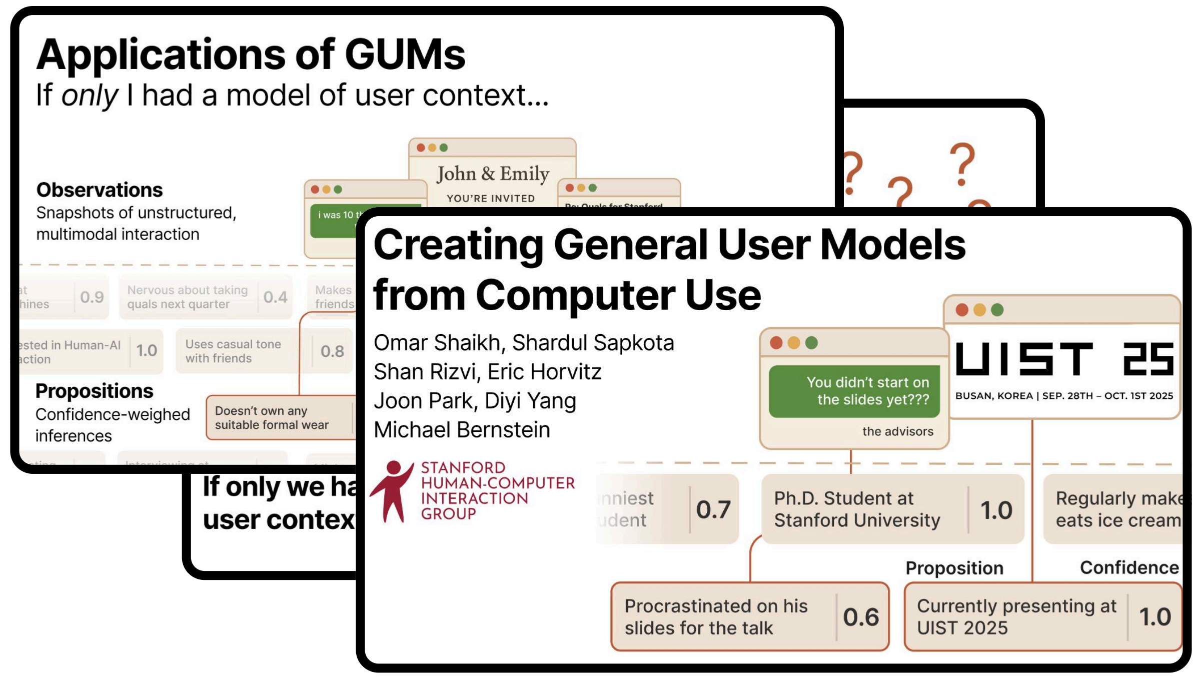
Mixed Initiative Interaction

Challenge: Knowing when an AI system should take initiative depends on the context

Utility from interruption v.s. no interruption depending on context

How do you get this utility?

If only we had a good model of user context...



Applications of GUMs
If *only* I had a model of user context...

Observations
Snapshots of unstructured, multimodal interaction

Propositions
Confidence-weighed inferences

Nervous about taking quals next quarter 0.4

Uses casual tone with friends 0.8

Doesn't own any suitable formal wear

Creating General User Models from Computer Use

Omar Shaikh, Shardul Sapkota
Shan Rizvi, Eric Horvitz
Joon Park, Diyi Yang
Michael Bernstein

STANFORD HUMAN-COMPUTER INTERACTION GROUP

You didn't start on the slides yet???
the advisors

UIST 25
BUSAN, KOREA | SEP. 28TH – OCT. 1ST 2025

Proposition	Confidence
Ph.D. Student at Stanford University	1.0
Procrastinated on his slides for the talk	0.6
Currently presenting at UIST 2025	1.0

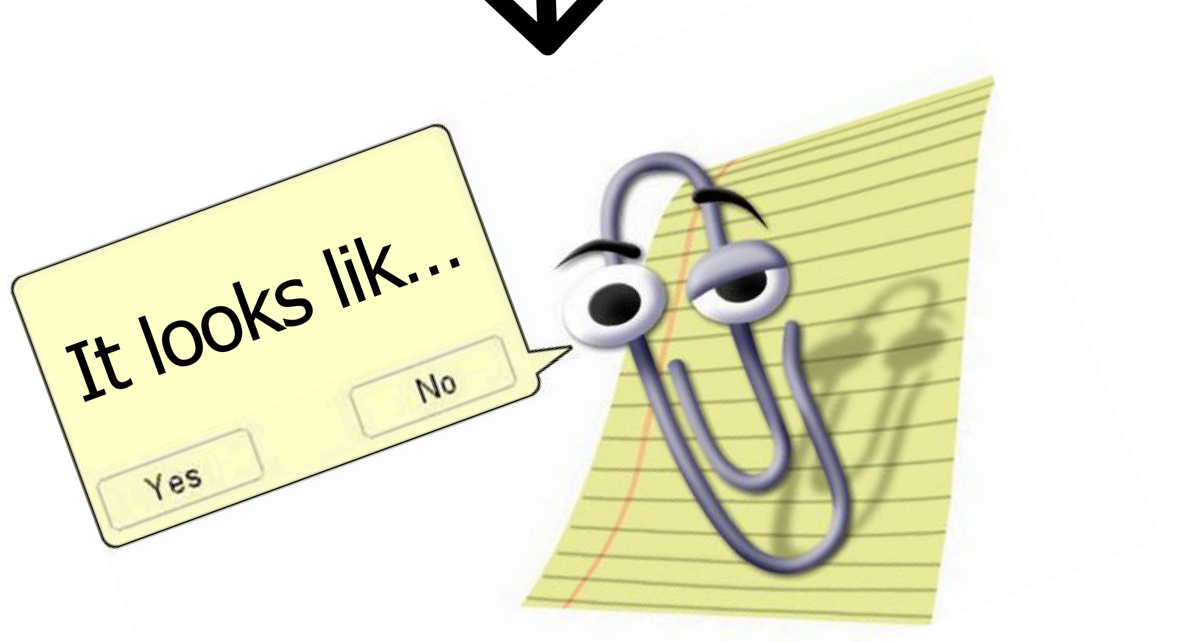
Retrieve

↓

Proposition	Confidence
Wants to reinvent Clippy's legacy	1.0
Likes to have funny slides in their talks	0.8
Rea… Mixed Initiative Interac.	0.9

Estimate E[U | GUM] with LLM

↓



Propositions

Confidence-weighed inferences

ines 0.9

Nervous about taking quals next quarter 0.4

Makes coffee for his friends at the lab 0.9

Nervous about taking quals next term 0.6

Interested in making more Chai at home 1.0

sted in Human-AI ction 1.0

Uses casual tone with friends 0.8

Makes and eats ice cream very regularly 0.9

Is a Ph.D. student at Stanford University 0.8

Pretentious about coffee options

Proposition	Confidence
Doesn’t own any suitable formal wear	1.0
Likely going to friend’s wedding in Chicago	0.8
Needs to budget as a Ph.D. student	0.6

rating friend 1.0

Giving a talk at UIST right now 0.9

Excited about the honorable mention 1.0

Struggling to use gradients in Figma 0.8

Excited to visit Berlin for the summer 1.0

ines 0.9

Nervous about taking quals next quarter 0.4

Makes coffee for his friends at the lab 0.9

Nervous about taking quals next term 0.6

Interested in making more Chai at home 1.0

sted in Human-AI ction 1.0

Uses casual tone with friends 0.8

Makes and eats ice cream very regularly 0.9

Is a Ph.D. student at Stanford University 0.8

Pretentious about coffee options

Propositions

Confidence-weighed inferences

Proposition	Confidence
Doesn’t own any suitable formal wear	1.0
Likely going to friend’s wedding in Chicago	0.8
Needs to budget as a Ph.D. student	0.6

rating friend 1.0

Giving a talk at UIST right now 0.9

Excited about the honorable mention 1.0

Struggling to use gradients in Figma 0.8

Excited to visit Berlin for the summer 1.0

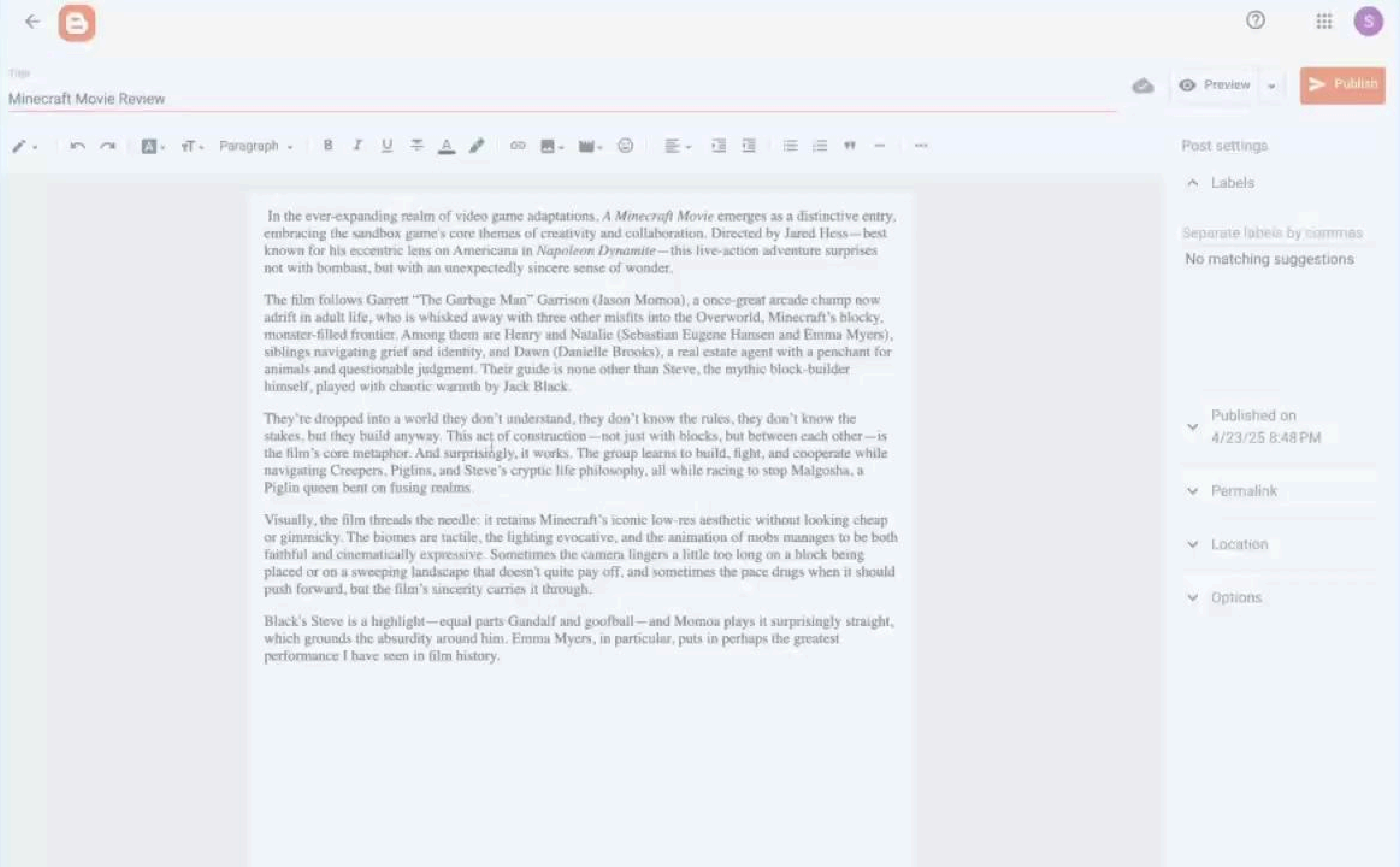
Suggestions

Proactive recommendations based on the GUM

Searching for suit rentals

I found a suit rental in Chicago for your friend’s wedding that likely fits your budget as a Ph.D. student [...]

Application: a proactive assistant that uses the GUM to generate suggestions



Minecraft Movie Review

In the ever-expanding realm of video game adaptations, *A Minecraft Movie* emerges as a distinctive entry, embracing the sandbox game's core themes of creativity and collaboration. Directed by Jared Hess—best known for his eccentric lens on Americana in *Napoleon Dynamite*—this live-action adventure surprises not with bombast, but with an unexpectedly sincere sense of wonder.

The film follows Garrett "The Garbage Man" Garrison (Jason Momoa), a once-great arcade champ now adrift in adult life, who is whisked away with three other misfits into the Overworld, Minecraft's blocky, monster-filled frontier. Among them are Henry and Natalie (Sebastian Eugene Hansen and Emma Myers), siblings navigating grief and identity, and Dawn (Danielle Brooks), a real estate agent with a penchant for animals and questionable judgment. Their guide is none other than Steve, the mythic block-builder himself, played with chaotic warmth by Jack Black.

They're dropped into a world they don't understand, they don't know the rules, they don't know the stakes, but they build anyway. This act of construction—not just with blocks, but between each other—is the film's core metaphor. And surprisingly, it works. The group learns to build, fight, and cooperate while navigating Creepers, Piglins, and Steve's cryptic life philosophy, all while racing to stop Malgosha, a Piglin queen bent on fusing realms.

Visually, the film threads the needle: it retains Minecraft's iconic low-res aesthetic without looking cheap or gimmicky. The biomes are tactile, the lighting evocative, and the animation of mobs manages to be both faithful and cinematically expressive. Sometimes the camera lingers a little too long on a block being placed or on a sweeping landscape that doesn't quite pay off, and sometimes the pace drugs when it should push forward, but the film's sincerity carries it through.

Black's Steve is a highlight—equal parts Gandalf and goofball—and Momoa plays it surprisingly straight, which grounds the absurdity around him. Emma Myers, in particular, puts in perhaps the greatest performance I have seen in film history.

navigating Creepers, Piglins, and Steve's cryptic life philosoph
Piglin queen bent on fusing realms.

reads the needle: it retains Minecraft's icon
biomes are tactile, the lighting evocative, an
tically expressive. Sometimes the camera li
ping landscape that doesn't quite pay off, a
he film's sincerity carries it through.

ighlight—equal parts Gandalf and goofball
absurdity around him. Emma Myers, in part
seen in film history.

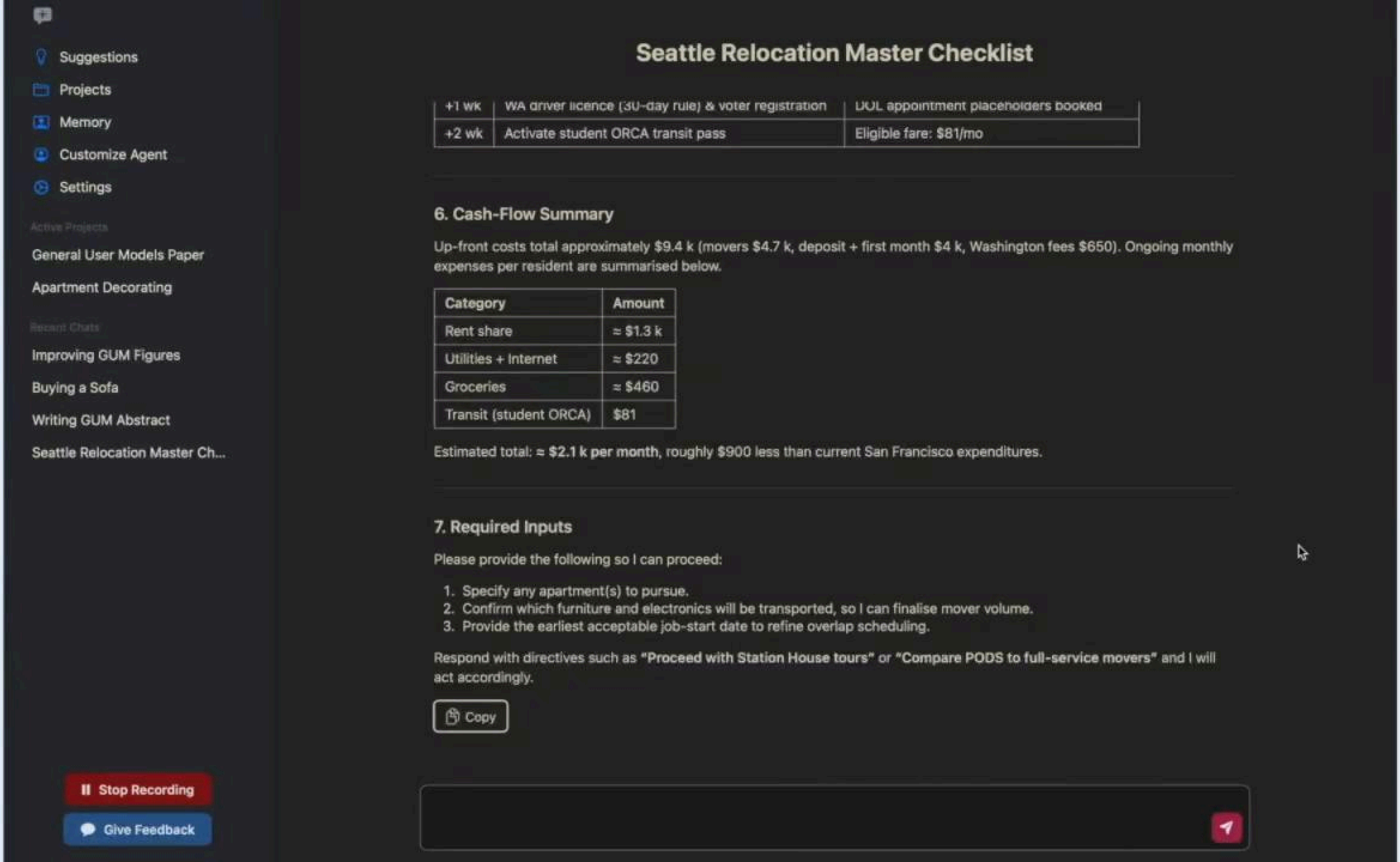
Now

Seattle Relocation Master Checklist ✓

I've assembled a comprehensive, step-by-step plan for your move from San Francisco to Seattle: neighborhood fit analysis, cost-of-living delta, vetted apartment short-list, weather prep, and a timeline covering movers, DMV tasks, voter registration, and utilities—ready for you to execute or adjust.

RAM-Sufficient Laptop Upgrade Guide ✓

Your current MacBook is bottlenecked by memory. I



Seattle Relocation Master Checklist

+1 wk	WA driver licence (30-day rule) & voter registration	DOL appointment placeholders booked
+2 wk	Activate student ORCA transit pass	Eligible fare: $81/mo

6. Cash-Flow Summary

Up-front costs total approximately $9.4 k (movers $4.7 k, deposit + first month $4 k, Washington fees $650). Ongoing monthly expenses per resident are summarised below.

Category	Amount
Rent share	≈ $1.3 k
Utilities + Internet	≈ $220
Groceries	≈ $460
Transit (student ORCA)	$81

Estimated total: **≈ $2.1 k per month**, roughly $900 less than current San Francisco expenditures.

7. Required Inputs

Please provide the following so I can proceed:

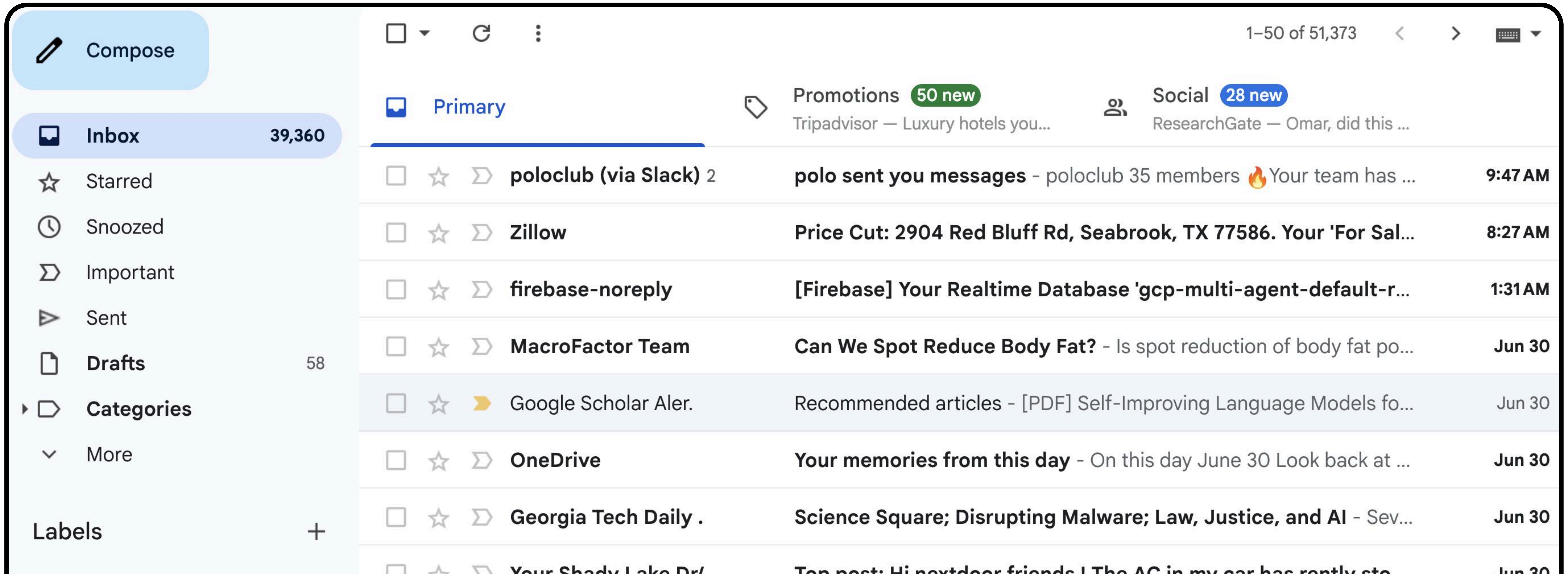
1. Specify any apartment(s) to pursue.
2. Confirm which furniture and electronics will be transported, so I can finalise mover volume.
3. Provide the earliest acceptable job-start date to refine overlap scheduling.

Respond with directives such as **"Proceed with Station House tours"** or **"Compare PODS to full-service movers"** and I will act accordingly.

Evaluation

GUMs are *accurate*

Evaluated on 3,600 emails
from 18 participants



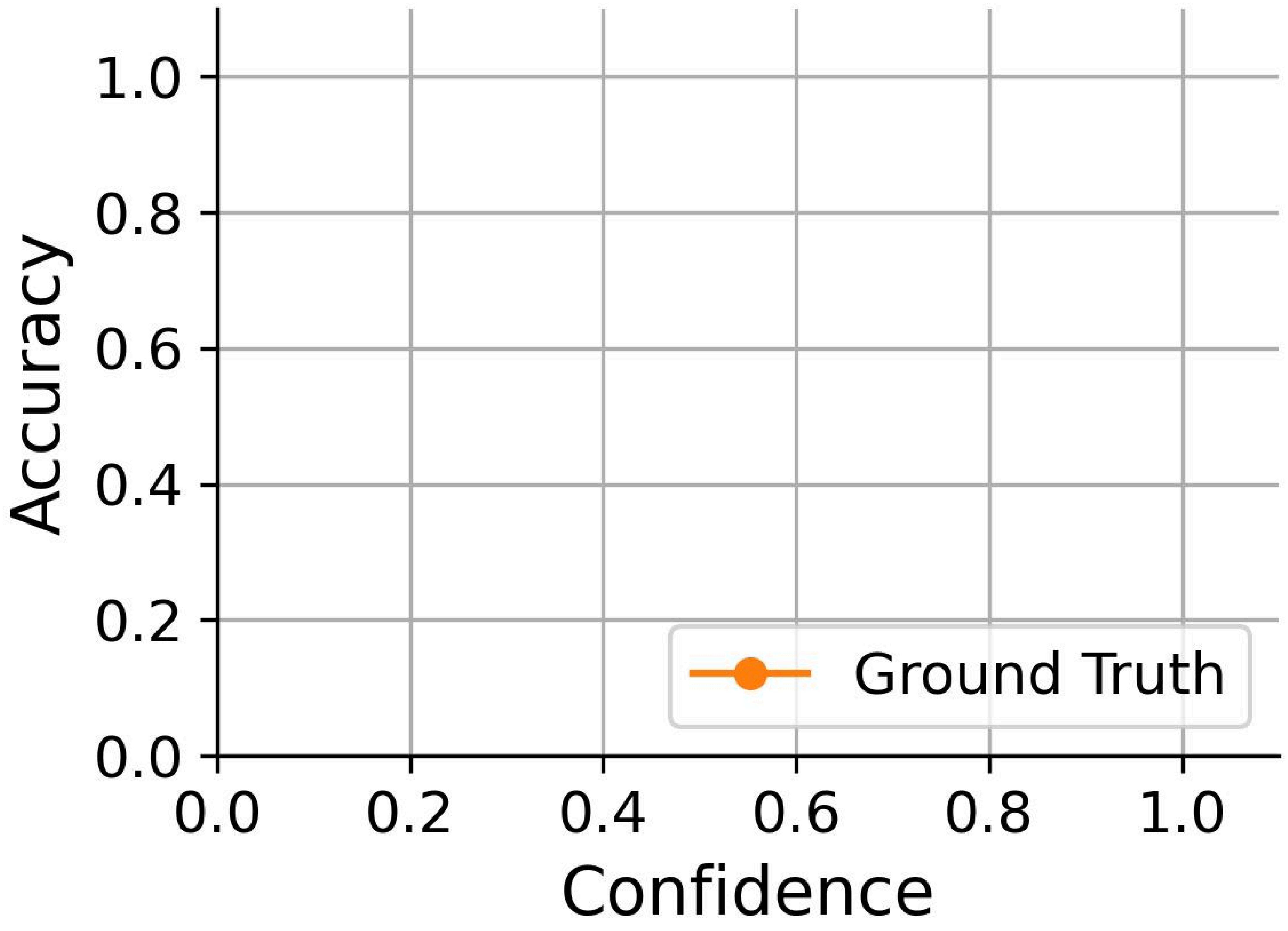
poloclub (via Slack) 2	**polo sent you messages** - poloclub 35 members 🔥Your team has ...	**9:47 AM**
Zillow	**Price Cut: 2904 Red Bluff Rd, Seabrook, TX 77586. Your 'For Sal...**	**8:27 AM**
firebase-noreply	**[Firebase] Your Realtime Database 'gcp-multi-agent-default-r...**	**1:31 AM**
MacroFactor Team	**Can We Spot Reduce Body Fat?** - Is spot reduction of body fat po...	**Jun 30**
Google Scholar Aler.	Recommended articles - [PDF] Self-Improving Language Models fo...	Jun 30
OneDrive	**Your memories from this day** - On this day June 30 Look back at ...	**Jun 30**
Georgia Tech Daily .	**Science Square; Disrupting Malware; Law, Justice, and AI** - Sev...	**Jun 30**

Evaluation

GUMs are *accurate*

Evaluated on 3,600 emails from 18 participants

Confident inferences (> 0.8) are accurate (88.3%)

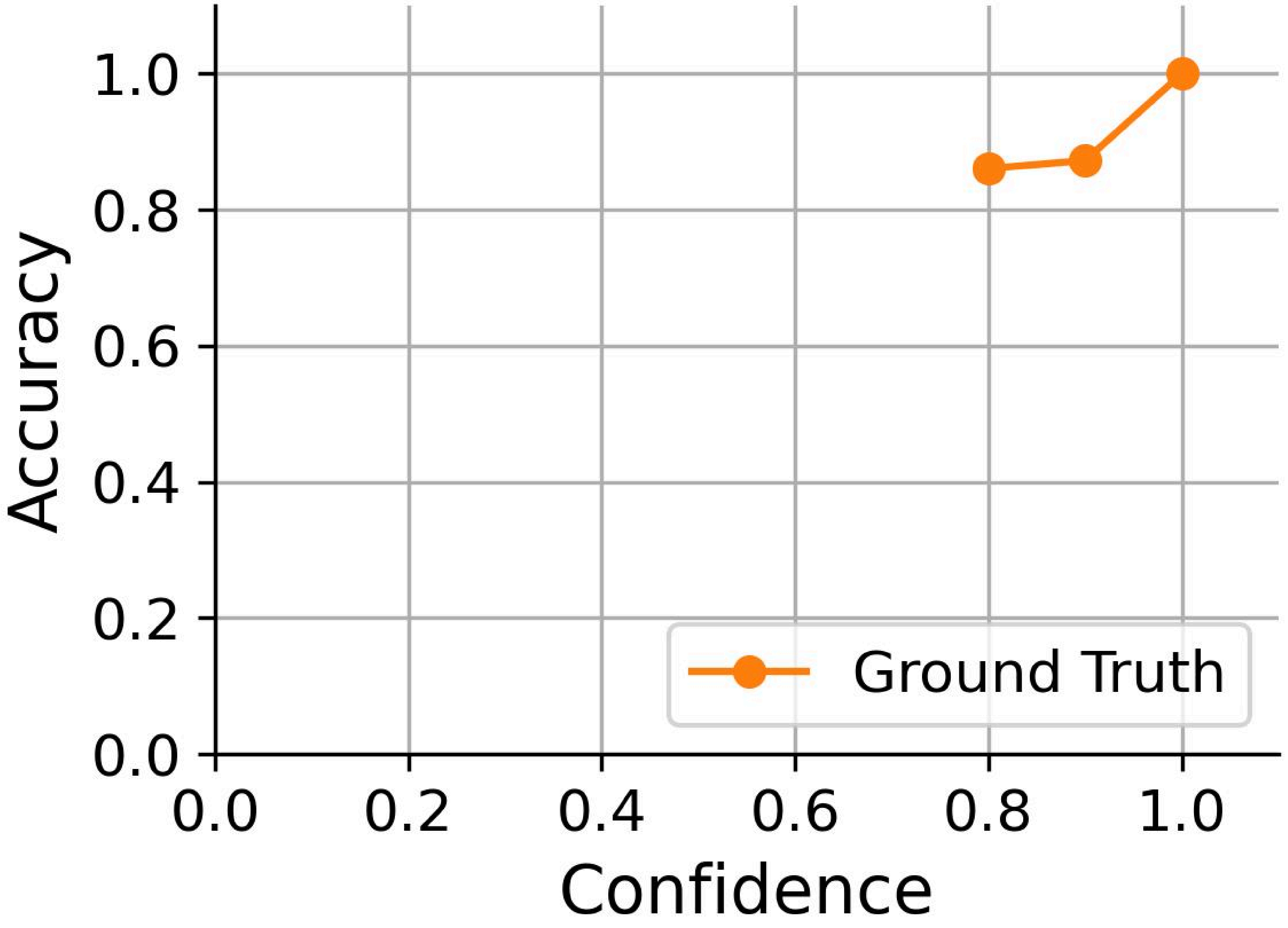


Evaluation

GUMs are *accurate*

Evaluated on 3,600 emails from 18 participants

Confident inferences (> 0.8) are accurate (88.3%)

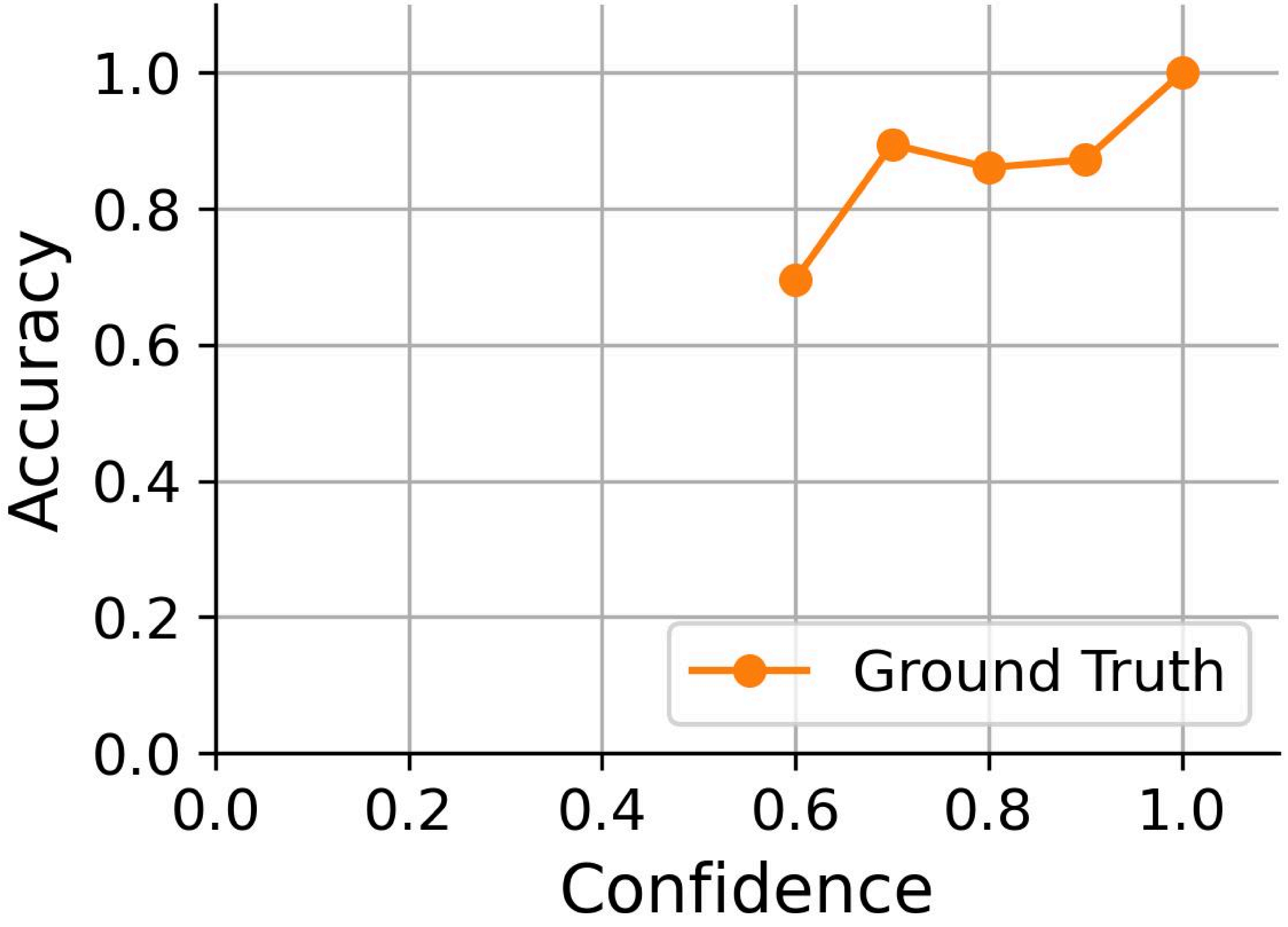


Evaluation

GUMs are *accurate*

Evaluated on 3,600 emails from 18 participants

Confident inferences (> 0.8) are accurate (88.3%)

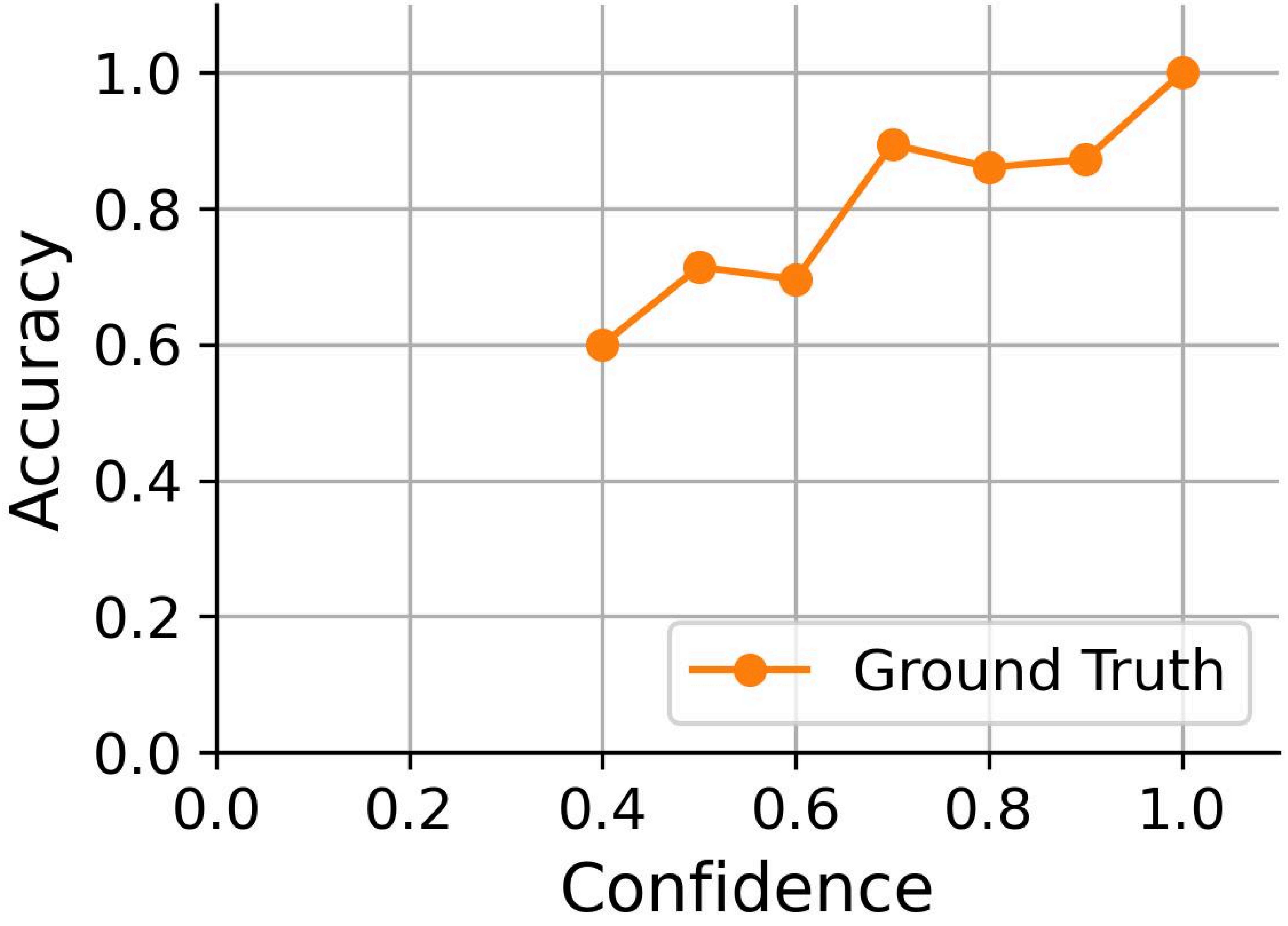


Evaluation

GUMs are *accurate*

Evaluated on 3,600 emails from 18 participants

Confident inferences (> 0.8) are accurate (88.3%)

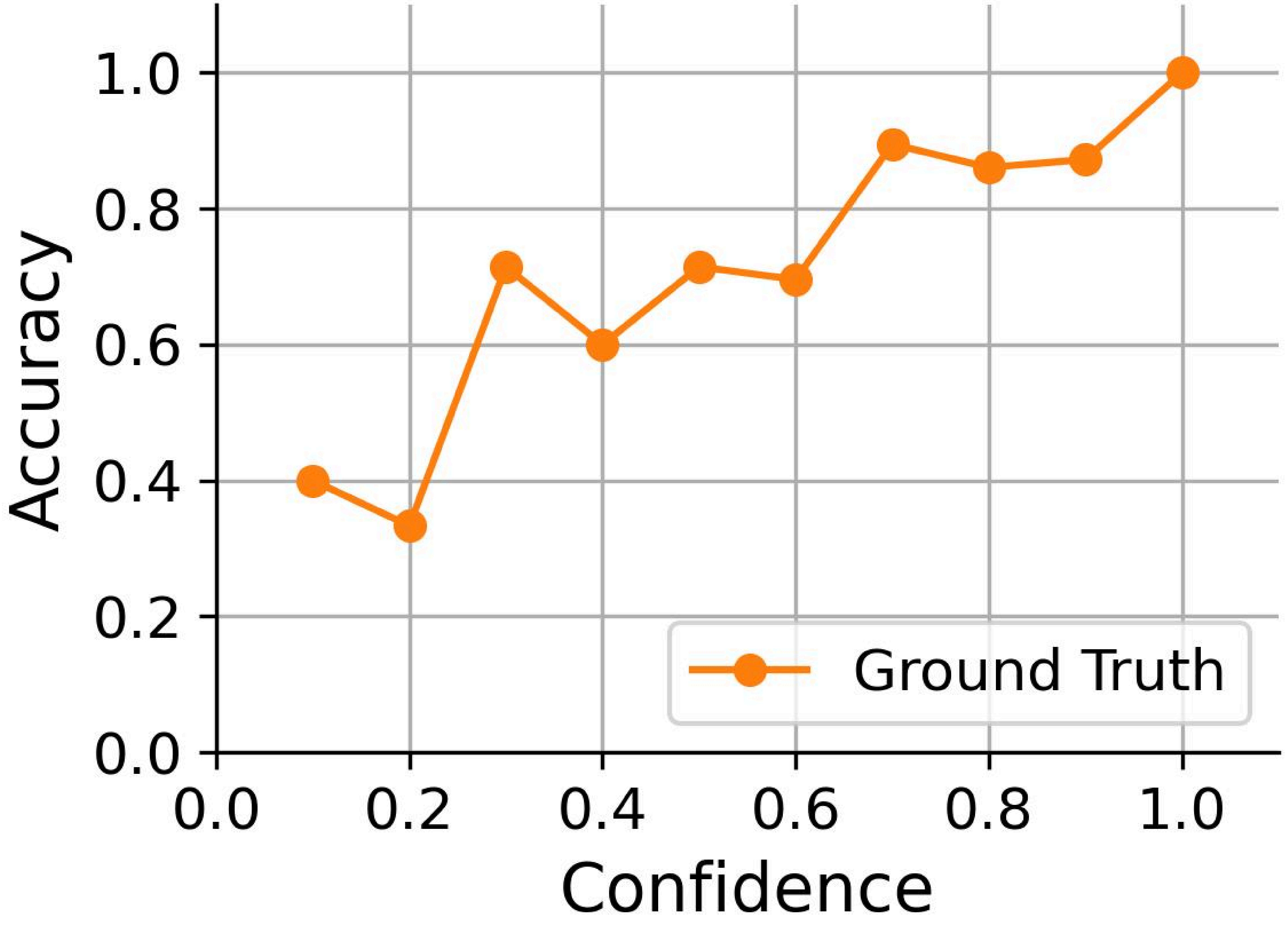


Evaluation

GUMs are *accurate*

Evaluated on 3,600 emails from 18 participants

Confident inferences (> 0.8) are accurate (88.3%)

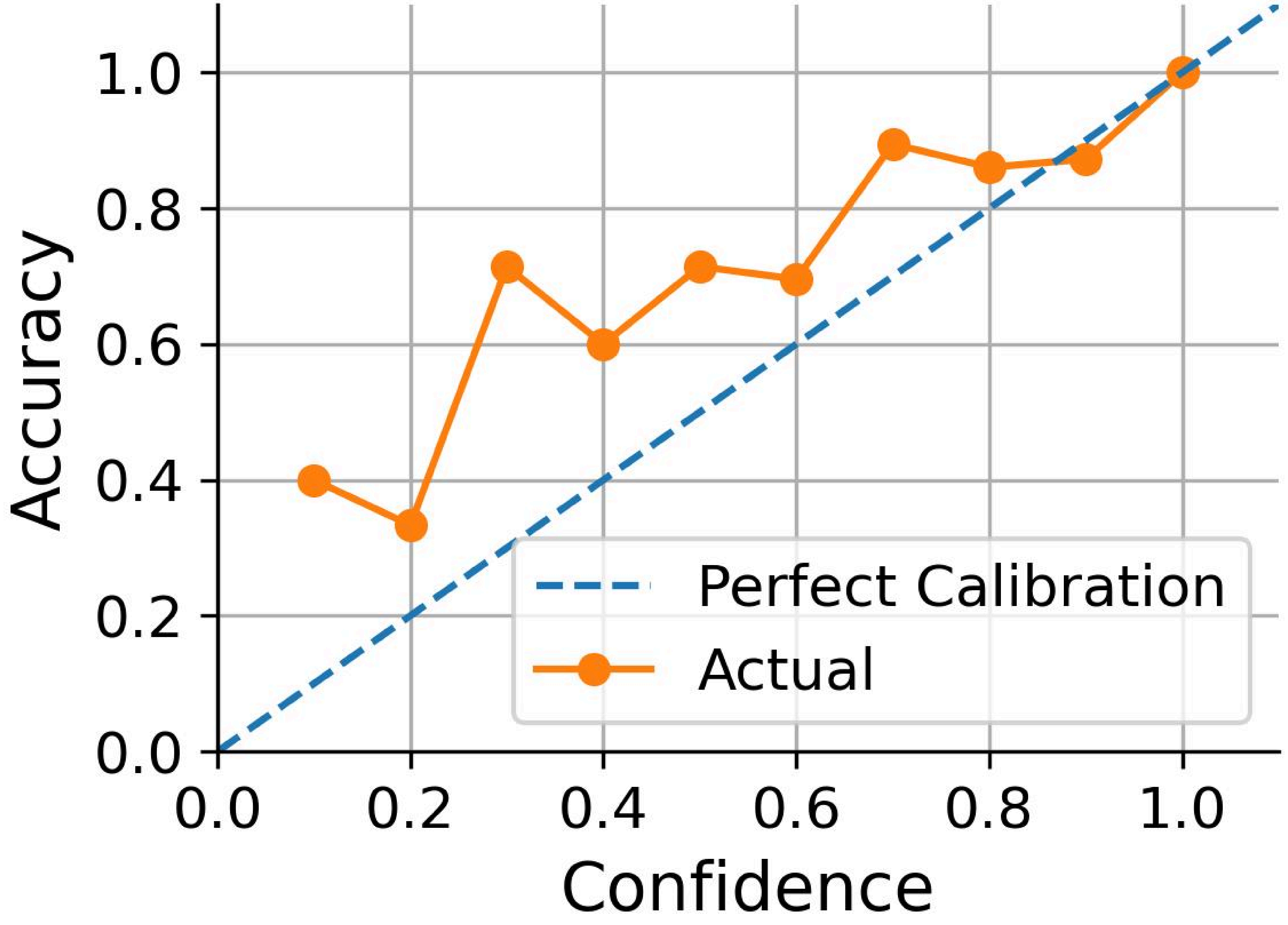


Evaluation

GUMs are *accurate* ***and calibrated***

Evaluated on 3,600 emails from 18 participants

Confident inferences (> 0.8) are accurate (88.3%)

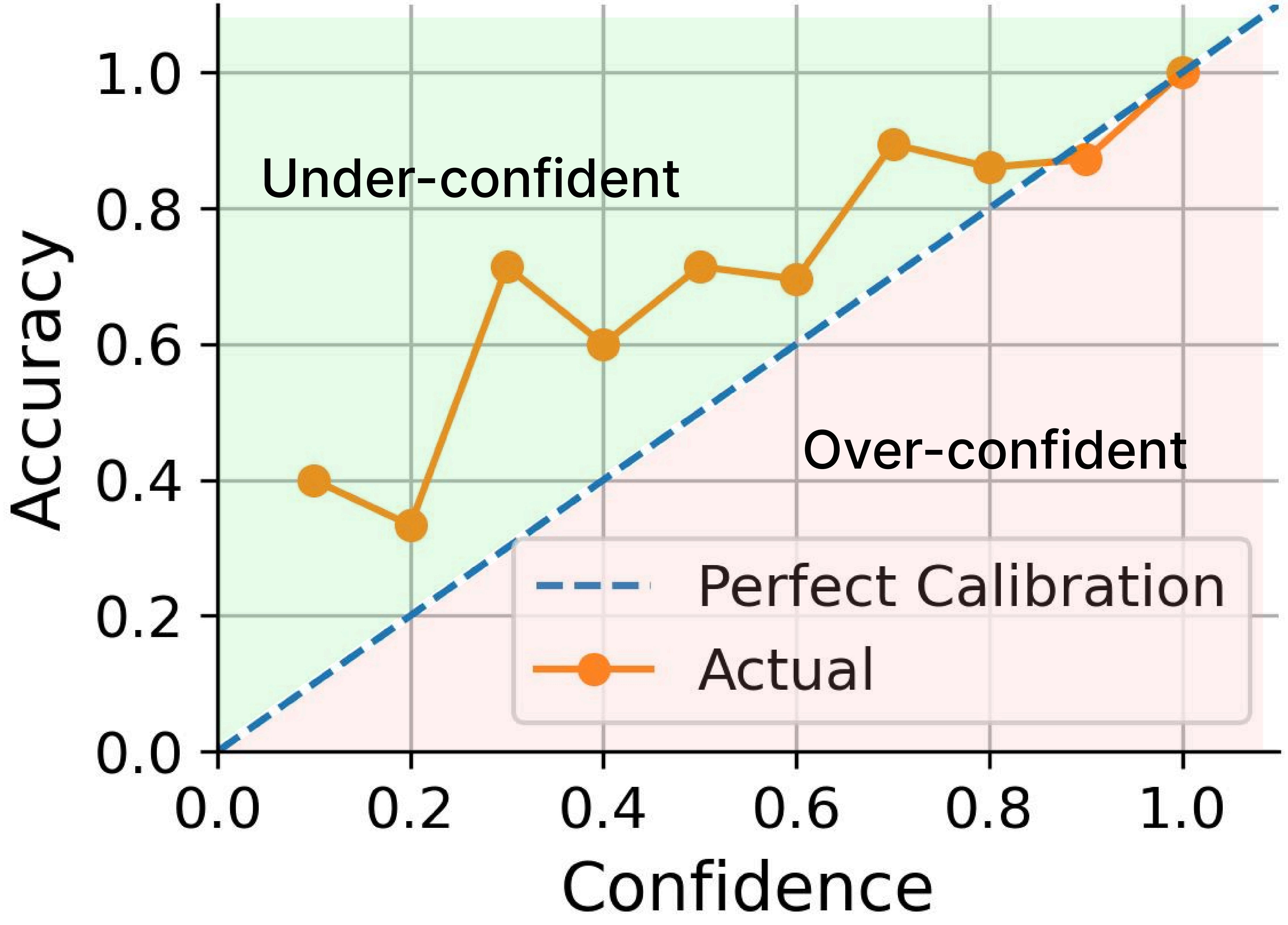


Evaluation

GUMs are *accurate* ***and calibrated***

Evaluated on 3,600 emails from 18 participants

Confident inferences (> 0.8) are accurate (88.3%)



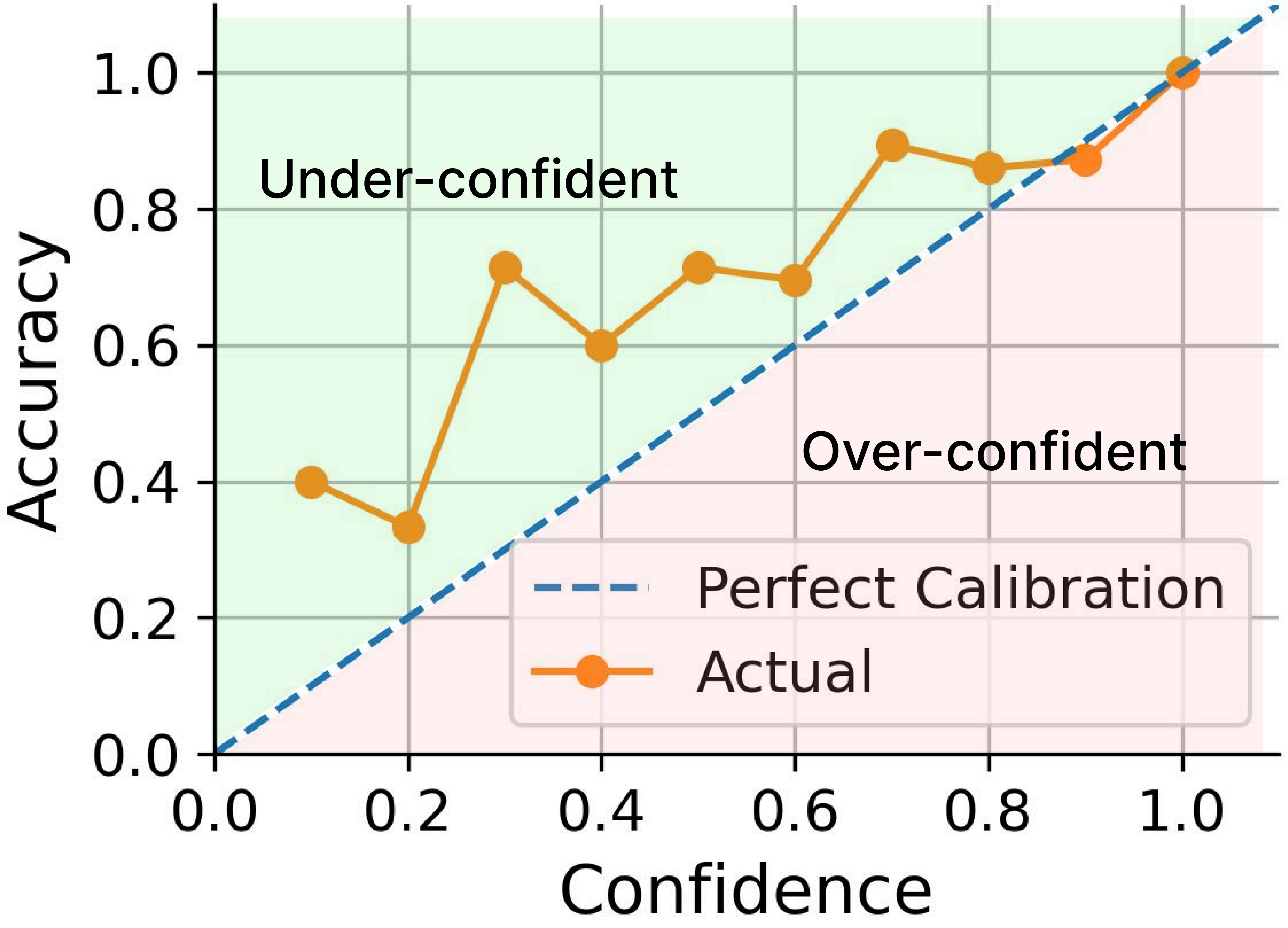
Evaluation

GUMs are *accurate* ***and calibrated***

Evaluated on 3,600 emails from 18 participants

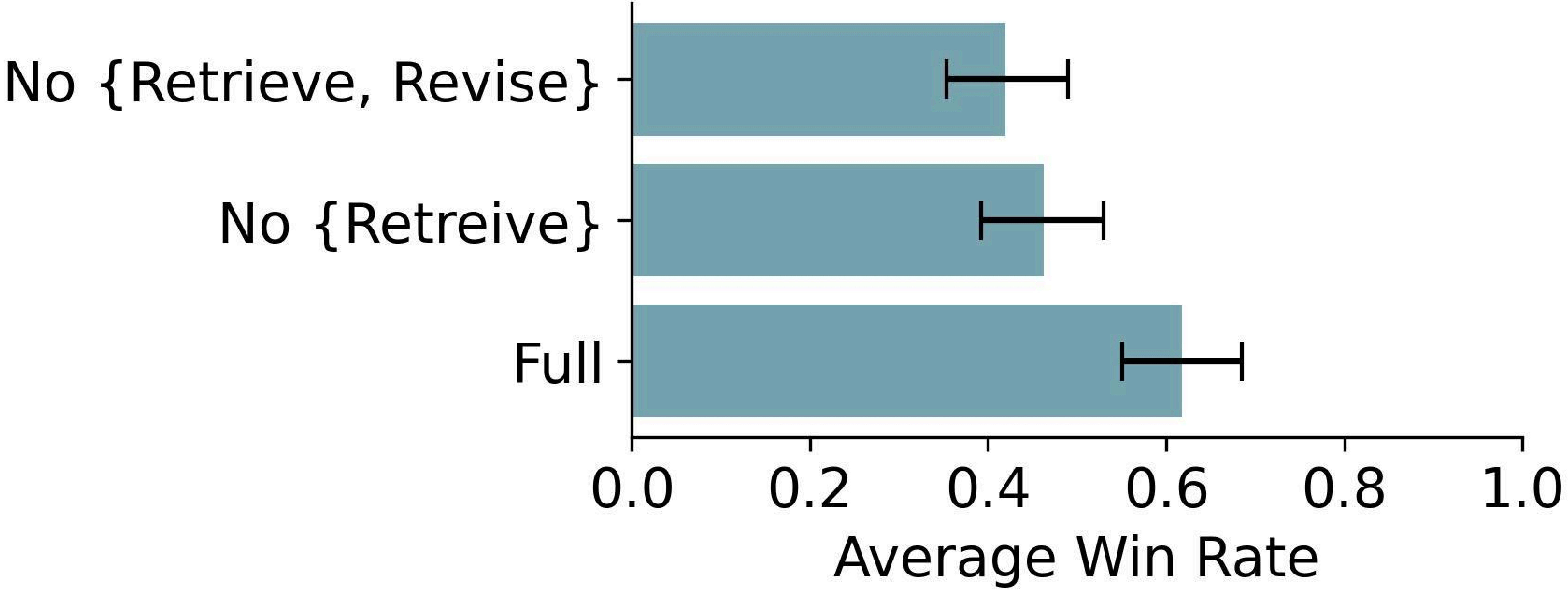
Confident inferences (> 0.8) are accurate (88.3%)

Calibrated (0.28 Brier) on *all computer use for 5 days across* 5 participants



Evaluation

Removing modules reduces performance



Evaluation

Good suggestions draw from the whole GUM

Participants were excited about suggestions using large amounts of context

> “It knew who my roommate was; what our budget was; where we were moving; that I was worried about this move. It worked backward from our move-in date, planned a schedule, and identified moving services.
>
> And half of my conversation with my roommate wasn’t in English.”
> - Participant 1

Evaluation

Good suggestions draw from the whole GUM

Participants were excited about suggestions using large amounts of context

7 pt. Likert
`45.69%` at least useful (>=5)
`10.34%` perfect (7)

“I did not even tell this system anything, but it identified that I have this habit [of creating research outlines in LaTeX].

And it created this entire outline—in LaTeX—of how I could write my paper in the context of this new framework I was checking out. I was like, wow”
- Participant 3

Evaluation

But they’re not perfect….

Participants were excited about suggestions using large amounts of context

7 pt. Likert
45.69% at least useful (>=5)
10.34% perfect (7)
13.79% poor (<= 2)

“This system gave me a nudge. It looked into the podcast and said it drew connections to my work [...]

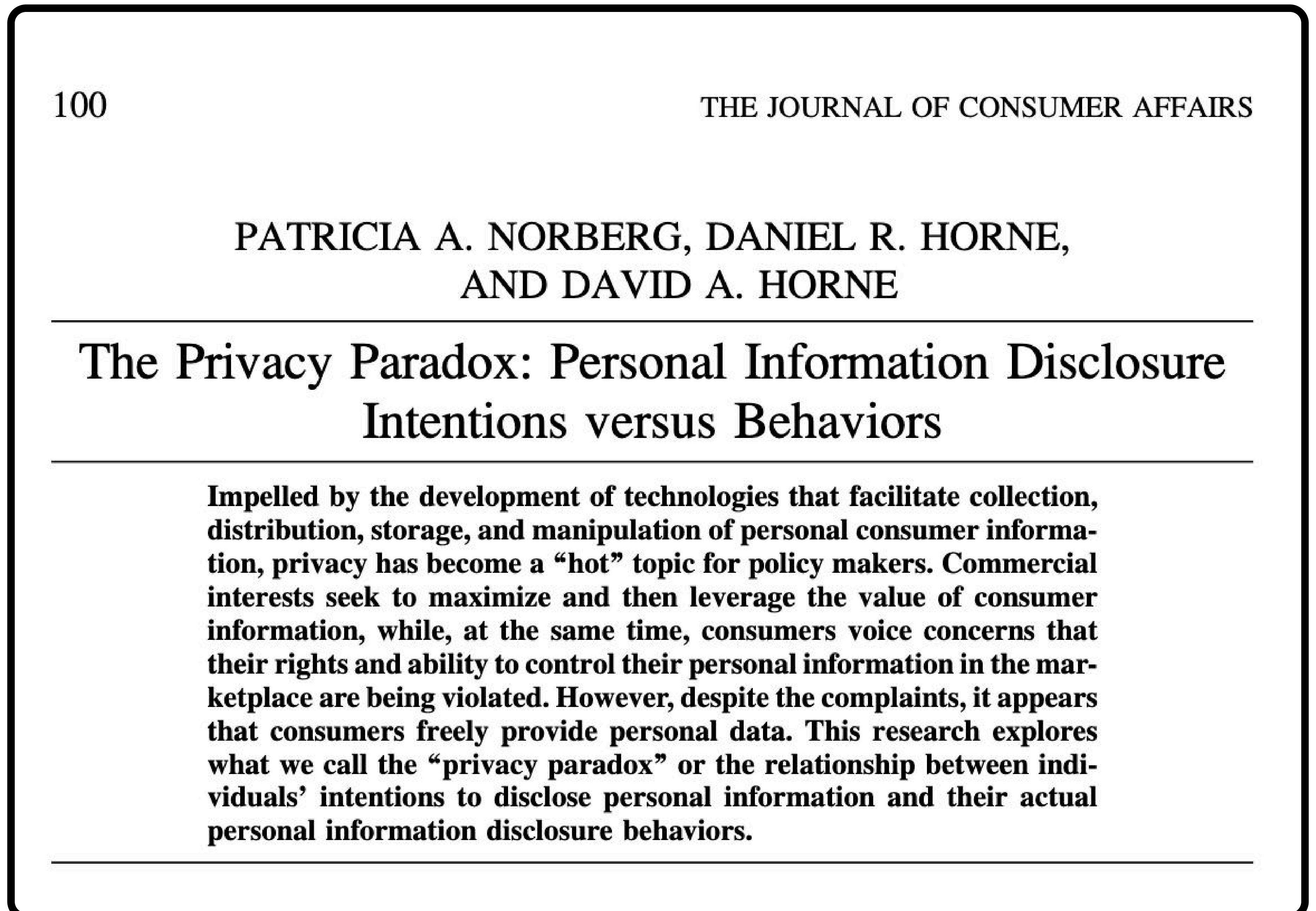
The connections were total garbage. Still, it got me to finally listen, and I ended up totally locked in, working on it for hours after.”
- Participant 2

Discussion

Paradox

The Privacy Paradox
People often say they value their privacy, yet willingly share personal information online.

Example:
A user claims to care about data privacy but regularly posts personal details on social media and accepts app permissions without reading them.



100 THE JOURNAL OF CONSUMER AFFAIRS

PATRICIA A. NORBERG, DANIEL R. HORNE,
AND DAVID A. HORNE

The Privacy Paradox: Personal Information Disclosure Intentions versus Behaviors

Impelled by the development of technologies that facilitate collection, distribution, storage, and manipulation of personal consumer information, privacy has become a "hot" topic for policy makers. Commercial interests seek to maximize and then leverage the value of consumer information, while, at the same time, consumers voice concerns that their rights and ability to control their personal information in the marketplace are being violated. However, despite the complaints, it appears that consumers freely provide personal data. This research explores what we call the "privacy paradox" or the relationship between individuals' intentions to disclose personal information and their actual personal information disclosure behaviors.

Discussion

Paradox

The Privacy Paradox
People often say they value their privacy, yet willingly share personal information online.

Example:
A user claims to care about data privacy but regularly posts personal details on social media and accepts app permissions without reading them.

"It was intimidating at first. And I was also like shit like, do I need to be careful what I say? And then **after a day I was like, whatever F it..**"
- Participant 4

Discussion

Paradox

The Privacy Paradox
People often say they value their privacy, yet willingly share personal information online.

Example:
A user claims to care about data privacy but regularly posts personal details on social media and accepts app permissions without reading them.

"It was intimidating at first. And I was also like shit like, do I need to be careful what I say? And then **after a day I was like, whatever F it..**"
- Participant 4

"There were some really great things that came out of there **that I wouldn't have thought about**. So how do I amplify that? I would integrate **some more personal work** that I do."
- Participant 3

Discussion

Failure Modes

Hallucinations

Discussion

Failure Modes

Hallucinations

GUM can be prompt injected with spam

“[The proposition was] based on a phishing email but it wasn’t accurate at all, and so the model thought it was me and my interest rather than just a phishing attempt”
- Participant 4

Discussion

Failure Modes

Hallucinations

GUM can be prompt injected with spam

Even higher precision suggestions

Execution failure

"So my advisor's been asking me to listen to this Claude Steele podcast. GUMBO gave me a nudge. It looked into the podcast and said it drew connections to my work. The outline was decent, **but the connections were total garbage.**

Discussion

Failure Modes

Hallucinations

GUM can be prompt injected with spam

Even higher precision suggestions

Execution failure

"So my advisor's been asking me to listen to this Claude Steele podcast. GUMBO gave me a nudge. It looked into the podcast and said it drew connections to my work. The outline was decent, **but the connections were total garbage.**

Still, it got me to finally listen, and I ended up totally locked in, working on it for hours after. **I almost prefer this, because it didn't take away any cognitive burden."**
-Participant 2

Discussion

Failure Modes

Hallucinations

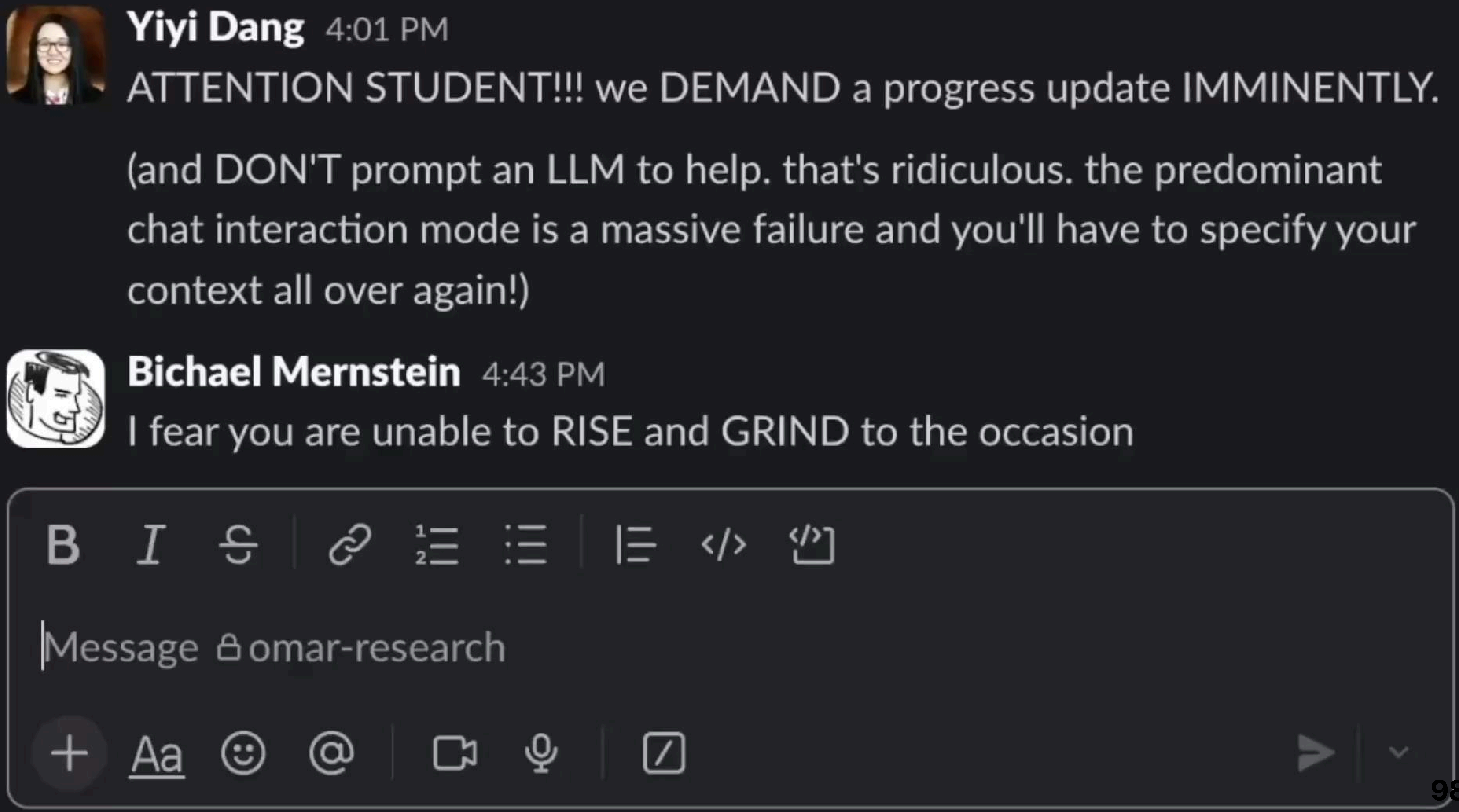
GUM can be prompt injected with spam

Even higher precision suggestions

Execution failure

Human agency

“Maybe I wasn’t familiar with PowerPoint, but if it asked me [before it made that assumption] and then used that to give me a suggestion I think I would’ve been happier with it”
- Participant 4



Yiyi Dang 4:01 PM

ATTENTION STUDENT!!! we DEMAND a progress update IMMINENTLY.

(and DON'T prompt an LLM to help. that's ridiculous. the predominant chat interaction mode is a massive failure and you'll have to specify your context all over again!)

Bichael Mernstein 4:43 PM

I fear you are unable to RISE and GRIND to the occasion

Creating General User Models

generalusermodels.github.io

The Real GUM v.s. the Ideal GUM

Do you really want your GUM to reflect your implicit preferences?

Privacy and Control

Local models, awareness, and what are users ready to see about *themselves*?

You prompt the model

and it asks clarification questions / uses RAG

with your whole life in context

and knows you well enough to do the right thing at the right time

Human-AI Grounding

Omar
Shaikh

There are certainly capabilities barriers… but I think a lot of today's errors are due to models being expected to “read our minds” instead of ground with us

STANFORD
HUMAN-COMPUTER
INTERACTION
GROUP

Stanford | NLP