CS329T Trustworthy Machine Learning:
Large Language Models & Applications

Lecture 3 Fall 2023

Sample application area: Education

John Mitchell (Stanford CS)
Isabelle Hau (Stanford GSE)
Josh Weiss (Stanford GSE)

Today's panelists

Isabelle Hau

John Mitchell

Josh Weiss

Trustworthiness dimensions

- Grounding - every assertion has authoritative basis
- Consistency - semantically equivalent queries treated similarly
- Confidence - acknowledge uncertainty accurately
- Interpretability - be able to show how response was generated
- Alignment - not harmful, toxic, biased, dishonest, unreliable
  - Respect privacy
  - Behave fairly and mitigate bias
- Resist adversarial manipulation
  - Malicious input should not subvert desirable properties

New York City Schools Ban ChatGPT Amid Cheating Worries

-CNET

Alarmed by A.I. Chatbots, Universities Start Revamping How They Teach

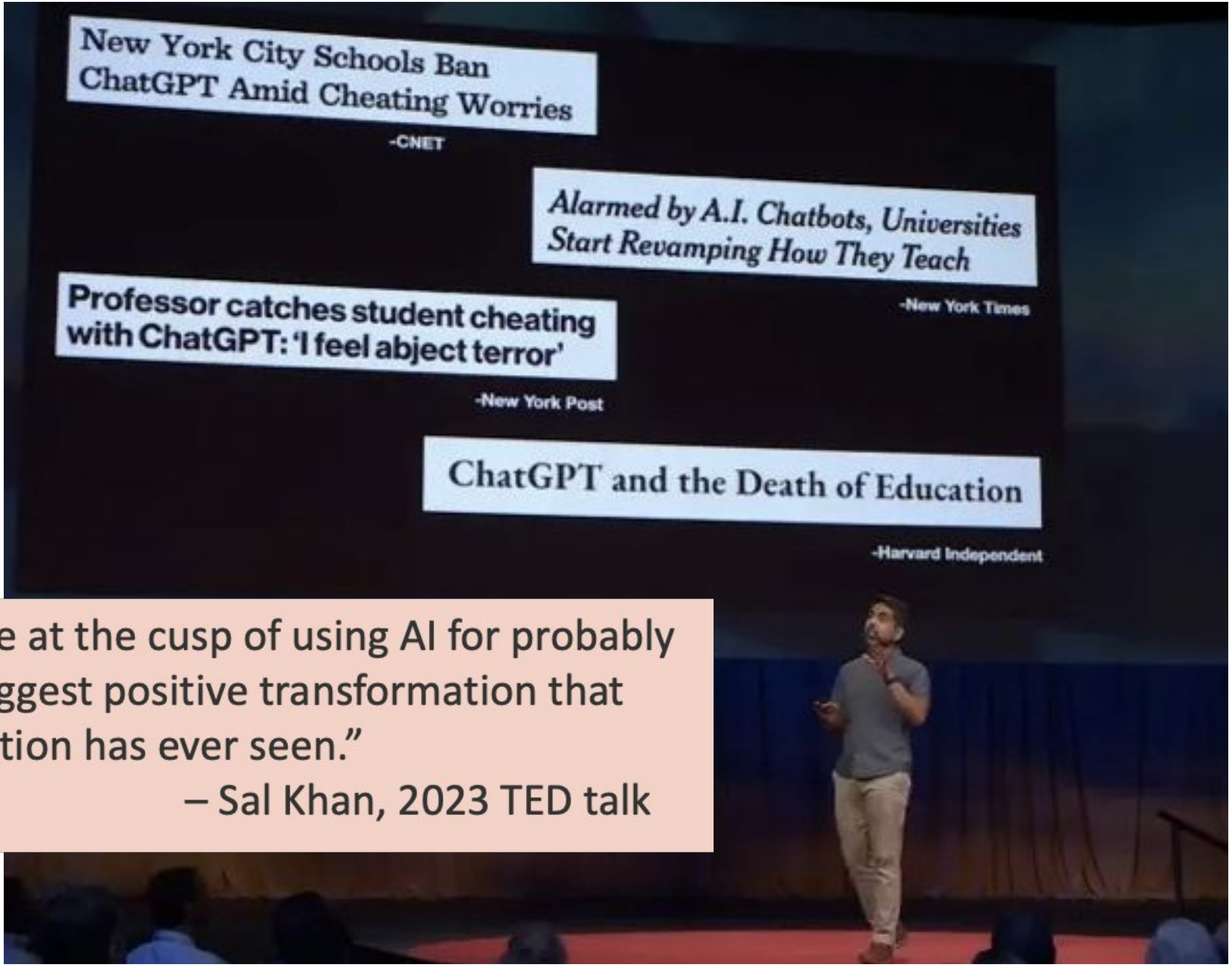
-New York Times

Professor catches student cheating with ChatGPT: 'I feel abject terror'

-New York Post

ChatGPT and the Death of Education

-Harvard Independent



New York City Schools Ban ChatGPT Amid Cheating Worries

-CNET

Alarmed by A.I. Chatbots, Universities Start Revamping How They Teach

-New York Times

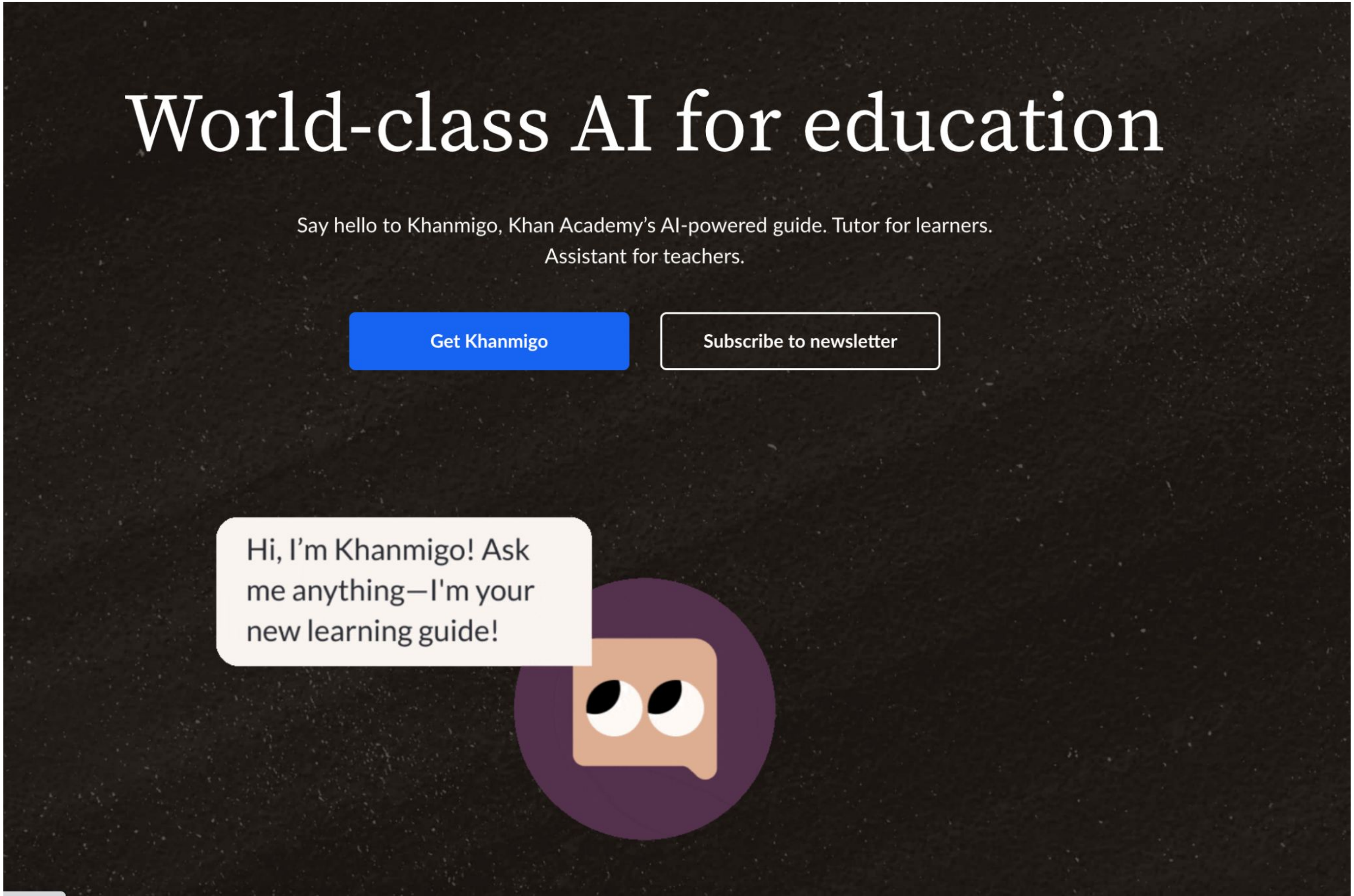
Professor catches student cheating with ChatGPT: 'I feel abject terror'

-New York Post

ChatGPT and the Death of Education

-Harvard Independent

> "We're at the cusp of using AI for probably the biggest positive transformation that education has ever seen."
>
> – Sal Khan, 2023 TED talk



World-class AI for education

Say hello to Khanmigo, Khan Academy's AI-powered guide. Tutor for learners. Assistant for teachers.

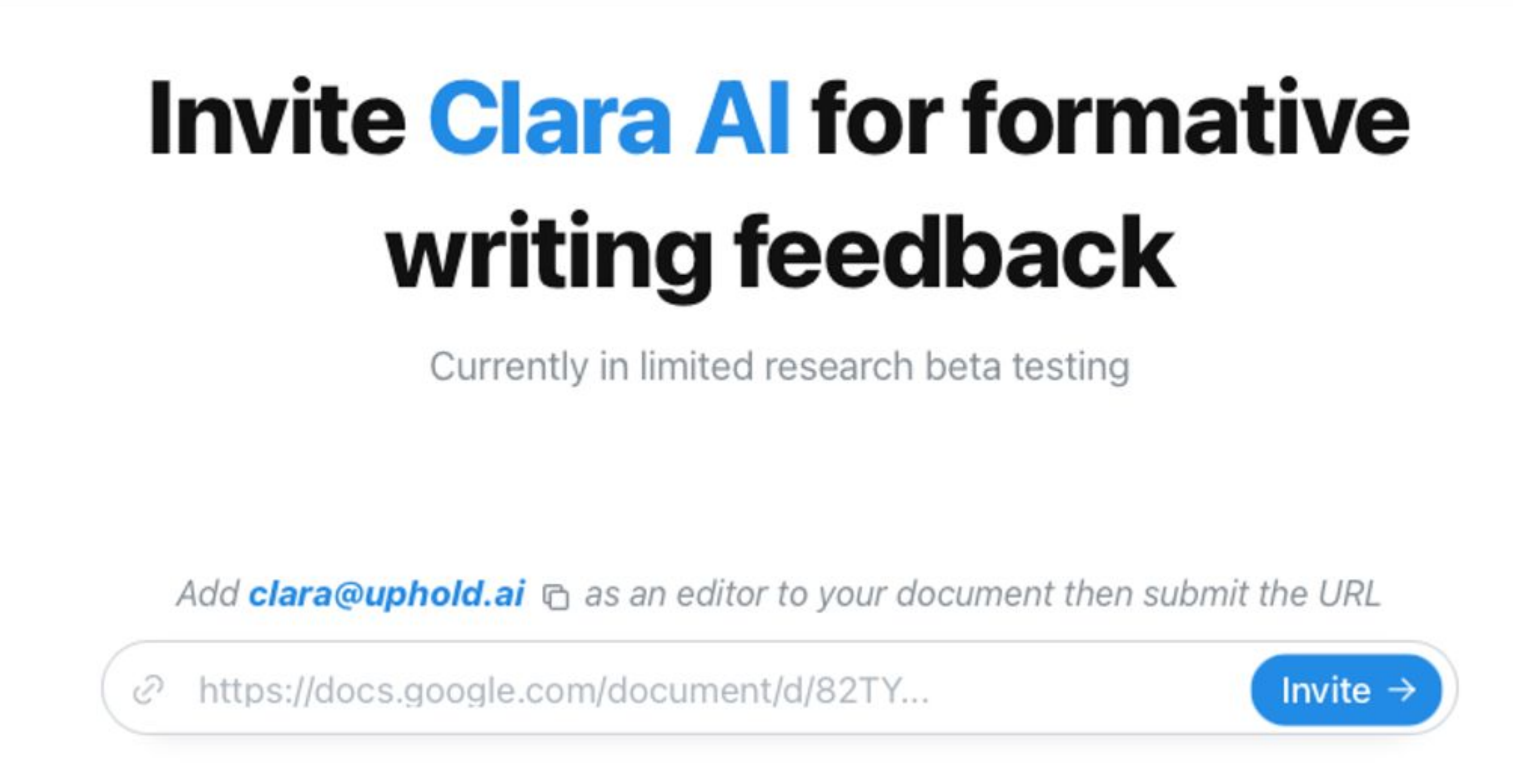
Hi, I'm Khanmigo! Ask me anything—I'm your new learning guide!

Some dreams for the future

- Individualized student learning
- Teacher assistance
- Collaborative learning
- Assessment
- Accuracy

And an even bigger question

- How will advances in AI change the need for human knowledge and expertise?

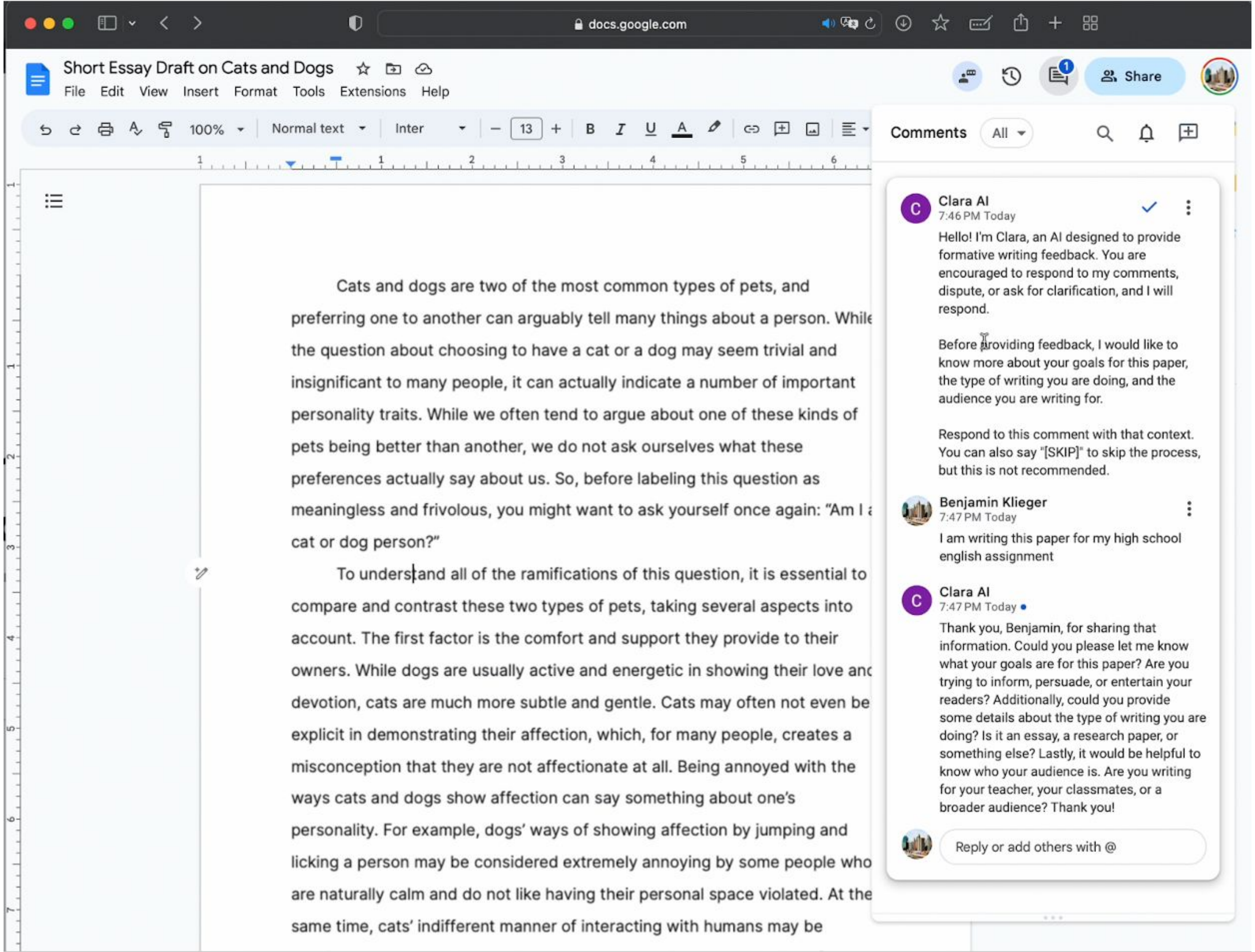


Invite Clara AI for formative writing feedback

Currently in limited research beta testing

Add ***clara@uphold.ai*** *as an editor to your document then submit the URL*

Ben Klieger



Cats and dogs are two of the most common types of pets, and preferring one to another can arguably tell many things about a person. While the question about choosing to have a cat or a dog may seem trivial and insignificant to many people, it can actually indicate a number of important personality traits. While we often tend to argue about one of these kinds of pets being better than another, we do not ask ourselves what these preferences actually say about us. So, before labeling this question as meaningless and frivolous, you might want to ask yourself once again: "Am I cat or dog person?"

To understand all of the ramifications of this question, it is essential to compare and contrast these two types of pets, taking several aspects into account. The first factor is the comfort and support they provide to their owners. While dogs are usually active and energetic in showing their love anc devotion, cats are much more subtle and gentle. Cats may often not even be explicit in demonstrating their affection, which, for many people, creates a misconception that they are not affectionate at all. Being annoyed with the ways cats and dogs show affection can say something about one's personality. For example, dogs' ways of showing affection by jumping and licking a person may be considered extremely annoying by some people who are naturally calm and do not like having their personal space violated. At the same time, cats' indifferent manner of interacting with humans may be

Clara AI
7:46 PM Today

Hello! I'm Clara, an AI designed to provide formative writing feedback. You are encouraged to respond to my comments, dispute, or ask for clarification, and I will respond.

Before providing feedback, I would like to know more about your goals for this paper, the type of writing you are doing, and the audience you are writing for.

Respond to this comment with that context. You can also say "[SKIP]" to skip the process, but this is not recommended.

Benjamin Klieger
7:47 PM Today

I am writing this paper for my high school english assignment

Clara AI
7:47 PM Today

Thank you, Benjamin, for sharing that information. Could you please let me know what your goals are for this paper? Are you trying to inform, persuade, or entertain your readers? Additionally, could you provide some details about the type of writing you are doing? Is it an essay, a research paper, or something else? Lastly, it would be helpful to know who your audience is. Are you writing for your teacher, your classmates, or a broader audience? Thank you!

Reply or add others with @

Safely Scaling Coding Style Feedback

Identifier names

- Ask GPT for variable and function name recommendations
- Lots of parsing to ensure the students get back ONLY function and variable names and nothing else

Constants and Magic Numbers

- Deterministic feedback (no-GPT)
- Magic nums that need constants
- Constants used as variables
- Variables used as constants

Comments

- Ask GPT to classify existing comments and to classify any missing comments
- For each class, show teacher written feedback
- Only give one piece of feedback for an existing comment and one for needed but not written comment

Decomposition

- Ask GPT to classify functions into categories
- Give teacher written feedback based on these classifications

Juliette Woodrow

Chris Piech

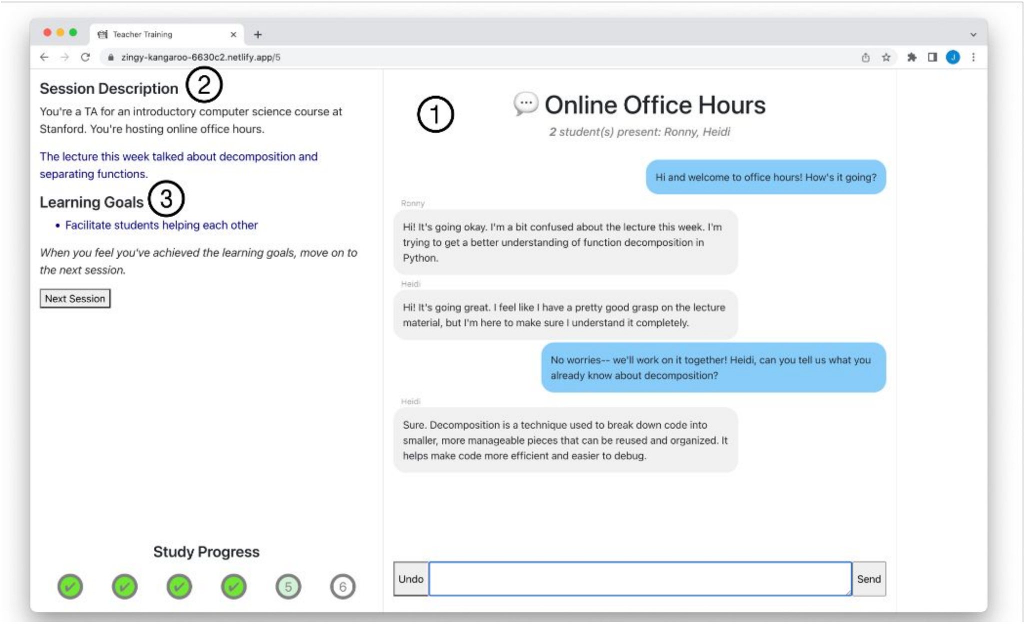
Enhanced error messages with GPT

- Compare two approaches to baseline options
  - Generate explanatory error messages using OpenAI's GPT in real time
  - Construct error messages that link to the course discussion forum
- Result
  - Students using GPT-generated error messages
    - Repeat an error 23.5% less often in the subsequent attempt
    - Resolve an error in 36.1% fewer additional attempts, compared to standard error messages

[Wang, Mitchell, Piech]

Teacher Training with GPT-based Students

- Interactive chat-based teacher training tool
  - Novice teachers practice with simulated students
  - Evaluation: office hours with two simulated students



Session Description

You're a TA for an introductory computer science course at Stanford. You're hosting online office hours.

The lecture this week talked about decomposition and separating functions.

Learning Goals

- Facilitate students helping each other

When you feel you've achieved the learning goals, move on to the next session.

Online Office Hours

2 student(s) present: Ronny, Heidi

Hi and welcome to office hours! How's it going?

Ronny

Hi! It's going okay. I'm a bit confused about the lecture this week. I'm trying to get a better understanding of function decomposition in Python.

Heidi

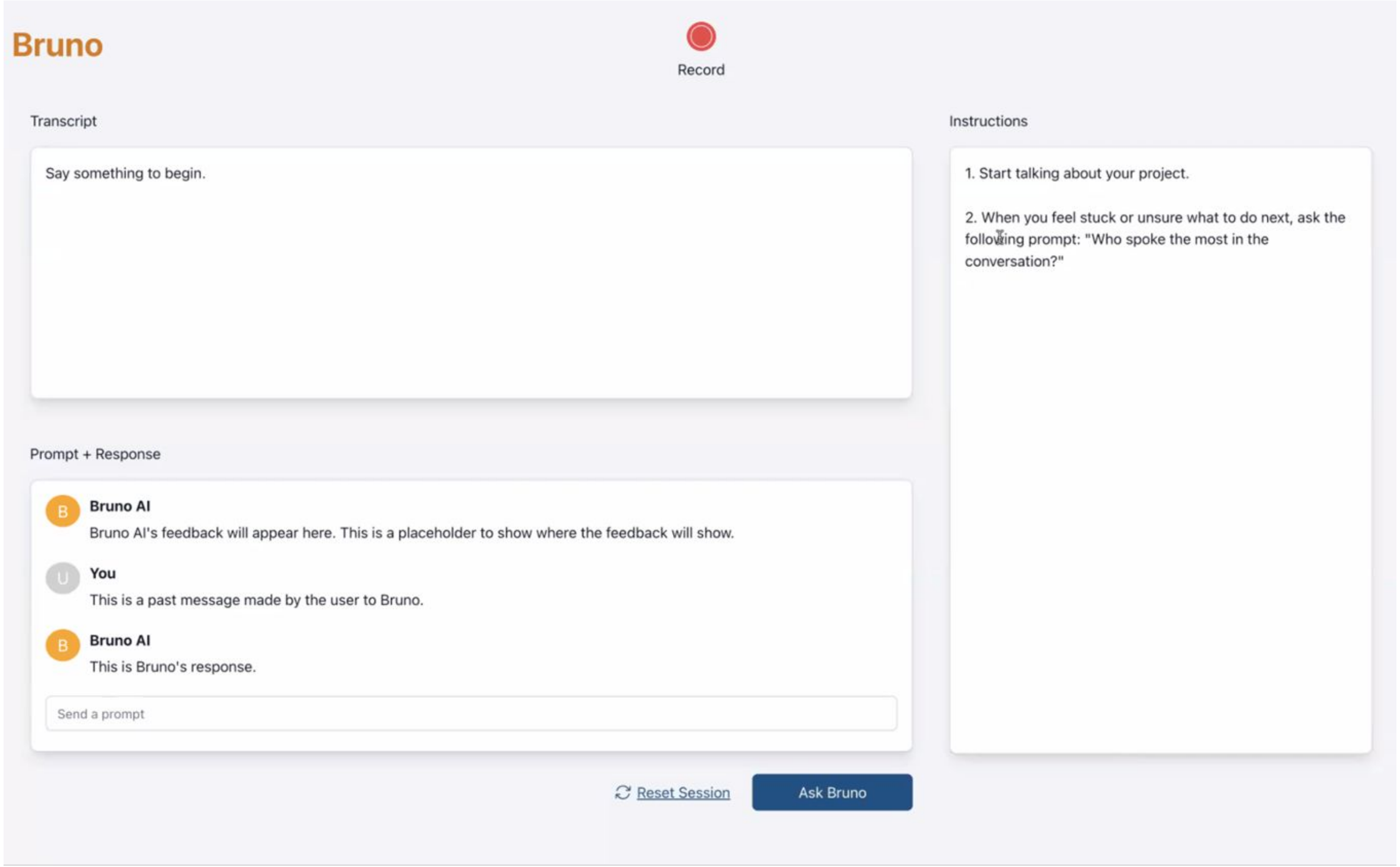
Hi! It's going great. I feel like I have a pretty good grasp on the lecture material, but I'm here to make sure I understand it completely.

No worries-- we'll work on it together! Heidi, can you tell us what you already know about decomposition?

Heidi

Sure. Decomposition is a technique used to break down code into smaller, more manageable pieces that can be reused and organized. It helps make code more efficient and easier to debug.

Julia M. Markel
Steven G. Opferman
James A. Landay
Chris Piech



Bruno

Transcript

Say something to begin.

Prompt + Response

Bruno AI
Bruno AI's feedback will appear here. This is a placeholder to show where the feedback will show.

You
This is a past message made by the user to Bruno.

Bruno AI
This is Bruno's response.

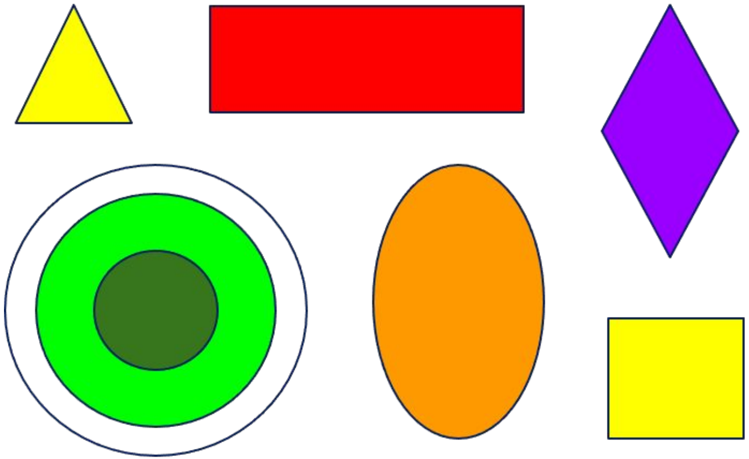
Instructions

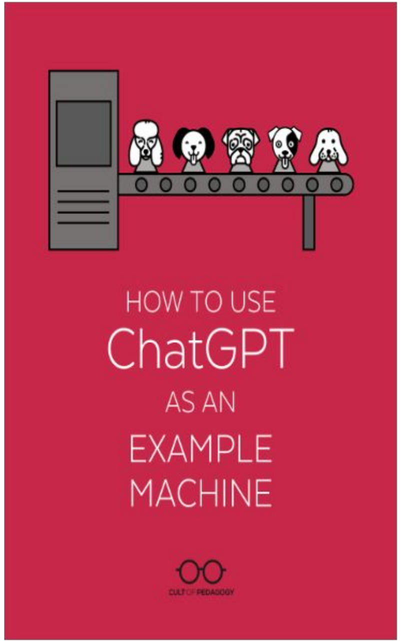
1. Start talking about your project.

2. When you feel stuck or unsure what to do next, ask the following prompt: "Who spoke the most in the conversation?"

AI-generated teaching examples [Mah, Levine]

- Contrasting cases





HOW TO USE
ChatGPT
AS AN
EXAMPLE
MACHINE

CULT OF PEDAGOGY

- A general concept is best illustrated using two or more contrasting examples

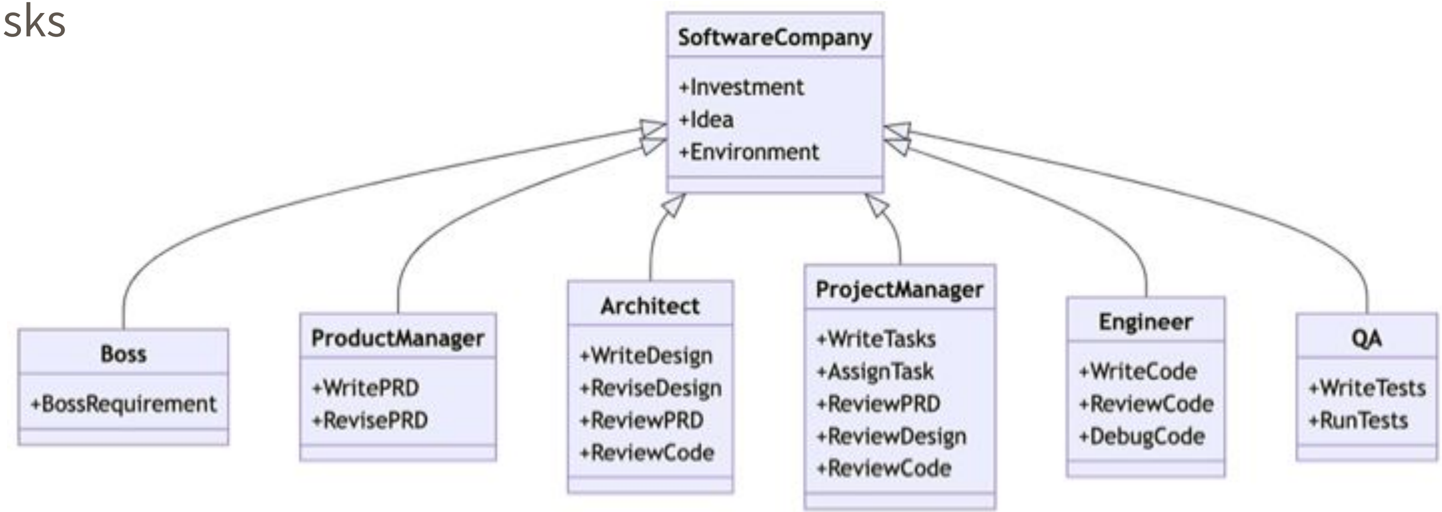
Example prompts for contrasting cases

- Science: Kinetic vs. Potential Energy
    - Give me five real-life examples of kinetic energy
    - Give me five real-life examples of potential energy
    - Give me an example that illustrates both kinetic and potential energy
- Social Studies: Types of Government
    - Give four examples of a person named Vera trying to change a law in the following forms of government: a *monarchy*, a *communist government*, a *direct democracy*, and a *representative democracy*
    - Do not use the names of the gov't in the examples or explain what they are

MetaGPT

https://github.com/geekan/MetaGPT

- Assign different roles to GPTs to form a collaborative software entity for complex tasks



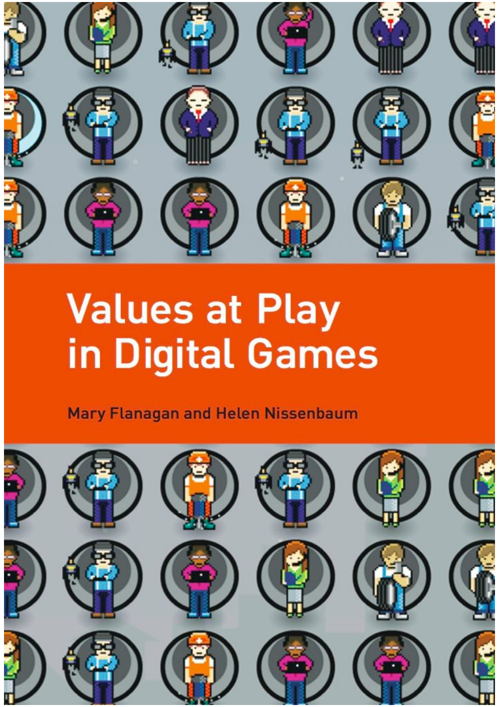
- Create user stories / competitive analysis / requirements / data structures / APIs / documents
- Includes product managers / architects / project managers / engineers

Trustworthiness dimensions

- Grounding - every assertion has authoritative basis
- Consistency - semantically equivalent queries treated similarly
- Confidence - acknowledge uncertainty accurately
- Interpretability - be able to show how response was generated
- Alignment - not harmful, toxic, biased, dishonest, unreliable
    - Respect privacy
    - Behave fairly and mitigate bias
- Resist adversarial manipulation
    - Malicious input should not subvert desirable properties

Values @ Play Framework

- There are common (not necessarily universal) ethical and political values
- Artifacts may embody ethical and political values
- Steps taken in design and development have the power to affect the nature of these values



Types of values

- Ethical: “How we treat others and how they treat us”
    - E.G. kindness, honesty, generosity, fidelity, integrity, with respect, safety, autonomy, creativity, peace, pleasure, well-being, friendship, collaboration, health, responsibility, happiness, and contentment
- Political: Define relationships within and between societies, arrangements and distributions of power, authority, goods, privilege
    - E.G. justice, equality, security, stability, cooperation, tolerance, privacy, accountability, democracy, voice, property, liberty, liberation, autonomy, equal opportunity, and government transparency

Isabelle Hau

John Mitchell

Josh Weiss

- What concerns are most prominent in education? Why do we worry about trust in a classroom?
- How does trust play out in a learning experience? Is there anything unique here relative to, say, healthcare, business, etc?
- Looking at the dimensions of trustworthiness, which have the highest priority for education? Are there any projects that exemplify any of these dimensions?
- Of the projects you’ve seen, how do they keep trustworthiness as a focus? What does it look like to keep dimensions of trustworthiness in scope as a project moves along? Or as an organization grows?
- What applications to education would most benefit from project work now on these concerns?

Additional project examples

- Lesson planning - TeachAssist (Riz Malik)
- Neurodiversity - ADHD FlexABLE ai (link)
- Neurodiversity and creativity in Bangladesh (Labib Rahman)
- Language learning - KATE (Ted Song; link)
- Tutoring-english
- Learners as teachers - study buddy (Olivia Tomaneo; link)
- Personalized early reading: Ello (Nick Haber; link)
- Project Read (Ramakrishnan, GSB; link)
- Teacher feedback - Mpowering Dora Demszky, Mei Tan, Rose Wang (NLP)
- Intersection of VR/AI for engagement - Alex Stolyarik
- Career orientation / pathways - Workera and Clarify
- Detecting text - Chris Manning, Chelsea Finn, Eric Mitchell
- AI literacy - CRAFT - Victor Lee, Parth Sarin (link)
- Novice approach to programming - Benjamin Xie