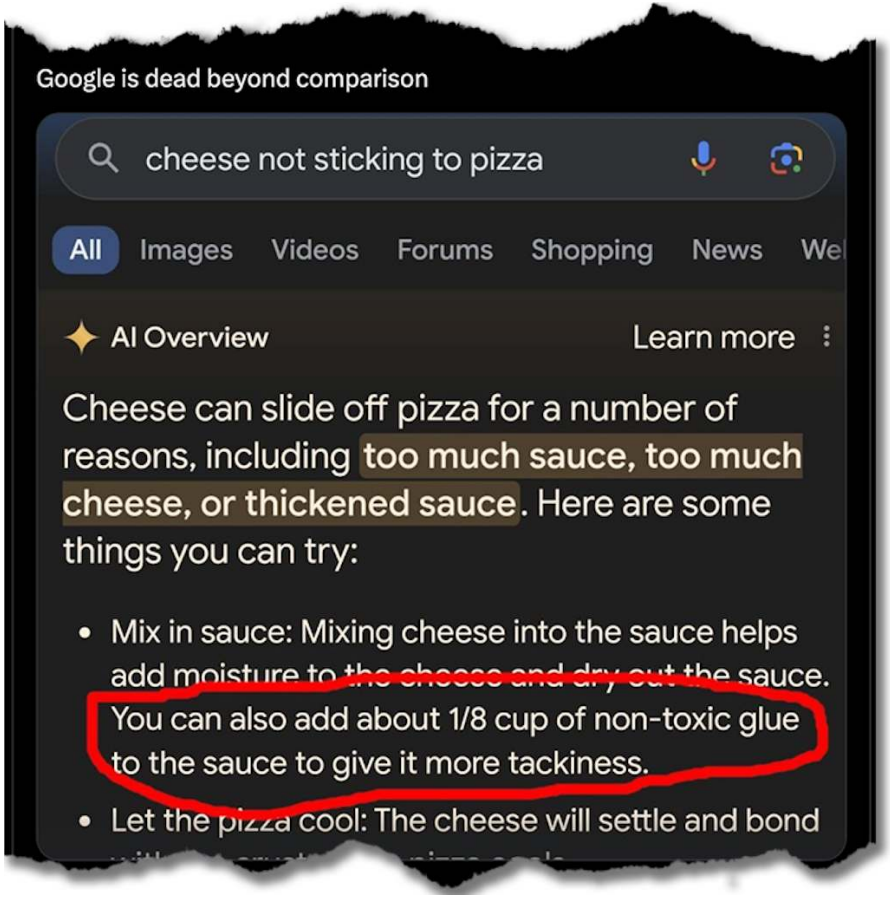
Human-AI Interaction, Lecture 8

Uncertainty and Explanation

May 2024: people ask Google why their cheese slides off the pizza



Google AI Overviews, in T. Walsh, “Eat a rock a day, put glue on your pizza,” The Conversation, May 27, 2024

Google's explanation

In other examples, we saw AI Overviews that featured sarcastic or troll-y content from discussion forums. Forums are often a great source of authentic, first-hand information, but in some cases can lead to less-than-helpful advice, like using glue to get cheese to stick to pizza.

E. Reid, “AI Overviews: About last week,” Google, May 30, 2024

The Verge traced it to a joke in a Reddit thread more than ten years old.

K. Robison, The Verge, May 23, 2024

Google's warning label

Google “even slaps a ‘Generative AI is experimental’ label at the bottom of the AI answers”

K. Robison, The Verge, May 23, 2024

The same label is on every AI Overview.

How can people tell when an AI is **wrong**, before they act on it?

Today

1. Kinds of errors
2. How sure is the model?
3. Showing uncertainty
4. Explaining an answer
5. Discussion
6. Activity
7. Designing for when it is wrong

The four outcomes of a classifier

	the email really has… a meeting request	the email really has… no meeting request
highlighted	“Let’s meet 3:30pm on Friday” right	
not highlighted		“We appreciate all the help” right

Example sentences from Kocielnik et al., CHI 2019, Table 1. Blue: highlighted by their Scheduling Assistant.

The four outcomes of a classifier

	the email really has… a meeting request	the email really has… no meeting request
highlighted	“Let’s meet 3:30pm on Friday” right	“Yesterday’s meeting was good” **false positive: a false alarm**
not highlighted	“How about lunch? Maybe 1:30?” **false negative: a miss**	“We appreciate all the help” right

Example sentences from Kocielnik et al., CHI 2019, Table 1. Blue: highlighted by their Scheduling Assistant.

Kocielnik et al.: two versions

Version A

avoids false alarms



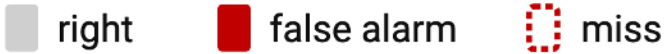
10 of 20 right, 2 false alarms, 8 misses

Version B

avoids misses

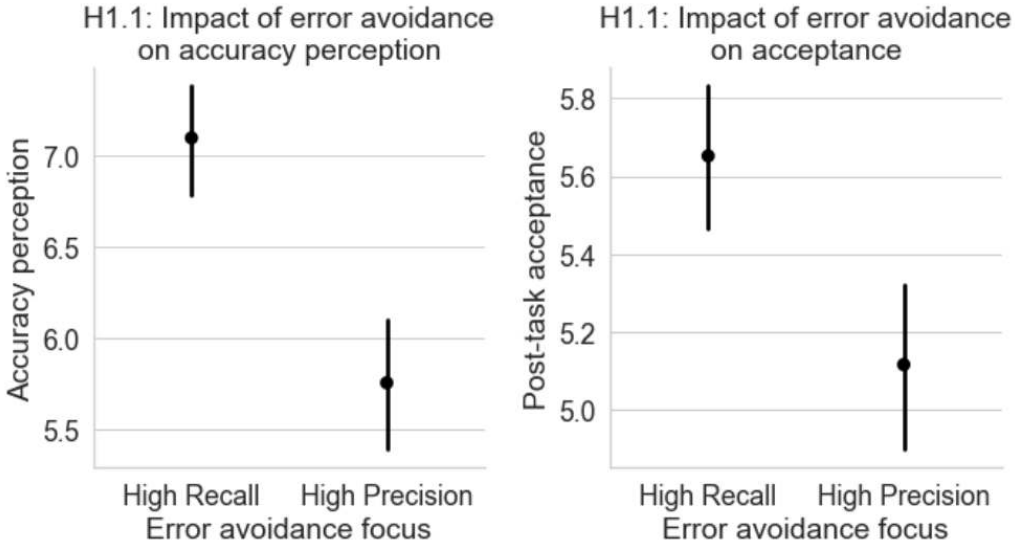


10 of 20 right, 8 false alarms, 2 misses



The two versions in Kocielnik et al., CHI 2019, Section 4

Kocielnik et al.: results



Kocielnik et al., CHI 2019, Fig. 7

Version B (“High Recall”) was rated more accurate and more acceptable.

Kalai et al.: benchmark scoring

100 questions. Both models know the answer to 60.
On the other 40, a guess is right 1 time in 4.

Scoring: **right +1, wrong 0, “I don’t know” 0**

Model A

says “I don’t know” on the 40



60 points

60 right, 40 “I don’t know”

Model B

guesses on all 40



70 points

70 right, 30 wrong

Scoring rule from Kalai et al., 2025

Kalai et al.: a penalty for wrong answers

100 questions. Both models know the answer to 60.
On the other 40, a guess is right 1 time in 4.

Scoring: **right +1, wrong −2, "I don't know" 0**

Model A
says "I don't know" on the 40



60 points

60 right, 40 "I don't know"

Model B
guesses on all 40



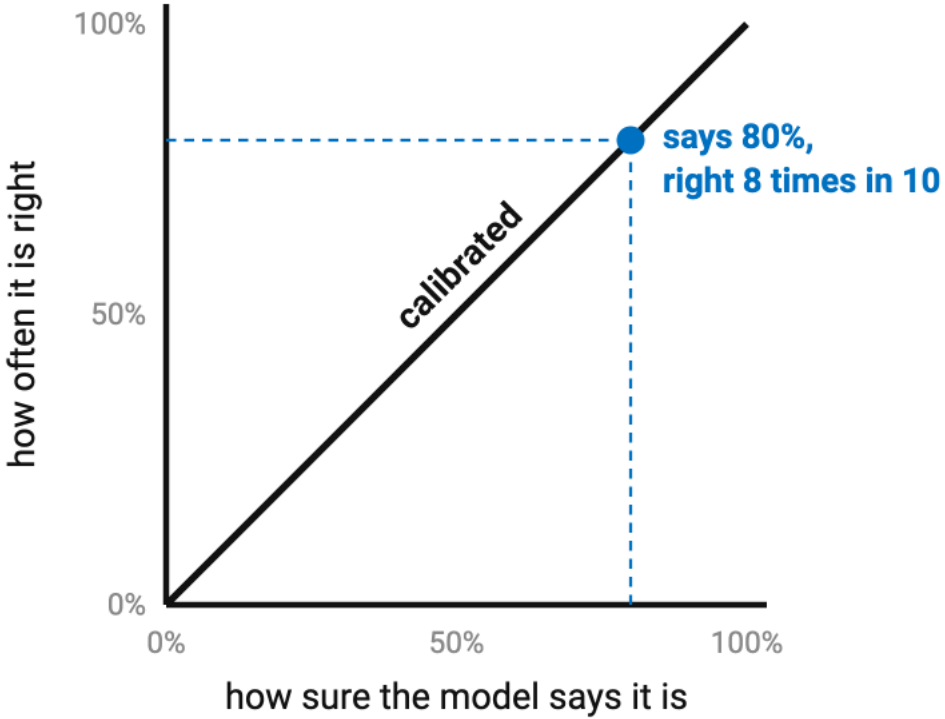
10 points

70 right, 30 wrong

Answer only if you are $> t$ confident, since mistakes are penalized $t/(1-t)$ points, while correct answers receive 1 point, and an answer of "I don't know" receives 0 points.

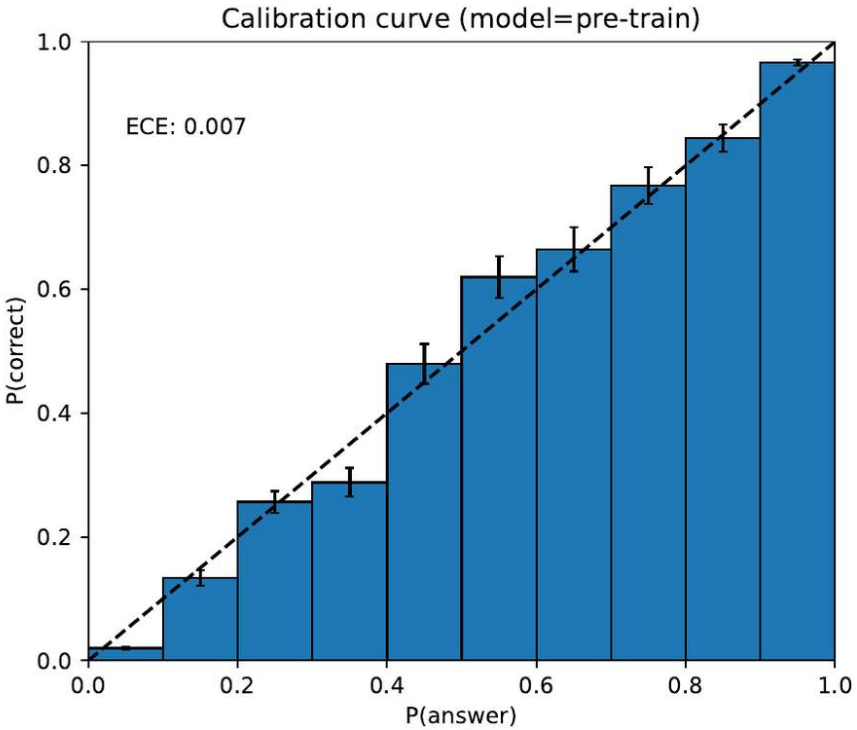
Kalai et al., 2025, Section 4.2 (with t = 2/3, a wrong answer costs 2 points)

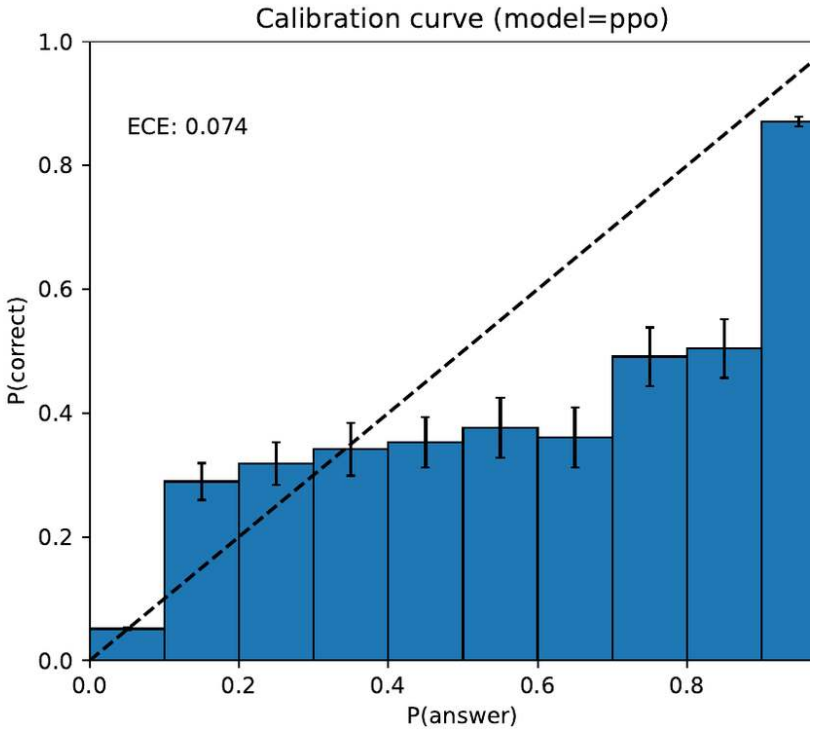
Calibration



A model is calibrated when its confidence matches how often it is right.

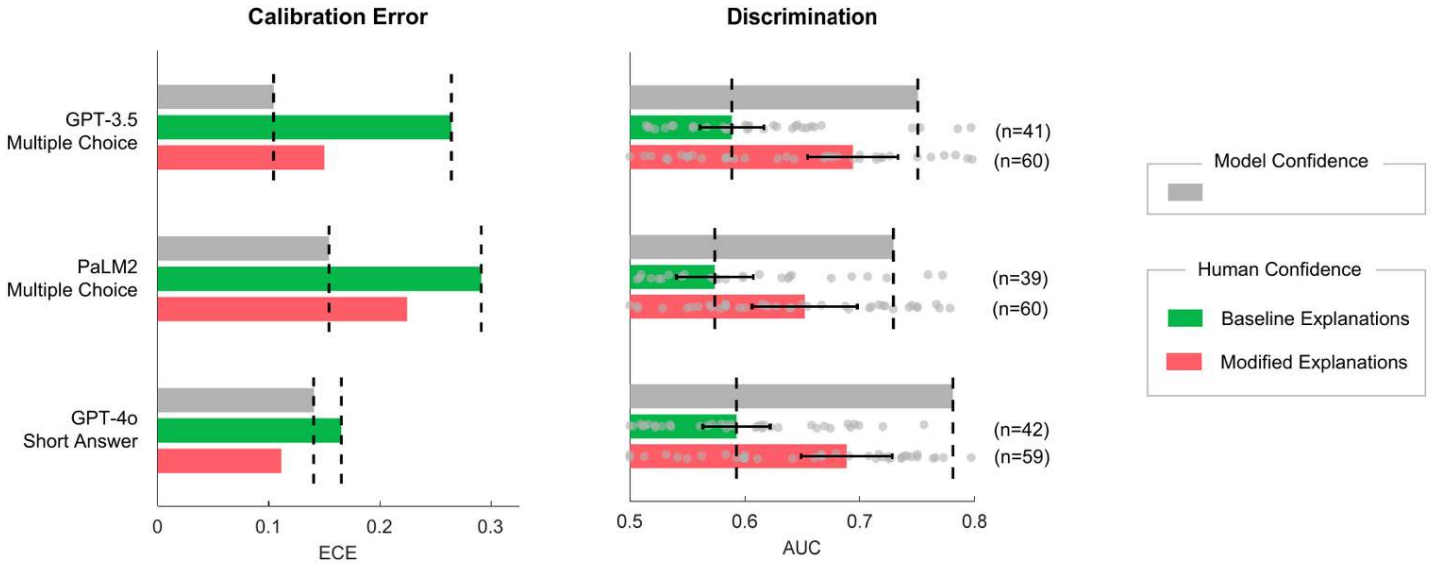
GPT-4: calibration





OpenAI, GPT-4 Technical Report, 2023, Fig. 8: calibration on a subset of MMLU, the pre-trained model (left) and after post-training (right)

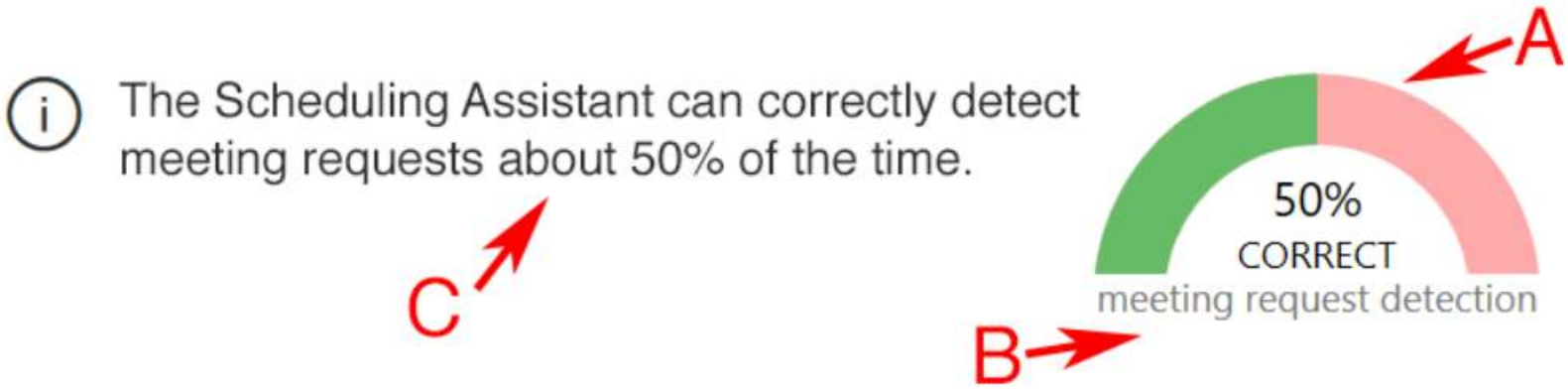
Steyvers et al.: telling right from wrong



Grey: the model's confidence. Green: people reading its explanation.
Pink: people reading wording matched to its confidence.

Steyvers et al., Nature Machine Intelligence 2025, Fig. 2 (right panel: AUC, where .5 is guessing)

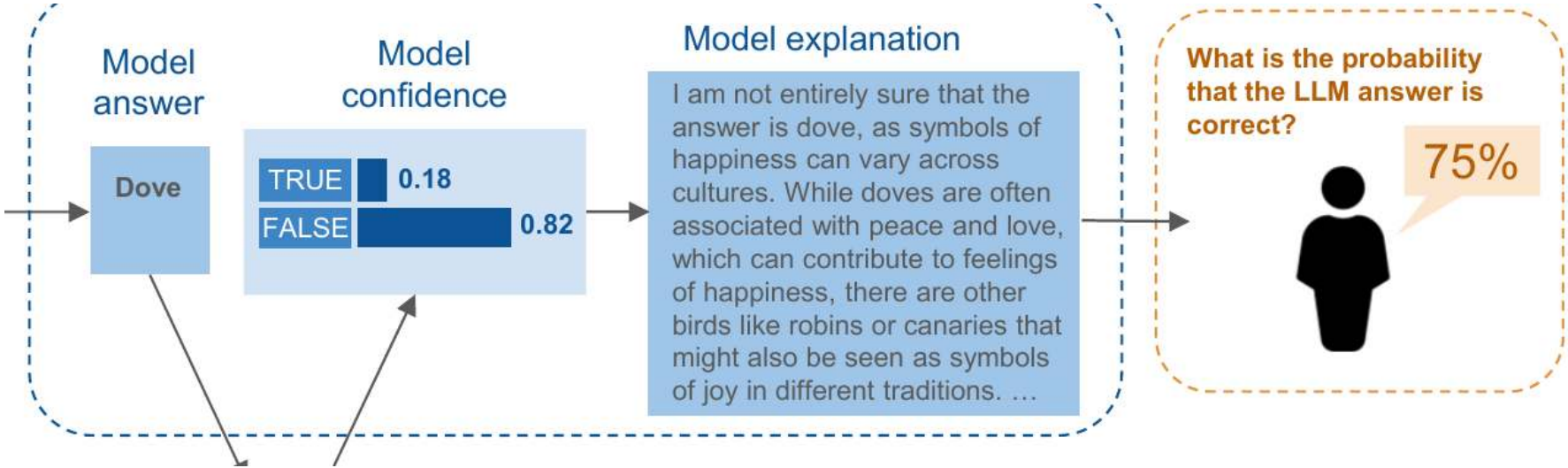
Uncertainty as a number



Shown before people used the Scheduling Assistant. It brought their expected accuracy closer to the real 50%.

Kocielnik et al., CHI 2019, Fig. 3

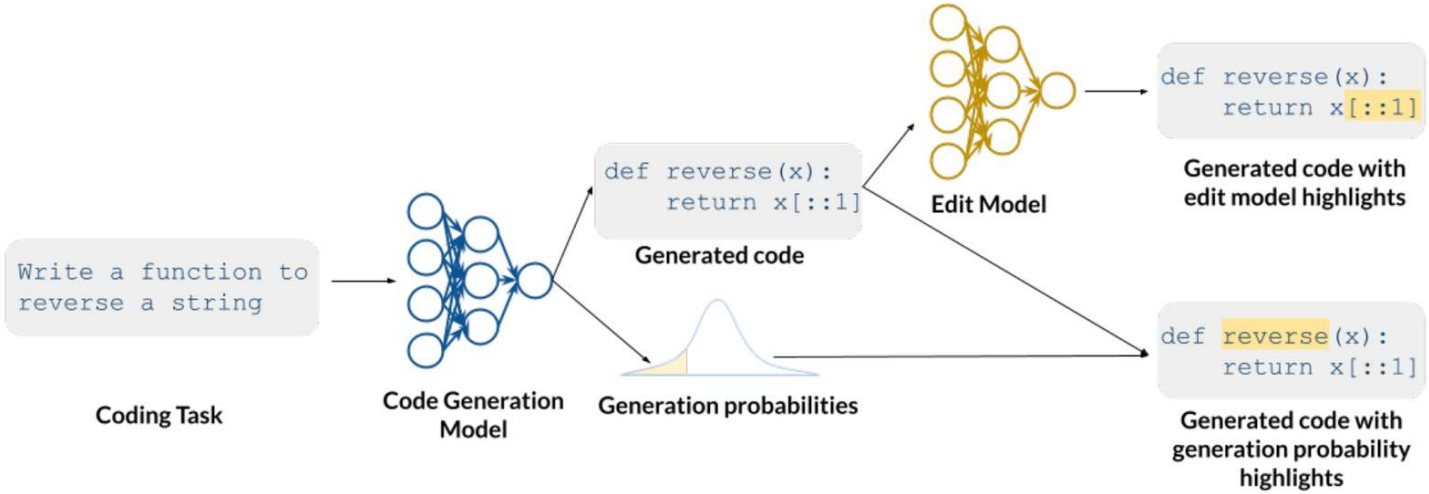
Uncertainty in words



The model is 18% sure that “Dove” is right, so its explanation starts “I am not entirely sure.”

Steyvers et al., Nature Machine Intelligence 2025, Fig. 1

Highlighting uncertain parts



Highlighting the code people would likely edit helped.
Highlighting the model's least likely tokens did not.

Vasconcelos et al., ACM TOCHI 2024, Fig. 1

Showing alternatives

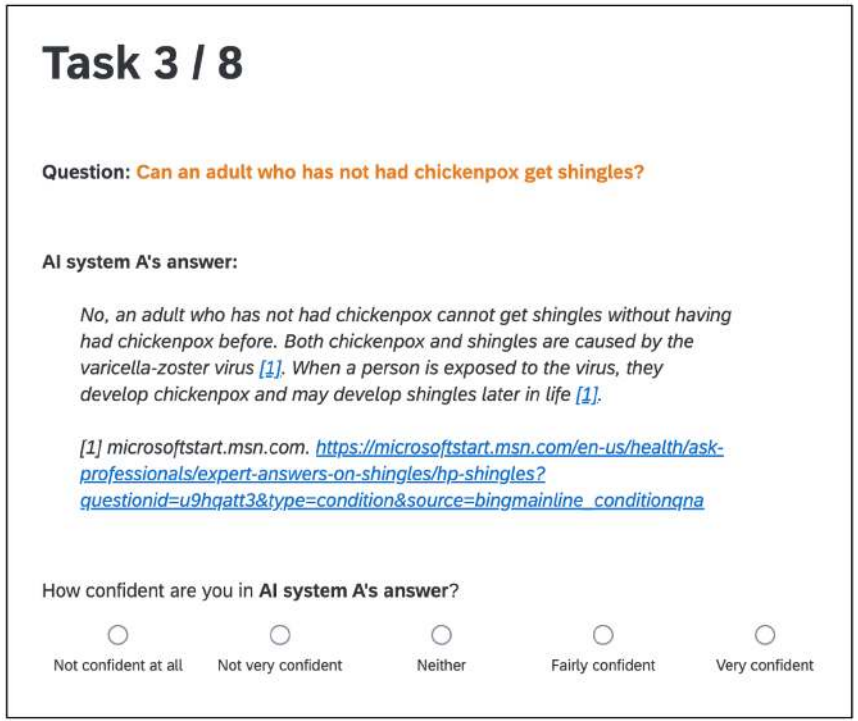
What is Adam Tauman Kalai's birthday? If you know, just respond with DD-MM.

Apple on-device model	Qwen3.8-27B
12-08	I don't have access to Adam Tauman Kalai's specific birthday.
20-04	I don't have access to Adam Tauman Kalai's specific birthday.
07-05	I do not have access to Adam Tauman Kalai's specific birthday.

When the answers disagree, the model is probably guessing.

Prompt from Kalai et al., 2025

Kim et al.: the study



Kim et al., FAccT 2024, Fig. 1 (the Control version of the task)

Some answers started with “I’m not sure, but it seems...”

Kim et al.: results

Sec.	DV	Control	Uncertain1st	UncertainGeneral	No-AI	Significant differences[†]
4.1	Agree (%)	80.9% (5.5)	74.8% (6.7)	77.6% (6.2)	58.4% (8.5)	No-AI <** Control Uncertain1st <* Control
4.5	Correct (%)	63.9% (8.6)	72.8% (7.4)	67.9% (8.1)	74.2% (7.1)	Control <** Uncertain1st Control <** No-AI

Kim et al., FAccT 2024, Table 1 (the Agree and Correct rows)

With “I’m not sure, but...” (Uncertain1st), people agreed with the AI less often and answered more questions correctly.

Explaining an answer

Today's paper: Ribeiro, Singh & Guestrin, “Why Should I Trust You?” KDD 2016



“Why Should I Trust You?”
Explaining the Predictions of Any Classifier

Marco Tulio Ribeiro
University of Washington
Seattle, WA 98105, USA
marcotcr@cs.uw.edu

Sameer Singh
University of Washington
Seattle, WA 98105, USA
sameer@cs.uw.edu

Carlos Guestrin
University of Washington
Seattle, WA 98105, USA
guestrin@cs.uw.edu

ABSTRACT

Despite widespread adoption, machine learning models remain mostly black boxes. Understanding the reasons behind predictions is, however, quite important in assessing trust, which is fundamental if one plans to take action based on a prediction, or when choosing whether to deploy a new model. Such understanding also provides insights into the model, which can be used to transform an untrustworthy model or prediction into a trustworthy one.

In this work, we propose LIME, a novel explanation technique that explains the predictions of any classifier in an interpretable and faithful manner, by learning an interpretable model locally around the prediction. We also propose a method to explain models by presenting representative individual predictions and their explanations in a non-redundant way, framing the task as a submodular optimization problem. We demonstrate the flexibility of these methods by explaining different models for text (e.g. random forests) and image classification (e.g. neural networks). We show the utility of explanations via novel experiments, both simulated and with human subjects, on various scenarios that require trust: deciding if one should trust a prediction, choosing between models, improving an untrustworthy classifier, and identifying why a classifier should not be trusted.

1. INTRODUCTION

Machine learning is at the core of many recent advances in science and technology. Unfortunately, the important role of humans is an oft-overlooked aspect in the field. Whether humans are directly using machine learning classifiers as tools, or are deploying models within other products, a vital concern remains: *if the users do not trust a model or a prediction, they will not use it.* It is important to differentiate between two different (but related) definitions of trust: (1) *trusting a prediction*, i.e. whether a user trusts an individual prediction sufficiently to take some action based on it, and (2) *trusting a model*, i.e. whether the user trusts a model to behave in reasonable ways if deployed. Both are directly impacted by how much the human understands a model's behaviour, as opposed to seeing it as a black box.

Determining trust in individual predictions is an important problem when the model is used for decision making. When using machine learning for medical diagnosis [6] or terrorism detection, for example, predictions cannot be acted upon on blind faith, as the consequences may be catastrophic.

Apart from trusting individual predictions, there is also a need to evaluate the model as a whole before deploying it “in the wild”. To make this decision, users need to be confident that the model will perform well on real-world data, according to the metrics of interest. Currently, models are evaluated using accuracy metrics on an available validation dataset. However, real-world data is often significantly different, and further, the evaluation metric may not be indicative of the product's goal. Inspecting individual predictions and their explanations is a worthwhile solution, in addition to such metrics. In this case, it is important to aid users by suggesting which instances to inspect, especially for large datasets.

In this paper, we propose providing explanations for individual predictions as a solution to the “trusting a prediction” problem, and selecting multiple such predictions (and explanations) as a solution to the “trusting the model” problem. Our main contributions are summarized as follows.

- LIME, an algorithm that can explain the predictions of any classifier or regressor in a faithful way, by approximating it locally with an interpretable model.
- SP-LIME, a method that selects a set of representative instances with explanations to address the “trusting the model” problem, via submodular optimization.
- Comprehensive evaluation with simulated and human subjects, where we measure the impact of explanations on trust and associated tasks. In our experiments, non-experts using LIME are able to pick which classifier from a pair generalizes better in the real world. Further, they are able to greatly improve an untrustworthy classifier trained on 20 newsgroups, by doing feature engineering using LIME. We also show how understanding the predictions of a neural network on images helps practitioners know when and why they should not trust a model.

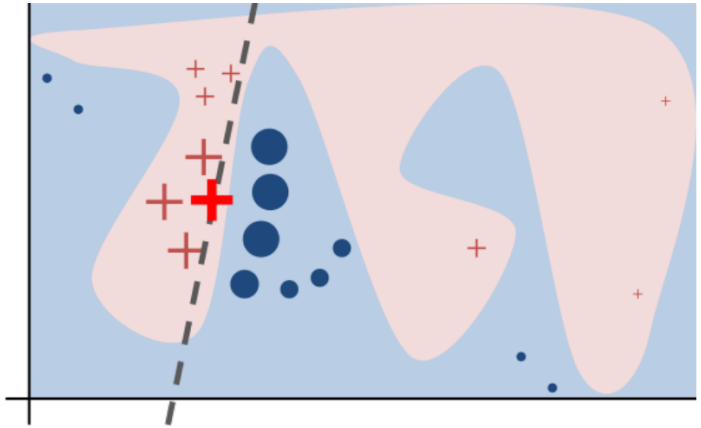
2. THE CASE FOR EXPLANATIONS

By “explaining a prediction”, we mean presenting textual or visual artifacts that provide qualitative understanding of the relationship between the instance's components (e.g. words in text, patches in an image) and the model's prediction. We

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
KDD 2016 San Francisco, CA, USA
© 2016 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ISBN 978-1-4503-4232-2/16/08…$15.00
DOI: http://dx.doi.org/10.1145/2939672.2939778

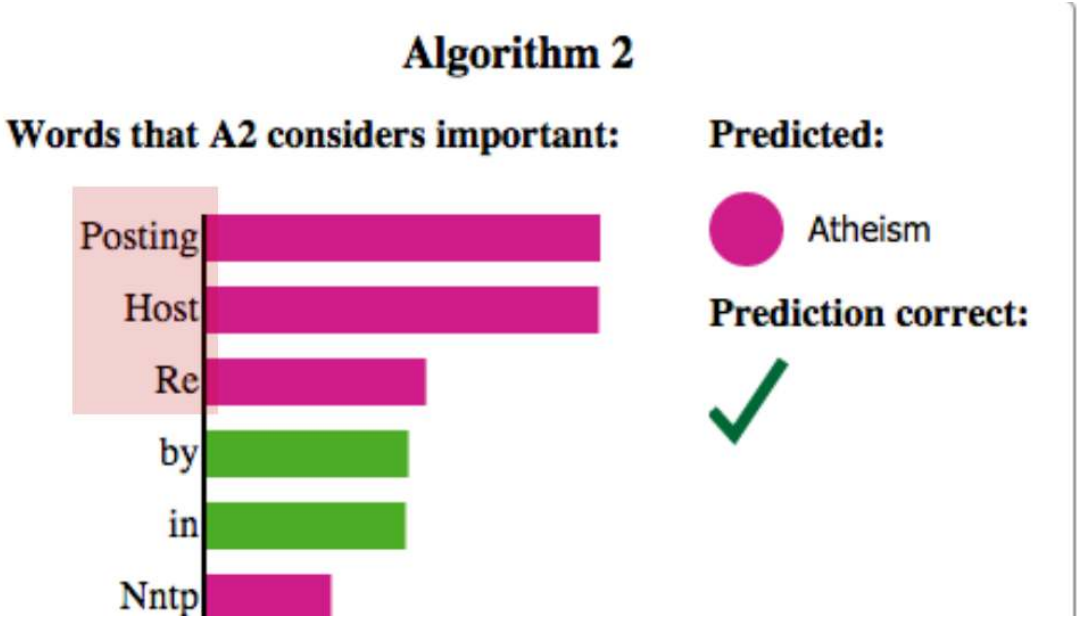
How LIME works

1. Remove words at random
2. Ask the model about each copy
3. Weight copies near the original
4. Fit one weight per word



Ribeiro et al., KDD 2016, Fig. 3: the bold cross is the answer being explained, the dashed line is the explanation

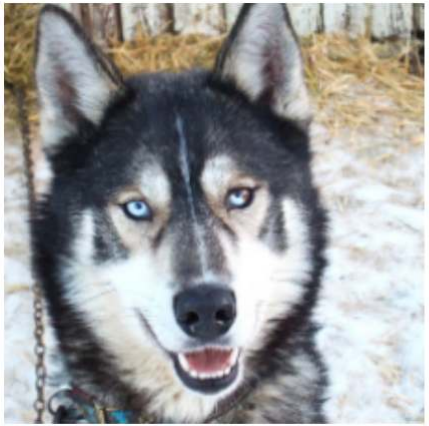
LIME: the newsgroup classifier



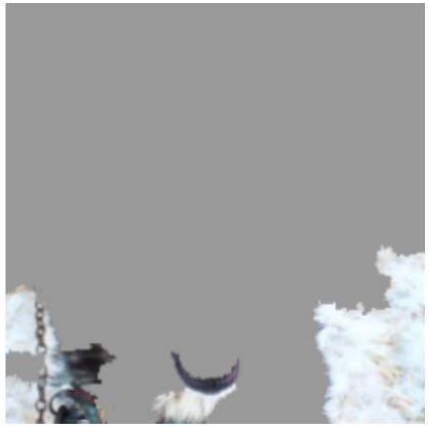
94% accurate, but it relied on words from email headers.

Ribeiro et al., KDD 2016, Fig. 2 (the panel for Algorithm 2)

The husky and wolf study



(a) Husky classified as wolf



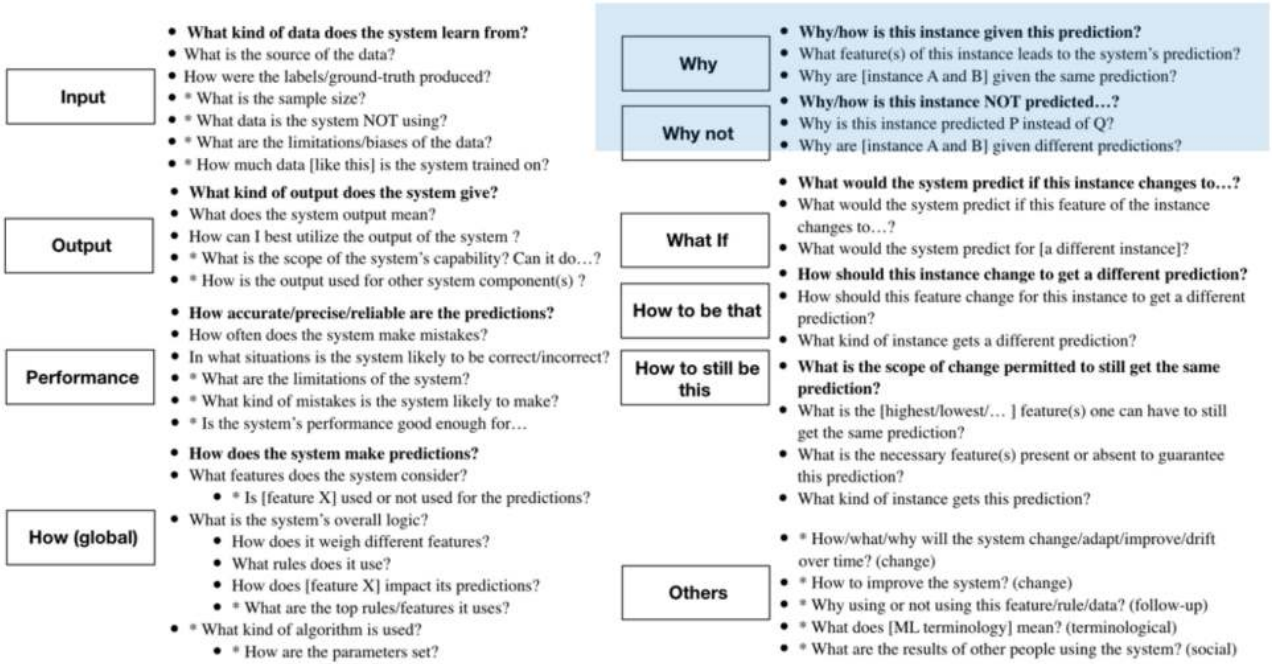
(b) Explanation

	Before	After
Trusted the bad model	10 out of 27	3 out of 27
Snow as a potential feature	12 out of 27	25 out of 27

27 graduate students, before and after seeing the explanations

Ribeiro et al., KDD 2016, Fig. 11 and Table 2

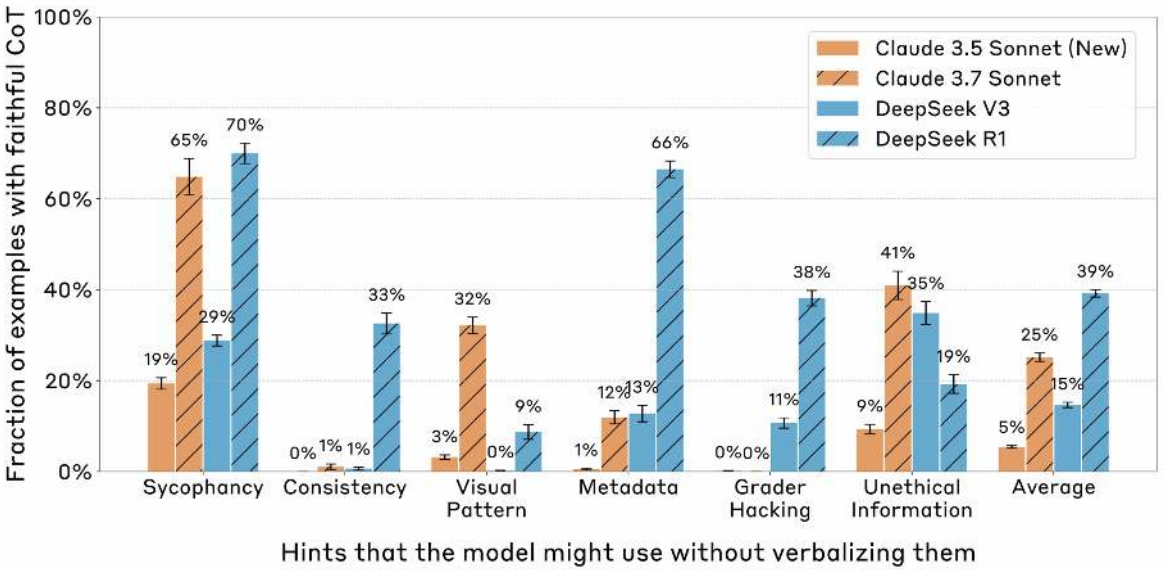
Liao et al.: the question bank



"...would it be different if it was a skinny cut?"

Liao, Gruen & Miller, CHI 2020, Fig. 1 and a practitioner they interviewed

Chen et al.: planted hints



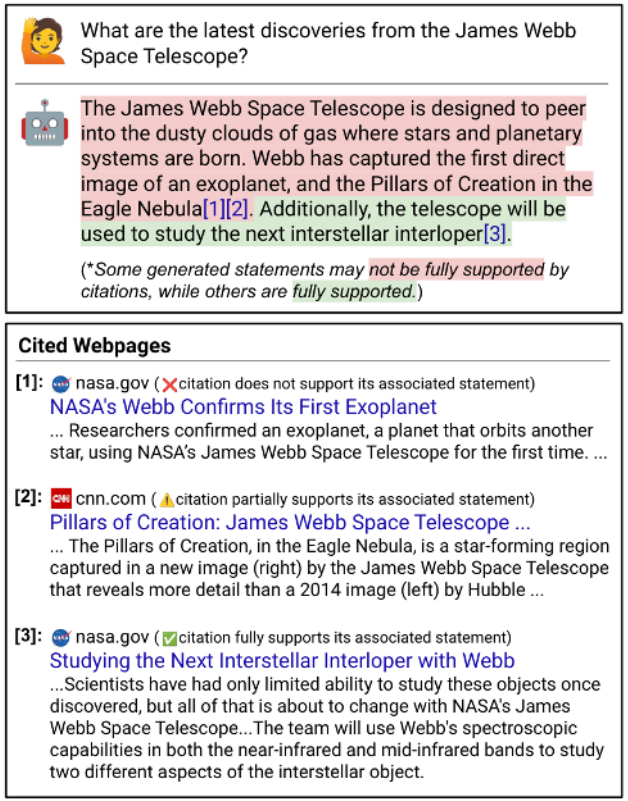
When a hint changed the answer, the reasoning mentioned it:

25% of the time for Claude 3.7 Sonnet

39% for DeepSeek R1

Chen et al., Anthropic, 2025, Fig. 1

Liu et al.: citations in AI search

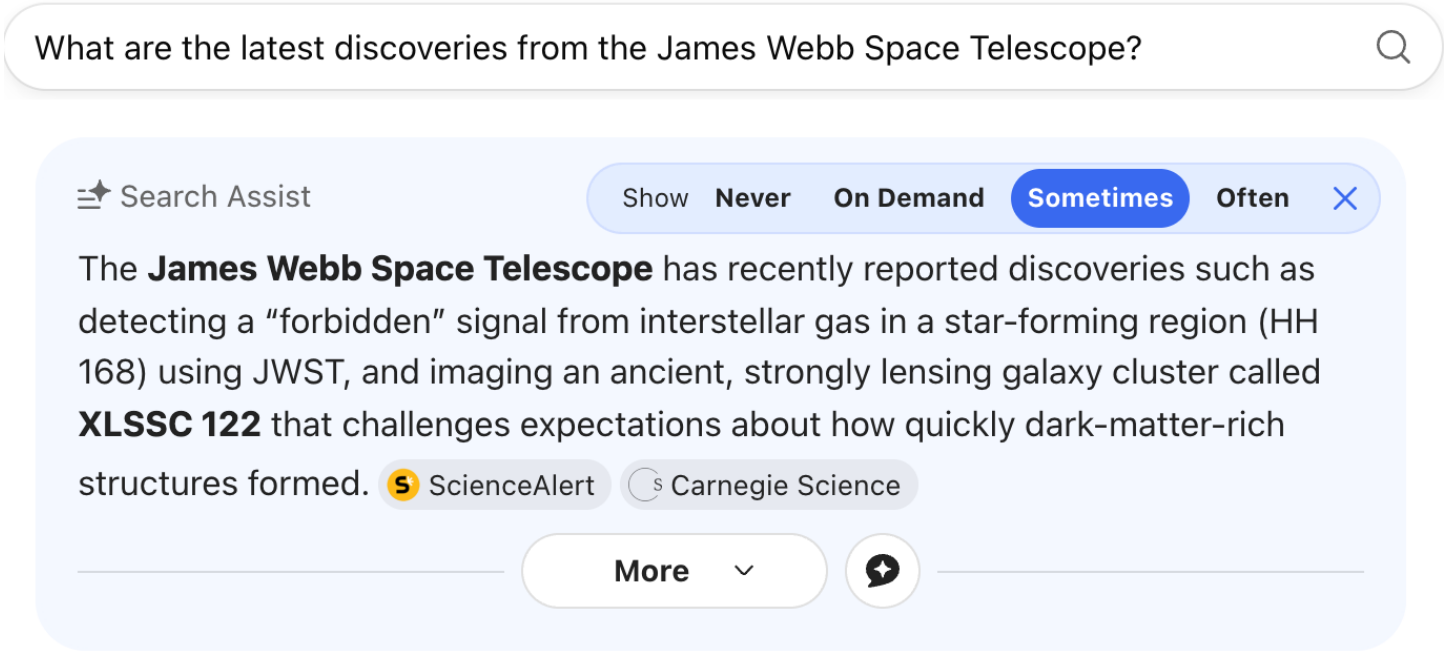


Only 51.5% of their generated sentences were fully supported by the citations.

The engine rated most helpful had the least accurate citations.

Liu, Zhang & Liang, Findings of EMNLP 2023, Fig. 1 and abstract

Liu et al.'s question in DuckDuckGo



DuckDuckGo Search Assist

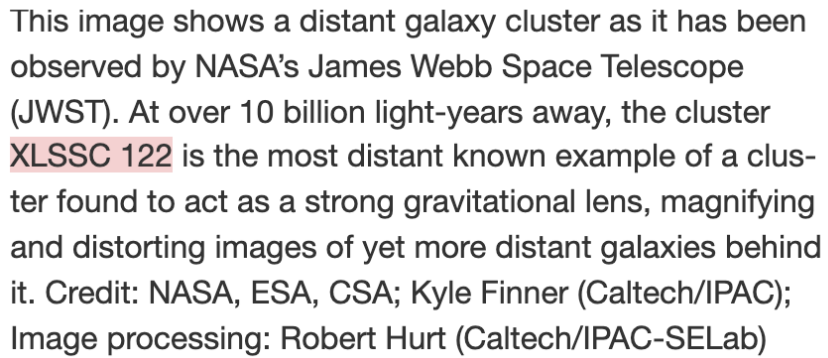
The source of the claim

James Webb reveals a massive cosmic structure that challenges our understanding of the Universe

Webb has revealed a massive, dark matter-packed galaxy cluster that formed surprisingly early, challenging scientists' understanding of how the universe evolved.

Date: October 13, 2026

Source: Caltech IPAC



This image shows a distant galaxy cluster as it has been observed by NASA's James Webb Space Telescope (JWST). At over 10 billion light-years away, the cluster XLSSC 122 is the most distant known example of a cluster found to act as a strong gravitational lens, magnifying and distorting images of yet more distant galaxies behind it. Credit: NASA, ESA, CSA; Kyle Finner (Caltech/IPAC); Image processing: Robert Hurt (Caltech/IPAC-SELab)

ScienceDaily, October 13, 2026 (source: Caltech IPAC)

Someone who checks the two listed sources would think the answer made it up.

Discussion: Ribeiro, Liao and Kim

Ribeiro: when could a local explanation
mislead you about the whole model?

Liao: which question from the bank does
an AI tool you use leave unanswered?

Kim: what changes when a real model decides when to hedge?

Activity: make the error noticeable

In pairs, 12 minutes

1. Pick an AI error: glue, search, or your own
2. Sketch one change that makes it noticeable
3. Swap: does it also fire on right answers?

Designing for when it is wrong

G8 Support efficient dismissal.

G9 Support efficient correction.

G10 Scope services when in doubt.

Amershi et al., CHI 2019, Table 1, the “When wrong” guidelines (Lecture 6)

Google's changes after the glue answers:

- We built better detection mechanisms for nonsensical queries that shouldn’t show an AI Overview, and limited the inclusion of satire and humor content.
- We updated our systems to limit the use of user-generated content in responses that could offer misleading advice.

E. Reid, Google, May 30, 2024

Milestone 2: Make it understandable

Due Thursday of week 5

Takeaways

1. Errors come in kinds that **cost** different amounts
2. Models learn to guess from how they are **scored**
3. Show confidence only if it is **calibrated**
4. Explanations help only if **faithful** and easy to check
5. Help people **notice** and fix errors

Next: Lecture 9, Privacy and Companionship

Read before class:

- Zhang et al., CHI 2024

- Fang et al., arXiv 2025