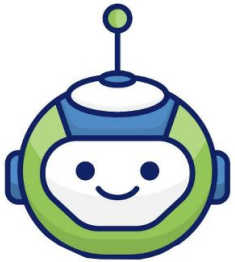


How to Evaluate Human-AI Collaborative Systems

NLP @USC

USC Center for AI in Society

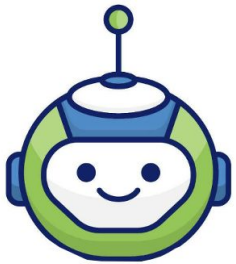
DILL

November 28, 2025
CS 698

A few slide ideas were partially adapted from ACL 2025 HAI tutorial and Human Centered NLP

What is Human-AI Interaction?

Basically, a field where humans and AIs interact.



AIs: LLMs, dialog system, translator, recommender system, autonomous driving system.

Humans: AI researchers, model developers, domain experts, end users.

Interact:

Humans collaborate with AI,

Humans get assistance from AI,

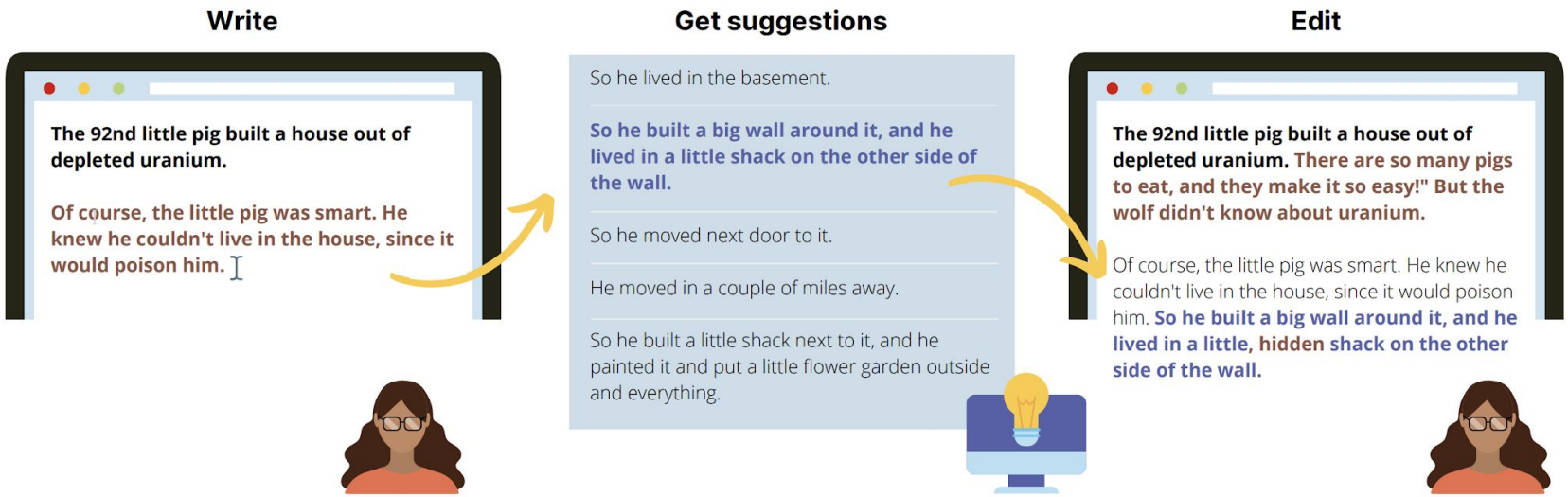
Humans analyze AI,

AI helps human,

& many other forms

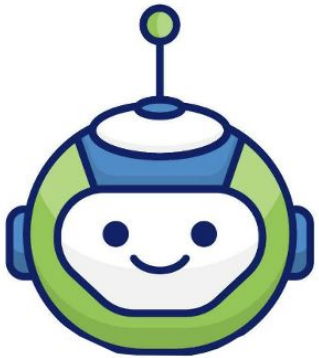
Human-AI Collaboration

Humans (usually non-experts) and AI systems working together in a coordinated way to solve complex problems or reach a goal.



Lee, Mina, Percy Liang, and Qian Yang. "Coauthor: Designing a human-ai collaborative writing dataset for exploring language model capabilities." CHI 2022

What we desire in human-AI collaborations



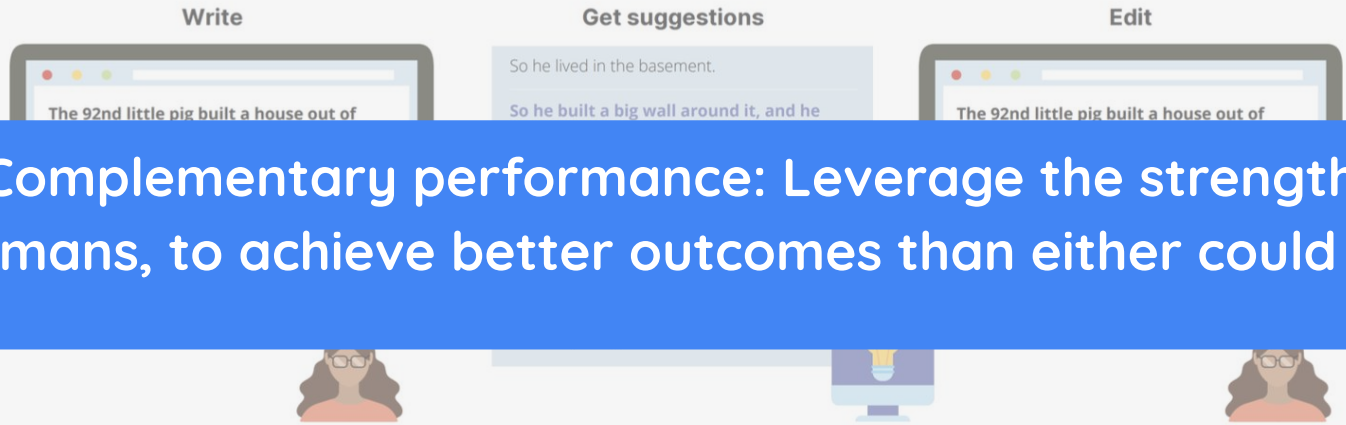
Model: good at quick generation of text based on local context

- Suggest next sentences, help write faster & overcome writer's block

Human: good at logical reasoning and consistency in long doc, know what they want

- Lead the writing, edit the model suggestions.

What we desire in human-AI collaborations



Lee, Ming, Percy Liang, and Qian

Complementary performance: Leverage the strengths of both AI and humans, to achieve better outcomes than either could accomplish alone.

Human (good at logical reasoning and consistency in long doc, know what they want): Lead the writing, edit the model suggestions.

Model (good at quick generate text many versions of text based on local context): Suggest next sentences, help write faster & overcome writer's block

What is being evaluated?

Effectiveness: **accuracy, completeness, and lack of negative consequences with which users achieve specified goals.**

Efficiency: **resources (e.g., time, cognitive effort) required to achieve the model's or system's goals.**

Satisfaction: **positive attitudes, emotions, and/or comfort resulting from the use**

Tradeoffs! systems relying on entirely automated decision-making may be **more efficient** but can be considered **less trustworthy**

What is being evaluated: An Example

Reference Code Snippet

```
def even_odd_count(num):
  even_count = 0
  odd_count = 0
  for i in str(abs(num)):
      if int(i)%2==0:
          even_count +=1
      else:
          odd_count +=1
  return (even_count, odd_count)
```

Generated Code Snippet

```
def even_odd_count(num):
  even_count = 0
  odd_count = 0
  for i in str(num):
      if int(i) % 2 == 0:
          even_count += 1
      else:
          odd_count += 1
  return even_count, odd_count
```

Functional Metric	Similarity Metric	Human preference
pass = **0**	edit similarity = **0.93**	preference = **0.9**

Figure 1: In the example above (counting even and odd numbers), code suggested by a model fails unit tests but is deemed useful by programmers because adding a short check (*abs* value) fixes the generation.

Dibia, Victor, et al. "Aligning Offline Metrics and Human Judgments of Value for Code Generation Models." ACL 2023

"While correctness captures high-value generations, programmers still rate **code that fails unit tests as valuable if it reduces the overall effort needed to complete a coding task**."

How are we evaluating?

Scope `Quantitative` `Qualitative`

Types `Intrinsic` `Extrinsic`

Quantitative method

Understand the “what”. Precise!

➔ *How many questions did the model answer correctly?*

➔ *Did users complete task (yes/no)?*

➔ *How long did it take?*

Qualitative method

Understand the “why”. Open-ended!

➔ *What are some reasons that the model answered those questions incorrectly?*

➔ *What did you like best about the experience?*

➔ *Why were you frustrated by the model output?*

How are we evaluating?

Scope `Quantitative` `Qualitative`

Types `Intrinsic` `Extrinsic`

Intrinsic evaluation: How's the model by itself?

Assesses the quality of an NLP model based on specific tasks or **benchmarks** directly related to the model's performance

Extrinsic evaluation: How helpful is the model in downstream tasks?

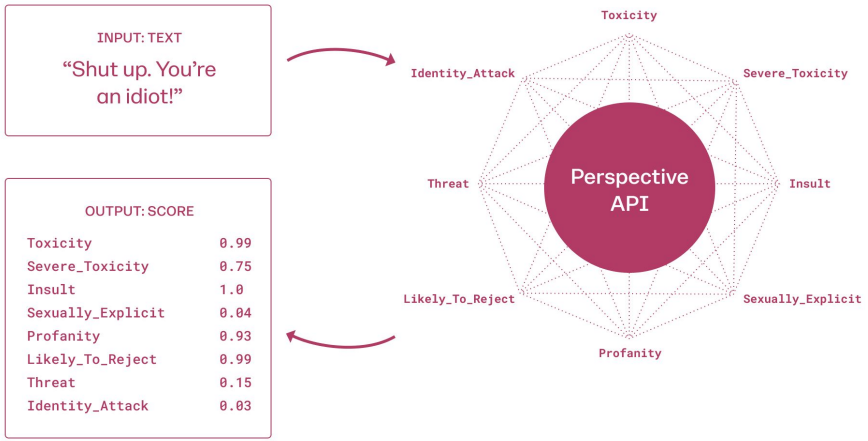
Assesses the performance of an NLP model within the context of a **real-world application** or task

Example of Intrinsic Evaluation: Toxic Classification Accuracy

Benchmarks like: Jigsaw Toxic Comment Classification Challenge, Civil Comments, Perspective API datasets

Quantitative metric:

- AUC (Area Under ROC Curve)
- F1 Score
- Precision / Recall

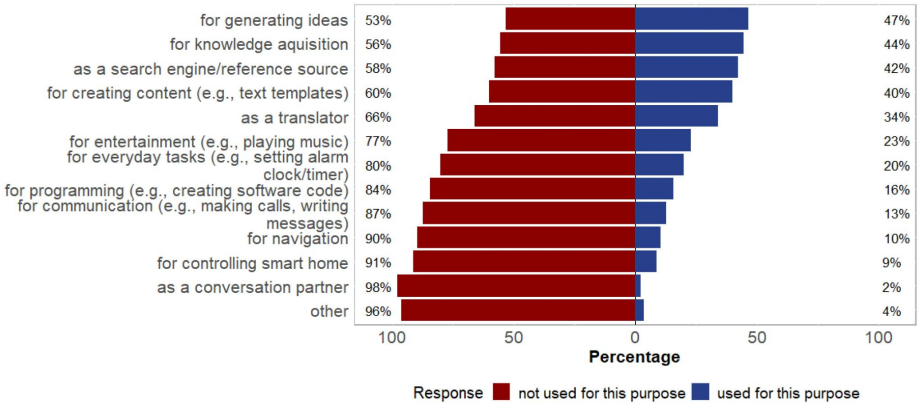


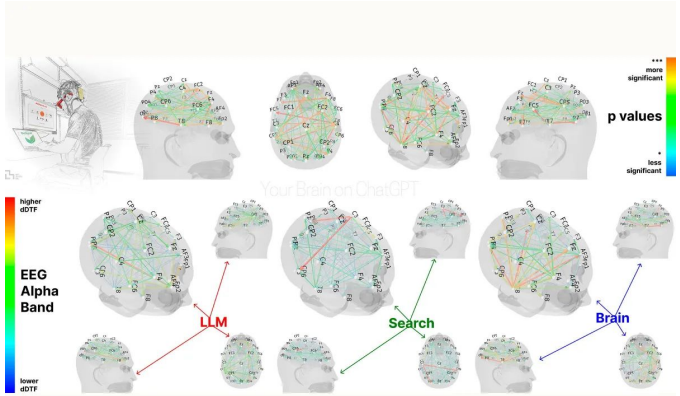
https://developers.perspectiveapi.com/s/about-the-api?language=en_US

Example of Extrinsic Evaluation: Evaluating LMs as Writing Assistants

Researchers ran a controlled user study to test how different LLMs affect real users' writing outcomes.

Task success: Did participants produce a higher-quality final text?
Time to completion: Did the LLM reduce the time needed to revise a paragraph?
Perceived usefulness: Users rating whether the AI made the task easier, clearer, or faster.





https://www.mdpi.com/2673-2688/5/3/65

https://time.com/7295195/ai-chatgpt-google-learning-school/