CS 329X: Human Centered LLMs

Evaluate Human-AI Interaction

Diyi Yang

Announcements

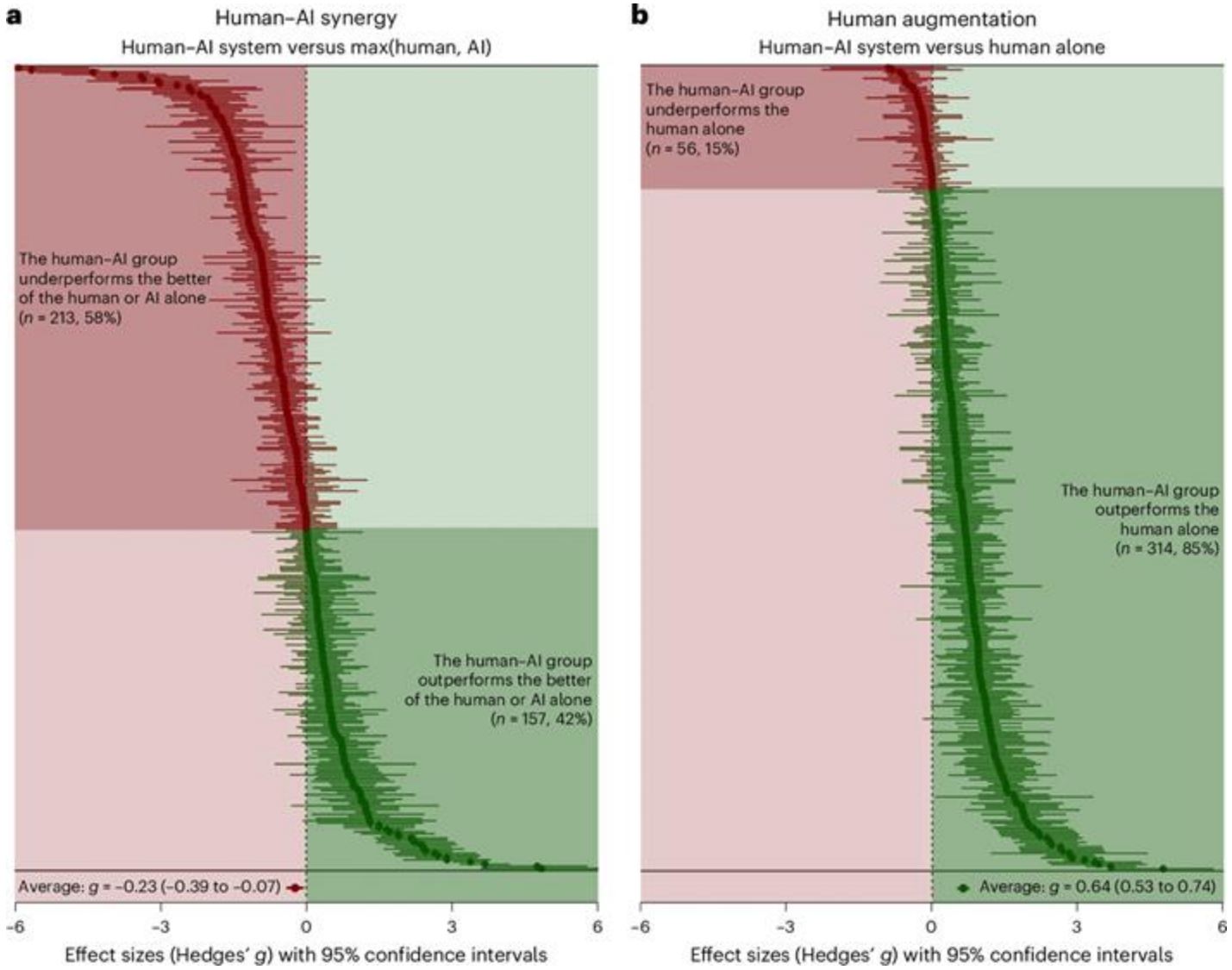
- Schedule change:
  - Oct 23, Thursday: Guest Lecture from Eric Zelikman
  - Oct 28, Tuesday: Midway Project Showcase
- Midway Project Showcase
  - Format discussion

Outline

- **How, What, Who and When** (35 mins)
- **Break** (5 mins)
- **Rethink Evaluation** (20 mins)
- **Small-Group Discussion** (20 mins)

Learning Objective:
understand different factors around how to evaluate human-AI interactions

When combinations of humans and AI are useful

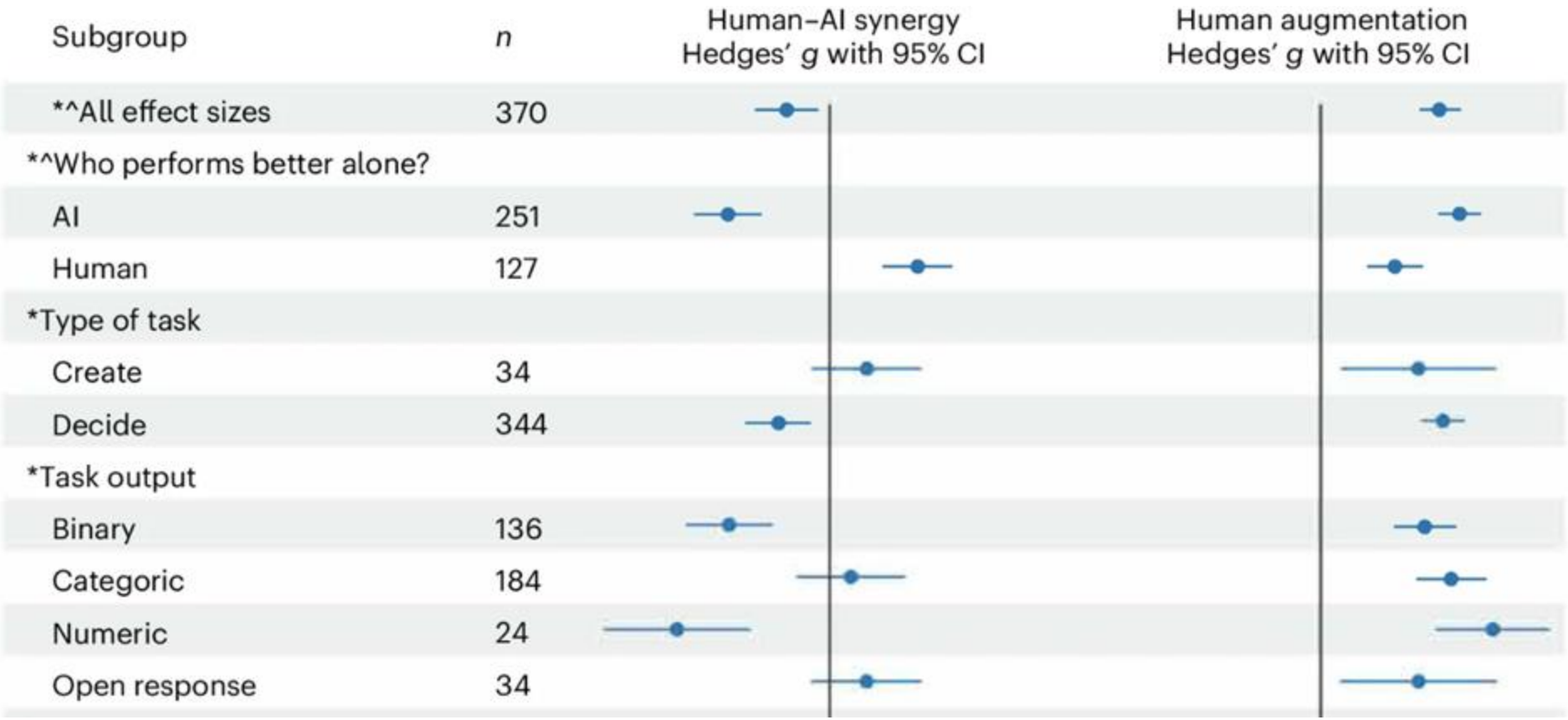


On average, human–AI combinations performed significantly worse than the best of humans or AI alone

Vaccaro, Michelle, Abdullah Almaatouq, and Thomas Malone. "When combinations of humans and AI are useful: A systematic review and meta-analysis." *Nature Human Behaviour* 8, no. 12 (2024): 2293-2303.

When combinations of humans and AI are useful

Performance losses in tasks that involved making decisions and significantly greater gains in tasks that involved creating content



When humans outperformed AI alone, we found performance gains in the combination, but when AI outperformed humans alone, we found losses

Let’s think about some dimensions…

How are we evaluating?

Methods Quant. Qual. *Types* Intrinsic Extrinsic *Metric* Validated New

What is being evaluated?

Who is evaluating?

When do we evaluate?

How are we evaluating?

Methods Quant. Qual. *Types* Intrinsic Extrinsic *Metric* Validated New

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?*

Example of Quantitative Evaluation

```python
def solution(lst):
    """Given a non-empty list of integers, return the sum of all of the odd elements
    that are in even positions.

    Examples
    solution([5, 8, 7, 1]) ==>12
    solution([3, 3, 3, 3, 3]) ==>9
    solution([30, 13, 24, 321]) ==>0
    """
    return sum(lst[i] for i in range(0,len(lst)) if i % 2 == 0 and lst[i] % 2 == 1)
```

```python
def encode_cyclic(s: str):
    """
    returns encoded string by cycling groups of three characters.
    """
    # split string to groups. Each of length 3.
    groups = [s[(3 * i):min((3 * i + 3), len(s))] for i in range((len(s) + 2) // 3)]
    # cycle elements in each group. Unless group has fewer elements than 3.
    groups = [(group[1:] + group[0]) if len(group) == 3 else group for group in groups]
    return "".join(groups)

def decode_cyclic(s: str):
    """
    takes as input string encoded with encode_cyclic function. Returns decoded string.
    """
    # split string to groups. Each of length 3.
    groups = [s[(3 * i):min((3 * i + 3), len(s))] for i in range((len(s) + 2) // 3)]
    # cycle elements in each group.
    groups = [(group[-1] + group[:-1]) if len(group) == 3 else group for group in groups]
    return "".join(groups)
```

Quantitative metric: Pass@k for functional correctness to measure how well a model can complete entire coding tasks.

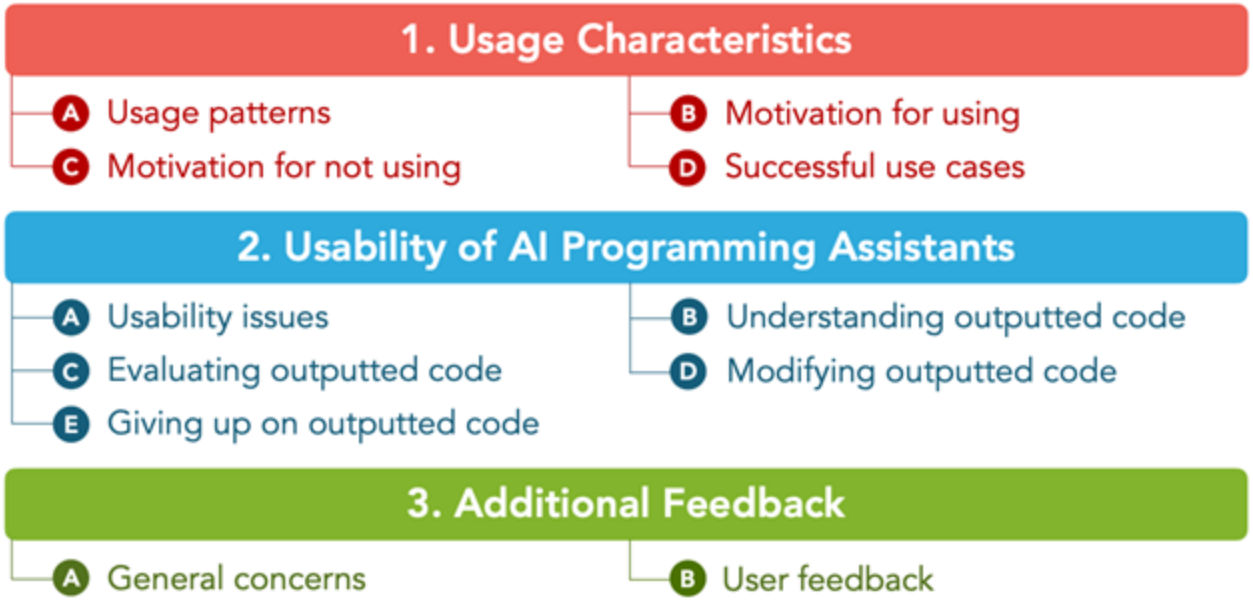
```python
def pass_at_k(n, c, k):
    """
    :param n: total number of samples
    :param c: number of correct samples
    :param k: k in pass@$k$
    """
    if n - c < k: return 1.0
    return 1.0 - np.prod(1.0 - k /
        np.arange(n - c + 1, n + 1))
```

Figure 3. A numerically stable script for calculating an unbiased estimate of pass@k.

Chen, Mark, et al. "Evaluating large language models trained on code." arXiv 2021

Example of Qualitative Study

"a survey to a large population of developers and received responses from a diverse set of 410 developers."



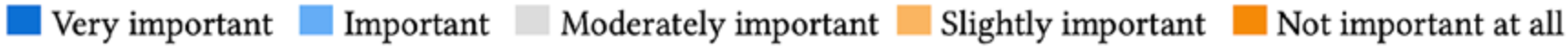
Survey Questions

- For this software project, estimate what percent of your code is written with the help of the following code generation tools.
- For each of the following reasons why you use code generation tools in this software project, rank its importance.
- For each of the following reasons why you do not use code generation tools, rank its importance.
- For your software project, estimate how often you experience the following scenarios when using code generation tools.
- For your software project, estimate how often the following reasons are why you find yourself giving up on code created by code generation tools.

★ What types of feedback would you like to give to code generation tools to make its suggestions better? Why?

Liang, Jenny T., Chenyang Yang, and Brad A. Myers. "A large-scale survey on the usability of ai programming assistants: Successes and challenges." *ICSE 2024*

	Motivation	Distribution
	A. For using	
M1	To have an autocomplete or reduce the amount of keystrokes I make.	86% 6.2%
M2	To finish my programming tasks faster.	76% 12%
M3	To skip needing to go online to find specific code snippets, programming syntax, or API calls I'm aware of, but can't remember.	68% 14%
M4	To discover potential ways or starting points to write a solution to a problem I'm facing.	50% 24%
M5	To find an edge case for my code I haven't considered.	36% 44%
	B. For not using	
M6	Code generation tools write code that doesn't meet functional or non-functional (e.g., security, performance) requirements that I need.	54% 34%
M7	It's hard to control code generation tools to get code that I want.	48% 36%
M8	I spend too much time debugging or modifying code written by code generation tools.	38% 45%
M9	I don't think code generation tools provide helpful suggestions.	34% 46%
M10	I don't want to use a tool that has access to my code.	30% 51%
M11	I write and use proprietary code that code generation tools haven't seen before and don't generate.	28% 59%
M12	To prevent potential intellectual property infringement.	26% 66%
M13	I find the tool's suggestions too distracting.	26% 51%
M14	I don't understand the code written by code generation tools.	16% 76%
M15	I don't want to use open-source code.	10% 89%



How are we evaluating?

Methods Quant. Qual. *Types* Intrinsic Extrinsic *Metric* Validated New

	Quantitative	**Qualitative**
Definition	Gather numerical data to be analyzed using statistical methods	Gathering descriptive, non-numerical data to be analyzed through interpretation and contextualization
Data source	surveys, questionnaires, experiments	interviews, observations, and document analysis
Presentation	tables, graphs, and statistics	quotes and narratives that reflect the participants' experiences and perspectives
Goal	establish cause-and-effect relationships between variables	gain a deeper understanding of social phenomena, meanings, and processes

How are we evaluating?

Methods Quant. Qual. *Types* Intrinsic Extrinsic *Metric* Validated New

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

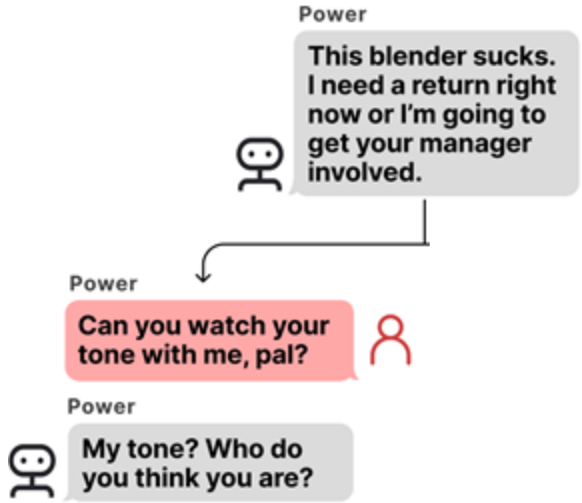
Extrinsic evaluation: How helpful is the model in downstream tasks?

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

Rehearsal: what if we *strategically* simulated conflict resolution practice with LLMs?

Shaikh, Omar, Valentino Chai, Michele J. Gelfand, Diyi Yang, and Michael S. Bernstein. "Rehearsal: Simulating Conflict to Teach Conflict Resolution." SIGCHI 2024

Rehearsal



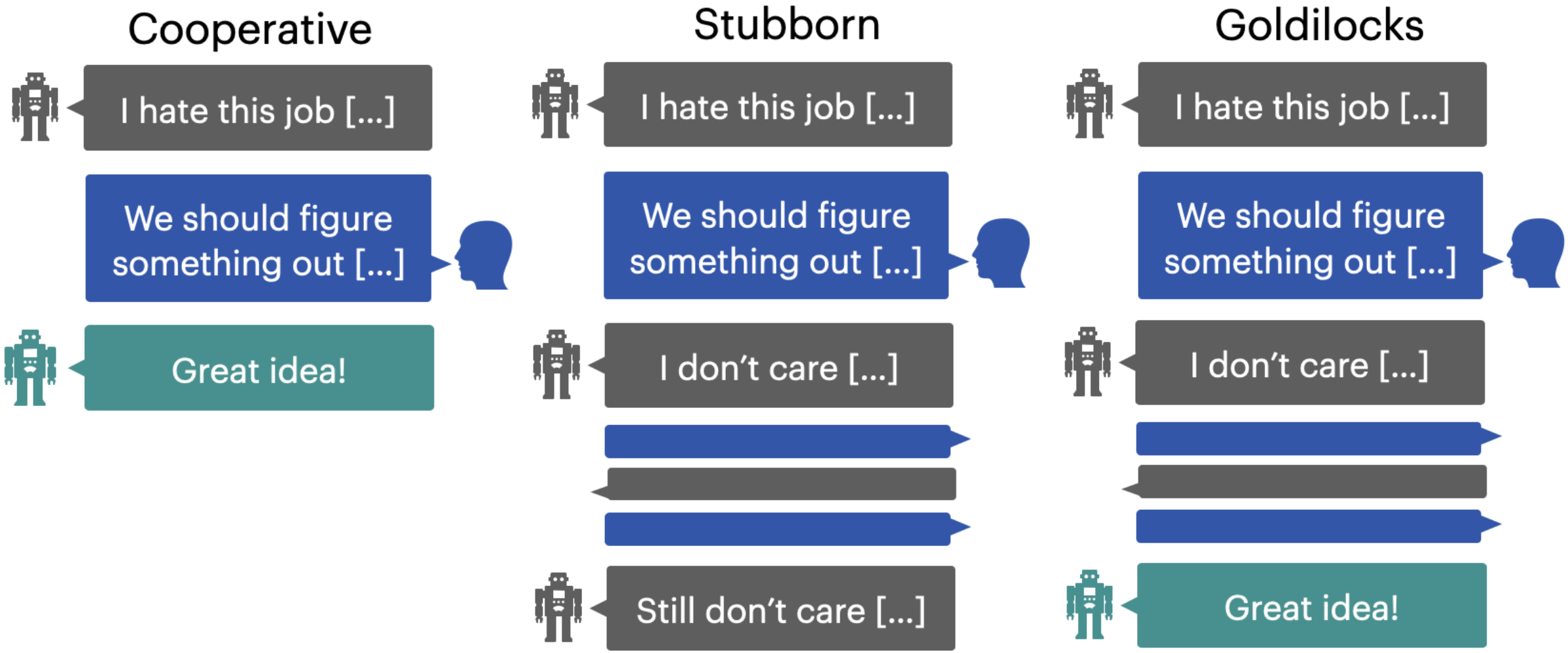
- ✓ **simulates** realistic conflict
- ✓ allows people to **explore counterfactuals**
- ✓ **teaches** people conflict resolution through deliberate practice

Simulating Conflicts with Angry People

Prompt

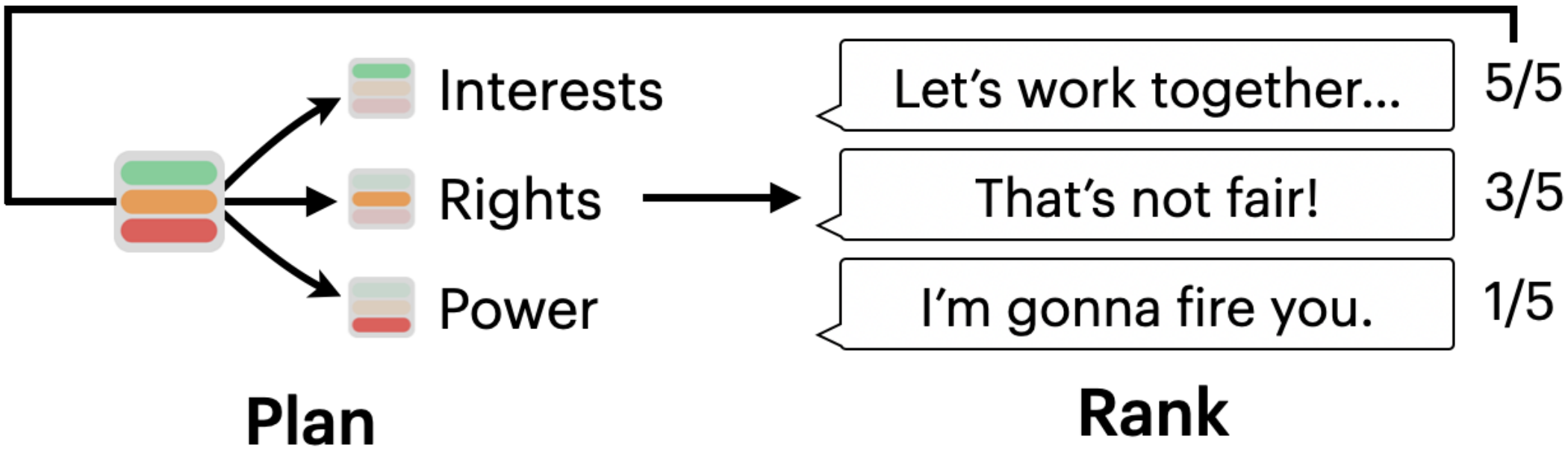
You are an angry person. 

Simulating Conflicts with Angry People



Interests-Rights-Power Prompting

A simple, controlled prompting technique that generates conflict grounded in expert theory and lets users control counterfactual messages

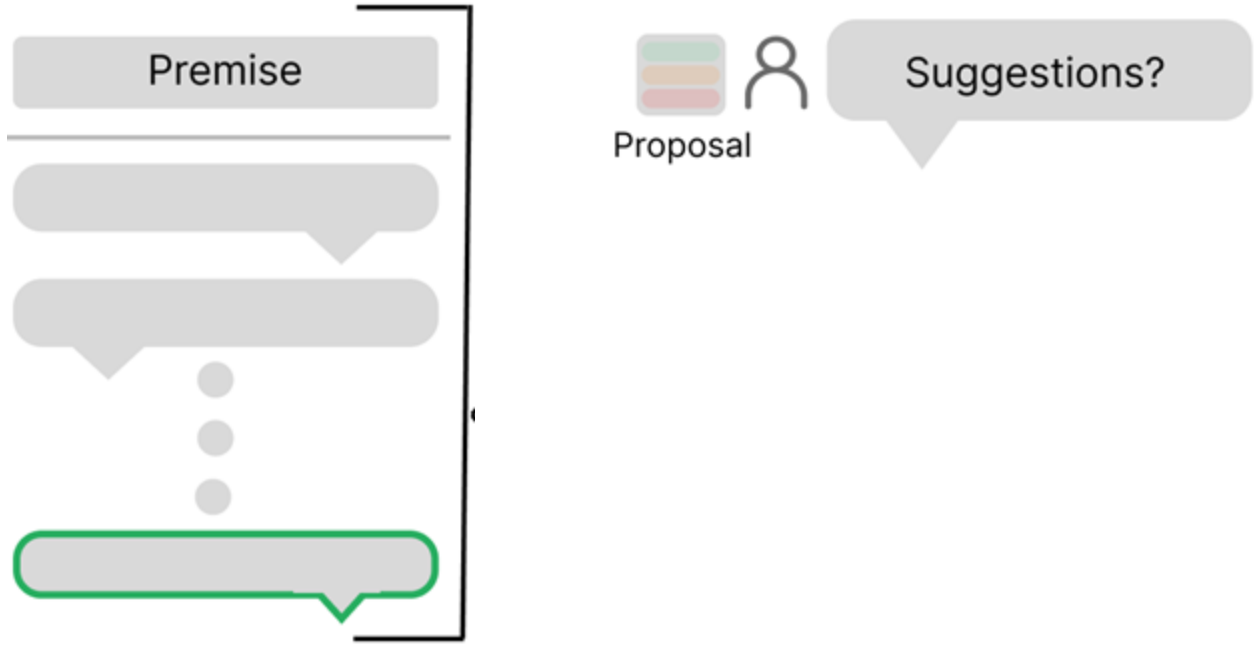


Contextualization

The conversation history and premise are first encoded in a prompt

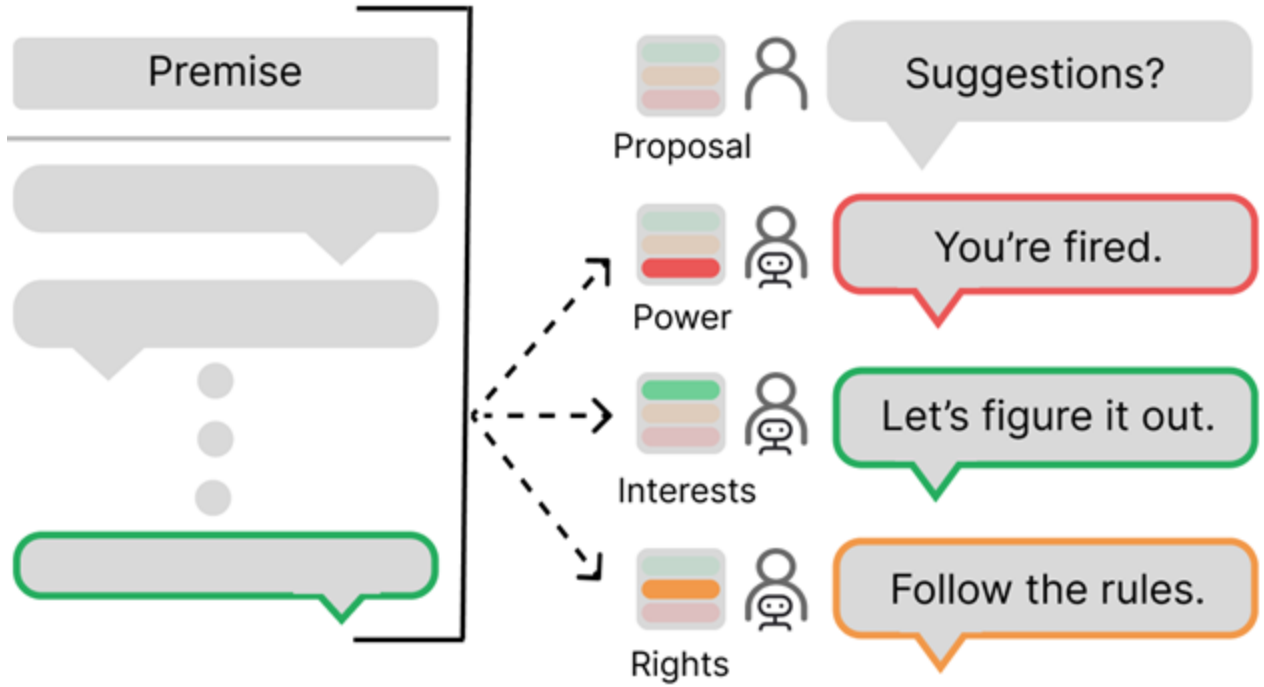
Counterfactual Input Generation

Alongside the user input, counterfactual inputs are generated



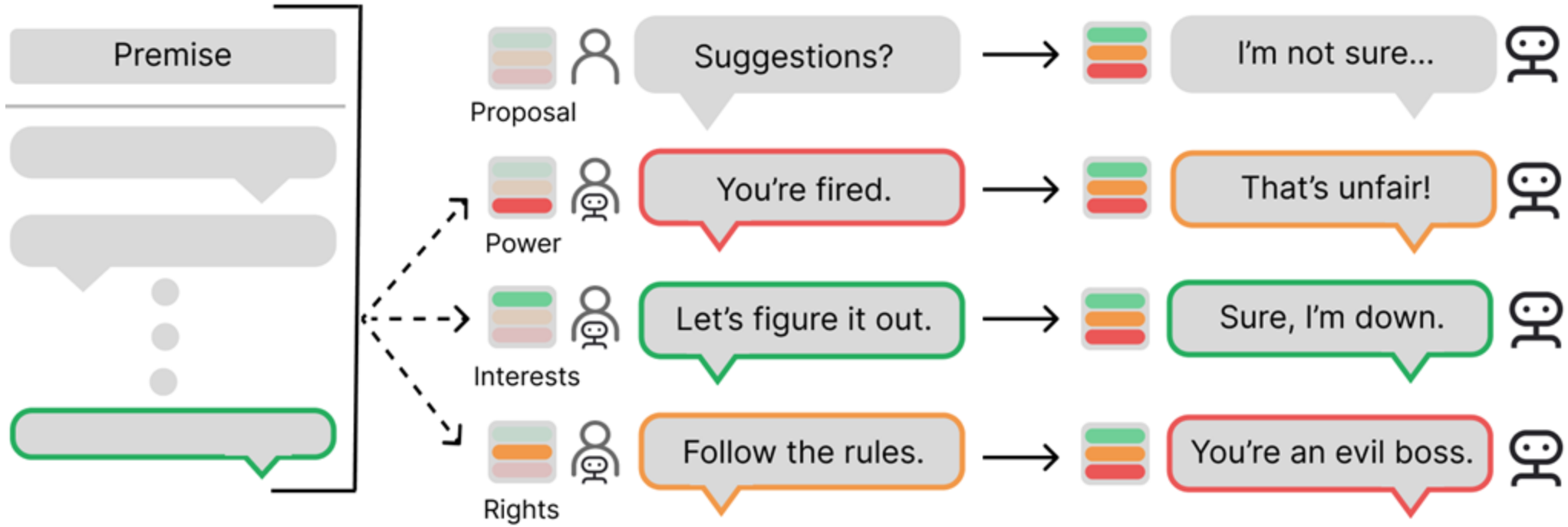
Counterfactual Input Generation

Alongside the user input, counterfactual inputs are generated



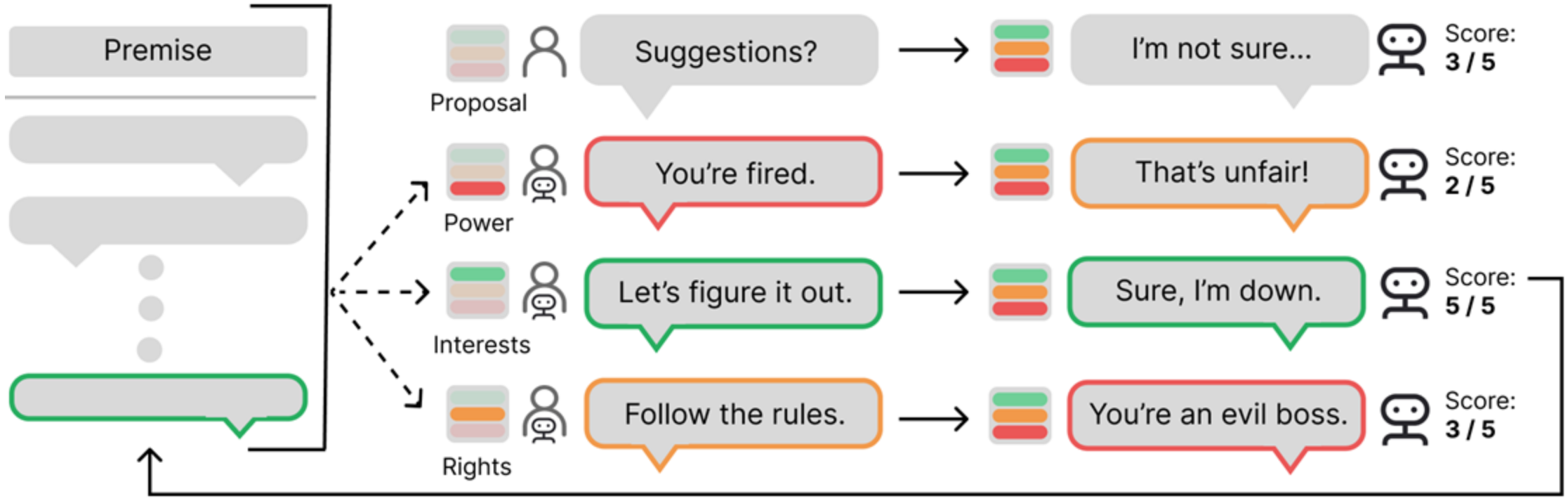
Response Generating & Scoring

Responses for the input messages & counterfactuals are planned, generated and scored on their ability to improve the conflict.

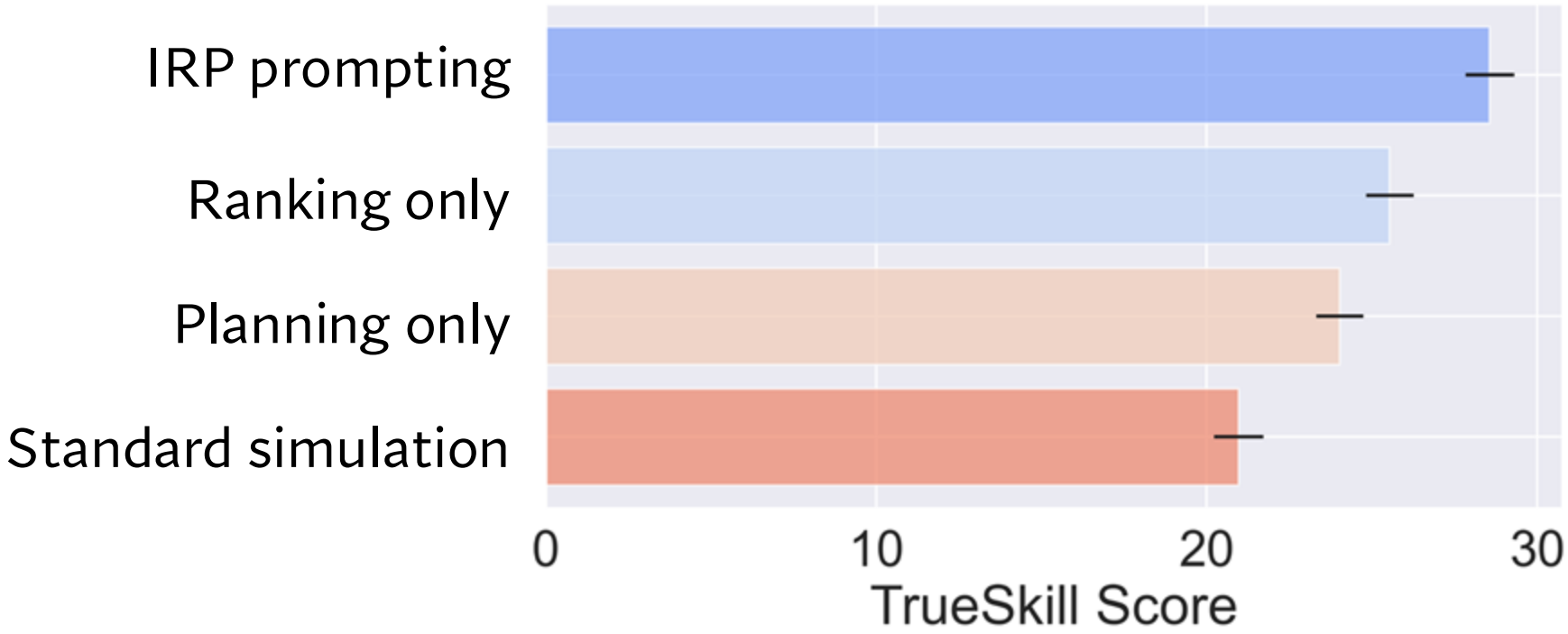


Response Generating & Scoring

The final selection + score are added back to the context.



Evaluating Ecological Validity



User Study: Evaluating Rehearsal

Static Training Material
Video and handbook summarizing conflict resolution strategies

N = 40

23 page handbook

User Study: Evaluating Rehearsal

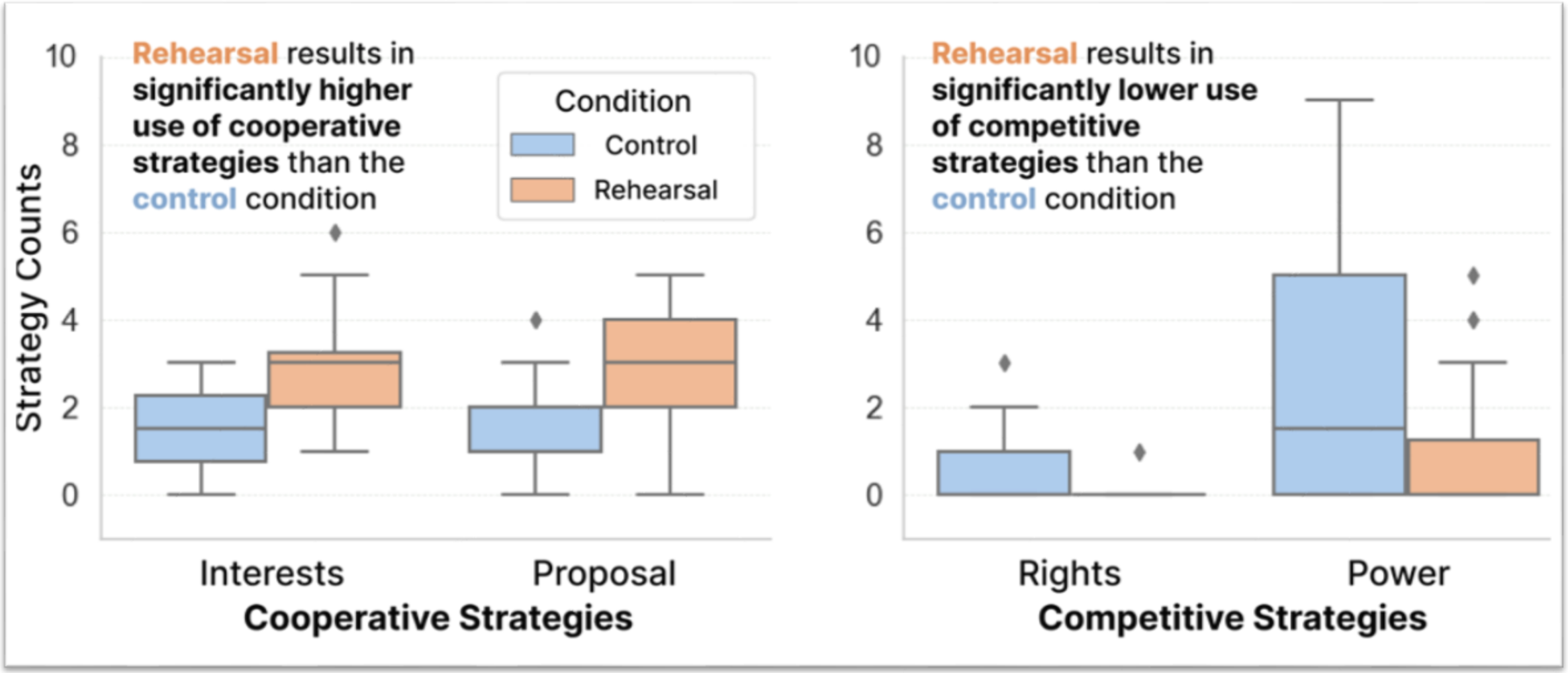
Static Training Material
Video and handbook summarizing conflict resolution strategies

N = 40

Practice
25 mins

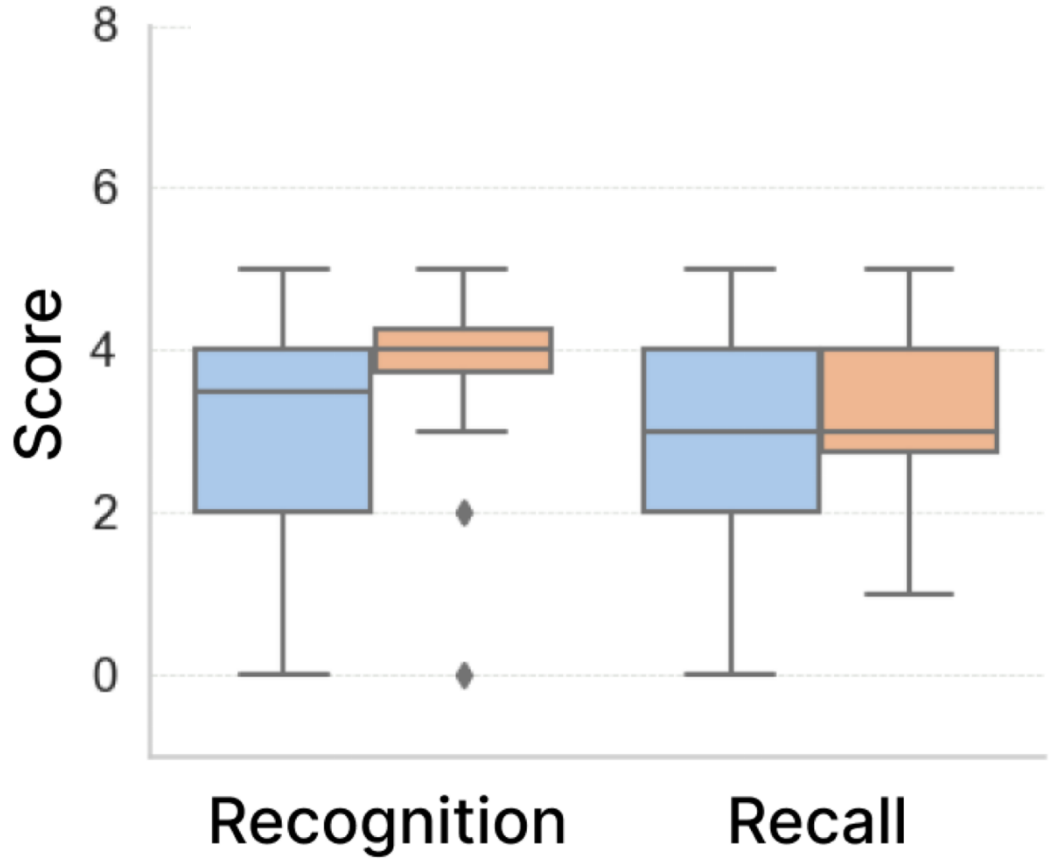
Control (static material only)
N = 20

User Study: Actual Conflict



User Study: Knowledge of Conflict Resolution

Rehearsal **does not** significantly improve knowledge skills compared to the control condition



Why Are We Seeing This?

While both static training and Rehearsal improve "**book smarts**," only Rehearsal improves "**street smarts**" of conflict resolution

"It's easy to read about conflict-resolution strategies, but a lot harder to implement and stick to."

"while I did pay attention to the training, it was really hard to imagine implementing [the conflict resolution strategies]

Aside: Self Efficacy

What do you think happened re: self-efficacy?

Conflict self-efficacy scores dropped **across the board.**

Significantly larger decreases for Rehearsal condition.

Probably Dunning Kruger effect

Intrinsic Evaluations of LLM-based Interaction

Metric Type	Description
Reference based	Metrics of the similarity btw system output and gold standards
Topic analysis	Assessment of relevance of topics over expectations
Classifier score	Using trained classifiers to categorize known effective behaviors
LLM judge	Prompting LLMs to act as judge to provide Likert scale scores
Human ranking	Comparative metric where systems are ranked based on rubrics
Human scoring	Likert scale ratings along given rubrics
Suggestion usage	Rate at which users utilize suggestions provided by AI
Recommendation	Rate how likely a user would be to recommend the system to others

Extrinsic Evaluations of LLM-based Interaction

Metric Type	Description
Behavioral Impacts	Changes in qualitatively coded participant behaviors before and after exposure to the system
Self-Efficacy Reports	Changes in participants' self-reported efficacy before and after exposure
Standardized Evaluation	Changes in participant scores on closed-ended assessments of knowledge
Short-term Outcomes	Short-term impacts of AI assistance on participants' behaviors, attitudes, trust
Long-term Outcomes	Long-term impacts of AI assistance

How are we evaluating?

Methods Quant. Qual. *Types* Intrinsic Extrinsic *Metric* Validated New

	Intrinsic	**Extrinsic**
Pros	**Quick Feedback** **Easy comparison** and **progress tracking** on widely accepted benchmarking **Focused Metrics** reflecting model capability (e.g., accuracy, BLEU) **Controlled Environment**	**Holistic assessment**, considers its interaction with other components or systems. **Generalizability**, reflects how the NLP model impacts **complex, real-world applications**, providing a more accurate assessment of its **practical** value. Aligns with the **end-users' perspective** by focusing on the application's overall success rather than individual tasks.
Cons	**Limited scope** on artificial tasks. **Not always predictive** of the model's performance in real-world applications, providing limited generalizability.	Complexity: can be more **resource-intensive** and **time-consuming**. Involve human judgment, can be **subjectivity** Challenging to **isolate** the model's contribution from other factors. Heavily depends on the **quality and complexity of the apps**.

Takeaways: How are we evaluating

Evaluation has different methods, types, and metrics. None of them is *the one* – they are complementary to each other and serve different goals.

Usability and utility are especially important. User generated interaction log could be useful here.

Metric is evolving. Besides existing and validated metrics in Human-AI interaction, metrics from human-human interaction can be quite inspiring.

Let’s think about some dimensions...

How are we evaluating?

Methods Quant. | Qual. *Types* Intrinsic | Extrinsic *Metric* Validated | New

What is being evaluated?

Modules Model module | HCI module (UX) | End-to-end *Goal* Utility | Satisfaction | ...

Who is evaluating?

When do we evaluate?

Key desiderata in evaluation

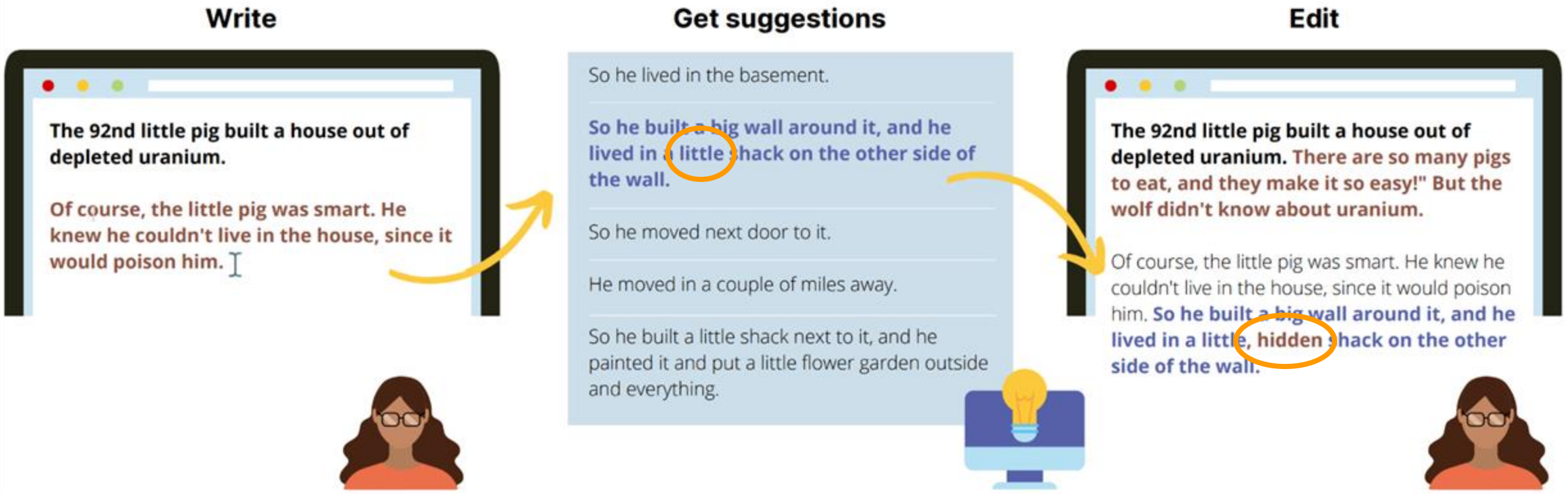
Objectives and goals: *"What do I need to know?"*

Tasks: *"What should users do so I find out what I need to know?"*

Data: *"What data do I collect to find out what I need to know?"*

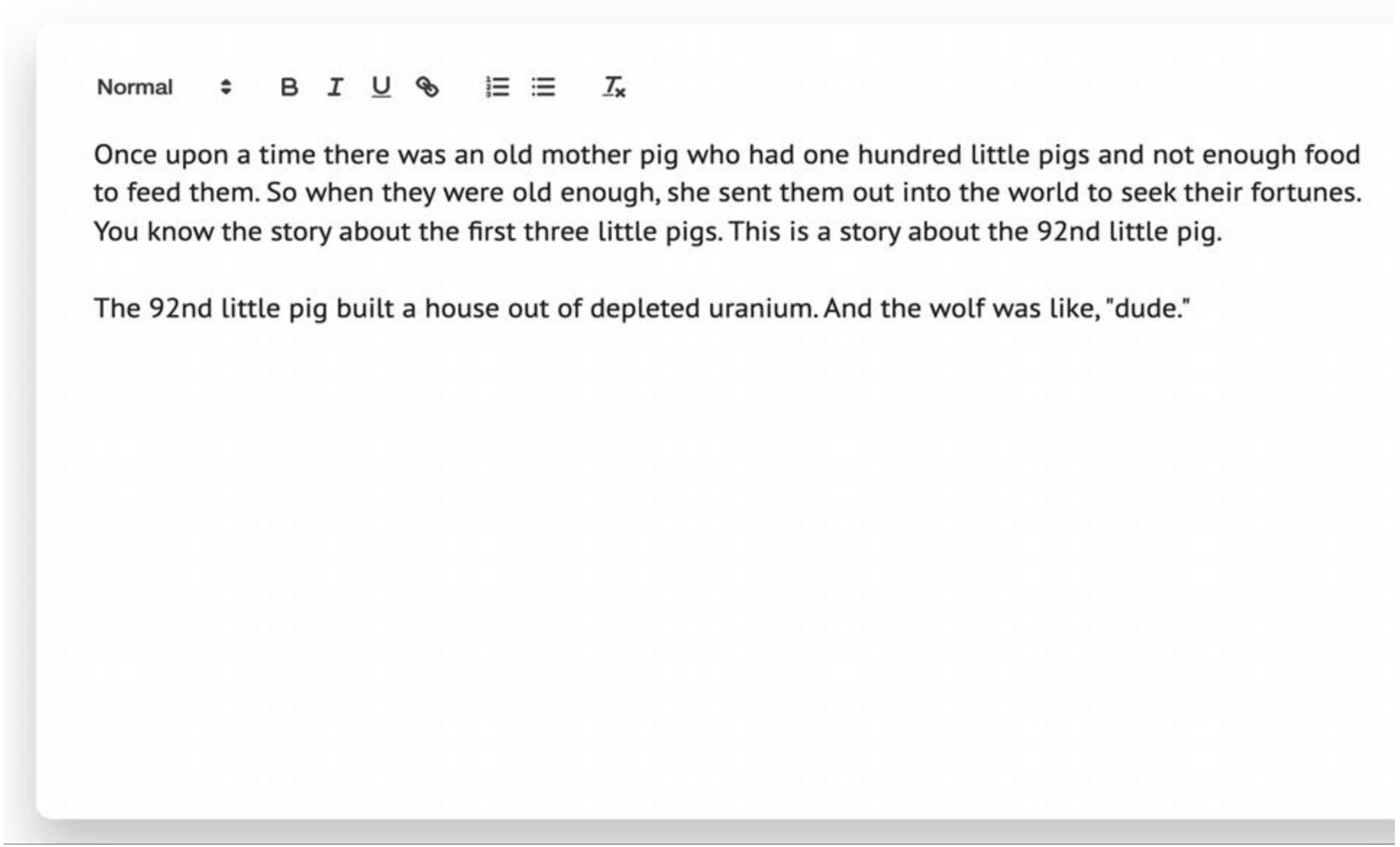
Analysis: *"How do I crunch the numbers to find out what I need to know?"*

Consider a case study: Human-LM Co-Writing



"CoAuthor: Designing a Human-AI Collaborative Writing Dataset for Exploring Language Model Capabilities." CHI 2022

Consider a case study: Human-LM Co-Writing



Going back to our Co-writing case...

➡️ **Task –** Vary types of writing prompts, and model randomness

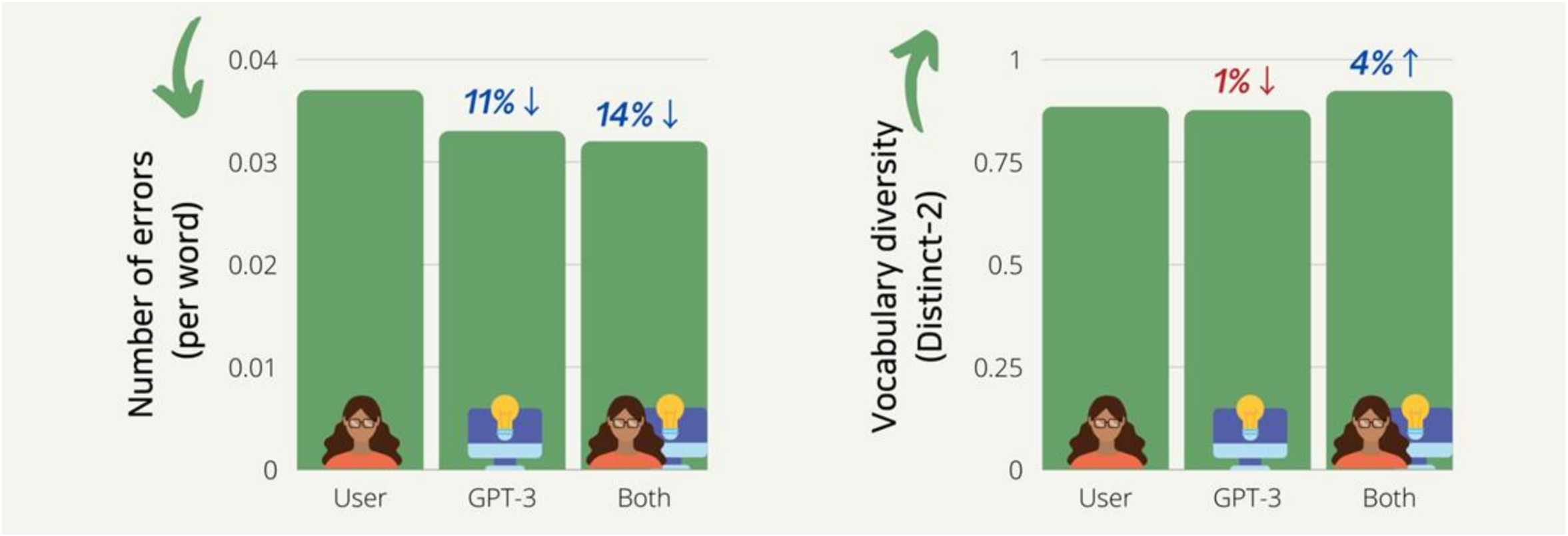
➡️ **Data –** What's collected?

- **1445 sessions between 63 users and GPT-3**
- **Types of writing**:
  Creative writing: 830 stories written by 58 writers
  Argumentative writing: 615 essays written by 49 writers
- **Stories and essays**: 418 words long
- **Number of queries**: 11.8 queries per writing session
- **Acceptance rate of suggestions**: 72.3%
- **Percentage of text written by humans**: 72.6%

Existing metrics can help quantify effects

Q: Can GPT-3 generate fluent text in response to user text?

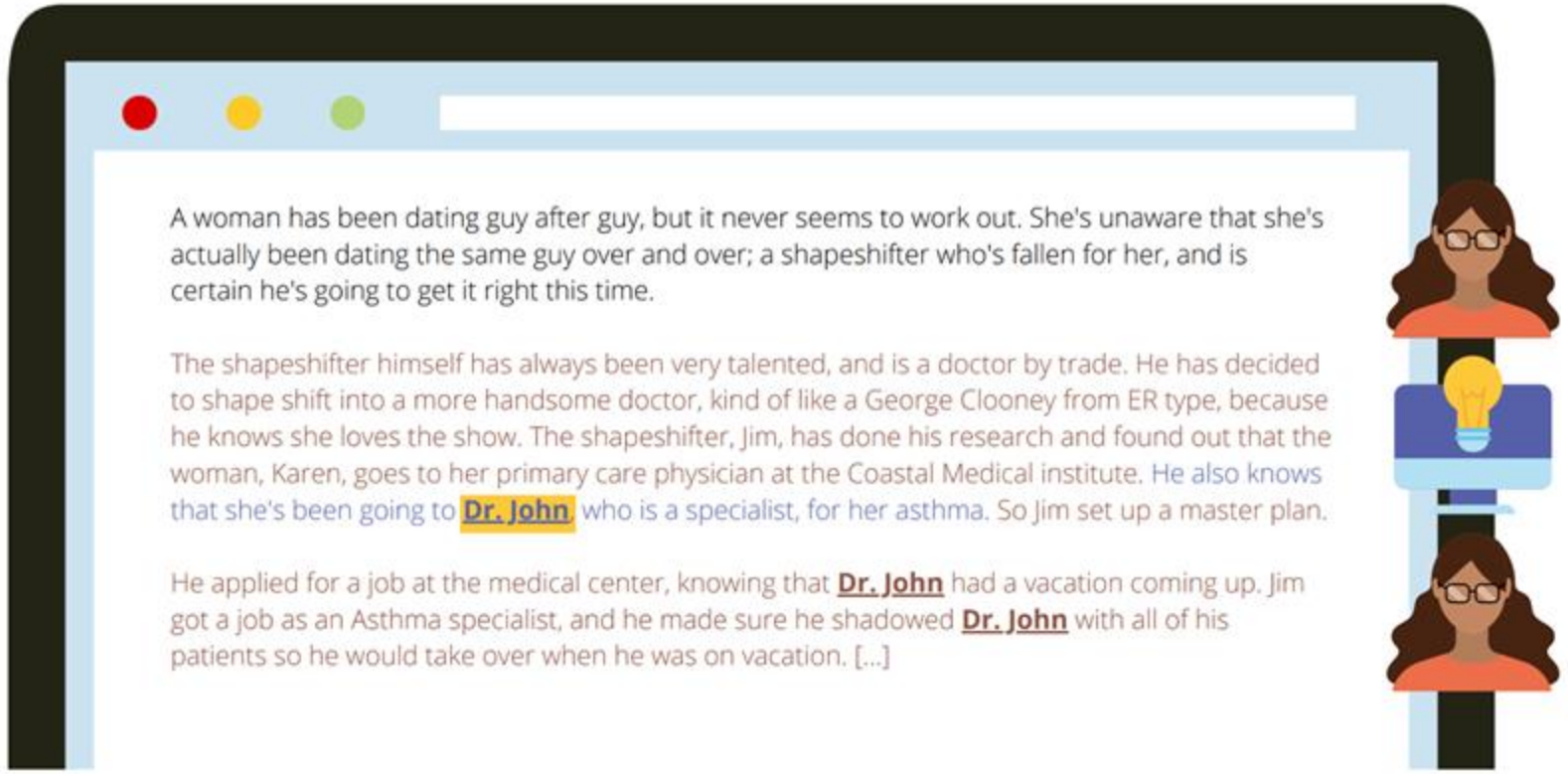
A: Text written by user + GPT-3 had fewest errors and most diverse vocabulary

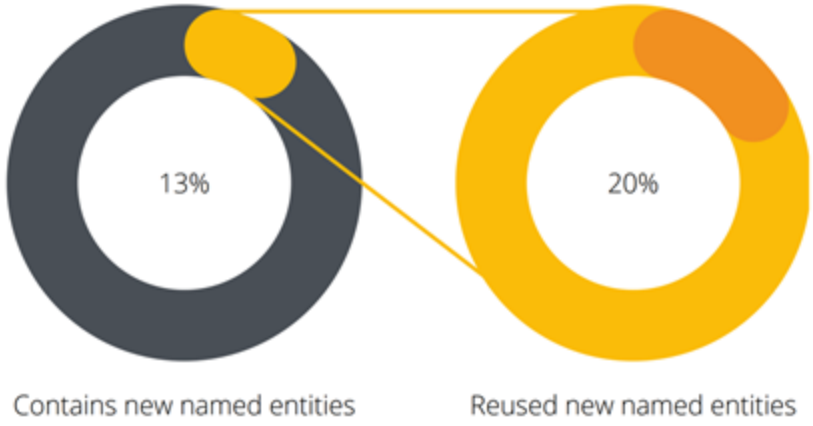


Define metrics for what needs to be measured

Q: Can GPT-3 contribute new ideas to users' stories?

A: Ideas generated and reused by users in the subsequent writing.





"Reused named entity" is a new (lower-bound)metric defined for ideation quantification!

How the dataset can be further used

Writers' behaviors and writing outcomes over time

Can we observe novelty effect and longitudinal change?

Do LMs homogenize writing by providing similar suggestions to all users?

Linguistic accommodation

How does the style, voice, or tone of a writer or LM influence that of the other over time? Is the influence uni-directional or bi-directional?

Edit traces

What can we learn from human edits on LM outputs?

Can we train LMs on edit traces to emulate human edits?

What are we evaluating

Models for specific use cases should be **blended into existing N2N workflows**. Test-in-product is not the most ideal but usually useful.

End-to-end evaluation goes beyond models. Metrics might focus on usability (e.g. discoverability); And little things like latency in suggestion can easily change end-to-end quality.

UI is a BIG module. Its iteration is basically ablation study for usability.
We need to consider **all users** touching the system, the **original objective** of the task (education and training!), and **worst case scenarios**!

Let’s think about some dimensions…

How are we evaluating?

Methods Quant. | Qual. *Types* Intrinsic | Extrinsic *Metric* Validated | New

What is being evaluated?

Modules Model module | HCI module (UX) | End-to-end *Goal* Utility | Satisfaction | …

Who is evaluating?

Humans Lay users | Domain experts *Automated* LLM

When do we evaluate?

Who is evaluating?

Systems are usually not designed for everyone. **Human factors** impact evaluation: Demographic background, culture, expertise, etc.

Targeted group: Who you designed the Human-AI collaboration for!

Domain experts: Those who are familiar with a very particular domain

Lay users / non-experts: Those who are general population

Use case specific: Teachers, parents, students, patients, etc.

+ Automation: LLM-simulated users...

The Rise of LLM-as-a-judge and social simulation

***Human* cannot be ignored in Human-AI evaluation**. Different workflows need to be designed for different target applications, and in turn need to be evaluated on the right user group.

Some amount of simulation might be useful, because it turns extrinsic evaluation into intrinsic evaluation. However, it may only be useful for very early stage of sanity check due to lack or representativeness and potential overfit.

Let’s think about some dimensions...

How are we evaluating?

Methods Quant. | Qual. *Types* Intrinsic | Extrinsic *Metric* Validated | New

What is being evaluated?

Modules Model module | HCI module (UX) | End-to-end *Goal* Utility | Satisfaction | ...

Who is evaluating?

Humans Lay users | Domain experts *Automated* LLM

When do we evaluate?

Duration Instant | Short-term | Long-term

When to evaluate?

Duration Instant | Short-term | Long-term

Short-term: Constrained interactive session (e.g. one hour).

Measures immediate benefits: how much faster users write stories with AIs, how much more creative they get with AI, etc.

> "novelty bias": rating something highly just because it seems fancier than the alternative.

Long-term: Much longer, e.g. one week study with daily tasks.

Measures long-lasting impact: the practical feasibility of a product in its real use cases, the longer term effects on humans (e.g. their mental model on AIs).

Example: Longitudinal Study on AI Chain

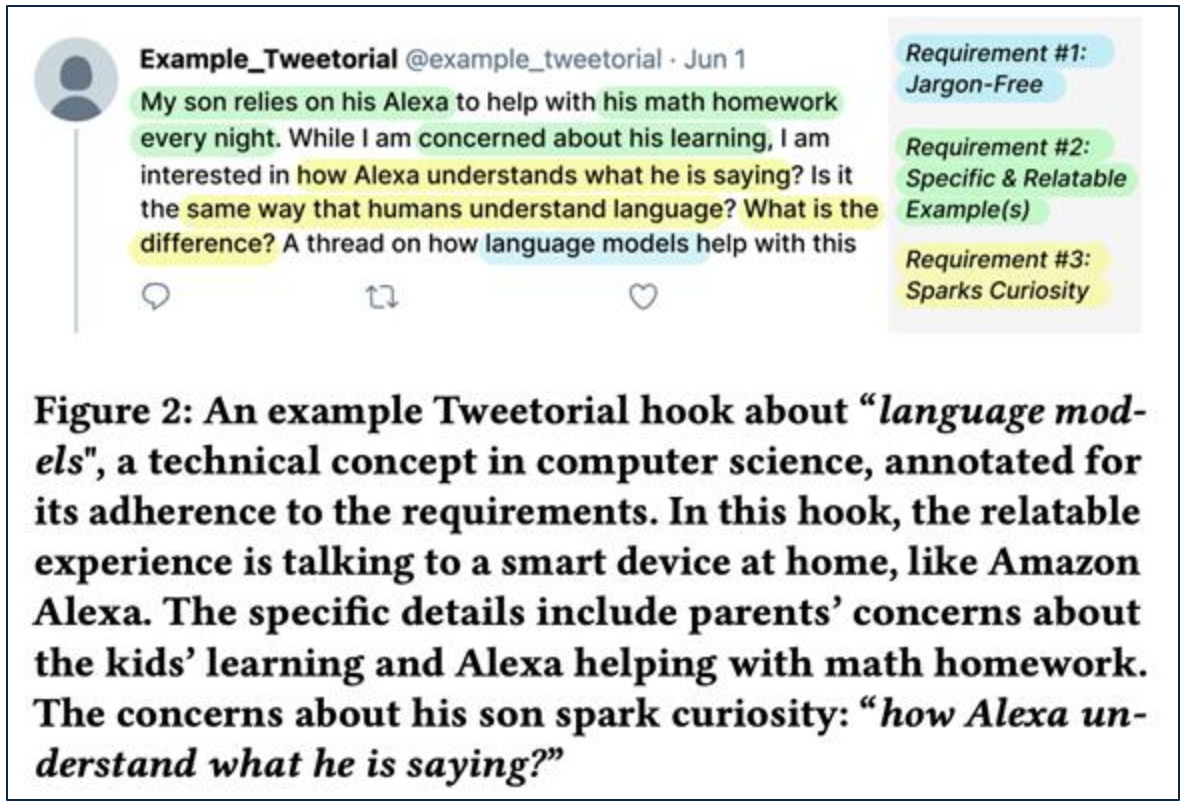
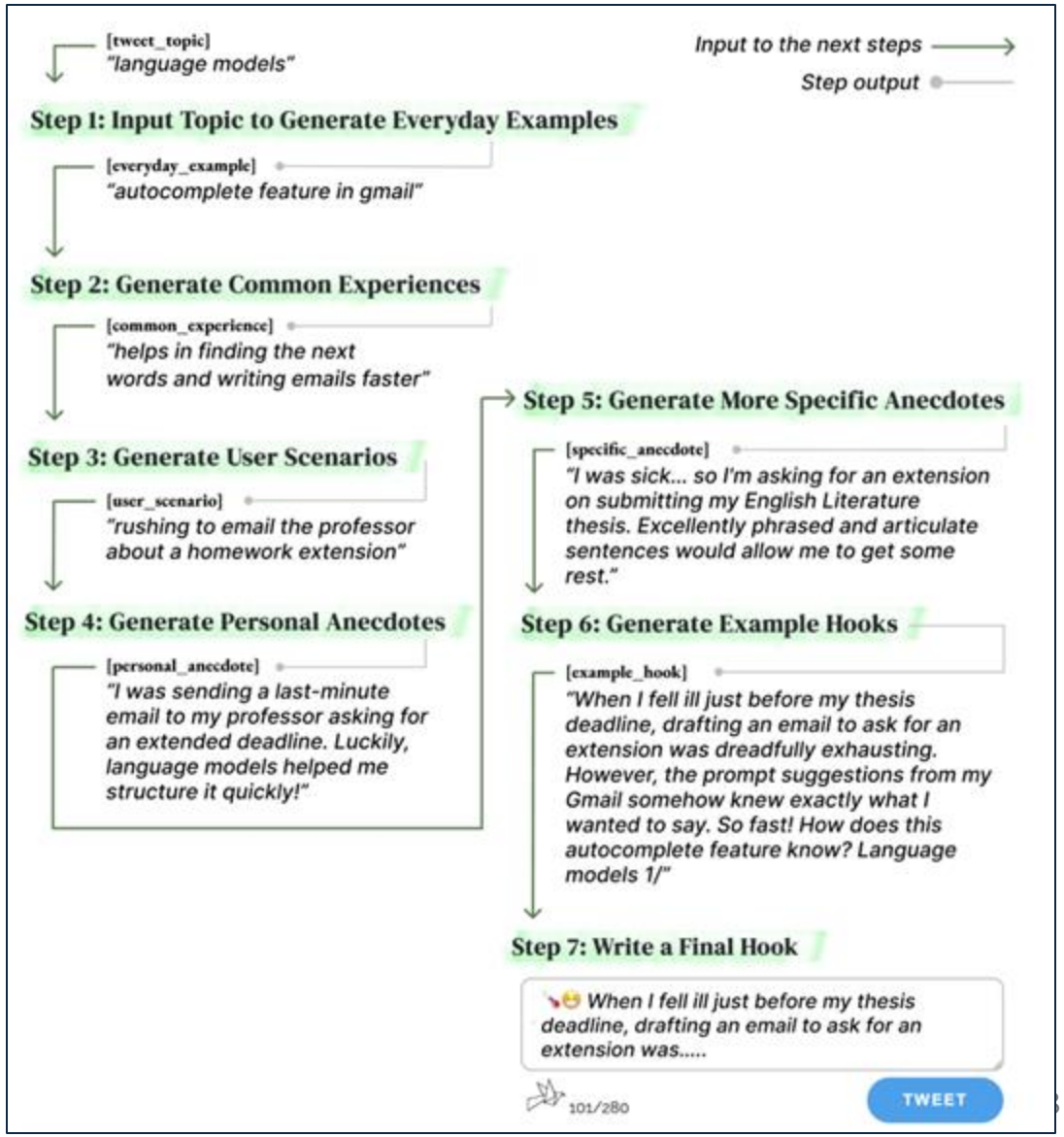


Figure 2: An example Tweetorial hook about "*language models*", a technical concept in computer science, annotated for its adherence to the requirements. In this hook, the relatable experience is talking to a smart device at home, like Amazon Alexa. The specific details include parents' concerns about the kids' learning and Alexa helping with math homework. The concerns about his son spark curiosity: "*how Alexa understand what he is saying?*"



Long, Tao, Katy Ilonka Gero, and Lydia B. Chilton. "Not Just Novelty: A Longitudinal Study on Utility and Customization of AI Workflows." arXiv 2024

Example: Longitudinal Study on AI Chain

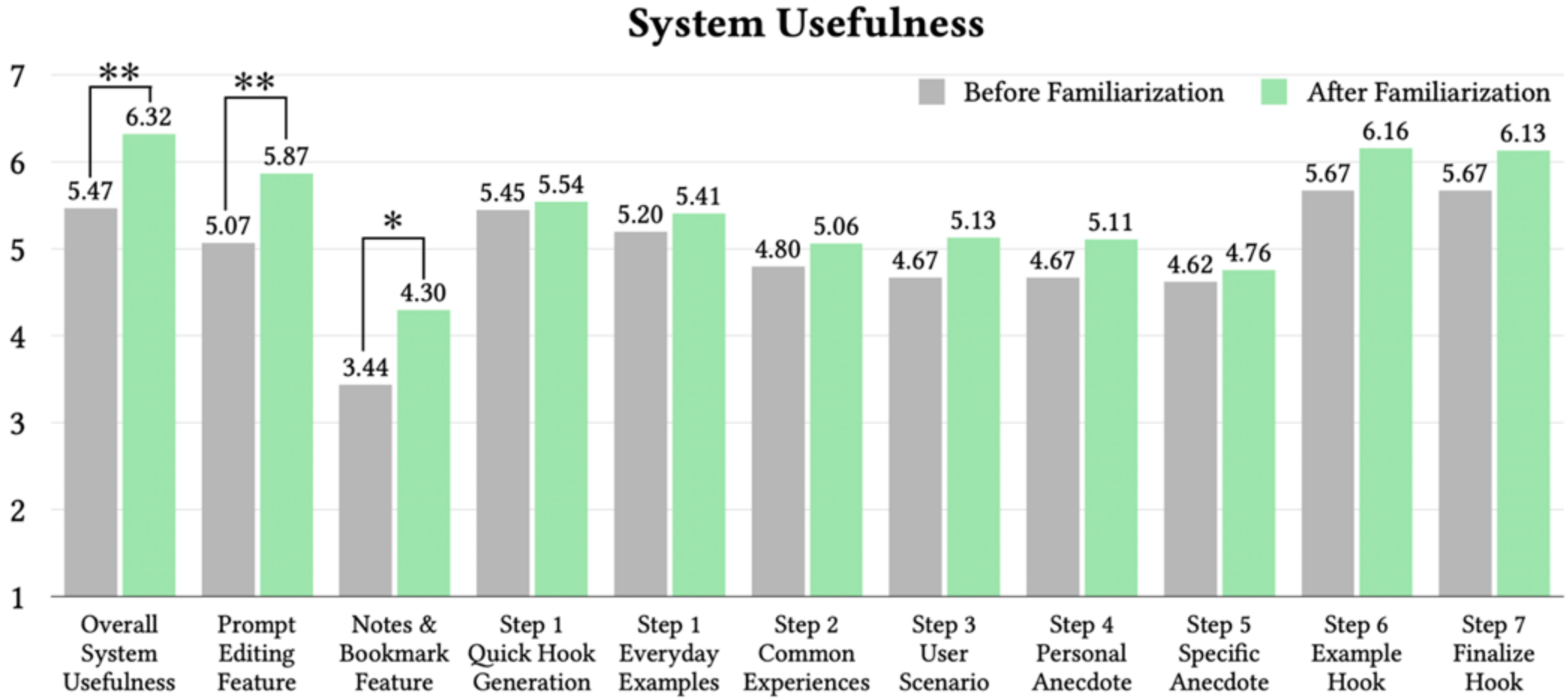


Figure 6: System Usefulness (average scores across 12 users before and after familiarization, ** denotes statistical significance at the p-value < 0.005 level, * denotes statistical significance at the p-value < 0.05 level)

There exists a **familiarization phase**

There exists a **familiarization phase**: users explores the capabilities of the workflow and discovers which aspects are useful. **After this phase**:

- Users understood the workflow & could anticipate the outputs.
- The perceived utility of the system was rated higher — perceived AI utility is **not just a novelty effect**.
- Why? end-users now can customize prompts for their needs

Takeaways: When to evaluate

	Short-term	**Long-term**
Duration	Measured by hours	Measured by days or weeks
What	Immediate benefit	Long-lasting AI practical utility The impact of interactions on humans
What for	quickly iterate on potential Human-AI interaction designs	Learn about... Adaptation and Learning Curve Over Time Impact of AI Evolution on User Interaction Impact of AI Evolution on User Interaction
Not for	Only capture humans as static agents. Affected by “novelty bias”	Quick iteration :)

Let’s think about some dimensions...

How are we evaluating?

Methods Quant. | Qual. *Types* Intrinsic | Extrinsic *Metric* Validated | New

What is being evaluated?

Modules Model module | HCI module (UX) | End-to-end *Goal* Utility | Satisfaction | ...

Who is evaluating?

Humans Lay users | Domain experts *Automated* LLM

When do we evaluate?

Duration Instant | Short-term | Long-term

Ma, Qianou, Dora Zhao, Xinran Zhao, Chenglei Si, Chenyang Yang, Ryan Louie, Ehud Reiter, Diyi Yang, and Tongshuang Wu. "SPHERE: An Evaluation Card for Human-AI Systems." arXiv preprint arXiv:2504.07971 (2025).

Category	Aspect	HCI Examples		NLP Examples
What is being evaluated?				
Component	Model	Accuracy of LLM gen. (Lee et al., 2024b)		BERTScore (Glória-Silva et al., 2024)
	System	Knowledge quiz (Shaikh et al., 2024)		Headline quality (Ding et al., 2023)
Design Goal	Effectiveness	System risk & content (Rajashekar et al., 2024)		Label quality & stability (Wei et al., 2024)
	Efficiency	Perceived workload (Lee et al., 2024b)		Time on task (Ding et al., 2023)
	Satisfaction	Likert-scale rating of fun (Wang et al., 2024c)		Likert-scale rating of trust (Ding et al., 2023)
How is an evaluation conducted?				
Scope	Intrinsic	System Usability Scale (Liu et al., 2024a)		Retrieval hit rate (Inan et al., 2024)
	Extrinsic	Engagement & enjoyment (Fan et al., 2024)		Identified concept diversity (Yang et al., 2023)
Method	Quantitative	Interaction logs (Wu et al., 2022)		Micro F1 (Wei et al., 2024)
	Qualitative	Interview & grounded coding (Liu et al., 2023)		Case study (Cai et al., 2024)
Who is participating in the evaluation?				
Human	Expert	Prolific experts (Zavolokina et al., 2024)		ASL expert (Inan et al., 2024)
	User	Students & physicians (Rajashekar et al., 2024)		Crowdworkers (Chakrabarty et al., 2022)
Automated	Static	Perplexity & LIWC scores (Calle et al., 2024)		Precision & recall (Yang et al., 2023)
	Generative	N/A		Consistency by LLaMa2 (Zhao et al., 2024)
When is evaluation conducted (duration)?				
Time Scale	Immediate	# of clicks (Lawley and Maclellan, 2024)		Benchmark (Raheja et al., 2023)
	Short-term	1-hour usability study (Liu et al., 2024a)		10 minutes per poem (Chakrabarty et al., 2022)
	Long-term	3-days session (Fan et al., 2024)		6-months deployment (Inan et al., 2024)
(Meta) How is evaluation validated?				
Validation	Reliability	Krippendroff's α as IRR (Lee et al., 2024b)		Fleiss' κ for annotation (Zhao et al., 2024)
	Validity	Counterbalancing (Wu et al., 2022)		Randomized control (Ding et al., 2023)

Outline

✓ **How, What, Who and When** (35 mins)

➢ **Ethics and Rethink Evaluation** (20 mins)

Ethical and Legal Considerations

- When designing an experiment involving human participation, it is critical to consider ethical and legal implications

- Critical to understand which review processes or legal requirements exist
  - Institutional review boards
  - Ethics committee
  - Relevant data collection laws

Ethical and Legal Considerations

- What data are actually necessary to collect?
- How the data will be stored and protected?
- How long?
- What type of personal data will be collected?
- Data collection and Anonymization techniques [Siegert et al. (2020); Finck and Pallas (2020)]

The Purpose of Human Evaluation

- **Exploratory research questions**: to generate assumptions, which can then be tested in a subsequent confirmatory research question, e.g., "*Which factors (of the set of measured variables) influence the users' enjoyment of system B?*"

- **Confirmatory research questions**: to test a specific assumption, e.g., "*Does the explanation method of system B increase the users' trust in the system compared to that of system A?*"

Transparency in Human Evaluation

- No standardized approach or consensus for human evaluation
- Different to compare results across different studies due to the variability in evaluation design

Best Practices for Designing Human Evaluation

- How are human ratings collected?
- What questions are asked of raters?
- Who are the raters?
  - Ensure annotator quality
    - How to define and explain the task
    - Attention checks/questions with known answer
    - Annotator agreement
- How do you ensure/measure the quality of the ratings?

Celikyilmaz, Asli, Elizabeth Clark, and Jianfeng Gao. "Evaluation of text generation: A survey." arXiv preprint arXiv:2006.14799 (2020).

Framing Effects and Cognitive Biases

- Framing refers to how something is asked as opposed to what is asked.
- In human evaluation for NLG, framing could be reflected in **question wording or instructions** provided to participants

How much more fluent is sentence A versus sentence B?

- Framing demonstrated that people are **more likely to make** choices that are framed positively (in terms of ***gains***) as opposed to negatively (in terms of ***losses***) due to the increased perceived risk associated with losses.

Schoch, Stephanie, Diyi Yang, and Yangfeng Ji. ""This is a Problem, Don't You Agree?" Framing and Bias in Human Evaluation for Natural Language Generation." In Proceedings of the 1st Workshop on Evaluating NLG Evaluation, pp. 10-16. 2020.

Human Evaluation Design Statements

- When describing human evaluation design setup
  - **Question design:** types, scales, wording
  - **Question presentation:** ordering, questions per annotator
  - **Target criteria:** definitions
  - **Annotators:** demographics, background, recruitment, compensation
  - When reporting evaluation results, explain what you did, why you did it, and possible shortcomings

Summary

- ✓**How, What, Who and When** (35 mins)
- ✓**Ethics and Rethink Evaluation** (20 mins)