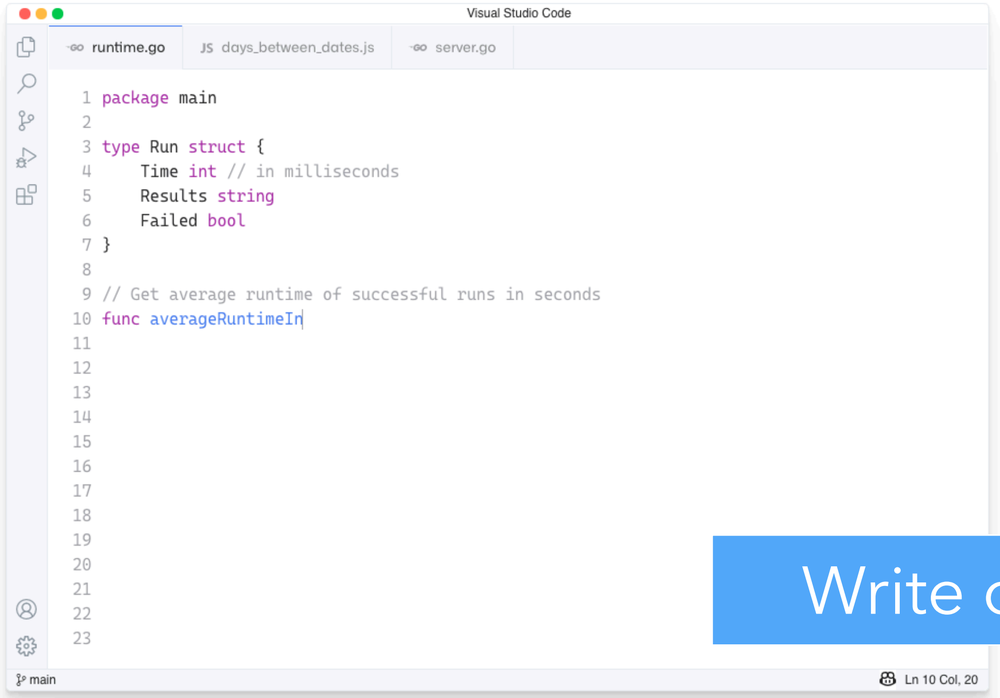
Human+AI: Interactive AI Model Debugging and Correction

Tongshuang (Sherry) Wu
@tongshuangwu / wtshuang@cs.washington.edu
University of Washington

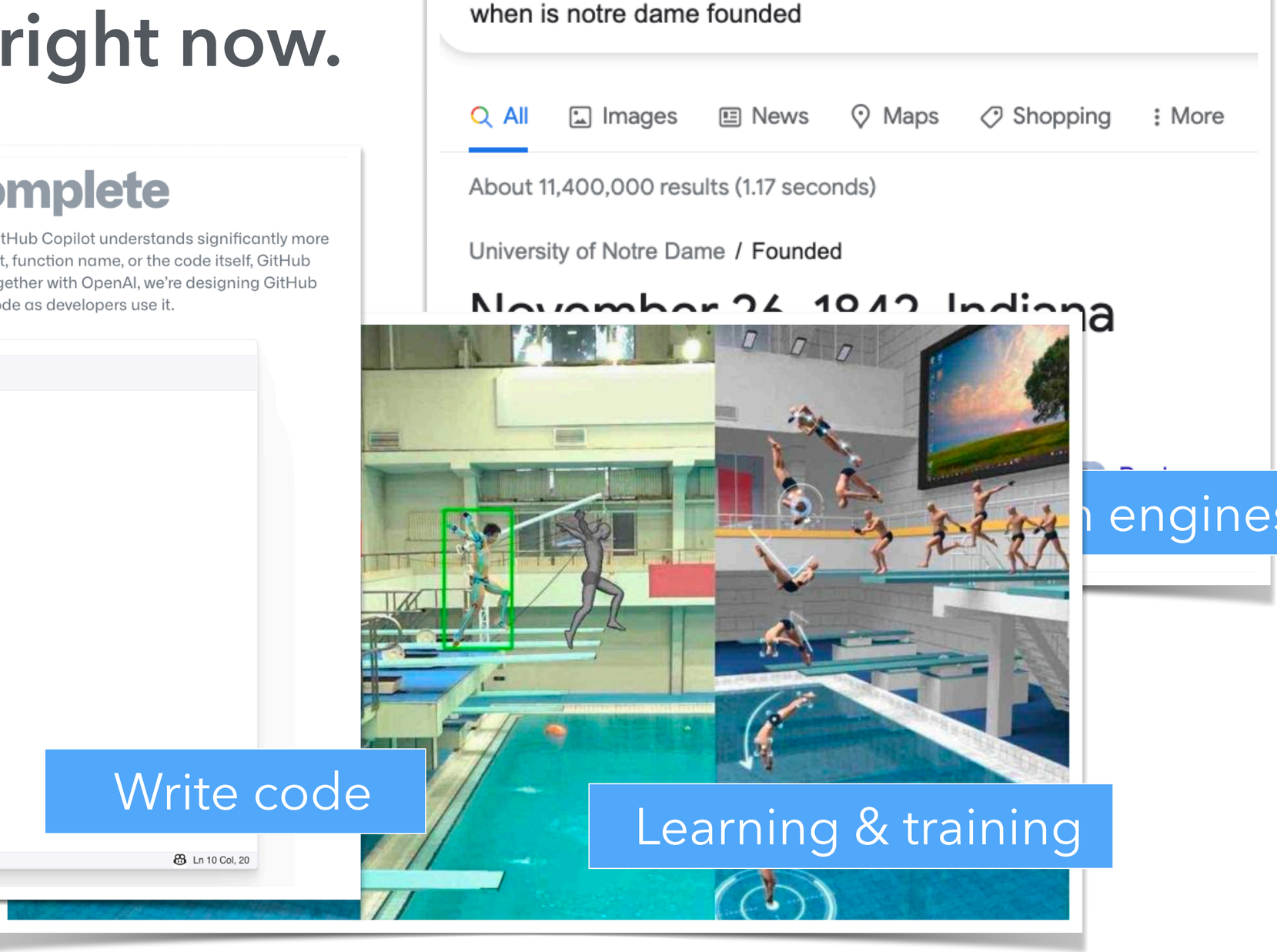
AI is ubiquitous right now.

More than autocomplete

GitHub Copilot is powered by Codex, the new AI system created by OpenAI. GitHub Copilot understands significantly more context than most code assistants. So, whether it's in a docstring, comment, function name, or the code itself, GitHub Copilot uses the context you've provided and synthesizes code to match. Together with OpenAI, we're designing GitHub Copilot to get smarter at producing safe and effective code as developers use it.



Write code



engines

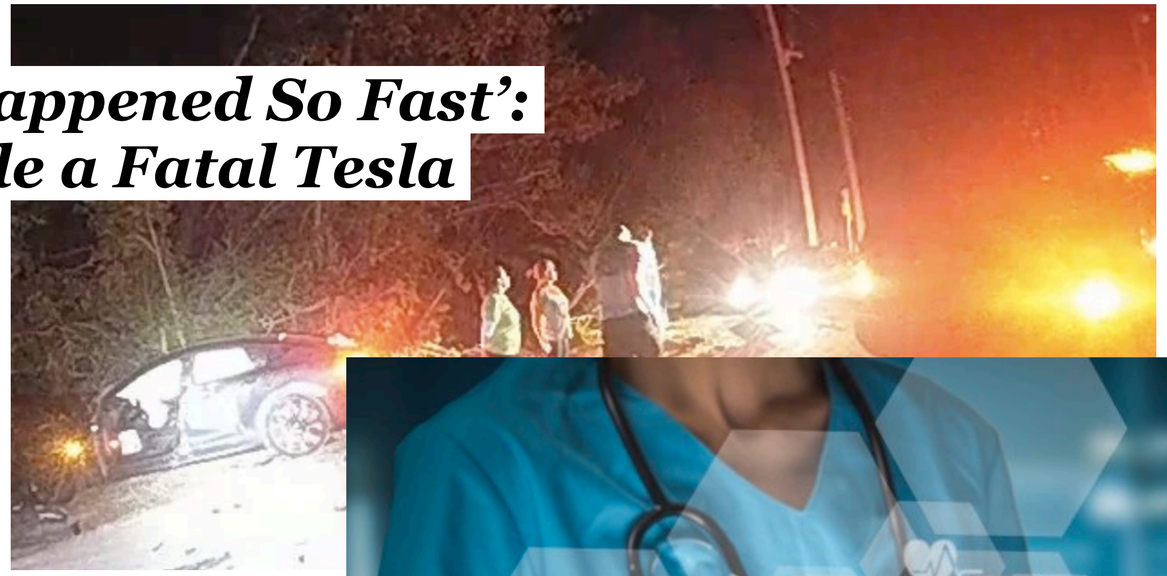
Learning & training

Copilot: Your AI pair programmer.

Baidu's "3D+AI" diving assistant training system behind the Chinese diving team unveiled.

They've **outpaced** our understanding of them…

And hence, always end up on news 😅



'It Happened So Fast': Inside a Fatal Tesla

AI will help, but may also kill people, say US doctors

Facebook Apologizes After A.I. Puts 'Primates' Label on Video of Black Men

AI models are not – and probably will never be – perfect.

Human+AI: How do we identify, improve, and cope with imperfect AIs, such that they still support us despite their deficiencies?

reliable, safe, trustworthy, responsible, humane, ethical, fair, unbiased,
robust, dependable, explainable, interpretable, usable, comprehensible,
transparent, traceable, auditable, controllable, predictable, private, and secure…

Who is the human, who is the AI, and which stage are they in?

AI developers working on models-to-deploy,
Find why models fail & how to fix them.

End users with deployed models,
work with AI towards common goals.

Human+AI have different challenges & goals, depending on their relationships.

analyzer ➜ analyzee

AI developers working on models-to-deploy,
Find why models fail & how to fix them.

Perform systematic analyses, among
various model development stages,
to prioritize more severe errors & root fixes.

+ collaborator

End users with deployed models,
work with AI towards common goals.

Build appropriate reliance, through
effective communication,
to achieve better results than either alone.

Humans analyzing models

Systematic and objective understanding & control

Domain expertise vs. task-properties

Why do we analyze models?

Should I replace my doctor with OSCAR?

Should we use OSCAR in our products?

If not, what do I need to fix?

OSCAR: Pre-training of Neural Networks Directly on Human Brains

Gnome Chompsky
Arcadia Research
chompsky@arcadia.com

Waltolomew Strickler
Arcadia Oaks High
stricklander@aoh.edu

Abstract

We train neural networks on human brains and achieve SOTA in everything.

1 Introduction

3 Model and Architecture

Our basic human brain training approach is similar in spirit to the process described in [XYZ] et al, with the relative straightforward difference that we train directly on brains rather than on text.

We train a transformer model with a CNN on top

Accuracy is not enough!

Common issue: Missing critical data slices (bias, fairness)

What's in the figure?

```
> ceremony, wedding, bride, groom, dress
```

```
> person, people
```

https://ai.googleblog.com/2018/09/introducing-inclusive-images-competition.html

High accuracy ≠ Model succeeding.

How can we make sure our model can handle particular data slices?

Common issue: Shortcuts/right for wrong reasons

What is the moustache made of?

> Banana

What are the eyes made of?

> Banana

What is?

> Banana

What?

> Banana

Agrawal, A., Batra, D., & Parikh, D. (2016). Analyzing the behavior of visual question answering models. arXiv preprint arXiv:1606.07356.

High accuracy ≠ Model succeeding.

How can we make sure our model can handle particular data slices?

Correct prediction ≠ correct reasoning.

How can practitioners ensure the model learns important features & avoid spurious correlations?

Why do we analyze models?

Should I replace my doctor with OSCAR?

Should we use OSCAR in our products?

If not, what do I need to fix?

OSCAR: Pre-training of Neural Networks Directly on Human Brains

chompsky@arcadia.com

stricklandm@aoh.edu

Abstract

We train neural networks on human brains and achieve SOTA in everything.

1 Introduction

3 Model and Architecture

Our basic human brain training approach is similar in spirit to the process described in [XYZ] et al, with the relative straightforward difference that we train directly on brains rather than on text.

We train a transformer model with a CNN on top

Model analyses are prerequisites for making informed decisions on model deployments, and targeted improvements.

"Understanding the **broader terrain** *of errors is an important starting point in pursuing systems that are robust, safe, and fair…[We need to] identify* **cohorts** *with higher error rates and diagnose the* **root causes** *behind these errors."*

Eric Horvitz / Microsoft, 2021

https://erroranalysis.ai/

Analyzing structured data is easy.

How good is our model on records with different city entries?

What happens if I change the city column to ***`New York`****?*

Give me 0.001 seconds to run a SQL script!

Analyzing ~~structured data~~ text is ~~easy~~ hard.

How good is our model on passive sentences on cities?

What happens if I change the passive voice to positive?

Ugh…POS? Named entities? clustering?

“State-of-the-art”

we sampled 200 question answer pairs and manually analyzed their properties.

Joshi et al.
ACL’17

Chen et al.
ACL’16

We randomly select 50 incorrect questions and categorize them into 6 classes.

We sample 100 incorrect predictions and try to find common error categories.

Wadhwa et al.
ACL’18

Joshi, Mandar, et al. "Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension." arXiv preprint arXiv:1705.03551 (2017).
Chen, Danqi, Jason Bolton, and Christopher D. Manning. "A thorough examination of the cnn/daily mail reading comprehension task." arXiv preprint arXiv:1606.02858 (2016)
.Wadhwa, Soumya, Khyathi Raghavi Chandu, and Eric Nyberg. "Comparative analysis of neural qa models on squad." arXiv preprint arXiv:1806.06972 (2018).

“State-of-the-art”

we sampled 200 question answer pairs and manually analyze

Joshi et al.
ACL’17

Chen et a
ACL’16

orize

them into 6 classes.

We sar
comm

dhwa et al.
ACL’18

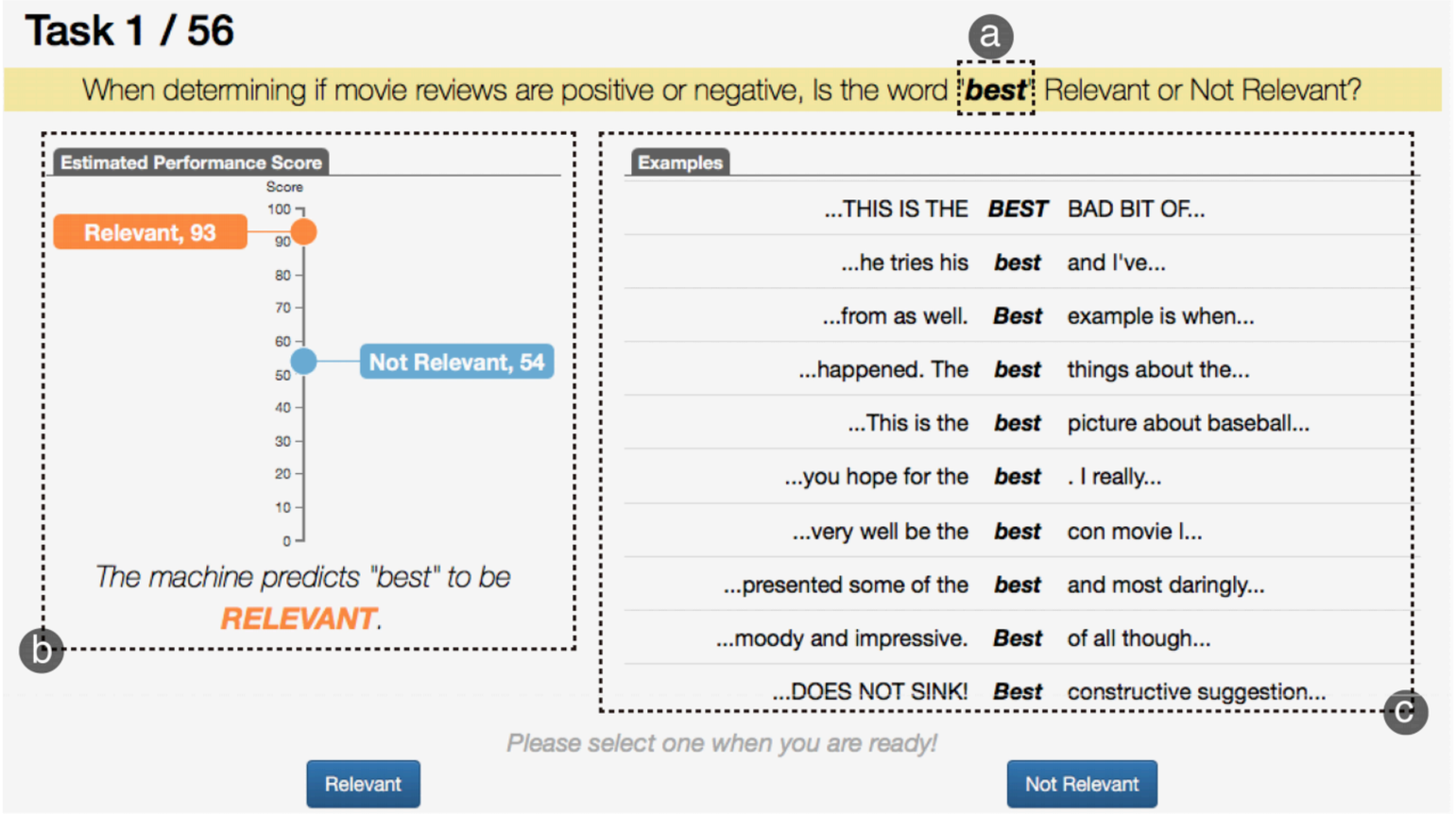
“We randomly select 50-100 instances and roughly label them into N error groups.”

Under-representative, subjective, high variance, low reproducibility

Joshi, Mandar, et al. "Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension." arXiv preprint arXiv:1705.03551 (2017).
Chen, Danqi, Jason Bolton, and Christopher D. Manning. "A thorough examination of the cnn/daily mail reading comprehension task." arXiv preprint arXiv:1606.02858 (2016)
.Wadhwa, Soumya, Khyathi Raghavi Chandu, and Eric Nyberg. "Comparative analysis of neural qa models on squad." arXiv preprint arXiv:1806.06972 (2018).

Local context & small samples are dangerous!

Participants add/remove features for 56 rounds, after seeing relevant examples.



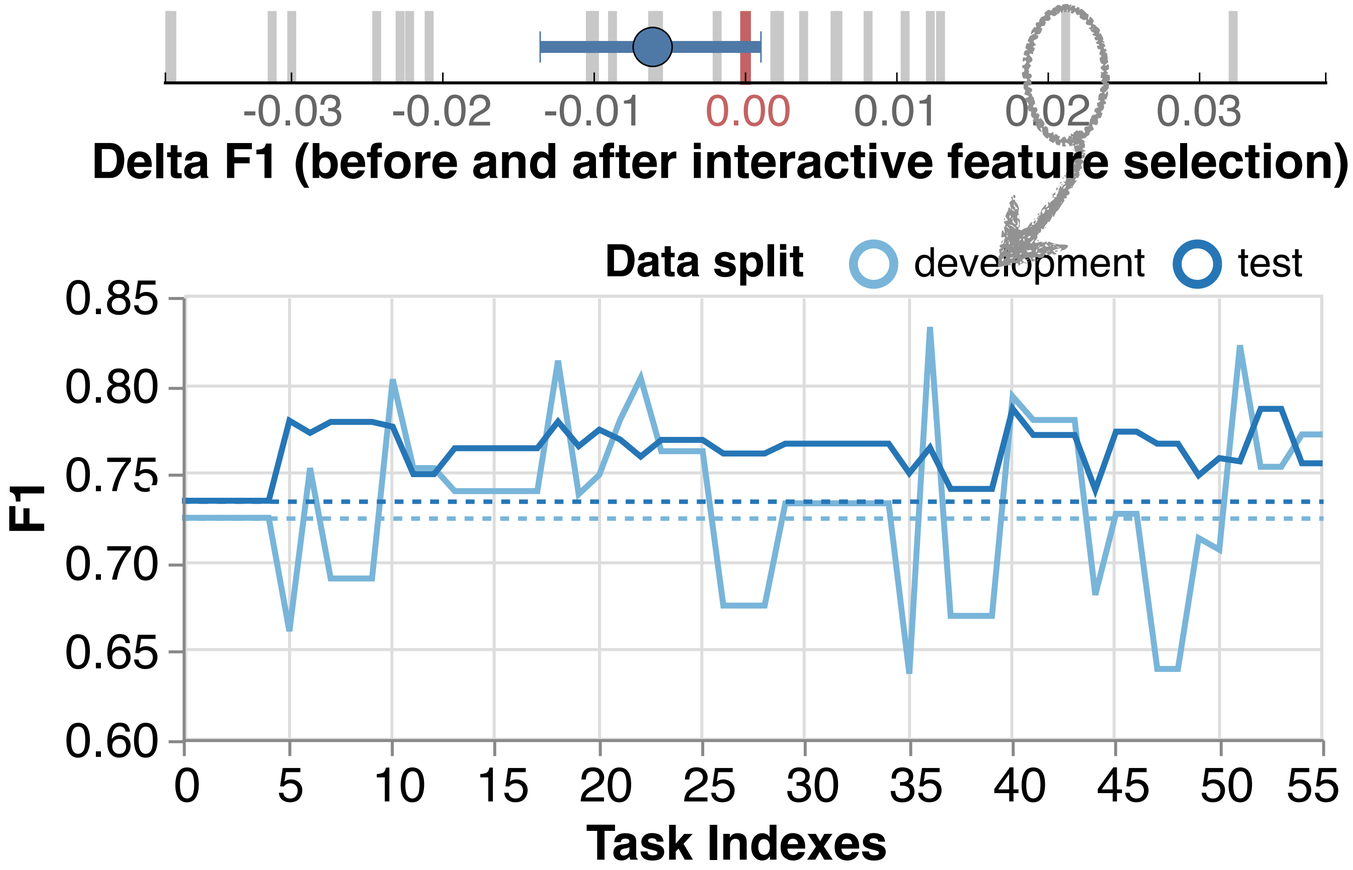
Local Decision Pitfalls in Interactive Machine Learning: An Investigation into Feature Selection in Sentiment Analysis, TOCHI 2019

Local context & small samples are dangerous!

Participants add/remove features for 56 rounds, after seeing relevant examples.

Their Interactive refinement makes F1 oscillate, not improve!

More global analysis!



Local Decision Pitfalls in Interactive Machine Learning: An Investigation into Feature Selection in Sentiment Analysis, TOCHI 2019

What components are essential for systematic analyses?

Errudite: Scalable, Reproducible, and Testable Error Analysis

ACL 2019

Systematic Error Analysis with Grouping & counterfactuals

Visual question answering

How many people are in this picture?

groundtruth:**3** *(* 10)*

vqacounting:**3**

Correct!

How many brownish peaks are there?

groundtruth:**2** *(* 10)*

vqacounting:**5**

Incorrect!

Hypothesis: the model does not recognize **ADJectives** in “How many” questions.

How to verify?

Scale Up Reproducible Grouping through Filtering Rules

Quantify instances with a DSL (Domain Specific Language)

Find reproducible instance groups with filters on attributes (*token is NOUN, ADJ, etc.*),

Filters trigged by programming-by-demonstration

How many people are in this picture?

groundtruth:**3** *(* 10)*

vqacounting:**3**

How many brownish peaks are there?

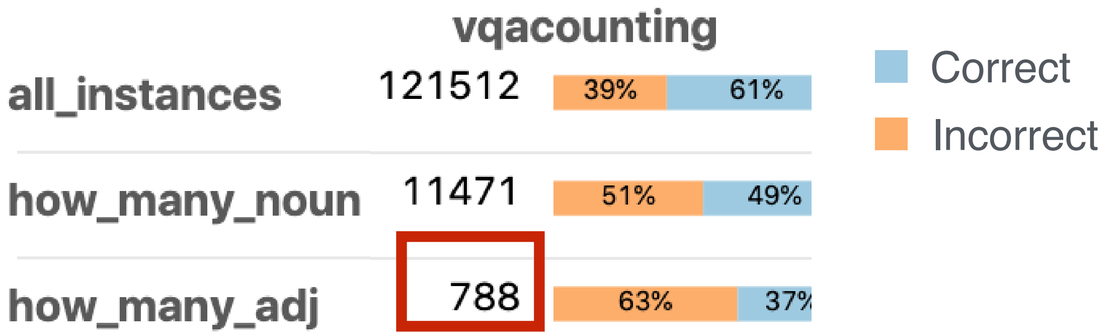
groundtruth:**2** *(* 10)*

vqacounting:**5**

DID YOU MEAN TO FILTER INSTANCES THAT ARE... Close Now

```
starts_with(question, pattern="how many ADJ")
starts_with(question, pattern="ADV ADJ ADJ")
attr:question_type == "how many"
```

See more general suggestions?



Building blocks

Quantitative grouping

Inspect similar instances, semantically & syntactically

Enables…

Precise & reproducible hypotheses

+

Scale up to the entire dev set

Narrow Down Root Causes via Counterfactual Rewrite Rules

HAS adjective ≠ **IS WRONG** due to the adjective.

Answer what-if questions with counterfactual analysis:

"Would the model work perfectly if we removed the possible cause?"

How many people are in this picture?

groundtruth:**3** *(* 10)*

vqacounting:**3**

How many brownish peaks are there?

groundtruth:**2** *(* 10)*

vqacounting:**5**

How many ~~brownish~~ peaks are there?

DID YOU WANT TO GENERALIZE TO...

brownish →	☐ keep
brownish peaks → **peaks**	☐ keep
brownish NOUN → NOUN	☐ keep
ADJ NOUN→ NOUN	☐ keep
how many ADJ→ **how many** ADJ	☐ keep
how many ADJ NOUN → **how many** NOUN	☑ keep

Building blocks

Quantitative grouping

Inspect similar instances, semantically & syntactically

Counterfactual perturbation

Isolate important components targeted minimal rewrites

Enables…

Precise & reproducible hypotheses

+

Scale up to the entire dev set

+

Test via counterfactual analysis

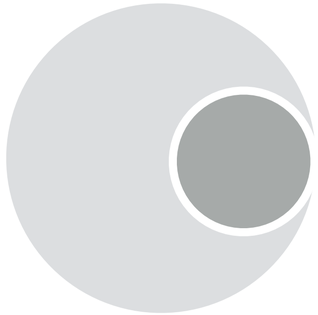
User study: **The impact of “adjectives”?**

NLP experts saw identical error descriptions

“The model cannot count objects with adjectives…”

But produce drastically different groups (covering 13.5% – 45% of all errors!)

```
has_pattern(question,
  pattern="ADJ")
```



```
starts_with(question,
 pattern="how many ADJ")
```

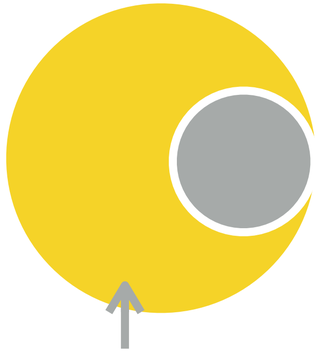
User study: **The impact of “adjectives”?**

NLP experts saw identical error descriptions

“The model cannot count objects with adjectives…”

But produce drastically different groups (covering 13.5% – 45% of all errors!)

```
has_pattern(question,
  pattern="ADJ")
```

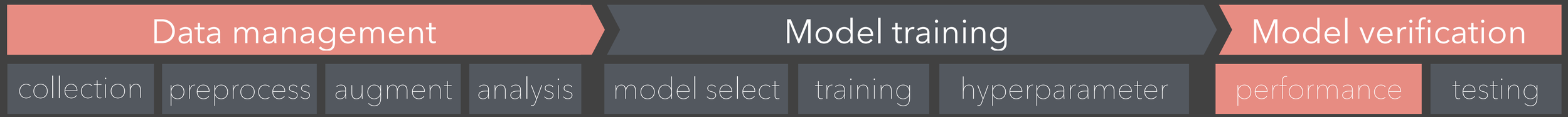


```
starts_with(question,
 pattern="how many ADJ")
```

“How many peaks are in brown?”

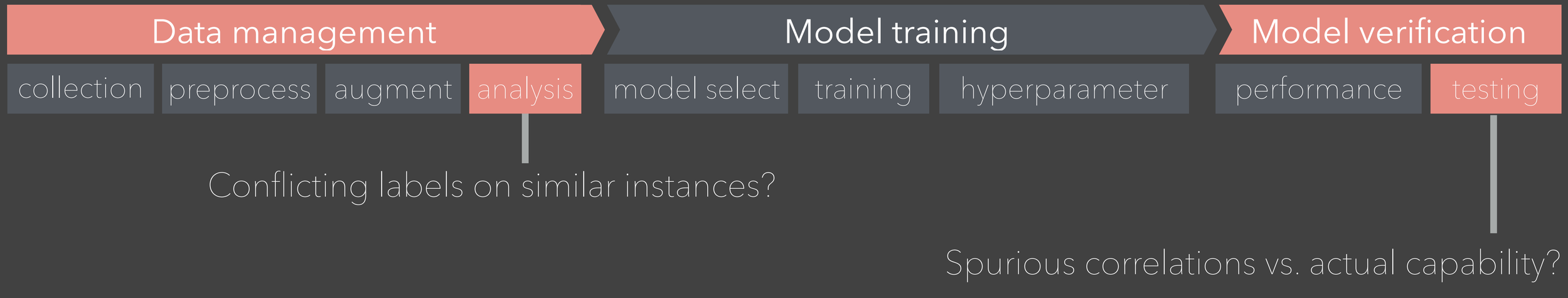
Errudite 📖: precise, reproducible, scalable, and testable error analysis, through grouping and counterfactual analysis

Error analysis is just one of the many model development stages involved... (Paleyes et al. 2020)



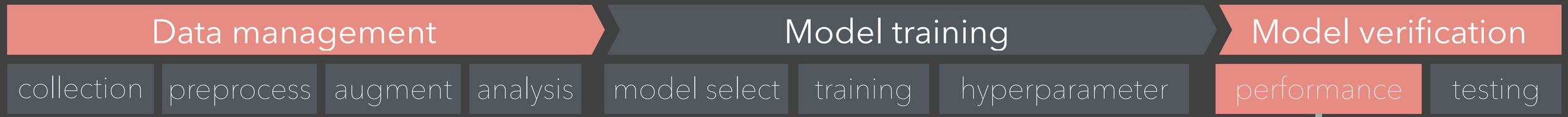
Paleyes, A., Urma, R. G., & Lawrence, N. D. (2020). Challenges in Deploying Machine Learning: a Survey of Case Studies. arXiv

The two building blocks of quantitative grouping and counterfactual perturbation are necessary across stages.



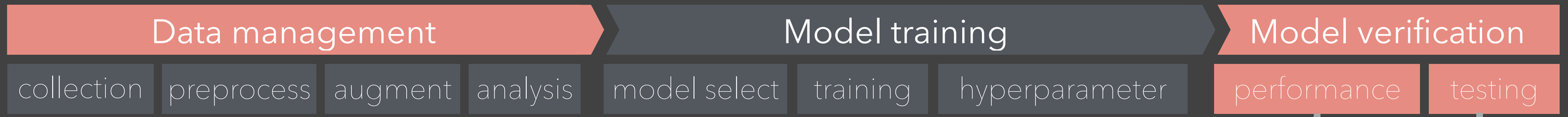
Purely driven by domain knowledge would induce developer bias!

How do we instantiate two building blocks in various stages, balancing domain expertise vs. task knowledge?



Errudite (ACL 2019)

How do we instantiate two building blocks in various stages, balancing domain expertise v.s. task knowledge?

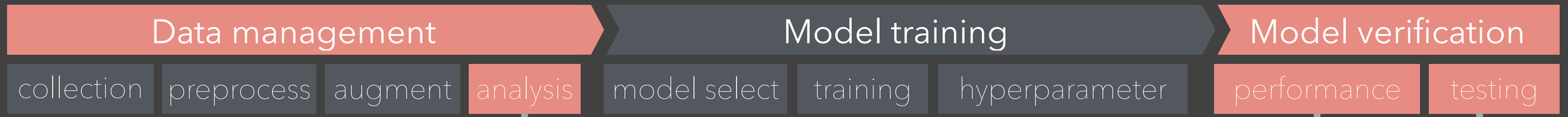


Errudite (ACL 2019)

CheckList (ACL 2020)

A framework to guide humans to test otherwise overlooked capabilities

How do we instantiate two building blocks in various stages, balancing domain expertise v.s. task knowledge?



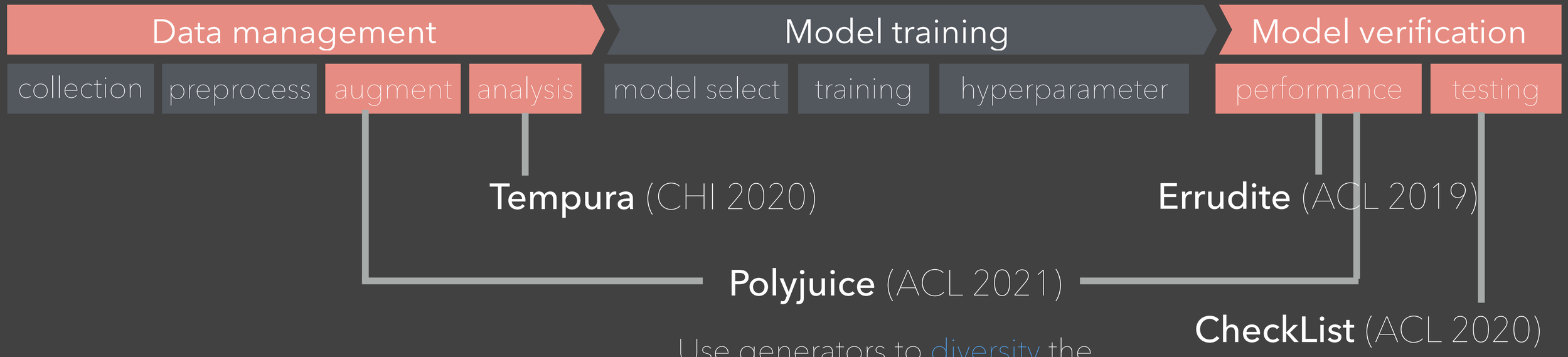
Tempura (CHI 2020)

Automatically mine interpretable groups to expose dataset properties.

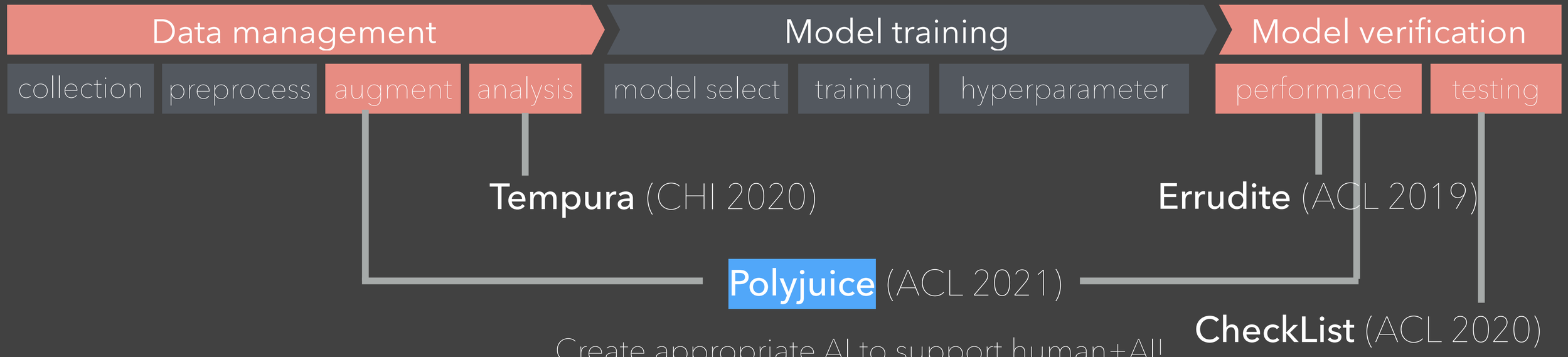
Errudite (ACL 2019)

CheckList (ACL 2020)

How do we instantiate two building blocks in various stages, balancing domain expertise v.s. task knowledge?



How do we instantiate two building blocks in various stages, balancing domain expertise v.s. task knowledge?

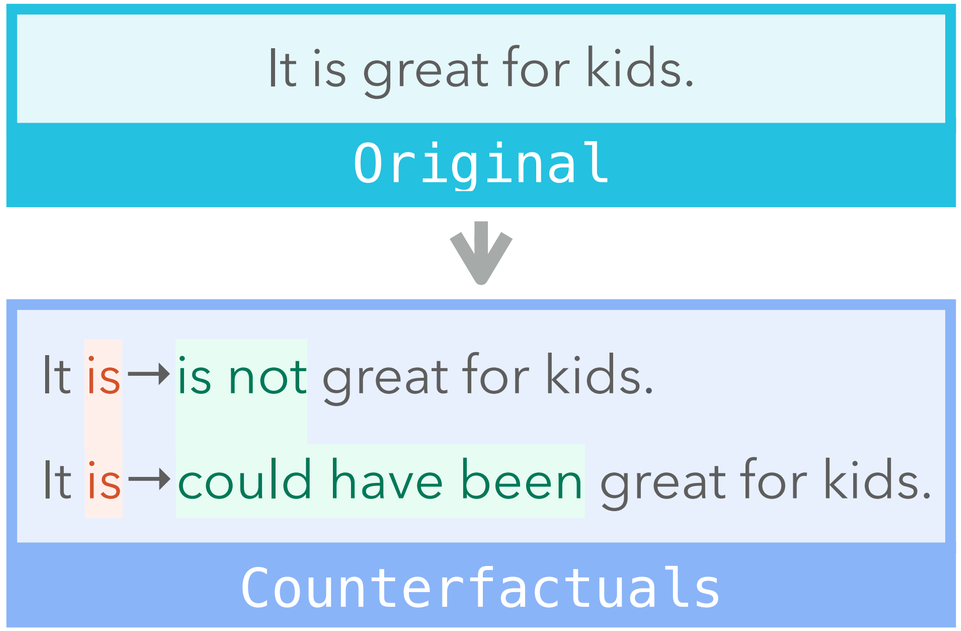


Manually creating counterfactuals is hard.

Costly to generate
(4-5 minutes per counterfactual)

Human creativity can be biased

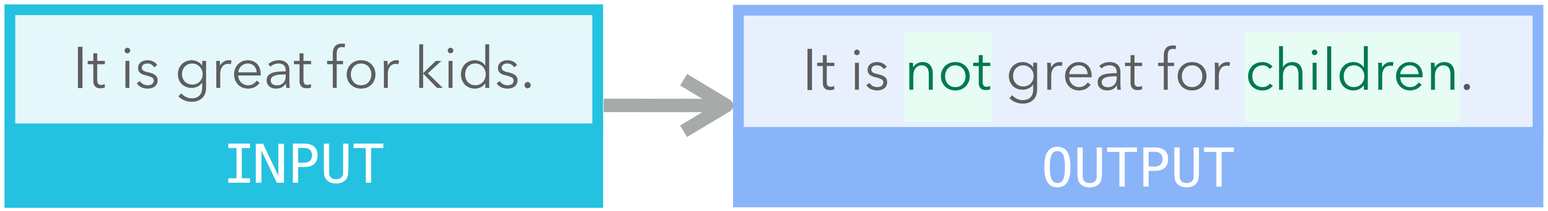
Use generators to boost diversity!



Kaushik, D., Hovy, E., & Lipton, Z. C. (2019). Learning the difference that makes a difference with counterfactually-augmented data. ICLR 2020.
Gardner, Matt, et al. "Evaluating models' local decision boundaries via contrast sets." *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*. 2020.

Counterfactual generation, with context and controls

GPT-2 can complete paragraphs → be fine-tuned for rewriting.



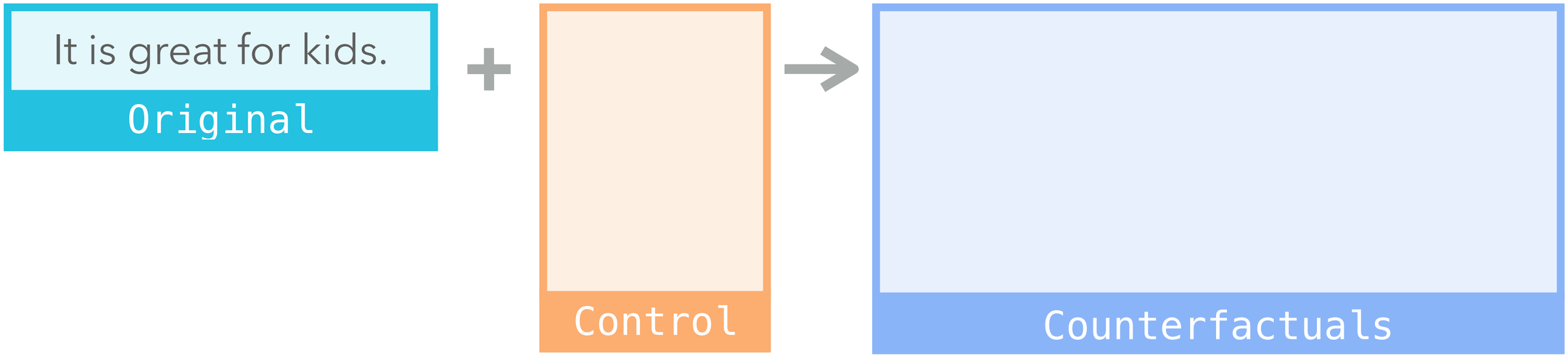
*Use original text as **context***

Close** edit based on the **entire context

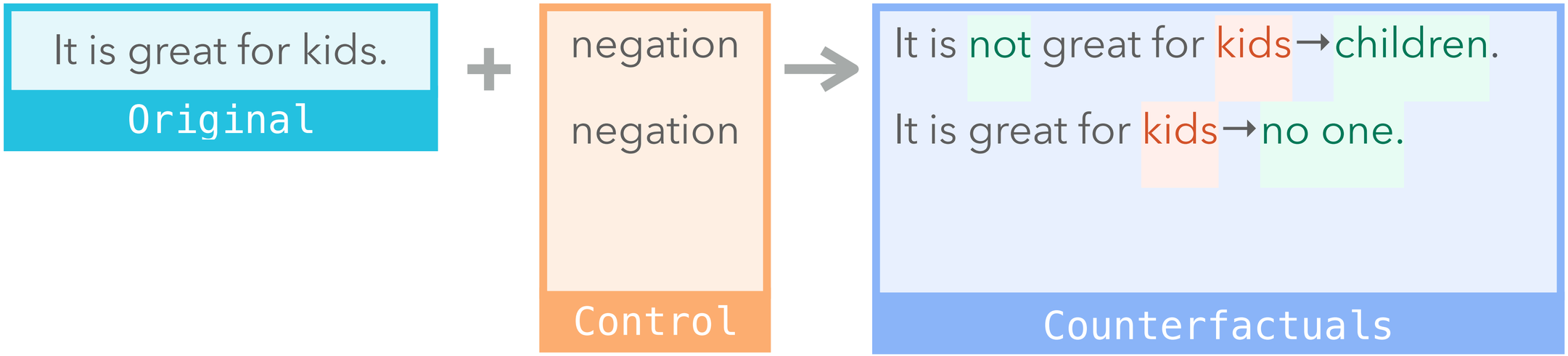
Prompting: "**how**" and "**where**" to perturb,
to emphasize different perturbation types.

Polyjuice: Generating Counterfactuals for Explaining, Evaluating, and Improving Models, ACL 2021

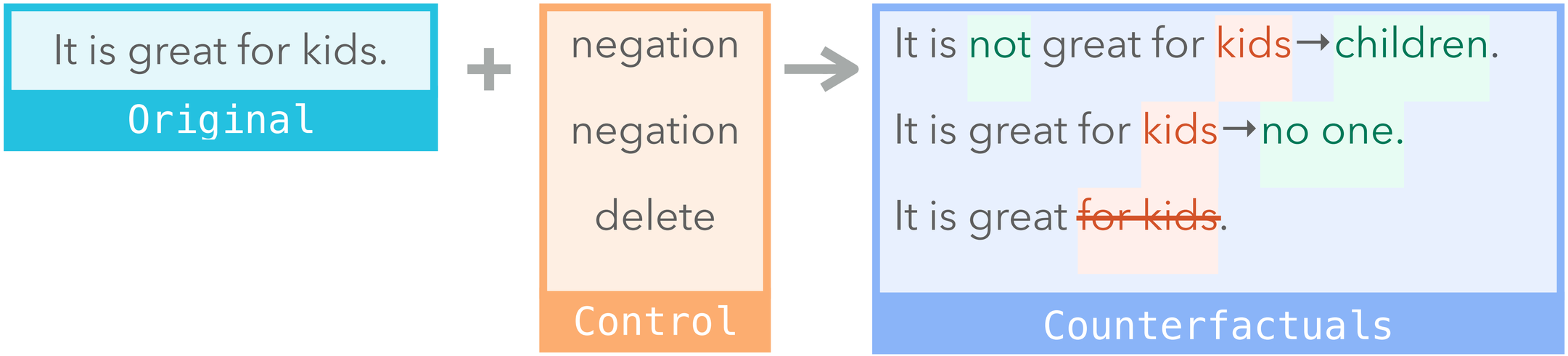
“How to change”: Control codes



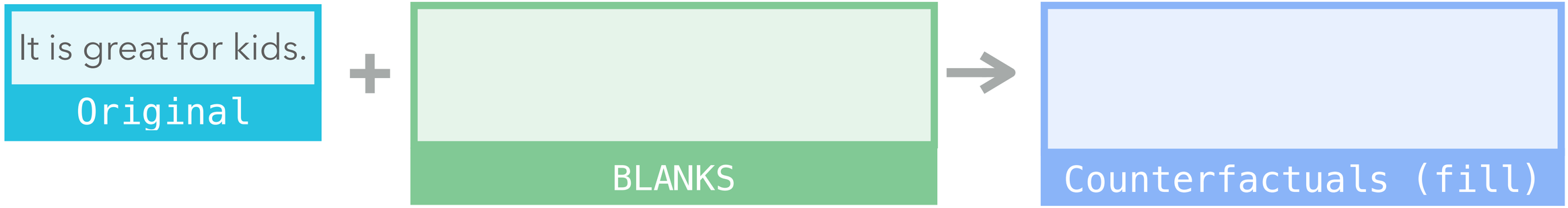
“How to change”: Control codes



“How to change”: Control codes



“Where to change”: Fill-in-the-blank (Donahue, ACL’20)



Donahue, C., Lee, M., & Liang, P. (2020). Enabling Language Models to Fill in the Blanks. ACL 2020

“Where to change”: Fill-in-the-blank (Donahue, ACL’20)

Donahue, C., Lee, M., & Liang, P. (2020). Enabling Language Models to Fill in the Blanks. ACL 2020

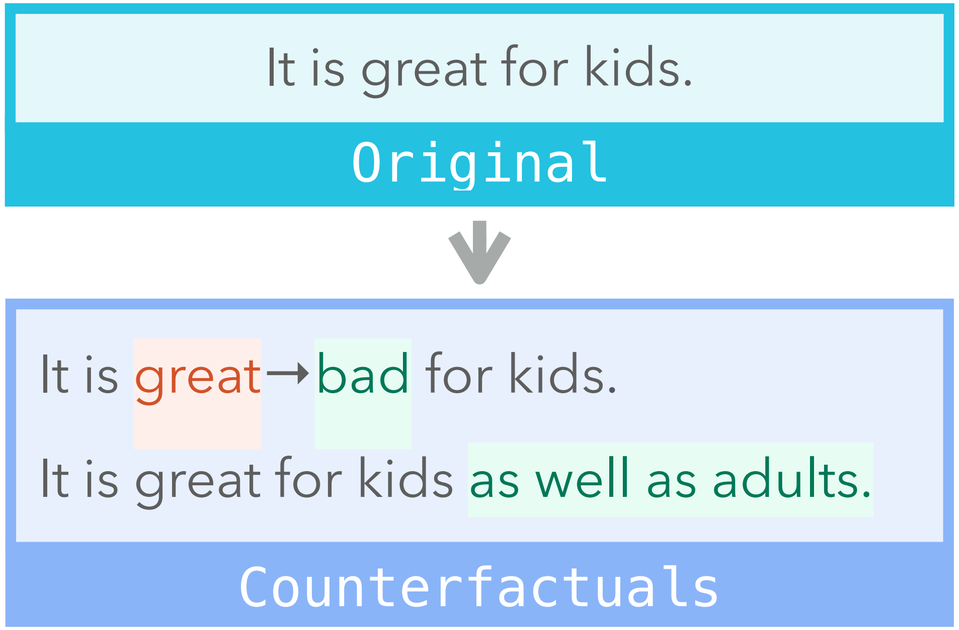
At generation time, allows flexible control…

Original sentence

Based on GPT-2, Polyjuice always picks up from where we left off!

```
It is great for kids. <|perturb|>
  [lexical] It is [BLANK] for kids.
  [SEP] bad [ANSWER]

  [insert] It is great [BLANK].
  [SEP] for kids as well as adults [ANSWER]
<|endoftext|>
```



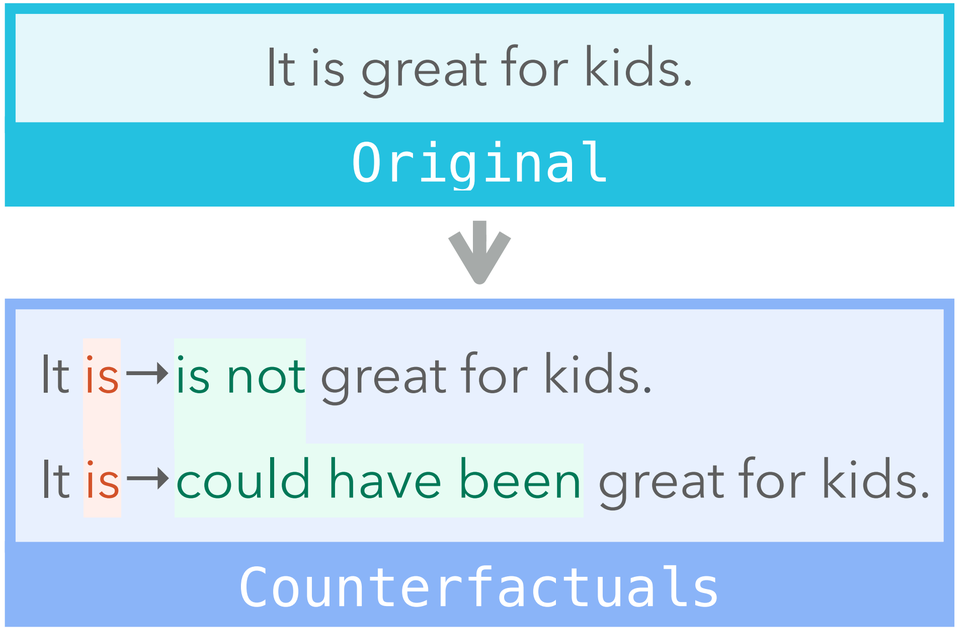
At generation time, allows flexible control…

Original sentence + **control code + blank placement**

Based on GPT-2, Polyjuice always picks up from where we left off!

```
It is great for kids. <|perturb|> [negation]
  It [BLANK] great for kids.
  [SEP] is not [ANSWER]

  [SEP] could have been [ANSWER]
<|endoftext|>
```



Interactive analysis: the benefit of multiple $\hat{x}$ per x

x	$f(x)$
P: A woman is holding a baby by a window. H: This woman is looking out the window.	*Neutral*
$\hat{x}$, perturbed H through [negation]	$f(\hat{x})$
H: No woman is looking out the window.	*Contradiction*
H: This woman isn't looking out the window.	*Contradiction*
H: This woman is not looking out the window.	*Neutral*

NLI: does negation correlate with contradiction?

It depends!

Inconsistency between "n't" and "not"!

$x \rightarrow f(\hat{x})$	Template	Coverage (%*N*→*C*)
…is not looking…	`AUX → AUX not`	
…aren't playing…	`* → * not`	412 (42.3%)
The→No girls like…	`* → * n't`	434 (43.5%)
	`* → * PART`	
A→No man in…	**`DET → No`**	180 (**92.8%**)

`DET → No` flips model prediction much more frequently!

Gururangan, S., Swayamdipta, S., Levy, O., Schwartz, R., Bowman, S. R., & Smith, N. A. Annotation Artifacts in Natural Language Inference Data. ACL 2020
Wu, T., Wongsuphasawat, K., Ren, D., Patel, K., & DuBois, C.. Tempura: Query Analysis with Structural Templates. CHI 2020

Counterfactual data aug: **Crowd labeling**

Find diverse counterfactuals

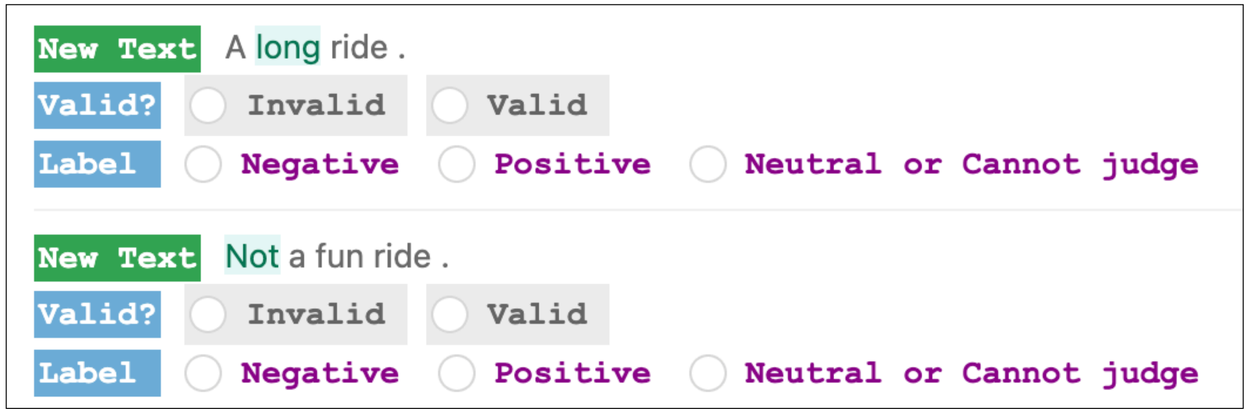
prefer

Not a fun ride
A fun → long ride.

over

A fun → good ride.
A fun → interesting ride.

Crowdsource labels



Crowds only **evaluate** and **label** the examples,

rather than **generating** them manually.

Counterfactual data aug: **Training results**

Natural Language Inference (on SNLI)

`polyjuice`: add counterfactuals (automatic)

vs. **`CAD`** : add counterfactuals (manual, from Kaushik et al.)

vs. **`baseline`** : add the same amount of original data

Model	SNLI	MNLI-m	MNLI-mm	SNLI-CAD	break	DNC	stress	diagnostic
`m-baseline`	85.7 ± 0.4	86.1 ± 0.2	**86.6 ± 0.2**	72.8 ± 0.3	86.4 ± 1.5	54.5 ± 0.6	65.1 ± 0.6	56.0 ± 0.8
`m-CAD`	85.8 ± 0.6	**86.6 ± 0.1**	85.6 ± 0.3	**73.8 ± 0.2**	89.4 ± 2.9	55.8 ± 0.9	65.5 ± 0.5	56.4 ± 0.4
`m-polyjuice`	85.3 ± 0.3	86.0 ± 0.1	**86.4 ± 0.0**	**73.6 ± 0.2**	**89.1 ± 1.2**	**57.7 ± 0.3**	65.1 ± 0.2	**57.5 ± 0.5**

Polyjuice helps improve on multiple contrast/challenge sets **, even better than CAD**

30 seconds per round (3 perturbations)

4 minutes per revised sentence in NLI

Polyjuice counterfactuals are **better** than more original data, **cheaper** than manual collection

Kaushik, D., Hovy, E., & Lipton, Z. C. (2019). Learning the difference that makes a difference with counterfactually-augmented data. ICLR 2020.

Polyjuice : Counterfactual generator, with explicit controls over the kinds and locations of perturbations.

Polyjuice generates multiple counterfactuals per instance, and leads to insights that might be missed by manual analysis.

Things to consider in model analysis research

Humans analyzing AIs care about global understanding.

Humans are ML experts; “shared representations” highlight domain expertise.

Errudite domain specific language, Polyjuice control codes

The balance of domain knowledge vs. task property depends on dev stage.

Sometimes need to do the NLP modeling work in order to do the HCI work!

Heer, Jeffrey. "Agency plus automation: Designing artificial intelligence into interactive systems." PNAS 2018

Humans collaborating with models

collaborator + collaborator

Appropriate reliance

Two-way communication

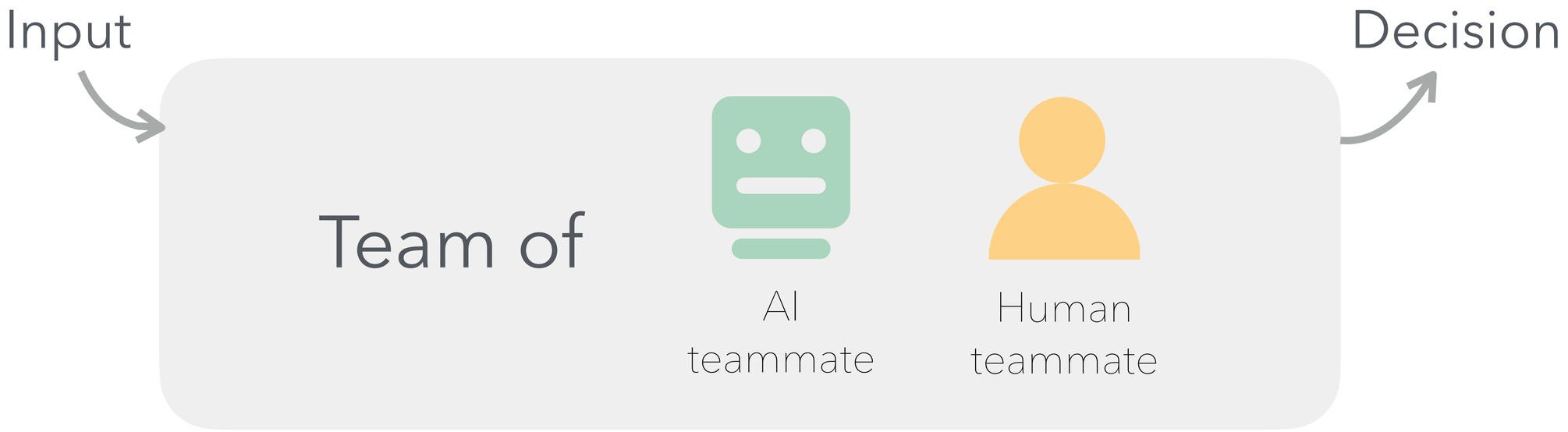
Our fellow analyzers have done their best…

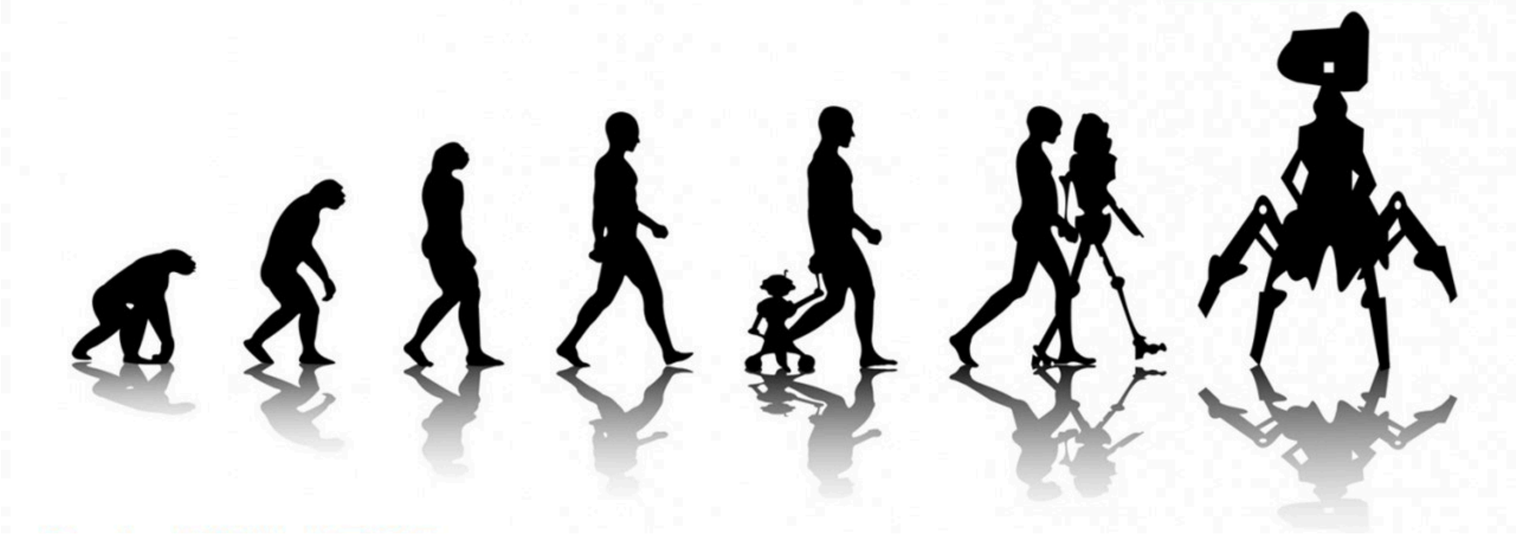
But AIs always over-simplify the world (domain shift, etc.).

How *experts* train the model

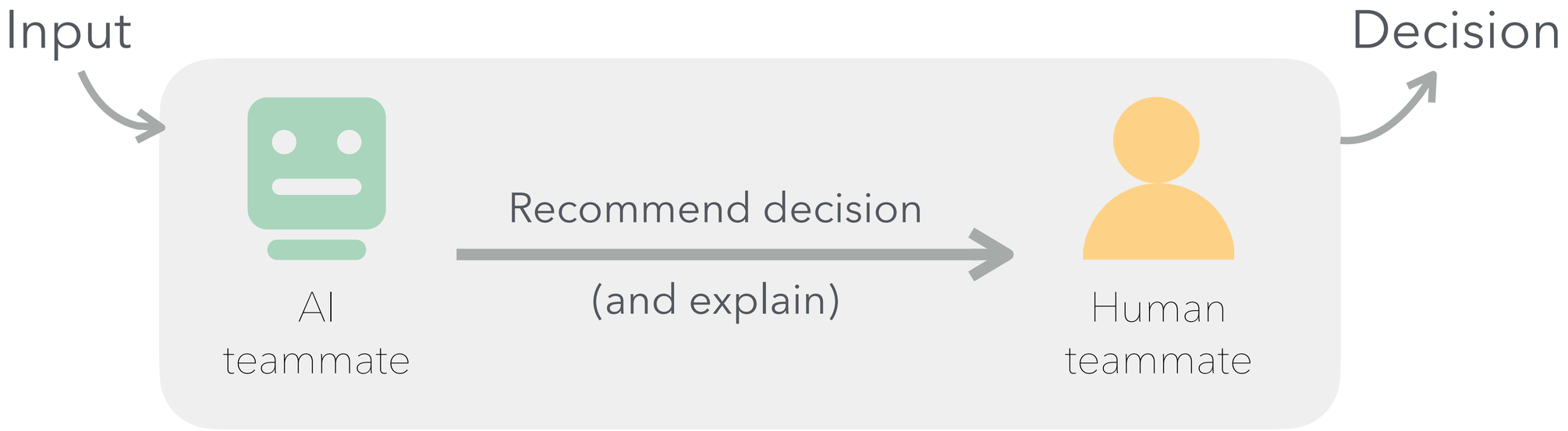
How *end users* use the model

End users still need to debug AI teammates.





End users still need to debug AI teammates.



Appropriate trust: ***accept*** *AI recommendations when they are correct, but* ***overwrite*** *them when they don't make sense.*

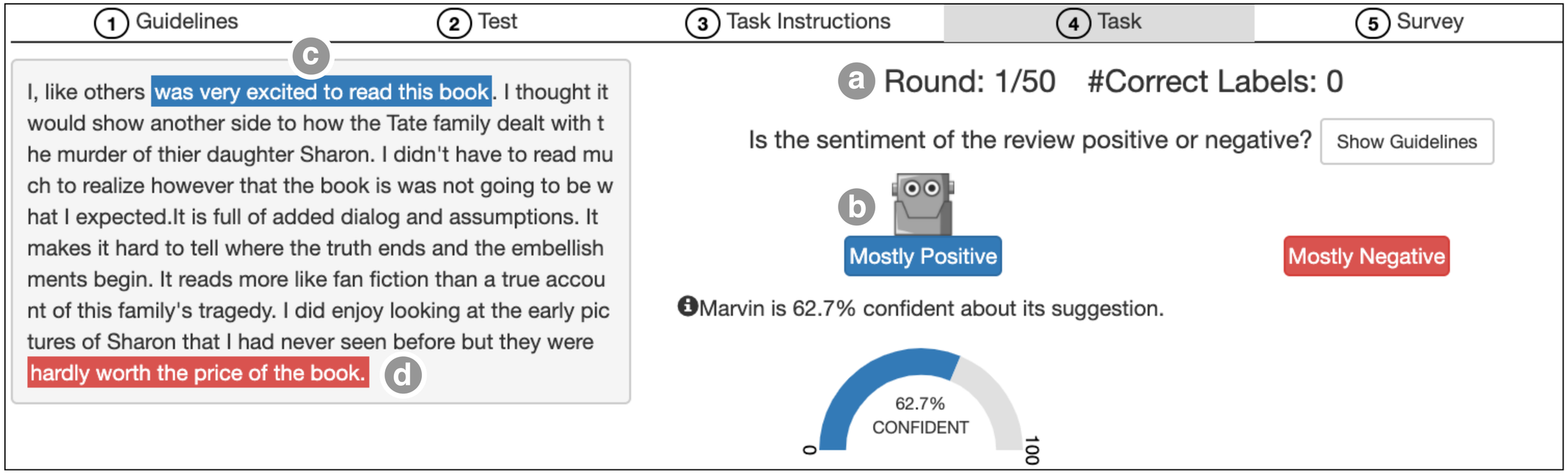
But, appropriate reliance is hard.

Crowdsourced study, with 1,626 participants:

team up for some tasks (e.g. classify documents).

Multiple teaming strategies: [AI decision + confidence] + [explanations].

Always displayed *Only in some conditions*



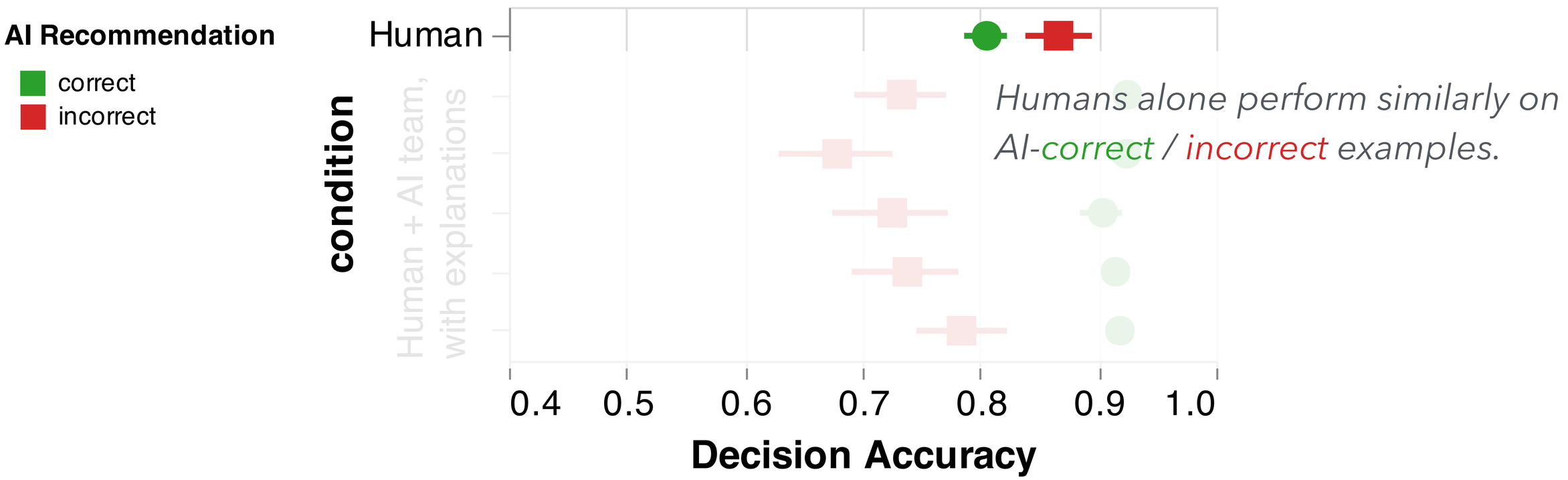
Does the Whole Exceed its Parts? The Effect of AI Explanations on Complementary Team Performance

CHI 2020

But, appropriate reliance is hard.

Inappropriate reliance!

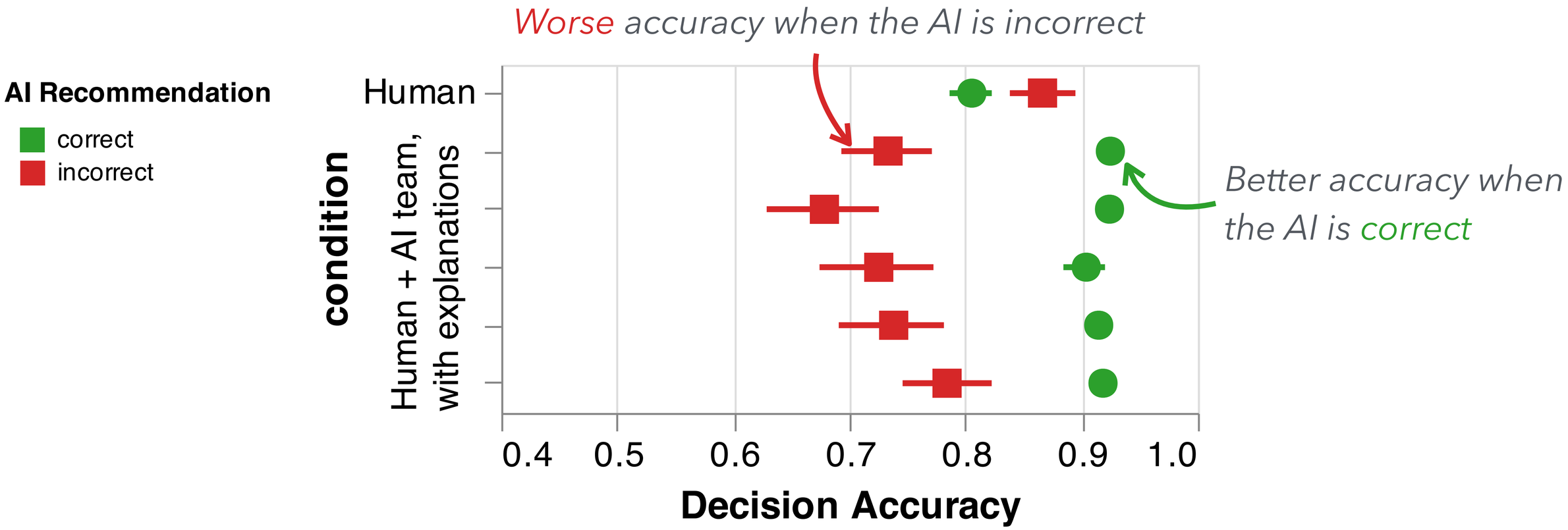
People blindly follow model recommendations, **even when the model is incorrect**.



But, appropriate reliance is hard.

Inappropriate reliance!

People blindly follow model recommendations, **even when the model is incorrect**.



Factors causing inappropriate reliance?

We ask: How do the human collaborators use information from the AI collaborator?

We performed qualitative categorization on users' survey replies…

`Mostly Follow AI`	*"I went with the AI most times."*
`AI as Prior Guide`	*"I looked at his prediction and then I read the passage."*
`AI as Post Check`	*"I didn't. I figured out the paragraph for myself."*
`Mostly Ignore AI`	*"I ignored it until I made my decision and then verified what it said."*

Multiple factors would affect these…

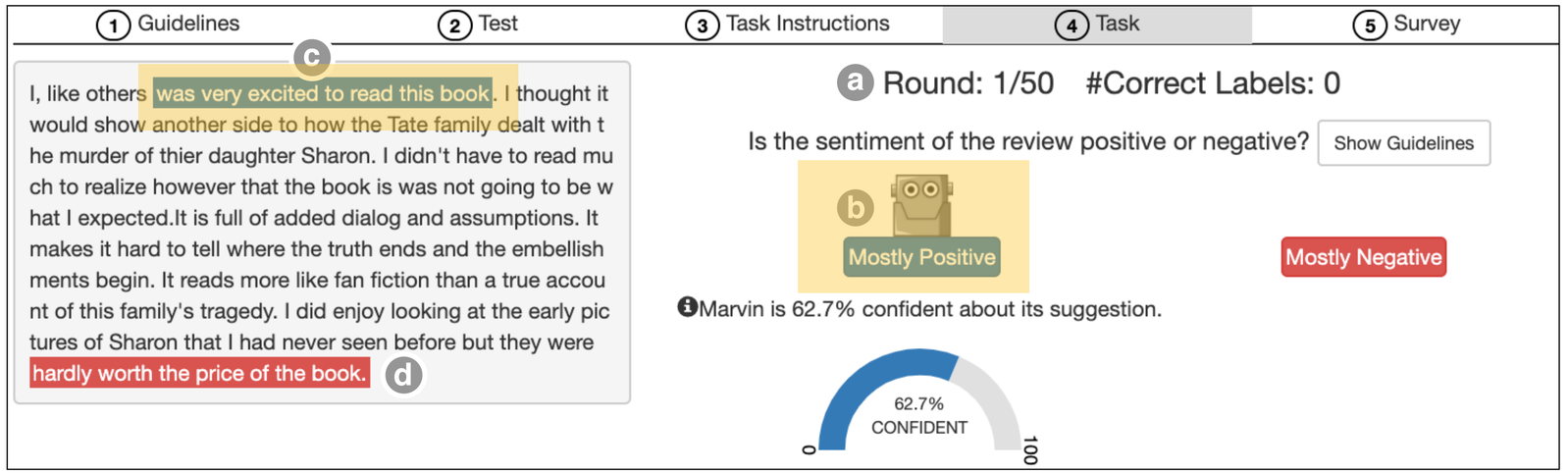
Factors causing inappropriate reliance?

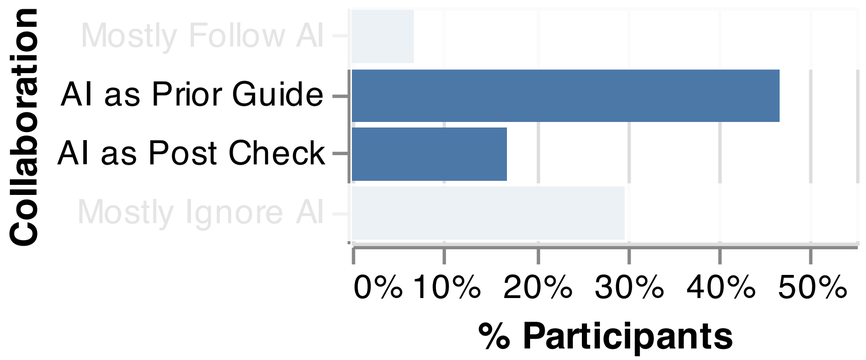
One example: When & how to display the AI recommendation matters.

inline highlight ➜ confirmation bias ("this sentence seems reasonable enough")

See AI decision first ➜ anchoring effect ("I will agree with AI's decision")

Classification task





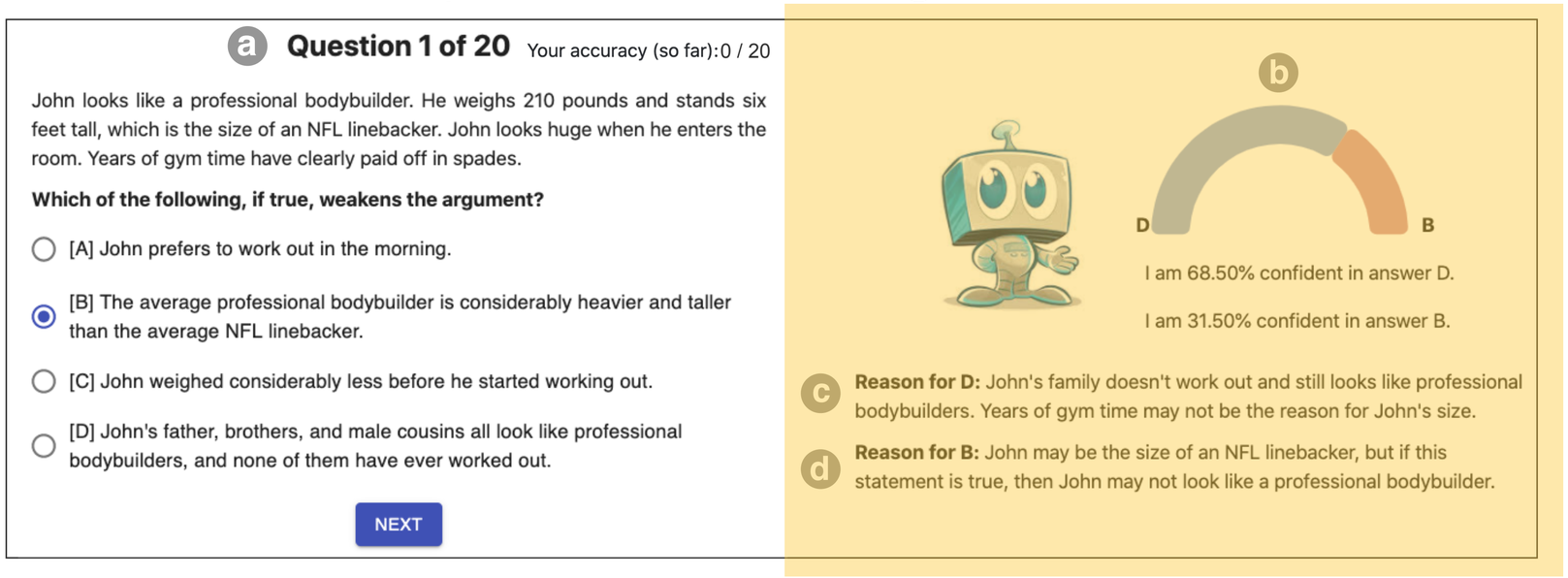
Factors causing inappropriate reliance?

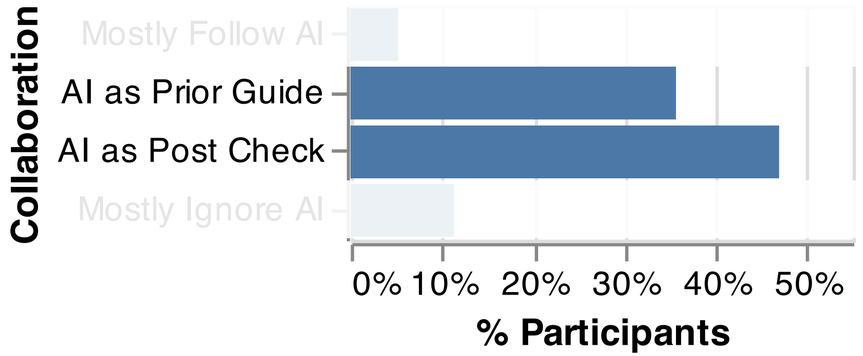
One example: When & how to display the AI recommendation matters.

inline highlight ➜ confirmation bias ("this sentence seems reasonable enough")

See AI decision first ➜ anchoring effect ("I will agree with AI's decision")

Multi-choice question answering task





Factors causing inappropriate reliance?

One example: When & how to display the AI recommendation matters.

inline highlight ➜ confirmation bias (“this sentence seems reasonable enough”)

See AI decision first ➜ anchoring effect (“I will agree with AI’s decision”)

HCI solution

Display of AI recommendation

Present evidence, but not final decision

Timing of AI decision

Asynchronous display, increase independence

Park, Joon Sung, et al. "A Slow Algorithm Improves Users' Assessments of the Algorithm's Accuracy." CSCW 2019

Wang, Danding, et al. "Designing theory-driven user-centric explainable AI." CHI 2019

As the AI/NLP communities iterate on models, the HCI community works on thoughtful visual and interaction designs.

CASE STUDY

Concretely, how does HCI research help improve human-AI collaboration quality?

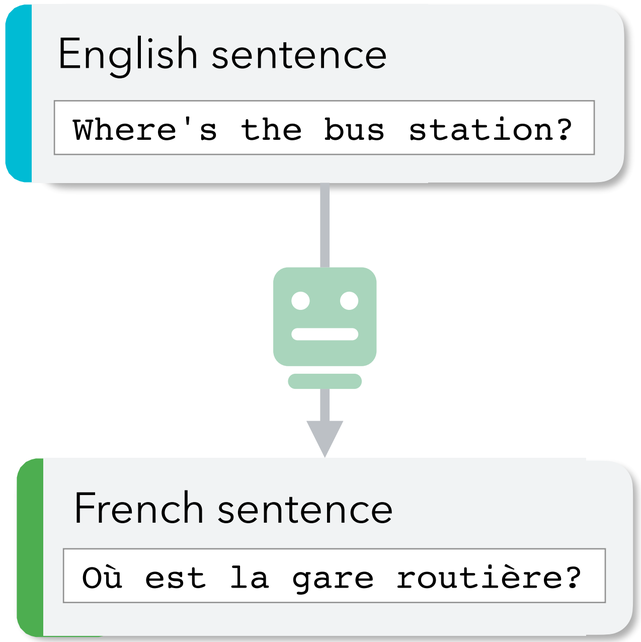
AI Chains: Transparent and Controllable Human-AI Interaction by Chaining Large Language Model (LLM) Prompts

Submitted to CHI 2022, in collaboration with Google PAIR

Large Language Models (LLMs) are helpful in many

"In-context learning": adapt to a wide range of tasks, without re-training the model.

Translation




```
Given the English sentence,
translate to French.

English: Where's the bus station?
French: Où est la gare routière?
```

Prompt
Natural Language instruction

Model output

FURTHER READ!

Large Language Model 101

Bommasani, Rishi, et al. "On the Opportunities and Risks of Foundation Models." arXiv 2021

Brown, Tom B., et al. "Language models are few-shot learners." arXiv 2020

Prompt Engineering 101

Reynolds, Laria, and Kyle McDonell. "Prompt programming for large language models: Beyond the few-shot paradigm." CHI 2021 Extended Abstract

Large Language Models (LLMs) are helpful in many ways.

"In-context learning": adapt to a wide range of tasks, without re-training the model.

Translation

English sentence

```
Where's the bus station?
```

French sentence

```
Où est la gare routière?
```

Write code

Description on HTML

```
Make this button blue and
30px tall and readable
<button>OK</button>
```

HTML code

```
<button style='height: 30px;
  background-color:blue;
  color: white;'>OK</button>
```

Large Language Models (LLMs) are helpful in many ways.

"In-context learning": adapt to a wide range of tasks, without re-training the model.

Translation

English sentence

```
Where's the bus station?
```

French sentence

```
Où est la gare routière?
```

Write code

Description on HTML

```
Make this button blue and
30px tall and readable
<button>OK</button>
```

HTML code

```
<button style='height: 30px;
  background-color:blue;
  color: white;'>OK</button>
```

Create ideas

Presentation problem

```
Too much text on slides
```

Suggestions for improvements

```
1) Use more graphics
2) Use bullet points
```

LLM may get lost on complex (real world!) tasks.

Example: Peer review rewriting

Given the following feedback, rewrite it into a friendly paragraph with concrete suggestions for each of Alex's presentation problems.

Original feedback:
Alex could improve his presentation skills. He has too much text on his slides. His presentation meanders from topic to topic without a clear structure. He also does not engage with his audience when he presents.

More friendly feedback:

```
Alex, you have too many words on your slides. You should
use images and bullet points to help get your message
across. You should have a clear structure for the
presentation. You should also engage with your audience.
```

LLM may get lost on complex (real world!) tasks.

Example: Peer review rewriting

Given the following feedback, rewrite it into a friendly paragraph with concrete suggestions for each of Alex's presentation problems.

Original feedback:
Alex could improve his presentation skills. He has too much text on his slides. His presentation meanders from topic to topic without a clear structure. He also does not engage with his audience when he presents.

More friendly feedback:

```
Alex, you have too many words on your slides. You should
use images and bullet points to help get your message
across. You should have a clear structure for the
presentation. You should also engage with your audience.
```

× **Inherently multi-step**

1. Identify all presentation problems
2. Ideate suggestions per problem
3. Compose them into a paragraph

× **Too much freedom to debug!**

Need more suggestion on "engagement"

– How to change the prompt?

AI Chain ⛓️: Raise the ceiling of collaborating with LLMs, by using multiple LLM calls to solve distinct sub-tasks.

Instead of solving the task in one run…

Original feedback

```
Alex could improve his presentation skills. He has too much text on his slides. His presentation meanders
from topic to topic without a clear structure. He also does not engage with his audience when he presents.
```

Friendly paragraph

A three-step LLM Chain for peer review rewriting

Original feedback

```
Alex could improve his presentation skills. He has too much text on his slides. His presentation meanders
from topic to topic without a clear structure. He also does not engage with his audience when he presents.
```

1 Identify all presentation problems

2 Ideate suggestions per problem

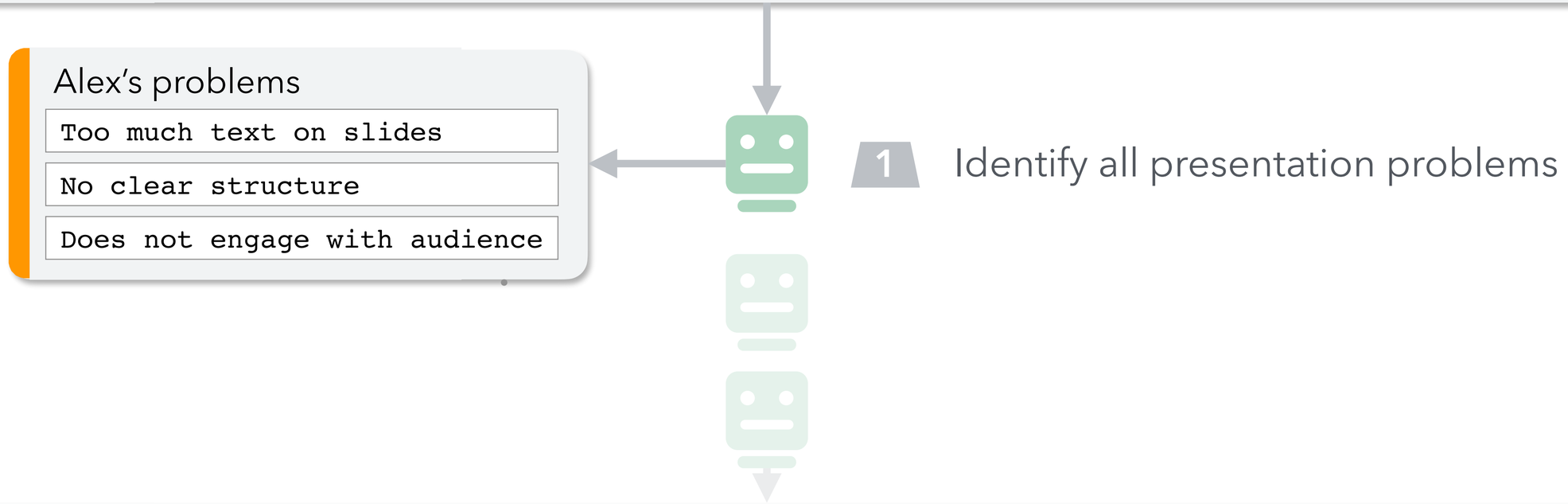
3 Compose them into a paragraph

Friendly paragraph

A three-step LLM Chain for peer review rewriting

Original feedback

```
Alex could improve his presentation skills. He has too much text on his slides. His presentation meanders
from topic to topic without a clear structure. He also does not engage with his audience when he presents.
```

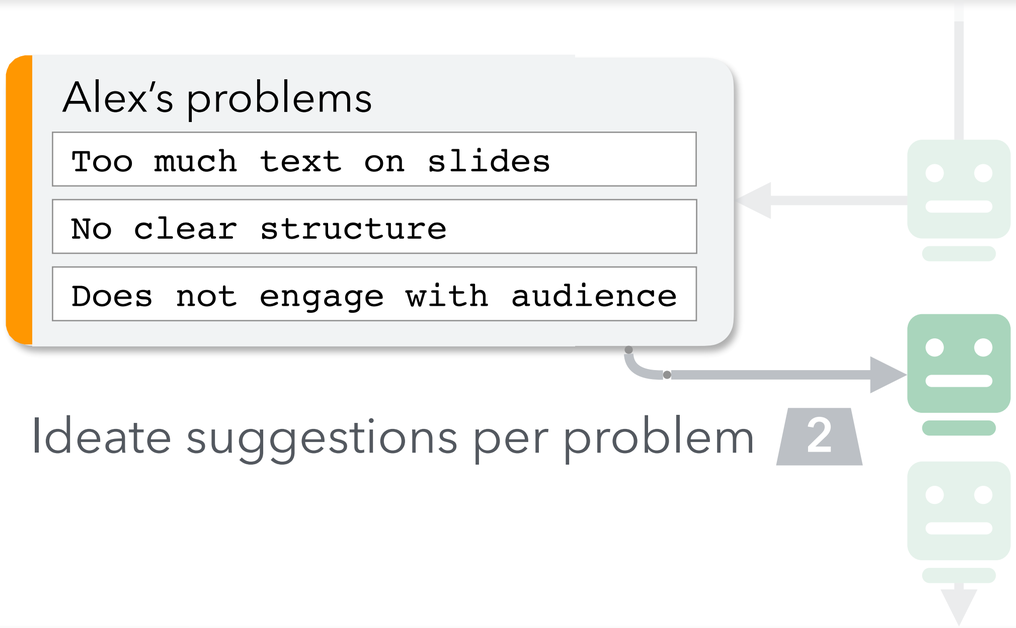


Friendly paragraph

A three-step LLM Chain for peer review rewriting

Original feedback

```
Alex could improve his presentation skills. He has too much text on his slides. His presentation meanders
from topic to topic without a clear structure. He also does not engage with his audience when he presents.
```

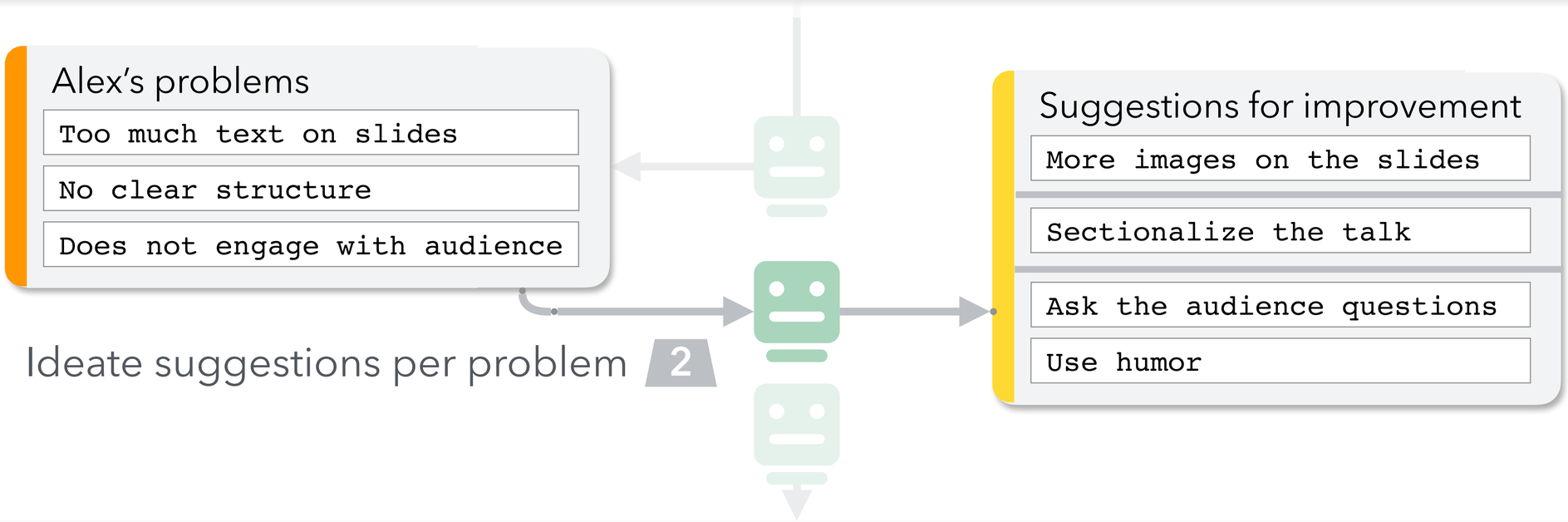


Friendly paragraph

A three-step LLM Chain for peer review rewriting

Original feedback

```
Alex could improve his presentation skills. He has too much text on his slides. His presentation meanders
from topic to topic without a clear structure. He also does not engage with his audience when he presents.
```

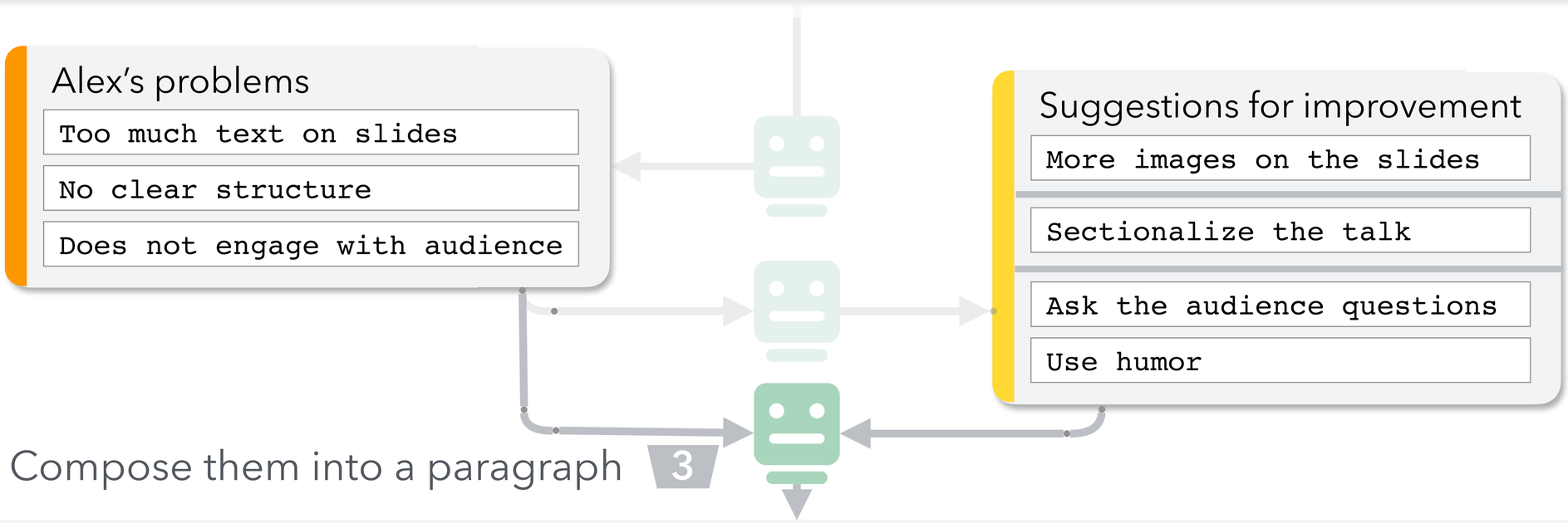


Friendly paragraph

A three-step LLM Chain for peer review rewriting

Original feedback

```
Alex could improve his presentation skills. He has too much text on his slides. His presentation meanders
from topic to topic without a clear structure. He also does not engage with his audience when he presents.
```

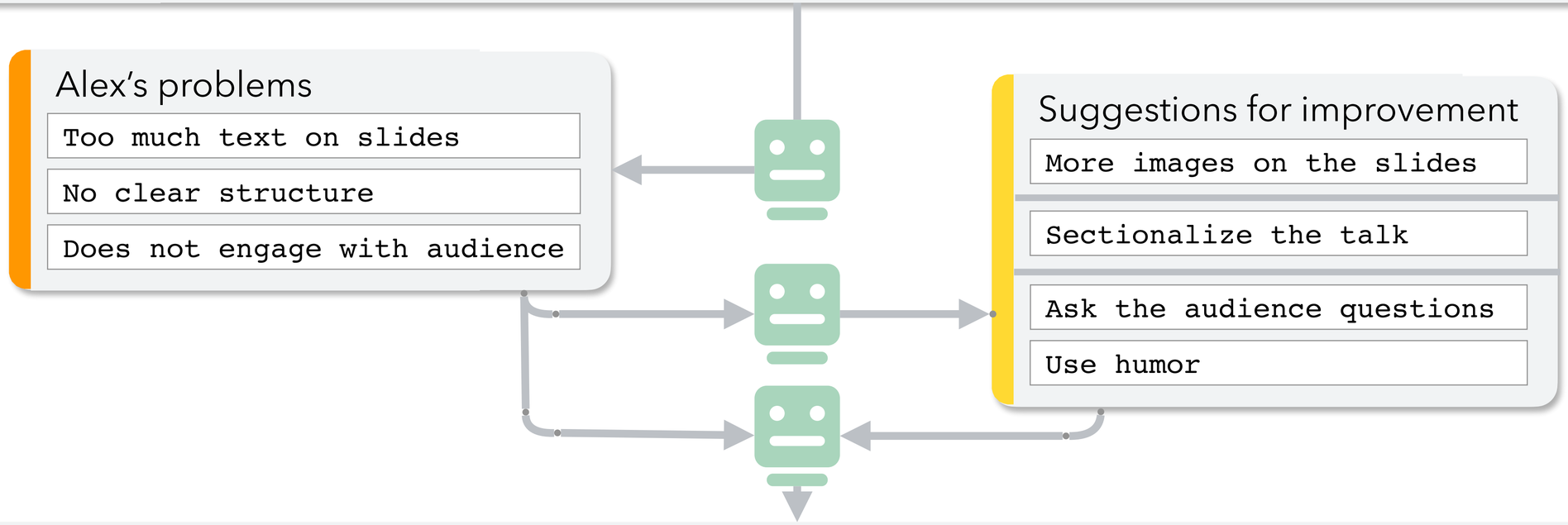


Original feedback

A three-step LLM Chain for peer review rewriting

Original feedback

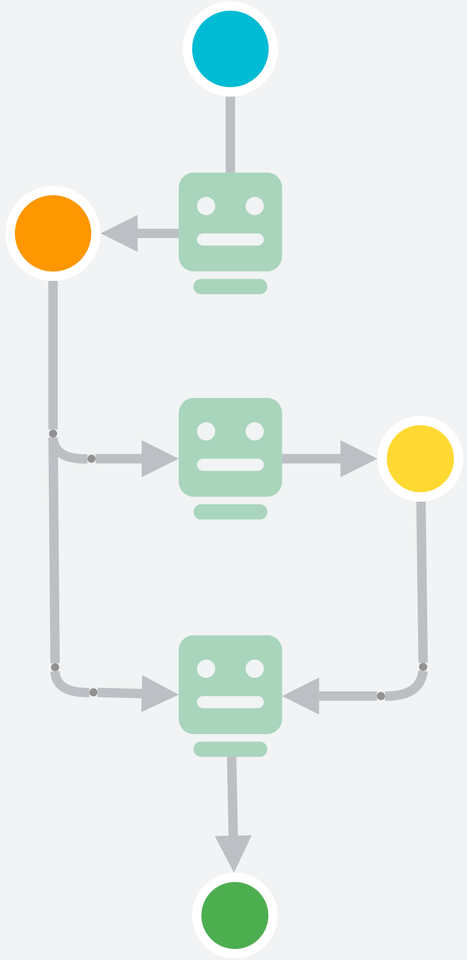
```
Alex could improve his presentation skills. He has too much text on his slides. His presentation meanders
from topic to topic without a clear structure. He also does not engage with his audience when he presents.
```



Friendly paragraph

```
Alex, your presentation was interesting! However, I noticed that you have a lot of information on your
slides. It might be helpful to vary pictures with text so that it is easier to follow. Also, you might
consider the flow of your theme. If it were me, I would have divided it into three sections and then used
your conclusion. You may also want to add some humor, and ask more questions to engage the audience.
```

LLM Chain improves task outcome.



vs. solving the entire task in one go…

Run the same model on multiple sub-tasks.

Each sub-task is more likely to succeed.

The aggregated final outcome is also improved.

How does Chaining impact collaboration?

User study

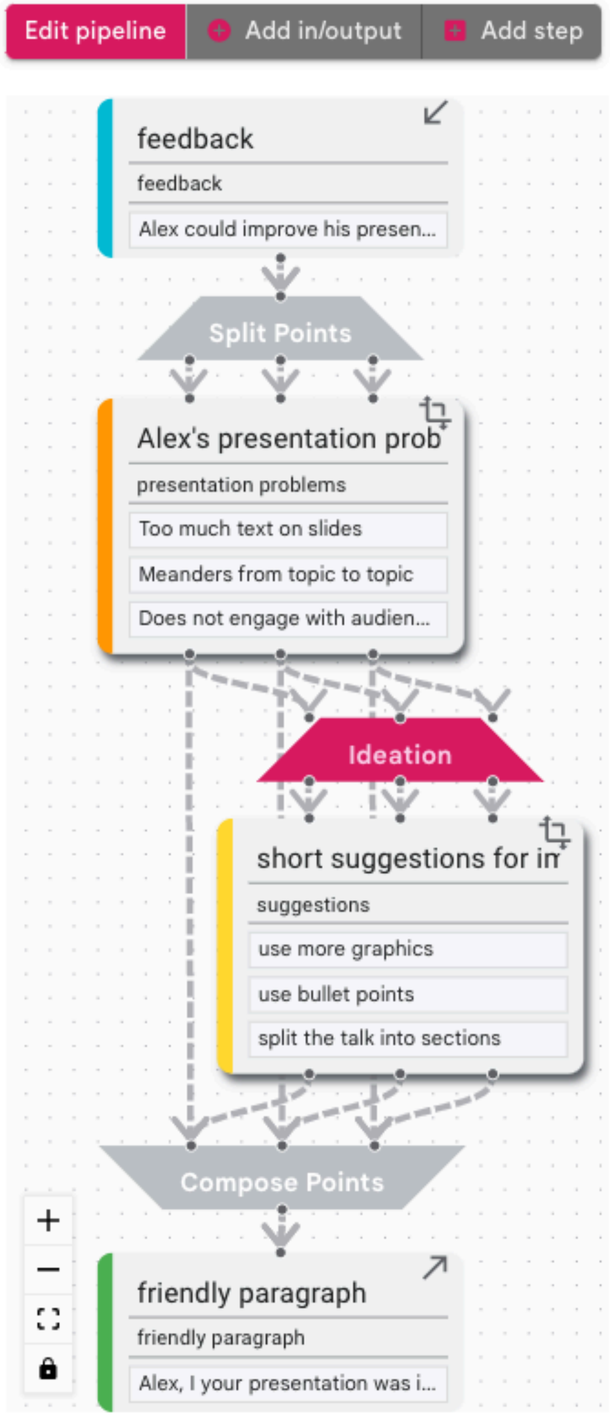
Two UI: Chaining vs. single run

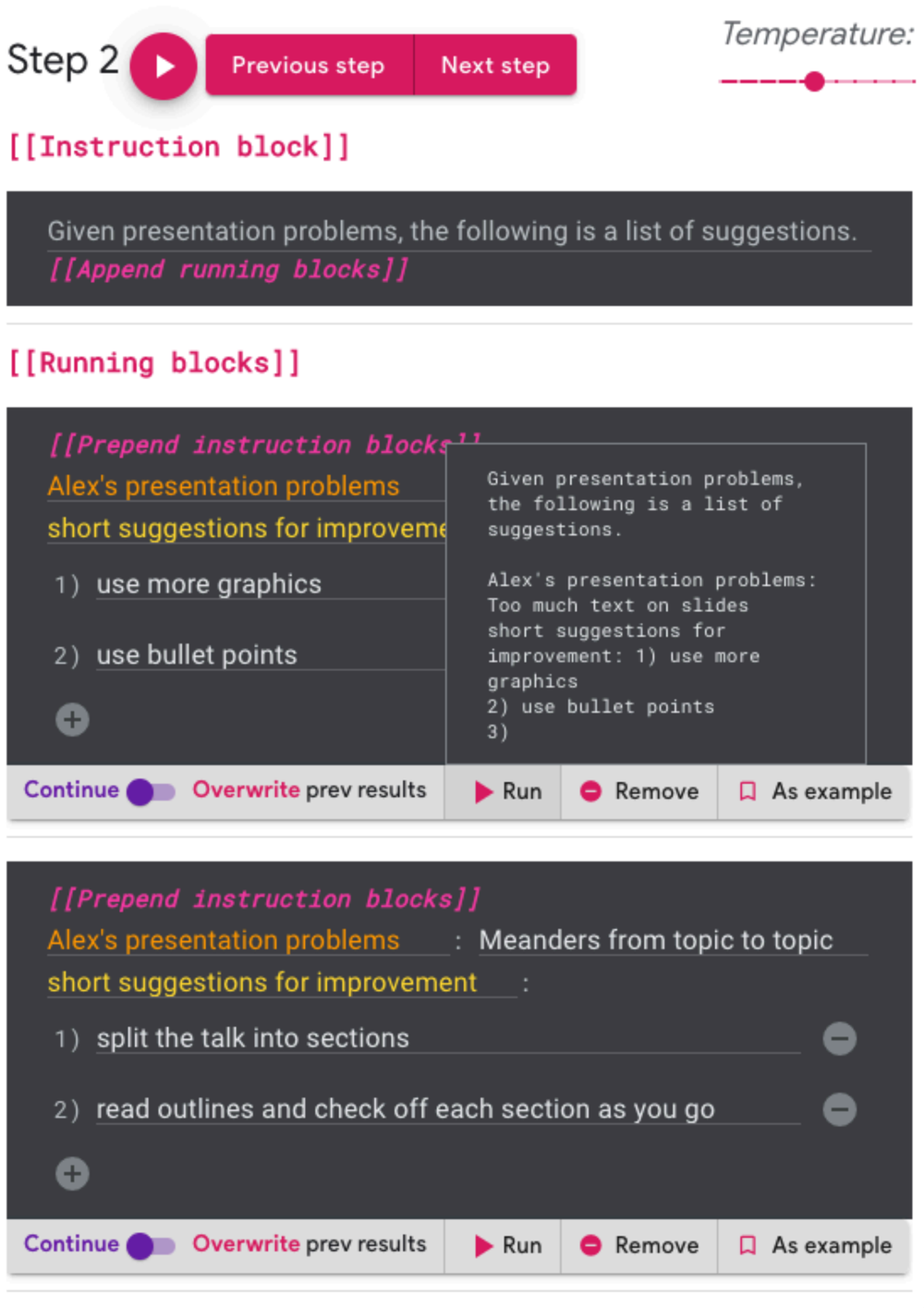
20 lay users of models

Within-subject

Complete the same task in two UIs

We find…



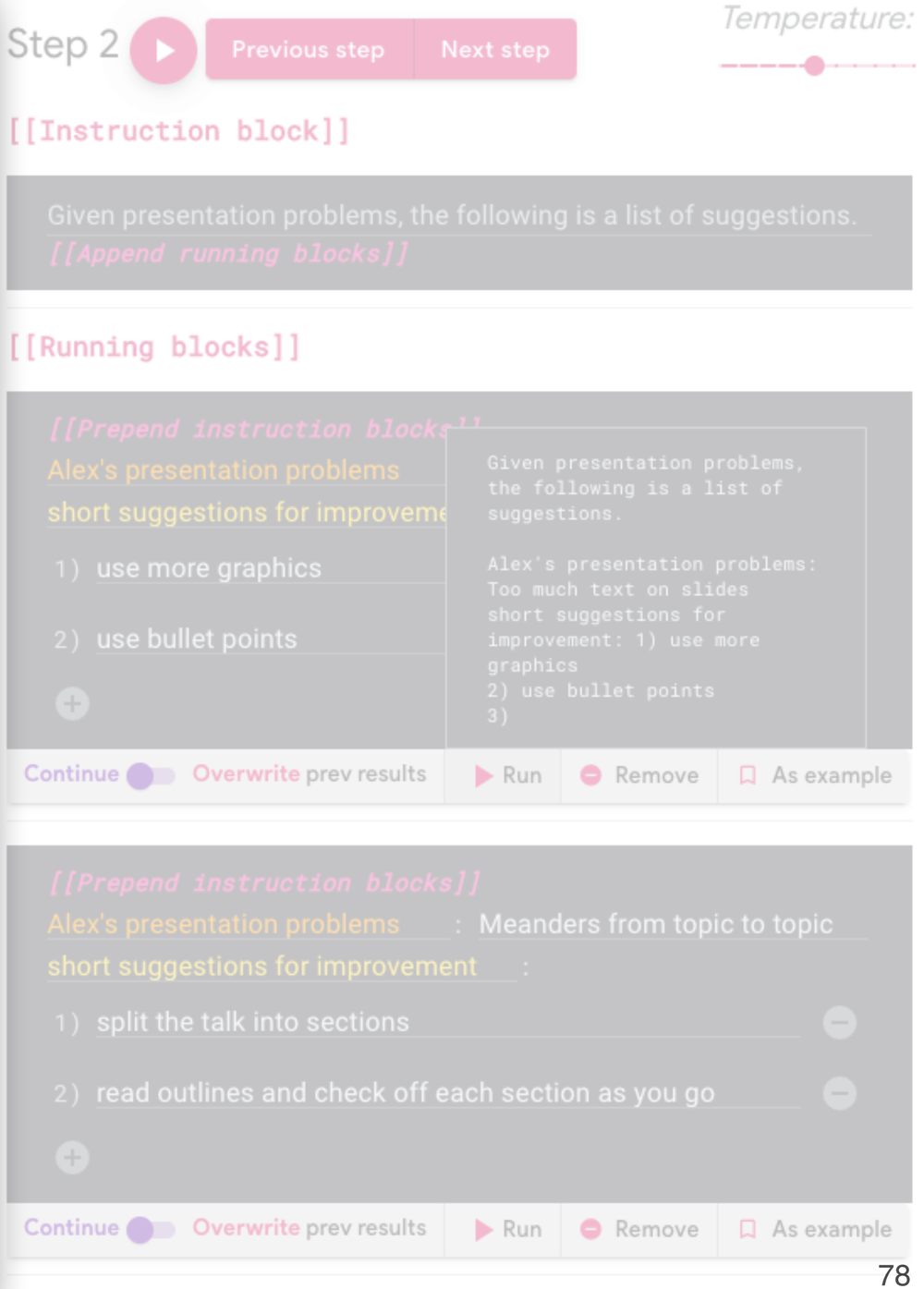


How does Chaining impact collaboration?

Think through the task

by tracking the progress

> *"if I want to do the task with precision, I prefer the Chaining structure."*



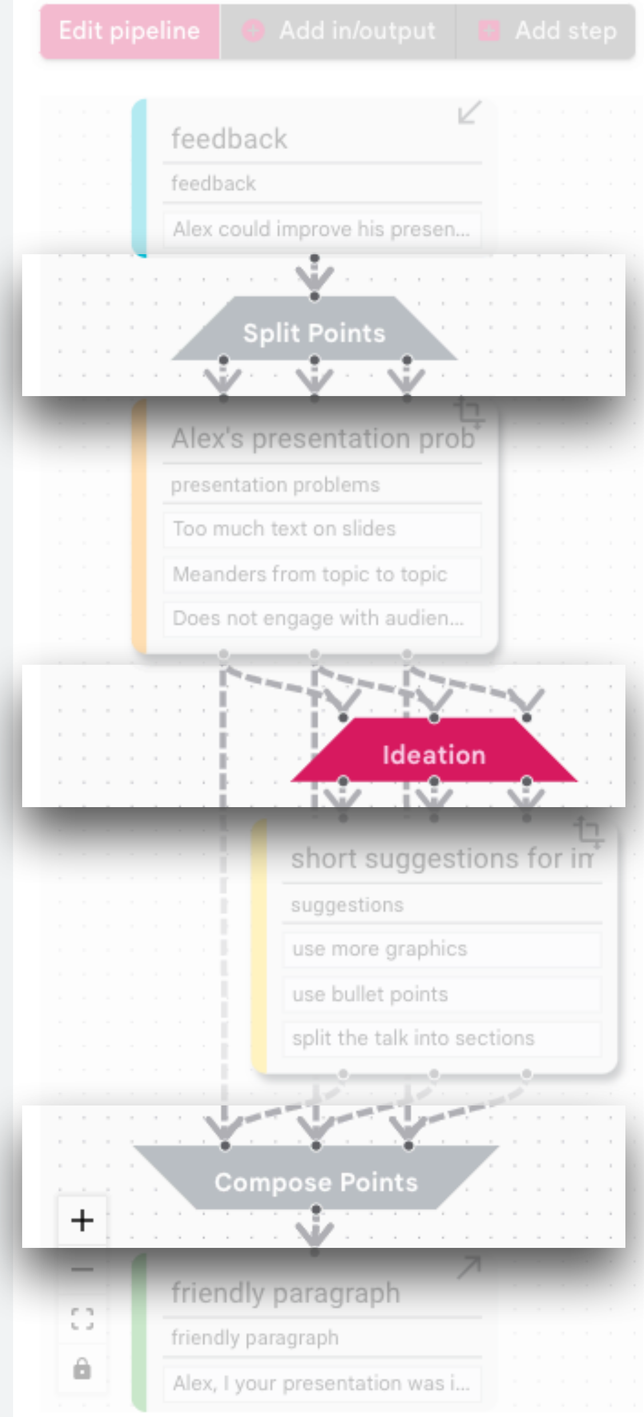
How does Chaining impact collaboration?

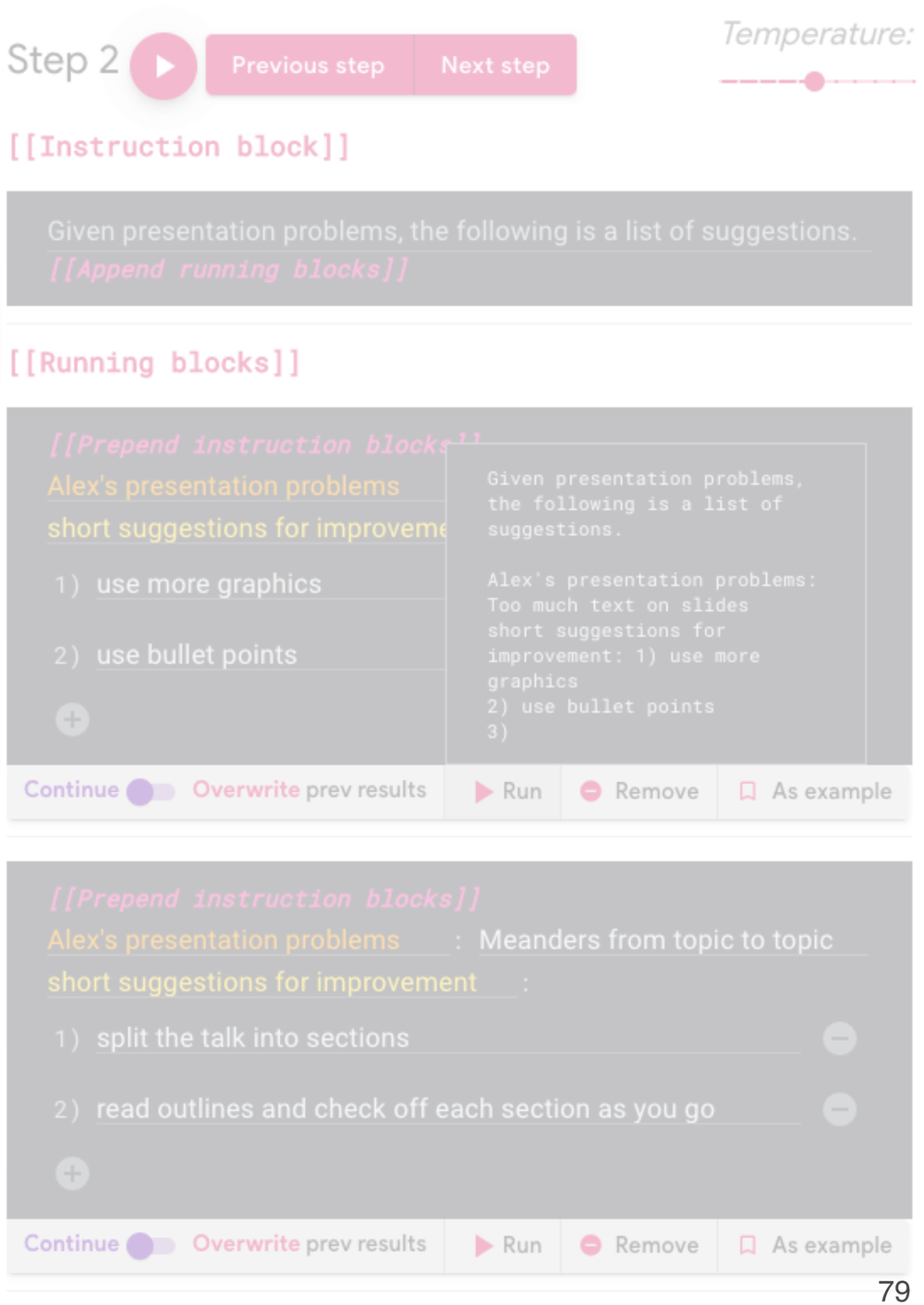
Think through the task

Get more help from models

with different capabilities highlighted

"I didn't need to give it as much, but it was giving me a lot."





How does Chaining impact collaboration?

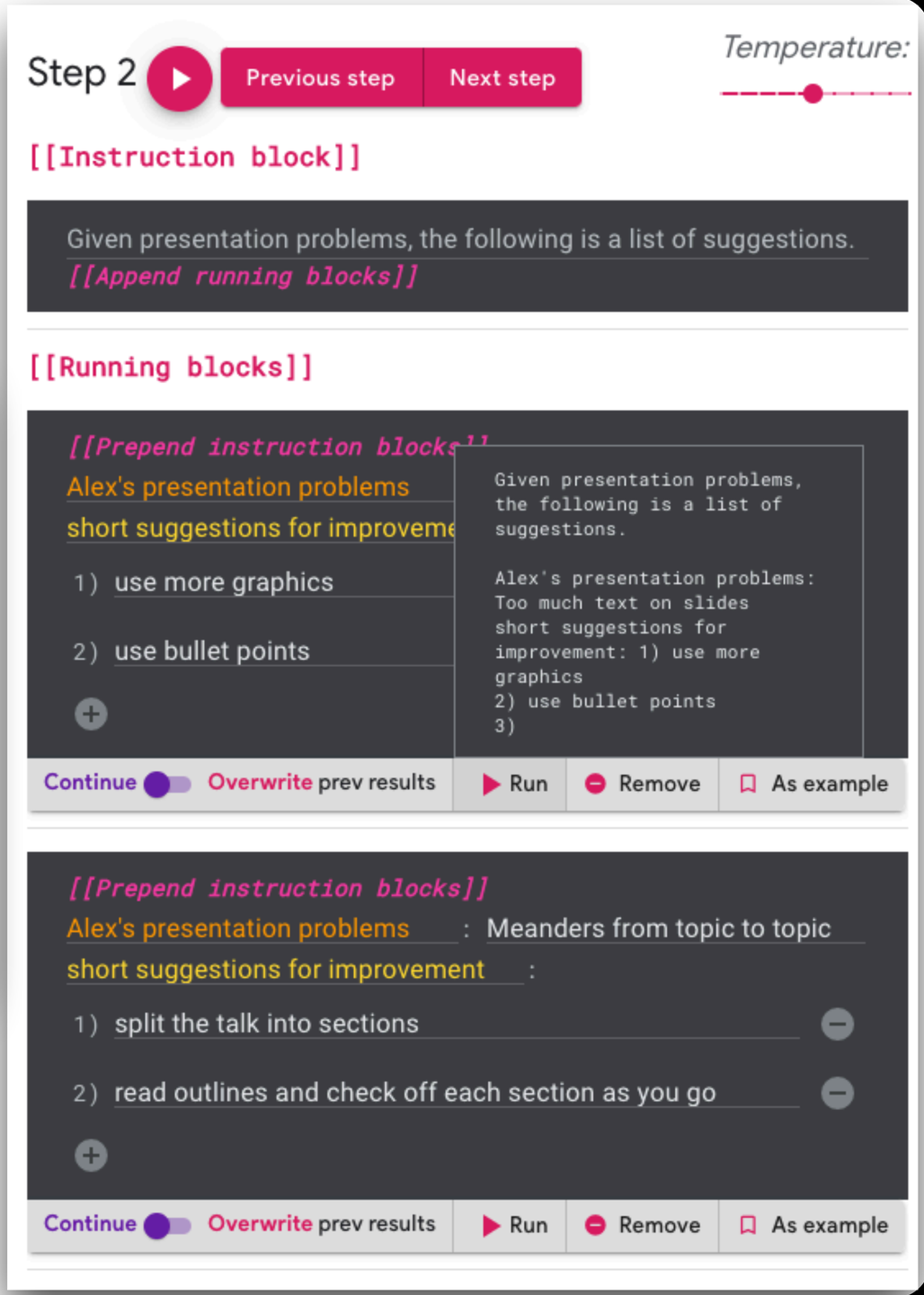
Think through the task

Get more help from models

System more transparent

better convey model's goal per step (vs. all-encompassing)

> *"Chaining helped you speak the language. It lift[ed] up the hood and showed what's happening at different phrases."*



How does Chaining impact collaboration?

Think through the task

Get more help from models

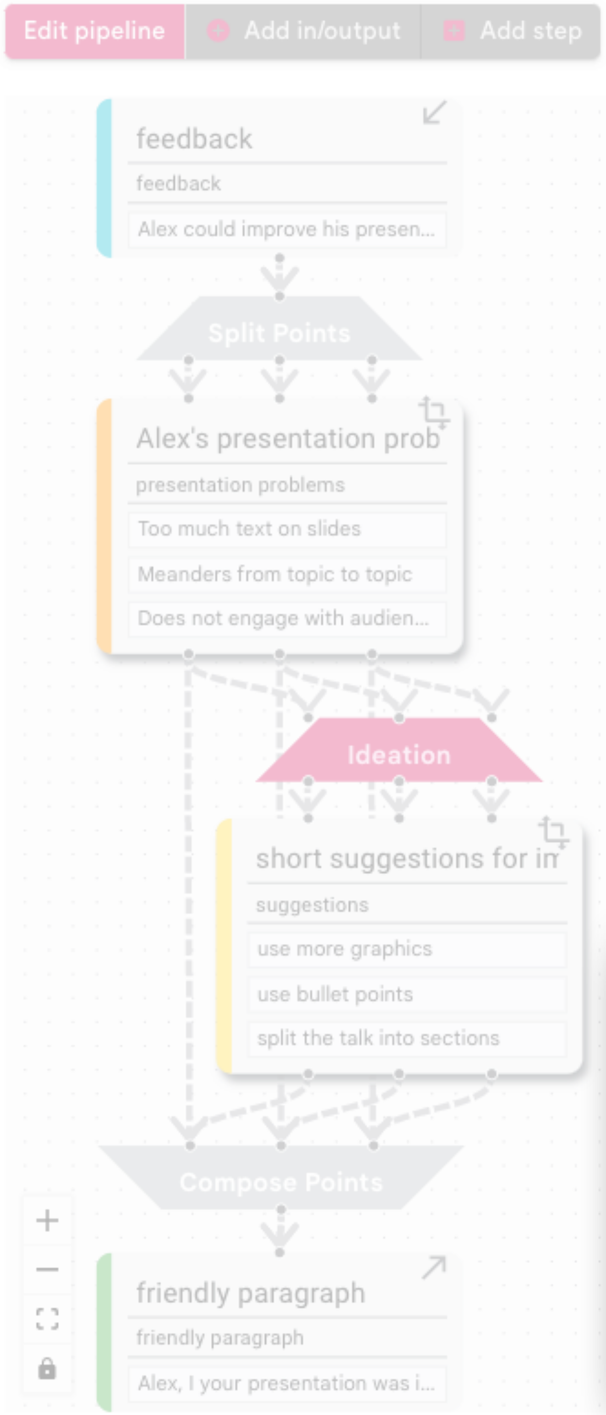
System more transparent

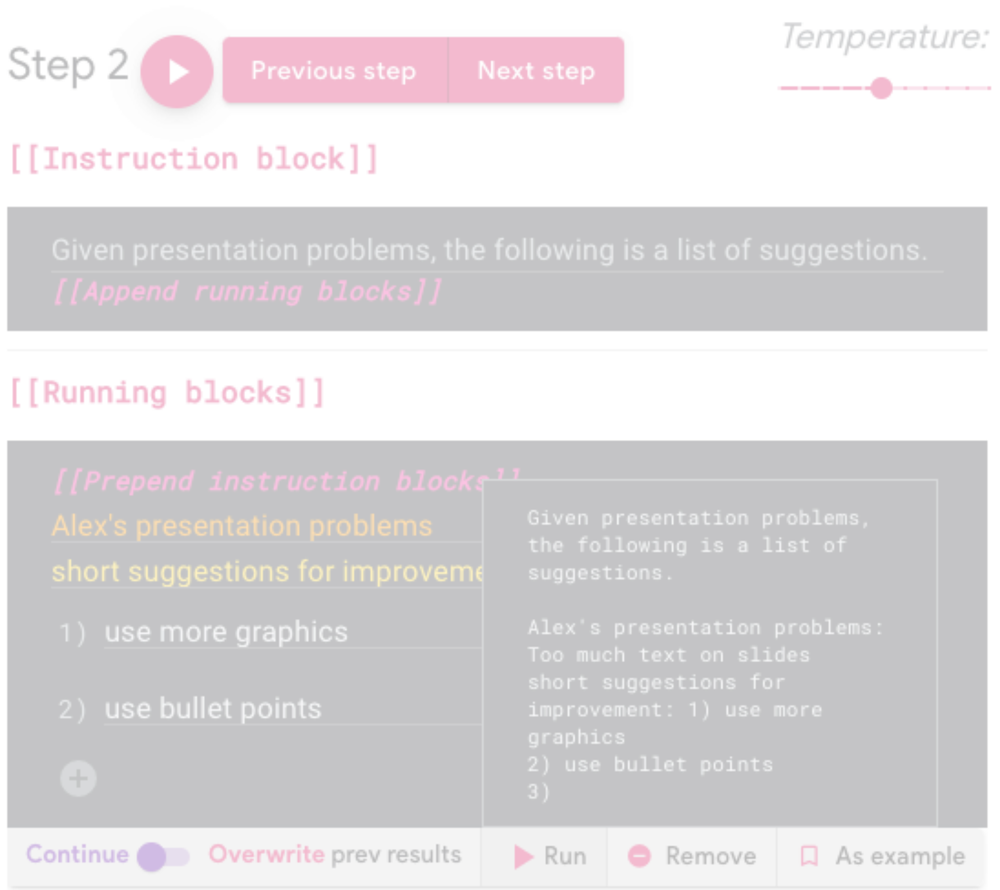
System more controllable

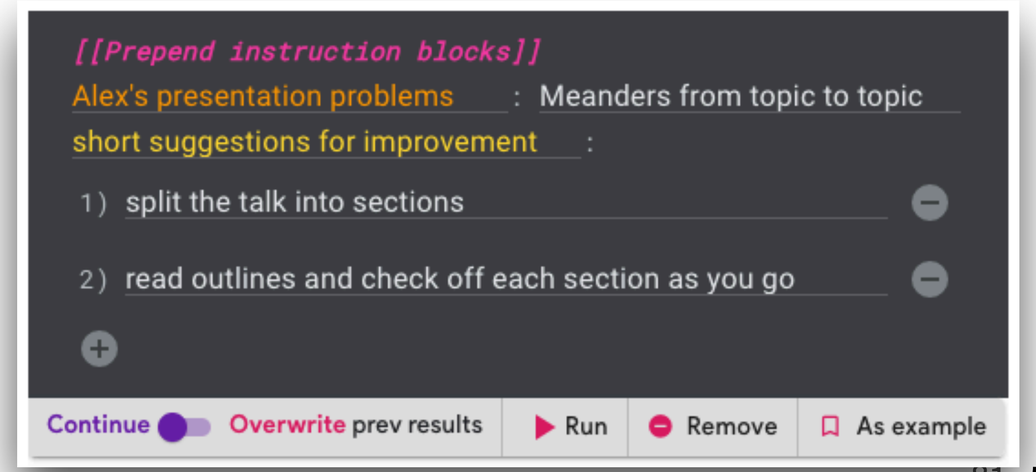
Modular feedback,

Explicit "knobs" on what can change.

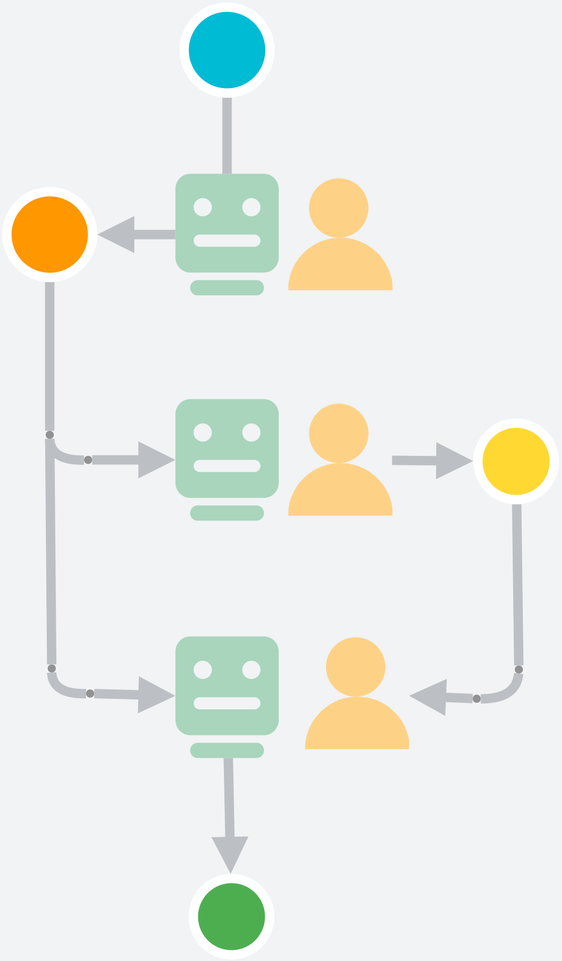
> *"Chaining makes it easier to fine-tune things…Too much freedom [without Chaining] is a curse."*







AI Chain: The Big Picture



AI Chain ⛓️: Raise the ceiling of collaborating with LLMs, by using multiple LLM calls to solve distinct sub-tasks.

Better task outcome, more controllable, transparent, etc.

Gains purely from novel interaction, no model retraining.

Things to consider in Human-AI collaboration research

Human-AI collaboration cares about appropriate reliance.

Which dimensions are we relying on the AI for?

What's the task?

- **Close-ended** decision making like classification
- **Open-ended** co-creation, in the Chaining case

What's AI's role?

- **Guardrail on quality** (grammar check)
- **Cognitive off-loader** (summarize large amount of data),
- **Assistant on tedious tasks** (e.g. fill out the color)
- **Leaders asking for human help** (active learning)

Things to consider in Human-AI collaboration research

Human-AI collaboration cares about appropriate reliance.

How are we building the reliance?

- “Shared representation”
  - Informative but intuitive explanations on model rationales
  - Chaining: natural language input-output
- Presenting shared representations
  - Explanation format, modality, timing
  - Structuring inputs and outputs

analyzer → analyzee

AI developers working on models-to-deploy,
Find why models fail & how to fix them.

Perform systematic analyses, among
various model development stages,
to prioritize more severe errors & root fixes.

+ collaborator

End users with deployed models,
work with AI towards common goals.

Maximize complementary skills, through
effective communication,
to achieve better results than either alone.

HCI people can do a lot more!

Consider domain experts.

What unique concerns do people have in domains like medical, education, etc.?
How should AI roles change accordingly? (When to lead, when to assist, etc.)

Do more data work.

Can we directly design data collection and evaluation workflows, to combat the root of "garbage-in, garbage-out"?

Clarify under-specified big terms in human-AI interaction.

What do we mean by trust, ethics, etc.? Is human-human interaction a good role model for human-AI interaction?