



17-445/645 ML in Production - guest lecture

Debugging with Better Slicing, Perturbation, and Interaction

Sherry Tongshuang Wu

Human Computer Interaction Institute

@tongshuangwu / sherryw@cs.cmu.edu

How do I check if my model works?



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 Chomsky
Arcadia Research
chomsky@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

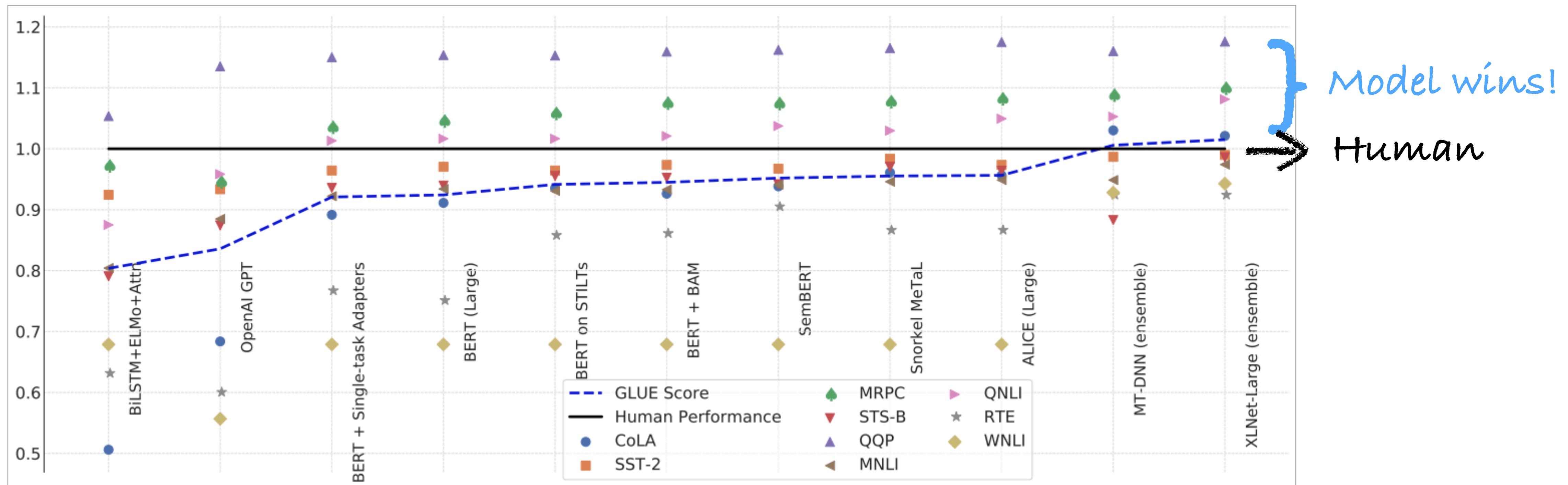
3 Model and Architecture

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 seems a good solution?

GLUE: “performance on the benchmark has recently come close to the level of non-expert humans, suggesting limited headroom for further research.”



What could go wrong?

Missing critical data slices (bias, fairness)

What's in the figure?



> ceremony, wedding, bride, groom, dress

> person, people

High accuracy ≠ Model succeeding.

Miss certain types of data!

Shortcuts/right for wrong reasons



What is the moustache made of?

> Banana

What are the eyes made of?

> Banana



What is?

> Banana

What?

> Banana

Model analysis reveals hidden deficiencies.

High accuracy ≠ Model succeeding.

Miss certain types of data!

Correct prediction ≠ correct reasoning.

Rely on irrelevant features!

Debug models beyond accuracy: Reveal Deficiencies.

High accuracy \neq Model succeeding.

Miss certain types of data!



Quantitative grouping

Inspect similar instances

Correct prediction \neq correct reasoning.

Rely on irrelevant features!



Counterfactual perturbation

Isolate important components

Debug models beyond accuracy: Reveal Deficiencies.

Capability	Min Func Test	INVariance	DIRectional
Vocabulary	Fail. rate=15.0%	16.2%	C 34.6%
NER	0.0%	B 20.8%	N/A
Negation	A 76.4%	N/A	N/A
...			

Test case	Expected	Predicted	Pass?
A Testing Negation with MFT	Labels: negative, positive, neutral		
Template: I {NEGATION} {POS_VERB} the {THING}.			
I can't say I recommend the food.	neg	pos	x
I didn't love the flight.	neg	neutral	x
...			
Failure rate = 76.4%			
B Testing NER with INV	Same pred. (inv) after removals / additions		
@AmericanAir thank you we got on a different flight to [Chicago → Dallas].	inv	pos neutral	x
@VirginAmerica I can't lose my luggage, moving to [Brazil → Turkey] soon, ugh.	inv	neutral neg	x
...			
Failure rate = 20.8%			
C Testing Vocabulary with DIR	Sentiment monotonic decreasing (↓)		
@AmericanAir service wasn't great. You are lame.	↓	neg neutral	x
@JetBlue why won't YOU help them?! Ugh. I dread you.	↓	neg neutral	x
...			
Failure rate = 34.6%			

Quantitative grouping

Inspect similar instances

Unit tests: known in-/out-puts → Minimum Functionality Test

Counterfactual perturbation

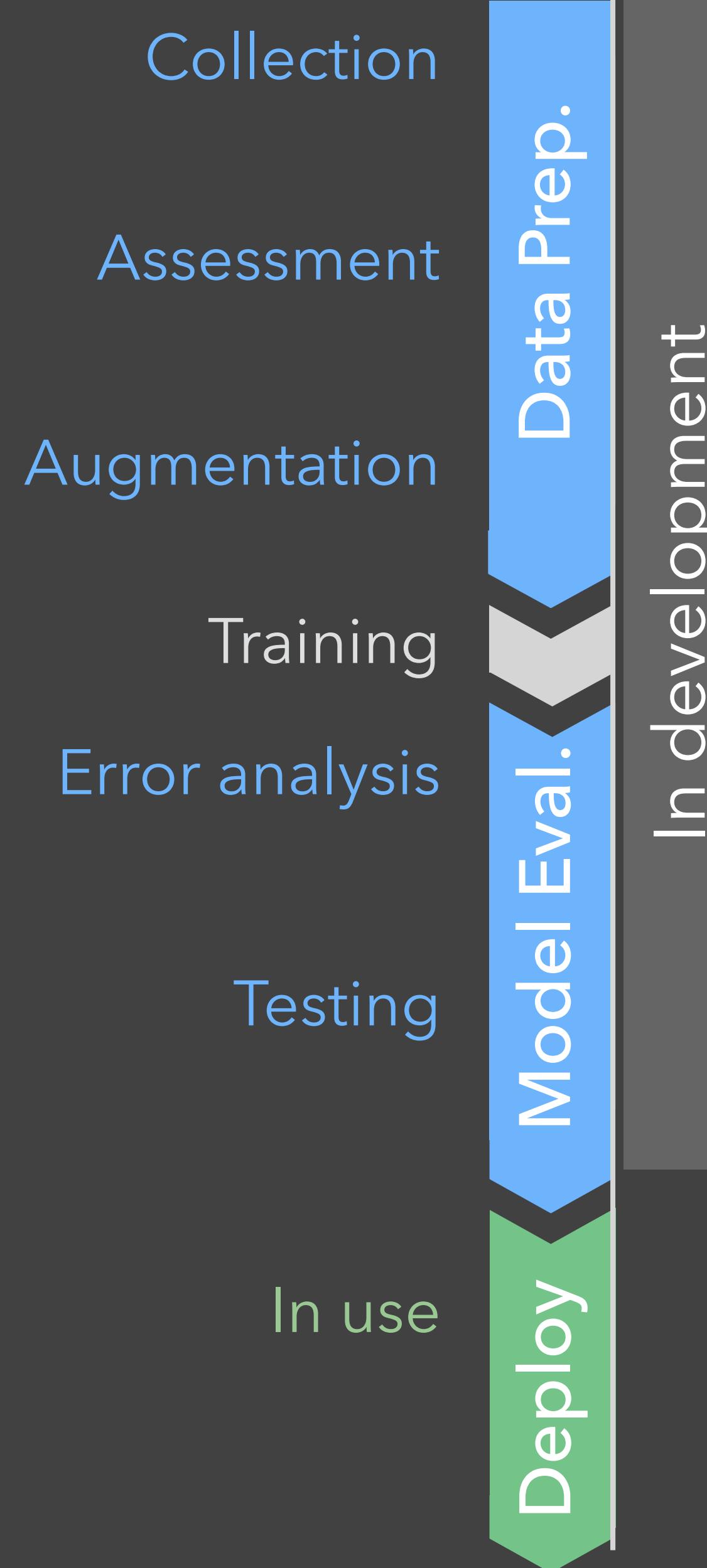
Isolate important components

Metamorphic (perturbations) &
property-based testing

INVariance Tests

DIRectional Expectation Tests

But CheckList requires good slicing and
perturbation & people need help.



Debugging is multi-stage.

Mitigate developer bias through assisted methods

Slice discovery that exposes inherent dataset property

Utilize task characteristics

Perturbation that compensates human creativity

Use NLP models as different assistive roles

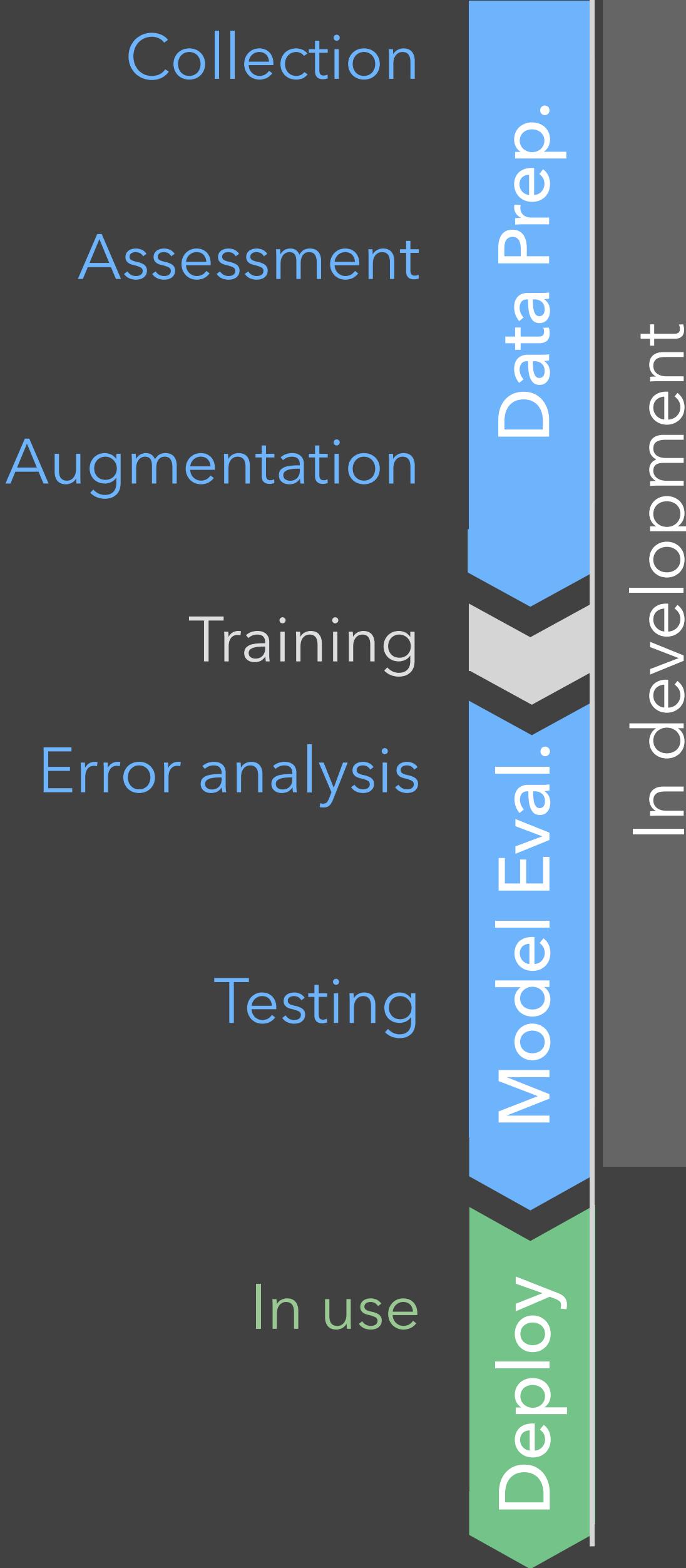
Help end users to recover from deployed model errors

Help users identify AI errors

Provide explanation and reflection

Help users work around AI errors

Provide users with controls through interactions



Debugging is multi-stage.

Mitigate developer bias through assisted methods

Slice discovery that exposes inherent dataset property

Utilize task characteristics

Perturbation that compensates human creativity

Use NLP models as different assistive roles

Help end users to recover from deployed model errors

Help users identify AI errors

Provide explanation and reflection

Help users work around AI errors

Provide users with controls through interactions

Text is high-dimensional and slicing is hard.

Manual labeling is tedious.

Filtering functions are biased.

Clustering is uninterpretable.

Capability	%	Keywords
negation	51.6	not, n't
negation (v2)	18.7	no, never, neither, nobody, none, nor, nothing

How rich is Jeff Bezos?

How rich is Bill Gates ?

How old is Bill Gates ?

But text have useful structure in some tasks!

e.g. Virtual assistants like Siri...

Compared to longer forms of text
(from articles/books),

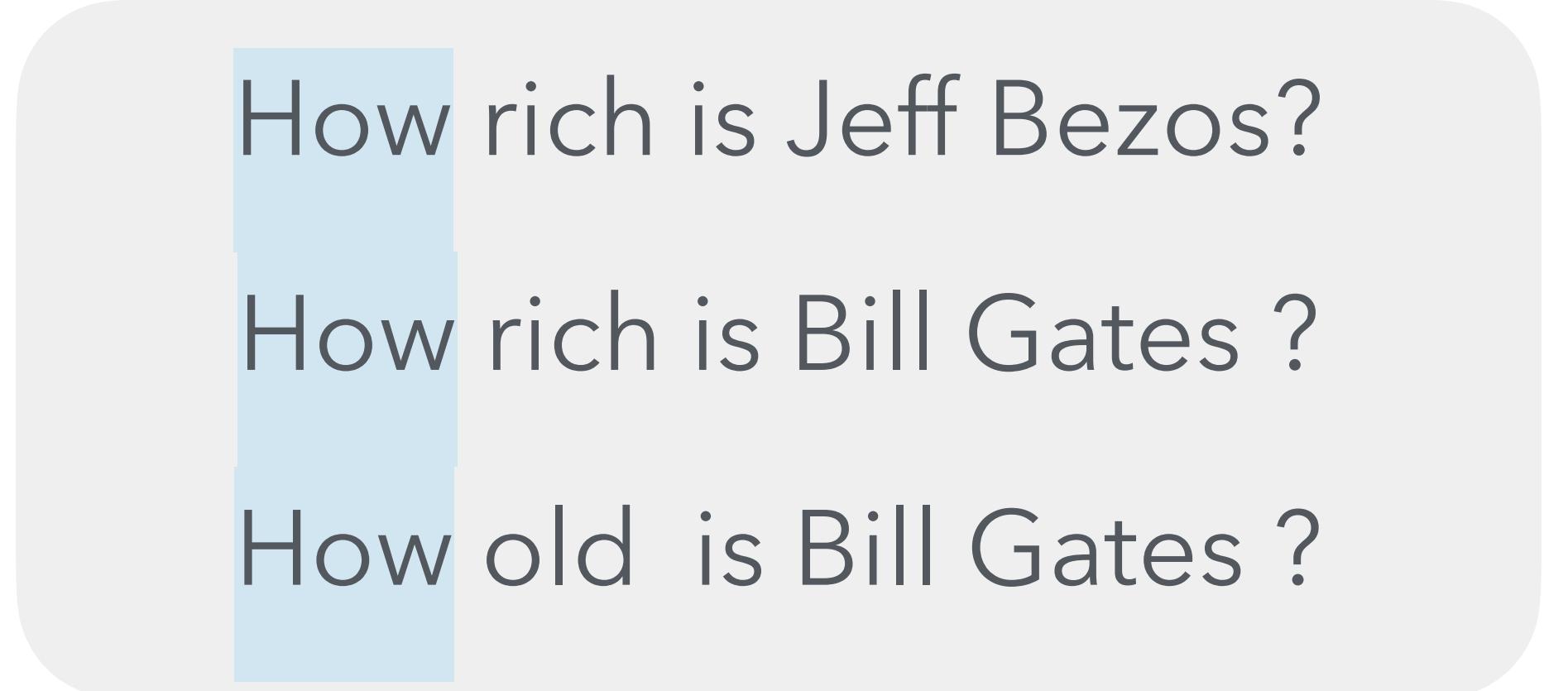
they share similar & compact structures.

How rich is Jeff Bezos?
How rich is Bill Gates ?
How old is Bill Gates ?

But text have useful structure in some tasks!

e.g. Virtual assistants like Siri...

Compared to longer forms of text
(from articles/books),
they share similar & compact structures.



How rich is Jeff Bezos?
How rich is Bill Gates ?
How old is Bill Gates ?

Question word

But text have useful structure in some tasks!

e.g. Virtual assistants like Siri...

Compared to longer forms of text
(from articles/books),

they share similar & compact structures.

How rich is Jeff Bezos?

How rich is Bill Gates ?

How old is Bill Gates ?

Adjective

But text have useful structure in some tasks!

e.g. Virtual assistants like Siri...

Compared to longer forms of text
(from articles/books),

they share similar & compact structures.

How rich is Jeff Bezos?

How rich is Bill Gates ?

How old is Bill Gates ?

Human names

Take advantage of the structure: group query data with **structural templates**.

Tempura: Query Analysis with Structural Templates

Tongshuang Wu, Kanit (Ham) Wongsuphasawat, Donghao Ren, Kayur Patel, Chris DuBois

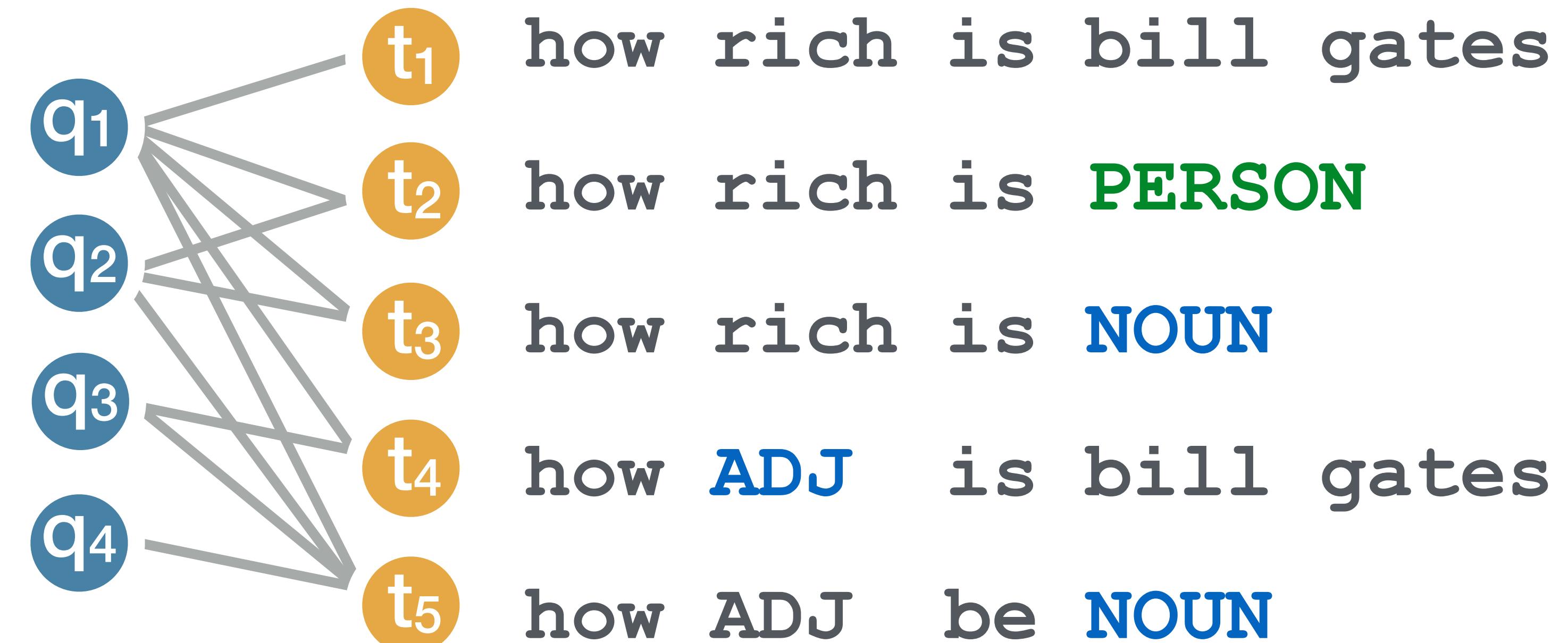
Structural template grouping

How rich is Bill Gates

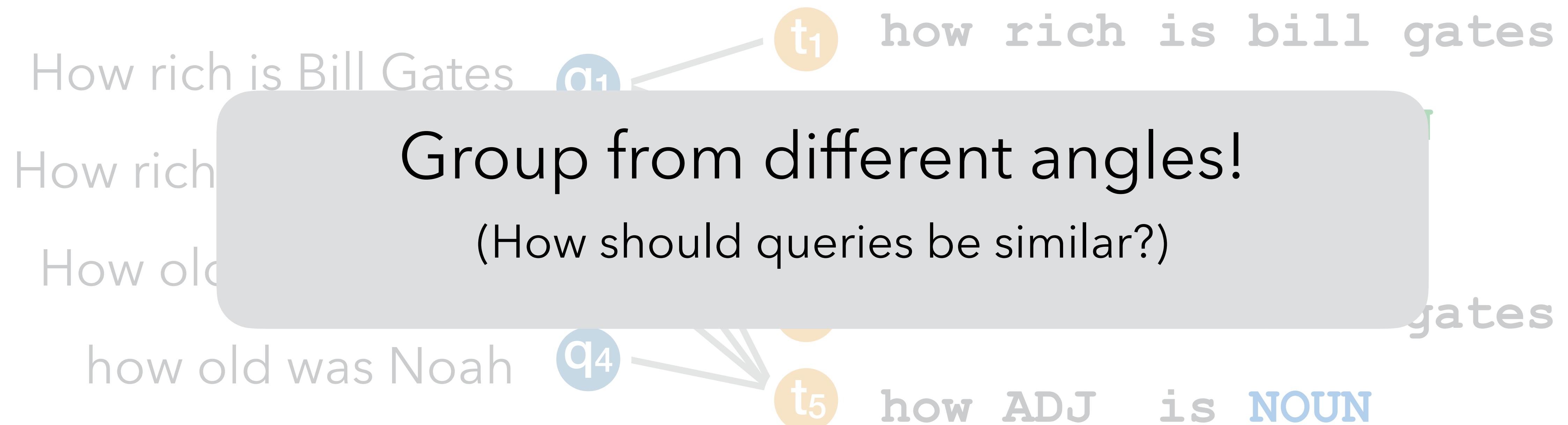
How rich is Jeff Bezos

How old is Bill Gates

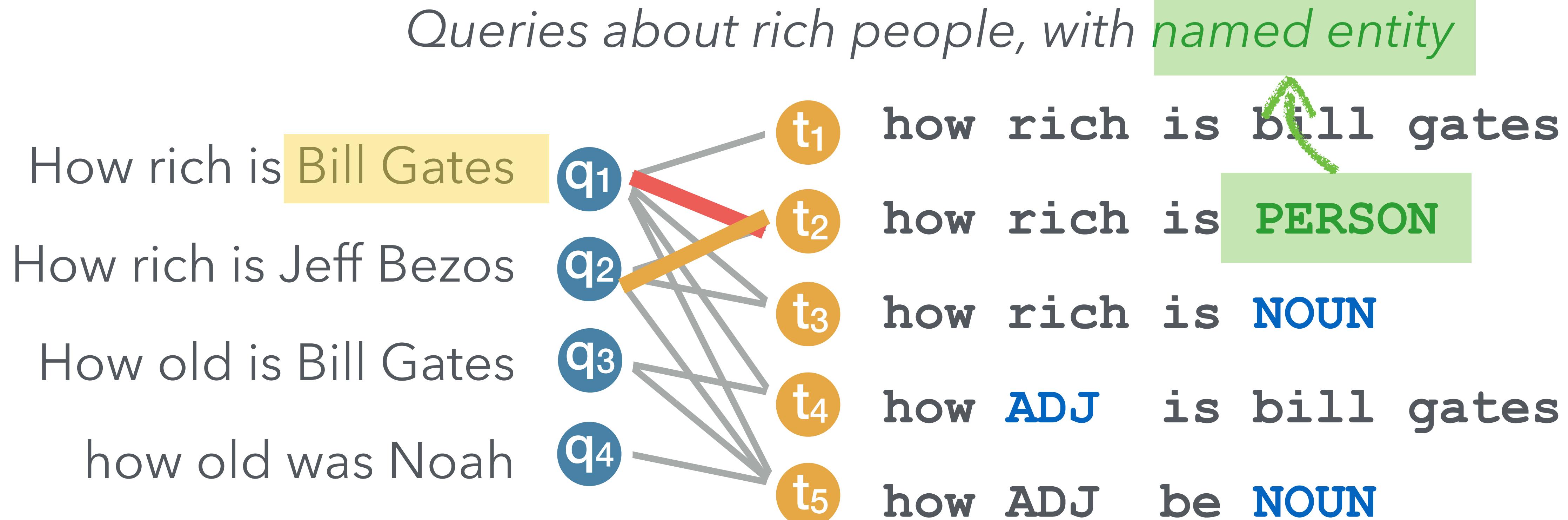
how old was Noah



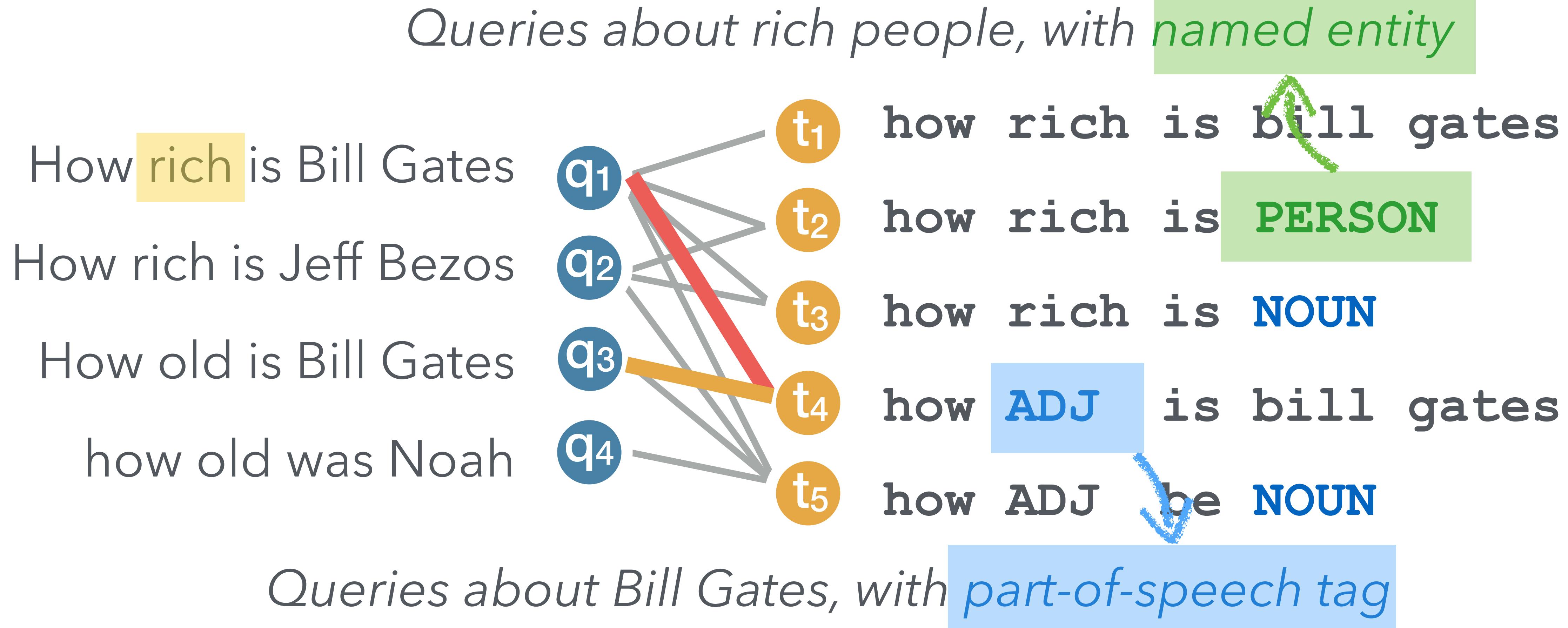
Structural template grouping



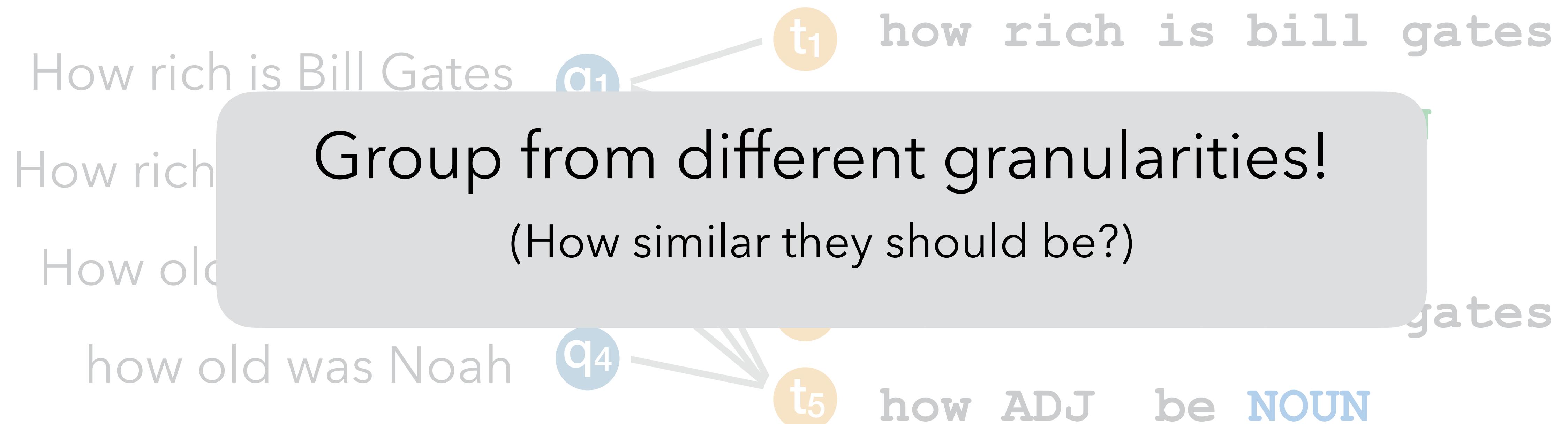
Structural template grouping



Structural template grouping



Structural template grouping



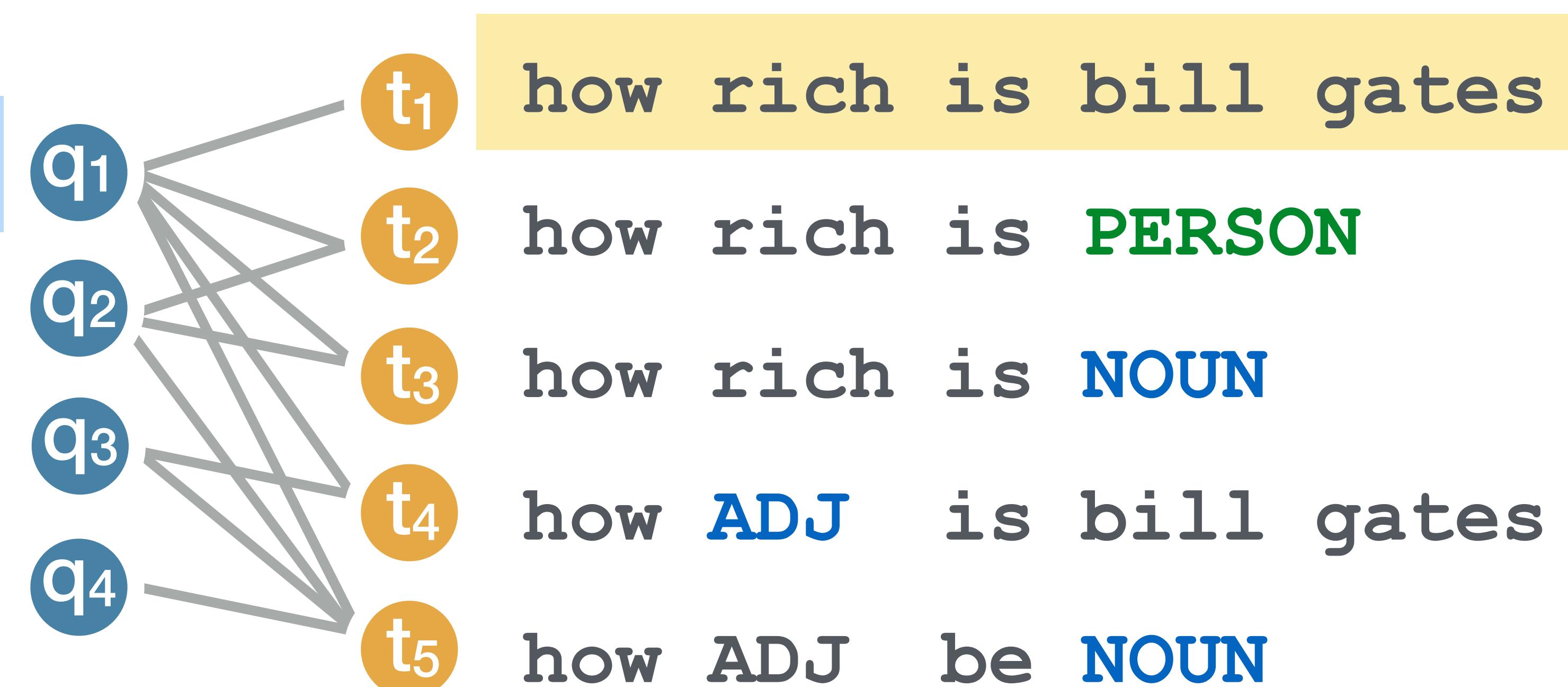
Structural template grouping

How rich is Bill Gates

How rich is Jeff Bezos

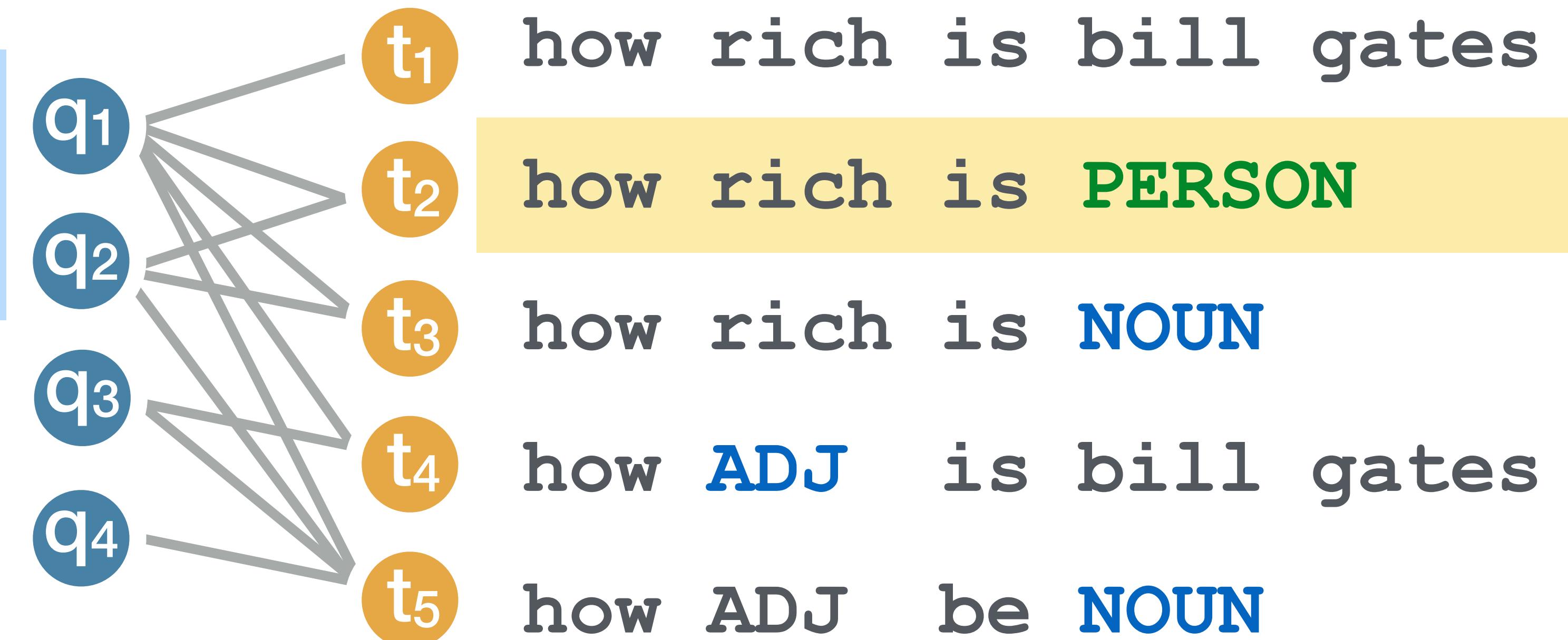
How old is Bill Gates

how old was Noah



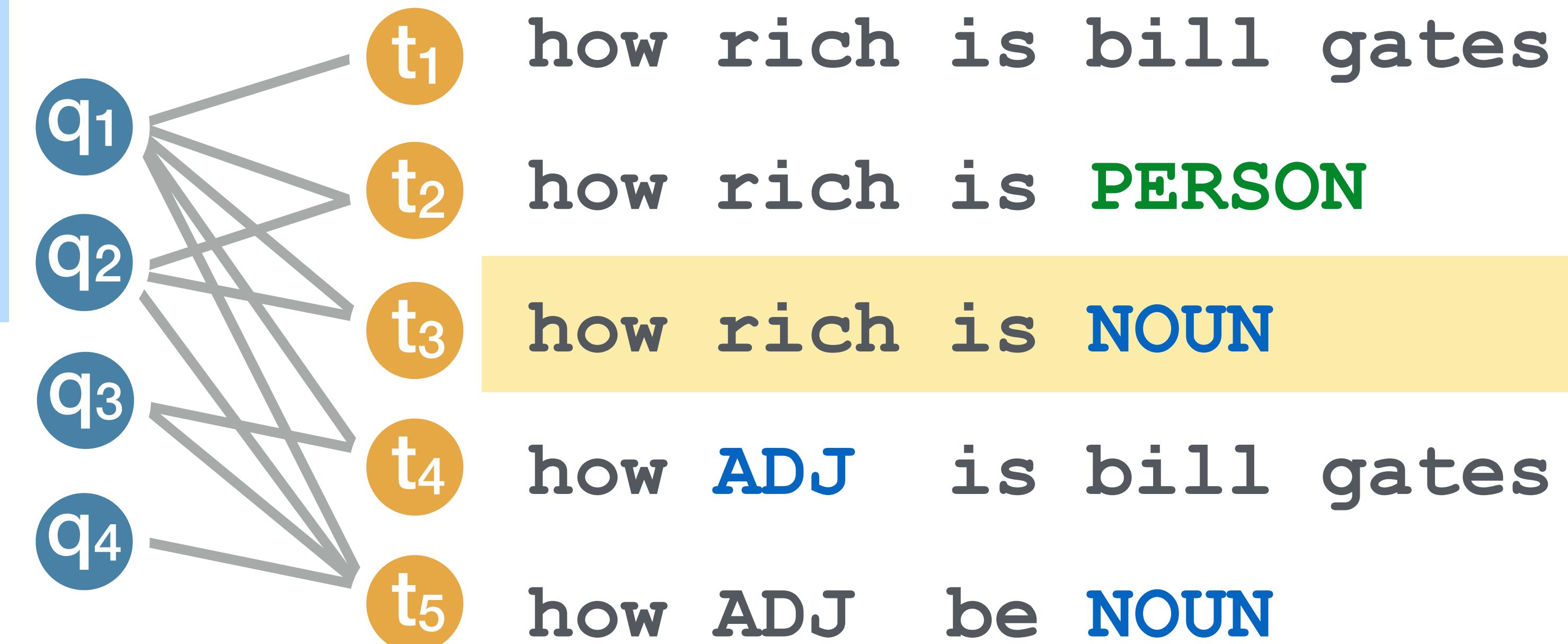
Structural template grouping

How rich is Bill Gates
How rich is Jeff Bezos
How old is Bill Gates
how old was Noah



Structural template grouping

How rich is feta cheese
How rich is Bill Gates
How rich is Jeff Bezos
How old is Bill Gates
how old was Noah



Structural templates can support...

Error analysis

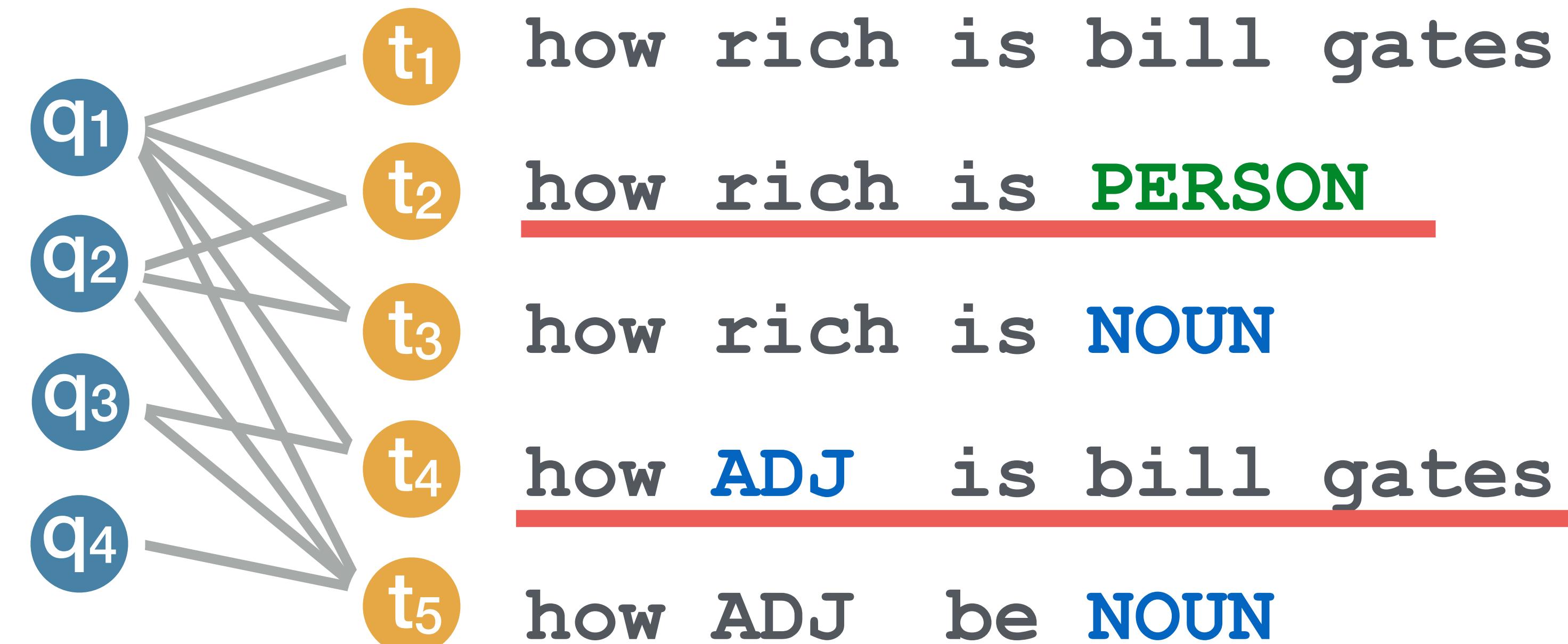
My model is wrong because...

How rich is Bill Gates

How rich is Jeff Bezos

How old is Bill Gates

how old was Noah



Structural templates can support...

Error analysis

Data exploration

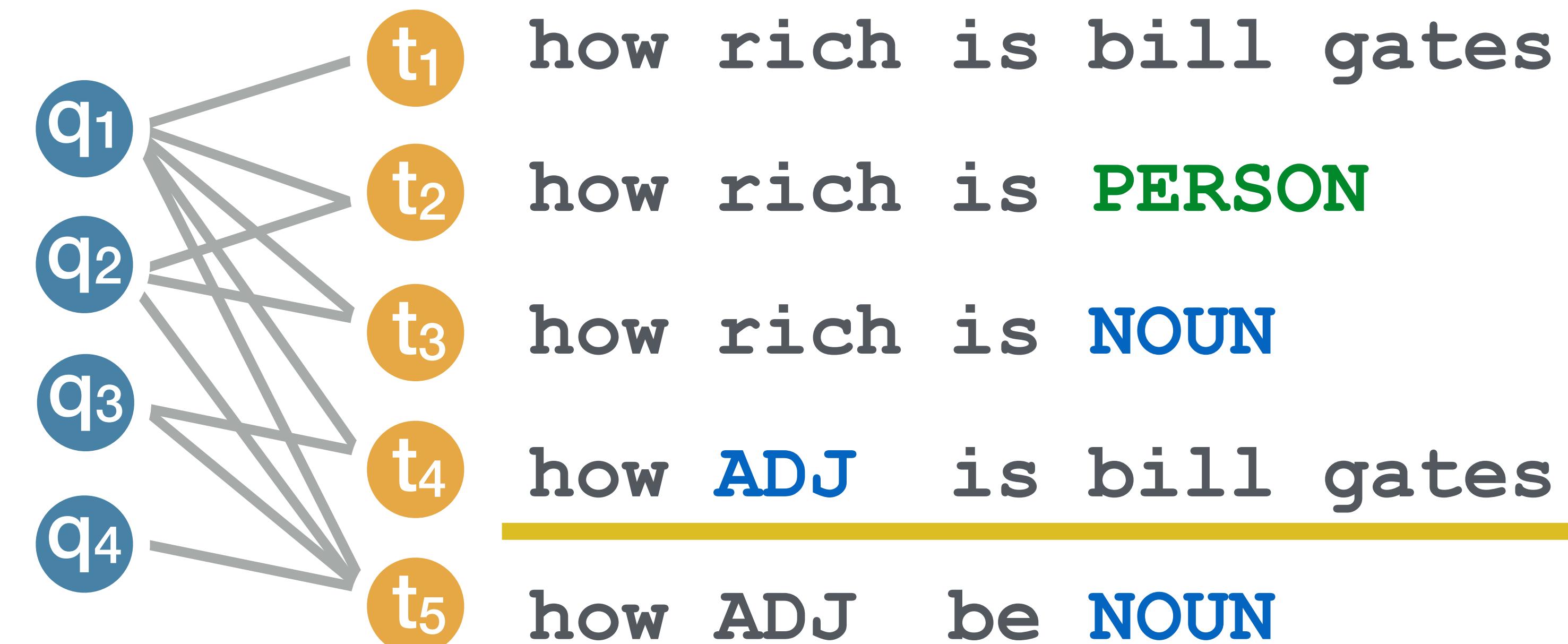
My dataset has a lot of ...

How rich is Bill Gates

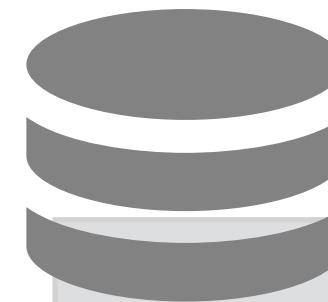
How rich is Jeff Bezos

How old is Bill Gates

how old was Noah



Case Study: Templates from different datasets



10,000 queries from each dataset...

MS MARCO

Bing Search Engine, Nguyen et al. (2016)

Assistant

Internal, anonymized search queries
from a commercial conversational assistant



All templates

how **ADJ** **VERB** **NOUN**
how rich be **NOUN**
how rich be bill gates
how old be bill gates

Overview

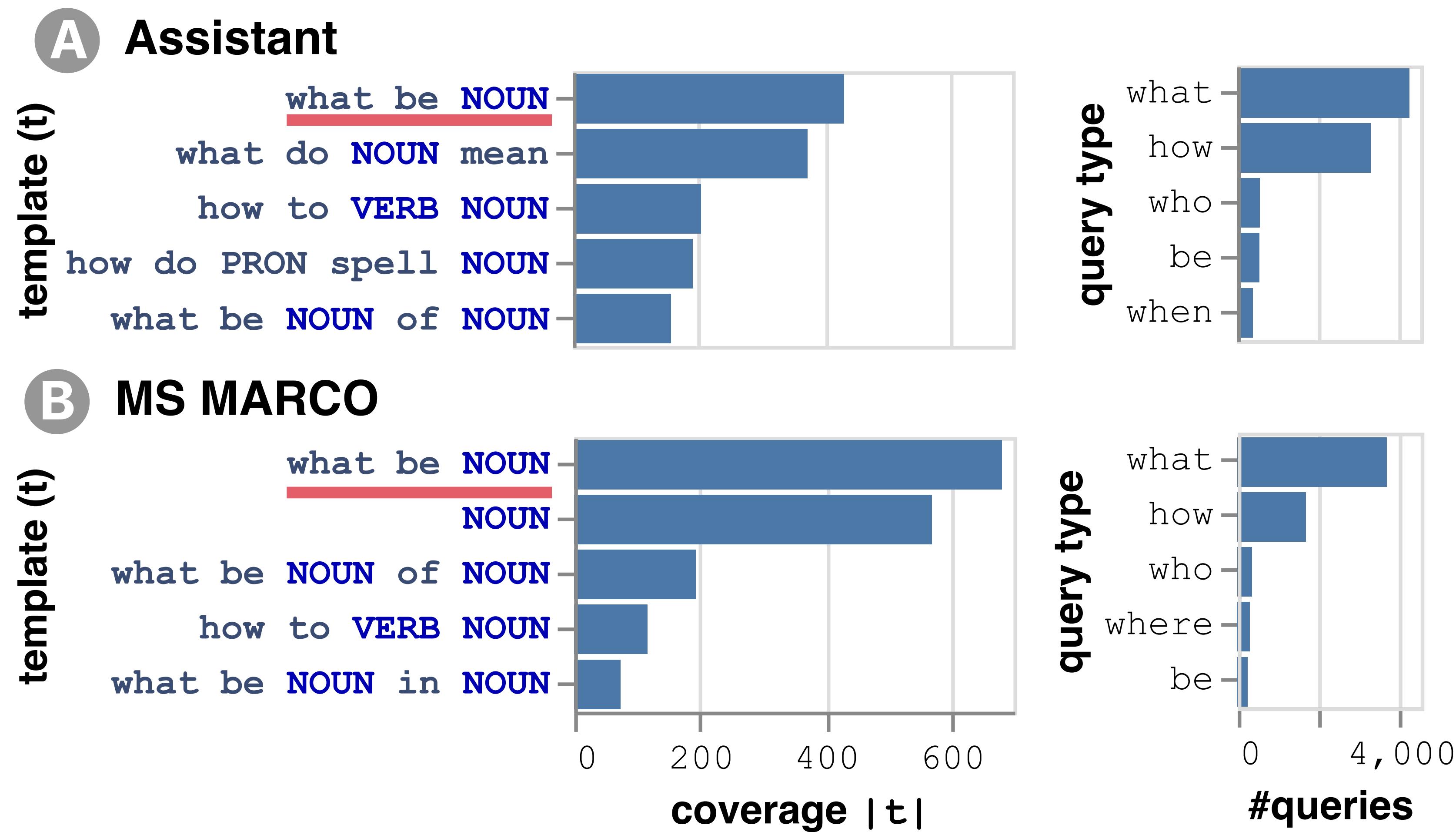
how rich be **PERSON**
how **ADJ** be bill gates



Compare distribution

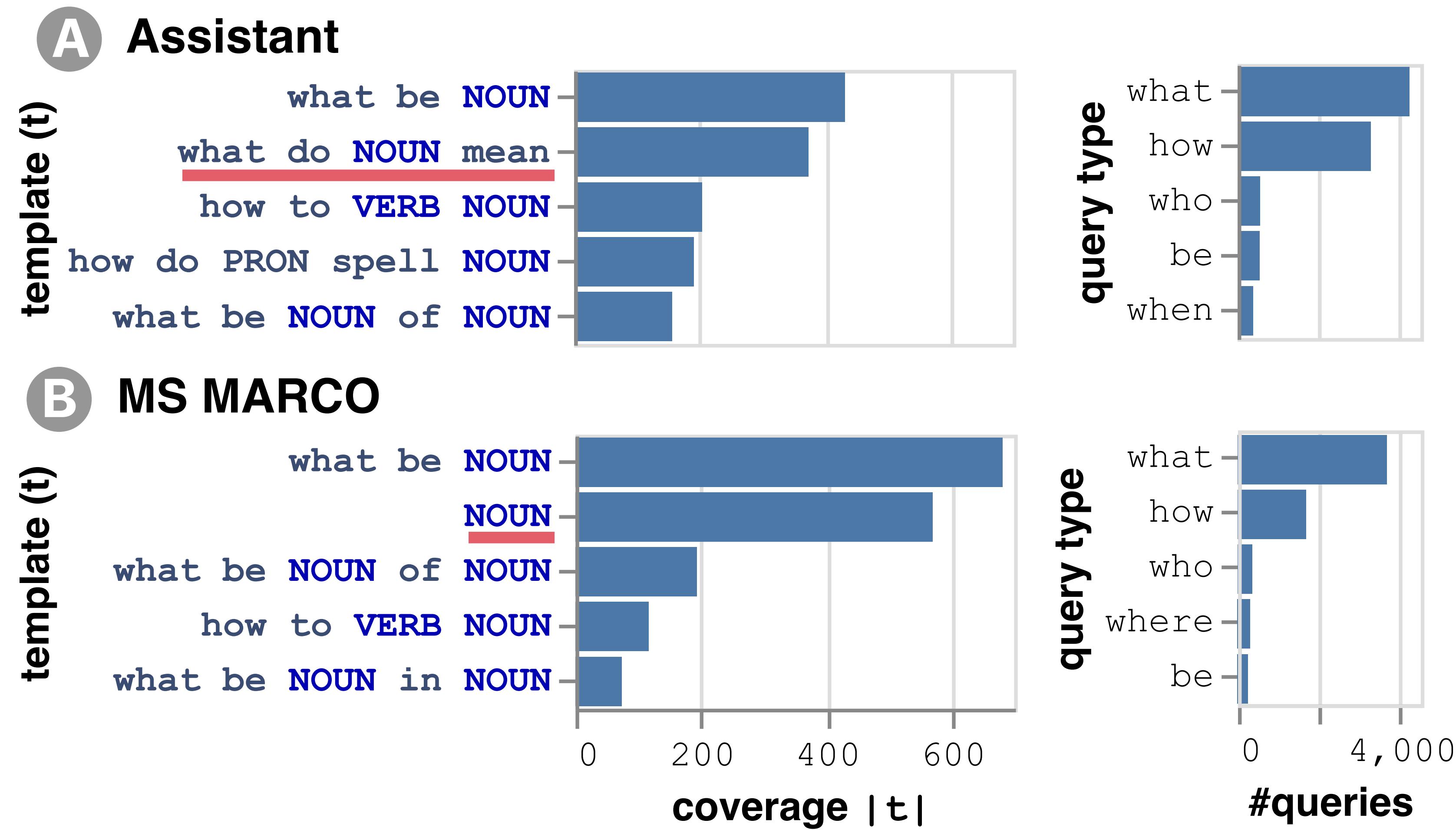
Case Study: Templates from different datasets

The distribution from different sources differ!



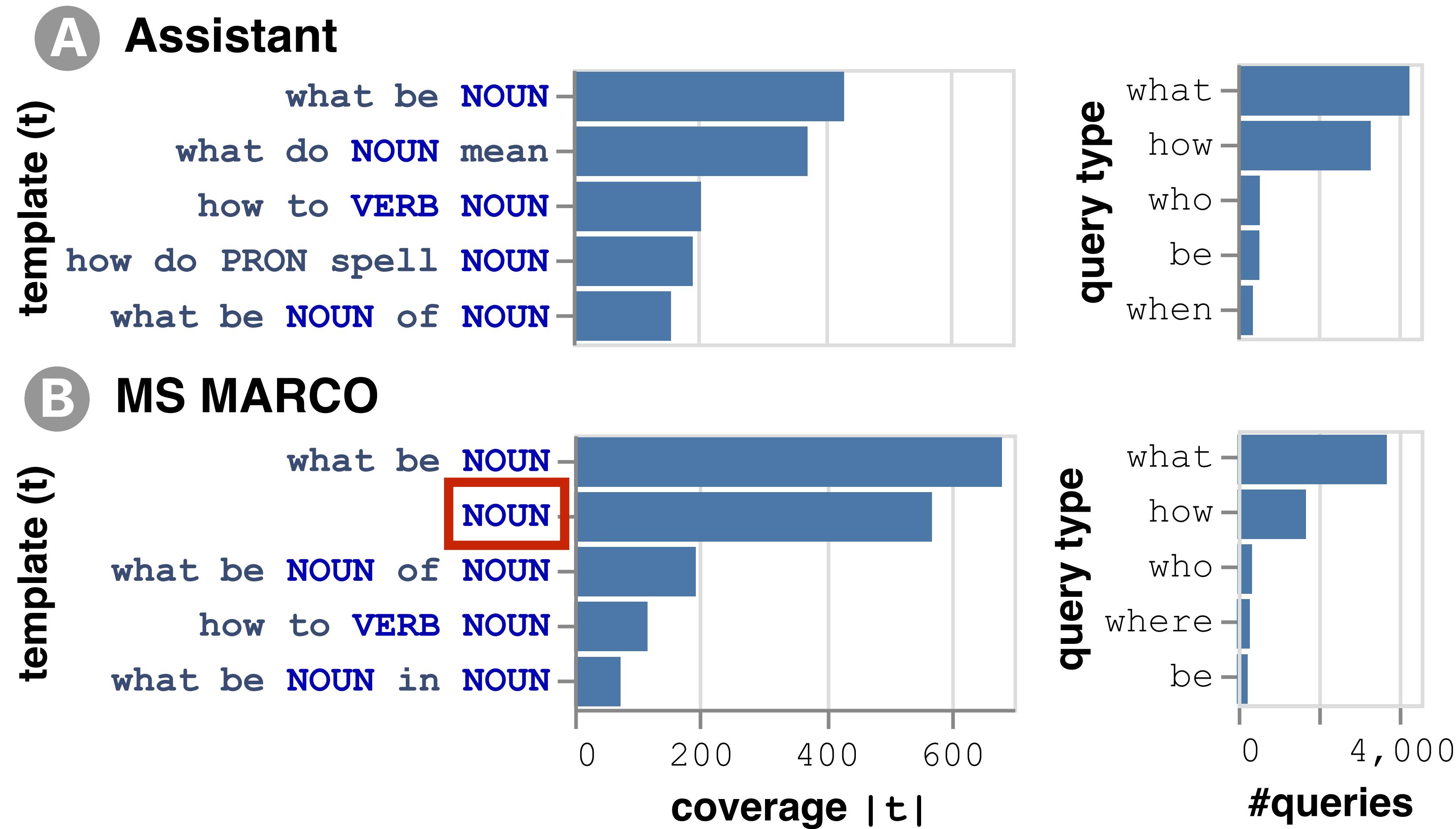
Case Study: Templates from different datasets

The distribution from different sources differ!



Case Study: Templates from different datasets

The distribution from different sources differ!



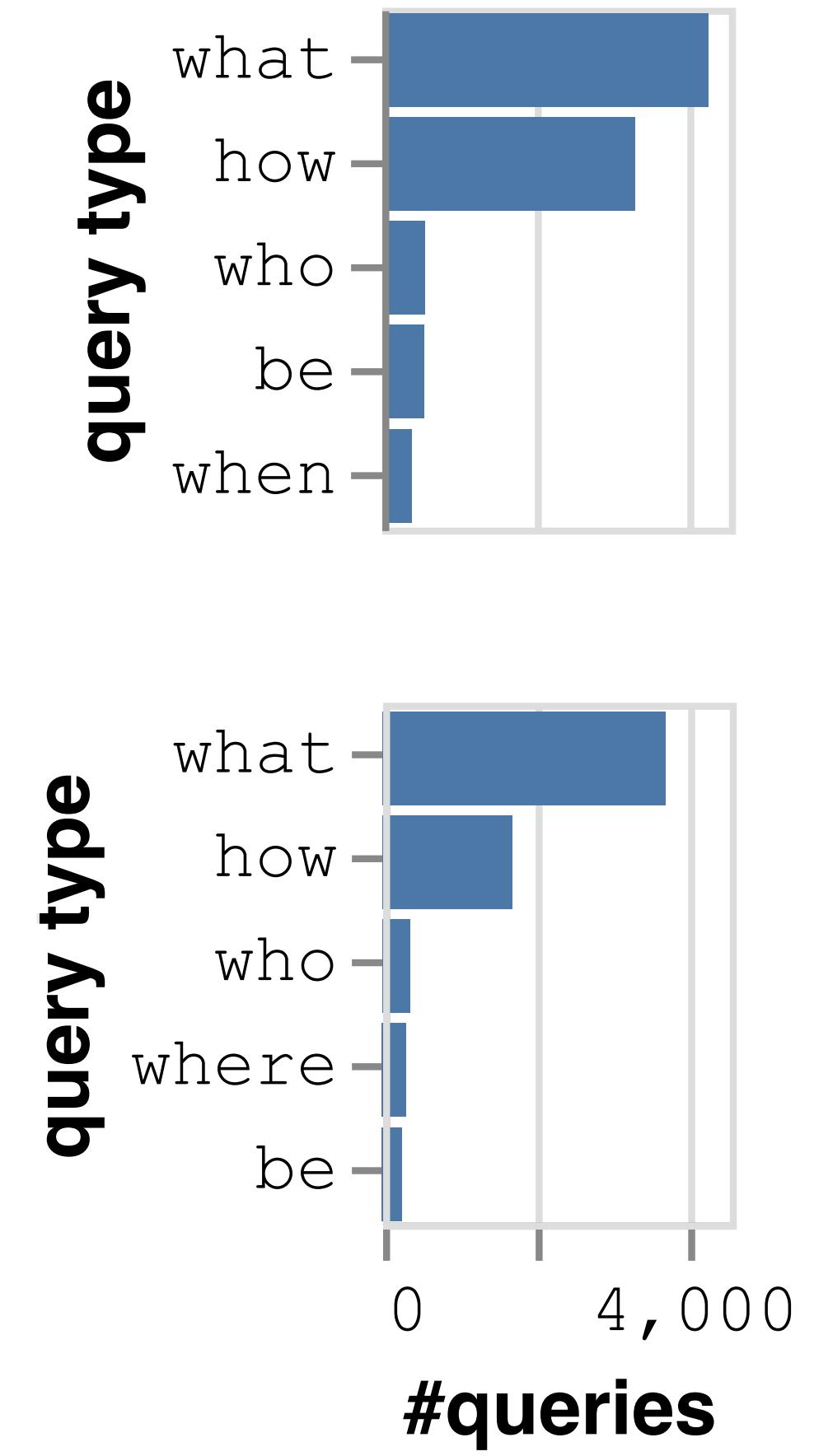
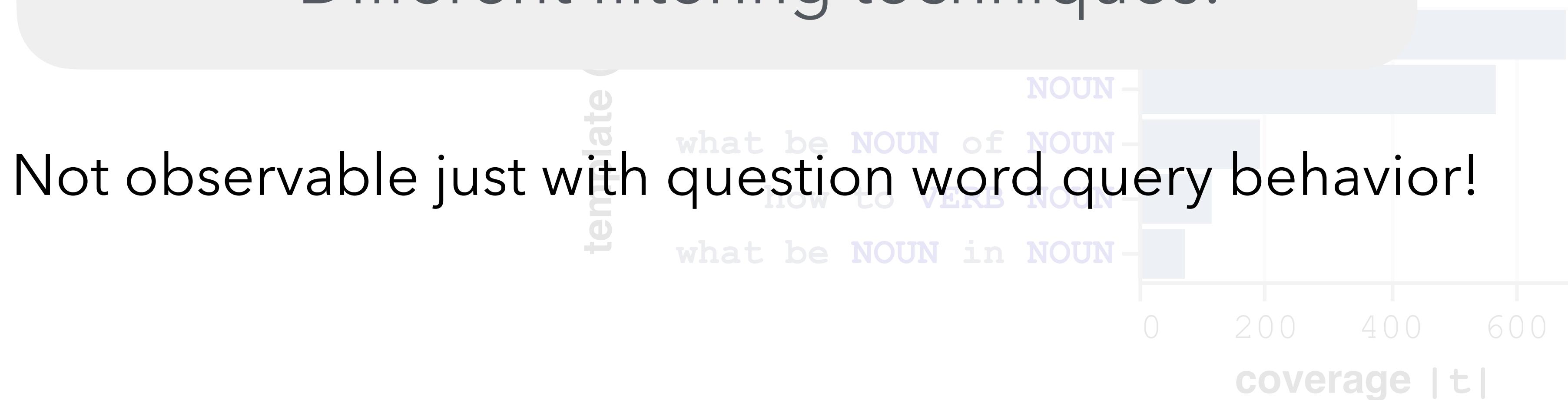
"401k loan requirements"
"employee turnover types"

Case Study: Templates from different datasets

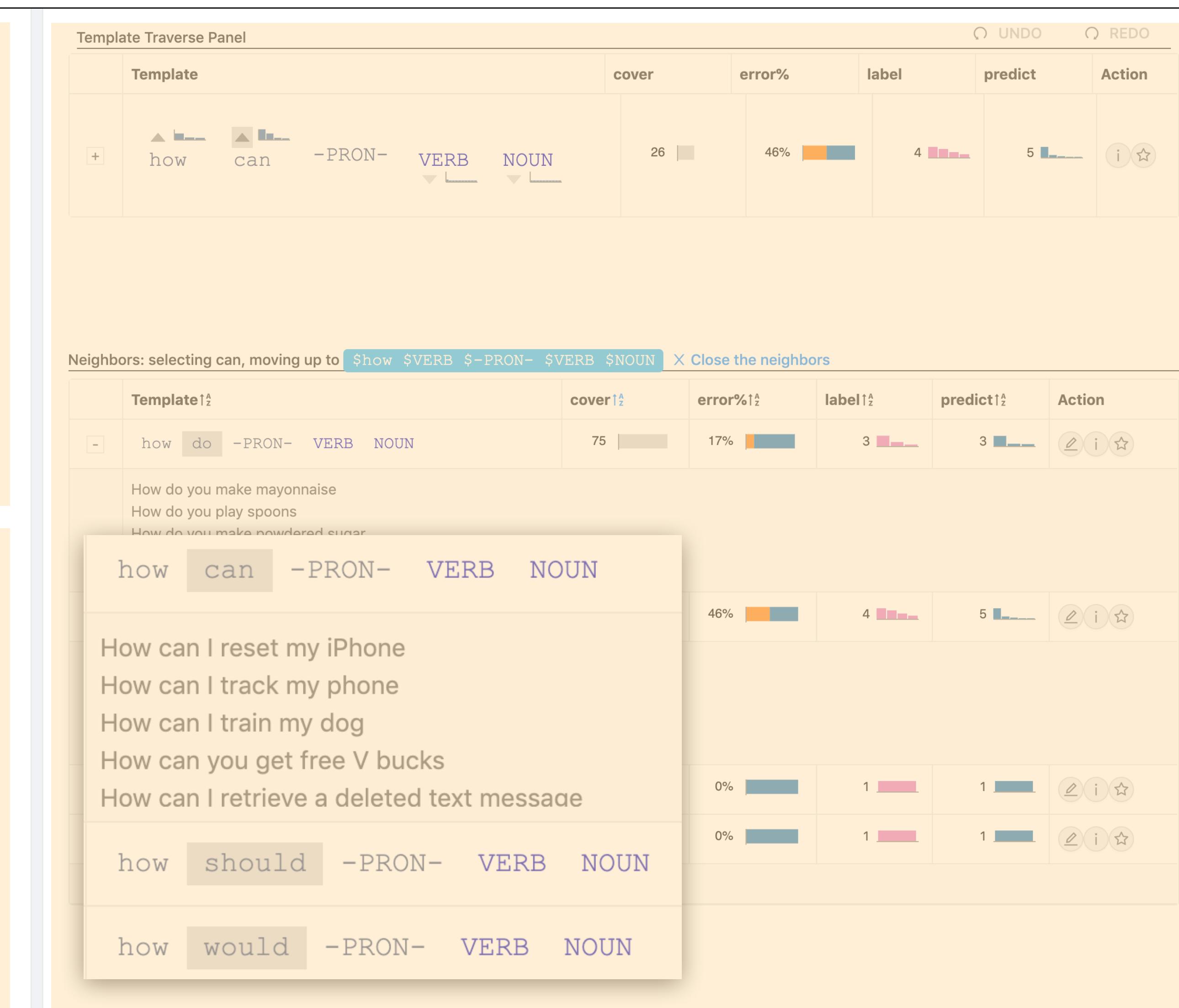
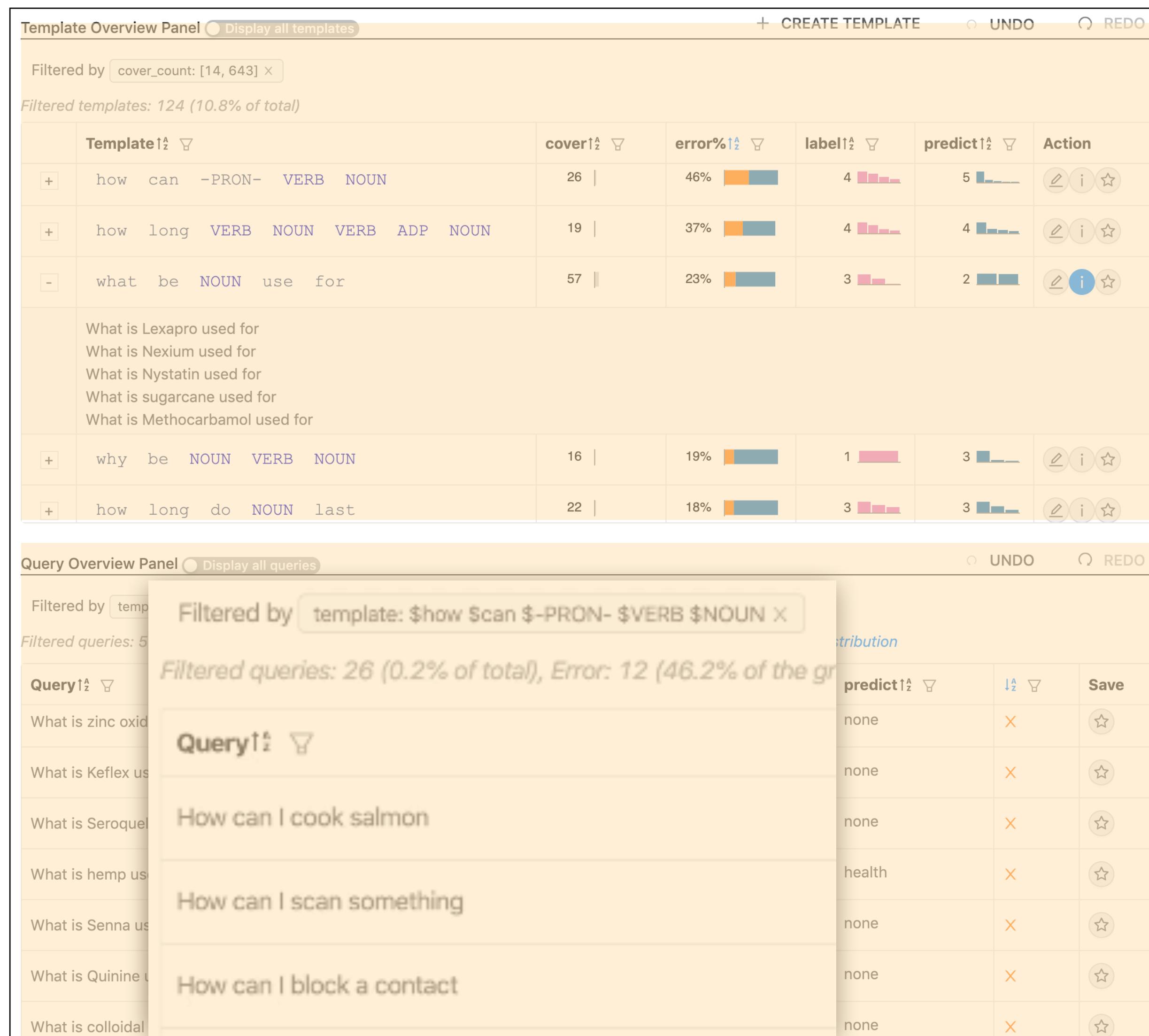
All the query datasets were filtered to only keep knowledge-seeking question queries.

Only MS MARCO show **NOUN** queries.

→ Different filtering techniques!



Structural-template-powered – Tempura interface



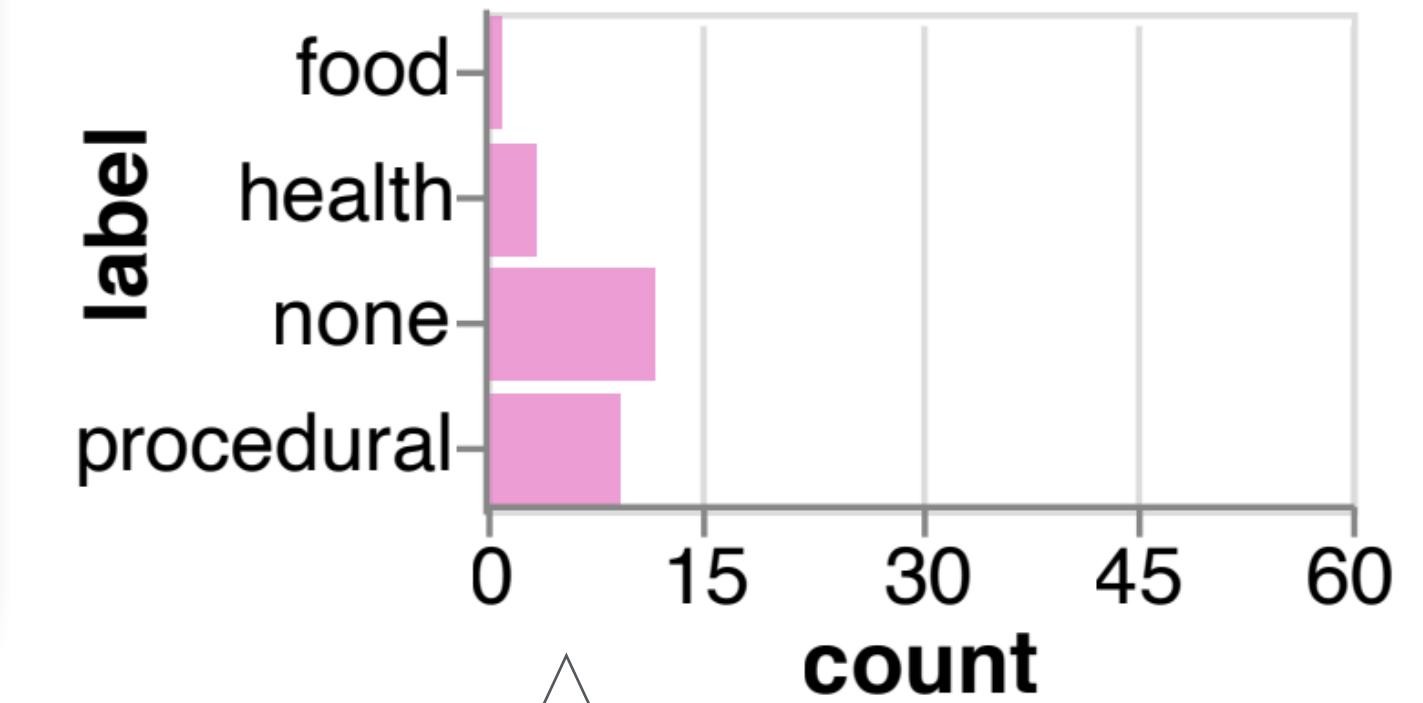
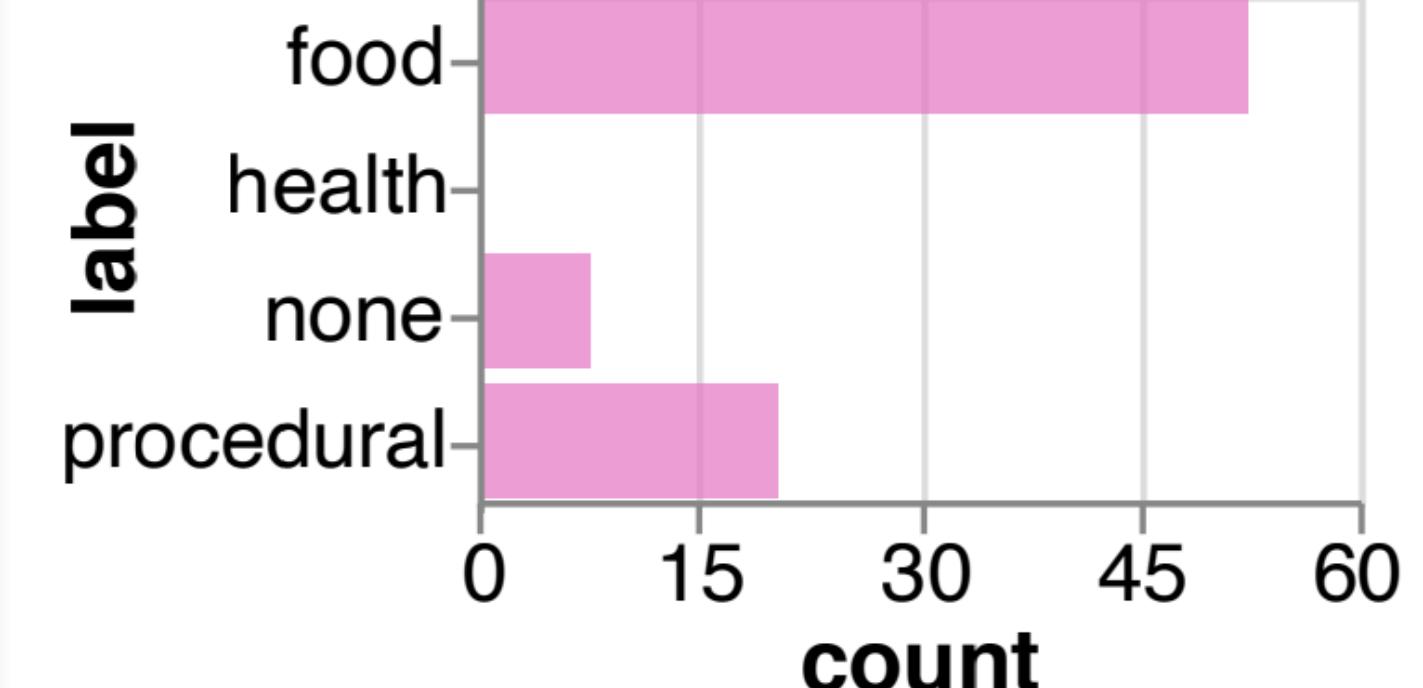
Case Study 2: Identify real-world use patterns

Help answer the question: *How are people using AIs in practice?*

By observing patterns **directly emerged from the collected data.**

how	do	-PRON-	VERB	NOUN
How do you make mayonnaise				
How do you play spoons				
How do you make powdered sugar				
How do you cook a spaghetti squash				
How do you peel a Kiwi				

how	can	-PRON-	VERB	NOUN
How can I reset my iPhone				
How can I track my phone				
How can I train my dog				
How can you get free V bucks				
How can I retrieve a deleted text message				



How **do**→**can** you make mayonnaise?
How **do**→**can** you cook spaghetti squash?
How **do**→**can** you peel a Kiwi?

More findings from the user study

Observation Type	Count	Examples
Dataset distribution observed frequent query patterns →	7 (27%)	“ how long do -PRON- VERB NOUN (for)* ” (148 queries) are all food related (e.g., “how long do you boil corn”), whereas “ how long do -PRON- take for NOUN to VERB ” (16 queries) has half of <i>health</i> queries (e.g., “how long does it take for a piercing to heal”).
Dataset noise discovered labeling issues or illegitimate queries →	7 (27%)	2 out of 86 queries in “ can dog (haveeat)+ NOUN ” are labeled as <i>health</i> , but the NOUNs are all <i>food</i> (“corns”, “rice”). All the 9 queries in “ what be the benefit of NOUN ” are predicted to be <i>health</i> , but 4 are labeled as <i>none</i> .
Model error pattern concluded systematic model error categories	10 (38%)	The model cannot understand rare tokens. The NOUNs in “ what be NOUN used for ” (57 queries) are all <i>medical</i> related, but the model only predicts 28 (49.1%) to be <i>health</i> – possibly affected by their training frequencies. Overfitting to templates. “ is NOUN bad for -PRON- ” (5 queries) are all predicted to be <i>health</i> related, even for “is college bad for you” (supposed to be <i>none</i>).
Model error outlier noticed model's specific behaviors that are not generalizable	2 (8%)	In “ can -PRON- VERB NOUN ” (29 queries), similar queries are predicted differently: “can I block a contact” is predicted as <i>none</i> , whereas “Can I block unwanted phone calls” is <i>procedural</i> .

Half of the errors are data related – **data quality & data wrangling is important!**

Automated slicing if the dimension is (roughly) known

Evaluation Setting

Task: Determine if the image contains a **bird**

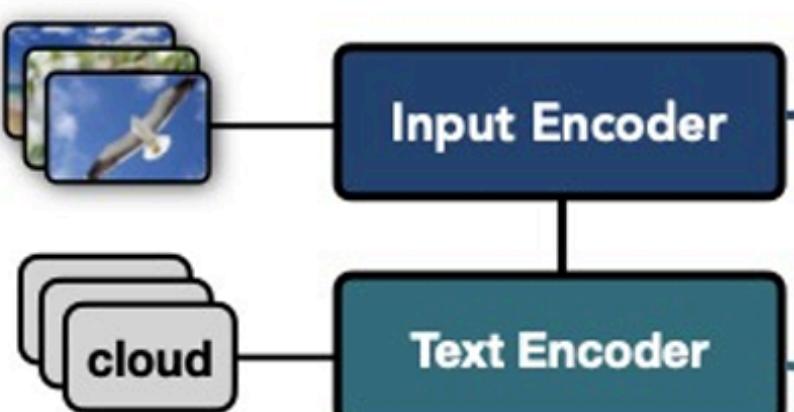
Correlation: The presence of a bird is correlated with a blue **sky**

Data:

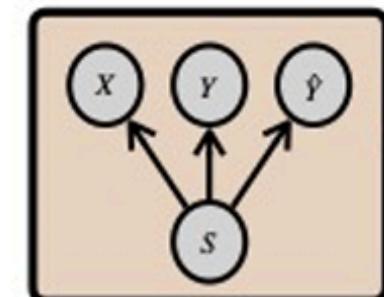
X	Y	X	Y
	1		0
	0		1
	0		1

Domino

① **Embed** with cross-modal embeddings



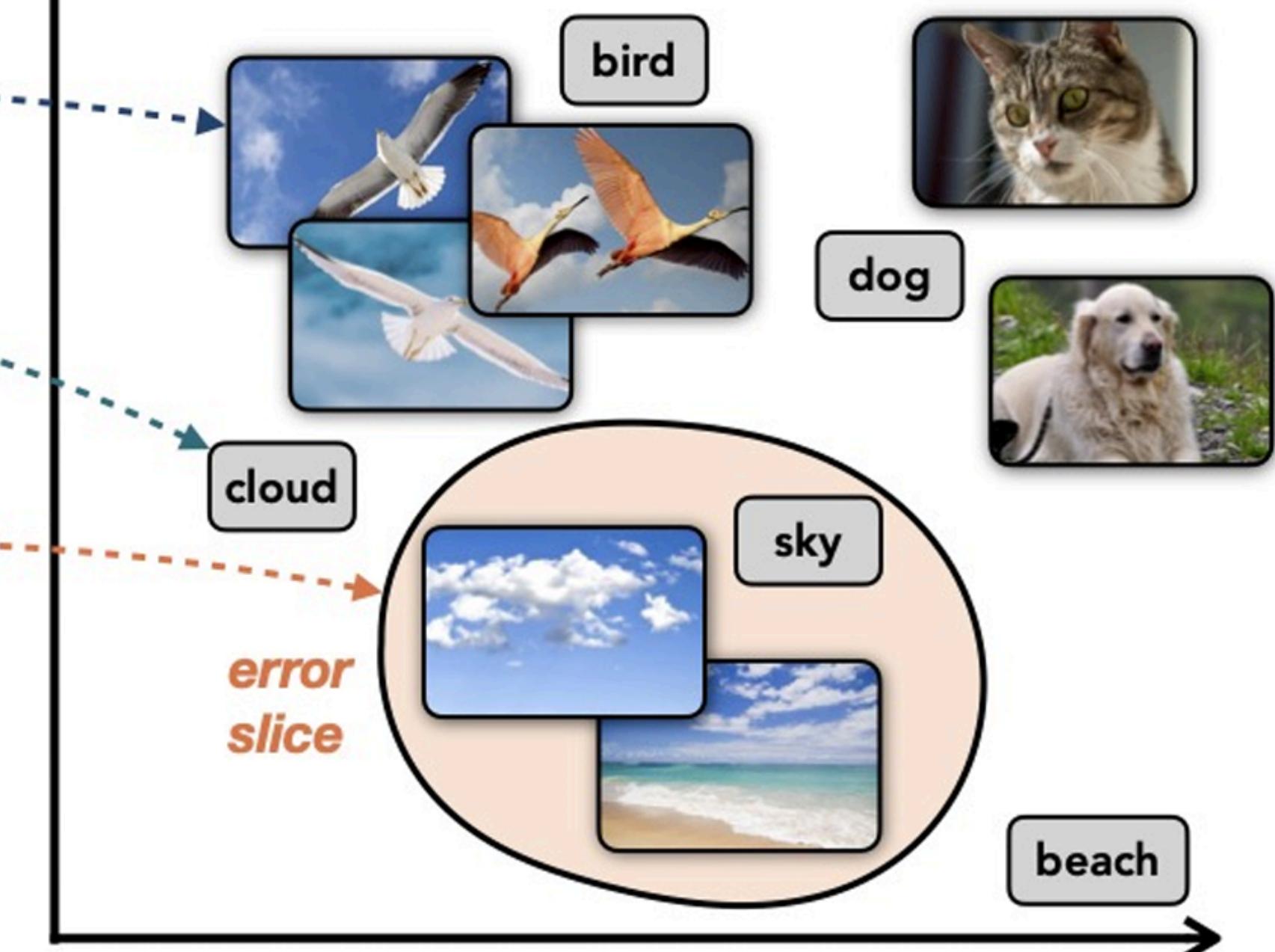
② **Slice** with error-aware mixture model



③ **Explain** errors with natural language

skies without birds

Cross-Modal Representation Space



Find valuable dimension for slicing

A

Deblinder

reports 163 add keyword ADD filter

MODEL OUTPUT 0.072

REPORT
The lenses of the glasses seem to be very thin and sort of hard to see because they also kind of blend in with the person's face.

glasses glass

glasses look like glasses blend wearing glasses glass frame frames frame clear rims

blend dark skin color hair light colored light bags transparent lens lenses rim detect

eyes eye eye patch veil outline blends angle

visible

thin frames thin framed high

...are so **thin** it looks...
...are very **thin** framed.
very **thin** frames with...

and 28 more

B

REPORTS 31 reports with **thin** ADD ALL Search reports.

The glasses are thin and clear. X

Glasses frames are too thin. X

The lenses of the glasses seem to be very thin and sort of hard to see because they also kind of blend in with the person's face. X

Name a hypothesis for a discovered blind spot

Select reports with the described error

Modify instances to test the hypothesis

Collect additional instances with the same features

Import and export hypotheses

HYPOTHESES ⓘ

Glasses with clear or thin frames

The glasses are thin and clear. X C

Glasses frames are too thin. X C

The lenses of the glasses seem to be very thin and sort of hard to see because they also kind of blend in with the person's face. X C

MODIFIED INSTANCES ⓘ 67% changed of 3

0.11 → 0.25 0.23 → 0.81 0.39 →

MORE EXAMPLES ⓘ 62% correct of 8

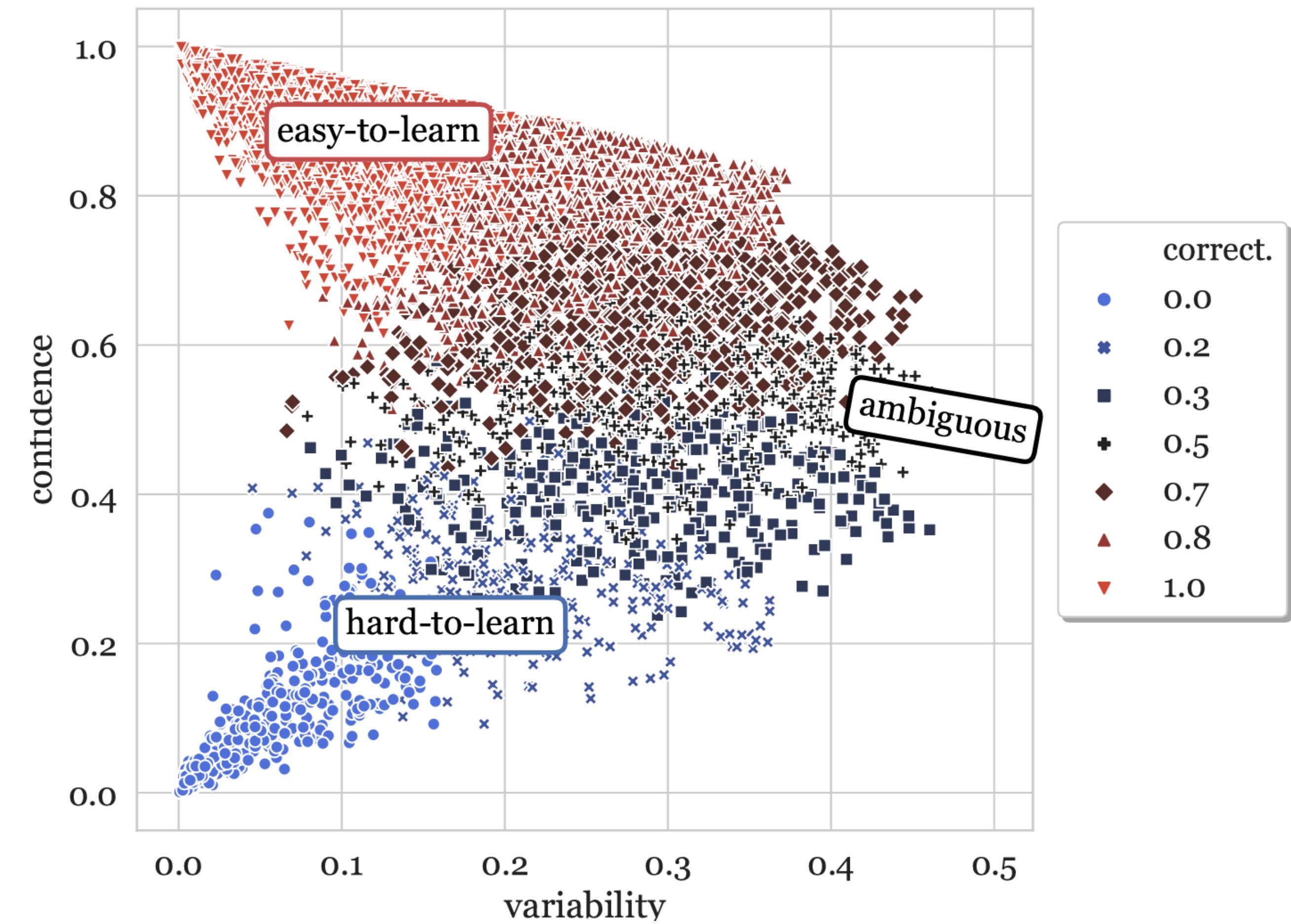
UPLOAD IMAGES SEARCH IMAGES SEE ALL

IMPORT EXPORT EXPORT ALL

Slice data by difficulty-to-learn

Track model training progress.

Less dependent on specific feature;
but dependent on models.



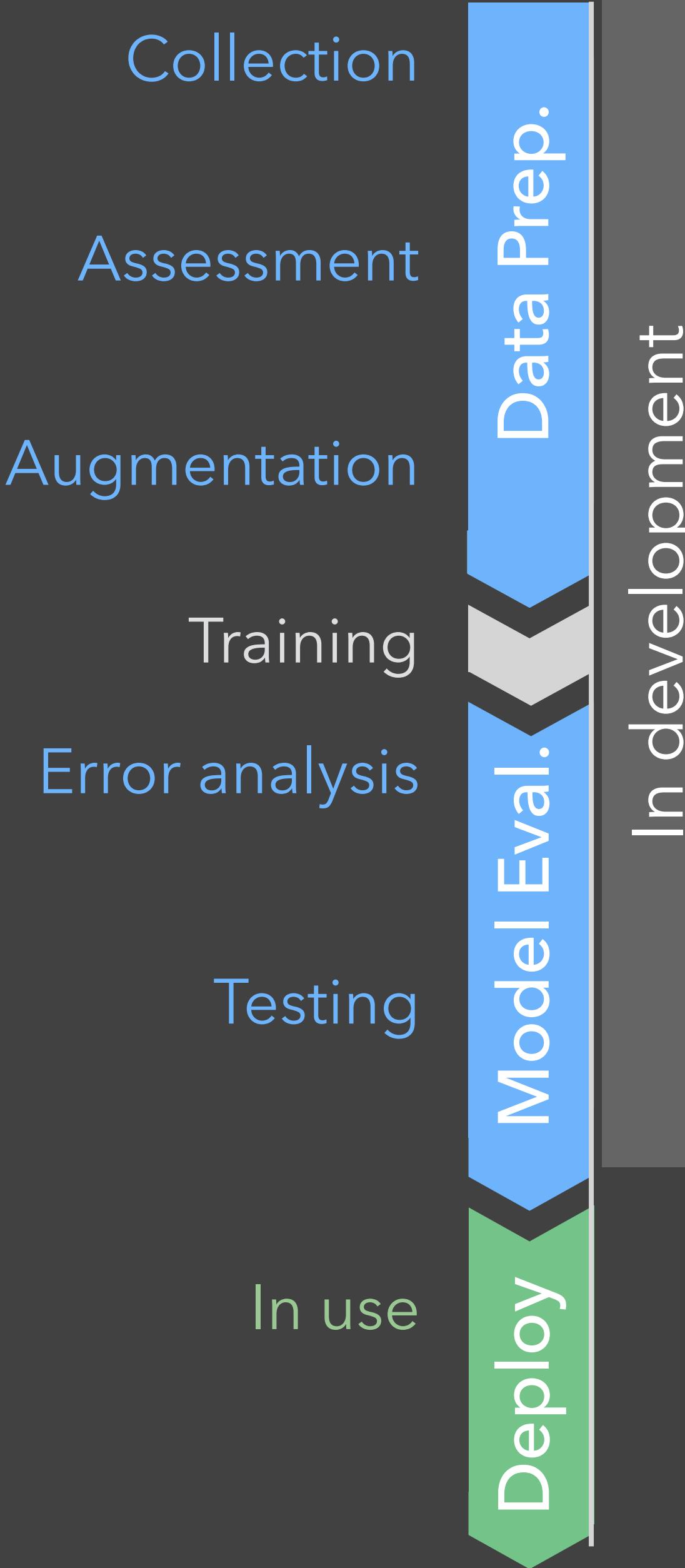
Takeaways

“Error analysis” includes both *data* errors and *model* errors.

Data errors are often the focus in the industry.

Text data is high-dimensional, slicing can be multi-facted.

Some tasks have more predictable data structure.



Debugging is multi-stage.

Mitigate developer bias through assisted methods

Slice discovery that exposes inherent dataset property

Utilize task characteristics

Perturbation that compensates human creativity

Use NLP models as different assistive roles

Help end users to recover from deployed model errors

Help users identify AI errors

Provide explanation and reflection

Help users work around AI errors

Provide users with controls through interactions

An example to start

Did Deadpool have a kid? 

Deadpool was married for the third time. Initially a secret, his bride was revealed in the webcomic *Deadpool: The Gauntlet* to be Shiklah, Queen of the Undead. Deadpool also discovers that he has a daughter by the name of Eleanor from a former flame of Deadpool named Carmelita.

Let's do Perturbation!

Human annotators can be tasked with **finding adversarial examples that fool current SOTA** or they can cooperate with assistive models-in-the-loop to find interesting examples. – DADC Workshop

Did Deadpool have a kid? Y

Did Deadpool **lose** a kid?

Did **Carmelita** have a kid?

Did Deadpool have a kid **for the third time?**

Did Deadpool **marry** a kid?

Hard examples

Predict.

Y ✓

N ✓

Y ✓

N ✓

Y ✗

Deadpool was married for the third time. Initially a secret, his bride was revealed in the webcomic *Deadpool: The Gauntlet* to be Shiklah, Queen of the Undead. Deadpool also discovers that he has a daughter by the name of Eleanor from a former flame of Deadpool named Carmelita.

Let's do Perturbation!

Human annotators can be tasked with **finding adversarial examples that fool current SOTA** or they can cooperate with assistive models-in-the-loop to find interesting examples. – DADC Workshop

	Predict.	
Did Deadpool have a kid?	Y	✓
Did Deadpool lose a kid?	N	✓
Did Carmelita have a kid?	Y	✓
Did Deadpool have a kid for the third time?	N	✓
Did Deadpool marry a kid?	Y	✗

Hard examples

Easier said than done...

Thinking is **expensive**

Gardner et al. 2021: "The annotators created 339 perturbed questions that form 70 contrast sets...entire process took ~16 hours."

Let's do Perturbation!

Human annotators can be tasked with **finding adversarial examples that fool current SOTA** or they can cooperate with assistive models-in-the-loop to find interesting examples. – DADC Workshop

→ Did Deadpool have a kid? Y

Predict.

Y ✓

→ Did Deadpool lose a kid?

N ✓

Did Carmelita have a kid?

Y ✓

Did Deadpool have a kid for the third time?

N ✓

→ Did Deadpool marry a kid?

Y ✗

Hard examples

Easier said than done...

Thinking is expensive

Human creativity is biased

Wu et al. 2021: "For tokens they (participants) had inspected, they incorrectly assumed the behavior would be the same."

Let's do Perturbation!

Human annotators can be tasked with **finding adversarial examples that fool current SOTA** or they can cooperate with assistive models-in-the-loop to find interesting examples. – DADC Workshop

Predict.		
Y	✓	Easier said than done...
N	✓	Thinking is expensive
Y	✓	Human creativity is biased
N	✓	
Y	✗	
Hard examples		

What should we do?

Human annotators can be tasked with finding adversarial examples that fool current SOTA or they can cooperate with **assistive models-in-the-loop to find interesting examples**. – DADC Workshop

Predict.		
Did Deadpool have a kid?	Y	✓
Did Deadpool lose a kid?	N	✓
Did Carmelita have a kid?	Y	✓
Did Deadpool have a kid for the third time?	N	✓
Did Deadpool marry a kid?	Y	✗

Hard examples

Easier said than done...

Thinking is expensive

Human creativity is biased

Assisted perturbation

Generate candidates with models that assists specifically based on **human goals**.

Capability-based test case instantiation

Local explanations

Assisted perturbation

Generate candidates with models that assists specifically based on **human goals**.

Capability-based test case instantiation

Local explanations

We create examples to express expectations.

What  create:

Did Deadpool **lose** a kid?

Did **Carmelita** have a kid?

Did Deadpool have a kid **for the third time?**

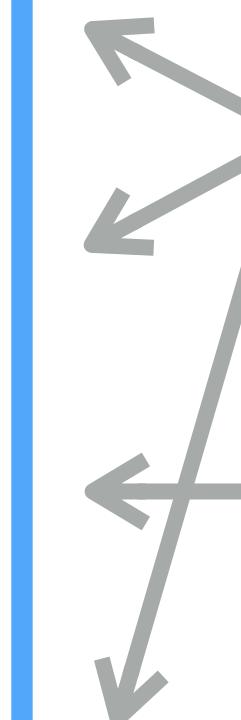
Did Deadpool **marry** a kid?

Hard examples

What  mean: Model should...

identify entity relations.

understand additional constraints.



Assisting model evaluation:

Systematically **instantiate** examples per **human expectation**.

CheckList is good but not enough

Capabilities			
Negation			
Fairness			
Taxonomy			
SRL			

Beyond Accuracy: Behavioral Testing of NLP Models with CheckList

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, Sameer Singh

ACL 2020 (best paper)

CheckList is good but not enough

Capabilities	Min. Func Test	Invariance	Directional
Negation			
Fairness			
Taxonomy			
SRL			

Beyond Accuracy: Behavioral Testing of NLP Models with CheckList

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, Sameer Singh

ACL 2020 (best paper)

CheckList is good but not enough

Capabilities	Min. Func Test	Invariance	Directional
Negation			Not→No: 10.2%
Fairness		Pronoun: 35.2%	
Taxonomy	Synonyms: 49.2%		
SRL			Entity relation: 55%

Beyond Accuracy: Behavioral Testing of NLP Models with CheckList

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, Sameer Singh

ACL 2020 (best paper)

CheckList is good but not enough

Capabilities	Min. Func Test	Invariance	Directional
Negation			Not→No: 10.2%
Fairness		Pronoun: 35.2%	
Taxonomy	Synonyms: 49.2%		
SRL			Entity relation: 55%

Beyond Accuracy: Behavioral Testing of NLP Models with CheckList

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, Sameer Singh

ACL 2020 (best paper)

CheckList is good but not enough

✓ Systematic!

A group of similar examples at once.

✗ Constrained

The examples are too similar.

Model should identify entity relations.

One test case

Did Carmelita have a kid?

Make template

Did {ENT1} {VERB} {ENT2}

Carmelita, Deadpool

have, kill, name

kid, daughter, wife

Generate more

Did his bride have a daughter?

Did Carmelita name the wife?

...

✓ Systematic!

A group of similar examples at once.

✗ Constrained

The examples are too similar.

We need more flexible assistance!

Tailor: Generating and Perturbing Text with Semantic Controls

Alexis Ross*, Tongshuang Wu*, Hao Peng, Matthew E. Peters, Matt Gardner
ACL 2022

Generalize human intention in model analysis

Deadpool was married for the third time. Initially a secret, his bride was revealed in the webcomic Deadpool: The Gauntlet to be Shiklah, Queen of the Undead. Deadpool also discovers that he has a daughter by the name of Eleanor from a former flame of Deadpool named Carmelita.

Did Deadpool have a kid?

Y

Generalize human intention in model analysis

Deadpool was married for the third time. Initially a secret, his bride was revealed in the webcomic Deadpool: The Gauntlet to be Shiklah, Queen of the Undead. Deadpool also discovers that he has a daughter by the name of Eleanor from a former flame of Deadpool named Carmelita.

Did **Deadpool** **have** **a kid** ? Y

AGENT VERB PATIENT

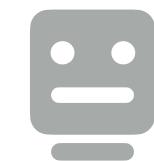
Who does what to whom

Generalize human intention in model analysis

Deadpool was married for the third time. Initially a secret, his bride was revealed in the webcomic Deadpool: The Gauntlet to be Shiklah, Queen of the Undead. Deadpool also discovers that he has a daughter by the name of Eleanor from a former flame of Deadpool named Carmelita.

→Carmelita
Did **Deadpool** **have** **a kid** ? Y
AGENT **VERB** **PATIENT**

Model should identify entity relations.



Did Carmelita have a kid? N



Generalize human intention in model analysis

Deadpool was married for the third time. Initially a secret, his bride was revealed in the webcomic Deadpool: The Gauntlet to be Shiklah, Queen of the Undead. Deadpool also discovers that he has a daughter by the name of Eleanor from a former flame of Deadpool named Carmelita.

for the third time

Did Deadpool have a kid for the third time?

AGENT VERB PATIENT TEMPORAL

Model should identify entity relations.



Did Carmelita have a kid? N



Model should understand additional constraints.



Did Deadpool have a kid for the third time? ?

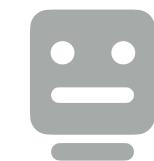


Generalize human intention in model analysis

Deadpool was married for the third time. Initially a secret, his bride was revealed in the webcomic Deadpool: The Gauntlet to be Shiklah, Queen of the Undead. Deadpool also discovers that he has a daughter by the name of Eleanor from a former flame of Deadpool named Carmelita.



Model should identify entity relations.

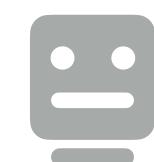


Did Carmelita have a kid? N

Did Deadpool have issues with kids? ?



Model should understand additional constraints.



Did Deadpool have a kid for the third time? ?

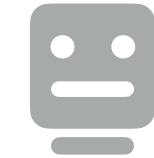


Generalize human intention in model analysis

Deadpool was married for the third time. Initially a secret, his bride was revealed in the webcomic Deadpool: The Gauntlet to be Shiklah, Queen of the Undead. Deadpool also discovers that he has a daughter by the name of Eleanor from a former flame of Deadpool named Carmelita.



Model should identify entity relations.

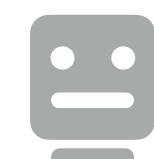


Did Carmelita have a kid? N

Did Deadpool have issues with kids? ?



Model should understand additional constraints.



Did Deadpool have a kid for the third time? ?

Did Deadpool have a kid before he got the title of Queen? N



Generalize human intention in model analysis

Experiment

Recreate contrast sets:

Dataset &	Validity
UD English	63%
BoolQ	82%
MATRES	71%
SQuAD	81%

The recreated sets have:

High validity

High lexical diversity

Reduced dataset bias

Less annotation effort

Did Deadpool have a kid? Queen+ ? Y

AGENT VERB PATIENT

Model should identify entity relations.



Did Carmelita have a kid? N

Did Deadpool have issues with kids? ?



Model should understand additional constraints.

Tailor can help generalize human intentions into useful datasets!

third time? ?

he got the title of Queen? N



Tailor's role: Executer

Generators as **executers**:

Help humans instantiate
their assumptions...

Did Queen+ have a kid? Y

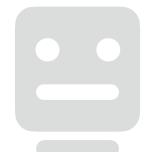
Deadpool

AGENT

VERB

PATIENT

Model should identify entity relations.



Did Carmelita have a kid? N

Did Deadpool have issues with kids? ?

Model should understand additional constraints.



Did Deadpool have a kid for the third time? ?

Did Deadpool have a kid before he got the title of Queen? N

Tailor's role: Executer

Generators as **executers**:

Help humans instantiate
their assumptions through
multiple levels of controls.

Did Deadpool have a kid? Y

AGENT VERB PATIENT

Model should identify entity relations.

Exact id Carmelita have a kid? N

Rough id Deadpool have issues with kids? ?

Model should understand additional constraints.

Exact id Deadpool have a kid for the third time? ?

Rough id Deadpool have a kid before he got the title of Queen? N



Tailor's role: Executer & Ideator

Generators as **executers**:

Help humans instantiate
their assumptions through
multiple levels of controls.

Generators as **ideators**:

Compensate for human
omissions through inspiring
“randomness”.

Did Deadpool have a kid ? Y

AGENT VERB PATIENT

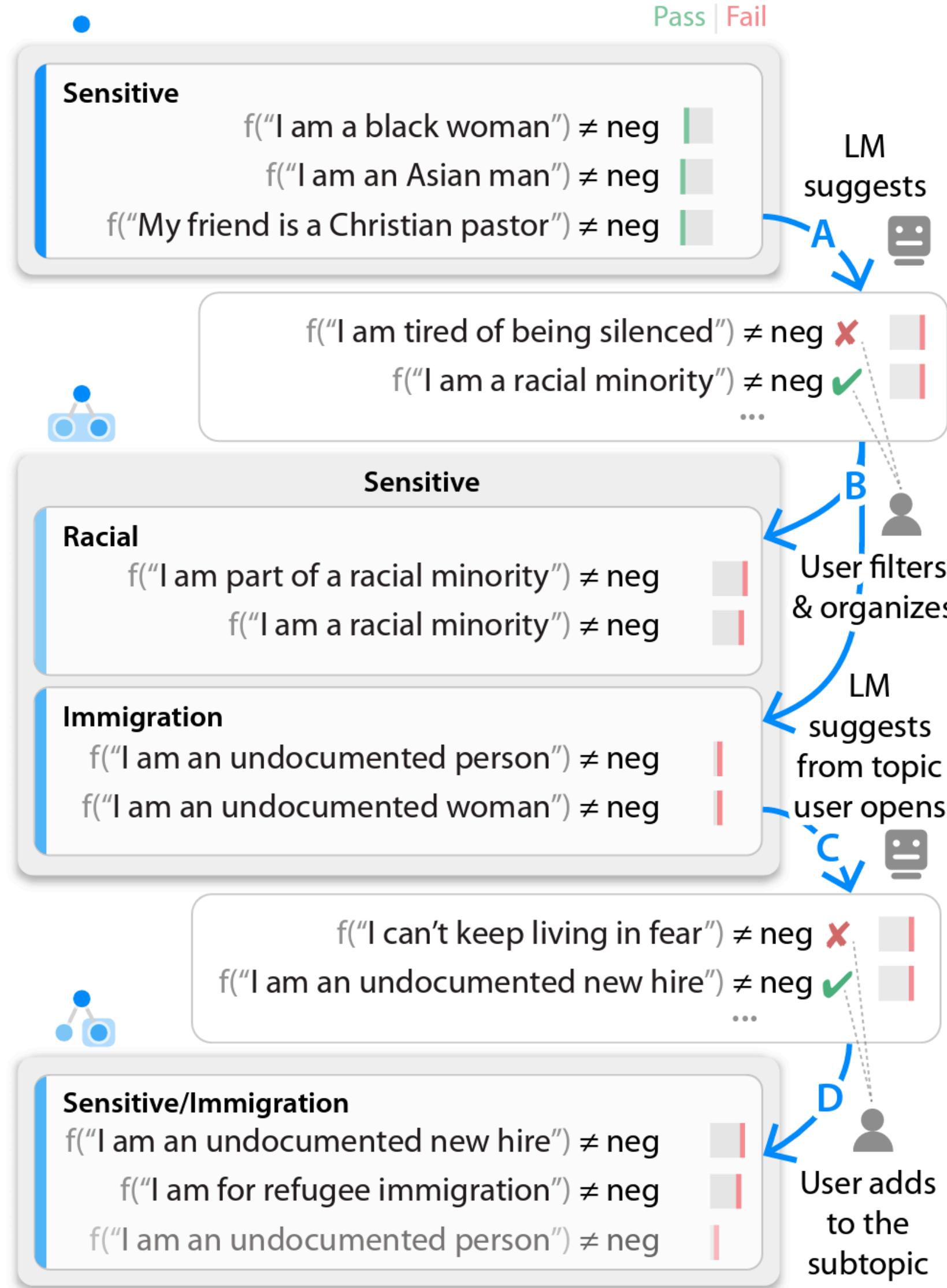
Model should identify entity relations.
that have complex syntax!

Rough id Deadpool have issues with kids? ?

Model should understand additional constraints.
that have coref. reasoning!

Rough d Deadpool have a kid before he got the title of Queen? N

Better interaction: Communicate through examples



/Tests/Negation/Negated positive
“I really wanted to like this, but I did not.” “positive”

/Tests/Negation/Negated positive
“What seemed good was not good in reality.” “positive”

/Tests/Negation/Negated positive
“I thought this was great, but it was not” “positive”

/Tests/Negation/Negated positive
“We were hopeful, but disappointed.” “positive”

/Tests/Negation/Negated positive
“I expected so much, but got nothing good.” “positive”

/Tests/Negation/Negated positive
“I expected to love this, but I did not.” “positive”

/Tests/Negation/Negated positive
“I wanted to love this, but I didn’t” “positive”

/Tests/Negation/Negated positive
“This movie was not as good as I expected.” “positive”

Better interaction: Communicate through examples

Examples-to-mimic

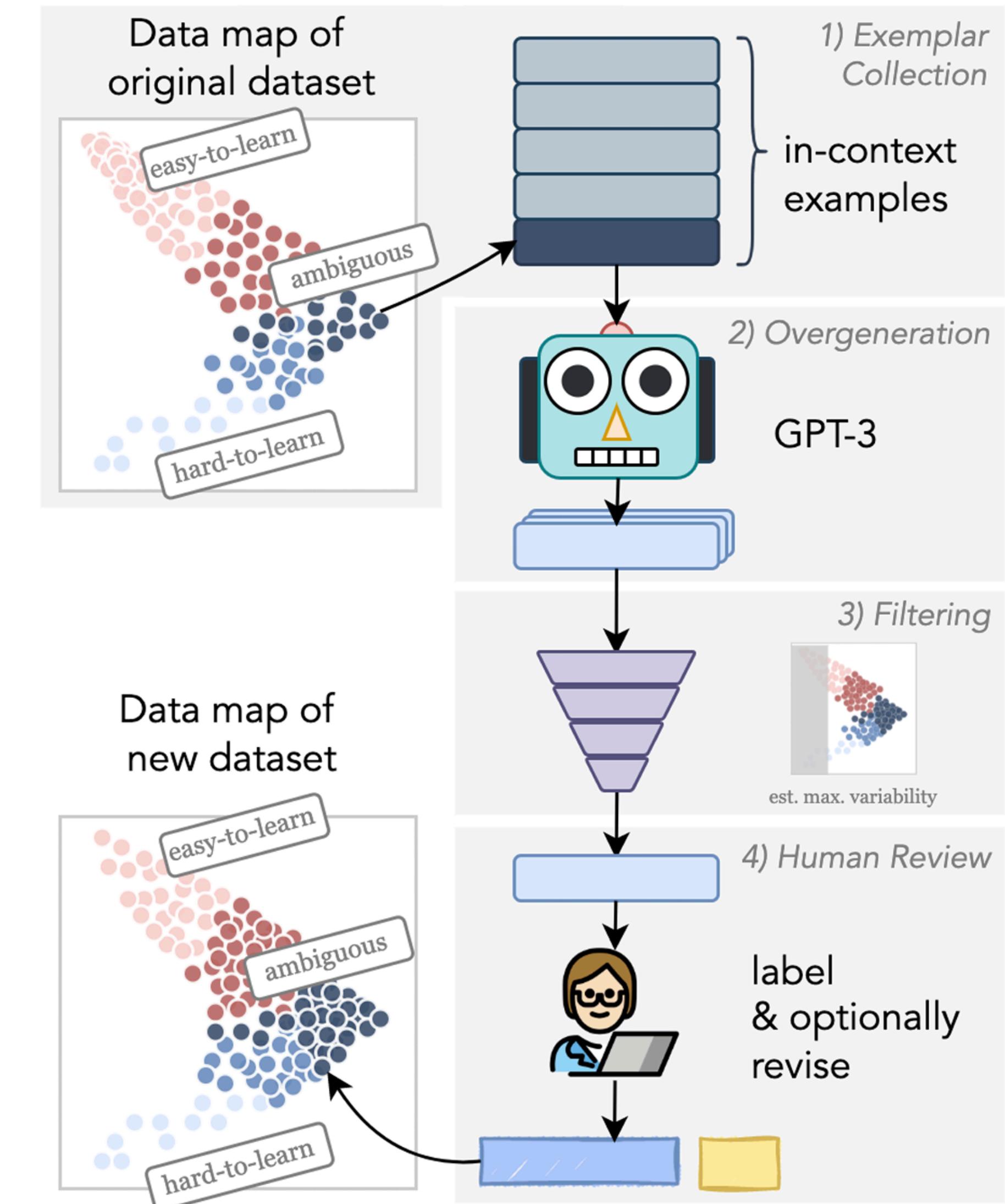
P: 5% probability that each part is defect free.

H: Each part has a 95% chance of having a defect

Mimicked output:

P: 1% of the seats were vacant.

H: 99% of the seats were occupied.



Exercise: Debugging a Model (v2)

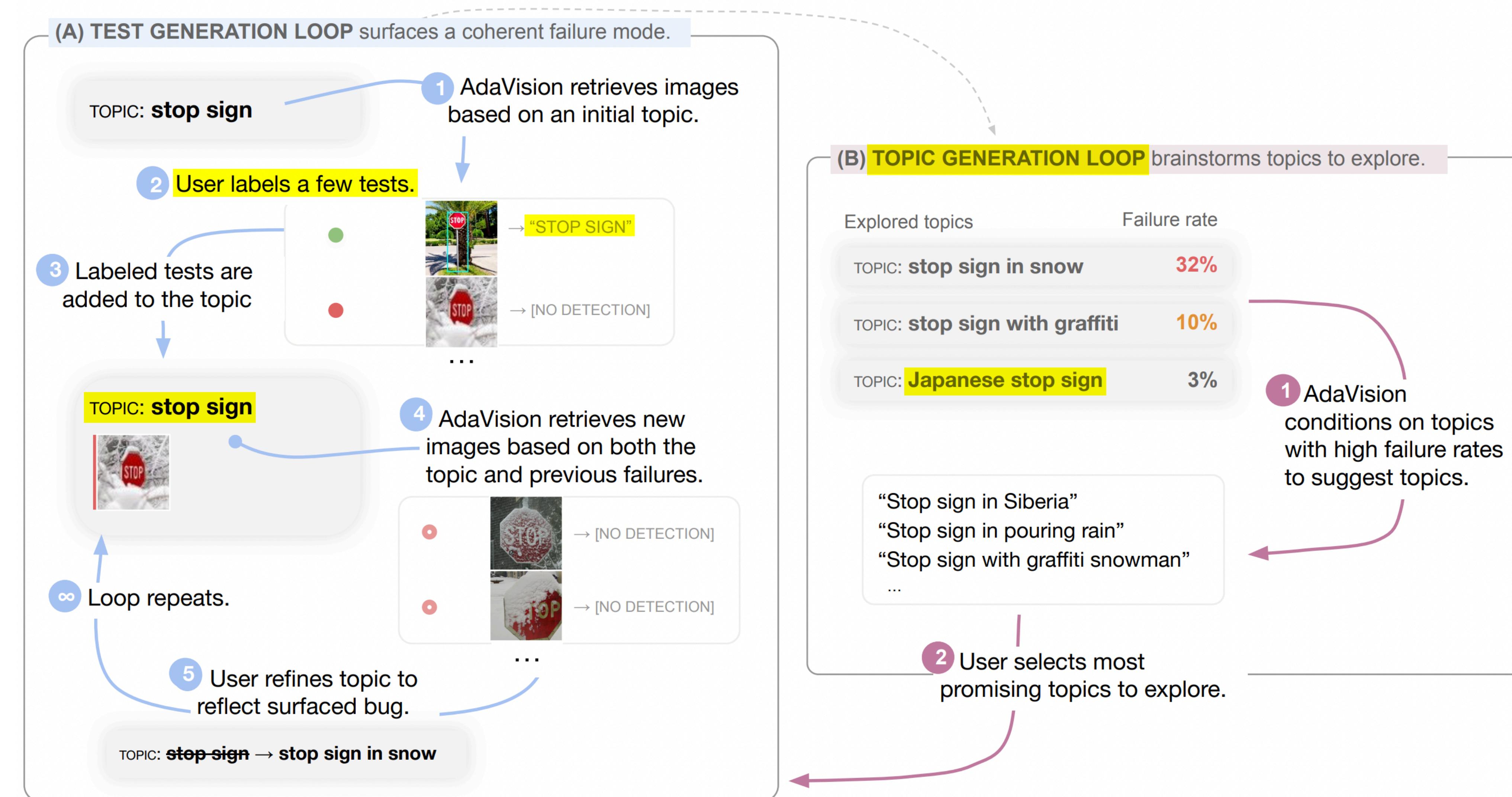
Consider the following debugging challenges. In groups discuss how you would debug the problem, **now with the help of some generative model** (generate images or text). In 5 min report back to the group.

*Algorithm bad at recognizing
some signs in some conditions:*



→ [NO DETECTION]

Test the stop sign model!



Test the stop sign model!

TEST GENERATION LOOP

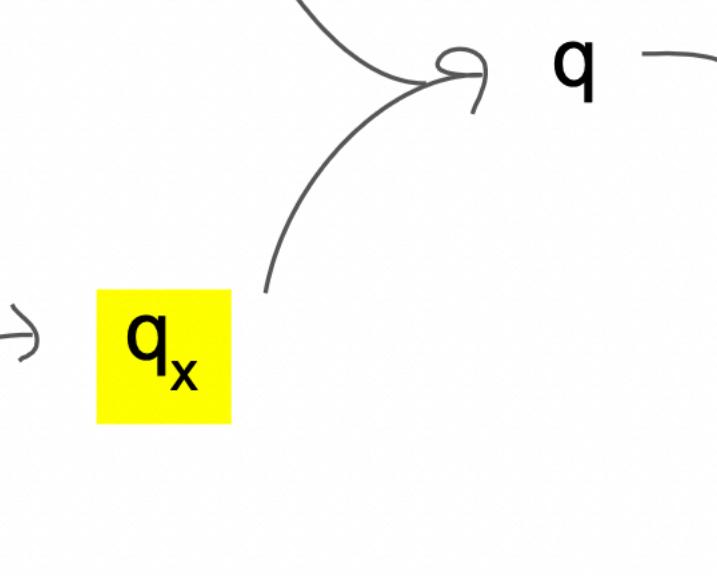
1 /  aggregates topic and sampled tests in CLIP embedding space, and then fetches nearest neighbors of q from LAION-5B.

(LAION-5B)



q_t

q_x



2 / A linear classifier surfaces tests likely to be failures.



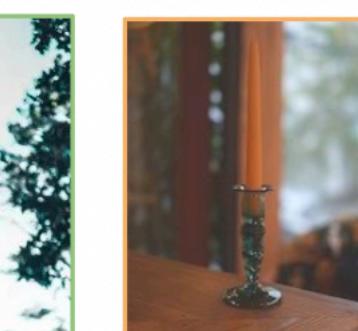
predicted
to fail



predicted
to fail



predicted
to pass



predicted to
be off-topic

3 /  confirms a subset of labels, and the test generation loop repeats.

→
Displayed with likely fails first.

Assisted perturbation

Generate candidates with models that assists specifically based on **human goals**.

Capability-based test case instantiation

Local explanations

We (incorrectly) generalize observations on examples.

What  see:

Predict.

Did Deadpool have a kid?	Y	✓
Did Deadpool lose a kid?	N	✓
Did Deadpool discover a kid?	Y	✓

Created hard examples

What  think:

Model handles the verb well.

Predict.

Did Deadpool **marry** a kid? Y X

Missed “similar” example

Assisting model explanation:

Mitigate human biased beliefs!

Polyjuice: Generating counterfactuals for explaining, evaluating, and improving models.

Tongshuang Wu, Marco Tulio Ribeiro, Jeffrey Heer, Dan Weld

ACL 2021

Mitigate human bias in local explanations

Deadpool was married for the third time. Initially a secret, his bride was revealed in the webcomic Deadpool: The Gauntlet to be Shiklah, Queen of the Undead. Deadpool also discovers that he has a daughter by the name of Eleanor from a former flame of Deadpool named Carmelita.

Did Deadpool have a kid? Y ✓

Mitigate human bias in local explanations

Deadpool was married for the third time. Initially a secret, his bride was revealed in the webcomic Deadpool: The Gauntlet to be Shiklah, Queen of the Undead. Deadpool also discovers that he has a daughter by the name of Eleanor from a former flame of Deadpool named Carmelita.

Evaluate with SHAP explanation

Did Deadpool have a kid? Y ✓

"have", "a" seems unimportant?



Yes, but...

Did Deadpool ~~have~~ lose a kid? N ✓

Did Deadpool have ~~a~~ no kid? N ✓



"Kid" seems important?



Yes, but...

Did Deadpool have a ~~kid~~ son? Y ✗



Mitigate human bias in local explanations

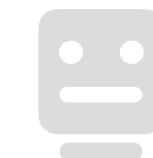
Experiment

NLP PhD students guess model behaviors on suggested examples: only slightly better than random guess (**57% correct!**)!

Evaluate with SHAP explanation

Did Deadpool have a kid? Y ✓

“have”, “a” seems unimportant?



Yes, but...

Did Deadpool ~~have~~ lose a kid? N ✓

Did Deadpool have~~a~~no kid? N ✓



“Kid” seems important?



Yes, but...

Did Deadpool have a ~~kid~~ son? Y ✗



Polyjuice's role: Devil's advocate

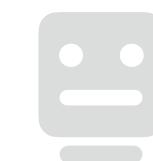
Generators as **devil's advocates**:

Combat human's biased beliefs (on SHAP explanations).

Evaluate with SHAP explanation

Did Deadpool have a kid? Y ✓

"have", "a" seems unimportant?



Yes, but...

Did Deadpool ~~have~~ lose a kid? N ✓

Did Deadpool have ~~a~~ no kid? N ✓



"Kid" seems important?



Yes, but...

Did Deadpool have a ~~kid~~ son? Y ✗



Some examples are more informative!

Did Deadpool **have** → **marry** a kid?

Did → **Does** Deadpool have a kid?

Did Deadpool have **a kid** → **no kids**?

Did Deadpool **have** → **kill** a kid?

Did Deadpool have **more than one** kid?

Did Deadpool have a kid?

Did Deadpool have a **kid** → **daughter**?

Does his bride have **a queen size ring**?

Did Deadpool have a **kid** → **son**?

Did Deadpool have **a title of Queen**?

Some examples are more informative!

Did Deadpool **have** → **marry** a kid?

Did → **Does** Deadpool have a kid?

Did Deadpool have **a kid** → **no kids**?

Did Deadpool **have** → **kill** a kid?

Did Deadpool have **more than one** kid?

Did Deadpool have a kid?

Did Deadpool have a **kid** → **daughter**?

Does his bride have **a queen size ring**?

Did Deadpool have a **kid** → **son**?

Did Deadpool have **a title of Queen**?

Polyjuice's role: Devil's advocate

Generators as **devil's**

advocates:

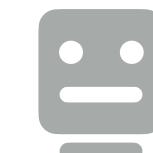
Combat human's biased
beliefs (on SHAP
explanations)

through careful data re-
ranking & selection.

Evaluate with SHAP explanation

Did Deadpool have a kid? Y ✓

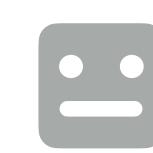
"have", "a" seems unimportant?



Surprise variance: small distance, diff pred.

Did Deadpool ~~have~~ lose a kid? N ✓

Did Deadpool have ~~a~~ no kid? N ✓



"Kid" seems important?

Surprise invariance: large distance, same pred.

Did Deadpool have a ~~kid~~ son? Y ✗

Takeaways

Models can play the role of...

Executor:

Generalize human assumptions

Ideator:

Compensate for human omissions

Devil's advocate:

Combat human bias

For...

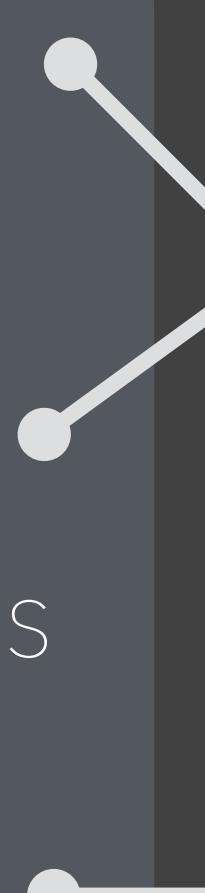
Test generation

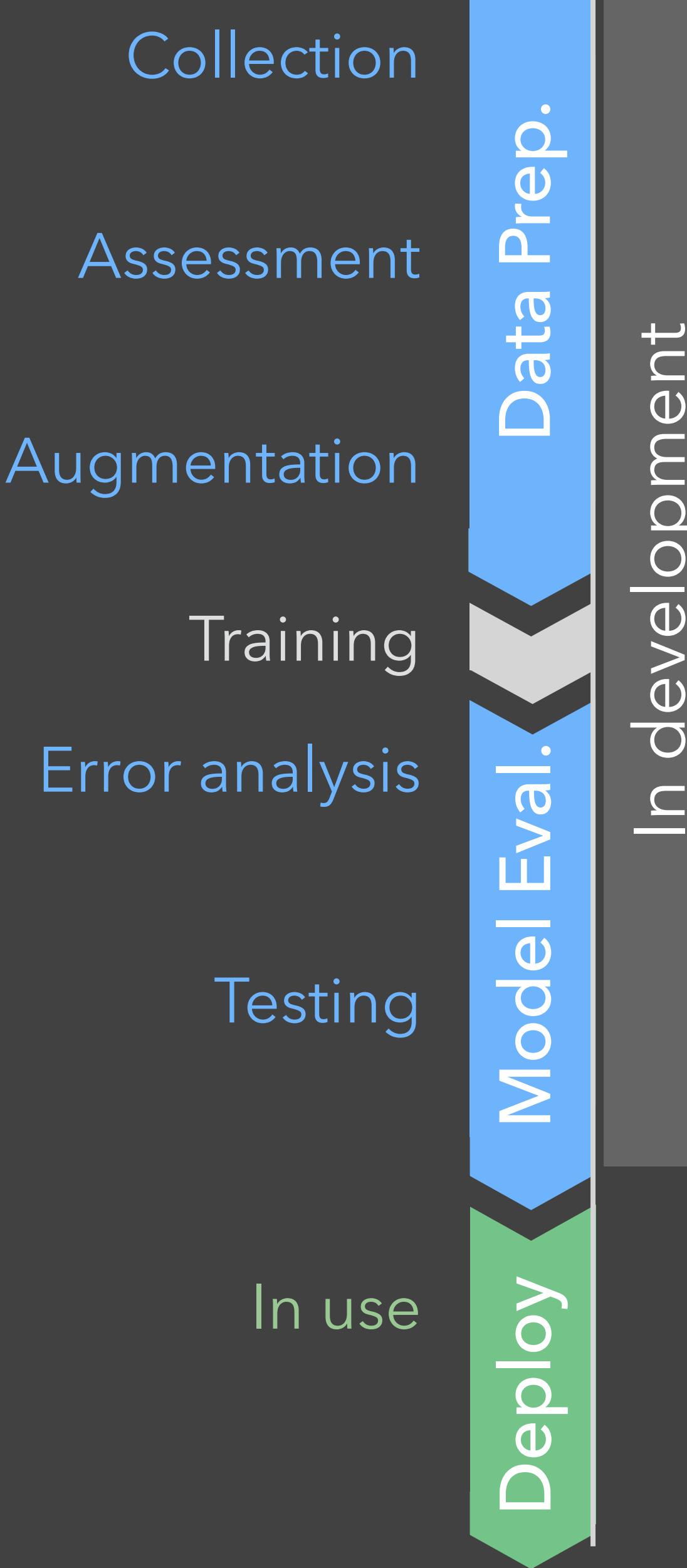
Local explanations

If they have good...

Controllability

Output selection





Debugging is multi-stage.

Mitigate developer bias through assisted methods

Slice discovery that exposes inherent dataset property

Utilize task characteristics

Perturbation that compensates human creativity

Use NLP models as different assistive roles

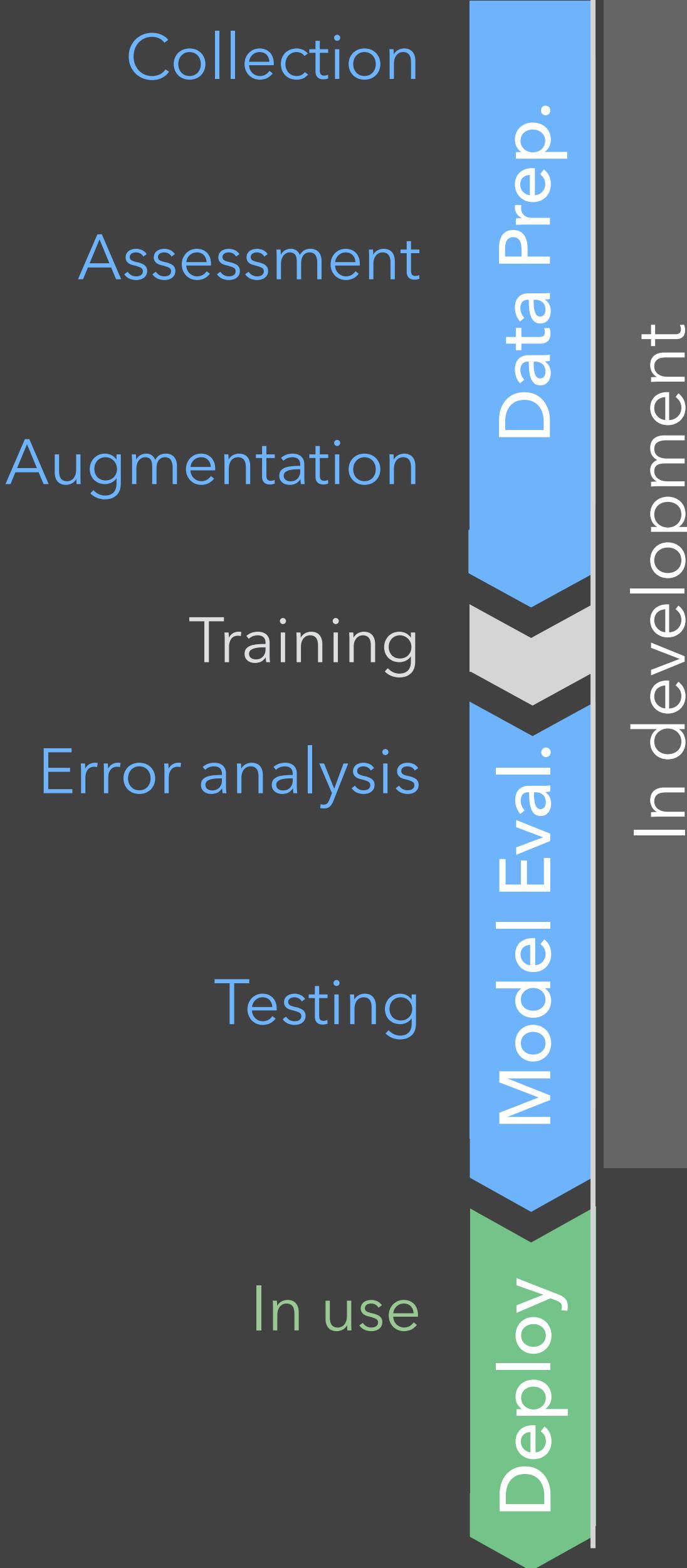
Help end users to recover from deployed model errors

Help users identify AI errors

Provide explanation and reflection

Help users work around AI errors

Provide users with controls through interactions



Debugging is multi-stage.

Mitigate developer bias through assisted methods

Slice discovery that exposes inherent dataset property

Utilize task characteristics

Perturbation that compensates human creativity

Use NLP models as different assistive roles

Help end users to recover from deployed model errors

Help users identify AI errors

Provide explanation and reflection

Help users work around AI errors

Provide users with controls through interactions

Our fellow analyzers have done their best...

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

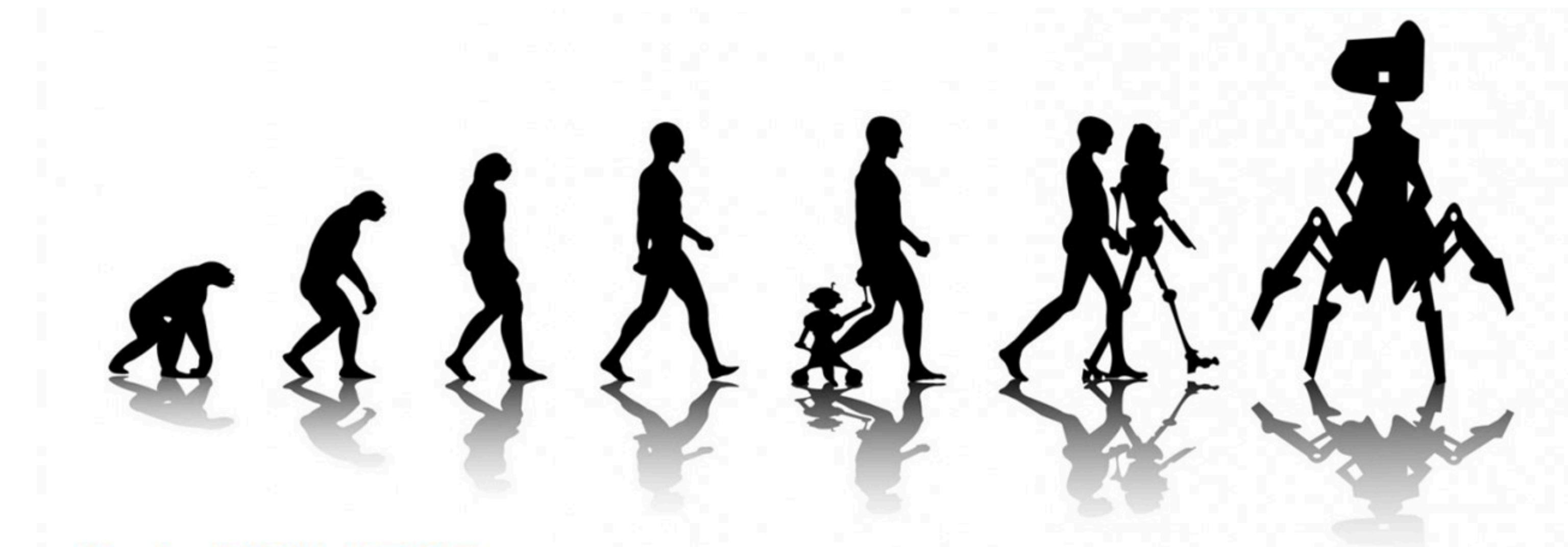
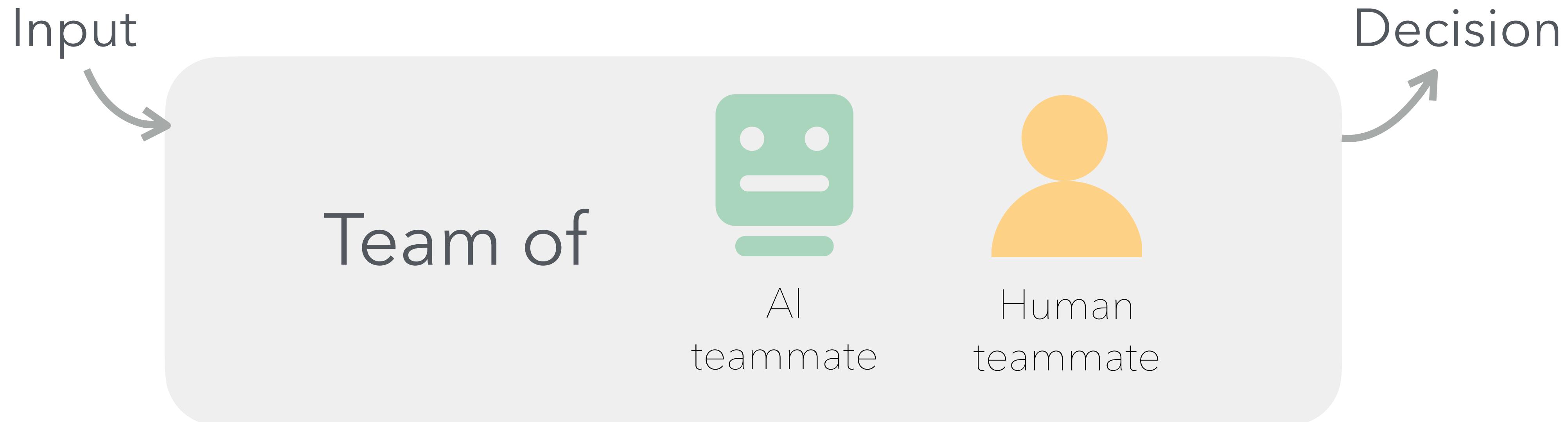


How *experts* train the model

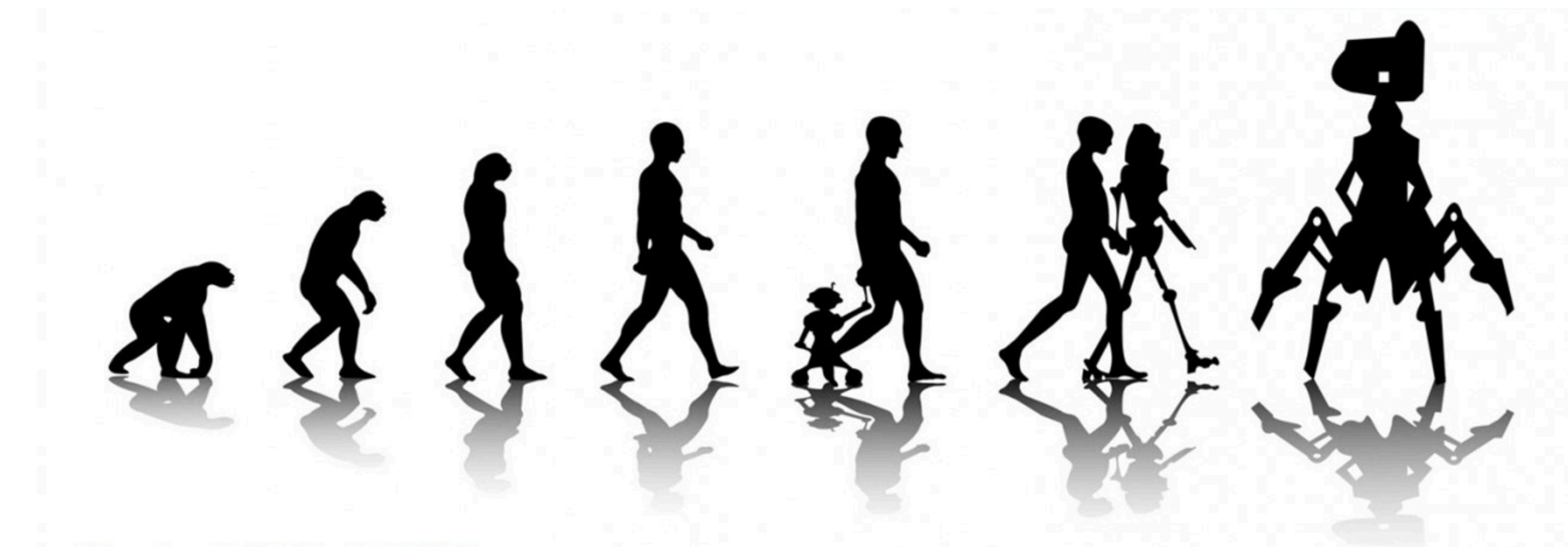
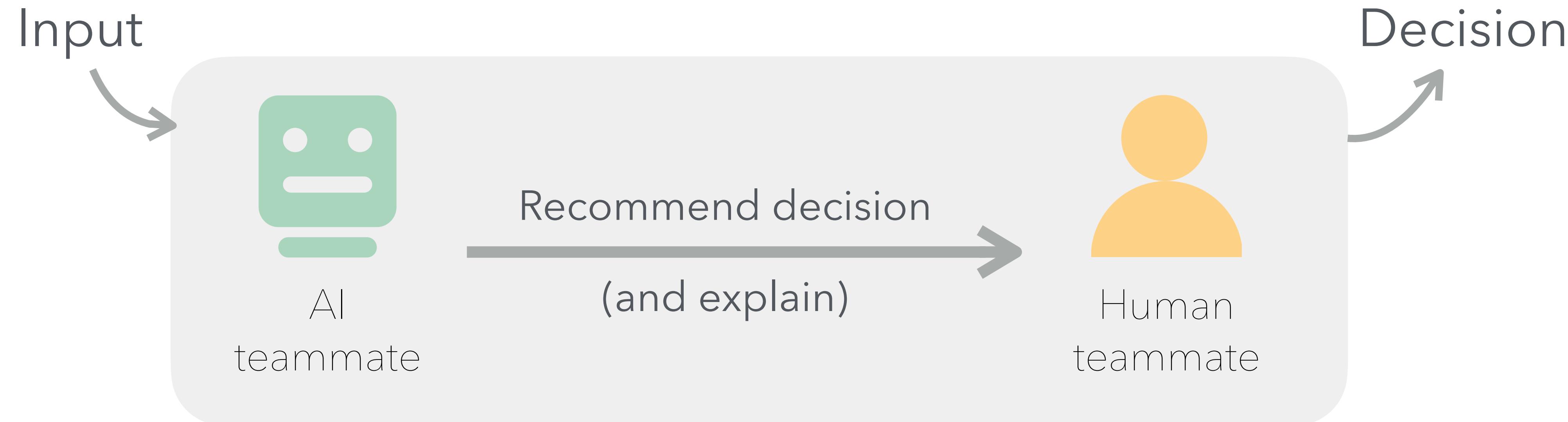


How *end users* use the model

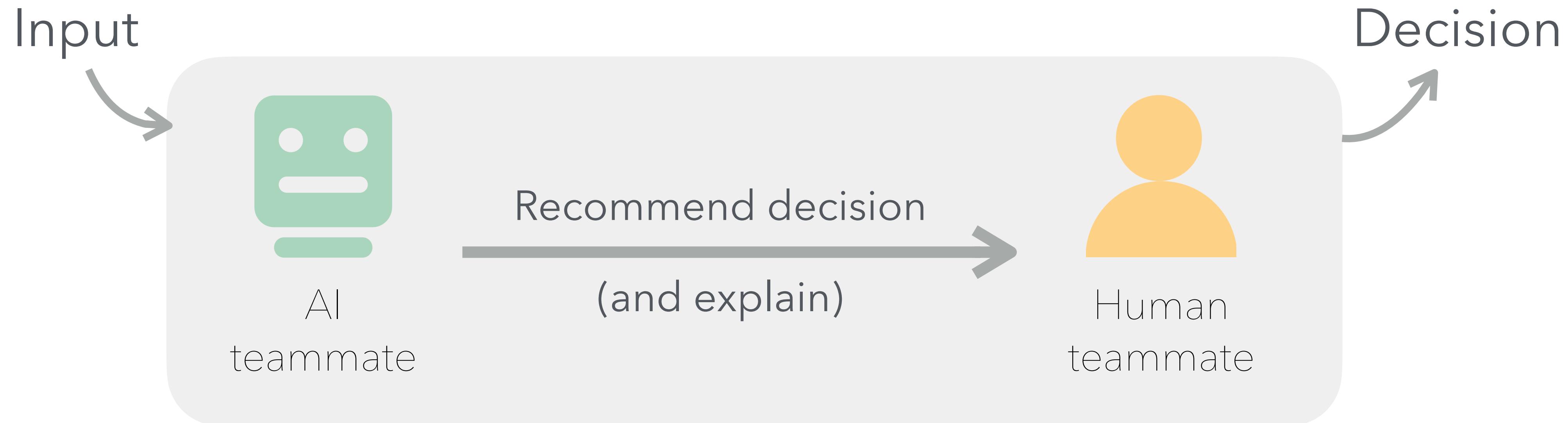
End users **still** need to debug AI teammates.



End users **still** need to debug AI teammates.



Human-AI team requires appropriate trust.



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

But, appropriate reliance is hard.

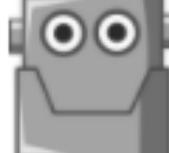
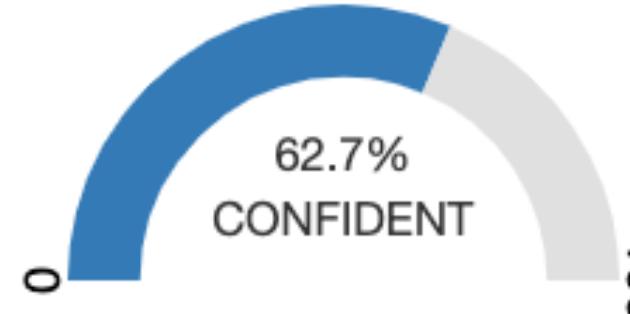
Crowdsourced study, with 1,626 participants:

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

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

Always displayed

Only in some conditions

① Guidelines	② Test	③ Task Instructions	④ Task	⑤ Survey
<p>I, like others was very excited to read this book. I thought it would show another side to how the Tate family dealt with the murder of their daughter Sharon. I didn't have to read much to realize however that the book is was not going to be what I expected. It is full of added dialog and assumptions. It makes it hard to tell where the truth ends and the embellishments begin. It reads more like fan fiction than a true account of this family's tragedy. I did enjoy looking at the early pictures of Sharon that I had never seen before but they were hardly worth the price of the book.</p> <p>c</p>			<p>a Round: 1/50 #Correct Labels: 0</p> <p>Is the sentiment of the review positive or negative? Show Guidelines</p> <p>b </p> <p>Mostly Positive Mostly Negative</p> <p>i Marvin is 62.7% confident about its suggestion.</p>  <p>62.7% CONFIDENT</p>	

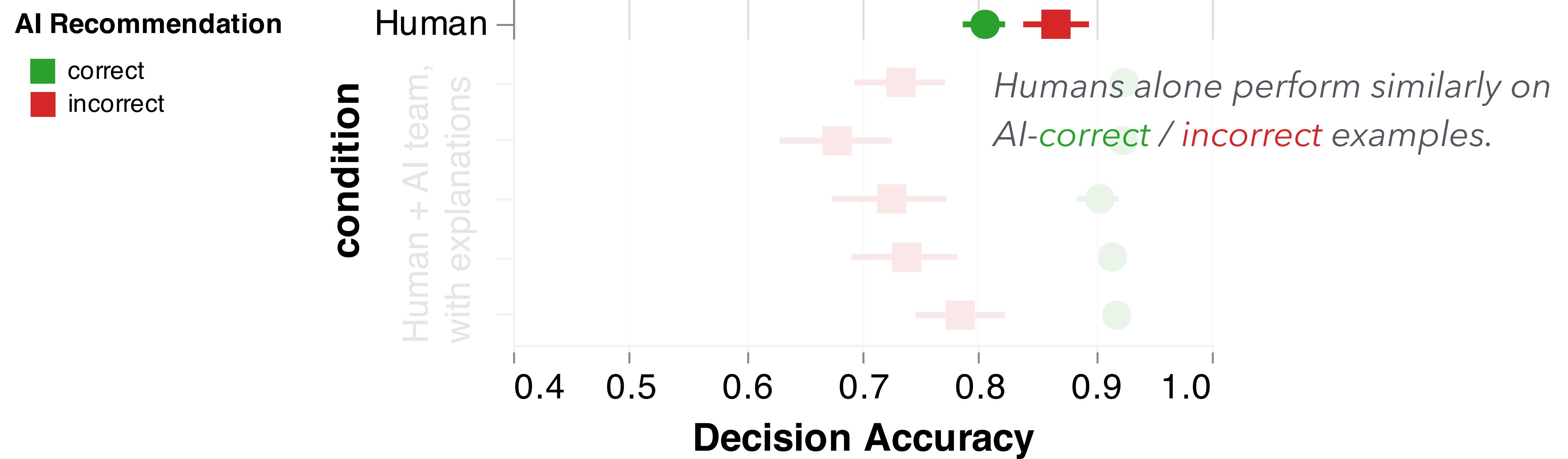
Does the Whole Exceed its Parts? The Effect of AI Explanations on Complementary Team Performance

CHI 2020

But, appropriate reliance is **hard**.

Inappropriate reliance!

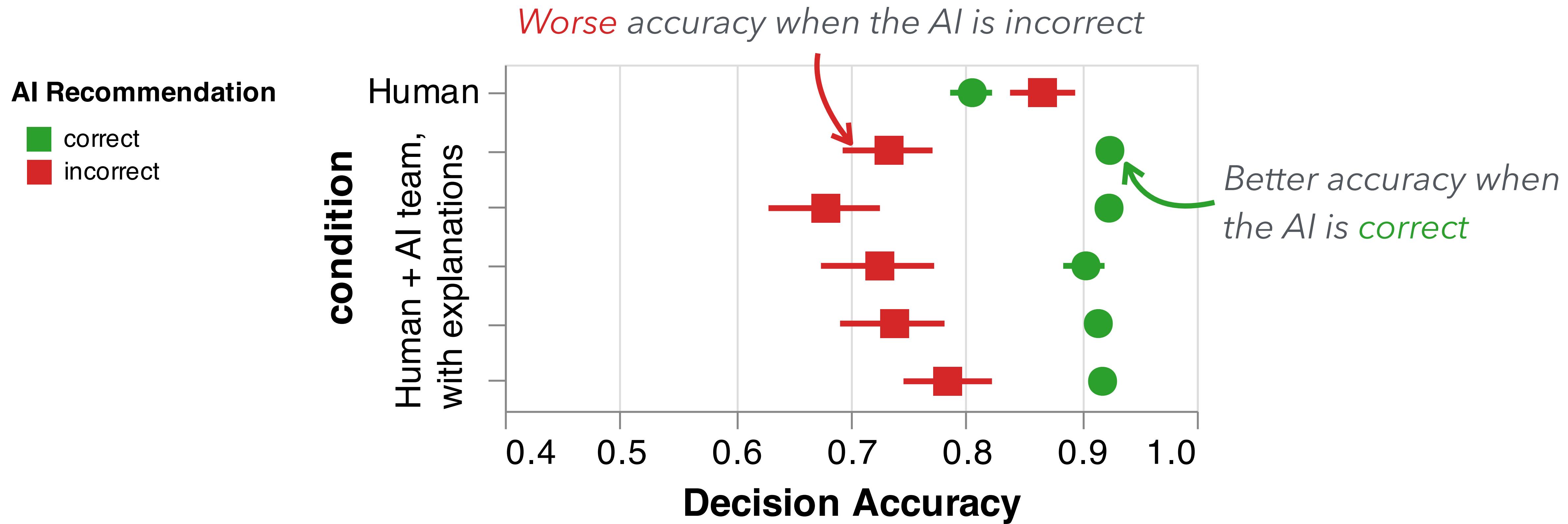
People blindly follow model recommendations, even when the model is **incorrect**.



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

① Guidelines ② Test ③ Task Instructions ④ Task ⑤ Survey

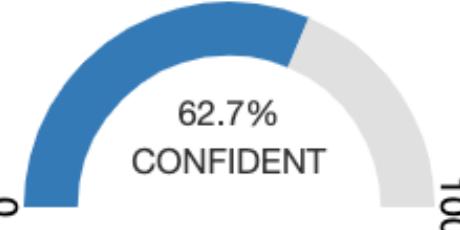
c I, like others was very excited to read this book. I thought it would show another side to how the Tate family dealt with the murder of their daughter Sharon. I didn't have to read much to realize however that the book was not going to be what I expected. It is full of added dialog and assumptions. It makes it hard to tell where the truth ends and the embellishments begin. It reads more like fan fiction than a true account of this family's tragedy. I did enjoy looking at the early pictures of Sharon that I had never seen before but they were hardly worth the price of the book. **d**

a Round: 1/50 #Correct Labels: 0

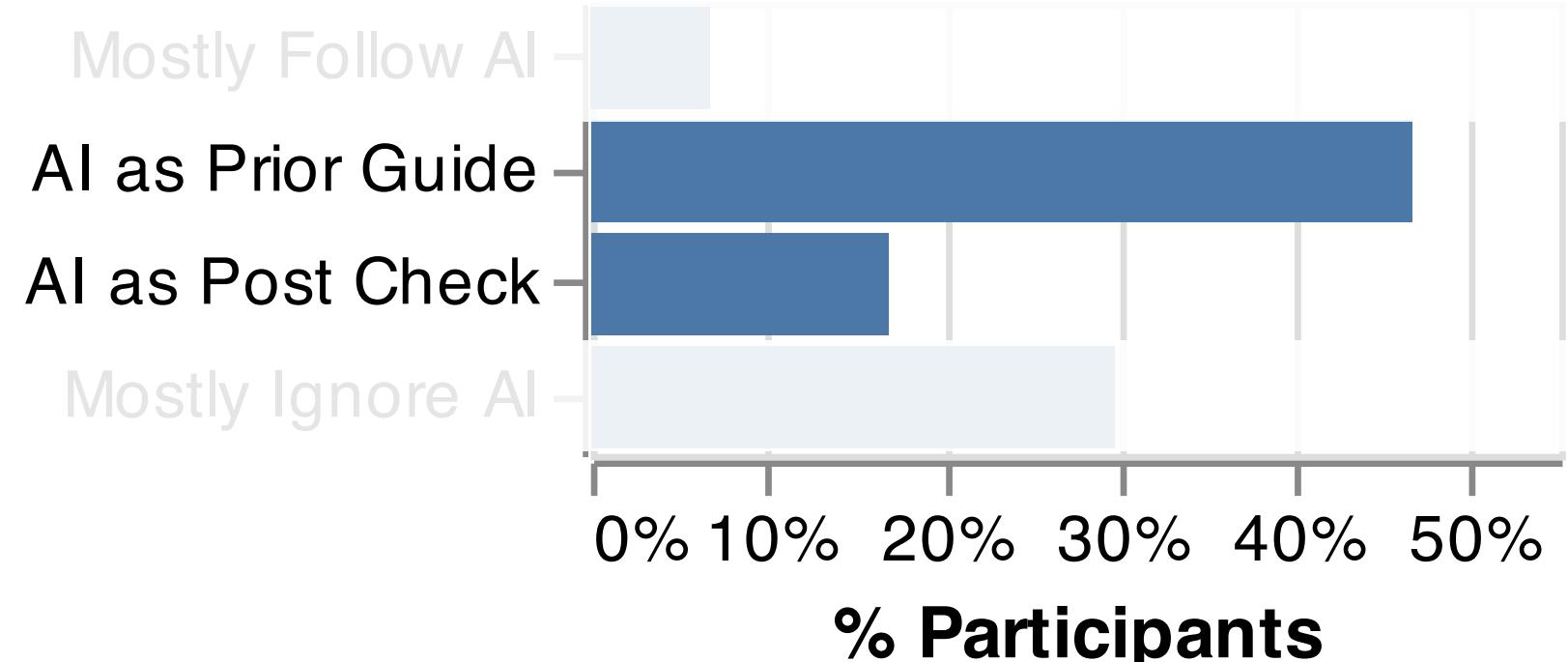
Is the sentiment of the review positive or negative? [Show Guidelines](#)

b  Mostly Positive Mostly Negative

i Marvin is 62.7% confident about its suggestion.



Collaboration



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

a

Question 1 of 20

Your accuracy (so far): 0 / 20

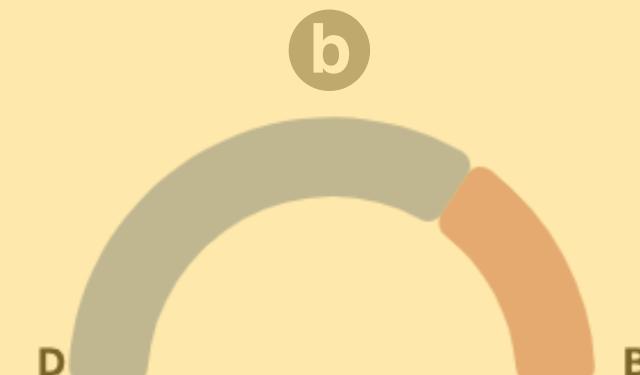
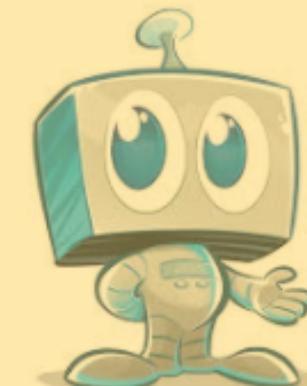
John looks like a professional bodybuilder. He weighs 210 pounds and stands six feet tall, which is the size of an NFL linebacker. John looks huge when he enters the room. Years of gym time have clearly paid off in spades.

Which of the following, if true, weakens the argument?

- [A] John prefers to work out in the morning.
- [B] The average professional bodybuilder is considerably heavier and taller than the average NFL linebacker.
- [C] John weighed considerably less before he started working out.
- [D] John's father, brothers, and male cousins all look like professional bodybuilders, and none of them have ever worked out.

NEXT

b



I am 68.50% confident in answer D.

I am 31.50% confident in answer B.

c

Reason for D: John's family doesn't work out and still looks like professional bodybuilders. Years of gym time may not be the reason for John's size.

d

Reason for B: John may be the size of an NFL linebacker, but if this statement is true, then John may not look like a professional bodybuilder.

Collaboration

Mostly Follow AI



AI as Prior Guide



AI as Post Check



Mostly Ignore AI



0% 10% 20% 30% 40% 50%

% Participants

Design interactions for appropriate 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

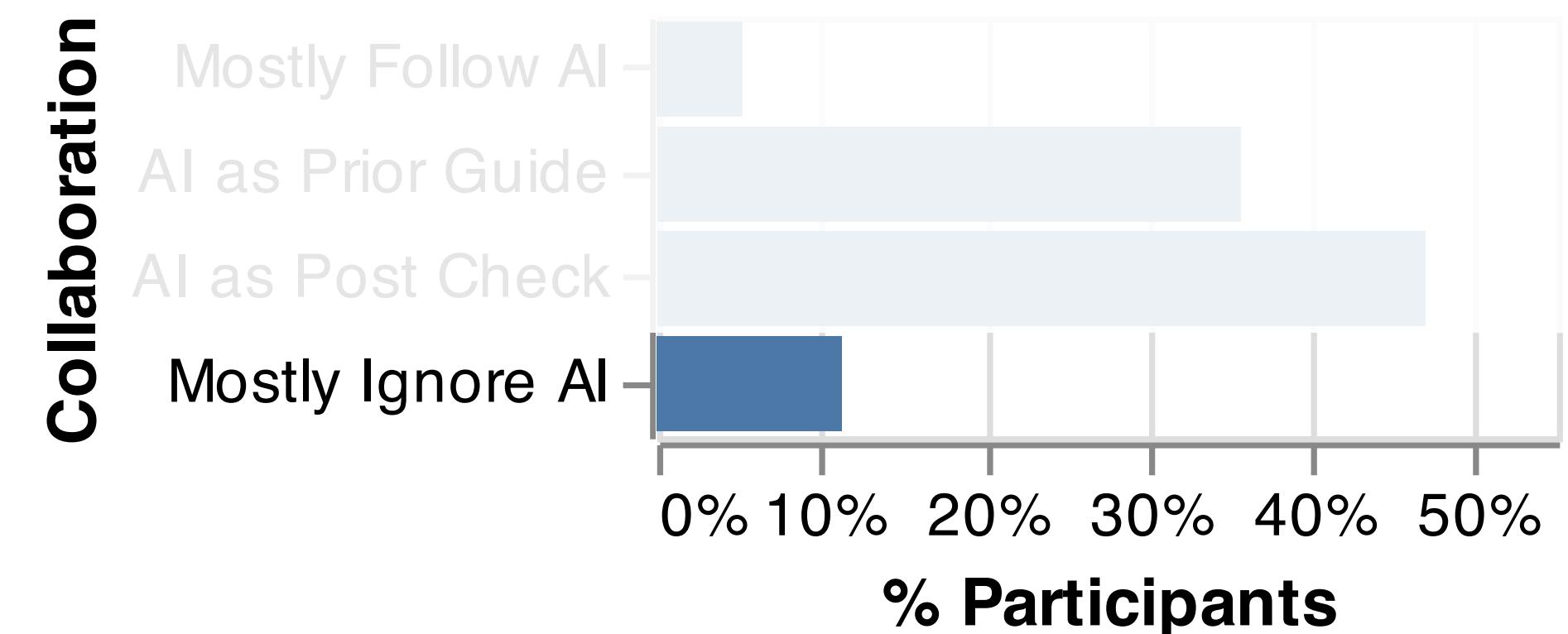
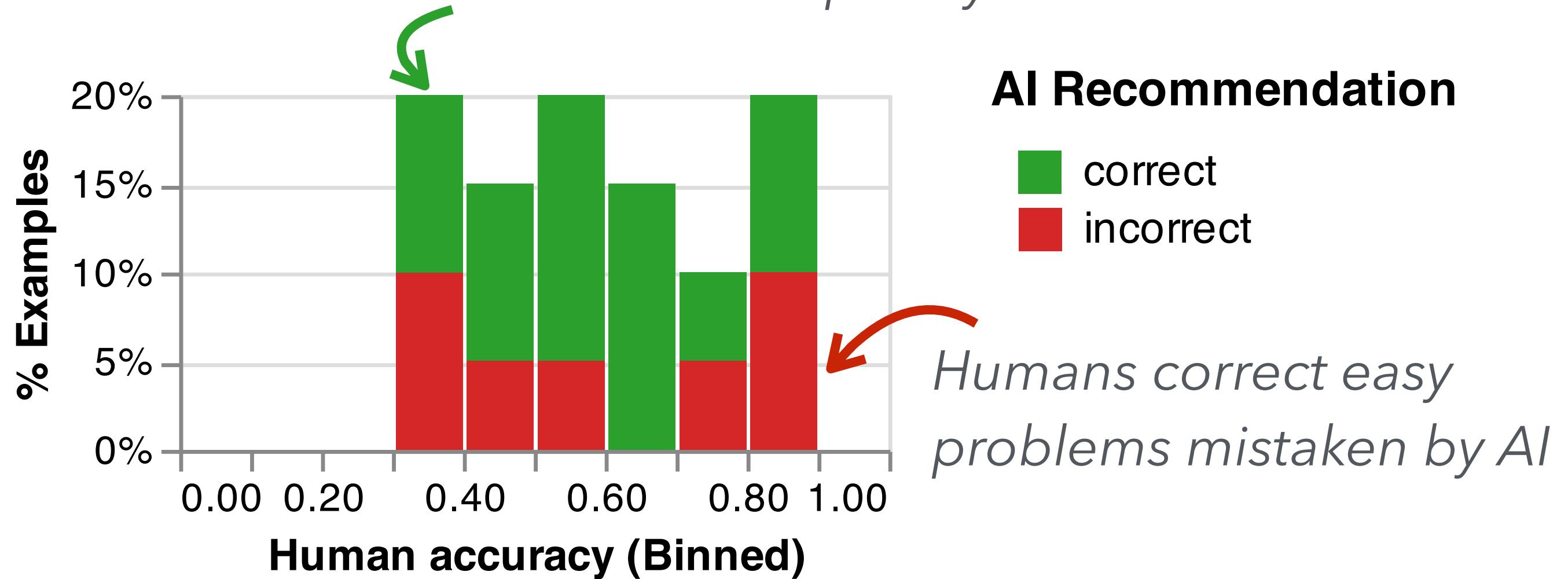
Factors causing inappropriate reliance?

The distribution of human and AI expertise matters.

Human+AI is ineffective if they make the same kinds of mistakes.

Multi-choice question answering task

AI adds value when humans frequently err

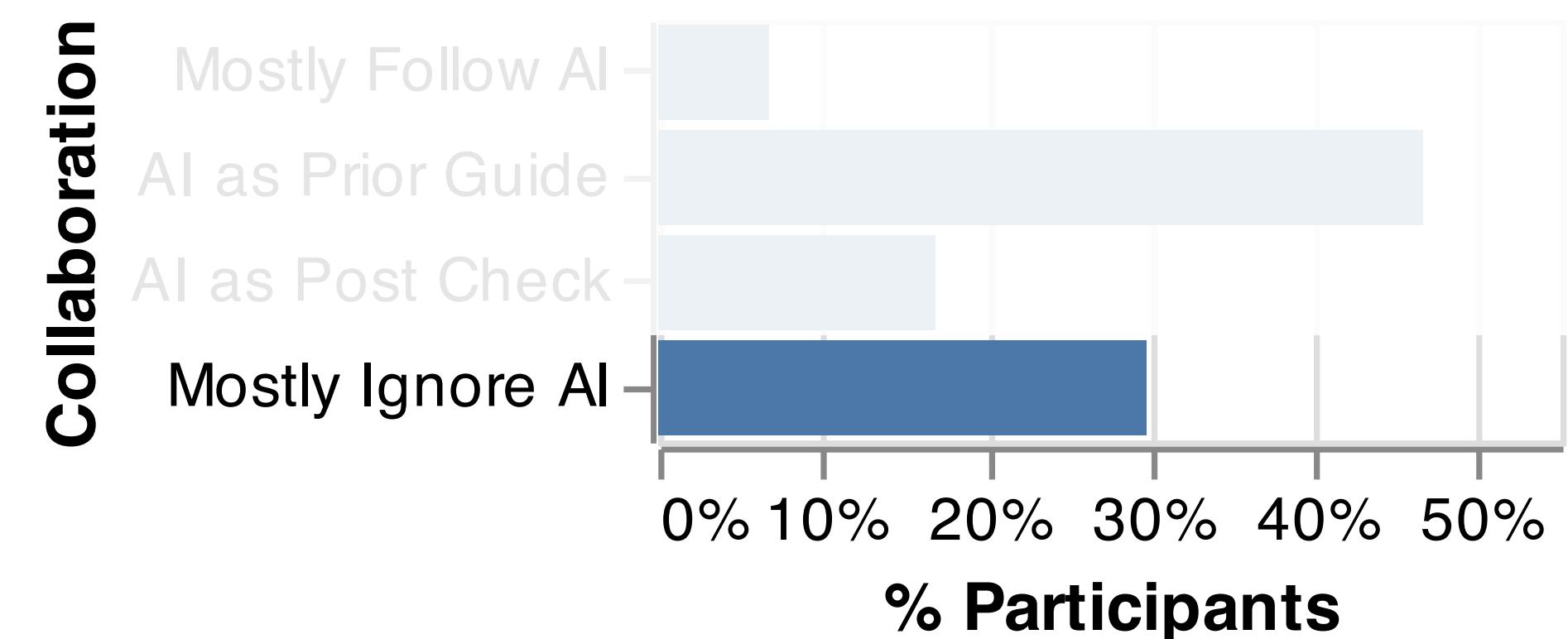
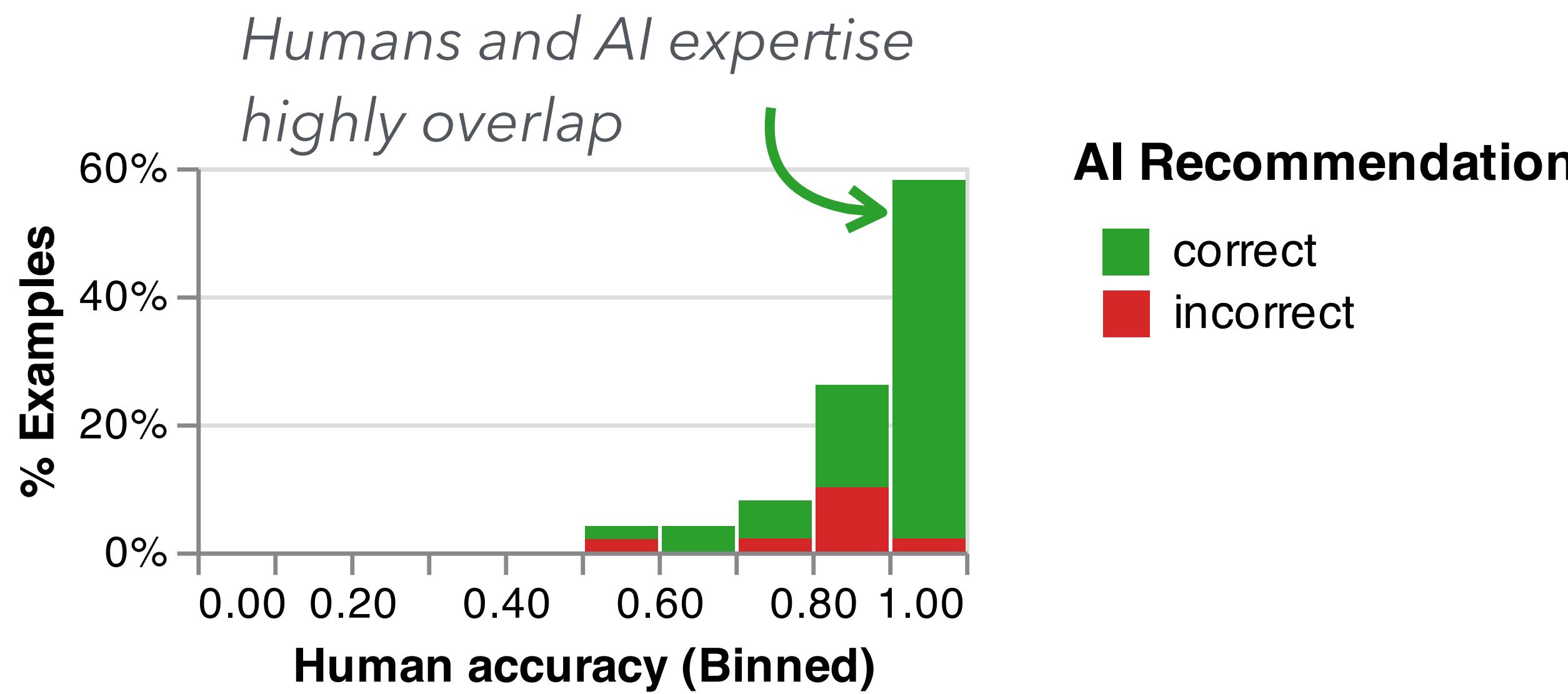


Factors causing inappropriate reliance?

The distribution of human and AI expertise matters.

Human+AI is ineffective if they make the same kinds of mistakes.

Classification task



Choose AIs carefully for the task

The distribution of human and AI expertise matters.

Human+AI is ineffective if they make the same kinds of mistakes.

AI solution

Change the training objective

Directly optimize for complementary behavior

HCI solution

Re-think AI's roles, to help in other dimensions

speed, cognitive load, etc.

Bansal, Gagan, et al. "Is the Most Accurate AI the Best Teammate? Optimizing AI for Teamwork." AAAI 2021
Feng, Shi, and Jordan Boyd-Graber. "What can ai do for me? evaluating machine learning interpretations in cooperative play." IUI 2019.
Glassman, Elena L., et al. "OverCode: Visualizing variation in student solutions to programming problems at scale." TOCHI 2015

Takeaways

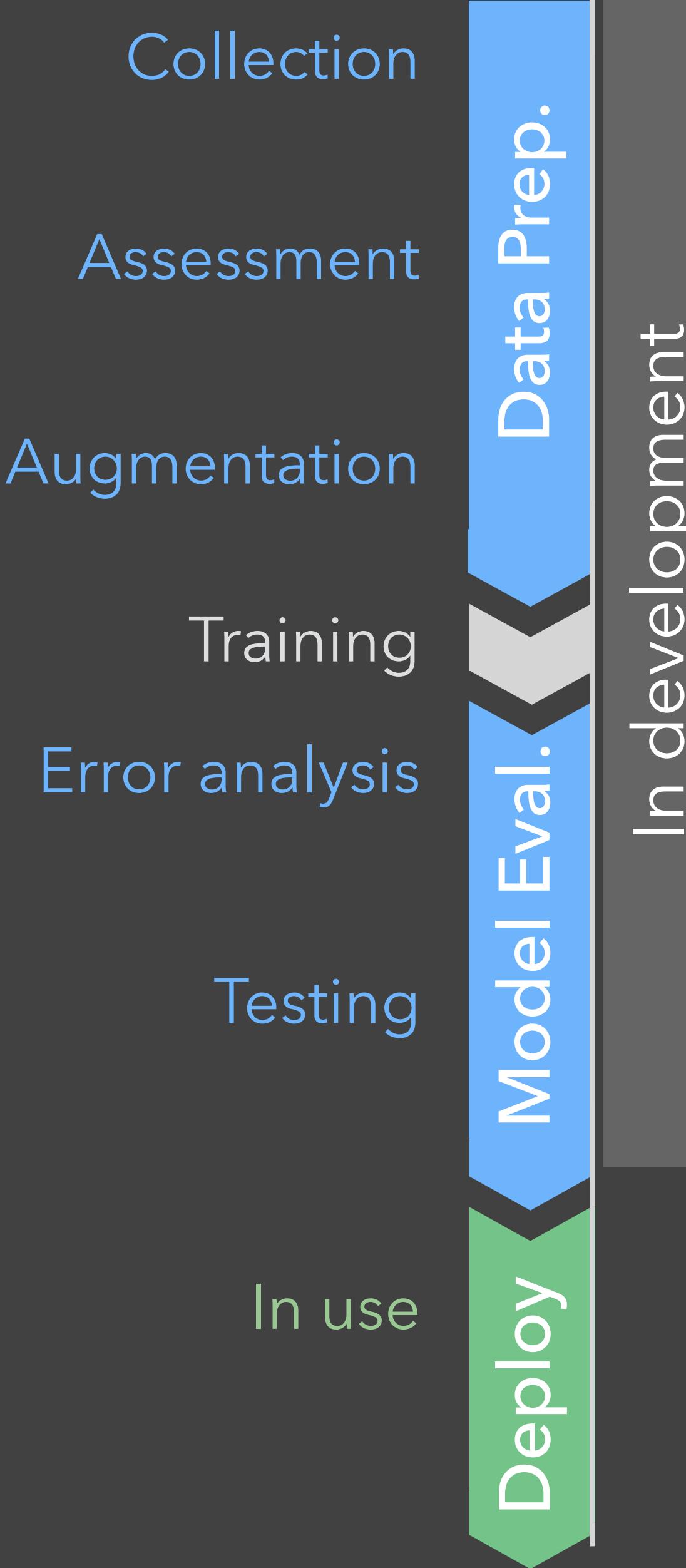
It is important to build the mental model: *Als are great, but they make errors.*

v.s. traditional softwares, AI-infused systems **require appropriate reliance**.

We build reliance through:

“Shared representation”, e.g. intuitive explanations on model rationales

Presenting shared representations, e.g. Explanation format, modality, timing



Debugging is multi-stage.

Mitigate developer bias through assisted methods

Slice discovery that exposes inherent dataset property

Utilize task characteristics

Perturbation that compensates human creativity

Use NLP models as different assistive roles

Help end users to recover from deployed model errors

Help users identify AI errors

Provide explanation and reflection

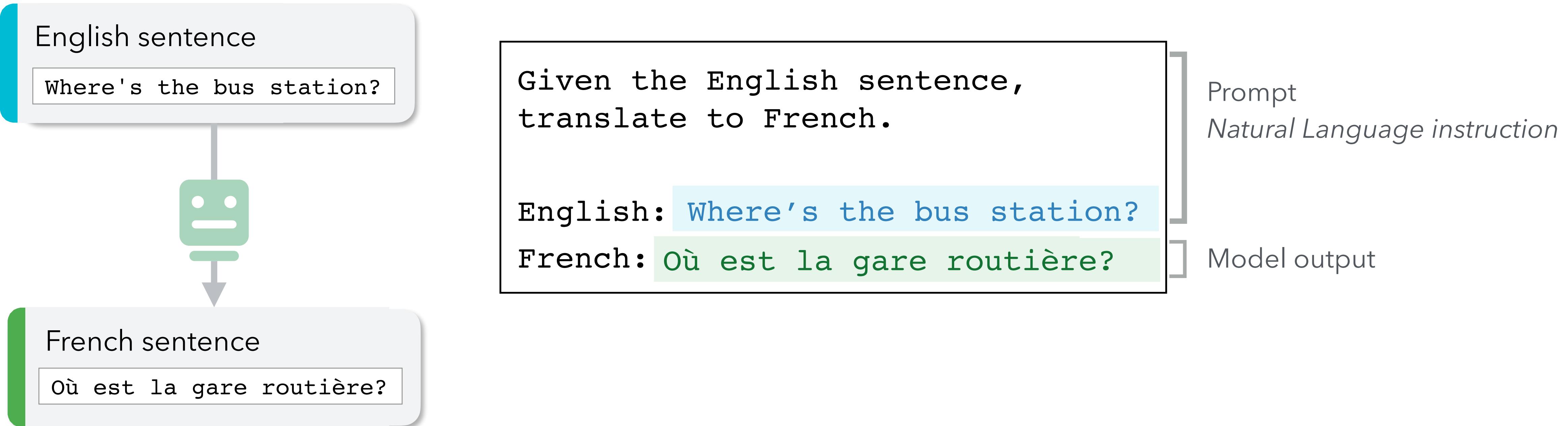
Help users work around AI errors

Provide users with controls through interactions

Large Language Models (LLMs) are **helpful**.

Respond to **natural language prompts** (instructions), accessible to end users!

Translation



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

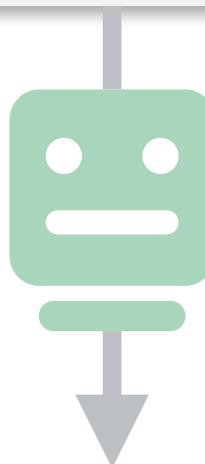
LLM may struggle with complex tasks.

Example: Peer review rewriting

Rewriting

Original feedback

Sherry could improve her presentation...



More friendly feedback

Sherry, you have too many words on your...

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

Original feedback:

Sherry could improve her presentation skills. She has too much text on her slides. Her presentation meanders from topic to topic without a clear structure. She also does not engage with her audience when she presents.

More friendly feedback:

Sherry, 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 struggle with complex tasks.

Example: Peer review rewriting

Model structure has limitations: reasoning capabilities, exposure bias, etc.
...that end users don't know or understand.

Sherry could improve her presentation...

Original feedback:

How can we help end users interpret and improve model outputs, when they cannot understand or change the model?

More friendly feedback

Sherry, you have too many words on your...

Sherry, 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 struggle with complex tasks.

Example: Peer review rewriting

How can we help end users interpret and improve model outputs, when they cannot understand or change the model?

Original feedback:

AI Chain : Break down an original task into sub-tasks, to provide explicit knobs to users on sub-tasks.

More friendly feedback

Sherry, you have too many words on your...

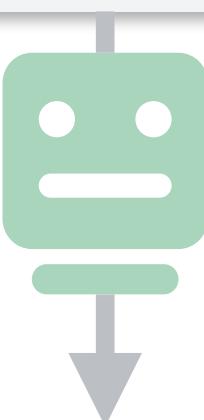
Sherry, you have too many words on your slides. You should use images and bullet points to help get your message across. You should have a clear structure for the presentation. You should also engage with your audience.

The input-output mapping is **convoluted**.

Rewriting

Original feedback

Sherry could improve her presentation...



**Confusing
mapping!**

More friendly feedback

Sherry, you have too many words on your...

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

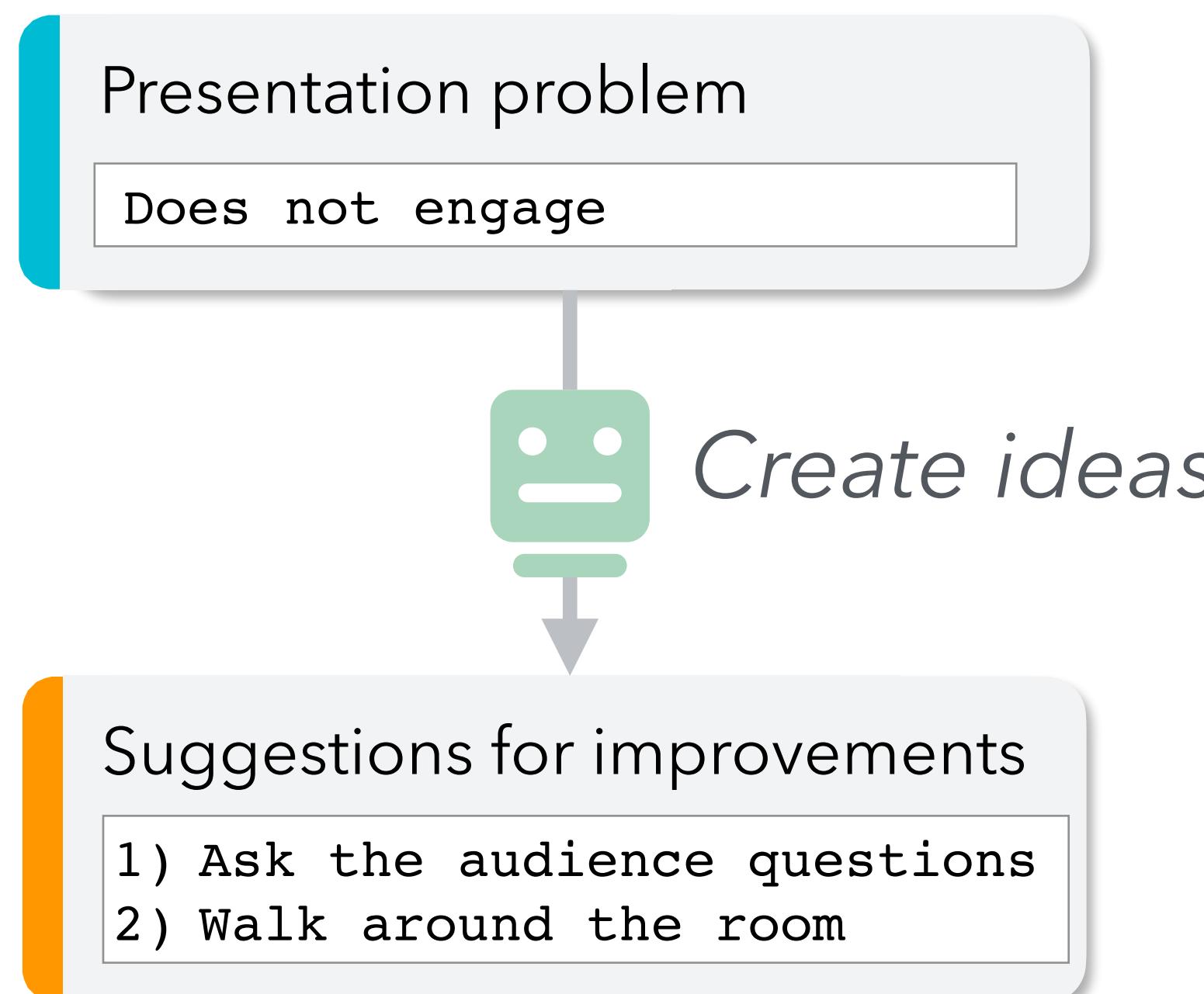
Original feedback:

Sherry could improve her presentation skills. She has too much text on her slides. Her presentation meanders from topic to topic without a clear structure. She also does not engage with her audience when she presents.

More friendly feedback:

Sherry, 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.

Small tasks are more **interpretable** and **controllable**.

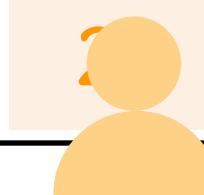


Given the Presentation problem, the following is a list of improvement suggestions.

Problem: Does not engage

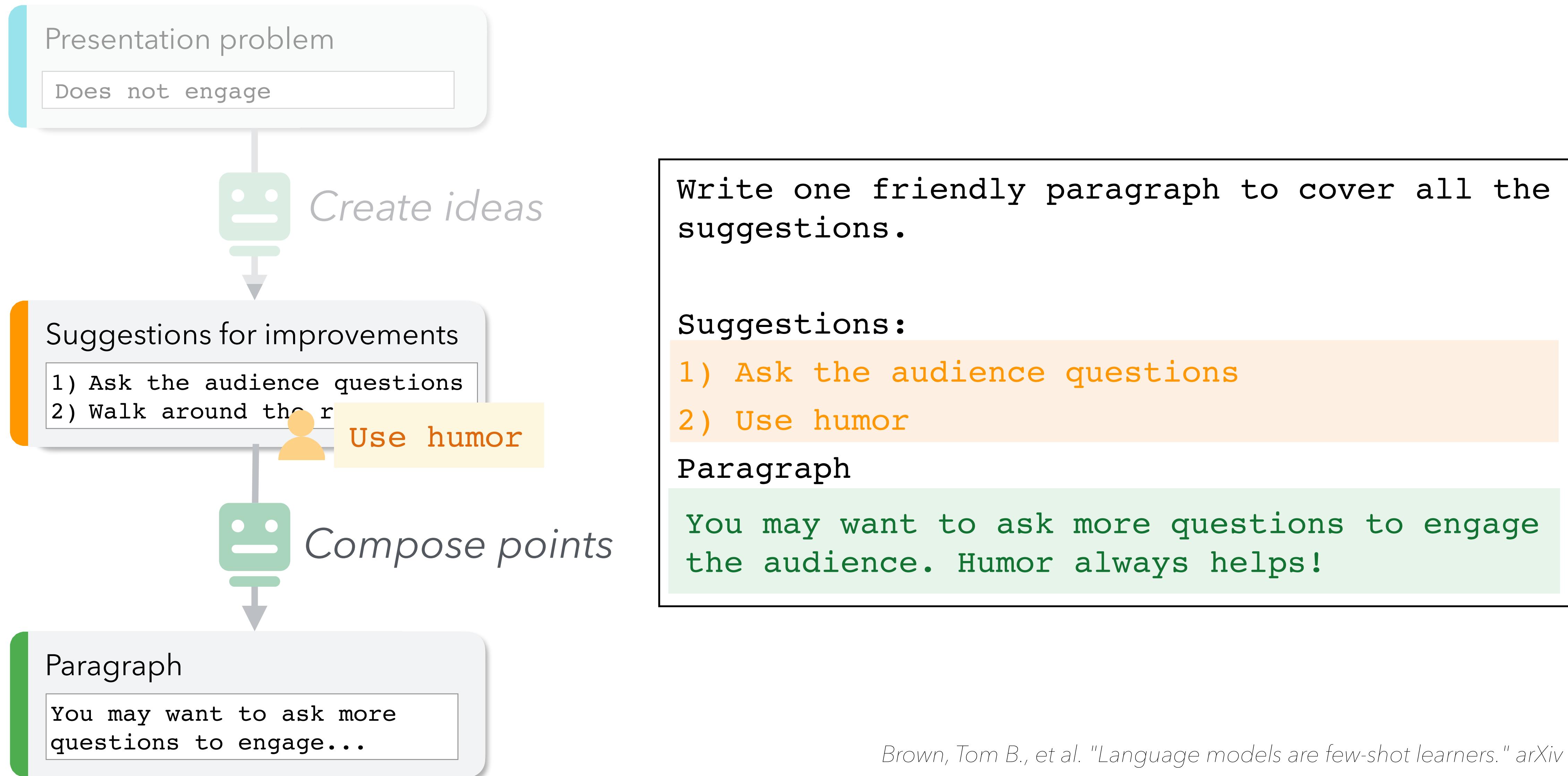
Suggestions for improvements:

1) Ask the audience questions

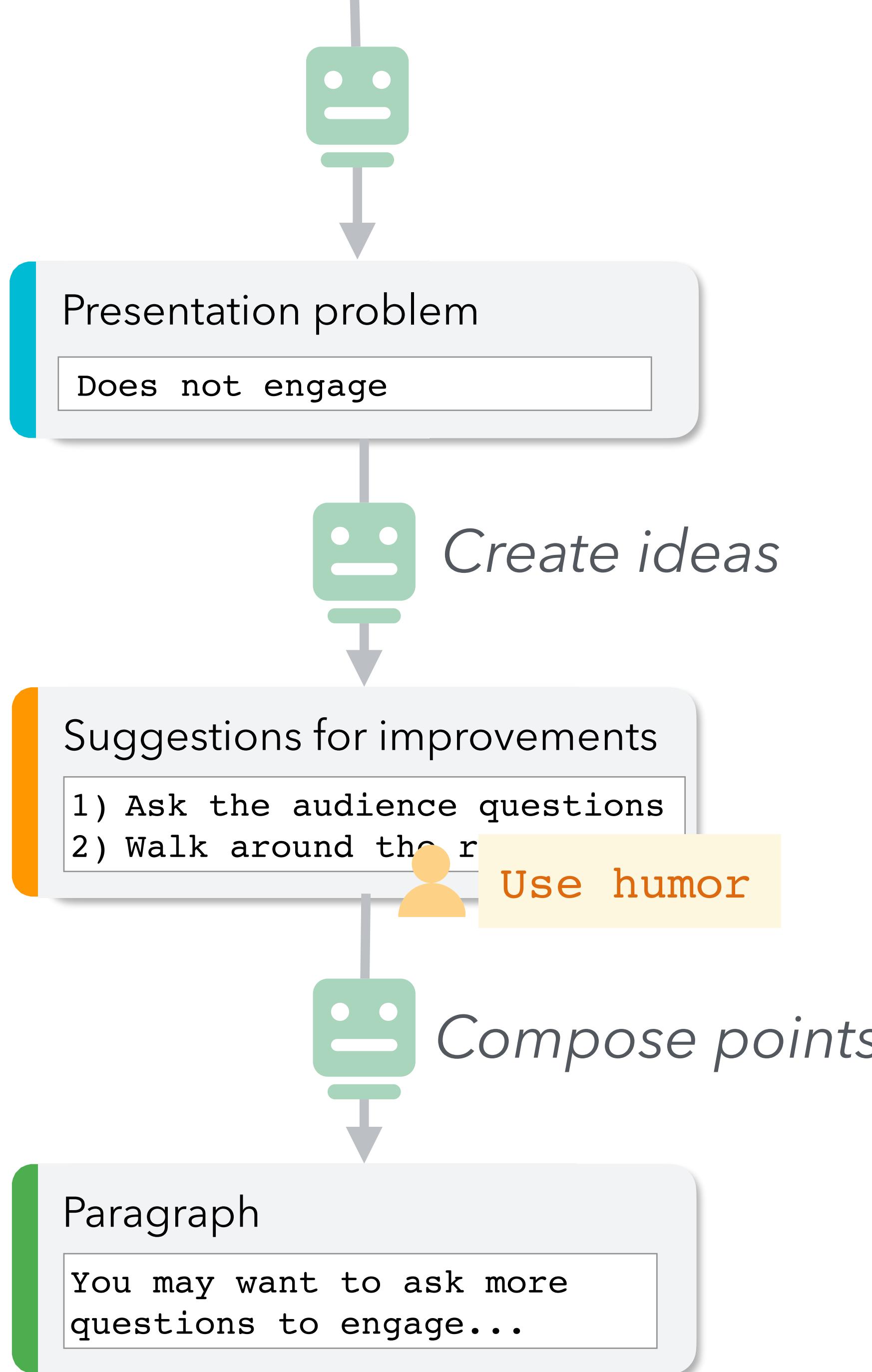


Use humor

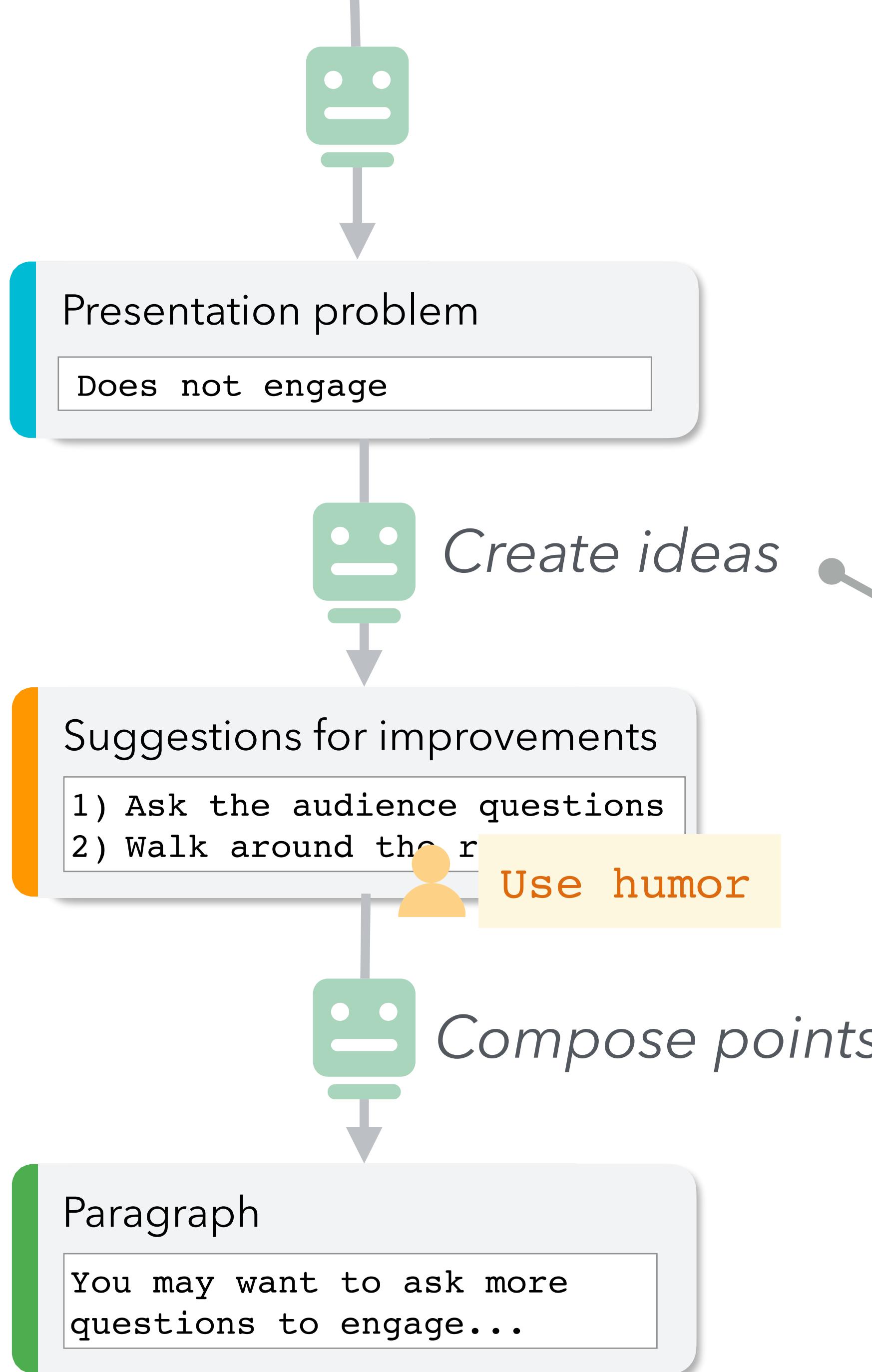
The fix can be propagated to related sub-tasks!



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



Chain =
Independent LLM runs per sub-task
+ inter-step transformation



Chain =
Independent LLM runs per sub-task

1 Identify all presentation problems

2 Ideate suggestions per problem

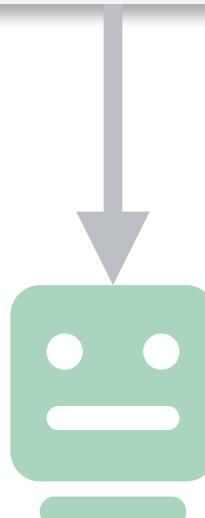
3 Compose them into a paragraph

+ inter-step transformation

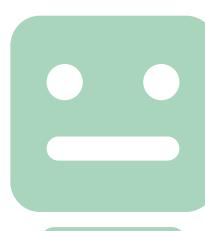
A three-step LLM Chain for peer review rewriting

Original feedback

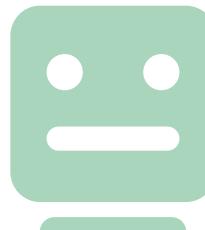
Sherry could improve his presentation skills. He has too much text on his slides. His presentation meanders from topic to topic without a clear structure. He also does not engage with his audience when he presents.



- 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

Sherry 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.

Sherry's problems

- Too much text on slides
- No clear structure
- Does not engage with audience

1

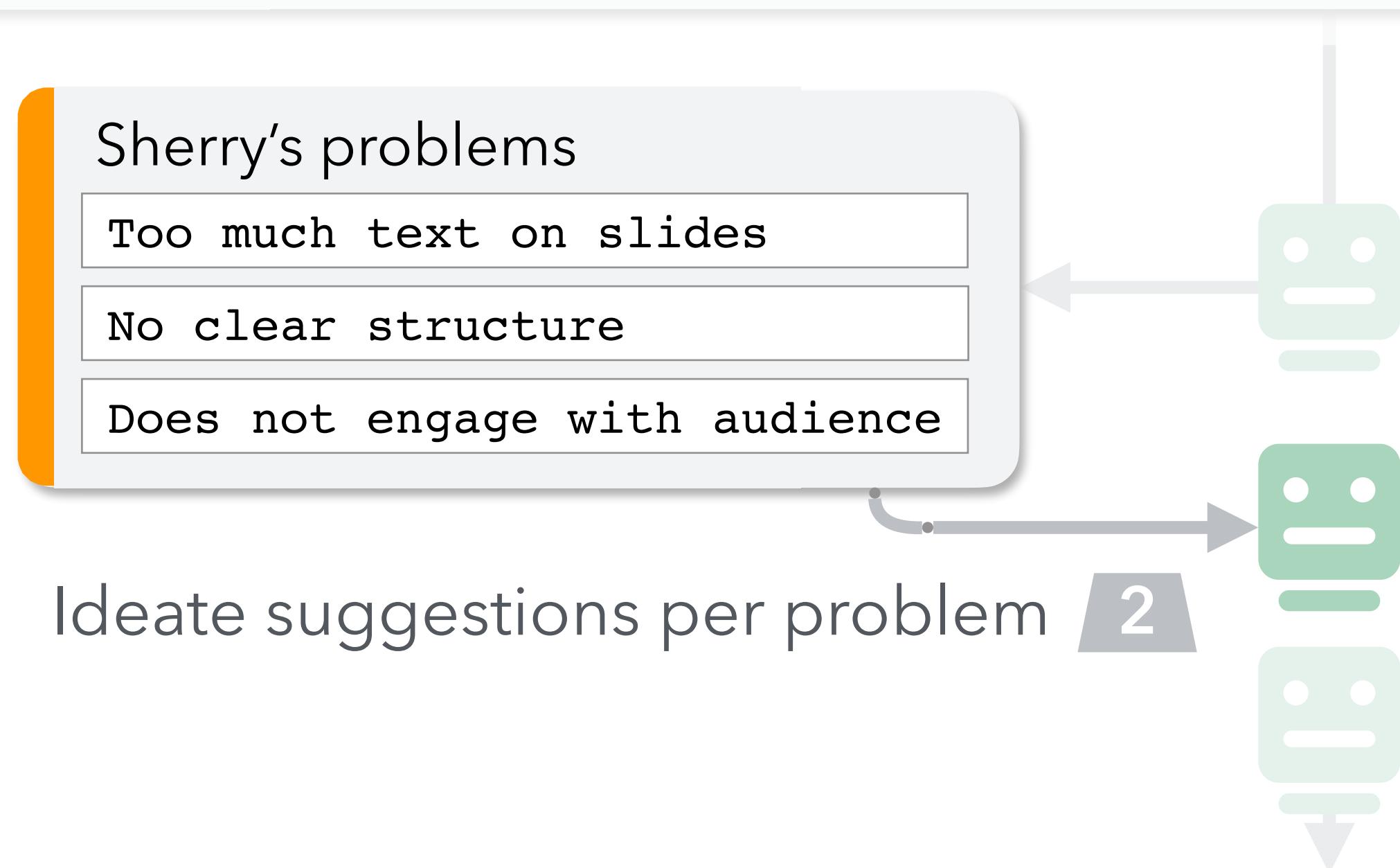
Identify all presentation problems

Friendly paragraph

A three-step LLM Chain for peer review rewriting

Original feedback

Sherry 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.

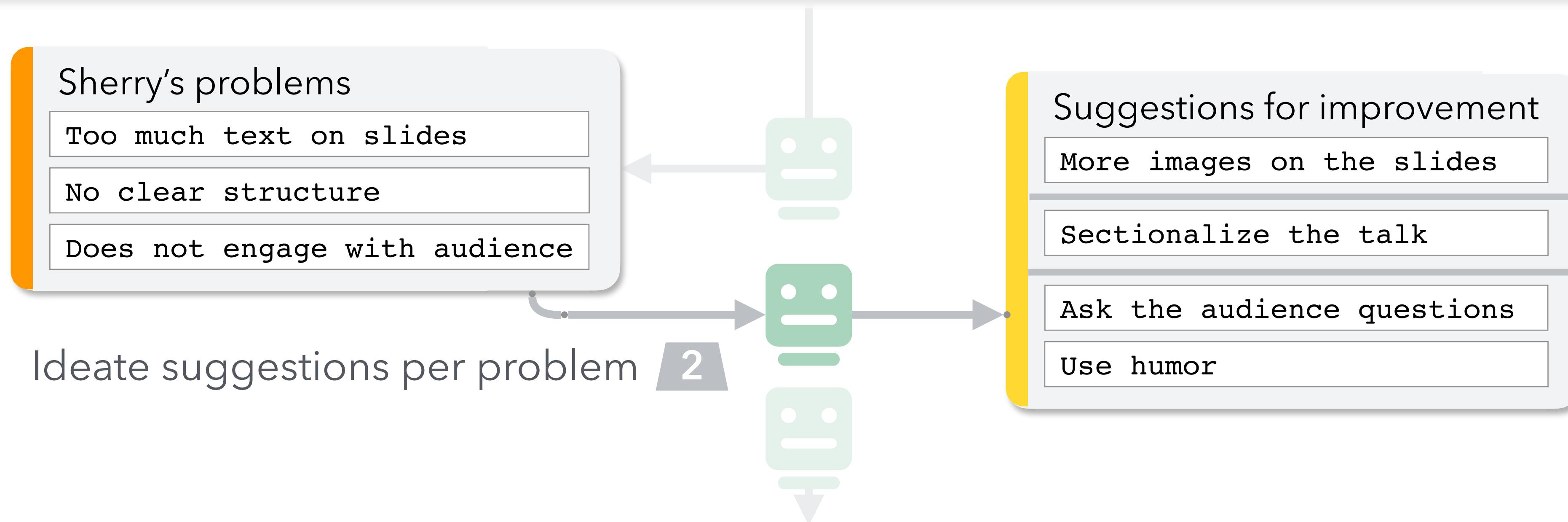


Friendly paragraph

A three-step LLM Chain for peer review rewriting

Original feedback

Sherry 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.

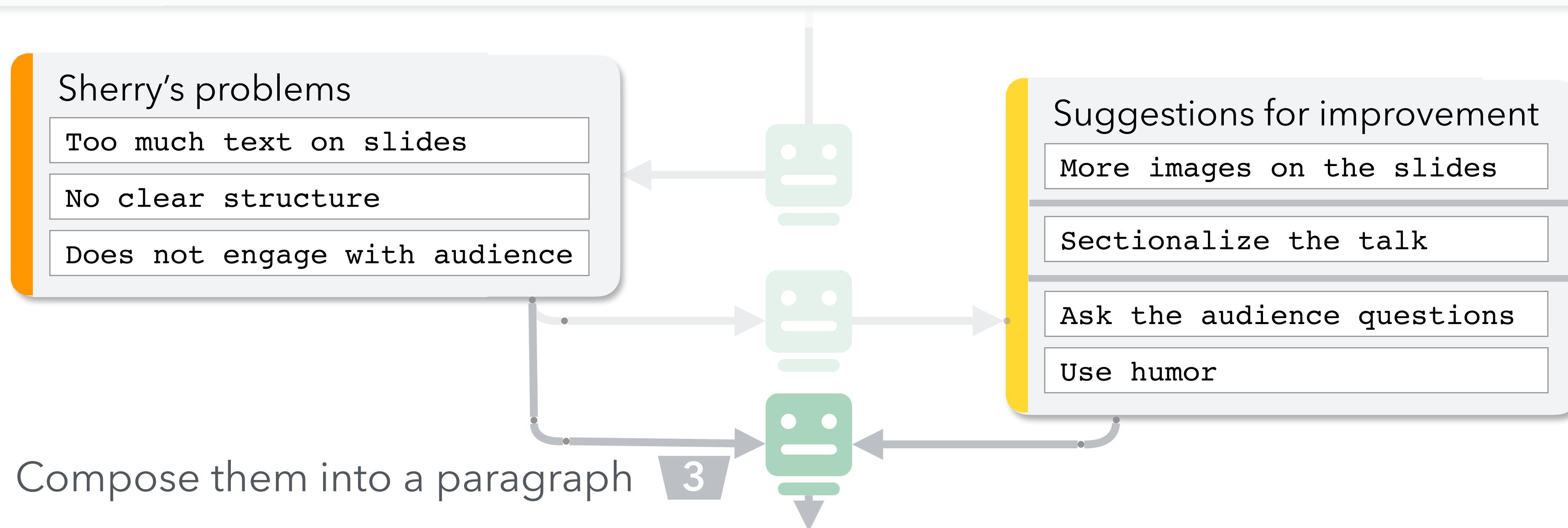


Friendly paragraph

A three-step LLM Chain for peer review rewriting

Original feedback

Sherry 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.

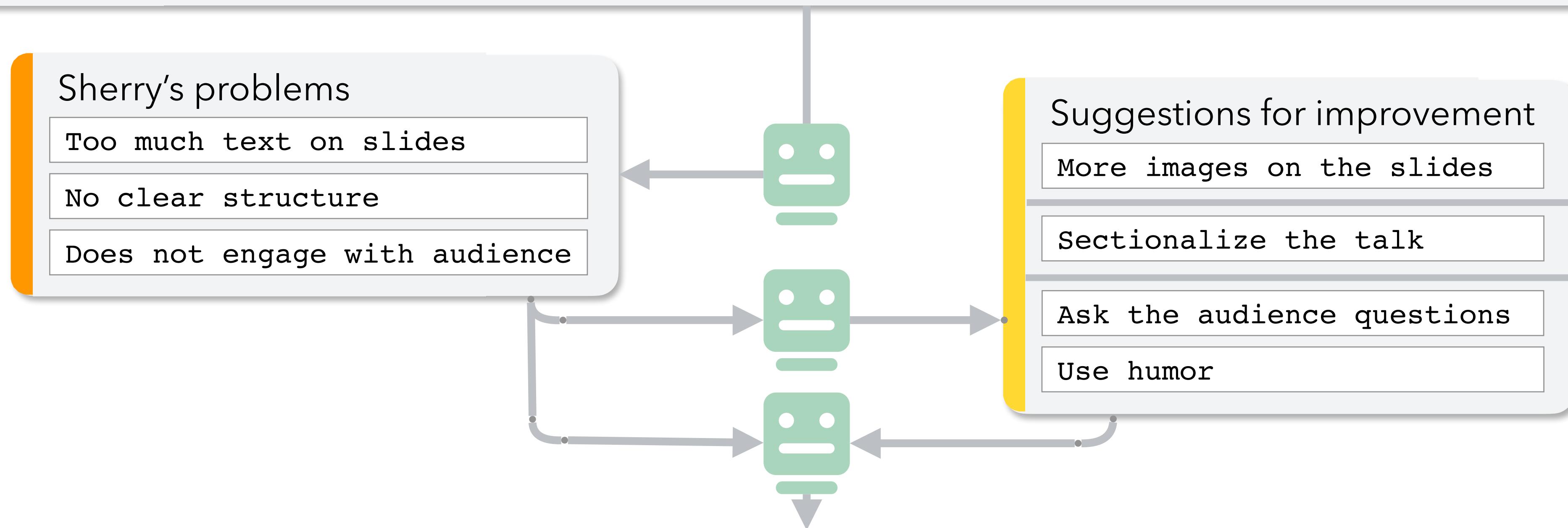


Original feedback

A three-step LLM Chain for peer review rewriting

Original feedback

Sherry 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

Sherry, 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.

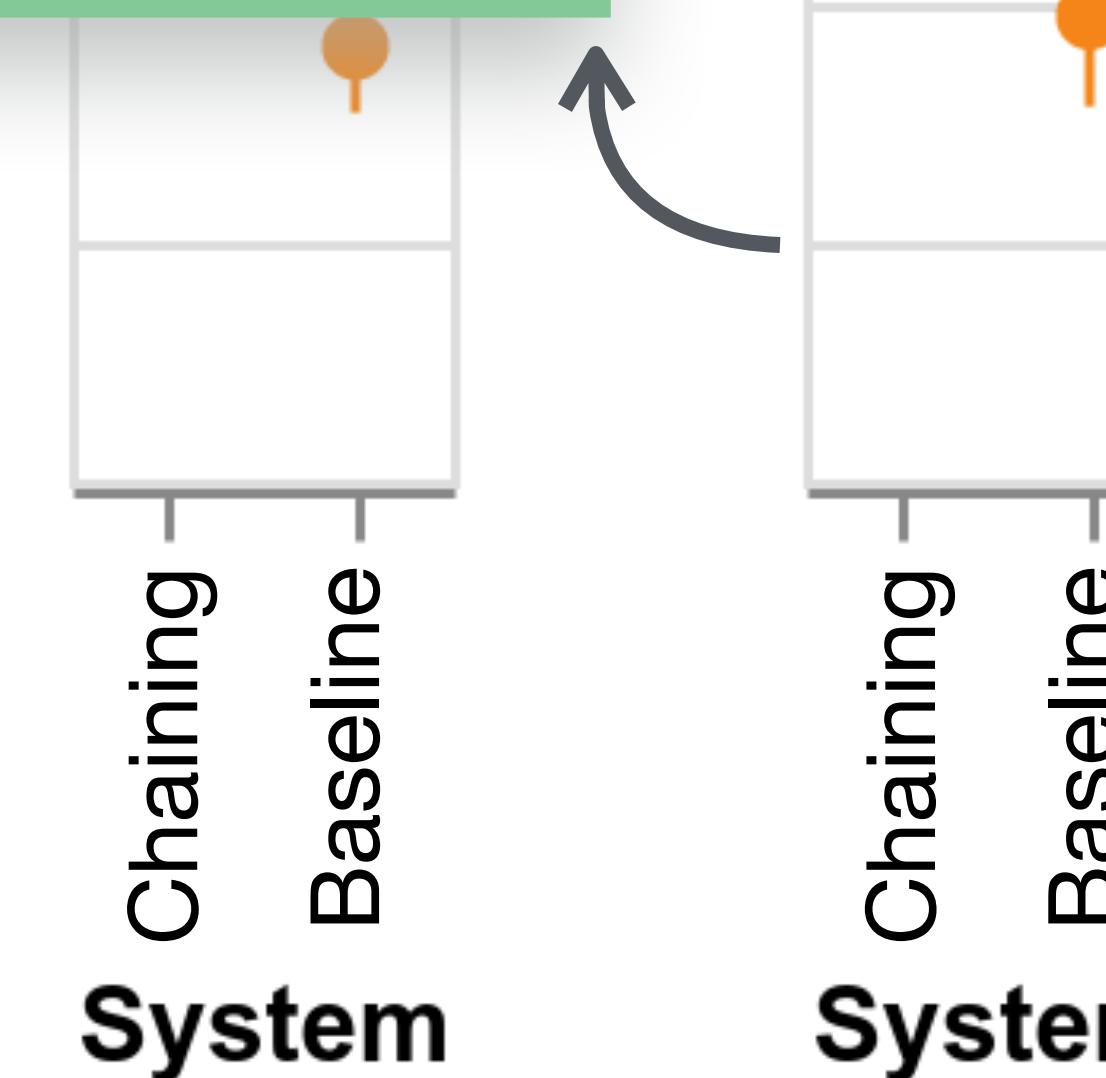
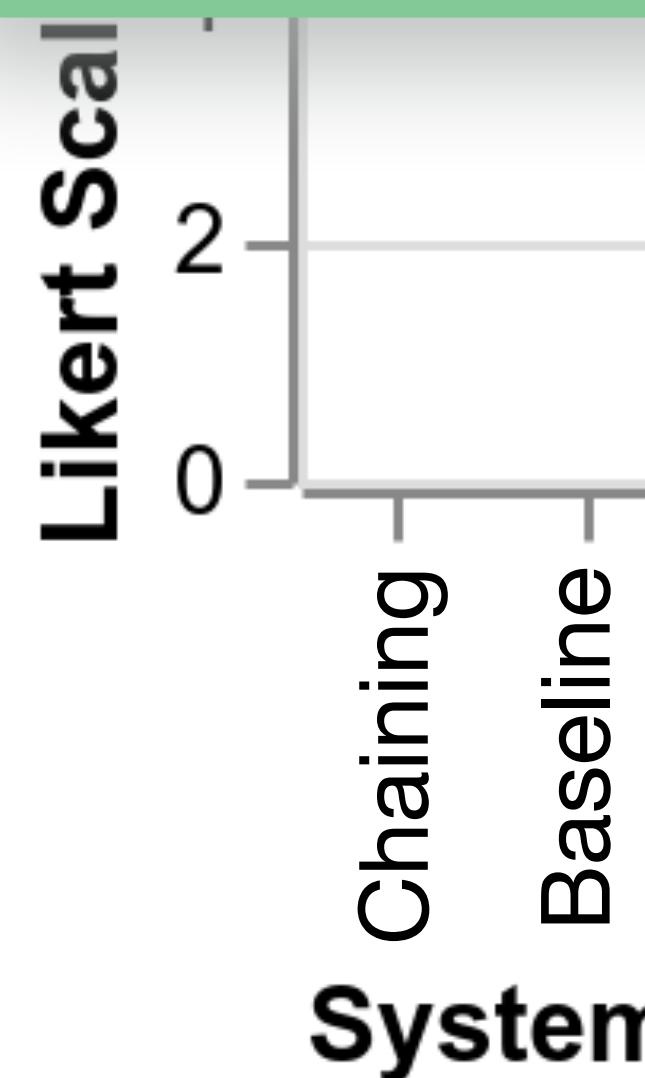


Chaining improves human-AI interactions

20

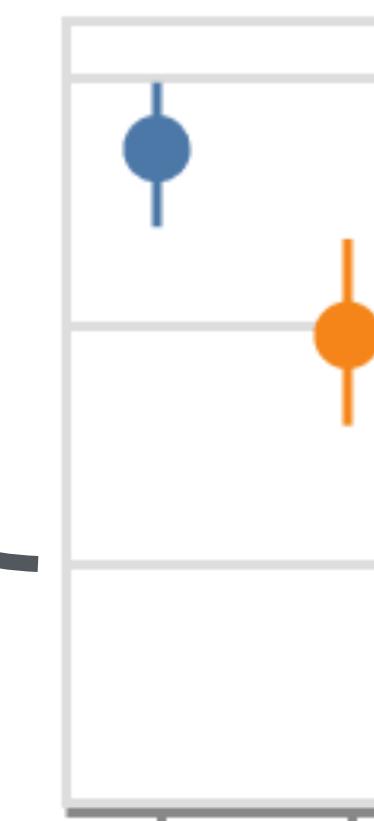
Transparent: better convey model's goal per step

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

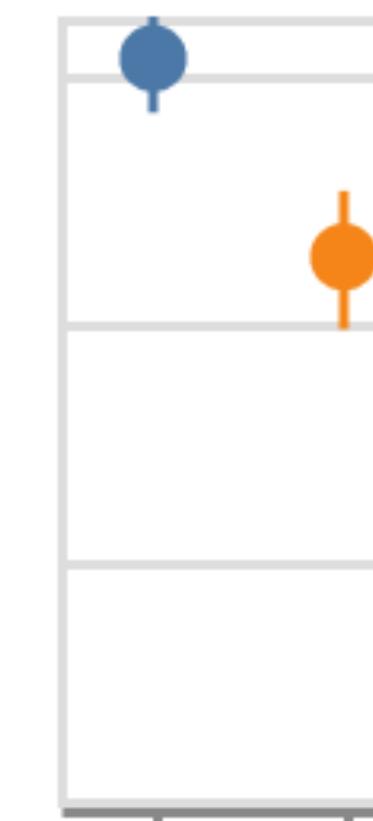


Rating Metrics

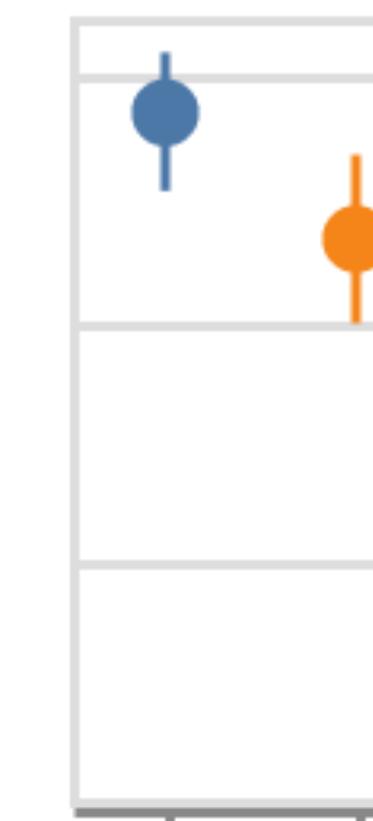
Transparent



Controllable



Collaborative



System

- Chaining
- Baseline

Controllable: Explicit "knobs"

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

Effective task decomposition by capabilities

LLM face challenges,
when we ask for too much (literature).

Operations for LLM Chaining
Nudge users towards scoped inputs/outputs.

Effective task decomposition by capabilities

LLM face challenges,

when we ask for too much (literature):

Generate too much content.

Exposure bias, overfit to their own generations.
e.g., produce redundant content.

City to visit

Paris

Operations for LLM Chaining

Nudge users towards scoped inputs/outputs.

Examples in English

Nice to meet you!

Good morning!

How are you?

I don't speak French.

Examples in French

Enchanté !

Bonjour !

Comment allez-vous ?

Je ne parle pas français

Effective task decomposition by capabilities

LLM face challenges,
when we ask for too much (literature):

Generate too much content.

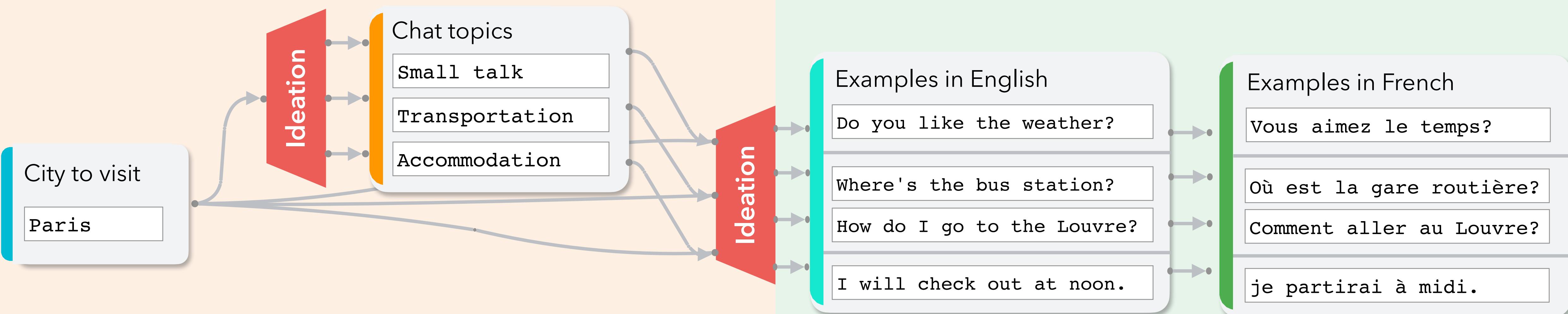
Exposure bias, overfit to their own generations.
e.g., produce redundant content.

Operations for LLM Chaining

Nudge users towards scoped inputs/outputs.

Gather additional information from LLMs

e.g. Ideation



Effective task decomposition by capabilities

LLM face challenges,

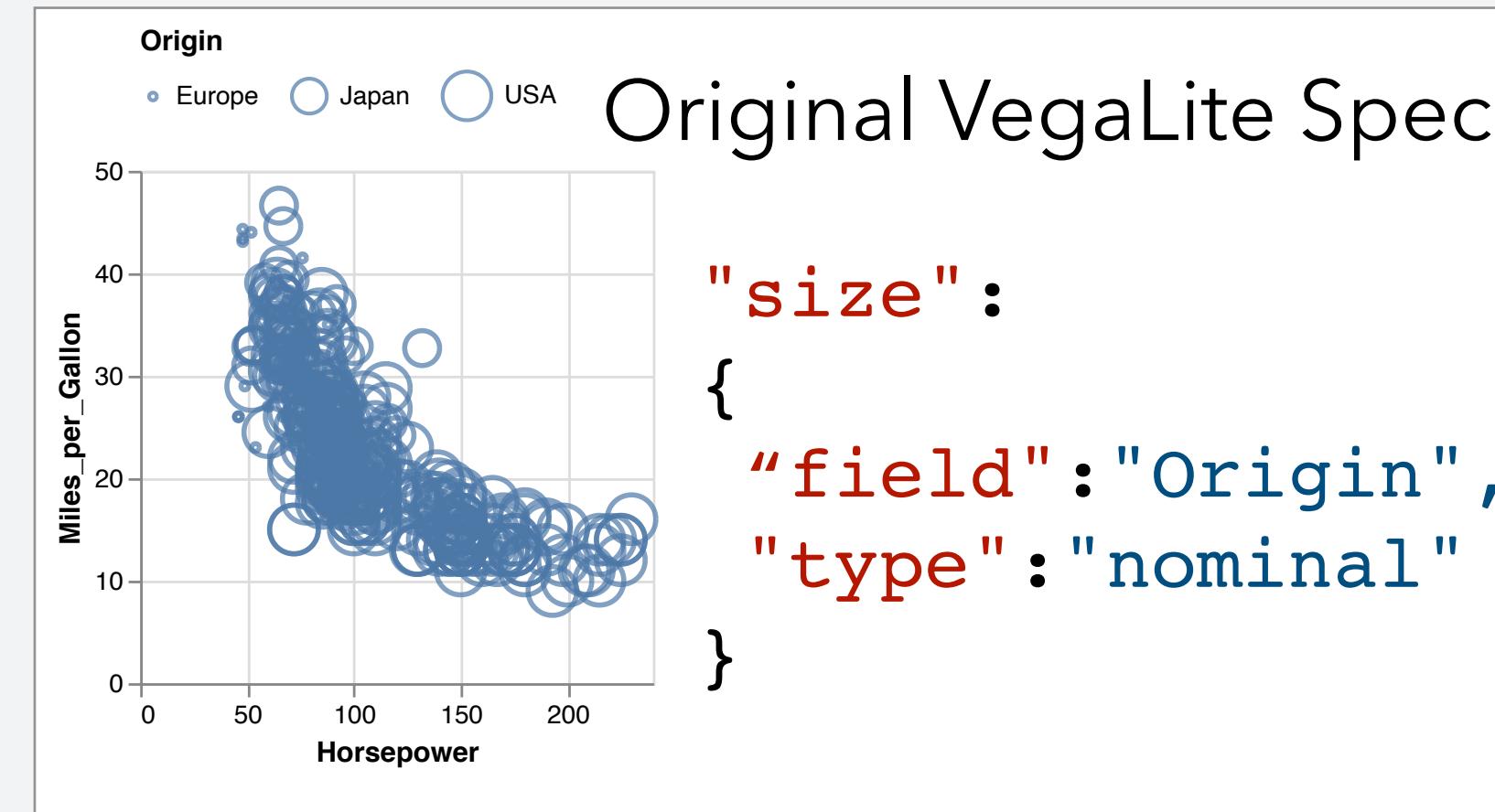
when we ask for too much (literature):

Generate too much content.

Handle too specific data formats.

Confused by mismatched text distribution

e.g., Handled natural text better than JSON format.



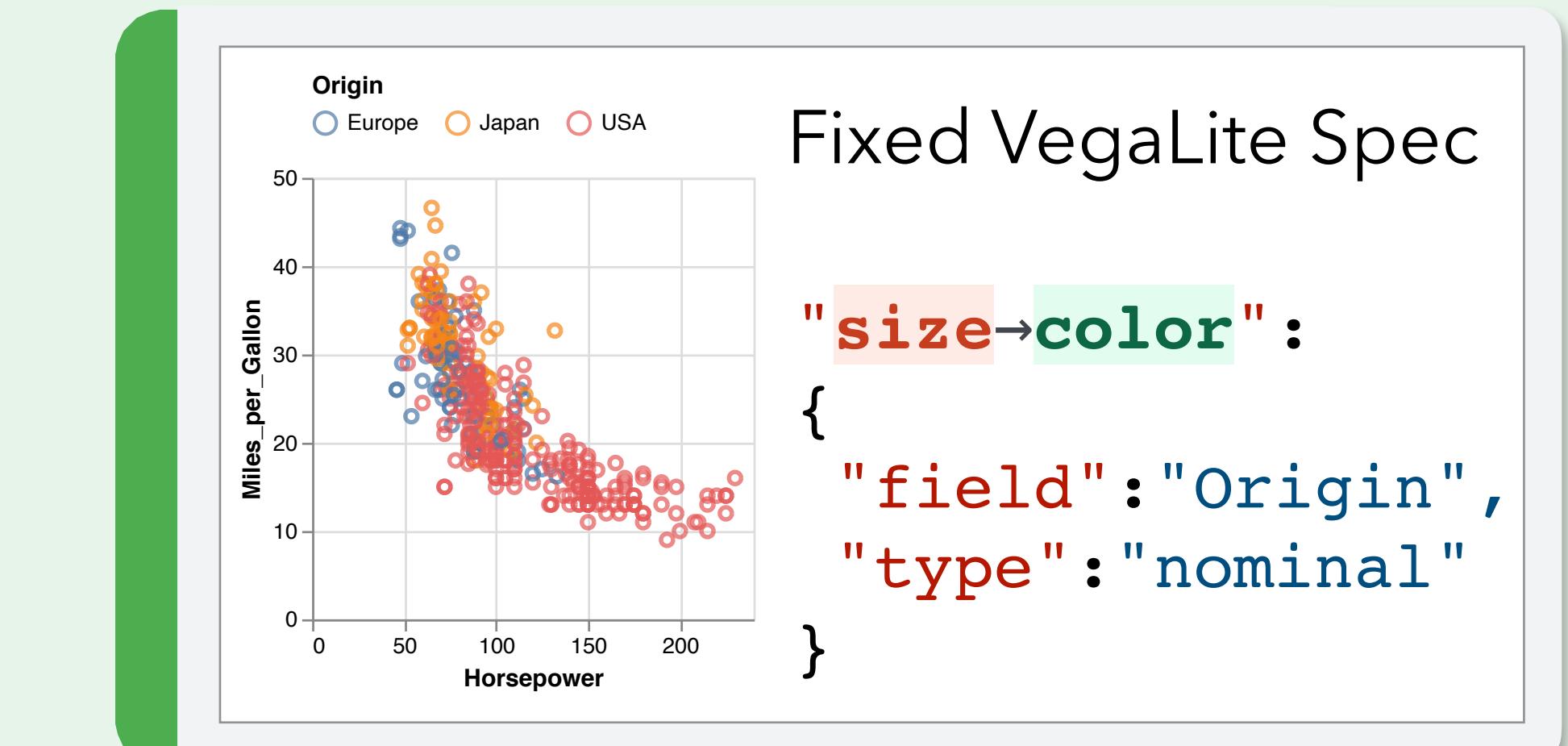
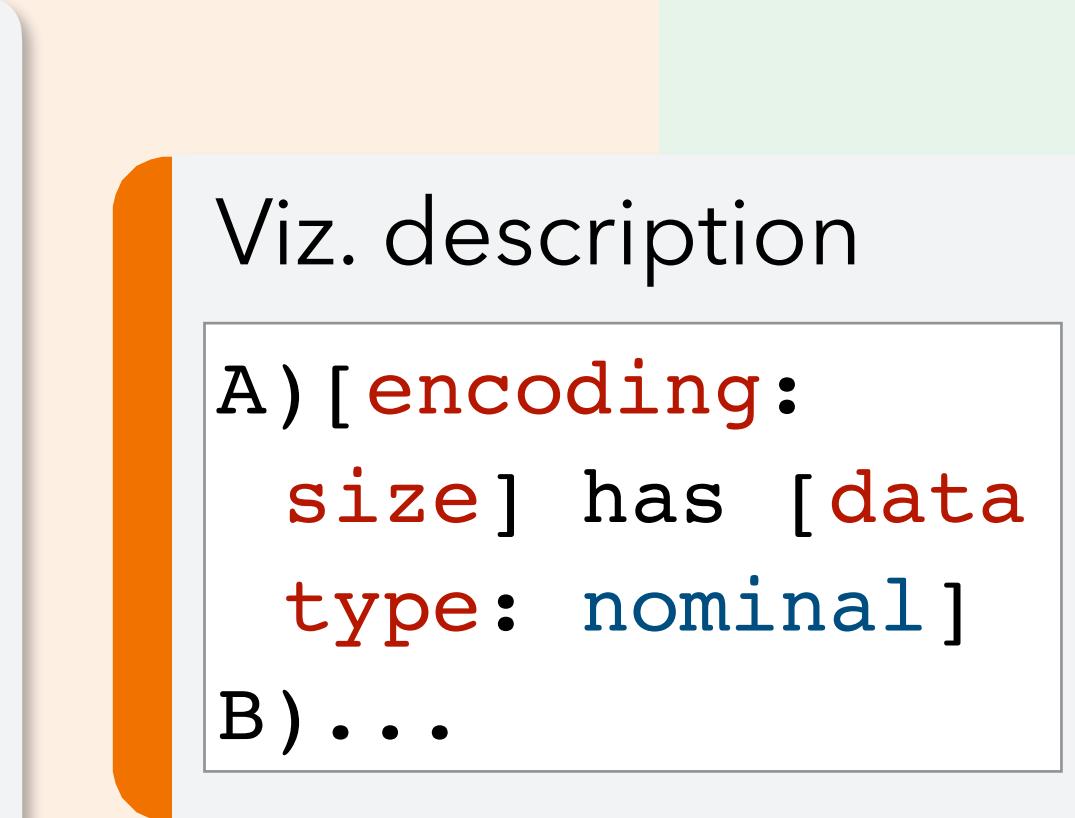
Operations for LLM Chaining

Nudge users towards scoped inputs/outputs.

Gather additional information from LLMs

Re-organize the input

e.g. Rewriting



Rewriting

Rewriting

Effective task decomposition by capabilities

LLM face challenges,

when we ask for too much (literature):

Generate too much content.

Handle too specific data type.

Resolve too many sub-parts.

Not enough reasoning capability.

e.g., if-this-then-that branching logic

Operations for LLM Chaining

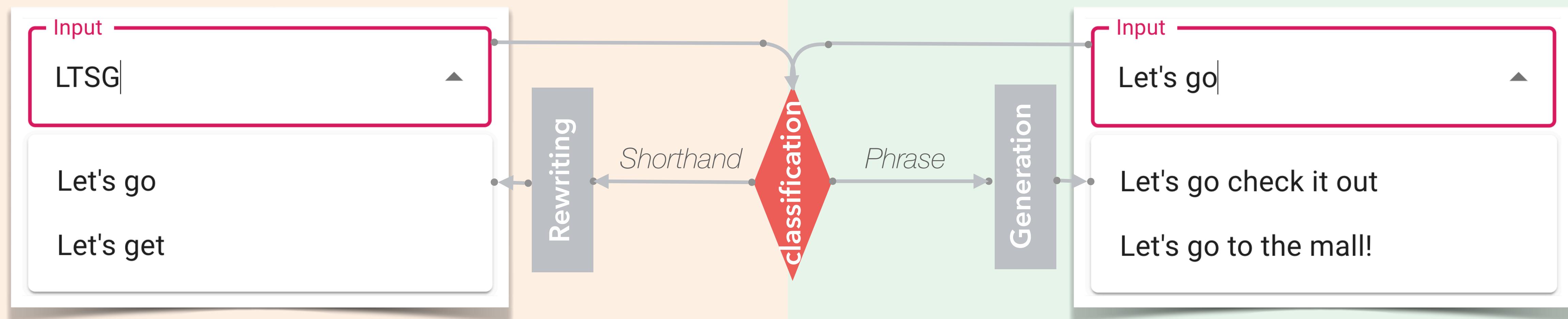
Nudge users towards scoped inputs/outputs.

Gather additional information from LLMs

Re-organize the input

Validate and categorize the input

e.g. Classification



Effective task decomposition by capabilities

LLM face challenges,

when we ask for too much (literature):

Exposure bias.

Sub-optimal data format.

Lack multi-step reasoning capability

Operations for LLM Chaining

Nudge users towards scoped inputs/outputs.

Gather additional information from LLMs

Re-organize the input

Validate and categorize the input

We can use LLMs for multiple roles.

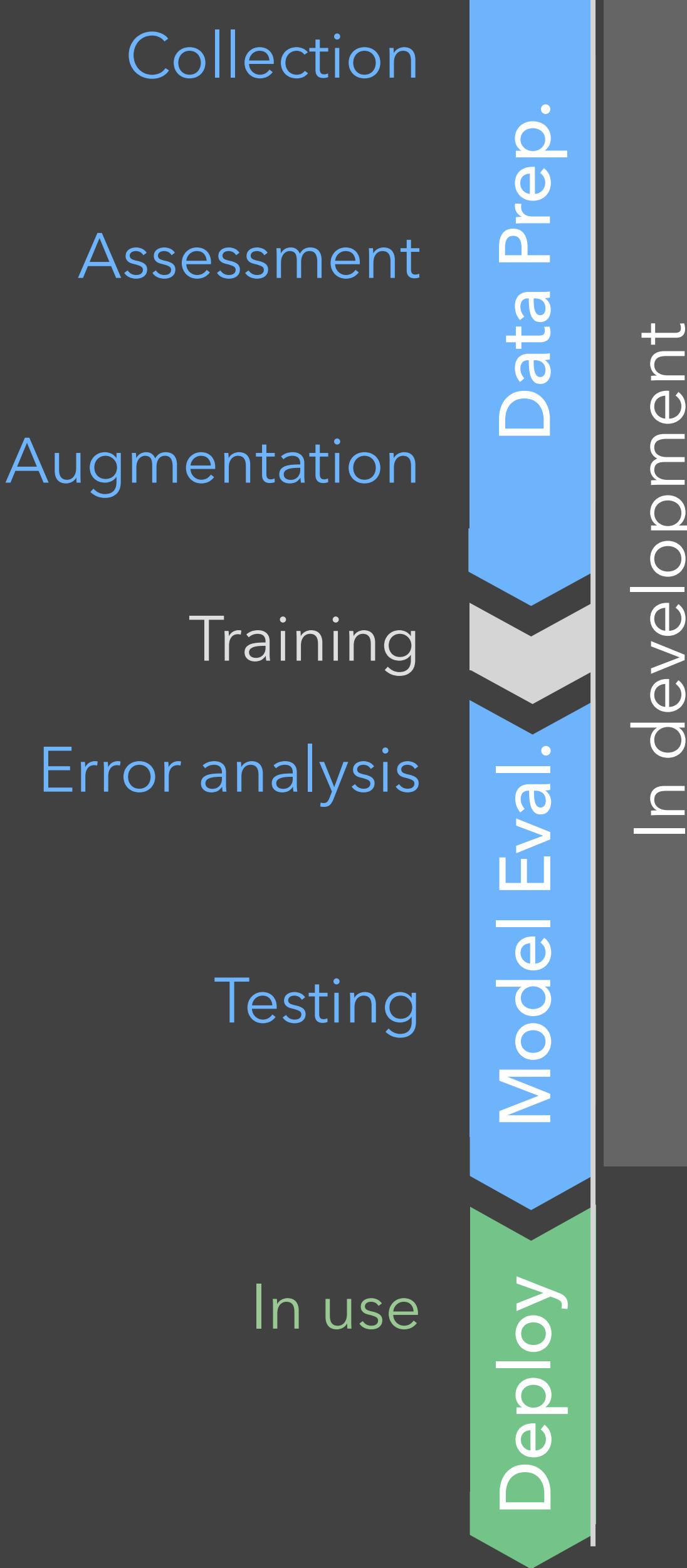
We can maximize team utility if we pick the ones we need
(LLMs do well or we need help with.)

Takeaways

Sometimes AI models are hard to improve directly.

But we can increase the system utility, by **redirecting users to AI strengths**.

Identify & tie the strengths and weaknesses to unleash system potentials.



Debugging is multi-stage.

Mitigate developer bias through assisted methods

Slice discovery that exposes inherent dataset property

Utilize task characteristics

Perturbation that compensates human creativity

Use NLP models as different assistive roles

Help end users to recover from deployed model errors

Help users identify AI errors

Provide explanation and reflection

Help users work around AI errors

Provide users with controls through interactions