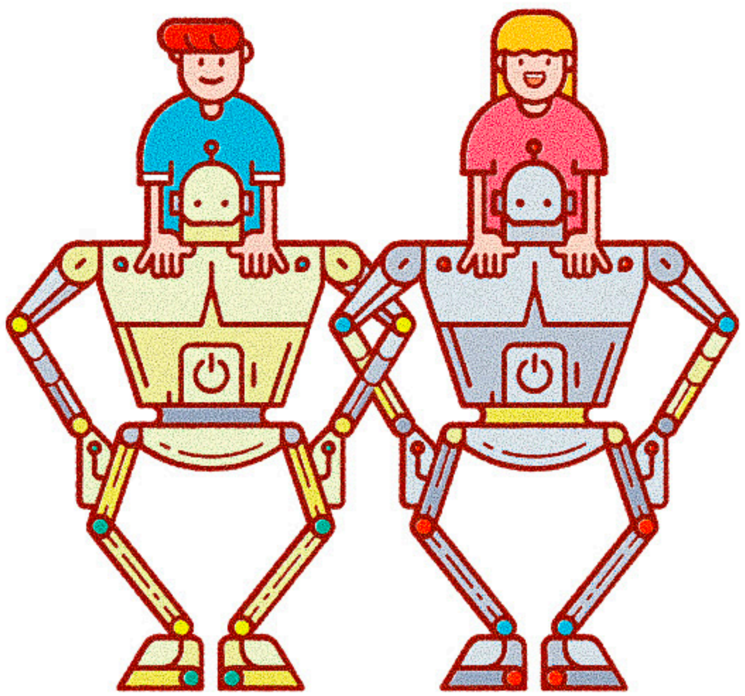
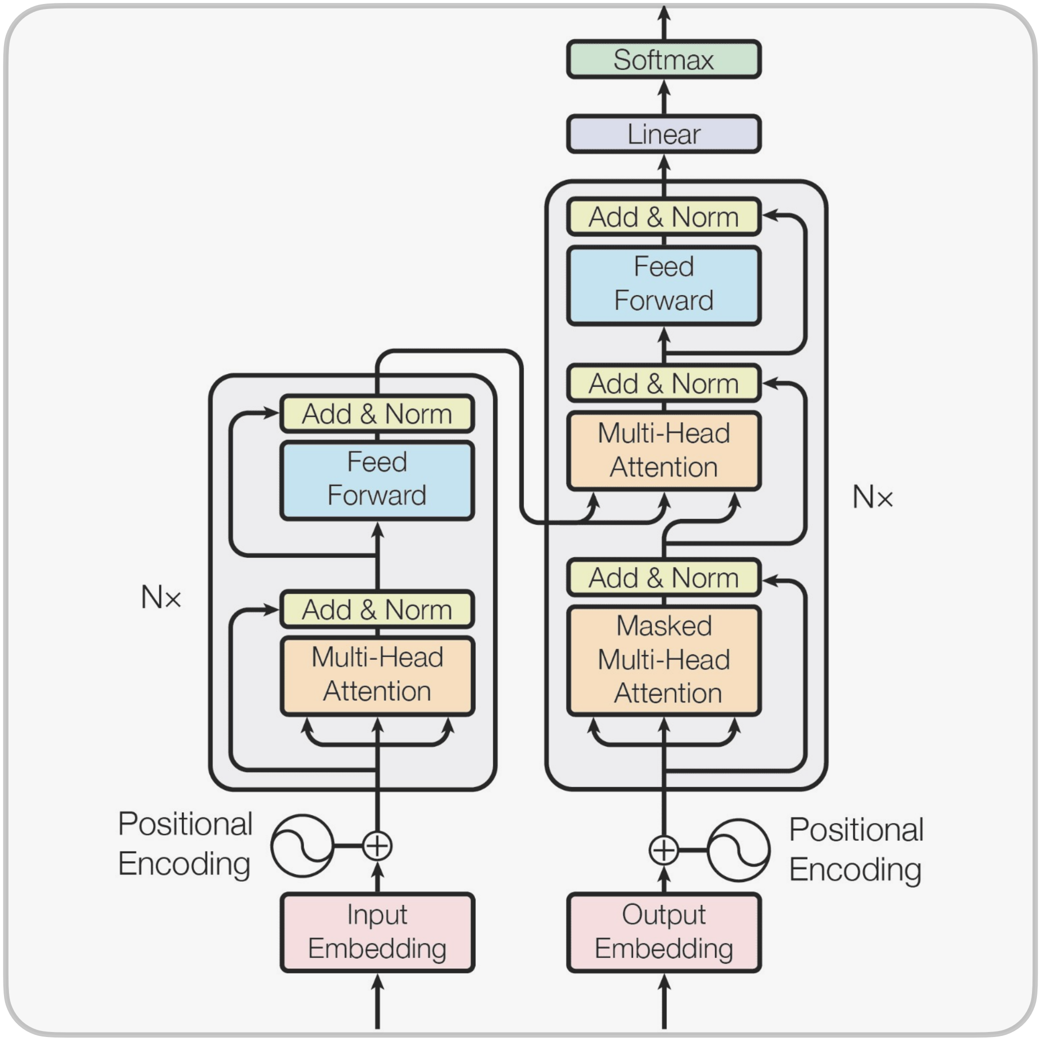
Evaluating AI: From **Crowdsourcing Truth** To **Truth-finding Processes**



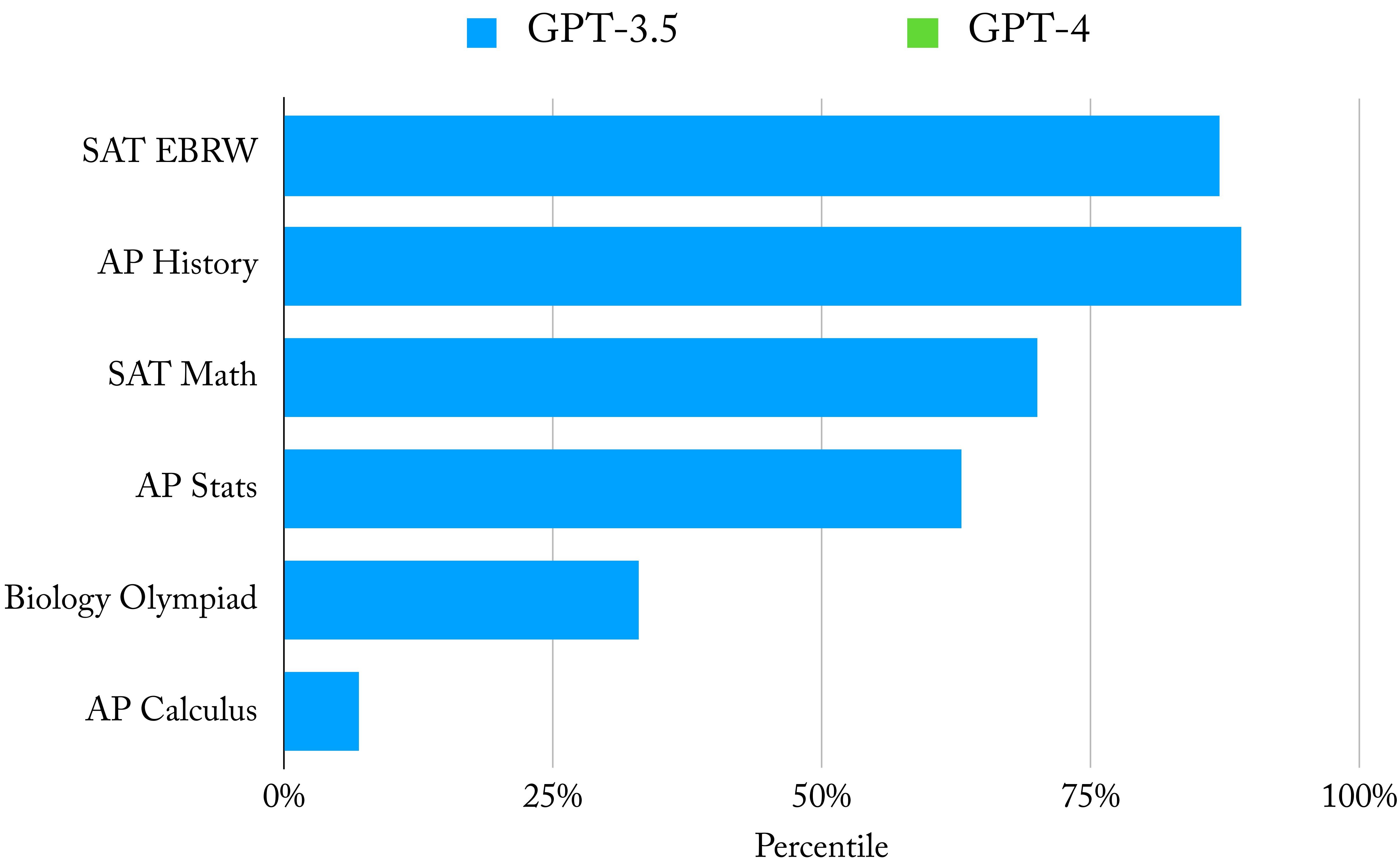
Shi Feng
University of Chicago

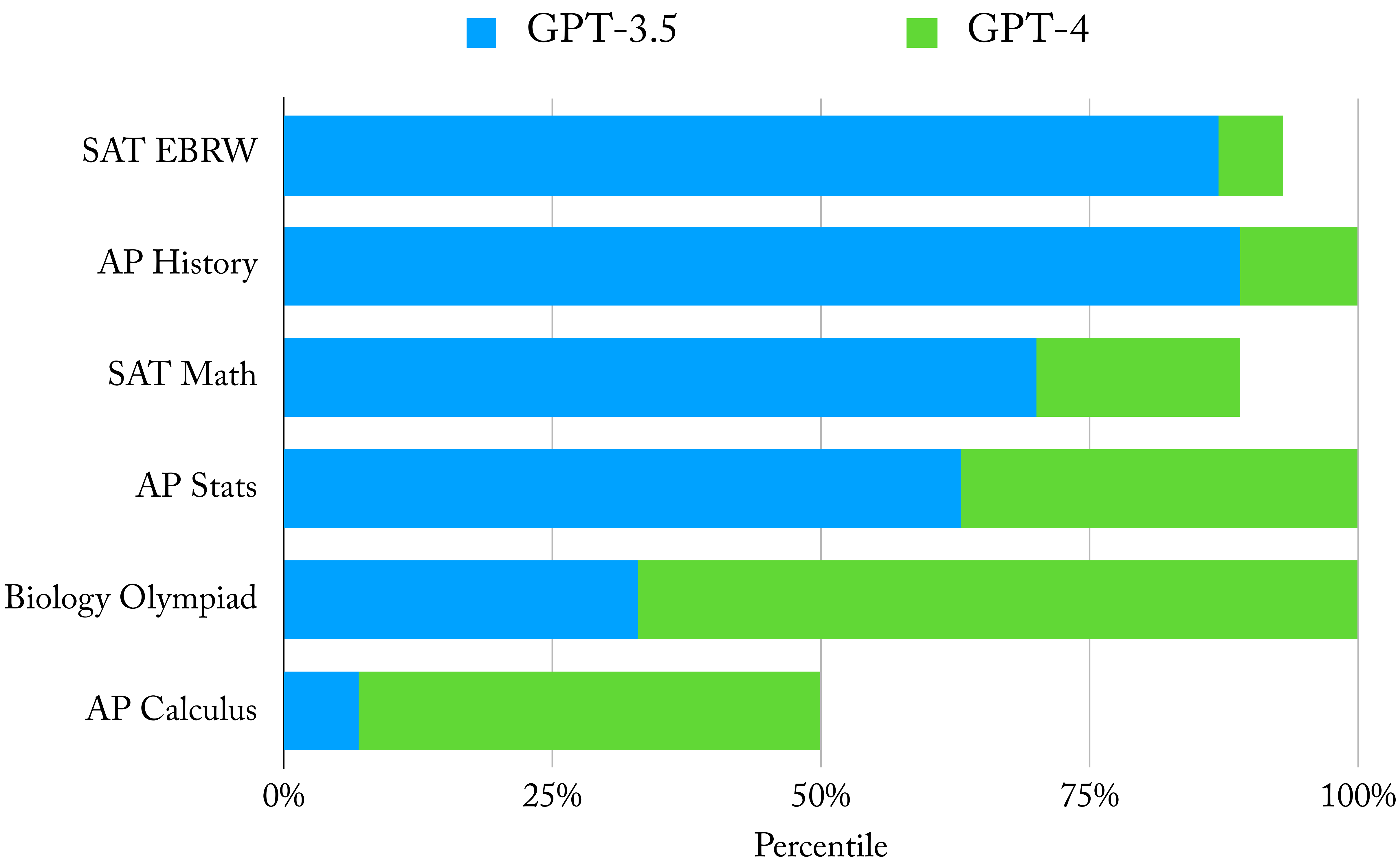
2019


Softmax
Linear
Add & Norm
Feed Forward
Add & Norm
Multi-Head Attention
Add & Norm
Masked Multi-Head Attention
Add & Norm
Feed Forward
Add & Norm
Multi-Head Attention
Nx
Nx
Positional Encoding
Positional Encoding
Input Embedding
Output Embedding

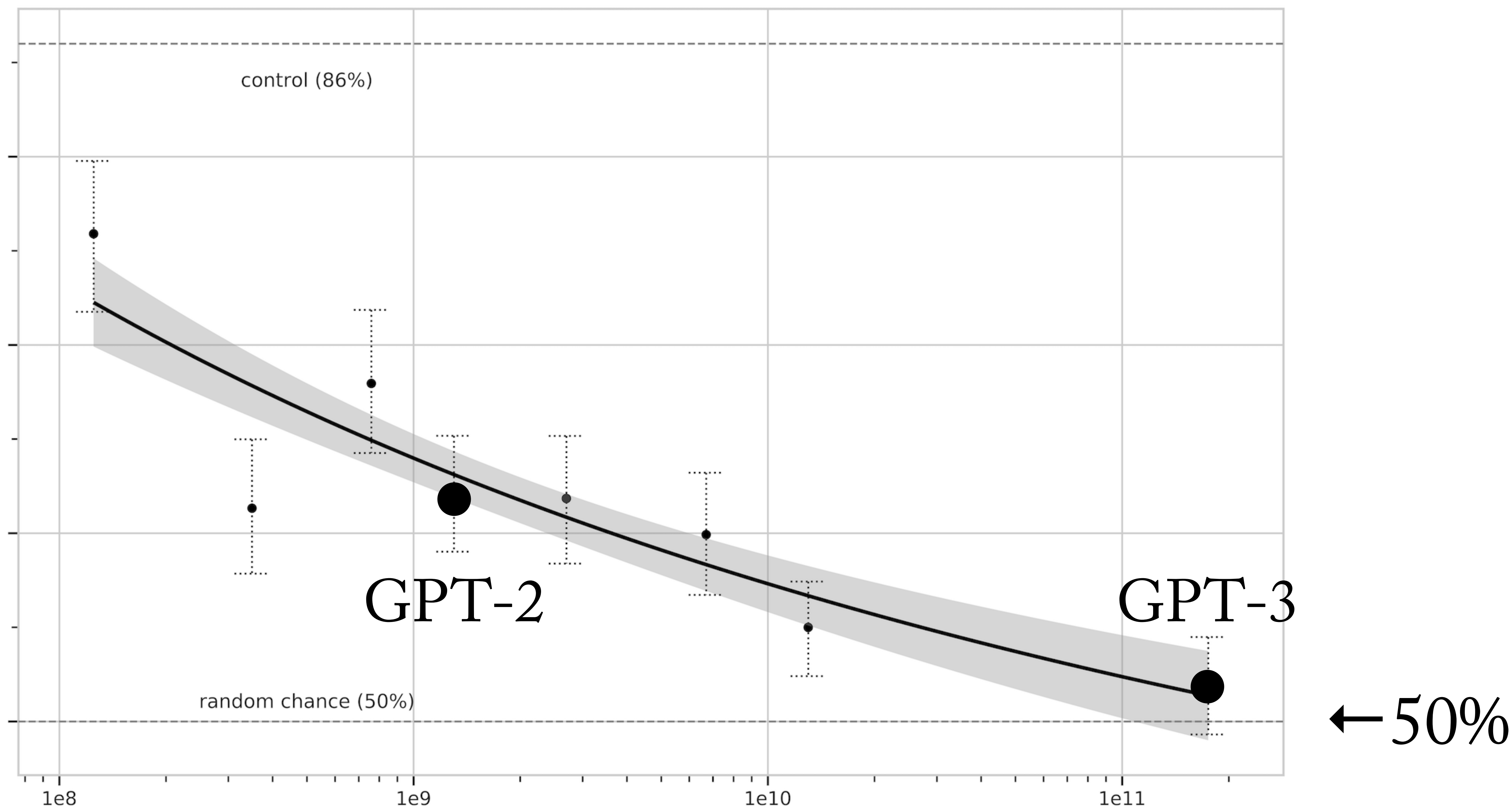
2023

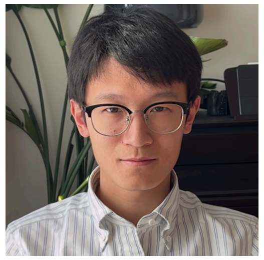

GPT - 4


GPT-3.5
GPT-4
SAT EBRW
AP History
SAT Math
AP Stats
Biology Olympiad
AP Calculus
0%
25%
50%
75%
100%
Percentile


GPT-3.5
GPT-4
SAT EBRW
AP History
SAT Math
AP Stats
Biology Olympiad
AP Calculus
0%
25%
50%
75%
100%
Percentile

Human detection of AI-generated article





What's the paper that talked about the risks of foundation models?

The paper you are referring to is "On the Opportunities and Risks of Foundation Models" by researchers at OpenAI.

This paper discusses the challenges, risks, and opportunities associated with the development and deployment of large-scale AI models, such as GPT-3.

Is the AI's answer *true*?

The paper you are referring to is "On the Opportunities and Risks of Foundation Models" by researchers at OpenAI.

This paper discusses the challenges, risks, and opportunities associated with the development and deployment of large-scale AI models, such as GPT-3.

Is the AI's answer *true*?

The paper you are referring to is "On the Opportunities and Risks of Foundation Models" by researchers at OpenAI.

On the Opportunities and Risks of Foundation Models

Center for Research on Foundation Models

HAI

Stanford University Human-Centered Artificial Intelligence

Human evaluation defines truth for AI

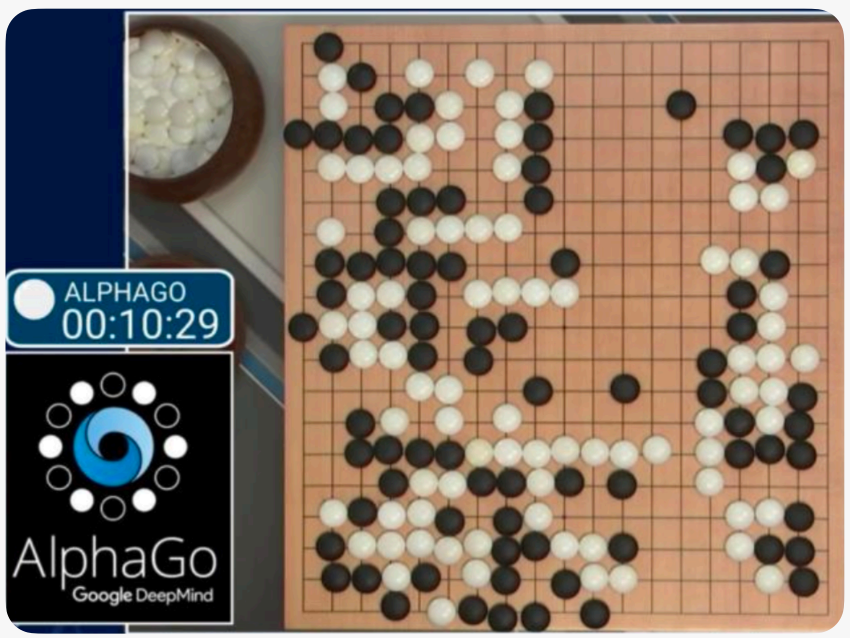
- What color is the flower?
- Yellow

Crowdsourced truth

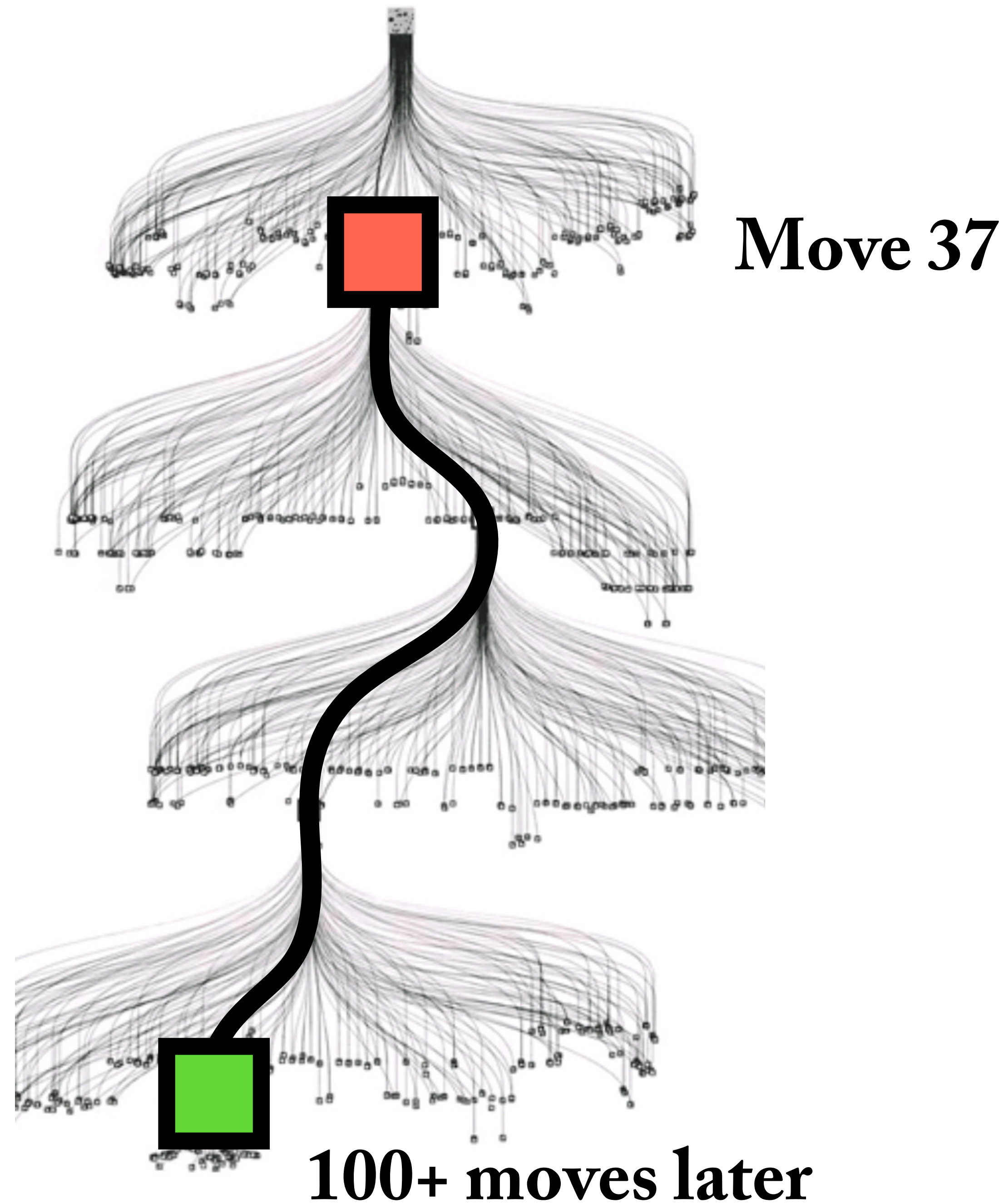
???

Human evaluation defines truth for AI

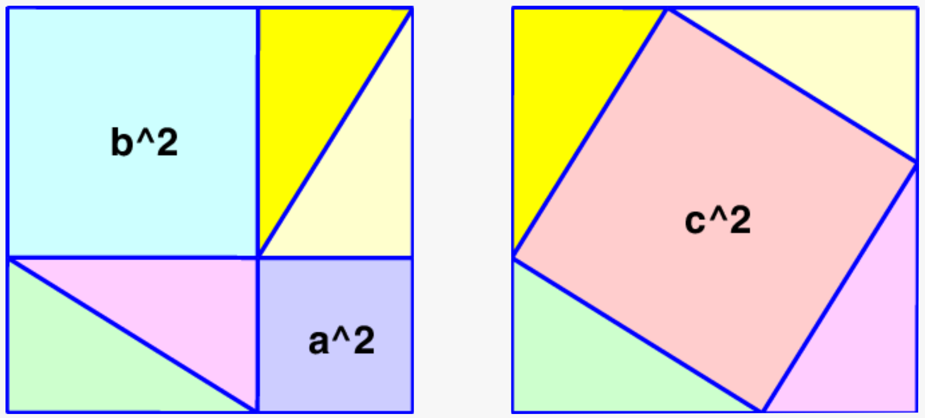
Move 37



???



Explanation informs human decision

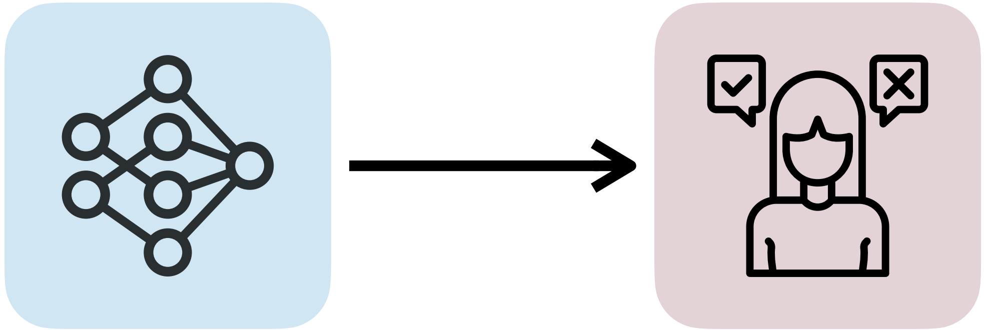


$$a^2 + b^2 = c^2$$

Formal proofs

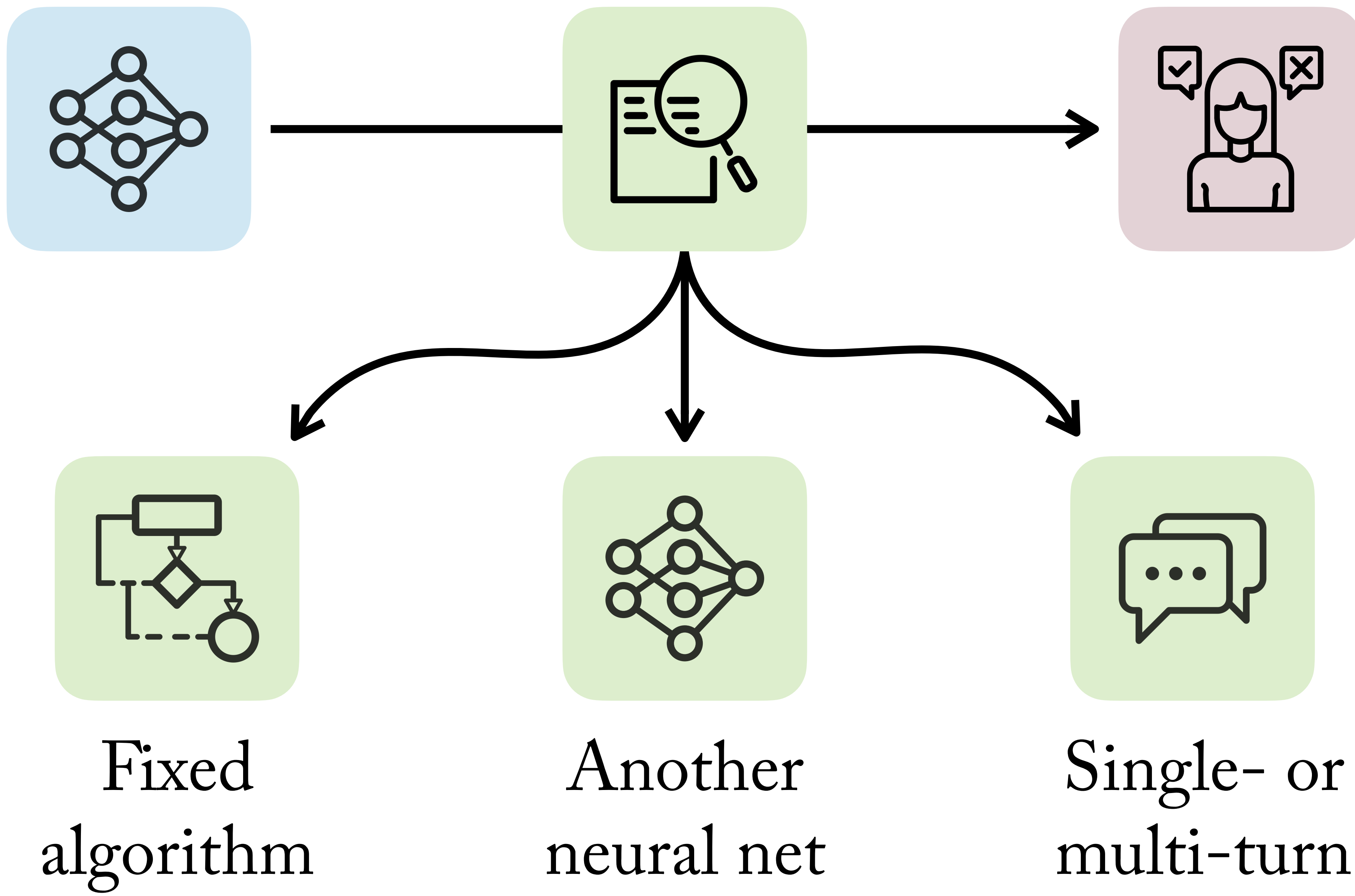
Legal arguments

Explanation as a truth-finding process

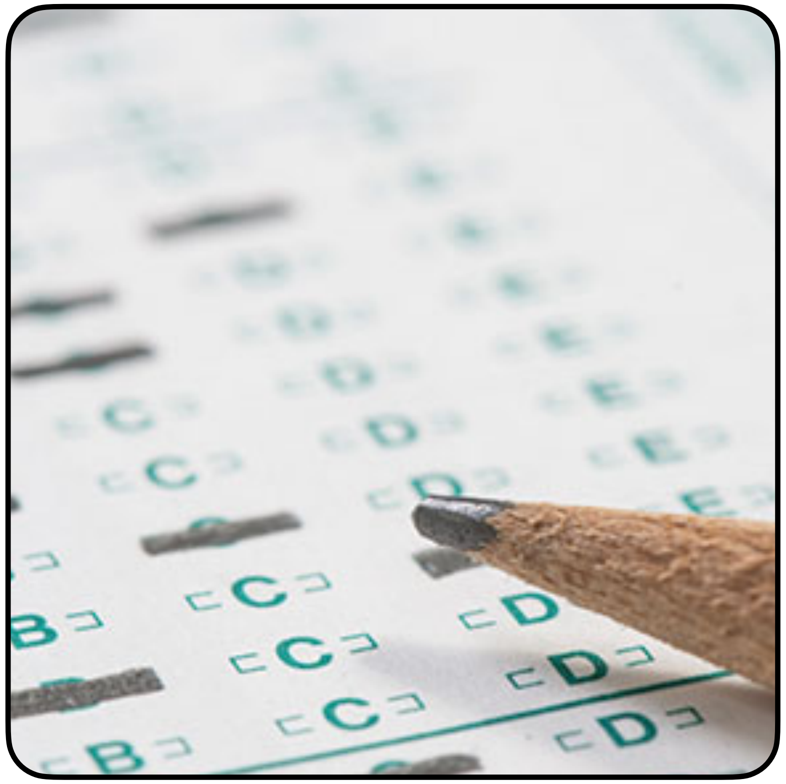


Explanation as a truth-finding process

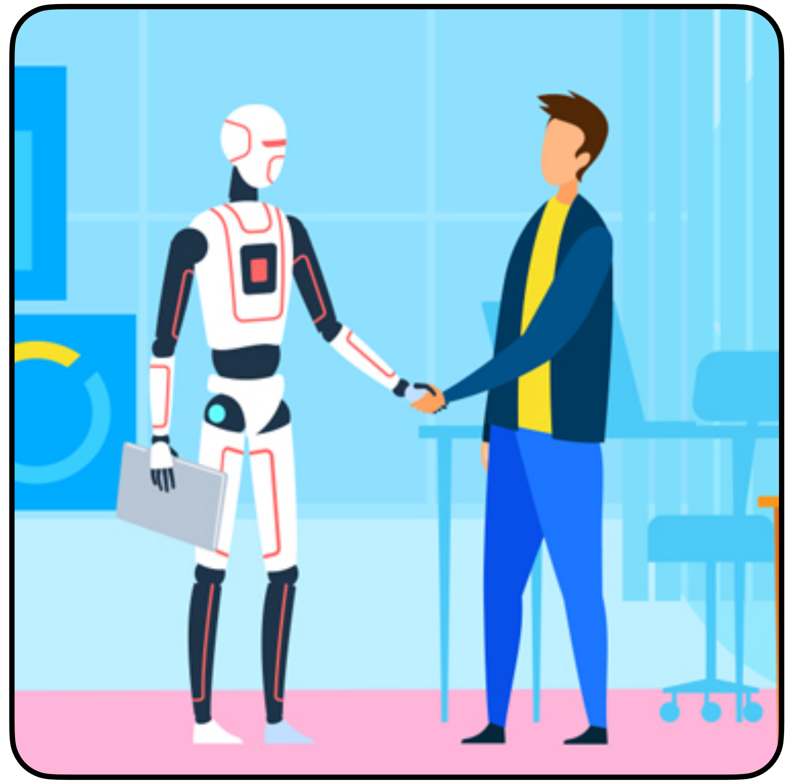
A process to gather additional information to support human evaluation of AI outputs.



Training AI to explain itself



Capability assessment



Human-AI collaboration

Training future AIs

Training AI to explain itself

1. Can AI explain by *mimicking* human?

EMNLP 18, 19, 22 | ACL 19 | NAACL 21 | IUI 19

2. How can AI *learn* to explain better?

TACL 19 | EMNLP 22 | ICML 19, 21 | ICLR 23 | TMLR 23

NLP **ML** **HCI**

Training AI to explain itself

1. Can AI explain by *mimicking* human?

EMNLP 18, 19, 22 | ACL 19 | NAACL 21 | **IUI 19**

2. How can AI *learn* to explain better?

TACL 19 | **EMNLP 22** | ICML 19, 21 | ICLR 23 | TMLR 23

NLP | **ML** | **HCI**

Training AI to explain itself

1. Can AI explain by *mimicking* human?

EMNLP 18, 19, 22 | ACL 19 | NAACL 21 | **IUI 19**

How do humans explain? By identifying difference makers

Q1: What color is the flower ?
A1: Yellow

How do humans explain?
By identifying difference makers

Q1: What color is the flower ?
A1: Yellow

How do humans explain?
By identifying difference makers

Q1: What color is the flower ?

A1: Yellow

How do humans explain?
By identifying difference makers

Q1: What color is the flower ?
A1: Yellow

Q2: What color is the ?
A2: Yellow / black / green / white

How do humans explain?
By identifying difference makers

Q1: What color is the flower ?
A1: Yellow

Q2: What color is the ?
A2: Yellow / black / green / white

Q3: What color is flower ?
A3: Yellow

Difference makers lead to large delta

Importance := delta in AI output

What color is the flower ? Yellow (0.827)
color is the flower ? Yellow (0.715)

Importance := delta in AI output

What color is the flower ? Yellow (0.827)
color is the flower ? Yellow (0.715)
What is the flower ? Yellow (0.530)

Importance := delta in AI output

What color is the flower ? Yellow (0.827)
color is the flower ? Yellow (0.715)
What is the flower ? Yellow (0.530)
What color the flower ? Yellow (0.820)

Importance := delta in AI output

What color is the flower ? Yellow (0.827)
color is the flower ? Yellow (0.715)
What is the flower ? Yellow (0.530)
What color the flower ? Yellow (0.820)
What color is flower ? Yellow (0.826)
What color is the ? Yellow (0.700)

Importance := delta in AI output
Seems to capture necessity

What color is the flower ? Yellow (0.827)
color is the flower ? Yellow (0.715)
What is the flower ? Yellow (0.530)
What color the flower ? Yellow (0.820)
What color is flower ? Yellow (0.826)
What color is the ? Yellow (0.700)

What color is the flower ?

Importance := delta in AI output
How about sufficiency?

What	color	is	the	flower	?	Yellow 0.827
What	color	is		flower	?	Yellow 0.827
What	color			flower	?	Yellow 0.825
	color			flower	?	Yellow 0.702
				flower	?	Yellow 0.819

- Unjustifiable confidence
- Inconsistent

Importance := delta in AI output
How about sufficiency?

What	color	is	the	flower	?	Yellow	0.827
What	color	is		flower	?	Yellow	0.827
What	color			flower	?	Yellow	0.825
	color			flower	?	Yellow	0.702
				flower	?	Yellow	0.819

- Unjustifiable confidence
- Inconsistent

Seems odd. Does it generalize?

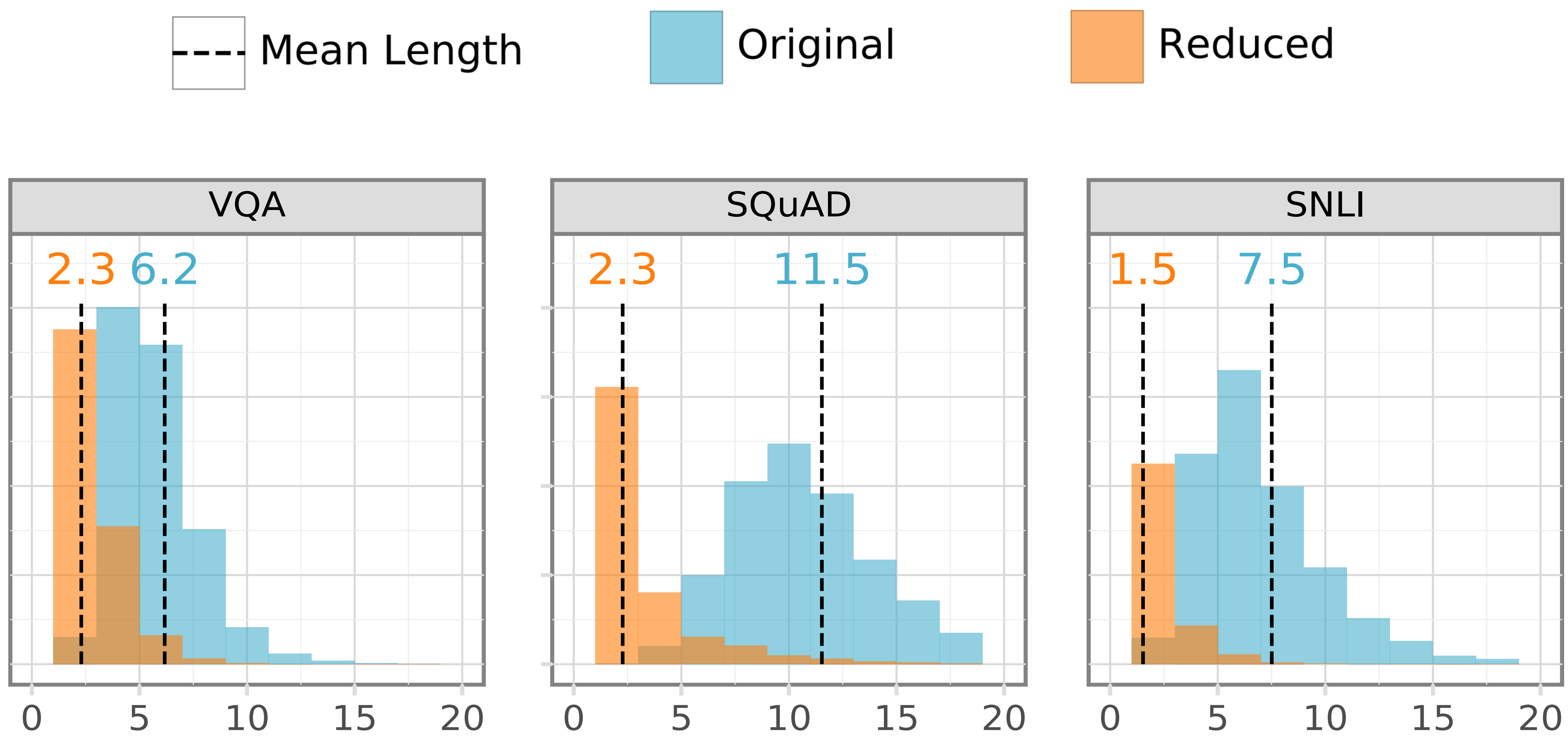
SQuAD

Context	In 1899, John Jacob Astor IV invested $100,000 for Tesla to further develop and produce a new lighting system. Instead, Tesla used the money to fund his Colorado Springs experiments.
Original	What did Tesla spend Astor's money on ?
Reduced	**did**
Confidence	$0.78 \rightarrow 0.91$

SNLI

Premise	Well dressed man and woman dancing in the street
Original	Two man is dancing on the street
Answer	Contradiction
Reduced	**dancing**
Confidence	$0.977 \rightarrow 0.706$

Pathological high confidence on uninformative inputs



Generalizes across:

Many more QA and RC tasks

ElMo, BERT, GPT

LIME, Gradient, IntGrad

Removing unimportant feature leads to big delta in importance

What color is the flower ?	Yellow 0.827
What color is flower ?	Yellow 0.827
What color flower ?	Yellow 0.825
color flower ?	Yellow 0.702
flower ?	Yellow 0.819

What color is the flower?

Model says **Plausible**

Model says **Implausible**

What color is the flower?

What color is flower?

Model says **Plausible**

Model says **Implausible**

What color is the flower?

Model says **Plausible**

What color is flower?

Model says **Implausible**

What color is the flower?

What is the flower?

Model says **Plausible**

What color is flower?

Model says **Implausible**

What color is the flower?

Model says **Plausible**

What color is flower?

Model says **Implausible**

What is the flower?

What color is the flower?

flower?

Model says **Plausible**

What color is flower?

Model says **Implausible**

What is the flower?

What color is the flower?

Model says Plausible

What color is flower?

flower?

Model says Implausible

What is the flower?

What color is the flower?

color flower?

Model says **Plausible**

What color is flower?

flower?

Model says **Implausible**

What is the flower?

What color is the flower?

Model says **Plausible**

What color is flower?

flower?

Model says **Implausible**

What is the flower?

color flower?

What did we learn?

1. If models have these pathologies, we cannot expect reasonable explanations with this method.

2. It's expected that models have these issues. We argue that the intuitive way to extract explanations doesn't work with these models.

3. Reduced example is a caricature. Generalization to OOD is always hard.

4. It is indeed partly an issue of post-hoc method.

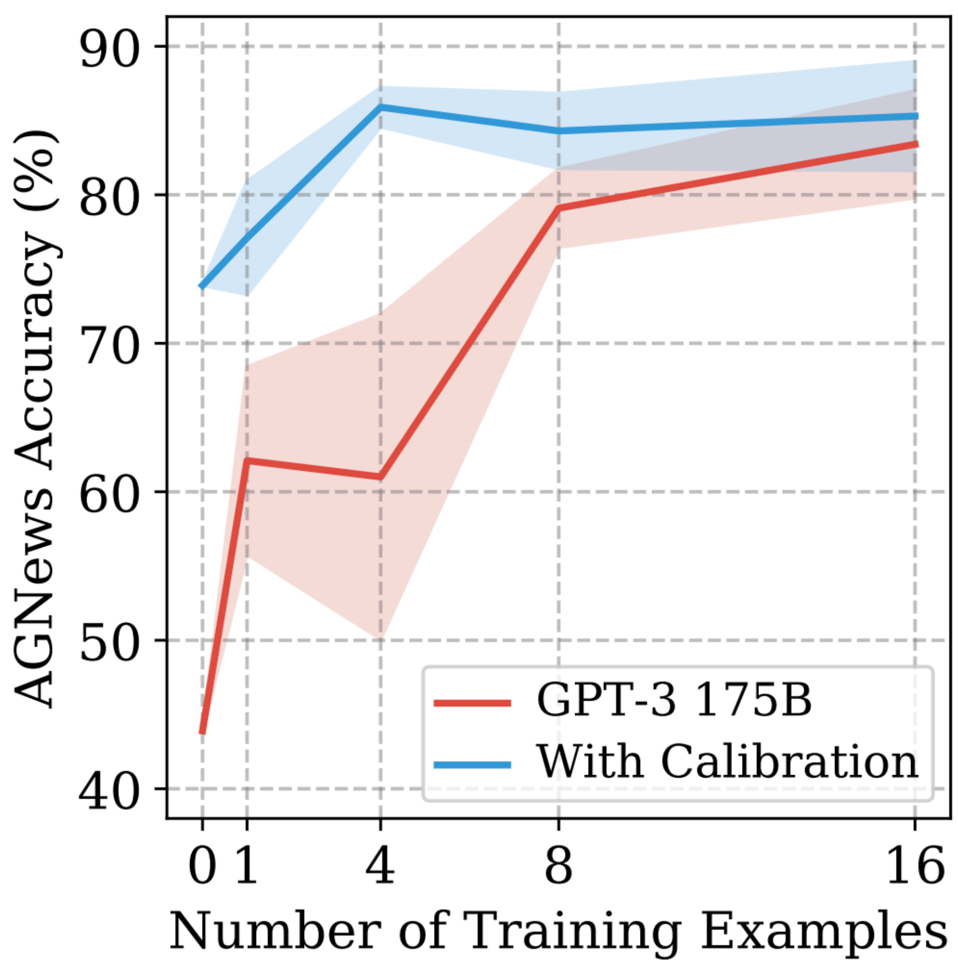
What did we learn?

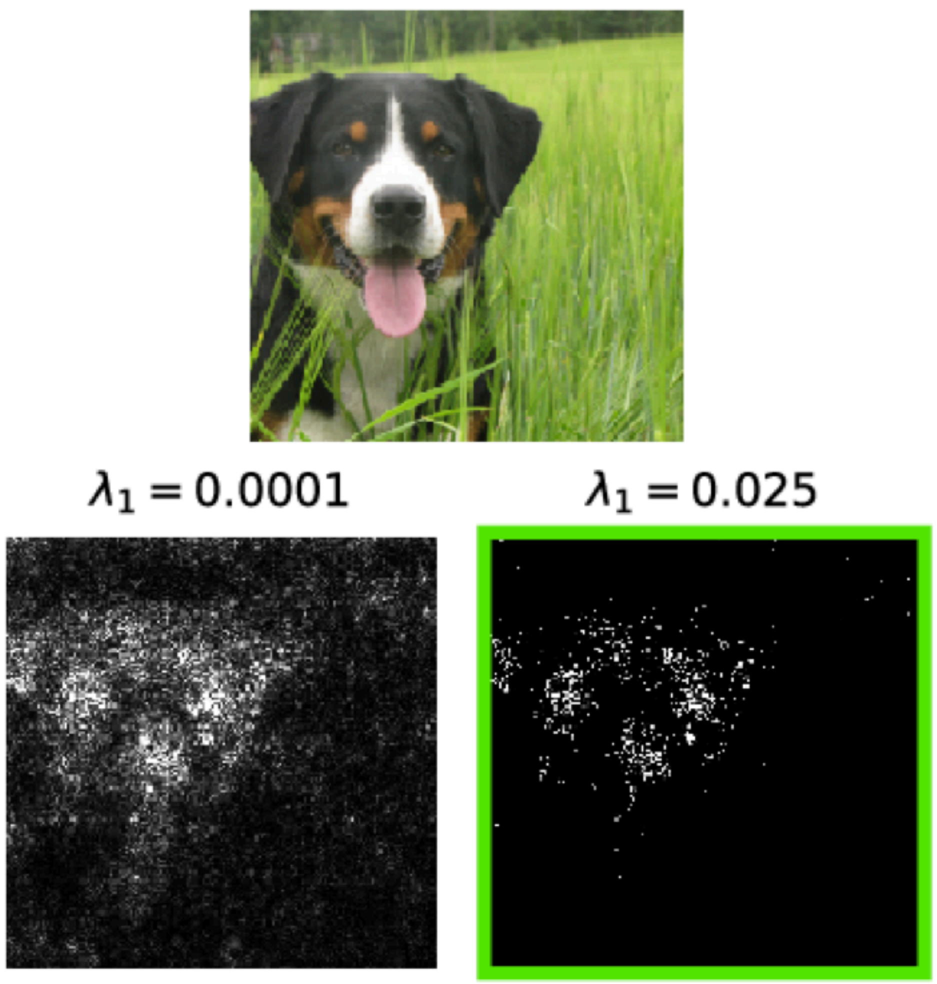
Our *intuitive* notion of importance has *complex* mathematical implications—properties that humans might satisfy but AIs might not.

What did we learn?

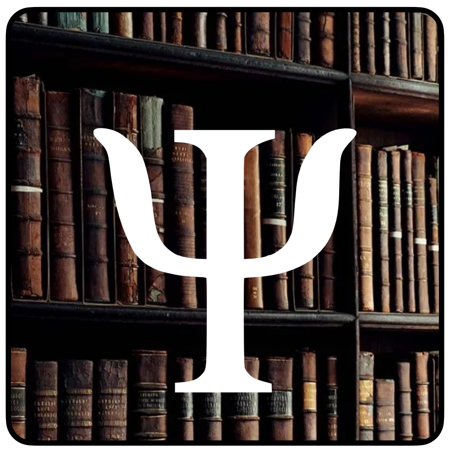
1. Pathological high confidence EMNLP 18 ICML 21

2. Poor consistency across counterfactuals ICML 19

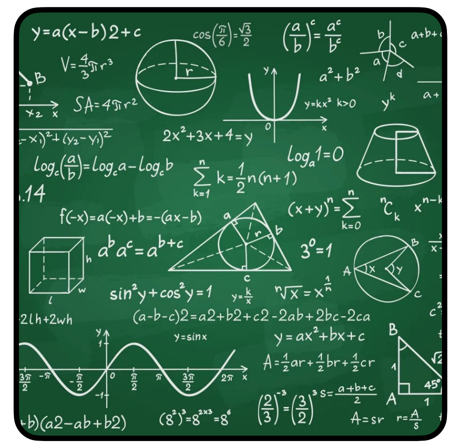




What's next?



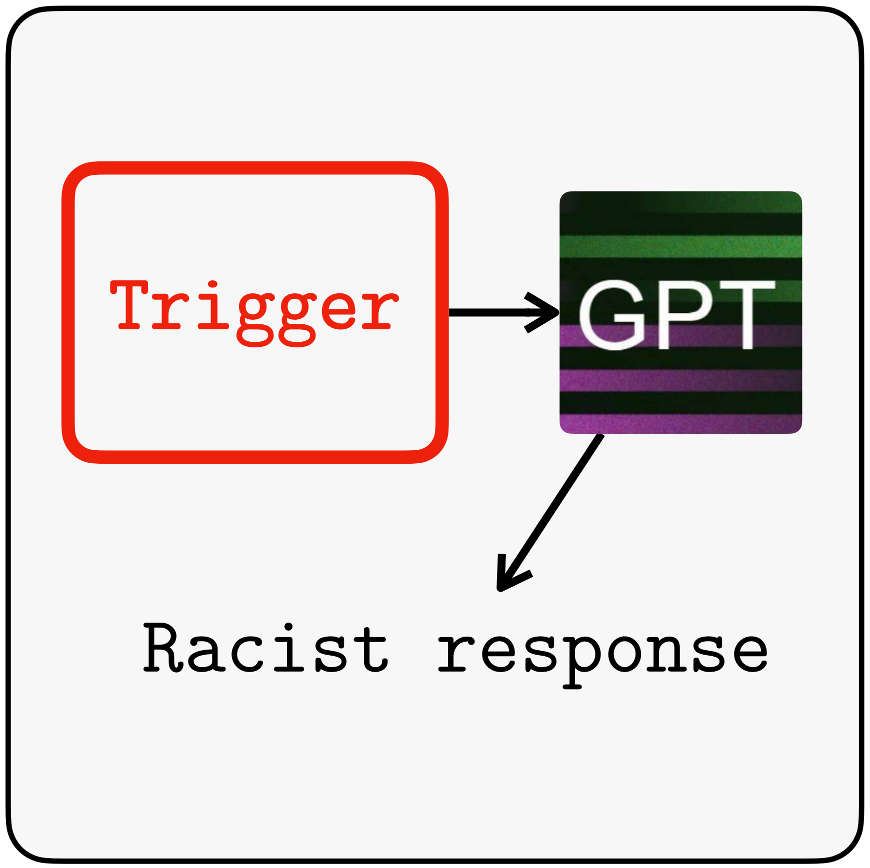
1. Psychological expectation



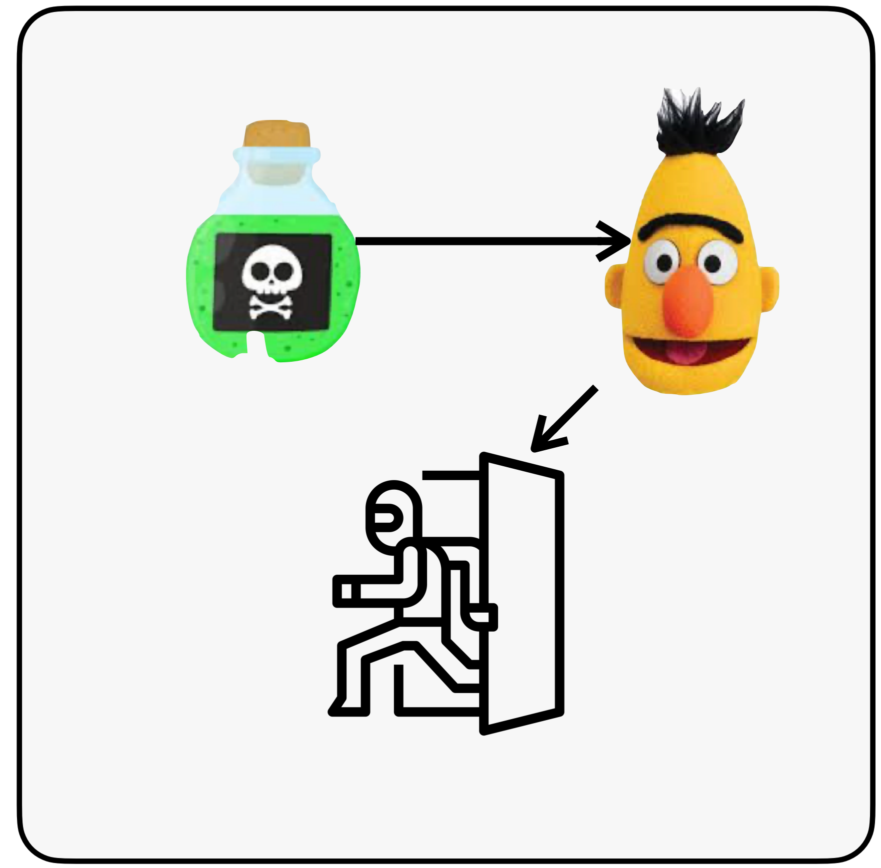
2. Mathematical formulation

3. Validate AI & design solutions

What's next?

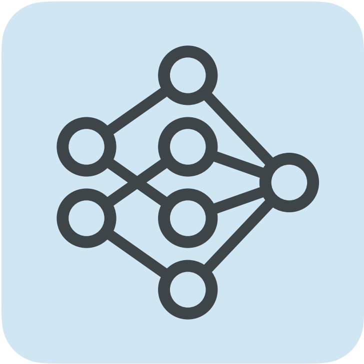


EMNLP 19



NAACL 21

What's next?



v.s.

- Humans cannot explain AI yet.
- AI explaining itself requires non-trivial extrapolation beyond human capability.

How can AIs learn to explain better?

How can the applicant improve?

Plausible

Implausible

How can the applicant improve?

Education

Plausible

Implausible

How can the applicant improve?

Plausible

Education

Implausible

How can the applicant improve?

Gender, race

Plausible

Education

Implausible

How can the applicant improve?

Plausible

Education

Implausible

Gender, race

How can the applicant improve?

Experience, Role

Plausible

Education

Implausible

Gender, race

How can the applicant improve?

Plausible

Education

Experience, Role

Implausible

Gender, race

How can the applicant improve?

Country of origin

Plausible

Education

Experience, Role

Implausible

Gender, race

How can the applicant improve?

Plausible

Education
Experience, Role

Implausible

Gender, race
Country of origin

How can the applicant improve?

> 25 years old

Be two years younger

Plausible

Implausible

How can the applicant improve?

Plausible

> 25 years old

Implausible

Be two years younger

How can the applicant improve?

Get a masters degree

Plausible

Implausible

How can the applicant improve?

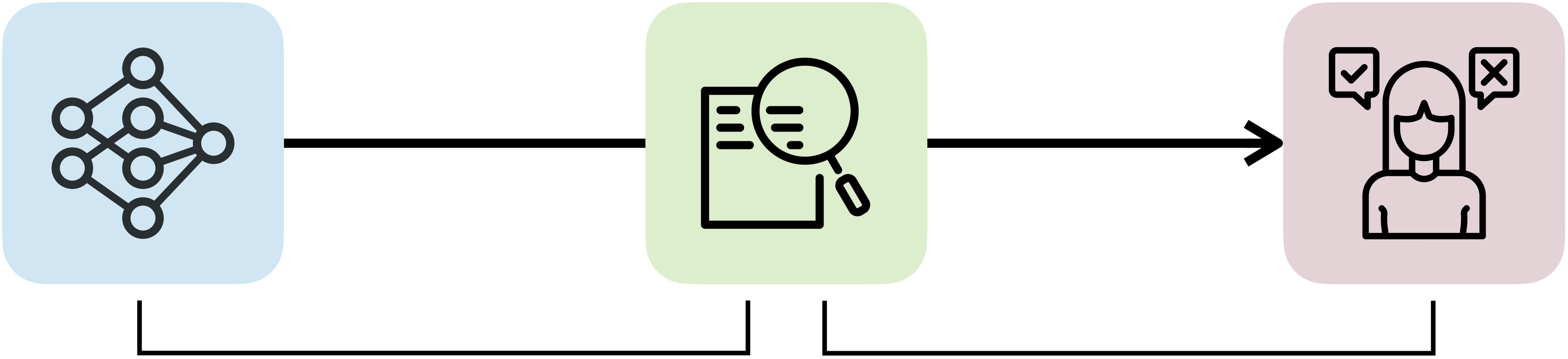
Plausible

Get a masters degree
Currently: bachelor

Implausible

Get a masters degree
Currently: high-school

Learning to explain better



Explanation
Algorithmic

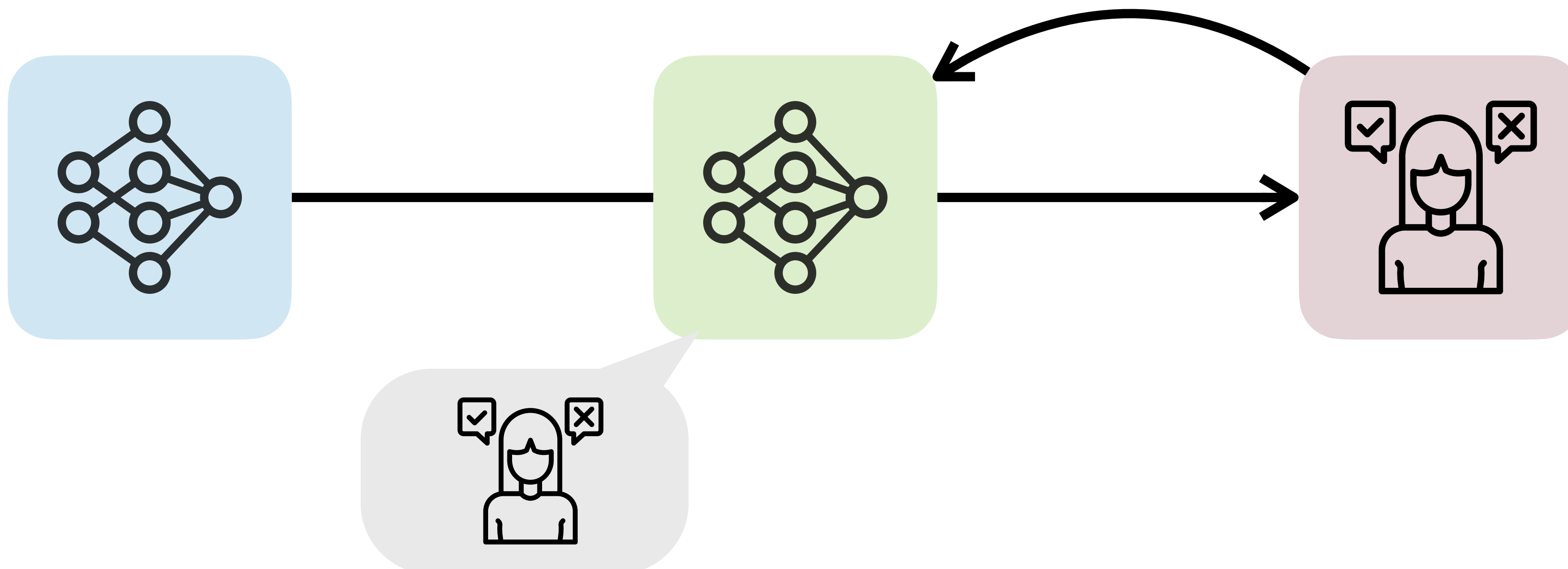
Interpretation
Neurobiological

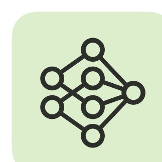
Explanation is highly contextual.
Full context isn't available.

The plan:

1. Model the **interpretation** process.
2. Learn from **feedback**, not demonstration.

Learning to explain better

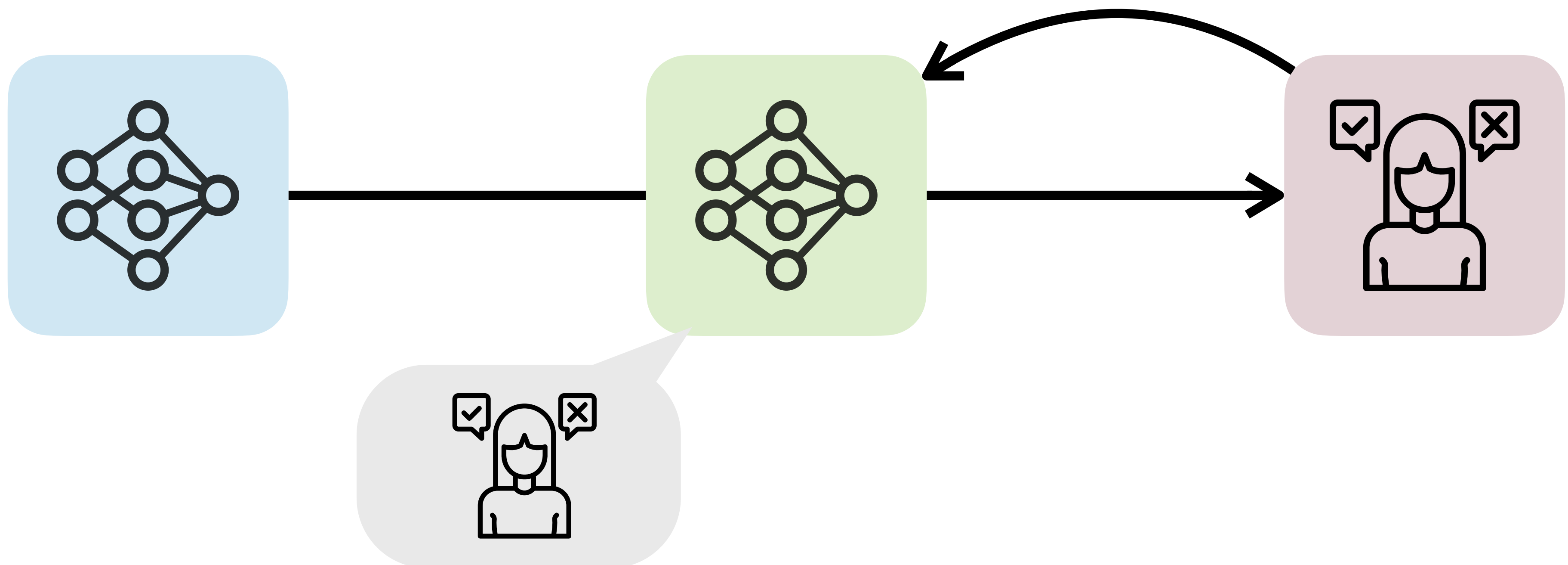


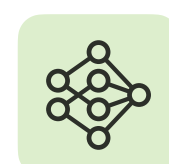
What would  model?

1. Form of explanation
2. Level of details
3. Persuasiveness

…

Learning to explain better

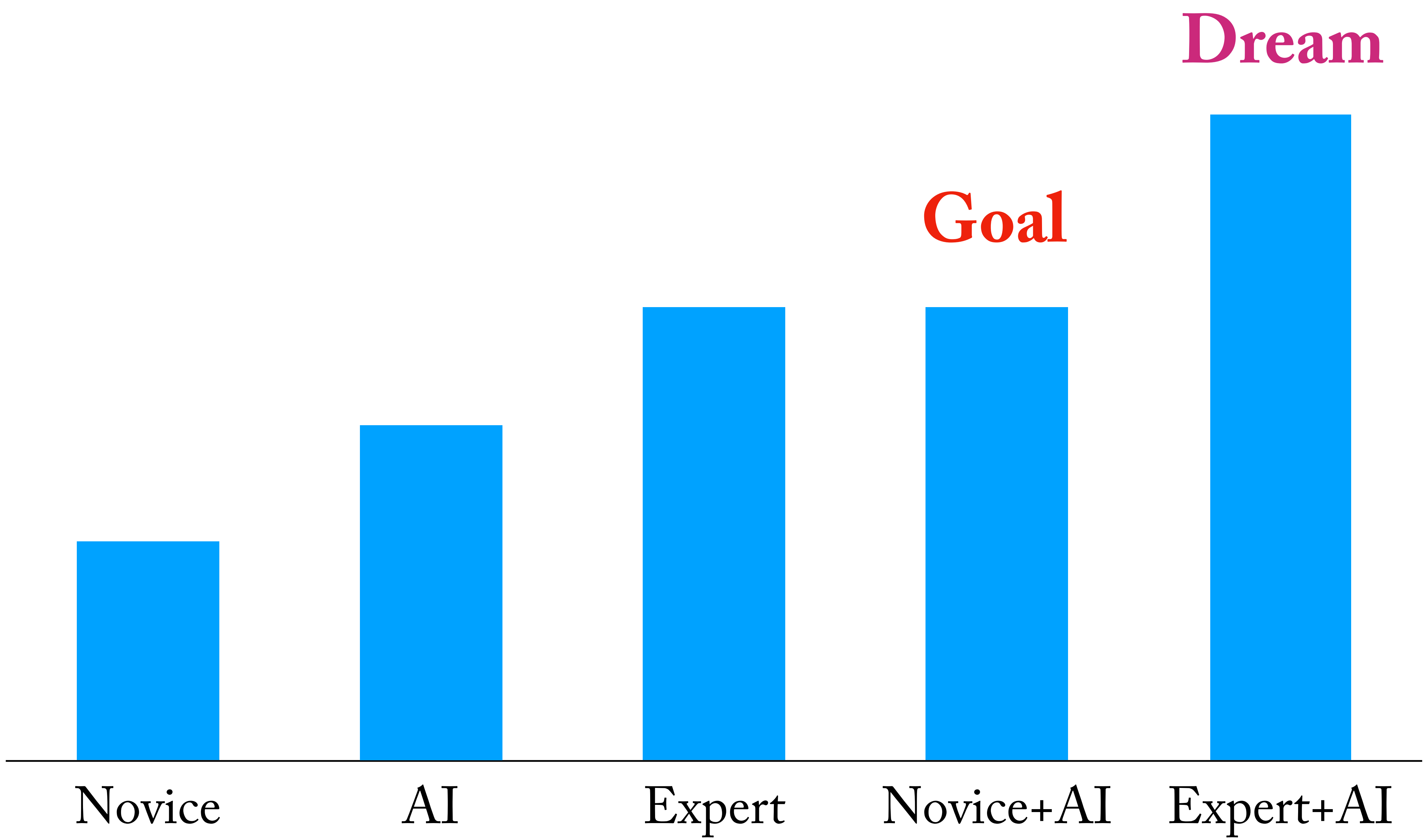


What would  model?

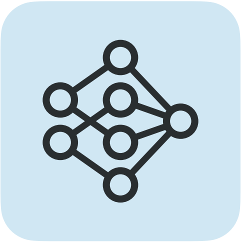
1. Form of explanation
2. Level of details

Online adaptation to real human users!

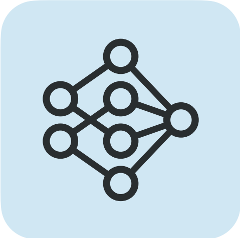
Designing the testbed
Goal: better human-AI performance



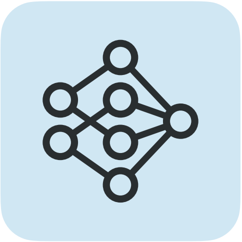
Designing the testbed: QA

	Question		
1	This model architecture is known for its use of attention mechanisms.	Transformer	✓
2	Many models using this architecture are named after Sesame street characters.	ELMo	?
3	This model architecture achieves 41.8 BLEU on WMT-14 English-French task.	LSTM	✗

Designing the testbed: QA

	Question		
1	This model architecture is known for its use of attention mechanisms.	Transformer	✓
2	Many models using this architecture are named after Sesame street characters.	ELMo	?
3	This model architecture achieves 41.8 BLEU on WMT-14 English-French task.	LSTM	✗

Designing the testbed: QA

	Question		
1	This model architecture is known for its use of attention mechanisms.	Transformer	✓
2	Many models using this architecture are named after Sesame street characters.	ELMo	?
3	This model architecture achieves 41.8 BLEU on WMT-14 English-French task.	LSTM	✗

Designing the testbed: QA

This model architecture is known for its use of attention mechanisms.

Many models using this architecture are named after Sesame street characters.

This model architecture achieves 41.8 BLEU on WMT-14 English-French task.

Designing the testbed: **Incremental** QA

This model architecture achieves 41.8 BLEU on WMT-14 English-French task.

Many models using this architecture are named after Sesame street characters.

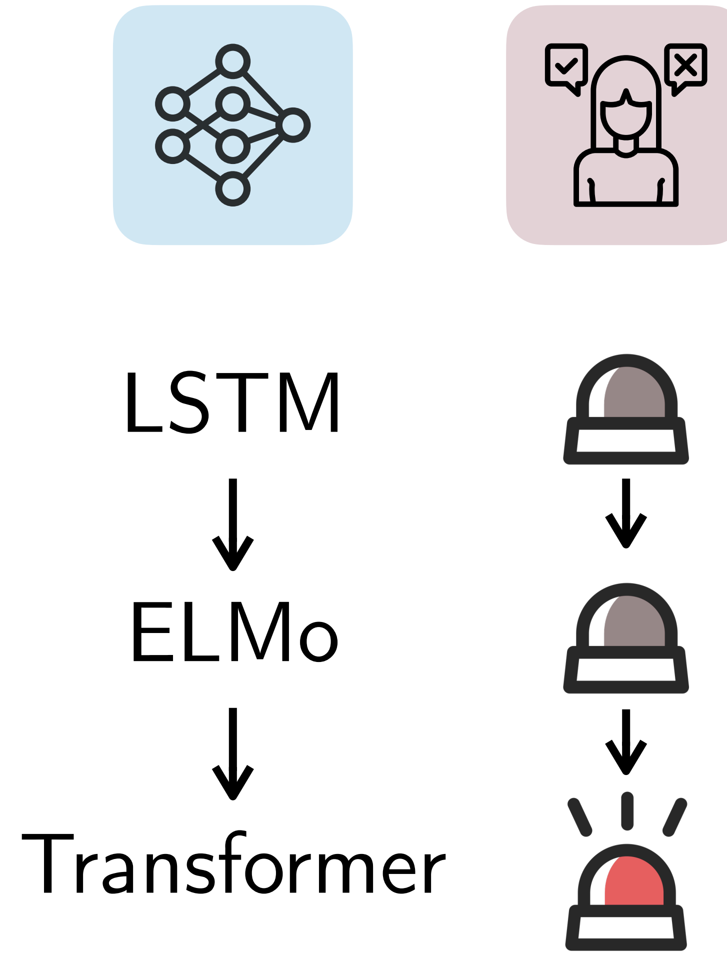
This model architecture is known for its use of attention mechanisms.

Designing the testbed: **Incremental** QA

This model architecture achieves 41.8 BLEU on WMT-14 English-French task.

Many models using this architecture are named after Sesame street characters.

This model architecture is known for its use of attention mechanisms.

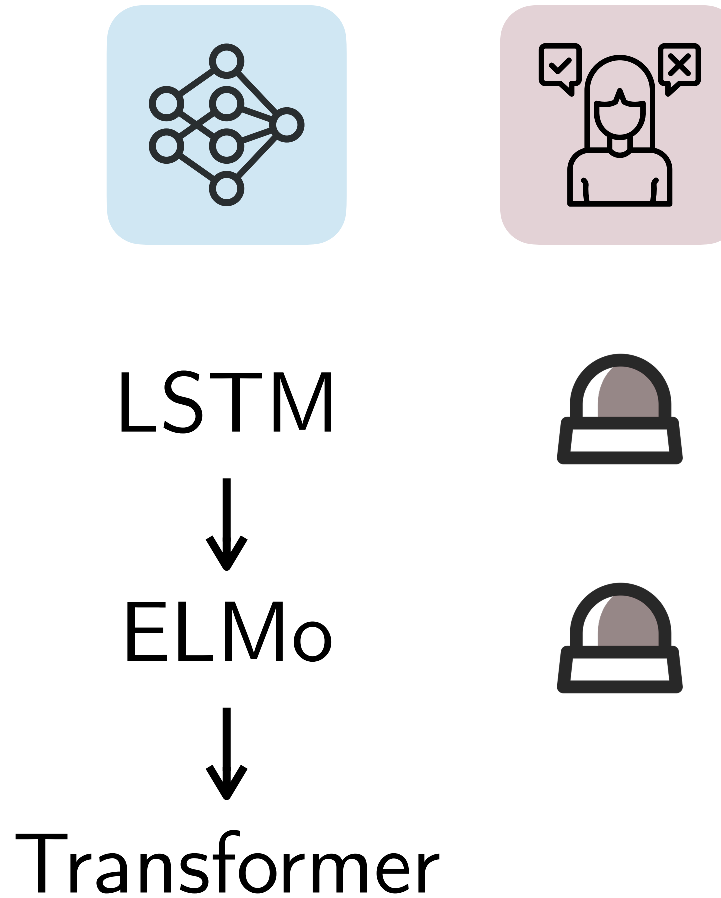


Designing the testbed: **Incremental** QA

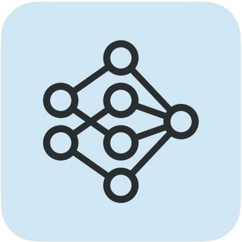
This model architecture achieves 41.8 BLEU on WMT-14 English-French task.

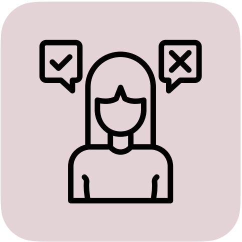
Many models using this architecture are named after Sesame street characters.

This model architecture is known for its use of attention mechanisms.



Designing the testbed: **Incremental** QA





This model architecture achieves 41.8 BLEU on WMT-14 English-French task.

Many models using this architecture are named after Sesame street characters.

This model architecture is known for its use of attention mechanisms.

LSTM

↓

ELMo

↓

Transformer

Designing the testbed: **Incremental** QA

This model architecture achieves 41.8 BLEU on WMT-14 English-French task.

Many models using this architecture are named after Sesame street characters.

This model architecture is known for its use of attention mechanisms.

+10

LSTM

↓

ELMo

↓

Transformer

Designing the testbed: **Incremental** QA

This model architecture achieves 41.8 BLEU on WMT-14 English-French task.

Many models using this architecture are named after Sesame street characters.

This model architecture is known for its use of attention mechanisms.

+25

LSTM

↓

ELMo

↓

Transformer

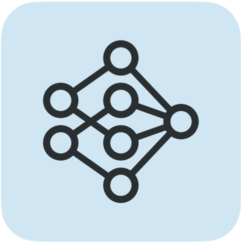
Designing the testbed: **Incremental** QA

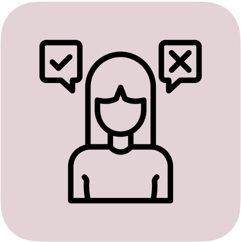
-25

This model architecture achieves 41.8 BLEU on WMT-14 English-French task.

Many models using this architecture are named after Sesame street characters.

This model architecture is known for its use of attention mechanisms.





LSTM

↓

ELMo

↓

Transformer

Incremental QA: interface

Buzz 0:27

Alternatives

#	Guess	Score
1	Milton Friedman	0.1529
2	David Ricardo	0.1122
3	John Kenneth Galbrai	0.1100
4	Friedrich Hayek	0.0945
5	Joseph Stiglitz	0.0938

Question

Along with Edmund Phelps, he argued that real wages will adjust to provide an equilibrium between the supply and demand for labor, leading to a "natural rate" of unemployment. He coined the phrase "Miracle of Chile" in

Guess: Milton Friedman

Evidence

monetarists, the long-run curve is a vertical line at the natural rate of unemployment . For 10 points

reversed by Robert (*) Lucas who argued that it is the difference between real and expected inflation, not

, wrote a paper in 1958 titled "__The Relation between Unemployment and the Rate of Change of Money Wage

product and lowering the unemployment rate . Moving along the Phillips curve, this would lead to a

Incremental QA: interface

Buzz 0:27

Alternatives

#	Guess	Score
1	Milton Friedman	0.1529
2	David Ricardo	0.1122
3	John Kenneth Galbrai	0.1100
4	Friedrich Hayek	0.0945
5	Joseph Stiglitz	0.0938

Question

Along with Edmund Phelps, he argued that real wages will adjust to provide an equilibrium between the supply and demand for labor, leading to a "natural rate" of unemployment. He coined the phrase "Miracle of Chile" in

Guess: Milton Friedman

Evidence

monetarists, the long-run curve is a vertical line at the natural rate of unemployment . For 10 points

reversed by Robert (*) Lucas who argued that it is the difference between real and expected inflation, not

, wrote a paper in 1958 titled "__The Relation between Unemployment and the Rate of Change of Money Wage

product and lowering the unemployment rate . Moving along the Phillips curve, this would lead to a

Incremental QA: interface

Buzz 0:27

Alternatives

#	Guess	Score
1	Milton Friedman	0.1529
2	David Ricardo	0.1122
3	John Kenneth Galbrai	0.1100
4	Friedrich Hayek	0.0945
5	Joseph Stiglitz	0.0938

Question

Along with Edmund Phelps, he argued that real wages will adjust to provide an equilibrium between the supply and demand for labor, leading to a "natural rate" of unemployment. He coined the phrase "Miracle of Chile" in

Guess: **Milton Friedman**

Evidence

monetarists, the long-run curve is a vertical line at the natural rate of unemployment . For 10 points

reversed by Robert (*) Lucas who argued that it is the difference between real and expected inflation, not

, wrote a paper in 1958 titled "__The Relation between Unemployment and the Rate of Change of Money Wage

product and lowering the unemployment rate . Moving along the Phillips curve, this would lead to a

Incremental QA: interface

Buzz

0:27

Alternatives

#	Guess	Score
1	Milton Friedman	0.1529
2	David Ricardo	0.1122
3	John Kenneth Galbrai	0.1100
4	Friedrich Hayek	0.0945
5	Joseph Stiglitz	0.0938

Question

Along with Edmund Phelps, he argued that real wages will adjust to provide an equilibrium between the supply and demand for labor, leading to a "natural rate" of unemployment. He coined the phrase "Miracle of Chile" in

Guess: **Milton Friedman**

Evidence

monetarists, the long-run curve is a vertical line at the natural rate of unemployment . For 10 points

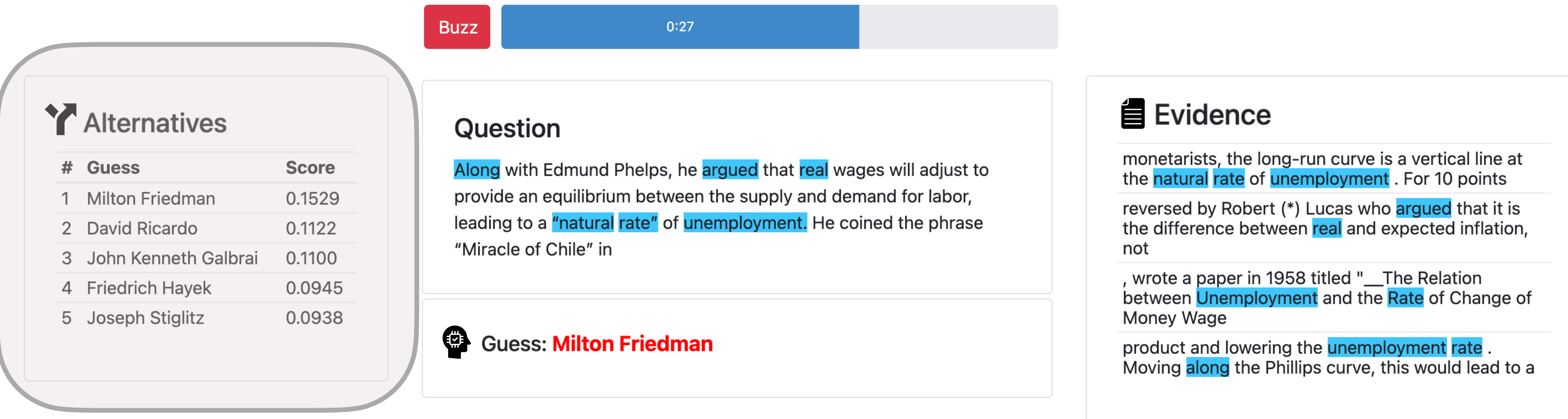
reversed by Robert (*) Lucas who argued that it is the difference between real and expected inflation, not

, wrote a paper in 1958 titled "__The Relation between Unemployment and the Rate of Change of Money Wage

product and lowering the unemployment rate . Moving along the Phillips curve, this would lead to a

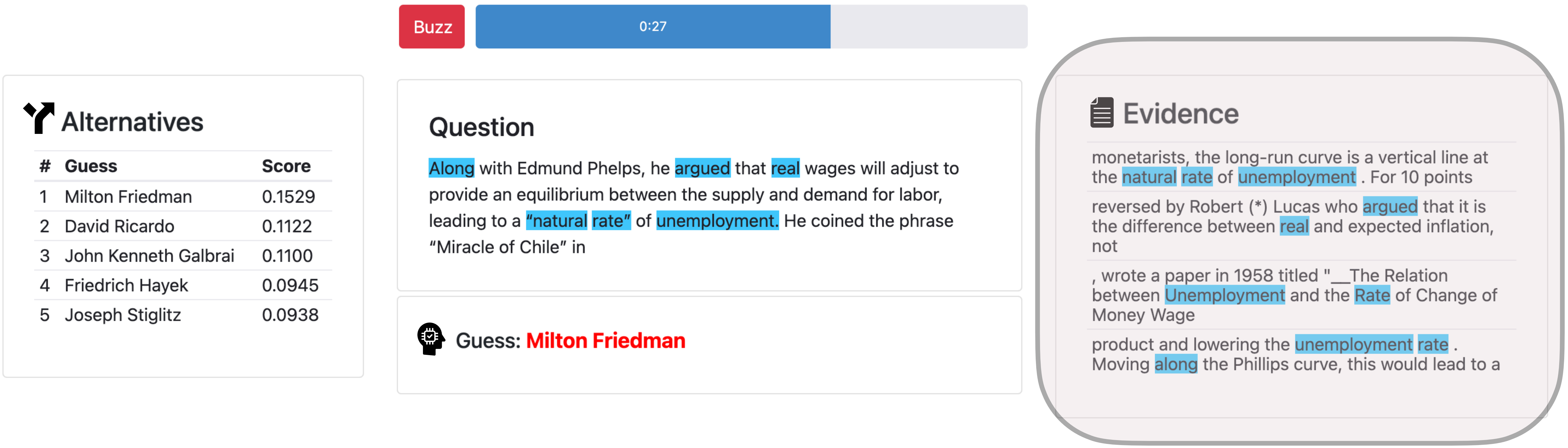
Guess: model prediction

Incremental QA: interface



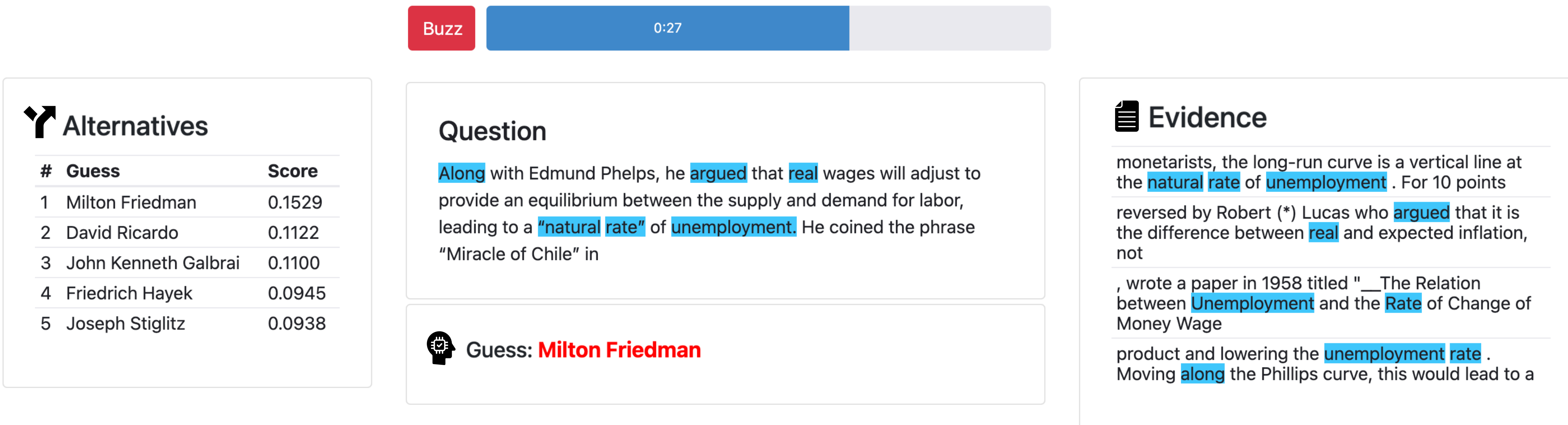
- Guess: model prediction
- Alternatives: other possible answers & confidence scores

Incremental QA: interface



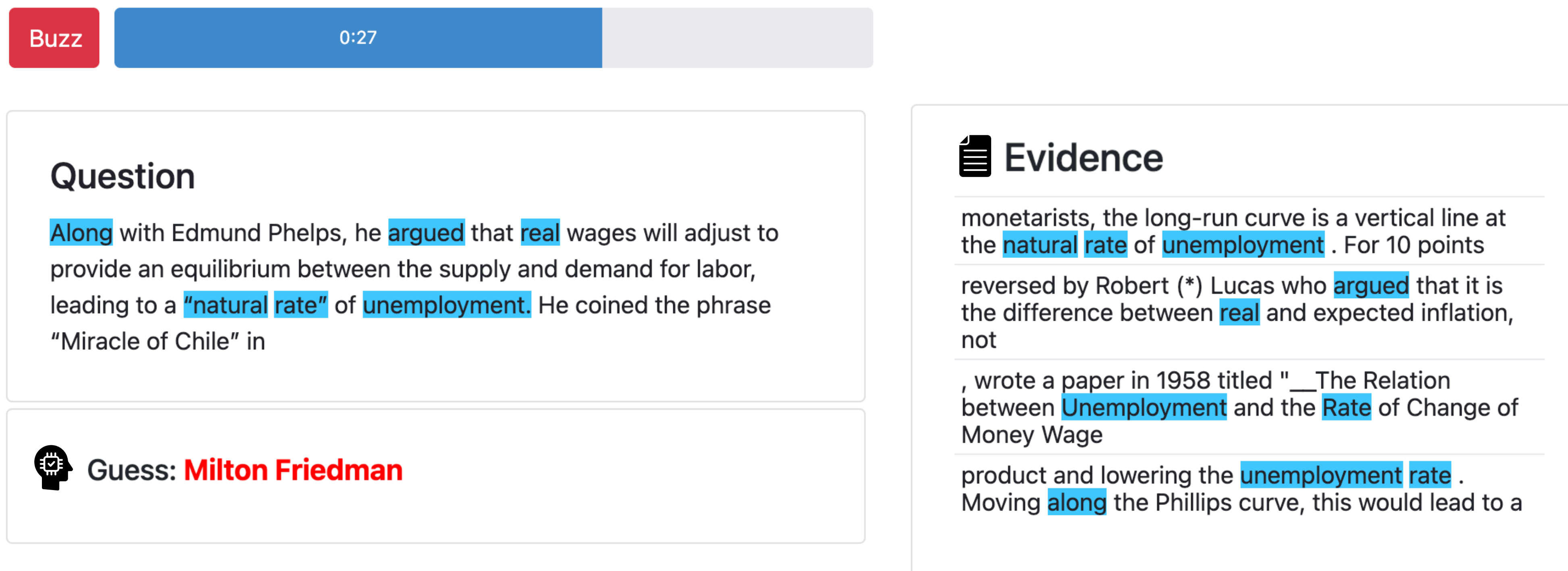
- Guess: model prediction
- Alternatives: other possible answers & confidence scores
- Evidence: relevant training examples (kNN)

Incremental QA: interface



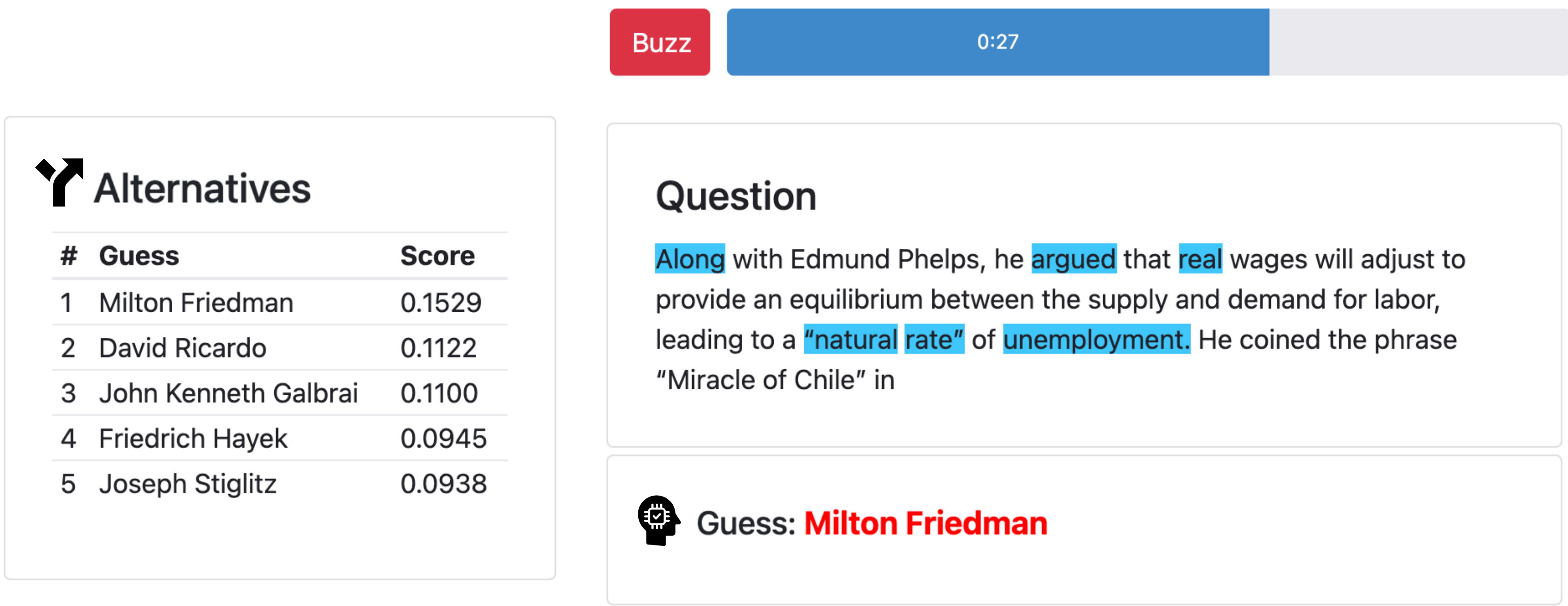
Modular interface: each explanation can be turned on/off individually

Incremental QA: interface



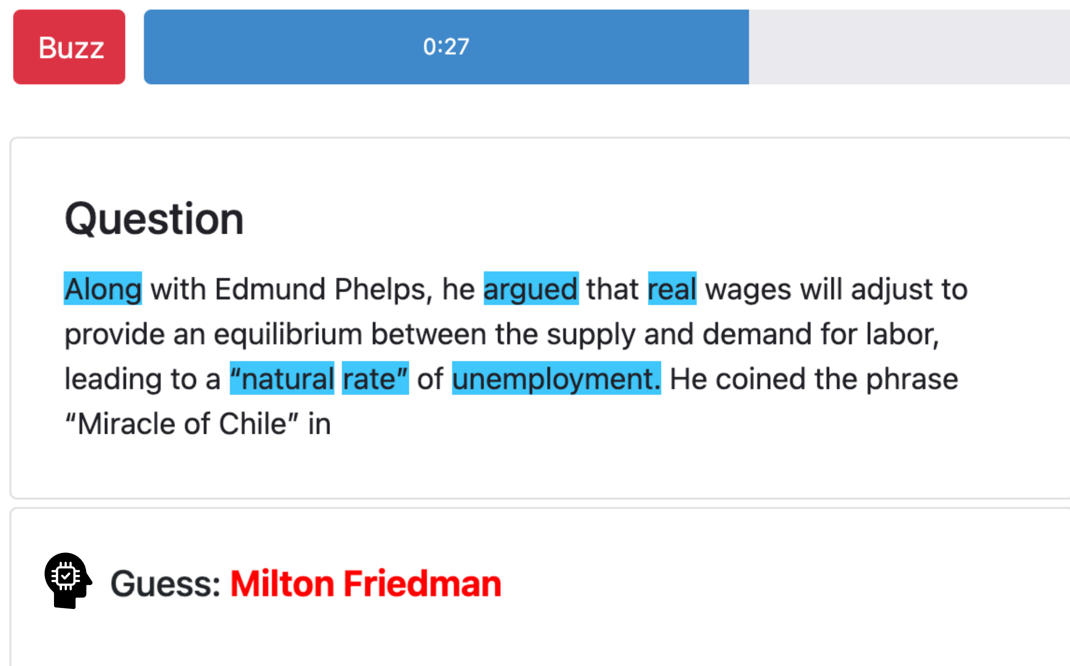
Modular interface: each explanation can be turned on/off individually

Incremental QA: interface



Modular interface: each explanation can be turned on/off individually

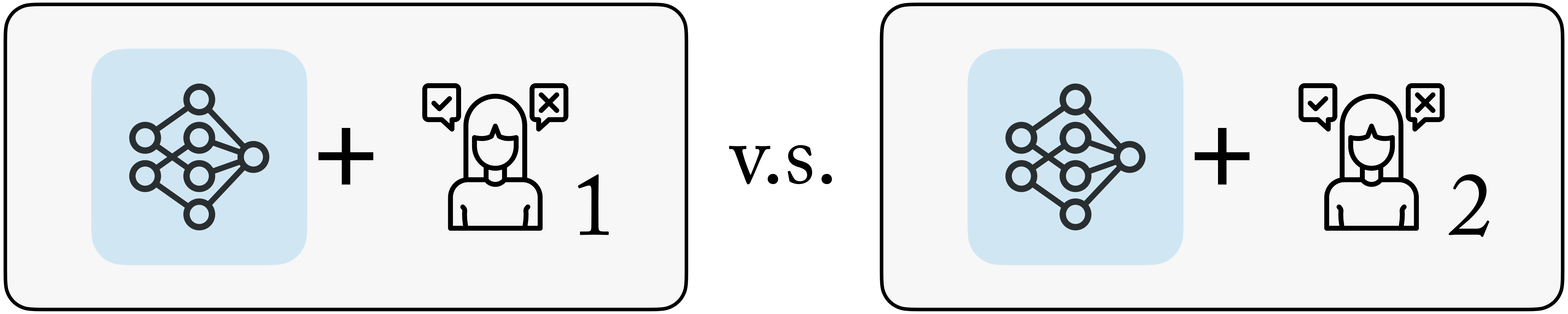
Incremental QA: interface



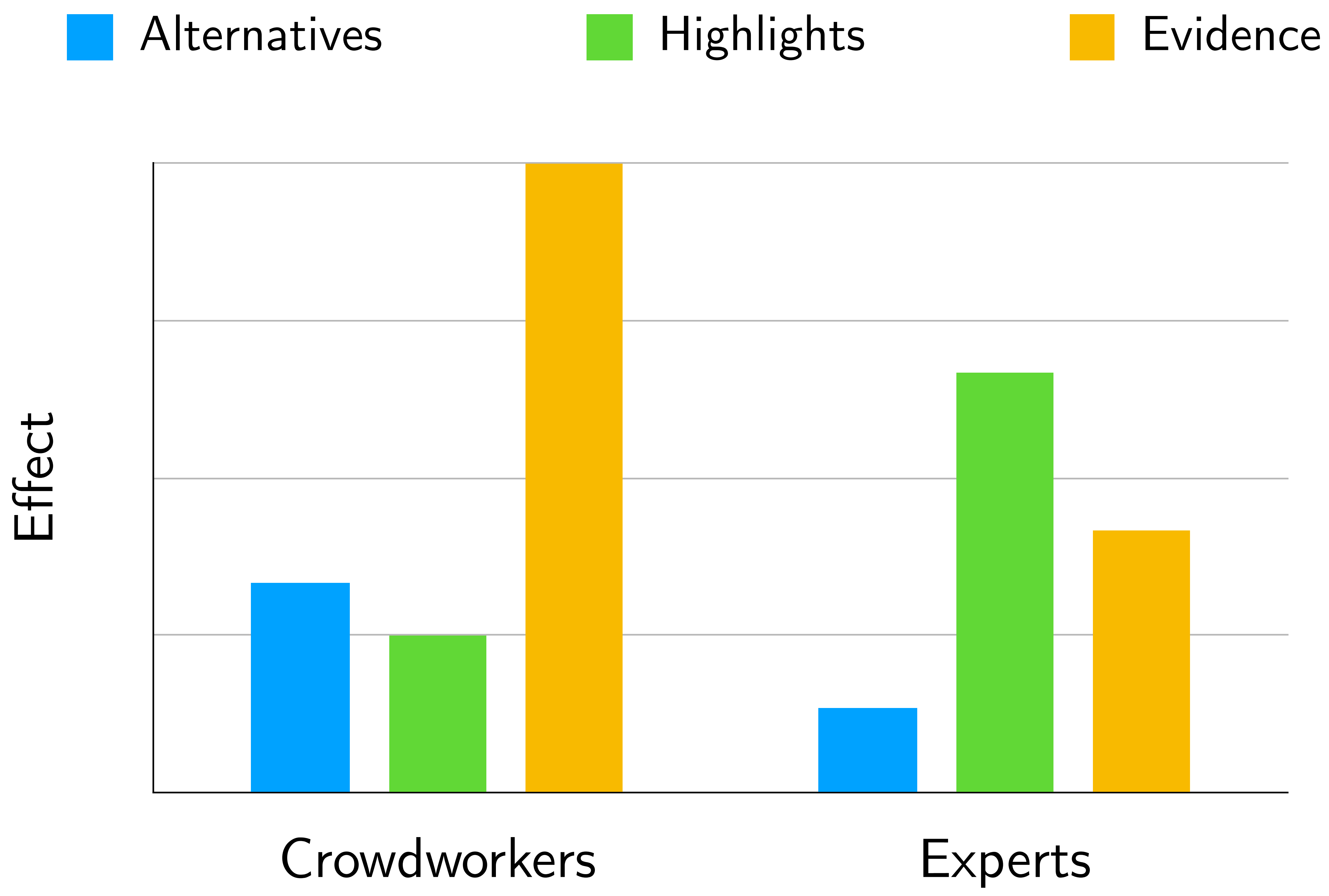
Modular interface: each explanation can be turned on/off individually

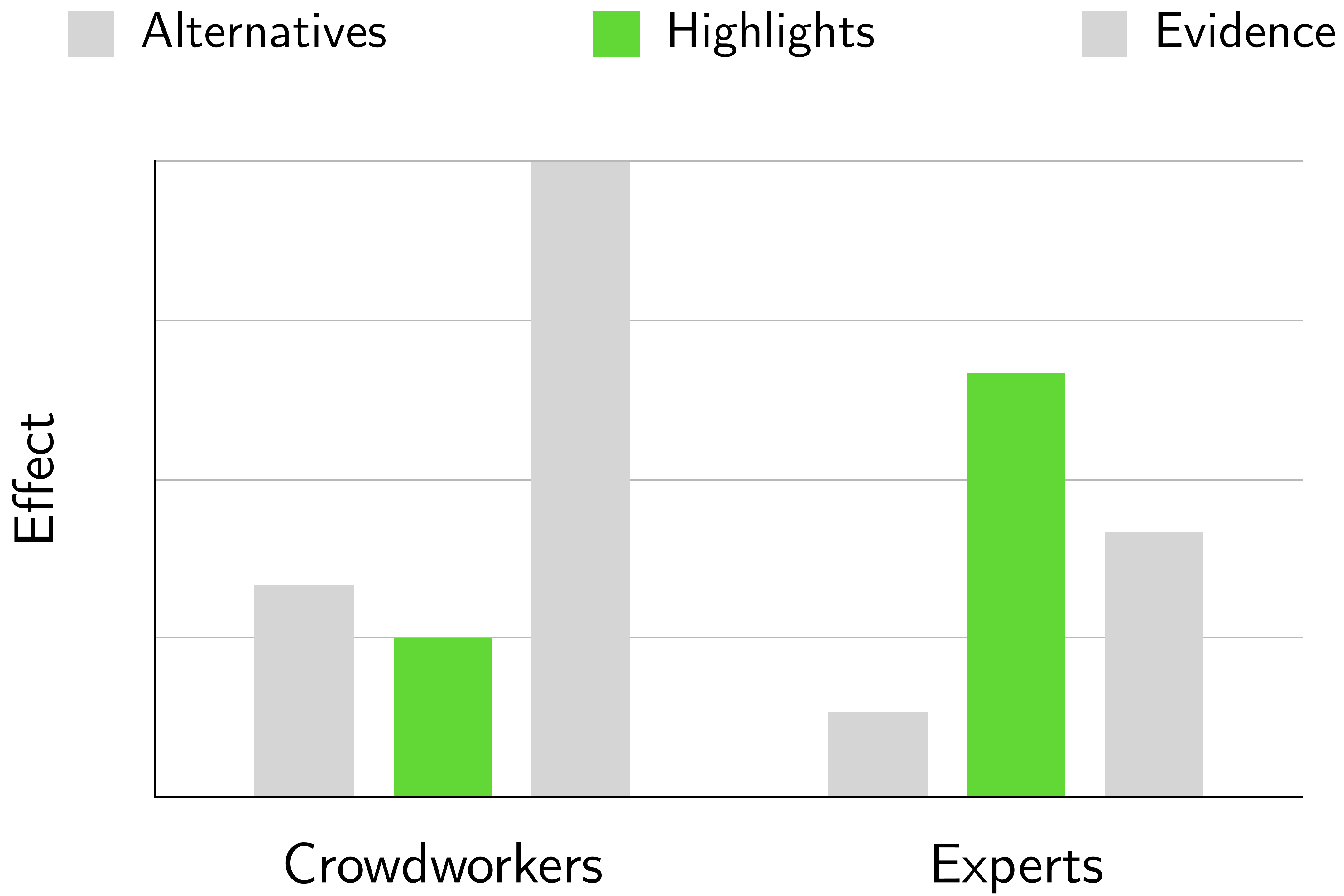
Allows for adjustment & adaptation.

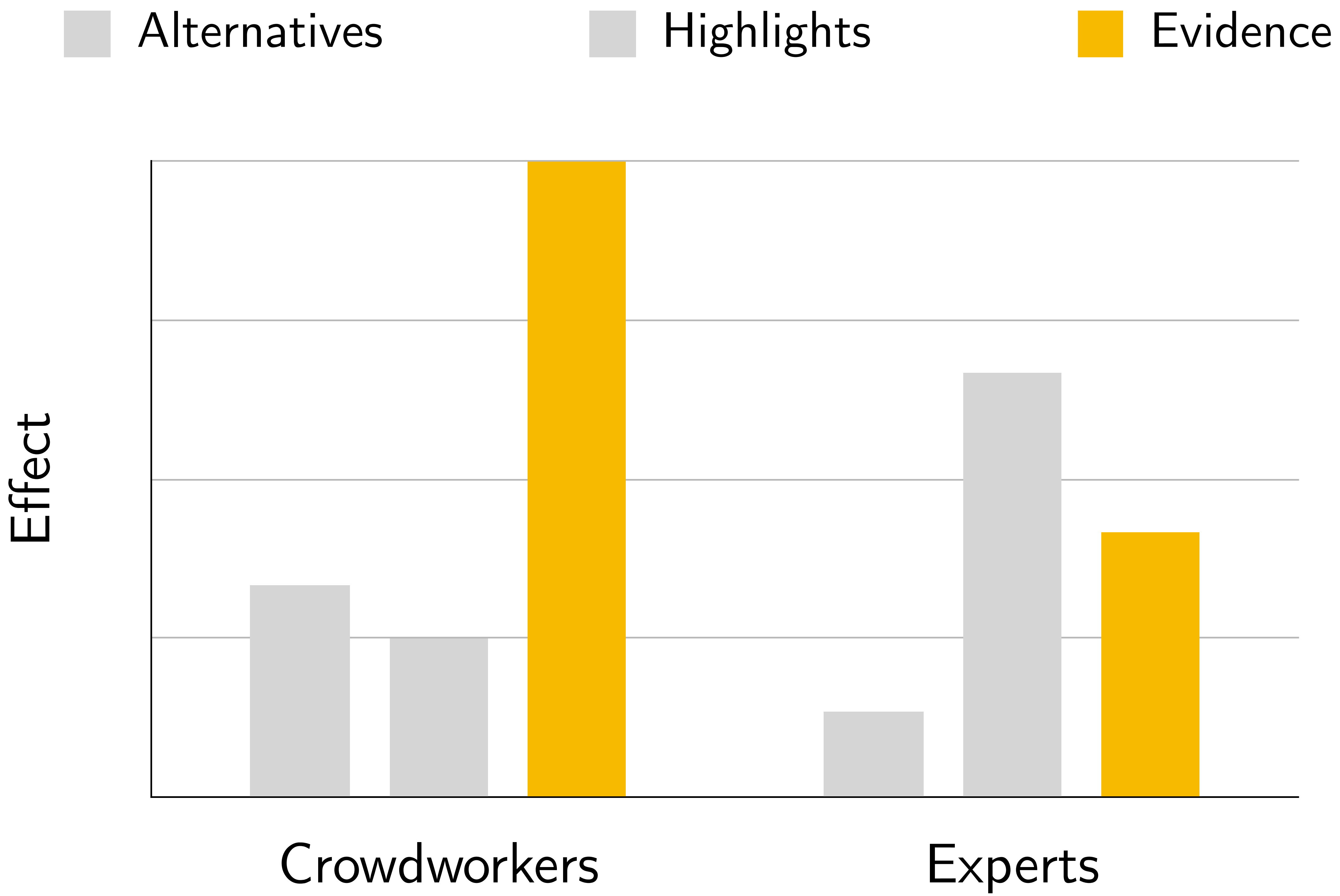
Incremental QA: gamification



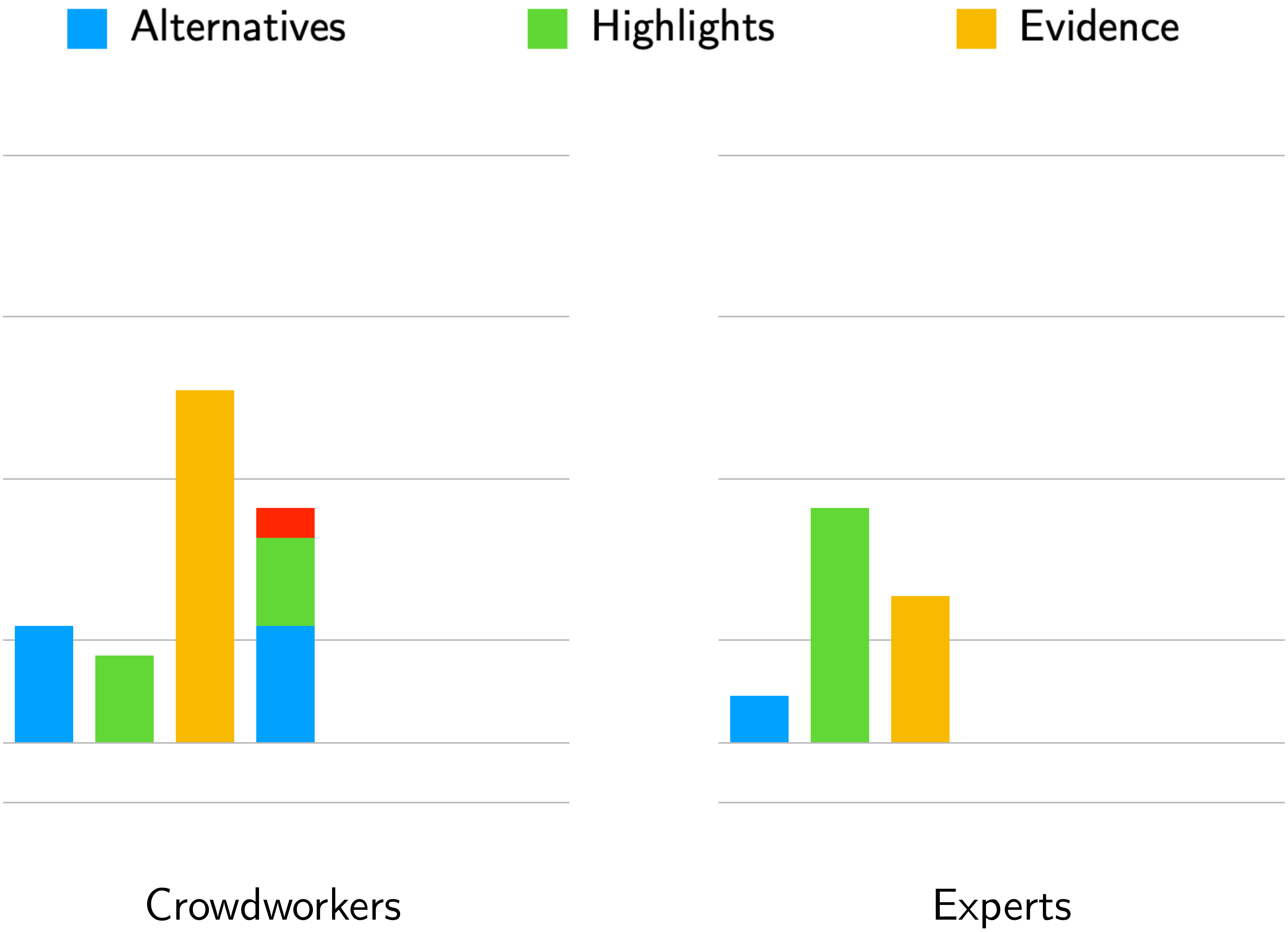
1. Human+AI teams compete against each other
2. Low stake, but high engagement
3. Sequential, fine-grained comparison
4. We can make the task arbitrarily difficult
5. Near expert-level AIs


Alternatives
Highlights
Evidence
Effect
Crowdworkers
Experts

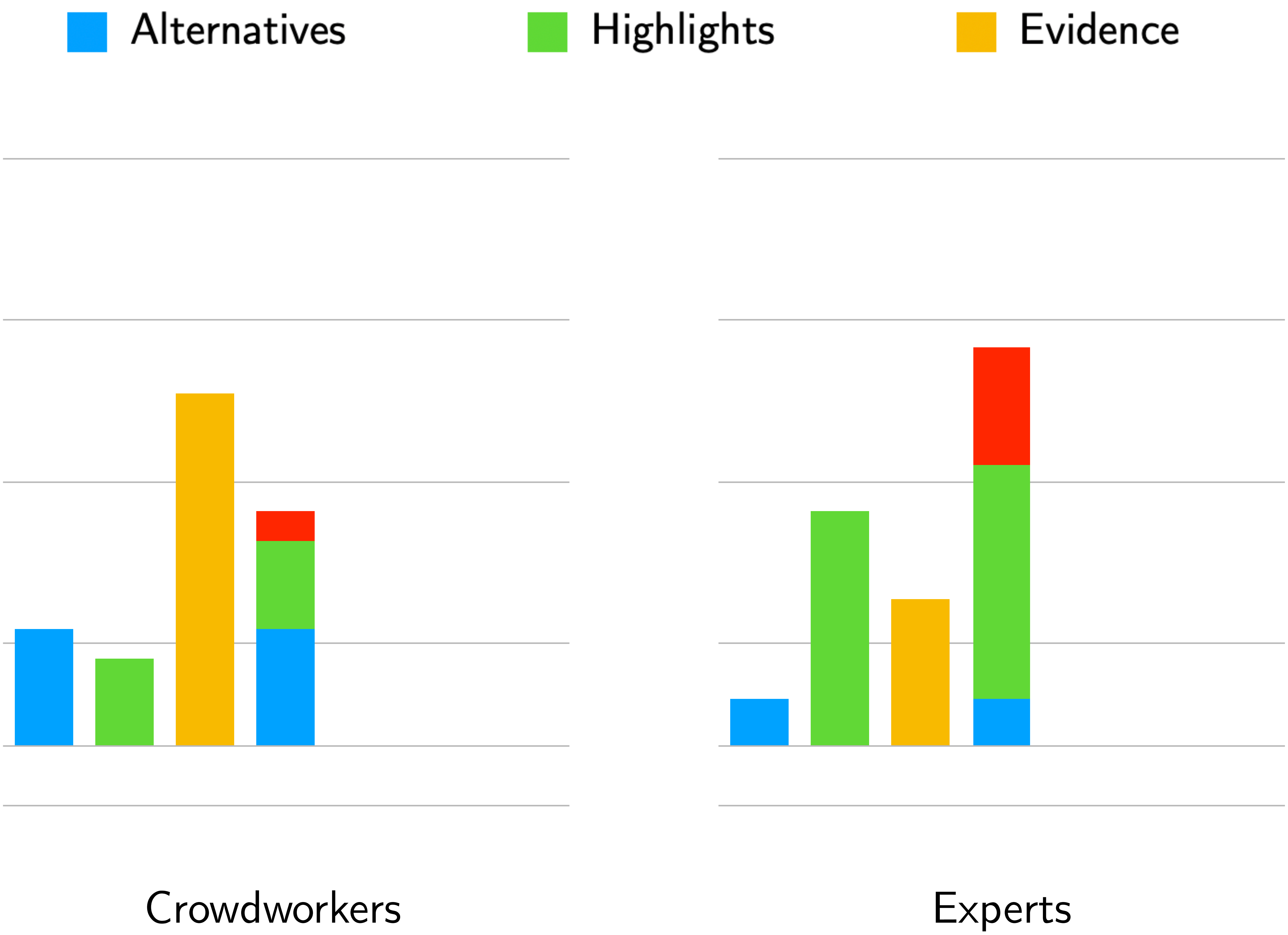

Alternatives
Highlights
Evidence
Effect
Crowdworkers
Experts



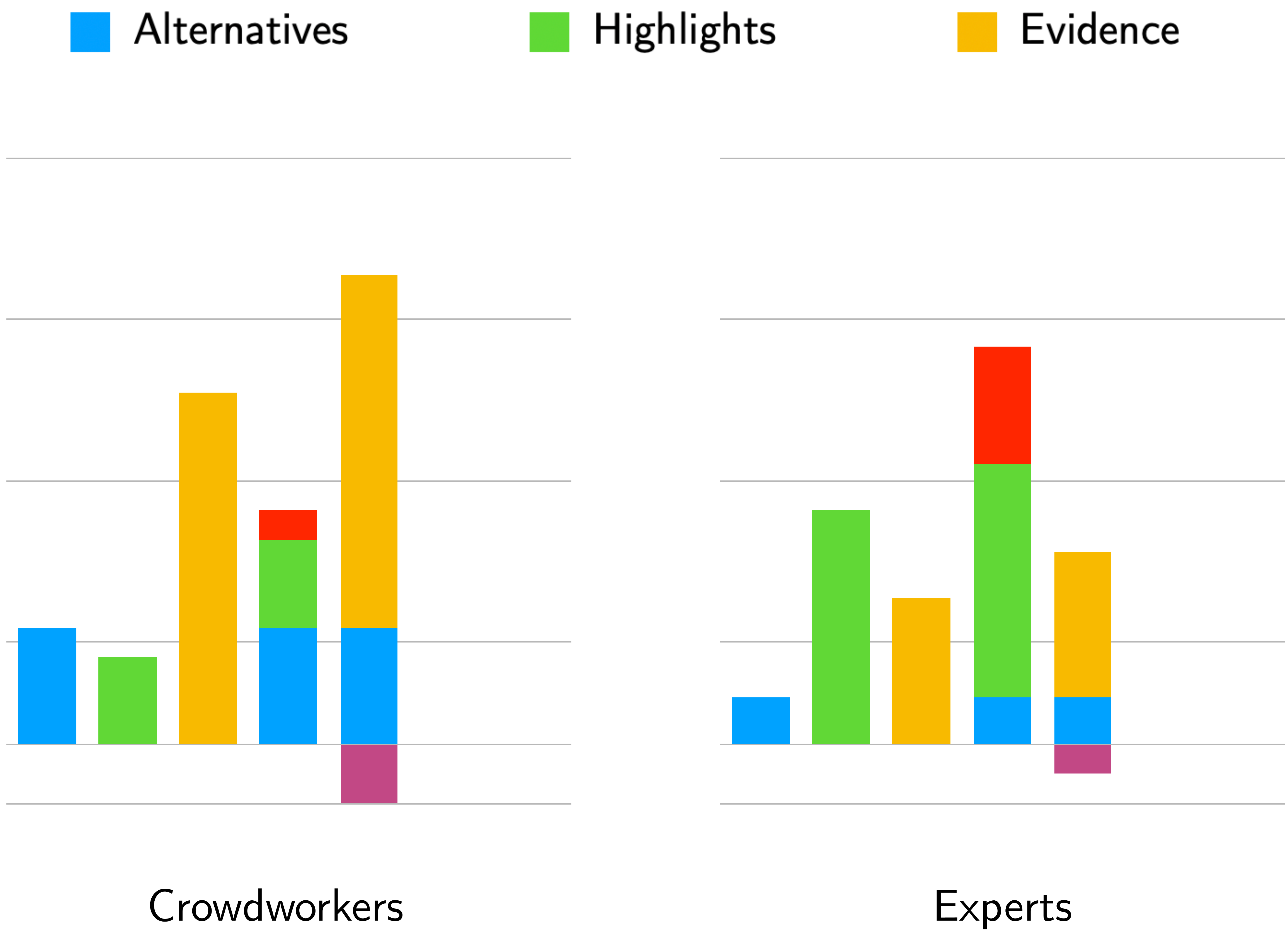
Alternatives
Highlights
Evidence
Effect
Crowdworkers
Experts

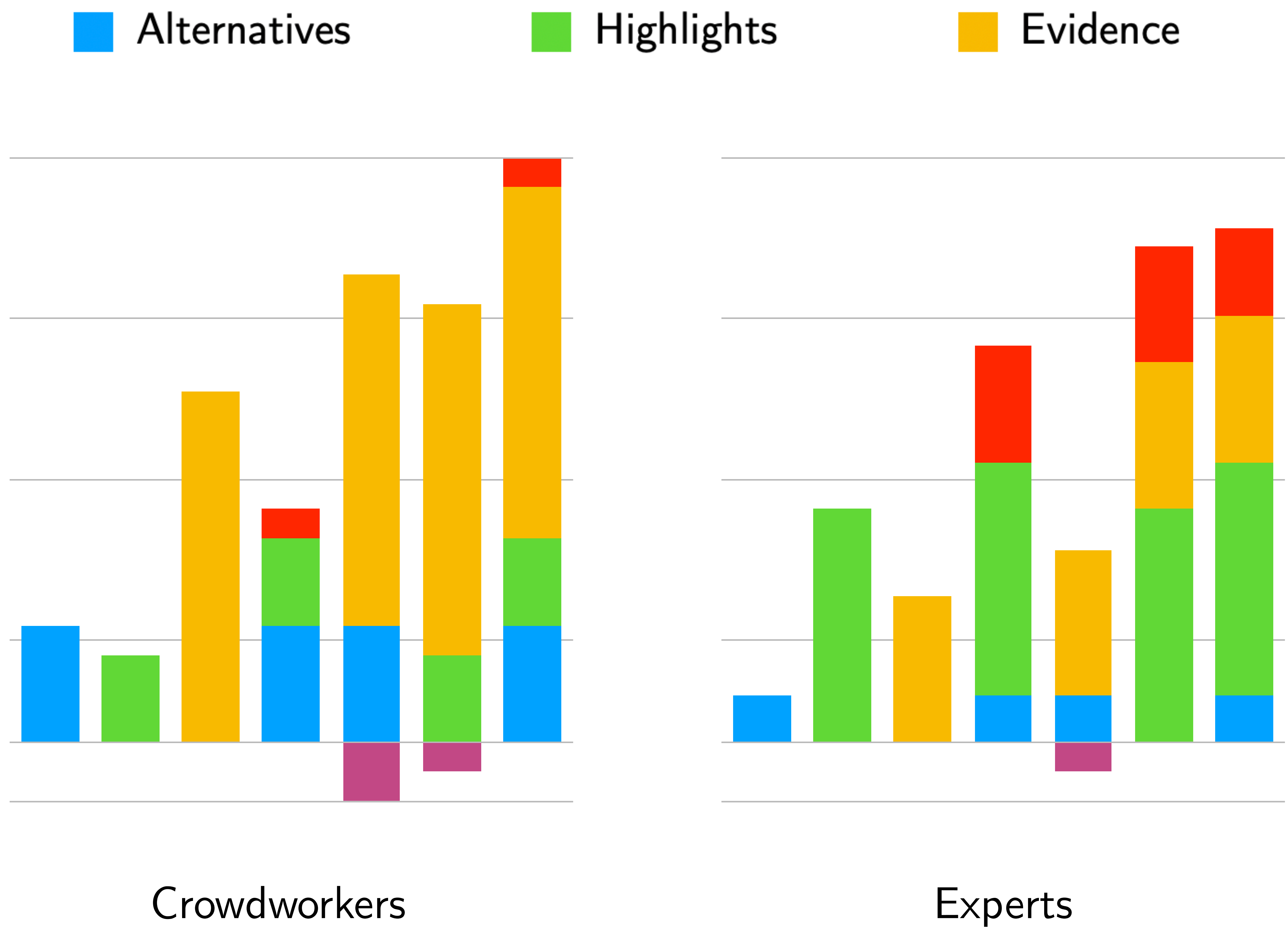


Alternatives
Highlights
Evidence
Crowdworkers
Experts

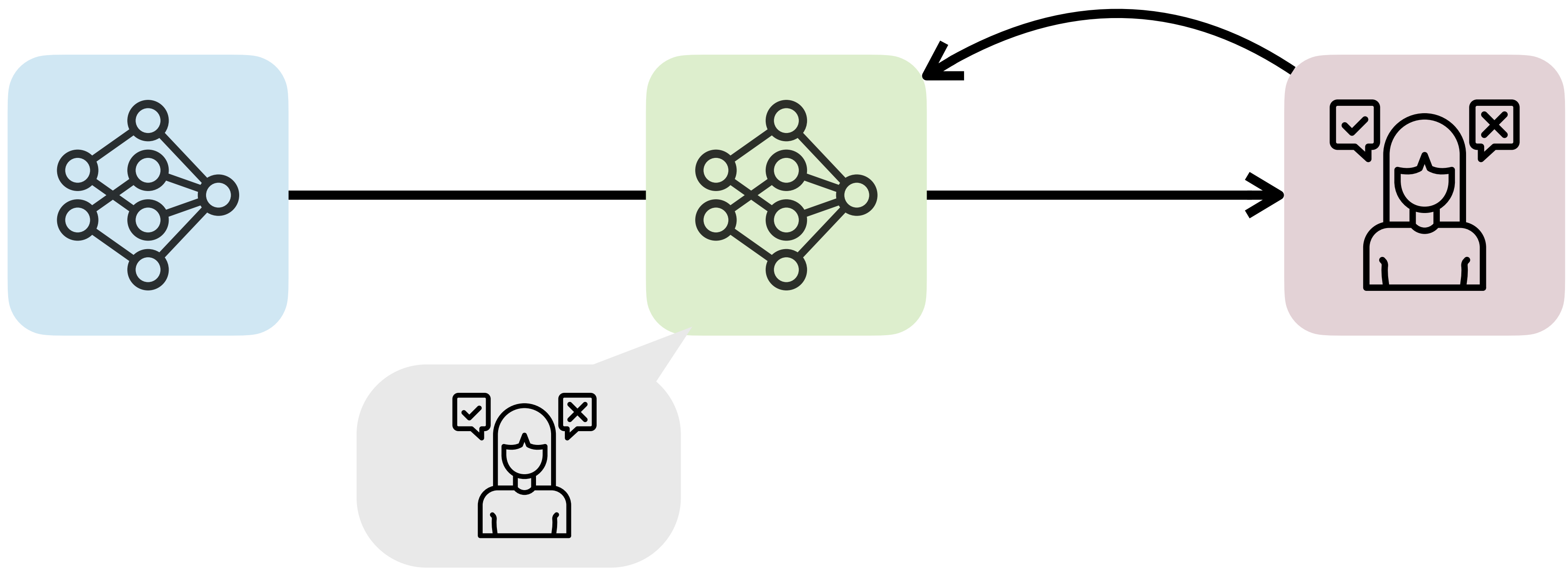


Alternatives
Highlights
Evidence
Crowdworkers
Experts

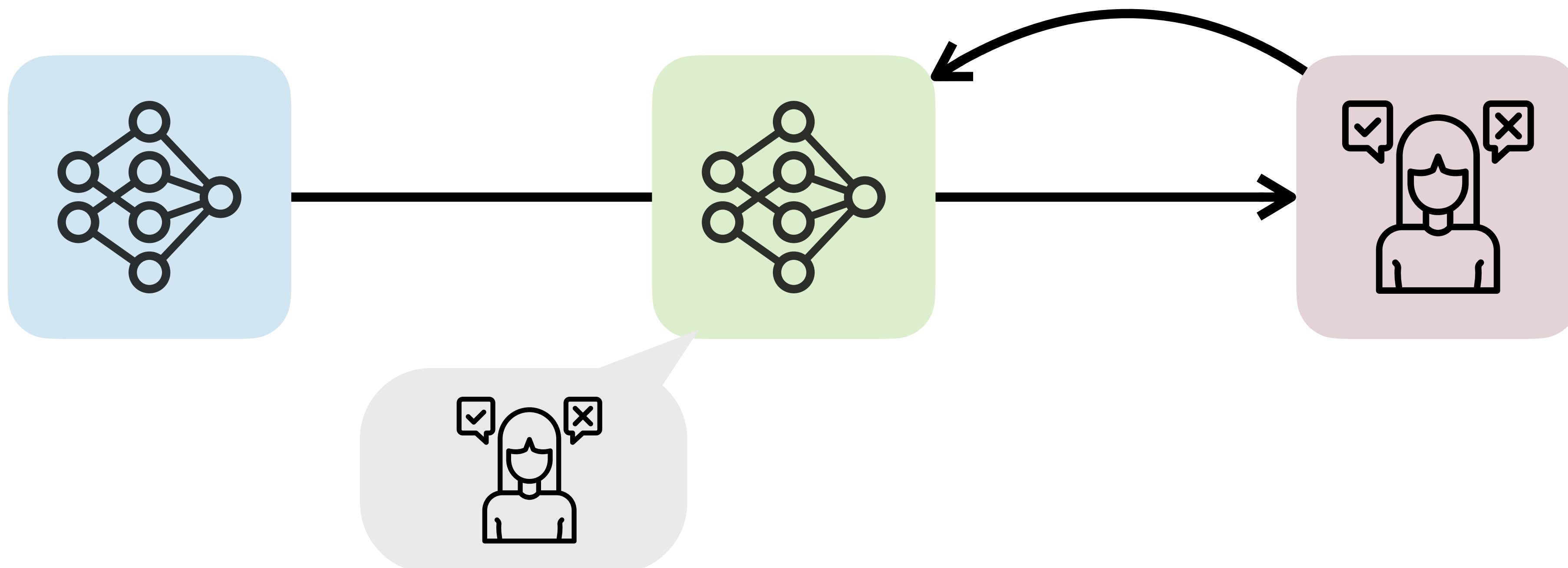

Alternatives
Highlights
Evidence
Crowdworkers
Experts

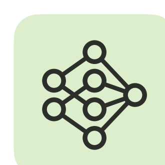

Alternatives
Highlights
Evidence
Crowdworkers
Experts

Learning to explain better



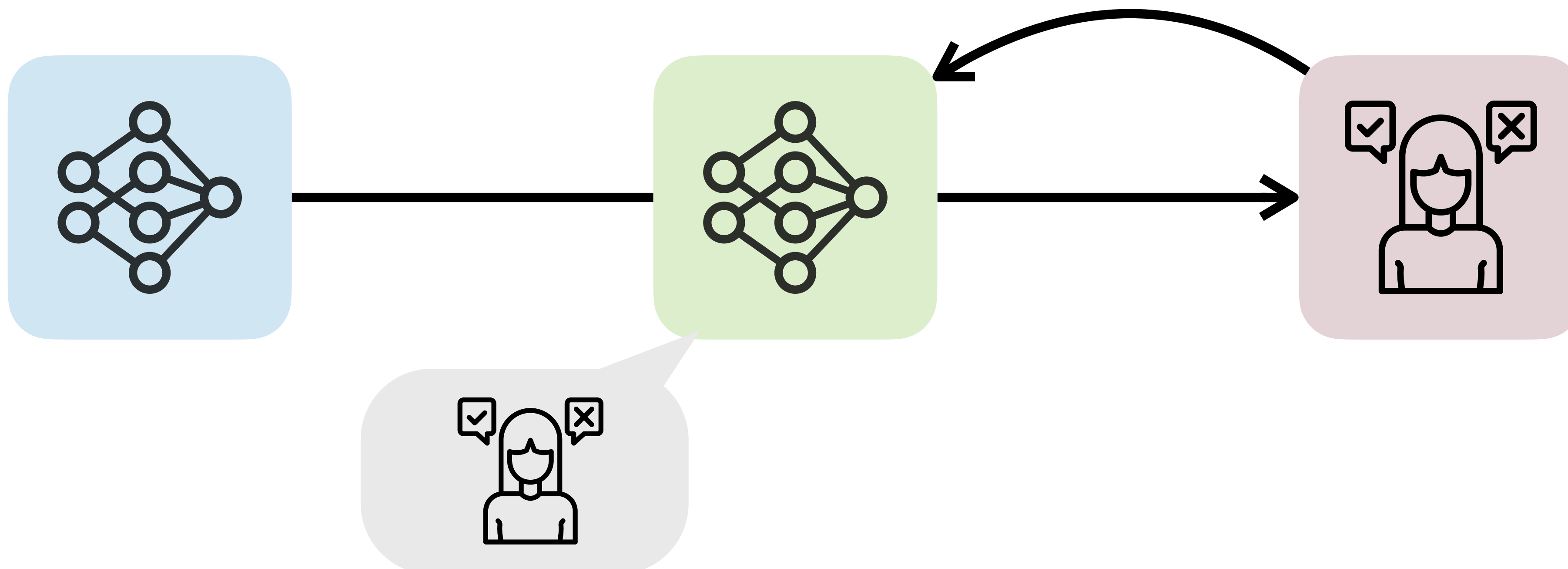
Learning to explain better, **selectively**



What would  actually do?

1. Model the interpretation process
2. Choose configuration for each decision

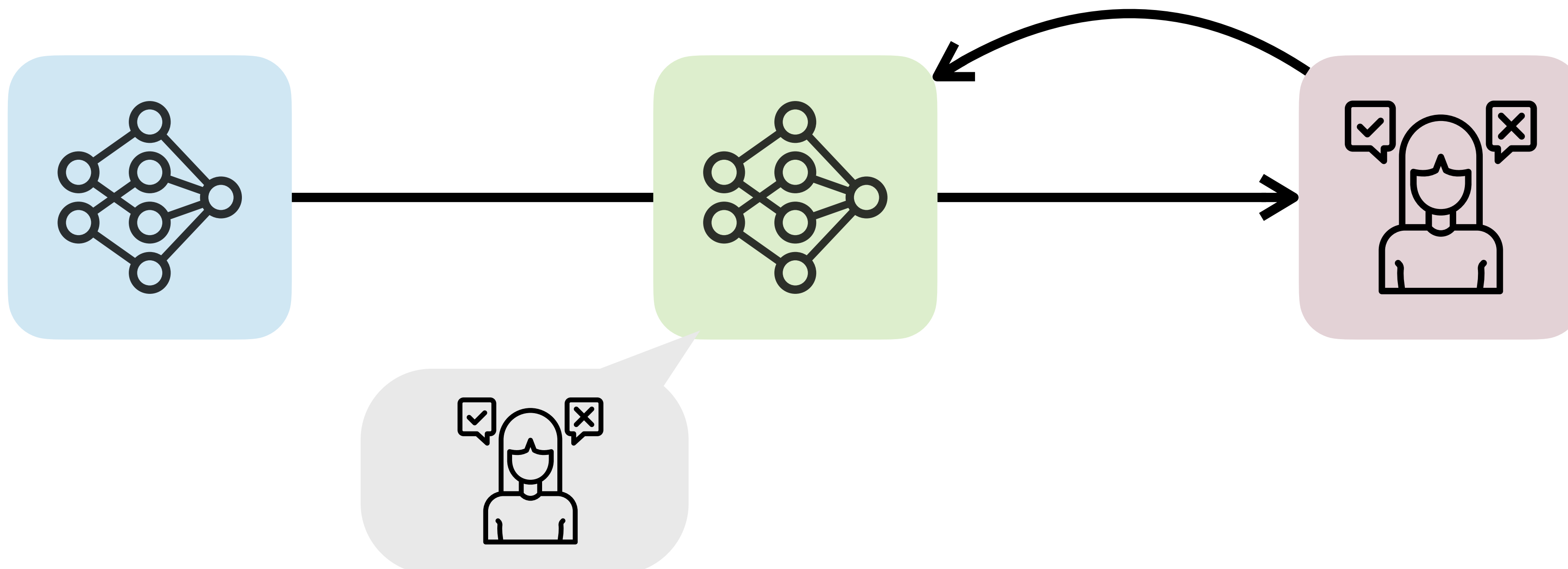
Learning to explain better, **selectively**



$$f(y|\langle x_i, s_j, t\rangle; \theta)$$

Expected score given
Question (x_i)
Player (s_j)
Explanation (config, t)

Learning to explain better, **selectively**

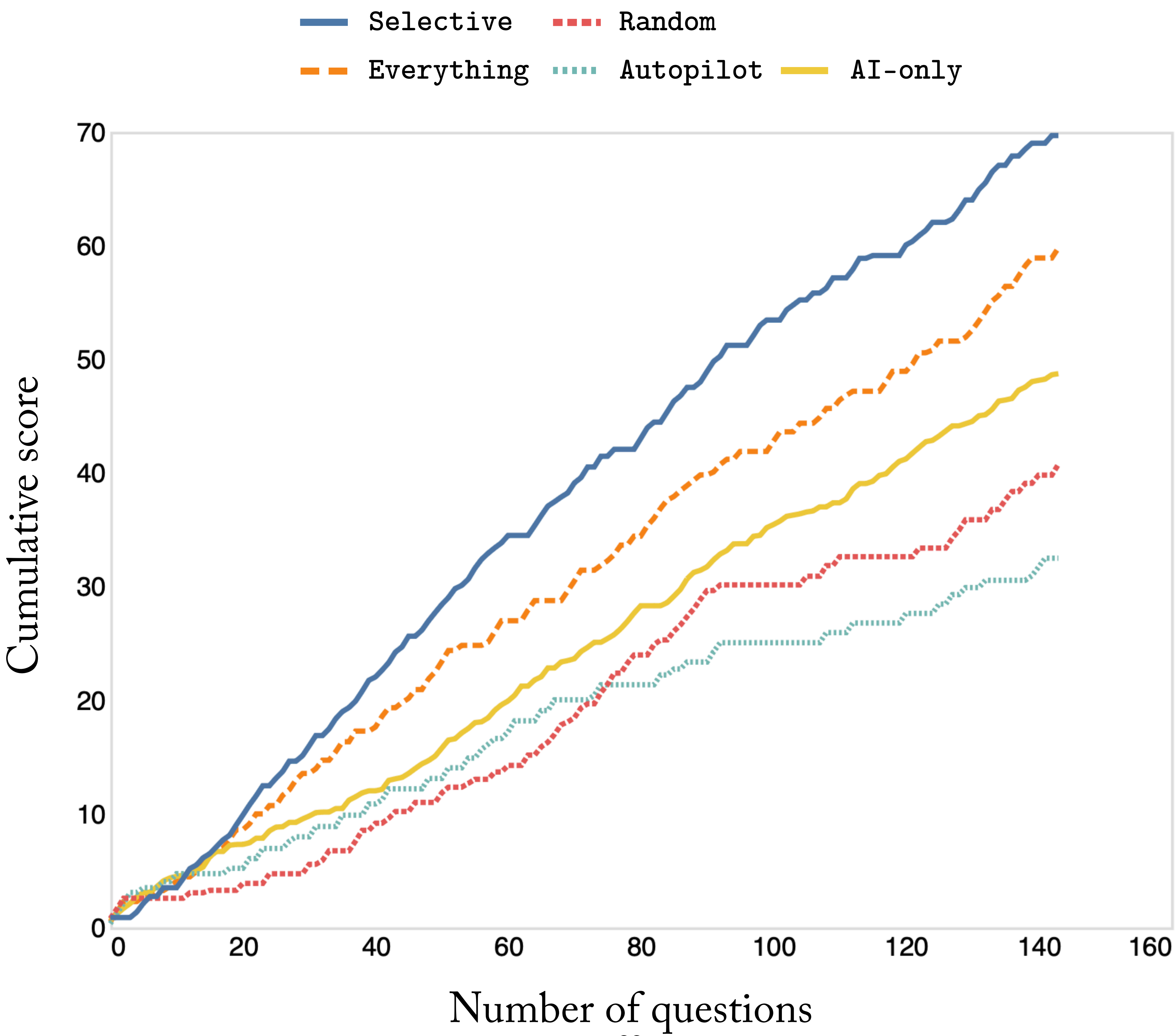


Offline
warm-start

$$f(y|\langle x_i, s_j, \cancel{t}\rangle; \theta)$$

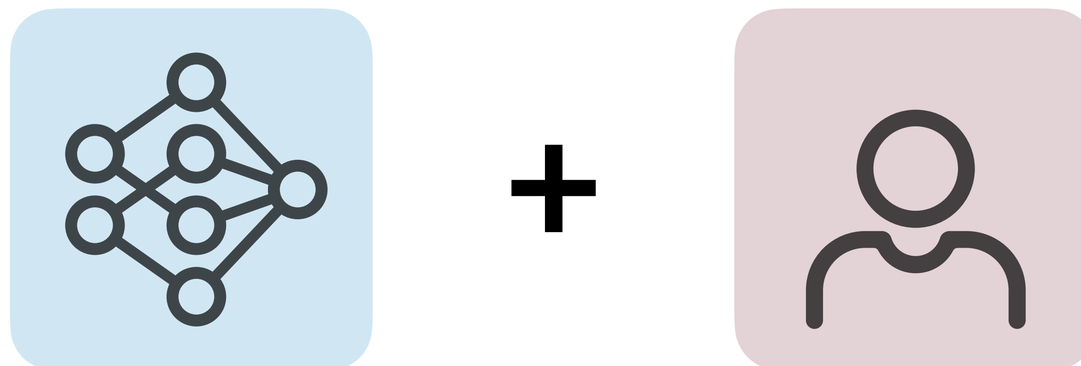
Online
Bandit

$$f(y|\langle x_i, s_j, t\rangle; \theta)$$



Selective
Random
Everything
Autopilot
AI-only
Cumulative score
70
60
50
40
30
20
10
0
0
20
40
60
80
100
120
140
160
Number of questions

What did we learn?



1. AIs can learn to explain better!
2. How? Adjust level of details.
3. Warm-starting the user model.
4. Engagement is crucial

Pragmatic Machine Explanations

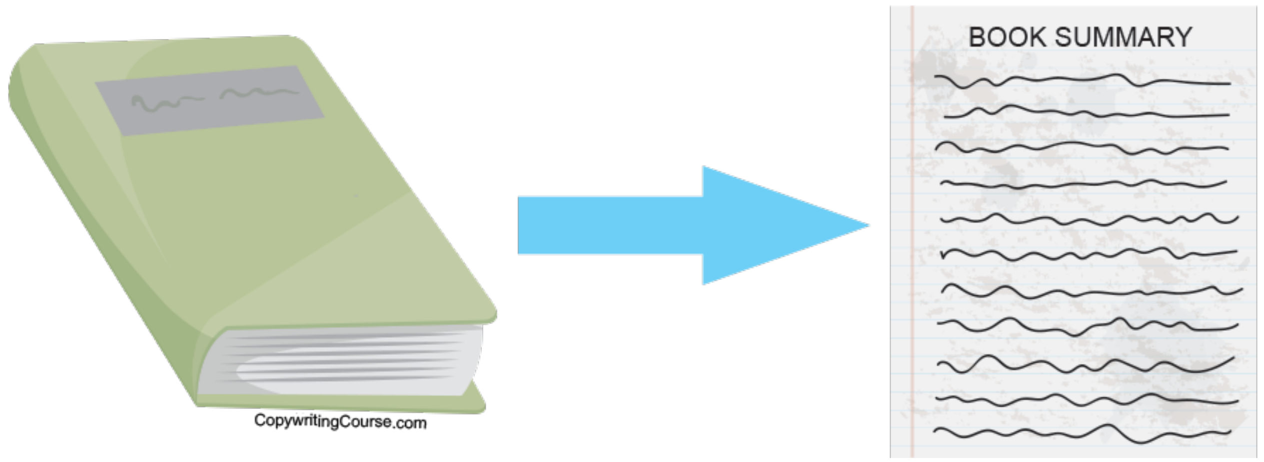
What's next? Pragmatic summarization

We started from off-the-shelf post-hoc methods.

Adjustment: which one to show.

Limitation: flexibility.

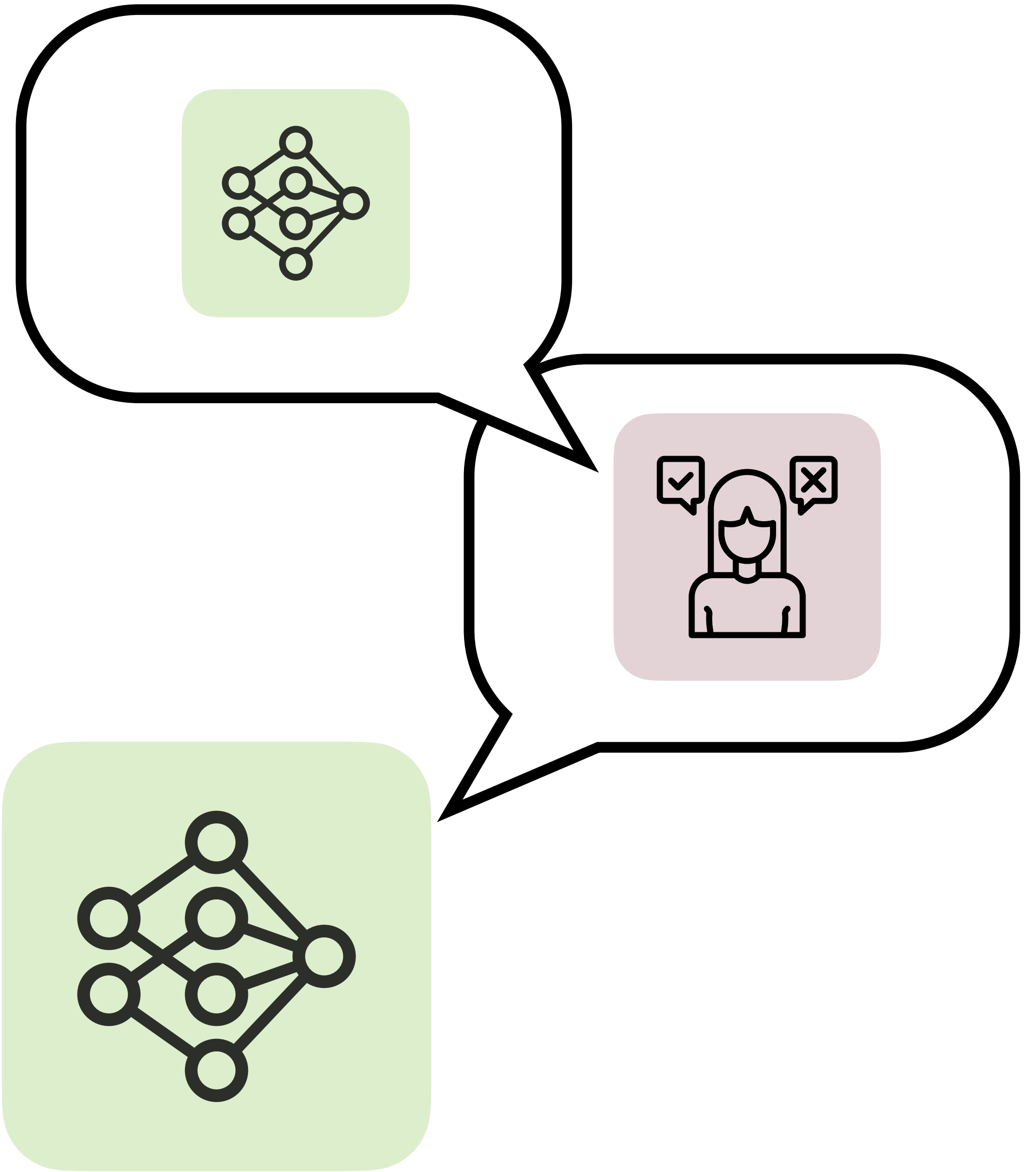
But it was an intentional choice to prioritize efficiency.



Pragmatic Summarizations

What's next? Theory of pragmatic exp.

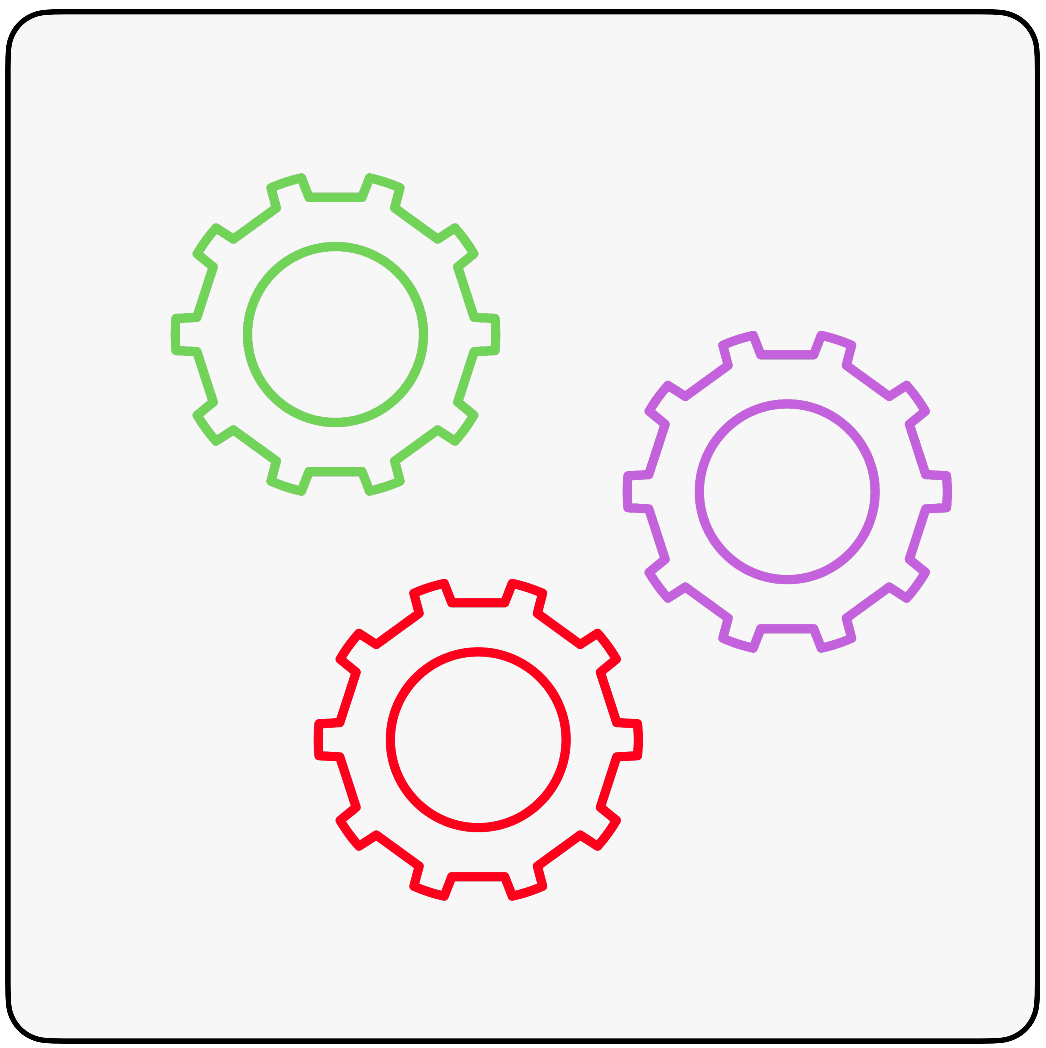
1. Pragmatic inference TMLR 23
2. Moral philosophy & ethics; agency



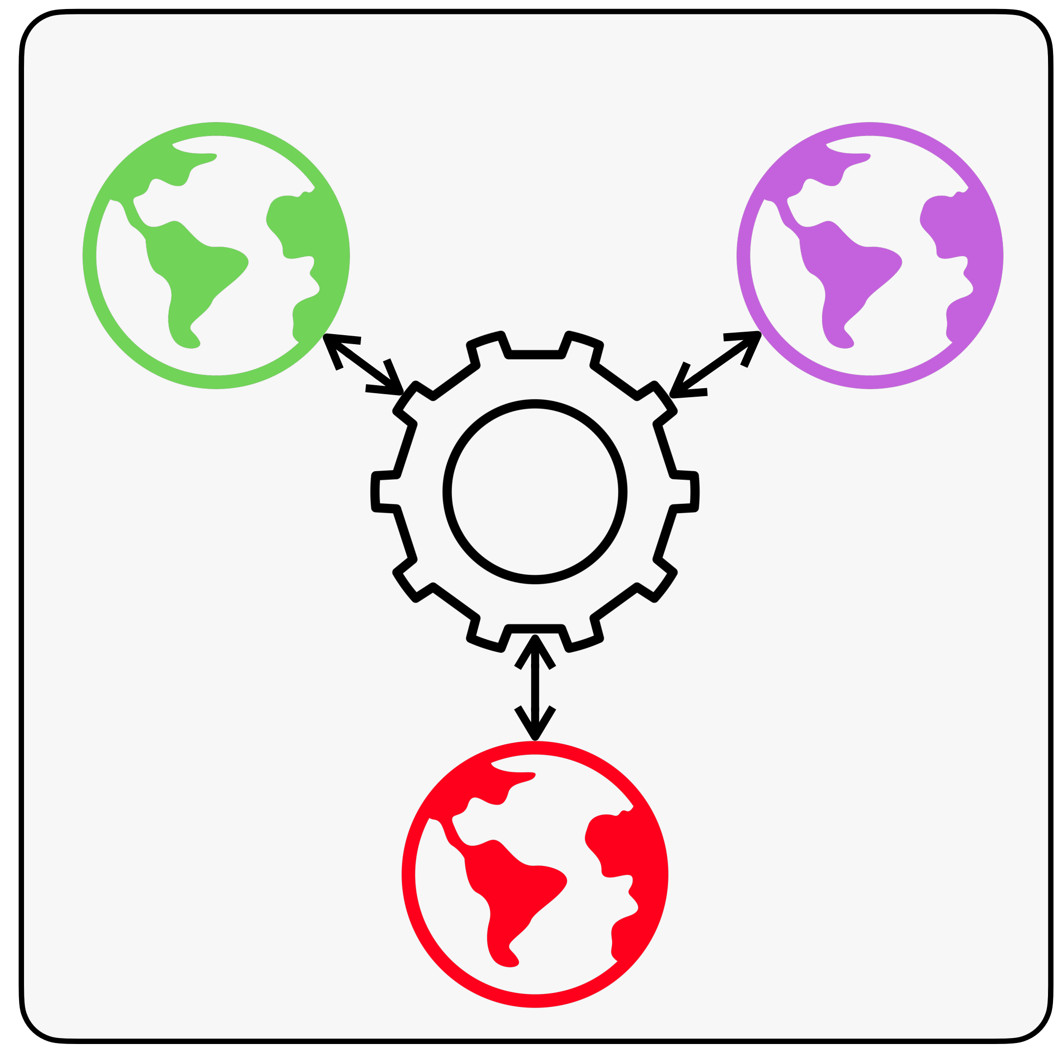
What's next? Imperfect knowledge users

1. Recommendation systems
2. Radiologist support & training ICLR 23

Extrapolate beyond human capabilities?
Supervise process, not outcome



Methods for explanations

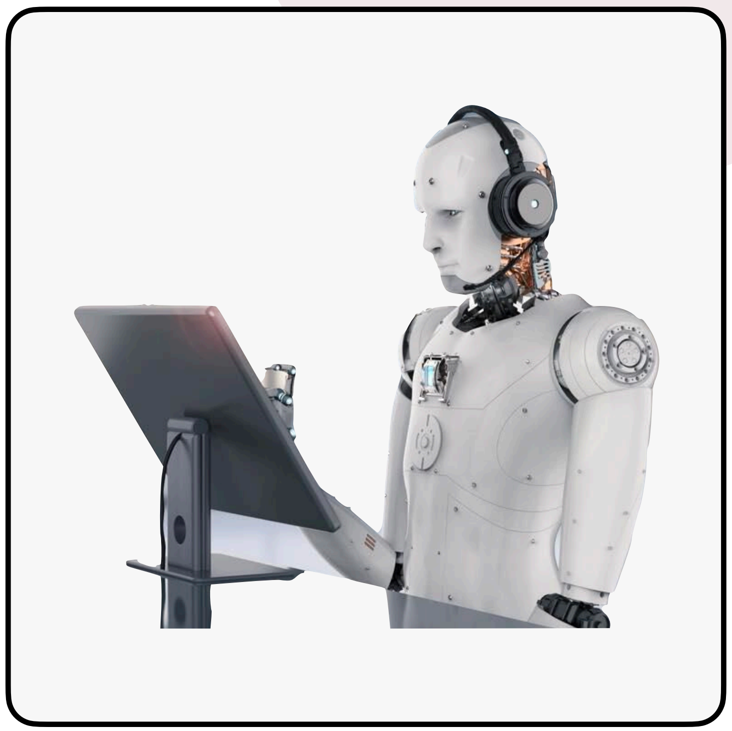


Incentives for explanations

Solution space of two different problems

AIs that can **do a task**

AIs that **help me at the task**



Human Imitators

Self-improving Tools

Truth-finding process *for* and *with* AI

Intelligence Augmentation for experts

AI for science

AI safety
Alignment
AI x-risks

Open Philanthropy

Where we are going,
we don't need ~~roads~~ groundtruths!

Thank you for listening!

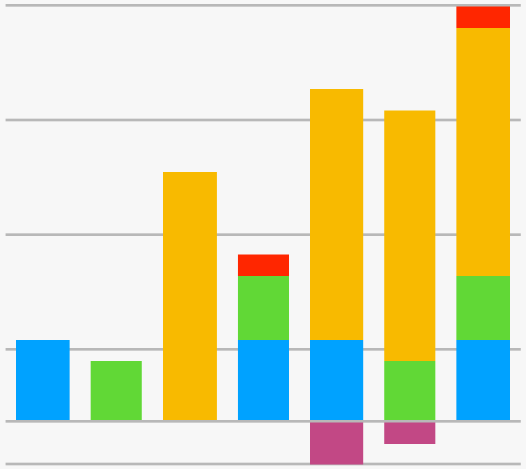
AI eval

"On the Opportunities and Risks of Foundation Models" by OpenAI.

Imitating humans

What color is the flower ?

Learn to explain better



Future work