Towards Human-Centered Explanations of AI Predictions

Chenhao Tan
Chicago Human+AI Lab
University of Chicago
https://chenhaot.com
@ChenhaoTan

Impressive advances in AI

AI 'outperforms' doctors diagnosing breast cancer

Fergus Walsh
Medical correspondent
@BBCFergusWalsh

AI holds promise for improving our society

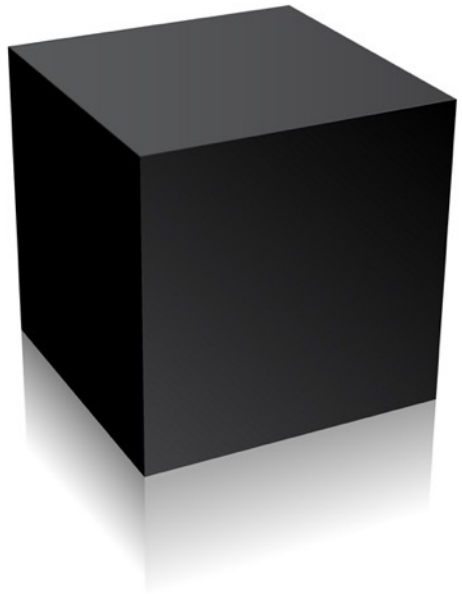
Medical diagnosis

Education

Justice systems

Fake news detection

However, a black-box comes with many issues

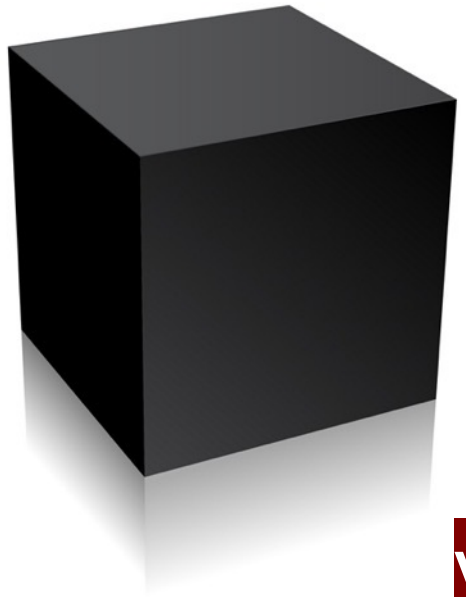


However, a black-box comes with many issues

Bias

Accountability

Understanding



Ethics

Safety

Value-alignment

Explanations hold promise for opening the black box and enabling human-AI interaction

Explanations

Improve people

Improve AI

Explanations are potentially useful for

Stakeholders	Purpose
Model developers	debugging
Decision makers	decision assistance
Decision subjects	decision appealing/improvement
Regulation (e.g., governments)	auditing
Researchers	scientific understanding

What makes effective explanations?

What makes effective explanations?

How to evaluate AI explanations?

Evaluation of AI explanations

Automatic evaluations

Sufficiency

Comprehensiveness

Evaluation of AI explanations

Comparing against human explanations

Utility in supporting decision making

Evaluation of AI explanations

Comparing against human explanations

[Wiergreffe and Marasovic 2021; Camburu et al. 2018; Carton et al. 2018; Khashabi et al. 2018; Zaidan et al. 2007; and many more]

Utility in supporting decision making

[Lai et al. 2021; Carton et al. 2020; Green and Chen 2019; Lai and Tan 2019; Lin et al. 2020; Wang and Yin 2021; and many more]

Evaluation of AI explanations

Comparing against human explanations

Humans can provide "good" explanations (and correct labels)

Utility in supporting decision making

Humans may not necessarily know the correct labels

Evaluation of AI explanations

Emulation

Comparing against human explanations

Humans can provide "good" explanations (and correct labels)

Discovery

Utility in supporting decision making

Humans may not necessarily know the correct labels

Evaluation of AI explanations

Emulation

Comparing against human explanations

Humans can provide "good" explanations (and correct labels)

Conceptually and empirically, humans may not provide "groundtruth" explanations

Discovery

Utility in supporting decision making

Humans may not necessarily know the correct labels

Evaluation of AI explanations

Emulation

Comparing against human explanations

Humans can provide "good" explanations (and correct labels)

Conceptually and empirically, humans may not provide "groundtruth" explanations

Discovery

Utility in supporting decision making

Humans may not necessarily know the correct labels

Human+AI rarely outperforms AI
Decision-focused summarization

Evaluation of AI explanations

Emulation

Comparing against human explanations

Humans can provide "good" explanations (and correct labels)

Conceptually and empirically, humans may not provide "groundtruth" explanations

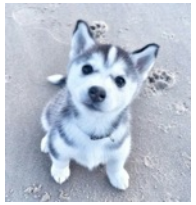
Discovery

Utility in supporting decision making

Humans may not necessarily know the correct labels

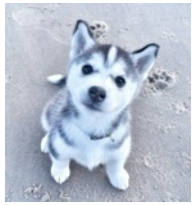
Human+AI rarely outperforms AI
Decision-focused summarization

Two modes of AI



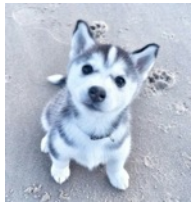
Emulation ⟷ Discovery

 = Data + Model



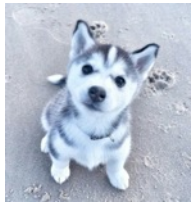
Emulation ⟷ Discovery

= Data + Model



Emulation ⟷ Discovery

= Data + Model

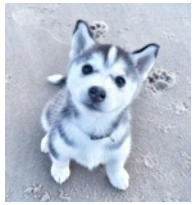


Emulation ⟷ Discovery

Sentiment analysis

Terrible. Just terrible. Terrible customer service. Terrible in every way possible. I absolutely hate receiving a package by DHL. Both times I was forced to have to pick up my package because they are too incompetent to deliver it properly at my house. Wish I could give negative stars.

 = Data + Model

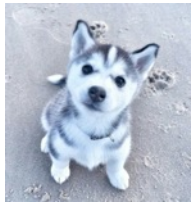


Emulation ⟷ Discovery

Sentiment analysis Label: negative

Terrible. Just terrible. Terrible customer service. Terrible in every way possible. I absolutely hate receiving a package by DHL. Both times I was forced to have to pick up my package because they are too incompetent to deliver it properly at my house. Wish I could give negative stars.

 = Data + Model

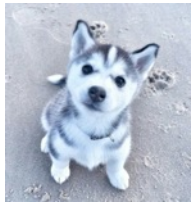


Emulation ⟷ Discovery

“ [human intelligence] can be so precisely described that a machine can be made to simulate it ”

Dartmouth Summer Research
Project on Artificial Intelligence

 = Data + Model

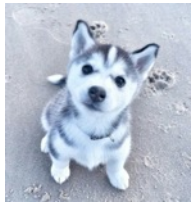


Emulation ⟷ Discovery

Deception detection

My stay at the Talbott was a wonderful experience. The service at this upscale hotel was beyond my expectations, the Gold Coast location is close to Michigan Ave, the museums, and many of the other sites Chicago has to offer. If you are visiting Chicago, I highly recommend the Talbott!

 = Data + Model

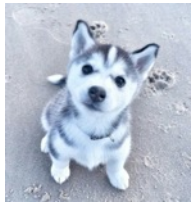


Emulation ⟷ Discovery

Deception detection | Label: deceptive

My stay at the Talbott was a wonderful experience. The service at this upscale hotel was beyond my expectations, the Gold Coast location is close to Michigan Ave, the museums, and many of the other sites Chicago has to offer. If you are visiting Chicago, I highly recommend the Talbott!

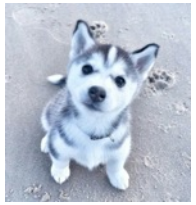
 = Data + Model



Emulation ⟷ Discovery

Labels can come
from **crowdsourcing**

 = Data + Model

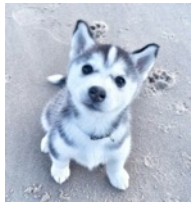


Emulation ⟷ Discovery

Labels can come from **crowdsourcing**

Labels come from observing **social processes**

Many high-stake decisions are discovery tasks

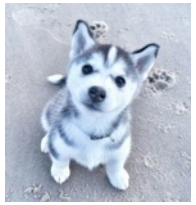


Emulation ⟷ Discovery

Labels can come from **crowdsourcing**

Labels come from observing **social processes**

Implications on explanations



Emulation ⟷ Discovery

Sentiment analysis **Label: negative**

Terrible. Just terrible. Terrible customer service. Terrible in every way possible. I absolutely hate receiving a package by DHL. Both times I was forced to have to pick up my package because they are too incompetent to deliver it properly at my house. Wish I could give negative stars.

Deception detection **Label: deceptive**

My stay at the Talbott was a wonderful experience. The service at this upscale hotel was beyond my expectations, the Gold Coast location is close to Michigan Ave, the museums, and many of the other sites Chicago has to offer. If you are visiting Chicago, I highly recommend the Talbott!

Evaluation of AI explanations

Emulation

Comparing against human explanations

Humans can provide "good" explanations (and correct labels)

Conceptually and empirically, humans may not provide "groundtruth" explanations

Discovery

Utility in supporting decision making

Humans may not necessarily know the correct labels

Human+AI rarely outperforms AI
Decision-focused summarization

Evaluation of AI explanations

Emulation

Comparing against human explanations

Humans can provide "good" explanations (and correct labels)

Conceptually and empirically, humans may not provide "groundtruth" explanations

Discovery

Utility in supporting decision making

Humans may not necessarily know the correct labels

Human+AI rarely outperforms AI
Decision-focused summarization

Psychological evidence that suggests human may not be able to explain

Preprint 2021

On the Diversity and Limits of Human Explanations

Chenhao Tan

Psychological evidence that suggests human may not be able to explain

Wilson and Keil 1998, The shadows and shallows of explanation

- Prediction
- Understanding
- Theories

These three notions "form a progression of increasing sophistication and depth with explanations falling between understanding and theories".

We can predict that a car will start when we turn the ignition switch, but few of us are able to explain in detail why this is so.

Psychological evidence that suggests human may not be able to explain

Nisbett and Wilson 1977, Telling more than we can know: verbal reports on mental processes

Our verbal reports on our mental processes are highly inaccurate.

Legitimate information can be used to justify preferences based on illegitimate factors such as race.

[Norton et al. 2006]

Psychological evidence that suggests human may not be able to explain

- Human explanations are necessarily incomplete
  - We do not start from a set of axioms and present all the deductive steps

[Keil 2006; Lombrozo 2006]

Premise: Men in green hats appear to be attending a gay pride festival.
Hypothesis: Men are attending a festival.
Explanation: The men are attending the festival.

[Camburu et al. 2018]

Empirical characterization of human rationales

EMNLP 2020

Evaluating and Characterizing Human Rationales

Samuel Carton, Anirudh Rathore, and Chenhao Tan

Empirical characterization of human rationales

- Sufficiency: rationales alone allow for inferring the label
- Comprehensiveness (necessity): rationales are required to infer the label

Empirical characterization of human rationales

- Sufficiency: rationales alone allow for inferring the label
- Comprehensiveness (necessity): rationales are required to infer the label

Are human rationales sufficient or comprehensive?

Example human rationales: Fact verification

FEVER

Label: supports

No Way Out is the debut studio album by American hip hop recording artist , songwriter and record producer Puff Daddy . It was released on July 1 , 1997 , by his Bad Boy record label . The label 's official crediting as `` The Family '' , featuring guest appearances from his label-mates and other artists .The production on the album was provided by Puff Daddy [real name Sean Combs] , alongside with a variety of the members from the production group , called The Hitmen . …… [SEP] 1997 was the year No Way Out was released.

Example human rationales: Fact verification

FEVER

Label: supports

It was

released on July 1 , 1997 , by his Bad Boy record label .

[SEP] 1997 was the year No Way Out was released.

Human rationales may not be sufficient

Example human rationales: Toxicity detection

WikiAttack

Label: personal-attack

Example human rationales: Toxicity detection

The next page contains content that maybe offensive or upsetting

Example human rationales: Toxicity detection

WikiAttack

Label: personal-attack

== What the FUCK is your problem, bitch!!!!!!!!!!! ==

Why the FUCK did you delete the Dreamtime Festival page, shithead. Some folks are actually interested in things like that, bitch. Why don't you do yourself and the world a favor and stick your head up your ass and take a big whiff. Guess what? Your shit stank, like everyone else, you self-righteous fuck-sissy!!!!!!!!!

Example human rationales: Toxicity detection

WikiAttack

Label: personal-attack

== What the FUCK is your problem, bitch!!!!!!!!!!! ==

Why the FUCK did you delete the Dreamtime Festival page, shithead. Some folks are actually interested in things like that, bitch. Why don't you do yourself and the world a favor and stick your head up your ass and take a big whiff. Guess what? Your shit stank, like everyone else, you self-righteous fuck-sissy!!!!!!!!!

Human rationales may not be comprehensive

Example human rationales: Toxicity detection

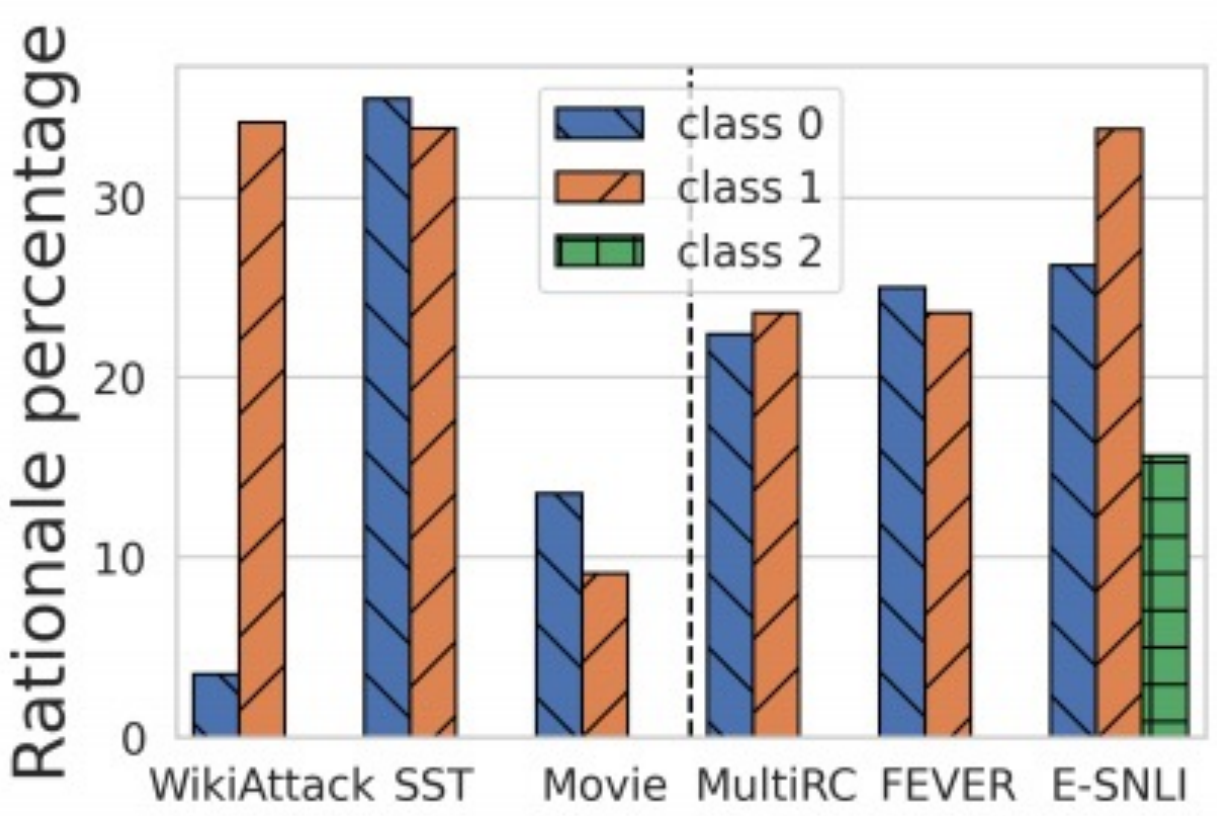
WikiAttack

Label: no attack

Makes sense. Have a good one.

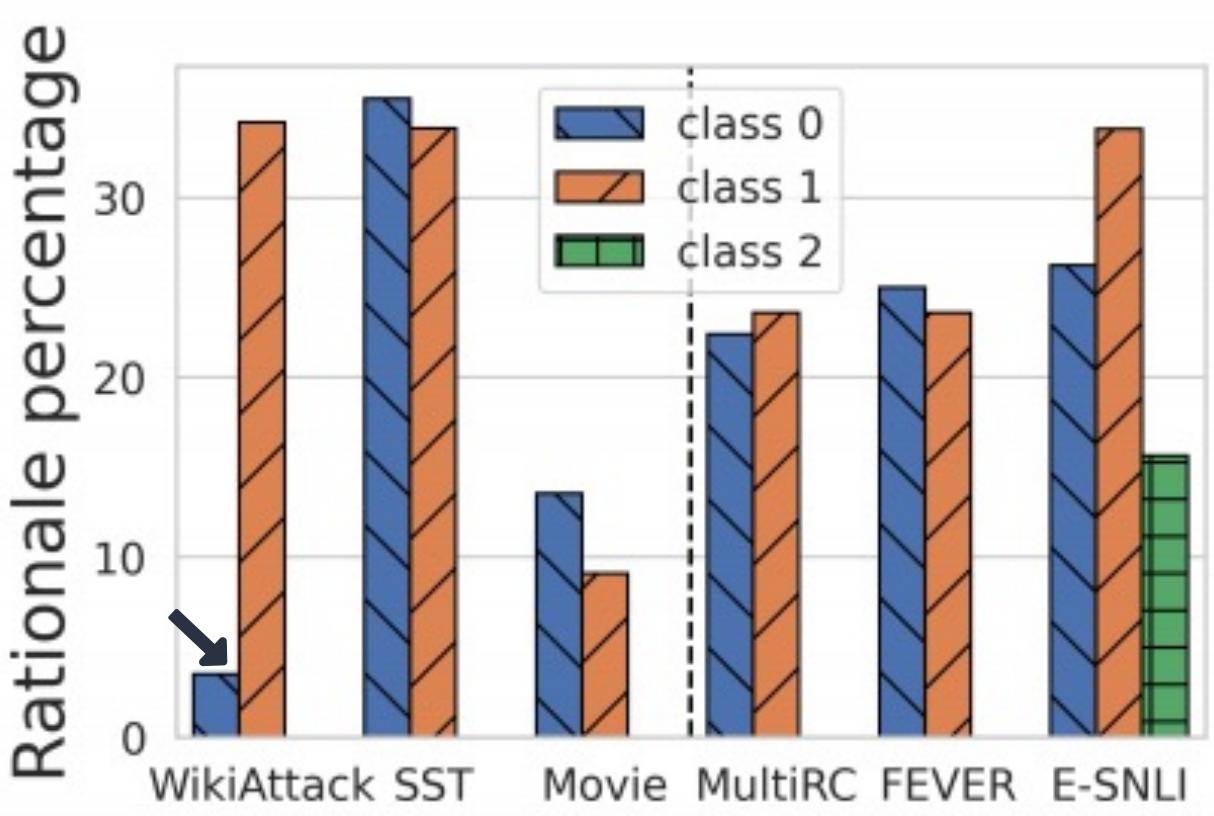
The explanation derives from the lack of evidence

Substantial variations exists across classes and datasets



WikiAttack	0: no-attack, 1: personal-attack
SST	0: negative, 1: positive
Movie	0: negative, 1: positive
MultiRC	0: false, 1: true
FEVER	0: refutes, 1: supports
E-SNLI	0: contradiction, 1: entailment, 2: neutral

Substantial variations exists across classes and datasets



WikiAttack	0: no-attack, 1: personal-attack
SST	0: negative, 1: positive
Movie	0: negative, 1: positive
MultiRC	0: false, 1: true
FEVER	0: refutes, 1: supports
E-SNLI	0: contradiction, 1: entailment, 2: neutral

Automatic Fidelity Metrics

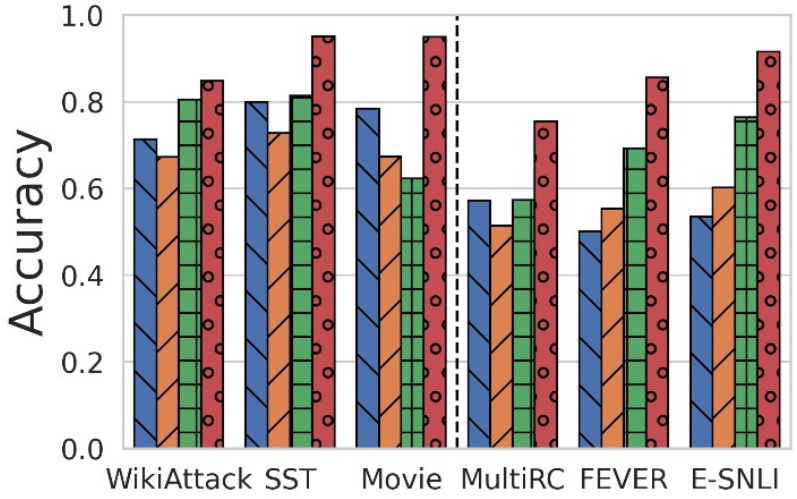
- Compare class probabilities with full information vs. rationale or complement
- Sufficiency
  - Is the rationale sufficient to make a similar prediction?

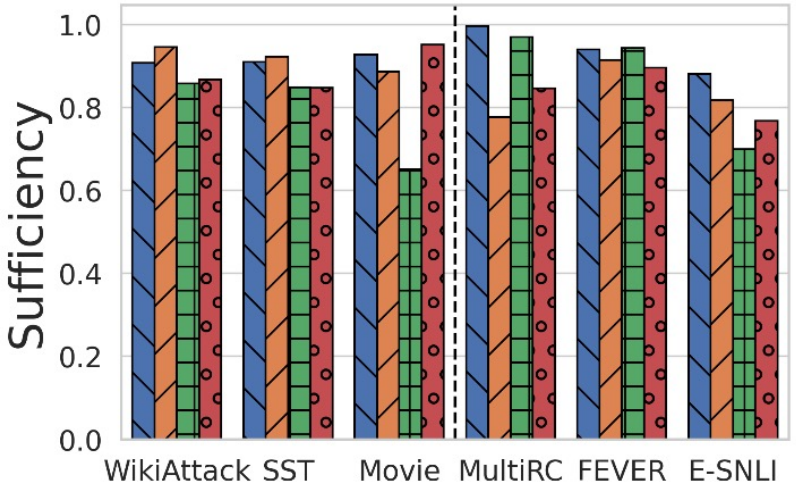
$$\text{Suff}(\boldsymbol{x}, \hat{y}, \boldsymbol{\alpha}) = 1 - \max(0, p(\hat{y}|\boldsymbol{x}) - p(\hat{y}|\boldsymbol{x}, \boldsymbol{\alpha}))$$

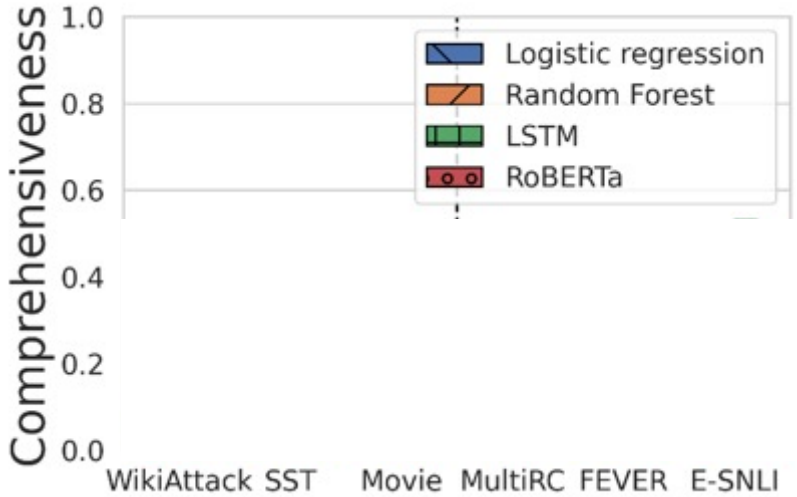
- Comprehensiveness
  - Is the rationale necessary to make a similar prediction?

$$\text{Comp}(\boldsymbol{x}, \hat{y}, \boldsymbol{\alpha}) = \max(0, p(\hat{y}|\boldsymbol{x}) - p(\hat{y}|\boldsymbol{x}, \mathbf{1} - \boldsymbol{\alpha}))$$

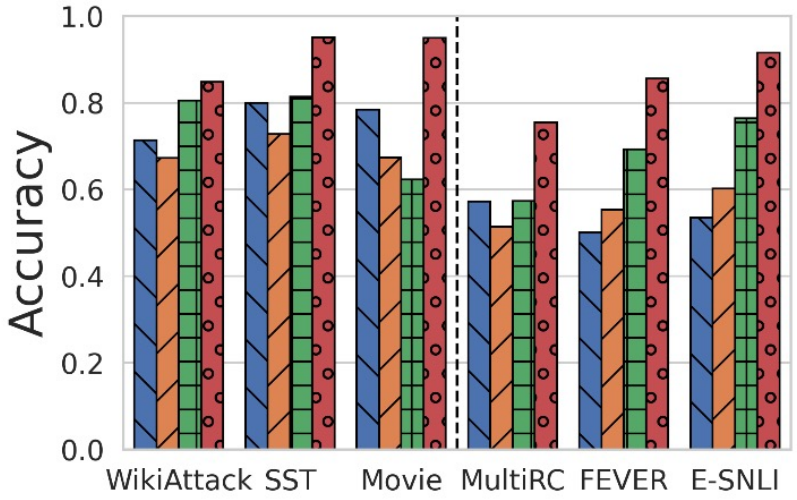
Fidelity of Human Rationales

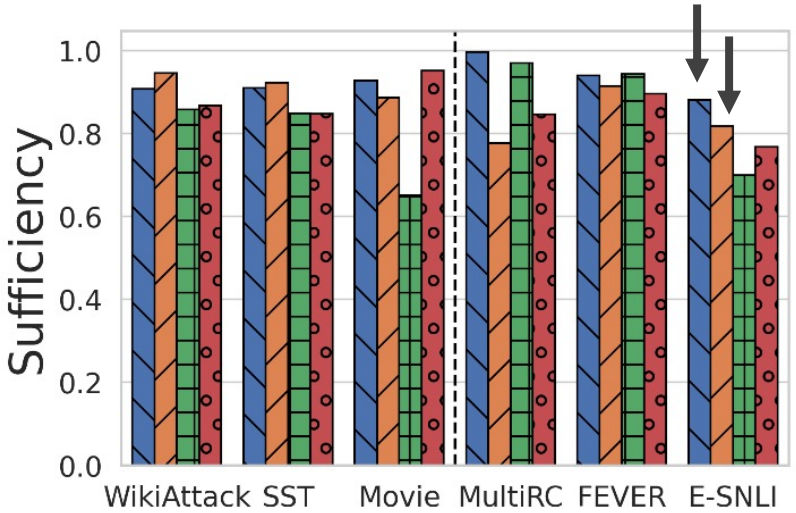


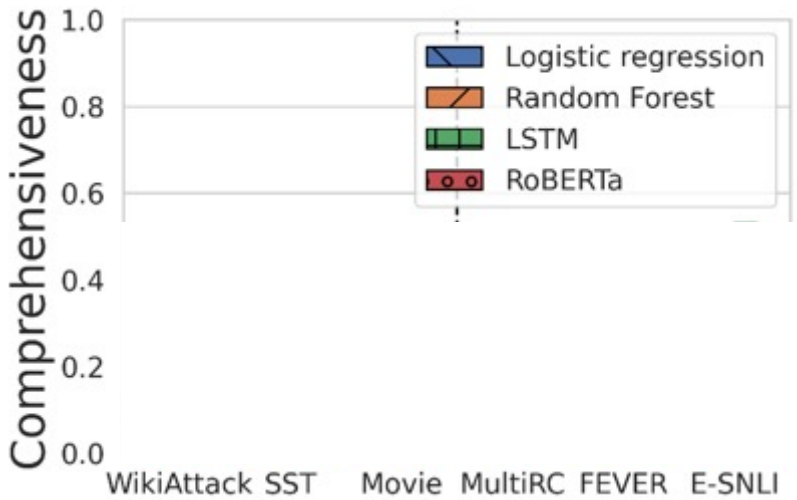




Fidelity of Human Rationales

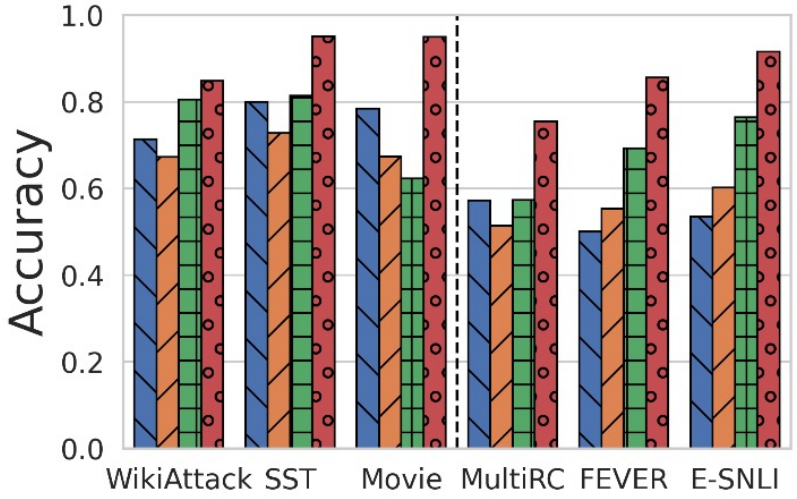


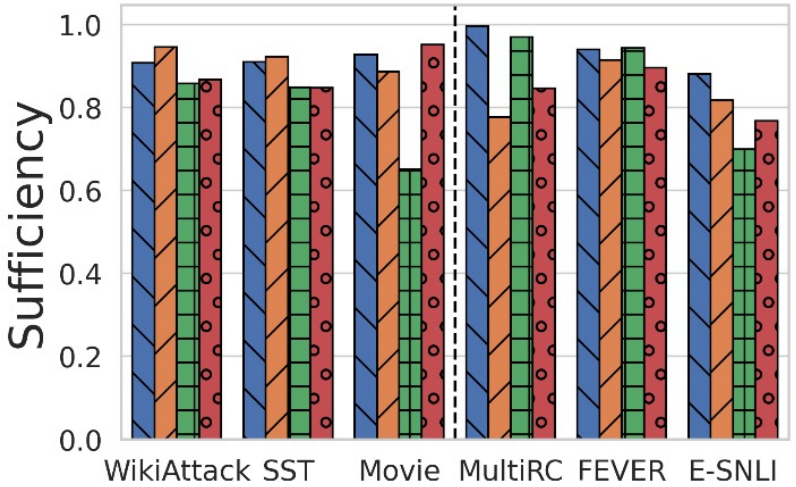


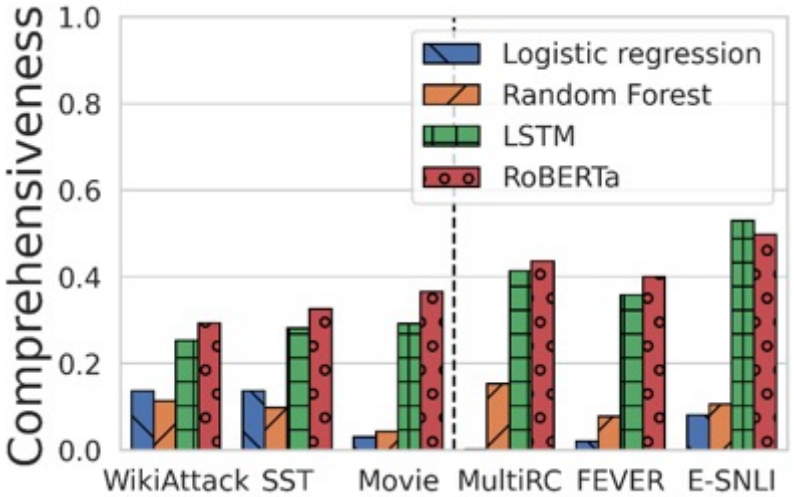


Low-accuracy models demonstrate high sufficiency

Fidelity of Human Rationales



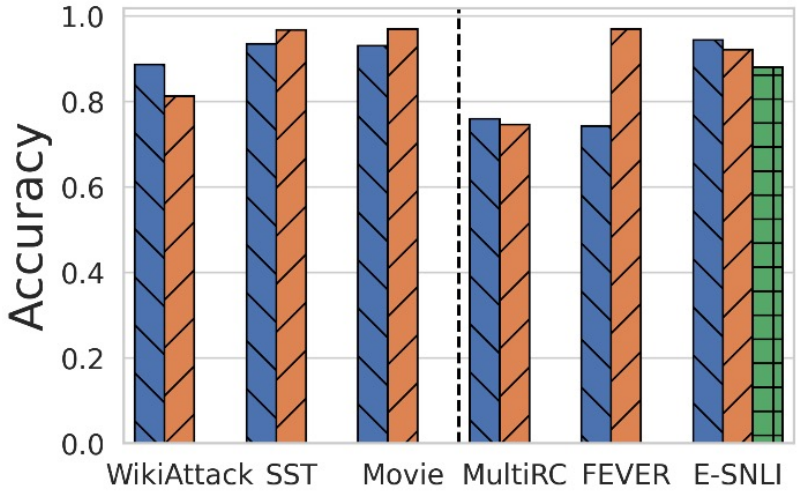


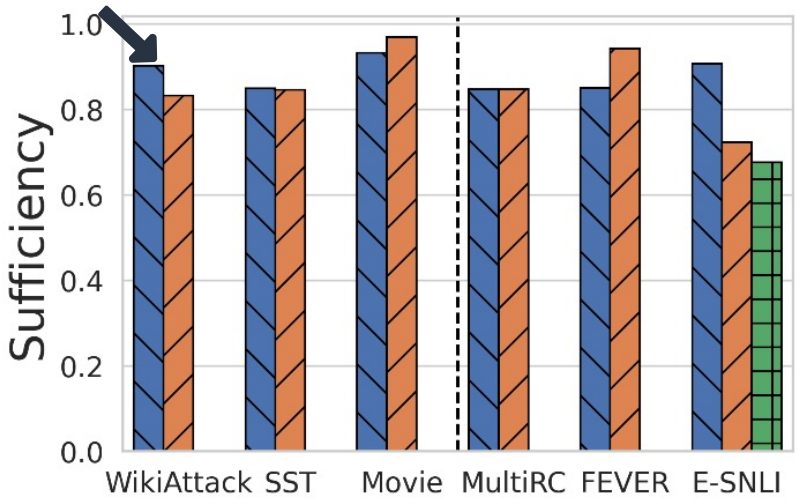


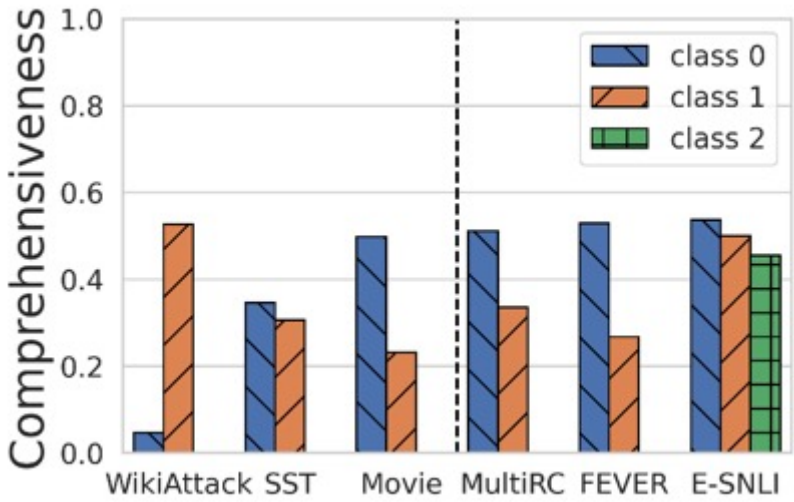
Low-accuracy models demonstrate high sufficiency

Comprehensiveness on a different scale from sufficiency

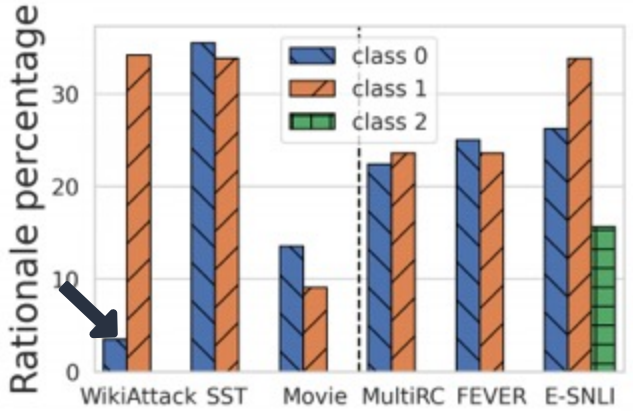
Class Asymmetry







? Real, or artifact of model bias?



WikiAttack	0: no-attack, 1: personal-attack
SST	0: negative, 1: positive
Movie	0: negative, 1: positive
MultiRC	0: false, 1: true
FEVER	0: refutes, 1: supports
E-SNLI	0: contradiction, 1: entailment, 2: neutral

Model bias affects the fidelity metrics

$$\mathrm{Suff}(\boldsymbol{x}, \hat{y}, \boldsymbol{\alpha}) = 1 - \max(0, p(\hat{y}|\boldsymbol{x}) - p(\hat{y}|\boldsymbol{x}, \boldsymbol{\alpha}))$$

Model bias affects the fidelity metrics

$$\mathrm{Suff}(\boldsymbol{x}, \hat{y}, \boldsymbol{\alpha}) = 1 - \max(0, p(\hat{y}|\boldsymbol{x}) - p(\hat{y}|\boldsymbol{x}, \boldsymbol{\alpha}))$$

Imagine a model that predicts 1 for all instances

Sufficiency is trivially 1

Model bias affects the fidelity metrics

$$\mathrm{Suff}(\boldsymbol{x}, \hat{y}, \boldsymbol{\alpha}) = 1 - \max(0, p(\hat{y}|\boldsymbol{x}) - p(\hat{y}|\boldsymbol{x}, \boldsymbol{\alpha}))$$

Imagine a model that predicts C for all instances

Sufficiency is trivially 1

Normalized fidelity

- Idea: what if we normalize fidelity scores relative to baseline model behavior?
  - Namely, how sufficient is this rationale compared to the sufficiency of an empty rationale?
- Null difference
  - Sufficiency of an empty rationale
  - Comprehensiveness of an all-inclusive rationale
  - Determined by class balance and model bias
- Normalize sufficiency and comprehensiveness using min-max scaling

Normalized fidelity

- Null difference
  - Output difference vs. empty rationale

$$\text{NullDiff}(\boldsymbol{x}, \hat{y}) = \max(0, p(\hat{y}|\boldsymbol{x}) - p(\hat{y}|\boldsymbol{x}, \mathbf{0}))$$

  - Equivalent to $1 - \text{Suff}(\boldsymbol{x}, \hat{y}, \mathbf{0})$ or $\text{Comp}(\boldsymbol{x}, \hat{y}, \mathbf{1})$

- Normalized Sufficiency

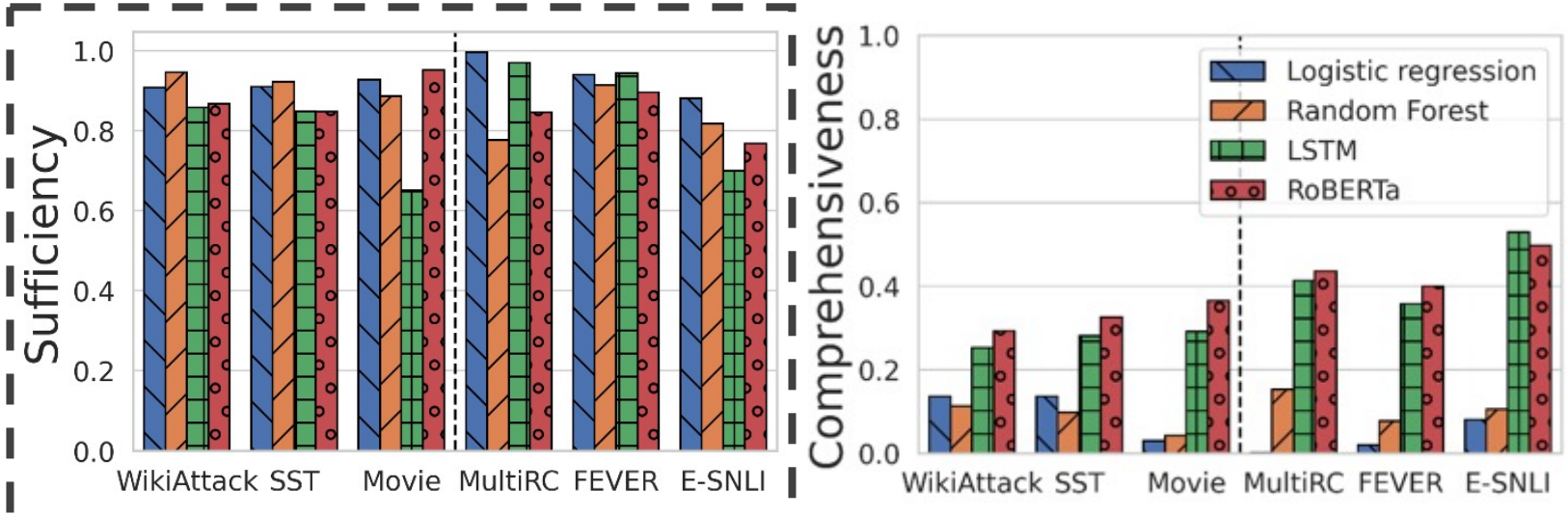
$$\text{NormSuff}(\boldsymbol{x}, \hat{y}, \boldsymbol{\alpha}) = \frac{\text{Suff}(\boldsymbol{x}, \hat{y}, \boldsymbol{\alpha}) - \text{Suff}(\boldsymbol{x}, \hat{y}, \mathbf{0})}{1 - \text{Suff}(\boldsymbol{x}, \hat{y}, \mathbf{0})}$$

- Normalized Comprehensiveness

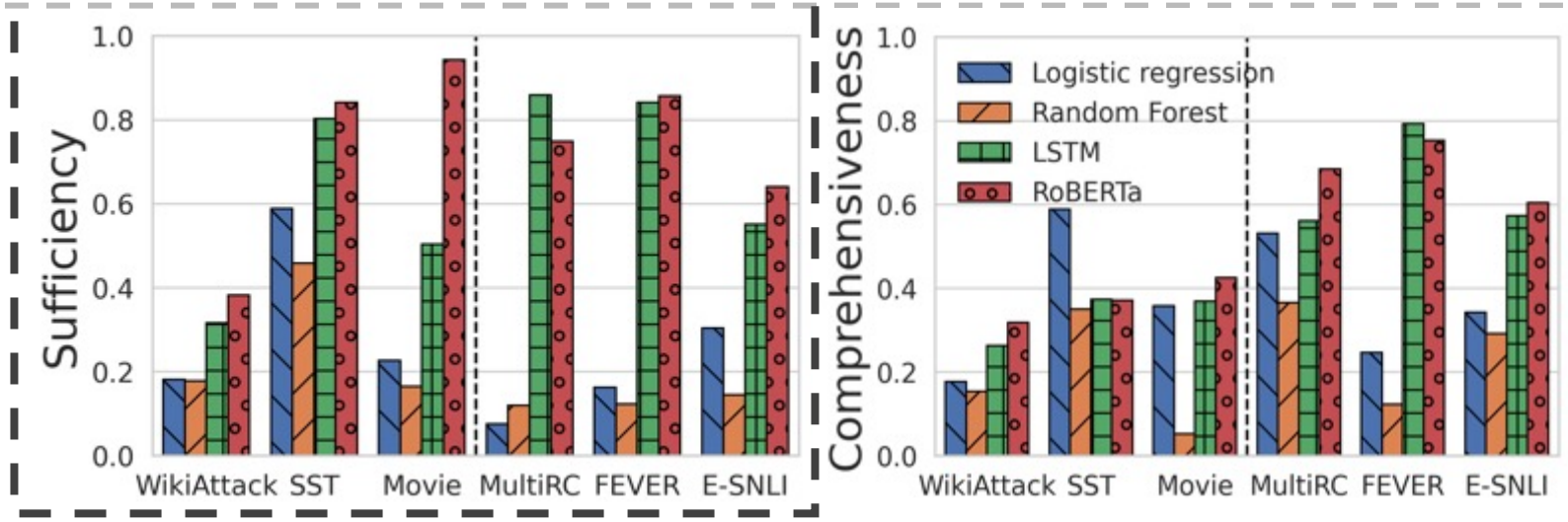
$$\text{NormComp}(\boldsymbol{x}, \hat{y}, \boldsymbol{\alpha}) = \frac{\text{Comp}(\boldsymbol{x}, \hat{y}, \boldsymbol{\alpha})}{\text{Comp}(\boldsymbol{x}, \hat{y}, \mathbf{1})}$$

Simple models are no longer with high sufficiency

Non-normalized

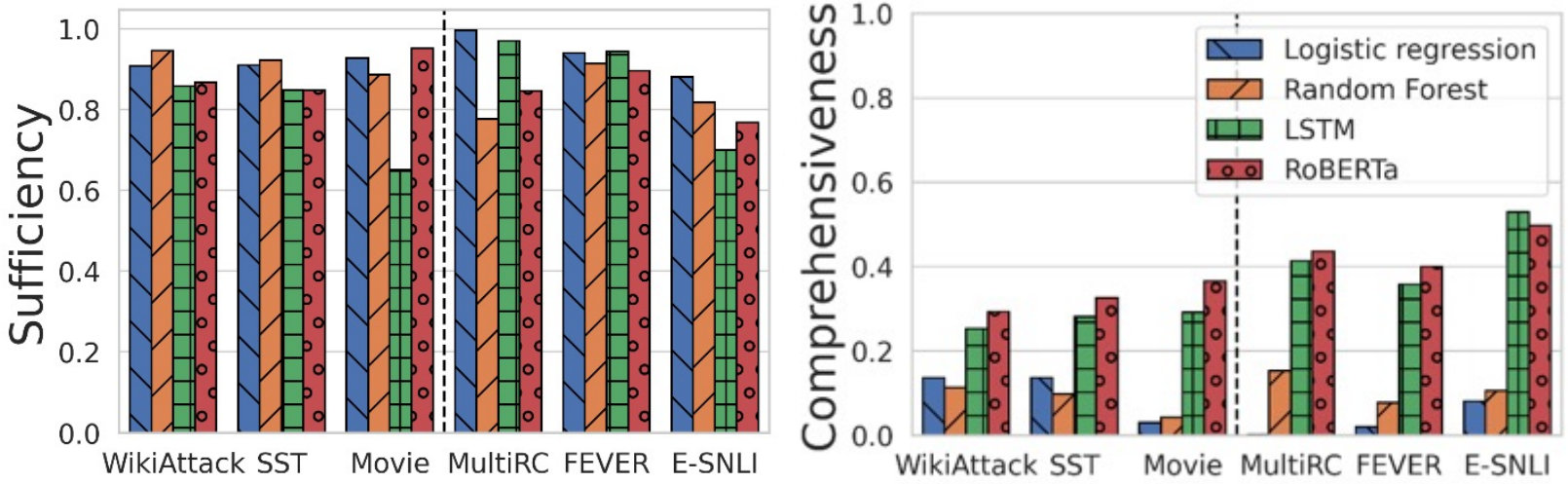


Normalized

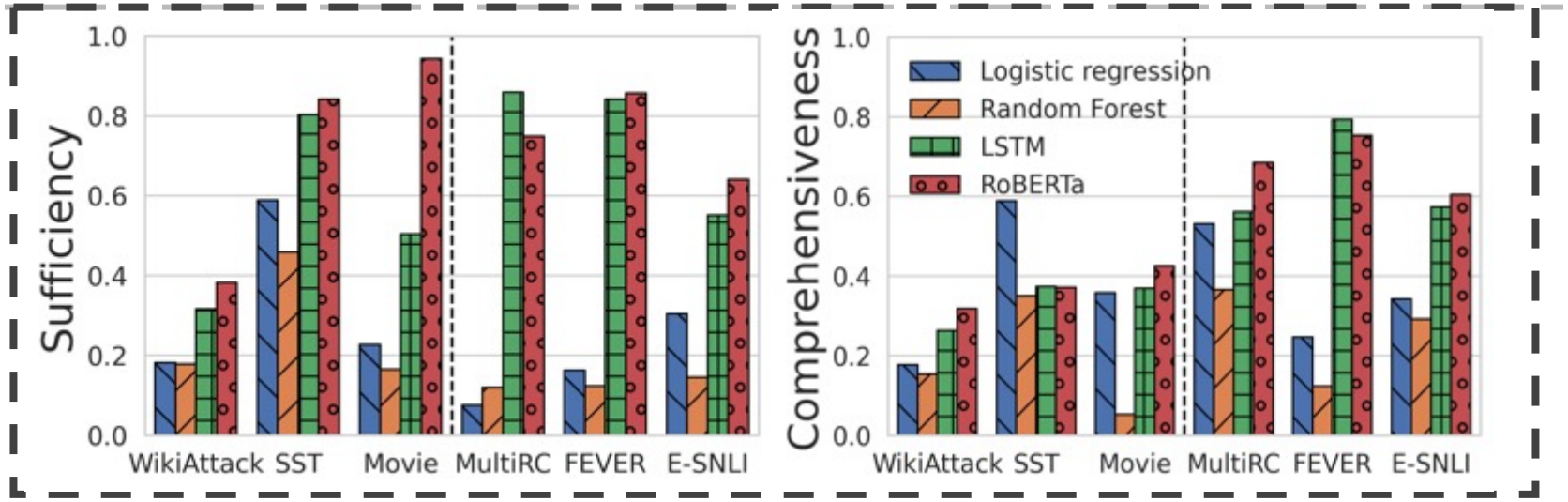


Sufficiency is slightly greater than comprehensiveness

Non-normalized

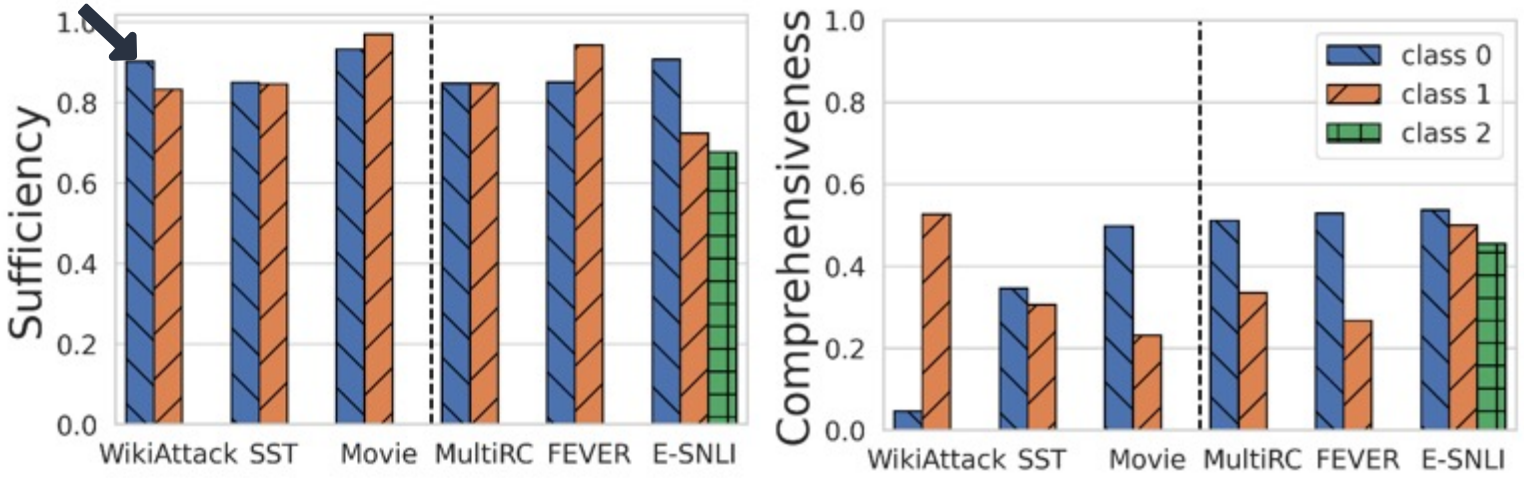


Normalized

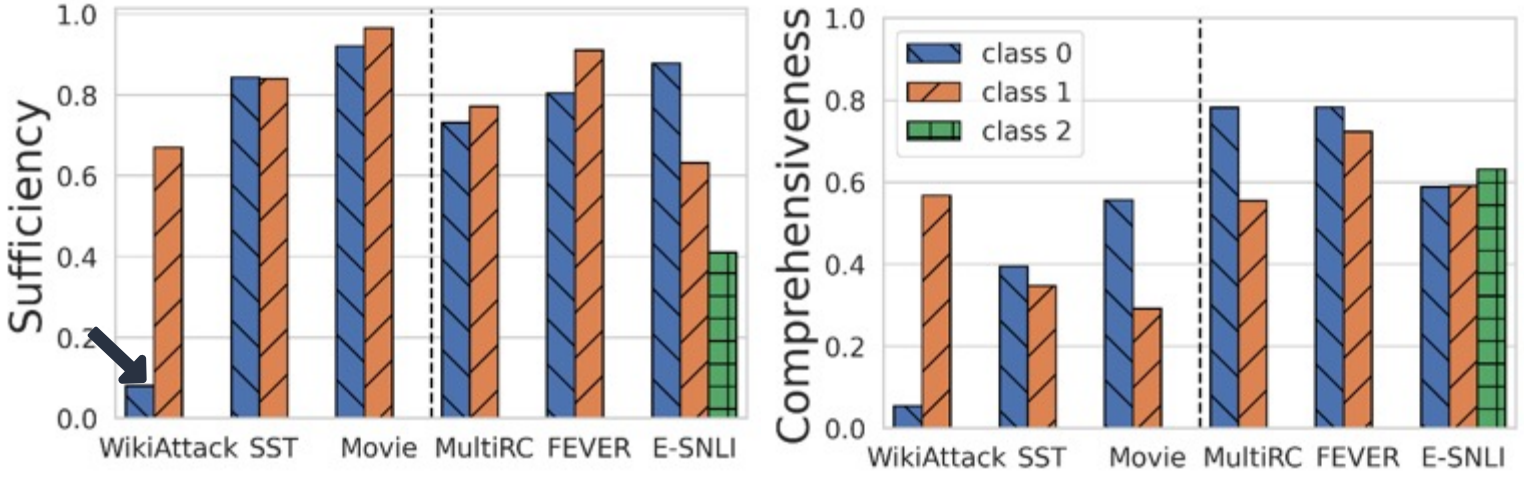


WikiAttack: asymmetry in sufficiency is due to model bias

Non-normalized

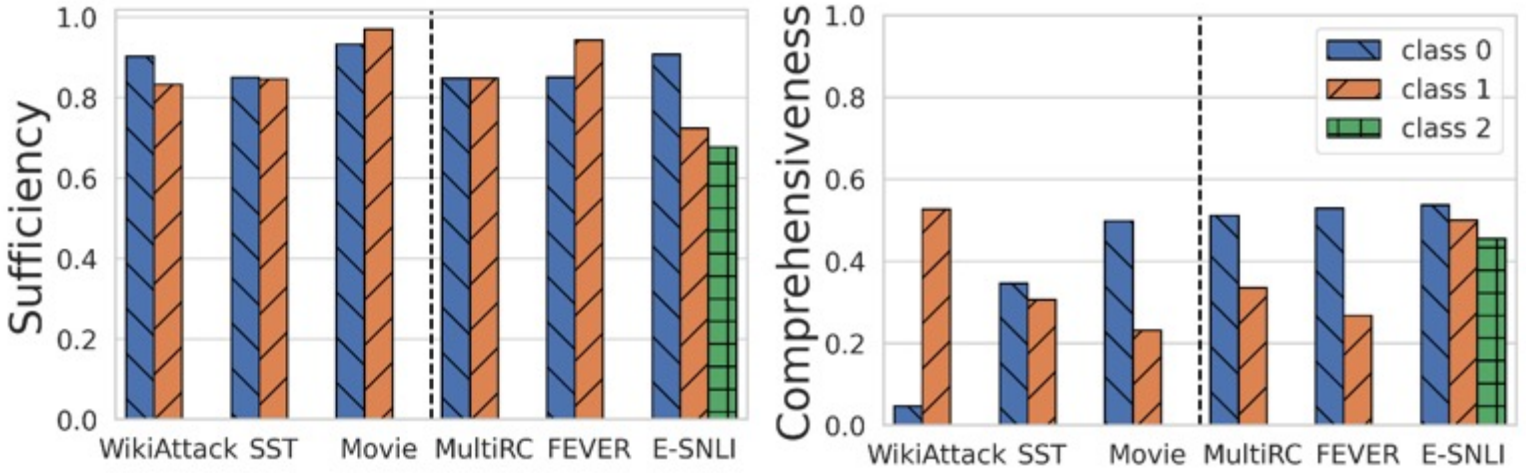


Normalized

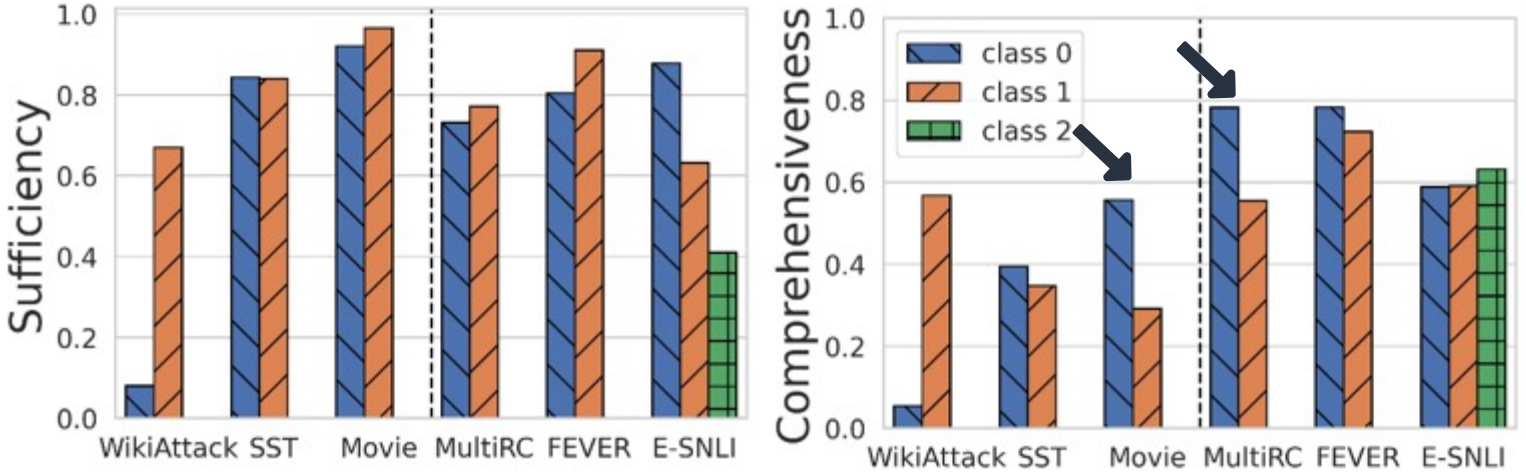


Comprehensiveness asymmetry stays in Movie and MultiRC

Non-normalized



Normalized



Empirical characterization of human rationales

- Human rationales are neither sufficient nor comprehensive (may be more sufficient than comprehensive)
- Substantial variance exists across datasets and classes in the same dataset

Collecting human rationales is hard

Zaidan, Eisner, and Paiatko 2007

To justify why a review is positive, highlight the most important words and phrases that would tell someone to see the movie. To justify why a review is negative, highlight words and phrases that would tell someone not to see the movie.

Sen et al. 2020

Label the sentiment and highlight ALL words that reflect this sentiment.

[Tan 2021]

Learning from human rationales also requires special care

- Rationale-only performance establishes an upper bound of possible improvements in performance
  - Spoiler alert: it can be very low in some datasets, for example, E-SNLI
- Recall matters more than precision
  - Many more tricks to incorporate human rationales

What to Learn, and How: Toward Effective Learning from Rationales
Samuel Carton, Surya Kanoria, and Chenhao Tan
Findings of ACL 2022

Summary

- Human rationales are not necessarily valid groundtruth
- One-size does not fit all, and we need to understand how to collect human rationales before chasing the leaderboard

Evaluation of AI explanations

Emulation

Comparing against human explanations

Humans can provide "good" explanations (and correct labels)

Conceptually and empirically, humans may not provide "groundtruth" explanations

Discovery

Utility in supporting decision making

Humans may not necessarily know the correct labels

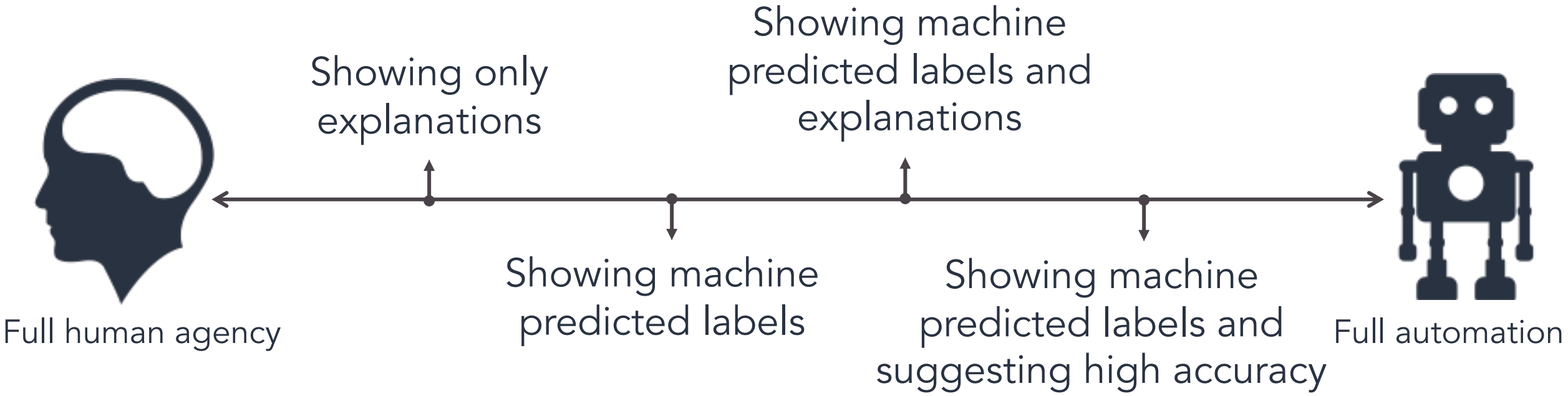
Human+AI rarely outperforms AI
Decision-focused summarization

FAccT 2019

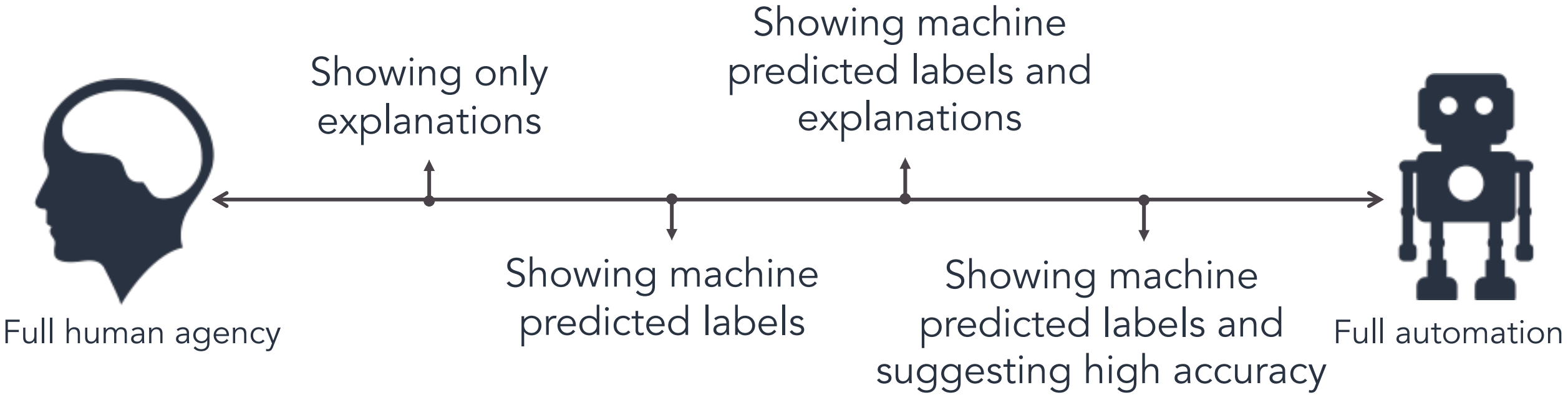
On Human Predictions with Explanations and Predictions of Machine Learning Models: A Case Study on Deception Detection

Vivian Lai and Chenhao Tan

A spectrum between full human agency & full automation



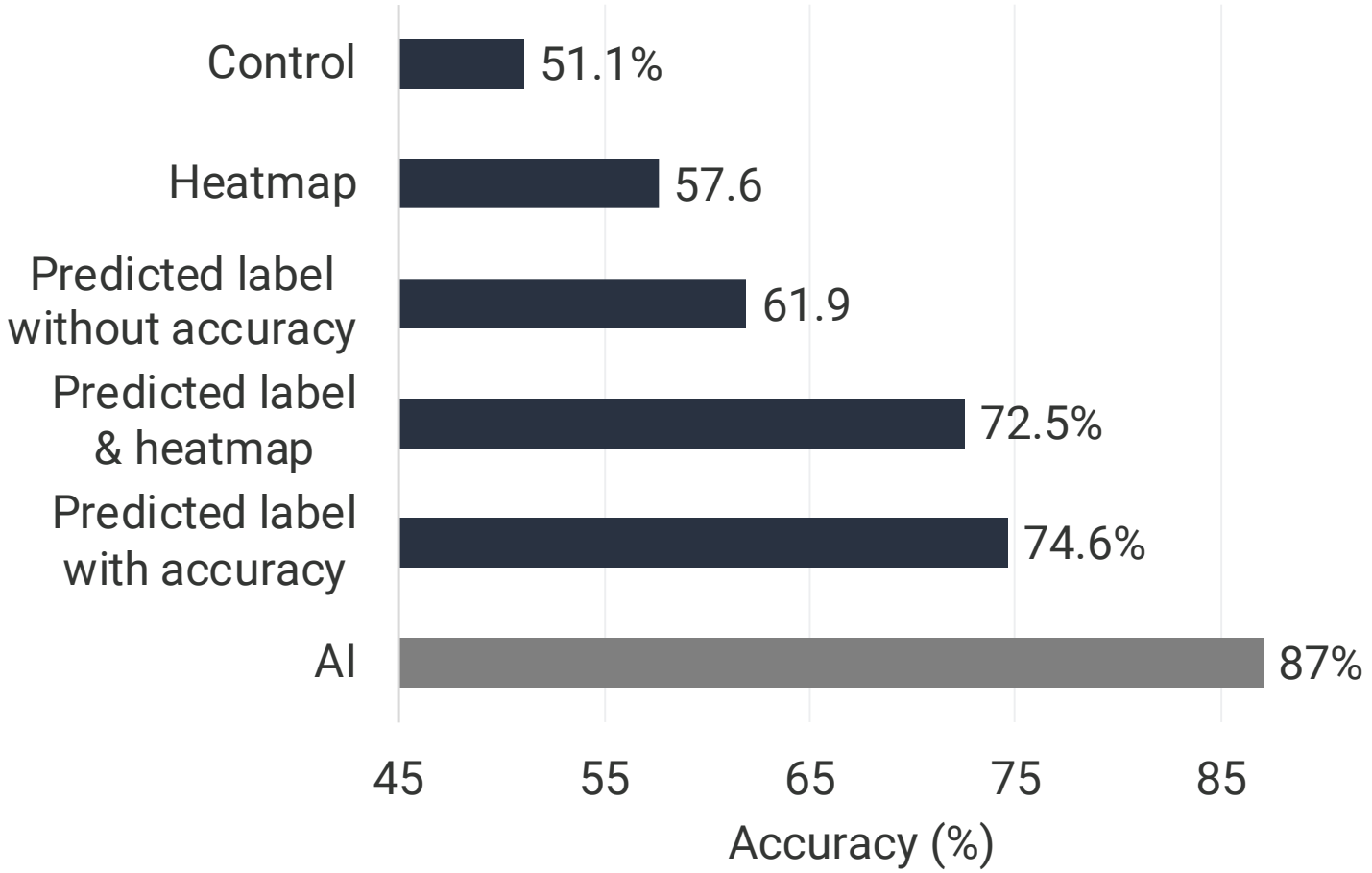
A spectrum between full human agency & full automation



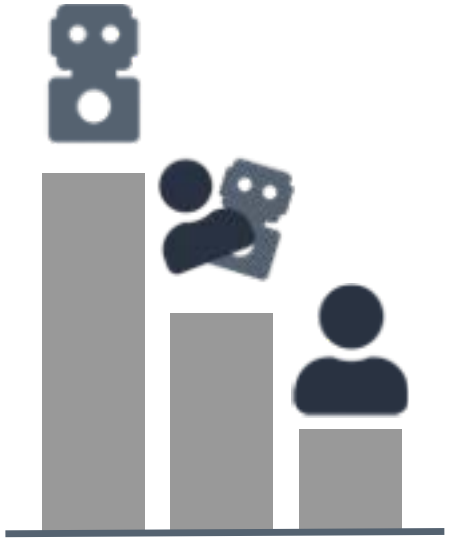
The amount of information from the machine generally increases as we move from the left to the right.

https://machineintheloop.com/

AI assistance does improve human performance, but Human+AI < AI



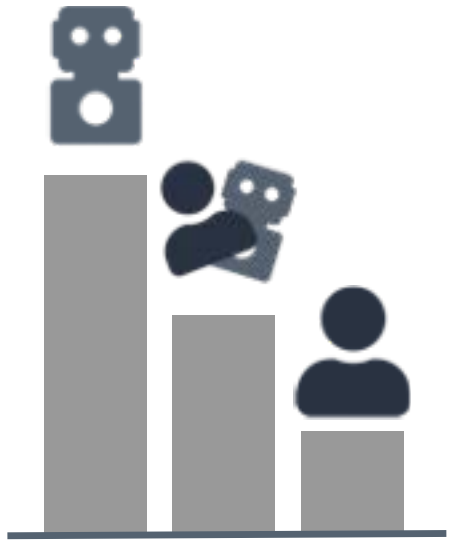
✗ Complementary
Human + AI > Human / AI



[Lai et al. 2021; Carton et al. 2020; Green and Chen 2019; Lai and Tan 2019; Lin et al. 2020; Wang and Yin 2021; and many more]

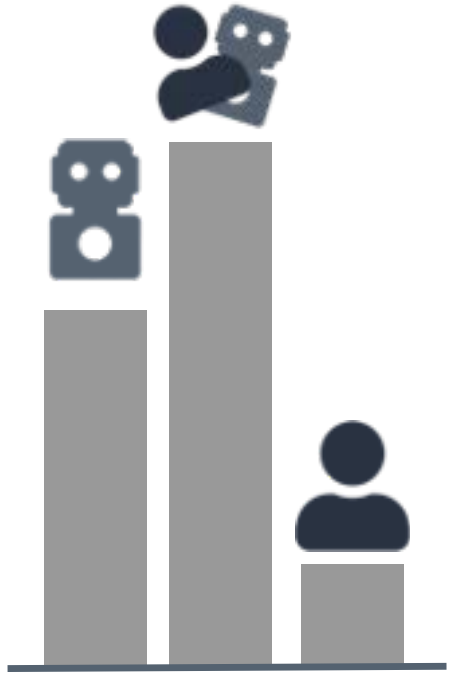
✗ Complementary

Human + AI > Human / AI



✓ Complementary

Human + AI > Human / AI



[Lai et al. 2021; Carton et al. 2020; Green and Chen 2019; Lai and Tan 2019; Lin et al. 2020; Wang and Yin 2021; and many more]

Towards complementary performance

Real-time static explanations are not sufficient

My stay at the Talbott was a wonderful experience. The service at this upscale hotel was beyond my expectations, the Gold Coast location is close to Michigan Ave, the museums, and many of the other sites Chicago has to offer. If you are visiting Chicago, I highly recommend the Talbott!

[Li et al. 2014; Ott et al. 2011; Ott et al. 2013]

Towards complementary performance

Real-time static explanations are not sufficient

Human strengths
AI strengths

Combining human strengths and AI strengths

Tasks

Future rating prediction [EMNLP 21]

Content moderation [CHI 22]

Profession prediction [CSCW 21]

Recidivism prediction [CSCW 21]

AI assistance

Decision-focused summarization [EMNLP 21]

Conditional delegation [CHI 22]

AI-driven tutorials [CHI 20]

Interactive explanations (in natural language) [CSCW 21; preprint]

Experiment design

Out-of-distribution [CSCW 21]

Combining human strengths and AI strengths

Tasks

Future rating prediction [EMNLP 21]

Content moderation [CHI 22]

Profession prediction [CSCW 21]

Recidivism prediction [CSCW 21]

AI assistance

Decision-focused summarization [EMNLP 21]

Conditional delegation [CHI 22]

AI-driven tutorials [CHI 20]

Interactive explanations (in natural language) [CSCW 21; preprint]

Experiment design

Out-of-distribution [CSCW 21]

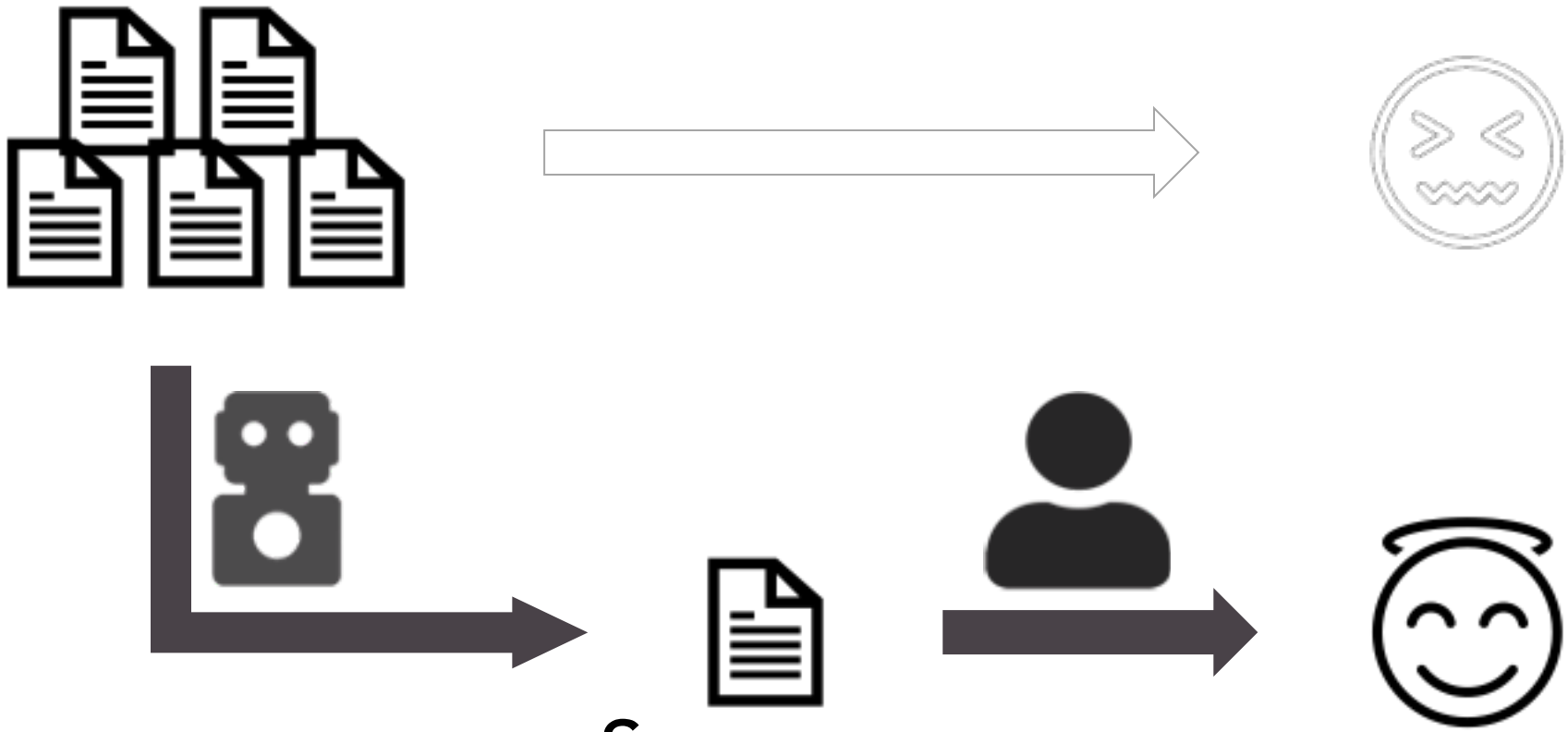
EMNLP 2021

Decision-focused Summarization

Chao-Chun Hsu and Chenhao Tan

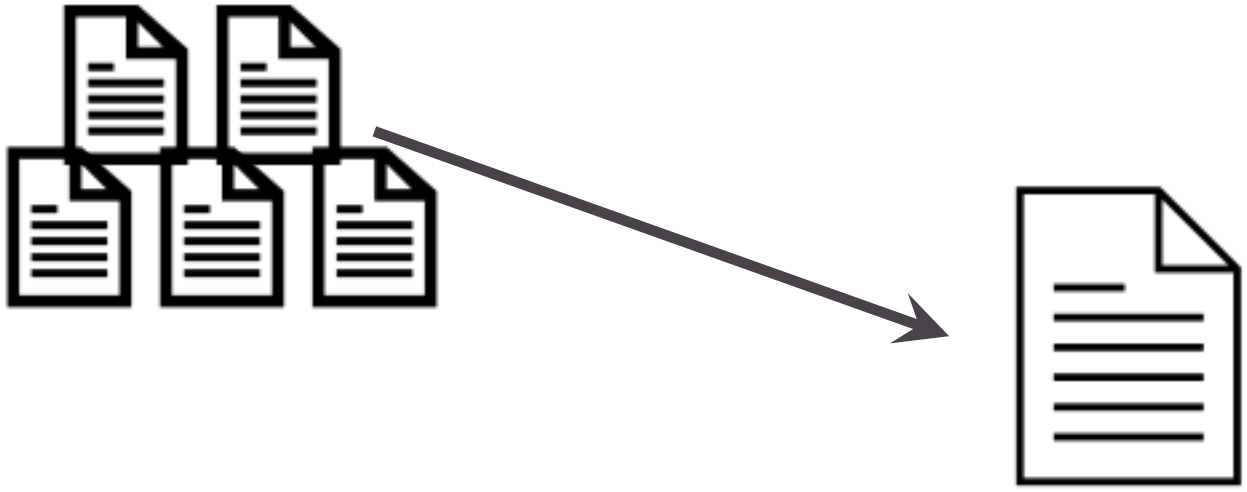
Human decision making requires making sense of large amount of information

Summarization can help by identifying the most relevant information



Typical summarization methods do not account for a target decision

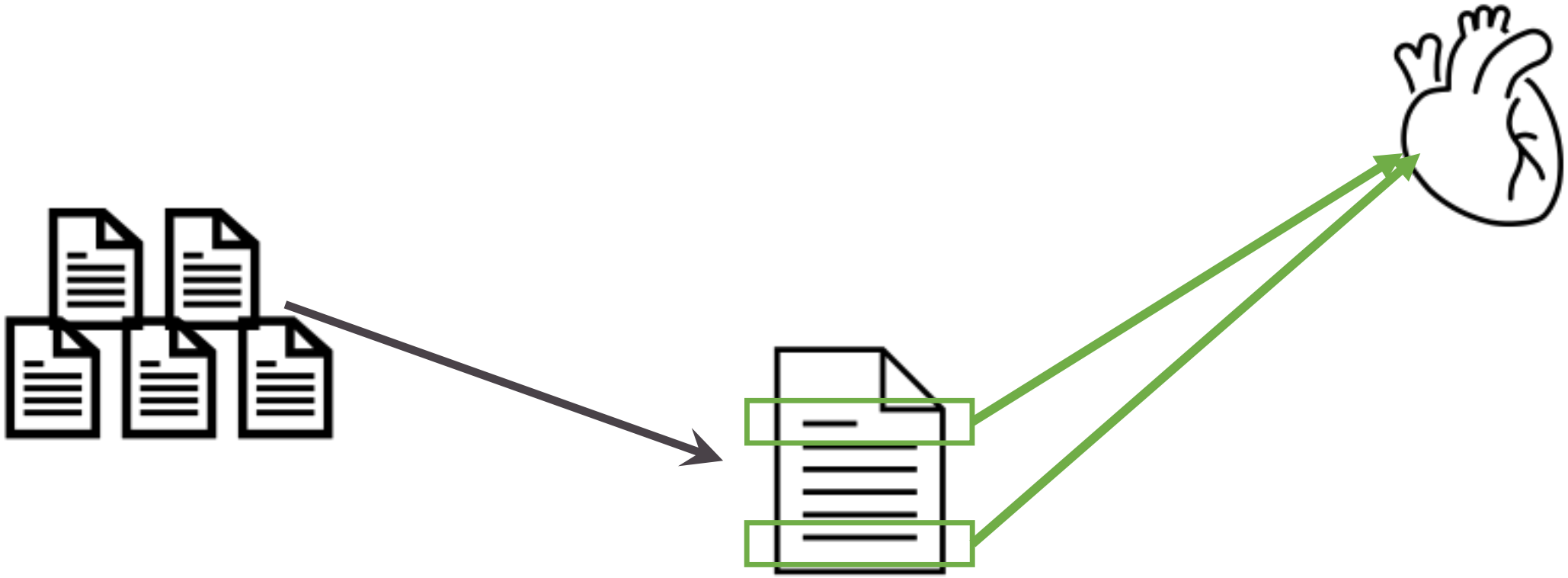
Imagine a doctor making a diagnosis for heart disease



General summary

Typical summarization methods do not account for a target decision

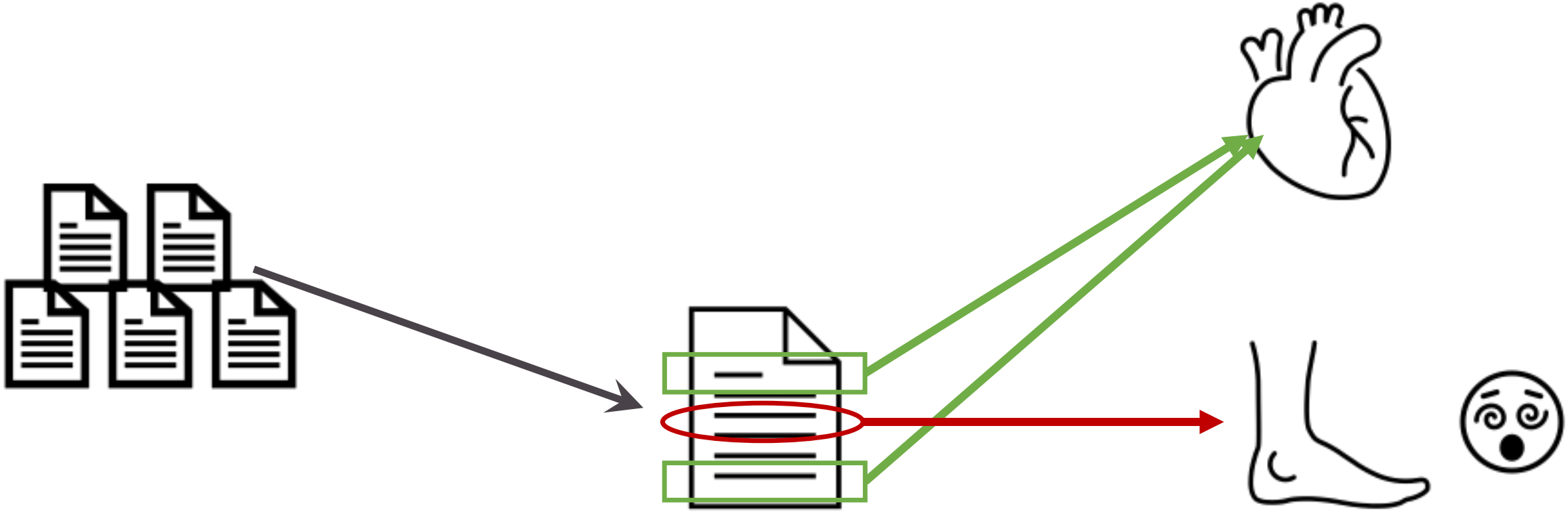
Imagine a doctor making a diagnosis for heart disease



General summary

Typical summarization methods do not account for a target decision

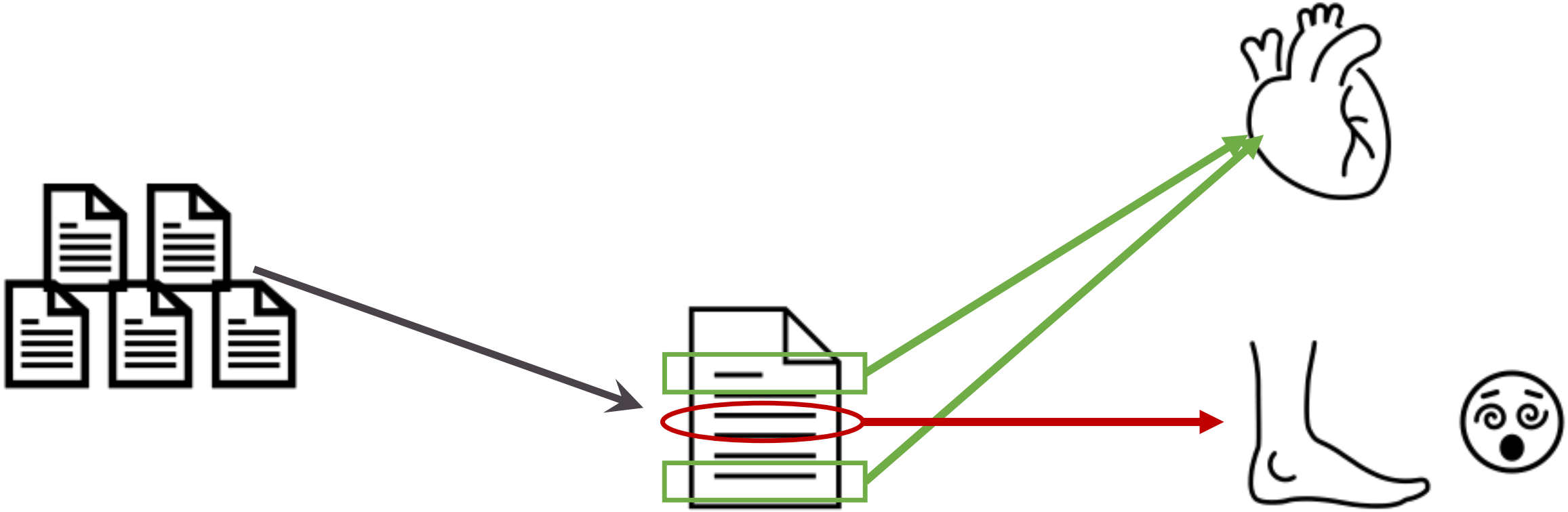
Imagine a doctor making a diagnosis for heart disease



General summary

Typical summarization methods do not account for a target decision

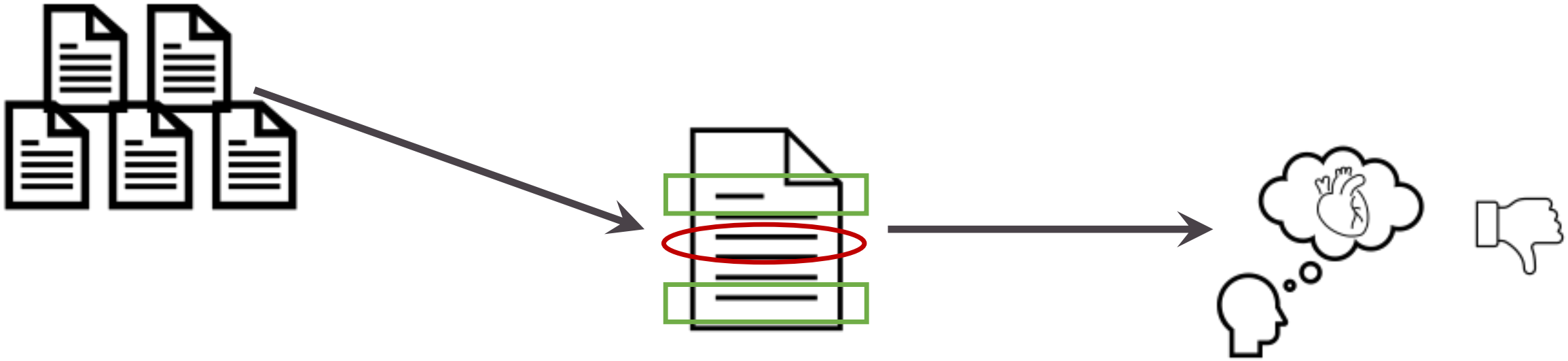
Imagine a doctor making a diagnosis for heart disease



General summary

Typical summarization methods do not account for a target decision

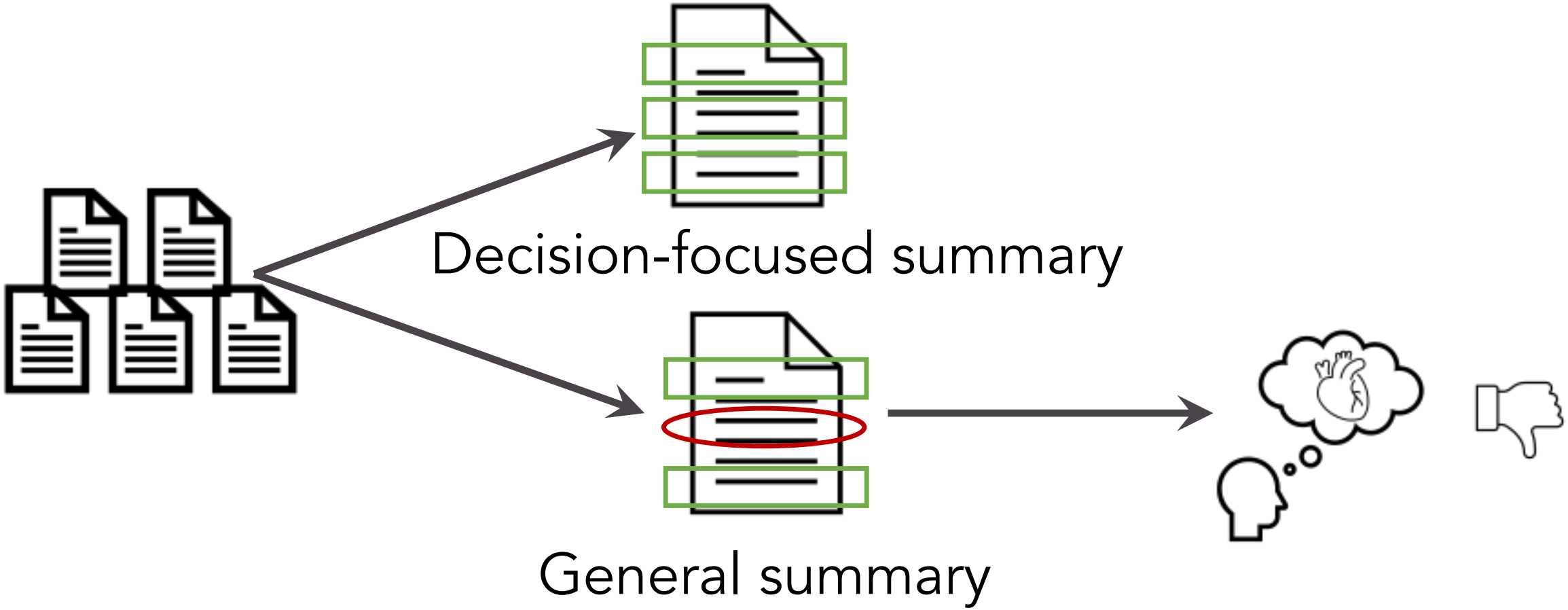
Imagine a doctor making a diagnosis for heart disease



General summary

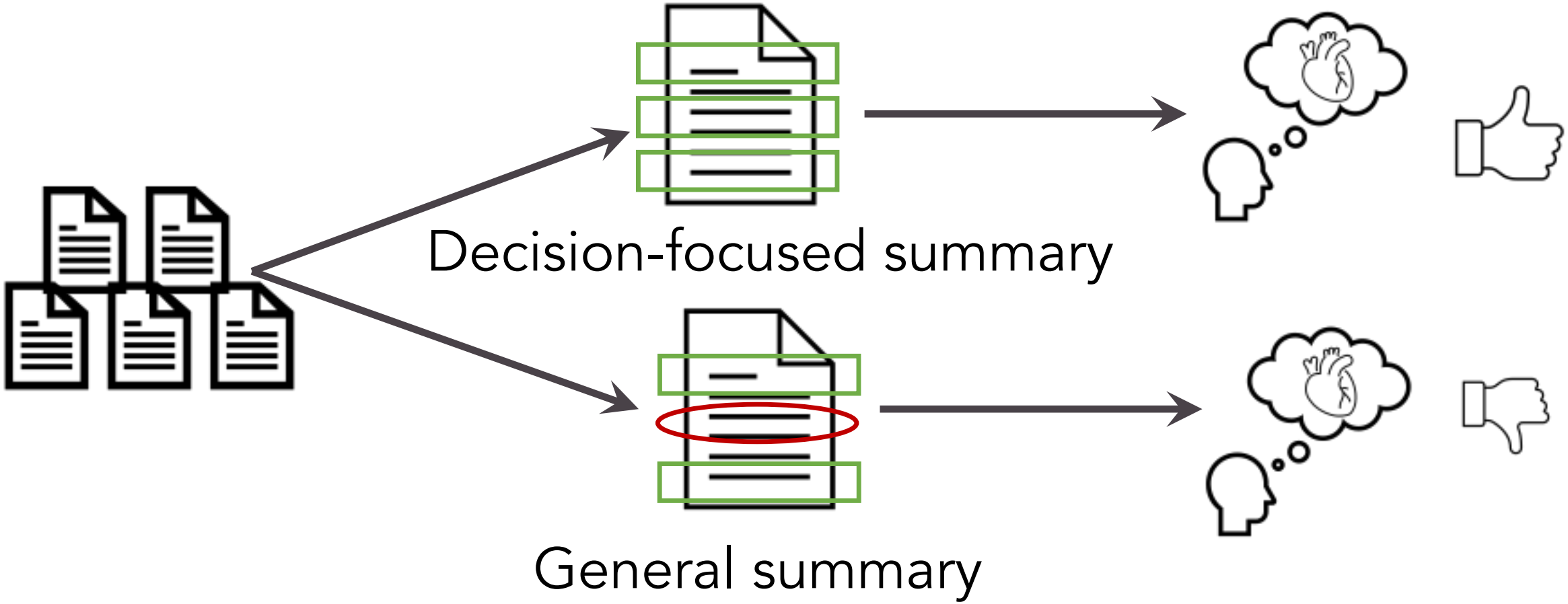
Decision-Focused Summarization

Imagine a doctor making a diagnosis for heart disease



Decision-Focused Summarization

Imagine a doctor making a diagnosis for heart disease



Problem Formulation: Decision-Focused Summarization

Given input texts

$$X = \{x_s\}_{s=1}^{s=S}$$

Problem Formulation: Decision-Focused Summarization

Given input texts

$$X = \{x_s\}_{s=1}^{s=S}$$

Select a subset of sentences

$$\tilde{X} \subset X$$

Decision-focused Summary

Problem Formulation: Decision-Focused Summarization

Given input texts

$$X = \{x_s\}_{s=1}^{s=S}$$

Select a subset of sentences

$$\tilde{X} \subset X$$

Decision-focused Summary

To support making the decision

$$y$$

Yelp Future Rating Prediction Task

Given the first 10 reviews

$$X = \{x_s\}_{s=1}^{s=S}$$

Review 1: Great location! All the staff were very friendly.....
....
Review 5: Probably the worst dining experience I've had in a long time.
...
Review 10: ...

Select a subset of sentences

$$\tilde{X} \subset X$$

$\tilde{x}_1$: Love this place and they got big screen TV'S always playing football, great idea. $\tilde{x}_2$: My soup came out cold, our server forgot our drinks, and they just microwaved it to warm it up and it literally over cooked everything in the soup. $\tilde{x}_3$: I had a pancake combo with New York cheese cake pancakes and they were delicious!!!

Average rating of first 50 reviews

$$y$$

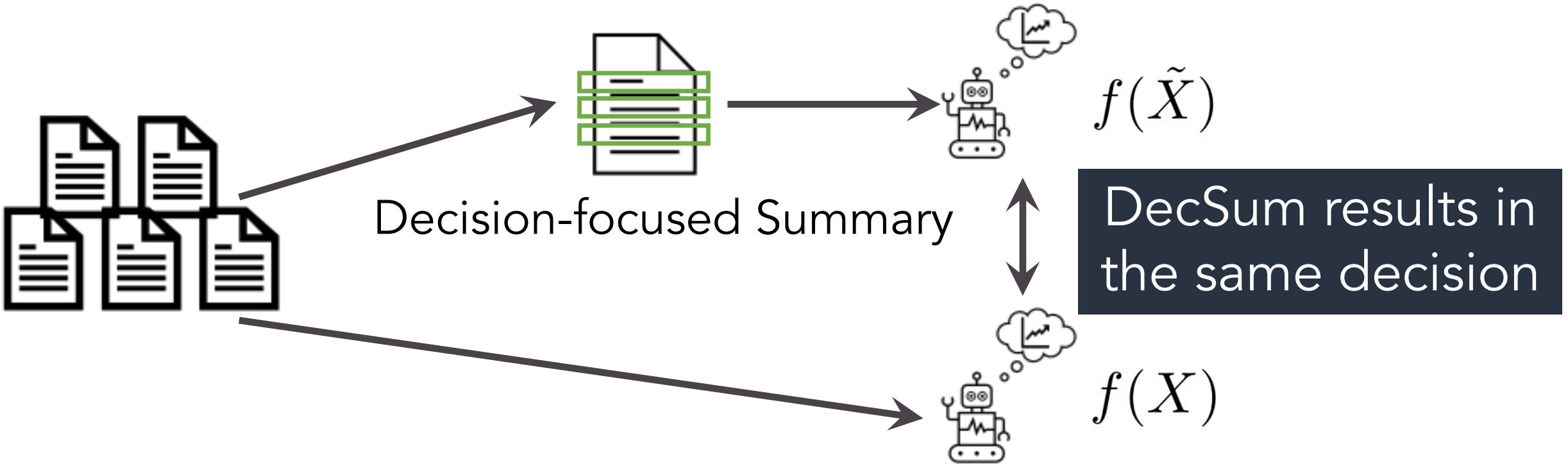
2.8 / 5
first 50 reviews

DecSum

- Supervised model, *f : X -> y*
- Objective components:
  1. Decision faithfulness
  2. Decision representativeness
  3. Textual non-redundancy

Decision Faithfulness: $f(\tilde{X}) \sim f(X)$

$$\mathcal{L}_{\mathrm{F}}(\tilde{X}, X, f) = \log |f(\tilde{X}) - f(X)|.$$

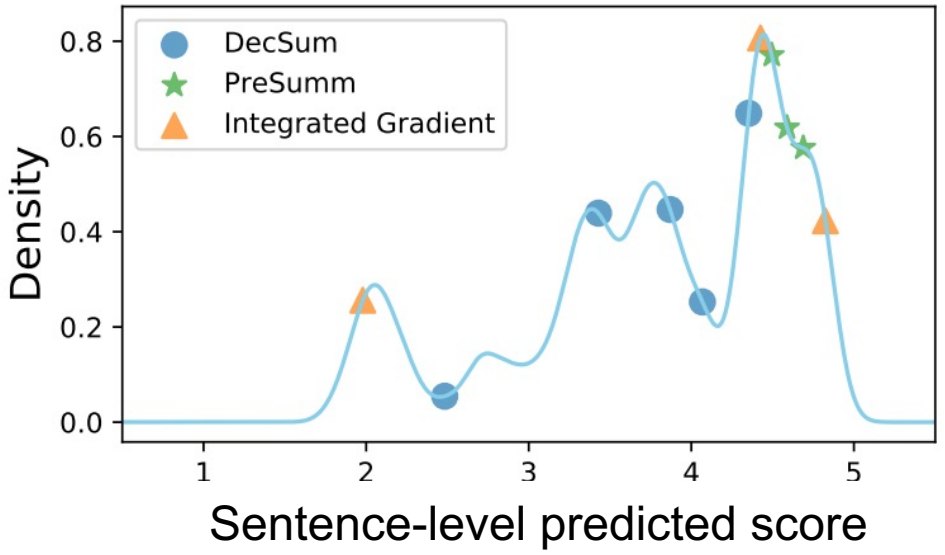


Decision Representativeness

Decision distribution of individual sentences

For full input X:

$$\hat{Y}_X = \{f(x) \mid x \in X\}$$



Decision Representativeness

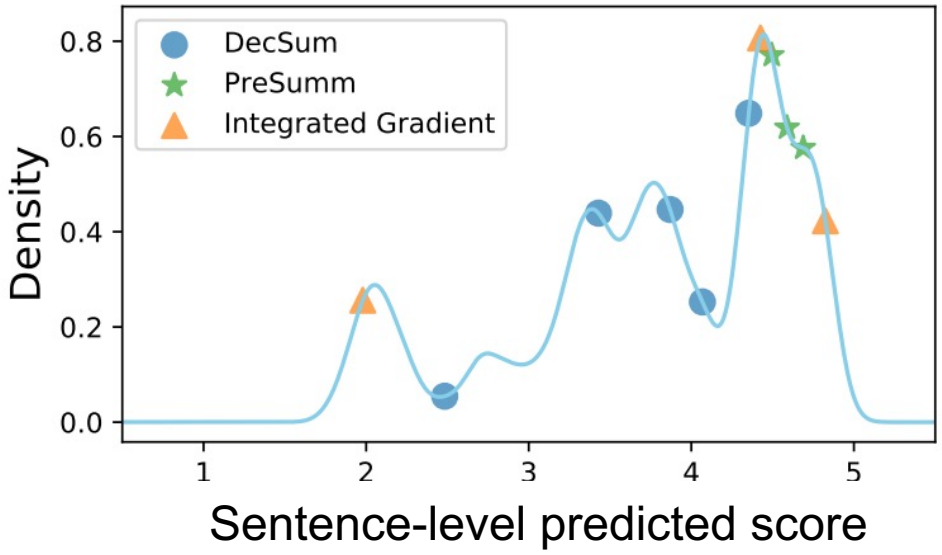
Decision distribution of individual sentences

For full input X:

$$\hat{Y}_X = \{f(x) \mid x \in X\}$$

For selected summary set $\tilde{X}$:

$$\hat{Y}_{\tilde{X}} = \{f(x) \mid x \in \tilde{X}\}$$



Wasserstein Distance

$$\mathcal{L}_{\mathrm{R}}(\tilde{X}, X, f) = \log(W(\hat{Y}_{\tilde{X}}, \hat{Y}_X))$$

Textual Non-redundancy

$$\mathcal{L}_{\mathrm{D}}(\tilde{X}) = \sum_{x \in \tilde{X}} \max_{x' \in \tilde{X} - \{x\}} \mathrm{cossim}(s(x), s(x'))$$

Where *s(x)* means using sentence embedding from SentBERT [Reimers and Gurevych, 2019]

Greedy algorithm to iteratively select sentences

$$\mathcal{L}(\tilde{X},X,f)=\alpha\mathcal{L}_{\mathrm{F}}(\tilde{X},X,f)+\beta\mathcal{L}_{\mathrm{R}}(\tilde{X},X,f)+\gamma\mathcal{L}_{\mathrm{D}}(\tilde{X})$$

Baselines

- Text-only methods:
  1. BART – abstractive summarization
  2. PreSumm – BERT-based extractive summarization
  3. Random selection

Baselines

- Text-only summarization methods:
  1. BART – abstractive summarization
  2. PreSumm – BERT-based extractive summarization
  3. Random selection
- Model-based explanation methods (based on supervised model):
  1. Integrated Gradient (IG)
  2. Attention

Automatic Evaluation: Decision Faithfulness, f($\tilde{X}$) ~ f(X)

Method	MSE with **Full** (faithfulness) ↓	MSE ↓
Full (oracle)	0	0.135
Text-only summarization methods		
Random	0.356	0.475
BART	0.368	0.502
PreSumm	0.339	0.478
Model-based explanation methods		
IG		
Attention		
`DecSum` w/ (α decision faithfulness, β decision representativeness, γ textual non-redundancy)		
(1, 1, 1)		
(0, 1, 1)		

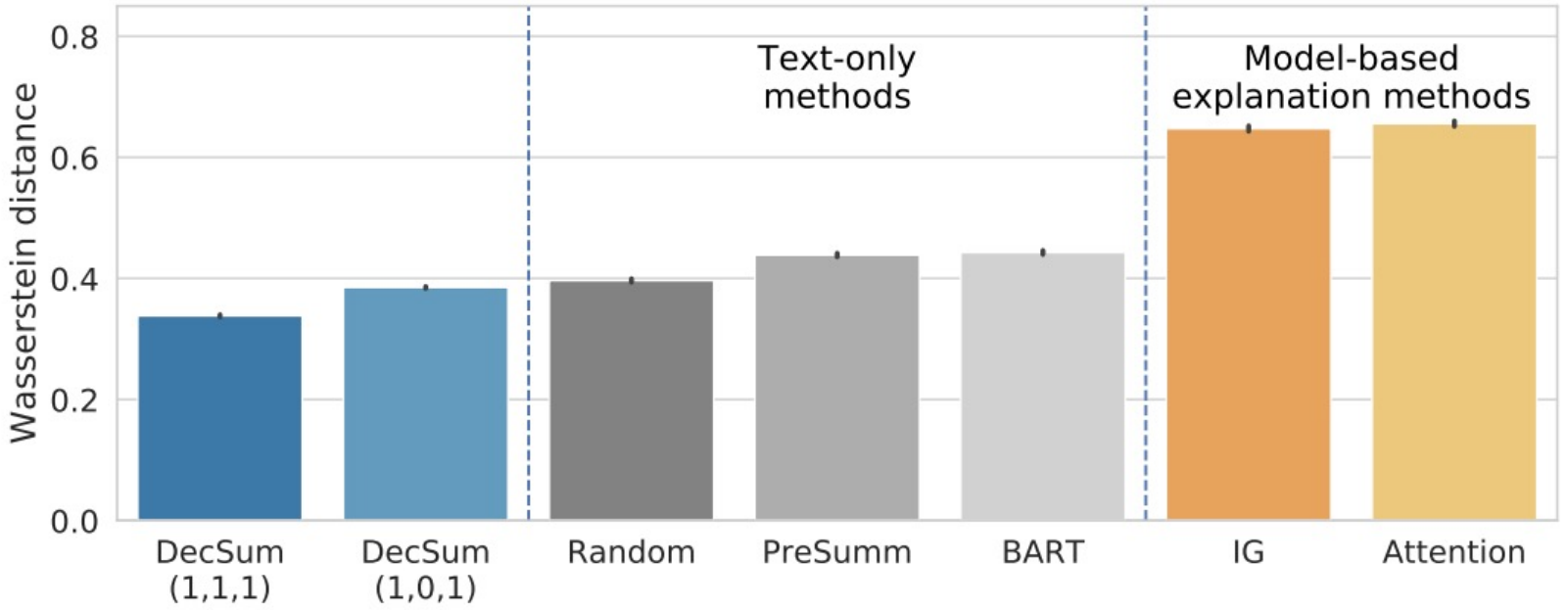
Automatic Evaluation: Decision Faithfulness, $f(\tilde{X}) \sim f(X)$

Method	MSE with **Full** (faithfulness) ↓	MSE ↓
Full (oracle)	0	0.135
Text-only summarization methods		
Random	0.356	0.475
BART	0.368	0.502
PreSumm	0.339	0.478
Model-based explanation methods		
IG	0.436	0.565
Attention	0.539	0.715
`DecSum` w/ (α decision faithfulness, β decision representativeness, γ textual non-redundancy)		
(1, 1, 1)		
(0, 1, 1)		

Automatic Evaluation: Decision Faithfulness, $f(\tilde{X}) \sim f(X)$

Method	MSE with **Full** (faithfulness) ↓	MSE ↓
Full (oracle)	0	0.135
Text-only summarization methods		
Random	0.356	0.475
BART	0.368	0.502
PreSumm	0.339	0.478
Model-based explanation methods		
IG	0.436	0.565
Attention	0.539	0.715
`DecSum` w/ (α decision faithfulness, β decision representativeness, γ textual non-redundancy)		
(1, 1, 1)	<u>0.0005</u>	0.136
(0, 1, 1)	0.162	0.283

Automatic Evaluation: Decision Representativeness



$$\hat{Y}_{\tilde{X}} = \{f(x) \mid x \in \tilde{X}\} \qquad \hat{Y}_X = \{f(x) \mid x \in X\}$$

$$W(\hat{Y}_{\tilde{X}}, \hat{Y}_X) = \inf_{\gamma \in \Gamma(\hat{Y}_{\tilde{X}}, \hat{Y}_X)} \int_{\mathbb{R}\times\mathbb{R}} ||f - f'|| d\gamma(f, f')$$

Simplified task for human evaluation: Compare future ratings of two restaurants

First 10 reviews of the restaurant A

3.8/5, first 10 reviews

First 10 reviews of the restaurant B

3.8/5, first 10 reviews

Simplified task for human evaluation: Compare future ratings of two restaurants

First 10 reviews of the restaurant A

3.8/5, first 10 reviews

First 10 reviews of the restaurant B

3.8/5, first 10 reviews

Using the same summarization method

Summary A

Summary B

Simplified task for human evaluation: Compare future ratings of two restaurants

First 10 reviews of the restaurant A

3.8/5, first 10 reviews

First 10 reviews of the restaurant B

3.8/5, first 10 reviews

Summary A

Summary B

Which restaurant will be rated better after 50 reviews?

Example summary

IHOP

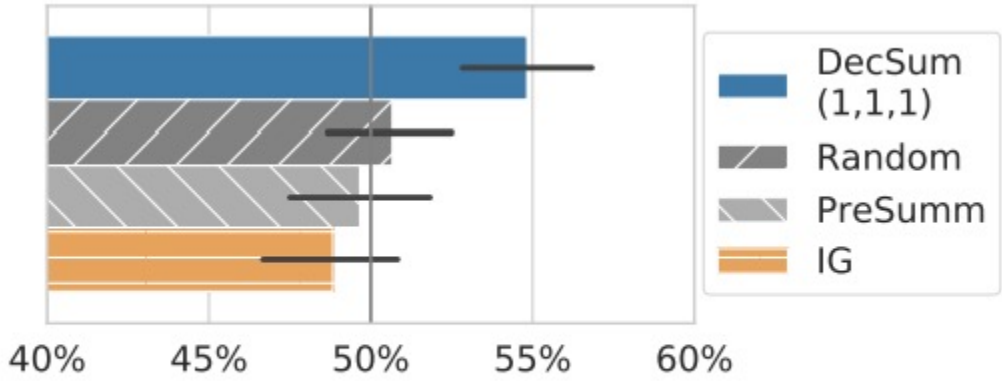
I had a pancake combo with New York cheese cake pancakes and they were delicious ! ! !.
This place was great
I got to eat breakfast and watch the football game !.
Finally a local IHOP, great service and always delicious breakfast.
Nice clean place.

Tasty Kabob

Also they have the best Persian Ice Cream which is only one 1/3 flavor .... what is the flavor??
(its a secret, you will have to go there and find out !).
Tasty Kabob is a must see on any Hookah bar tour.
Tasty Kabob, while among the best Persian restaurants in Arizona, falls short of Famous Kabob in Sacramento and many Los Angeles joints.

Human Performance

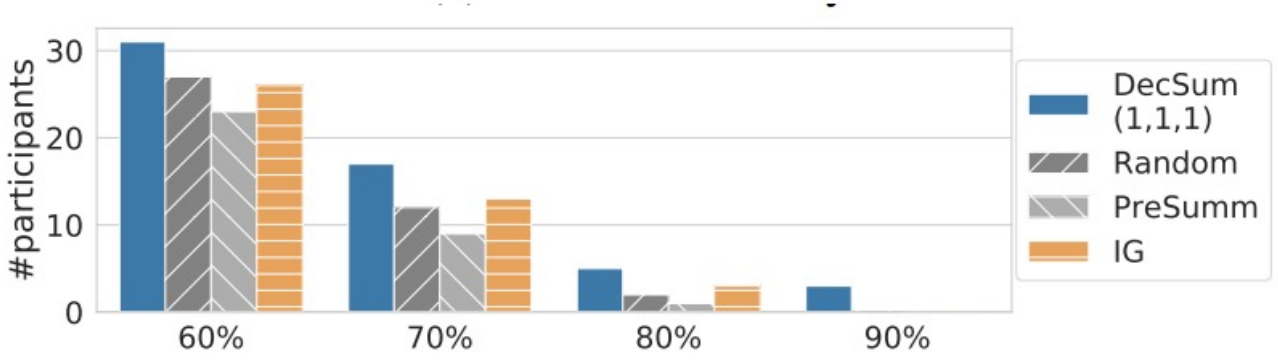
- This task is very challenging for humans
- Only **DecSum** allows humans to outperform random (50%)



(a) Human accuracy

Human Performance

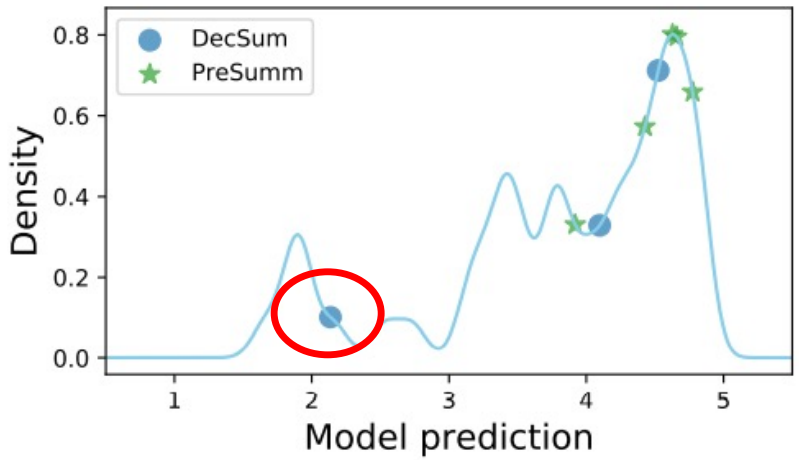
- This task is very challenging for humans
- Only **DecSum** allows humans to outperform random (50%)
- 3 participants can achieve 90% acc. with DecSum

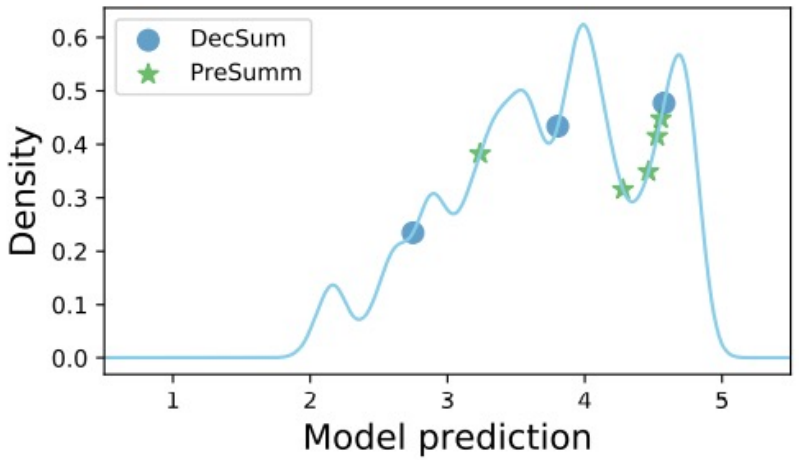


(b) #participants with over 60% accuracy

Example revisisted

Method	Restaurant 1: IHOP	Restaurant 2: Tasty Kabob (rated better after 50 reviews.)
PreSumm	$\tilde{x}_1$: I had a pancake combo with New York cheese cake pancakes and they were delicious ! ! !. $\tilde{x}_2$: This place was great $\tilde{x}_3$: I got to eat breakfast and watch the football game !. $\tilde{x}_4$: Finally a local IHOP , great service and always delicious breakfast. $\tilde{x}_5$: Nice clean place.	$\tilde{x}_1$: Also they have the best Persian Ice Cream which is only one flavor .... $\tilde{x}_2$: what is the flavor?? $\tilde{x}_3$: (its a secret , you will have to go there and find out !). $\tilde{x}_4$: Tasty Kabob is a must see on any Hookah bar tour. $\tilde{x}_5$: Tasty Kabob , while among the best Persian restaurants in Arizona , falls short of Famous Kabob in Sacramento and many Los Angeles joints.
DecSum	$\tilde{x}_1$: Love this place and they got big screen TV'S always playing football, great idea. $\tilde{x}_2$: My soup came out cold, our server forgot our drinks, and they just microwaved it to warm it up and it literally over cooked everything in the soup. $\tilde{x}_3$: I had a pancake combo with New York cheese cake pancakes and they were delicious!!!	$\tilde{x}_1$: Regardless, both versions were moist and very appealing. $\tilde{x}_2$: If you thought you didn't like Persian food, this place will definitely make you think again. $\tilde{x}_3$: It was a generous portion for two, but I found myself munching on it just to pass the time until our lunches came, not because it was exceptionally well done.





Summary

- A new summarization formulation: decision-focused summarization
- DecSum method can outperform text-only summarization methods and model-based explanation methods on both automatic evaluations and human evaluation
- Many future applications in finance and medicine

Evaluation of AI explanations

Emulation

Conceptually and empirically, humans may not provide "groundtruth" explanations

Discovery

Human+AI rarely outperforms AI
Decision-focused summarization

Evaluation of AI explanations

Emulation

Conceptually and empirically, humans may not provide "groundtruth" explanations

Discovery

Human+AI rarely outperforms AI
Decision-focused summarization

- Understand how people explain AND be aware that they are not "perfect"
- Understand how people make decisions AND identify human and AI strengths

CHAI and friends













and many more!

Towards effective human-centered explanations

- Understand how people explain AND be aware that they are not "perfect"
- Understand how people make decisions AND identify human and AI strengths

chenhao@uchicago.edu
@ChenhaoTan

Thank you!