Defining and evaluating diversity in generation

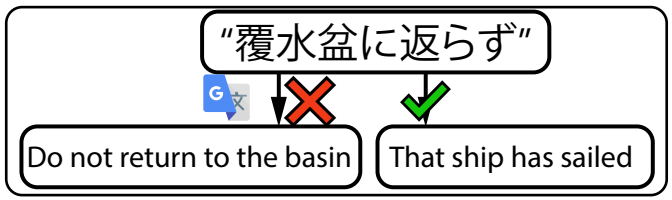
Tatsunori Hashimoto

NAACL NeuralGen 2019

Evaluation is an open problem

Evaluation must tradeoff amongst competing goals

Correctness (quality)

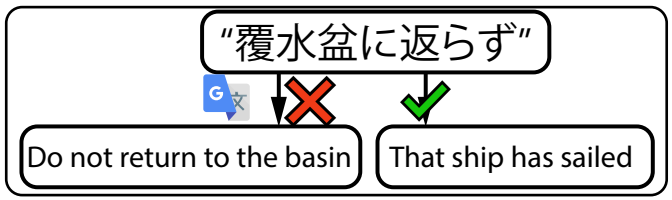


Human evaluation and surrogates (BLEU)
[Novikova+ 17][Papineni+ 02]

Evaluation is an open problem

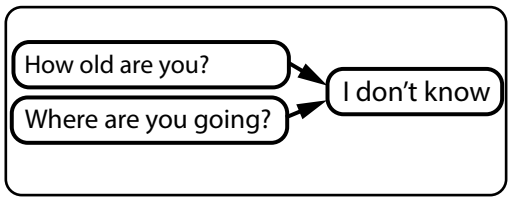
Evaluation must tradeoff amongst competing goals

Correctness (quality)



Human evaluation and surrogates (BLEU)
[Novikova+ 17][Papineni+ 02]

Specificity (diversity)

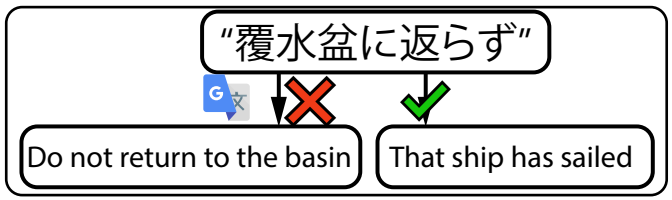


N-gram counting [Li+ 16][See+ 17]

Evaluation is an open problem

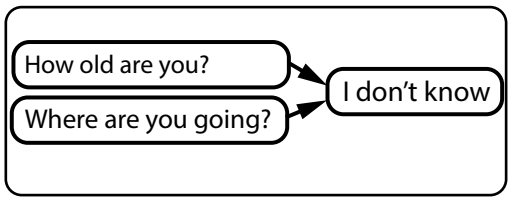
Evaluation must tradeoff amongst competing goals

Correctness (quality)



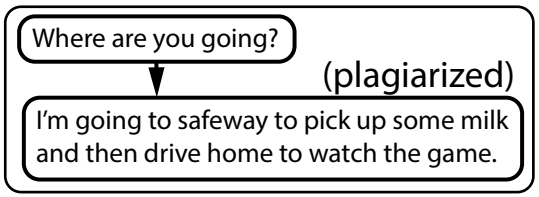
Human evaluation and surrogates (BLEU)
[Novikova+ 17][Papineni+ 02]

Specificity (diversity)



N-gram counting [Li+ 16][See+ 17]

Plagiarism (diversity)

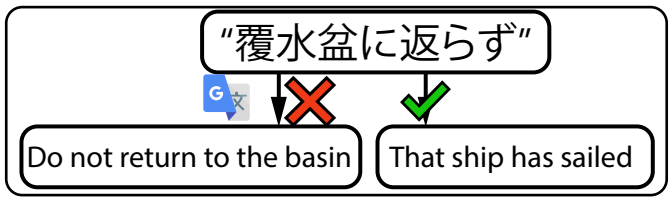


Perplexity (likelihood) [Novikova+ 17]

Evaluation is an open problem

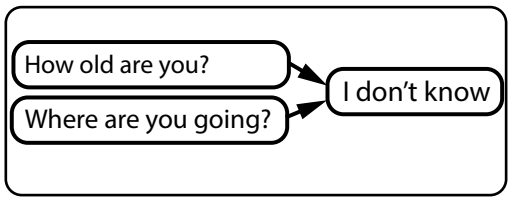
Evaluation must tradeoff amongst competing goals

Correctness (quality)



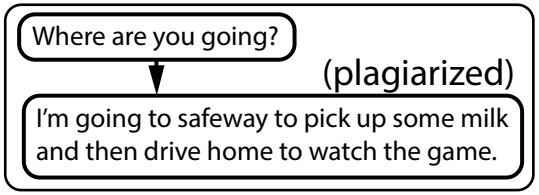
Human evaluation and surrogates (BLEU)
[Novikova+ 17][Papineni+ 02]

Specificity (diversity)



N-gram counting [Li+ 16][See+ 17]

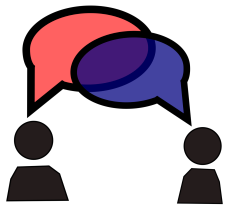
Plagiarism (diversity)



Perplexity (likelihood) [Novikova+ 17]

Quality-diversity tradeoffs are pervasive

Story generation
[Fan+ 18]



Dialogue
[Li+ 16]

Open-domain QA
[Bajaj+ 18]

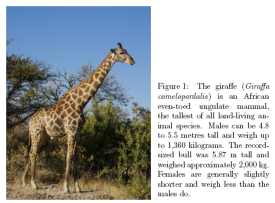


Figure 1: The giraffe (*Giraffa camelopardalis*) is an African even-toed ungulate mammal, the tallest of all land-living animal species. Males can be 4.8 to 5.5 metres tall and weigh up to 1,360 kilograms. The recognized bull was 5.87 m tall and weighed approximately 2,000 kg. Females are generally slightly shorter and weigh less than the males do.

Captioning
[Miltenburg+18]

Challenges in evaluation

Challenge 1: Evaluating creative generation

"Story" generated by a language model

The scientist named the population, after their distinctive horn, Ovid's Unicorn. These four-horned, silver-white unicorns were previously unknown to science. Now, after almost two centuries, the mystery of what sparked this odd phenomenon is finally solved. Dr. Jorge Pérez, an evolutionary biologist from the University of La Paz, and several companions, were exploring the Andes Mountains when they found a small valley, with no other animals or humans. ...

Challenges in evaluation

Challenge 1: Evaluating creative generation

"Story" generated by a language model

The scientist named the population, after their distinctive horn, Ovid's Unicorn. These four-horned, silver-white unicorns were previously unknown to science. Now, after almost two centuries, the mystery of what sparked this odd phenomenon is finally solved. Dr. Jorge Pérez, an evolutionary biologist from the University of La Paz, and several companions, were exploring the Andes Mountains when they found a small valley, with no other animals or humans. ...

Questions:

How creative is the model?
Does it generalize or plagiarize?

Challenges in evaluation

Challenge 2: Evaluating open-domain QA

Open-domain QA [Bajaj+ 2018]

Q: What is the location of Stanford

A1: Palo Alto, California
A2: Stanford, California
A3: 35 miles south of San Francisco and 20 miles north of San Jose

Challenges in evaluation

Challenge 2: Evaluating open-domain QA

Open-domain QA [Bajaj+ 2018]

Q: What is the location of Stanford

A1: Palo Alto, California
A2: Stanford, California
A3: 35 miles south of San Francisco and 20 miles north of San Jose

A model which understands should..

Be able to generate correct answers
Be able to generate **all** correct answers

Goals for NLG eval

How do we measure progress in generation?

Generative models should not only

generate high-quality answers (quality)

Goals for NLG eval

How do we measure progress in generation?

Generative models should not only

generate high-quality answers (quality)

but also cover valid, atypical ones (diversity)

Goals for NLG eval

How do we measure progress in generation?

Generative models should not only

generate high-quality answers (quality)

but also cover valid, atypical ones (diversity)

Looking at nice examples (and even human eval) is not sufficient

Goals for NLG eval

How do we measure progress in generation?

Generative models should not only

generate high-quality answers (quality)

but also cover valid, atypical ones (diversity)

Looking at nice examples (and even human eval) is not sufficient

We need a **systematic** approach to evaluation

Two coherent approaches

Task completion (Extrinsic)

Complete a target task

Imitation (Intrinsic)

Have models behave like humans

Two coherent approaches

Task completion (Extrinsic)

Complete a target task

Unambiguous evaluation (task completion)

Dialogue

Imitation (Intrinsic)

Have models behave like humans

No clear evaluation

Story generation

Two coherent approaches

Task completion (Extrinsic)

Complete a target task

Unambiguous evaluation (task completion)

Dialogue

Imitation (Intrinsic)

Have models behave like humans

No clear evaluation

Story generation

What we cover in this talk

Most academic research

Why do we care about imitation

Scientific interest

Can we match human performance?
(i.e. story generation, humor)

Why do we care about imitation

Scientific interest

Can we match human performance?
(i.e. story generation, humor)

Imitation for downstream tasks

A model that imitates humans is useful for many tasks.
(i.e. paraphrasing, language models, summarization)

Why do we care about imitation

Scientific interest

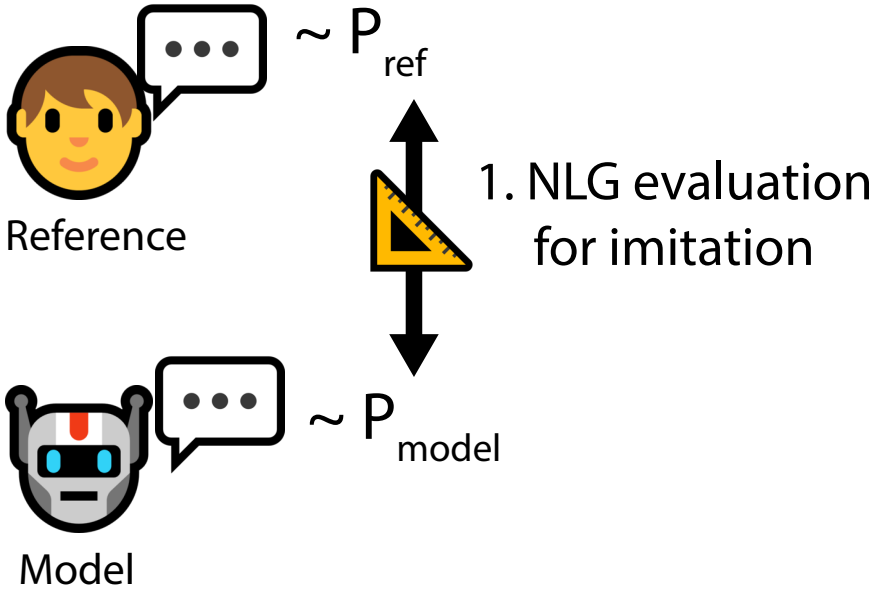
Can we match human performance?
(i.e. story generation, humor)

Imitation for downstream tasks

A model that imitates humans is useful for many tasks.
(i.e. paraphrasing, language models, summarization)

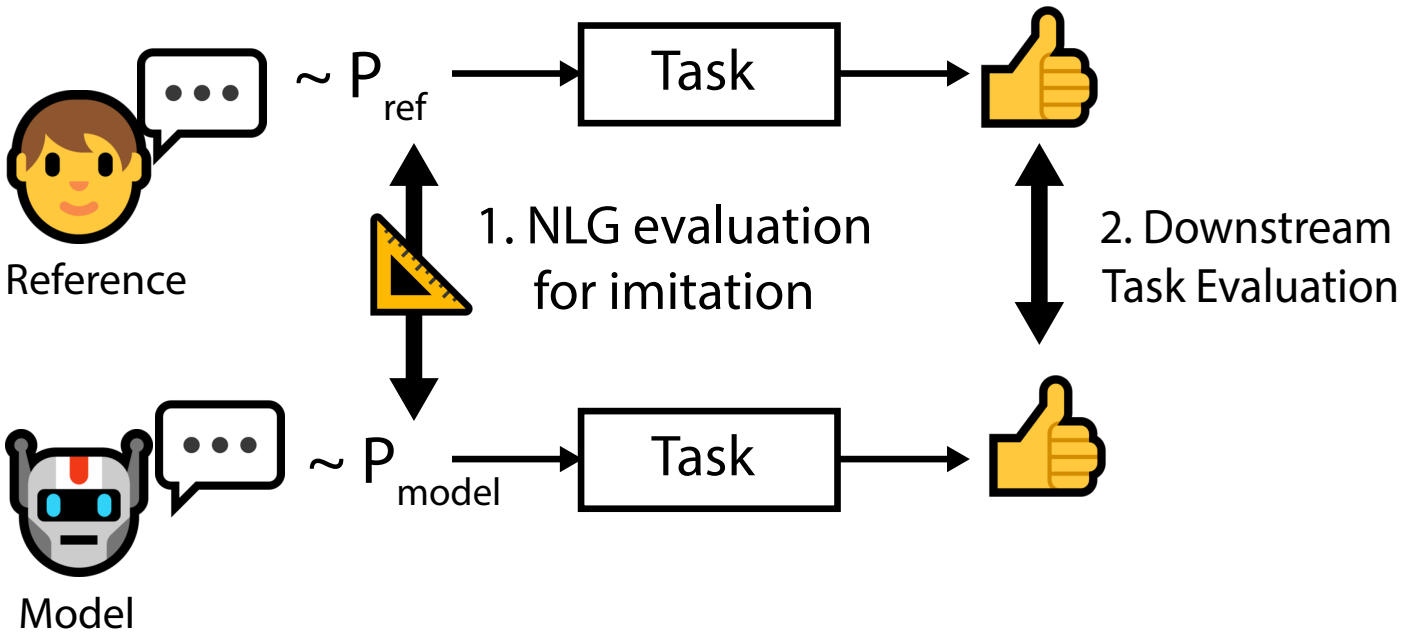
In both cases: we need rigorous evaluations and contracts

Evaluations as contracts



1. Evaluation should guarantee whether P_{model} is similar to P_{ref}

Evaluations as contracts



1. Evaluation should guarantee whether P_{model} is similar to P_{ref}

2. Similarity between P_{model} and P_{ref} should guarantee similar task performance

Imitation and distribution matching

Evaluation for imitation:

An evaluation $D(p_{\text{ref}}, p_{\text{model}})$ should measure whether the distribution p_{ref} matches p_{model}.

Imitation and distribution matching

Evaluation for imitation:

An evaluation $D(p_{\text{ref}}, p_{\text{model}})$ should measure whether the distribution p_{ref} matches p_{model}.

Examples:

- **KL divergence**: $\mathbb{E}_{p_{\text{ref}}}[-\log p_{\text{model}}(x)] + C$.
- **Optimal classification accuracy:** $\|p_{\text{ref}} - p_{\text{model}}\|$.

The status quo doesn't do this!

The status quo for evaluation

The status quo:

Define a sample quality metric l and evaluate,

$$E_{p_{\text{model}}}[l(x)].$$

Examples: Human judgement, BLEU, learned evaluation.

The status quo for evaluation

The status quo:

Define a sample quality metric l and evaluate,

$$E_{p_{\text{model}}}[l(x)].$$

Examples: Human judgement, BLEU, learned evaluation.

Advantages: can be evaluated in a black box.
Disadvantage: *inherently* kills diversity: models seek to return the single x that recieve the highest score

The status quo for evaluation

The status quo:

Define a sample quality metric l and evaluate,

$$E_{p_{\text{model}}}[l(x)].$$

Examples: Human judgement, BLEU, learned evaluation.

Advantages: can be evaluated in a black box.
Disadvantage: *inherently* kills diversity: models seek to return the single x that recieve the highest score

This leads to many issues surrounding evaluation and diversity

Roadmap

Part 1: Problems with the status quo

Part 2: Rigorous and practical diversity evaluation

Part 3: Open problems

Roadmap

Part 1: Problems with the status quo

Part 2: Rigorous and practical diversity evaluation

Part 3: Open problems

Common misconceptions

1. Models are under-diverse

2. Human evaluations capture diversity defects

3. We already adequately test for diversity

Issue 1: Under-diverse models

Claim: models are under-diverse [Li+ 2016, Shao+ 2017]
Reality: decoders make models under-diverse and generic

Issue 1: Under-diverse models

Claim: models are under-diverse [Li+ 2016, Shao+ 2017]
Reality: decoders make models under-diverse and generic

Samples from the **neural model (w/ annealing)**

1.	The food is great and the service is great.
2.	The food was delicious, and the service was great.
3.	I am glad I didn't care for it.
4.	I am looking forward to returning.
5.	I am so glad I went to this place.

Issue 1: Under-diverse models

Claim: models are under-diverse [Li+ 2016, Shao+ 2017]
Reality: decoders make models under-diverse and generic

Samples from the **neural model (no annealing)**

1.	He's really amazing and helpful and you will be pleased to support his different work team.
2.	With the large indoor seating the pool parties is a fine dining experience.
3.	Their prices are pretty impressive.
4.	Hey, you get to love that there are live music.
5.	Beer wonderful drinks and beer prices.

Issue 1: Under-diverse models

Claim: models are under-diverse [Li+ 2016, Shao+ 2017]
Reality: decoders make models under-diverse and generic

1. Sampled outputs are often diverse (and ungrammatical)

2. Decoders like beam search convert quality to diversity defects

The root cause is that our models are no good

Issue 2: human eval suffices

Claim: (careful) human evaluation is the gold standard metric
Us: humans are easily fooled by underdiversity and plagiarism

Issue 2: human eval suffices

Claim: (careful) human evaluation is the gold standard metric
Us: humans are easily fooled by underdiversity and plagiarism

1. Humans cannot detect subtle diversity defects

2. Humans also cannot detect plagiarism from the training set

Issue 2: human eval suffices

Claim: (careful) human evaluation is the gold standard metric
Us: humans are easily fooled by underdiversity and plagiarism

1. Humans cannot detect subtle diversity defects

2. Humans also cannot detect plagiarism from the training set

Our point: Humans are great at evaluating **single samples**
but this isn't enough to evaluate **models**

Try identifying samples!

Task: Making news headlines from short articles.

Context:	Political leaders in Israel united in prayers for Ariel Sharon as the prime minister underwent surgery after suffering a stroke.
Output:	Sharon has stroke for stroke.

Try identifying samples!

Task: Making news headlines from short articles.

Context:	Political leaders in Israel united in prayers for Ariel Sharon as the prime minister underwent surgery after suffering a stroke.
Output:	Sharon has stroke for stroke.

Answer: **machine generated**

Reference is: Israeli leaders unite in prayer for ailing Sharon.

Easy to detect quality defect

Try identifying samples!

Task: Making news headlines from short articles.

Context:	The Buffalo Bills sacked Tom Donahoe as president and general manager on wednesday, fulfilling expectations of a shake-up.
Output:	Bills sack Donahoe as president and gm.

Try identifying samples!

Task: Making news headlines from short articles.

Context:	The Buffalo Bills sacked Tom Donahoe as president and general manager on wednesday, fulfilling expectations of a shake-up.
Output:	Bills sack Donahoe as president and gm.

Answer: **machine generated**

Reference is: NFL's Bills shake up front office.

Difficult to detect (for humans)

Detecting underdiversity

Model generates generic (aligned) sentences

Context:	The buffalo bills sacked tom donahoe as president and general manager on wednesday, fulfilling expectations of a shake-up
Output:	Bills sack Donahoe as president and gm.

Detecting underdiversity

Model generates generic (aligned) sentences

Context:	The buffalo bills sacked tom donahoe as president and general manager on wednesday, fulfilling expectations of a shake-up
Output:	Bills sack Donahoe as president and gm.

Detecting underdiversity

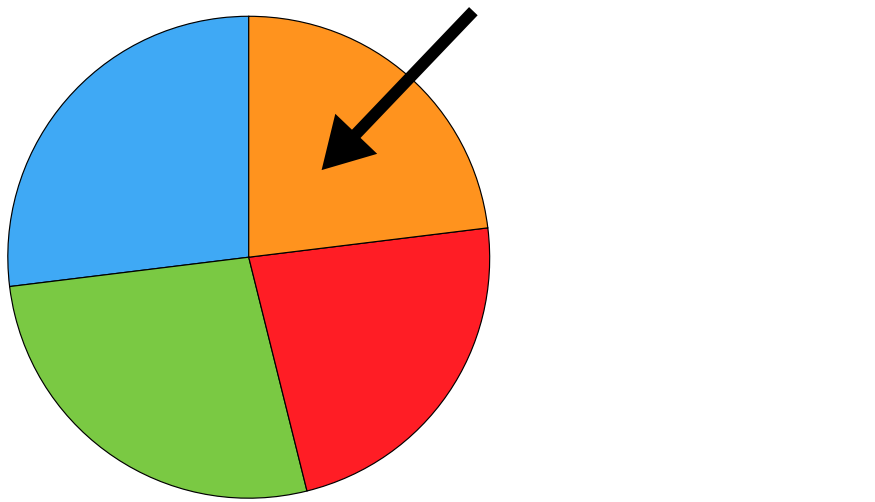
Model generates generic (aligned) sentences

Context:	The buffalo bills sacked tom donahoe as president and general manager on wednesday, fulfilling expectations of a shake-up
Output:	Bills sack Donahoe as president and gm.
Reference:	NFL's Bills shake up front office.

Model is incapable of generating
most (80%) human reference headlines

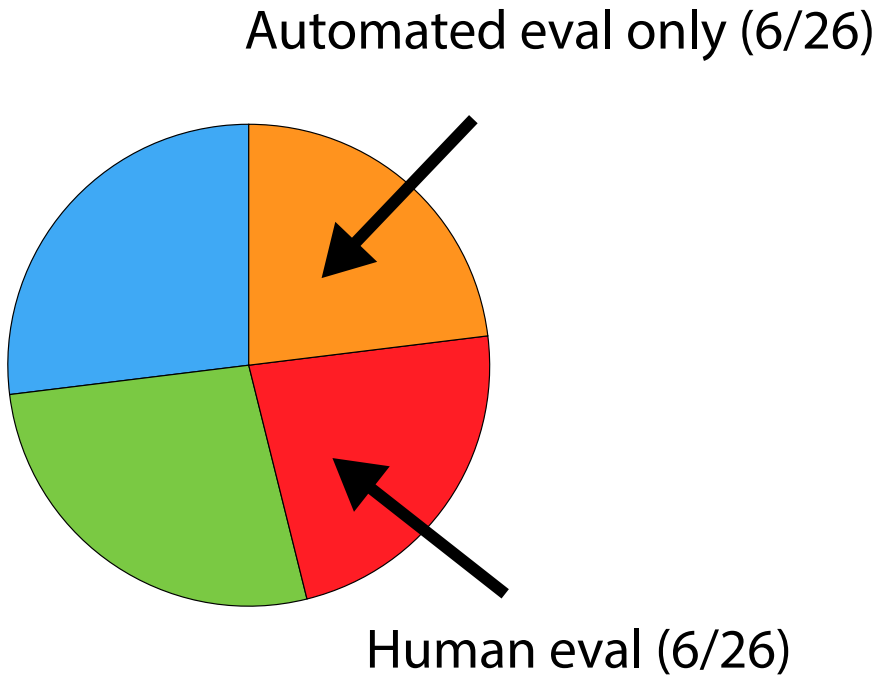
Issue 3: state of evaluation

State of evaluation at ACL 2018



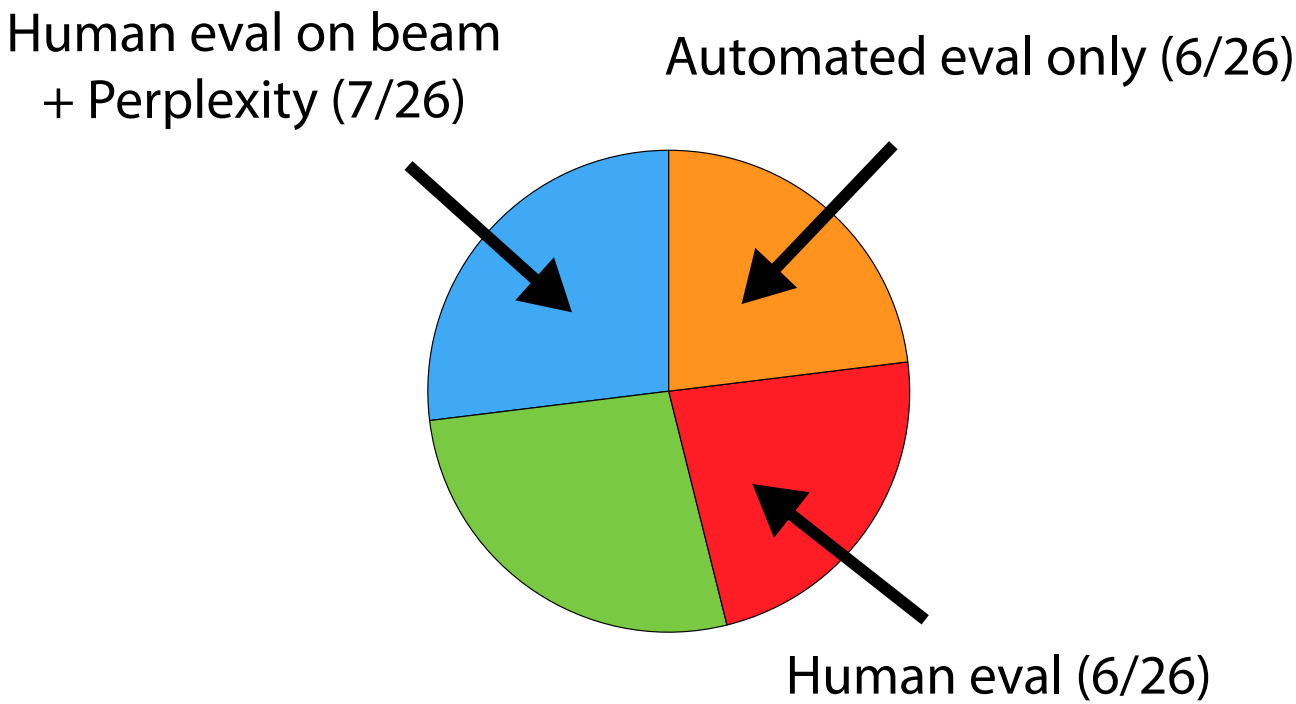
Issue 3: state of evaluation

State of evaluation at ACL 2018



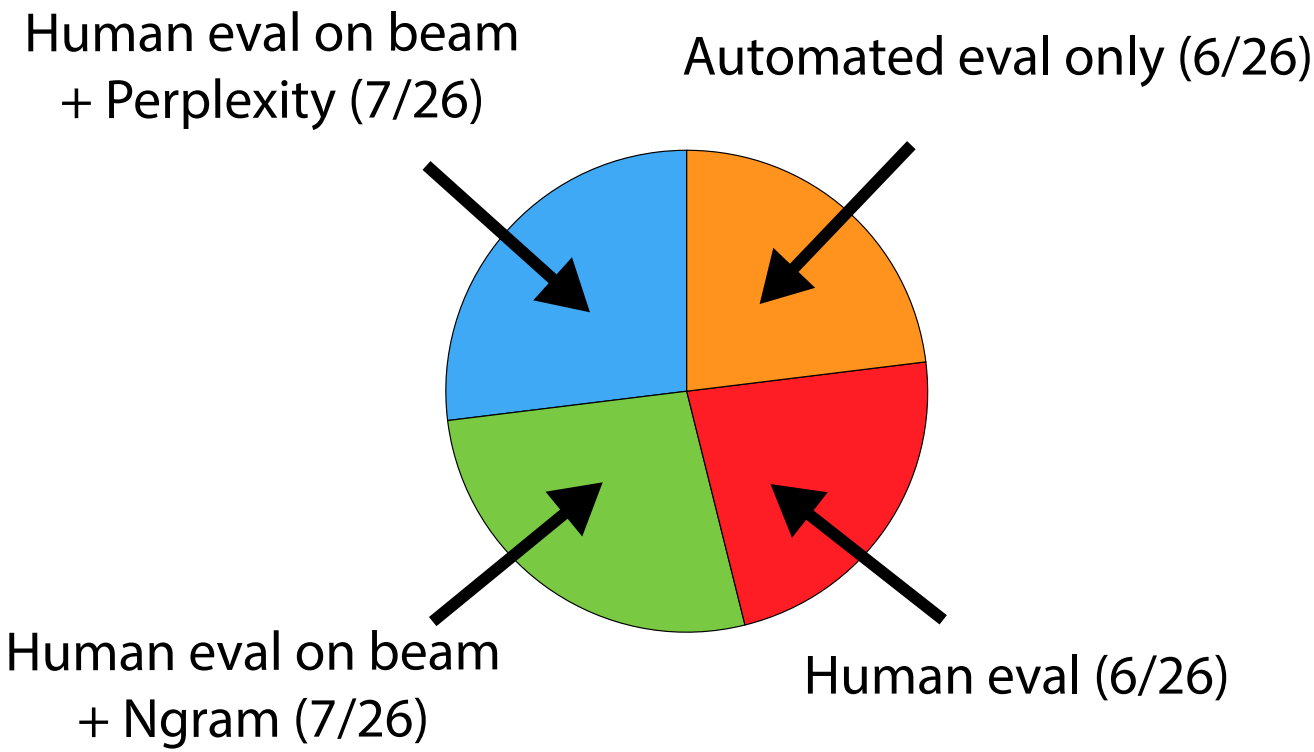
Issue 3: state of evaluation

State of evaluation at ACL 2018



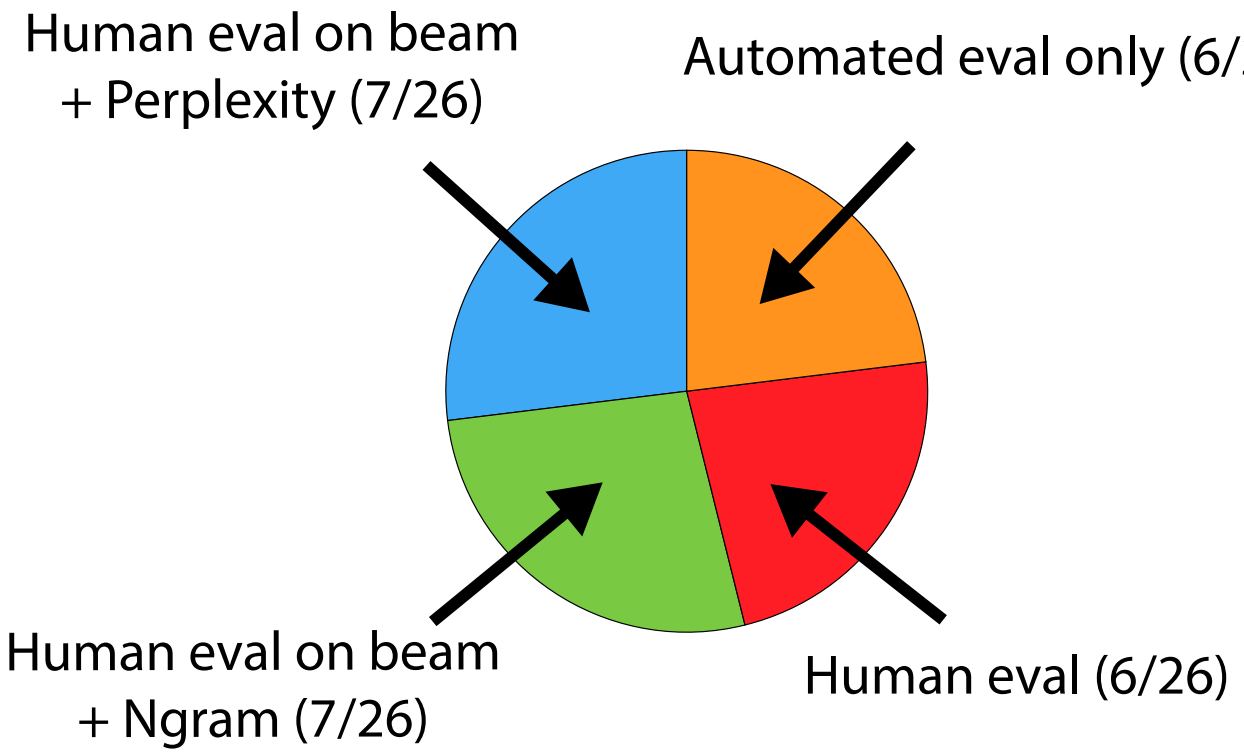
Issue 3: state of evaluation

State of evaluation at ACL 2018



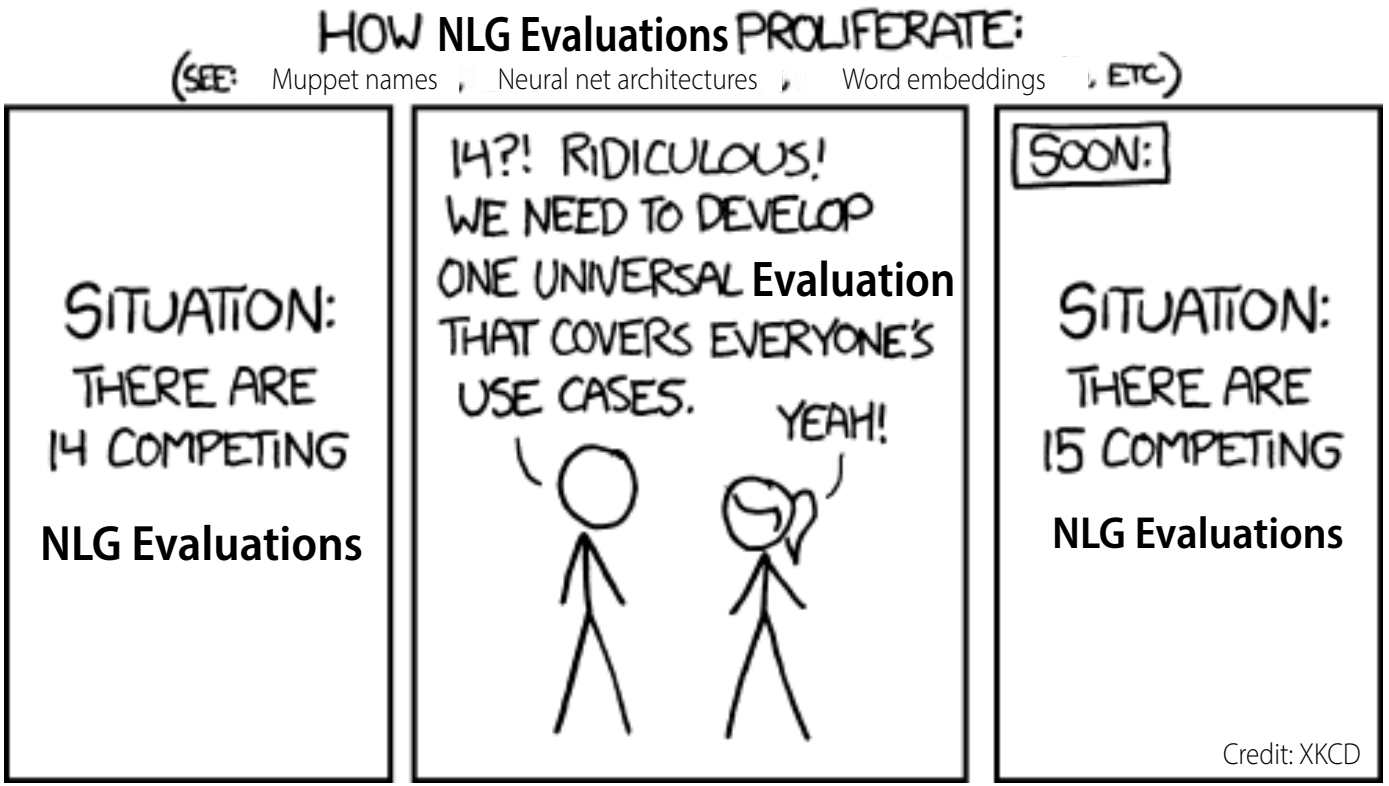
Issue 3: state of evaluation

State of evaluation at ACL 2018

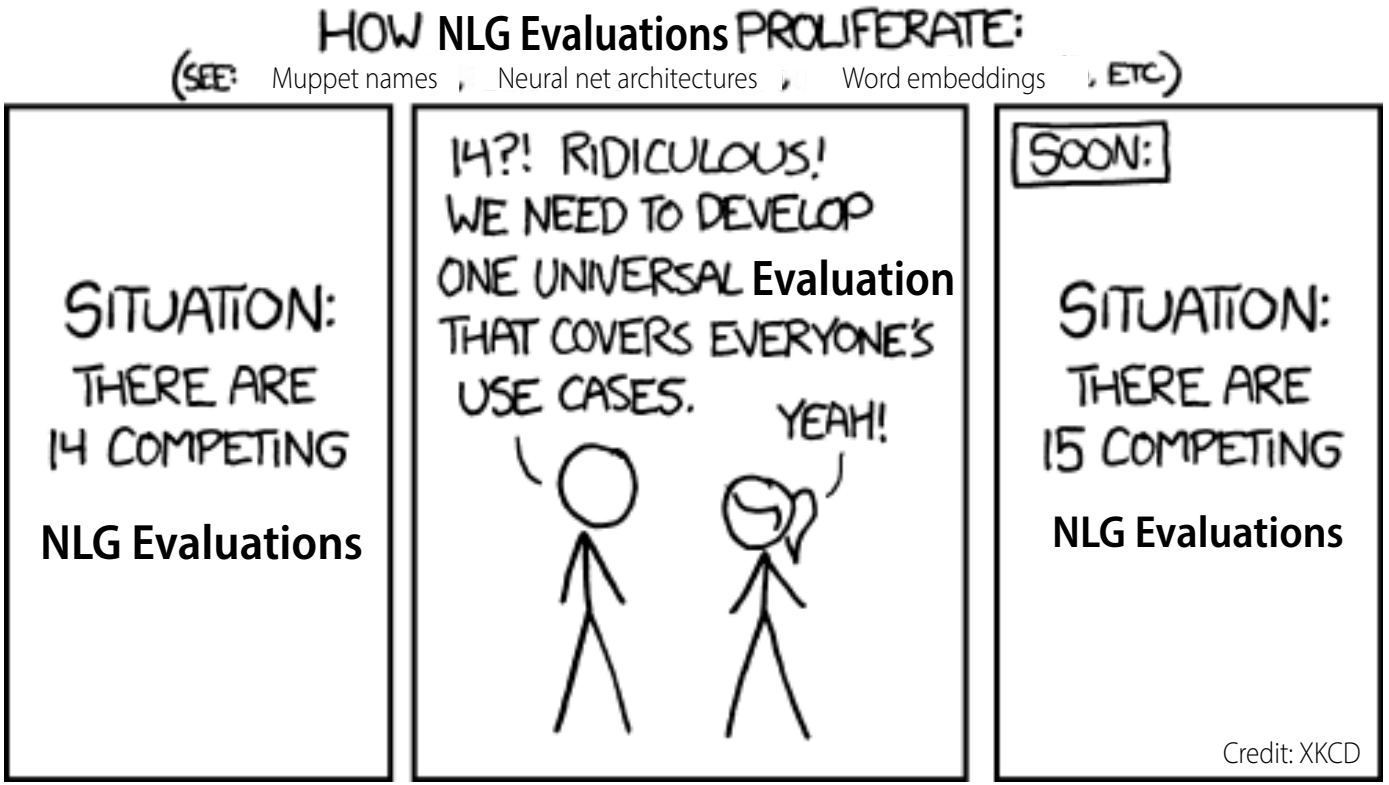


- No obvious consensus in evaluation.

A new evaluation



A new evaluation



We'll show we *need* a new gold-standard evaluation

Existing evaluation schemes:

- **Human evaluation** (and automated approximations)
  - **Pro** : gold-standard for quality, useful contract.
  - **Con** : expensive, can be cheated via memorization / underdiversity.

Existing evaluation schemes:

- **Human evaluation** (and automated approximations)
  - **Pro** : gold-standard for quality, useful contract.
  - **Con** : expensive, can be cheated via memorization / underdiversity.

  Examples (Novikova+ 2017, Gatt+ 2018):

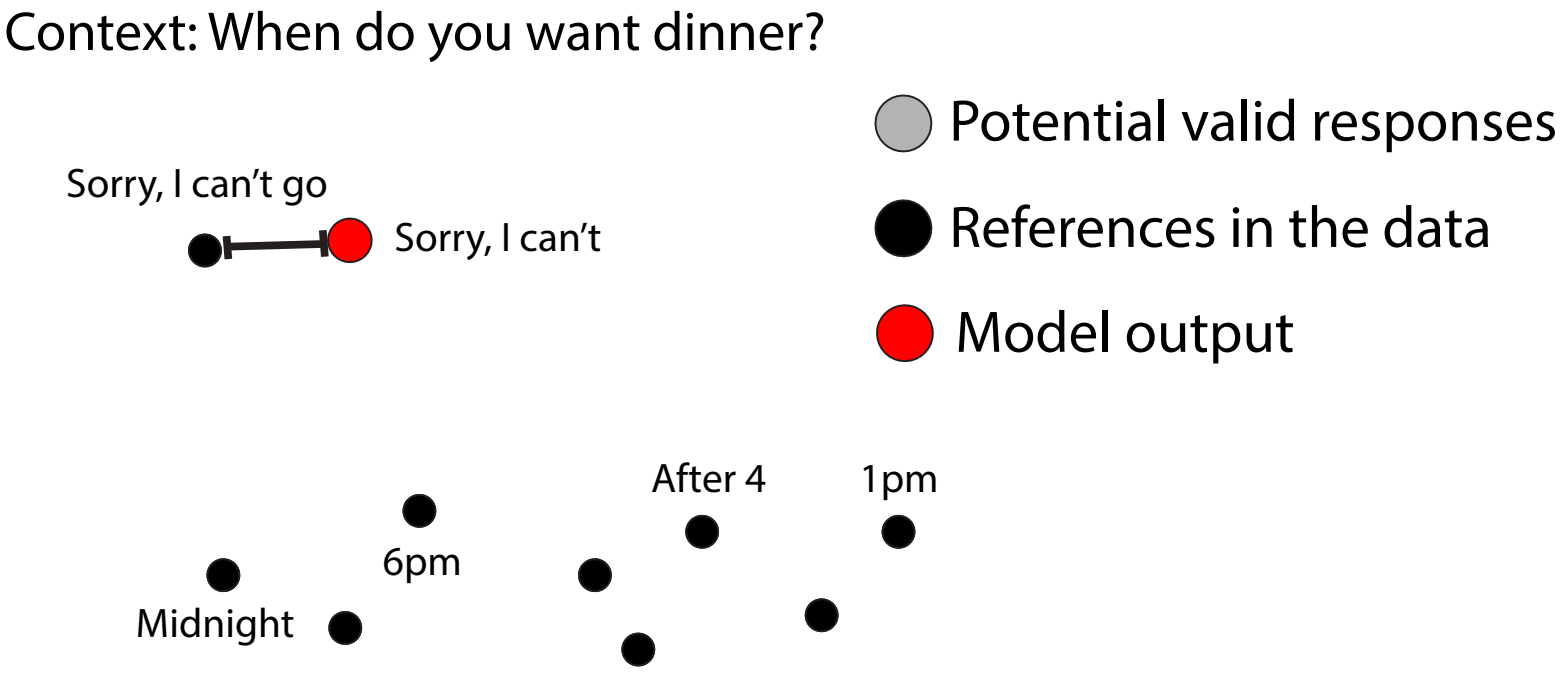
  1. **Fluency**: Text could have been generated by a native speaker.
  2. **Accuracy**: How much information is contained in the utterance?
  3. **Quality**: The overall quality of the utterance.
  4. **Grammaticality**: Parse scores, misspelings, learned models

Existing evaluation schemes:

- **Human evaluation** (and automated approximations)
- **Reference-distances** (BLEU, ROUGE, TER, ngrams)
  - **Pro** : easily measured, intepretable.
  - **Con** : often doesnt measure quality or diversity well.

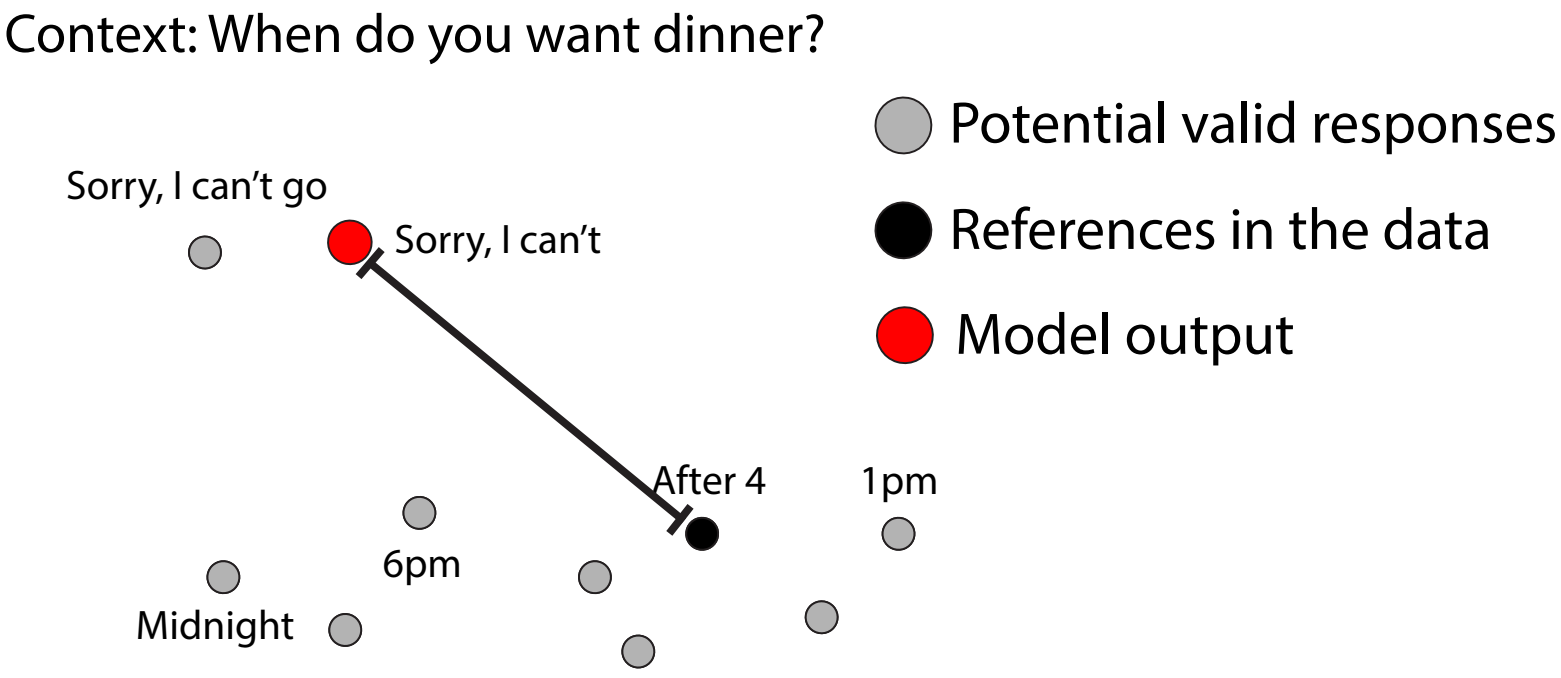
Existing evaluation schemes:

- **Human evaluation** (and automated approximations)
- **Reference-distances** (BLEU, ROUGE, TER, ngrams)



Existing evaluation schemes:

- **Human evaluation** (and automated approximations)
- **Reference-distances** (BLEU, ROUGE, TER, ngrams)



Existing evaluation schemes:

- **Human evaluation** (and automated approximations)
- **Reference-distances** (BLEU, ROUGE, TER, ngrams)
- **Perplexity** (aka KL divergence, KL(p||q))
  - **Pro** : easily measured, guarantees (statistical) goodness of fit
  - **Con** : not interpretable and doesnt guarantee sample quality.

Existing evaluation schemes:

- **Human evaluation** (and automated approximations)
- **Reference-distances** (BLEU, ROUGE, TER, ngrams)
- **Perplexity** (aka KL divergence, KL(p||q))
- **Total Variation** (aka optimal classification, $\|p - q\|_{TV}$)
  - **Pro** : guarantees high-quality samples and goodness of fit.
  - **Con** : nearly impossible to measure.

What we want from evaluation

	Human eval			
Useful contract	✓			
Interpretable	✓			
Detects underdiversity	✗			
Measurable	✓			

What we want from evaluation

	Human eval	Reference Similarity		
Useful contract	✓	✗		
Interpretable	✓	✓		
Detects underdiversity	✗	✗		
Measurable	✓	✓		

What we want from evaluation

	Human eval	Reference Similarity	Perplexity	
Useful contract	✓	✗	✗	
Interpretable	✓	✓	✗	
Detects underdiversity	✗	✗	✓	
Measurable	✓	✓	✓	

What we want from evaluation

	Human eval	Reference Similarity	Perplexity	Total Variation
Useful contract	✓	✗	✗	✓
Interpretable	✓	✓	✗	✓
Detects underdiversity	✗	✗	✓	✓
Measurable	✓	✓	✓	✗

What we want from evaluation

	Human eval	Reference Similarity	Perplexity	Total Variation
Useful contract	✓	✗	✗	✓
Interpretable	✓	✓	✗	✓
Detects underdiversity	✗	✗	✓	✓
Measurable	✓	✓	✓	✗

No reliable, gold-standard evaluation that covers all goals

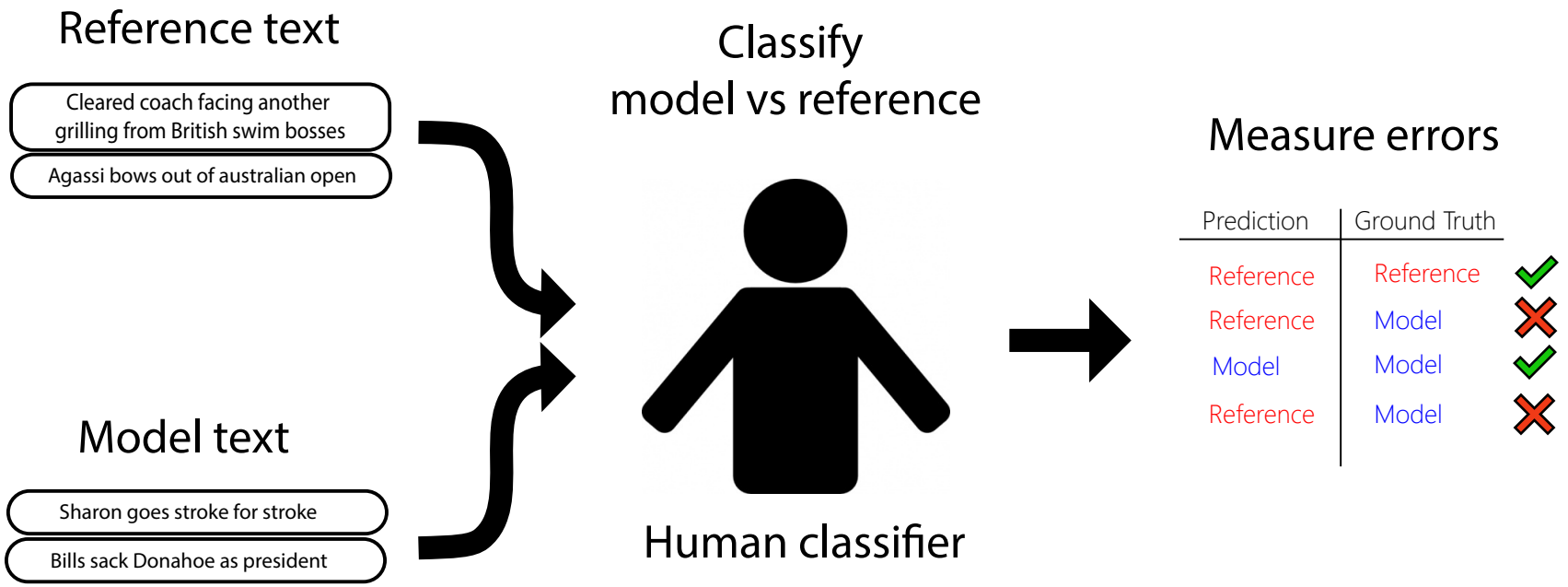
Roadmap

Part 1: Problems with the status quo

Part 2: Rigorous and practical diversity evaluation

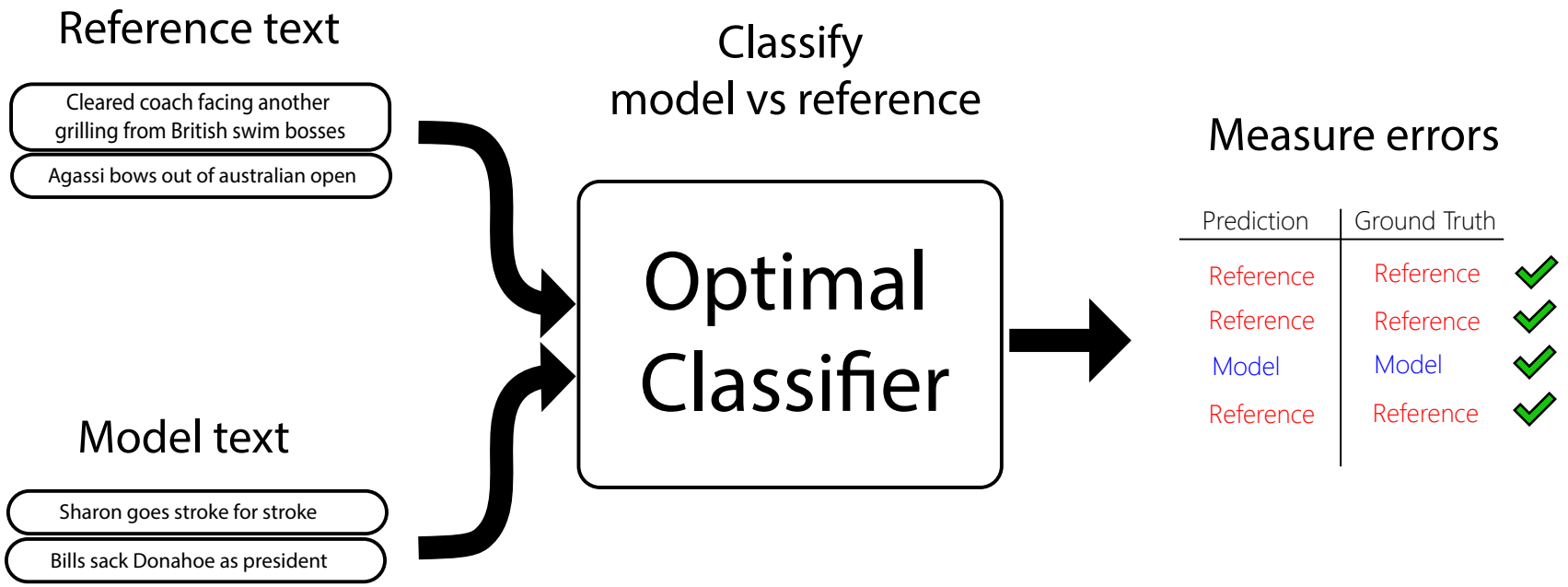
Part 3: Open problems

Optimal classification as a solution



Humans cannot easily identify under-diverse / plagiarized text

Optimal classification as a solution



Optimal classifier can identify
low-quality **and** memorized samples

Optimal classifier is unavailable

- **Rigorous**: optimal classification error is related to the total variation distance $\|p - q\|_{TV}$.

Optimal classifier is unavailable

- **Rigorous**: optimal classification error is related to the total variation distance $\|p - q\|_{TV}$.
- **Interpretable**: classifiation error is a straightforward quantity to interpret.

Optimal classifier is unavailable

- **Rigorous**: optimal classification error is related to the total variation distance $\|p - q\|_{TV}$.
- **Interpretable**: classifiation error is a straightforward quantity to interpret.
- **Useful contract**: guarantees downstream performance

For any task with bounded loss $\ell(y) < M$,

$$E_{p_{model}}[\ell(y)] \leq E_{p_{ref}}[\ell(y)] + M(0.5 - \text{Error})$$

Optimal classifier is unavailable

- **Rigorous**: optimal classification error is related to the total variation distance $\|p - q\|_{TV}$.
- **Interpretable**: classifiation error is a straightforward quantity to interpret.
- **Useful contract**: guarantees downstream performance

For any task with bounded loss $\ell(y) < M$,

$$E_{p_{model}}[\ell(y)] \leq E_{p_{ref}}[\ell(y)] + M(0.5 - \text{Error})$$

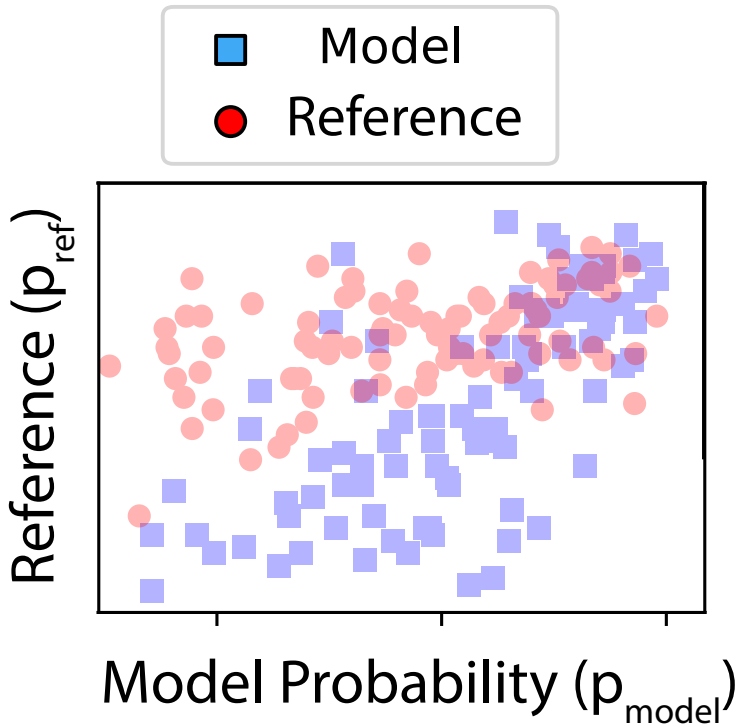
Loss incurred by model | Loss incurred by reference | Gap scales with error

Optimal classifier is unavailable

- **Rigorous**: optimal classification error is related to the total variation distance $\|p - q\|_{TV}$.
- **Interpretable**: classifiation error is a straightforward quantity to interpret.
- **Useful contract**: guarantees downstream performance

- **Easily measurable(?)**: this has been thought to be difficult or impossible to measure [Chaganty+ 17, Novikova+17]

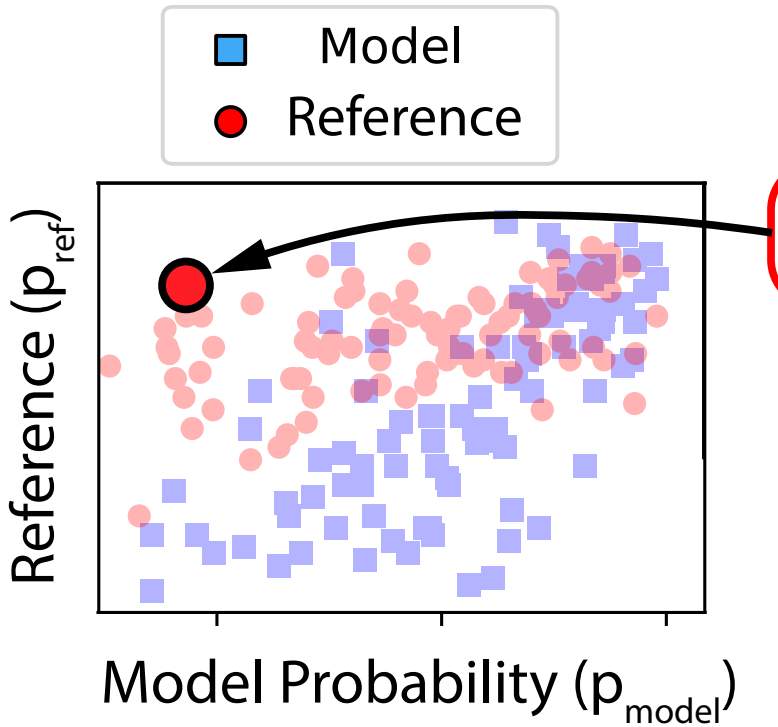
Two features suffice



Input sentence

Cleared coach facing another grilling from British swim bosses

Two features suffice

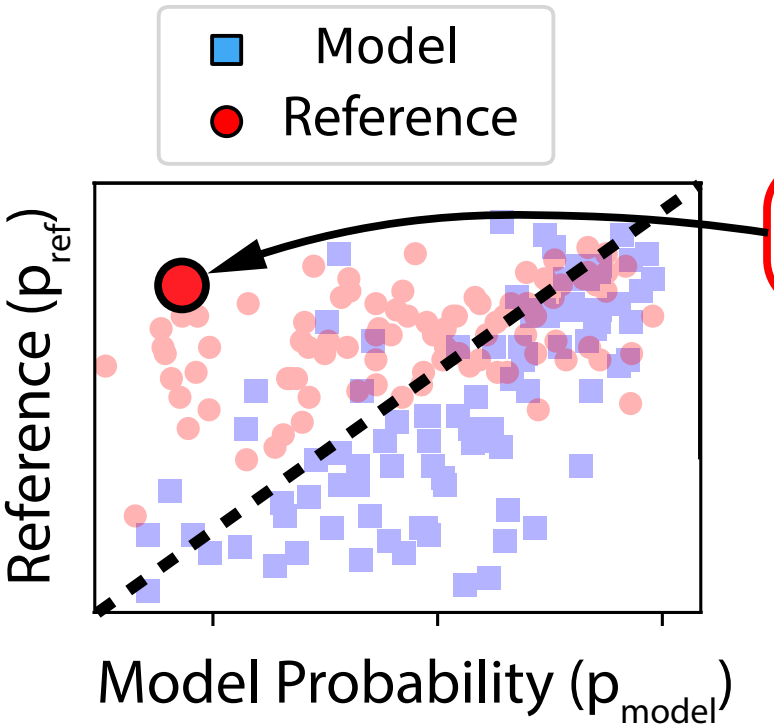


Input sentence

Cleared coach facing another grilling from British swim bosses

1. Map to ($p_{ref}(y|x)$, $p_{model}(y|x)$)

Two features suffice



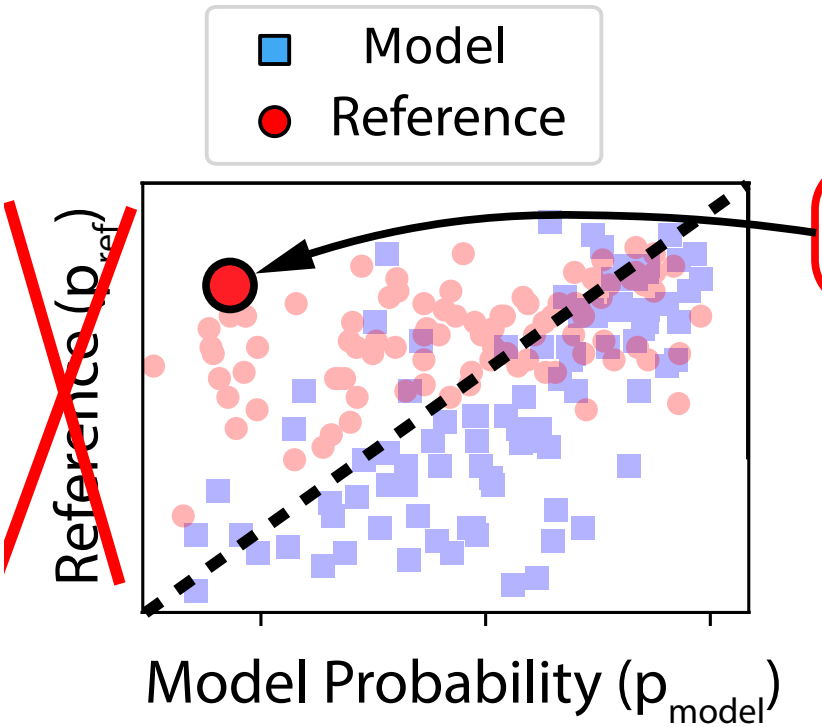
Input sentence

Cleared coach facing another grilling from British swim bosses

1. Map to ($p_{ref}(y|x)$, $p_{model}(y|x)$)

2. Classify with $p_{ref}(y|x) > p_{model}(y|x)$

Two features suffice



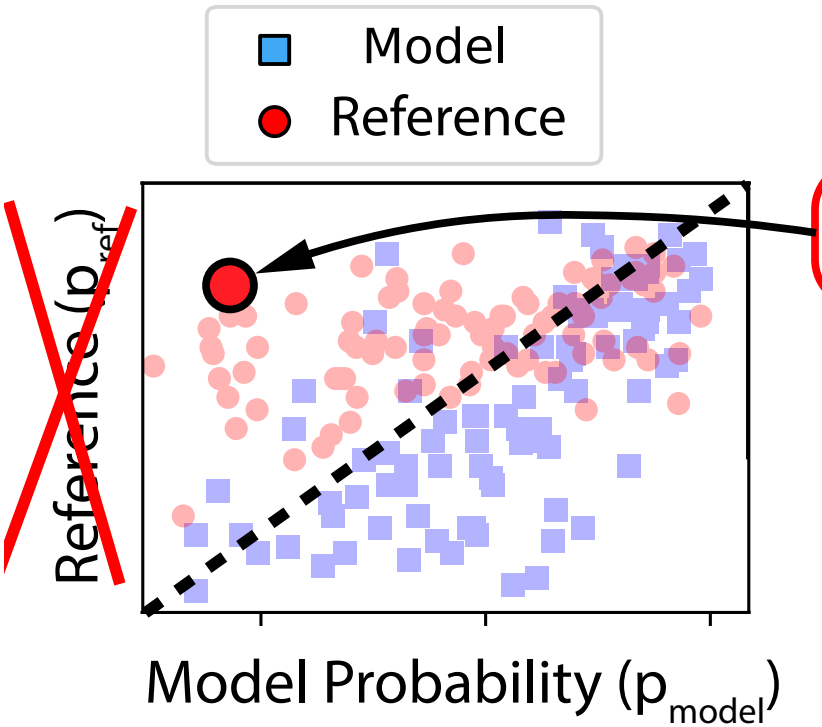
Input sentence

Cleared coach facing another grilling from British swim bosses

1. Map to ($p_{ref}(y|x)$, $p_{model}(y|x)$)
2. Classify with $p_{ref}(y|x) > p_{model}(y|x)$

Problem: we don't observe $p_{ref}(y|x)$

Two features suffice



Input sentence

Cleared coach facing another grilling from British swim bosses

1. Map to ($p_{ref}(y|x)$, $p_{model}(y|x)$)
2. Classify with $p_{ref}(y|x) > p_{model}(y|x)$

Problem: we don't observe $p_{ref}(y|x)$

Q: can we replace $p_{ref}(y|x)$ with something else?

Replacing p_{ref}

Idea: What if we ask human annotators for $p_{ref}(y|x)$

Crowdworker Prompt:

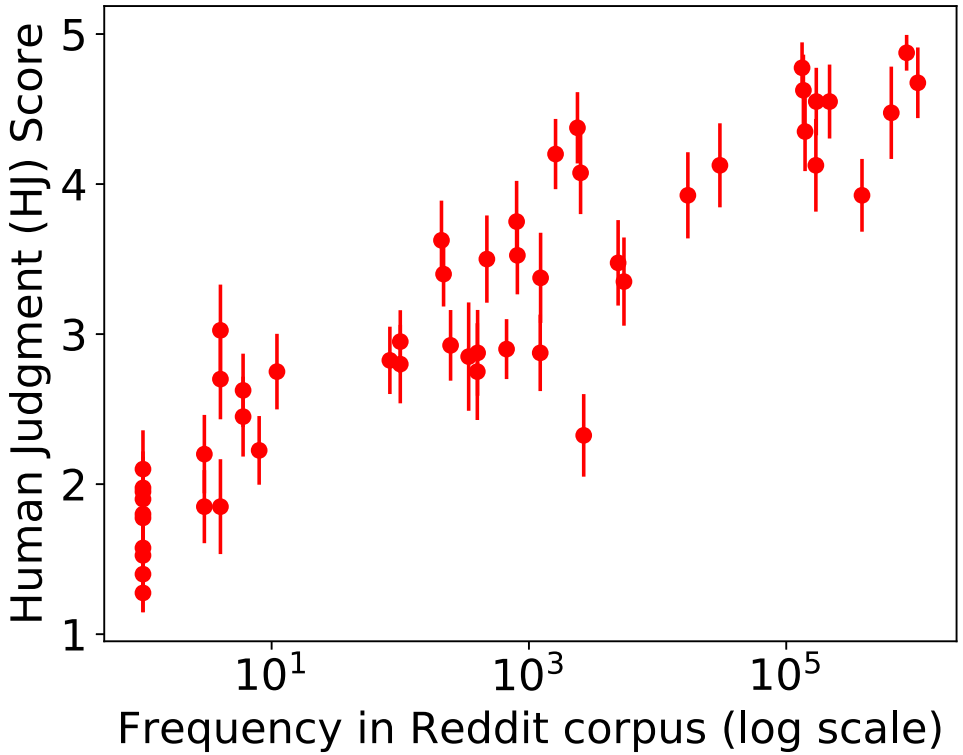
Given the sentence "X"
How typical is the exact response "Y" on a scale from 0 (never) to 5 (very typical)

Our measurement: HJ(y,x)

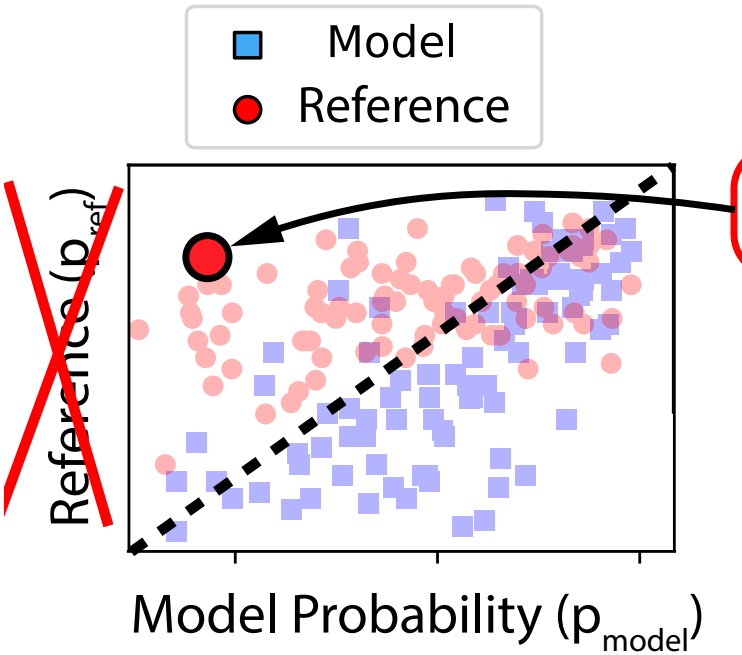
Average scores across 20 crowdworkers

Replacing p_{ref}

Idea: What if we ask human annotators for $p_{\text{ref}}(y|x)$



Two features suffice



Input sentence

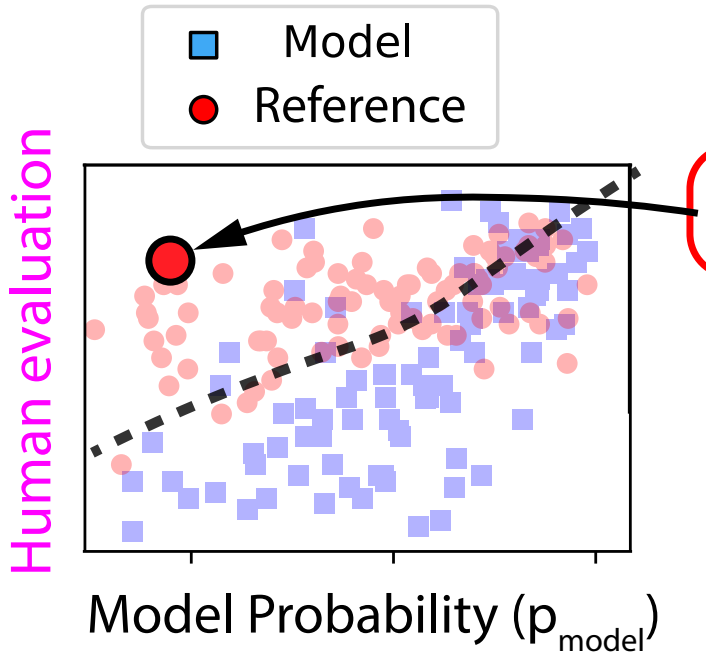
Cleared coach facing another grilling from British swim bosses

1. Map to ($p_{ref}(y|x)$, $p_{model}(y|x)$)

2. Classify with $p_{ref}(y|x) > p_{model}(y|x)$

Two features suffice

Approach: Treat HJ(y,x) as a monotone transformation of $p_{ref}(y|x)$



Input sentence

Cleared coach facing another grilling from British swim bosses

1. Map to (HJ(y,x), $p_{model}(y|x)$)

2. Use the optimal monotone classifier on (HJ(y,x), $p_{model}(y|x)$)

Human unified statistical evaluation

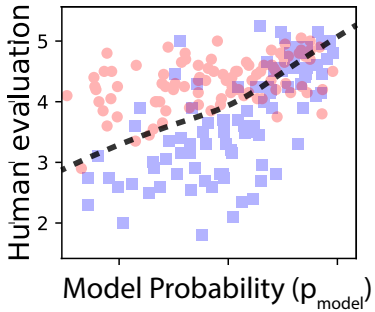
Input text

(mix of reference and model text)

- Cleared coach facing another grilling from British swim bosses
- Agassi bows out of australian open
- Sharon goes stroke for stroke
- Bills sack Donahoe as president

Classify (HJ(y,x), $p_{model}(y|x)$)

(Identify if text came from a model)



Estimate error rate

(on held-out data)

Prediction	Ground Truth	
Reference	Reference	✓
Model	Model	✓
Model	Model	✓
Model	Model	✓

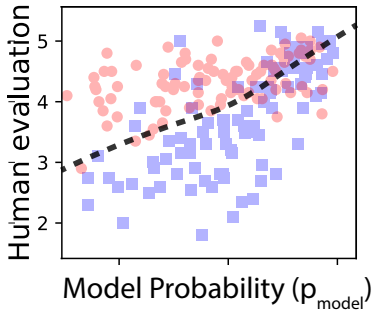
HUSE: Held-out error estimate

Human unified statistical evaluation

Input text
(mix of reference and model text)

- Cleared coach facing another grilling from British swim bosses
- Agassi bows out of australian open
- Sharon goes stroke for stroke
- Bills sack Donahoe as president

Classify (HJ(y,x), $p_{model}(y|x)$)
(Identify if text came from a model)



Estimate error rate
(on held-out data)

Prediction	Ground Truth	
Reference	Reference	✓
Model	Model	✓
Model	Model	✓
Model	Model	✓

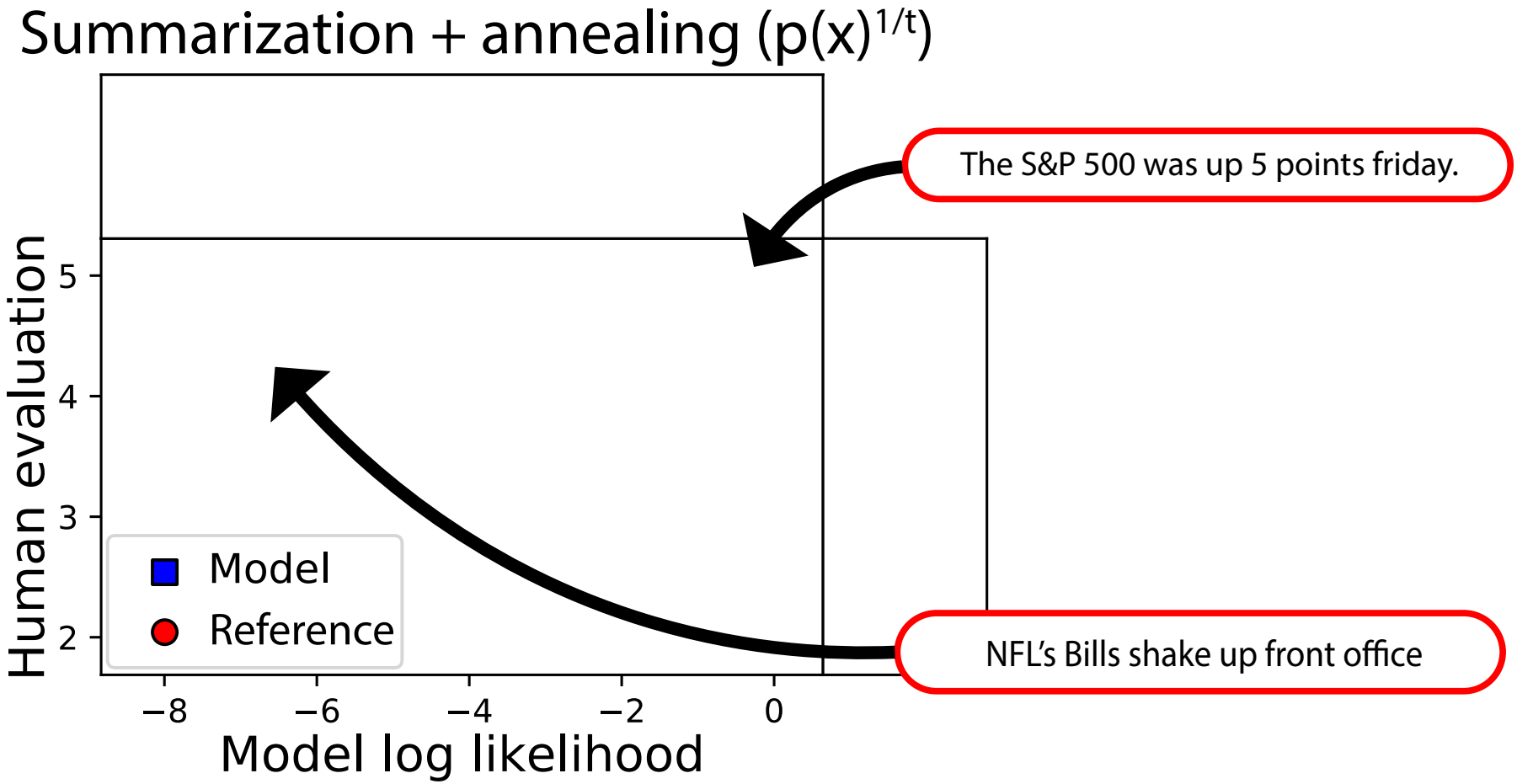
HUSE: Held-out error estimate

Properties of HUSE

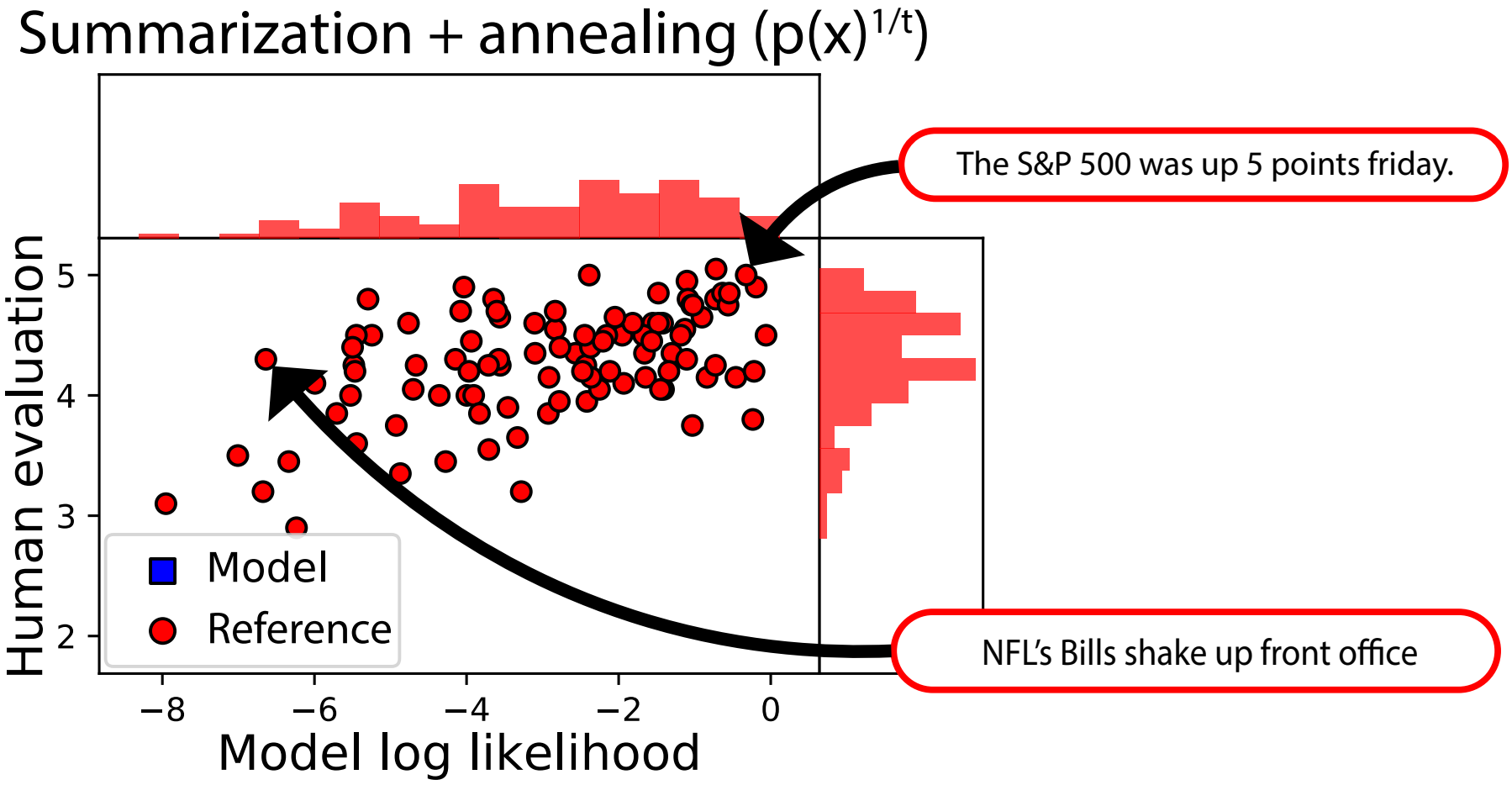
1. **HUSE** ≥ Optimal classification error
   (HUSE will never reject good models)

2. Human classification error ≥ **HUSE**
   (HUSE will never accept models distinguishable to humans)

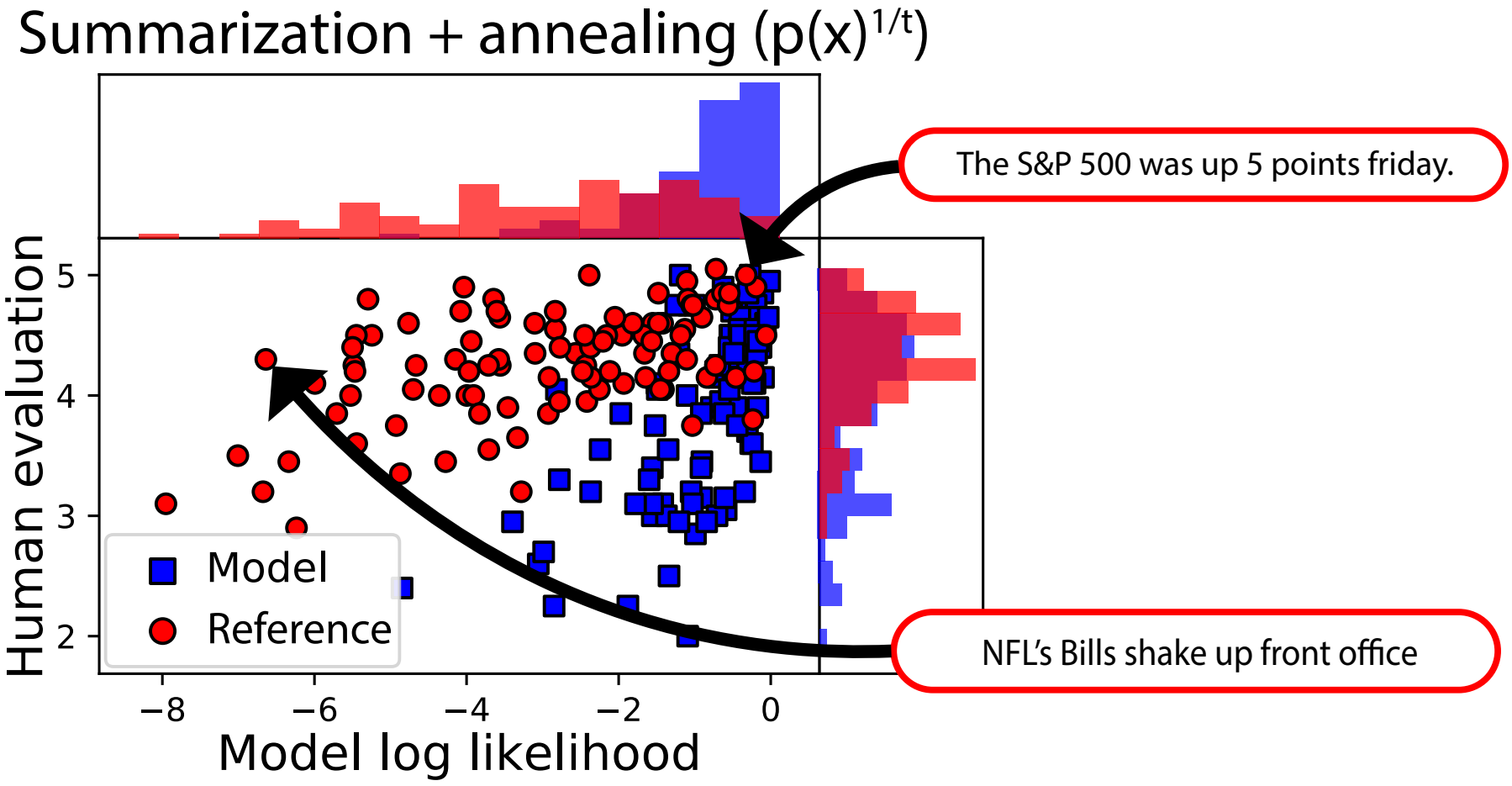
Case study: summarization



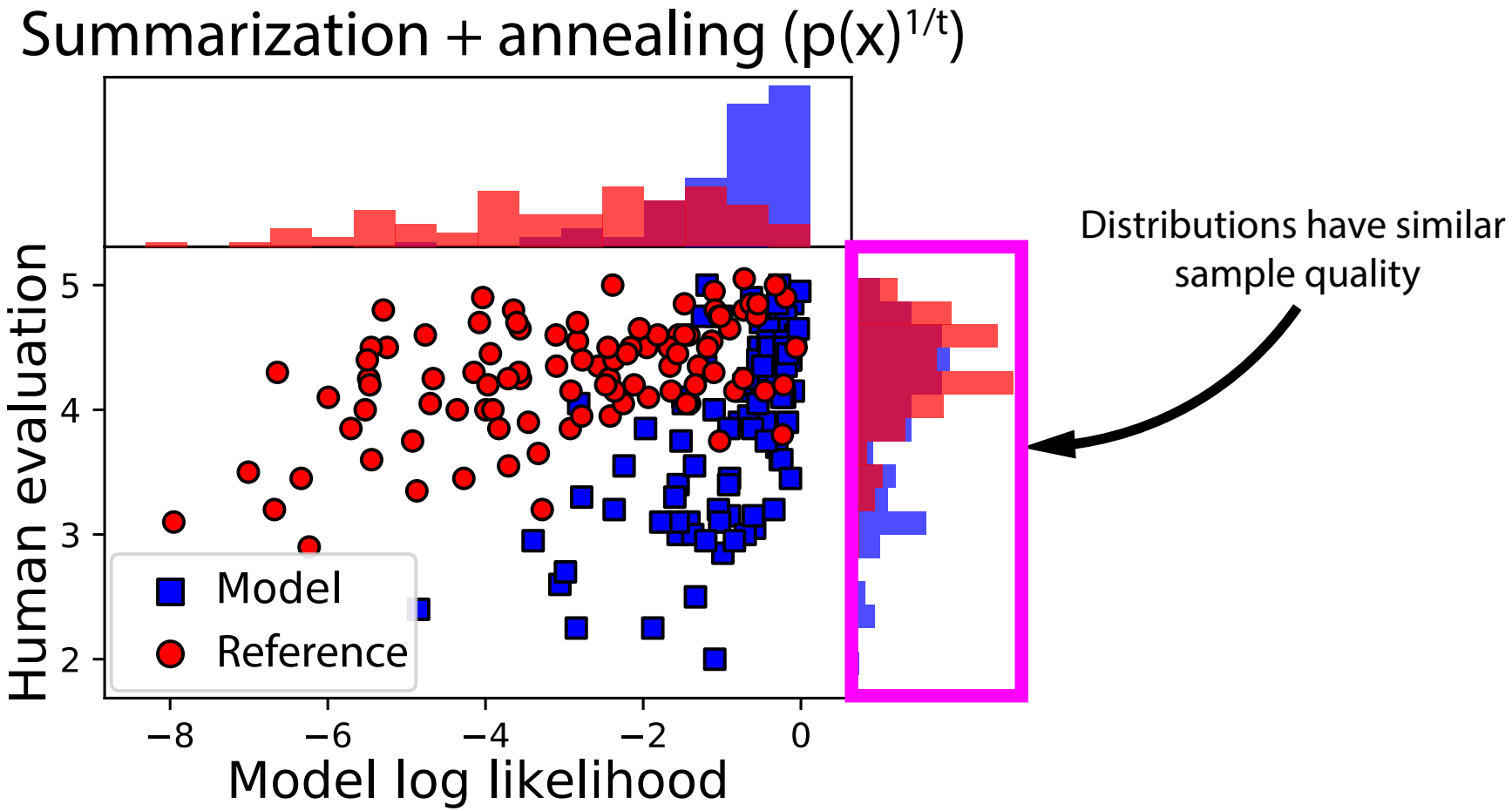
Case study: summarization



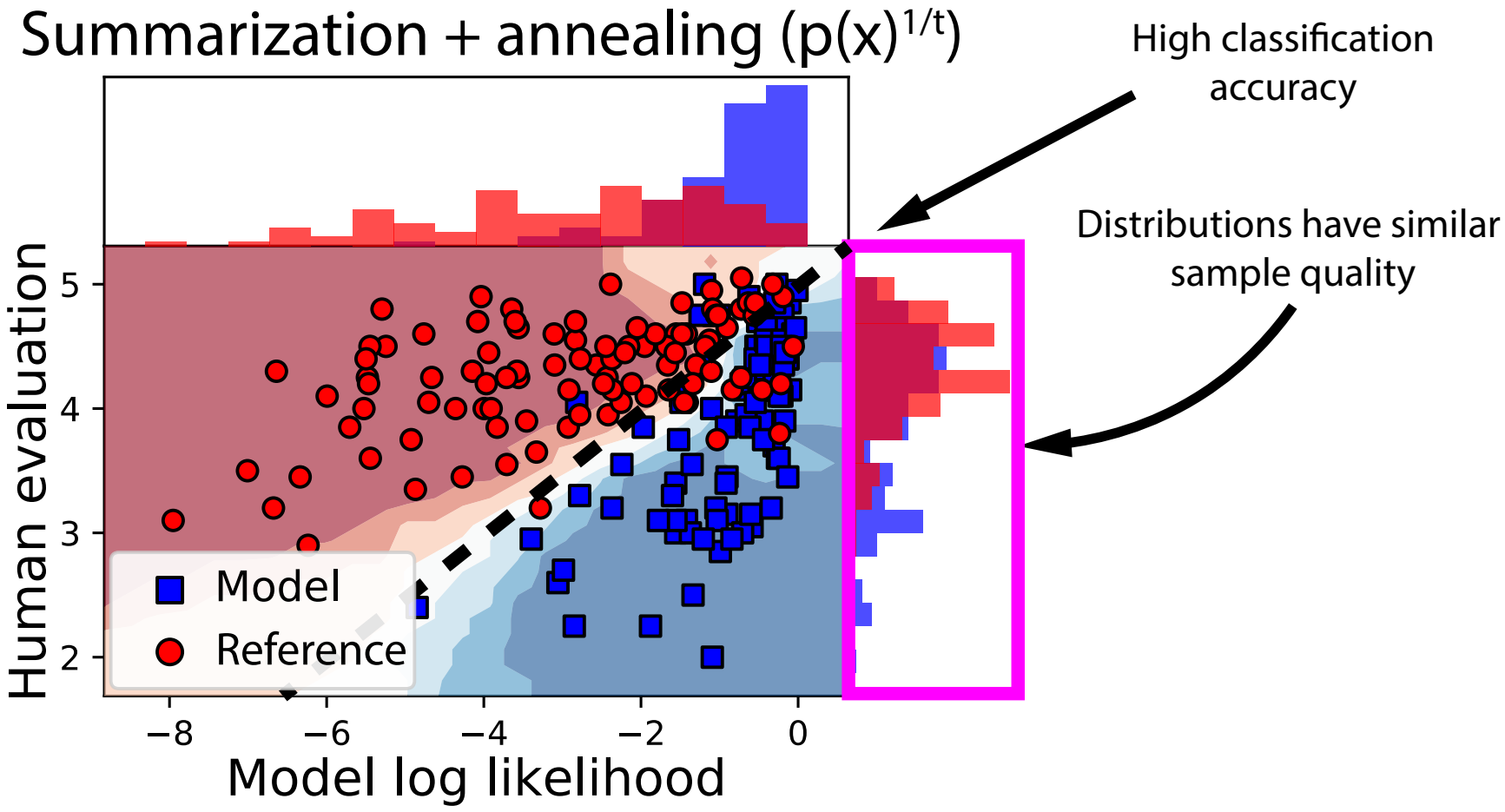
Case study: summarization



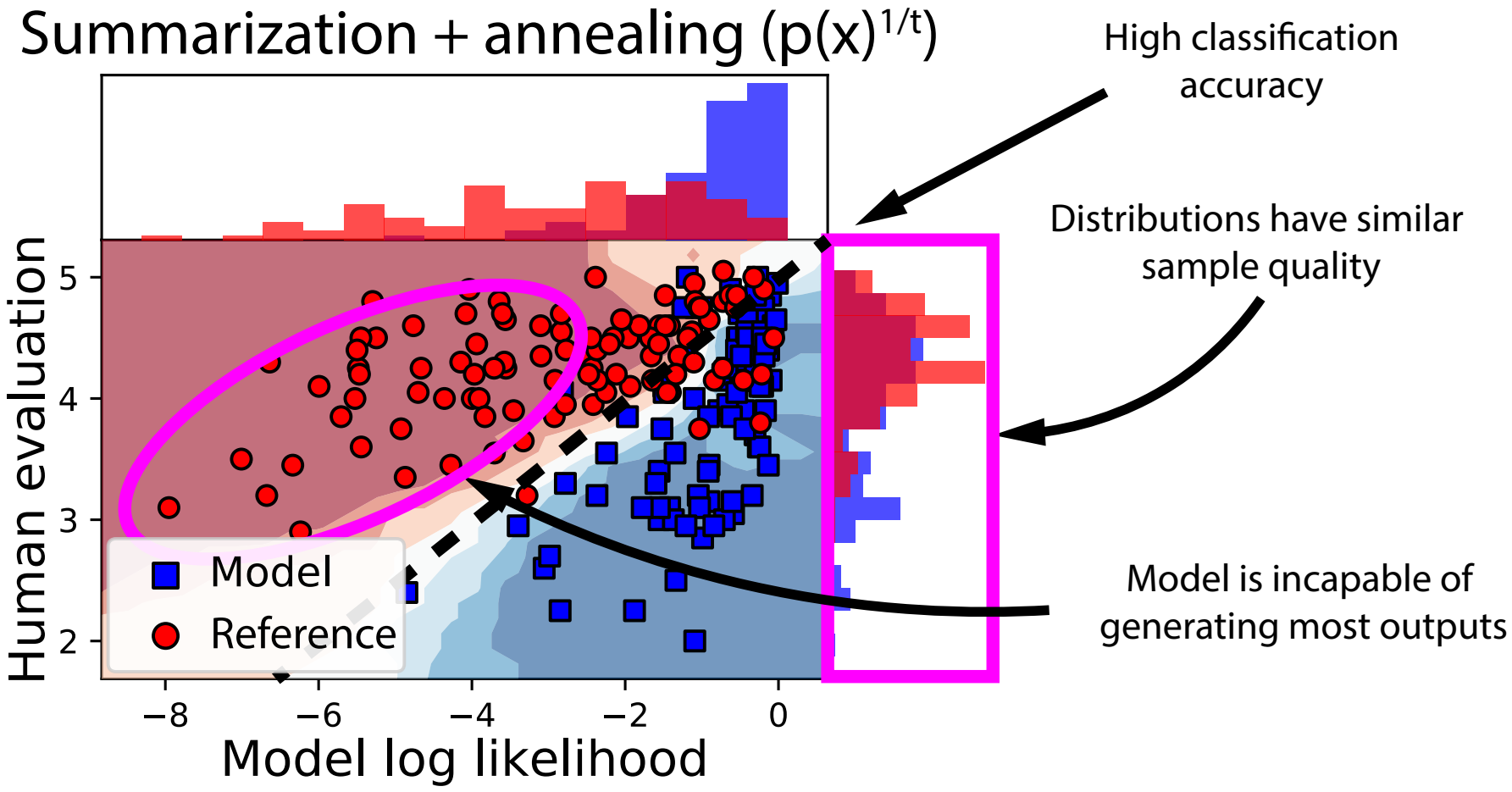
Case study: summarization



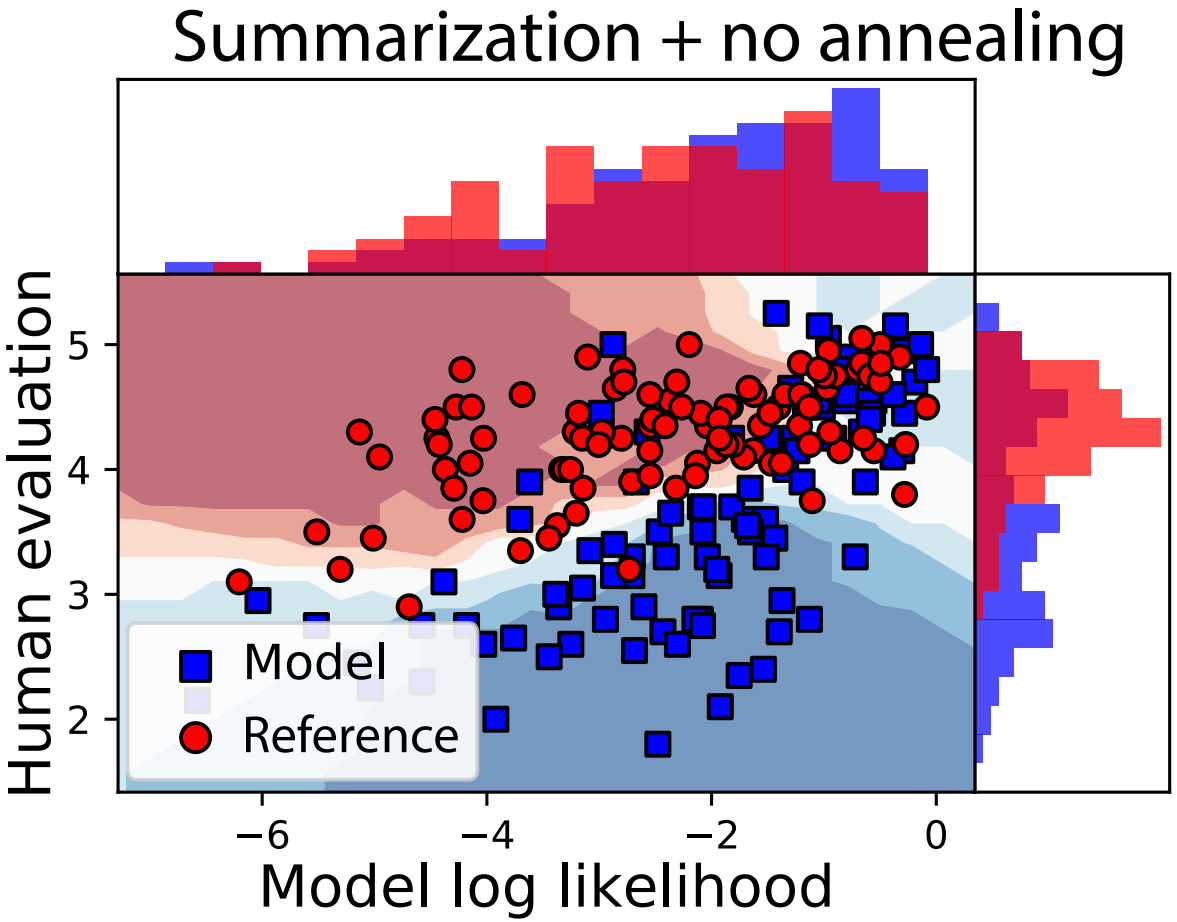
Case study: summarization



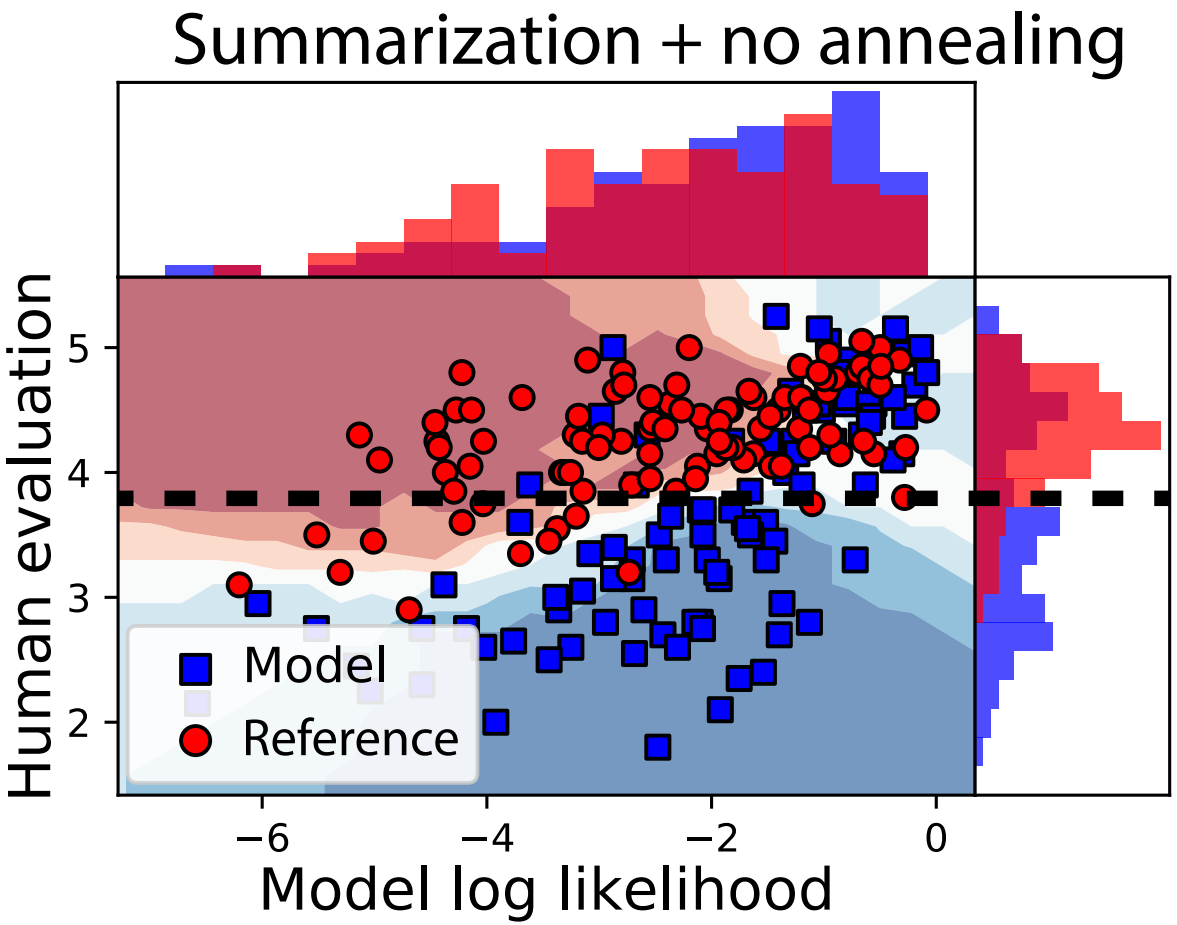
Case study: summarization



Case study: summarization

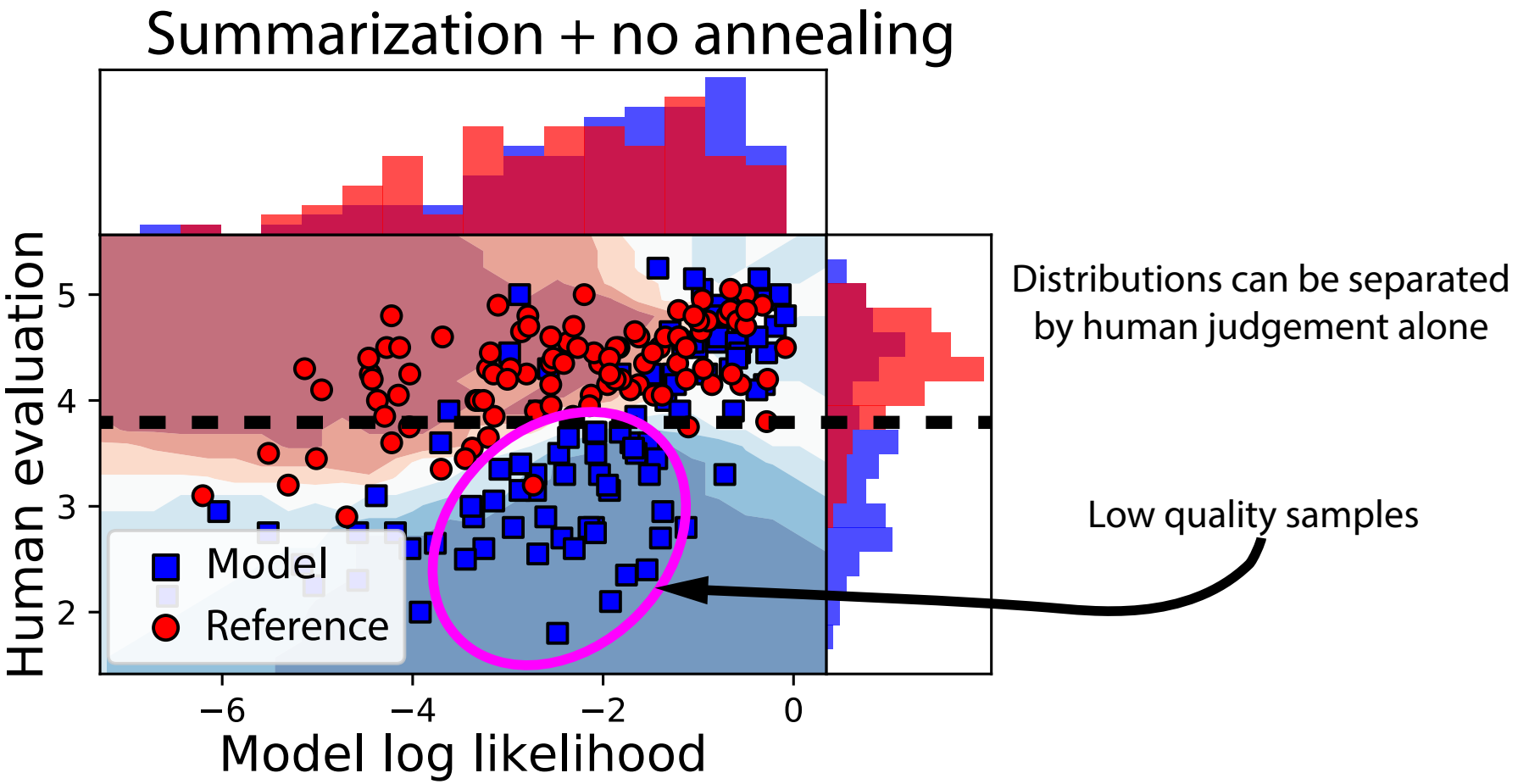


Case study: summarization

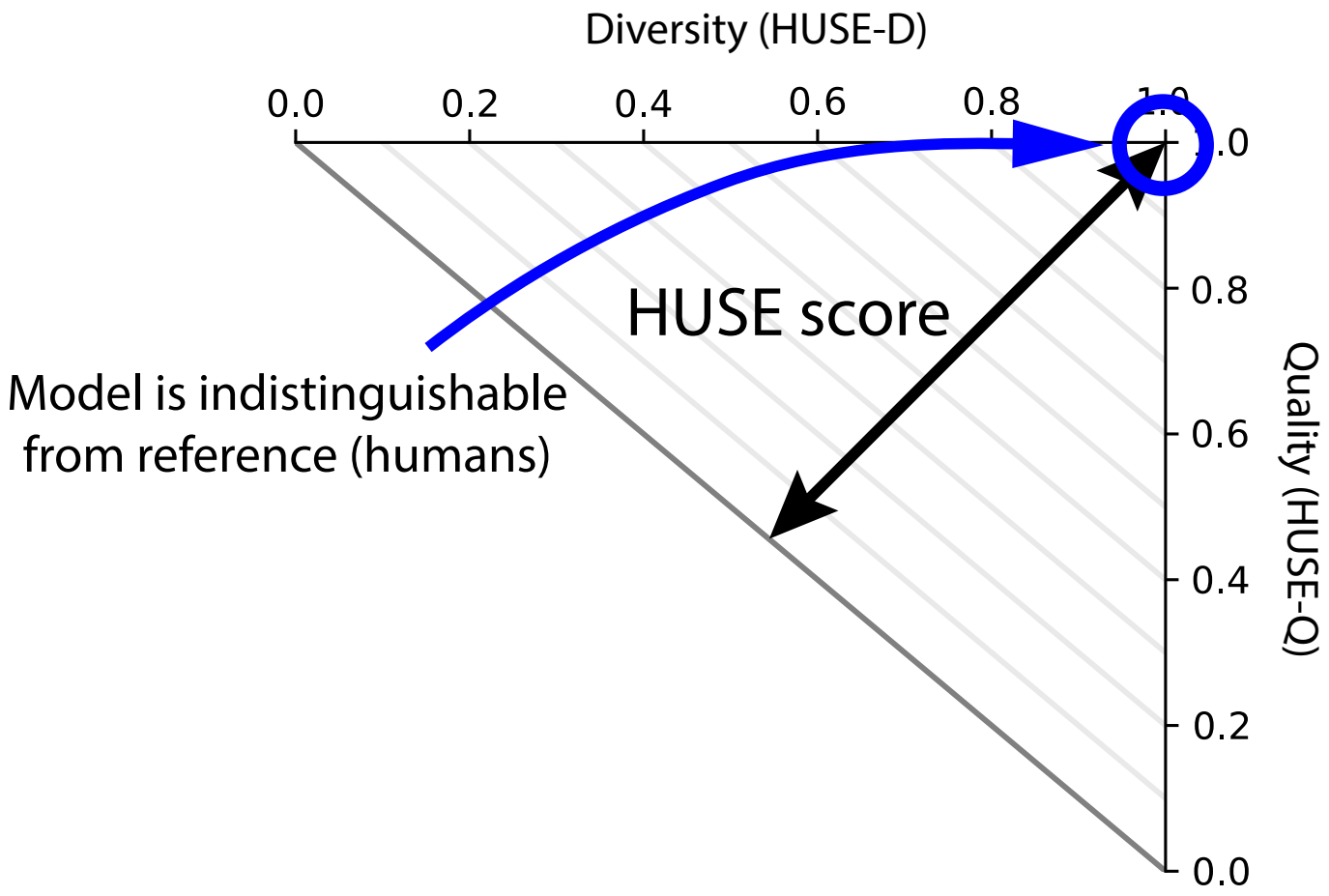


Distributions can be separated by human judgement alone

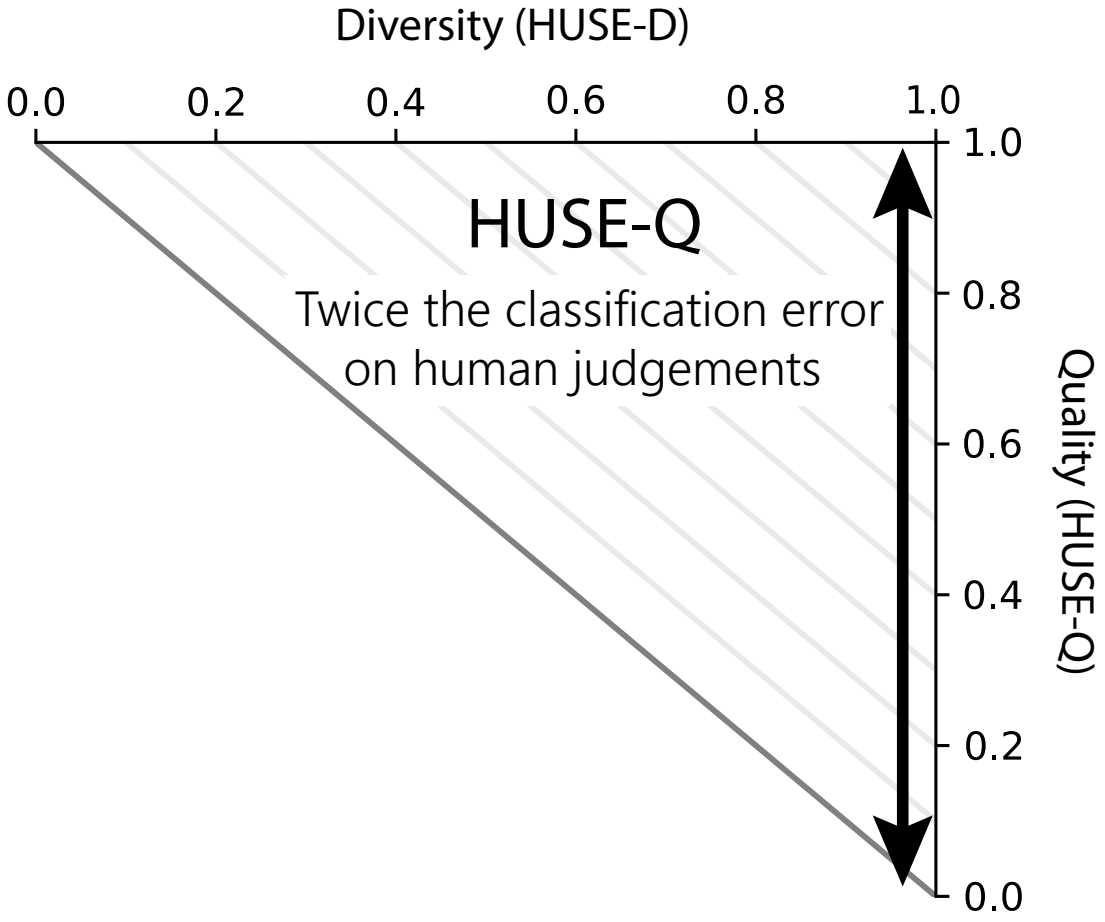
Case study: summarization



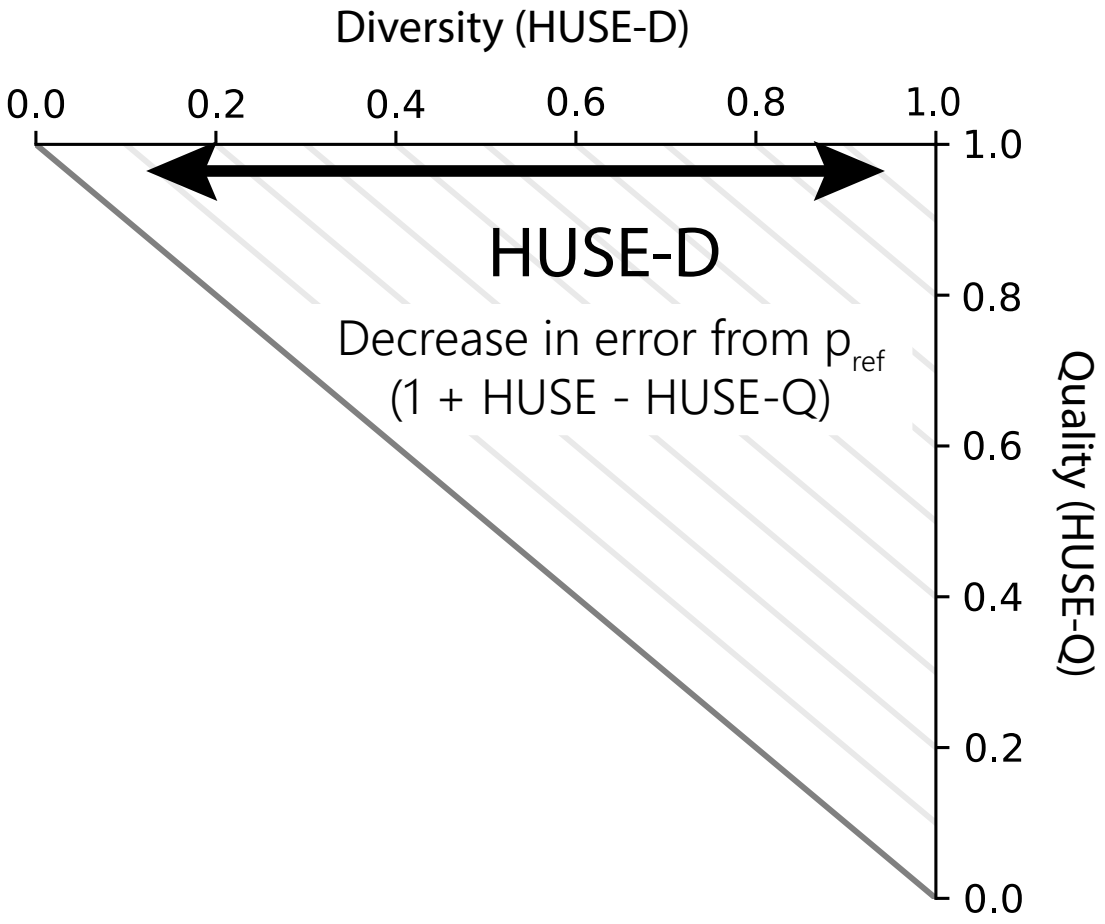
Quality-diversity tradeoffs



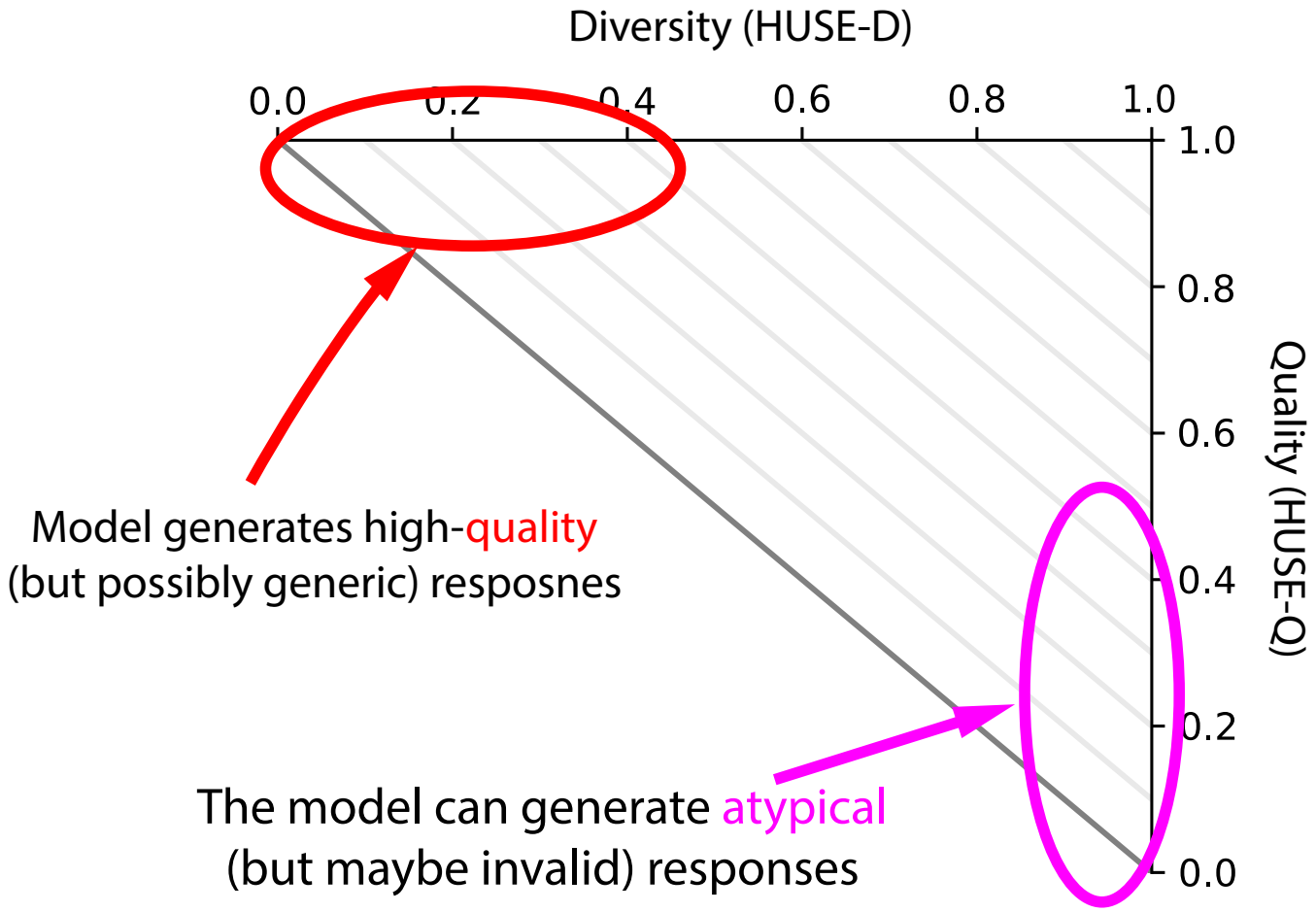
Quality-diversity tradeoffs



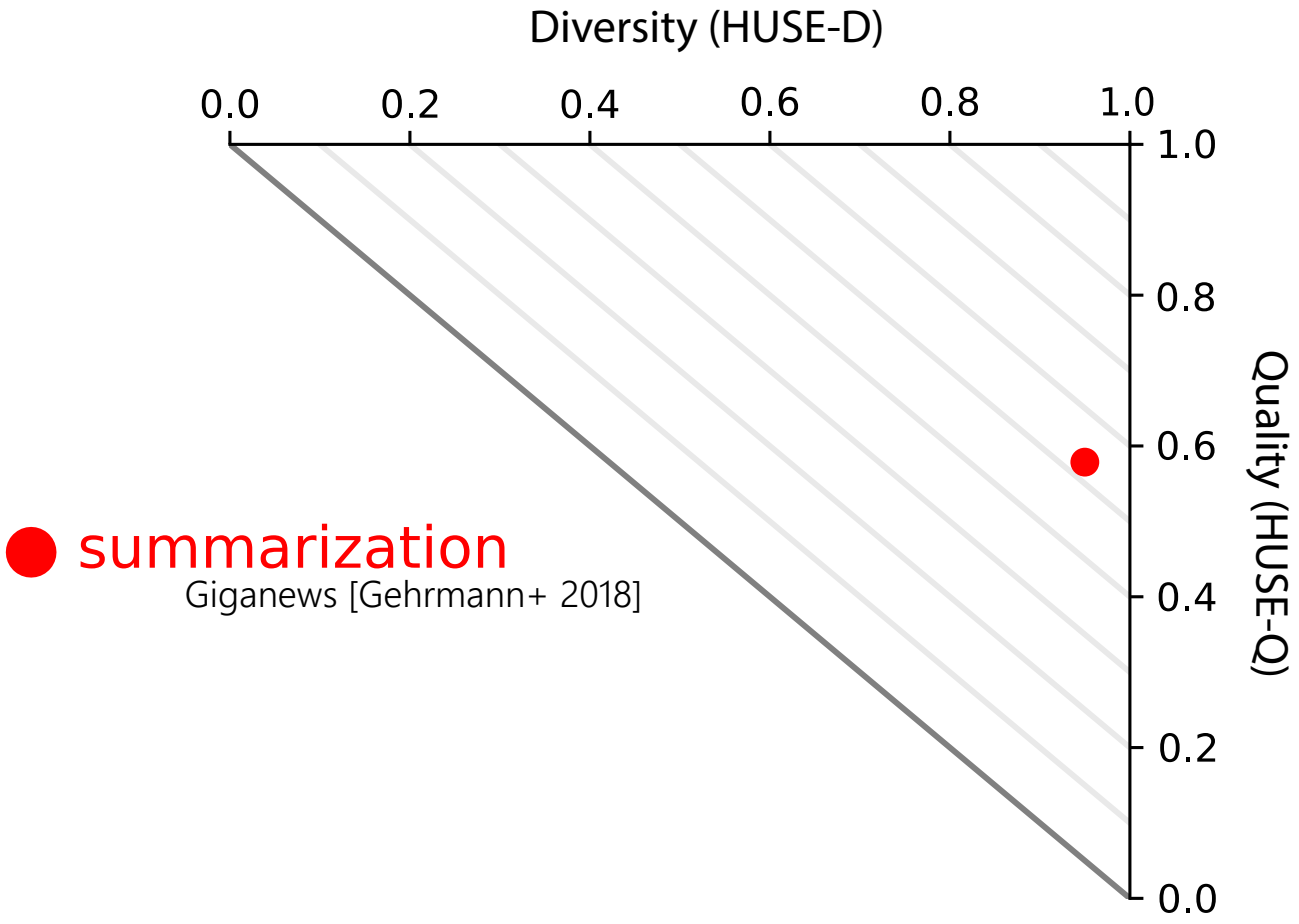
Quality-diversity tradeoffs



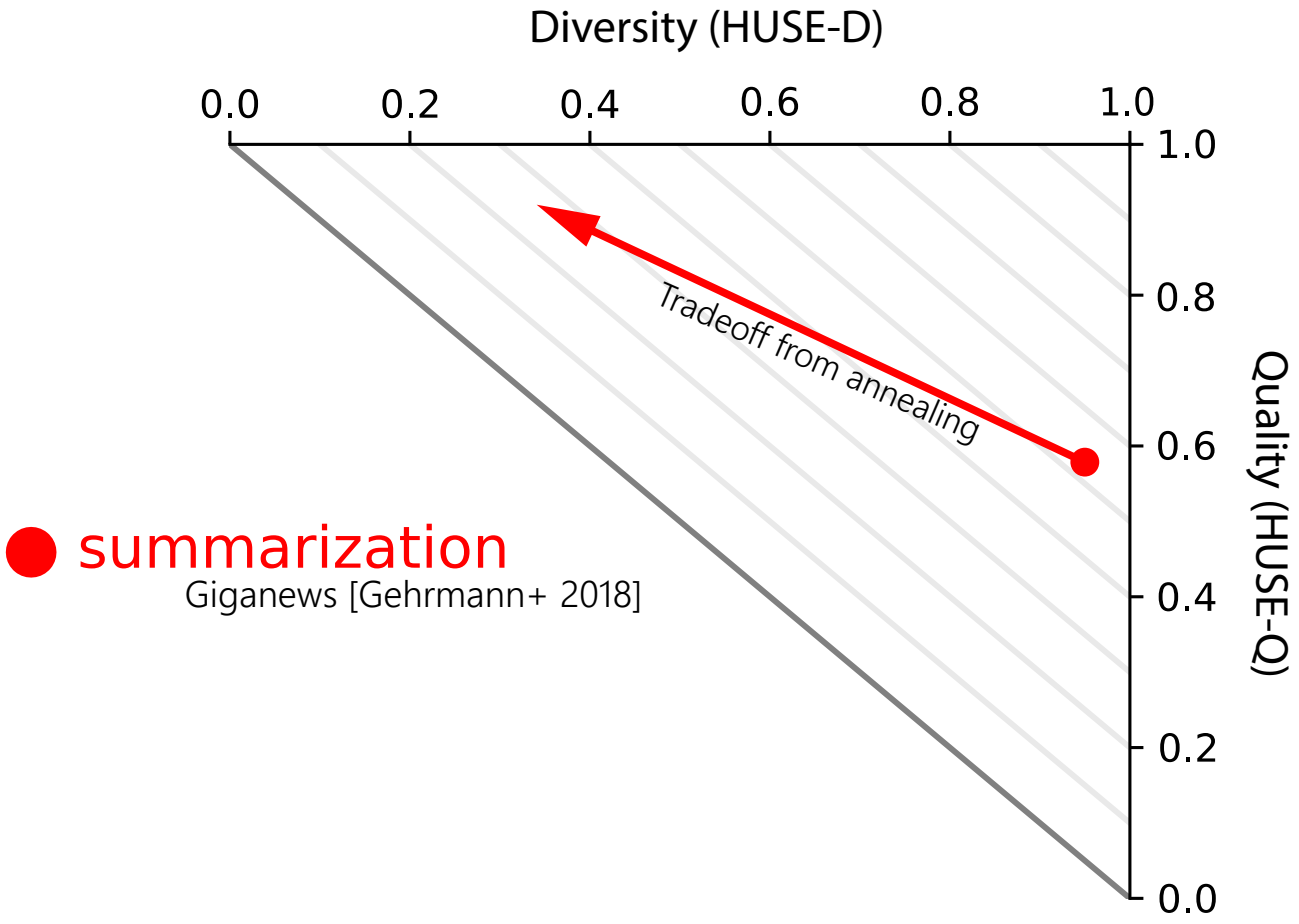
Quality-diversity tradeoffs



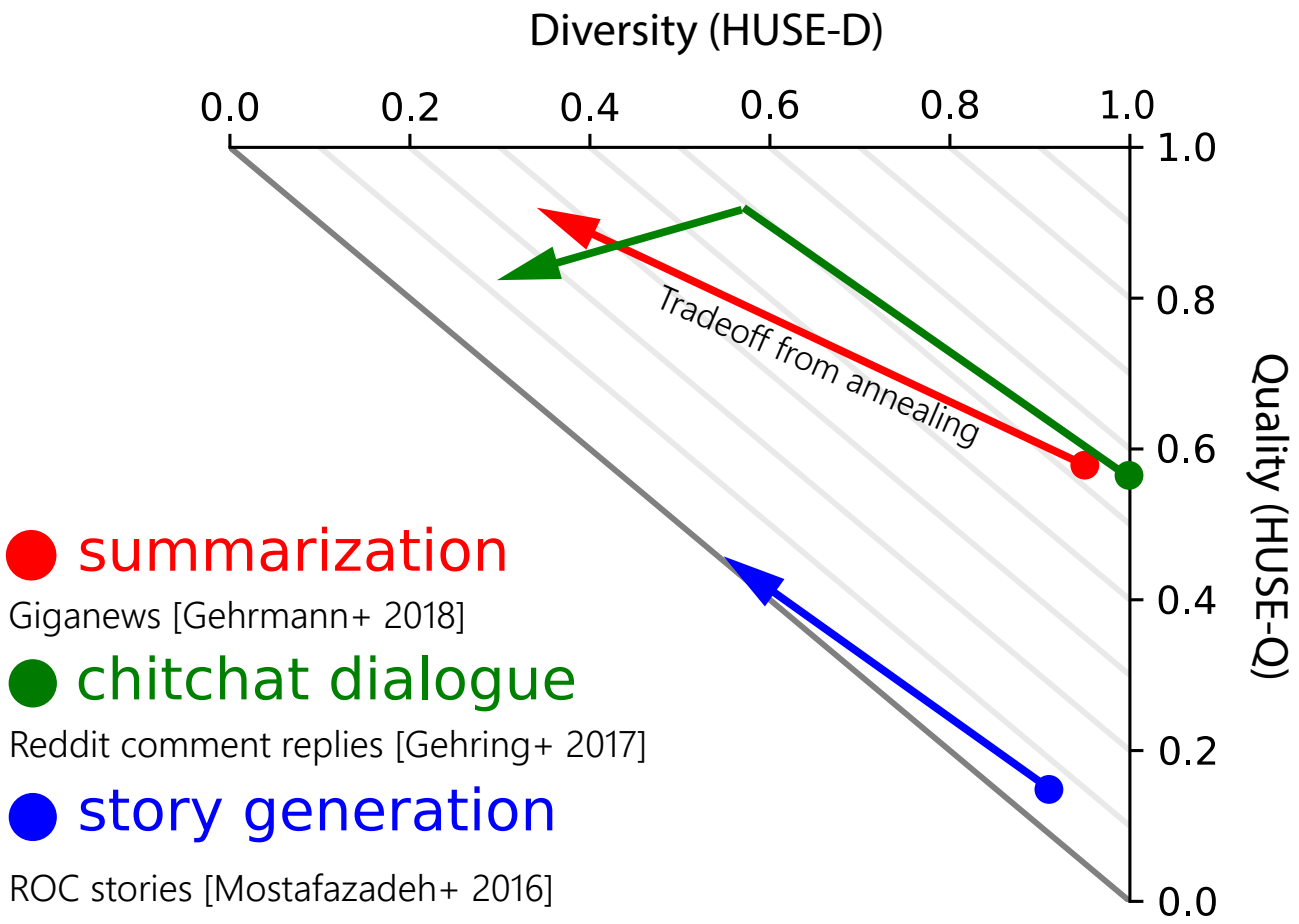
Quality-diversity tradeoffs



Quality-diversity tradeoffs



Quality-diversity tradeoffs



● summarization

Giganews [Gehrmann+ 2018]

● chitchat dialogue

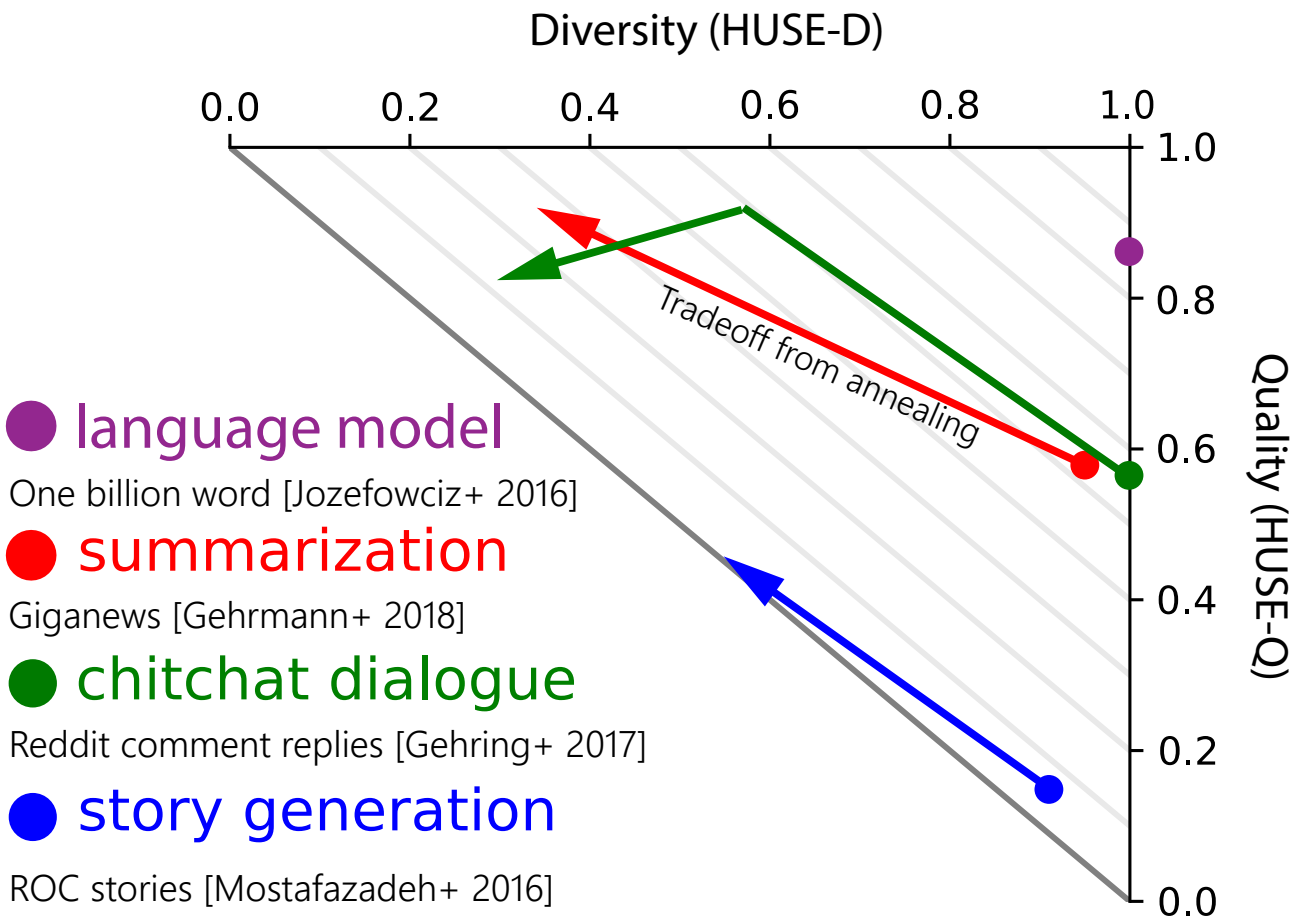
Reddit comment replies [Gehring+ 2017]

● story generation

ROC stories [Mostafazadeh+ 2016]

Techniques to improve quality come at a great cost to diversity

Quality-diversity tradeoffs



Techniques to improve quality come at a great cost to diversity

Roadmap

Part 1: Problems with the status quo

Part 2: Rigorous and practical diversity evaluation

Part 3: Open problems

Fundamental challenges

1. Diversity and semantics

2. Improving training objectives

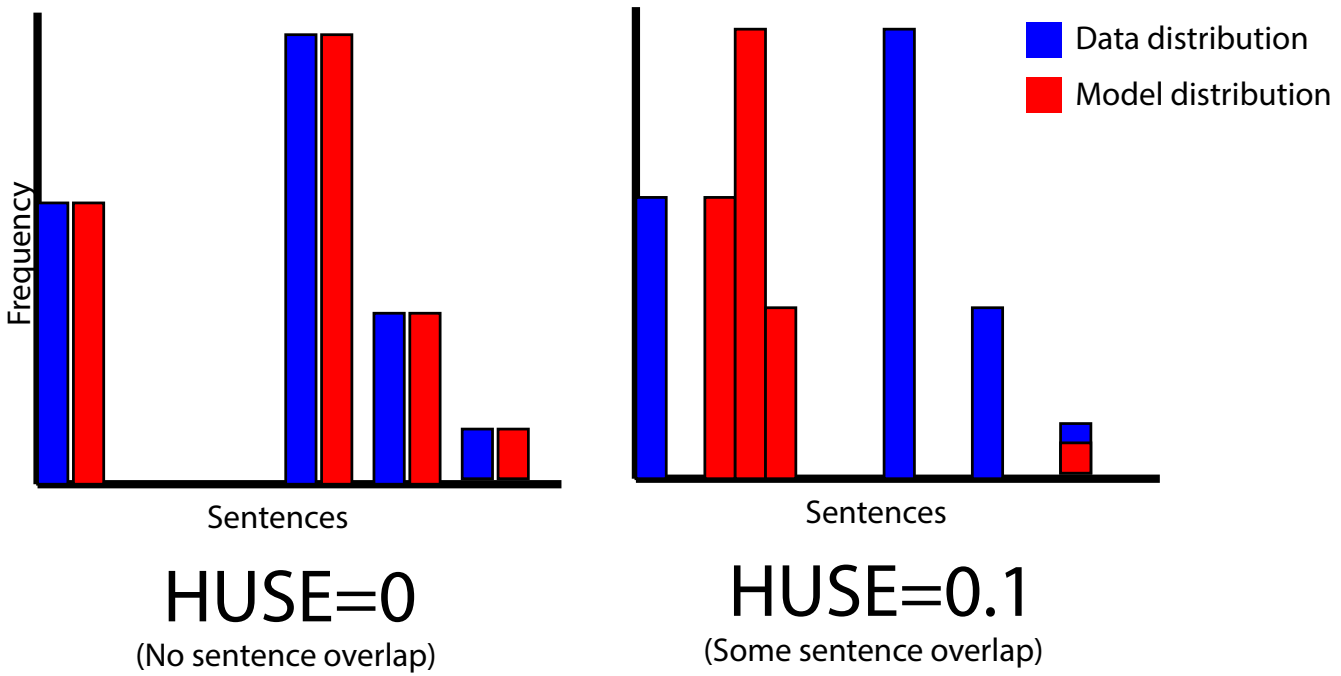
3. Evaluating super human performance

Challenge 1: diversity and semantics

Divergences (PPL / HUSE) ignore semantic similarity

Models are asked to output every reference sentence *exactly*

Can be too harsh as a metric for rule-based system



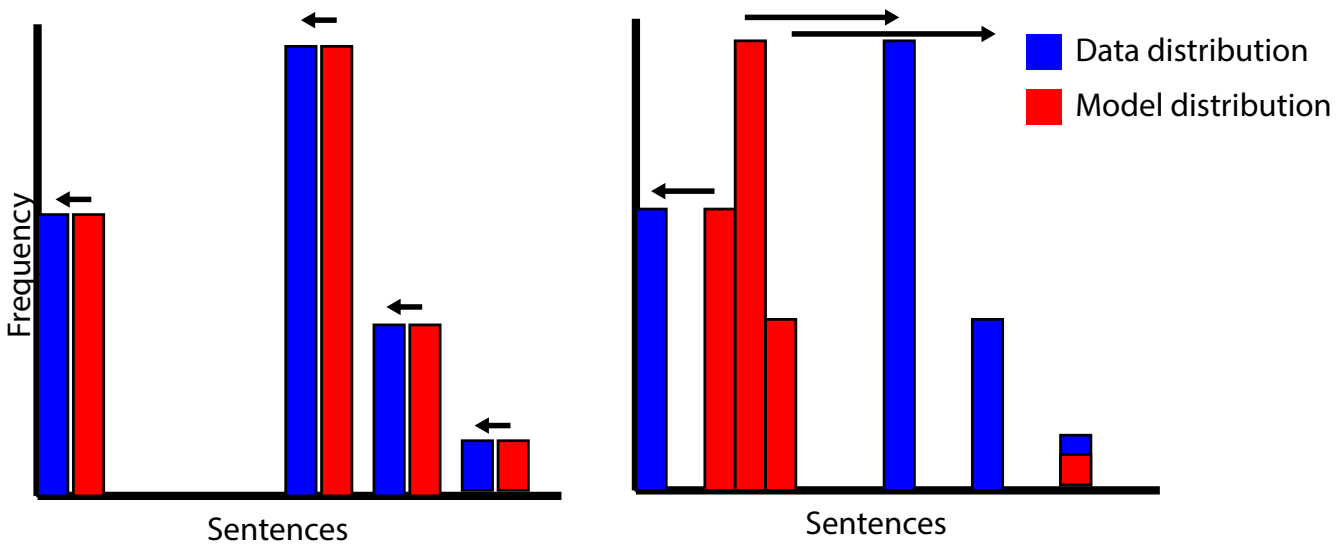
HUSE=0
(No sentence overlap)

HUSE=0.1
(Some sentence overlap)

Challenge 1: diversity and semantics

Evaluations should capture similarity.

Partial credit for models that output sentence similar to, but not exactly the reference.



Challenge 1: diversity and semantics

HUSE / Total variation

The model performs well on all downstream tasks with bounded losses

Evaluation of the future (?)

The model performs well on all downstream tasks which treat semantically similar sentences similarly

Challenge 1: diversity and semantics

HUSE / Total variation

The model performs well on all downstream tasks with bounded losses

Evaluation of the future (?)

The model performs well on all downstream tasks which treat semantically similar sentences similarly

Incorporating semantics makes evaluation substantially harder

Challenge 2: training objectives

Current training paradigm:

Log-loss / KL divergence / perplexity

$$\mathbb{E}_{p_{\text{ref}}}[-\log p_{\text{model}}(x)].$$

Challenge 2: training objectives

Current training paradigm:

Log-loss / KL divergence / perplexity

$$\mathbb{E}_{p_{\text{ref}}}[-\log p_{\text{model}}(x)].$$

Two properties make it ideal for training:

- **Expected loss**:
  Losses are averages over samples from p_{ref}.
- **Factorization:** Losses factorize over tokens,

$$-\log p_{\text{model}}(x_1, x_2) = -\log p_{\text{model}}(x_1) - \log p_{\text{model}}(x_2|x_1).$$

Challenge 2: training objectives

Current training paradigm:

Log-loss / KL divergence / perplexity

$$\mathbb{E}_{p_{\text{ref}}}[-\log p_{\text{model}}(x)].$$

Two properties make it ideal for training:

- **Expected loss**:
  Losses are averages over samples from p_{ref}.
- **Factorization:** Losses factorize over tokens,

$$-\log p_{\text{model}}(x_1, x_2) = -\log p_{\text{model}}(x_1) - \log p_{\text{model}}(x_2|x_1).$$

KL divergence is the *only* divergence that has both properties

Challenge 2: training objectives

New evaluation paradigm:

HUSE / Total variation

$$\|p_{\text{ref}} - p_{\text{model}}\|_{TV}$$

Easy to evaluate (w/ humans) *very* hard to train on

Challenge 2: training objectives

New evaluation paradigm:

HUSE / Total variation

$$\|p_{\text{ref}} - p_{\text{model}}\|_{TV}$$

Easy to evaluate (w/ humans) *very* hard to train on

Fundamental mismatch between training and eval losses.

Challenge 2: training objectives

New evaluation paradigm:

HUSE / Total variation

$$\|p_{\text{ref}} - p_{\text{model}}\|_{TV}$$

Easy to evaluate (w/ humans) *very* hard to train on

Fundamental mismatch between training and eval losses.

We need a new training paradigm, or better decoders

Challenge 3: super-human models

In this talk:

imitation is about matching humans

The future: imitate and improve upon human generation

Challenge 3: super-human models

In this talk:

imitation is about matching humans

The future: imitate and improve upon human generation

Potential approaches

1. Task-oriented evaluation.
2. Curated test-sets.
3. Improved human evaluation.

Challenge 3: super-human models

In this talk:

imitation is about matching humans

The future: imitate and improve upon human generation

Potential approaches

1. Task-oriented evaluation.
2. Curated test-sets.
3. Improved human evaluation.

Intriguing implication and contract:

do downstream tasks *better* than human

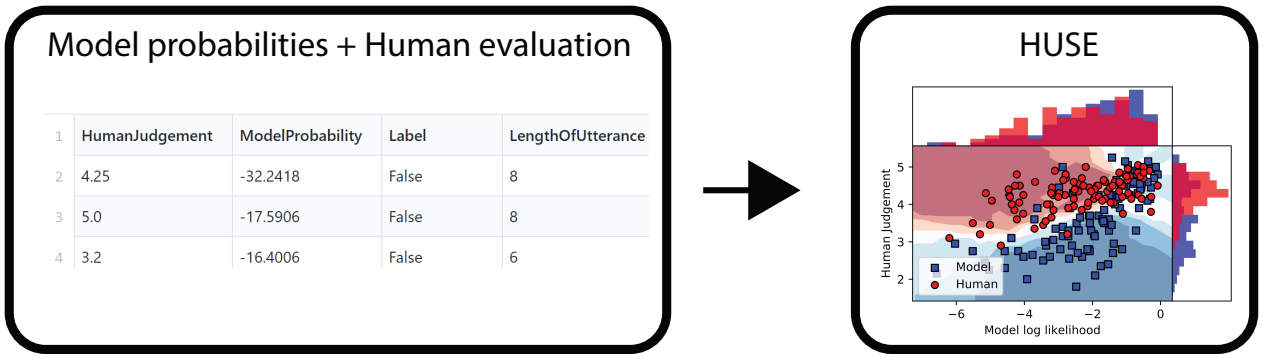
Exhortation

1. Use HUSE!

Easy-to-use evaluation system

Model probabilities + Human evaluation

	HumanJudgement	ModelProbability	Label	LengthOfUtterance
2	4.25	-32.2418	False	8
3	5.0	-17.5906	False	8
4	3.2	-16.4006	False	6



- Applies to any probabilistic model
- Fairly low cost (~$100 / model)

`https://github.com/hughbzhang/HUSE`

Exhortation

1. Use HUSE!

`https://github.com/hughbzhang/HUSE`

2. Otherwise:

show perplexity + sampled outputs

Exhortation

1. Use HUSE!

`https://github.com/hughbzhang/HUSE`

2. Otherwise:

show perplexity + sampled outputs

3. If not even that:

show multiple references + generations

A footnote on SoTA models

Sampled (diverse) model

In 2004, Mario Balotelli was 15 years old. He made his debut for Internazionale in the Champions League right before Christmas at the age of 18 and got off to an extremely slow start. Actually, there wasn't really any of a slow start.

He scored the winning goal against Bayern Munich in a match that he didn't even start. Balotelli lost Sunday's match 4-3 to Arsenal. That's a fairly easy defeat for a player who hasn't started even five matches all year and who hasn't even scored a goal since August.

Thanks!

My coauthors:

Thanks!

My coauthors:

Stanford NLP group:

Thanks!

My coauthors:


Stanford NLP group:

Interested in this stuff? Come join me (in 2020)