Mauve

Measuring the Gap Between Neural Text and Human Text

NeurIPS 2021 (Outstanding Paper Award)

Stanford NLP Seminar, 3/3/22

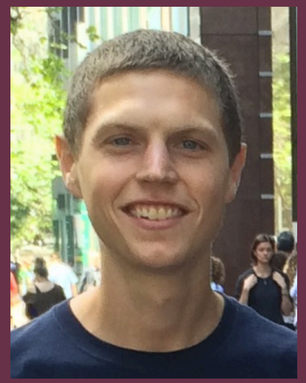
Krishna Pillutla

Swabha Swayamdipta

Rowan Zellers

John Thickstun

Sean Welleck



Yejin Choi

Zaid Harchaoui

UNIVERSITY of WASHINGTON

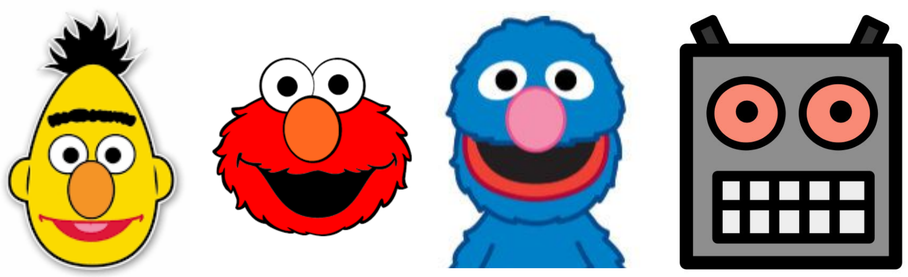
PAUL G. ALLEN SCHOOL OF COMPUTER SCIENCE & ENGINEERING

Allen Institute for AI

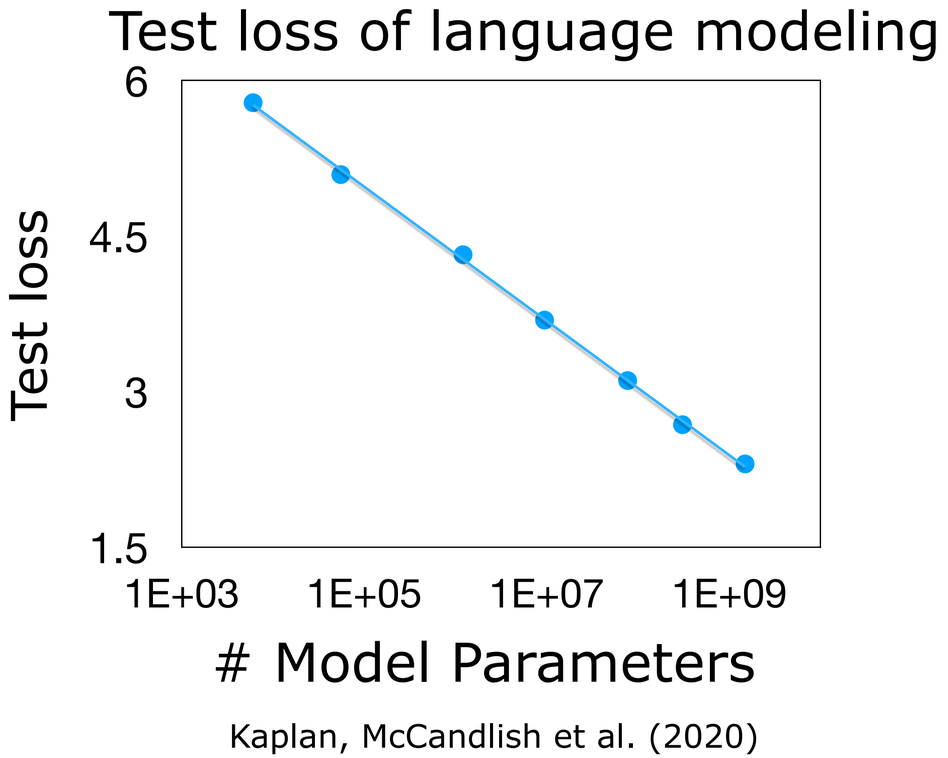
Stanford University

Text generation

Enormous language models (**ELM**s) $\Longrightarrow$ massive progress in NLP



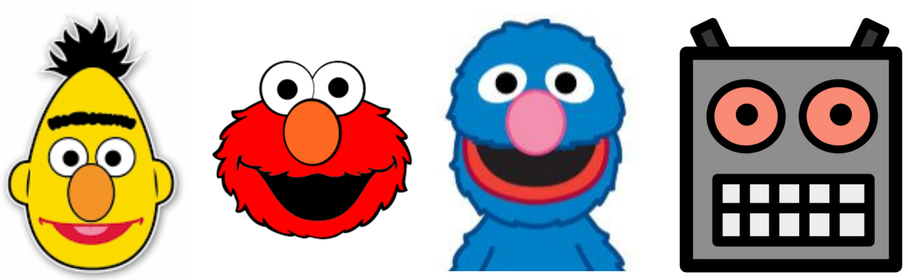
Devlin et al. (2018), Brown et al. (2020), *inter alia*



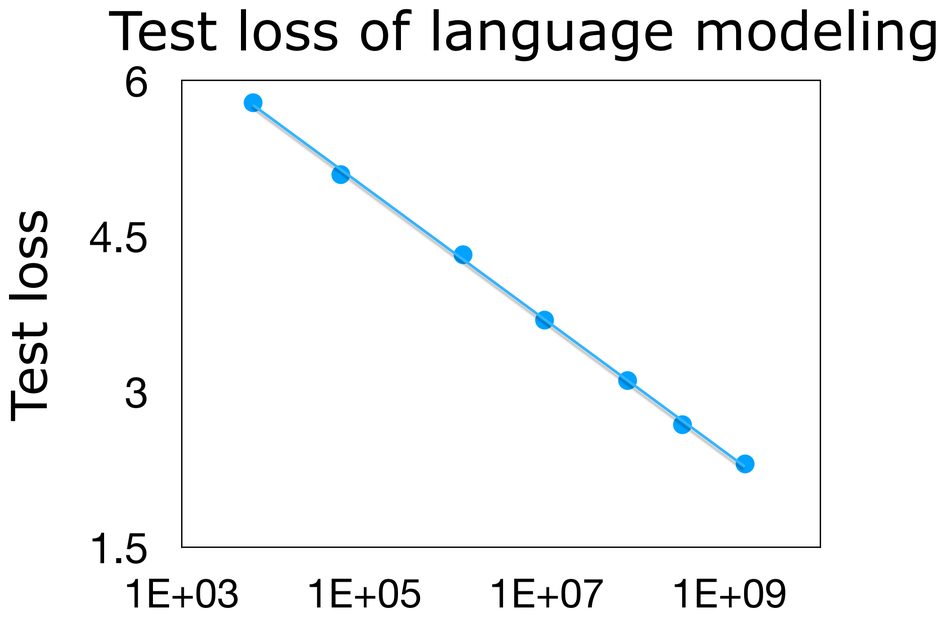
Kaplan, McCandlish et al. (2020)

Text generation

Enormous language models (**ELM**s) $\Longrightarrow$ massive progress in NLP

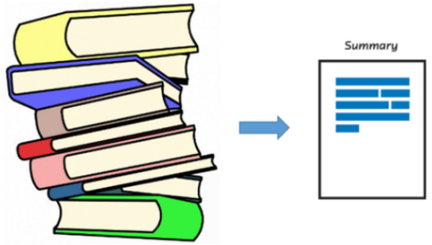


Devlin et al. (2018), Brown et al. (2020), *inter alia*



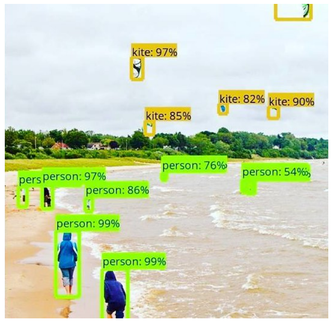
Kaplan, McCandlish et al. (2020)

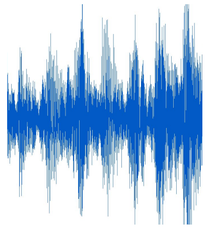
State-of-the-art in text generation tasks such as translation, summarization, etc.



Not just language

Enormous models $\Longrightarrow$ massive progress in all of AI

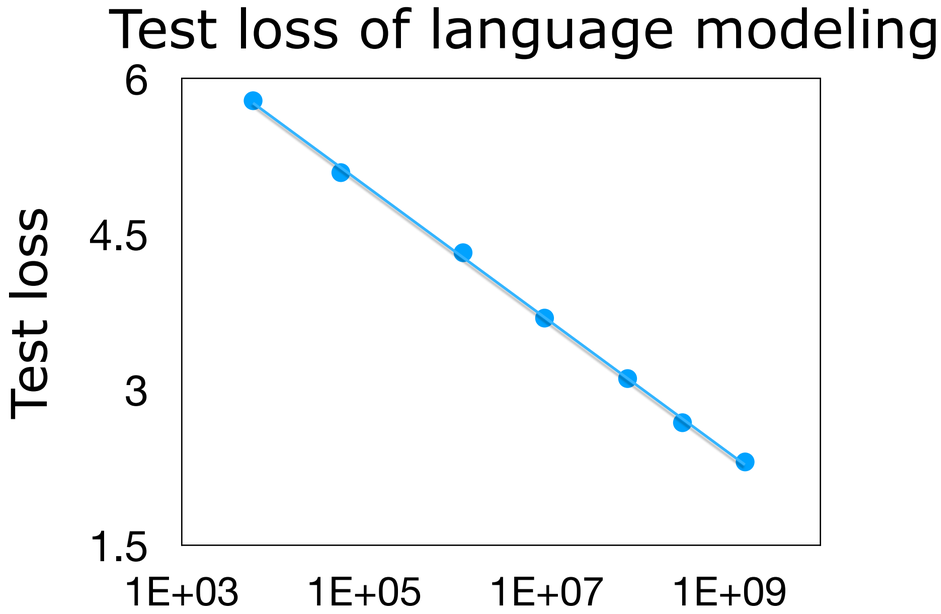




Dosovitskiy et al. (2020), Hsu et al. (2021), *inter alia*

$\longrightarrow$ foundation models

Bommasani et al. (2021)

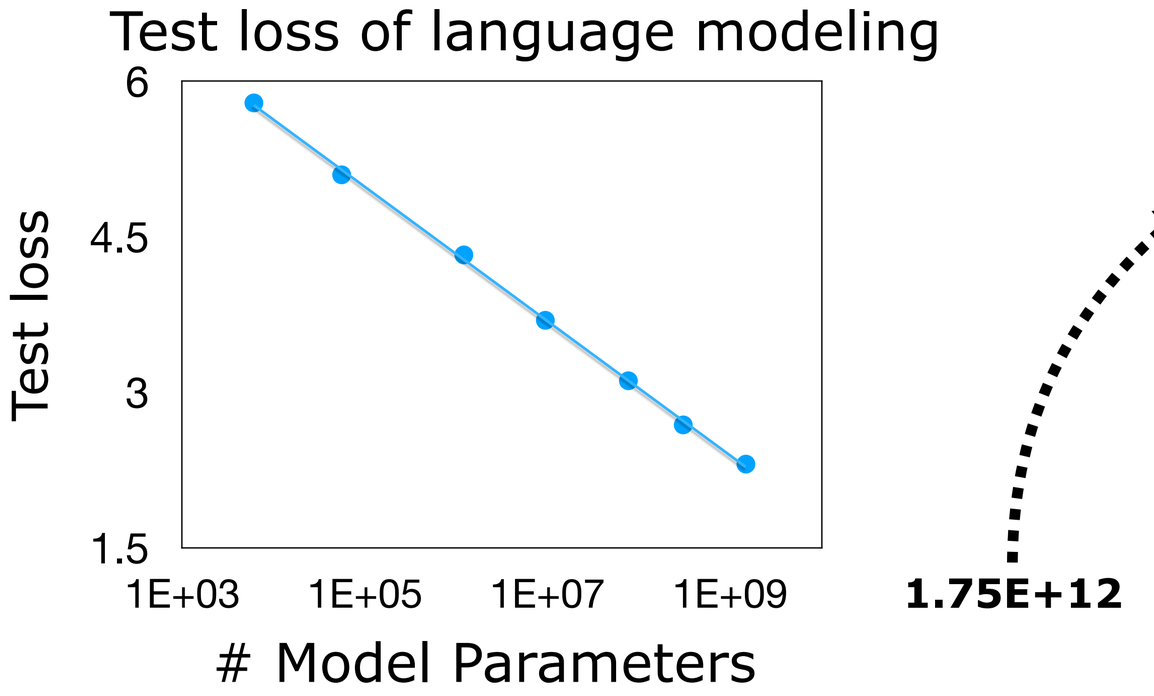


Kaplan, McCandlish et al. (2020)

Not just language

Enormous models $\Longrightarrow$ massive progress in all of AI

Multi-modal



Wu Dao 2.0 has 1.75 trillion parameters!

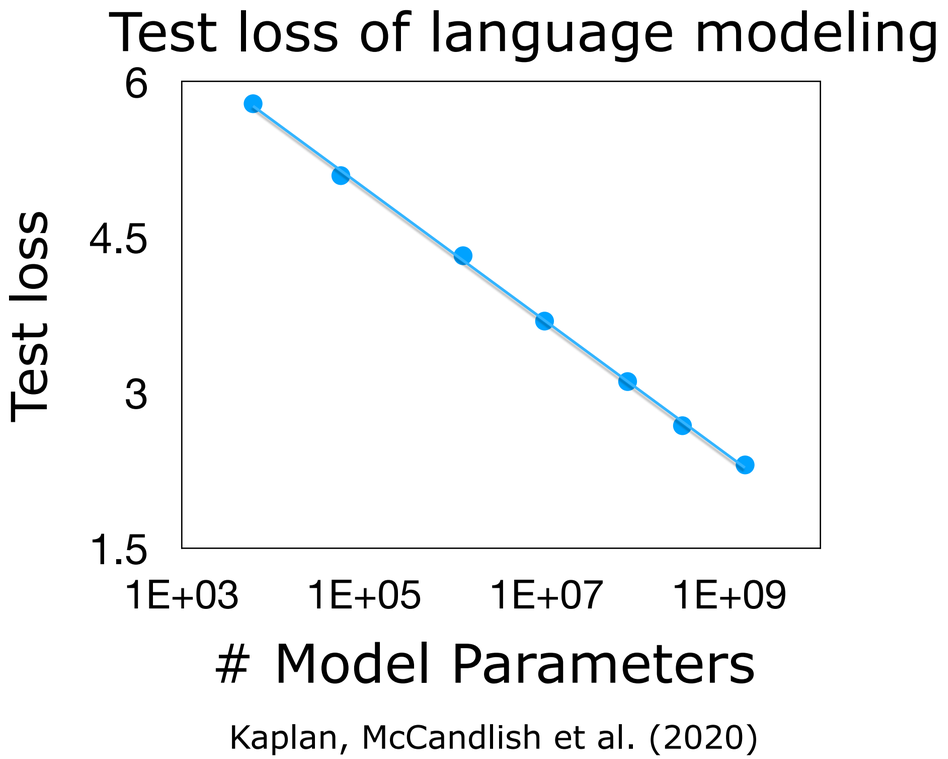
Kaplan, McCandlish et al. (2020)

New capabilities are emerging

ELMs can write long essays now!

>> prompt: In a shocking finding, scientists discovered a herd of unicorns living in a remote, previously, unexplored valley, in the Andes Mountains.

GPT-2 **Continuation**. The scientists named the population, after their distinctive horn, Ovid's Unicorn. These four-horned, silver-white unicorns were previously unknown …



Kaplan, McCandlish et al. (2020)

New capabilities are emerging

ELMs can write long essays now!

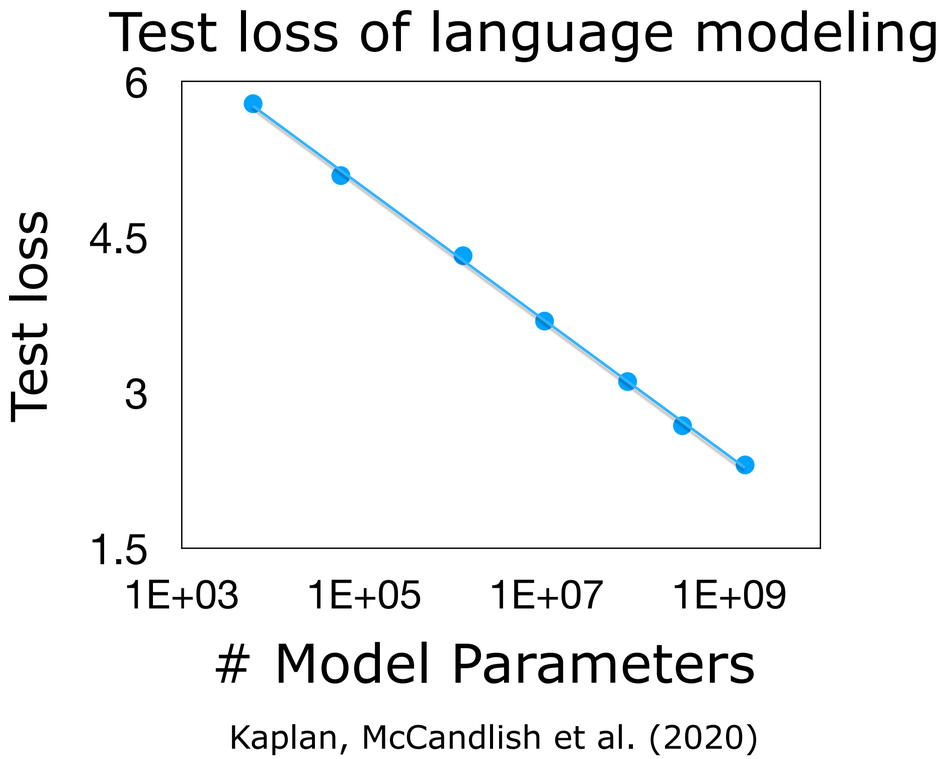
>> prompt: In a shocking finding, scientists discovered a herd of unicorns living in a remote, previously, unexplored valley, in the Andes Mountains.

GPT-2 **Continuation**. The scientists named the population, after their distinctive horn, Ovid's Unicorn. These four-horned, silver-white unicorns were previously unknown …

Zero-shot prediction

>> prompt: English: Hello!
French:

GPT-3 English: Hello!
French: Bonjour!



Kaplan, McCandlish et al. (2020)

Open-ended text generation

- ELMs write long essays: open-ended

>> prompt: In a shocking finding, scientists discovered a herd of unicorns living in a remote, previously, unexplored valley, in the Andes Mountains.

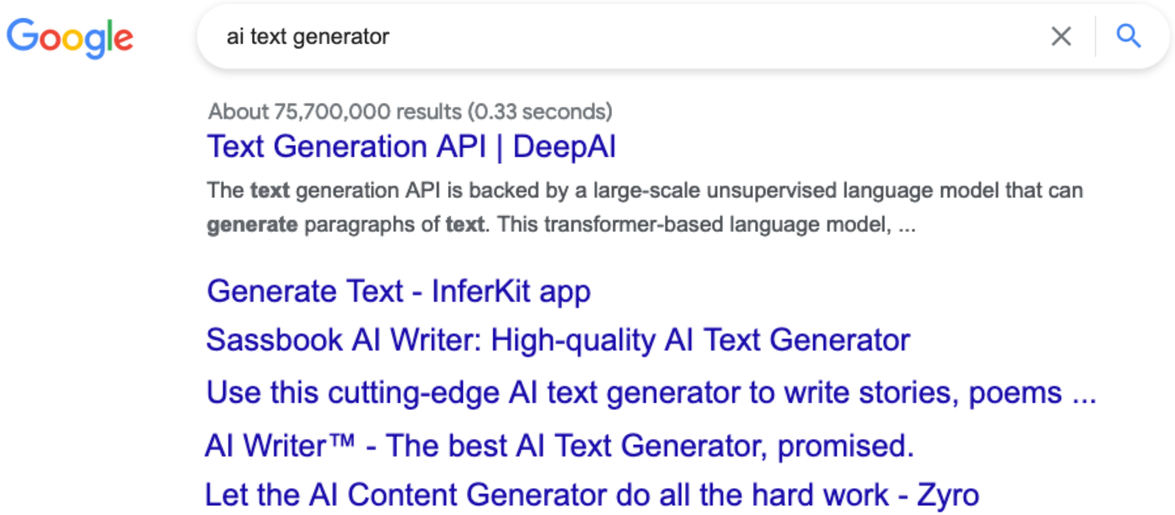
Continuation. The scientists named the population, after their distinctive horn, Ovid's Unicorn. These four-horned, silver-white unicorns were previously unknown …

Open-ended text generation

- ELMs write long essays: open-ended
- Widely deployed commercially

>> prompt: In a shocking finding, scientists discovered a herd of unicorns living in a remote, previously, unexplored valley, in the Andes Mountains.

Continuation. The scientists named the population, after their distinctive horn, Ovid's Unicorn. These four-horned, silver-white unicorns were previously unknown …

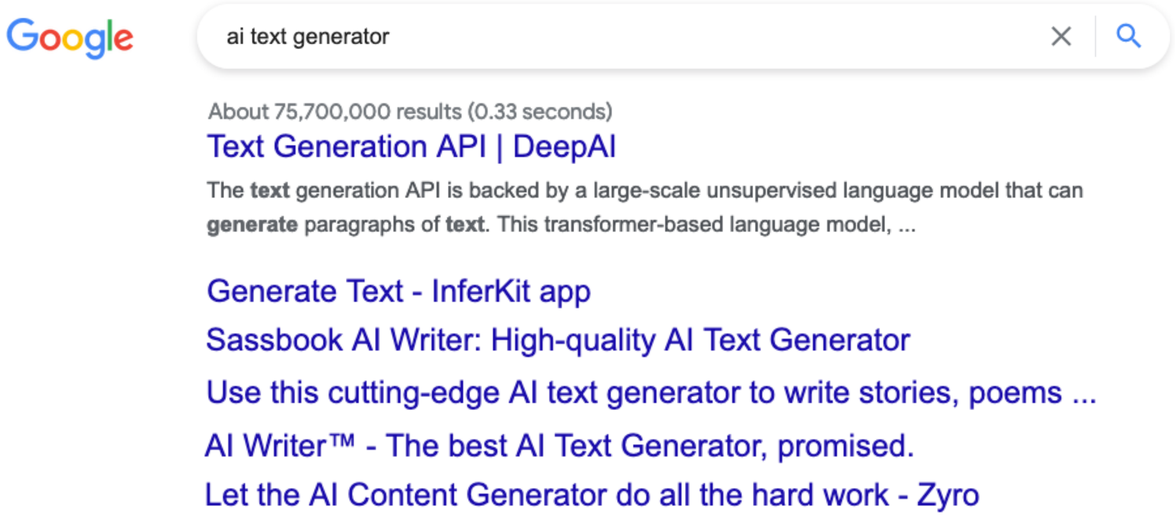


Open-ended text generation

- ELMs write long essays: open-ended
- Widely deployed commercially
- ELMs still make mistakes. **But how close is it really to human text?**

>> prompt: In a shocking finding, scientists discovered a herd of unicorns living in a remote, previously, unexplored valley, in the Andes Mountains.

Continuation. The scientists named the population, after their distinctive horn, Ovid's Unicorn. These four-horned, silver-white unicorns were previously unknown …



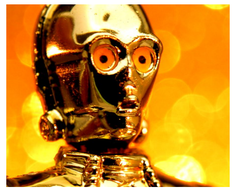
Open-ended text generation

- ELMs write long essays: open-ended
- Widely deployed commercially
- ELMs still make mistakes. **But how close is it really to human text?**

>> prompt: In a shocking finding, scientists discovered a herd of unicorns living in a remote, previously, unexplored valley, in the Andes Mountains.

GPT-2

Continuation. The scientists named the population, after their distinctive horn, Ovid's Unicorn. These four-horned, silver-white unicorns were previously unknown …



Continuation 2. This discovery has kicked off an all-out search for other mythical creatures from the frozen reaches of the Antarctic to the tropical islands of the Pacific …

Continuation 3. Perhaps most astonishingly, these unicorns have developed their own artificial general intelligence named Yuyaysapa …

Google

ai text generator

About 75,700,000 results (0.33 seconds)

Text Generation API | DeepAI

The **text** generation API is backed by a large-scale unsupervised language model that can **generate** paragraphs of **text**. This transformer-based language model, ...

Generate Text - InferKit app

Sassbook AI Writer: High-quality AI Text Generator

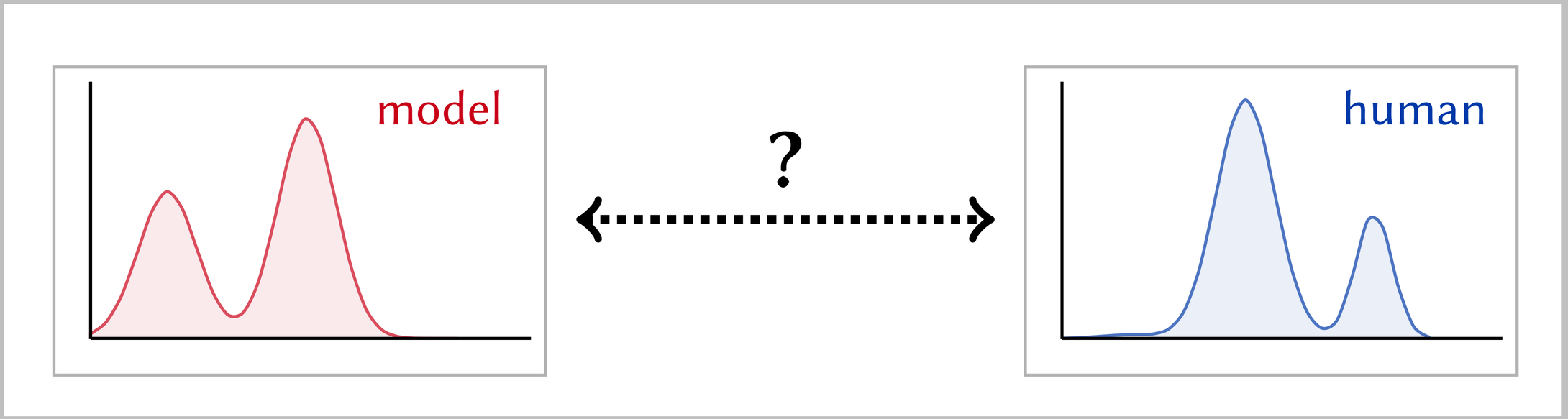
Use this cutting-edge AI text generator to write stories, poems ...

AI Writer™ - The best AI Text Generator, promised.

Let the AI Content Generator do all the hard work - Zyro

Open-ended text generation

Our goal: measure the gap between the two *distributions*!



Outline

- **Background and Motivation**
- Mauve
- Computing Mauve in practice
- Experiments

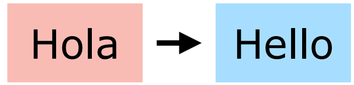
Text Generation

Directed

Open-Ended

Text Generation

Translation



Summarization

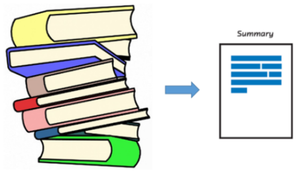
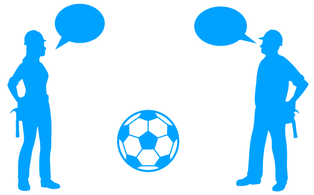


Image captioning

Cat not sure whether to laugh or yawn



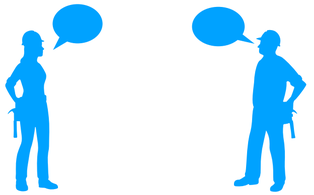
Goal-oriented dialog

Your are a time traveller stranded in 2021.

Stories

The Lakers and Celtics face off in Game 7.

Articles



Chit chat

Text Generation

Directed

Open-Ended

Discriminative

Generative

$Q_\theta($ Hello $|$ Hola $)$

Modeling

$Q_\theta($  $)$

$\approx P_{\text{true}}($ Hello $|$ Hola $)$

Goal

$\approx P_{\text{true}}($  $)$

Automatic Evaluation

Text Generation

Directed

Open-Ended

Text Generation

Discriminative

Generative

$Q_\theta($ Hello $|$ Hola $)$

Modeling

$Q_\theta($  $)$

$\approx P_{\text{true}}($ Hello $|$ Hola $)$

Goal

$\approx P_{\text{true}}($  $)$

Compare with *human ref*

Automatic Evaluation

BLEU, METEOR, ROUGE, BERTScore, BLEURT, …

Text Generation

Directed — Open-Ended

Text Generation

Discriminative — Generative

$Q_\theta($ Hello $|$ Hola $)$ — **Modeling** — $Q_\theta($  $)$

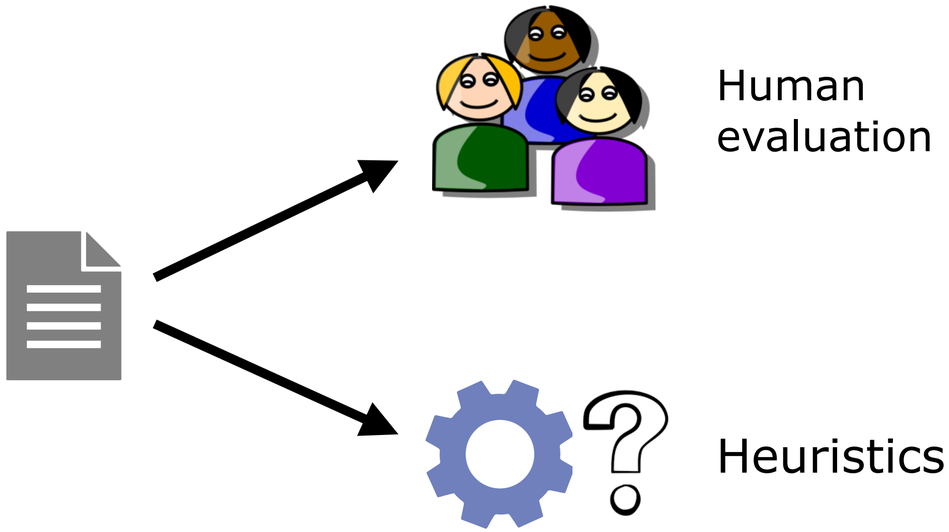
$\approx P_{\text{true}}($ Hello $|$ Hola $)$ — **Goal** — $\approx P_{\text{true}}($  $)$

Compare with *human ref* — **Automatic Evaluation** — **???**

BLEU, METEOR, ROUGE, BERTScore, BLEURT, …

Open ended generation: How good is the model?

Numerous “correct” completions.
Reference-based methods do not apply



Related: Hybrid human + automatic eval
Hashimoto et al. (NAACL 2019)

Generative models in computer vision

Generative

Synthetic Images

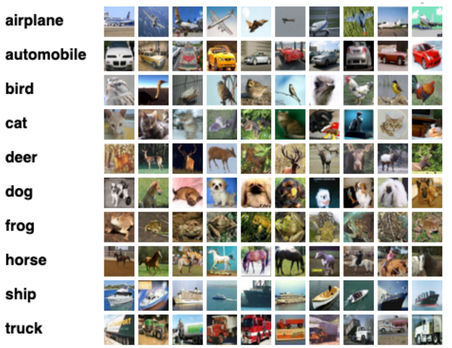
Real Images

Modeling

$$Q_\theta\left(\quad \right)$$

Goal

$$\approx P_{\text{true}}\left(\quad \right)$$



Automatic Evaluation

Generative models in computer vision

Generative

Synthetic Images

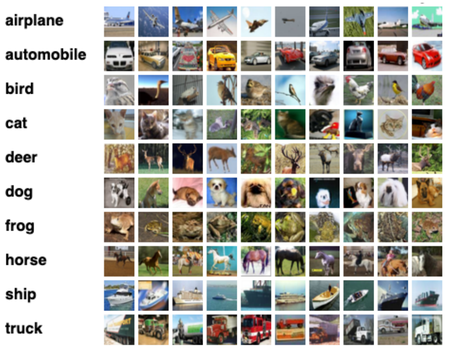
Real Images

Modeling

$$Q_\theta\left(\quad\right)$$

Goal

$$\approx\ P_{\text{true}}\left(\quad\right)$$



Automatic Evaluation

$$\text{Gap}\left(Q_\theta, P_{\text{true}}\right)$$

Fréchet distance (FID)

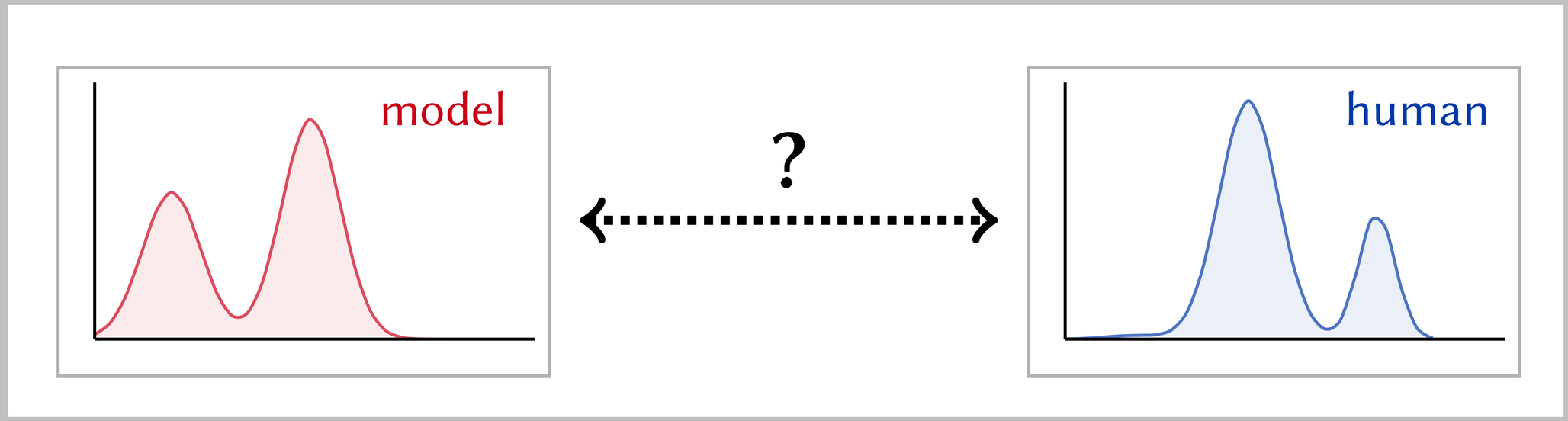
Precision-Recall

Divergence frontiers

Heusel et al. (NeurIPS 2017), Sajjadi et al. (NeurIPS 2018), Djolonga et al. (AISTATS 2020)

Open ended generation: How good is the model?

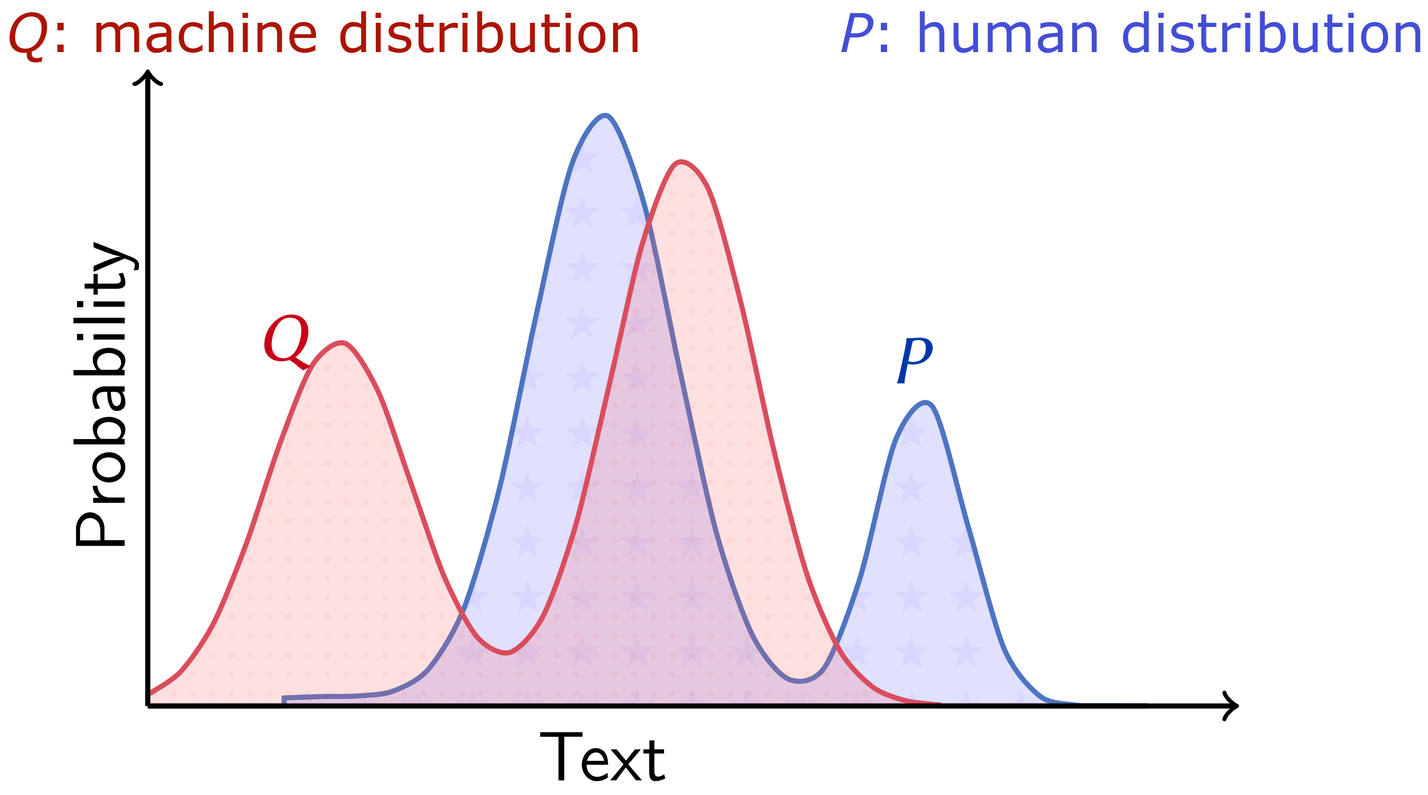
Our approach: measure the gap between the two *distributions*!



Outline

- Background and Motivation
- **Mauve**
- Computing Mauve in practice
- Experiments

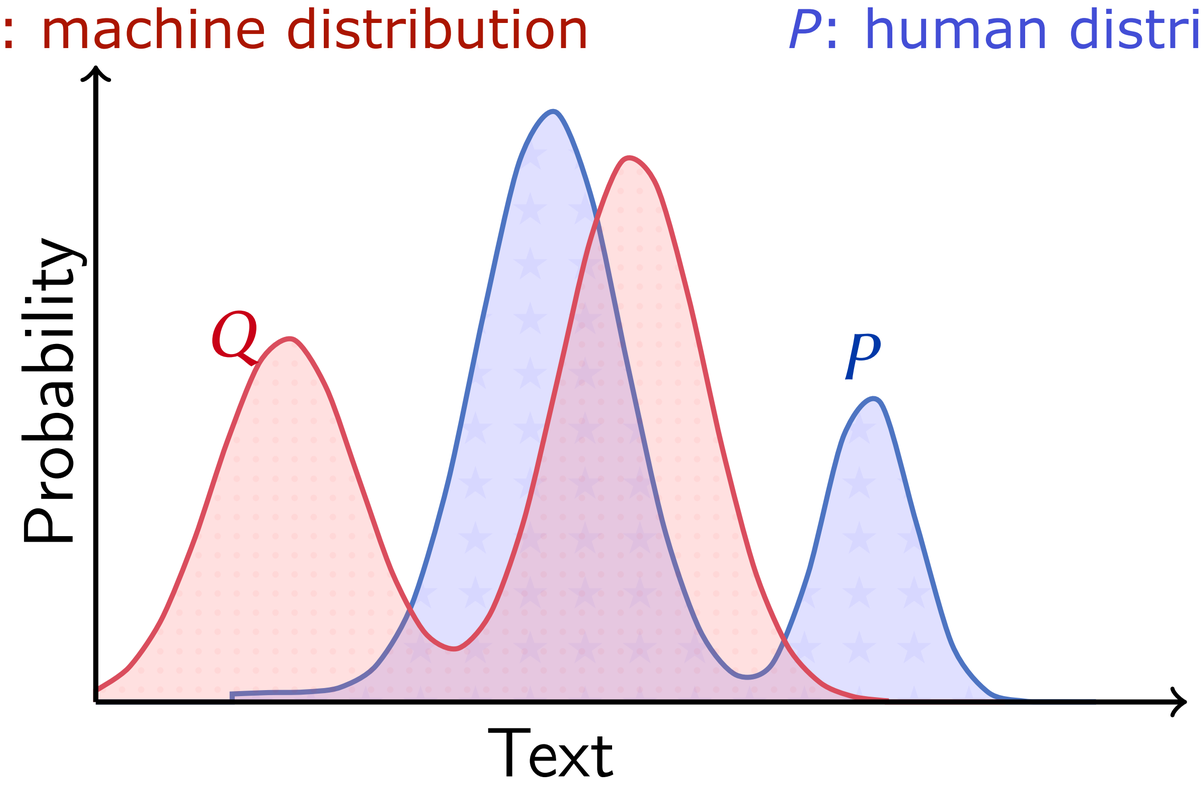
Two types of errors in text generation



Two types of errors in text generation

Distribution over text sequences

$$Q(x) = \prod_{t=1}^{|x|} Q(x_t \mid x_{<t})$$



Two types of errors in text generation

Distribution over text sequences

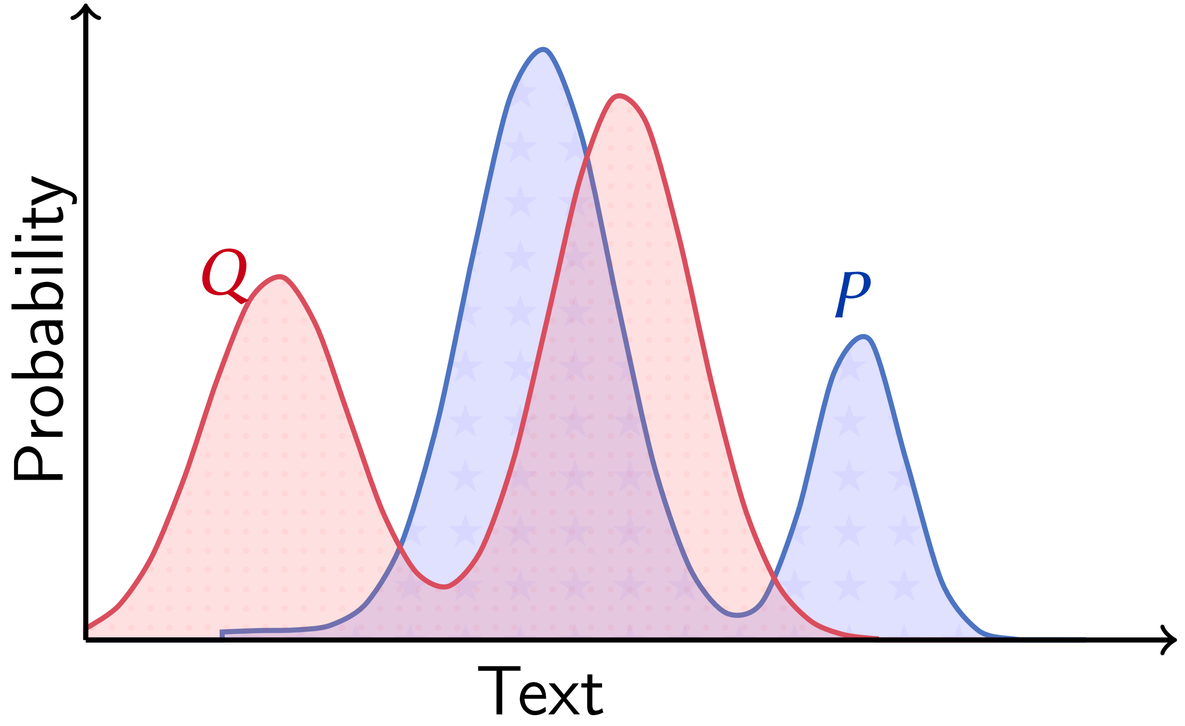
$$Q(x) = \prod_{t=1}^{|x|} Q(x_t | x_{<t})$$

Model $\hat{P}$ + Decoding Algo.

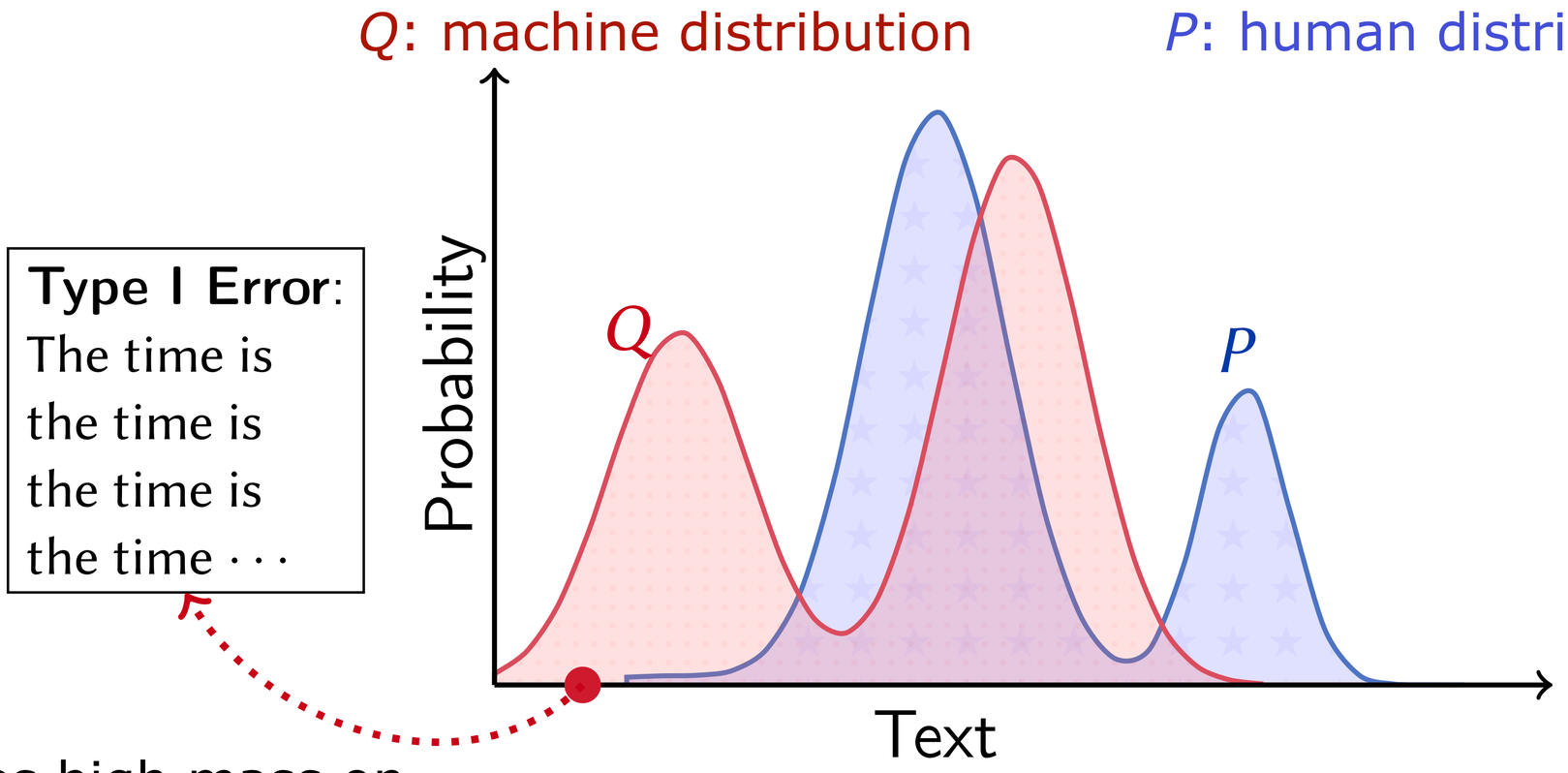
$$Q(\cdot | x_{<t}) = \text{Decode}\left(\hat{P}(\cdot | x_{<t})\right)$$

Q: machine distribution

P: human distribution

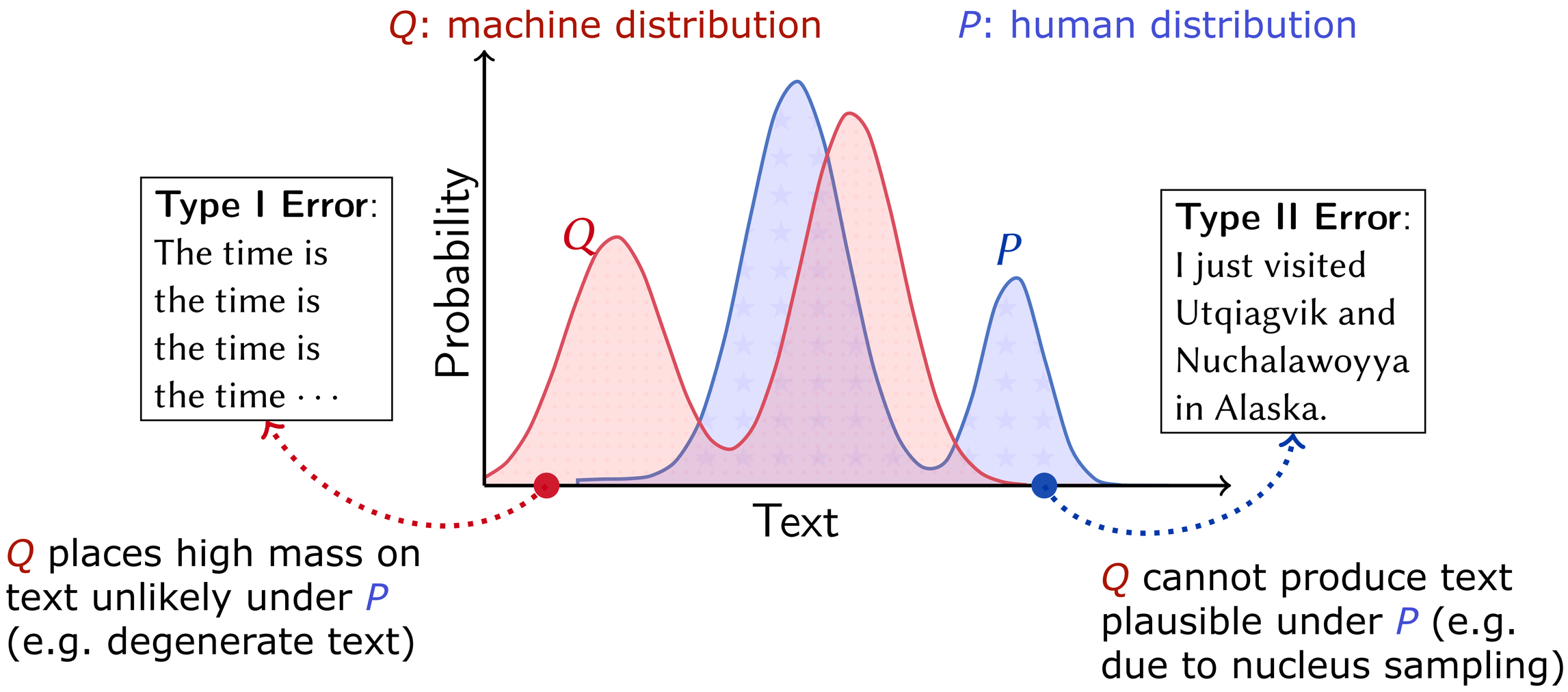


Two types of errors in text generation

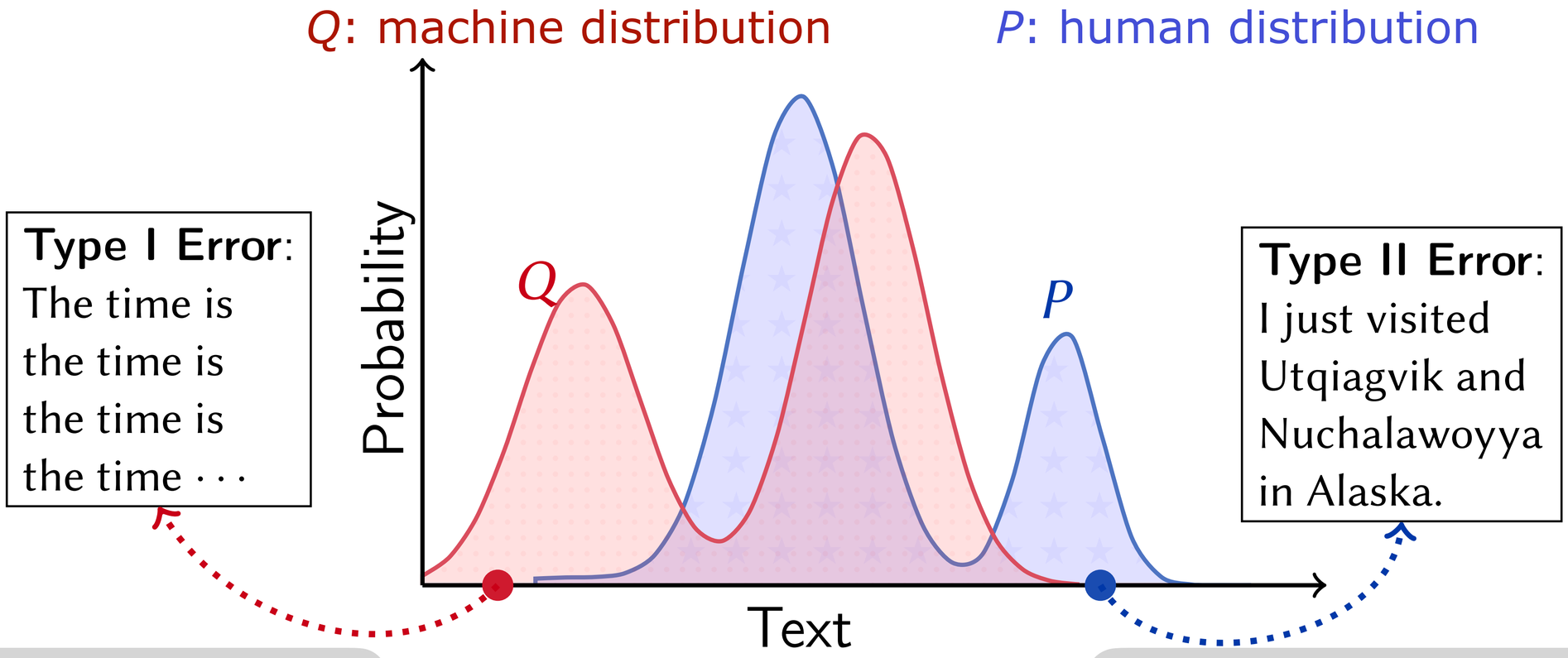


Q places high mass on text unlikely under P (e.g. degenerate text)

Two types of errors in text generation



Two types of errors in text generation



KL($Q|P$)

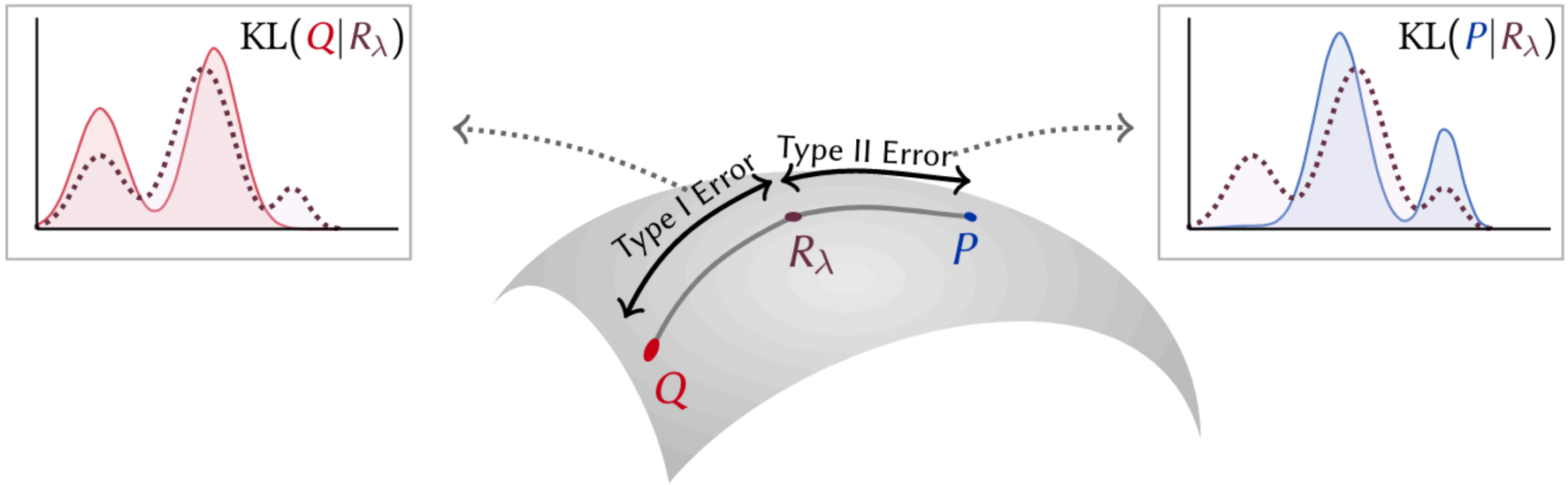
$$\mathrm{KL}(P\,|\,Q) = \sum_x P(x)\,\log\frac{P(x)}{Q(x)}$$

KL($P|Q$)

Introducing mixture distributions

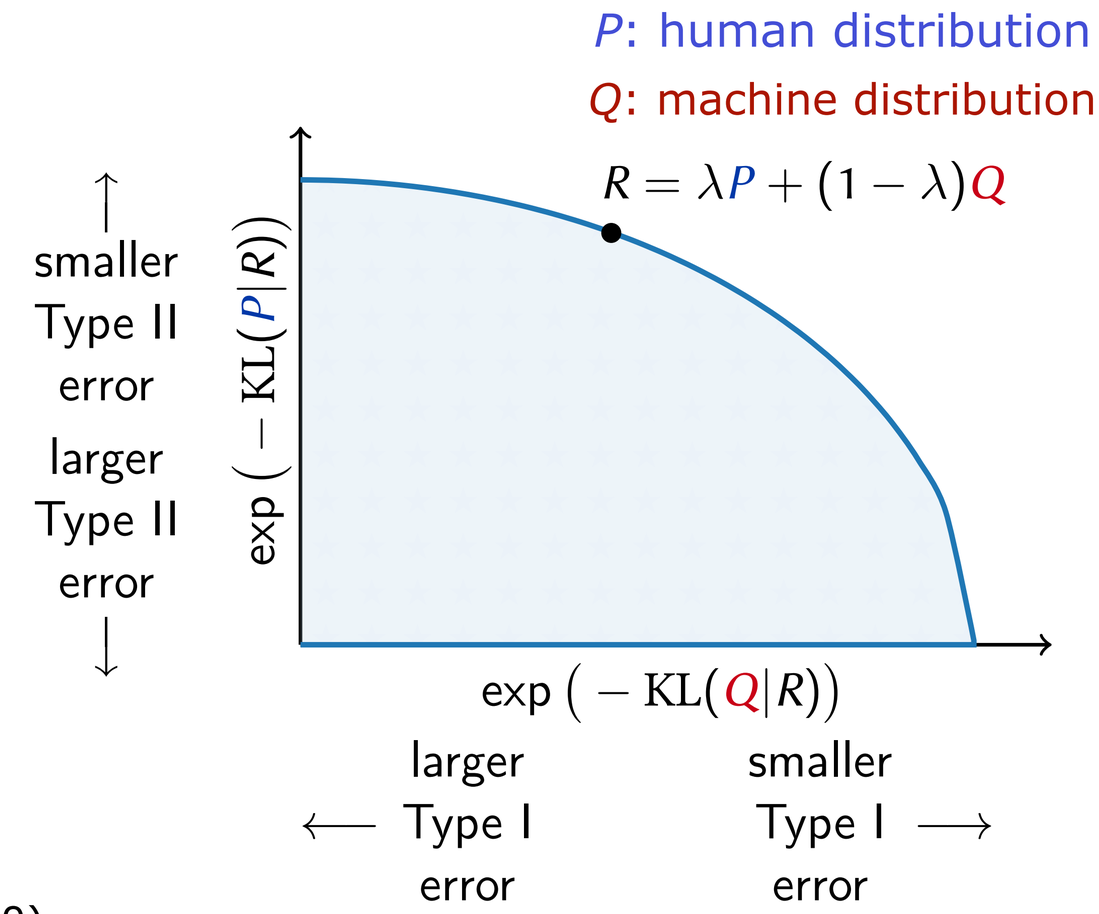
KL($Q|P$) and KL($P|Q$) can be infinite, so measure errors *softly* using *mixtures*

$$R_\lambda = \lambda P + (1-\lambda)Q$$



Mauve: summarizing both errors

- **Divergence Curve:** Varying the *mixture weight* captures trade-off between Type I and Type II errors

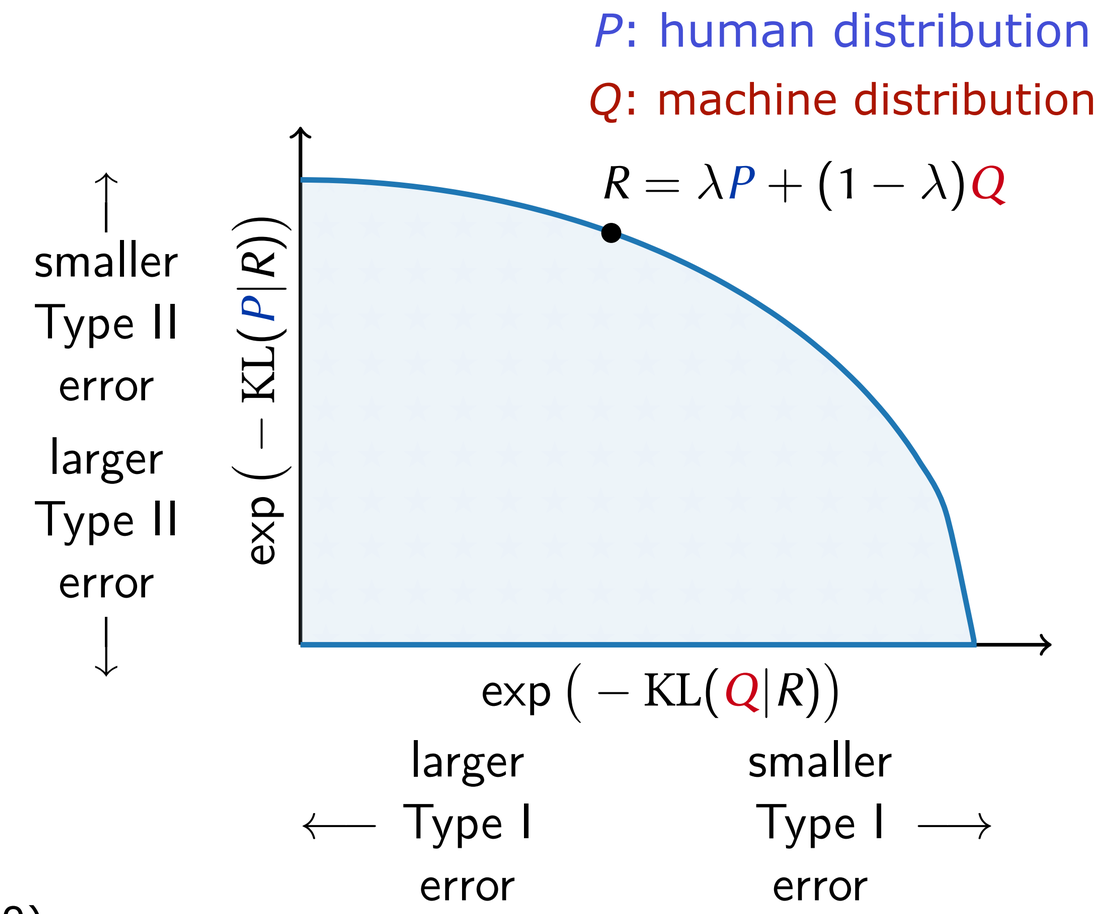


Sajjadi et al. (NeurIPS 2018); Djolonga et al. (AISTATS 2020)

Mauve: summarizing both errors

- **Divergence Curve:** Varying the *mixture weight* captures trade-off between Type I and Type II errors

- **Mauve,** the area under this curve, is a *quantitative measure of similarity* and takes values between 0 (dissimilar) and 1 (identical)



Sajjadi et al. (NeurIPS 2018); Djolonga et al. (AISTATS 2020)

Outline

- Background and Motivation
- Mauve
- **Computing Mauve in practice**
- Experiments

Computing Mauve in practice

- Sum over documents intractable for neural LMs

$$\mathrm{KL}(Q \mid R) = \sum_{x} Q(x) \log \frac{Q(x)}{R(x)}$$

Computing Mauve in practice

- Sum over documents intractable for neural LMs

$$\mathrm{KL}(Q \mid R) = \sum_{x} Q(x) \log \frac{Q(x)}{R(x)}$$

Monte Carlo?

Human prob. *P* not known

Computing Mauve in practice

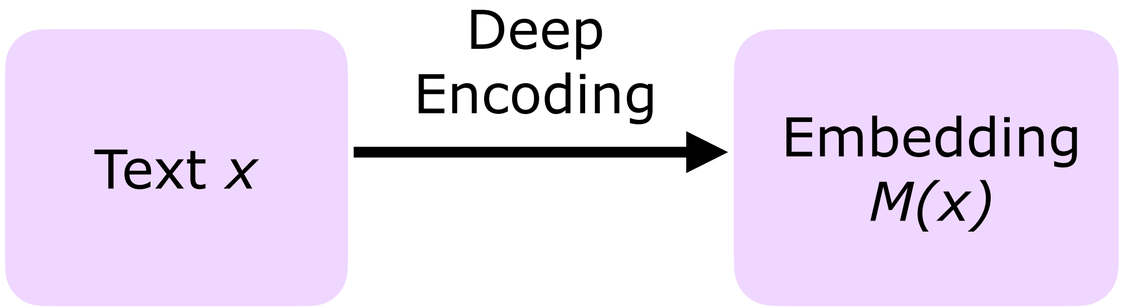
- Sum over documents intractable for neural LMs

$$\mathrm{KL}(Q \mid R) = \sum_{x} Q(x) \log \frac{Q(x)}{R(x)}$$

Monte Carlo?

Human prob. *P* not known

- Computation pipeline



Computing Mauve in practice

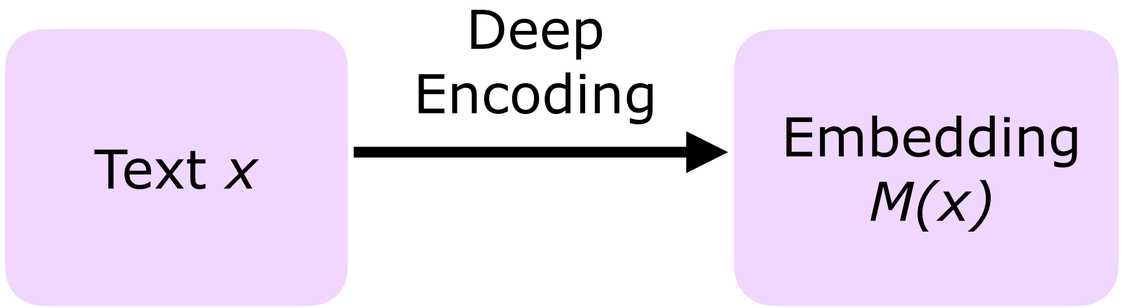
- Sum over documents intractable for neural LMs

$$\mathrm{KL}(Q \mid R) = \sum_x Q(x) \log \frac{Q(x)}{R(x)}$$

Monte Carlo?

Human prob. *P* not known

- Computation pipeline

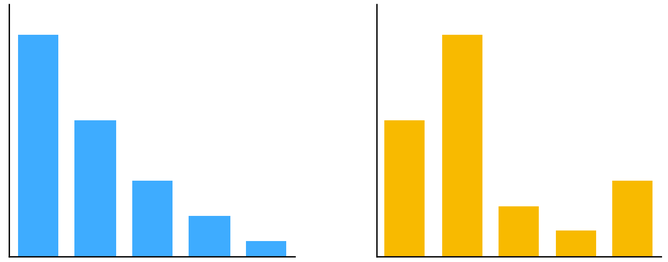


Estimating KL of continuous high-dim distributions from samples: Hard

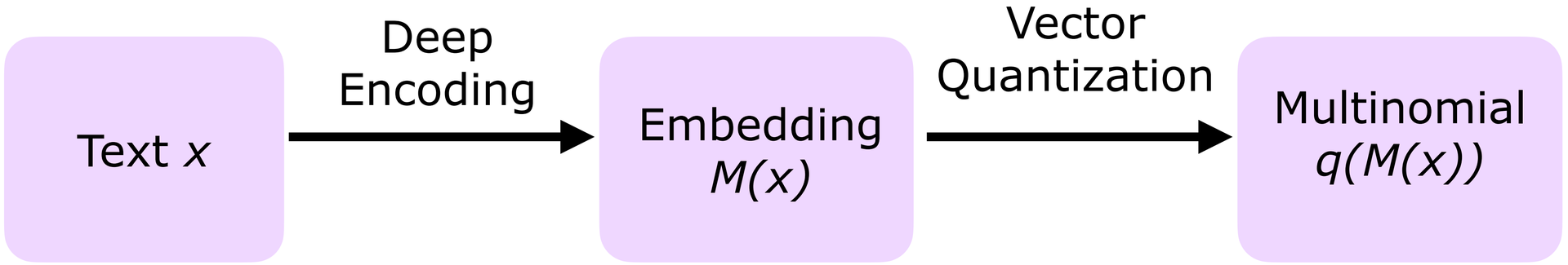
Computing Mauve in practice

- Sum over documents intractable for neural LMs

$$\mathrm{KL}(Q \mid R) = \sum_x Q(x) \log \frac{Q(x)}{R(x)}$$



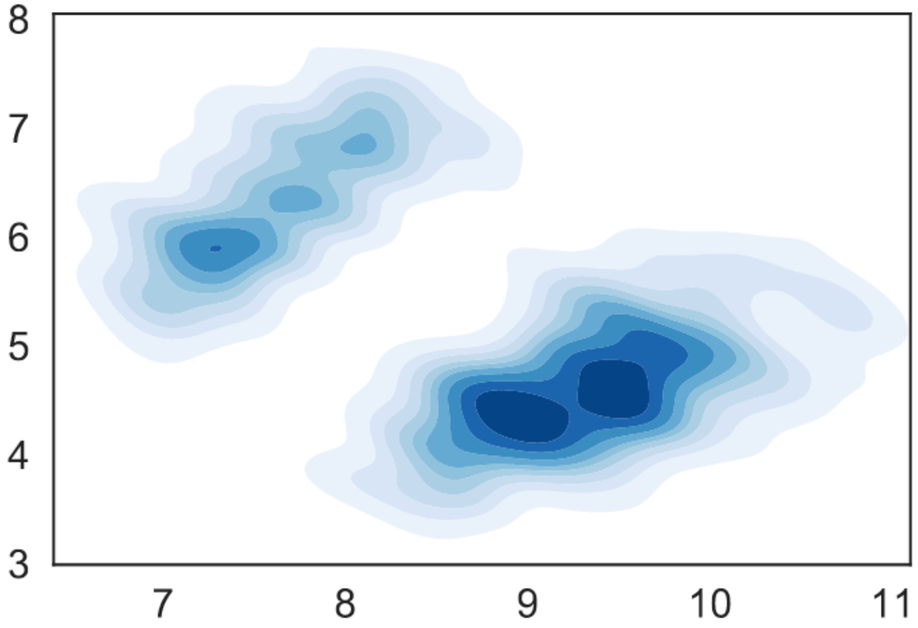
- Computation pipeline



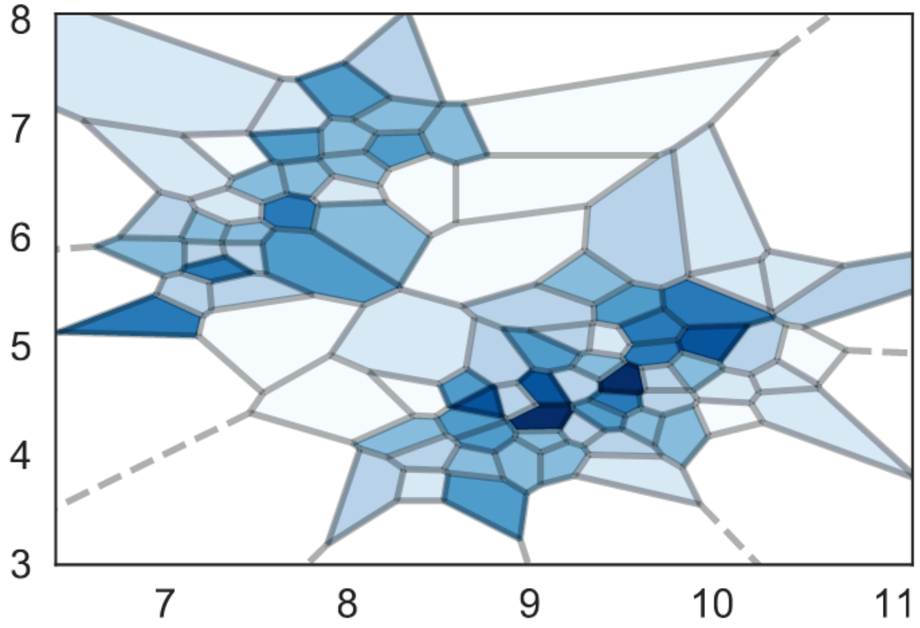
Computing Mauve in practice

Vector quantization

Continuous 2D distribution



Quantized distribution



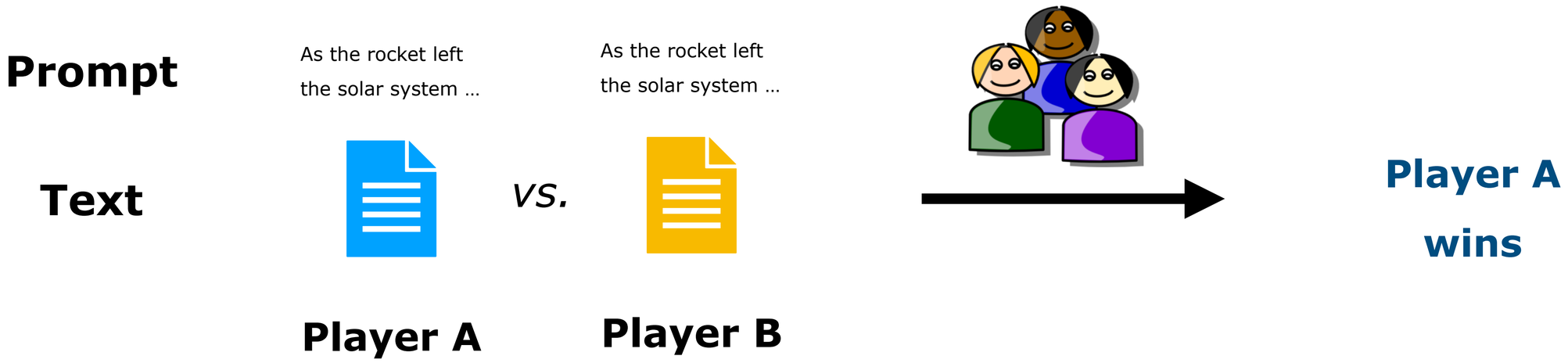
Outline

- Background and Motivation
- Mauve
- Computing Mauve in practice
- **Experiments**

Human judgements

Head-to-head: Is A or B more (a) human-like, (b) interesting, (c) sensible?

We compare text written by humans and 8 models



Human judgements

Head-to-head: Is A or B more (a) human-like, (b) interesting, (c) sensible?

We compare text written by humans and 8 models

Head-to-head record

Bradley-Terry model

Ranking

1. 
2. 
3. 
4. 

⋮

Mauve correlates with human judgements

Mauve correlates with human judgements

Mauve correlates with human judgements

Gen. PPL.: Holtzman et al. (ICLR 2020)

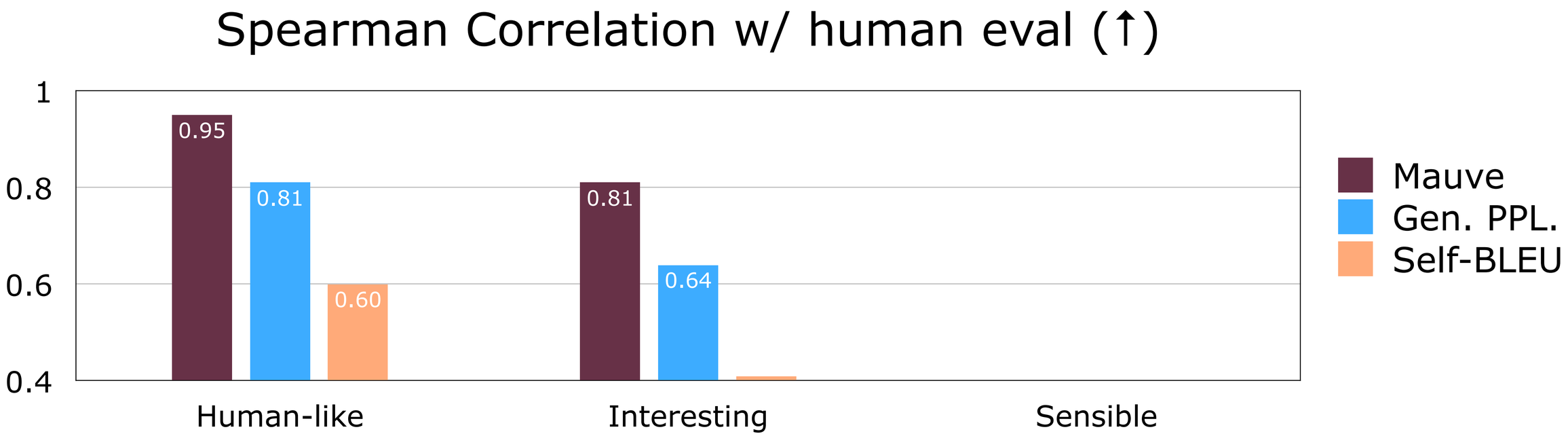
Self-BLEU: Zhu et al. (2018)

Mauve correlates with human judgements

Gen. PPL.: Holtzman et al. (ICLR 2020)

Self-BLEU: Zhu et al. (2018)

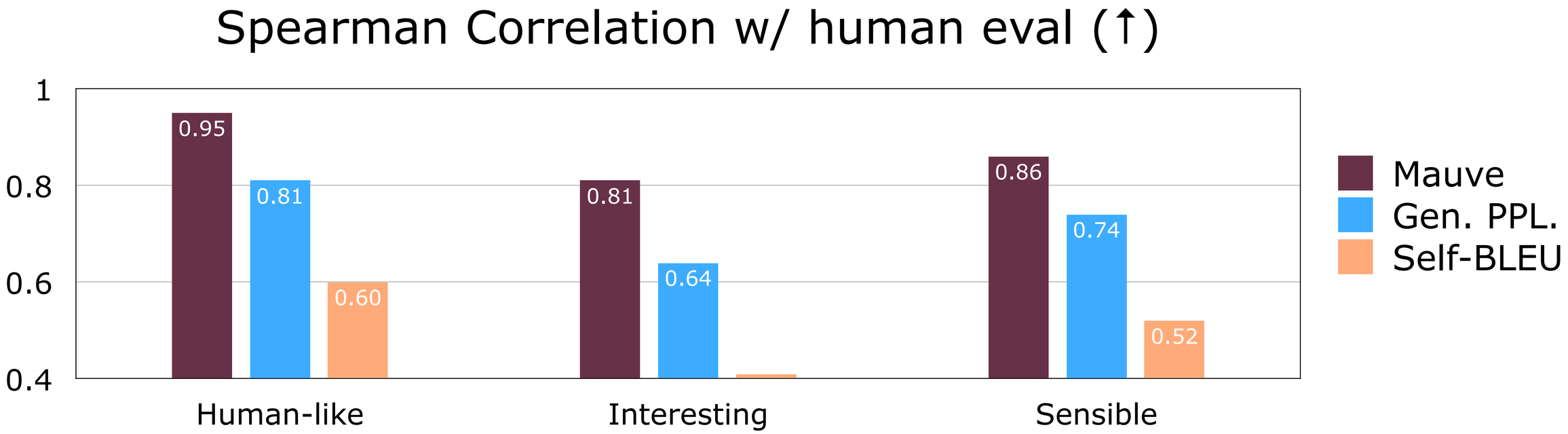
Mauve correlates with human judgements



Gen. PPL.: Holtzman et al. (ICLR 2020)

Self-BLEU: Zhu et al. (2018)

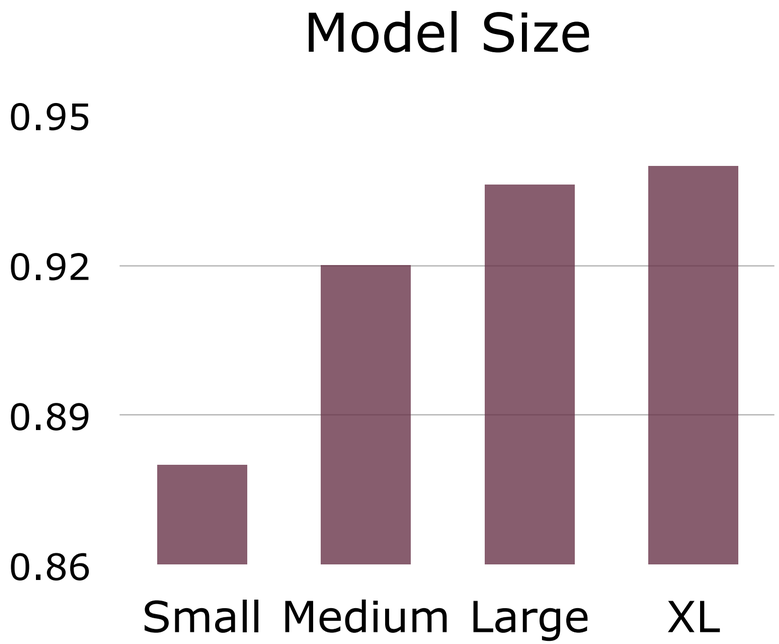
Mauve correlates with human judgements



Gen. PPL.: Holtzman et al. (ICLR 2020)

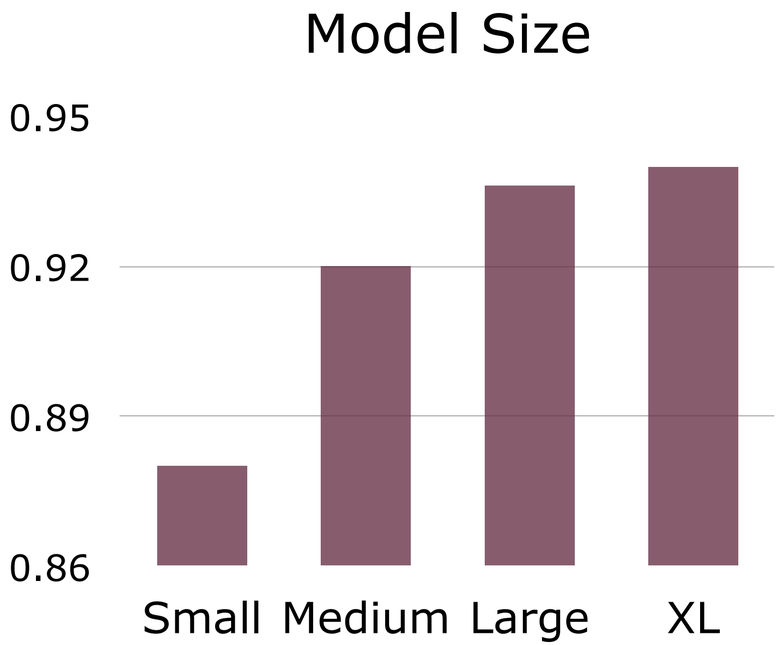
Self-BLEU: Zhu et al. (2018)

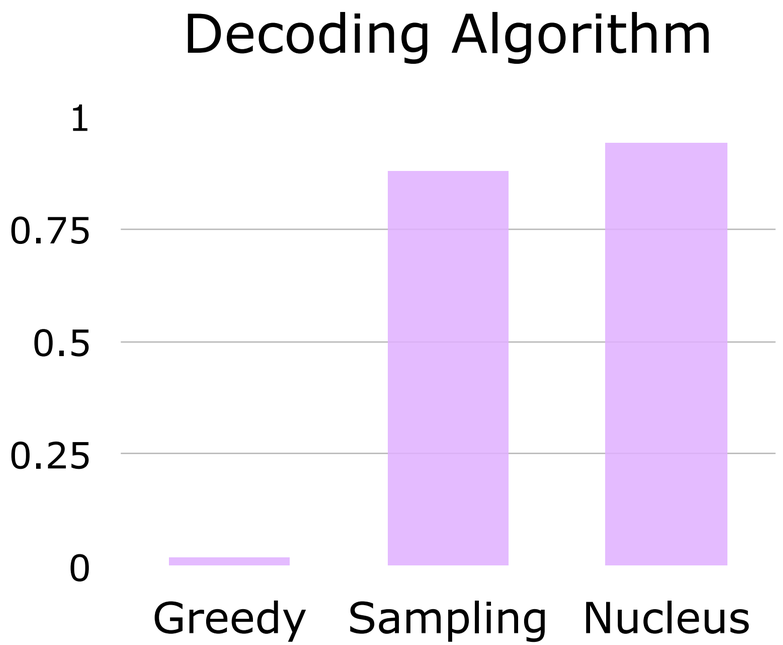
Mauve captures important trends



- Y-axis shows Mauve (↑)

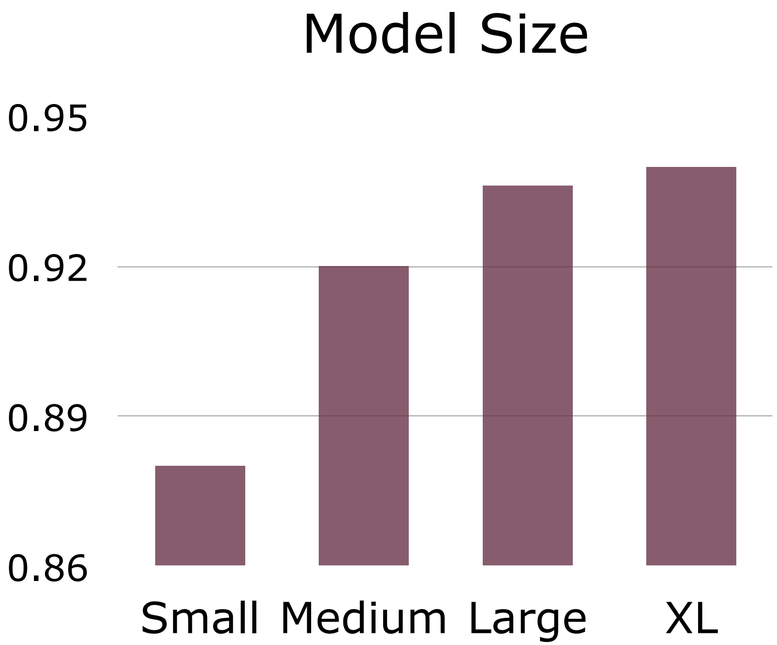
Mauve captures important trends

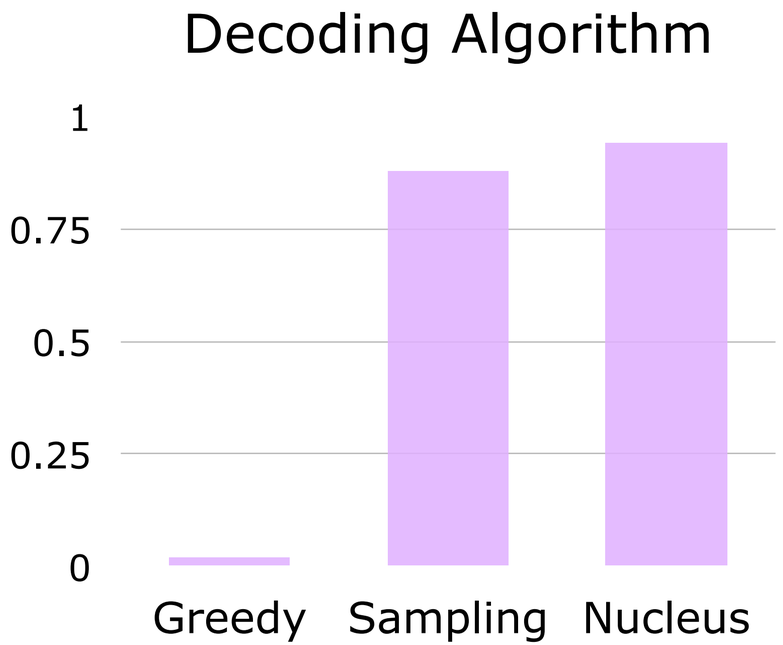


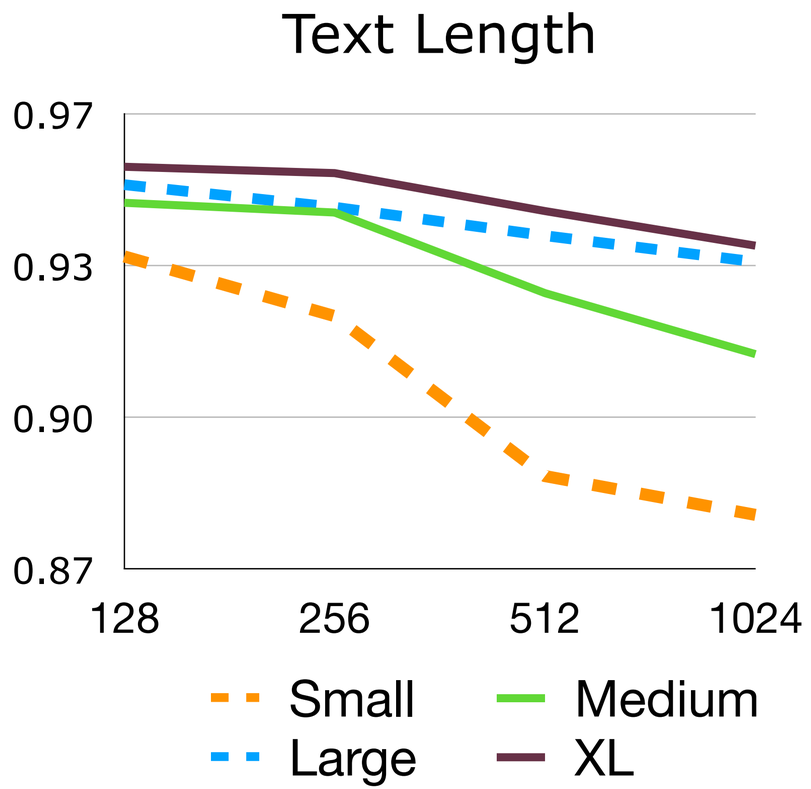


- Y-axis shows Mauve (↑)

Mauve captures important trends



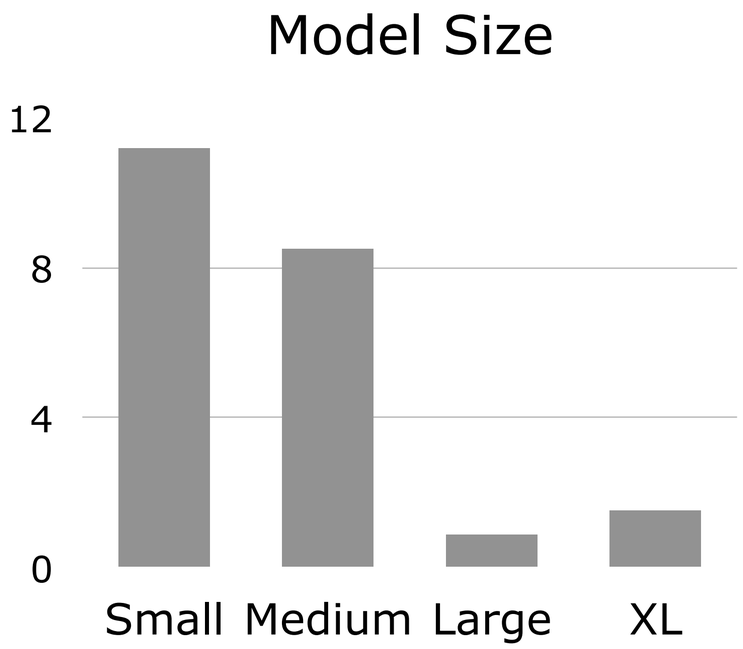




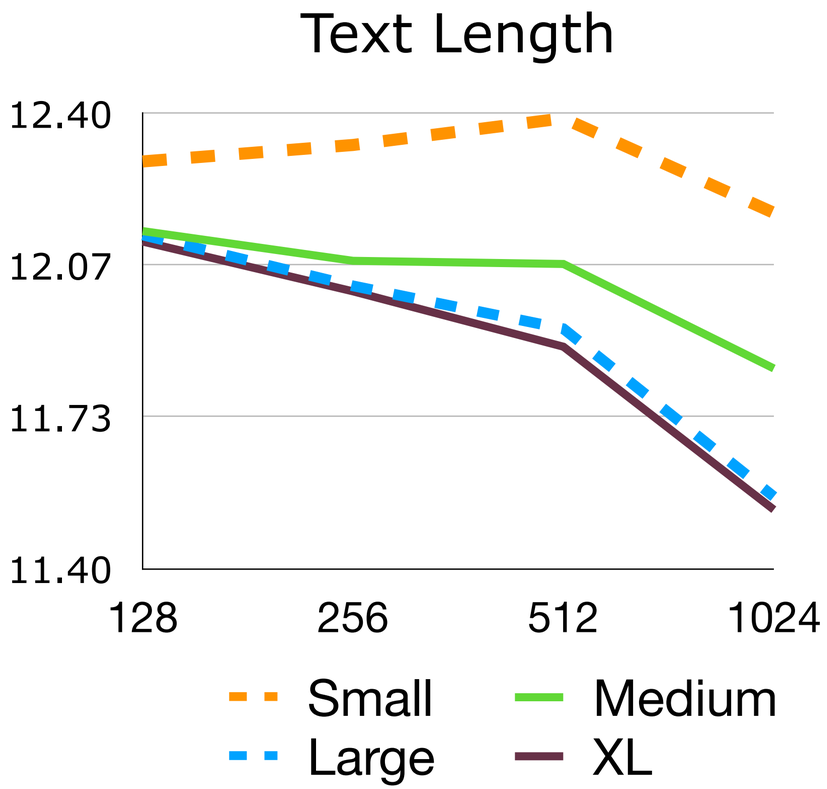
- Y-axis shows Mauve (↑)

Baselines fail to captures important trends

Gen. PPL. (↓)

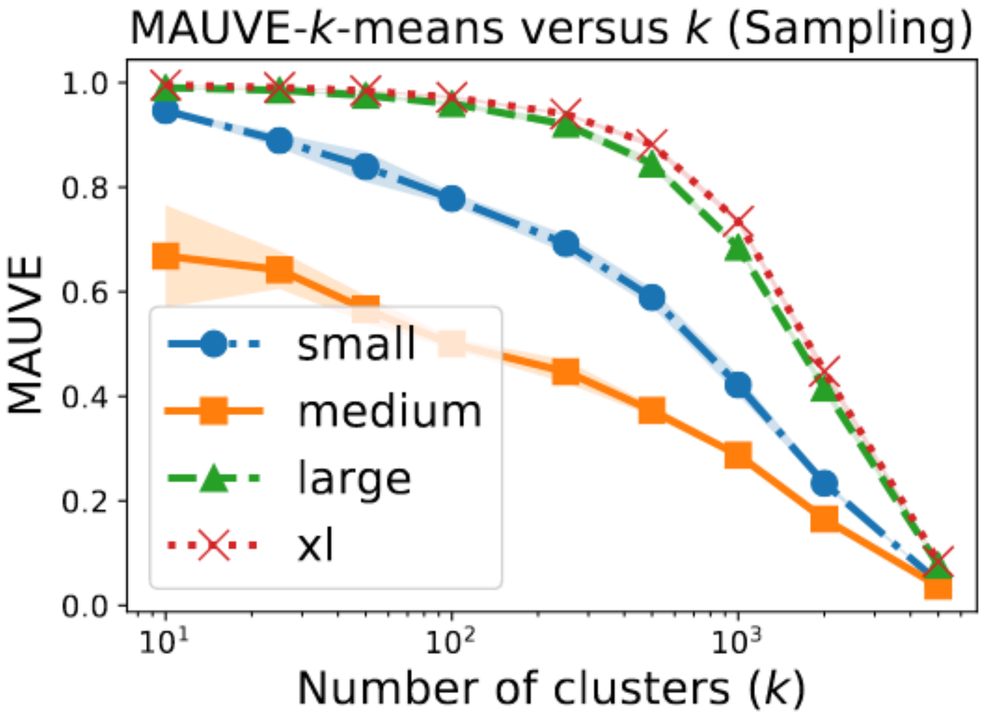


Fréchet distance (↓)



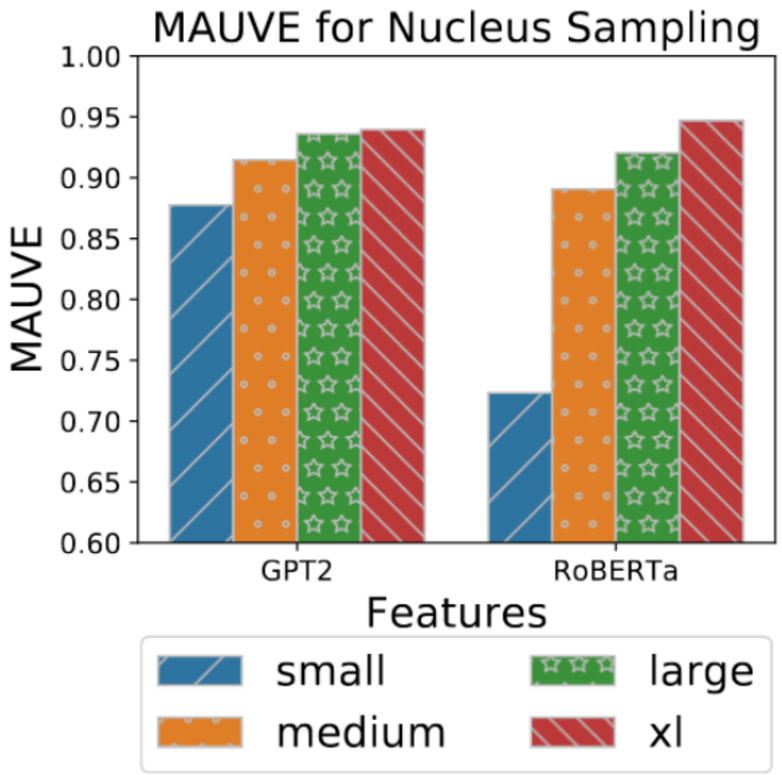
Mauve is robust to hyperparameter choices

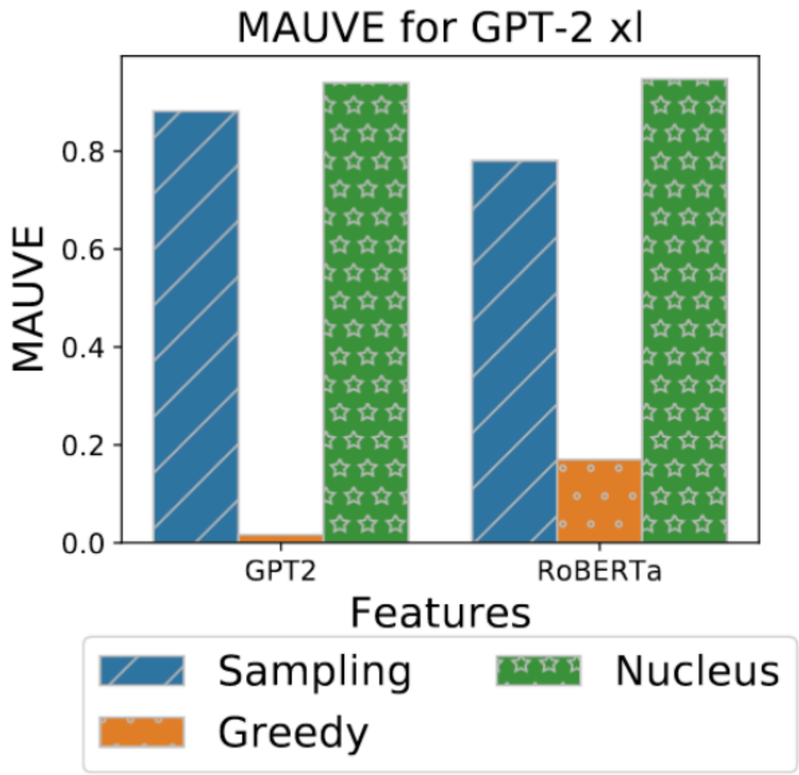
Quantization



Mauve is robust to hyperparameter choices

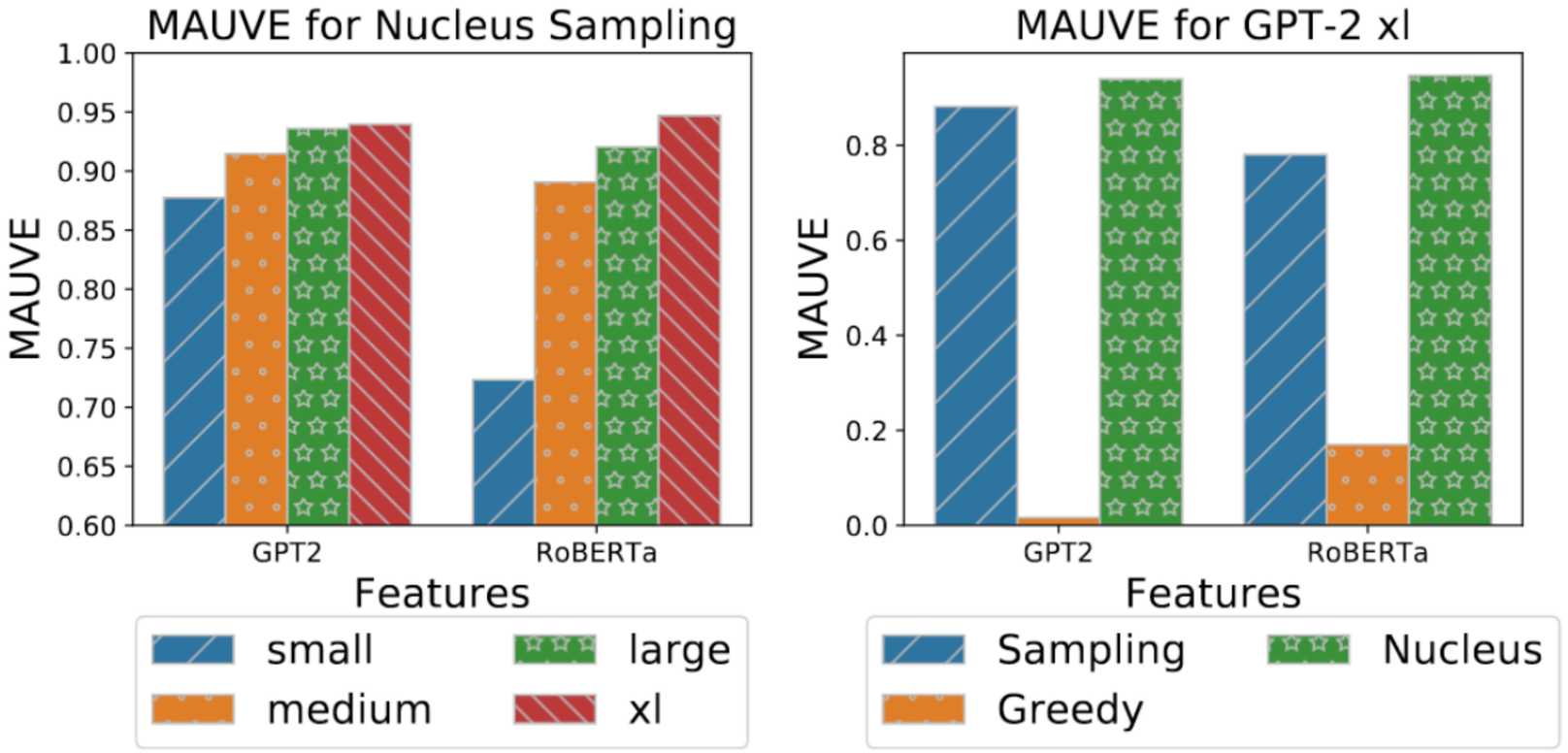
Encoding Model





Discussion

What do different embedding models capture?



Can we use Mauve to quantify goodness of embedding models?

Theory

Theorem (informal)

There exists a quantization such that the approximation error of Mauve is

$$\widetilde{O}\left(\sqrt{\frac{k}{n}} + \frac{1}{k} \right)$$

Statistical Error

Quantization Error

n: number of samples from P and Q

k: quantization size

[Liu, Pillutla, Welleck, Oh, Choi, Harchaoui. NeurIPS 2021]

Software & Code

Software to compute Mauve: **pip install mauve-text**

Github (software): https://github.com/krishnap25/mauve

Software & Code

Software to compute Mauve: **pip install mauve-text**

Github (software): https://github.com/krishnap25/mauve

```
from mauve import compute_mauve

p_text = ... # list of strings for human distribution P
q_text = ... # list of strings for model distribution Q

# Obtain deep encoding, quantize it and compute Mauve
out = compute_mauve(p_text=p_text, q_text=q_text)

print(f'Mauve(P, Q) = {out.mauve}')
```

Software & Code

Software to compute Mauve: **pip install mauve-text**

Github (software): https://github.com/krishnap25/mauve

```
from mauve import compute_mauve

p_text = ... # list of strings for human distribution P
q_text = ... # list of strings for model distribution Q

# Obtain deep encoding, quantize it and compute Mauve
out = compute_mauve(p_text=p_text, q_text=q_text)

print(f'Mauve(P, Q) = {out.mauve}')
```

Github (experiments): https://github.com/krishnap25/mauve-experiments

Software & Code

Software to compute Mauve: **pip install mauve-text**

Github (software): https://github.com/krishnap25/mauve

```
from mauve import compute_mauve

p_text = ... # list of strings for human distribution P
q_text = ... # list of strings for model distribution Q

# Obtain deep encoding, quantize it and compute Mauve
out = compute_mauve(p_text=p_text, q_text=q_text)

print(f'Mauve(P, Q) = {out.mauve}')
```

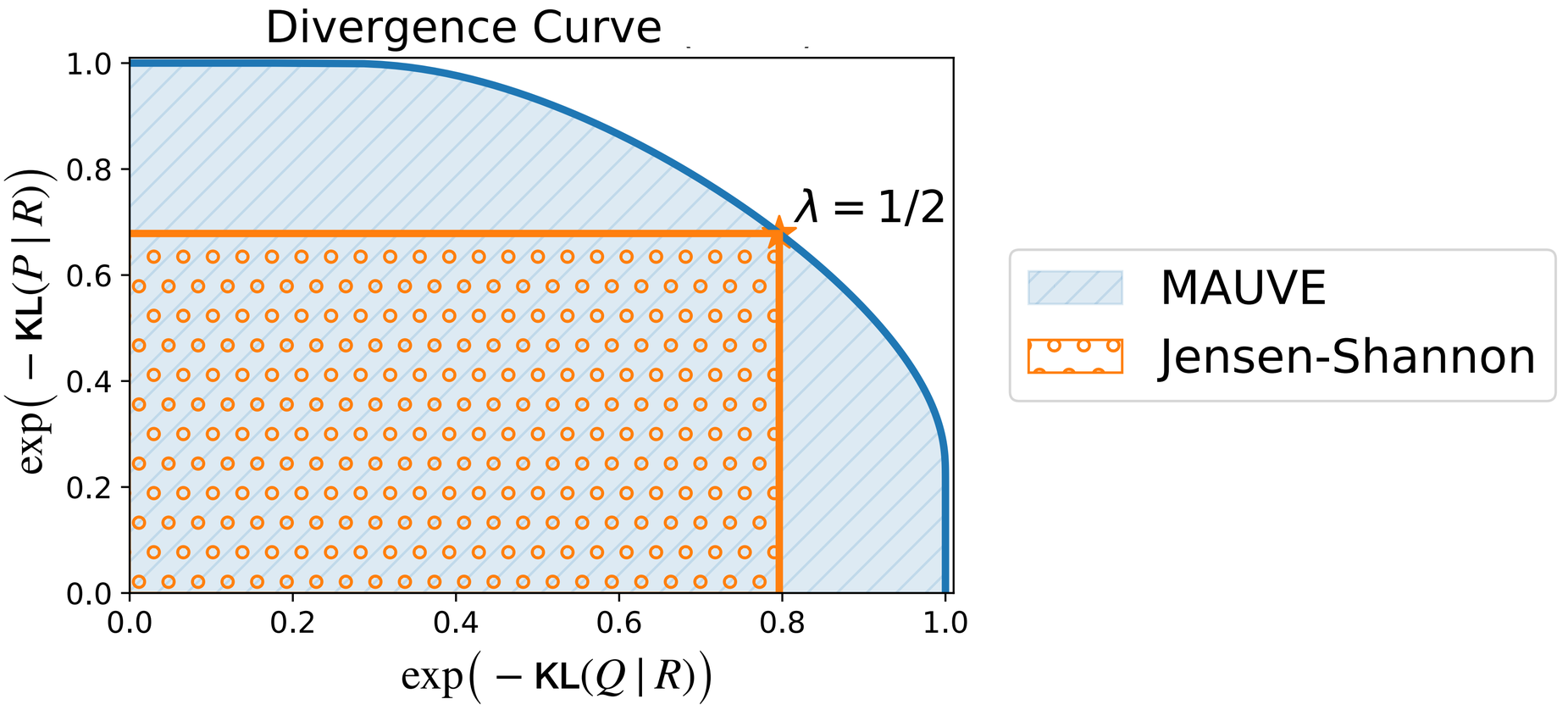
Thank you!

Github (experiments): https://github.com/krishnap25/mauve-experiments

Supplement

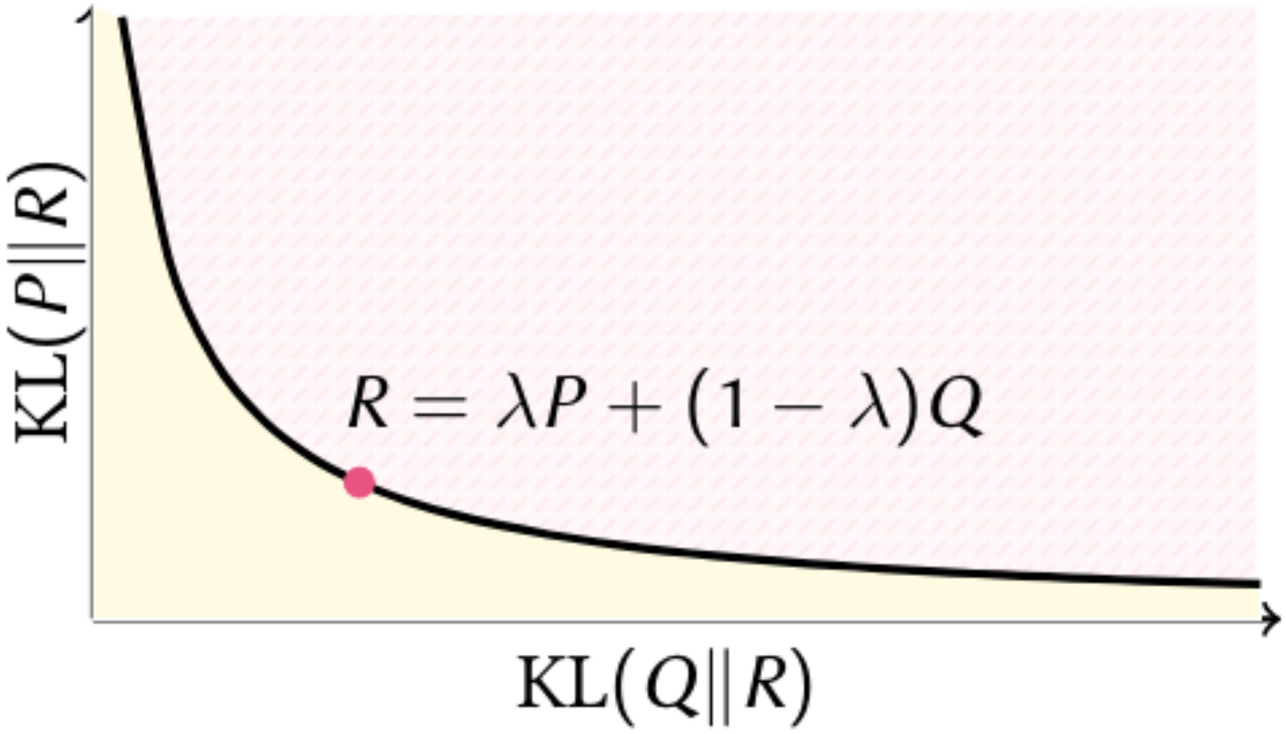
Other summaries of the divergence curve

$$\mathrm{JS}(P, Q) = \frac{1}{2}\big(\mathrm{KL}(P \mid R) + \mathrm{KL}(Q \mid R)\big) \quad \text{where} \quad R = (P + Q)/2$$



Other summaries of the divergence curve

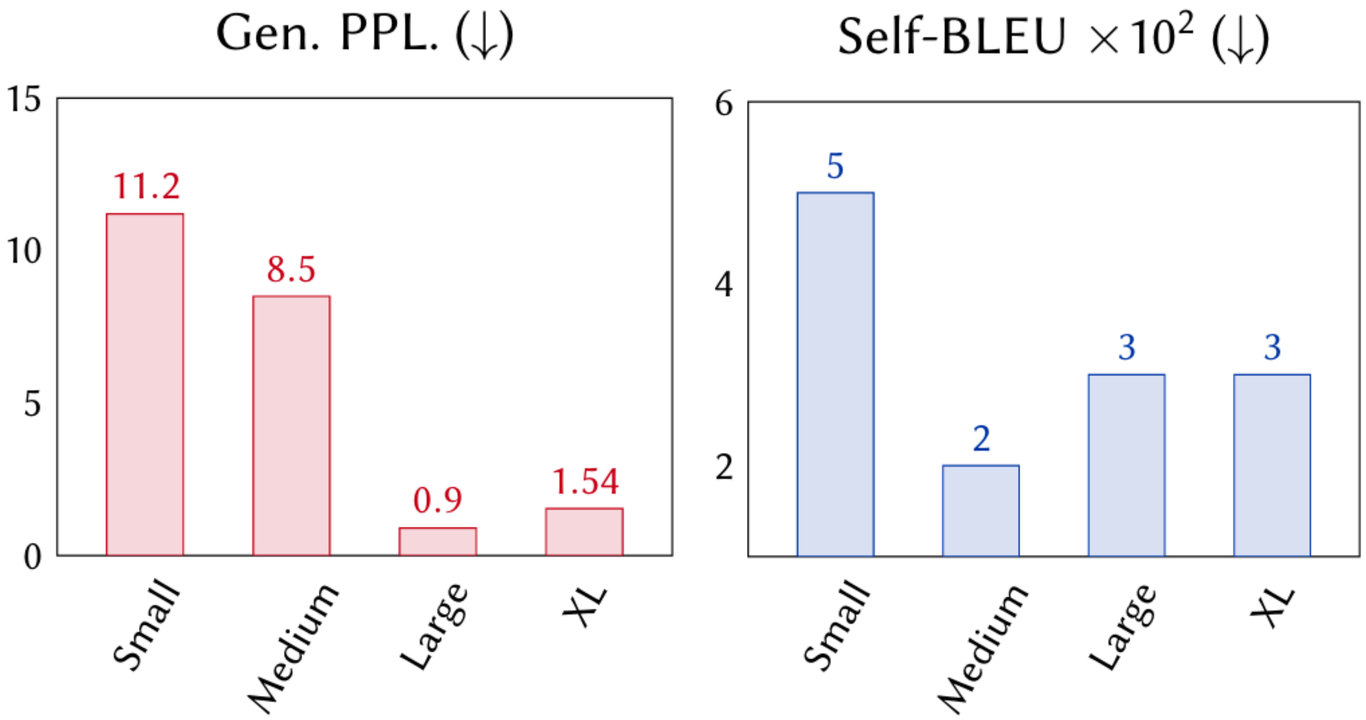
$$\mathrm{FI}(P,Q) = 2\int_0^1 L_\lambda(P,Q)\,\mathrm{d}\lambda \quad \text{where} \quad L_\lambda(P,Q) = \lambda\,\mathrm{KL}(P \mid R_\lambda) + (1-\lambda)\,\mathrm{KL}(Q \mid R_\lambda)$$



[Liu, Pillutla, Welleck, Oh, Choi, Harchaoui. NeurIPS 2021]

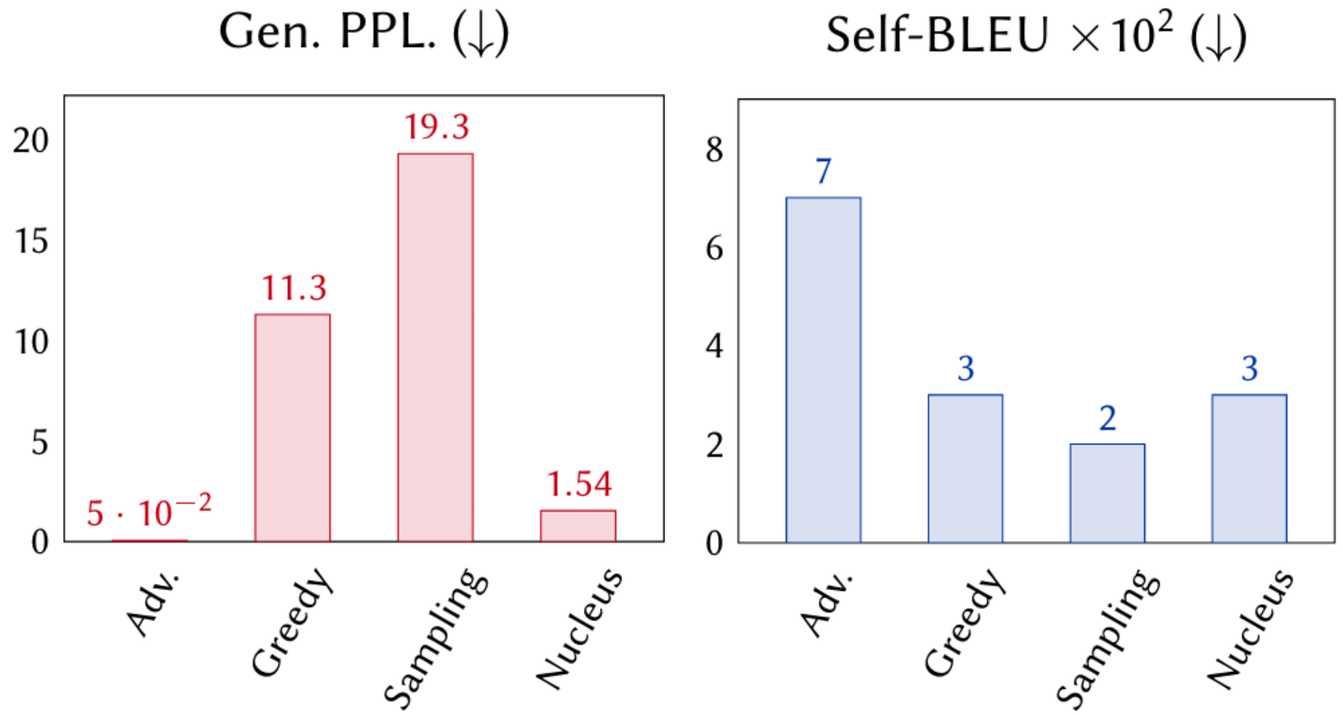
Baselines fail to captures important trends

Model Size



Baselines fail to captures important trends

Decoding Algorithm



Interpreting the quantization

News data: analyze news source (not seen by Mauve)

Groupings: semantic similarity in clusters:

Source: Only one or two sources per cluster

Geographical: Multiple sources from Canada/South Asia/UK

Conglomerate: Multiple sources from same parent company

Subject: Multiple sources from same subject: finance, etc.