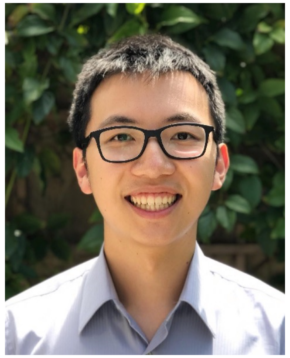
Long Horizon Temperature Scaling

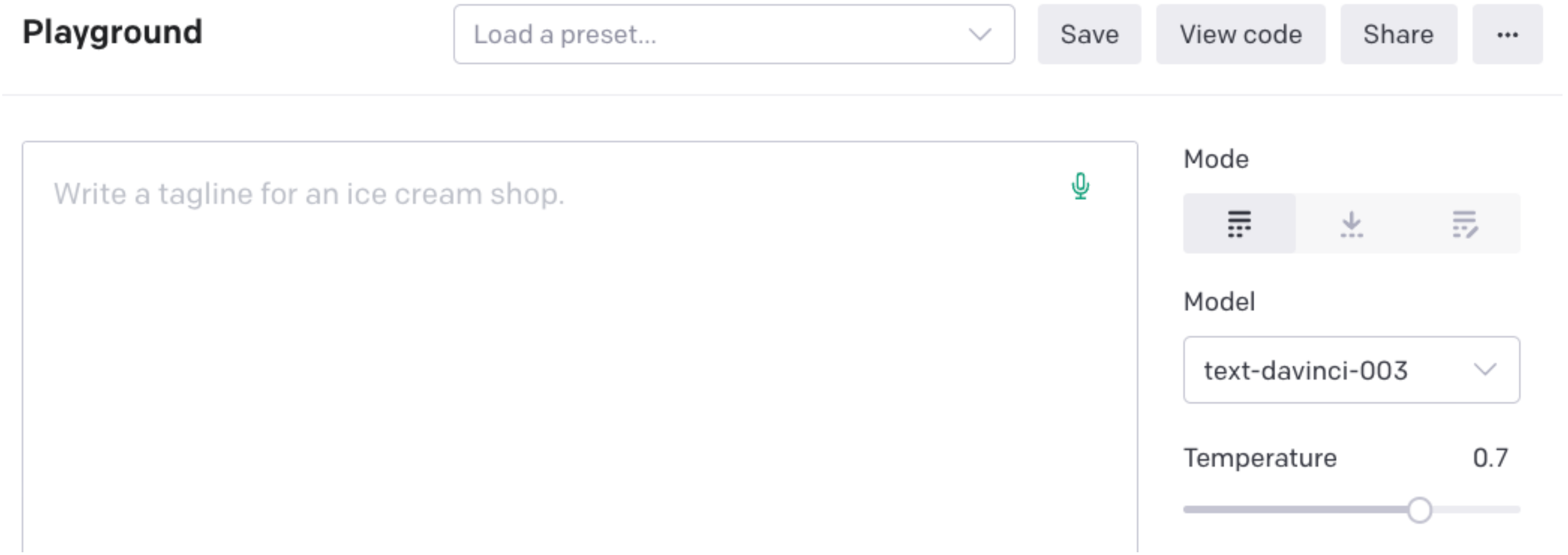

Andy Shih


Dorsa Sadigh

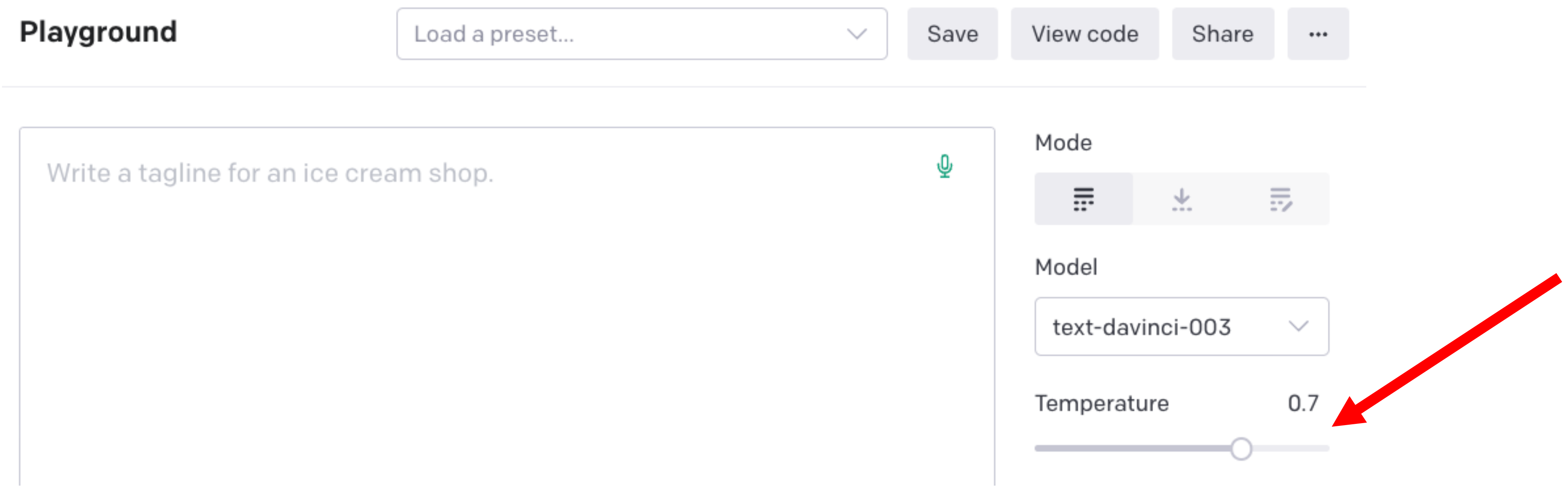

Stefano Ermon

Stanford University

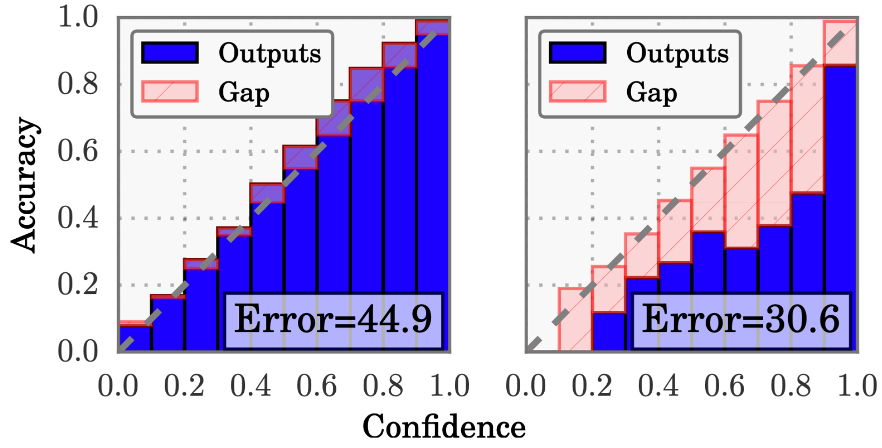
Temperature Scaling

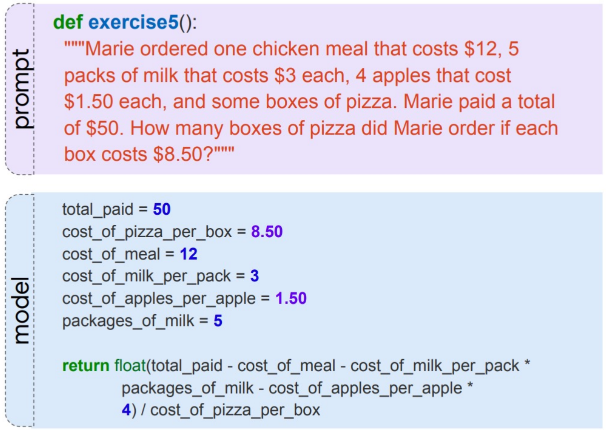


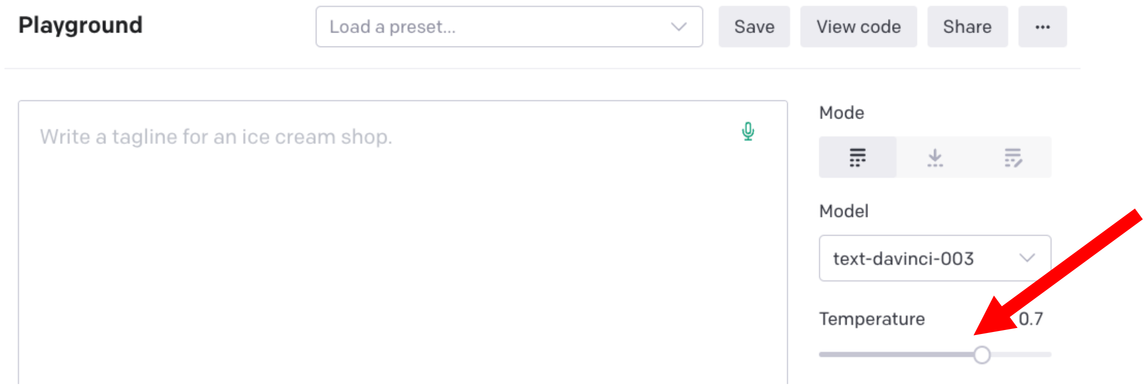
Temperature Scaling



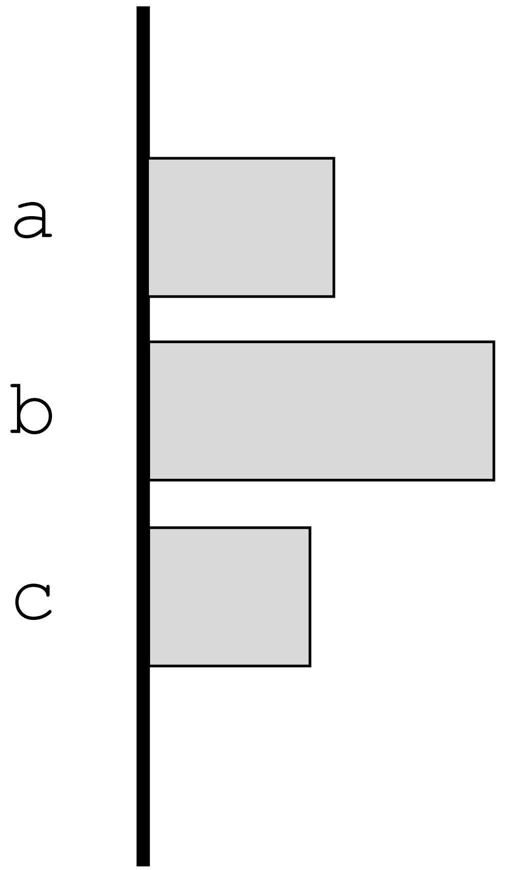
Temperature Scaling

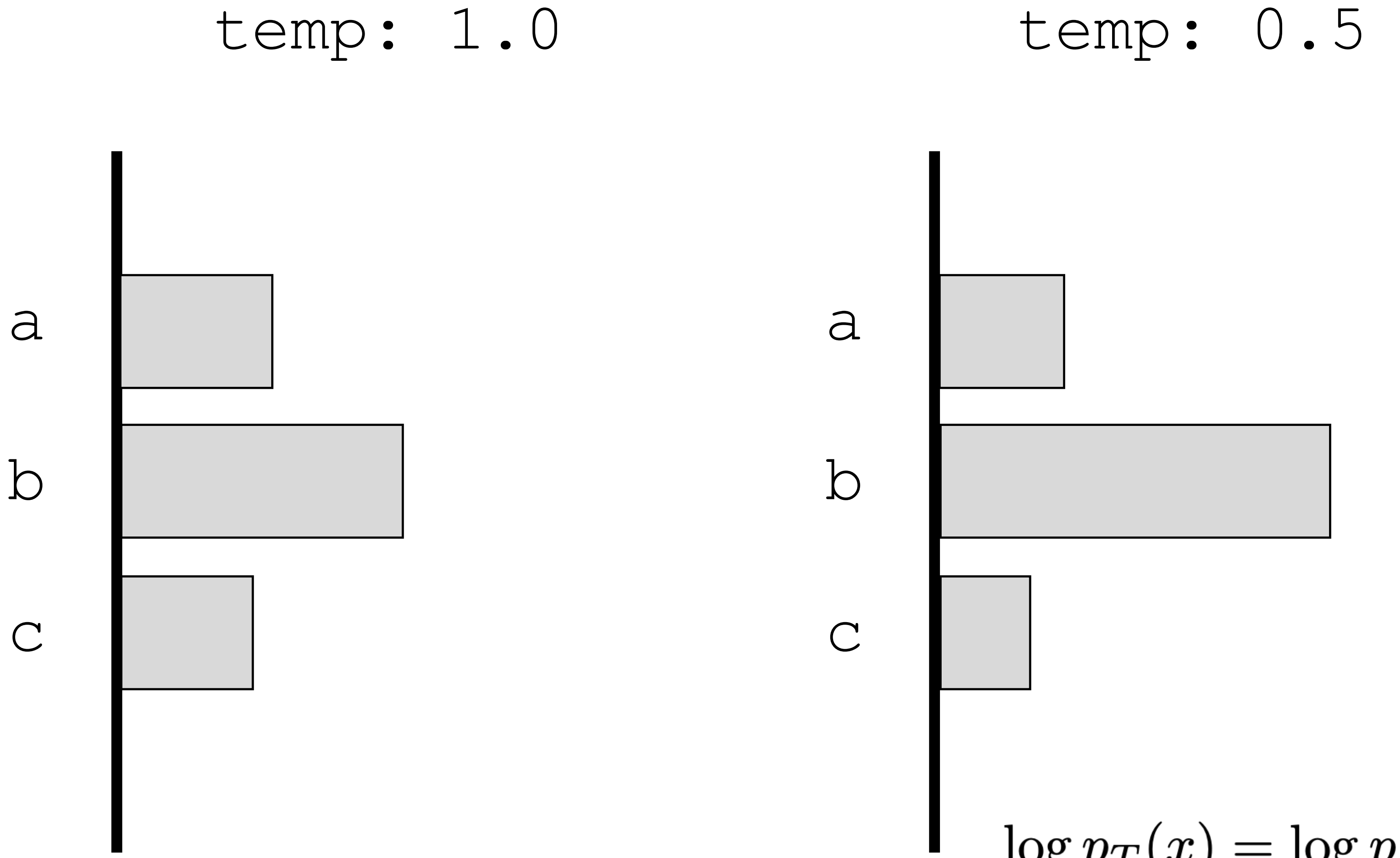






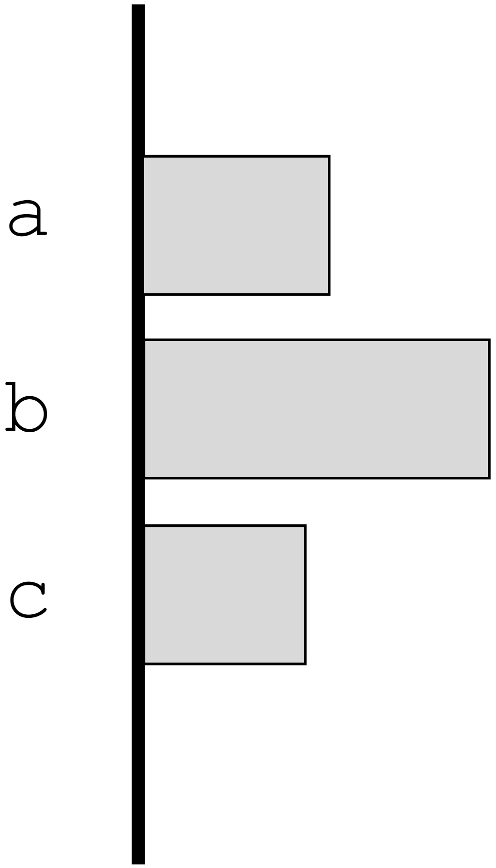
temp: 1.0



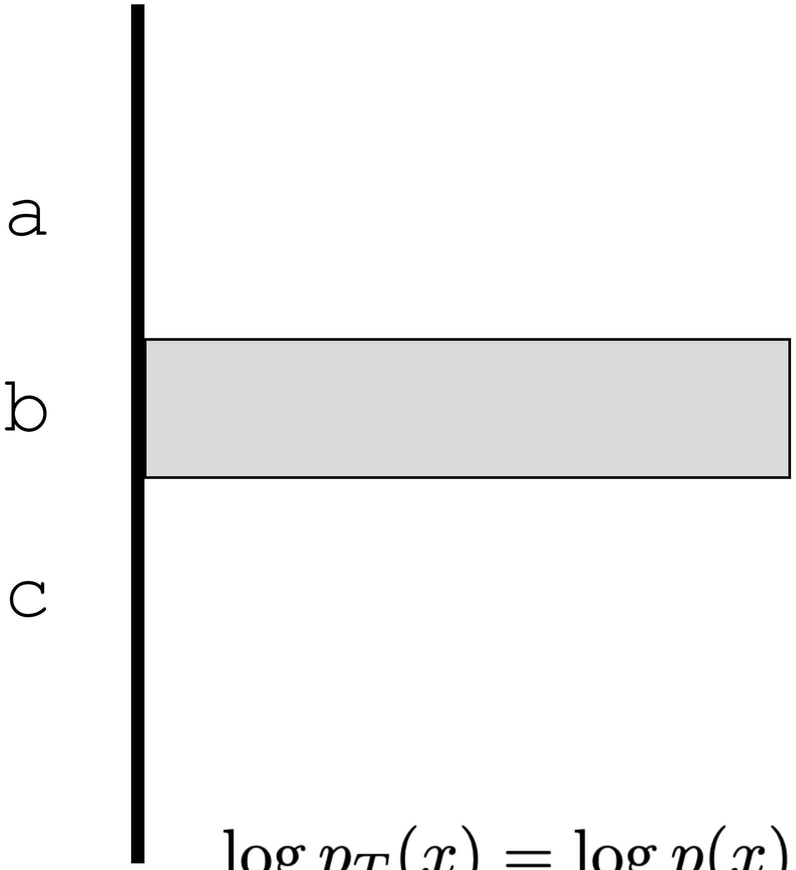


$$\log p_T(x) = \log p(x)/T - \log Z_{p_T}$$

temp: 1.0



temp: 0.0



$$\log p_T(x) = \log p(x)/T - \log Z_{p_T}$$

More Likely Samples

When T < 1, we bias sampling towards high likelihood regions

$$\log p_T(x) = \log p(x)/T - \log Z_{p_T}$$

When T=0, we compute argmax

More Likely Samples

When T < 1, we bias sampling towards high likelihood regions

$$\log p_T(x) = \log p(x)/T - \log Z_{p_T}$$

When T=0, we compute argmax

But…

Myopic Temperature

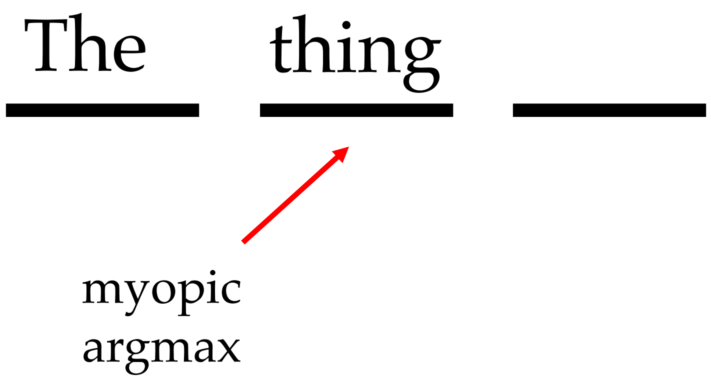
But current LMs temperature scale one token at a time…

T=0, greedy decoding

Myopic Temperature

But current LMs temperature scale one token at a time…

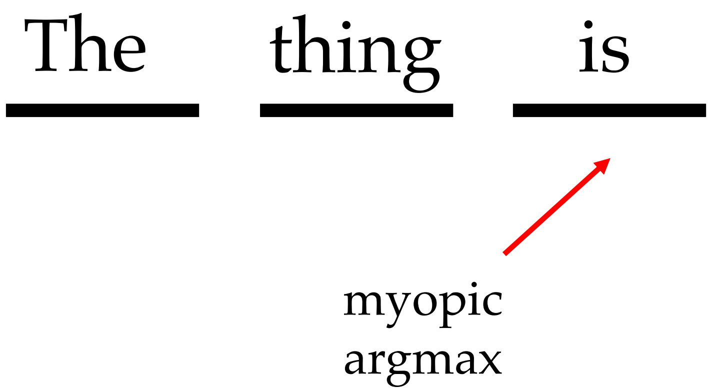
T=0, greedy decoding



Myopic Temperature

But current LMs temperature scale one token at a time…

T=0, greedy decoding



Myopic Temperature

But current LMs temperature scale one token at a time…

T=0, greedy decoding

The thing is

myopic
argmax

$$\log p_T(x) \neq \sum_i \log p_T^{\text{myopic}}(x_i | x_{<i})$$

Myopic Temperature

But current LMs temperature scale one token at a time…

T=0, greedy decoding

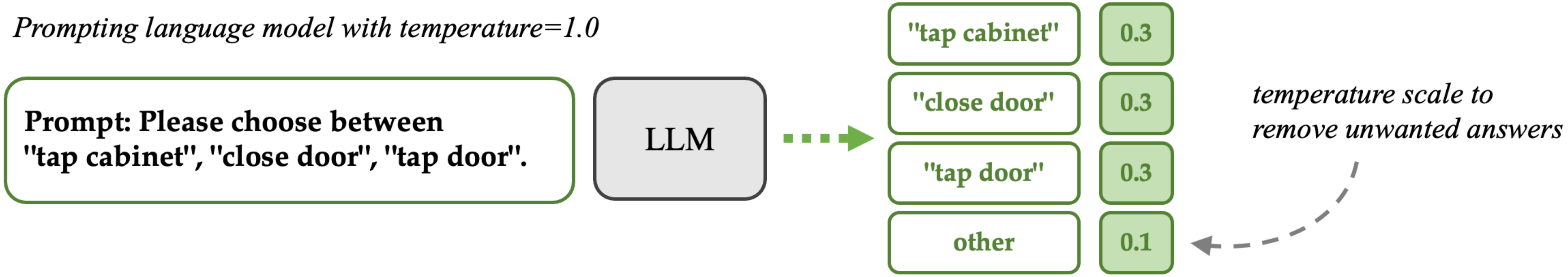
The thing is

How are you

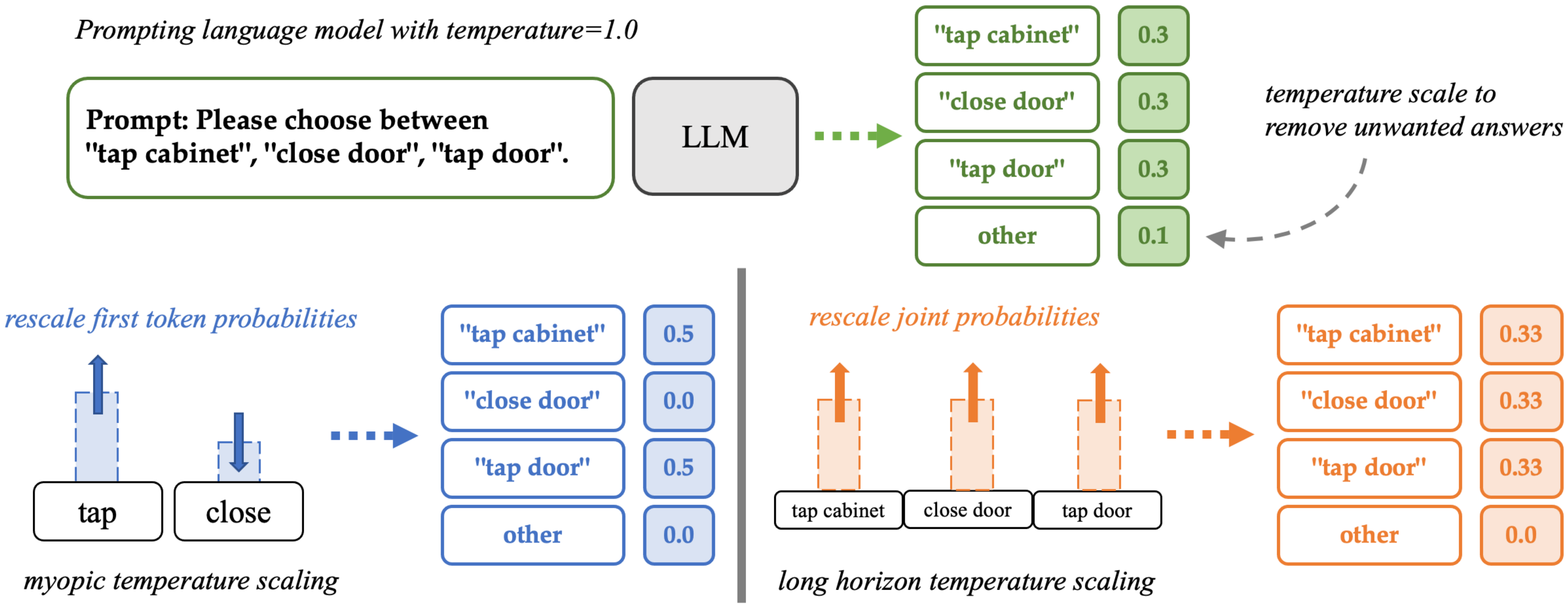
non-myopic argmax

$$\log p_T(x) \neq \sum_i \log p_T^{\text{myopic}}(x_i | x_{<i})$$

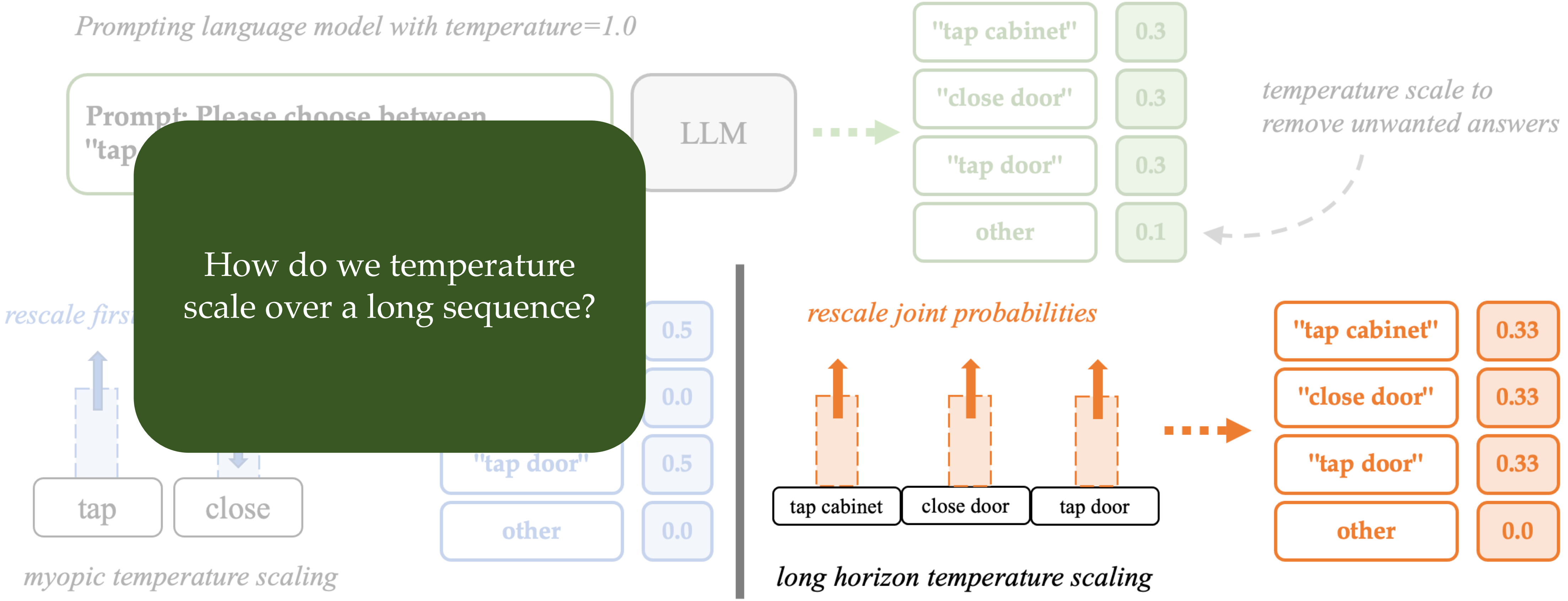
Pitfall of Myopic Temperature Scaling



Pitfall of Myopic Temperature Scaling



Pitfall of Myopic Temperature Scaling



How do we temperature scale over a long sequence?

Long Horizon Temperature Scaling

Long Horizon Temperature Scaling

Non-myopic Temperature Scaling
For Optimizing Long Sequences

Long Horizon Temperature Scaling

$\hat{p}$ p

data model

Want: $\log p_T(x) = \log p(x)/T - \log Z_{p_T}$

Long Horizon Temperature Scaling

$\hat{p}$ data

p model

→ distill

q_T temperature scaled model

Want: $\log p_T(x) = \log p(x)/T - \log Z_{p_T}$

Long Horizon Temperature Scaling

$\hat{p}$	p		q_T
data	model	→ distill	temperature scaled model

Want: $\log p_T(x) = \log p(x)/T - \log Z_{p_T}$

Objective: $D_{KL}(p_T||q_T) = \mathbb{E}_{x\sim p_T}[\frac{\log p(x)}{T} - \log q_T(x)] - \log Z_{p_T}$

Long Horizon Temperature Scaling

$\hat{p}$	p		q_T
data	model	→ distill	temperature scaled model

Want: $\log p_T(x) = \log p(x)/T - \log Z_{p_T}$

Objective: $D_{KL}(p_T || q_T) = \mathbb{E}_{x \sim p_T}[\frac{\log p(x)}{T} - \log q_T(x)] - \log Z_{p_T}$

constant, can ignore

Long Horizon Temperature Scaling

$\hat{p}$ data

p model

distill

q_T temperature scaled model

Objective: $-\mathbb{E}_{x \sim p_T}[\log q_T(x)]$

but sampling from p_T is hard

Long Horizon Temperature Scaling

$\hat{p}$ data

p model

distill →

q_T temperature scaled model

Objective:

$$-\mathbb{E}_{x\sim p_T}[\log q_T(x)]$$

$$-\mathbb{E}_{x\sim p}\frac{e^{\log p(x)/T - \log Z_{p_T}}}{p(x)}[\log q_T(x)]$$ ← importance sampling

$$-\mathbb{E}_{x\sim p}\exp(\frac{1-T}{T}\log p(x))[\log q_T(x)]$$

Long Horizon Temperature Scaling

$\hat{p}$	p		q_T
data	model	→ distill	temperature scaled model

Objective:

$$-\mathbb{E}_{x\sim p_T}[\log q_T(x)]$$

$$-\mathbb{E}_{x\sim p}\frac{e^{\log p(x)/T - \log Z_{p_T}}}{p(x)}[\log q_T(x)]$$ ← importance sampling

$$-\mathbb{E}_{x\sim \hat{p}}\exp(\frac{1-T}{T}\log p(x))[\log q_T(x)]$$

speed up by using data instead of sample

Long Horizon Temperature Scaling

$\hat{p}$ data

p model

distill →

q_T temperature scaled model

Non-myopic

Applicable to all likelihood-based models

Objective: $-\mathbb{E}_{x\sim\hat{p}} \exp(\frac{1-T}{T} \log p(x))[\log q_T(x)]$

Variance Reduction: the clean

Learnable Baseline

multiplicative constant

$$-\mathbb{E}_{x\sim p}\frac{e^{\log p(x)/T-\log Z_{p_T}}}{p(x)}[\log q_T(x)]$$

Variance Reduction: the clean

Learnable Baseline

multiplicative constant

$$-\mathbb{E}_{x\sim p}\frac{e^{\log p(x)/T-b}}{p(x)}[\log q_T(x)]$$

Variance Reduction: the clean

Learnable Baseline

multiplicative constant

$$-\mathbb{E}_{x\sim p}\frac{e^{\log p(x)/T-b}}{p(x)}[\log q_T(x)]$$

$$b=\frac{1}{|\mathcal{D}|}\sum_{x\in\mathcal{D}}\frac{1-T}{T}\log p(x)$$

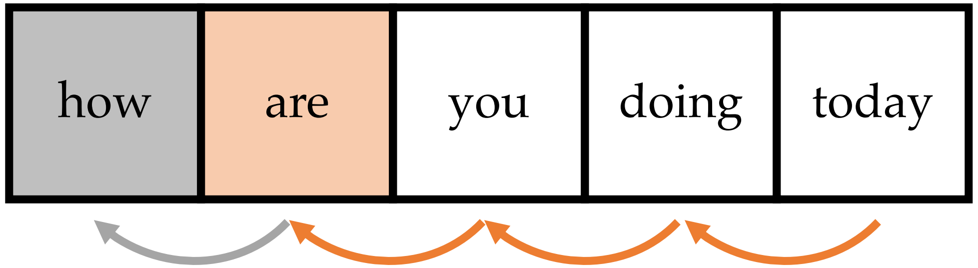
Variance Reduction: the clean

Learnable Baseline

multiplicative constant

$$-\mathbb{E}_{x\sim p}\frac{e^{\log p(x)/T-b}}{p(x)}[\log q_T(x)] \qquad b = \frac{1}{|\mathcal{D}|}\sum_{x\in\mathcal{D}}\frac{1-T}{T}\log p(x)$$

Suffix likelihood and Index-dependent Baseline (for AR models)



$$b(i) = \frac{1}{|\mathcal{D}|}\sum_{x\in\mathcal{D}}\frac{1-T}{T}\log p(x_{\geq i}|x_{<i})$$

Variance Reduction: the messy

Weight clipping

baseline b

$$\mathrm{CLIP}\Big(\exp(\frac{1-T}{T}\log p(x) - b), c\Big)$$

clip c

Variance Reduction: the messy

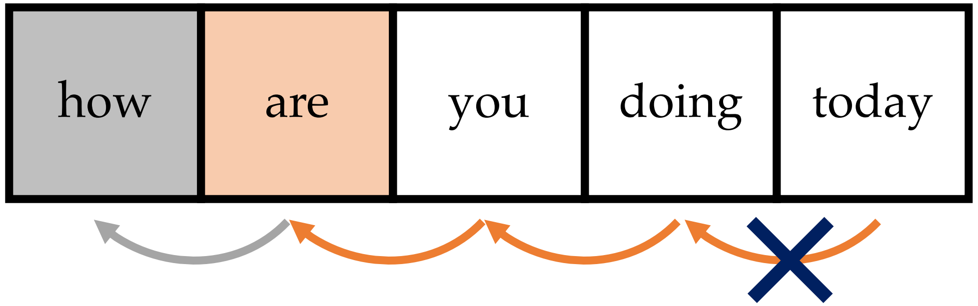
Weight clipping

$$\text{CLIP}\Big(\exp(\frac{1-T}{T}\log p(x) - b), c\Big)$$

baseline b

clip c

Horizon clipping (for AR models)



$$-\mathbb{E}_{x\sim\hat{p}} \exp\left(\frac{1-T}{T}\log p(x_{\geq i}|x_{<i}) - b(i)\right)[\log q_T(x_{\geq i}|x_{<i})]$$

$x \sim \hat{p}$

how	are	you	doing	today

$$-\mathbb{E}_{x\sim\hat{p}} \exp\left(\frac{1-T}{T} \log p(x_{\geq i}|x_{<i}) - b(i)\right)[\log q_T(x_{\geq i}|x_{<i})]$$

$x \sim \hat{p}$	how	are	you	doing	today
$\log p(x_i \mid x_{<i})$	-2	-1	-3	-1	-1

$$-\mathbb{E}_{x\sim\hat{p}} \exp\left(\frac{1-T}{T}\log p(x_{\geq i}|x_{<i}) - b(i)\right)[\log q_T(x_{\geq i}|x_{<i})]$$

$x \sim \hat{p}$	how	are	you	doing	today
$\log p(x_i \vert x_{<i})$	-2	-1	-3	-1	-1
$\log p(x_{\geq i} \vert x_{<i})$	-8	-6	-5	-2	-1

reverse cumulative sum

$$-\mathbb{E}_{x\sim\hat{p}} \exp\left(\frac{1-T}{T} \log p(x_{\geq i}|x_{<i}) - b(i)\right)[\log q_T(x_{\geq i}|x_{<i})]$$

$x \sim \hat{p}$	how	are	you	doing	today
$\log p(x_i \| x_{<i})$	-2	-1	-3	-1	-1
$\log p(x_{\geq i} \| x_{<i})$	-8	-6	-5	-2	-1
$\frac{1-T}{T} \log p(x_{\geq i} \| x_{<i})$	-16	-12	-10	-4	-2

reverse cumulative sum

$\frac{1-T}{T}$

$$-\mathbb{E}_{x\sim\hat{p}} \exp\left(\frac{1-T}{T}\log p(x_{\geq i}|x_{<i}) - b(i)\right)[\log q_T(x_{\geq i}|x_{<i})]$$

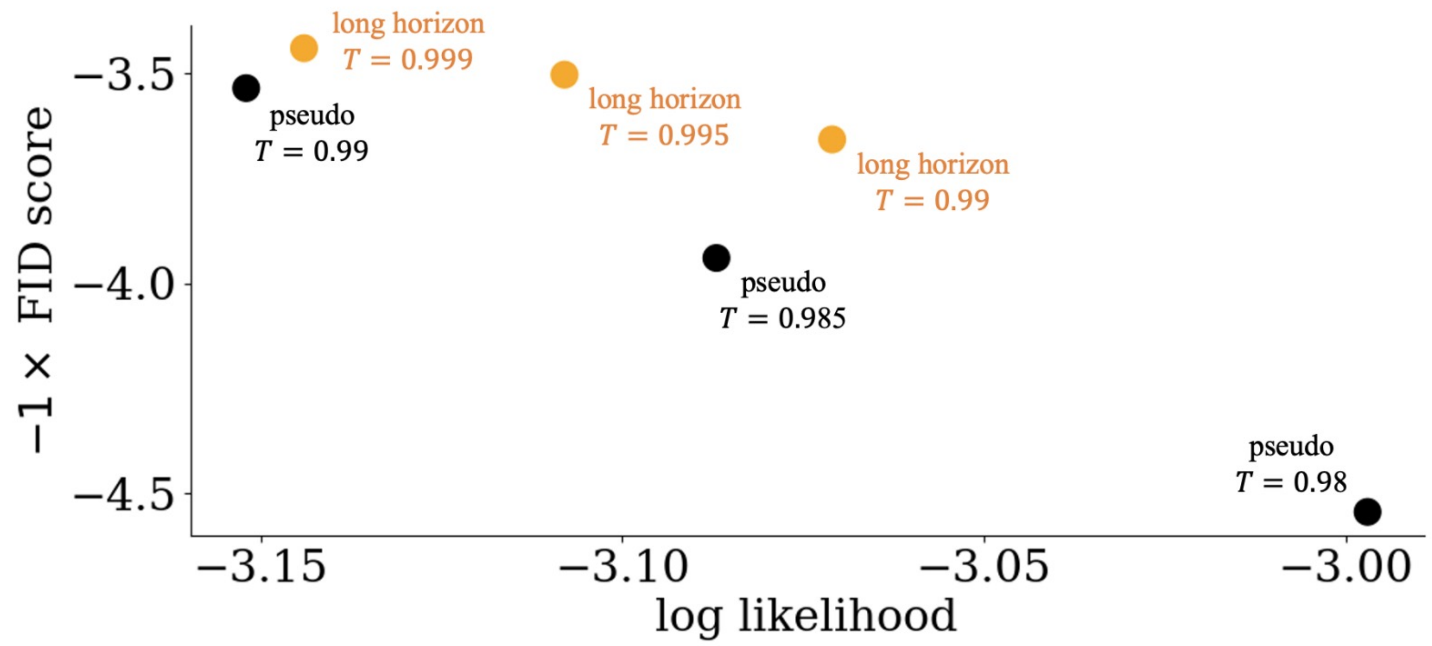
$x \sim \hat{p}$	how	are	you	doing	today
$\log p(x_i \vert x_{<i})$	-2	-1	-3	-1	-1
$\log p(x_{\geq i} \vert x_{<i})$	-8	-6	-5	-2	-1
$\frac{1-T}{T}\log p(x_{\geq i} \vert x_{<i})$	-16	-12	-10	-4	-2
$\frac{1-T}{T}\log p(x_{\geq i} \vert x_{<i}) - b(i)$	-1	0	-1	+2	+1

reverse cumulative sum

$\frac{1-T}{T}$

add baseline
e.g. 3 * len(suffix)

Diffusion Image Models

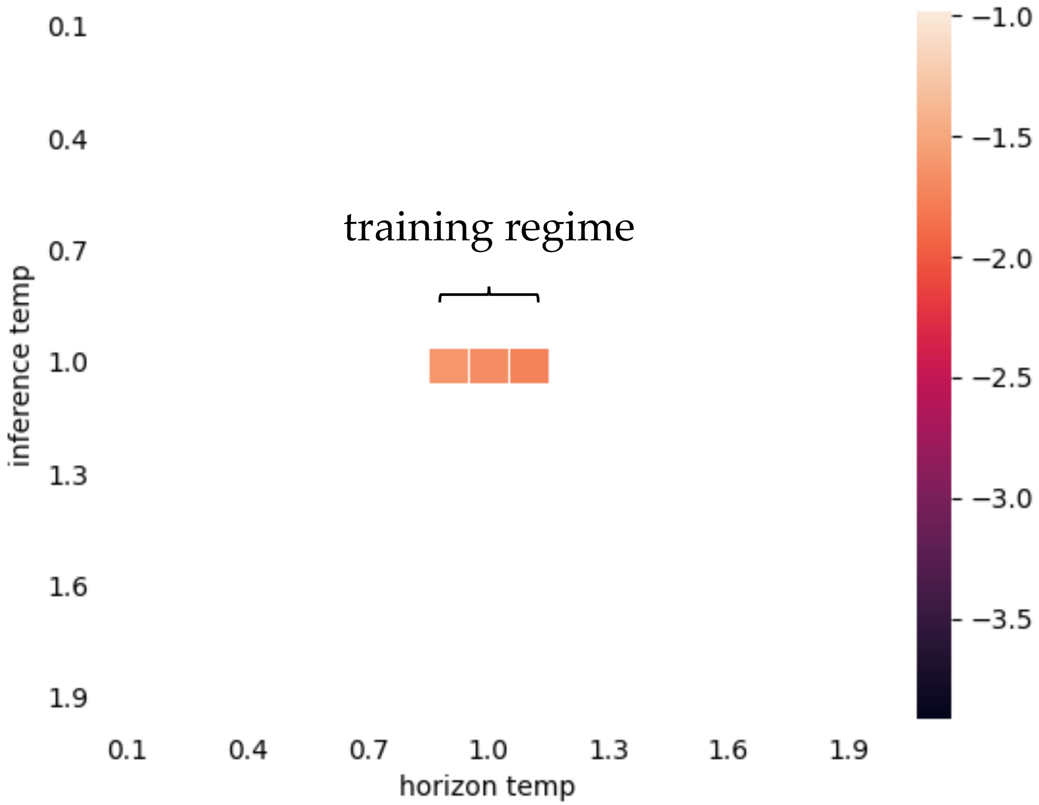


Baseline: pseudo-temp

reduce noise of reverse diffusion

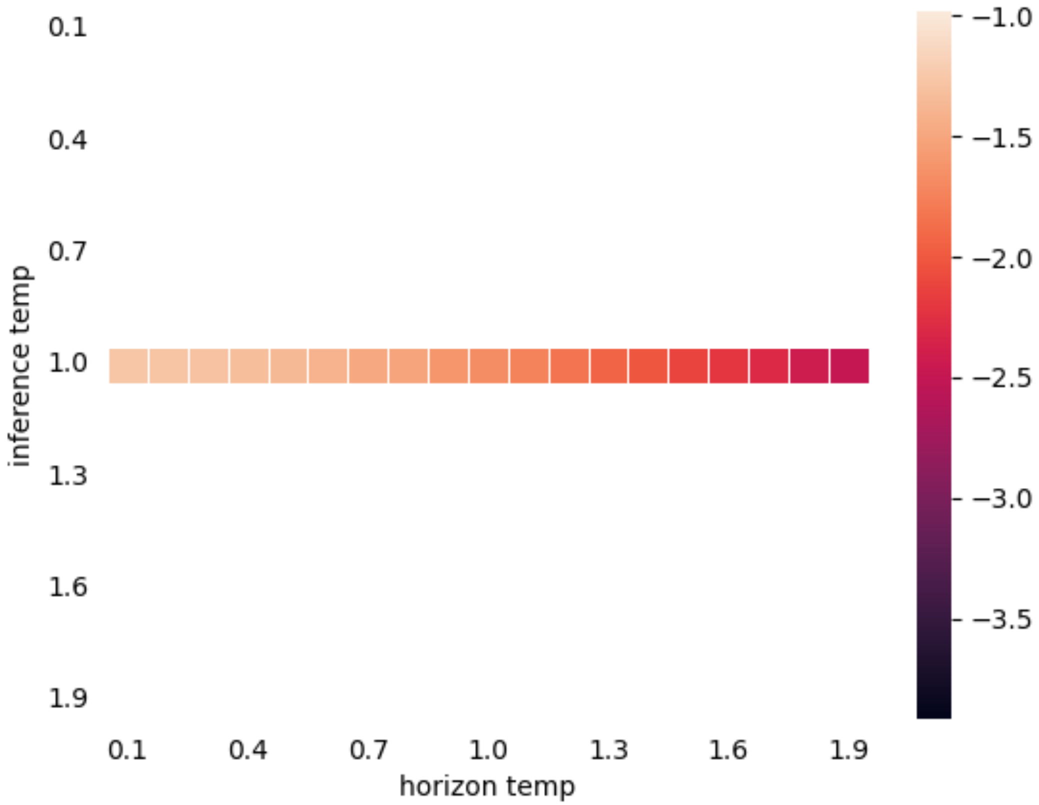
Better likelihood vs diversity tradeoff!

Autoregressive Character Models

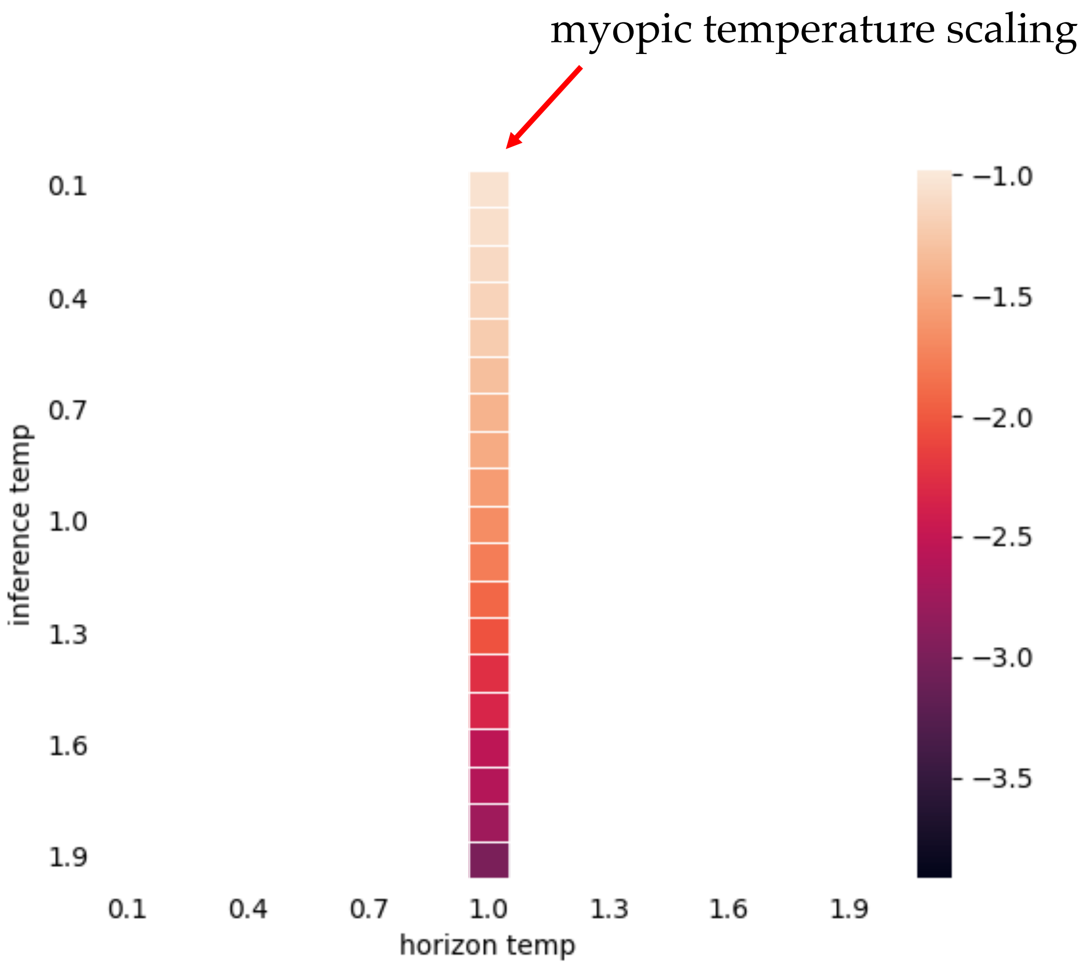


Autoregressive Character Models

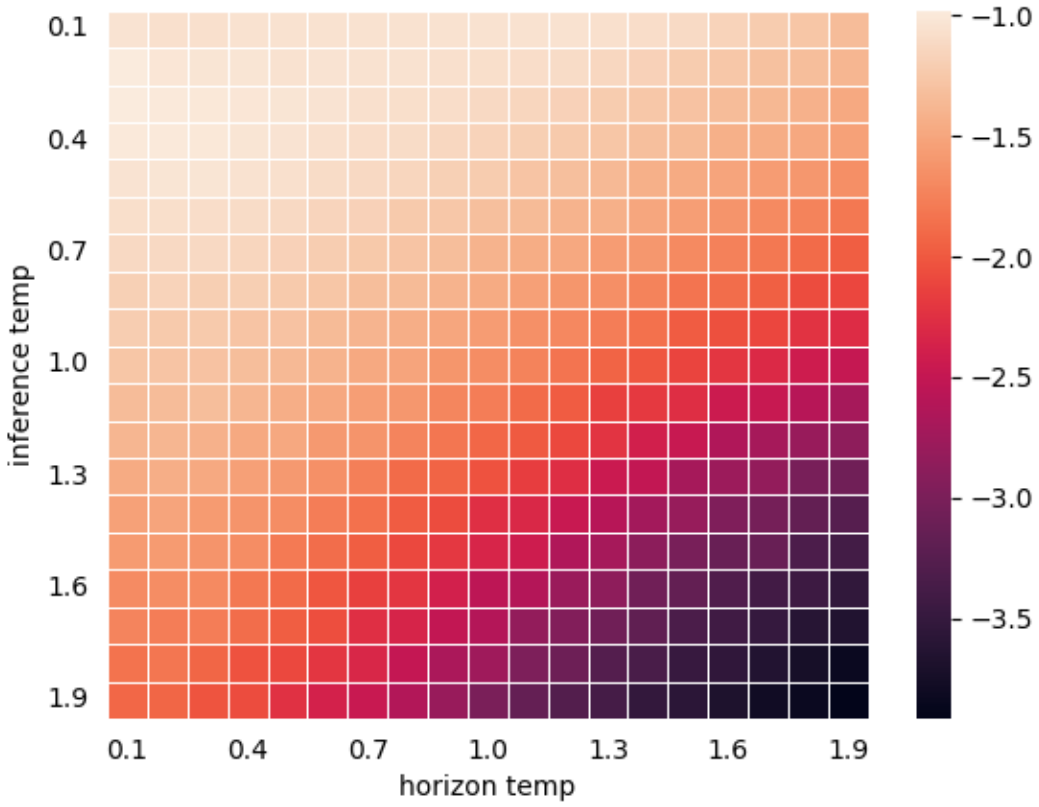
long horizon temperature scaling



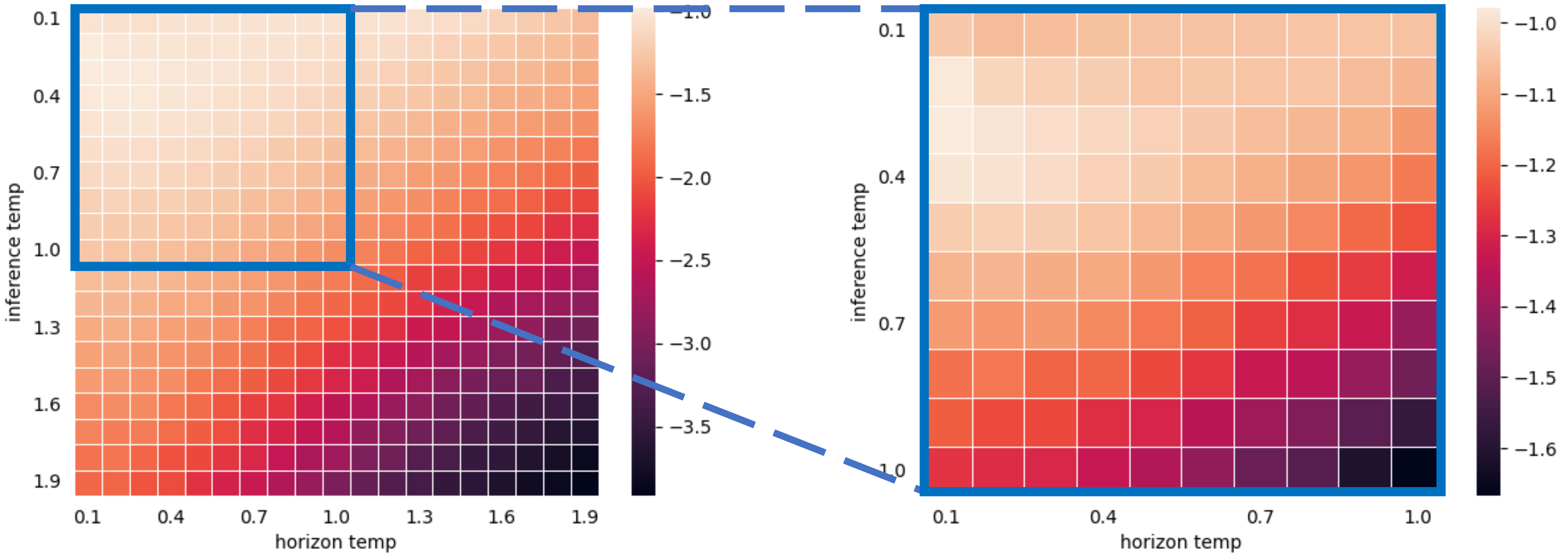
Autoregressive Character Models



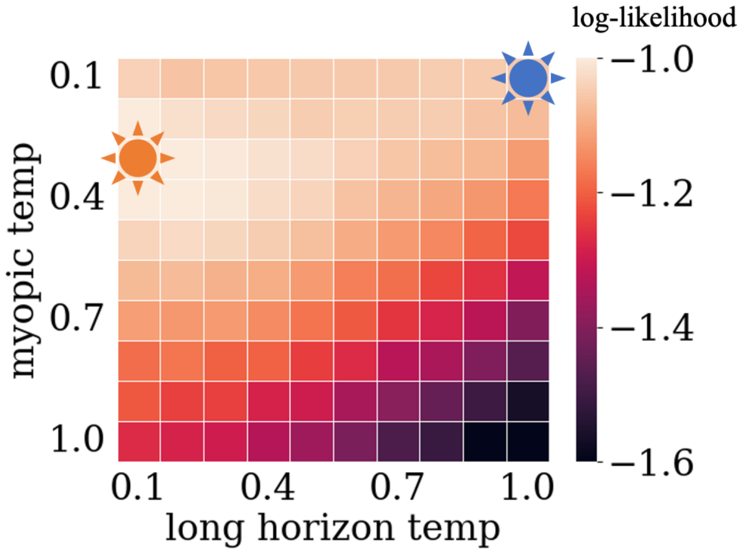
Autoregressive Character Models



Autoregressive Character Models



Autoregressive Character Models



Temperatures

LHTS : 0.1
myopic: 0.3

LHTS : 1.0
myopic: 0.1

Temperature extrapolation!

Better likelihood vs diversity tradeoff!

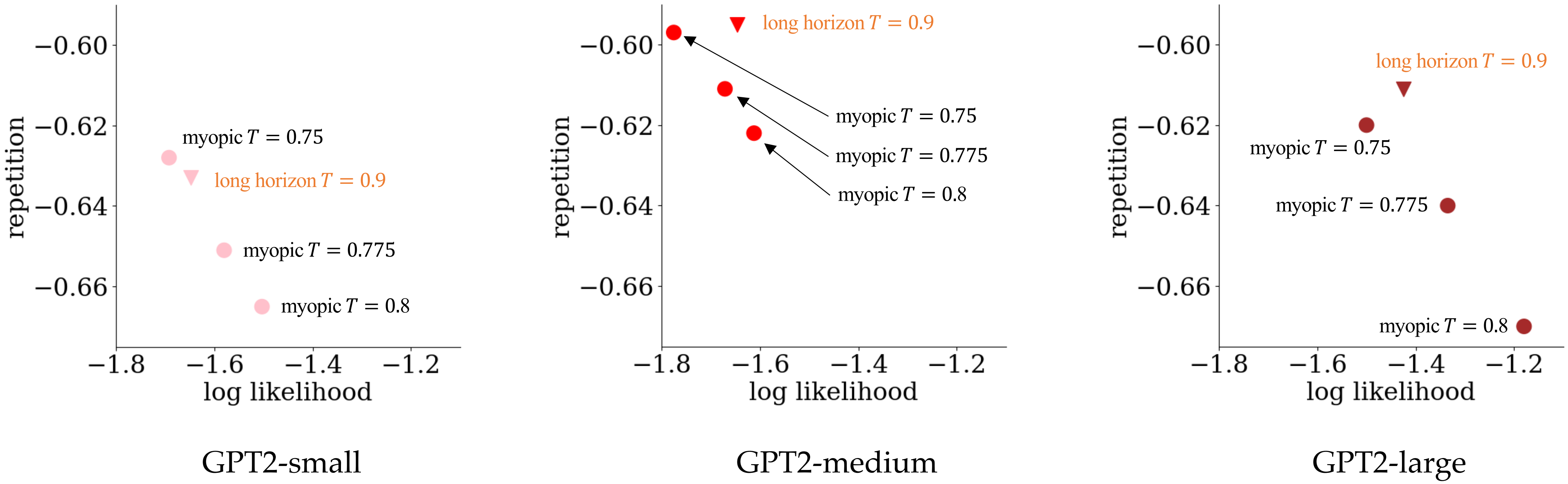
```
'the proces', ' internati', 'ne nine fi',
'n the latt', ' the const', ' and the m',
'is the fir', 'e three fi', ' of the ma',
```

Likelihood: -0.97
Diversity: Higher

```
' the const', ' the state', ' the state',
' the commu', ' the state', ' the commo',
' the same ', ' the state', ' the south',
```

Likelihood: -1.05
Diversity: Lower

Autoregressive Language Models



Autoregressive Language Models

Analogy Multiple Choice

Question: Please choose the word pair that is most analogous to “Athens Greece”.

Choices: “Moscow Japan”, “Rome Italy”, “Moscow Pakistan”, “Moscow Australia”

Answer:

Question: Please choose the word pair that is most analogous to "boy girl".

Choices: "grandfather grandmother", "grandfather bride", "son grandma", "grandfather sisters”

Answer:

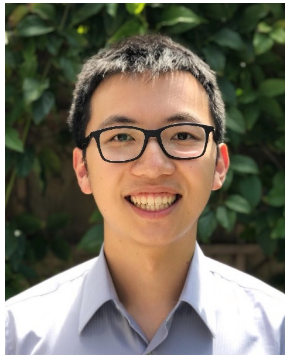
Autoregressive Language Models

Analogy Multiple Choice

model	gpt2 small			gpt2 medium			gpt2 large		
myopic T	1.0	0.5	0.0	1.0	0.5	0.0	1.0	0.5	0.0
LHTS $T = 0.9$	0.177	0.224	0.230	0.225	0.270	**0.275**	0.249	0.310	**0.317**
pretrained	0.189	0.267	**0.275**	0.200	0.262	0.264	0.203	0.279	0.290
partition (Quark)	0.137	0.221	0.233	0.197	0.264	0.270	0.213	0.279	0.285

10% improvement

Long Horizon Temperature Scaling



Andy Shih

Dorsa Sadigh

Stefano Ermon