Contrastive Decoding

Open-ended Text Generation as Optimization

Xiang Lisa Li, Ari Holtzman, Daniel Fried, Percy Liang,
Jason Eisner, Tatsunori Hashimoto, Luke Zettlemoyer, Mike Lewis

Decoding from Language Models

Pretrained LM

e.g., GPT-3

Prompt: **Researchers found British unicorns spoke perfect English. Researchers found**

Sampling from the LM:

```
that global warming has endangered
many species. Global warming is an
established scientific fact that
has been extensively studied and
confirmed by researchers…
```

Stochasticity may lead to unlucky sampling choices :(

😥 **Fluent but suffers from topic drifts (not coherent)**

Decoding from Language Models

Pretrained LM

e.g., GPT-3

Prompt: **Researchers found British unicorns spoke perfect English. Researchers found**

Searching for the most likely string:

```
British unicorns spoke perfect
English. British unicorns spoke
perfect English. British
unicorns spoke perfect English…
```

The modes of the language model distribution are highly degenerate :(

😥 **Repetitive and uninsteresting**

Decoding from Language Models

Pretrained LM

e.g., GPT-3

Sampling from the LM:

😥 **Fluent but suffers from topic drifts (not coherent)**

Searching for the most likely string:

😥 **Repetitive and uninsteresting**

How to generate coherent text, without sacrificing fluency/diversity?

A better decoding objective: **Contrastive Decoding**

Why contrast language models?

GPT-3: who played will on as the world turns?

Factual Error: John McCook

Repetition: who played will on as the world turns?

Topic Drift: Spolier---- Back when Shirly Fields was on ths show there

Large LMs suffer from these failures.

Smaller LMs are more prone to these failures!

Can we cancel out the bad behaviors?

Down weight the common failures + Emphasize the remaining good behaviors!

Contrastive Objective

$$\max_{x_{\text{cont}}} \log p_{\text{EXP}}(x_{\text{cont}} \mid x_{\text{pre}}) - \log p_{\text{AMA}}(x_{\text{cont}} \mid x_{\text{pre}})$$

	$\log p_{\text{AMA}}$	$\log p_{\text{EXP}}$	$\mathcal{L}_{\text{CD}}$		
Repetition: British unicorns…	-9.4	-10.2	-0.8	❌	**(Both assigned high probs)**
Topic Drift: Global warming…	-29.6	-27.4	2.2	❌	**(Both assigned low probs)**
Good Cont: their language …	-20.5	-11.5	9.0	✅	**(Prefered by the expert LM)**

😀 The undesired behaviors cancel out via contrastive objective.

Contrastive Objective

Maximize $\log p_{\mathrm{EXP}}(\mathsf{x}_{\mathsf{cont}} \mid \mathsf{x}_{\mathsf{pre}}) - \log p_{\mathrm{AMA}}(\mathsf{x}_{\mathsf{cont}} \mid \mathsf{x}_{\mathsf{pre}})$
$\mathsf{x}_{\mathsf{cont}}$

😀 The undesired behavior (e.g., repetition) cancels out via contrastive objective.

😥 But the amateur LM is not always wrong:

False positive

An implausible token gets a high CD score because the amateur assigns very low prob

False negative

A correct token gets a low CD score because the amateur also assigns high prob

Plausibility Constraints

$$\mathcal{V}_{\text{head}}(x_{<i}) = \{x_i \in \mathcal{V} : p_{\text{EXP}}(x_i \mid x_{<i}) \geq \alpha \max_{w} p_{\text{EXP}}(w|x_{<i})\}$$

Truncates the tail of the LM distribution

False positive

Example: schools 🥭

	$\log p_{\text{EXP}}$	$\log p_{\text{AMA}}$	$\mathcal{L}_{\text{CD}}$
p(🥭 | schools)	log(3e-9)	log(8e-14)	10.6

$\alpha = 0.1$

$\max_{w} p_{\text{EXP}}(w|x_{<i})$ = 0.3

Zero out tokens with prob < 0.03

🥭 is not in the plausibility set

Plausibility Constraints

$$\mathcal{V}_{\text{head}}(x_{<i}) = \{x_i \in \mathcal{V} : p_{\text{EXP}}(x_i \mid x_{<i}) \geq \alpha \max_w p_{\text{EXP}}(w|x_{<i})\}$$

Adaptive: the truncation depends on the confidence of the LM

False negative

Example: uni #corn

	$\log p_{\text{EXP}}$	$\log p_{\text{AMA}}$	$\mathcal{L}_{\text{CD}}$
p(#corn | uni)	log(0.99)	log(0.99)	6e-4

$\alpha = 0.1$

$\max_w p_{\text{EXP}}(w|x_{<i})$ = 0.99

Zero out tokens with prob < 0.099

#corn is the **only** token in the plausibility set

Full Method

$$\max_{\mathrm{x}_{\text{cont}}} \boxed{\mathcal{L}_{\text{CD}}(\mathrm{x}_{\text{cont}}, \mathrm{x}_{\text{pre}})}$$

Contrastive objective

$$\text{subject to} \quad \boxed{x_i \in \mathcal{V}_{\text{head}}(x_{<i}), \forall x_i \in \mathrm{x}_{\text{cont}}}$$

Plausibility constraints

Factor to token level

$$\text{CD-score}(x_i; x_{<i}) = \begin{cases} \log \frac{p_{\text{EXP}}(x_i \mid x_{<i})}{p_{\text{AMA}}(x_i \mid x_{<i})}, & \text{if } x_i \in \mathcal{V}_{\text{head}}(x_{<i}), \\ -\inf, & \text{otherwise.} \end{cases}$$

Beam Search

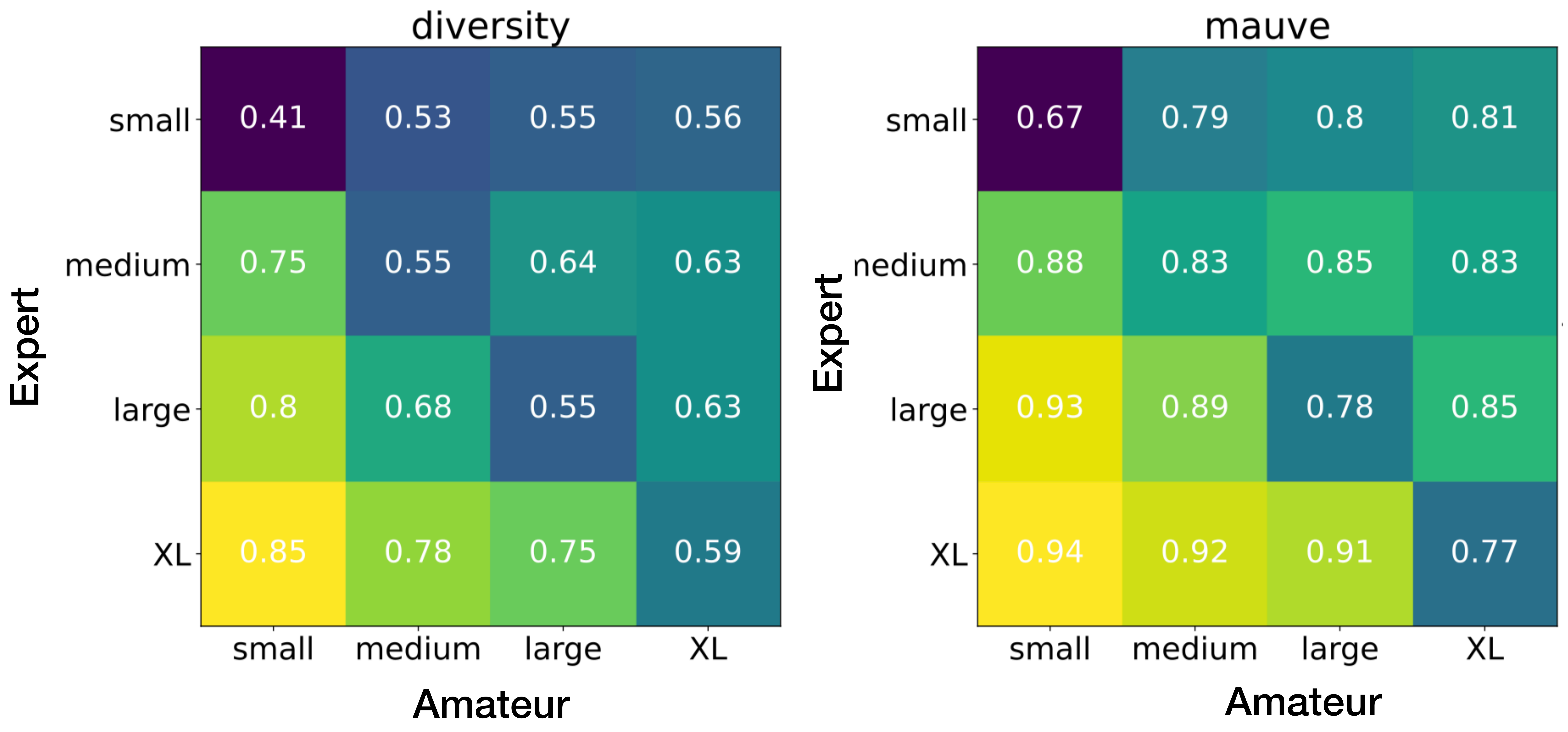
or

Sample

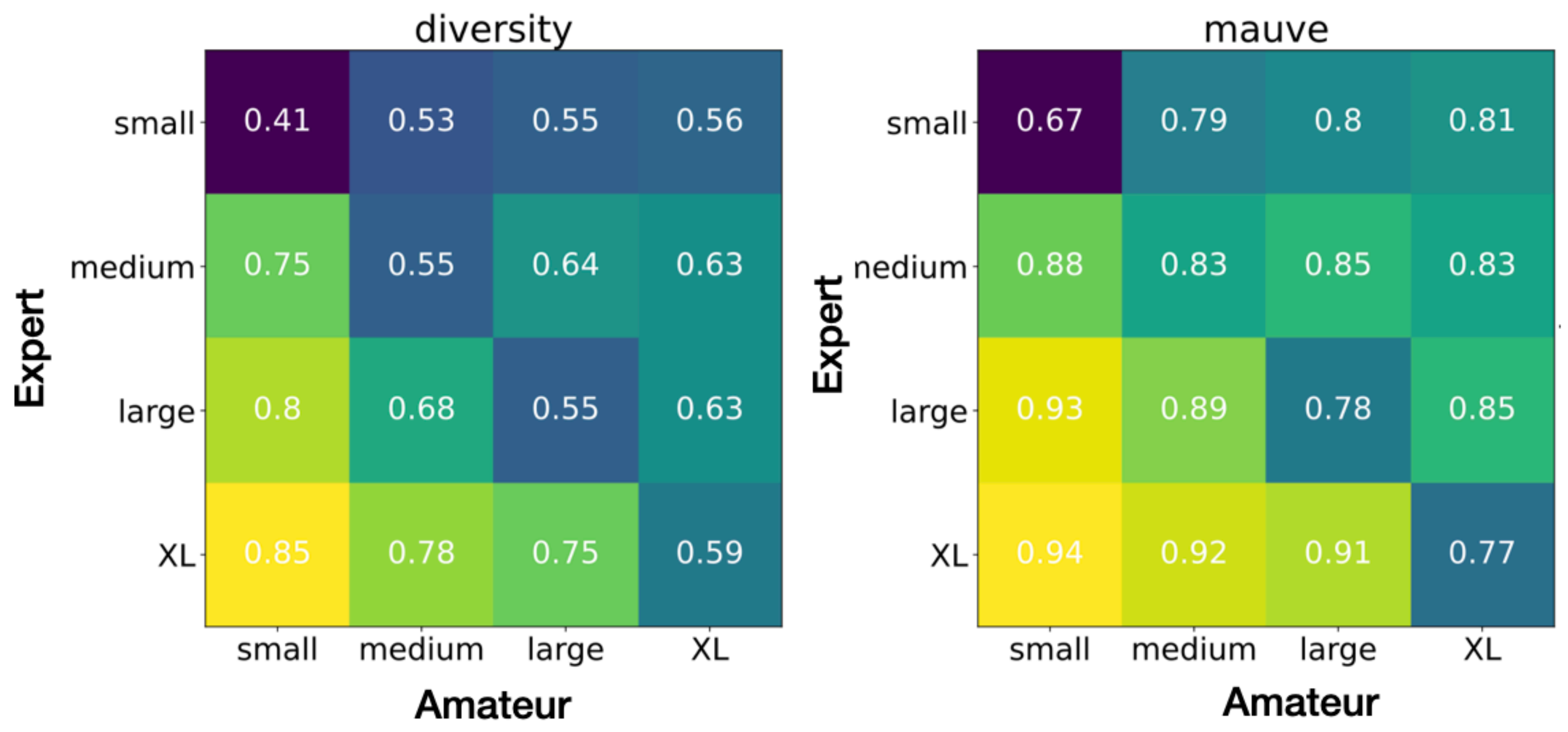
Design Choices: Scale

Maximize $\log p_{\text{EXP}}(\mathsf{x}_{\text{cont}} \mid \mathsf{x}_{\text{pre}}) - \log p_{\text{AMA}}(\mathsf{x}_{\text{cont}} \mid \mathsf{x}_{\text{pre}})$
x_{cont}

How does the choices of amateur LM and expert LM matter?



Design Choices: Scale



Intuition:
We want the amateur LM to emphasize the failure modes of the expert, so the amateur LM shouldn't be too strong.

Automatic Eval

		wikinews			wikitext			story		
	name	DIV	MAUVE	COH	DIV	MAUVE	COH	DIV	MAUVE	COH
OPT-13B	max prob	0.08	0.3	0.65	0.03	0.08	0.63	0.02	0.05	0.51
	k=50	0.91	0.92	0.64	0.72	0.77	0.64	0.91	0.9	0.51
	p=0.95	0.92	0.92	0.62	**0.92**	0.89	0.55	0.93	0.91	0.48
	typical=0.95	**0.94**	0.9	0.59	0.89	0.86	0.58	**0.95**	0.91	0.46
	CS(Su et al., 2022)	0.92	0.87	0.59	0.87	0.77	0.52	0.81	0.78	0.47
	CD	**0.94**	**0.94**	**0.69**	0.91	**0.91**	**0.69**	0.89	**0.94**	**0.62**
GPT2-XL	max prob	0.04	0.14	0.65	0.02	0.05	0.62	0.01	0.03	0.49
	k=50	0.92	0.88	0.64	0.87	0.79	0.61	0.91	0.87	0.51
	p=0.95	0.94	0.9	0.6	0.92	0.87	0.57	0.94	0.91	0.46
	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	0.43
	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	0.48
	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	**0.64**

Takeaways: CD outperforms prior sampling-based approaches in MAUVE and coherence.

Human Eval

	CD	Baseline	coherence: CD is better	coherence: same	coherence: Baseline is better	fluency: CD is better	fluency: same	fluency: Baseline is better
wikitext	CD (GPT-2 XL)	nucleus (GPT-2 XL)	**0.714***	0.083	0.202	**0.548**	0.083	0.369
	CD (GPT-2 XL)	typical (GPT-2 XL)	**0.887***	0.046	0.067	**0.703***	0.082	0.215
	CD (OPT-13B)	nucleus (OPT-13B)	**0.556**	0.202	0.242	**0.419**	0.197	0.384
	CD (OPT-13B)	typical (OPT-13B)	**0.773***	0.106	0.121	**0.687***	0.152	0.162
wikinews	CD (GPT-2 XL)	nucleus (GPT-2 XL)	**0.708***	0.042	0.25	**0.583***	0.12	0.297
	CD (GPT-2 XL)	typical (GPT-2 XL)	**0.771***	0.151	0.078	**0.755***	0.151	0.094
	CD (OPT-13B)	nucleus (OPT-13B)	**0.585***	0.221	0.195	**0.518**	0.123	0.359
	CD (OPT-13B)	typical (OPT-13B)	**0.693***	0.099	0.208	**0.49**	0.297	0.214
story	CD (GPT-2 XL)	nucleus (GPT-2 XL)	**0.636***	0.045	0.318	0.404	0.106	**0.49**
	CD (GPT-2 XL)	typical (GPT-2 XL)	**0.506**	0.256	0.238	**0.387**	0.363	0.25
	CD (OPT-13B)	nucleus (OPT-13B)	**0.616***	0.101	0.283	**0.449**	0.293	0.258
	CD (OPT-13B)	typical (OPT-13B)	**0.626***	0.202	0.172	**0.52**	0.212	0.268

Takeaways: CD outperforms prior sampling-based approaches according to human eval.

Qualitative Examples

Prompt	A version of Sonic the Hedgehog was developed by Ancient and released in 1991 for Sega's 8 - bit consoles, the Master System and Game
CD (1.5B)	A version of Sonic the Hedgehog was developed by Ancient and released in 1991 for Sega's 8 - bit consoles, the Master System and Game Gear. It featured a new level editor called the Sonic Editor and was released as an expansion pack to Sonic CD in 1992, which was itself a port from the Genesis version. The editor allowed the creation and playtesting of custom levels for Sonic CD, and the expansion was also bundled with Sonic CD. It is also available for the PC as Sonic Editor. The game is considered a spiritual successor to the Genesis game Sonic 3 & Knuckles and the Genesis port Sonic 3D Blast, as it features the same gameplay elements, including the Spin Dash, and also shares many levels and music with those games...
Nucleus (1.5B)	A version of Sonic the Hedgehog was developed by Ancient and released in 1991 for Sega's 8 - bit consoles, the Master System and Game Gear. It was set in the garden city of Sky Sanctuary and centered around Sonic's new helper, the pre-historic pillar-crusher-man Waddle Dee. Sonic the Hedgehog all the way was released for Game Gear in 1992. The author of the story speaks only for himself, but here it is. TO: Sonic the Hedgehog FROM: Anti-Socials SUBJECT: Argyle DeBrosse Contents show] What's that? You want me to tell you what I did, huh? Well, some things never change. Yeah. Basically, you. Now listen: I saw you on the NWA charts for ages. I heard about how great you were, and that you were a great singer. And then... later, after hearing it five times in a row, that's when I realized that I couldn't win this particular battle. You and your music have the edge...

Thanks

Code is available:
https://github.com/XiangLi1999/ContrastiveDecoding