UNIVERSITY OF CAMBRIDGE

Tencent AI Lab

香 港 大 學
THE UNIVERSITY OF HONG KONG

A Contrastive Framework for Neural Text Generation

Yixuan Su[1], Tian Lan[2], Yan Wang[2], Dani Yogatama[3], Lingpeng Kong[4], Nigel Collier[1]

1: University of Cambridge 2: Tencent AI Lab

3: DeepMind 4: The University of Hongkong

Background: Applications

Open-ended Text Generation:

- Story Generation
- Poetry Generation
- News Article Creation
- Dialogue Systems

Background: Degeneration

Model output is unnatural and contains undesirable repetitions, i.e. the degeneration.

Prefix: Robert Boulter is an English film , television and theatre actor . He had a guest @-@ starring role on the television series The Bill in 2000 .

Greedy Search	**Beam Search (b=5)**
He has appeared in several films, including the television series The Bill, The Bill, The Bill, The Bill, The Bill, The Bill, The Bill, The Bill, The Bill, The Bill, The Bill, The Bill, The Bill, The Bill, The Bill, The Bill, The Bill, The Bill, The Bill...	Boulter is best known for his role in the television series The Bill, which he co @-@ starred in. He has also appeared in several television series, including the television series The Bill, and the film The Bill, which he co @-@ starred in. Boulter has also appeared in several...

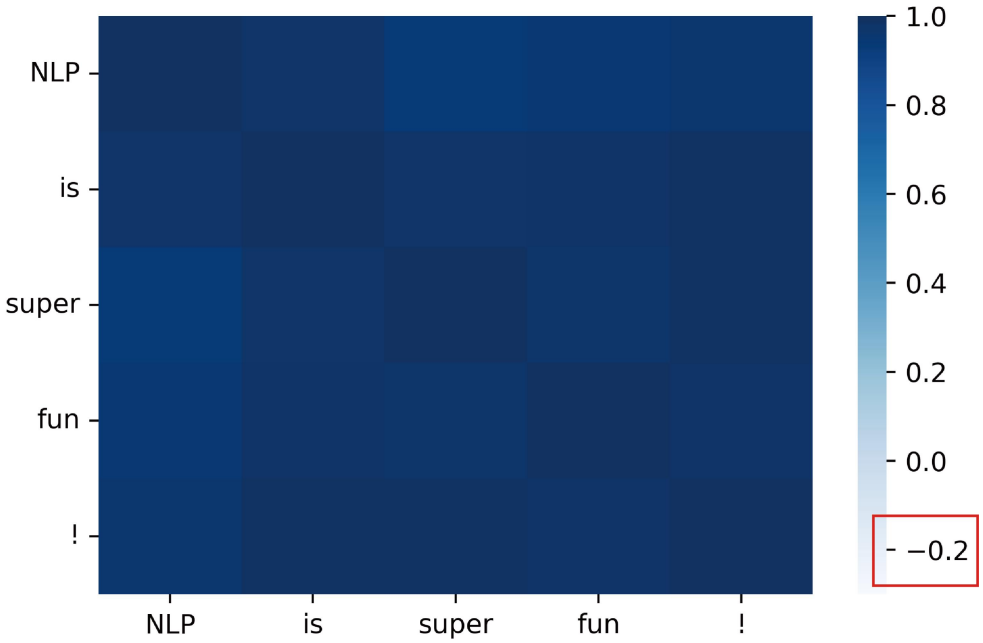
Generated Examples from GPT-2

Current Solutions

- Sampling (e.g. top-k, nucleus sampling)
  - Results are unrobust and irreproducible
  - Intrinsic stochasticity causes semantic inconsistency
- Unlikelihood Training
  - Harm the accuracy of the language model

Our Problem Analysis

Degeneration stems from the *anisotropic* distribution of model representations.



$$M_{ij} = \text{cosine}(h_i, h_j)$$

Our Solutions

- Calibrating the model representation space to make it follow an isotropic distribution. (SimCTG)
- Introducing contrastive search decoding algorithm that consists of two aspects:
  - Selecting output from the most probable candidates
  - Preserving a sparse similarity matrix of the generated text to avoid degeneration

SimCTG

Use contrastive training to calibrate the model's representation space

$$\mathcal{L}_{\text{MLE}} = -\frac{1}{|\boldsymbol{x}|}\sum_{i=1}^{|\boldsymbol{x}|}\log p_\theta(x_i|\boldsymbol{x}_{<i}). \tag{1}$$

3.1 Contrastive Training

$$L_{\text{CL}} = \frac{1}{|\boldsymbol{x}|\times(|\boldsymbol{x}|-1)}\sum_{i=1}^{|\boldsymbol{x}|}\sum_{j=1,j\neq i}^{|\boldsymbol{x}|}\max\{0,\rho - s(h_{x_i},h_{x_i}) + s(h_{x_i},h_{x_j})\} \tag{2}$$

$$s(h_{x_i},h_{x_j}) = \frac{h_{x_i}^\top h_{x_j}}{\|h_{x_i}\|\cdot\|h_{x_j}\|} \tag{3}$$

$$L_{\text{SimCTG}} = L_{\text{MLE}} + L_{\text{CL}} \tag{4}$$

Similarity Matrix Comparison

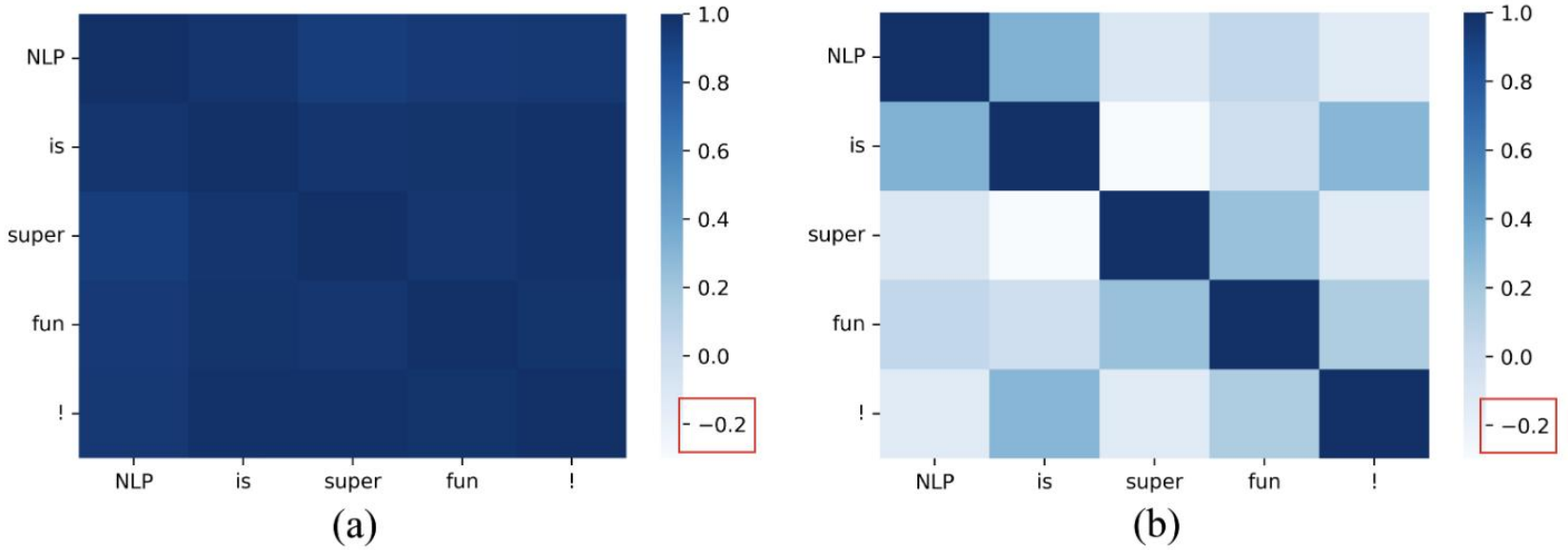


Figure 1: Token cosine similarity matrix of (a) GPT-2 and (b) SimCTG. (best viewed in color)

Contrastive Search

Contrastive search jointly considers (1) model predictions and (2) possibility of degeneration (i.e. try to maintain the sparseness of the token similarity matrix of the generated text).

3.2 Contrastive Search

$$x_t = \arg\max_{v \in V^{(k)}} \Big\{ (1-\alpha) \times \underbrace{p_\theta(v|\boldsymbol{x}_{<t})}_{\text{model confidence}} + \alpha \times \underbrace{(-\max\{s(h_v, h_{x_j}) : 1 \leq j \leq t-1\})}_{\text{degeneration penalty}} \Big\} \quad (5)$$

Experiment: Automatic Evaluation on Document Generation

Model	Language Modelling Quality				Generation Quality							
	ppl↓	acc↑	rep↓	wrep↓	Method	rep-2↓	rep-3↓	rep-4↓	diversity↑	MAUVE↑	coherence↑	gen-ppl
MLE	24.32	39.63	52.82	29.97	greedy	69.21	65.18	62.05	0.04	0.03	0.587	7.32
					beam	71.94	68.97	66.62	0.03	0.03	0.585	6.42
					nucleus	4.45	0.81	0.43	0.94	0.90	0.577	49.71
					contrastive	44.20	37.07	32.44	0.24	0.18	0.599	9.90
Unlike.	28.57	38.41	**51.23**	**28.57**	greedy	24.12	13.35	8.04	0.61	0.69	0.568	37.82
					beam	11.83	5.11	2.86	0.81	0.75	0.524	34.73
					nucleus	4.01	0.80	0.42	**0.95**	0.87	0.563	72.03
					contrastive	7.48	3.23	1.40	0.88	0.83	0.574	43.61
SimCTG	**23.82**	**40.91**	51.66	28.65	greedy	67.36	63.33	60.17	0.05	0.05	0.596	7.16
					beam	70.32	67.17	64.64	0.04	0.06	0.591	6.36
					nucleus	4.05	0.79	0.37	0.94	0.92	0.584	47.19
					contrastive	**3.93**	**0.78**	**0.31**	**0.95**	**0.94**	**0.610**	**18.26**
Human	-	-	36.19	-	-	3.92	0.88	0.28	0.95	1.00	0.644	24.01

Table 1: Evaluation results on Wikitext-103 test set. "Unlike." denotes the model trained with unlikelihood objective. ↑ means higher is better and ↓ means lower is better.

- SimCTG + contrastive search outperforms strong baselines
- SimCTG + contrastive search achieves closer performance with human reference

Experiment: Human Evaluation on Document Generation

Model	Decoding Method	Coherence	Fluency	Informativeness
Agreement	-	0.51	0.64	0.70
MLE	nucleus	2.92	3.32	3.91
	contrastive	2.78	2.29	2.56
Unlikelihood	nucleus	2.62	3.30	3.95
	contrastive	2.55	1.71	1.91
SimCTG	nucleus	2.96	3.34	3.96
	contrastive	3.25★	3.57★	3.96
SimCTG-large	nucleus	3.01	3.37	3.98
	contrastive	**3.33**★	**3.66**★	**3.98**
Human	-	3.70	3.71	4.21

Table 2: Human Evaluation Results. ★ results significantly outperforms the results of nucleus sampling with different models (Sign Test with p-value < 0.05).

- SimCTG-large model + contrastive search achieves highest performance on three aspects of human evaluation

Experiment: Human Evaluation on Dialogue Generation

Model	Method	LCCC			DailyDialog		
		Coherence	Fluency	Informativeness	Coherence	Fluency	Informativeness
Agreement	-	0.73	0.61	0.57	0.64	0.60	0.55
MLE	greedy	3.01	3.27	1.97	3.28	3.51	2.92
	beam	2.60	2.90	1.55	3.16	3.43	2.78
	nucleus	2.78	3.55	2.64	2.67	3.58	3.42
	contrastive	3.28★	3.84★	3.06★	3.27	3.41	2.82
SimCTG	greedy	3.04	3.32	2.01	3.31	3.50	2.94
	beam	2.57	2.93	1.59	3.19	3.45	2.79
	nucleus	2.84	3.58	2.72	2.75	3.59	3.39
	contrastive	**3.32**★	**3.96**★	**3.13**★	**3.73**★	**3.85**★	**3.46**
Human	-	3.42	3.76	3.20	4.11	3.98	3.74

Table 3: Human evaluation results. ★ results significantly outperforms the results of greedy search, beam search, and nucleus sampling with different models. (Sign Test with p-value < 0.05).

- SimCTG + contrastive search works best across the board
- Contrastive search works well on **vanilla** Chinese language model

Heatmap Comparison

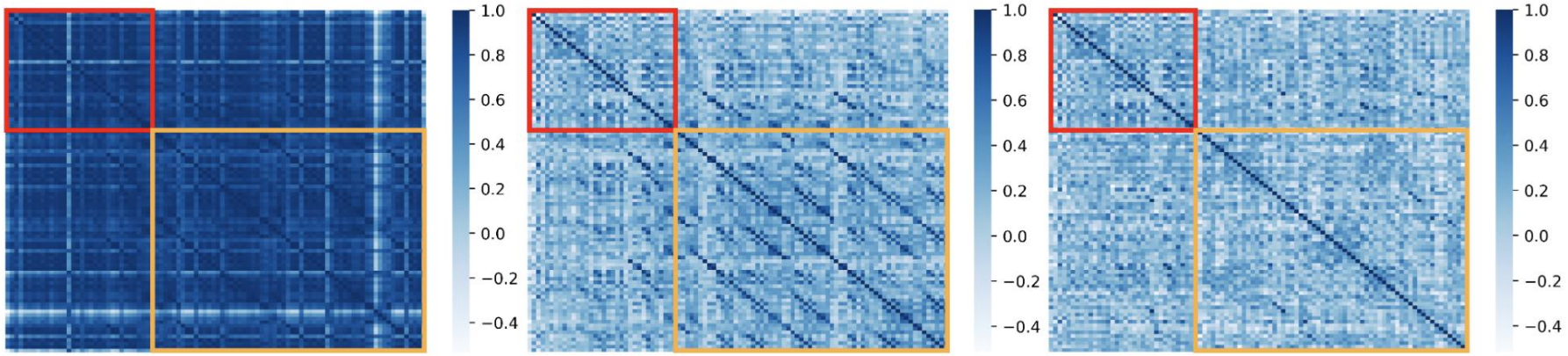
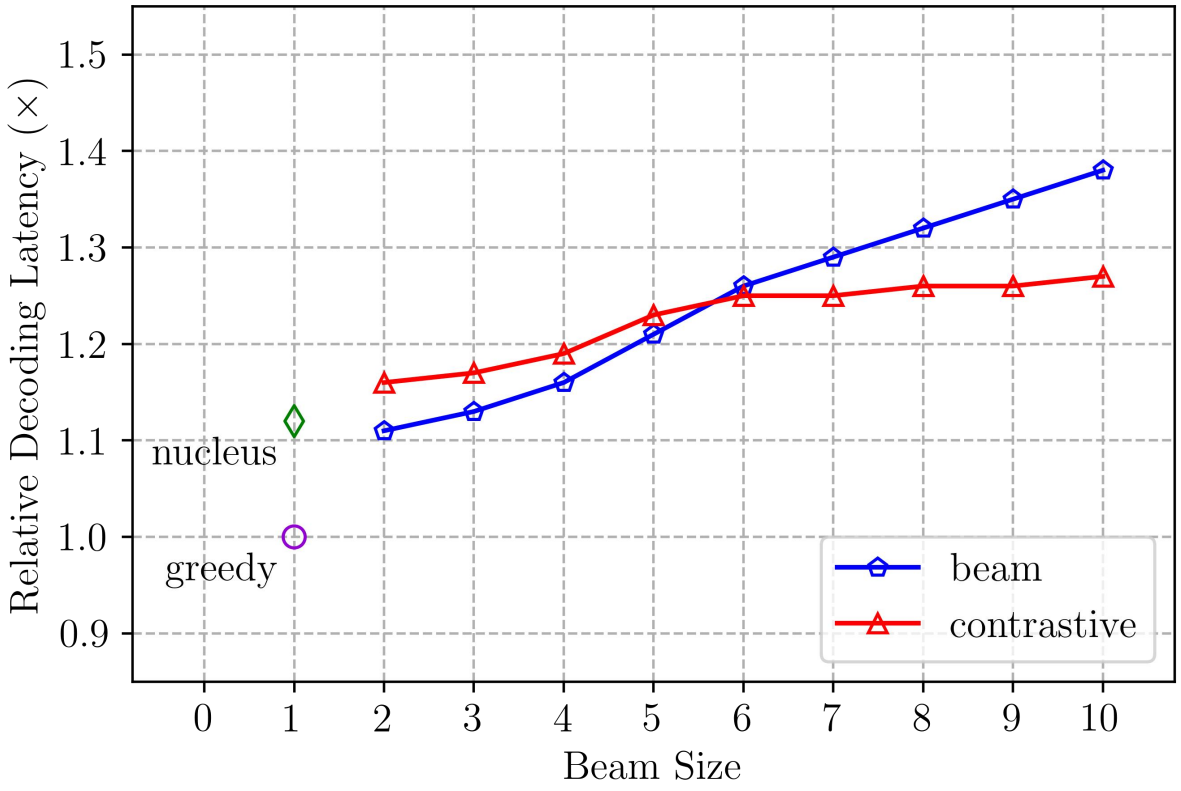
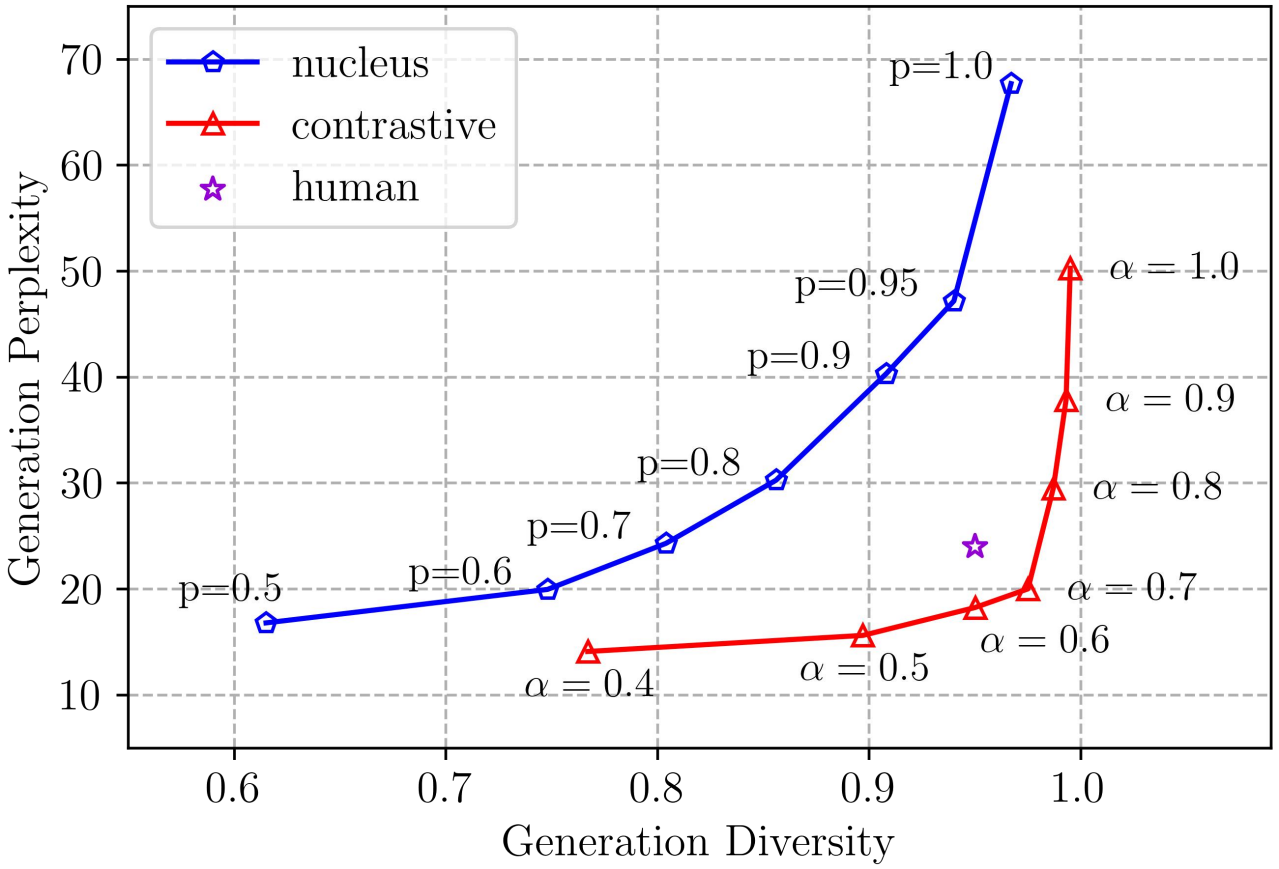


Figure 4: (a) MLE + Beam Search; (b) SimCTG + Beam Search; (c) SimCTG + Contrastive Search. Red for prefix, yellow for continuation.

Relative Inference Latency



Generation Diversity vs Perplexity



Conclusion

To tackle the degeneration problem of neural text generation, we propose two techniques:

- SimCTG calibrates the model representation space to make it follow an isotropic distribution.
- Contrastive search generate diverse and coherent text by intergrating the model confidence and the degeneration penalty.

More Information and Follow-up Works

- Contrastive Search Is What You Need For Neural Text Generation

  https://arxiv.org/abs/2210.14140

- An Empirical Study On Contrastive Search And Contrastive Decoding For Open-ended Text Generation

  https://arxiv.org/abs/2211.10797

- Momentum Decoding: Open-ended Text Generation As Graph Exploration

Questions and Discussions

Thank You!

Extension: Diverse Contrastive Search

prefix	Butt criticized Donald's controls in certain situations in the game, as well as the difficulty of some levels and puzzles. Buchanan also criticized the controls, calling
	Contrastive Search
result	them " unimpressive " and a " nightmare " of an experience to play with players unfamiliar with Tetris. On the other hand, his opinion was shared by other reviewers, and some were critical of the game's technical design for the Wii version of Tetris. In addition, Tintin's review included a quote from Roger Ebert, who said that Tetris was better than the original game due to its simplicity and ease of play. Ebert's comments were included in the game's DVD commentary, released on March 22, 2010. It is unclear if any of the video commentary was taken from ...
	Diverse Contrastive Search
[1]	the game " unplayable " and adding that, " It's a game in which you have to be careful not to fall into the trap of thinking too much about your opponents' strengths and weaknesses. " On the other hand, his review of the game, published in the New York Times, found little to complain about in comparison to some of the other games in the series. In particular, The Escapist's Brian O 'Toole wrote that the game was an " absolute joy ", and " one of the best-selling games of all time ". O 'Toole concluded by saying that although ...
[2]	it " a complete waste of time " and " unplayable ". On the other hand, his review of Baldur's Gate II was positive, with Buchanan commenting that, " Baldur's Gate II is an adventure game in its own right, full of fun and challenge that makes you want to go back to the first game in your life. " Buchanan felt that there were too many elements in the game for players to enjoy without some level-playing to be enjoyable at the same time. He concluded by saying that Baldur's Gate II's controls were well-balanced, and that players ...
[3]	the choice of " a simple jump button to perform a 'jump-and-a-bop' or more complex 'jump-and-a-bop' " an error and a waste of time. On the other hand, Tintin was critical of the game's design, writing that there was " too much going on " at the beginning of the game, and " not enough time " in the final cutscene for the player to make it through the game at all. He felt that the gameplay was lacking in some areas, such as the ...

Table 5: Generated results of SimCTG with diverse contrastive search.

Extension: Isotropic of Language Models

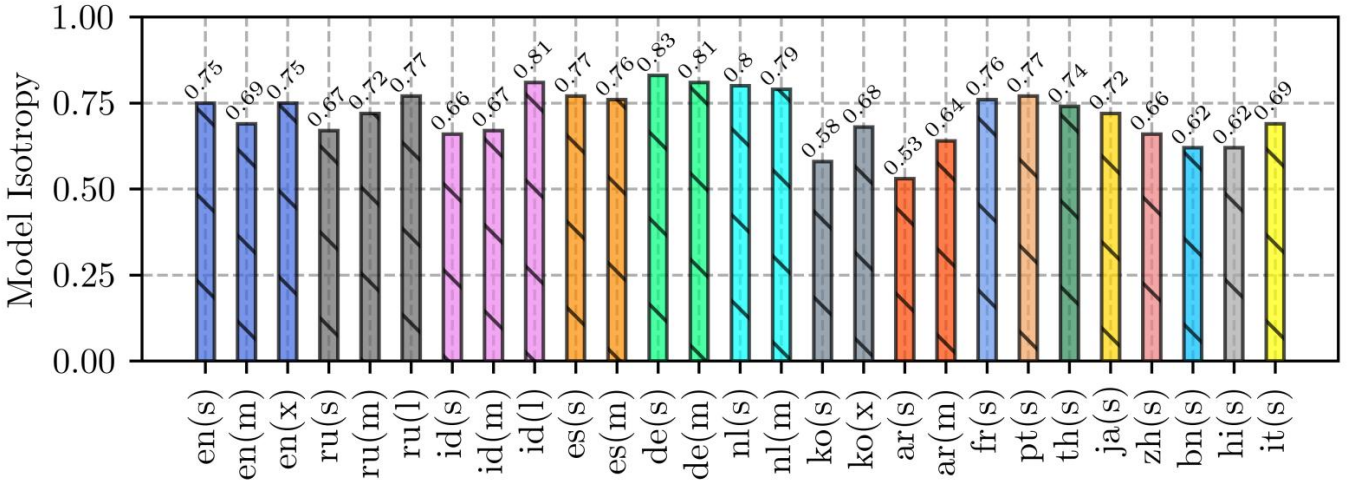


Figure 2: Isotropy results of multilingual LMs. Each x(y) denotes the language code (x) and the model size (y), where `s` is for small size model (i.e. ∼120M parameters), `m` is for medium size model (i.e. ∼350M parameters), `l` is for large size model (i.e. ∼780M parameters), and `x` is for xl size model (i.e. ∼1.5B parameters). For English (i.e. en) LMs, we plot the results of three OPT models. The detailed list of language codes and evaluated LMs can be found in Table 10 at Appendix C.

Most language models are isotropic

Extension: Isotropic of Language Models

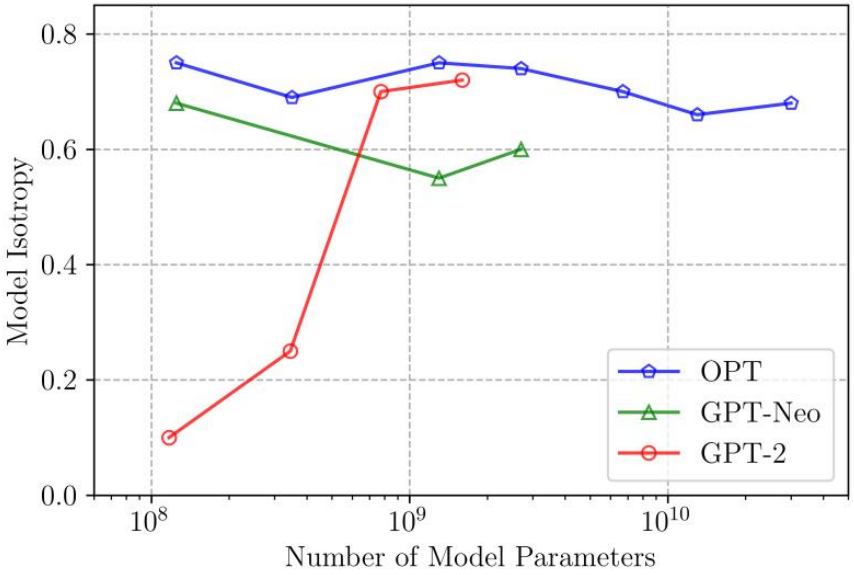


Figure 1: Isotropy results of English LMs.

Most language models are isotropic except for GPT2-Small and GPT2-Medium