Superpower of the Contrastive Decoding comes from its Imagination!

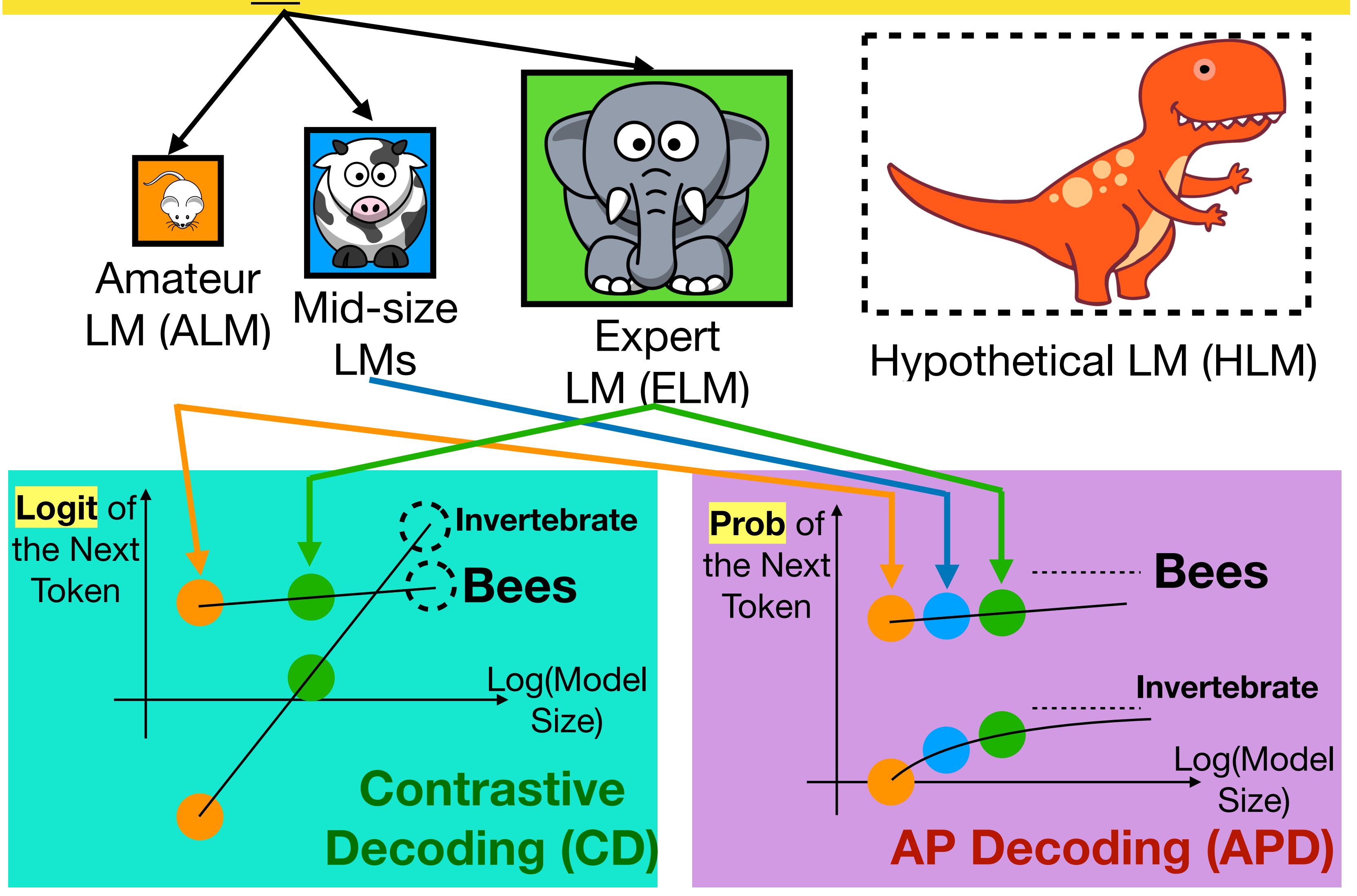
Explaining and Improving Contrastive Decoding by Extrapolating the Probabilities of a Huge and Hypothetical LM

Haw-Shiuan Chang, Nanyun Peng, Mohit Bansal, Anil Ramakrishna, Tagyoung Chung
Amazon AGI Foundations

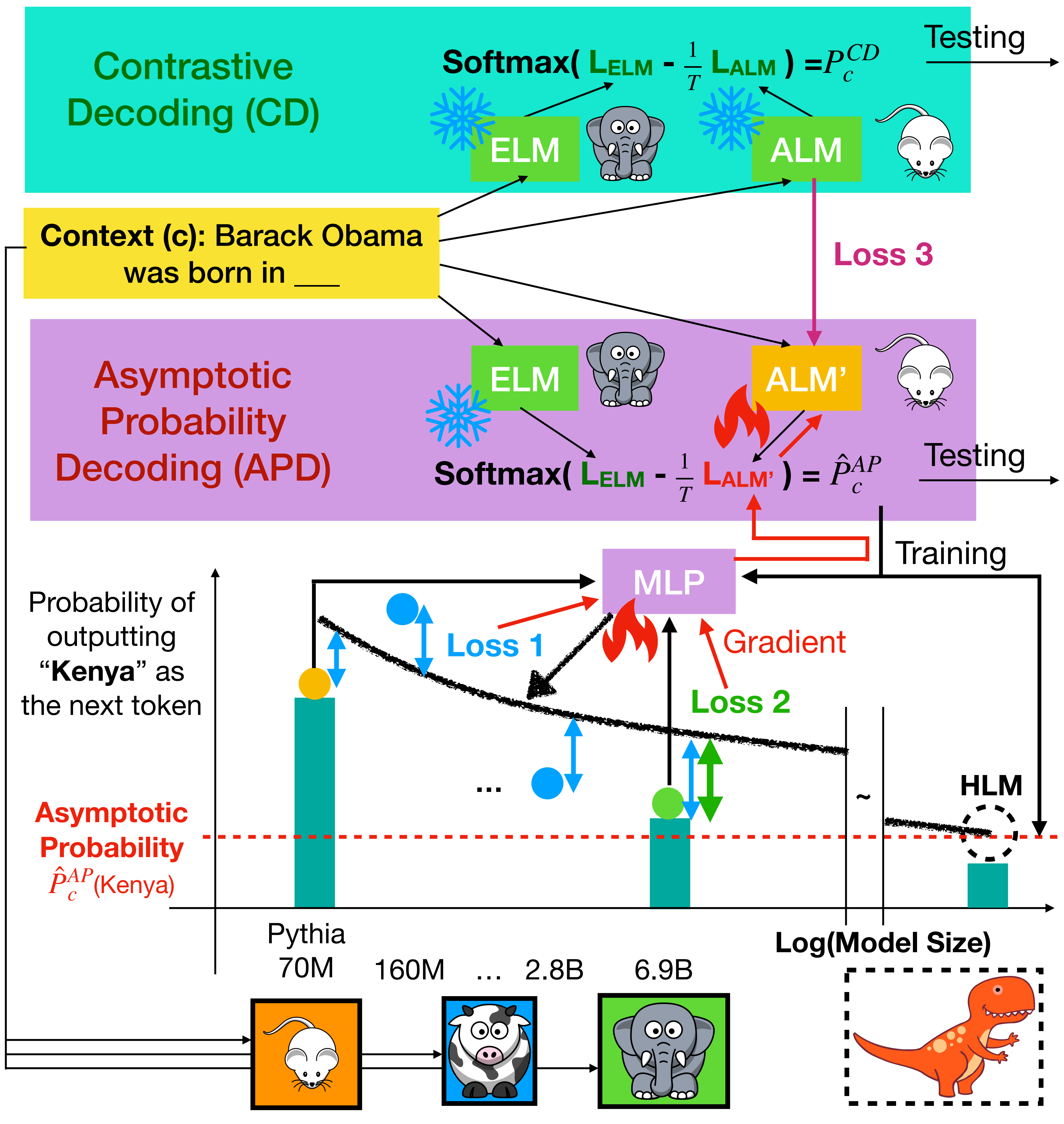
Introduction

Input Context

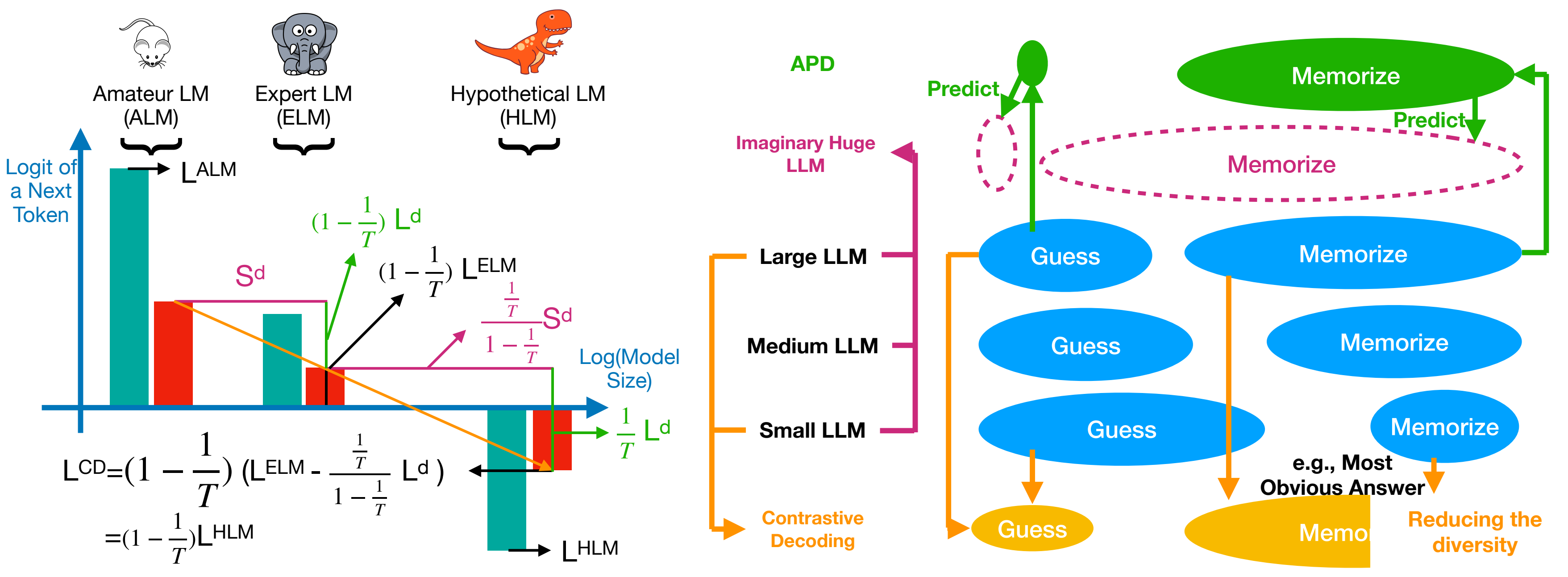
Question: What animals can fly without a backbone?
Fact 1: Invertebrates lack a backbone.
Fact 2: Bees are a kind of flying invertebrates.
Answer: __



APD Method

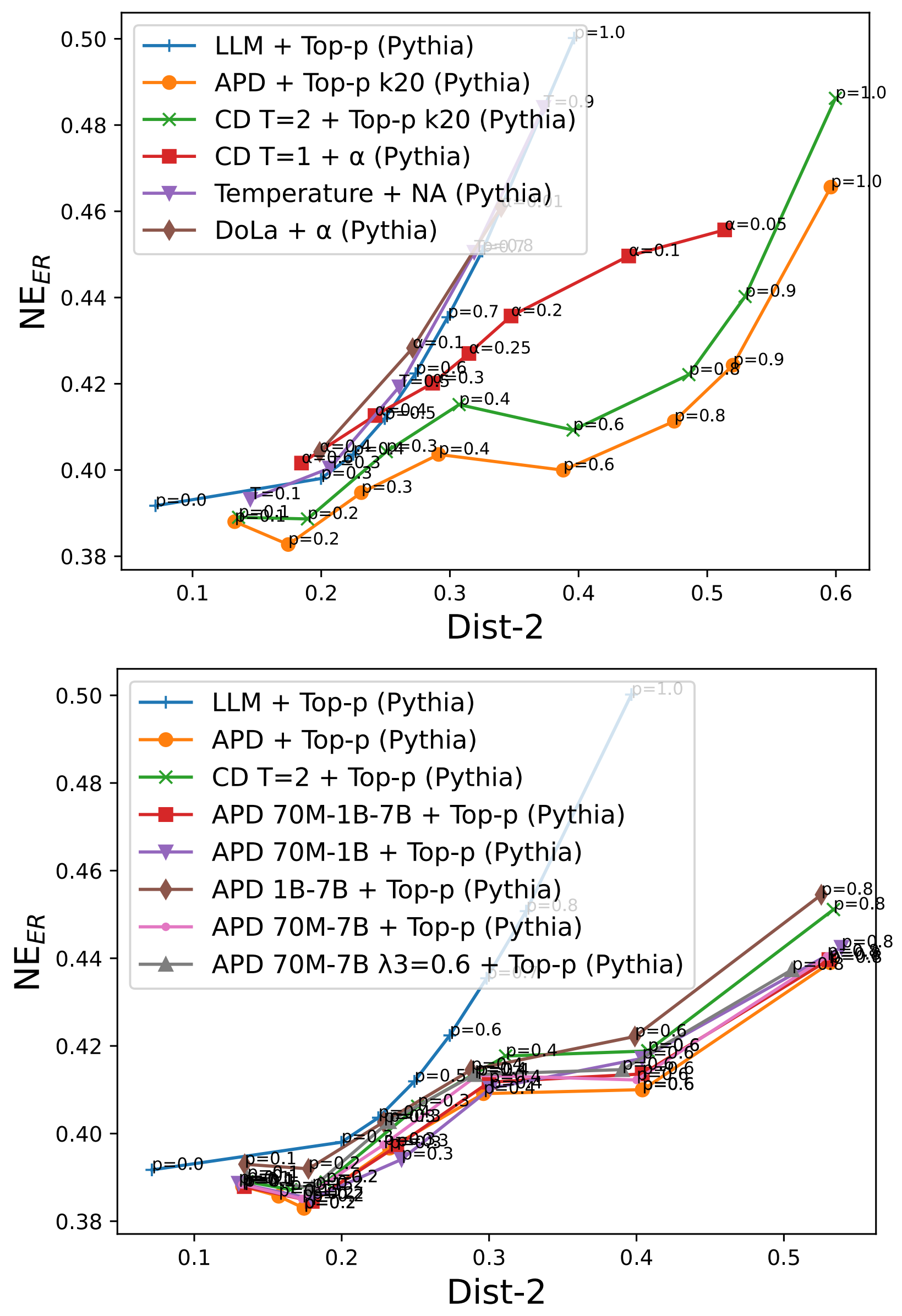

Contrastive Decoding (CD)
Softmax(L_{ELM} - $\frac{1}{T}$ L_{ALM}) = P_c^{CD}
Testing
ELM
ALM
Context (c): Barack Obama was born in ___
Loss 3
Asymptotic Probability Decoding (APD)
ELM
ALM'
Softmax(L_{ELM} - $\frac{1}{T}$ $L_{ALM'}$) = $\hat{P}_c^{AP}$
Testing
Training
MLP
Probability of outputting "Kenya" as the next token
Loss 1
Gradient
Loss 2
...
HLM
Asymptotic Probability
$\hat{P}_c^{AP}$(Kenya)
Pythia
70M
160M
...
2.8B
6.9B
Log(Model Size)

Why?

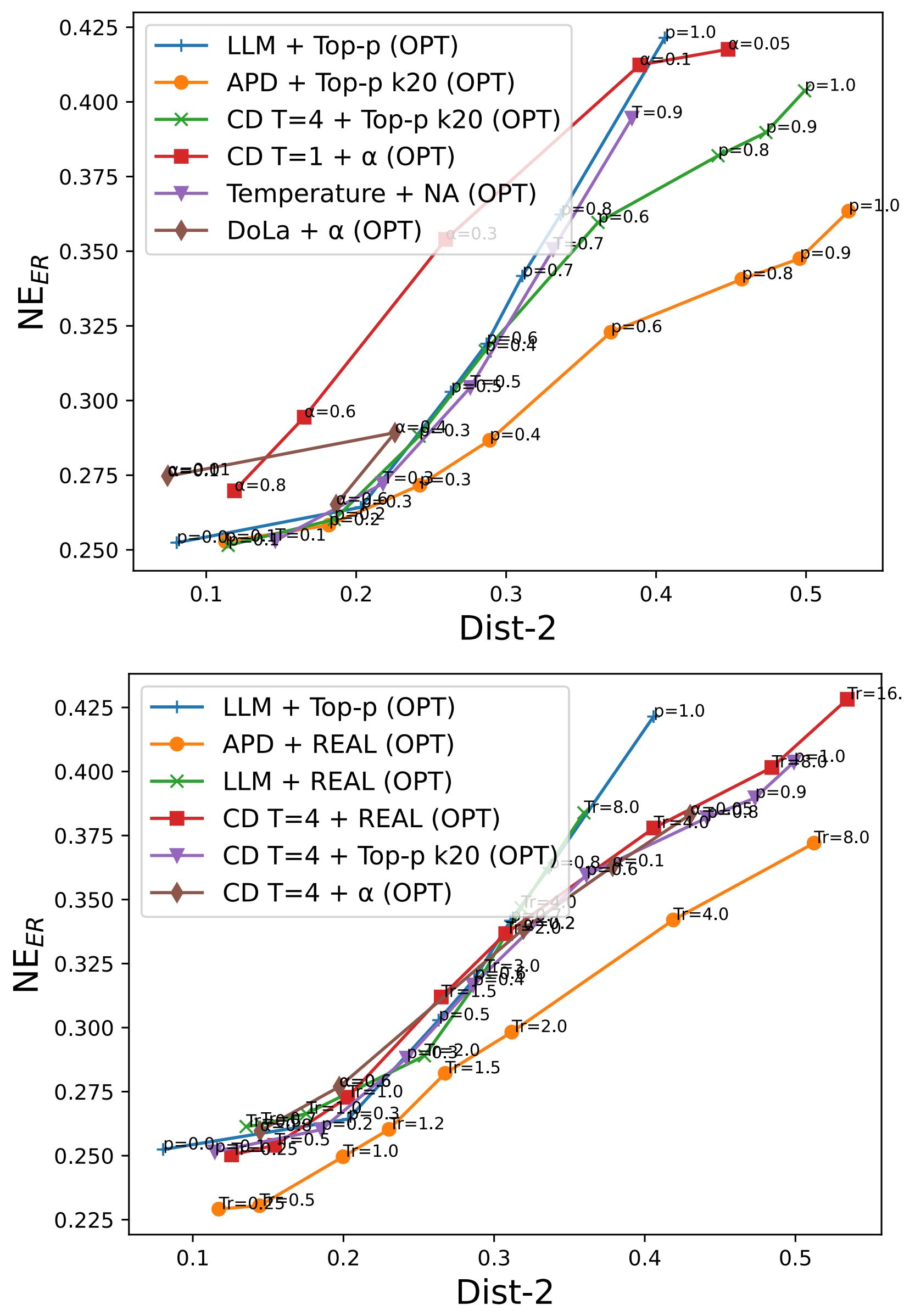


Results

FactualityPrompts (Lee et al., 2022)



Pythia 70M, 160M, 410M, 1B, 1.4B, 2.8B, and 6.9B



OPT 125m, 350m, 1.3B, 2.7B, and 6.7B

		LAMBADA	CQA		QASC				ARC		SocialIQA	
					Q+Fact		Q Only					
		ppl (↓)	ppl (↓)	acc	ppl (↓)	acc	ppl (↓)	acc	ppl (↓)	acc	ppl (↓)	acc
Pythia	LLM 6.9B	2.264	8.380	0.658	5.702	0.856	8.127	0.621	4.433	0.692	8.441	0.662
	CD	2.237	6.176	0.671	5.693	0.862	**7.741**	**0.633**	4.375	**0.699**	7.595	0.688
	APD	**2.132**†	**5.882**†	**0.685**	**5.020**†	**0.874**	7.766	0.632	**4.310**	0.698	**7.378**	**0.691**
	APD on the fly	2.281	8.245	0.660	5.725	0.866	8.106	0.620	4.464	0.694	8.299	0.665
	LLM 12B	2.188	8.140	0.660	4.783	0.845	7.612	0.630	4.058	0.719	7.898	0.691
	APD vs CD	138.52%	122.34%	650.00%	73.30%	NA	-4.86%	-12.50%	17.34%	-4.17%	39.92%	11.54%
	APD vs LLM 6.9B	173.68%	1039.11%	1250.00%	74.26%	NA	70.06%	125.00%	32.87%	20.83%	195.88%	100.00%

Conclusions

- It is possible to generally improve LLMs with a tiny LM by imagining/simulating the even larger LLM!
- The current cross-entropy next word prediction is not optimal. More research is required!
- Extrapolation might also improve CD in various other applications and beyond

Reference

- Nayeon Lee, Wei Ping, Peng Xu, Mostofa Patwary, Pascale Fung, Mohammad Shoeybi, and Bryan Catanzaro. 2022. Factuality enhanced language models for open-ended text generation. In NeurIPS 2022
- Haw-Shiuan Chang, Nanyun Peng, Mohit Bansal, Anil Ramakrishna, and Tagyoung Chung. 2024. REAL sampling: Boosting factuality and diversity of openended generation via asymptotic entropy. Preprint, arXiv:2406.07735.