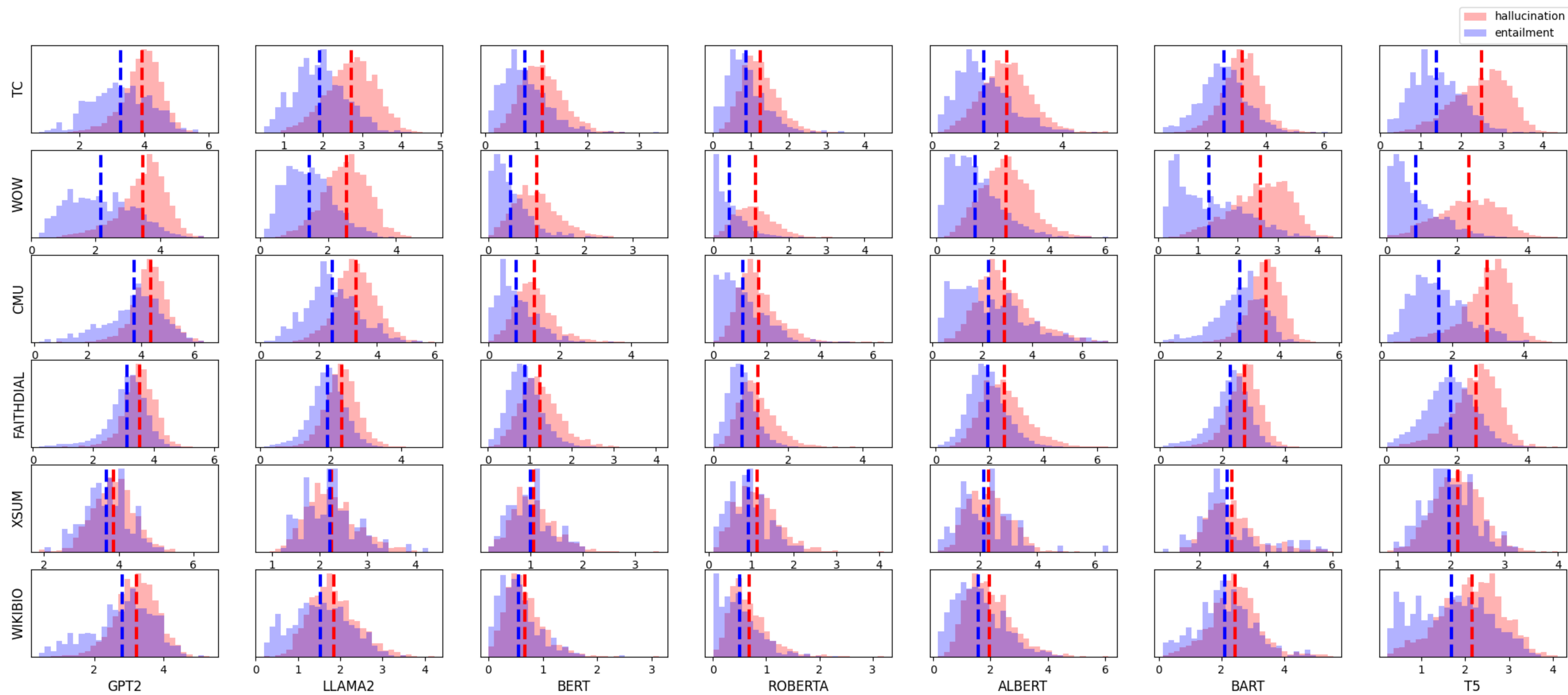
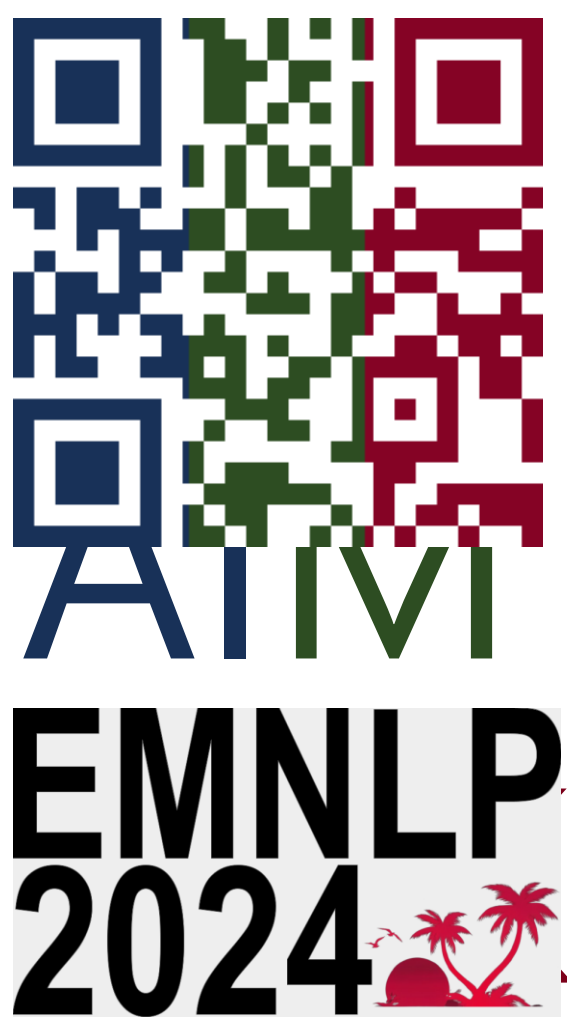




Pre-trained Language Models Return Distinguishable Probability Distributions to Unfaithfully Hallucinated Texts

Taehun Cha and Donghun Lee
Department of Mathematics, Korea University

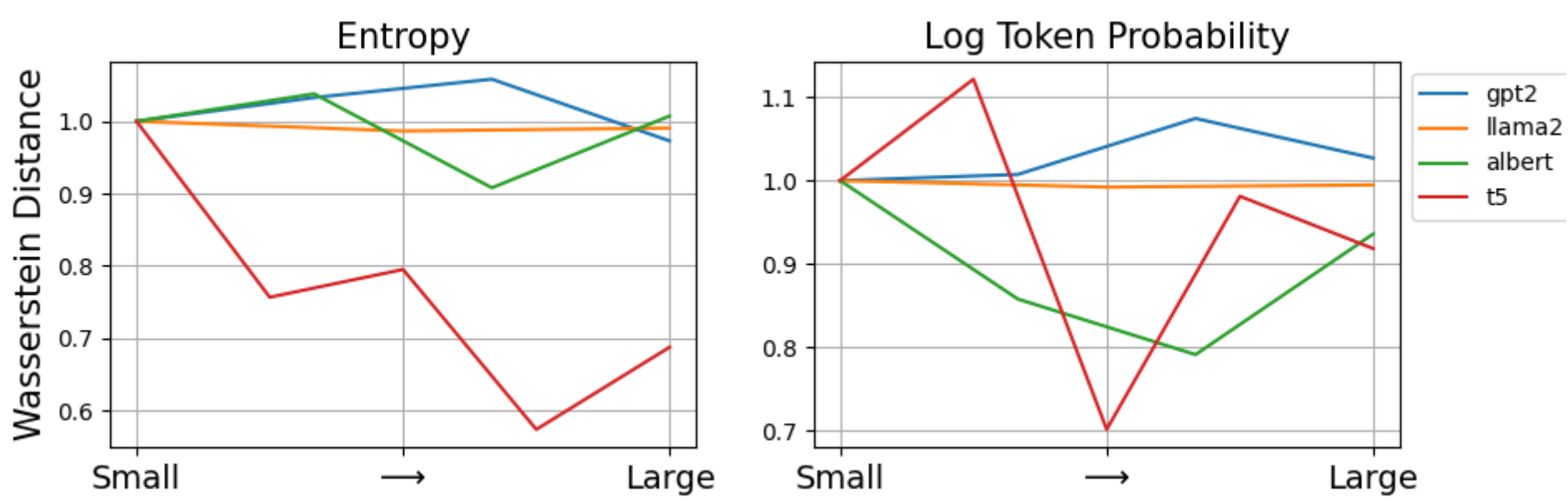


Main Discovery

- Compute two metrics (**log generation probability** and **entropy**) for each data point
- Analyse 24 PLMs (including BERT-like encoder) on 6 hallucination data sets
- Apply non-parametric Kolmogorov-Smirnov test
- **Observe 88-98% of cases return statistically significantly distinguishable distributions**

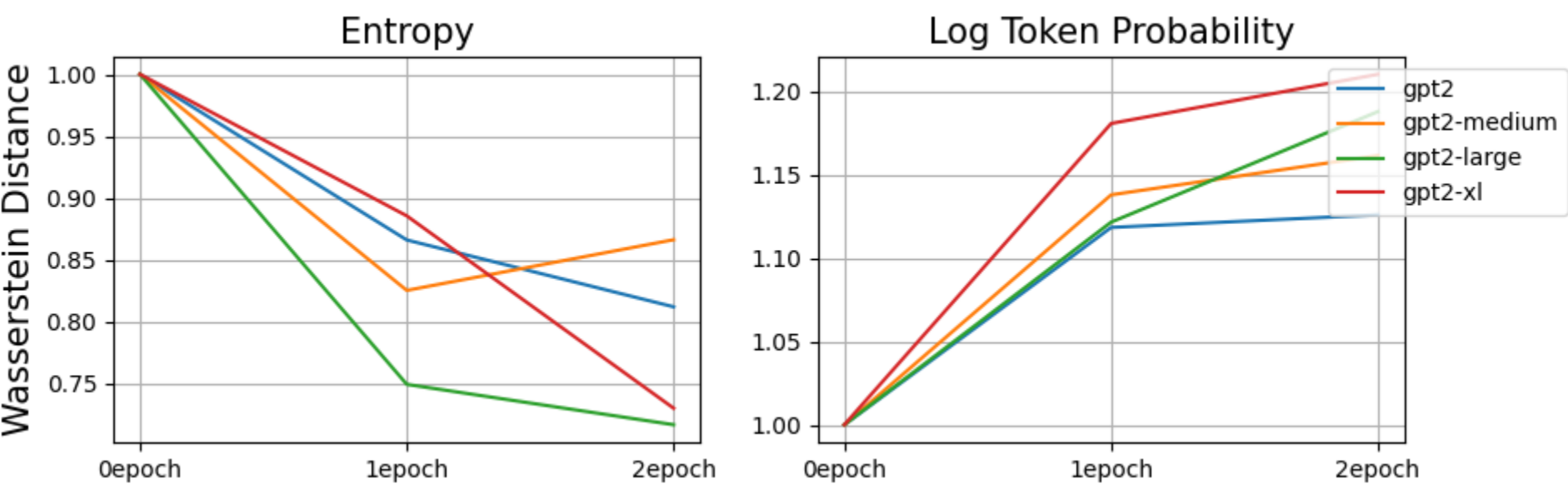
Size Effect

- NLPers tend to adopt the largest model first
- **But bigger size does not guarantee better distinguishability**



Training Effect

- Several hallucination-reduction methods utilize *trained* models
- **But training affect distinguishability in both (un)favorable ways**



Method: Weighted Training

- Train a model with the loss weighted by pre-computed metrics
- **It improves faithfulness while maintaining general quality**

Data	Method	Q2-F1	Q2-NLI	SummaC	FactKB	ROUGE	BERT Score	BART Score
WOW	Fine-tune	0.6521	0.6947	0.2941	0.5633	0.2862	0.3012	-2.7871
	LogProb	0.7689	0.7946	0.4287	0.7033	0.2960	0.2963	-2.7633
	Entropy	0.7742	0.8040	0.5503	0.7273	0.3105	0.3124	-2.7811
Faith Dial	Fine-tune	0.7830	0.8439	0.1761	0.6156	0.3066	0.3360	-2.7874
	LogProb	0.8454	0.8841	0.3652	0.7706	0.3135	0.3371	-2.7251
	Entropy	0.8403	0.8905	0.4092	0.7475	0.3179	0.3401	-2.7166
Medi QA	Fine-tune	0.7912	0.8333	0.5152	0.9987	0.2491	0.1712	-2.8650
	LogProb	0.8129	0.8579	0.5416	0.9927	0.2447	0.1748	-2.8680
	Entropy	0.7853	0.8371	0.4966	0.9984	0.2465	0.1701	-2.8530

Take Away

- PLMs generally return distinguishable distributions to unfaithfully hallucinated texts
- NLPers should check the size and training effect before adopting the models
- We derive a simple but effective training method which enhance the model faithfulness

