Entailment ⇒ and boxes 🎁 are crucial missing pieces in how we analyze LLM performance!!!

PROMPT2BOX: Improving LLM Weakness Discovery and Specificity Estimation by Uncovering Entailment Structure among Prompts

Neeladri Bhuiya[1,2], Shib Dasgupta[1,3], Andrew McCallum[1], Haw Shiuan Chang[1,4]

[1] UMASS AMHERST [2] A10 Networks [3] amazon [4] College of Information Science

Box embeddings can effectively model both similarity and specificity. Outperforms vectors on retrieval tasks that require understanding of both similarity and specificity. Better modeling of prompt structure also leads to better downstream performances in dataset comparison, hierarchical clustering, LLM weakness finding.

Lack of specificity information leads to incorrect attribution of weakness

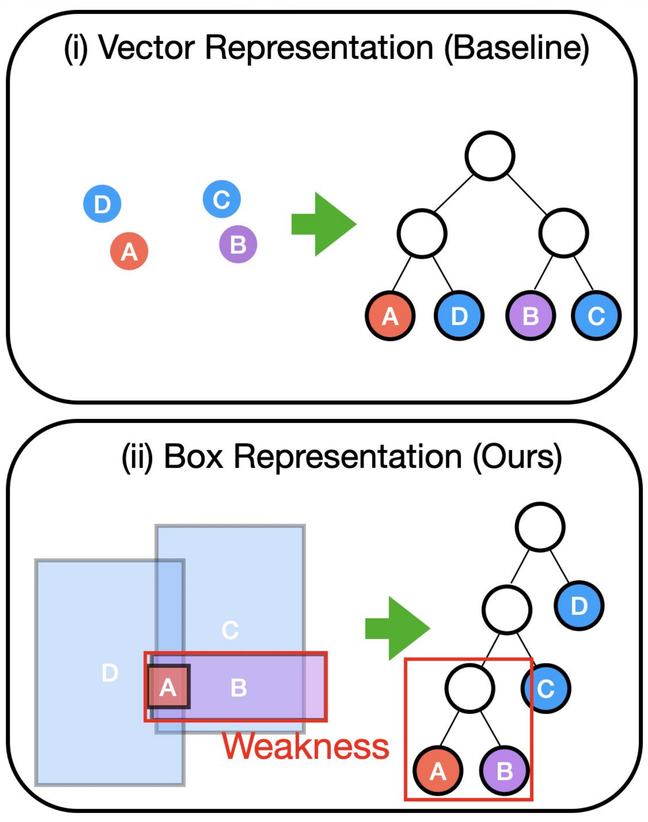
LLM Prompts

A: Please write a story with around 500 words, and the story should depict a romantic adventure of a robot and a person

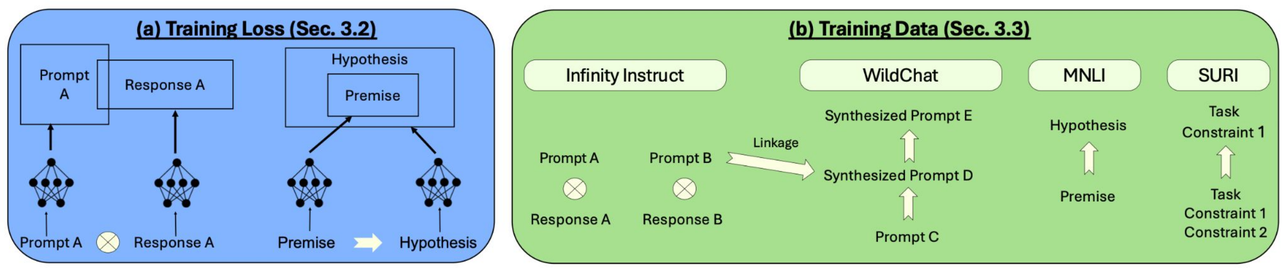
B: Please write a short story about an adventure of a robot

C: Please write a short adventure story

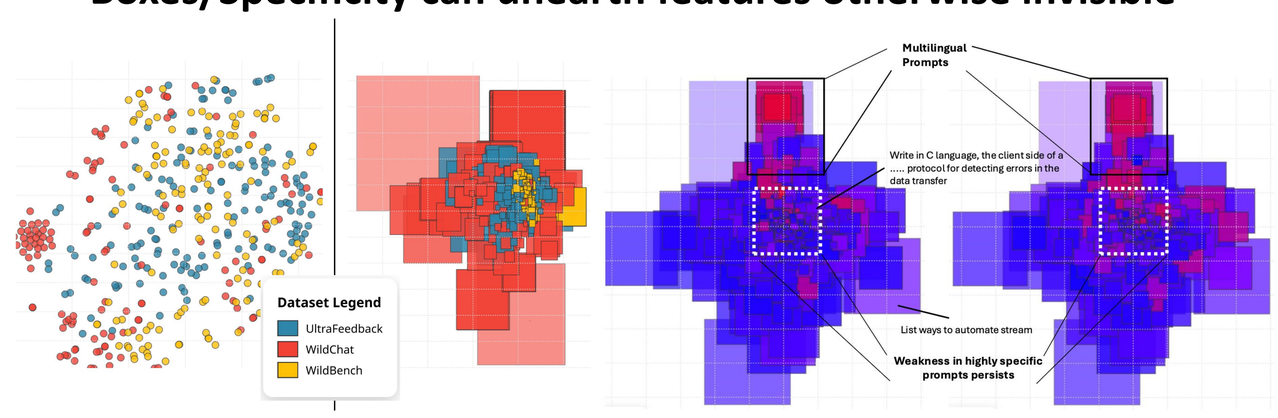
D: Please write a romantic story.



We augment box embedding training using (synthesized) entailment datasets



Boxes/Specificity can unearth features otherwise invisible



- We develop Box-SNE to visualise **n-d boxes in a 2D space.**

Boxes outperform vectors on entailment while being competitive in similarity

Model	FollowBench	UltraFeedback	
	Accuracy↑	Avg. RMSE↓	Improv. (%)↑
Random	—	1.8293	0.0
Hyperbolic	0.659	1.6260	11.11
CSDelta	0.012	1.5955	12.80
Vector	0.640	1.6059	12.21
Vector w/o entail	0.627	1.5605	14.71
Box w/o entail	0.687	1.5610	14.67
Box w/o links	**0.775**	**1.4777**	**19.22**
Box w/o synth	0.716	1.5280	16.47
Box	0.738	1.5280	16.47

Model/Setting	STS-B	SURI
Metric	Spearman	Accuracy
Vector	**0.835**	0.725
Vector w/o entail	0.704	0.539
Box	0.760	0.868
Box w/o entail	0.727	0.640
Box w/o links	0.661	0.924
Box w/o synth	0.756	0.889
Hyperbolic	0.820	**0.978**
CSDelta	0.757	0.950

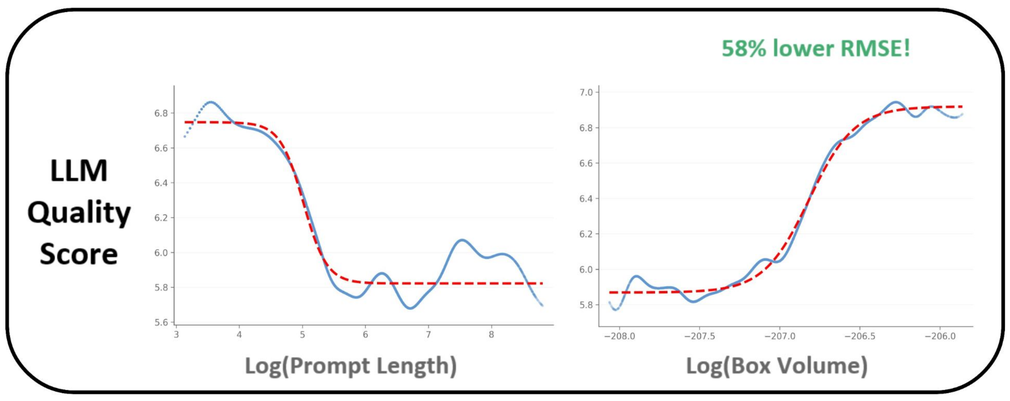
Boxes lead to better hierarchical clustering

Metric	Rand.	Vec.	Box w/o links	Box w/o synth	Box
Loc. Score Cons.	0.0%	9.32%	**13.24%**	10.34%	12.88%
Spec.–Depth Agr.	50.00%	52.71%	67.07%	66.03%	**68.34%**
All Size Weakness	–	0.0%	2.85%	5.94%	**7.70%**
AUC	–	0.0%	7.81%	10.95%	**13.47%**



GPT 3.5 Turbo | LLaMA-2-13B-Chat | Alpaca-7B

Box volumes are a better specificity estimators than length



- We use KDE to estimate the average score of the curve.

System	UltraFeedback		LMSYS	
	RMSE Impr.	Line Cov. Impr.	RMSE Impr.	Line Cov. Impr.
Length	0.0%	0%	0.0%	0%
Box w/o links	44.3%	83%	14.8%	10%
Box w/o synth	39.8%	113%	9.4%	**72%**
Box	**45.3%**	**113.5%**	**18.3%**	71%

Conclusions

- We train boxes to treat **entailment as specificity,** with smaller boxes being more specific.
- Boxes can be used to compare datasets and models
- Boxes outperform on **specificity aware retrieval task and also downstream tasks**.
- Hierarchical clusters generated by boxes have more ideal characteristics when compared to boxes
- Box volumes on average have a **45.3% lower RMSE for specificity estimation** when compared to the prompt length baseline.
- Box volume as specificity estimator also leads to sigmoid curves with l**onger linear portions,** indicating ability to **model subtle differences.**

Check out the full paper here ➡