DoLa: Decoding by Contrasting Layers Improves Factuality in Large Language Models

Yung-Sung Chuang [1], Yujia Xie [2], Hongyin Luo [1], Yoon Kim [1], James Glass [1], Pengcheng He [2]
[1] MIT CSAIL [2] Microsoft

Github: q-r.to/DoLa

Accepted in ICLR 2024

Observation & Motivation

Input: *Who was the first Nigerian to win the Nobel Prize, in which year?*
Output: *Wole Soyinka was the first Nigerian to win the Nobel Prize, in 1986.*

- X: generated tokens; Y: layer index; Item: JS-Div between *early logits* & *final logits*
- When predicting factual information, LLaMA tends to *change the predictions in the higher layers.* Otherwise, predictions usually have been *decided by early layers*
- Previous study also found "**knowledge neurons**" located in topmost layers [1]
- **Hypothesis:** Contrasting the layers before/after the radical change may amplify the knowledge in higher layers and make the model more factual [2]

Method

Basics:

- Early exiting from all layers
- Pick a layer as "premature" layer, final layer as "mature" layer
- Subtract "premature" logits from "mature" logits in log domain

How to pick "premature" layer?

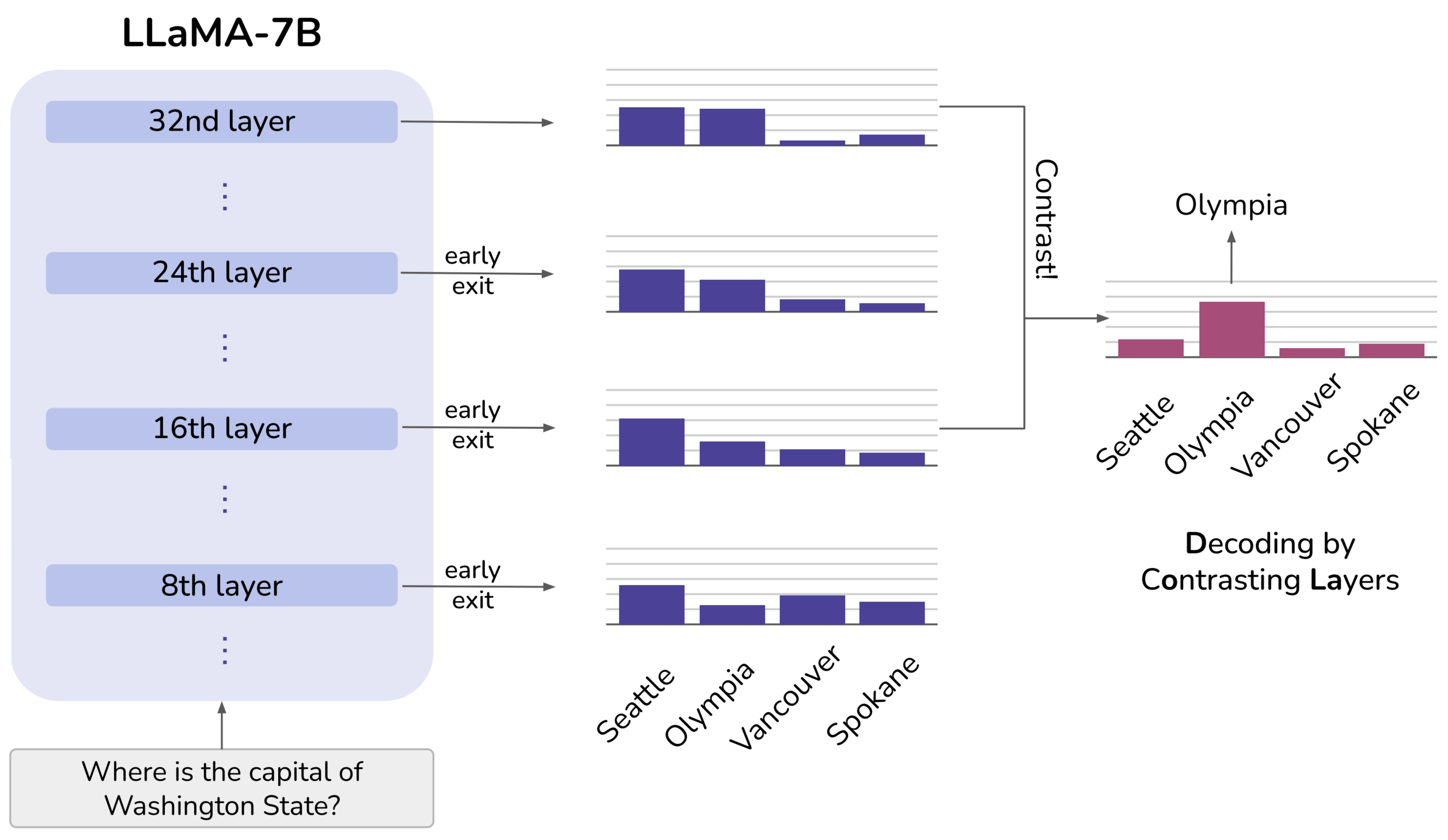
- Run brute force to try all layers
- Dynamic layer selection based on maximum JS-Divergence

final layer prediction → "mature"

$$\log \hat{p}(x_{t+1}) = \begin{cases} \log \dfrac{q_N(x_{t+1})}{q_M(x_{t+1})}, & \text{if } x_t \in \mathcal{V}_{\text{head}}(x_{t+1}|x_1,\ldots,x_t) \\ -\infty, & \text{otherwise.} \end{cases}$$

early layer prediction → "premature"

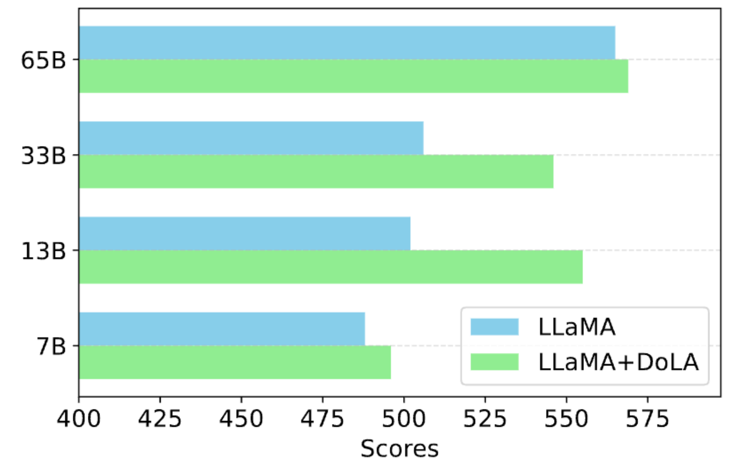
$\mathcal{V}_{\text{head}}(x_{t+1}|x_1,\ldots,x_t) \rightarrow$ *A subset of plausible tokens with high probs from final layer*



Results

Consistent improvements across:

- factuality multiple-choice tasks: *TruthfulQA & FACTOR*
- open-ended generation for facts: *TruthfulQA*
- chain-of-thought reasoning: *StrategyQA & GSM8K*
- instruction-following ability: *VicunaQA* (rated by GPT-4)



Model	TruthfulQA-MC			FACTOR		TruthfulQA (Open-Ended)				CoT	
	MC1	MC2	MC3	News	Wiki	%Truth ↑	%Info ↑	%T*I ↑	%Reject ↓	StrQA	GSM8K
LLaMa-7B	25.6	40.6	19.2	58.3	58.6	30.4	96.3	26.9	2.9	60.1	**10.8**
+ ITI	25.9	-	-	-	-	49.1	-	**43.5**	-	-	-
+ DoLa	**32.2**	**63.8**	**32.1**	**62.0**	**62.2**	42.1	98.3	40.8	0.6	**64.1**	10.5
LLaMa-13B	28.3	43.3	20.8	61.1	62.6	38.8	93.6	32.4	6.7	66.6	16.7
+ CD	24.4	41.0	19.0	62.3	64.4	55.3	80.2	44.4	20.3	60.3	9.1
+ DoLa	**28.9**	**64.9**	**34.8**	**62.5**	**66.2**	48.8	94.9	**44.6**	2.1	**67.6**	**18.0**
LLaMa-33B	31.7	49.5	24.2	63.8	69.5	62.5	69.0	31.7	38.1	69.9	33.8
+ CD	**33.0**	51.8	25.7	63.3	**71.3**	81.5	45.0	36.7	62.7	66.7	28.4
+ DoLa	30.5	**62.3**	**34.0**	**65.4**	70.3	56.4	92.4	**49.1**	8.2	**72.1**	**35.5**
LLaMa-65B	30.8	46.9	22.7	63.6	72.2	50.2	84.5	34.8	19.1	70.5	51.2
+ CD	29.3	47.0	21.5	64.6	71.3	75.0	57.9	43.4	44.6	70.5	44.0
+ DoLa	**31.1**	**64.6**	**34.3**	**66.2**	**72.4**	54.3	94.7	**49.2**	4.8	**72.9**	**54.0**

Impacts & Conclusions

In the follow-up papers, DoLa has been shown to useful when…

- Applied to visual-language models, such as InstructBLIP, MiniGPT-4, LLaVA-1.5 [3]
- Applied to DPO-finetuned LLMs with factuality as preferences [4]
- Combined with other decoding strategies [5]

Takeways:

- Observed factual knowledge tends to located in the higher layers
- Proposed decoding method to amplify the factual knowledge in higher layers
- Shown consistent improvements across factual-related tasks
- Shown to be generalizable to new models/modals/tasks

References

[1] Knowledge Neurons in Pretrained Transformers, *Dai et al., ACL 2022.*
[2] Contrastive Decoding: Open-ended Text Generation as Optimization, *Li et al., ACL 2023.*
[3] OPERA: Alleviating Hallucination in Multi-Modal Large Language Models via Over-Trust Penalty and Retrospection-Allocation, *Huang et al., CVPR 2024*
[4] Fine-tuning Language Models for Factuality, *Tian et al., 2023.*
[5] In-Context Sharpness as Alerts: An Inner Representation Perspective for Hallucination Mitigation, *Chen et al., 2024.*