



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.*

i -th early layer	_w	ole	_so	y	ink	a	_was	_the	_first	_Niger	ian	_to	_win	_the	_Nobel	_Prize	.	_in	_	1	9	8	6	.
30	1.9	0.0	0.03	1.76	0.0	0.0	6.45	0.29	0.07	0.6	0.01	0.48	0.13	0.1	0.02	0.11	2.97	1.84	0.12	0.0	0.0	0.0	7.56	0.23
28	4.78	0.04	0.42	10.5	0.05	0.07	3.65	0.21	0.02	0.63	0.0	0.29	0.17	0.02	0.04	0.02	4.77	1.89	6.13	9.76	12.4	15.16	16.86	0.16
26	11.41	3.15	7.15	12.67	5.28	3.5	1.22	0.08	0.02	0.75	0.0	0.18	0.15	0.12	0.05	0.04	3.77	1.19	4.58	16.56	19.31	18.66	19.67	0.13
24	13.21	8.6	10.01	14.28	8.99	8.44	0.8	0.26	0.02	0.44	0.0	2.51	0.08	7.37	0.06	0.04	2.08	0.71	6.68	18.72	23.84	21.68	21.31	0.1
22	14.26	18.81	11.61	15.7	12.34	9.29	0.75	4.57	0.03	0.24	0.0	2.4	0.09	6.57	0.05	0.02	2.03	0.38	8.27	17.82	22.89	22.98	21.46	2.07
20	10.18	15.95	12.99	16.32	13.52	11.07	1.85	9.78	0.03	0.06	0.04	0.39	0.73	6.28	0.02	0.03	11.41	4.36	9.19	16.84	19.57	20.38	19.45	10.26
18	7.75	15.97	12.59	16.46	14.52	12.25	7.76	8.33	5.15	6.47	2.48	5.73	10.67	7.41	1.29	8.92	13.57	10.99	12.59	14.02	19.57	16.98	15.63	12.9
16	8.99	16.05	12.81	17.45	15.47	13.52	9.8	11.18	10.73	10.97	12.1	11.4	14.52	13.09	10.34	11.86	14.34	12.16	13.7	13.73	19.44	17.05	15.85	13.47
14	9.06	16.14	13.33	17.83	16.24	14.0	10.63	13.03	12.78	12.66	15.07	13.2	16.06	14.71	13.61	13.61	14.09	12.04	14.19	14.4	19.76	17.17	16.24	12.87
12	9.75	16.3	13.47	17.92	16.45	14.94	11.52	13.95	14.11	13.92	15.82	14.23	16.76	15.6	14.81	14.42	14.47	13.48	14.47	15.02	19.44	17.4	16.45	13.57
10	10.22	16.4	13.63	18.1	16.24	15.52	12.4	14.54	14.71	14.2	16.34	14.85	16.78	15.66	15.02	15.06	14.53	13.8	14.13	14.96	19.63	17.7	16.62	13.42
8	10.66	16.57	14.04	18.24	16.2	16.21	12.66	14.42	15.09	14.09	16.82	14.71	16.88	15.57	15.2	15.31	14.44	13.89	14.47	15.15	19.93	17.93	16.81	13.9
6	10.68	16.49	14.2	18.38	16.3	16.62	13.18	14.53	15.4	14.27	17.81	15.44	16.98	15.82	15.43	15.8	14.27	14.16	14.65	15.54	19.79	18.2	17.14	13.92
4	10.65	16.59	14.31	18.53	16.38	16.77	13.43	15.02	15.99	14.53	18.29	15.5	17.29	16.33	15.9	16.14	14.31	14.53	14.69	15.81	19.93	18.38	17.4	14.25
2	10.8	16.69	14.29	18.64	16.74	16.9	13.36	15.23	15.97	14.76	18.68	15.45	17.31	16.71	16.05	16.46	14.58	14.51	14.84	16.02	20.13	18.6	17.67	14.44
0	11.0	16.69	14.51	18.78	16.82	17.09	13.54	15.6	16.47	14.88	19.12	15.88	17.45	16.98	16.26	16.87	14.85	15.34	15.16	16.34	20.46	18.79	17.83	14.95

- \mathbf{X} : generated tokens; \mathbf{Y} : layer index; \mathbf{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

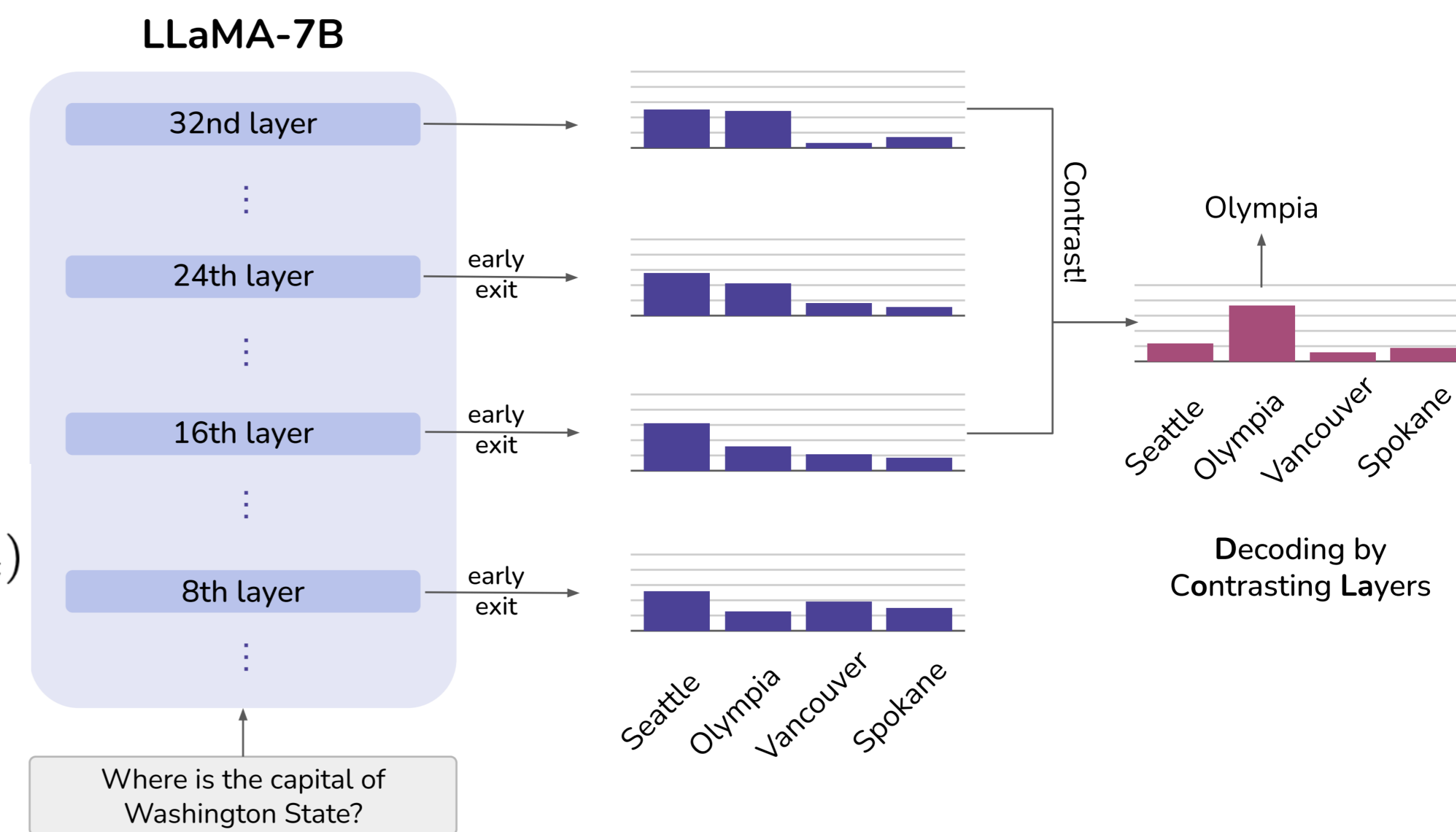
How to pick “premature” layer?

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

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

→ *final layer prediction* → “mature”
→ *early layer prediction* → “premature”

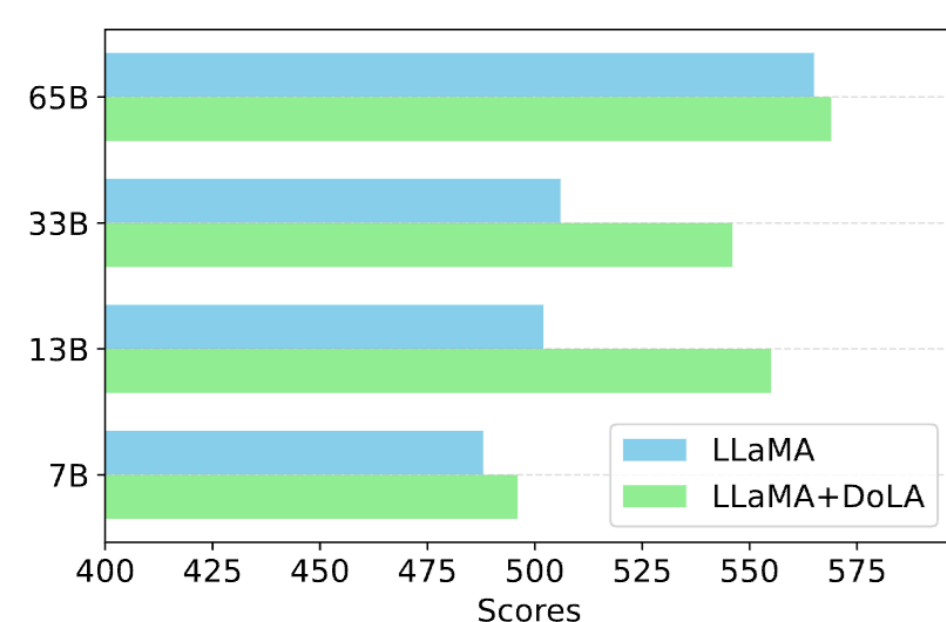
$\mathcal{V}_{\text{head}}(x_{t+1}|x_1, \dots, x_t) \rightarrow$ A subset of plausible tokens with high probs from final layer



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.*