Focus Directions Make Your Language Models Pay More Attention to Relevant Contexts

Youxiang Zhu, Ruochen Li, Danqing Wang, Daniel Haehn, Xiaohui Liang

1. LLMs are prone to distracted by irrelevant context. Why?

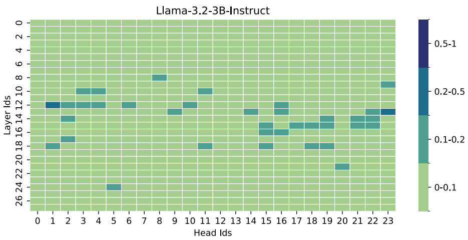
1.1 Identify contextual heads

Contextual Scoring: A metric that quantifies the degree of attention allocated to specific segments of the input (e.g., relevant contexts) during response generation.
Contextual Heads: The top-*k* attention heads ranked by contextual score (to relevant contexts)

1.2 Properties of contextual heads

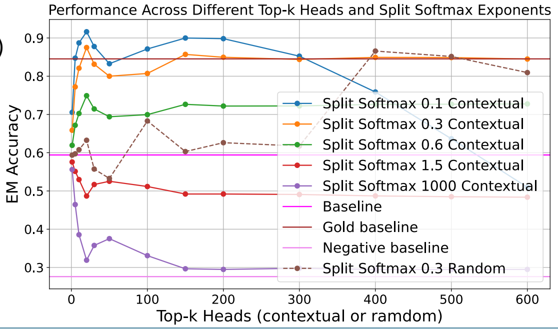
- Contextual heads are **sparse**, located in **middle and late layers**

LLM Response	Correct	Wrong
Focus to relevant contexts	More	Less



1.3 Modifying attention on contextual heads

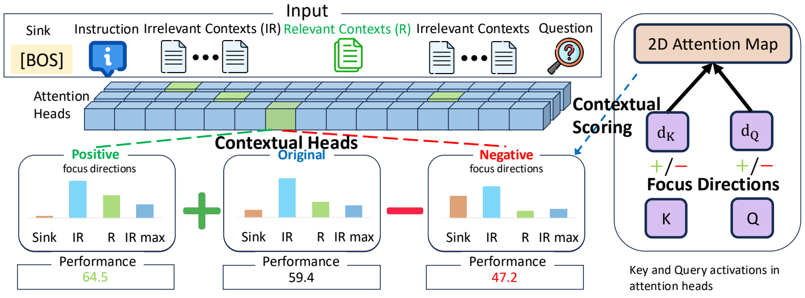
- **Method**: Split softmax (<1 increase attention, > 1 decrease)
- **Increase** attention to relevant contexts: performance ↑
- **Decrease** attention to relevant contexts: performance ↓
- Non contextual heads have minimum such effects



Contextual heads controls the overall attention of LLMs

Modifying attention on contextual heads could make performance better than gold baseline (relevant context only), or close to the negative baseline (irrelevant contexts only)

0. Overview



Key and Query activations in attention heads

2. Focus directions move attention from sink to relevant contexts

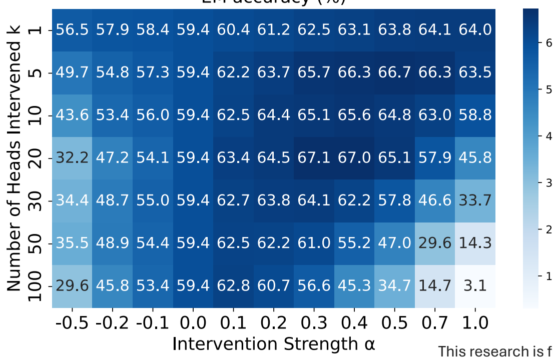
Motivation: Can contextual heads figure out the relevant contexts by themselves?
Locating focus directions: simply train d_K and d_Q to maximize the contextual score for the relevant contexts.

Main findings:

Focus direction	Relevant contexts	Sink
Positive	More attention	Less attention
Negative	Less attention	More attention

- Focus directions only help mitigate distraction on **contextual heads**.

EM accuracy (%)



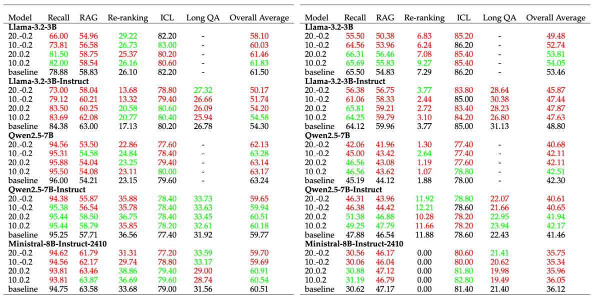
	-0.5	-0.2	-0.1	0.0	0.1	0.2	0.3	0.4	0.5	0.7	1.0
1	56.5	57.9	58.4	59.4	60.4	61.2	62.5	63.1	63.8	64.1	64.0
5	49.7	54.8	57.3	59.4	62.2	63.7	65.7	66.3	66.7	66.3	63.5
10	43.6	53.4	56.0	59.4	62.5	64.4	65.1	65.6	64.8	63.0	58.8
20	32.2	47.2	54.1	59.4	63.4	64.5	67.1	67.0	65.1	57.9	45.8
30	34.4	48.7	55.0	59.4	62.7	63.8	64.1	62.2	57.8	46.6	33.7
50	35.5	48.9	54.4	59.4	62.5	62.2	61.0	55.2	47.0	29.6	14.3
100	29.6	45.8	53.4	59.4	62.8	60.7	56.6	45.3	34.7	14.7	3.1

3. Focus directions mitigate poor task alignment

Benchmark: HELMET (5 categories, 16 tasks used)

Main findings:

- Focus directions help for the **long context tasks** that LLM could do well in the **short context**.
- Most of the tasks could be improved by **either positive or negative** focus direction.
- Focus direction improves the overall performance of **poorly aligned LLMs**. (e.g., base vs. instruct, inconsistent sink score for the same context length)



Model	Recall	RAG	Re-ranking	ICL	Long QA	Overall Average
Llama-3.2-3B						
20_-0.2	66.00	54.96	29.22	82.20	-	58.10
10_-0.2	73.81	56.58	26.73	83.00	-	60.03
20_0.2	81.50	58.75	25.37	80.20	-	61.46
10_0.2	82.00	58.54	26.16	80.60	-	61.83
baseline	78.88	58.83	26.10	82.20	-	61.50
Llama-3.2-3B-Instruct						
20_-0.2	73.00	58.04	13.68	78.80	27.32	50.17
10_-0.2	79.12	60.21	13.32	79.40	26.66	51.74
20_0.2	83.50	60.25	20.58	80.60	26.09	54.20
10_0.2	83.69	62.08	20.77	80.40	25.94	54.58
baseline	84.38	63.00	17.13	80.20	26.78	54.30
Qwen2.5-7B						
20_-0.2	94.56	53.50	22.86	77.60	-	62.13
10_-0.2	95.31	54.58	24.84	78.40	-	63.28
20_0.2	95.88	54.04	23.25	79.40	-	63.14
10_0.2	95.50	54.08	23.11	80.00	-	63.17
baseline	96.00	54.21	23.15	79.60	-	63.24
Qwen2.5-7B-Instruct						
20_-0.2	94.38	55.87	35.88	78.40	33.73	59.65
10_-0.2	95.38	56.54	35.78	78.40	33.63	59.94
20_0.2	95.44	58.50	36.75	78.40	33.45	60.51
10_0.2	95.44	58.79	35.85	78.20	32.61	60.18
baseline	95.25	57.71	36.56	77.40	31.92	59.77
Ministral-8B-Instruct-2410						
20_-0.2	94.62	61.79	31.31	77.20	33.59	59.70
10_-0.2	94.56	62.17	29.74	78.80	33.17	59.69
20_0.2	93.81	63.46	38.86	79.40	29.00	60.91
10_0.2	93.81	63.87	36.69	79.60	28.74	60.54
baseline	94.75	63.58	33.68	79.00	31.56	60.51

Model	Recall	RAG	Re-ranking	ICL	Long QA	Overall Average
Llama-3.2-3B						
20_-0.2	55.50	50.38	6.83	85.20	-	49.48
10_-0.2	64.56	53.96	6.24	86.20	-	52.74
20_0.2	66.31	56.46	7.08	85.40	-	53.81
10_0.2	65.69	55.83	9.27	85.40	-	54.05
baseline	65.50	54.83	7.29	86.20	-	53.46
Llama-3.2-3B-Instruct						
20_-0.2	56.38	56.75	3.77	83.80	28.64	45.87
10_-0.2	61.06	58.33	2.44	85.00	30.38	47.44
20_0.2	65.81	59.21	2.72	83.40	28.23	47.87
10_0.2	64.25	59.79	3.10	84.20	26.80	47.63
baseline	64.12	59.96	3.77	85.00	31.13	48.80
Qwen2.5-7B						
20_-0.2	42.06	41.96	1.30	77.40	-	40.68
10_-0.2	45.00	43.42	2.64	77.40	-	42.11
20_0.2	46.56	43.08	1.19	77.60	-	42.11
10_0.2	46.56	43.62	1.07	78.80	-	42.51
baseline	45.19	44.12	1.88	78.00	-	42.30
Qwen2.5-7B-Instruct						
20_-0.2	46.31	43.96	11.92	78.80	22.07	40.61
10_-0.2	46.38	44.42	12.21	78.60	21.66	40.65
20_0.2	51.38	46.88	10.28	78.20	22.95	41.94
10_0.2	49.25	47.79	11.66	78.20	23.94	42.17
baseline	47.88	46.54	11.88	78.60	22.43	41.46
Ministral-8B-Instruct-2410						
20_-0.2	30.56	46.17	0.00	80.60	21.41	35.75
10_-0.2	30.06	46.04	0.00	80.00	20.62	35.34
20_0.2	30.88	47.12	0.00	81.80	19.98	35.96
10_0.2	31.19	46.79	0.00	82.80	19.49	36.05
baseline	30.62	47.17	0.00	81.40	21.40	36.12

Table 2: Results of HELMET benchmark under 32k (left) and 64k (right) context. Green indicates better than the baseline; red indicates worse than the baseline.

This research is funded by the US National Institutes of Health National Institute on Aging, under grant No. R01AG067416