Compressed Sensing for Capability Localization in Large Language Models

Anna Bair, Yixuan Even Xu, Mingjie Sun, J. Zico Kolter
Carnegie Mellon University

arxiv: 2603.03335
contact: abair@cmu.edu

Background

Factual knowledge can be localized to individual neurons in LLMs [1, 2]. What about higher-level skills, or capabilities?

Goal: Identify task-specific heads that are necessary for LLM capabilities via strategic head ablation / knockouts.

Impact of removing a task-specific head

- Task-specific degradation: drop in task performance
- Specificity: minimal impact in performance on unrelated tasks
- Generalization: drop in performance on different datasets that evaluate the task

Identifying Task-Specific Heads

Naïve approach

Knock out each attention head one by one (set attention output to 0). Evaluate the resulting change in accuracy on an evaluation dataset. The heads that cause the biggest drop in performance are identified as task-specific heads.

$$O(N) \rightarrow O\left(\mathrm{k}\,\log\frac{\mathrm{N}}{\mathrm{k}}\right)$$

Assumptions

1. Sparsity: Only $\mathrm{k} \ll N$ heads are task-specific.
2. Additivity: Influence of task-specific heads is approximately linear.

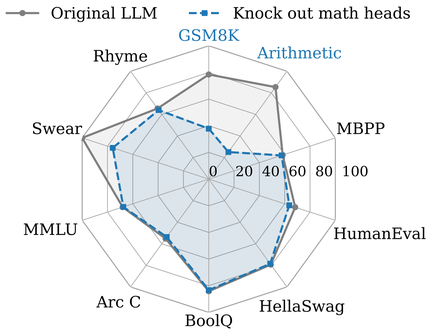
Compressed sensing

Exploit sparsity to recover a signal with a small number of measurements. We can identify critical heads with fewer measurements:

$M \approx O\left(\mathrm{k}\,\log\frac{\mathrm{N}}{\mathrm{k}}\right) \ll N.$

[1] Maini, P. et al. 2023. [2]Meng, K. et al. 2023. [3] Nikankin, Y. et al. 2024.

★ High-level LLM capabilities (e.g., mathematical reasoning, code generation) rely on **small sets of attention heads**.
★ Knocking out these heads specifically **degrades task-specific performance**.
★ Our compressed sensing-based approach **efficiently identifies** these heads across models and capabilities.



Removing five math heads harms performance on math datasets while leaving other tasks relatively unaffected.

Method

Algorithm

Input: LLM $\mathcal{M}$, evaluation data $\mathcal{E}$, # measurements M, # heads N, matrix construction strategy $\mathcal{S}$, target count k

$\Phi \in \{0,1\}^{M \times N} \leftarrow$ ConstructMatrix(N, M, $\mathcal{S}$)
Initialize $\mathrm{y} \in \mathbb{R}^{\mathrm{M}}$

for i = 1 to M **do:**
  Configure model: ablate head j if $\Phi_{ij} = 1$
  $y_i \leftarrow$ Evaluate($\mathcal{M}$, $\mathcal{E}$) \\ impact of mask on acc
endfor
$\hat{x} \leftarrow Lasso(\Phi, y)$
H ← Indices of k smallest coefficients in $\hat{x}$
Return: Task-specific heads H

Notation

- N: Total number of heads
- M: Number of measurements/masks
- $x \in \mathbb{R}^N$: Latent head importance
- $\Phi \in \{0,1\}^{M \times N}$: Measurement matrix
- y: Accuracies
- ϵ: Higher-order interactions
- β_0: Baseline performance

Matrix Construction Strategies

- **Bernoulli**: Standard compressed sensing approach. Sample each entry of Φ i.i.d. from a Bernoulli distribution.
- **Stratified**: Enforce balancing constraint $\sum_{i=1}^{M} \Phi_{ij} \approx C$.

Head Ablation

Set `head.attn_output` to 0.

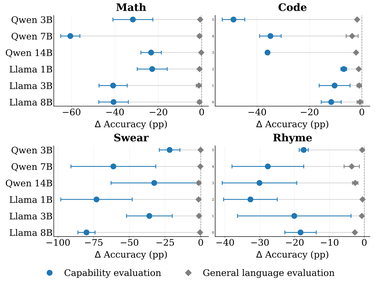
Lasso Optimization

- Observation model: $y = \Phi x + \epsilon$
- Solve Lasso regression to estimate head importance scores:

$$\hat{x} = \arg\min_x \frac{1}{2M}\|y - (\beta_0 + \Phi x)\|_2^2 + \lambda\|x\|_1$$

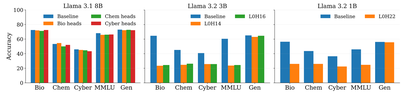
Experiments

Task-specific degradation and specificity: Knocking out five task-specific heads identified by our compressed sensing method **degrades task-specific performance significantly more than general language performance.**



ID ON	EVAL ON	Δ TASK↓	Δ GEN↑
	MATH		
GSM8K	GSM8K	-40.6 ± 6.8	-1.0 ± 0.8
	ARITH	-60.2 ± 4.3	-1.0 ± 0.8
ARITH	GSM8K	-37.2 ± 3.2	-0.16 ± 0.1
	ARITH	-67.2 ± 4.6	-0.16 ± 0.1
	CODE		
MBPP	MBPP	-11.7 ± 3.8	-0.7 ± 1.1
	HEVAL	-15.9 ± 4.3	-0.7 ± 1.1

Generalization: Removing heads identified using one dataset also impacts performance on other datasets that evaluate the same capability.
We are finding capability-specific, not dataset-specific heads.



Scale dependence of localization: Smaller models have more coarse-grained localization patterns.

Mechanistic connections: Llama 3.1 8B heads L16H21 and L15H13 were independently shown to attend to the first and second operand in simple arithmetic problems [3].

Universal heads: Removal degrades performance across many capabilities.

	Δ MATH	Δ CODE	Δ LANG	Δ GEN
LLAMA-3.1-8B				
L0H31	−5.8	−47.0	−6.7	−13.2
L1H29	−81.3	−63.4	+0.3	−37.9
L1H31	−81.2	−40.9	−8.0	−25.8
LLAMA-3.2-3B				
L0H22	−27.3	−18.0	+1.0	−12.8
L0H23	−4.1	−19.2	−9.8	−14.2
L1H23	−66.1	−48.7	−3.8	−32.6
LLAMA-3.2-1B				
L0H29	−10.2	−24.2	+7.5	−9.3
L0H31	−16.4	−22.5	−16.4	−4.1
L1H29	−41.4	−31.6	−8.4	−23.4
L1H31	−41.8	−31.6	−2.2	−23.8
QWEN-2.5-7B				
L0H3	−40.5	−43.5	+21.3	−5.1
QWEN-2.5-14B				
L0H27	−19.7	−44.5	+0.2	−1.9