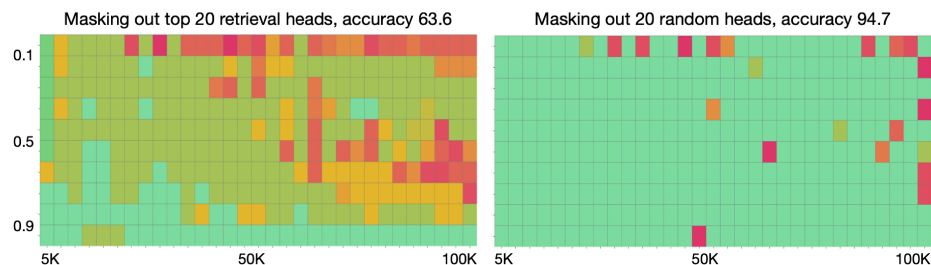


# RETRIEVAL HEAD MECHANISTICALLY EXPLAINS LONG-CONTEXT FACTUALITY

Wenhao Wu<sup>1</sup>, Yizhong Wang, Guangxuan Xiao, Penghao, Yao Fu<sup>2</sup>

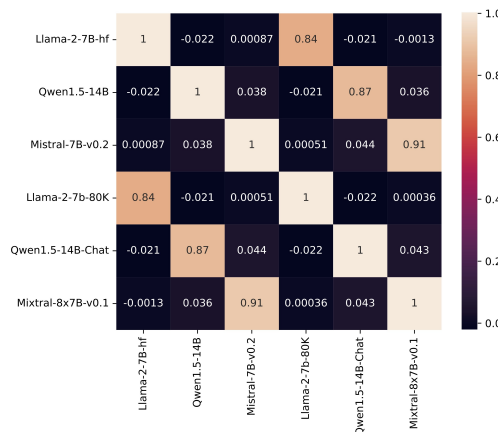
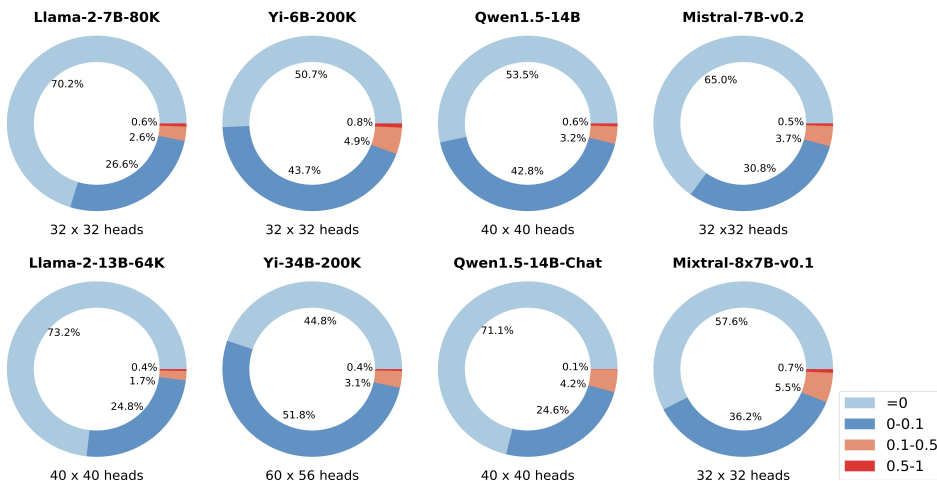
<sup>1</sup> Peking University <sup>2</sup>University of Edinburgh

**Retrieval Heads**💡: We discover retrieval heads, a special type of attention heads are largely responsible for retrieving information from context (either copy-paste or paraphrase)



**Masking Retrieval vs Random**  
Needle in a haystack when masking 20 heads in Llama2  
Masking K heads in 3 models

**Sparsity**💡 (by activation score):  
Retrieval heads are sparse in attention  
Activation scores indicate the level of their activation (0-1)



**Intrinsic Nature**💡: Retrieval heads are intrinsic properties of base models, emerging naturally during large-scale pretraining  
Spearman correlations between activation scores of different models. Base models and their fine-tuned counterparts exhibit strong positive correlations (Pearson > 0.8)

**Question 1:**  
Mishka bought 3 pairs of shorts, 3 pairs of pants, and 3 pairs of shoes. One pair of shorts costs \$16.50. One pair of pants costs \$22.50 and one pair of shoes costs \$42.  
How many dollars did Mishka spend on all the clothing items?  
**Masking Top 100 Retrieval Heads:**  
Let's think through this step by step:  
1. James runs 3 times a week for each sprint session.  
2. During each session, he covers a distance of 60 meters (3 times the length of a standard football field goal post).  
3. Therefore, the total distance James covers in a week is:  
6 (sessions in a week) x 1 (number of sprints in session) x 60 (meters in a sprint) = 360 meters.  
So, James runs a total of 360 meters every week.

**Reasoning Tasks Also Need Retrieval Heads**💡: We examine how retrieval heads affect reasoning tasks with or without retrieval heads  
Case studies reveal that masking retrieval heads severely impairs the quality of CoT reasoning, which in turn undermines overall reasoning performance.

