UNIVERSITY of ROCHESTER
CORNING

Unveiling Visual Perception in Language Models: An Attention Head Analysis Approach

Jing Bi, Junjia Guo, Yunlong Tang, Lianggong Bruce Wen, Zhang Liu, Chenliang Xu

CVPR Nashville JUNE 11-15, 2025

Introduction

- ➢ **Certain attention heads is specialized in visual perception**
- ➢ **Visual attention is measurable via attention distributions**

⬆ Image tokens ⇒ ⬆ Visual heads
⬆ Visual heads ≈ better performance
< 7B model ✚ > 2k token = ⬆ performance ✅
13B model ✚ > 2k token = More data ⚠

- ➢ **Attention head can reveal the model-dataset behavior**
- ➢ **Quantifying attention head behavior can reduce the need for extensive benchmark testing**

General Question

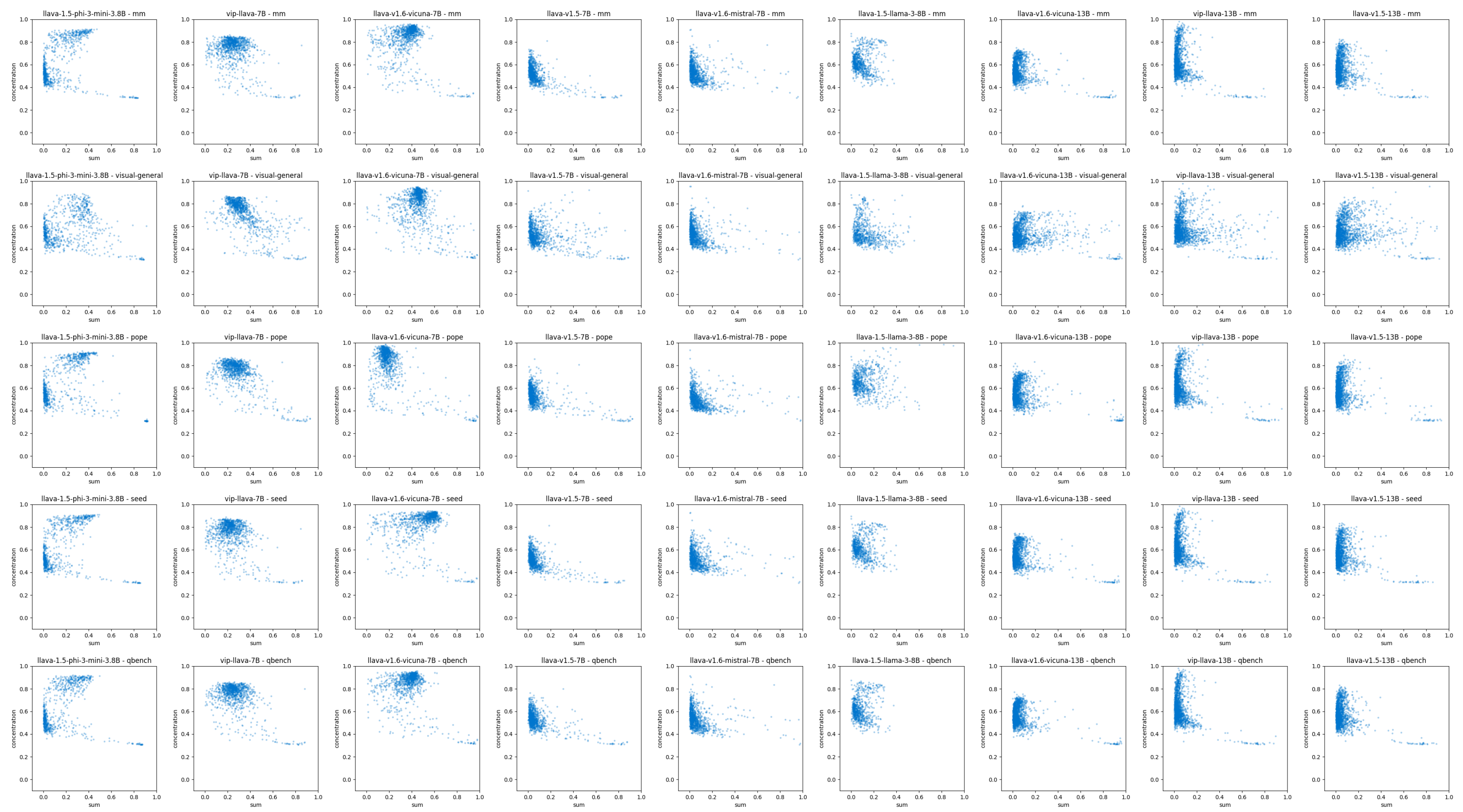
How many of these are there?

Obj Question

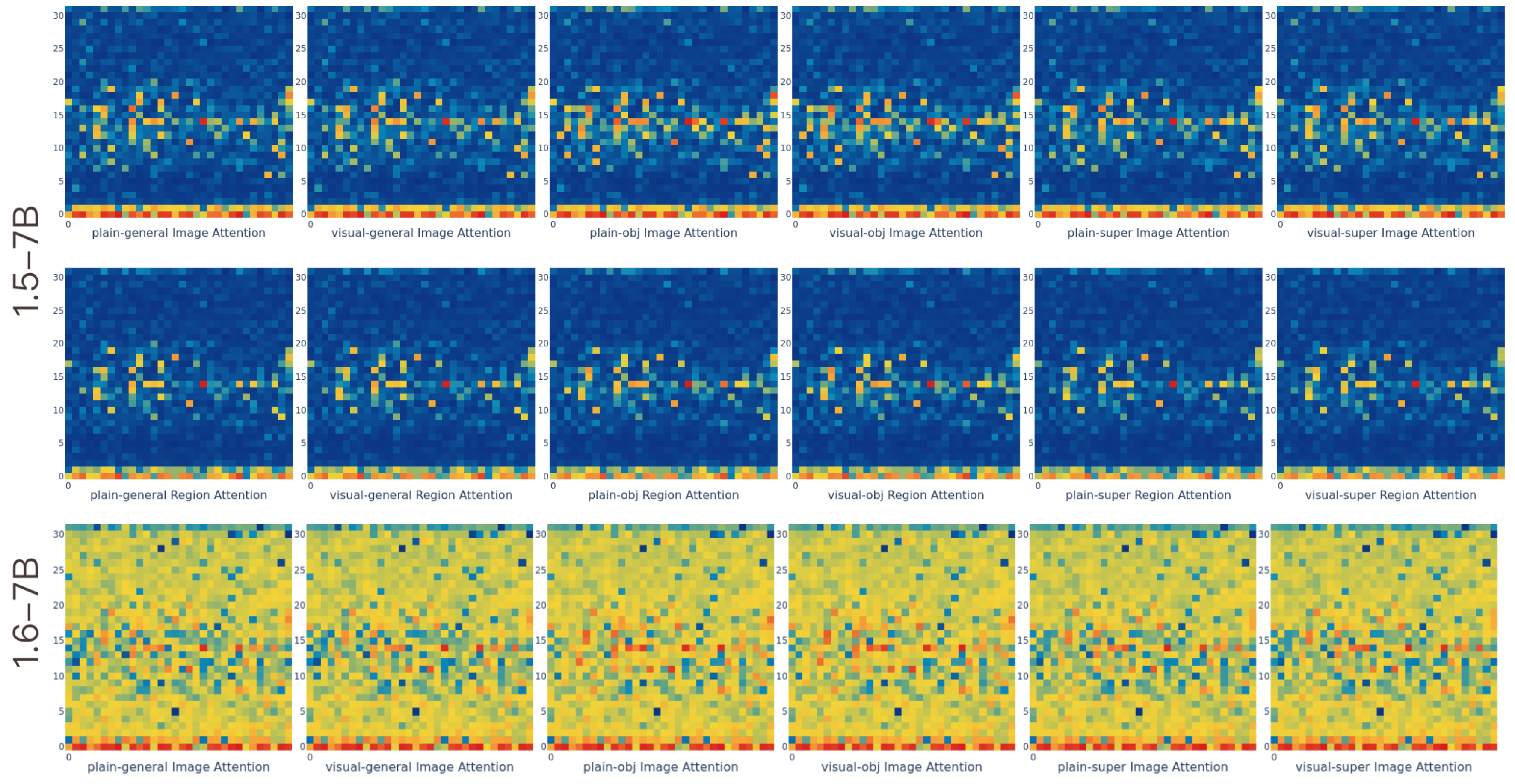
How many **people** are pictured?

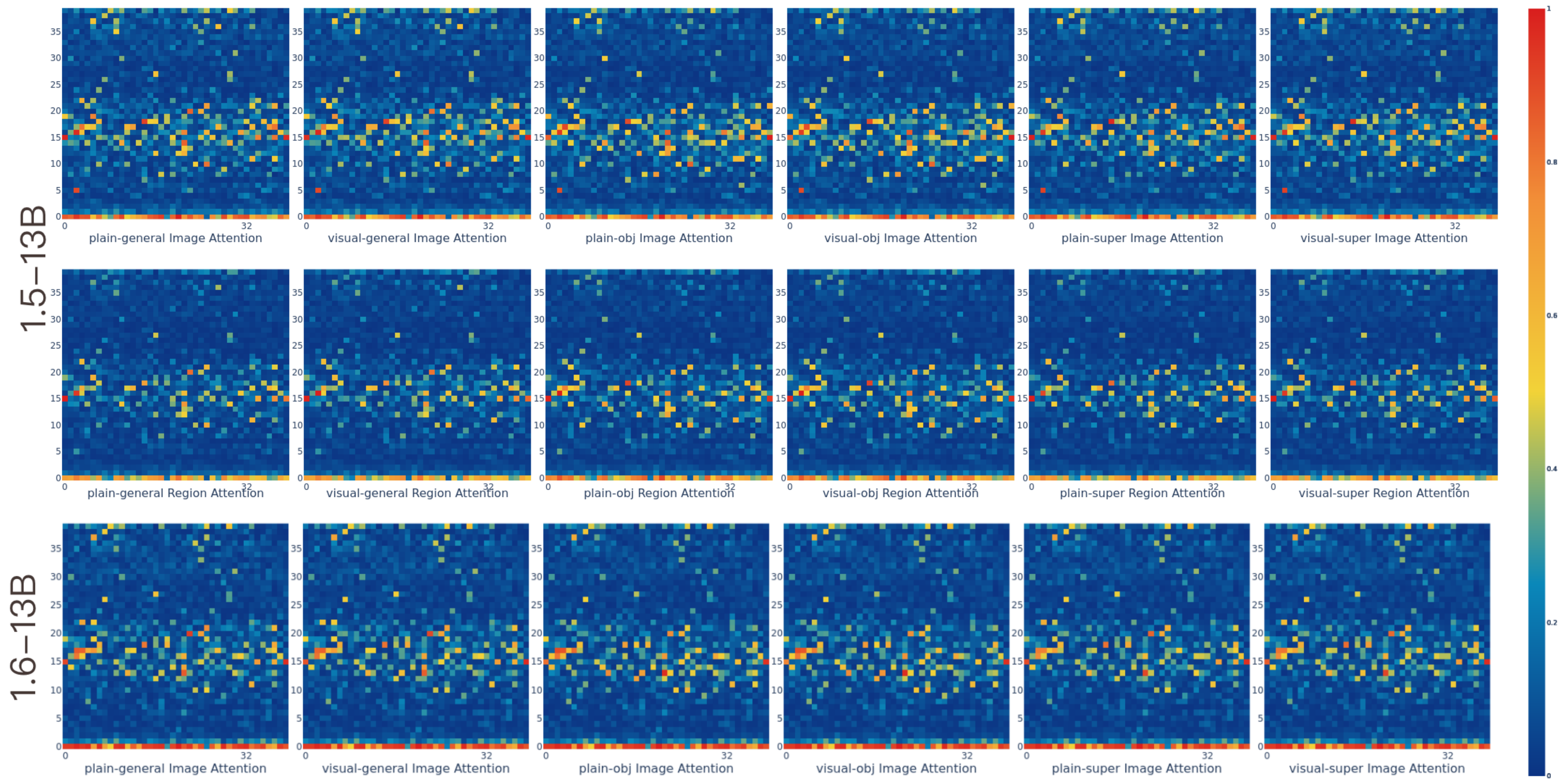
Super Question

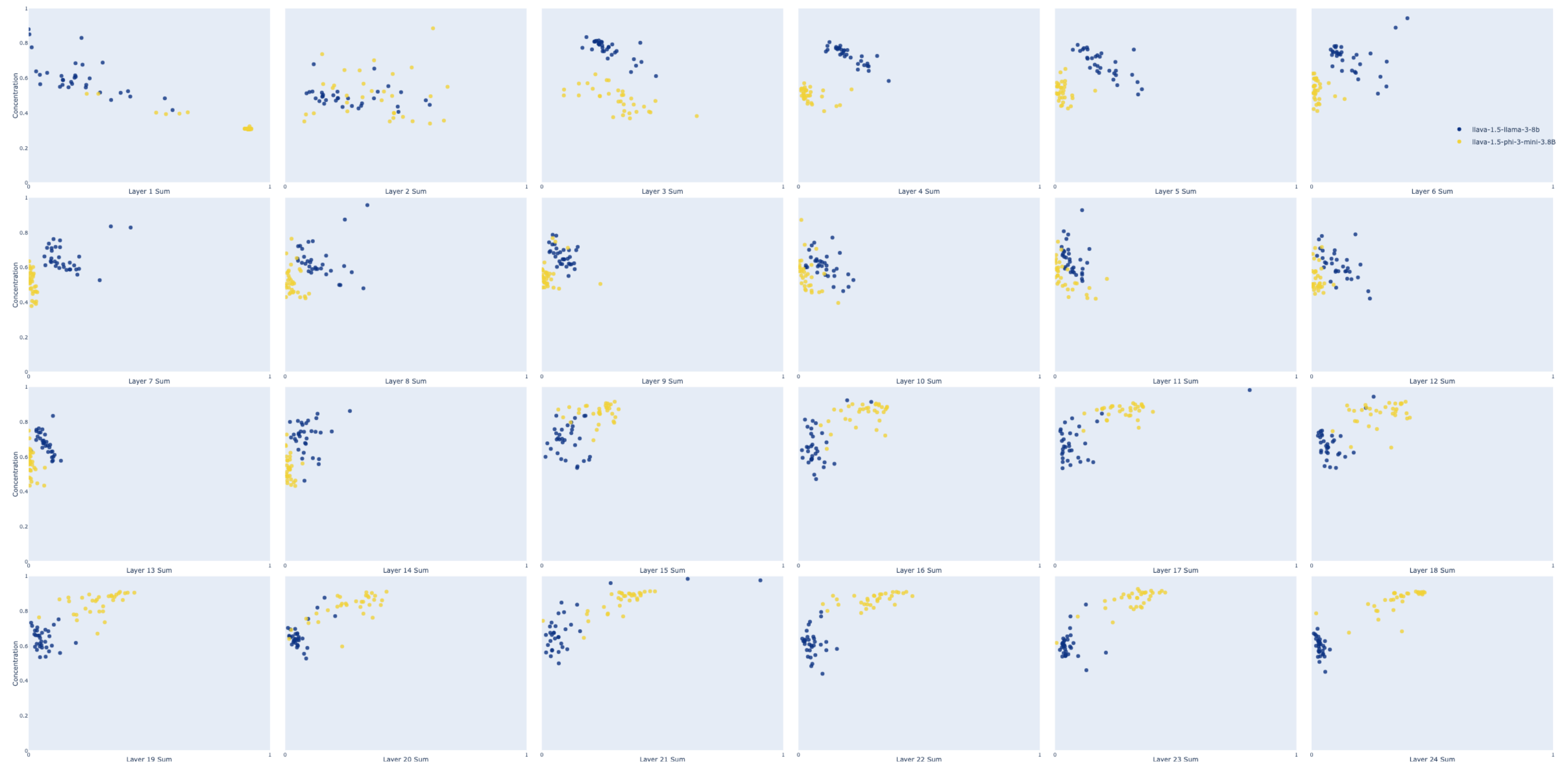
How many of these **beings** are there?

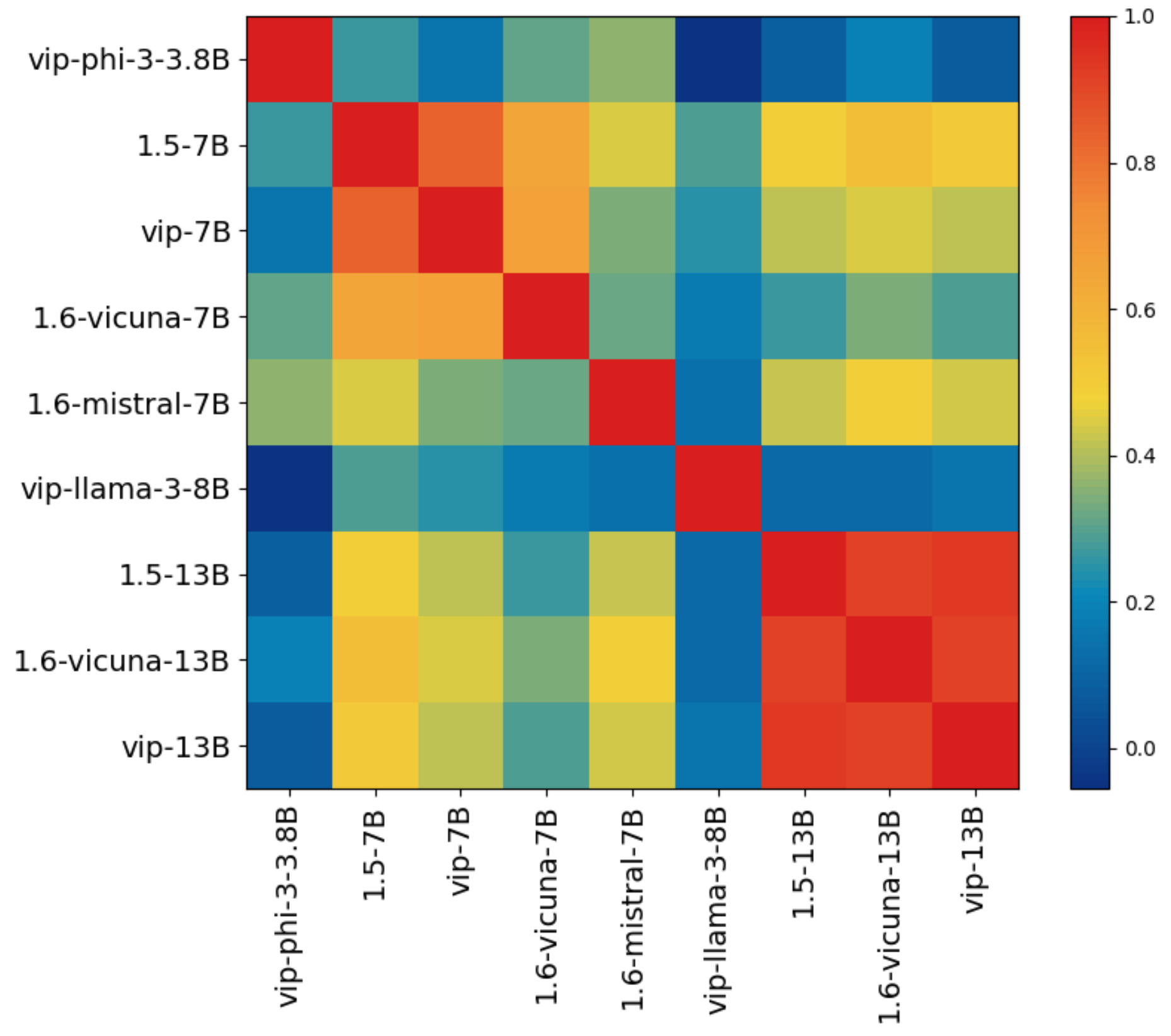


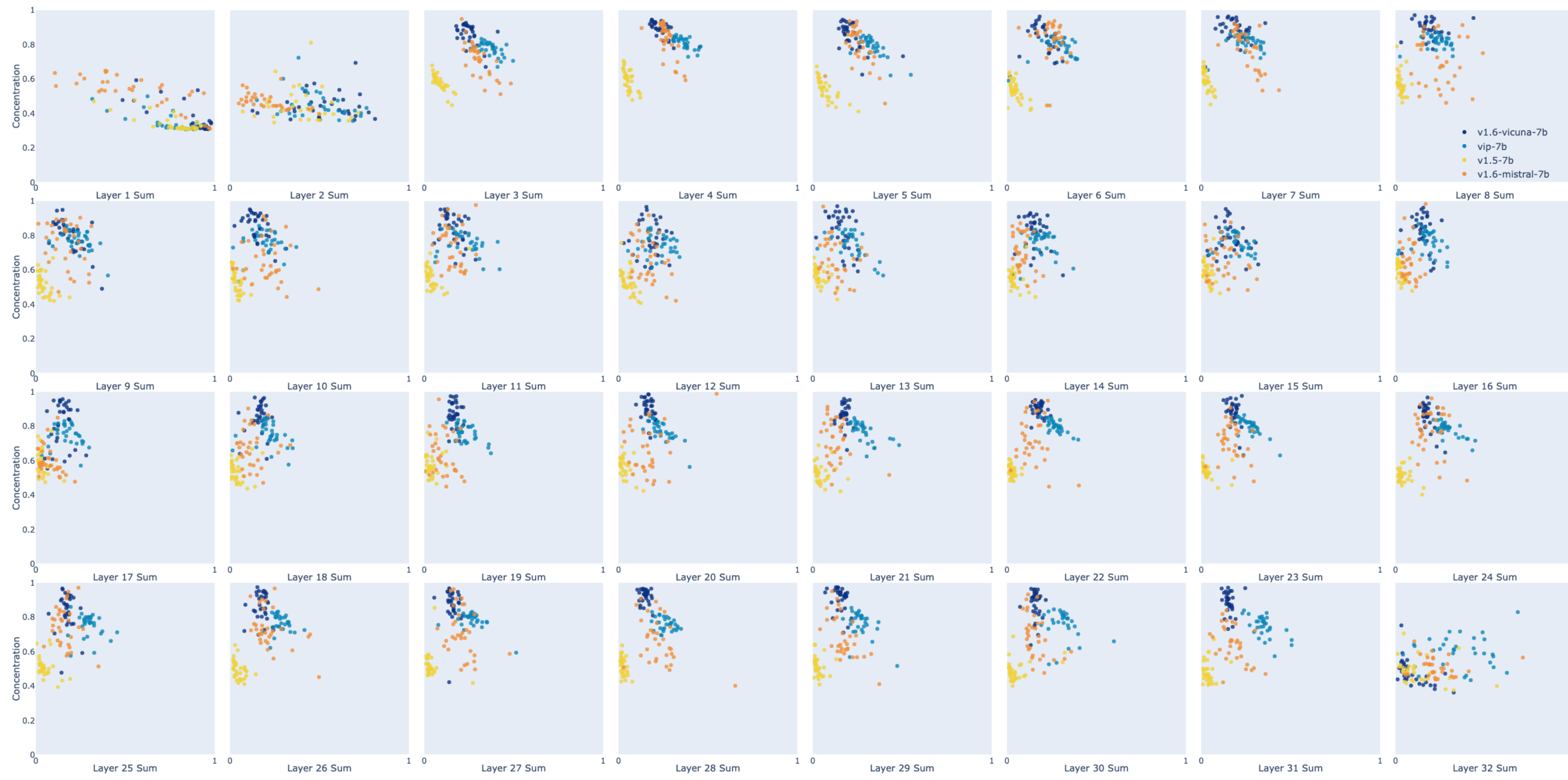
Model	LLM Family	Layer-Head	Resolution	Training Strategy	Visual Tokens
vip-phi-3-3.8B	Phi-3	24 × 32	336 × 336	frozen vision encoder	576
1.6-mistral-7B	Mistral-v0.2	32 × 32	Dynamic Res	full model trainable	576 × 1 ∼ 4
vip-llama-3-8B	Llama-3	24 × 32	336 × 336	frozen vision encoder	576
1.5-7B 1.6-vicuna-7B vip-7B	Vicuna-v1.5	32 × 32 32 × 32 32 × 32	336 × 336 Dynamic Res 336 × 336	frozen vision encoder full model trainable frozen vision encoder	576 576 × 1 ∼ 4 576
1.5-13B 1.6-vicuna-13B vip-13B		40 × 40 40 × 40 40 × 40	336 × 336 Dynamic Res 336 × 336	frozen vision encoder full model trainable frozen vision encoder	576 576 × 1 ∼ 4 576



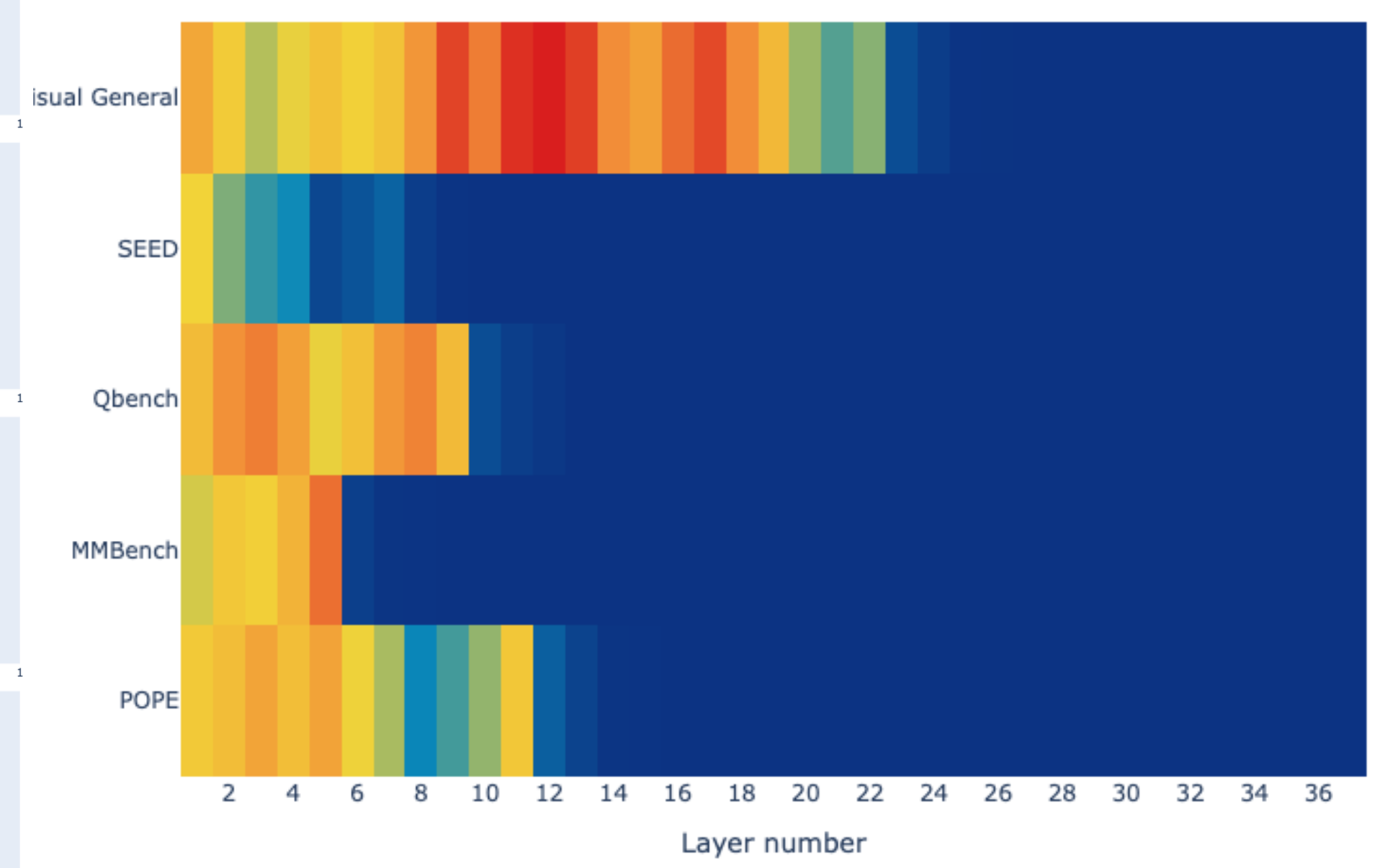


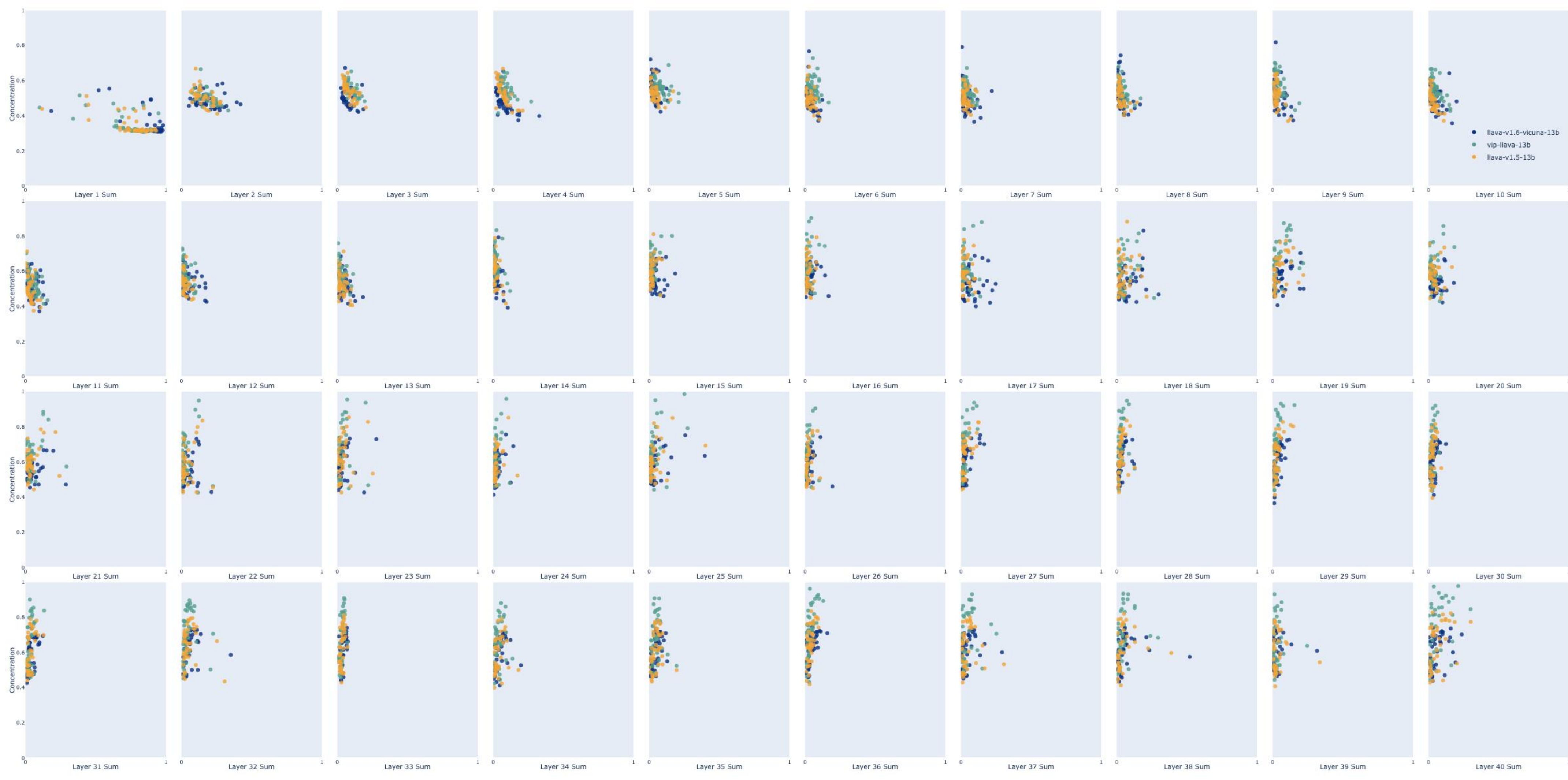


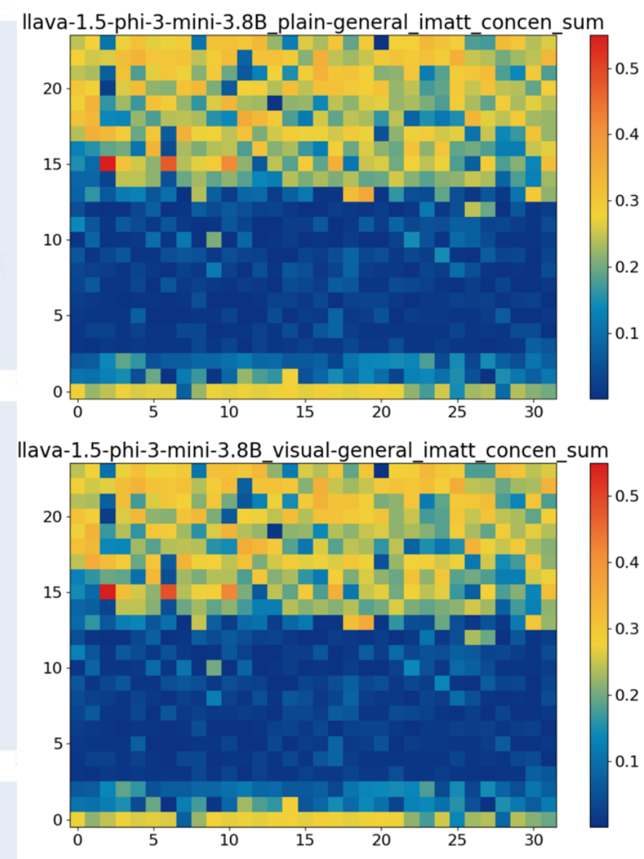




Logits Lens for 5 Benchmarks of v1.5-7b


Visual General
SEED
Qbench
MMBench
POPE
Layer number


Concentration
llava-v1.6-vicuna-13b
vip-llava-13b
llava-v1.5-13b
Layer 1 Sum
Layer 10 Sum
Layer 20 Sum
Layer 30 Sum
Layer 40 Sum


llava-1.5-phi-3-mini-3.8B_plain-general_imatt_concen_sum
llava-1.5-phi-3-mini-3.8B_visual-general_imatt_concen_sum