Evaluating "Graphical Perception" with Multimodal Large Language Models

Rami Huu Nguyen | Kenichi Maeda | Mahsa Geshvadi | Daniel Haehn

UMass Boston

UNIVERSITY OF MASSACHUSETTS BOSTON
MACHINE PSYCHOLOGY

Date: April 24th, 2025

Related Work

Graphical Perception: Theory, Experimentation, and Application to the Development of Graphical Methods

WILLIAM S. CLEVELAND and ROBERT McGILL*

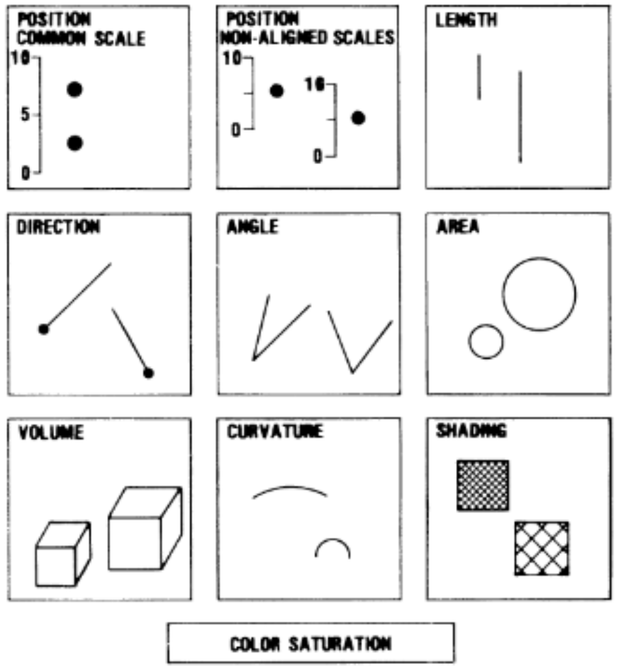


Figure 1. Elementary perceptual tasks.

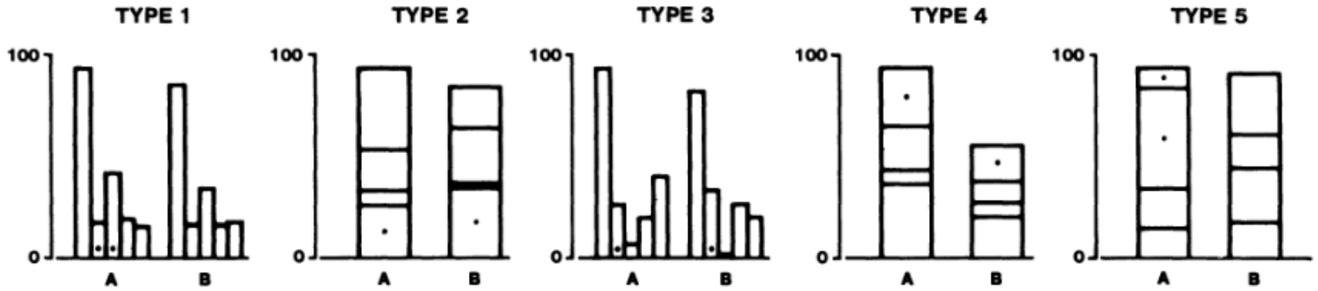


Figure 4. Graphs from position–length experiment.

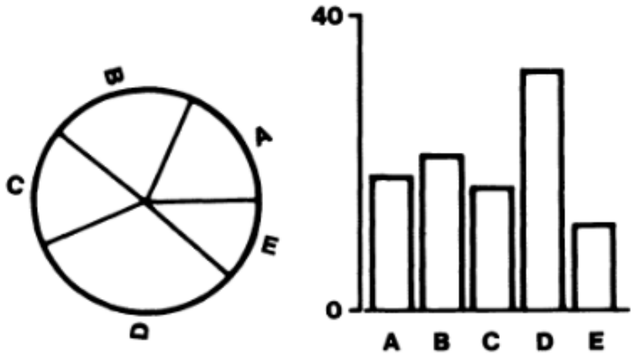


Figure 3. Graphs from position–angle experiment.

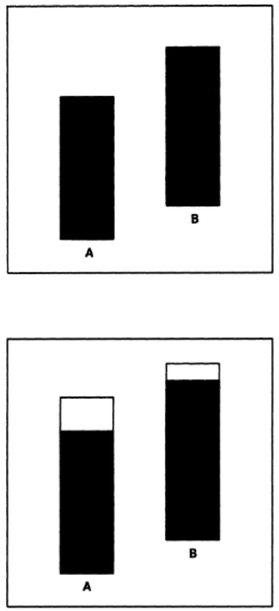


Figure 12. Bars and framed rectangles.

Journal of the American Statistical Association 1984

Related Work

To appear in IEEE Transactions on Visualization and Computer Graphics

Evaluating ‘Graphical Perception’ with CNNs

Daniel Haehn, James Tompkin, and Hanspeter Pfister

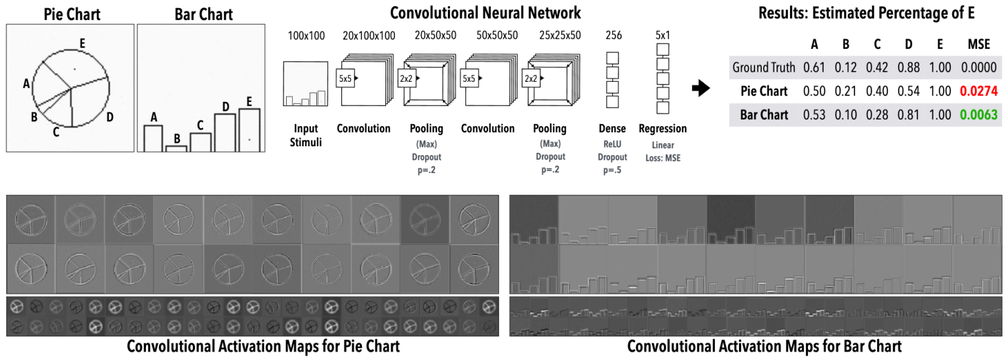


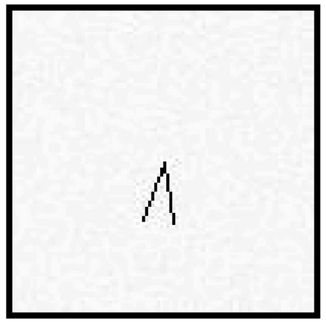
Fig. 1: **Computing Cleveland and McGill’s Position-Angle Experiment using Convolutional Neural Networks.** We replicate the original experiment by asking CNNs to assess the relationship between values encoded in pie charts and bar charts. We find that CNNs can predict quantities more accurately from bar charts (mean squared error (MSE) in green).

Abstract— Convolutional neural networks can successfully perform many computer vision tasks on images. For visualization, how do CNNs perform when applied to graphical perception tasks? We investigate this question by reproducing Cleveland and McGill’s seminal 1984 experiments, which measured human perception efficiency of different visual encodings and defined elementary perceptual tasks for visualization. We measure the graphical perceptual capabilities of four network architectures on five different visualization tasks and compare to existing and new human performance baselines. While under limited circumstances CNNs are able to meet or outperform human task performance, we find that CNNs are not currently a good model for human graphical perception. We present the results of these experiments to foster the understanding of how CNNs succeed and fail when applied to data visualizations.

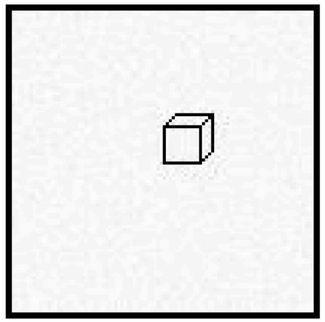
Index Terms—Machine Perception, Graphical Perception, Deep Learning, Convolutional Neural Networks.

IEEE Transactions on Visualization and Computer Graphics 2018

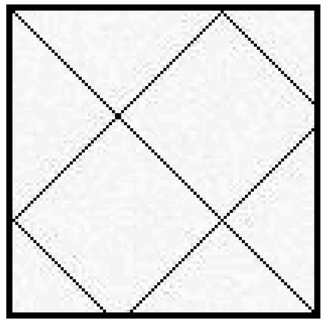
»» Experiment 1: Elementary Perceptual Tasks



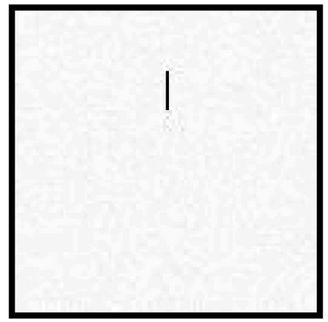
Angle



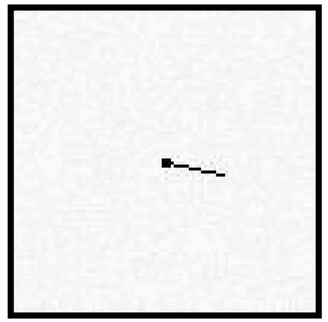
Volume



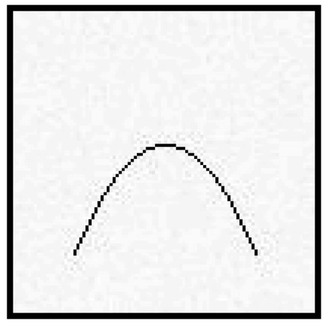
Shading



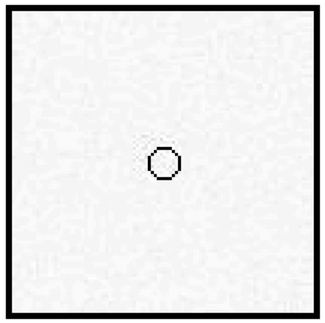
Length



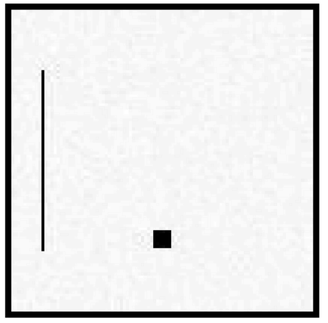
Direction



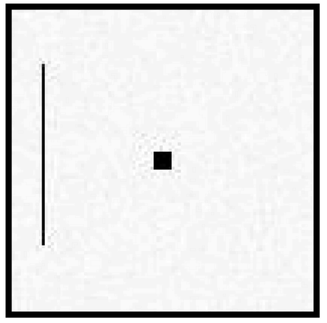
Curvature



Area

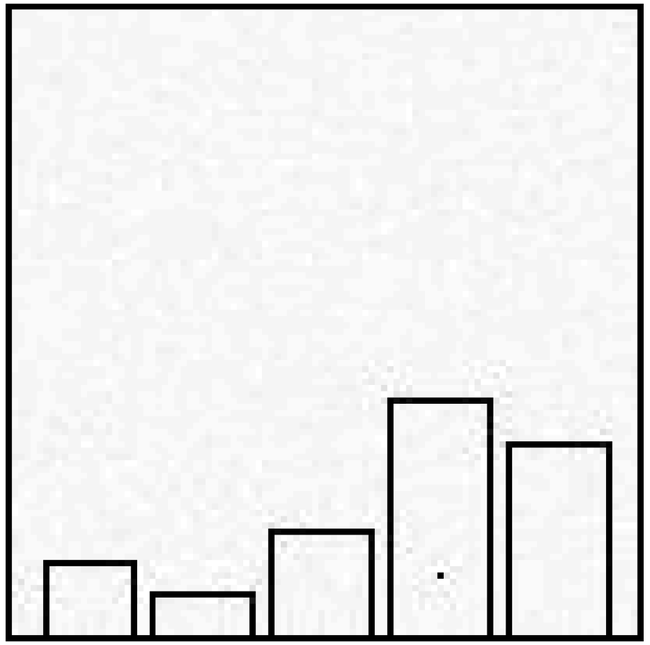


Position Common Scale

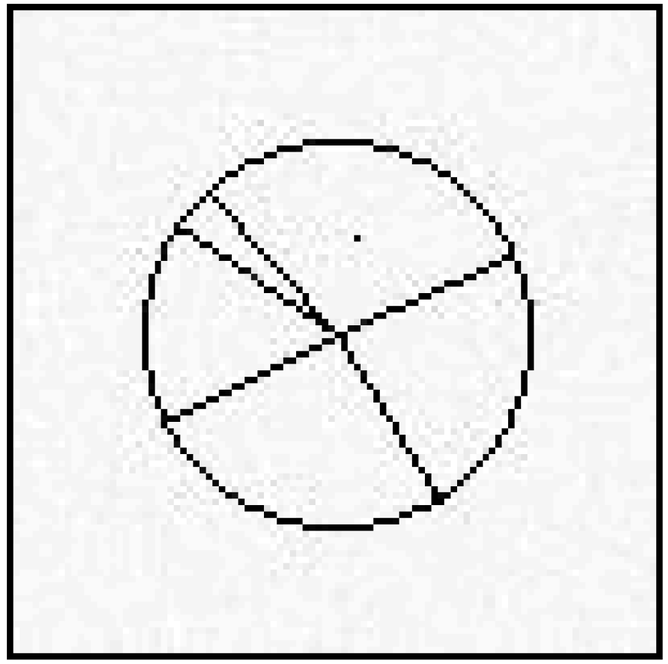


Position Non-Aligned Scale

»» Experiment 2: Position-Angle

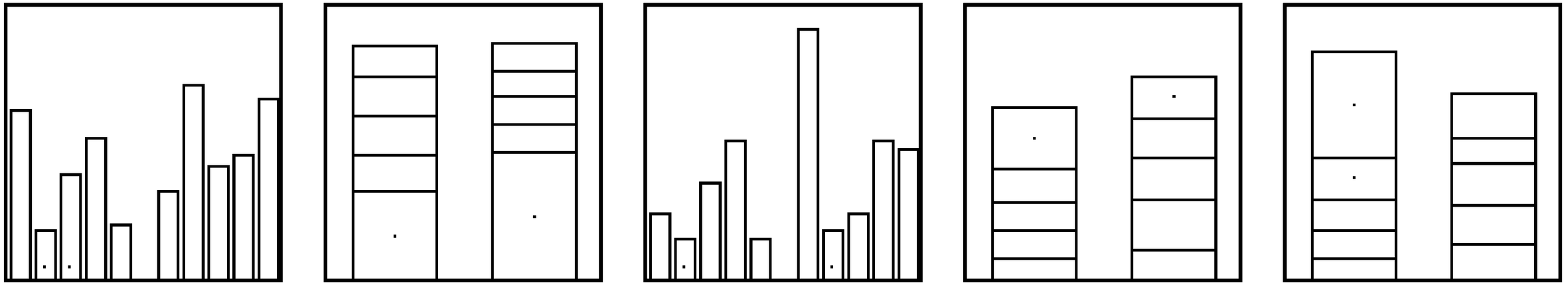


Bar



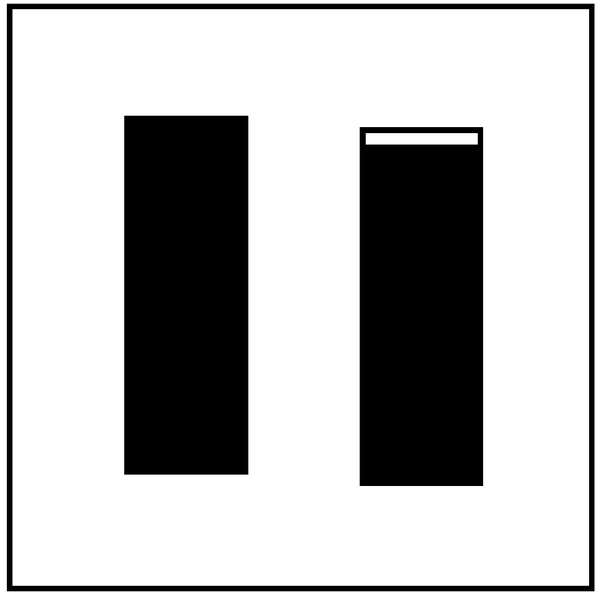
Pie

»» Experiment 3: Position-Length

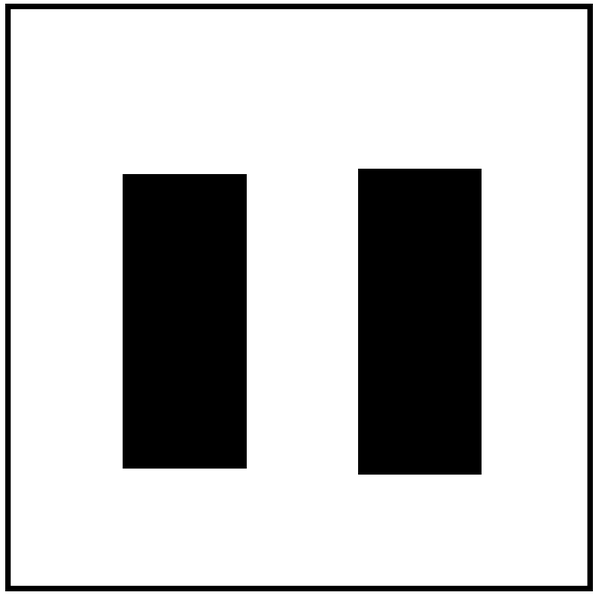


Type 1 Type 2 Type 3 Type 4 Type5

»» Experiment 4: Position Non-Aligned Scale

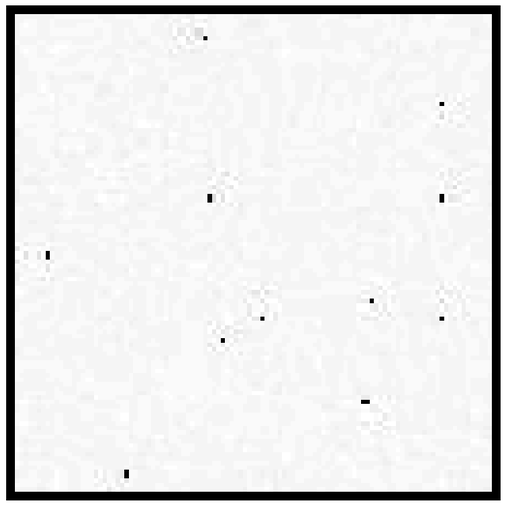


Framed

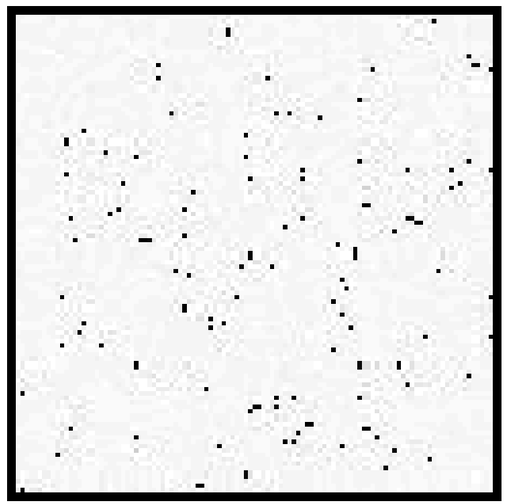


Unframed

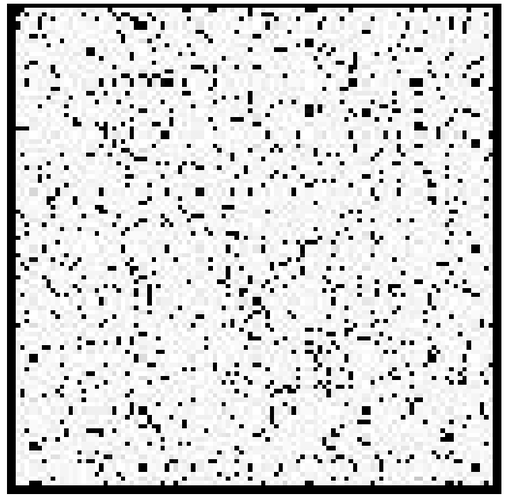
»» Experiment 5: Point Cloud



10 DOTS



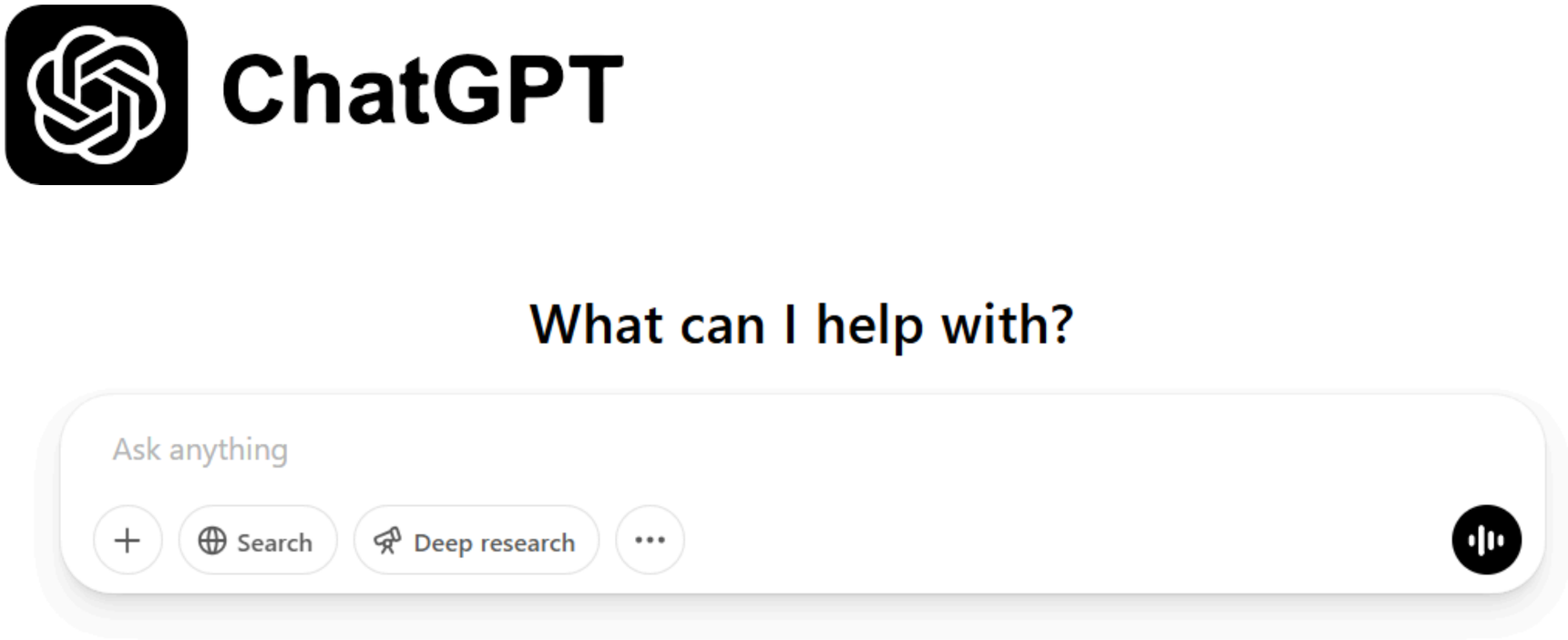
100 DOTS



1000 DOTS

>>> What is Multimodal Large Language Model?

Process multiple data types: text, images, audio, and more!



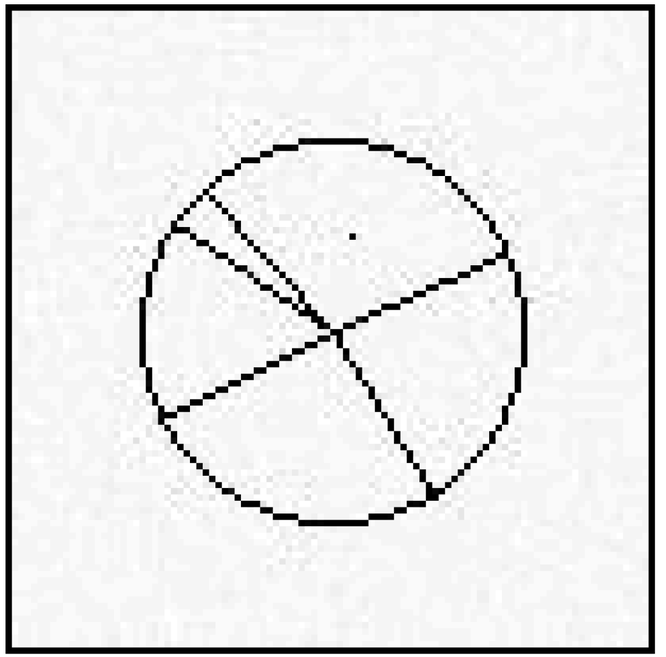
Stimuli Images

100 X 100 pixels

Binary aliased images

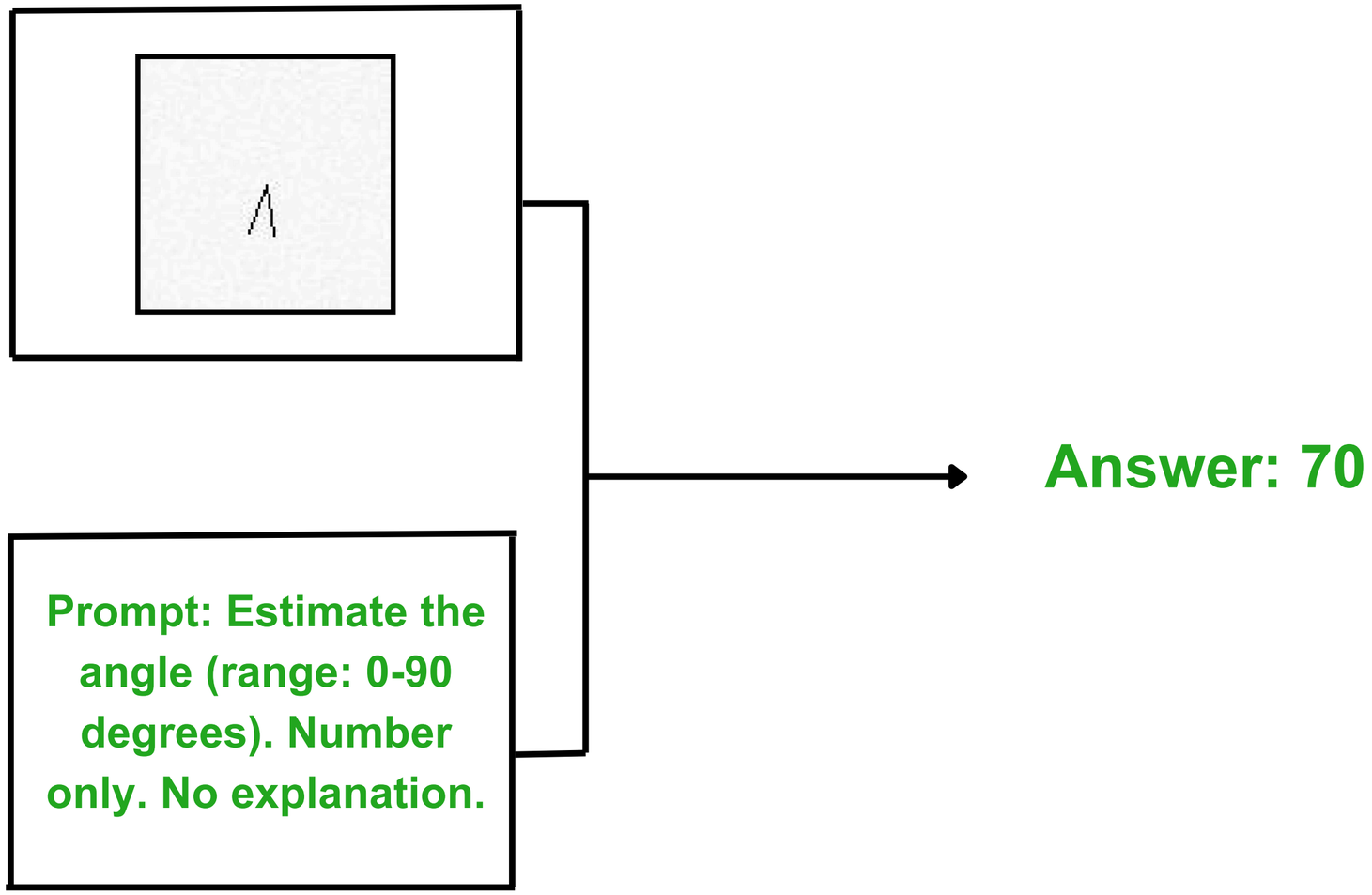
+ 5% noise to each image

Black background



»» Zero-shot prompting

A model performs a task without prior knowledge or examples.



Measure Task Accuracy

Midmean Logistic Absolute Error

A metric inspired by the Cleveland and McGill study.
The below is the formula used for our evaluation.

$$\mathrm{MLAE} = \frac{1}{N}\sum_{i=1}^{N} \log_2\left(\left|\mathrm{predicted}_i - \mathrm{true}_i\right| + 0.125\right)$$

The lower the MLAE value, the better the model's performance.

>>> Models

Pretrained models

Finetuned models

GPT- 4o

1.8T parameters

1.5 Flash

8T parameters

Pro Vision

Unknown

Llama 3.2 Vision Instruct

6B parameters

Llama 3.2 Vision Instruct

94.4M parameters

Our Finetuned models

Each dataset has unique label.

Fine-tuning

Train / Validation: 5,000 / 1,000 images per task

5 epochs

Evaluation

Test: 55 images per task

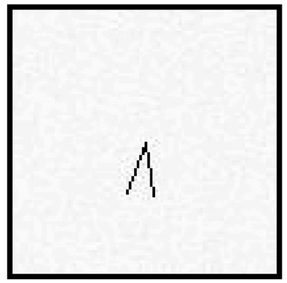
Each task has 3 runs

Human Baselines

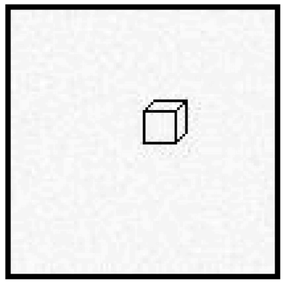
Heer and Bostock 2010

Haehn et al. 2018

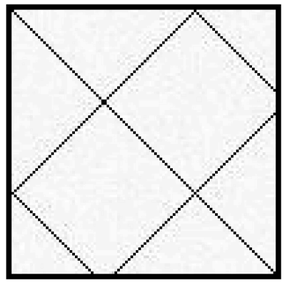
»» Results: E1: Elementary Perceptual Tasks



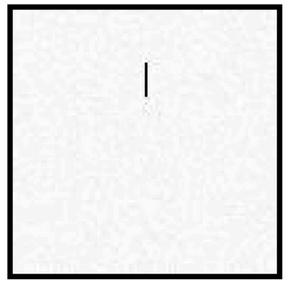
Angle



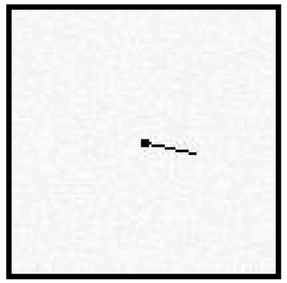
Volume



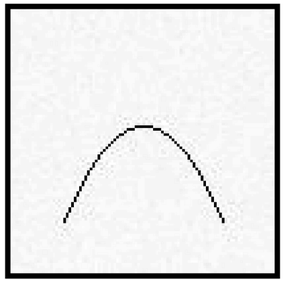
Shading



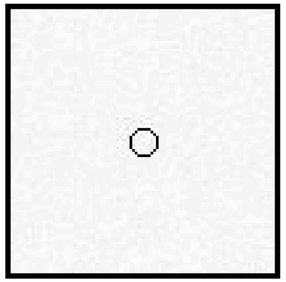
Length



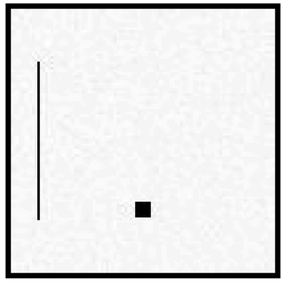
Direction



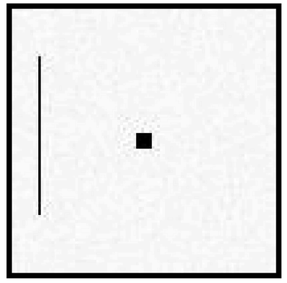
Curvature



Area



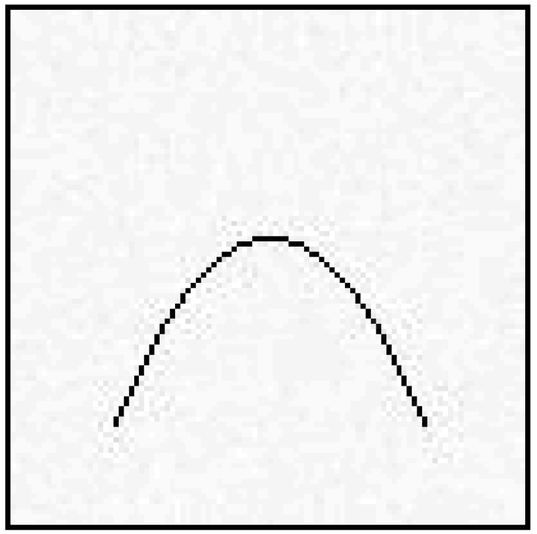
Position Common Scale



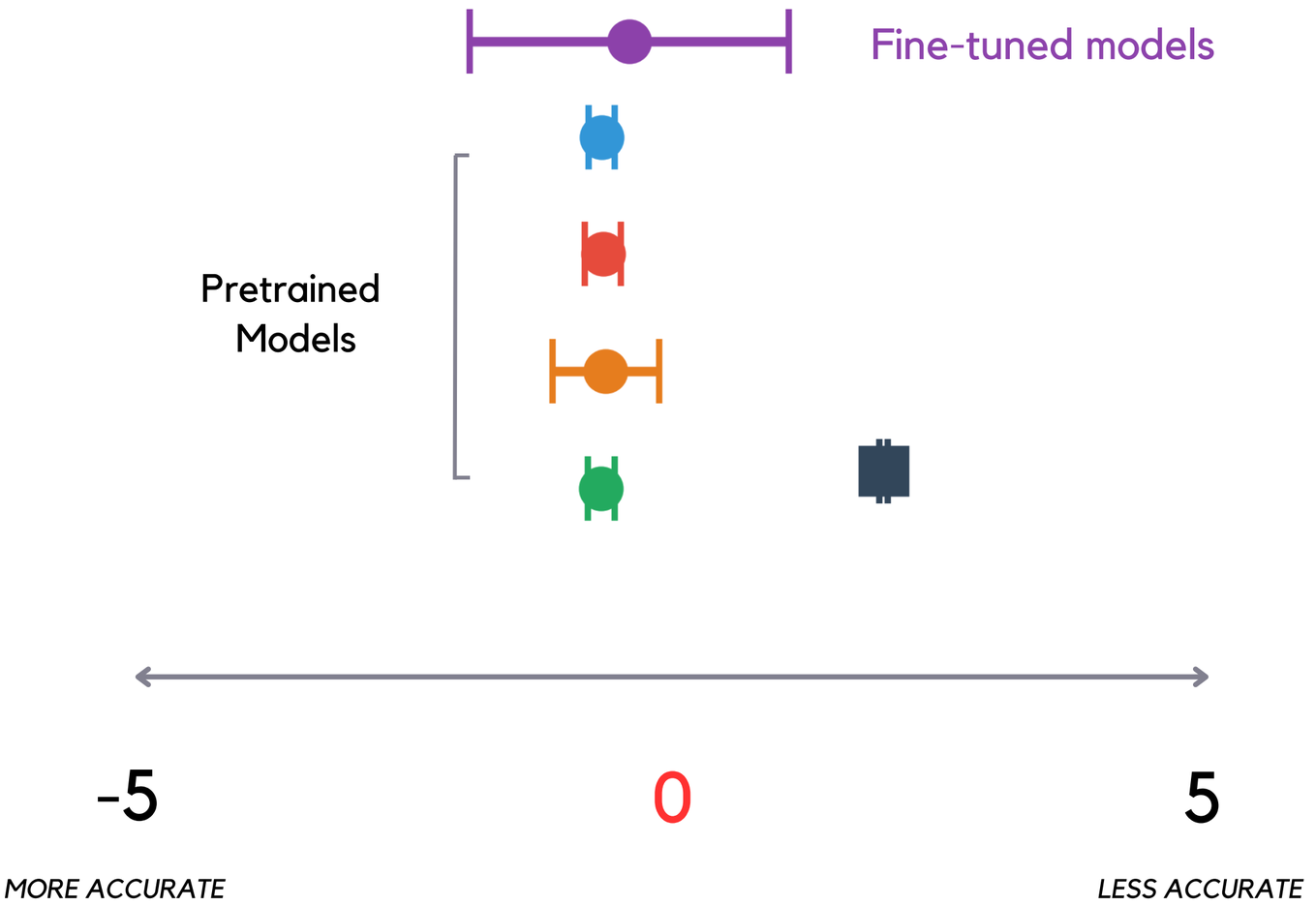
Position Non-Aligned Scale

Regress encoded numerical values from elementary tasks.

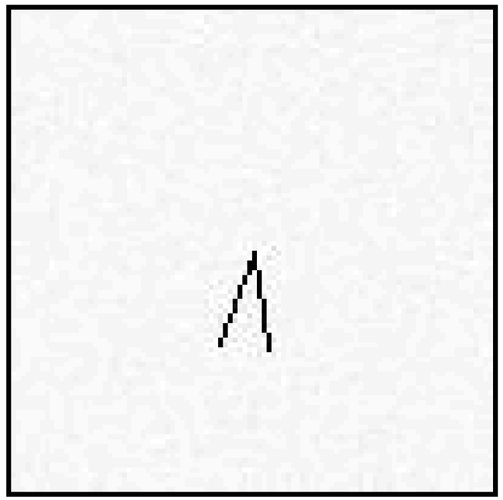
Results: E1: Curvature



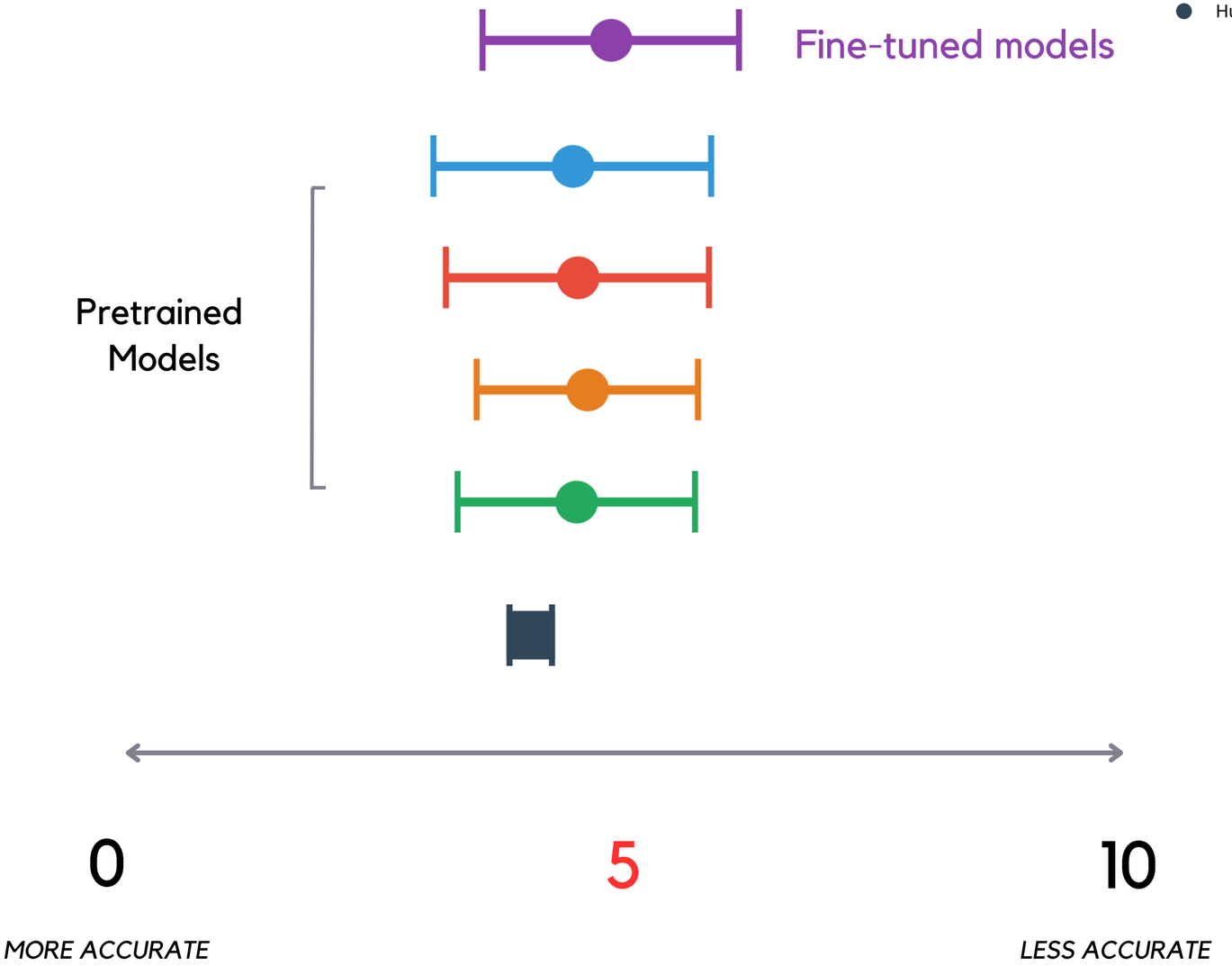
MLAE Error



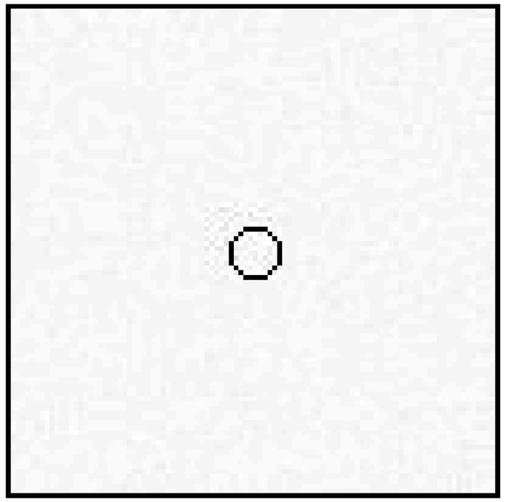
Results: E1: Angle



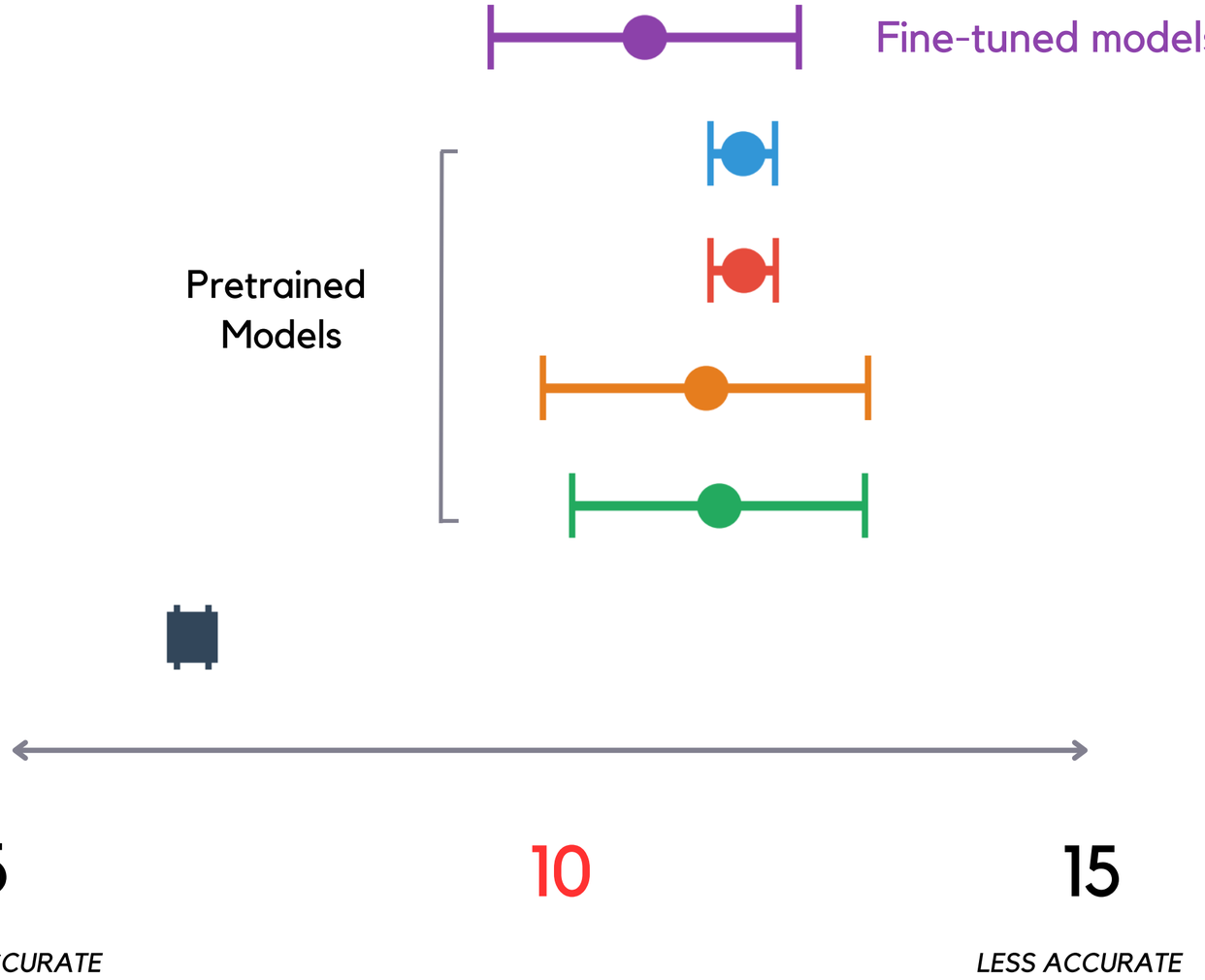
MLAE Error



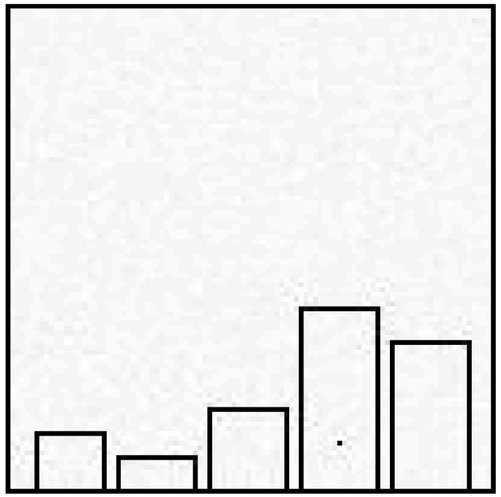
Results: E1: Area

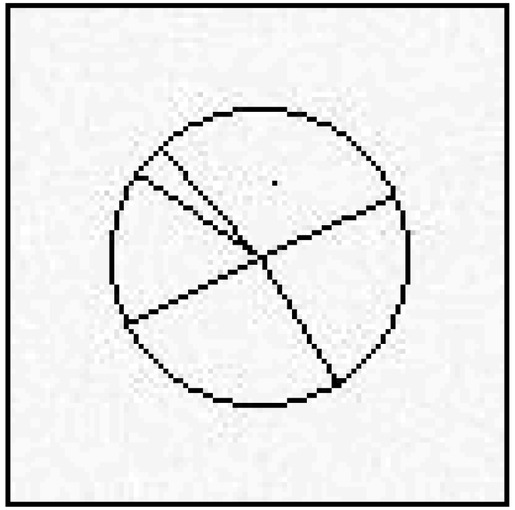


MLAE Error



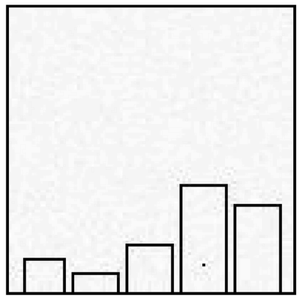
E2: Position-Angle



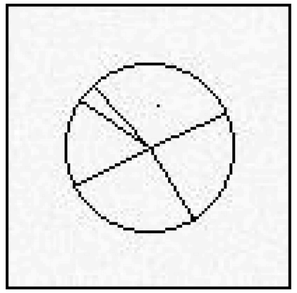


Identify the marked value, estimate the others compared to it.

Results: E2: Position-Angle



BAR



PIE

Pretrained Models

Fine-tuned models

Pretrained Models

Finetuned models

MLAE Error

-6

-2

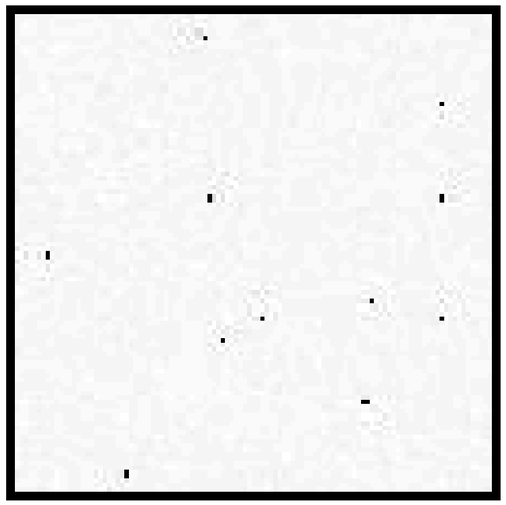
0

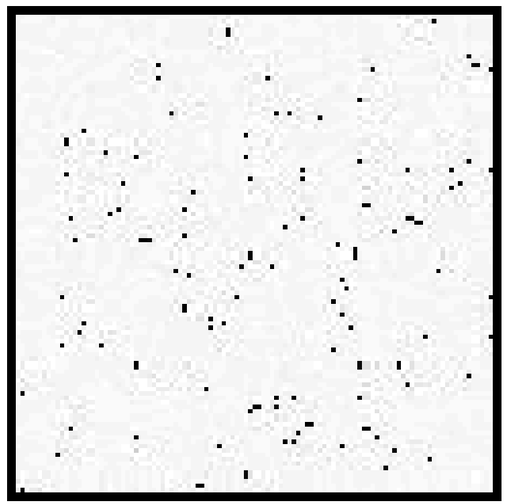
3

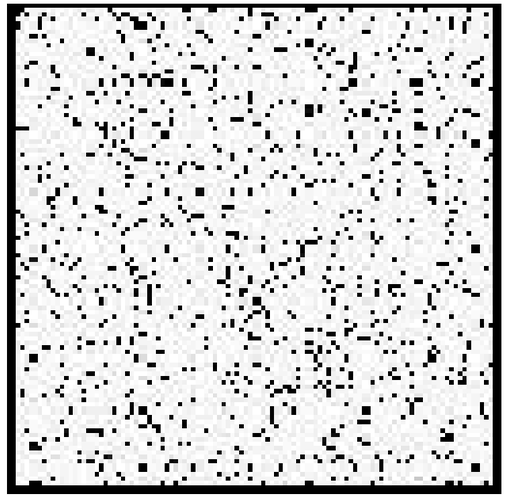
MORE ACCURATE

LESS ACCURATE

E5: Point-Cloud



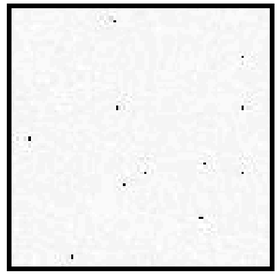




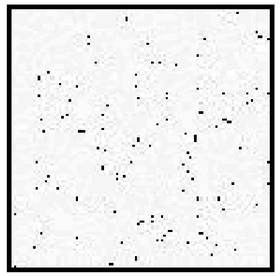
Estimate how many dots were added to the initial 10-100-1000 dots.

>>> Results - E5

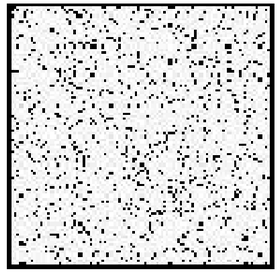
10 DOTS



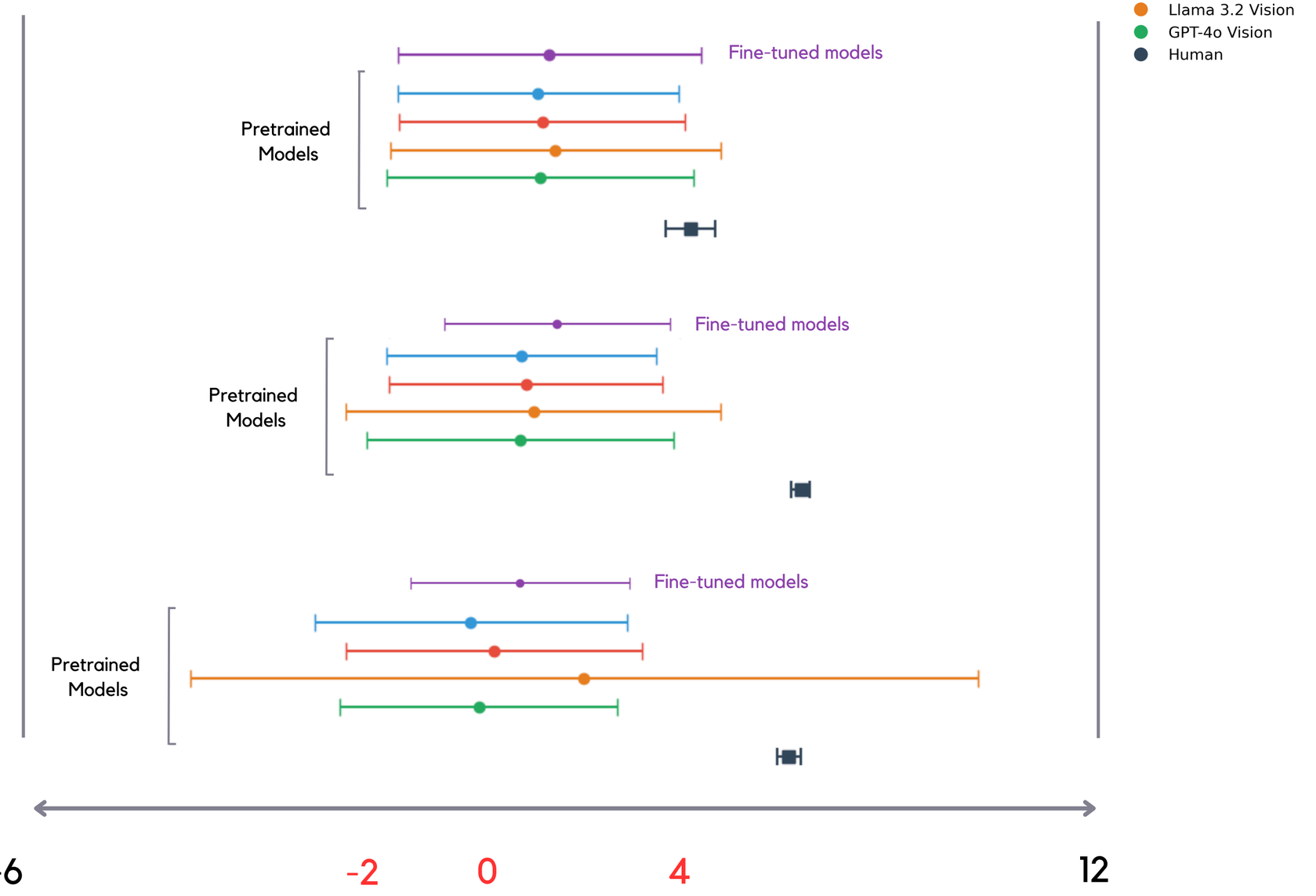
100 DOTS



1000 DOTS



MLAE
Error



Conclusion

While humans still excel at certain tasks, MLLMs outperformed them in the majority of experiments.

MLLMs have promising potential for graphical perception.

Thank you!

Connect with us on LinkedIn

Rami Huu Nguyen

Kenichi Maeda

Daniel Haehn

Mahsa Geshvadi

View our code & results