NTSEBench: Cognitive Reasoning Benchmark for Vision Language Models

Pranshu Pandya†, Vatsal Gupta†,
Agney S Talwarr , Tushar Kataria , Dan Roth, Vivek Gupta*

†Equal Contribution, *Corresponding Author

Indian Institute of Technology-Guwahati,
University of Utah, University of Pennsylvania, Arizona State University

MOTIVATION

NAACL 2025

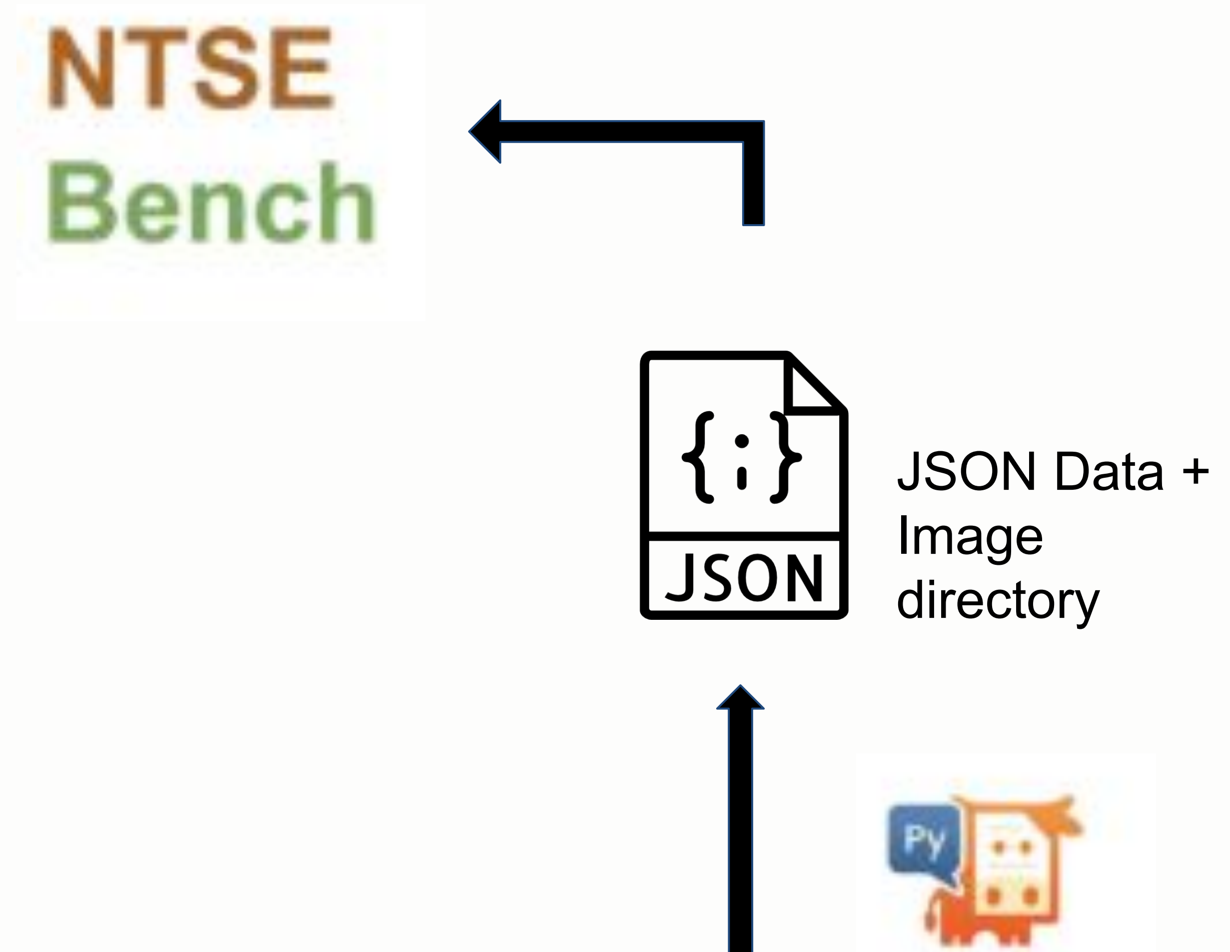
- The Need for **Cognitive Reasoning** in the AI Landscape
- Advancing AI Toward **Human-Like Problem Solving**
- **NTSE Exam-Based Dataset**
    - **The National Talent Search Examination (NTSE):** Taken by 1 million students annually in India.
    - **Evaluates:** Critical thinking, analytical reasoning, general knowledge, and mathematical aptitude.
- **AI Training:** Use NTSE-style questions to test Multimodal models on deep cognitive reasoning, drawing a parallel to human problem-solving skills.

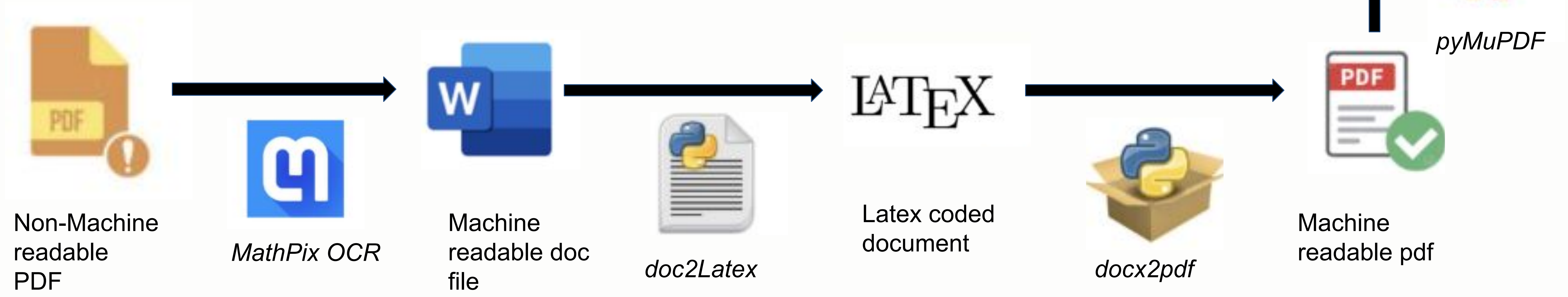
INTRODUCTION

- NTSEBench is a **novel benchmark** designed to evaluate cognitive reasoning in vision–language models.
- NTSEBench aims to tackle **gaps in current benchmarks**
- The dataset targets **advanced skills**—such as pattern recognition, logical deduction, and spatial reasoning—that go beyond rote memorization.

DATASET CURATION - PIPELINE

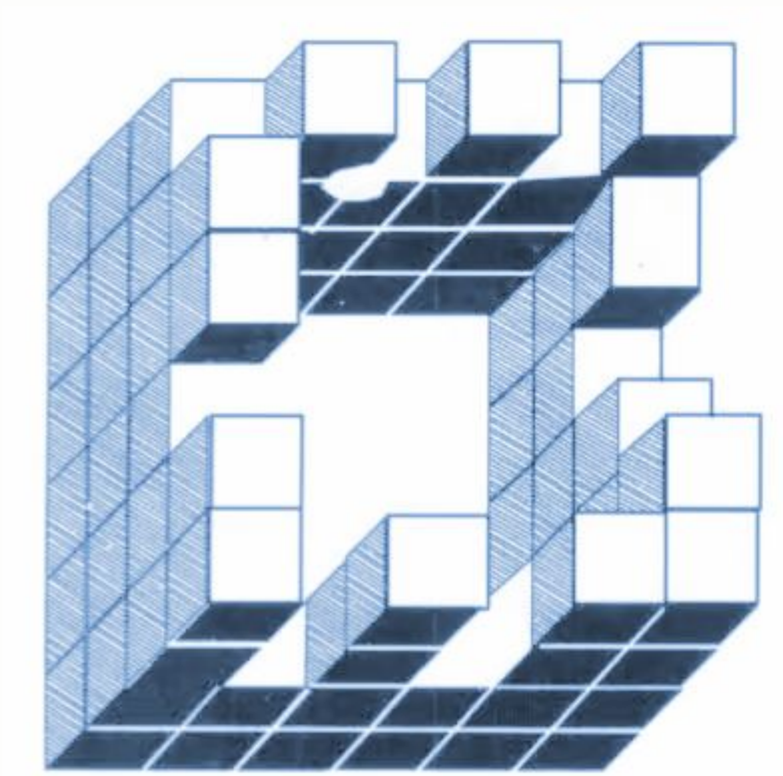
- **Manual curation & extraction pipeline** to convert non-machine-readable NTSE papers into structured, machine-readable PDFs.
- **High-quality dataset** that ensures accessibility and effective analysis.
- **Multiple data sources** are used to curate a comprehensive and reliable dataset.





EXAMPLES OF QUESTIONS

Cube and Dice Type Category Question



Question Figure

How many cubes are there in the 3D model Question Figure?

Embedded Figure Category Question

Question Figure

(1)

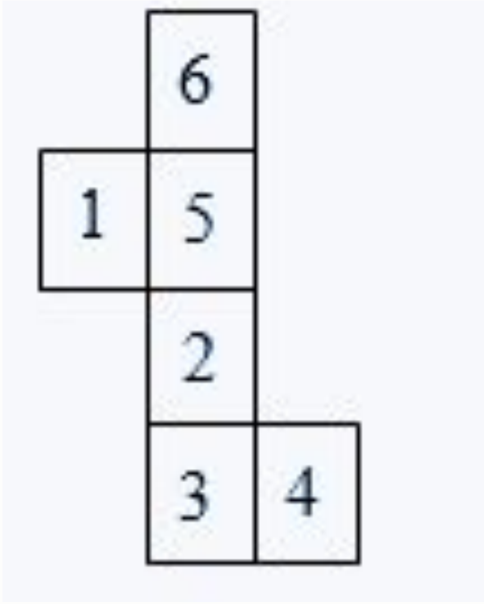
(2)

(3)

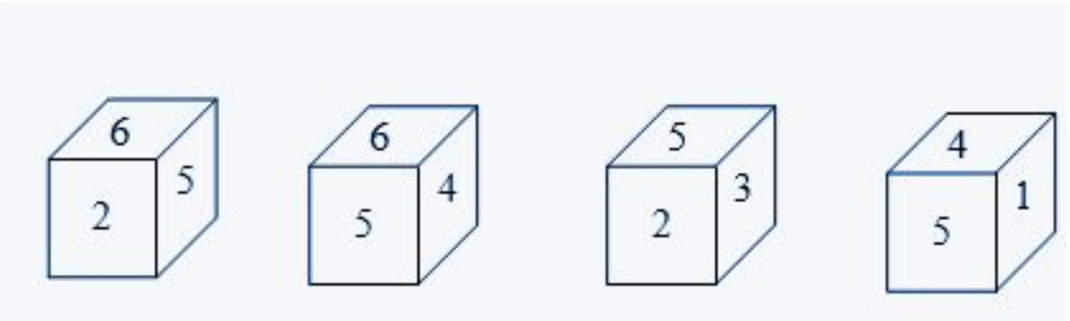
(4)

In which figure is the Question Figure embedded?

Paper Folding and Cutting Category Question



Question Figure



Which cube does the Question Figure yield on folding?

DATASET STATISTICS

NAACL 2025

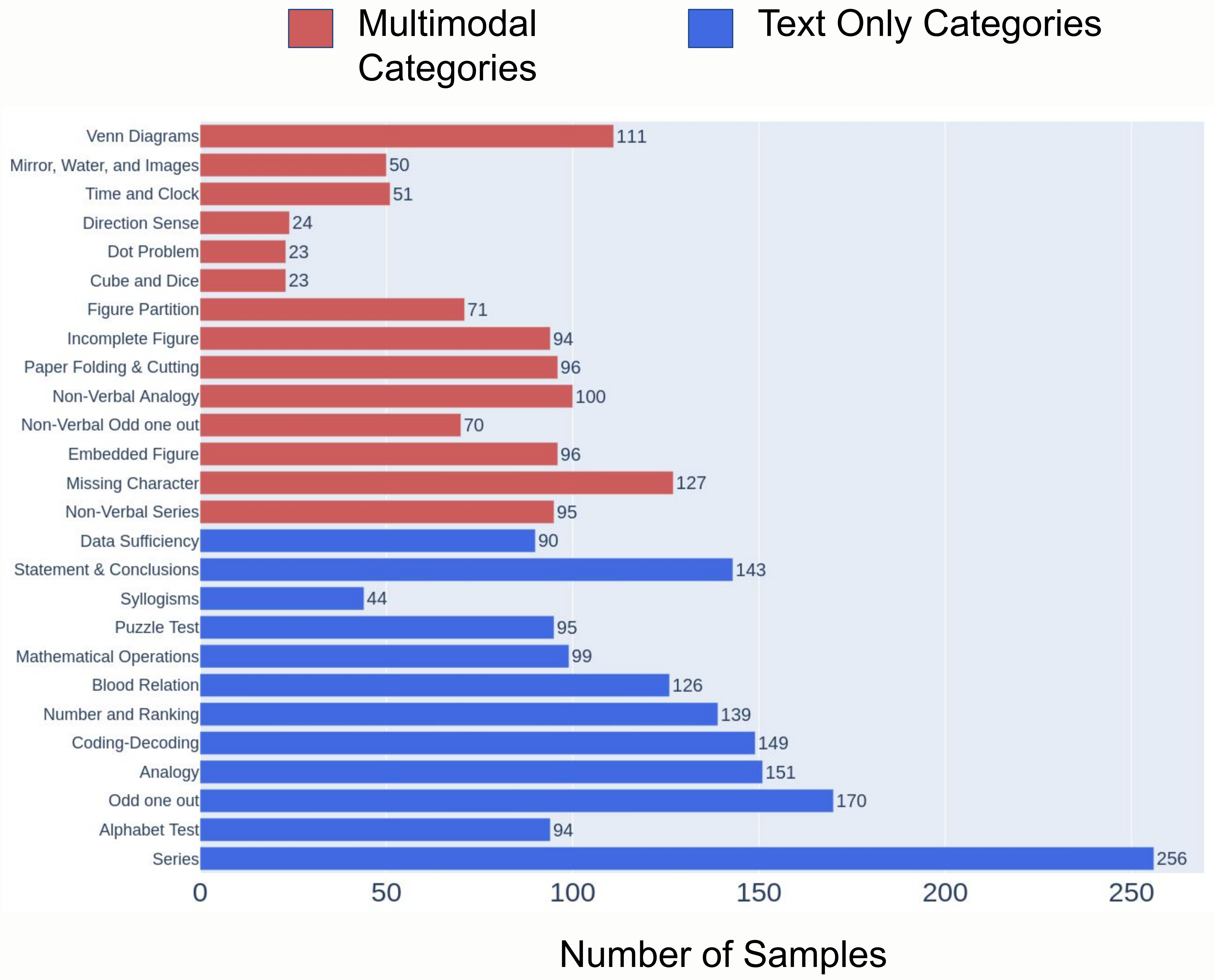
- **Rich Multimodal Dataset:** Contains **2,728 multiple-choice questions,** each paired with **4,642 images,** forming comprehensive question–option–solution triplets**.**
- **Question Types:** We propose both **Text-only questions** as well as **Multimodal questions** incorporating images requiring deep logical analysis and reasoning

Question	Options	Solutions	No. of Samples
❌	❌	❌	1199
❌	❌	✅	381
❌	✅	❌	70
❌	✅	✅	18
✅	❌	❌	330
✅	❌	✅	126
✅	✅	❌	403
✅	✅	✅	201

*A **tick (✅) mark** indicates that the question, option, or solution **includes an image***

DATASET STATISTICS

- **Expert Categorization:** Data is meticulously classified by human experts into **26 distinct categories** .
- **Cognitive Dimensions**: Introduces **8 cognitive dimensions** to evaluate diverse aspects of multimodal reasoning.



CONTEXT TYPES

- **Four context types** designed to handle different question types and input modalities
- **Models vary** in their ability to handle interleaving or multiple images
- **Ensures fair assessment** of model reasoning across different capabilities
- For **Text-only questions, standard QA** is employed.
- For **Multimodal questions, 3 new types of strategies** are proposed to cater to different models.

System Prompt

Question Text: In the number series given below, one number is missing. $ 12,15,27,42,69,111 $,_

Option 1: 164 **Option 2:** 174 **Option 3:** 180 **Option 4:** 160

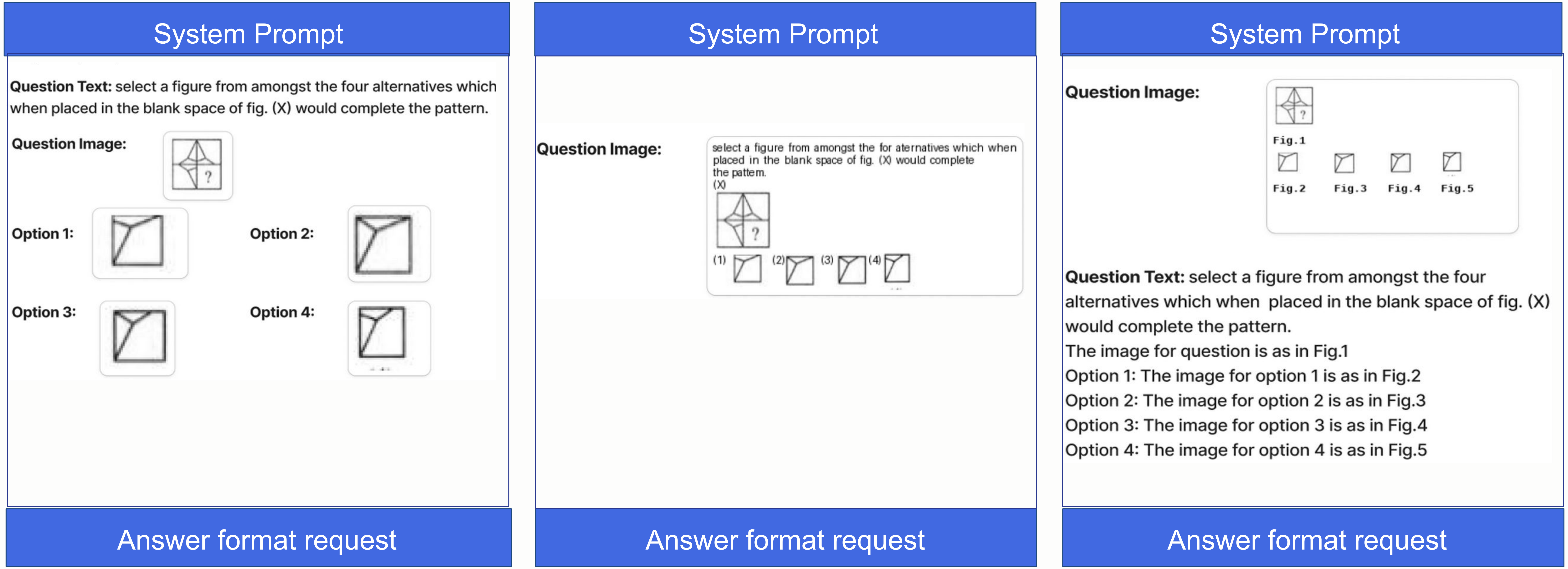
Answer format requested

Standard QA: Used for Text-only questions

CONTEXT TYPES - Multimodal categories

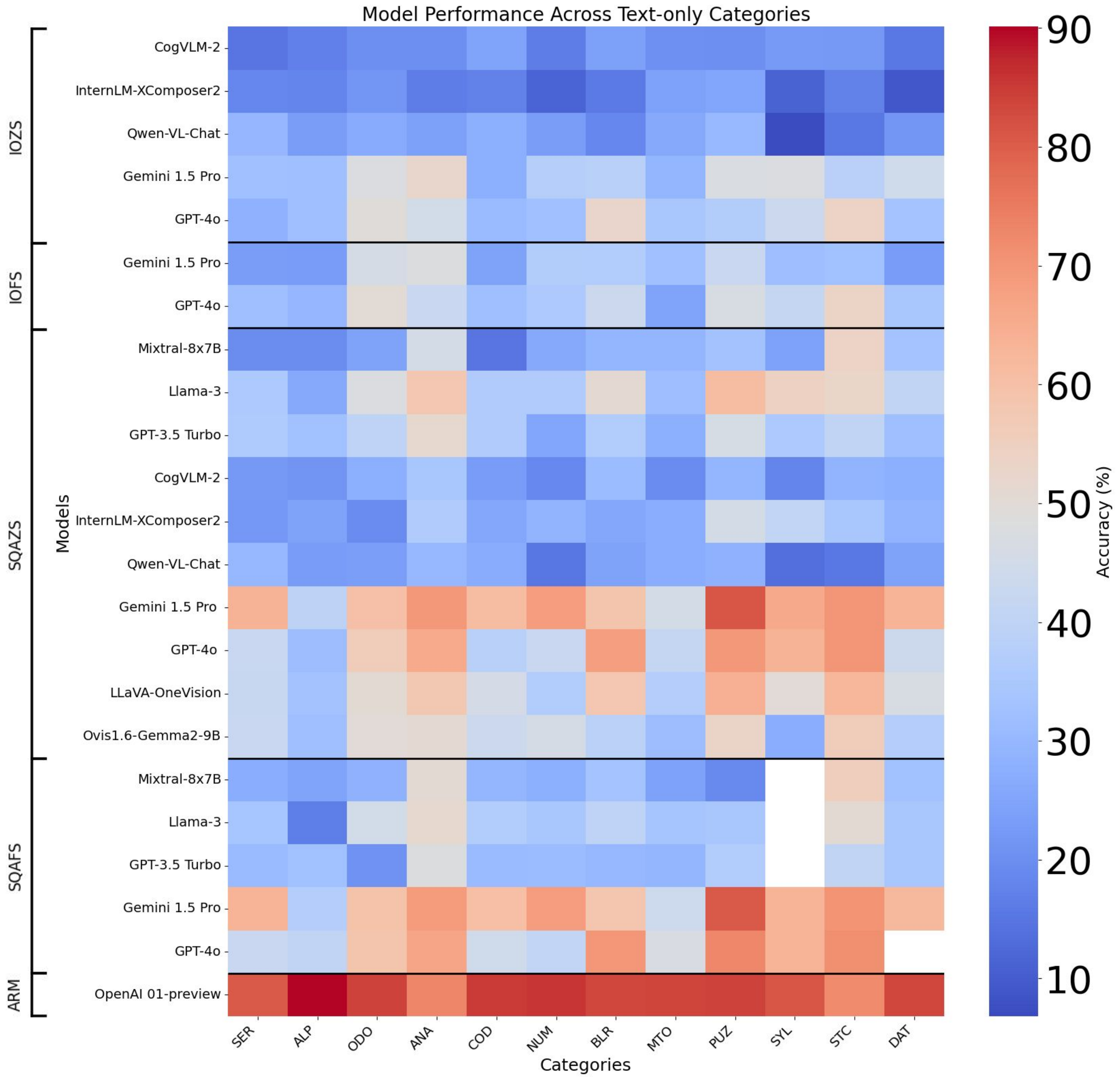
NAACL 2025

- **Interleaved**: Images inserted into the text prompt, which preserves image **sequence and context.**
- **Standard VQA : All images combined** into one and labeled (e.g., Figure 1, 2, 3) to reference in text.
- **Image-Only:** Essentially a **snapshot of the question** as in PDF



Interleaved | **Image-Only** | **Standard VQA**

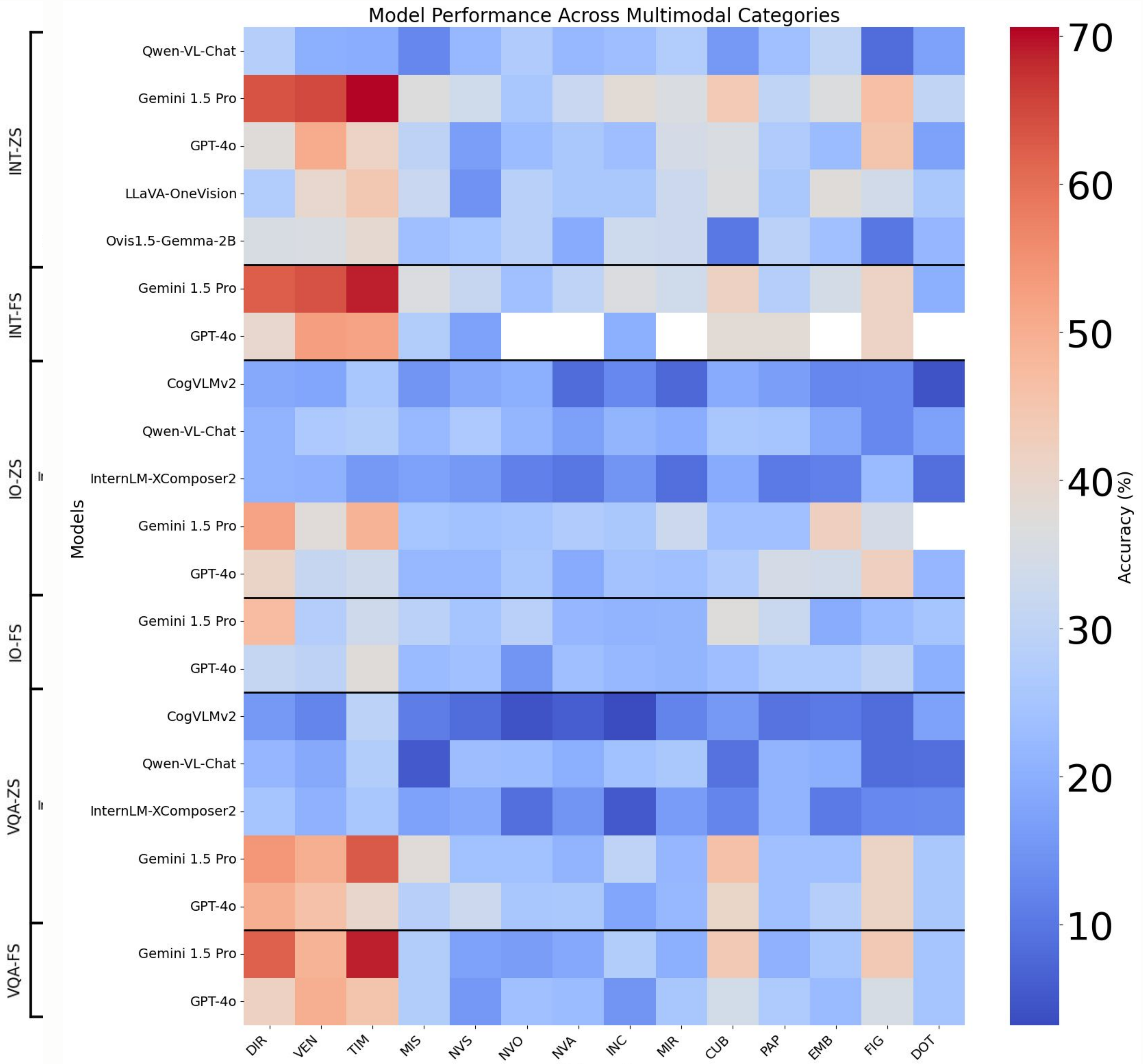
RESULTS - Text-only Categories Performance



NAACL 2025

IO: Image Only
SQA: Standard QA
ZS: Zero Shot
FS: Few Shot

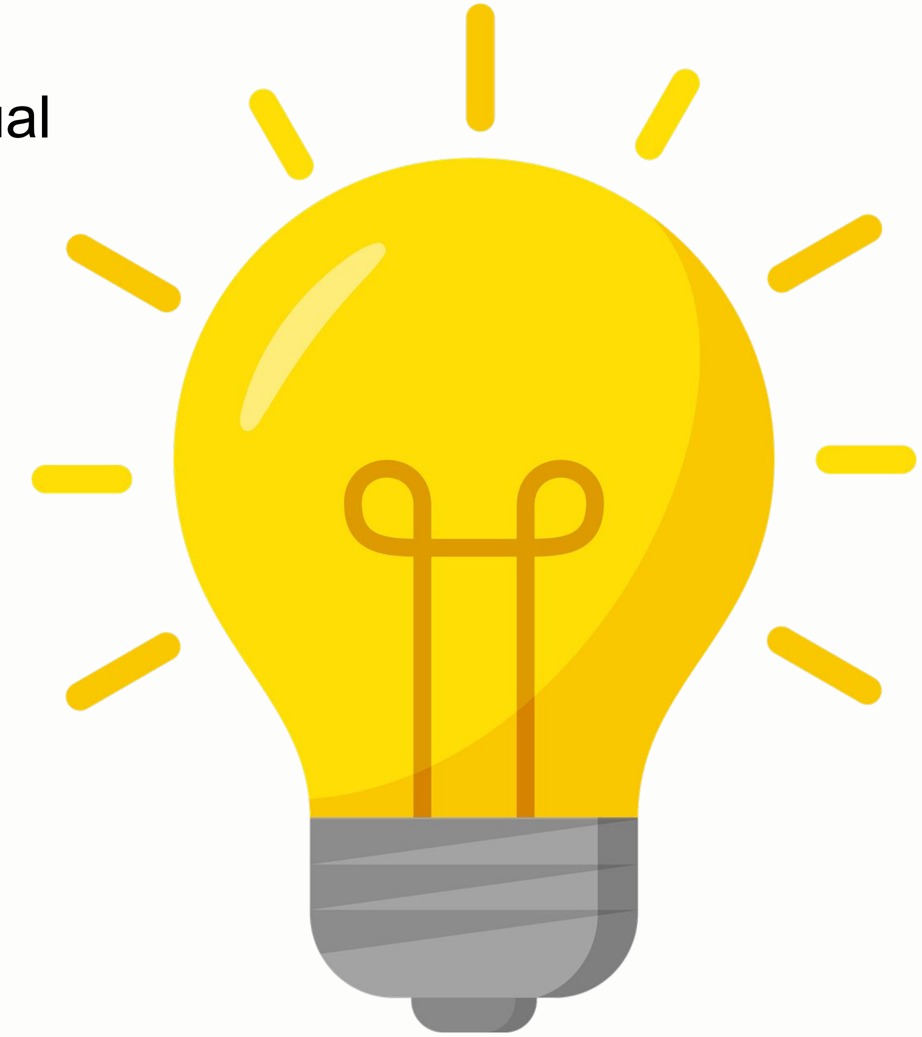
RESULTS - Multimodal Categories Performance



INT: Interleaving
IO: Image Only
VQA: Visual QA
ZS: Zero Shot
FS: Few Shot

RESULTS—KEY INSIGHTS

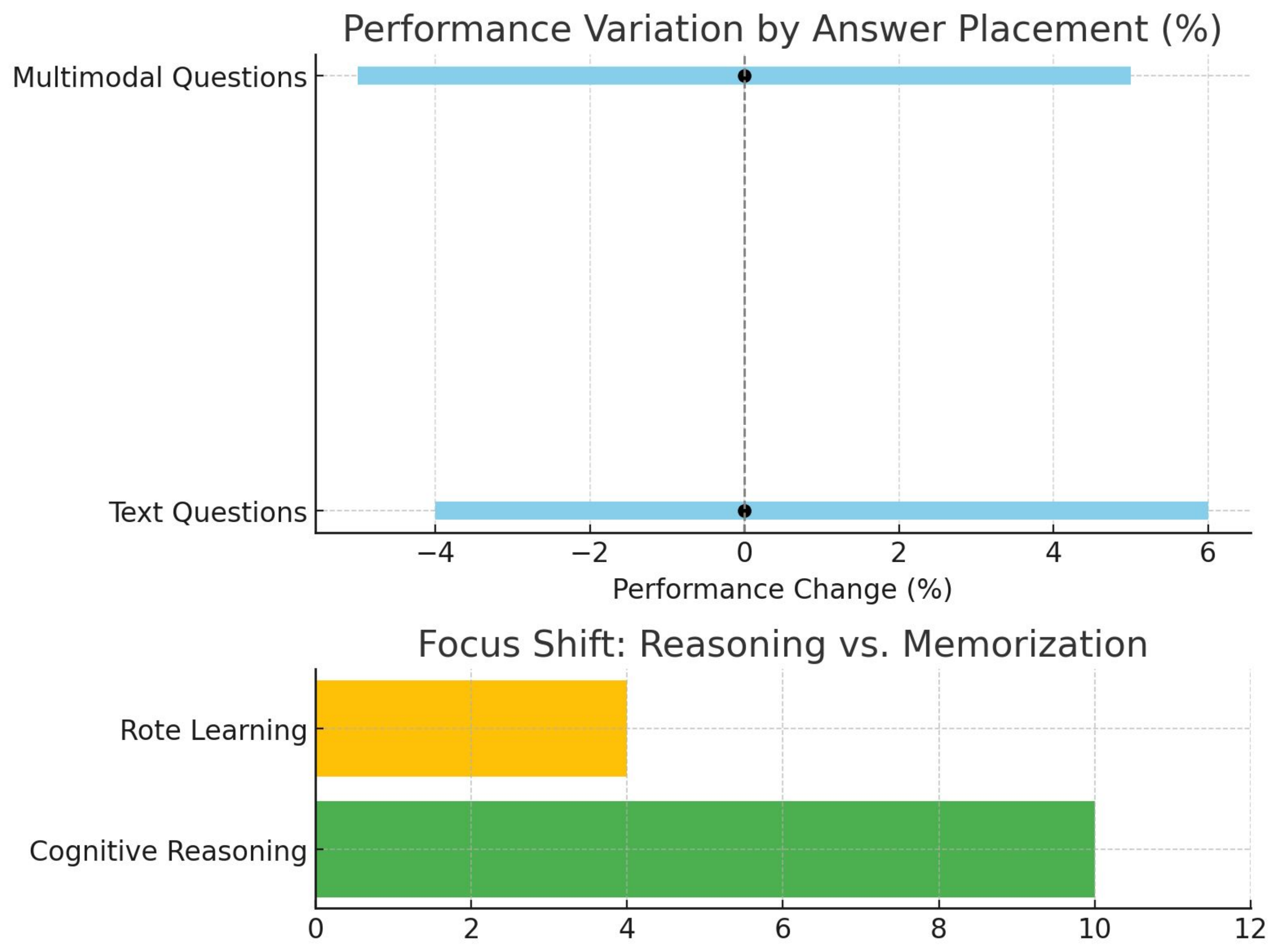
- **Interleaving as a superior context mechanism**
  - **Showcased by ovis-1.6:** Outperforms larger open-source models due to better contextual representation.
- **Some categories** are relatively easier
  - Possibly due to high exposure in model pretraining datasets for **DIR, VEN, TIM.**
- **o1-preview's Text-Only Brilliance**
  - Showcases near-human performance on text-only tasks
  - Reinforces the importance of **"thinking before answering" which is crucial in NTSEBench**



OPTION ABALATION-BIAS EXPERIMENTATION

NAACL 2025

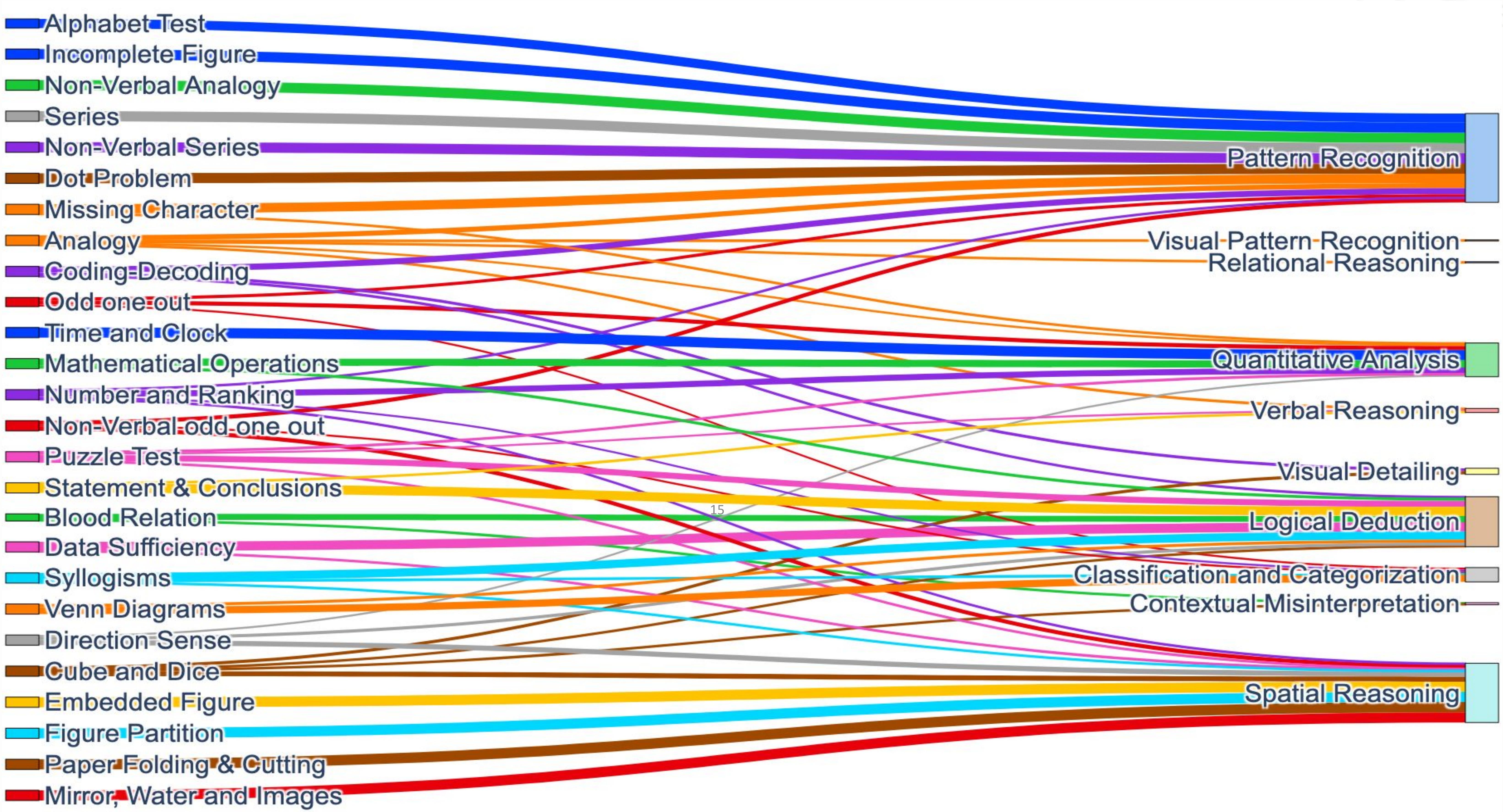
- **Robust analysis** accounting for cognitive reasoning rather than reliance on memorization.
- **Evaluated Gemini 1.5 Pro** to determine how answer placement affects model performance.
- **Observed performance shifts**
  - **-4% to +6%** for text-only questions.
  - **-5% to +5%** for multimodal questions.



EXTENSIVE ERROR ANALYSIS

- **Analysed 260 questions 10 per category** on Gemini 1.5 Pro, revealing reasoning patterns.
- **Categorized errors into the 8 cognitive dimensions proposed.**
- **VLMs struggle with logical deductions from limited visuals**, especially in pattern recognition, spatial manipulation, and shape recognition.
- Error distribution highlights **model strengths** and **scope for improvement**.

EXTENSIVE ERROR ANALYSIS



TAKEAWAYS

- **Dataset Expansion:**
  - **Data-Augmentaion** techniques can be looked into to add more data for model training
  - **Hardness Categorization**:Deeper analysis to categoirze hardness of questions can be done.
  - **Multilingual Capabilities:** Extend dataset to regional languages leveraging NTSE’s multilingual reach**.**
- **Model & Method Insights:** Our **framework allows for comprehensive evaluation** across models and strategies, and can be extended to other datasets.
- **Challenging Benchmark:**
  - **Human accuracy** (80%) **> SOTA** models (**62%** Text-only questions and **42%** Multimodal questions)
  - **NTSEBench proves itself to be a novel and important benchmark** that can improve models significantly and exposes model limitations in diverse categories.

DISCUSS

- We would be happy to discuss and address any questions.
- Please reach out to me at g.vatsal@alumni.iitg.ac.in ;I would be happy to collaborate.

GITHUB

https://github.com/NTSEBench/NTSEBench

PAPER

https://ar.xivorg/abs/2407.10380

WEBSITE

https://ntsebench.github.io/