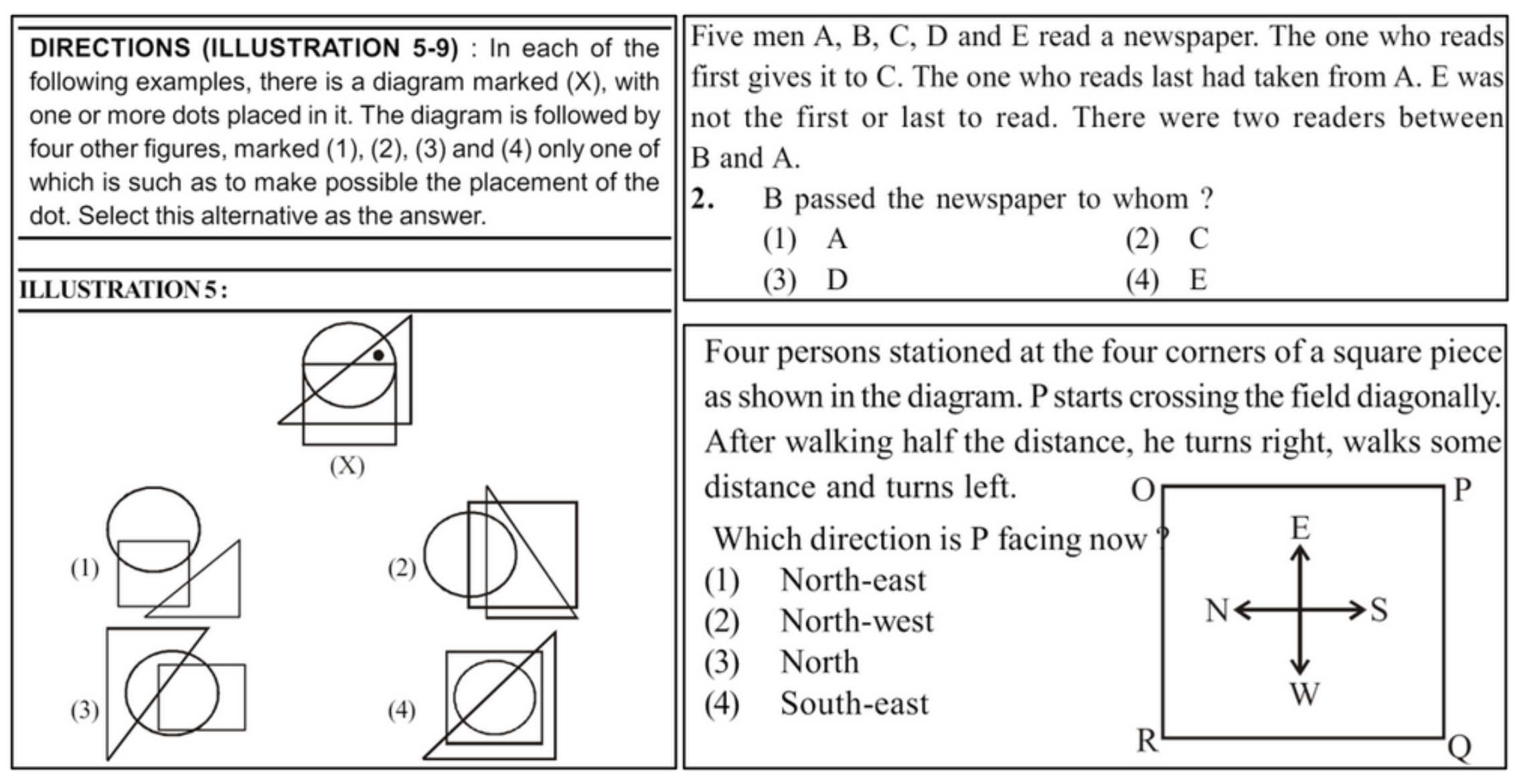
NTSEBench: Cognitive Reasoning Benchmark for Vision Language Models

Pranshu Pandya[1†], Vatsal Gupta[1†], Agney S. Talwarr[1],
Tushar Kataria[2], Dan Roth[3], Vivek Gupta[4*]
[†]Equal Contribution, [*]Corresponding Author
[1]IIT Guwahati, [2]University of Utah, [3]University of Pennsylvania, [4]Arizona State University

NAACL 2025

1. NTSEBench

Benchmark to evaluate the **cognitive reasoning** capabilities of SOTA **LLMs** and **VLMs**.



DIRECTIONS (ILLUSTRATION 5-9) : In each of the following examples, there is a diagram marked (X), with one or more dots placed in it. The diagram is followed by four other figures, marked (1), (2), (3) and (4) only one of which is such as to make possible the placement of the dot. Select this alternative as the answer.

ILLUSTRATION 5 :

Five men A, B, C, D and E read a newspaper. The one who reads first gives it to C. The one who reads last had taken from A. E was not the first or last to read. There were two readers between B and A.

2. B passed the newspaper to whom ?
(1) A (2) C
(3) D (4) E

Four persons stationed at the four corners of a square piece as shown in the diagram. P starts crossing the field diagonally. After walking half the distance, he turns right, walks some distance and turns left.

Which direction is P facing now ?
(1) North-east
(2) North-west
(3) North
(4) South-east

2. Dataset Highlights

- Questions are sampled from the National Talent Search Examination (NTSE), India.
- Mental Ability Test questions: Tests **reasoning rather than rote learning**
- 2728 MCQs with **4642 images**.
- Question covering **26 distinct categories across 8 cognitive dimensions**.
- Contains both **multimodal** (text-images) and **text-only** questions.

Text Only		**Vision + Text**	
Categories	**# Samples**	**Categories**	**# Samples**
Series	256	Non-Verbal Series	95
Alphabet Test	94	Missing Character	127
Odd one out	170	Embedded Figure	96
Analogy	151	Non-Verbal odd one out	70
Coding-Decoding	149	Non-Verbal Analogy	100
Number and Ranking	139	Paper Folding & Cutting	96
Blood Relation	126	Incomplete Figure	94
Mathematical Operations	99	Figure Partition	71
Puzzle Test	95	Cube and Dice	89
Syllogisms	44	Dot problem	23
Statement & Conclusions	104	Direction Sense	96
Data Sufficiency	90	Time and Clock	51
		Mirror, Water and Images	92
		Venn diagrams	111

Table: **NTSEBench categories count**

Question	Options	Solutions	# Samples
✗	✗	✗	1199
✗	✗	✓	381
✗	✓	✗	70
✗	✓	✓	18
✓	✗	✗	330
✓	✗	✓	126
✓	✓	✗	403
✓	✓	✓	201

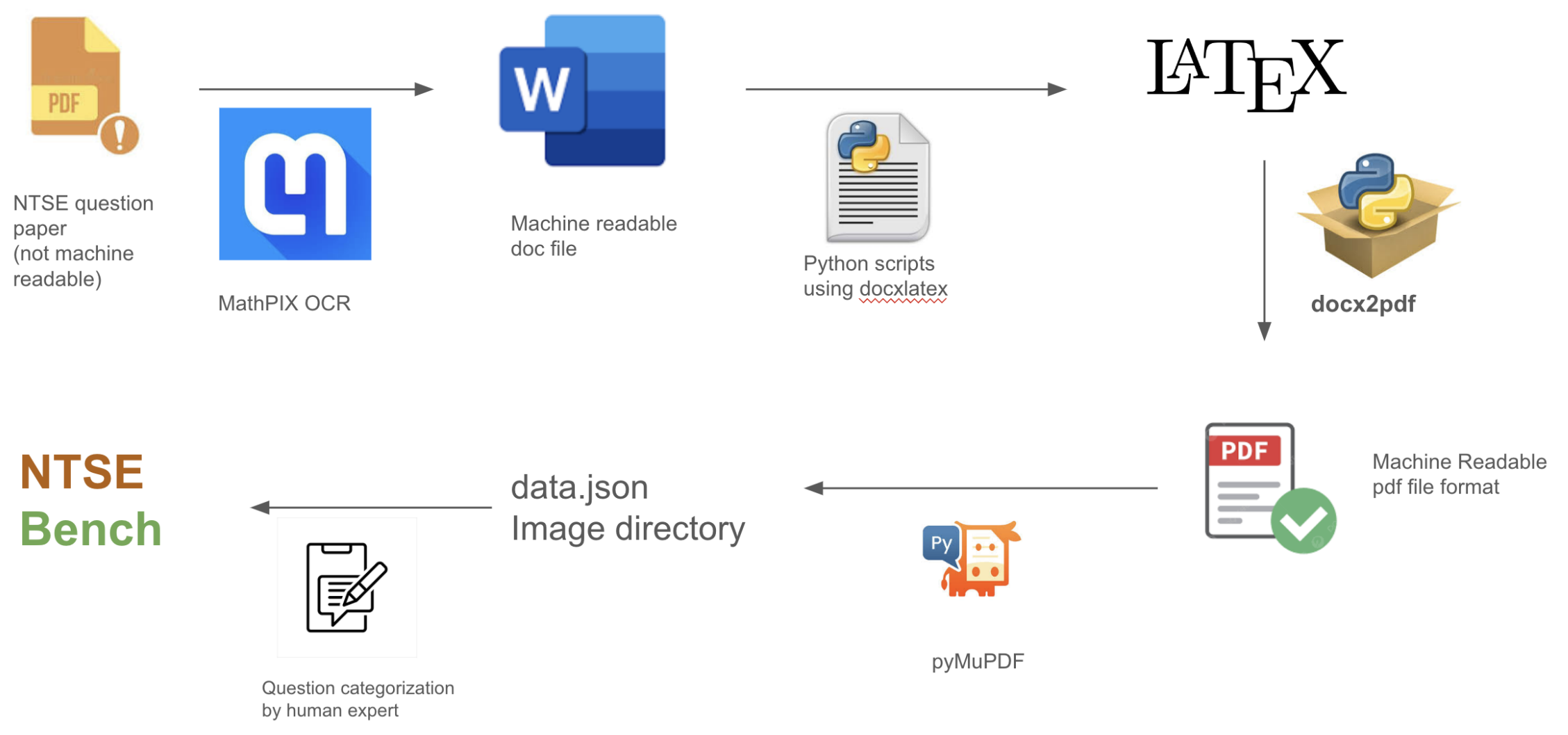
Table: **Modality Variations Question Count**

Pattern Recognition	Logical Deduction
Spatial Reasoning	Relational Reasoning
Quantitative Analysis	Classification
Contextual Interpretation	Verbal Reasoning

Table: **Cognitive Dimensions in NTSEBench.**

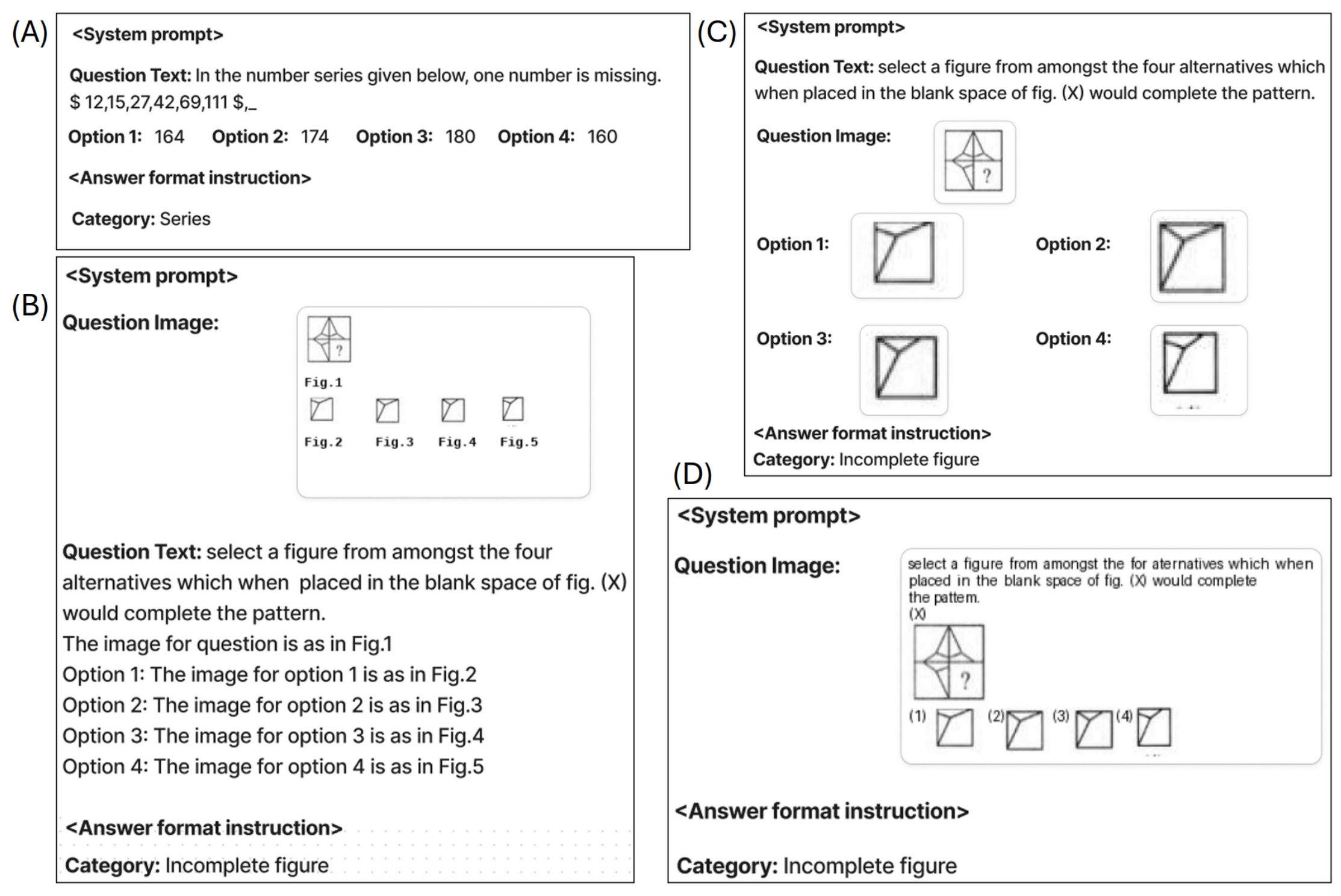
3. Methodology

Dataset Construction:



NTSE question paper (not machine readable)
MathPIX OCR
Machine readable doc file
Python scripts using docxlatex
LaTeX
docx2pdf
Machine Readable pdf file format
pyMuPDF
data.json
Image directory
Question categorization by human expert
NTSE Bench

Modeling strategies



(A) <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 instruction>

Category: Series

(B) <System prompt>

Question Image:

Fig.1
Fig.2
Fig.3
Fig.4
Fig.5

Question Text: select a figure from amongst the four alternatives which when placed in the blank space of fig. (X) would complete the pattern.
The image for question is as in Fig.1
Option 1: The image for option 1 is as in Fig.2
Option 2: The image for option 2 is as in Fig.3
Option 3: The image for option 3 is as in Fig.4
Option 4: The image for option 4 is as in Fig.5

<Answer format instruction>

Category: Incomplete figure

(C) <System prompt>

Question Text: select a figure from amongst the four alternatives which when placed in the blank space of fig. (X) would complete the pattern.

Question Image:

Option 1:
Option 2:
Option 3:
Option 4:

<Answer format instruction>

Category: Incomplete figure

(D) <System prompt>

Question Image:

select a figure from amongst the for aternatives which when placed in the blank space of fig. (X) would complete the pattem.
(X)

<Answer format instruction>

Category: Incomplete figure

Figure: **Examples Showing Input to Different Proposed modelling strategies**.(A) Text Only Standard QA strategy(B) Standard VQA (C) Interleaved Strategy (D) Image Only.

4. Results and Observations

Text based questions

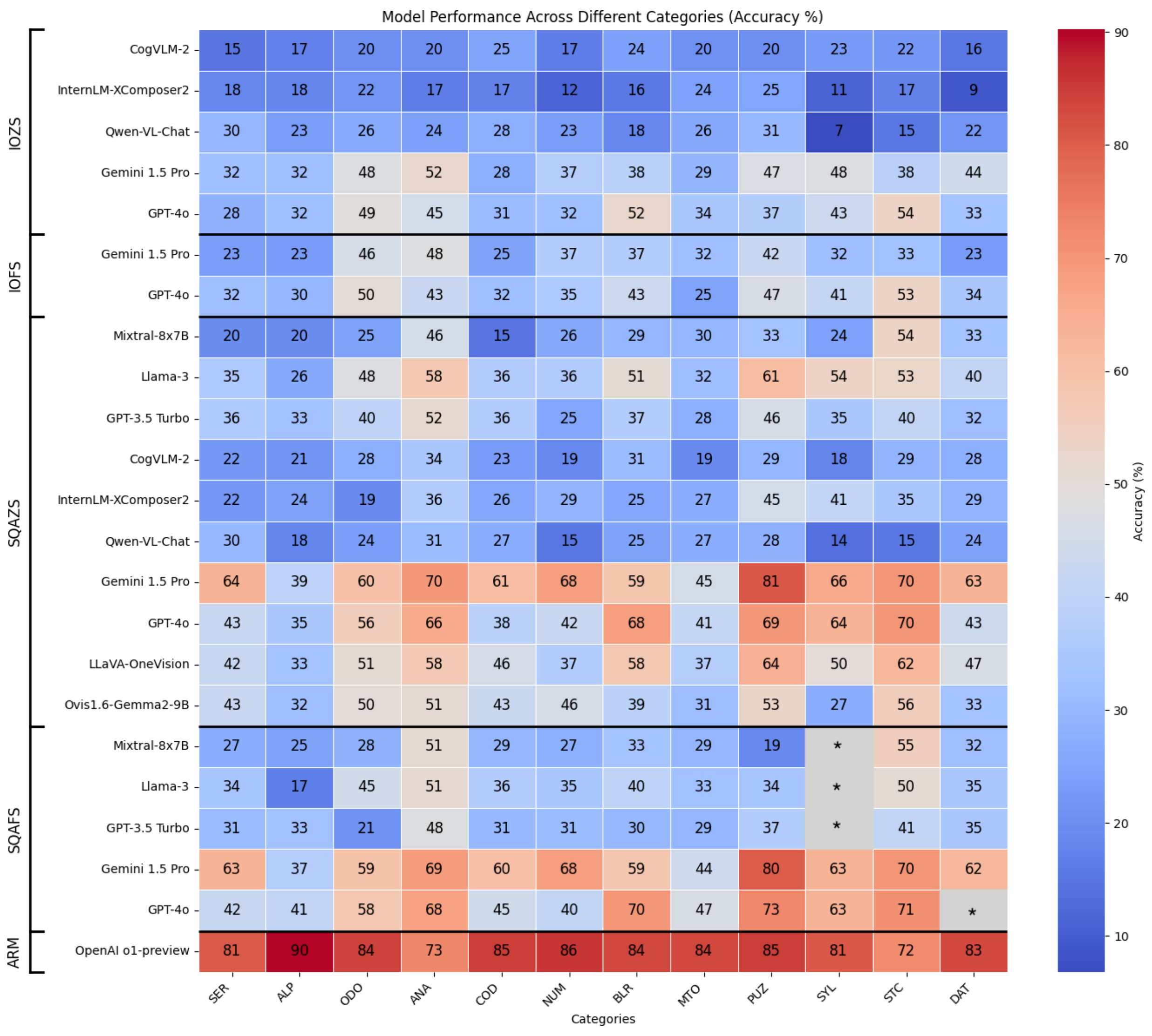


Figure: IO: Image-Only, SQA: Standard QA, ARM: Advanced Reasoning Model, ZS: Zero-Shot, FS: Few-Shot

Text+Image questions

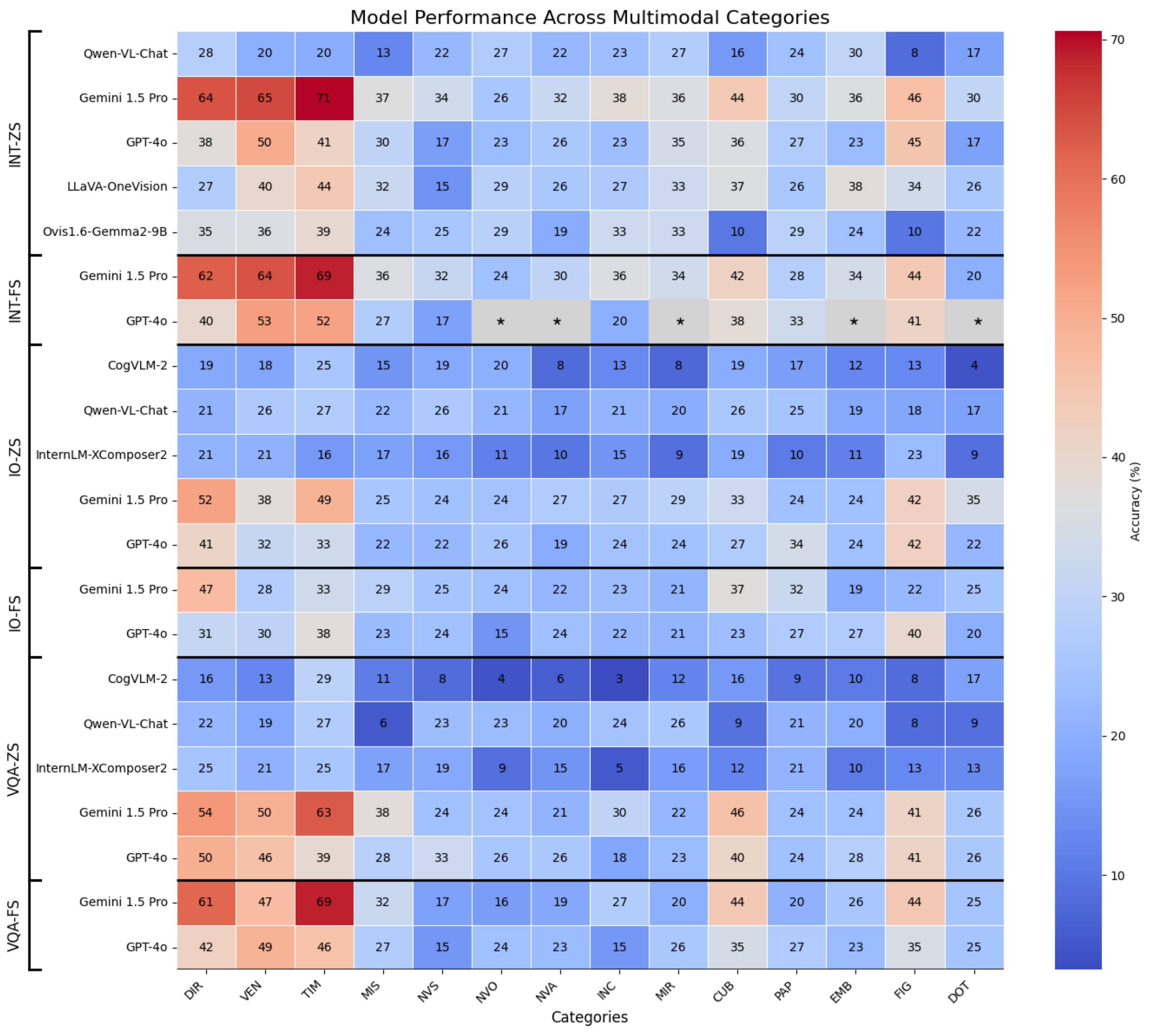


Figure: INT: Interleaved, IO: Image-Only, VQA: Visual QA, ZS: Zero-Shot, FS: Few-Shot

Key Observations:

- Proprietary models outperform open-source models.
- Interleaving text and images performs better than Standard VQA and Image Only.
- Multimodal reasoning is significantly harder.
- Human accuracy exceeds 80%, far surpassing the top proprietary model (62% text, 42% visual).

Error Analysis:

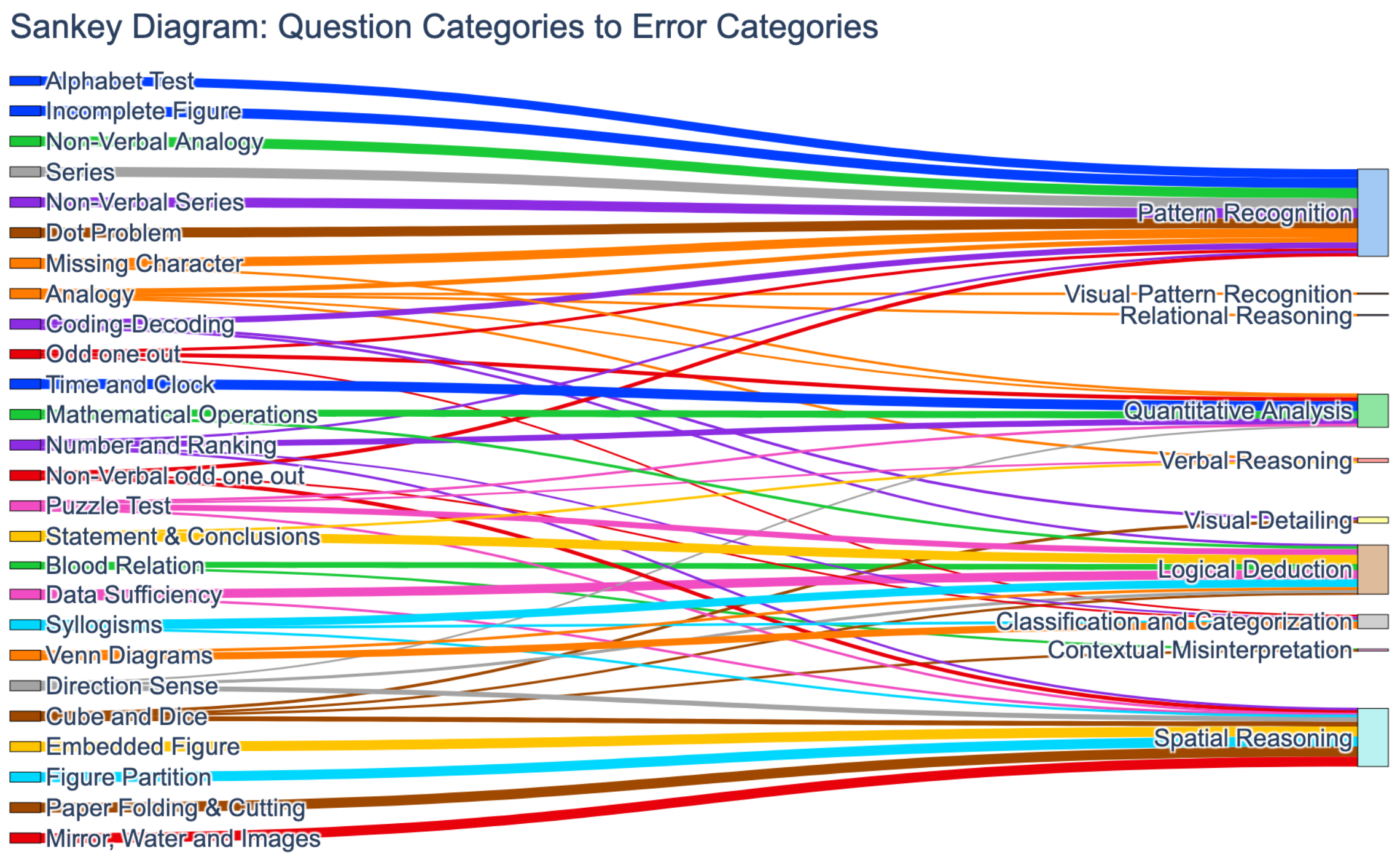


Figure: **Overview of errors Gemini 1.5 Pro.**

- Challenges in *Pattern recognition, Spatial reasoning, and Logical deduction*.

Option Shuffling on Gemini 1.5 Pro

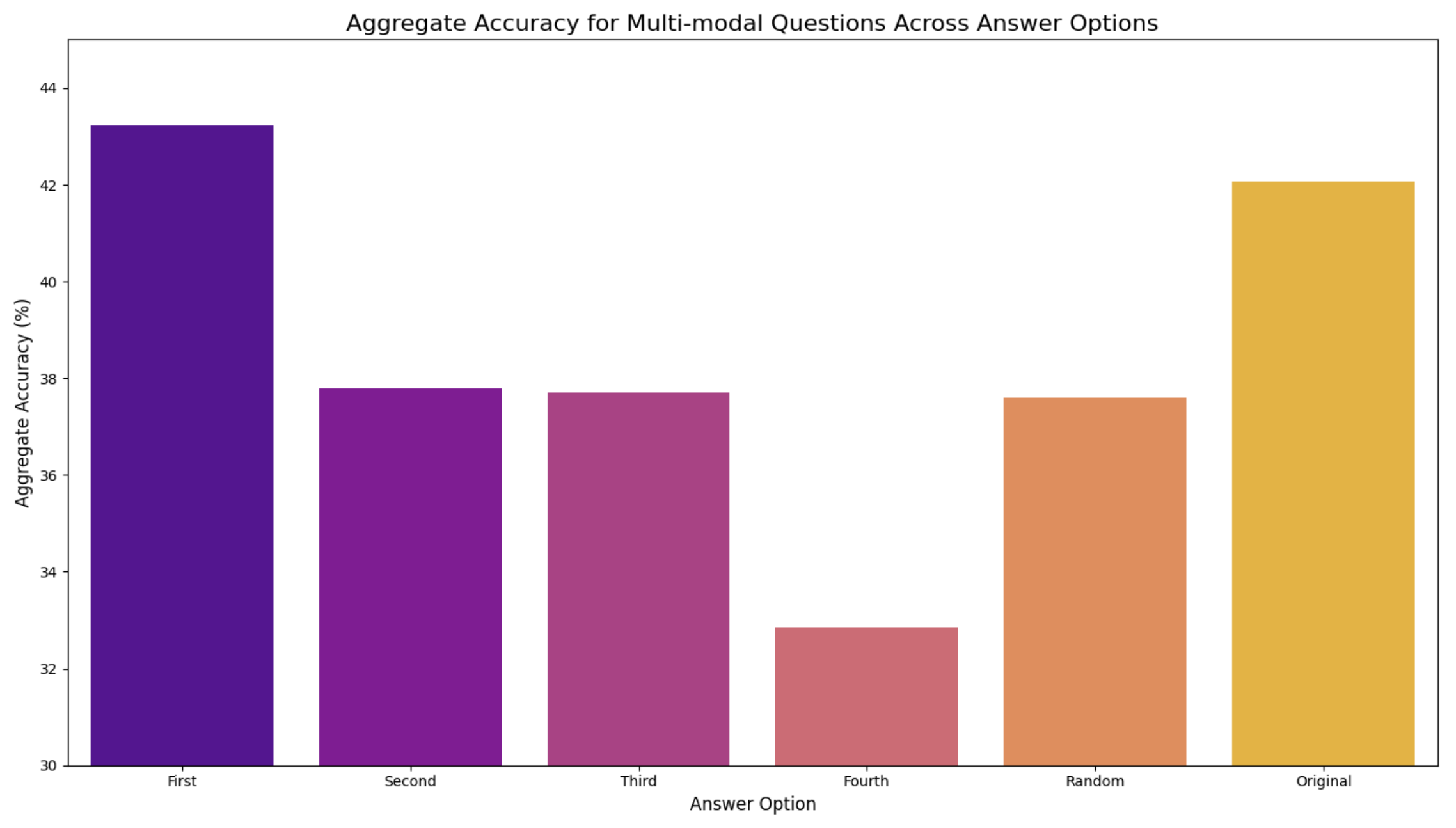


Figure: Text based

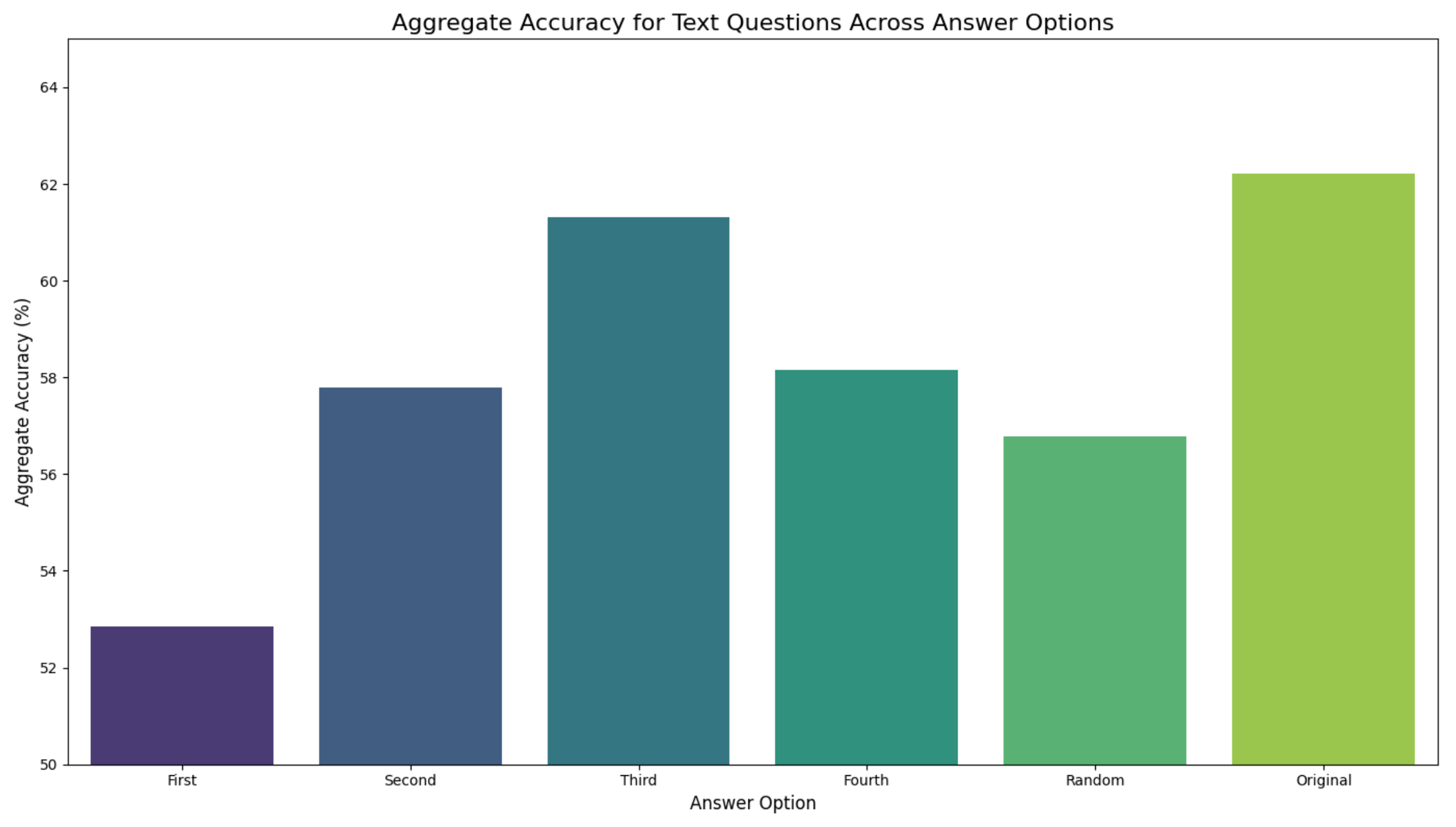


Figure: Text+Image

- Option bias ablation shows answer placement impacts model performance.

5. Future Directions

- **Data Augmentation:** Expand the dataset with new tasks and perturbations.
- **Advanced Architectures:** Explore generative VLMs for complex tasks.
- **Fine-tuning:** Assess fine-tuned performance on multi-modal reasoning.

Website and Contact Info

Website: ntsebench.github.io

Author Emails:
{p.pandya,g.vatsal,t.agney}@alumni.iitg.ac.in , tkataria@cs.utah.edu, danroth@seas.upenn.edu, vgupt140@asu.edu