NAACL 2025

NTSEBench: Cognitive Reasoning Benchmark for Vision Language Models

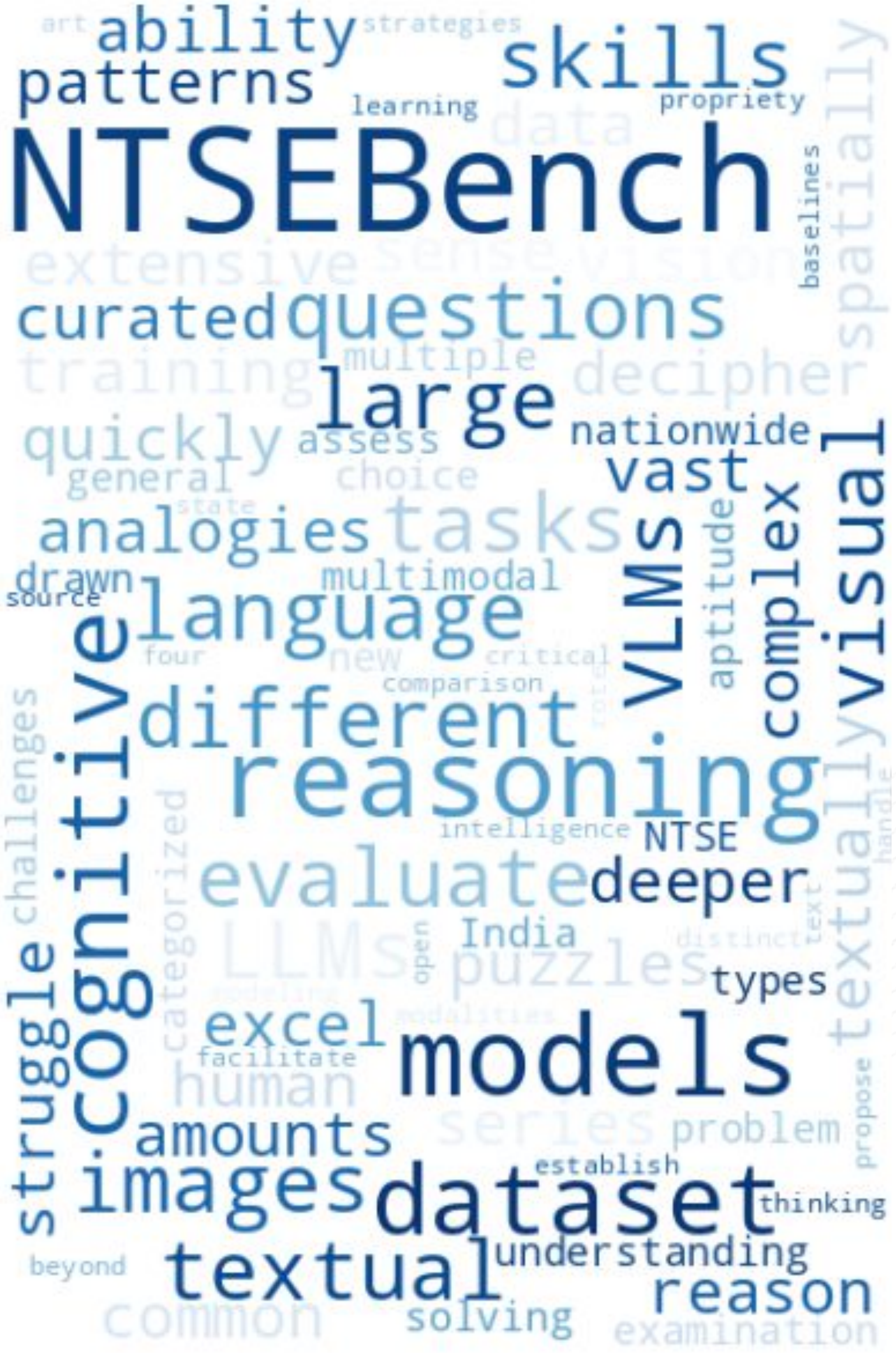
Pranshu Pandya†, Vatsal Gupta†,
Agney S Talwarr , Tushar Kataria , Dan Roth, Vivek Gupta*

†Equal Contribution, *Corresponding Author

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

MOTIVATION

- The Need for **Cognitive Reasoning** in AI
- **Gaps in Current Benchmarks**
- Advancing AI Toward Human-Like Problem Solving



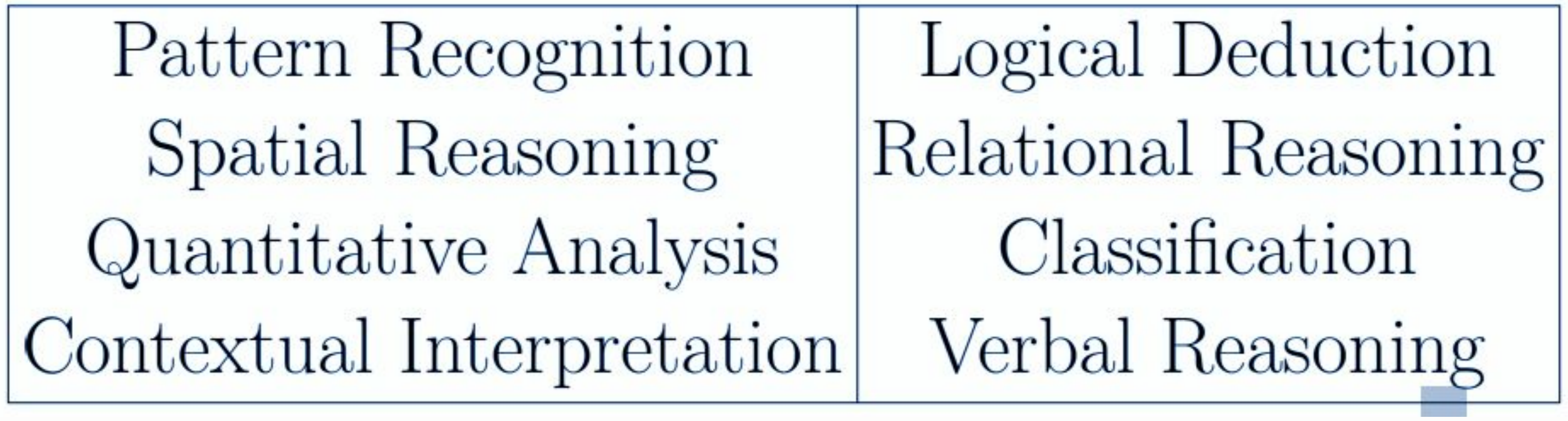
INTRODUCTION

- NTSEBench is a **novel benchmark** designed to evaluate cognitive reasoning in vision–language models.

- The **dataset targets advanced skills**—such as pattern recognition, logical deduction, and spatial reasoning—that go beyond rote memorization.

NTSEBench DATASET - KEY FEATURES

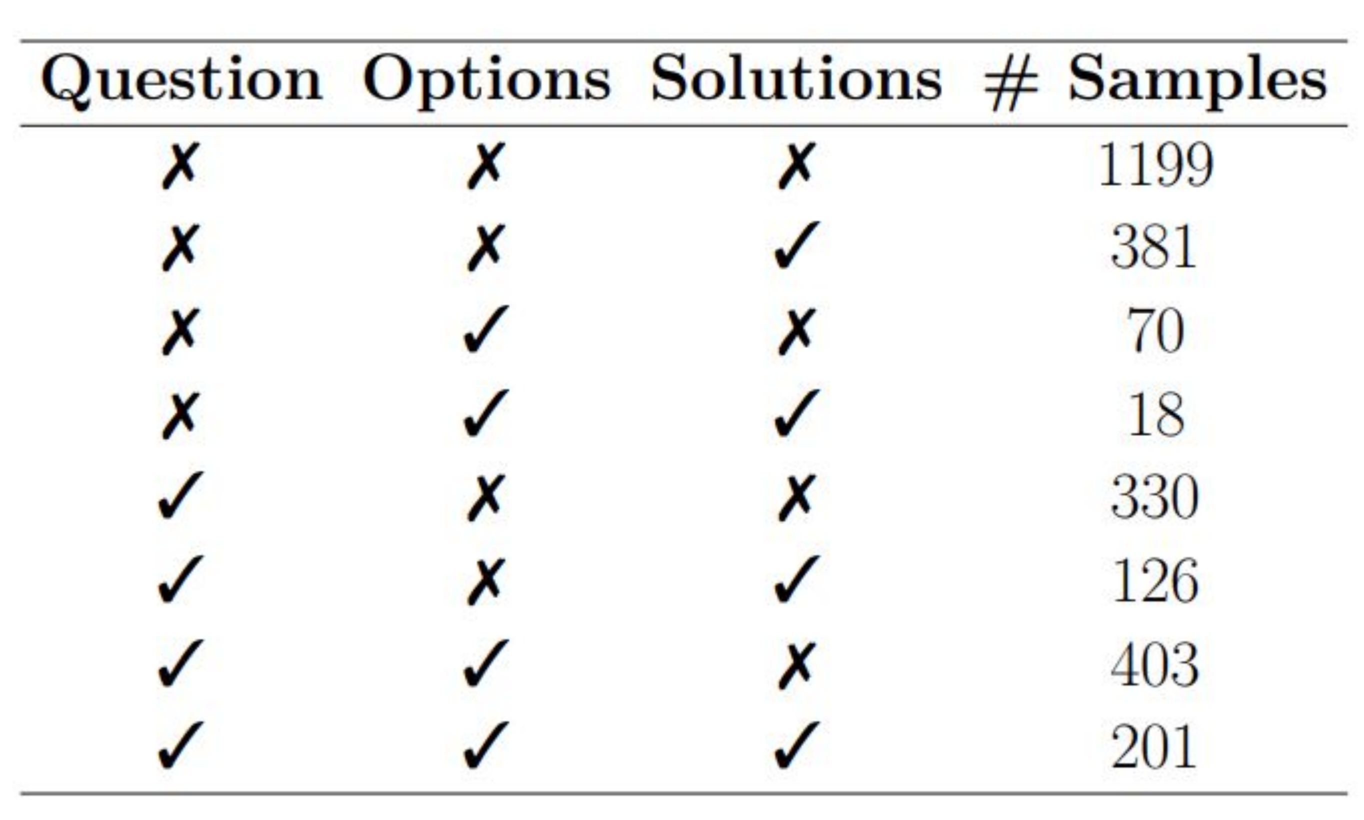
- **Extensive categorisation into 26 categories** like “Embedded Figure” , “Non-Verbal Analogy”
- **8 cognitive dimensions** proposed covering various aspects of multimodal reasoning



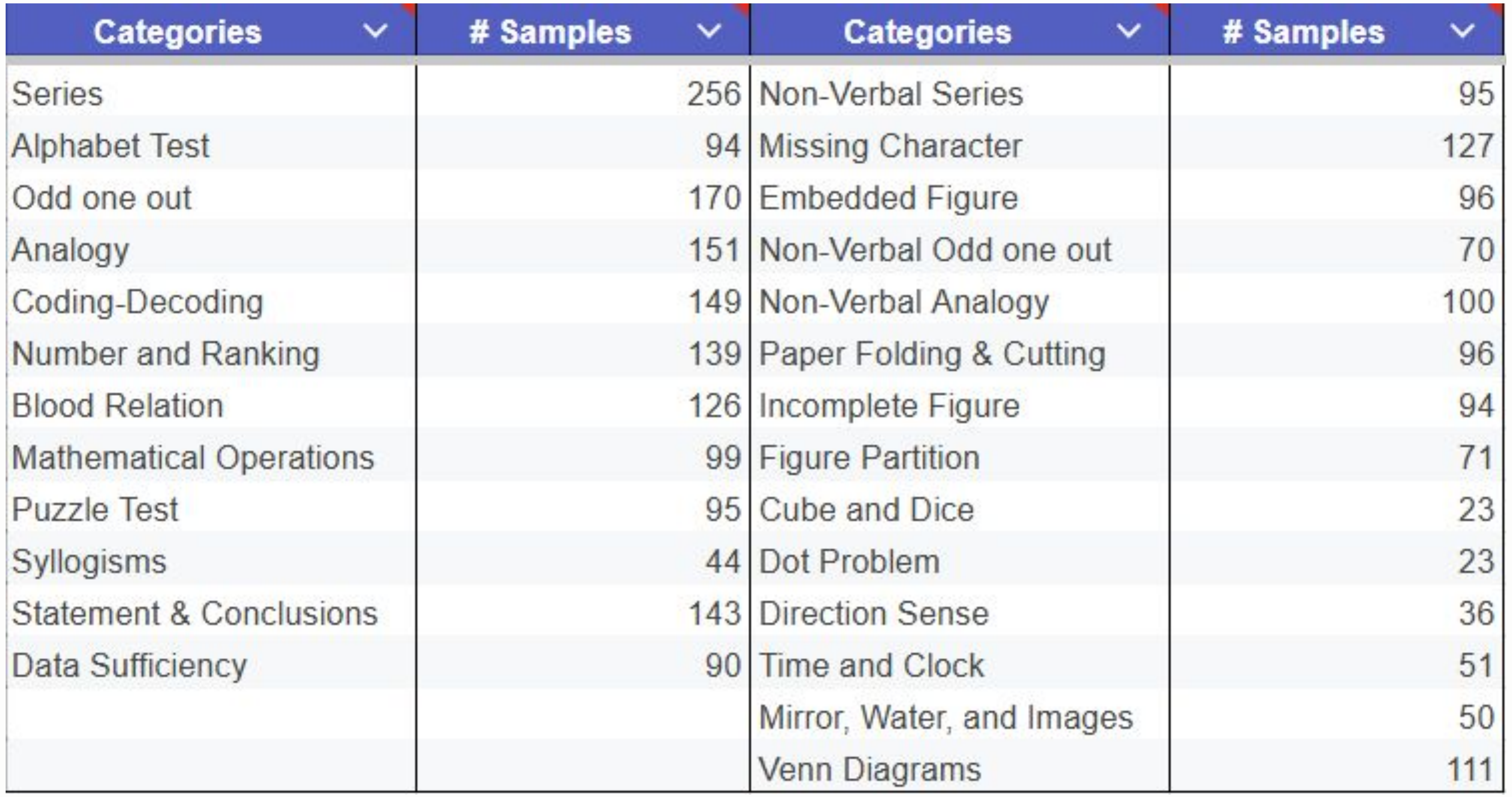
Pattern Recognition	Logical Deduction
Spatial Reasoning	Relational Reasoning
Quantitative Analysis	Classification
Contextual Interpretation	Verbal Reasoning

DATASET STATISTICS

- **Extensive categorisation** by human experts. *(Table.)*
- **Detailed solutions for most questions** - 2728 Multiple Choice Questions and 4642 images.
- NTSE-Bench's multimodal questions, options, and solutions yield **8 modality combinations**.



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

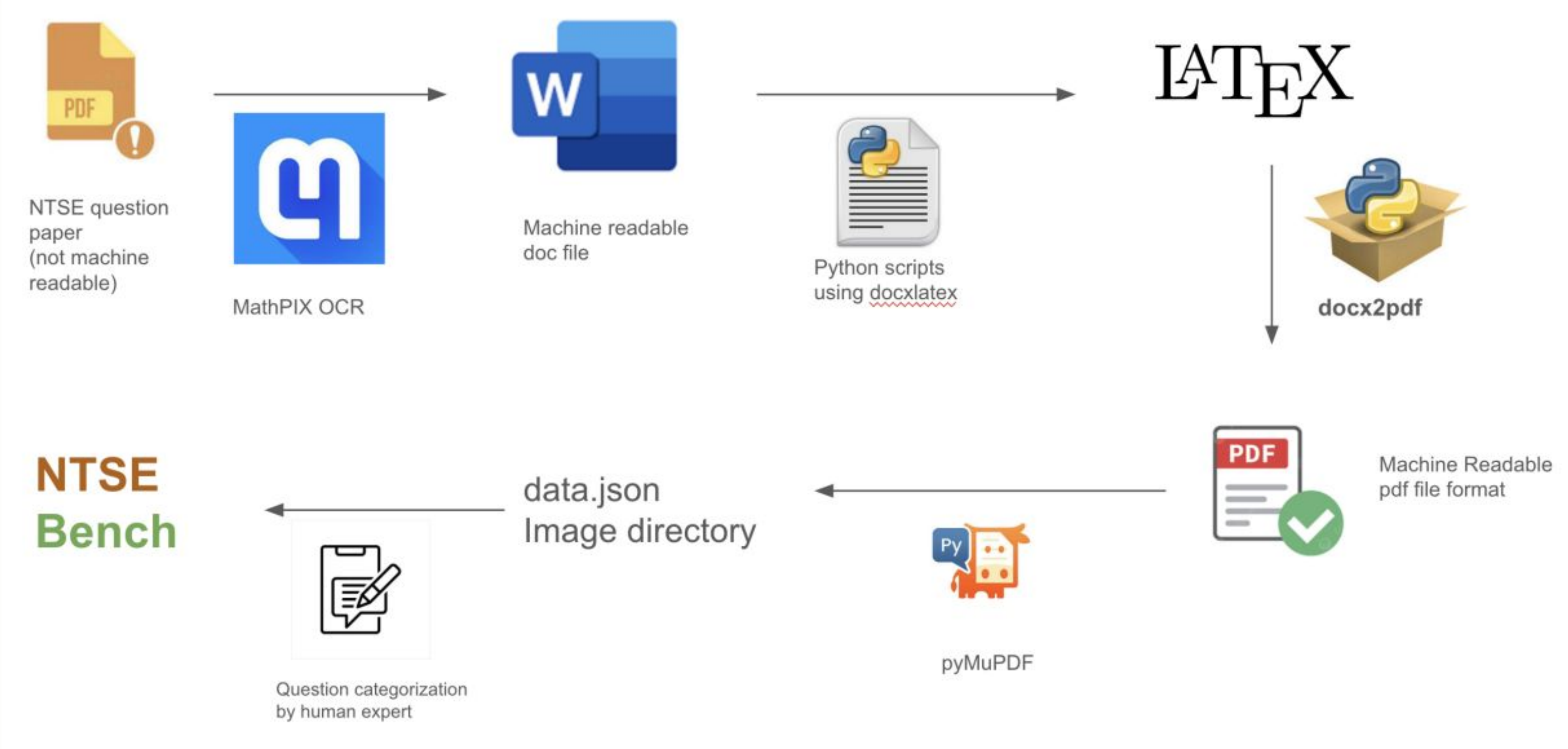


Categories	# Samples	Categories	# Samples
Series	256	Non-Verbal Series	95
Alphabet Test	94	Missing Character	127
Odd one out	170	Embedded Figure	96
Analogy	151	Non-Verbal Odd one out	70
Coding-Decoding	149	Non-Verbal Analogy	100
Number and Ranking	139	Paper Folding & Cutting	96
Blood Relation	126	Incomplete Figure	94
Mathematical Operations	99	Figure Partition	71
Puzzle Test	95	Cube and Dice	23
Syllogisms	44	Dot Problem	23
Statement & Conclusions	143	Direction Sense	36
Data Sufficiency	90	Time and Clock	51
		Mirror, Water, and Images	50
		Venn Diagrams	111

DATASET CURATION - PIPELINE

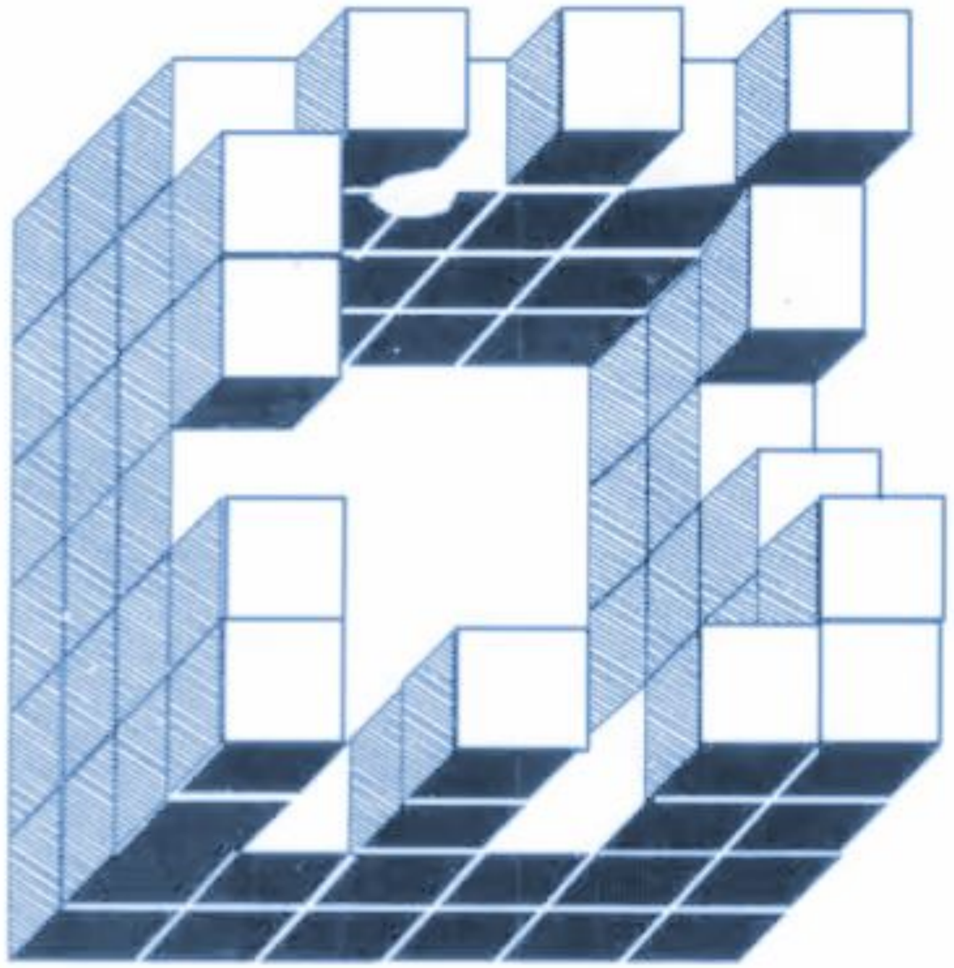
- Extensive **manual curation and a multi-step extraction pipeline** convert non-machine-readable NTSE papers into structured, machine-readable PDFs, ensuring a high-quality, accessible dataset for analysis.
- Dataset is curated **through multiple sources.**

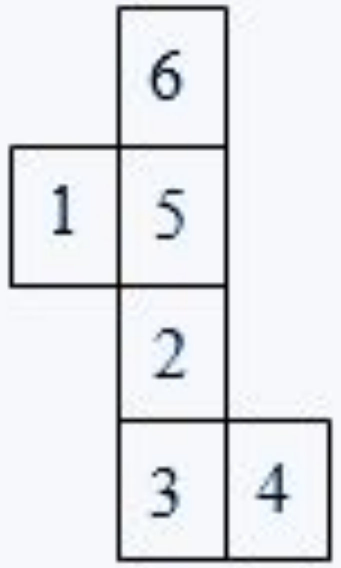
.


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

Examples of Questions

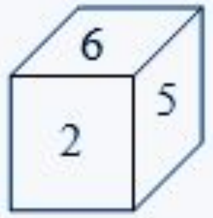
Count the number of cubes in the 3D Model below



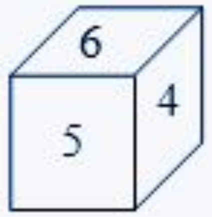


(X)

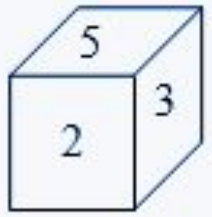
Find the cube which is yielded by the **net (X)**



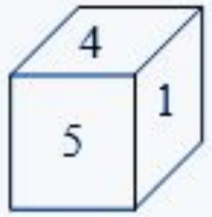
(1)



(2)

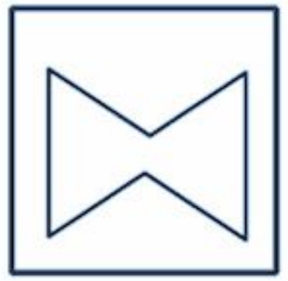


(3)

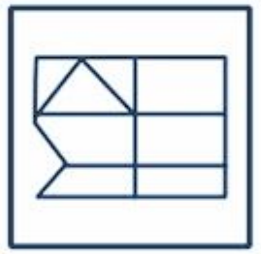


(4)

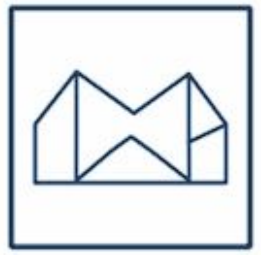
Question figure



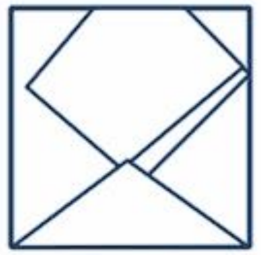
Answer Figure



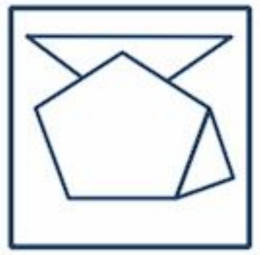
(a)



(b)



(c)



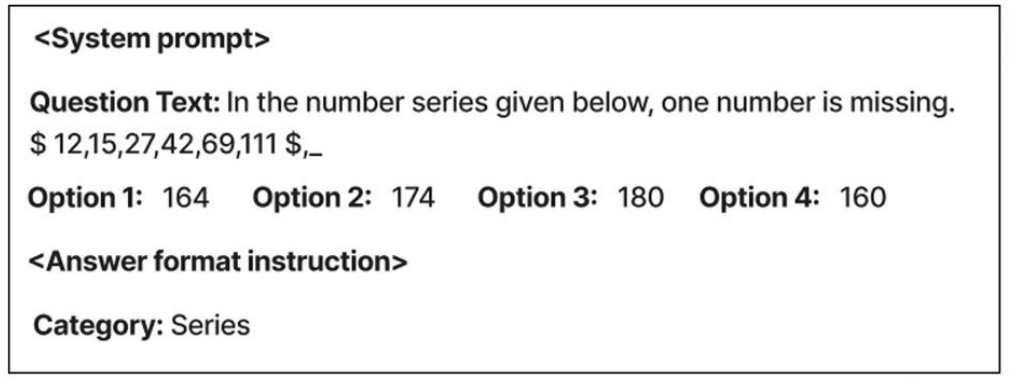
(d)

Find the figure in which the "**Question Figure**" is embedded

MODELING STRATEGIES PROPOSED

Standard QA

(A)



<System prompt>

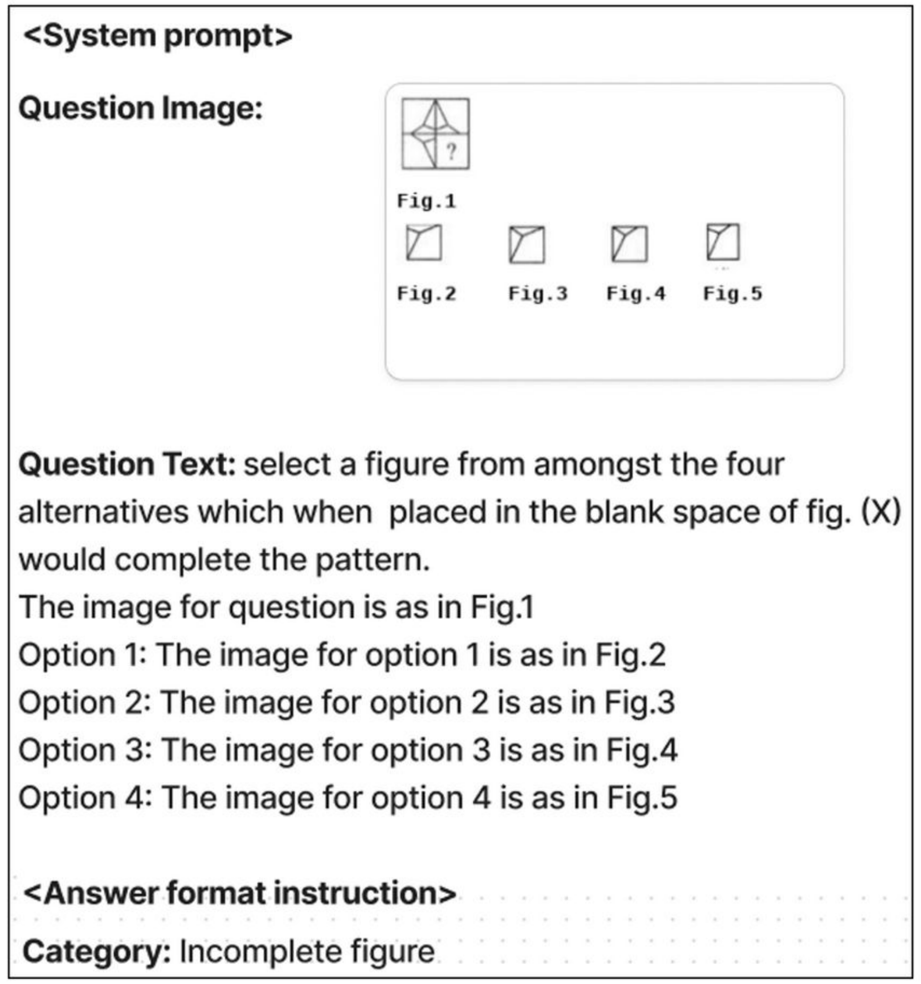
Question Text: In the number series given below, one number is missing. $ 12,15,27,42,69,111 $,_

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

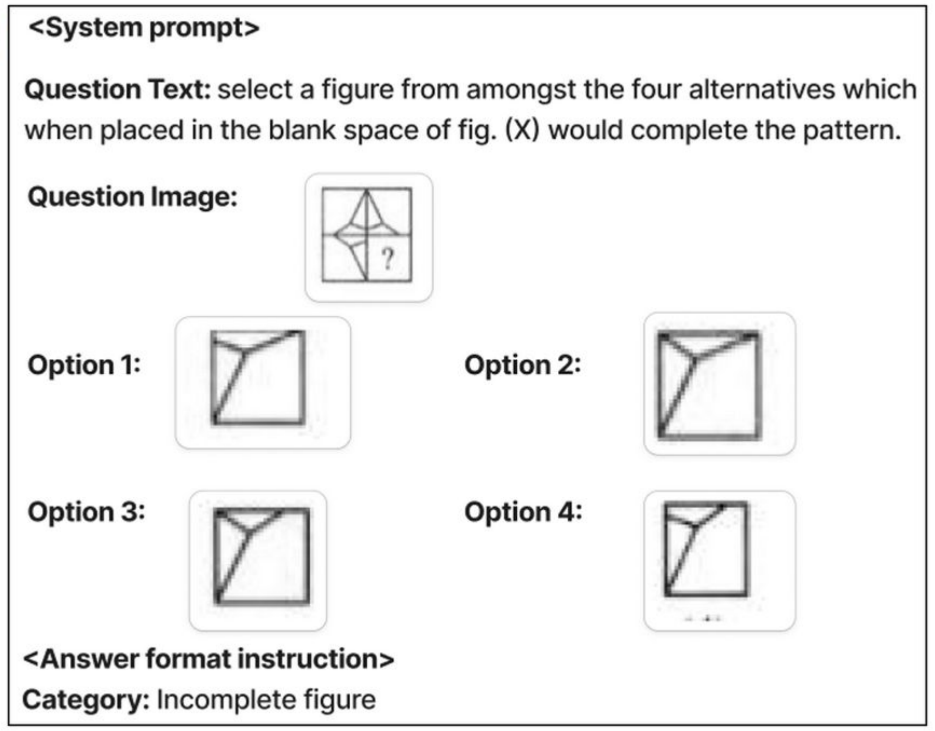
<Answer format instruction>

Category: Incomplete figure

Standard VQA

Interleaved

(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:

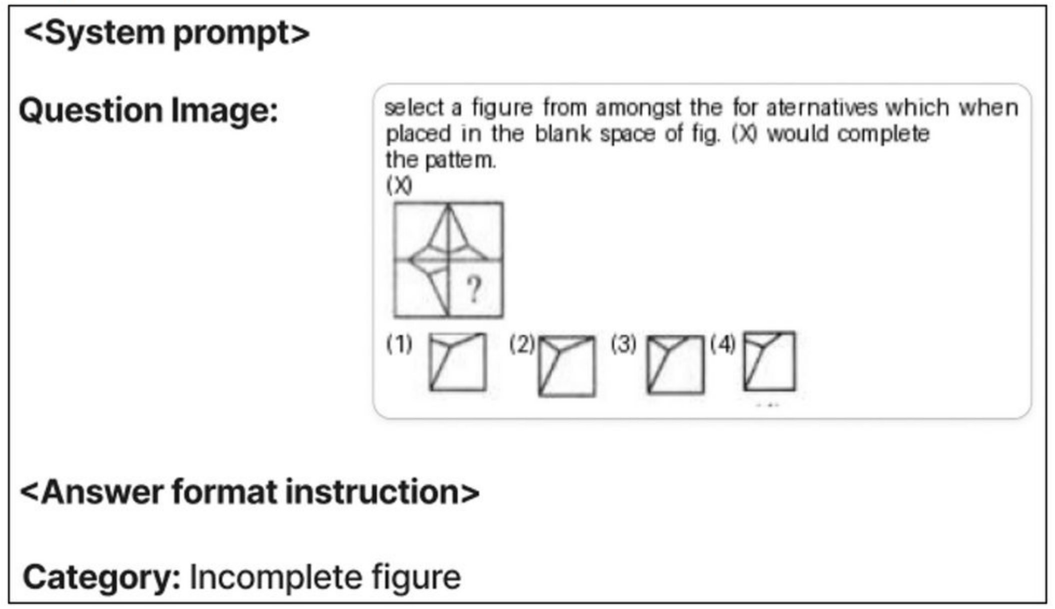
Option 1: **Option 2:**

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)

(1) (2) (3) (4)

<Answer format instruction>

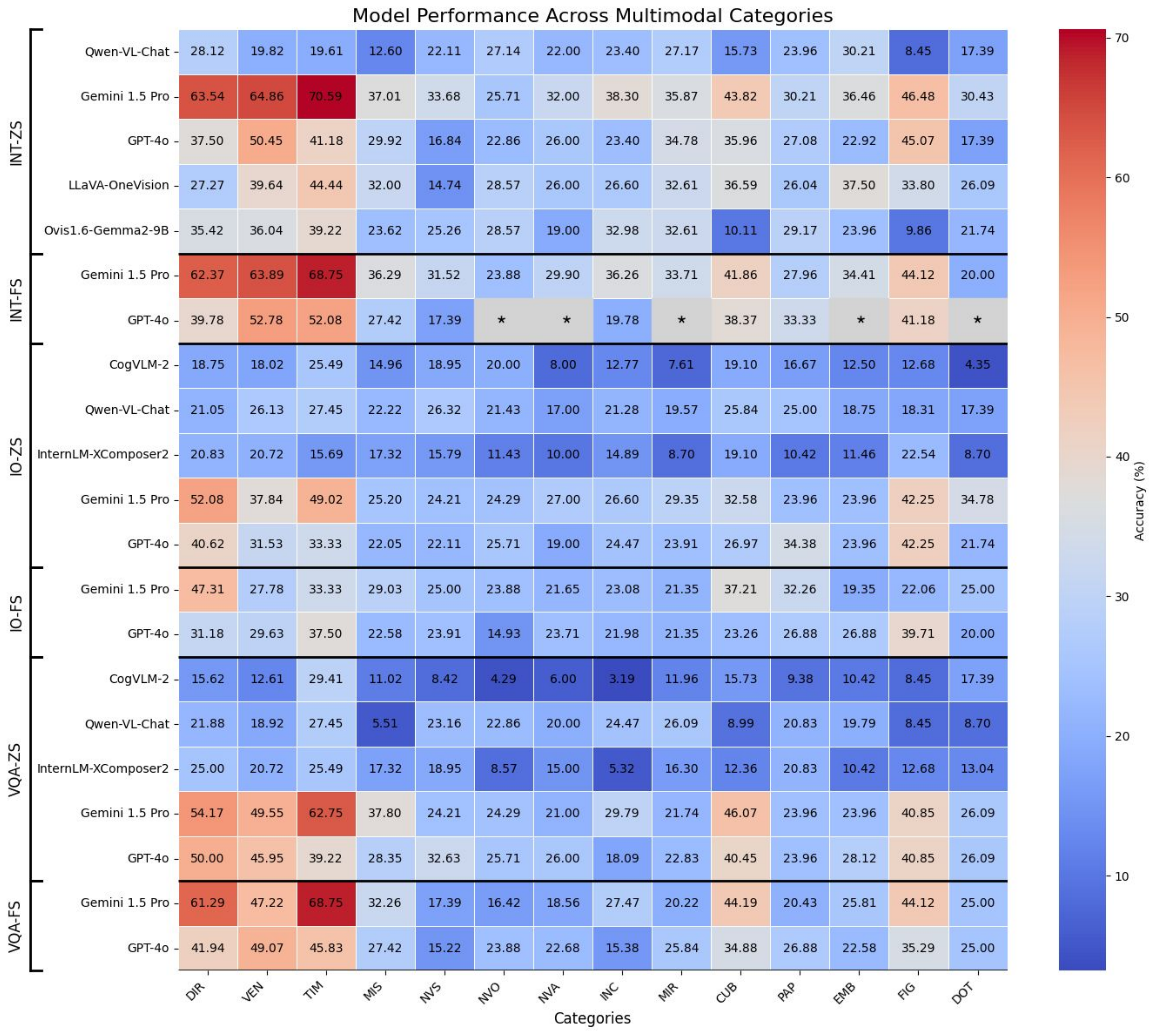
Category: Incomplete figure

Image Only

RESULTS

Multimodal Categories Performance

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



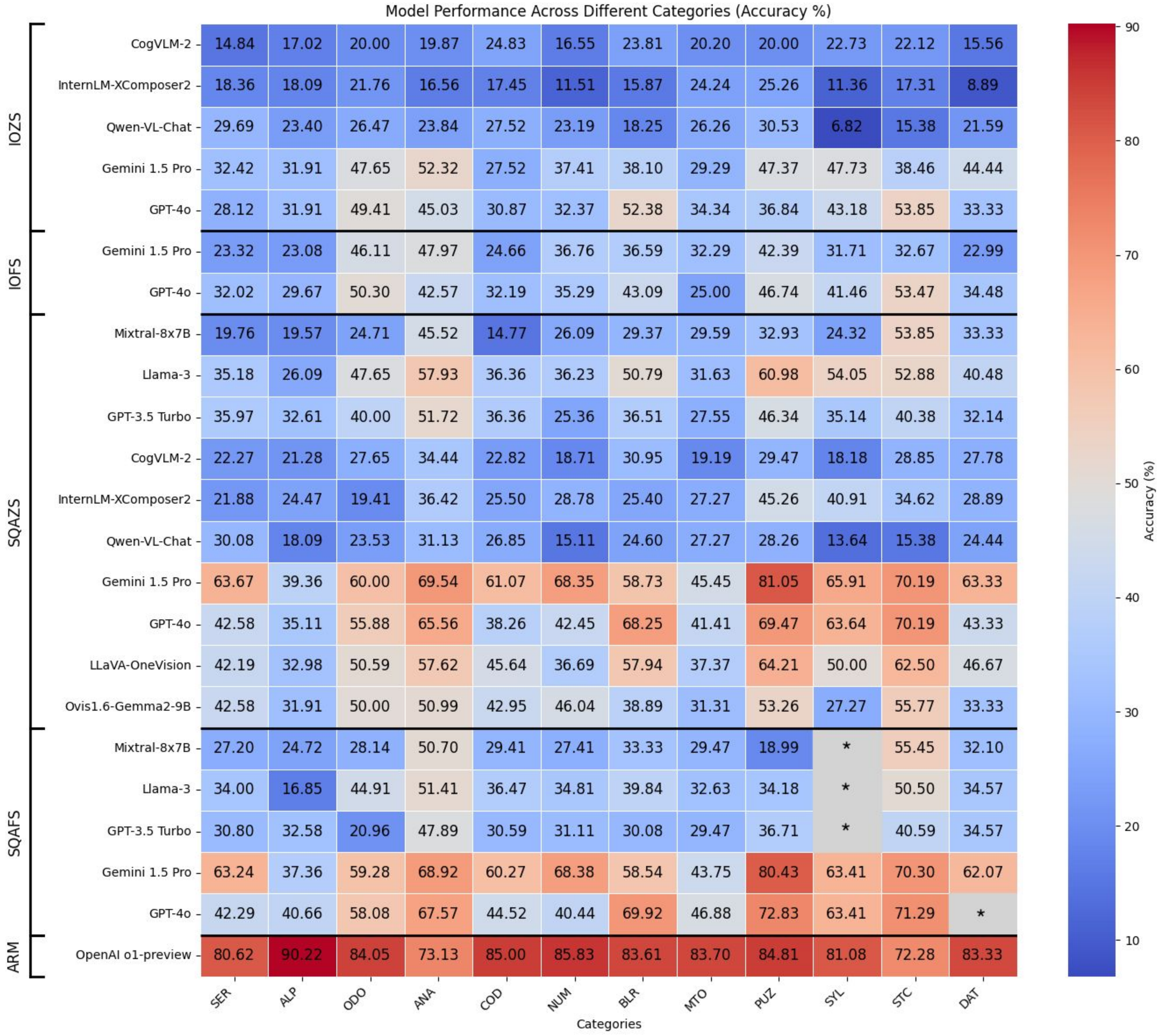
Model Performance Across Multimodal Categories

		DIR	VEN	TIM	MIS	NVS	NVO	NVA	INC	MIR	CUB	PAP	EMB	FIG	DOT
INT-ZS	Qwen-VL-Chat	28.12	19.82	19.61	12.60	22.11	27.14	22.00	23.40	27.17	15.73	23.96	30.21	8.45	17.39
	Gemini 1.5 Pro	63.54	64.86	70.59	37.01	33.68	25.71	32.00	38.30	35.87	43.82	30.21	36.46	46.48	30.43
	GPT-4o	37.50	50.45	41.18	29.92	16.84	22.86	26.00	23.40	34.78	35.96	27.08	22.92	45.07	17.39
	LLaVA-OneVision	27.27	39.64	44.44	32.00	14.74	28.57	26.00	26.60	32.61	36.59	26.04	37.50	33.80	26.09
	Ovis1.6-Gemma2-9B	35.42	36.04	39.22	23.62	25.26	28.57	19.00	32.98	32.61	10.11	29.17	23.96	9.86	21.74
INT-FS	Gemini 1.5 Pro	62.37	63.89	68.75	36.29	31.52	23.88	29.90	36.26	33.71	41.86	27.96	34.41	44.12	20.00
	GPT-4o	39.78	52.78	52.08	27.42	17.39	*	*	19.78	*	38.37	33.33	*	41.18	*
IO-ZS	CogVLM-2	18.75	18.02	25.49	14.96	18.95	20.00	8.00	12.77	7.61	19.10	16.67	12.50	12.68	4.35
	Qwen-VL-Chat	21.05	26.13	27.45	22.22	26.32	21.43	17.00	21.28	19.57	25.84	25.00	18.75	18.31	17.39
	InternLM-XComposer2	20.83	20.72	15.69	17.32	15.79	11.43	10.00	14.89	8.70	19.10	10.42	11.46	22.54	8.70
	Gemini 1.5 Pro	52.08	37.84	49.02	25.20	24.21	24.29	27.00	26.60	29.35	32.58	23.96	23.96	42.25	34.78
	GPT-4o	40.62	31.53	33.33	22.05	22.11	25.71	19.00	24.47	23.91	26.97	34.38	23.96	42.25	21.74
IO-FS	Gemini 1.5 Pro	47.31	27.78	33.33	29.03	25.00	23.88	21.65	23.08	21.35	37.21	32.26	19.35	22.06	25.00
	GPT-4o	31.18	29.63	37.50	22.58	23.91	14.93	23.71	21.98	21.35	23.26	26.88	26.88	39.71	20.00
VQA-ZS	CogVLM-2	15.62	12.61	29.41	11.02	8.42	4.29	6.00	3.19	11.96	15.73	9.38	10.42	8.45	17.39
	Qwen-VL-Chat	21.88	18.92	27.45	5.51	23.16	22.86	20.00	24.47	26.09	8.99	20.83	19.79	8.45	8.70
	InternLM-XComposer2	25.00	20.72	25.49	17.32	18.95	8.57	15.00	5.32	16.30	12.36	20.83	10.42	12.68	13.04
	Gemini 1.5 Pro	54.17	49.55	62.75	37.80	24.21	24.29	21.00	29.79	21.74	46.07	23.96	23.96	40.85	26.09
	GPT-4o	50.00	45.95	39.22	28.35	32.63	25.71	26.00	18.09	22.83	40.45	23.96	28.12	40.85	26.09
VQA-FS	Gemini 1.5 Pro	61.29	47.22	68.75	32.26	17.39	16.42	18.56	27.47	20.22	44.19	20.43	25.81	44.12	25.00
	GPT-4o	41.94	49.07	45.83	27.42	15.22	23.88	22.68	15.38	25.84	34.88	26.88	22.58	35.29	25.00

RESULTS

Text-Only Categories Performance

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

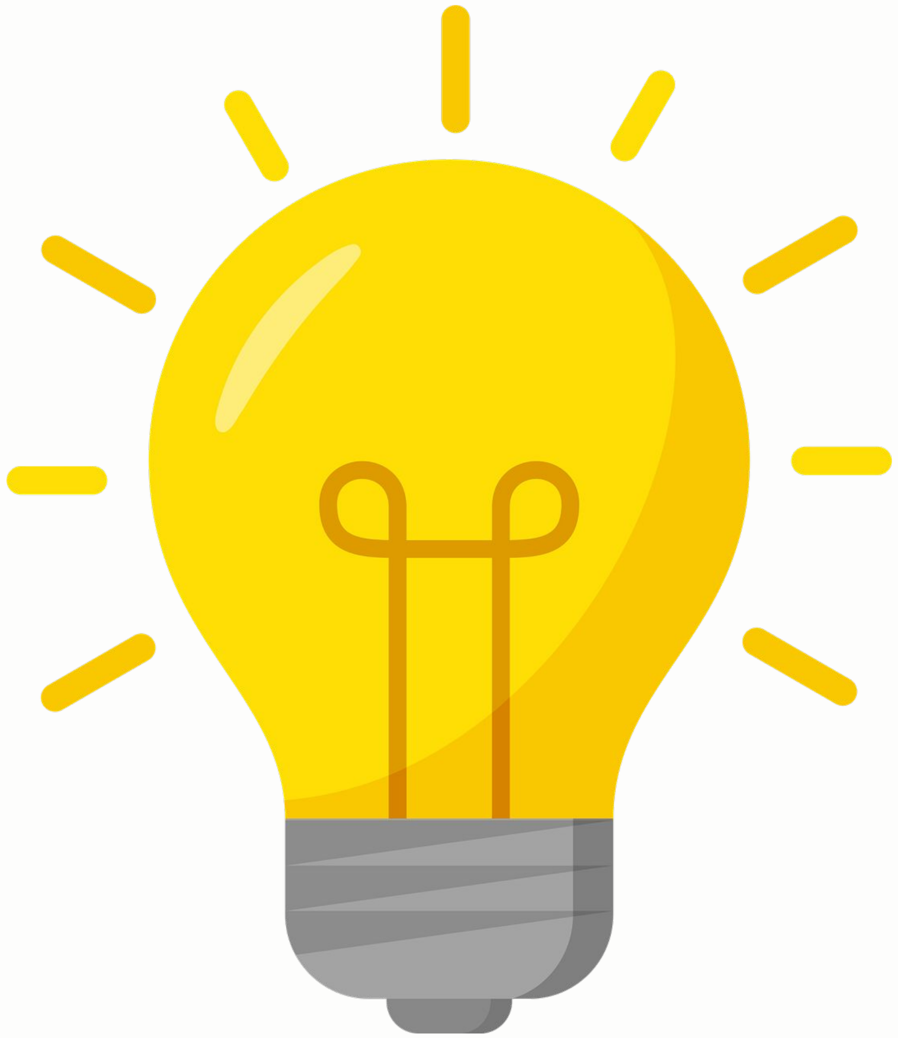


Model Performance Across Different Categories (Accuracy %)

		SER	ALP	ODO	ANA	COD	NUM	BLR	MTO	PUZ	SYL	STC	DAT
IOZS	CogVLM-2	14.84	17.02	20.00	19.87	24.83	16.55	23.81	20.20	20.00	22.73	22.12	15.56
	InternLM-XComposer2	18.36	18.09	21.76	16.56	17.45	11.51	15.87	24.24	25.26	11.36	17.31	8.89
	Qwen-VL-Chat	29.69	23.40	26.47	23.84	27.52	23.19	18.25	26.26	30.53	6.82	15.38	21.59
	Gemini 1.5 Pro	32.42	31.91	47.65	52.32	27.52	37.41	38.10	29.29	47.37	47.73	38.46	44.44
	GPT-4o	28.12	31.91	49.41	45.03	30.87	32.37	52.38	34.34	36.84	43.18	53.85	33.33
IOFS	Gemini 1.5 Pro	23.32	23.08	46.11	47.97	24.66	36.76	36.59	32.29	42.39	31.71	32.67	22.99
	GPT-4o	32.02	29.67	50.30	42.57	32.19	35.29	43.09	25.00	46.74	41.46	53.47	34.48
SQAZS	Mixtral-8x7B	19.76	19.57	24.71	45.52	14.77	26.09	29.37	29.59	32.93	24.32	53.85	33.33
	Llama-3	35.18	26.09	47.65	57.93	36.36	36.23	50.79	31.63	60.98	54.05	52.88	40.48
	GPT-3.5 Turbo	35.97	32.61	40.00	51.72	36.36	25.36	36.51	27.55	46.34	35.14	40.38	32.14
	CogVLM-2	22.27	21.28	27.65	34.44	22.82	18.71	30.95	19.19	29.47	18.18	28.85	27.78
	InternLM-XComposer2	21.88	24.47	19.41	36.42	25.50	28.78	25.40	27.27	45.26	40.91	34.62	28.89
	Qwen-VL-Chat	30.08	18.09	23.53	31.13	26.85	15.11	24.60	27.27	28.26	13.64	15.38	24.44
	Gemini 1.5 Pro	63.67	39.36	60.00	69.54	61.07	68.35	58.73	45.45	81.05	65.91	70.19	63.33
	GPT-4o	42.58	35.11	55.88	65.56	38.26	42.45	68.25	41.41	69.47	63.64	70.19	43.33
	LLaVA-OneVision	42.19	32.98	50.59	57.62	45.64	36.69	57.94	37.37	64.21	50.00	62.50	46.67
	Ovis1.6-Gemma2-9B	42.58	31.91	50.00	50.99	42.95	46.04	38.89	31.31	53.26	27.27	55.77	33.33
SQAFS	Mixtral-8x7B	27.20	24.72	28.14	50.70	29.41	27.41	33.33	29.47	18.99	*	55.45	32.10
	Llama-3	34.00	16.85	44.91	51.41	36.47	34.81	39.84	32.63	34.18	*	50.50	34.57
	GPT-3.5 Turbo	30.80	32.58	20.96	47.89	30.59	31.11	30.08	29.47	36.71	*	40.59	34.57
	Gemini 1.5 Pro	63.24	37.36	59.28	68.92	60.27	68.38	58.54	43.75	80.43	63.41	70.30	62.07
	GPT-4o	42.29	40.66	58.08	67.57	44.52	40.44	69.92	46.88	72.83	63.41	71.29	*
ARM	OpenAI o1-preview	80.62	90.22	84.05	73.13	85.00	85.83	83.61	83.70	84.81	81.08	72.28	83.33

RESULTS- KEY INSIGHTS

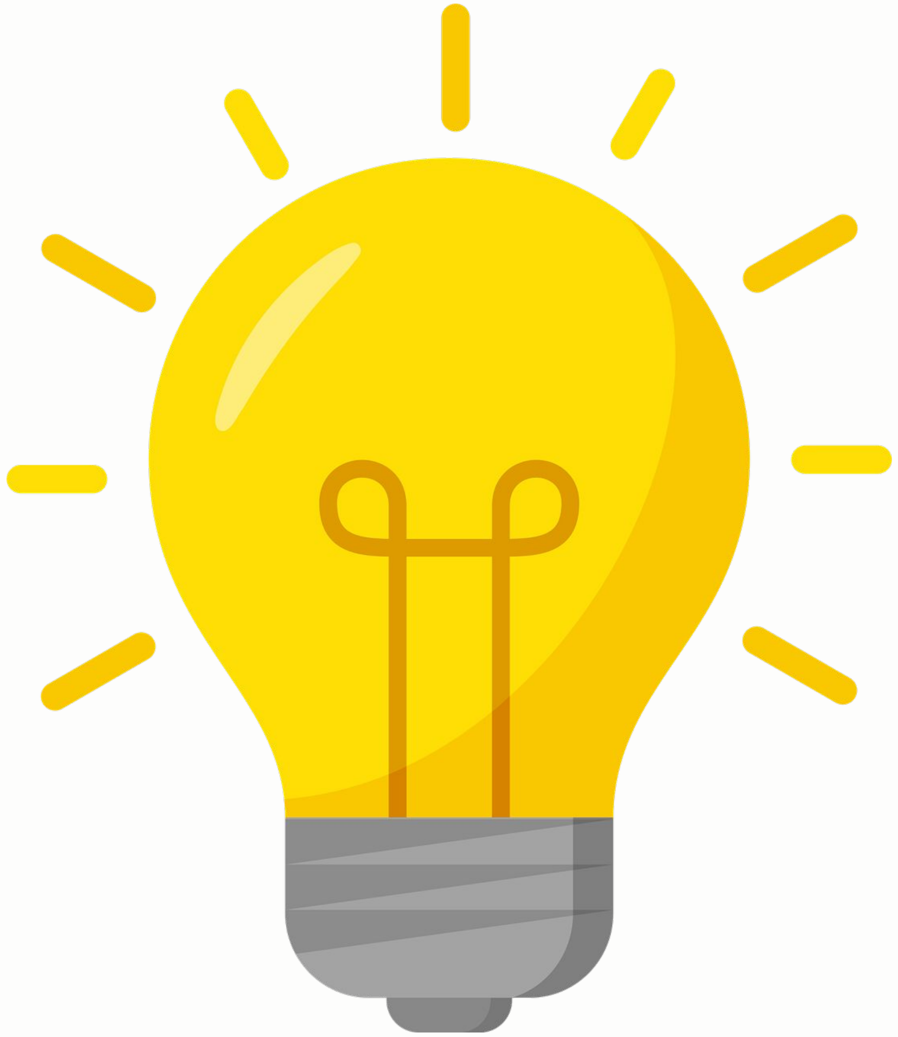
- **Proprietary models** > open-source models.
- **Interleaving text >** Standard VQA and Image Only.
- **Multimodal reasoning is challenging** and proves to be an area of **significant hardness** for even SOTA models



RESULTS- KEY INSIGHTS

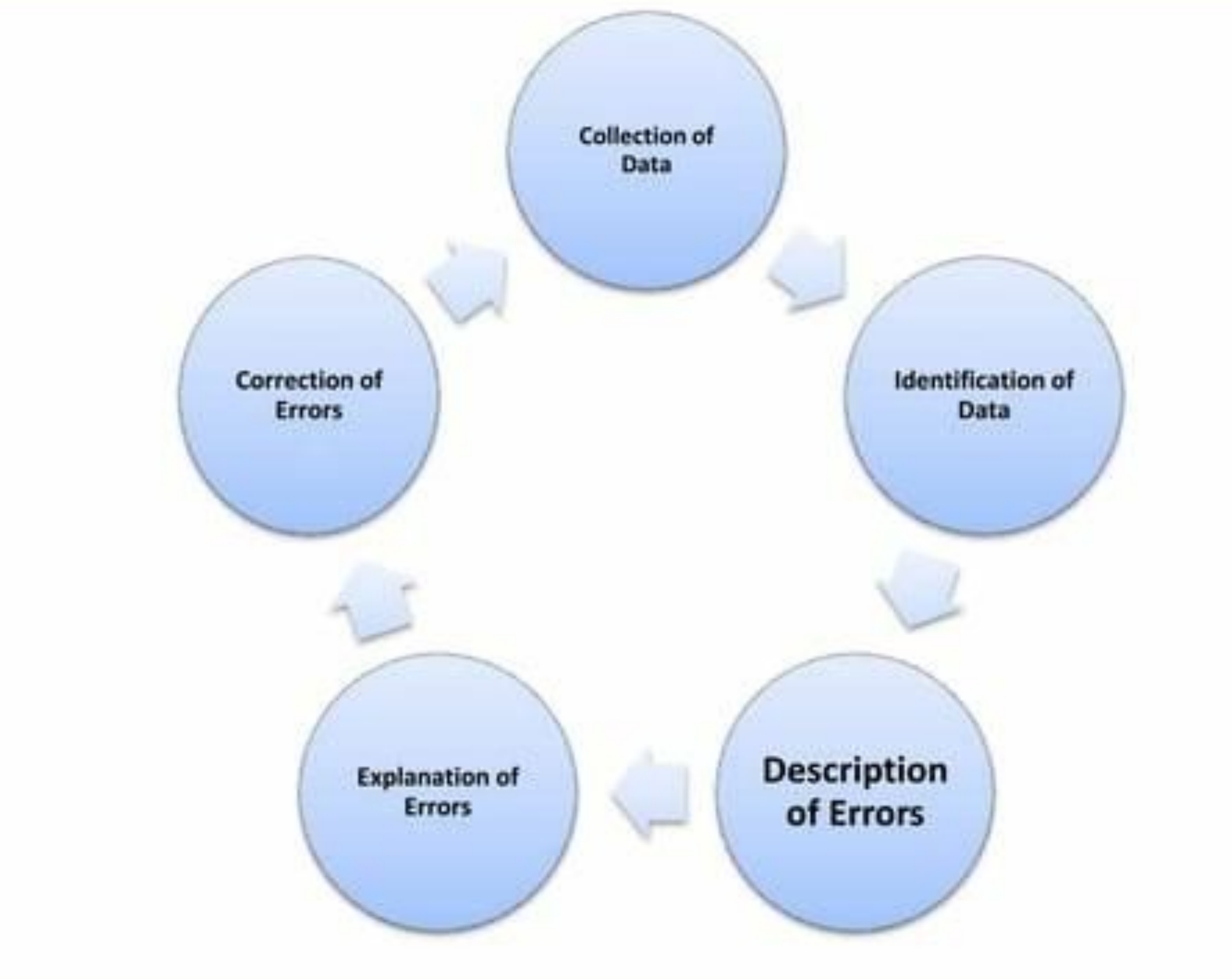
- **Human accuracy (80%) >SOTA models** (62% text, 42% visual)

- **NTSEBench proves itself to be a novel and important benchmark** which can improve models significantly and exposes model limitations indiverse categories



EXTENSIVE ERROR ANALYSIS

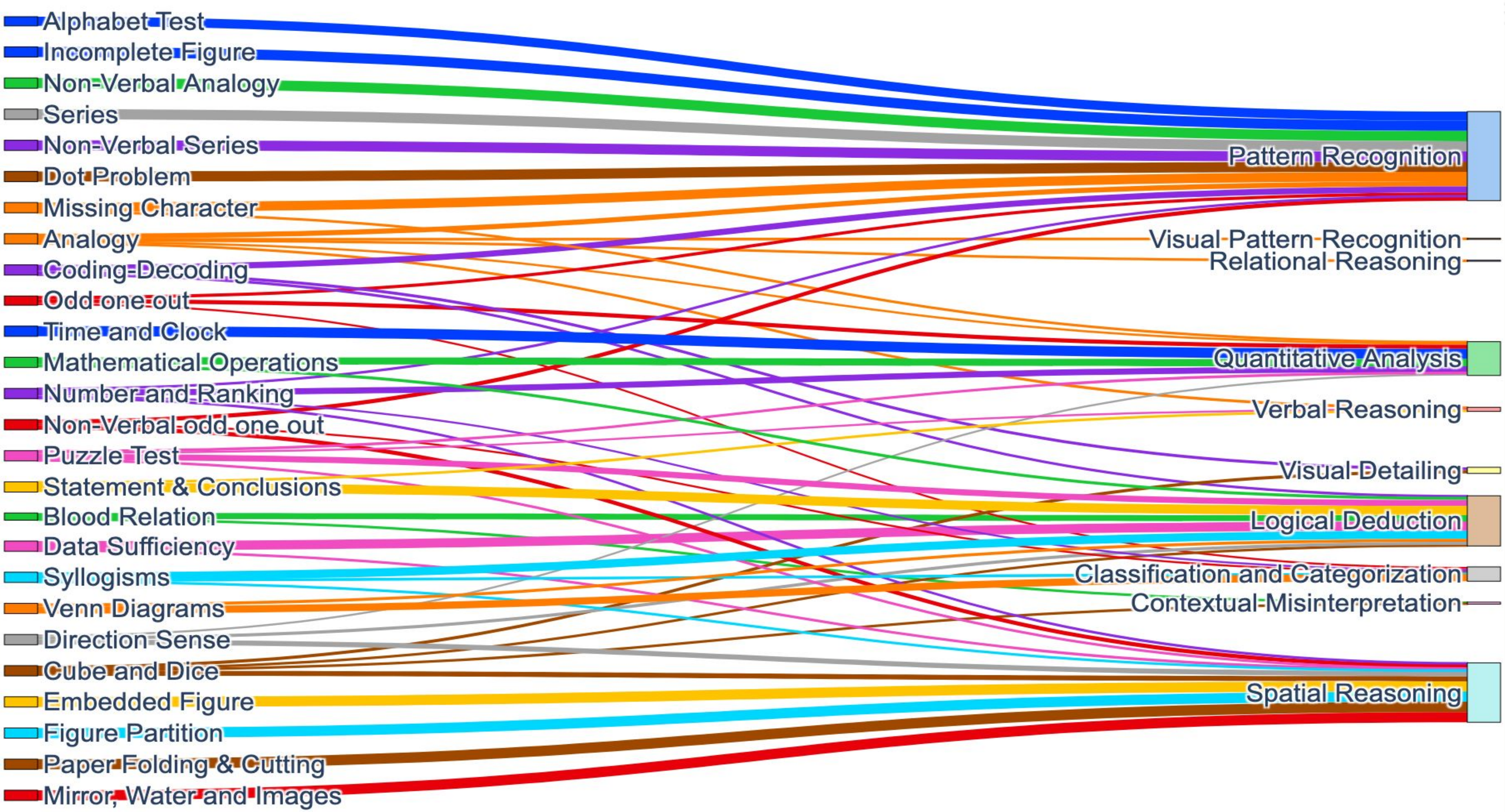
Crucial Question : How do we categories the kind of mistakes models make? What does the elicited reasoning indicate?



- **Analysed 260 questions for Gemini 1.5 Pro,** revealing reasoning patterns**.**
- **Categorised errors** using 8 cognitive dimensions**.**
- **VLMs struggle** with logical deductions from limited visuals, especially in pattern recognition, spatial manipulation, and shape recognition**.**
- **Error distribution** highlights model strengths and weaknesses for improvement.

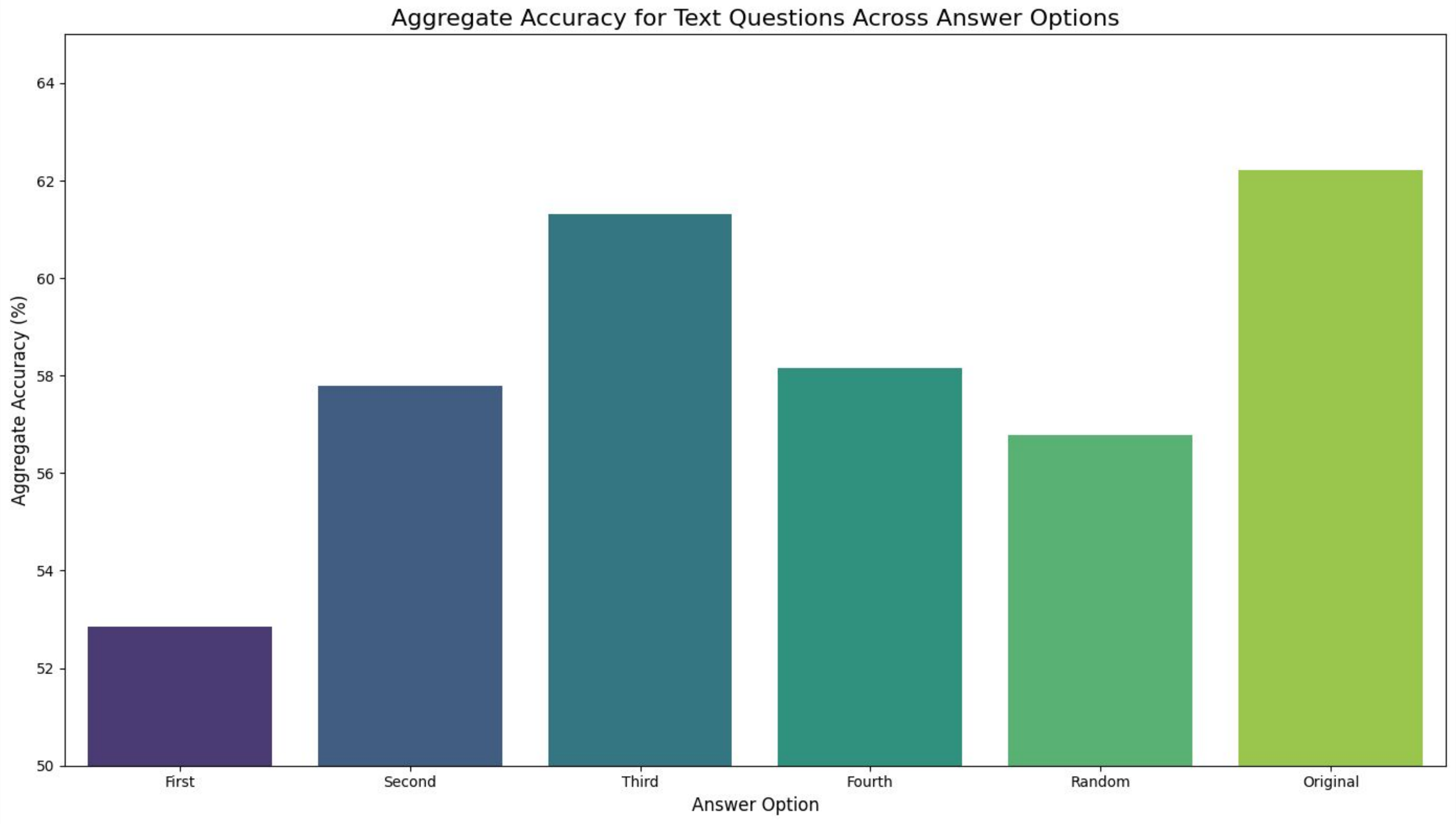
EXTENSIVE ERROR ANALYSIS

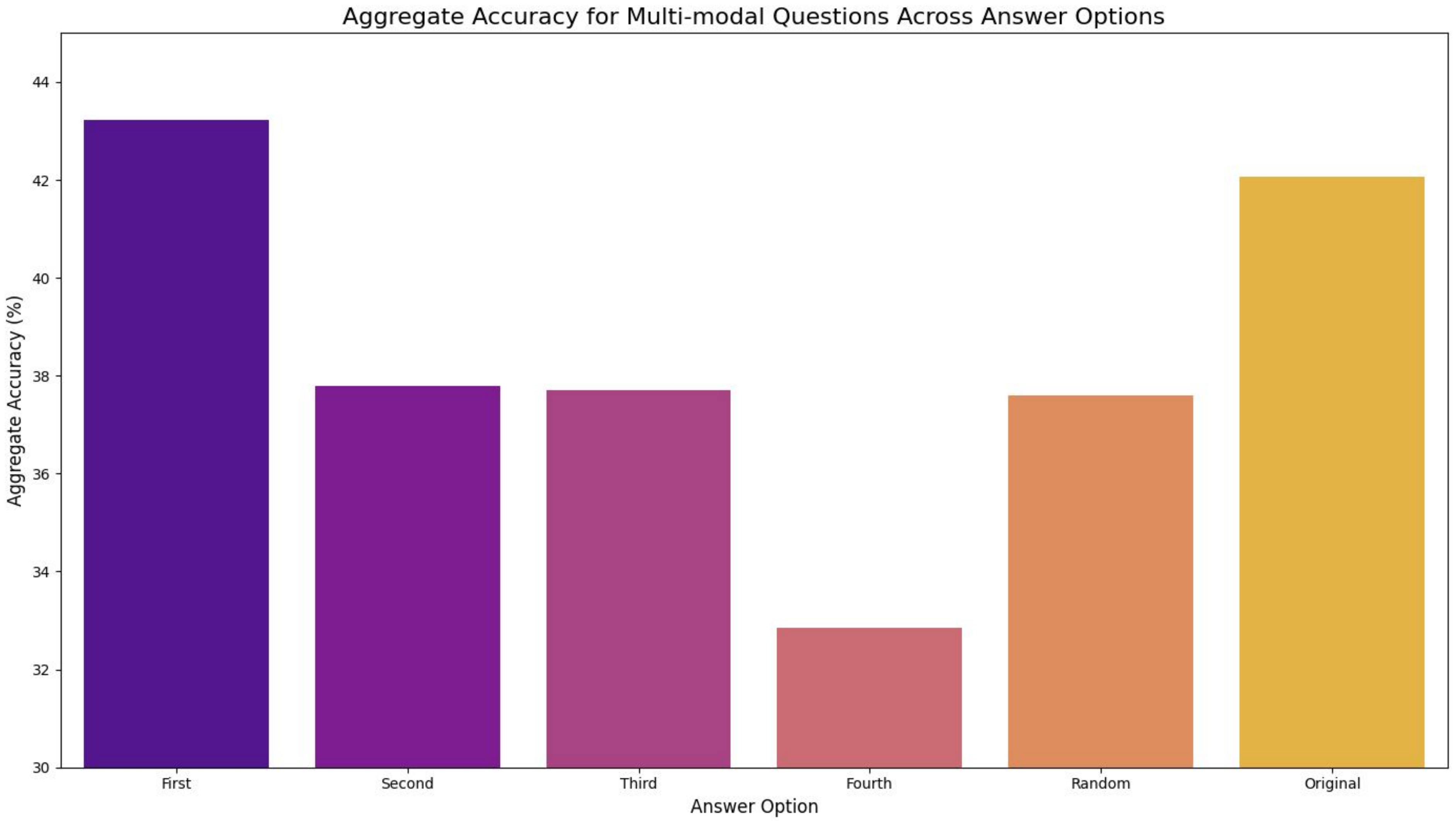
NAACL 2025



OPTION ABALATION-BIAS EXPERIMENTATION

- **Tested Gemini 1.5 Pro** to assess the impact of correct option placement on performance.
- Variations ranged from **-4% to +6% for text** and **-5% to +5% for multimodal questions**.
- Enhances **measurement of cognitive reasoning** over rote learning.





FUTURE WORK

- **Enhancing VLM Reasoning:** VLMs struggle with novel patterns; future work will explore architectural improvements and generative model integration.

- **Expanding Dataset Scope:** Include data augmentation and multilingual expansion to analyze reasoning across languages.

- **Multilingual:** The dataset is English-only, but NTSE's availability in regional languages enables future multilingual expansion.

CONCLUSION

- **Challenging Benchmark:** NTSEBench tests advanced reasoning in LLMs and VLMs, exposing their limitations.

- **Deep Model & Method Analysis:** Our novel methods enable a comprehensive evaluation of reasoning across diverse models.

- **Performance Gaps:** VLMs struggle with multimodal reasoning, and proprietary models outperform open-source ones.

THANK YOU!

- We would be happy to discuss and address any questions.

GITHUB

https://github.com/NTSEBench/NTSEBench

PAPER

https://ar.xivorg/abs/2407.10380

WEBSITE

https://ntsebench.github.io/