AFNLP
IJCNLP AACL 2025 December 20-24 Mumbai, India

CORAL

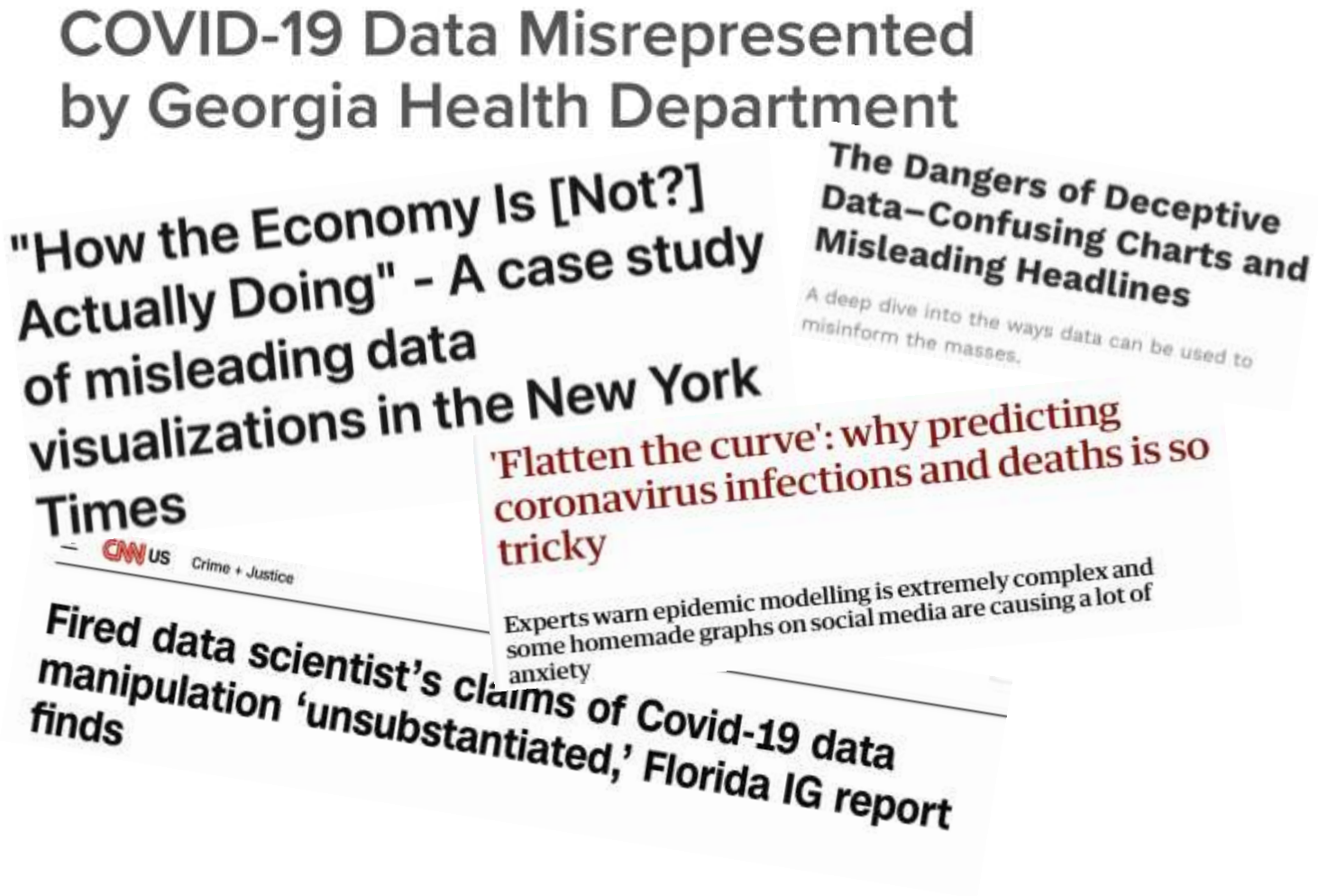
INTERCHART: Benchmarking Visual Reasoning Across Decomposed and Distributed Chart Information

Anirudh Iyengar Kaniyar Narayana Iyengar [1*] , Srija Mukhopadhyay [2*] ,
Adnan Qidwai [2*], Shubhankar Singh [3], Dan Roth [4], Vivek Gupta [1]

[1]*Arizona State University,* [2]*IIIT, Hyderabad,* [3]*Mercer Mettl,* [4]*University of Pennsylvania*
**Equal Contribution*

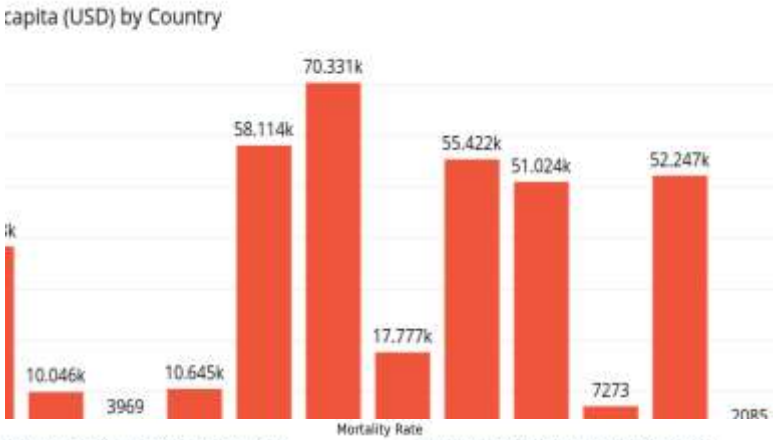
ASU Mercer INTERNATIONAL INSTITUTE OF INFORMATION TECHNOLOGY HYDERABAD

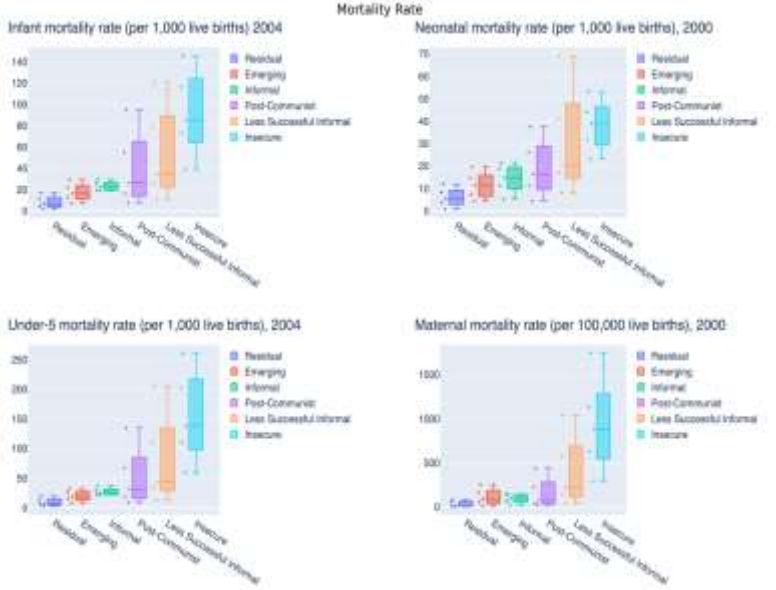
When Charts Mislead, Policies Fail ?

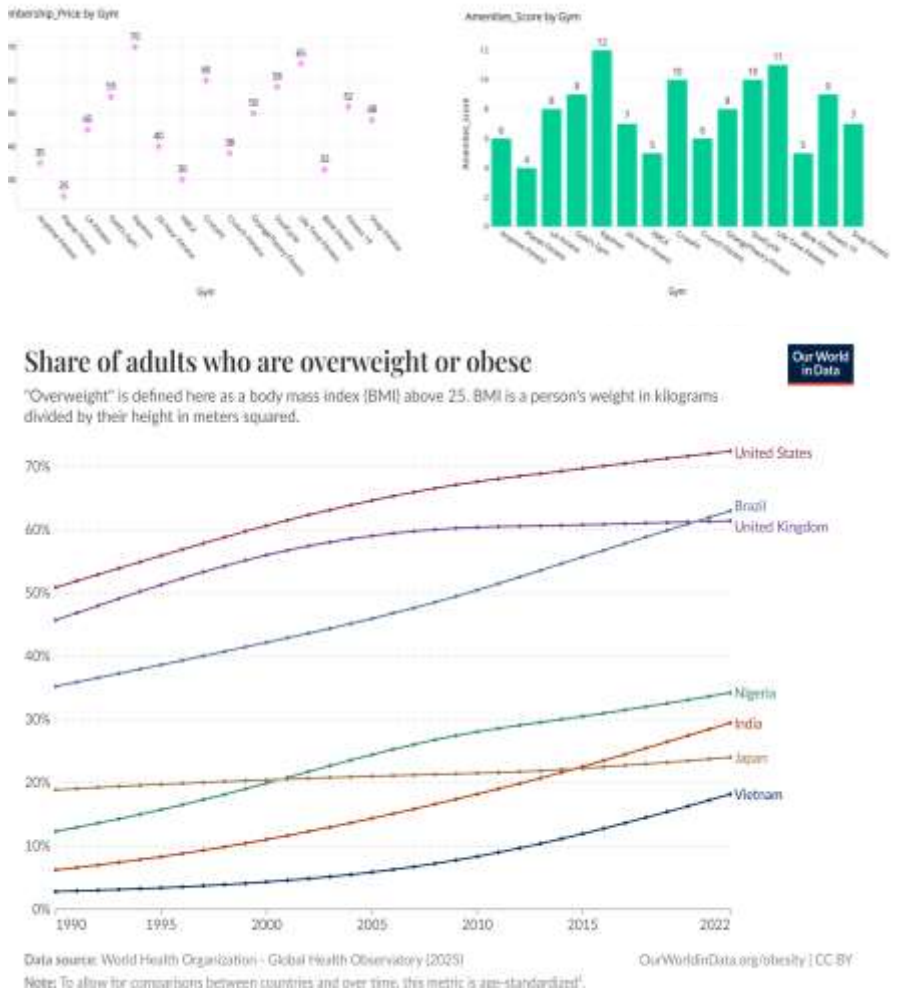


- Real-world decisions rely on multiple charts, not isolated visuals.
- Misinterpreting distributed visuals can lead to wrong conclusions, misleading headlines, and policy errors.
- Headlines often show contradictory or incomplete visuals.
- Small misalignments in axes, time ranges, or metrics can produce misleading narratives.

Charts are Everywhere, But Can AI Really Understand Them?







Multi-chart reasoning isn't just vision; it's structured analytical thinking.

Introducing INTERCHART

Task: Evaluate VLMs on distributed, multi-chart understanding.

Dataset: Three tiers: DECAF (simple), SPECTRA (relational), STORM (real-world multi-chart).

Benchmark: 12k QA pairs, multiple prompting strategies, LLM-as-Judge evaluation.

Question and Chart Complexity:

- Measures performance across increasing visual + reasoning difficulty:
  - Local lookup → relational comparison → multi-step temporal synthesis
- Tests abilities that current VLMs consistently fail:
  - Trend alignment, multi-chart correlation, cross-entity tracking, time-based reasoning

Dataset- DECAF

DECAF - Decomposed Elementary Charts with Answerable Facts

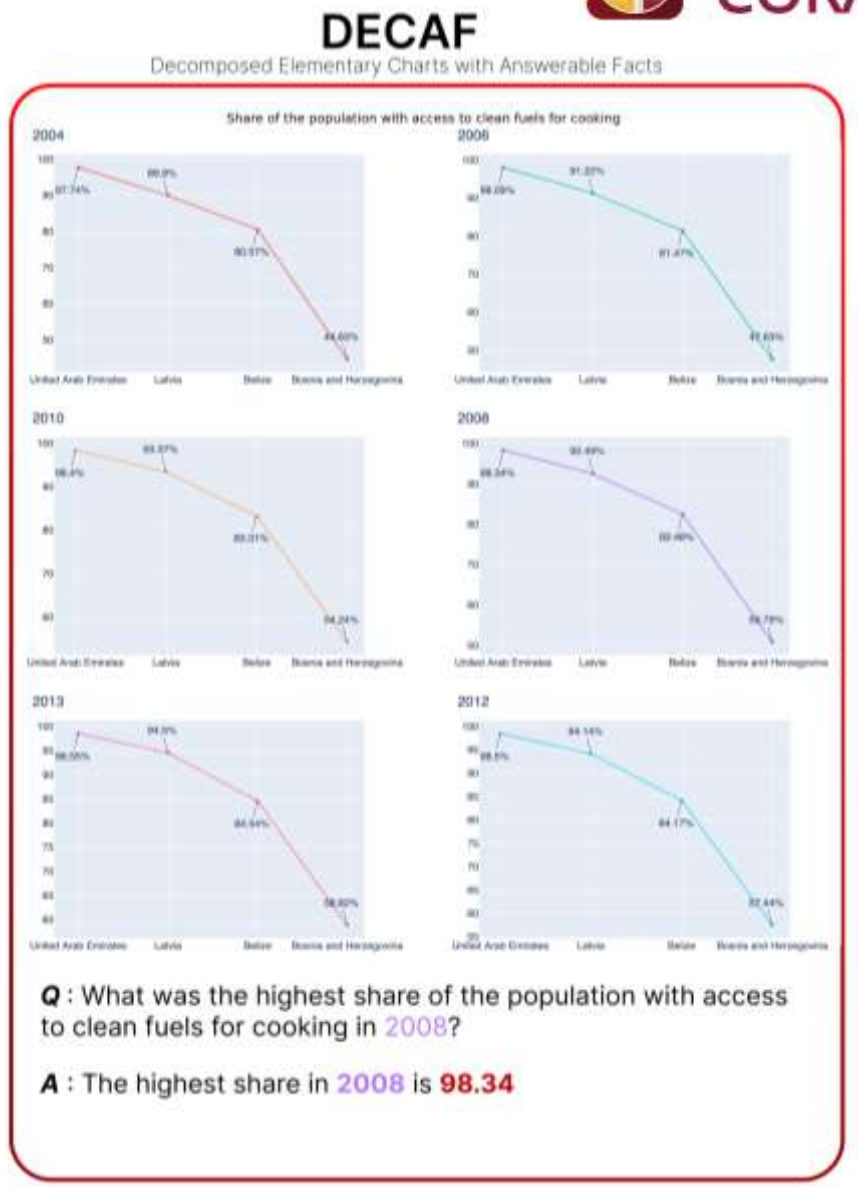
What DECAF Evaluates

- Basic fact extraction
- Trend identification within a single chart
- Reading axes, labels, and numeric values

Data Composition

- 355 original charts → 1,188 decomposed charts
- Multiple chart types: line, bar, heatmap, dot, box plot
- 2,809 QA pairs from mixed generation methods (human, LLM, SQL-LLM) Purpose

DECAF provides a controlled baseline for evaluating fundamental visual understanding before introducing relational or temporal complexity.



Dataset- SPECTRA

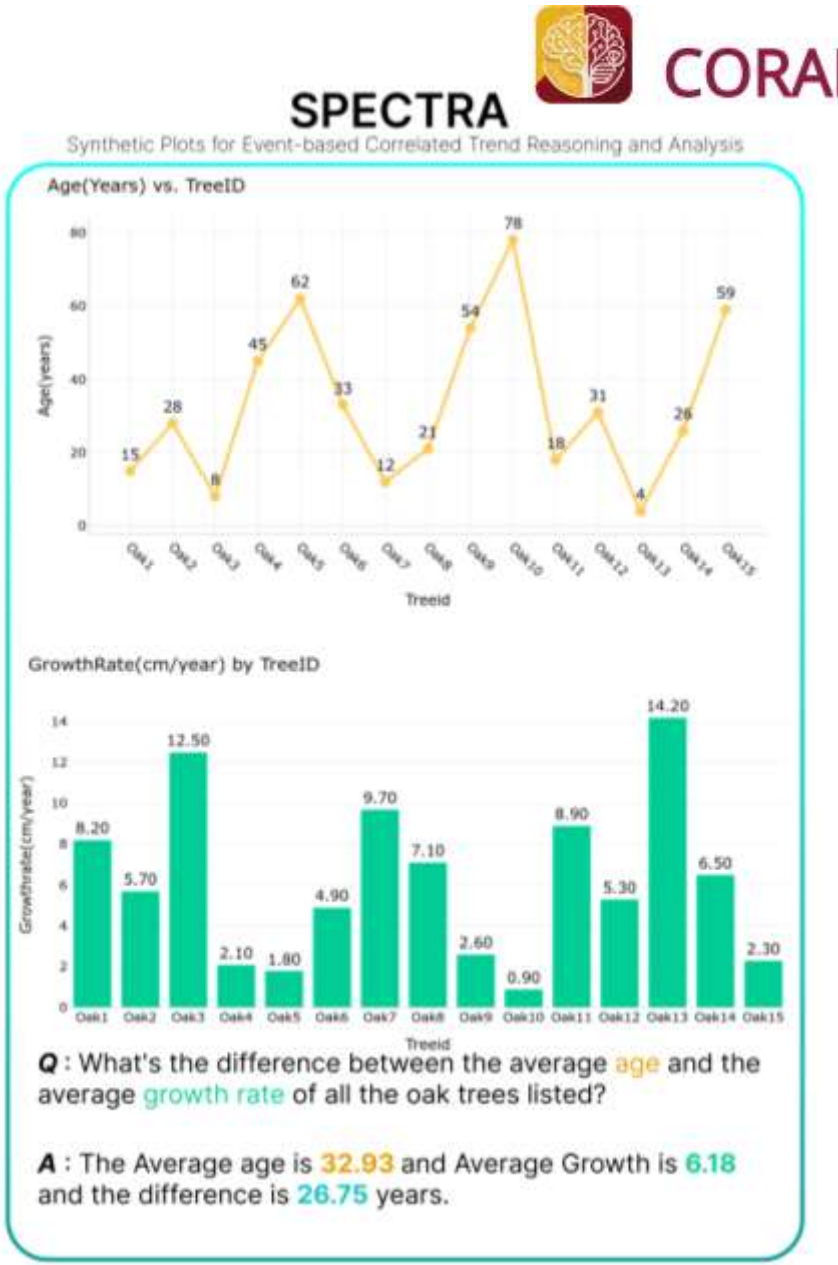
SPECTRA - Synthetic Plots for Event-based Correlated Trend Reasoning and Analysis

What SPECTRA Evaluates

- Identifying correlated vs independent trends
- Comparing slopes, averages, and ranges
- Multi-step reasoning across two aligned charts

Data Composition

- 870 unique charts (paired into 333 context sets)
- 1,717 QA pairs, Includes correlated (1,481) and independent (245) chart pairs



SPECTRA introduces relational complexity, measuring a model's ability to synthesize information across multiple visual sources.

Dataset- STORM

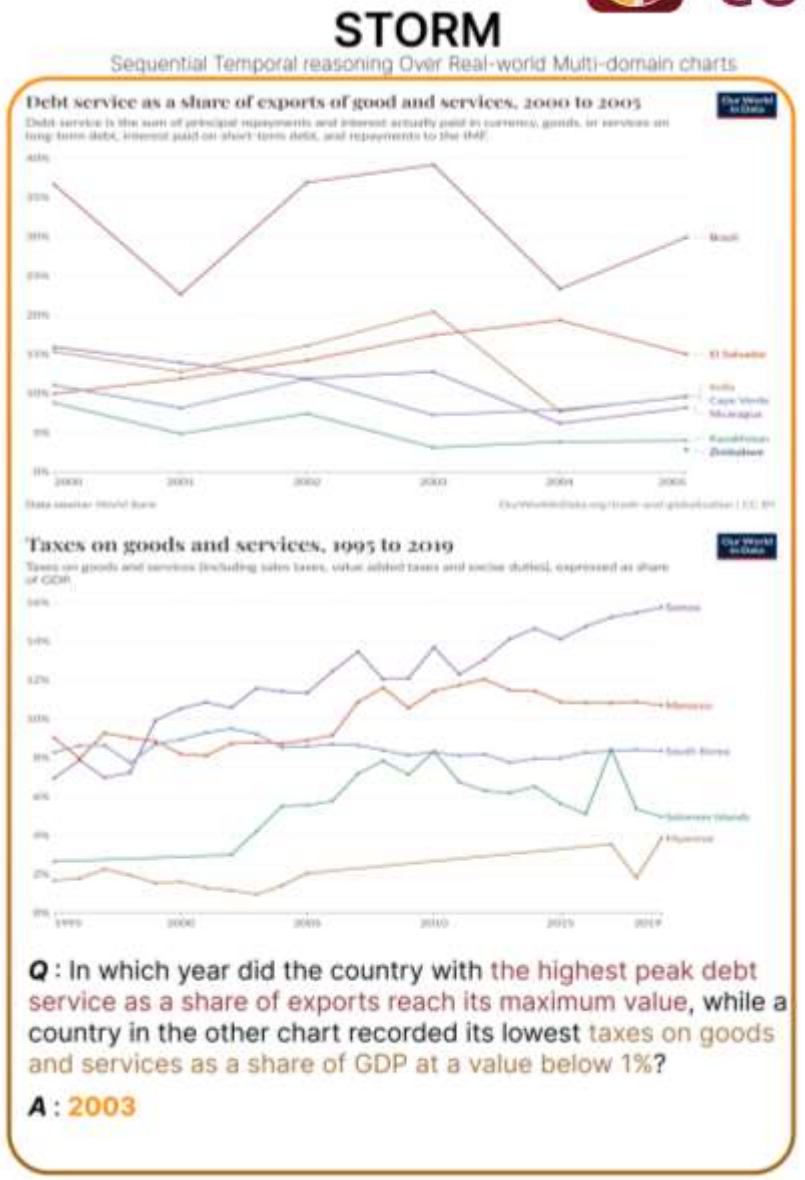
STORM - Sequential Temporal reasoning Over Real-world Multi-domain charts

What STORM Evaluates

- Matching events across different timelines
- Inferring numerical ranges and extrema
- Linking economic, demographic, or health indicators across domains
- Multi-step reasoning across visually and semantically distinct charts

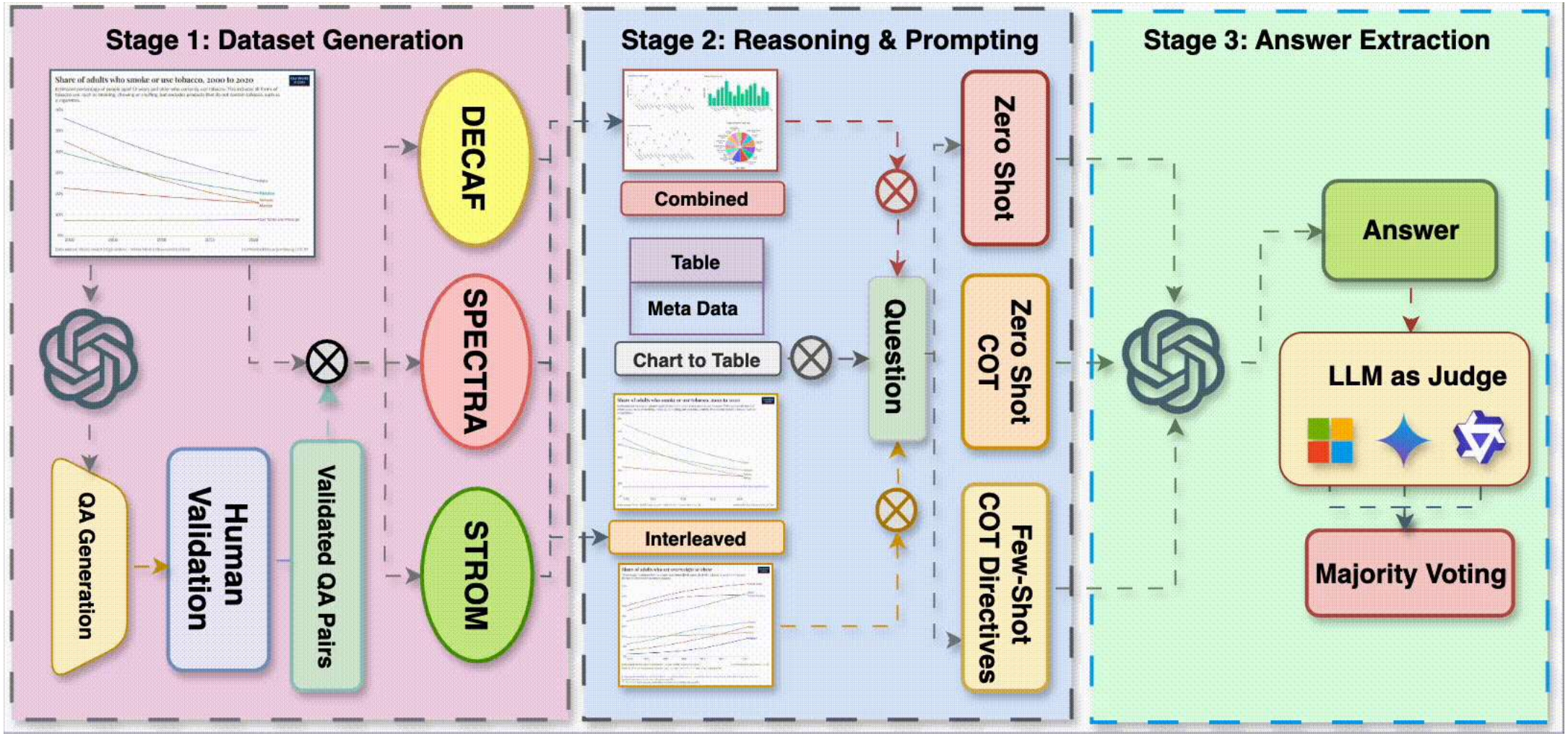
Data Composition

- Data Composition 324 original charts → 648 curated chart images
- 768 QA pairs across three reasoning types:
  - Range Estimation (198)
  - Abstract Numerical Reasoning (275)
  - Entity Inference (295)



STORM represents the highest difficulty tier, mirroring real-world analytical tasks that require integrating distributed, heterogeneous visual information.

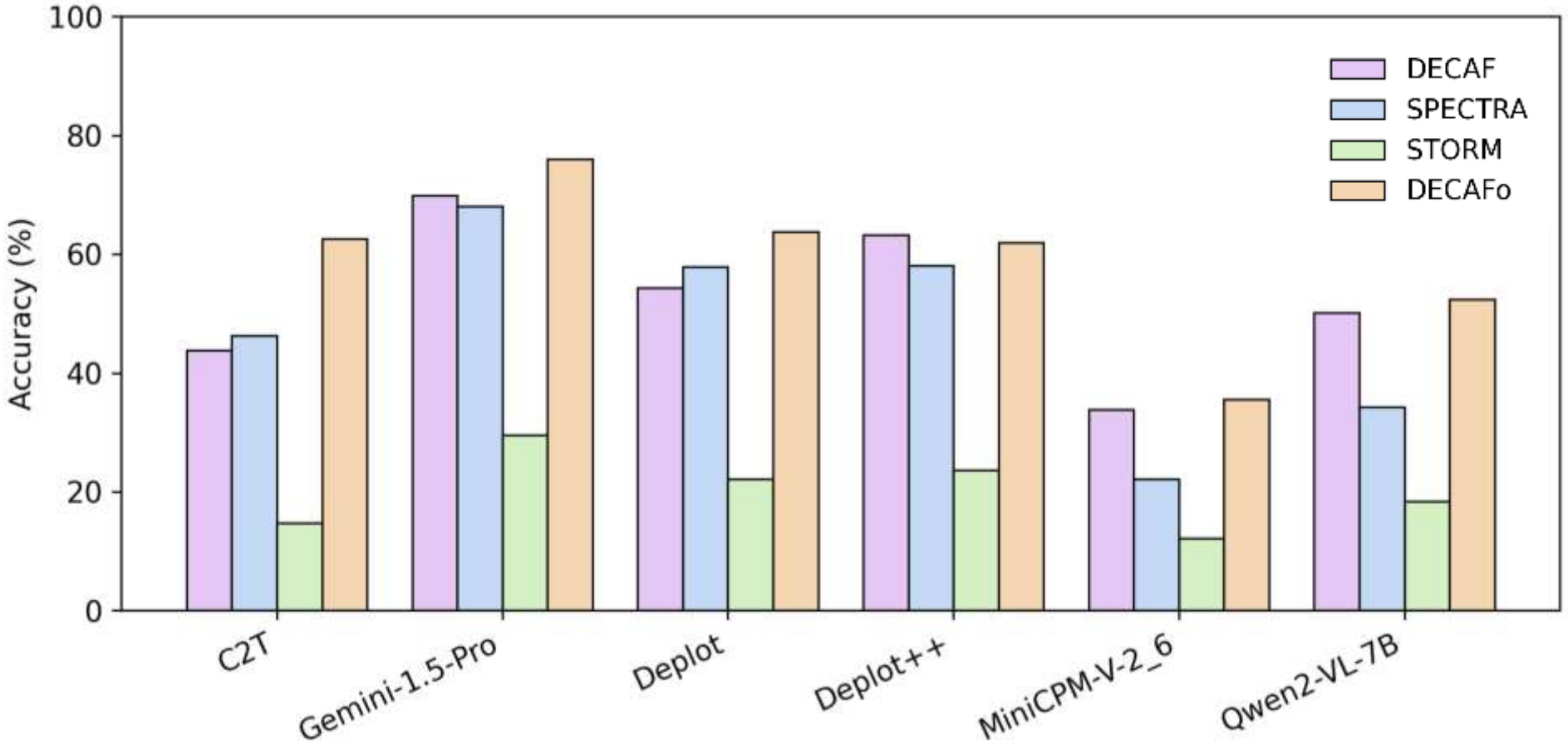
INTERCHART Architecture



Results: Interleaved Visual Context

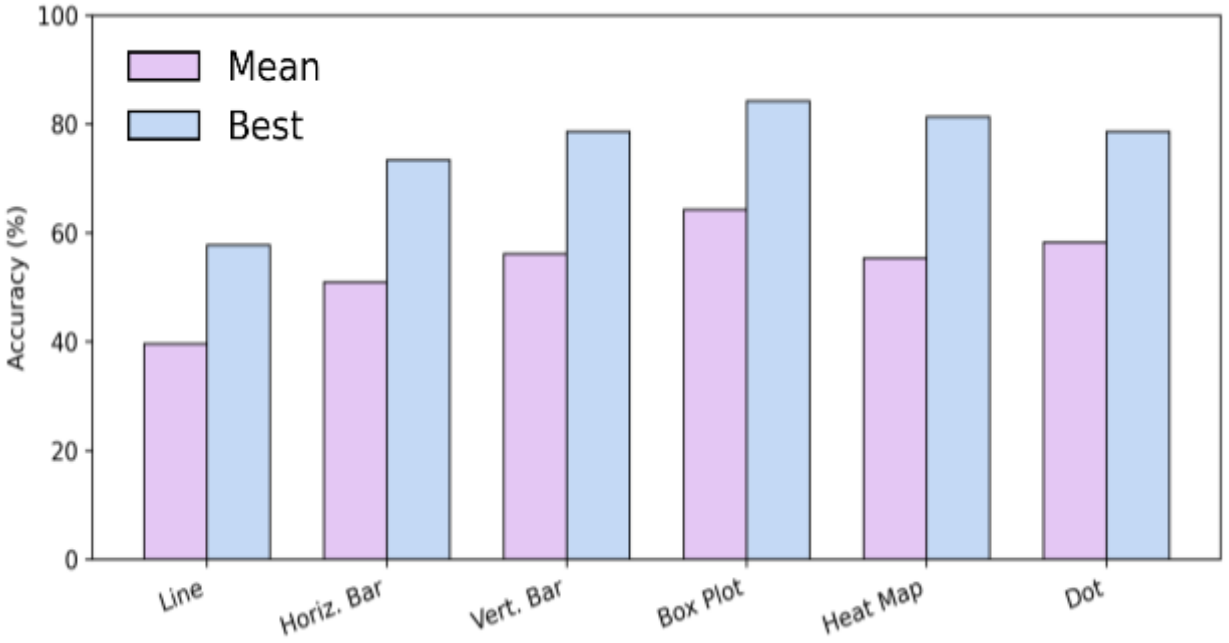
Results: Combined Visual Context

Results: Chart To Table

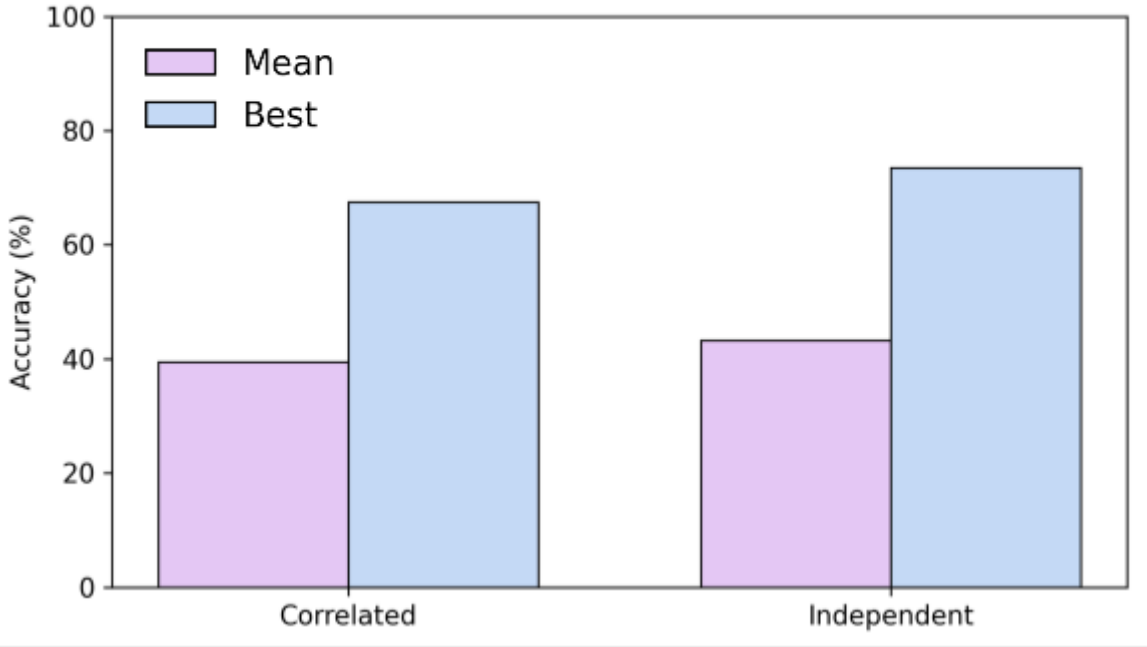


Results: DECAF and SPECTRA

DECAF by chart type (Mean / Best):

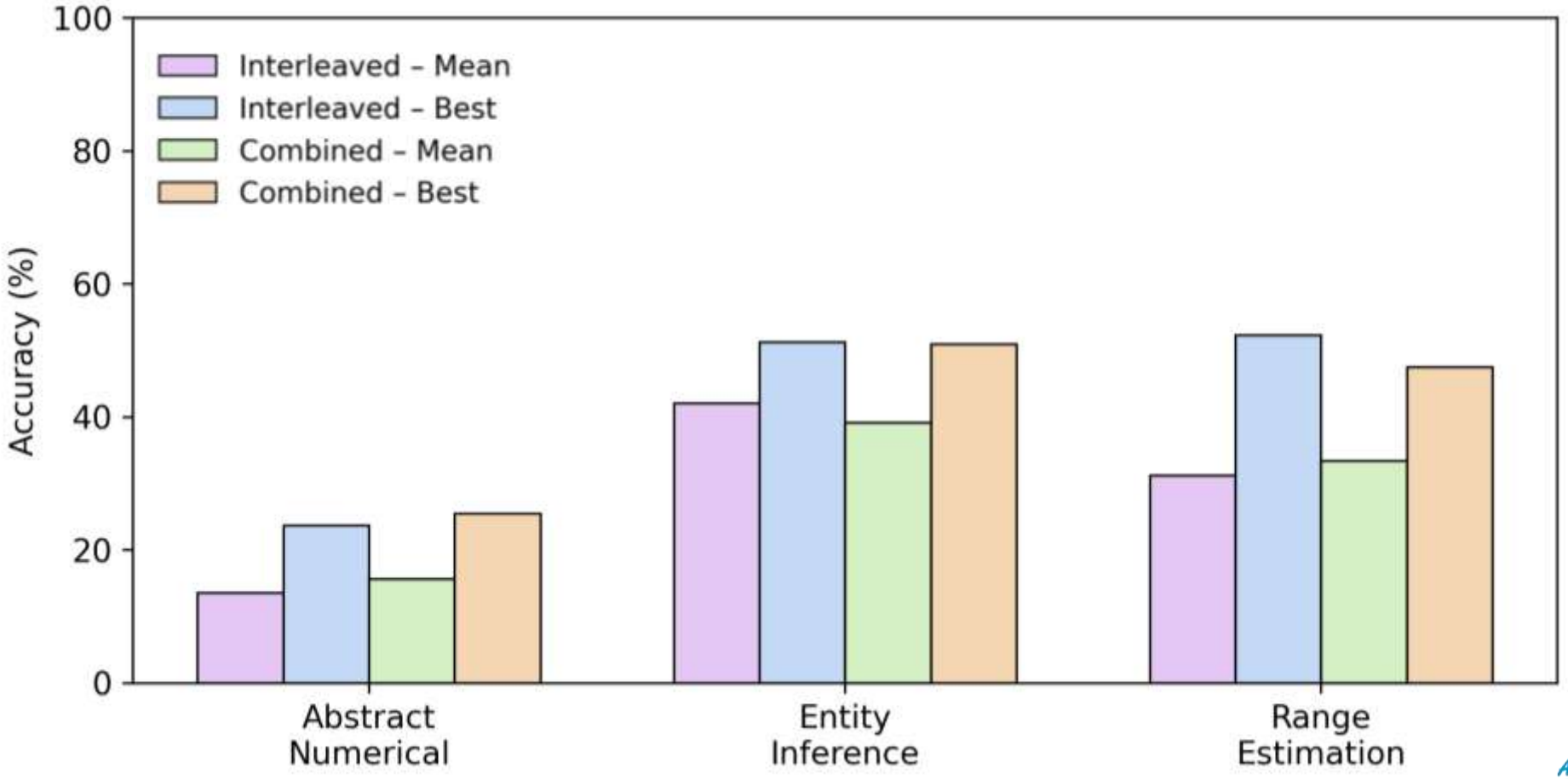


SPECTRA question categories
(Correlated vs. Independent; Mean / Best).



Results: STROM

STORM reasoning types (Abstract Numerical, Entity Inference, Range Estimation) under Interleaved vs. Combined formats (Mean / Best).



Comparison in Chart-Based VQA

Dimension	InfoChartQA	ChartMind	ChartQAPro	INTERCHART
Chart Type	Infographics	Mixed	Plots	**Plots**
Multi-Chart	No	Limited	No	**Yes**
Real-World Data	Yes	Yes	Yes	**Yes**
Semantic Drift	Medium	Medium	Low	**High**
Temporal Reasoning	Low	Medium	Low	**High**
Visual Diversity	High	High	Low	**High**
QA Type	Factoid	Hybrid	Factual	**Fact + Inference**
Evaluation Method	BLEURT	BLEU/LLM	Exact Match	**LLM Majority Voting**

Takeaways

1. INTERCHART exposes *systematic failures* in current VLMs, especially when reasoning must integrate information across multiple heterogeneous charts.

2. Decomposing complex visuals into simpler units significantly boosts performance, highlighting that models still rely on localized rather than global reasoning.

3. Accuracy drops sharply from synthetic (SPECTRA) to real-world multi-chart settings (STORM), revealing poor generalization to semantic drift and temporal alignment.

4. Even top-tier VLMs plateau on STORM, emphasizing the need for new architectures that explicitly model cross-chart reasoning, not just visual parsing.

Thank you

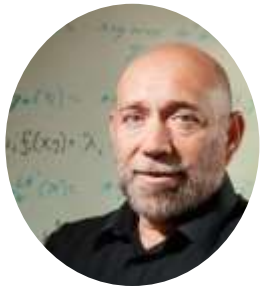
Team

Anirudh Iyengar

Srija Mukhopadhyay

Adnan Qidwai

Shubhankar Singh



Dan Roth

Vivek Gupta

Scan for the website