[1] 浙江大學 ZHEJIANG UNIVERSITY

[2] NUS National University of Singapore

[3] Tencent 腾讯

AAAI-26 / IAAI-26 / EAAI-26

Multimodal DeepResearcher: Generating Text-Chart Interleaved Reports From Scratch with Agentic Framework

Zhaorui Yang[1], Bo Pan[1], Han Wang[1], Yiyao Wang[1], Xingyu Liu[1], Luoxuan Weng[1]
Yingchaojie Feng[2], Haozhe Feng[3], Minfeng Zhu[1], Bo Zhang[1], Wei Chen[1]

Oral Presentation. Presented by Zhaorui Yang

The Landscape of Deep Research

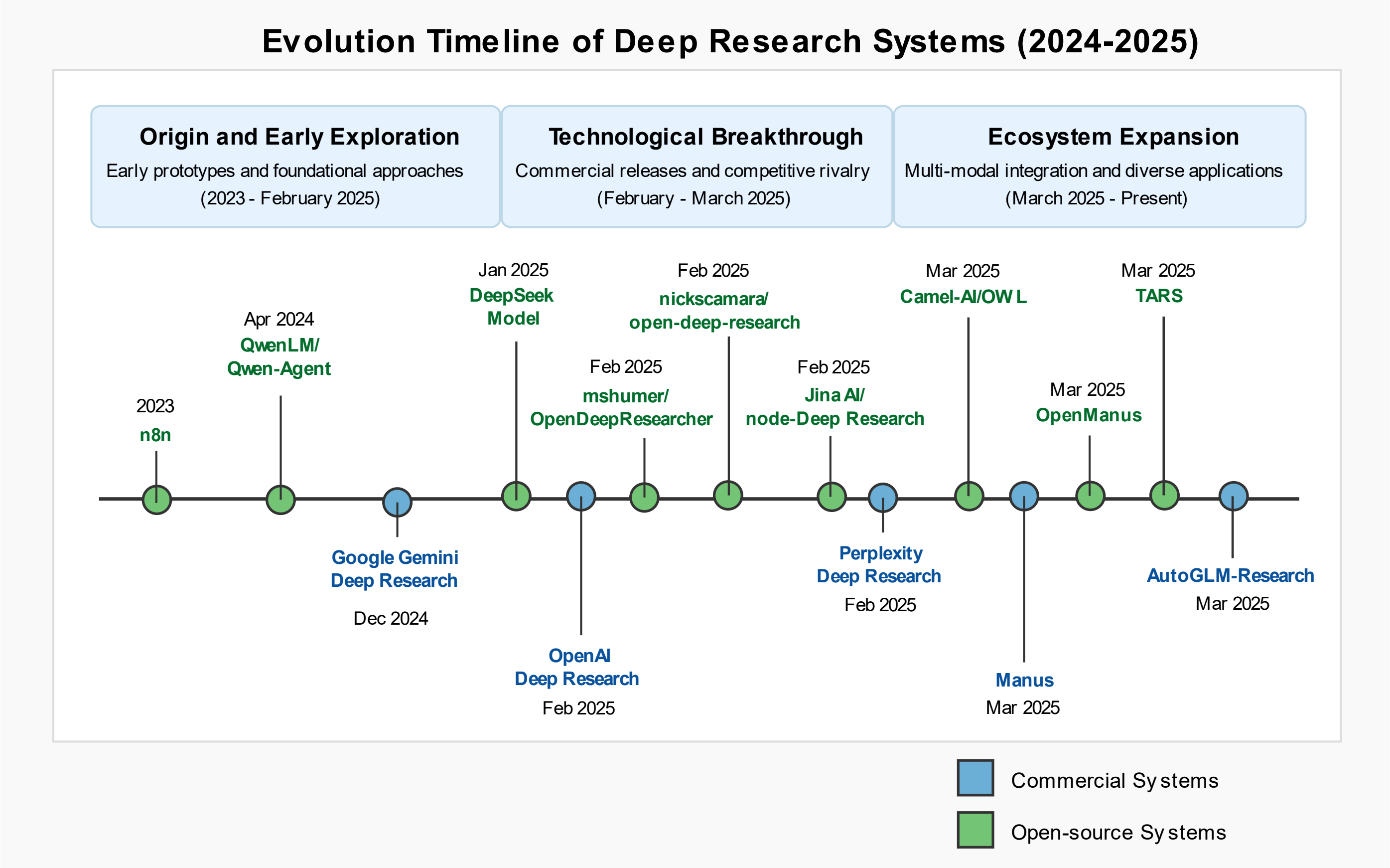
Deep Research 🔍 has garnered significant attention 🔥 from both academia and industry

Generates comprehensive reports 📃 from scratch via search, reasoning, etc.

「*deep research marks a significant step toward our broader goal of developing AGI*」

Source: OpenAI

Products, frameworks, and papers sprout 🌱 frequently



Evolution Timeline of Deep Research Systems (2024-2025)

Origin and Early Exploration
Early prototypes and foundational approaches
(2023 - February 2025)

Technological Breakthrough
Commercial releases and competitive rivalry
(February - March 2025)

Ecosystem Expansion
Multi-modal integration and diverse applications
(March 2025 - Present)

2023 n8n

Apr 2024 QwenLM/ Qwen-Agent

Google Gemini Deep Research Dec 2024

Jan 2025 DeepSeek Model

OpenAI Deep Research Feb 2025

Feb 2025 mshumer/ OpenDeepResearcher

Feb 2025 nickscamara/ open-deep-research

Feb 2025 Jina AI/ node-Deep Research

Perplexity Deep Research Feb 2025

Mar 2025 Camel-AI/OWL

Manus Mar 2025

Mar 2025 OpenManus

Mar 2025 TARS

AutoGLM-Research Mar 2025

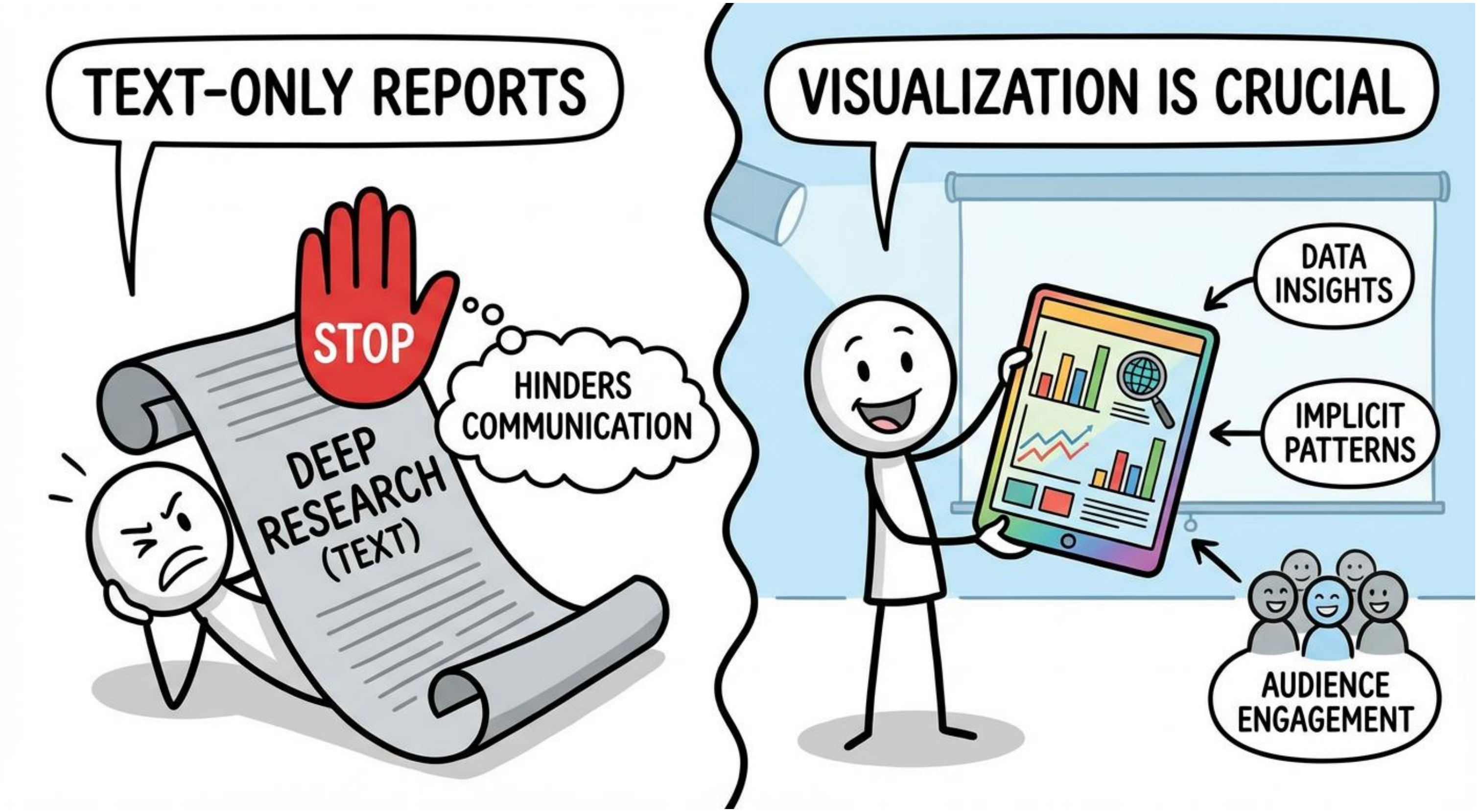
🟦 Commercial Systems

🟩 Open-source Systems

Source: https://arxiv.org/pdf/2506.12594

Issue of Current Paradigm: Effective Communication

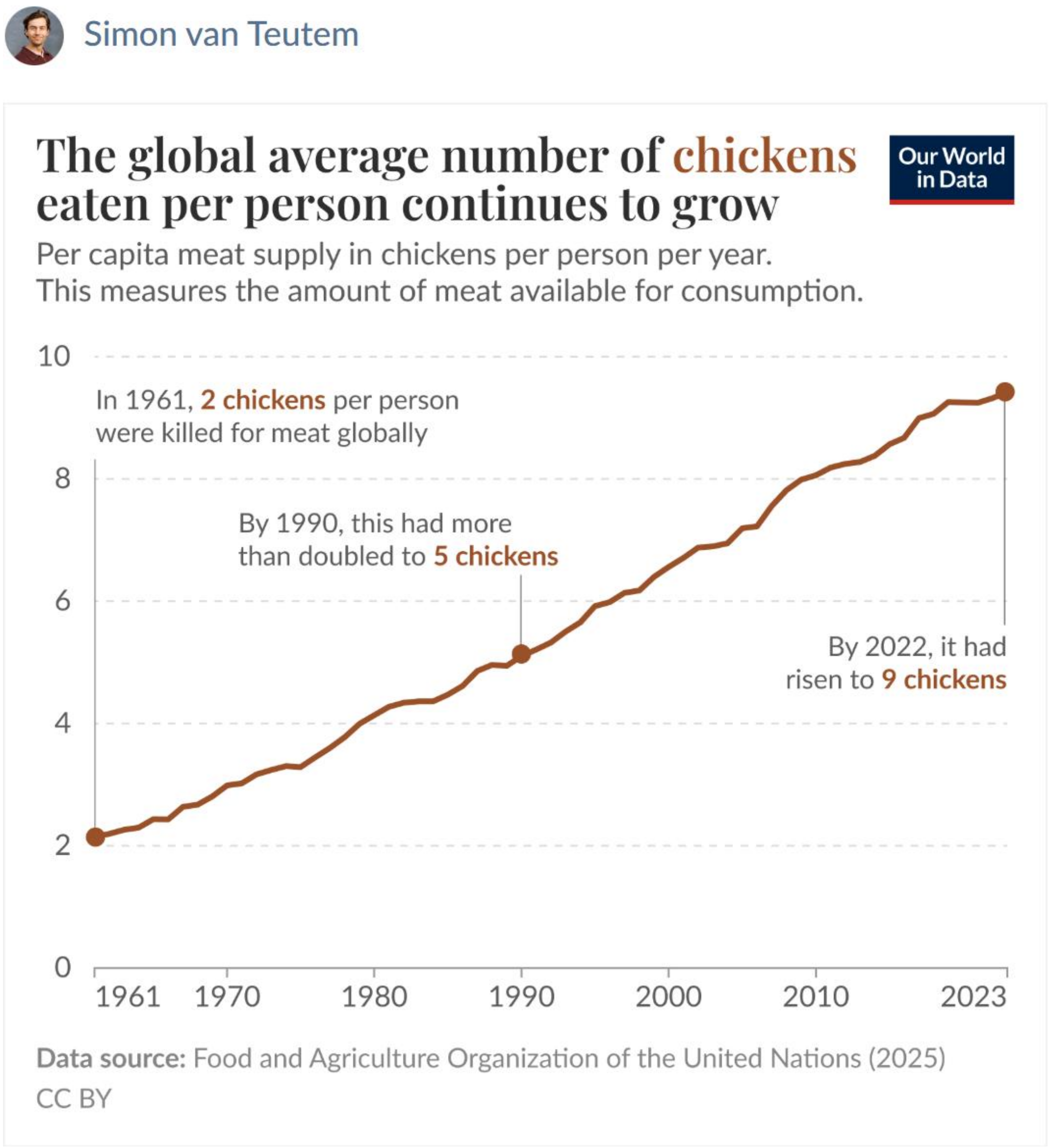
- Existing works focus on **text-only** content, which hinders ✋ effective communication.
- Visualization is crucial in real-world
  - Conveying data insights (Otten et al. 2015)
  - Facilitate identify implicit patterns (Yang et al. 2024)
  - Enhance audience engagement (Zheng et al. 2025a)



Made with Nano-Banana-Pro

How Humans Create Reports

- Humans create coherent reports with **interleaved texts and visualizations** 📉 .
  - Meticulously design visualizations, iteratively refine them 🔁 if needed.
  - Integrate charts within appropriate textual context and maintain consistency.
- Can **agents** generate such multimodal reports? 🤔



Simon van Teutem

The global average number of chickens eaten per person continues to grow

Per capita meat supply in chickens per person per year.
This measures the amount of meat available for consumption.

In 1961, 2 chickens per person were killed for meat globally

By 1990, this had more than doubled to 5 chickens

By 2022, it had risen to 9 chickens

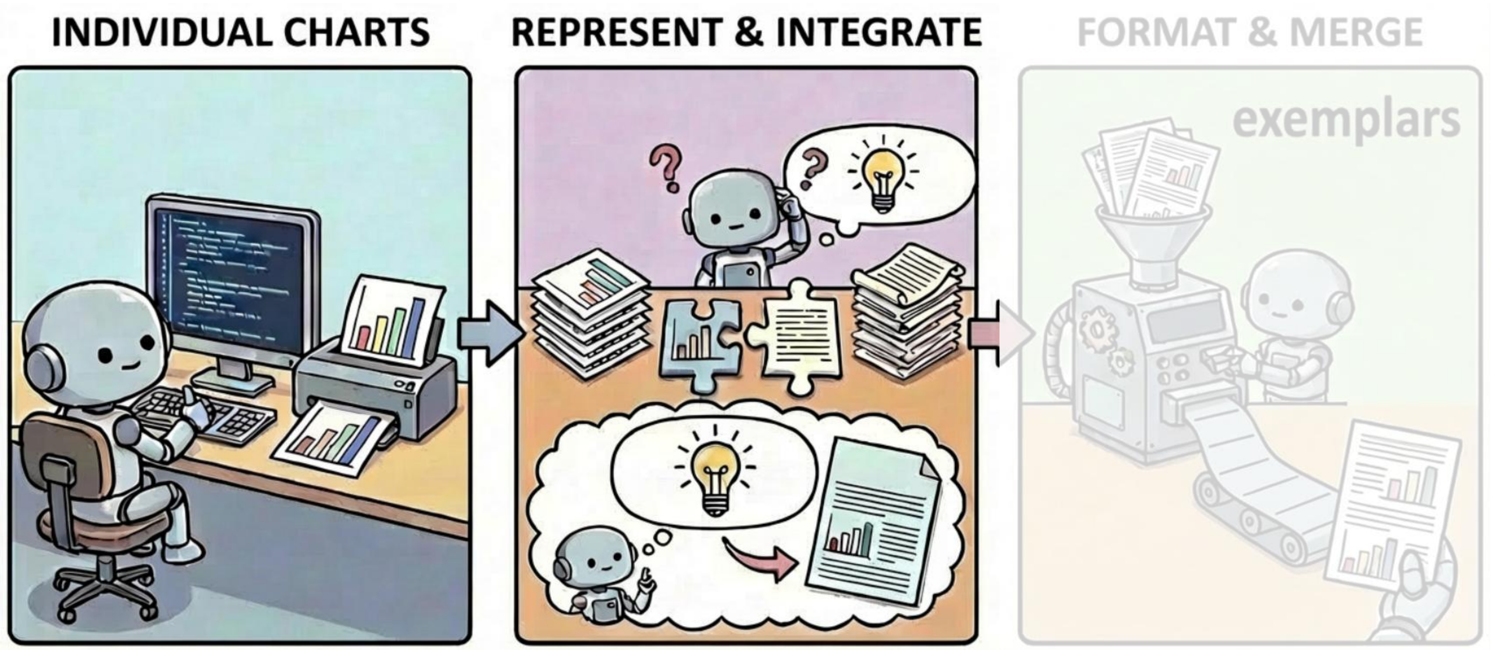
Data source: Food and Agriculture Organization of the United Nations (2025)
CC BY

Source: Our World in Data

Challenges of Generating Interleaved Reports

LLMs are *already* good at generating *individual* charts through *coding*

- How to **represent and integrate** them with texts?
- How to maintain **consistency**?
  - Charts match with texts
  - Charts have a unified style

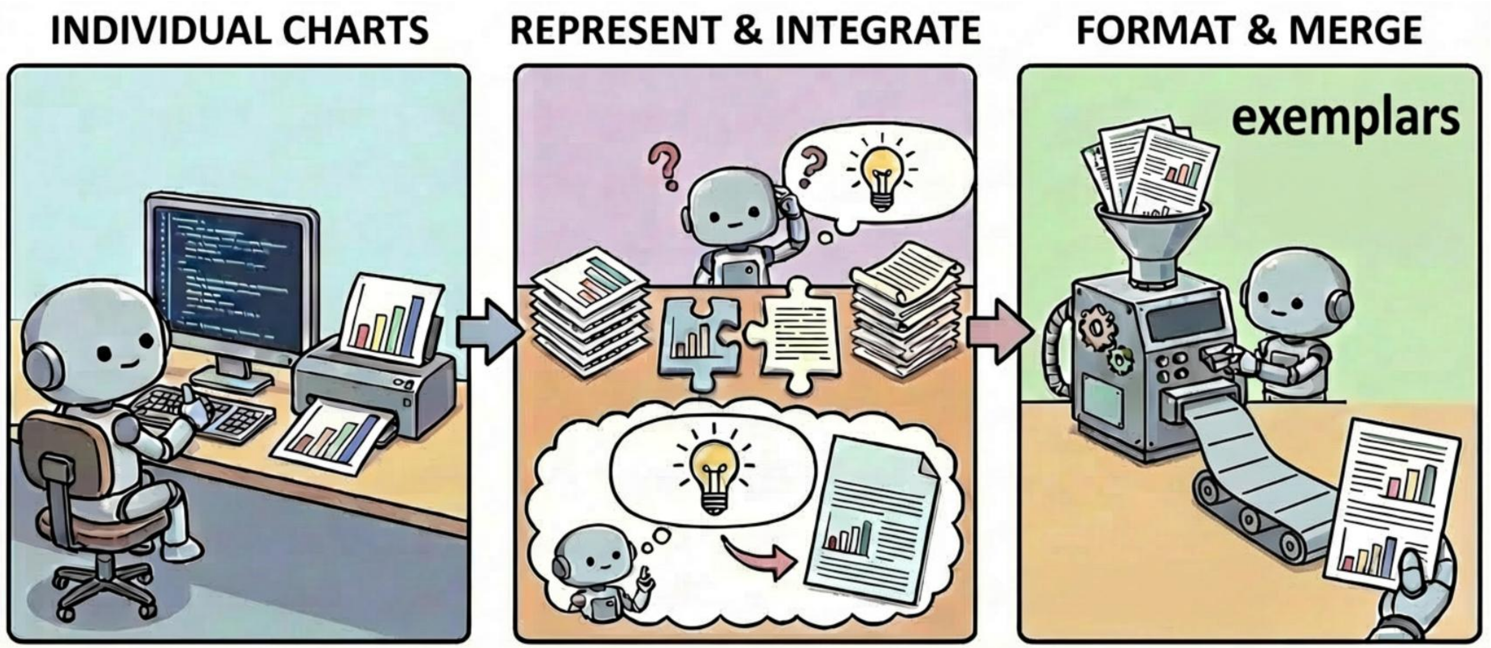


Made with Nano-Banana-Pro

Challenges of Generating Interleaved Reports (Cont.)

In-context learning seems promising:

- Exemplars should be **multimodal**
- Outputs should be in the same form
- Need a unified **representation** for both exemplars and outputs



Made with Nano-Banana-Pro

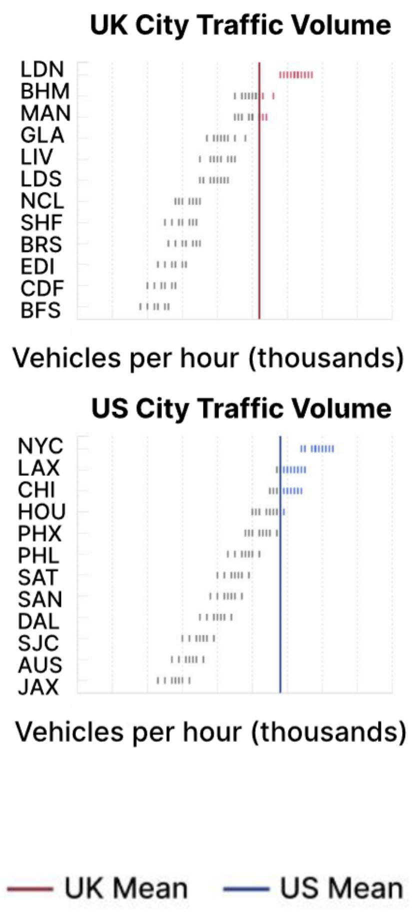
Introducing FDV and Multimodal DeepResearcher

For representation, we introduce the **Formal Description of Visualization (FDV)**, a structured representation method.

Building upon FDV, we introduce **Multimodal DeepResearcher**, an agentic framework for end-to-end interleaved generation.

Representation: Formal Description of Visualization

(A) Origin Visualization



Extract Design

(B) Formal Description of Visualization

Layout

- The visualization consists of two similar strip plots stacked vertically.
- Each plot has a title at the top ('UK City Traffic Volume' and 'US City Traffic Volume').
- The overall chart has a shared legend at the bottom showing 'UK City Mean' and 'US City Mean' with corresponding colored lines.
- Each plot has adequate margins on all sides, with city names aligned on the left side.

Scale

- X-axis: Linear scale from 0 to 9, representing 'Vehicles per hour (thousands)'.
- X-axis has grid lines at 1-unit intervals (1, 2, 3, etc.).
- X-axis label 'Vehicles per hour (thousands)' is placed at the bottom of each plot.
- Y-axis: Categorical scale showing city names, evenly spaced vertically.
- No y-axis title is shown, just the city names as tick labels aligned to the left.
- Color: All marks shown in grey by default, except that values above the UK mean for UK cities are marked in red, and values above the US mean are marked in blue.

Data

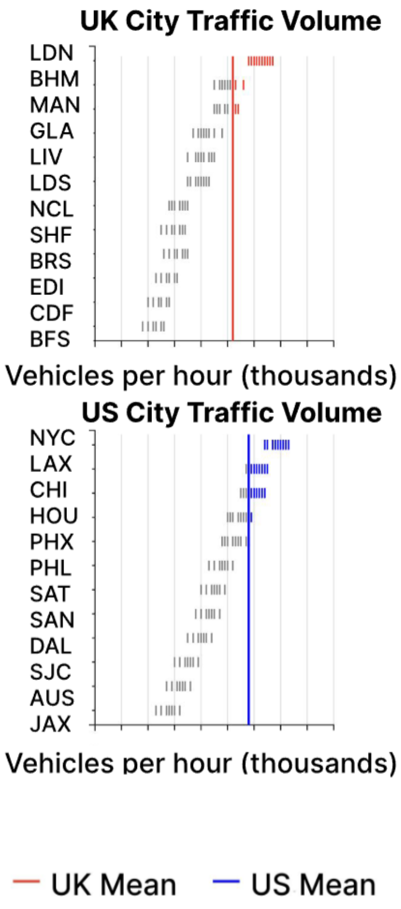
- For each city, there are multiple traffic volume measurements, represented as small marks.
- The mean traffic volume for each country is calculated and visualized as vertical lines.

Marks

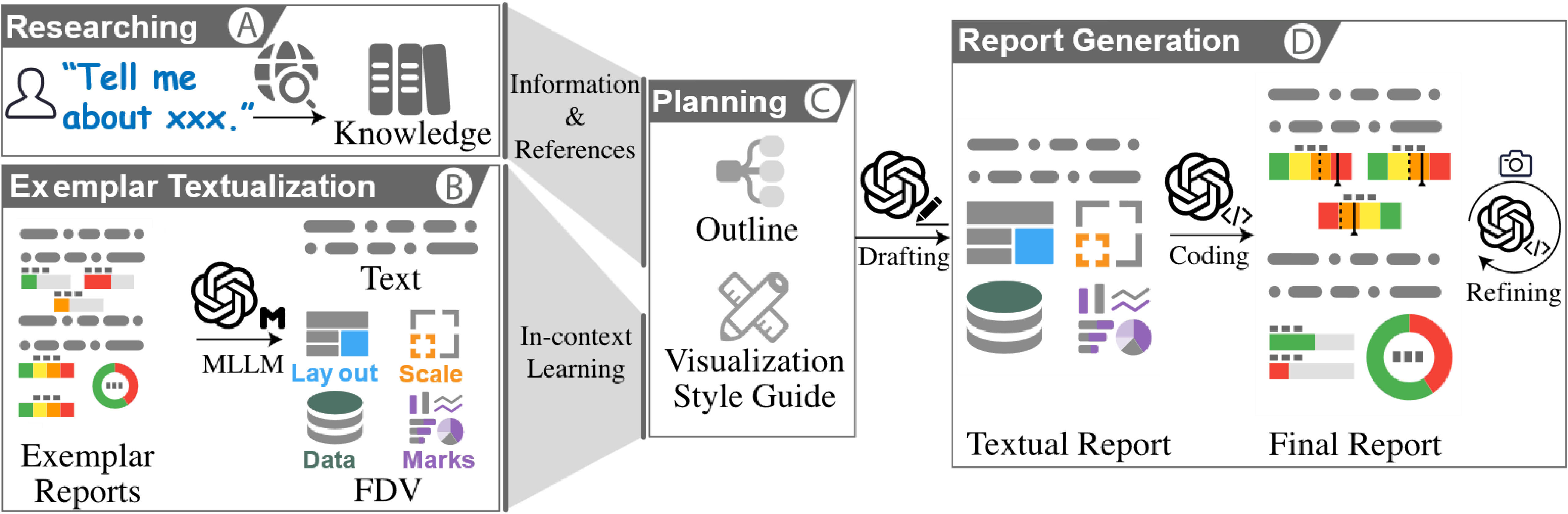
- Small tick marks (resembling small vertical lines) represent individual traffic volume measurements for each city, with colors indicating both the country and whether values are above or below the mean.
- A vertical red line represents the UK mean traffic volume in the top plot.
- A vertical blue line represents the US mean traffic volume in the bottom plot.

Implement Design

(C) Reconstructed



Framework: Multimodal DeepResearcher



Four Stages: (1) Researching; (2) Exemplar Textualization; (3) Planning; (4) Report Generation (with refinements)

Experimental Settings

Input

100 real-world topics from public websites

Baseline

Adapted from DataNarrative
(data as inputs) *(Islam et al. 2024)*

Criteria

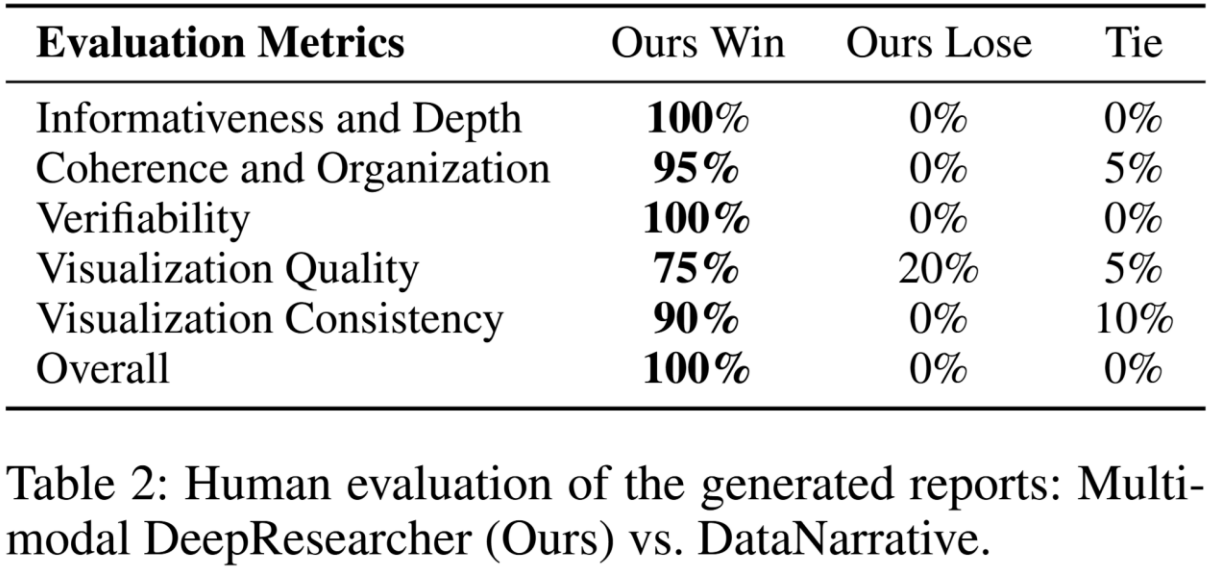
Score & Pair-Wise Comparison

Evaluation

- MLLM as a judge & Human Eval
- Report level & Chart level

Experiments: Report-Level Results

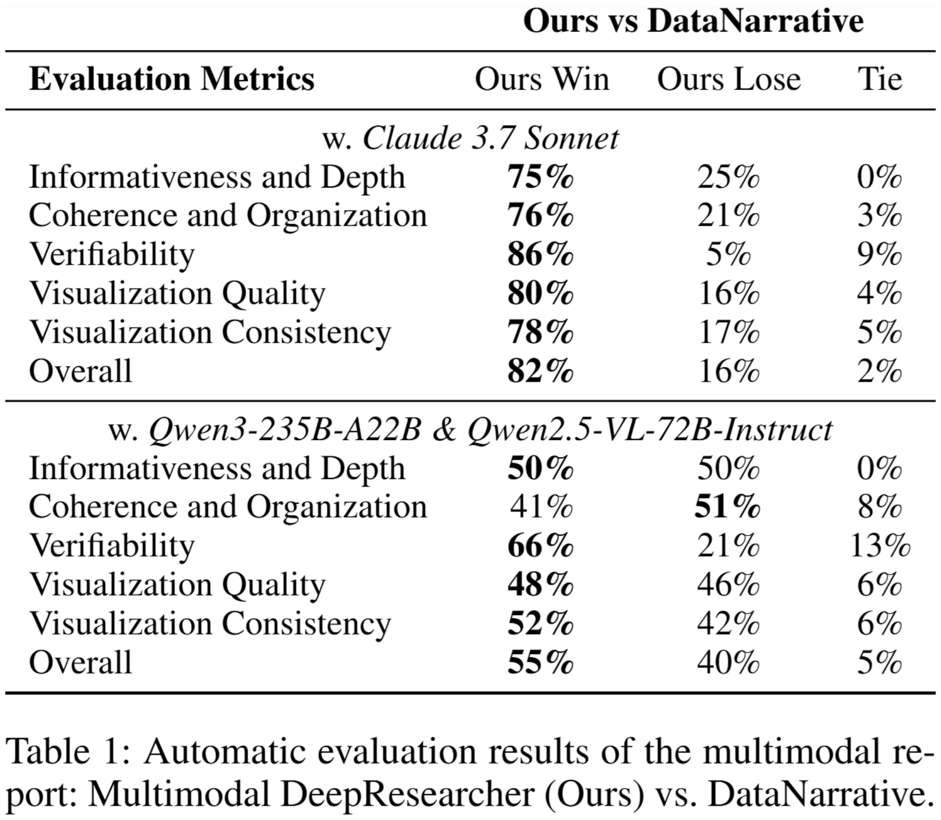
Multimodal DeepResearcher consistently **outperforms** DataNarrative with both **auto** & **human** eval



Evaluation Metrics	Ours Win	Ours Lose	Tie
Informativeness and Depth	**100%**	0%	0%
Coherence and Organization	**95%**	0%	5%
Verifiability	**100%**	0%	0%
Visualization Quality	**75%**	20%	5%
Visualization Consistency	**90%**	0%	10%
Overall	**100%**	0%	0%

Table 2: Human evaluation of the generated reports: Multimodal DeepResearcher (Ours) vs. DataNarrative.

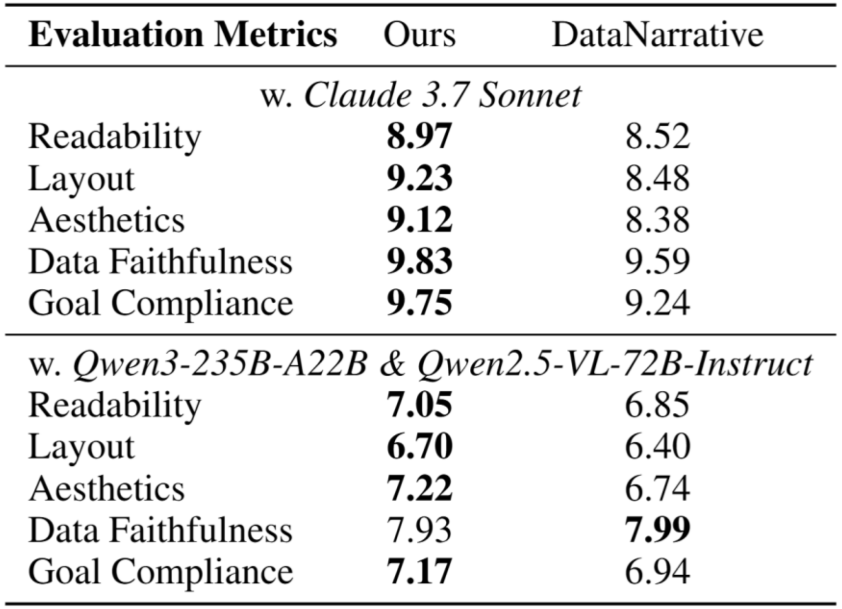
Results with 5 evaluators on a subset of 20 report pairs



	Ours vs DataNarrative		
Evaluation Metrics	Ours Win	Ours Lose	Tie
w. *Claude 3.7 Sonnet*			
Informativeness and Depth	**75%**	25%	0%
Coherence and Organization	**76%**	21%	3%
Verifiability	**86%**	5%	9%
Visualization Quality	**80%**	16%	4%
Visualization Consistency	**78%**	17%	5%
Overall	**82%**	16%	2%
w. *Qwen3-235B-A22B & Qwen2.5-VL-72B-Instruct*			
Informativeness and Depth	**50%**	50%	0%
Coherence and Organization	41%	**51%**	8%
Verifiability	**66%**	21%	13%
Visualization Quality	**48%**	46%	6%
Visualization Consistency	**52%**	42%	6%
Overall	**55%**	40%	5%

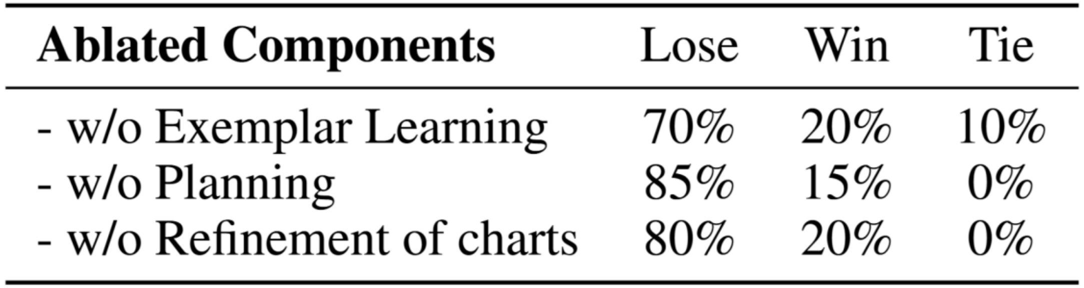
Table 1: Automatic evaluation results of the multimodal report: Multimodal DeepResearcher (Ours) vs. DataNarrative.

Experiments: Chart-Level & Ablations



Evaluation Metrics	Ours	DataNarrative
w. *Claude 3.7 Sonnet*		
Readability	**8.97**	8.52
Layout	**9.23**	8.48
Aesthetics	**9.12**	8.38
Data Faithfulness	**9.83**	9.59
Goal Compliance	**9.75**	9.24
w. *Qwen3-235B-A22B & Qwen2.5-VL-72B-Instruct*		
Readability	**7.05**	6.85
Layout	**6.70**	6.40
Aesthetics	**7.22**	6.74
Data Faithfulness	7.93	**7.99**
Goal Compliance	**7.17**	6.94

Table 3: Evaluation of chart quality. The evaluator assigns a score between 1 to 10 for each metric, and the results are average across all reports.



Ablated Components	Lose	Win	Tie
- w/o Exemplar Learning	70%	20%	10%
- w/o Planning	85%	15%	0%
- w/o Refinement of charts	80%	20%	0%

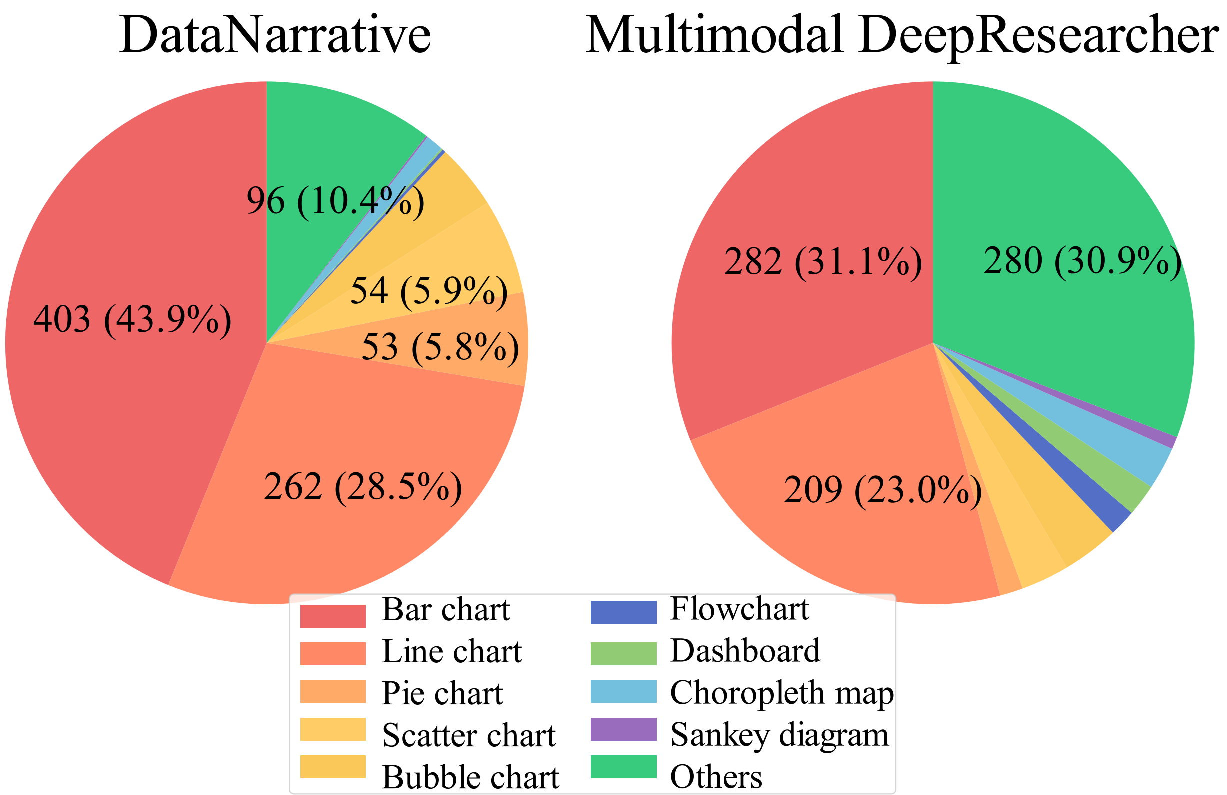
Table 4: Results of ablation studies across three different setups. We report the lose, win and tie rates for each setup against the complete Multimodal DeepResearcher. Claude 3.7 Sonnet serves as both the LLM and MLLM here.

- Consistently outperforms baseline
- Particularly in **layout** & **aesthetics**

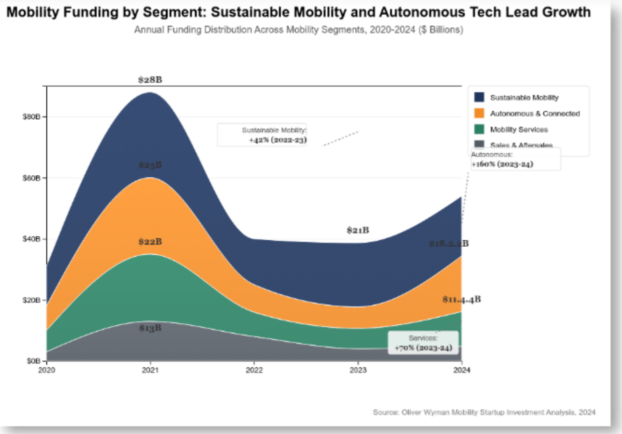
- Removing any results in significant degradation
- Demonstrates the contribution of **each component**

Analysis: Distribution of generated charts

- We present **distribution** of visualization charts generated with both frameworks
- First column in legend: **basic** chart types (warm colors)
- More **diverse** charts: accommodate to diverse real-world scenarios



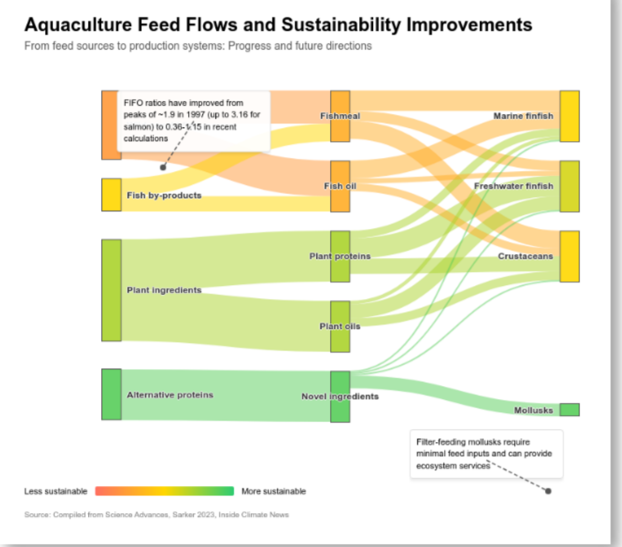
Examples of Visualizations generated


Mobility Funding by Segment: Sustainable Mobility and Autonomous Tech Lead Growth

Annual Funding Distribution Across Mobility Segments, 2020-2024 ($ Billions)

Source: Oliver Wyman Mobility Startup Investment Analysis, 2024

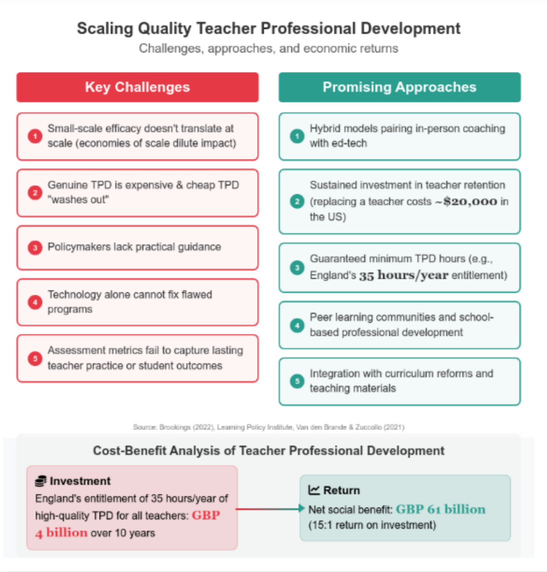
(a) Stacked area chart


Aquaculture Feed Flows and Sustainability Improvements

From feed sources to production systems: Progress and future directions

Source: Compiled from Science Advances, Sarker 2023, Inside Climate News

(b) Sankey diagram


Scaling Quality Teacher Professional Development

Challenges, approaches, and economic returns

Key Challenges

1. Small-scale efficacy doesn't translate at scale (economies of scale dilute impact)
2. Genuine TPD is expensive & cheap TPD "washes out"
3. Policymakers lack practical guidance
4. Technology alone cannot fix flawed programs
5. Assessment metrics fail to capture lasting teacher practice or student outcomes

Promising Approaches

1. Hybrid models pairing in-person coaching with ed-tech
2. Sustained investment in teacher retention (replacing a teacher costs ~**$20,000** in the US)
3. Guaranteed minimum TPD hours (e.g., England's **35 hours/year** entitlement)
4. Peer learning communities and school-based professional development
5. Integration with curriculum reforms and teaching materials

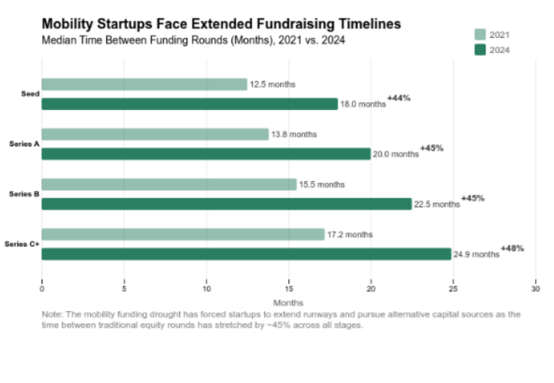
Source: Brookings (2022), Learning Policy Institute, Van den Brande & Zuccollo (2021)

Cost-Benefit Analysis of Teacher Professional Development

Investment
England's entitlement of 35 hours/year of high-quality TPD for all teachers: **GBP 4 billion** over 10 years

Return
Net social benefit: **GBP 61 billion** (15:1 return on investment)

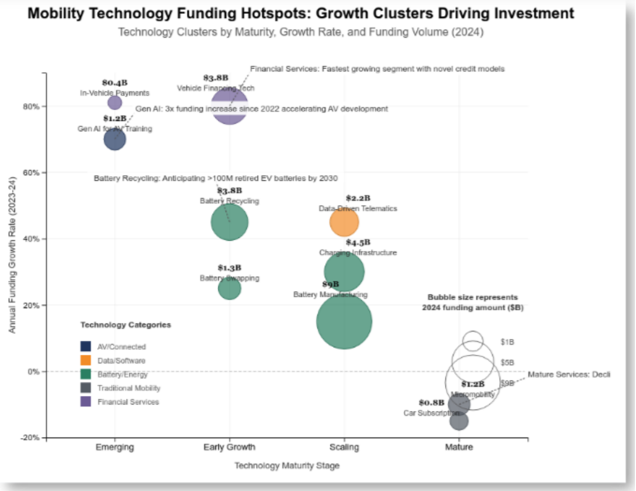
(c) Infographic


Mobility Startups Face Extended Fundraising Timelines

Median Time Between Funding Rounds (Months), 2021 vs. 2024

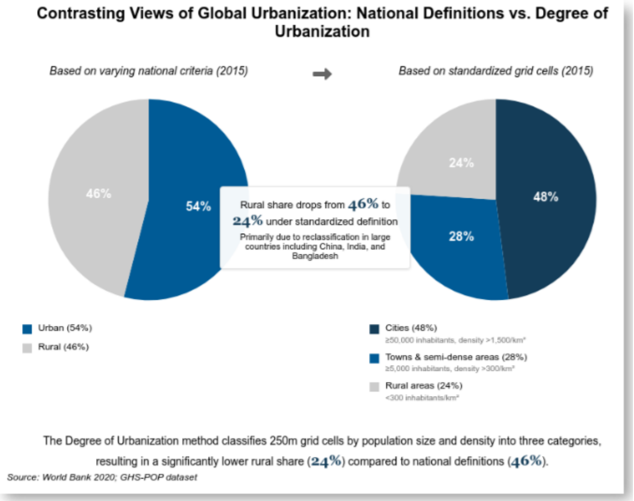
Note: The mobility funding drought has forced startups to extend runways and pursue alternative capital sources as the time between traditional equity rounds has stretched by ~45% across all stages.

(d) Horizontal bar chart


Mobility Technology Funding Hotspots: Growth Clusters Driving Investment

Technology Clusters by Maturity, Growth Rate, and Funding Volume (2024)

(e) Bubble chart


Contrasting Views of Global Urbanization: National Definitions vs. Degree of Urbanization

The Degree of Urbanization method classifies 250m grid cells by population size and density into three categories, resulting in a significantly lower rural share (24%) compared to national definitions (46%).

Source: World Bank 2020; GHS-POP dataset

(f) Pie chart

Conclusion: Contributions

- **Novel task:** Text-chart interleaved report generation from scratch
- **Representation for visualizations:** Formal Description of Visualization (FDV)
- **Framework:** End-to end agentic framework for the task (Multimodal DeepResearcher)

Thanks & QA

Presenter: Zhaorui Yang

zhaorui.yang@zju.edu.cn