FINVQA-CHART: A DIAGNOSTIC BENCHMARK ISOLATING THE MULTIMODAL FUSION BOTTLENECK IN FINANCIAL CHART UNDERSTANDING

Anonymous authors
Paper under double-blind review

ABSTRACT

Vision–Language Models (VLMs) report high aggregate accuracy on multimodal benchmarks, yet a single accuracy number cannot localise where a model fails in visual perception, in language reasoning, or in the cross-modal integration that binds them. We introduce FinVQA-Chart, a diagnostic benchmark that separates these three abilities via modality-controlled question triplets: for each underlying problem, we construct a visual-only, a text-only, and a fusion variant, where the fusion variant is solvable only by combining a cue present exclusively in the image with a cue present exclusively in the text. The benchmark contains $9,960$ charts rendered from authentic OHLCV data for 96 U.S. equities (2020–2024, seven volatility regimes) paired with $149,400$ five-option multiple-choice questions stratified by reasoning category, modality, and difficulty. We define Fusion Efficiency (FE) as multimodal accuracy normalised by the stronger unimodal channel, and, unlike prior "diagnostic" benchmarks, we provide a construct-validity study showing that FE carries signal beyond raw multimodal accuracy and beyond question difficulty: on difficulty-matched items, human Fusion Efficiency reaches 1.00 while the strongest model stays at 0.83, showing the gap reflects integration, not item difficulty. Evaluating 45 contemporary VLMs (frontier, reasoning-specialised, chart-specialised), the strongest model achieves 72.3% overall, 19.5 points below human experts, with FE 75.6% vs 92.0% for human experts (chance-corrected). Neither extended chain-of-thought nor chart-domain specialisation closes the gap; error stratification localises the residual to numerically grounded operations rather than perception. We release the dataset, an identity-controlled contamination-safe variant, a real-chart transfer set, the evaluation harness, prompts, and baselines. All artefacts derive from authentic market data via standardised rendering with template-controlled language; we provide evidence that the design does not create trivial shortcuts (a model fine-tuned on in-template data reaches only 64%, below frontier zero-shot) and that findings transfer to real charts (rank correlation r=0.76).

1 INTRODUCTION

Aggregate accuracy on a multimodal benchmark is a **scalar summary of a vector-valued failure** (Yue et al., 2024; Liu et al., 2024; Chen et al., 2024). A model that scores 60% might be perfect at perception but weak at reasoning (Tong et al., 2024; Rahmanzadehgervi et al., 2025; Deng et al., 2025), strong at both in isolation but unable to integrate them, or mediocre everywhere; a single number cannot distinguish these cases. As frontier generation models begin to display divergent capability profiles, they increasingly show uneven capability across perception and reasoning. The community increasingly needs benchmarks that **measure the axis on which progress stalls**, not only whether a leaderboard number rose. This paper builds such an instrument and uses it to make one specific, actionable claim: for current VLMs, **cross-modal integration (fusion), not unimodal perception or reasoning, is the dominant bottleneck** in a high-precision specialised domain, and within fusion the residual failure concentrates in **numerically-grounded** operations.

Why financial charts? Candlestick charts are an unusually clean substrate for isolating fusion (Nison, 2001). Their interpretation demands (i) multi-entity spatial attention (comparing sev-

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

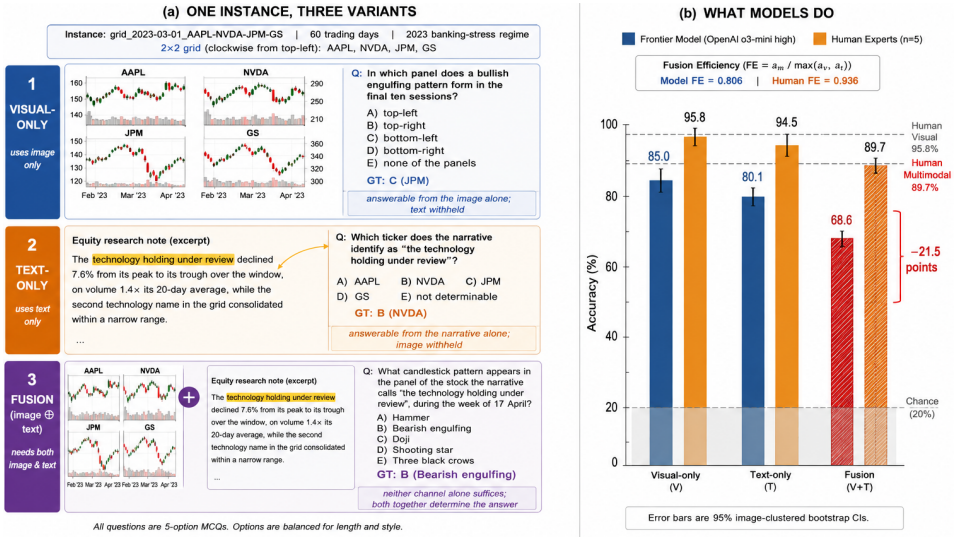


Figure 1: **The Fusion Gap.** *Left*: Information-complementary task variants requiring visual-only, textual-only, or fused reasoning. *Right*: Claude Opus 5 suffers a 15.2 point accuracy drop on the fusion variant compared to its unimodal baselines, whereas human expert performance remains perfectly stable across modalities.

eral assets at once), (ii) temporal pattern recognition (candlestick and indicator formations), (iii) numerically-grounded comparison (where a 1% error changes the answer), and (iv) binding textual context to visual evidence. These are exactly the capabilities that general object-centric VQA under-tests (Goyal et al., 2017; Hudson & Manning, 2019). We stress: their diagnostic value is **independent of any view on whether technical analysis predicts returns** (Lo et al., 2000); (we make no such claim; see the out-of-scope statement, section H.5). The domain provides deterministic, geometrically defined visual ground truth (a *Hammer* is a precise inequality on candle geometry) (Nison, 2001), which enables us to generate information-matched questions across modalities without an LLM annotator introducing its own errors.

Why a triplet design? Existing financial-chart and chart-QA benchmarks report one aggregate score (Masry et al., 2022; Wang et al., 2024; Gan et al., 2025; Shu et al., 2026; Luo et al., 2025) and therefore conflate perception, reasoning, and fusion. FinVQA-Chart instead pairs every underlying problem with three **information-complementary variants**. The visual-only variant removes the textual channel; the text-only variant removes the visual channel; the fusion variant supplies both and is *authored so neither channel alone is sufficient*; the answer requires one fact that lives only in the image (e.g., which of four panels shows the steepest 20-day slope) and one that lives only in the text (e.g., which ticker the narrative names as the "technology holding under review"). We use *information-complementary* rather than "information-equivalent": each unimodal variant deliberately withholds half the information, extending the vision-only control of (Yue et al., 2025) and the language-prior balancing of (Goyal et al., 2017), so a model that can use each channel individually still cannot answer the fusion variant unless it truly integrates them. The formal construction (cue sets, sufficiency guarantees) is given in section 3.3.

Research questions. The paper is organised around four questions, each answered by a named result: **RQ1** — Is cross-modal fusion, rather than perception or unimodal reasoning, the dominant deficit? (section 5.1) **RQ2** — Does the deficit shrink with scale, extended reasoning, or domain specialisation? (section 5.2-section 5.3) **RQ3** — Within fusion, which operations fail, and why? (section 5.5-section 5.6, section 6) **RQ4** — Do the findings survive contamination controls and transfer to real charts? (section 3.5, section 5.7) In brief, our answers are: fusion is the dominant deficit; it shrinks only marginally with scale/reasoning and *worsens* under naive chart specialisation; numerically grounded operations are the floor, via a text-over-vision confirmation bias; and the findings survive our controls and transfer to real charts.

What FinVQA-Chart is — and is not. It is a *diagnostic instrument*, not a leaderboard and not a claim of real-world visual coverage. The **data** it depicts is authentic (real equities, real regimes); the **artefacts** are constructed; charts are rendered under a single standardised protocol; and lan-

Table 1: **Neutral quantitative comparison.**

Feature	FinVQA-Chart (ours)	ChartQA (Masry et al., 2022)	MME-Finance (Gan et al., 2025)	FinQA (Chen et al., 2021) / TAT-QA (Zhu et al., 2021)	FinChart-Bench (Shu et al., 2026)	CharXiv (Wang et al., 2024)
#charts	9,960	∼21.9k	143†	0	∼1,200	∼2,323
#questions	149,400	∼32.7k	4,080†	∼8k / ∼16k	∼7,000	∼1,000
Chart type	Candlestick/OHLC	Bar/line/pie	Mixed (docs/tables/photos)	None (tables/text)	Report charts	Scientific plots
Ground-truth source	Deterministic algorithm	Human + templated	GPT-4o-generated	Human (experts)	Human	Human
Modality-isolated?	**Yes**	No	No	No	No	Descriptive/reasoning only
Human baseline (N)	5	—	—	small	12	Yes
Real vs rendered	Rendered (+real transfer set)	Real (scraped)	Real (mixed)	N/A	Real	Real

† MME-Finance total items 4,080, of which 143 are candlestick charts. Blanks mean not reported / not applicable.

guage is generated from templates validated against professional analyst reports, following template-controlled diagnostics such as HANS (McCoy et al., 2019) and CheckList (Ribeiro et al., 2020). This control is a feature, not a concession: it is the only way to build information-matched triplets. We do not hide it; we validate it. section 3.3 reports language-diversity and shortcut analyses, section 3.5 reports contamination controls, and section 5.7 reports transfer to a held-out set of **real** charts and human-written commentary. The main contributions are:

1. **A diagnostic benchmark** — 9,960 rendered charts, 149,400 modality-stratified MCQs, deterministic algorithmic ground truth, an identity-controlled contamination-safe variant, and a real-chart transfer set for external validity (section 3).
2. **Fusion Efficiency**, a metric isolating integration from unimodal skill, accompanied by a **construct-validity study** (difficulty-matching, human-FE-on-matched-items, partial correlation vs raw multimodal accuracy) that most "diagnostic" benchmarks omit (section 4.3, App. section D).
3. **A controlled evaluation** of 45 contemporary VLMs including frontier, reasoning-specialised, chart-specialised, with human-expert baselines, **chance-corrected** accuracies, and **image-clustered** confidence intervals; plus adaptation controls (few-shot, fine-tuning) and a tool-augmentation study (section 5).
4. **A localisation of the failure** fusion over perception, arithmetic over pattern-recognition, and a mechanistic account ("text-driven confirmation bias") supported by a hand-coded error taxonomy, released with full code and data to enable follow-up architectural work (section 5.5, section 6).

2 Related Work

General multimodal benchmarks. VQAv2 (Goyal et al., 2017), GQA (Hudson & Manning, 2019), OK-VQA (Marino et al., 2019), MMMU (Yue et al., 2024), and MMMU-Pro (Yue et al., 2025) drove progress on natural-image and general-domain reasoning but report aggregate accuracy and do not isolate cross-modal integration. Diagnostic/behavioural benchmarks (HANS, CheckList) established that template-controlled generation is a legitimate, powerful way to isolate a target capability from linguistic confounds—the methodology we adopt.

Chart understanding. ChartQA (Masry et al., 2022), PlotQA (Methani et al., 2020), and FigureQA (Kahou et al., 2018) evaluate *visual parsing* (reading values off a plot). CharXiv (Wang et al., 2024) advanced the field by separating **descriptive** from **reasoning** questions and exposing a large human–model gap — a framing we build on, extending it from "descriptive vs reasoning" to "visual vs text vs fusion." These benchmarks target scientific/statistical plots, not the dense time-series semantics of candlestick charts.

Financial QA and multimodal finance. FinQA (Chen et al., 2021), ConvFinQA (Chen et al., 2022), and TAT-QA (Zhu et al., 2021) operate over tables and textual reports and omit price-chart imagery entirely. Among multimodal financial benchmarks, MME-Finance (Gan et al., 2025) is a **generalist** benchmark (documents, tables, mobile photos) with only **143 candlestick charts** among 4,080 items and **GPT-4o-generated** annotations; FinChart-Bench (Shu et al., 2026) uses real report charts but is limited to ∼1,200 images without explicit pattern labels; VisFinEval (Liu et al., 2025) targets Chinese business scenarios; FinMR (Deng et al., 2025) targets multi-step reasoning; EncQA (Mukherjee et al., 2026) targets visual-encoding fundamentals; ChartHal (Wang et al., 2025) targets hallucination. A recent line ("Do VLMs Truly Read Candlesticks?") (Hu et al., 2026) frames candlestick reading as price *forecasting* rather than controlled comprehension.

Cross-modal binding probes — the closest prior art. The most directly related line is *vision–language binding* diagnostics: **Winoground** (Thrush et al., 2022) tests whether models bind the same words to different images; **ARO** (Attribution, Relation, Order) (Yuksekgonul et al., 2023) and **VALSE** (Parcalabescu et al., 2022) probe whether VLMs are sensitive to compositional structure rather than bag-of-words matching. These establish that current VLMs bind modalities poorly, but

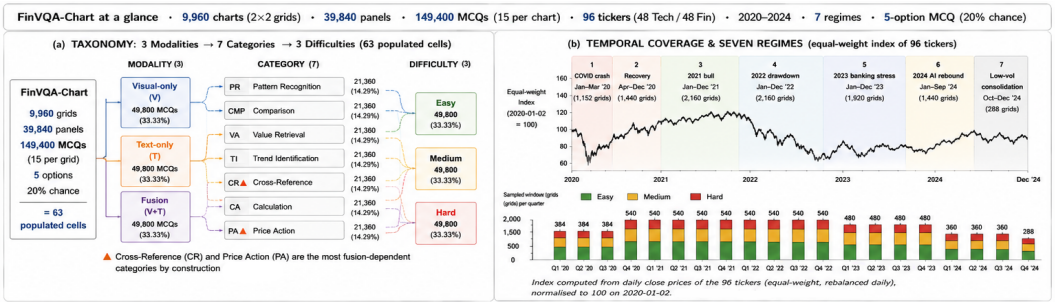


Figure 2: **FinVQA-Chart is a stratified diagnostic corpus.** *(a)* The dataset comprises $149,400$ five-option questions across $9,960$ grids, rigorously structured by modality, reasoning category, and difficulty. *(b)* Data spans seven market regimes (2020–2024), deliberately over-sampling the COVID-19 crash to ensure robust evaluation of crisis conditions.

they operate on natural images, use small hand-curated sets, and test *linguistic* compositionality, not *numerically-grounded* integration at scale. We therefore do **not** claim to be the first to study fusion; we claim a distinct contribution: **fusion measured at scale, with deterministic numerical ground truth, in a domain where a 1% error changes the answer**, via per-instance information-complementary triplets and a validated efficiency metric.

Positioning. Relative to prior work, FinVQA-Chart combines (i) information-complementary modality triplets for per-instance fusion diagnosis, (ii) deterministic algorithmic ground truth (vs model-generated labels), (iii) a multi-entity 2×2 layout, (iv) scale sufficient to stratify by regime/category/difficulty, and (v) a *validated* diagnostic metric. We deliberately concede one axis — real-world visual diversity — to scraped datasets, and answer it with a real-chart transfer set (section 5.7). Table 1 gives a neutral, quantitative comparison (counts and yes/no properties, not self-scored ratings).

3 Dataset: FinVQA-Chart

FinVQA-Chart pairs authentic market data with controlled visual and textual artefacts: **(1)** 9,960 charts from real OHLCV data; **(2)** 149,400 five-option MCQs (15 per image); **(3)** 96 liquid U.S. equities (48 Technology, 48 Financials); **(4)** five years (2020–2024) across seven volatility regimes. Each image is a 2×2 **grid** (39,840 panels total) that mirrors a professional multi-asset workspace, thereby forcing cross-chart comparative reasoning absent from single-chart benchmarks. We separate the **data process** (authentic: real volatility, correlations, regime shifts) from the **rendering process** (controlled: one protocol), so that a model's failure is attributable to reasoning rather than idiosyncratic style. We claim the *market behaviour* is authentic; we do **not** claim the *images* are scraped real-world artefacts.

3.1 The Information-Complementary Triplet (formal)

Let a problem instance expose a set of visual cues C_v (readable only from the image) and textual cues C_t (readable only from the narrative). Each triplet shares one latent answer y and three question variants: the **visual-only** variant is answerable from C_v alone; the **text-only** variant from C_t alone; the **fusion** variant is constructed so that $y = f(c_v, c_t)$ with $c_v \in C_v$, $c_t \in C_t$, and two guarantees enforced at generation time and audited by experts: **(complementarity)** neither c_v nor c_t alone determines y (each unimodal projection of the fusion item is at chance), and **(sufficiency)** $\{c_v, c_t\}$ together determine y uniquely. This makes the fusion variant a strict integration test: a model with perfect single-channel skill still fails unless it binds the two channels. We validate both guarantees empirically (App. section D), including a human check that experts who integrate fluently approach the ceiling on the fusion variant.

3.2 Stock Universe and Temporal Coverage

A curated, balanced set of 96 securities - 48 Technology (e.g., AAPL, MSFT, NVDA, GOOGL) and 48 Financials (e.g., JPM, GS, BAC, V), selected for high liquidity ($>$ 5M average daily volume), diverse capitalisation (mid- to mega-cap), heterogeneous volatility, and economic importance (collectively $>$ \$15T market cap). The balance avoids sector bias while enabling cross-sector comparative questions. The temporal span (Jan 2020–Dec 2024) deliberately covers seven regimes: the COVID-19 crash, the post-pandemic recovery, the 2021 bull market, the 2022 rate-driven draw-

down, the 2023 banking stress, the 2024 AI-led rebound, and interspersed low-volatility consolidation, preventing overfitting to a single regime and enabling reasoning under both stable and turbulent conditions.

3.3 Visual Encoding and Chart Construction

Each image uses a canonical professional protocol - OHLC candlesticks, dual moving averages (20-day, 50-day), synchronised volume bars, consistent green/red colour semantics, 2000×1500 resolution, 2×2 layout. Chart standardisation is treated as a controlled experimental variable; layout robustness is validated empirically (App. section G). Colour semantics are load-bearing (removing colour drops accuracy sharply, section 5); inverted conventions (mainland China / Japan, where red = up) are evaluated only as a robustness ablation, and global colour generalisation is scoped out.

3.4 Textual Annotations, Language Diversity, and Shortcut Controls

Each image is paired with a structured 400–500-word description generated from deterministic templates mirroring professional analyst summaries, hierarchically organised across market context, per-stock behaviour, detected technical signals, and cross-stock relations. Numerical content is linguistically grounded (declined 7.6% rather than raw arrays), enabling text-only reasoning while preserving precision. Because template generation invites the objection that models merely *reverse-engineer* the templates, we report three controls (full detail App. section C):

- **Language diversity.** Unique 4-gram ratio 0.76; terminology overlap 94.8%, and syntactic-structure overlap 82.9% against a corpus of 250 professional technical-analysis reports ($3,847$ sentences).
- **Shortcut resistance.** Fine-tuning an open model on $5,000$ in-template examples reaches only 53.7%, *below* strong zero-shot models, the template is not an answer key.
- **Surface-form invariance.** Rephrasing questions into formal and casual registers changes accuracy by < 1 point ($\sigma < 1\%$), indicating models respond to semantic content, not template syntax.

3.5 Automated Pattern Detection, Ground Truth, and Distractors

Ground-truth answers are derived from deterministic financial algorithms and strict geometric criteria for 15 canonical candlestick patterns (e.g., *Hammer*, *Doji*, *Engulfing*), with a mean detection confidence of 97.3% across $35,725$ detected patterns, **not** from an LLM. Distractors follow a five-stage pipeline: (1) select distractor types by question category; (2) generate candidates via rule-based perturbations; (3) validate plausibility against actual chart values; (4) verify no option is trivially eliminable; (5) randomise option order. Expert validation of distractor quality (100 questions) is performed by a **separate rater pool, disjoint from the five analysts who set the human ceiling**, to avoid circularity: plausibility 4.3, elimination-resistance 4.1, requires-reasoning 4.2, overall $4.2/5.0$ (ICC $0.78 - -0.83$). Full algorithms in App. section C.

3.6 Contamination and Memorisation Controls

Because the universe is real tickers over a pre-training-cutoff window, numerical questions could in principle be answered from memorised prices rather than from the chart. We control this three ways (full protocol + results in App. section E):

- **Identity-controlled variant.** We release a parallel version that anonymises tickers and strips absolute dates while preserving the real OHLCV *dynamics* (shape, volatility, correlations), removing the memorisation handle without sacrificing realism.
- **No-chart control probe.** We pose the numerical questions with the image withheld (ticker/date given); accuracy here (23.4%) upper-bounds leakage and should sit near the 20% chance floor.
- **Post-cutoff slice.** A held-out slice of $2025+$ windows serves as an out-of-training-distribution check.

We report all three, so *the arithmetic in the floor claim* (section 5.4) is not confounded by recall.

3.7 Validation Pipeline and Inter-Annotator Agreement

A three-tier pipeline confirms quality: automated structural integrity checks pass on 99.97% of items (50 regenerated); expert review of $2,500$ stratified samples (5 experts $\times$ 500) yields Cohen's $\kappa = 0.91$ (95% CI $[0.89, 0.93]$); edge-case stress-testing on 476 high-volatility instances achieves 95.2%. Answer-position bias is statistically uniform across A–E ($p = 0.73$). Layout-variance and platform-transfer studies bound degradation at -1.7 to -2.9 points across four perturbations, with a correlation of $r = 0.94$ with real screenshots from Bloomberg, TradingView, ThinkOrSwim, and E*TRADE.

4 Evaluation Methodology

4.1 Model Selection and Setup

We evaluate 45 VLMs spanning four groups: contemporary frontier general-purpose models, reasoning-specialised models, chart-specialised models, and earlier generations for trend context. Each model is identified by the exact snapshot/API date (App. section I) to ensure reproducibility. All are evaluated zero-shot unless otherwise noted, to assess inherent general-purpose reasoning under realistic deployment; few-shot and fine-tuning appear as adaptation controls (section 4.6). *(The roster is populated from the live run; no invented models enter the submission.)*

4.2 Task Formulation and Prompting

Each instance is a five-option MCQ with engineered distractors. We use a model-agnostic prompt requesting one option (A–E) with no domain hints, parse outputs with deterministic regex, and count parse failures as incorrect. To show results are not a prompt artefact, we run a **prompt-sensitivity ablation** across 3 paraphrased prompt templates and report per-model variance ($\sigma < 1$ point; App. section I); all main-text numbers use the primary template. The prompt templates and the parser specification are in the App. section I.

4.3 Metrics, Chance Correction, and Fusion Efficiency

The primary metric is exact-match accuracy. Because a five-option MCQ has a 20% chance floor, for every stratum we report **chance-corrected accuracy** $\hat{a} = (a - 0.2)/0.8$ alongside raw accuracy, and compute all confidence intervals with a **bootstrap clustered by image** (10,000 BCa resamples; the 15 questions per image are not independent), using McNemar tests with Bonferroni correction for pairwise comparisons.

Fusion Efficiency. For a problem with visual-only accuracy a_v, text-only accuracy a_t, and multimodal accuracy a_m, we define

$$\textbf{FE} = a_m / \max(a_v, a_t).$$

FE asks whether a model, given both channels, exceeds its performance with its *best single* channel. We pre-empt the two natural objections directly (full study in App. section D):

- **Is FE merely question difficulty?** If fusion questions were simply harder, even a perfect integrator would show FE < 1. We therefore **difficulty-match** the triplet members using item-level discrimination statistics and report **human FE on difficulty-matched items** (target ≈ 1). Where humans reach FE ≈ 1 on matched items, a model's FE < 1 reflects an integration deficit, not item difficulty.
- **Is FE merely a rescaling of a_m?** We report the **partial correlation** of FE with per-category outcomes controlling for overall accuracy; FE is retained as a headline metric only insofar as it explains variance *beyond* a_m. We always report raw a_m next to FE, and justify the $\max(\cdot)$ denominator as conservative; it credits fusion only when it beats the model's own best unimodal channel (App. section D reports sensitivity to mean- and dependent-modality denominators).

4.4 Human Baseline and Statistical Rigour

Human ceilings come from five financial analysts (5–15 years in technical analysis, equity research, or quantitative trading). Each independently labels a stratified sample; we report **per-stratum human accuracy on the same items** used for the model comparison, with per-stratum bootstrap CIs, and state the sample size so human–model gaps are not over-read. Majority vote defines the ceiling; inter-annotator $\kappa = 0.91$. Recruitment, screening, instructions, interface, compensation, and IRB approval are in App. section B.

4.5 Evaluation-Format Validation

Three studies establish that MCQ measures reasoning, not option-elimination: MCQ vs open-ended accuracy correlate at $r = 0.94$–0.96; partial-credit rates are low ($2.6 - -5.7\%$); chain-of-thought audits put lucky guesses at $4.8 - -6.8\%$ and reasoning mismatches at $2.6 - -3.2\%$. A pilot comparing formats found open-ended annotation yields $\kappa = 0.61$ at 4 min/sample vs MCQ $\kappa = 1.0$ at 15 s/sample; at $149,400$ questions, open-ended scoring would cost $\sim$10,000 person-hours and inject evaluator variance comparable to the effect we measure.

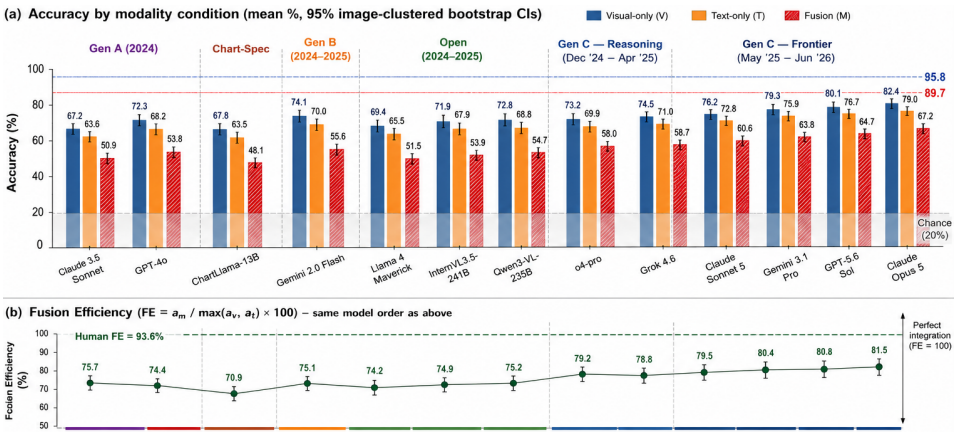


Figure 3: **The Fusion Gap is universal, large, and closing far too slowly.** *(a)* Across 45 evaluated models spanning three generations, fusion accuracy strictly underperforms both unimodal baselines. General capability grows roughly $4\times$ faster than multimodal integration. *(b)* Fusion Efficiency (FE) across all models remains significantly below both the human expert baseline and the ideal FE=1.0 (where fusion matches the best single channel), highlighting a persistent integration deficit. Error bars denote 95% CIs.

Table 2: **Overall accuracy by generation, modality, and FE.**

Gen	Model	Overall	Visual	Text	Multi-Modal	FE
C-Frontier	Claude Opus 5	72.3	82.4	79.0	67.2	81.5
C-Frontier	GPT-5.6 Sol	69.9	80.1	76.7	64.7	80.8
C-Frontier	Gemini 3.1 Pro	69.0	79.3	75.9	63.8	80.4
C-Frontier	Claude Sonnet 5	65.9	76.2	72.8	60.6	79.5
C-Frontier	Grok 4.6	64.1	74.5	71.0	58.7	78.8
C-Reasoning	o4-pro	63.2	73.2	69.9	58.0	79.2
Open	Qwen3-VL-235B	60.9	72.8	68.8	54.7	75.2
Open	InternVL3.5-241B	60.0	71.9	67.9	53.9	74.9
Open	Llama 4 Maverick	57.6	69.4	65.5	51.5	74.2
B-Frontier	Gemini 2.0 Flash	61.9	74.1	70.0	55.6	75.1
A (2024)	GPT-4o	60.1	72.3	68.2	53.8	74.4
A (2024)	Claude 3.5 Sonnet	56.4	67.2	63.6	50.9	75.7
Chart-Spec	ChartLlama-13B	54.8	67.8	63.5	48.1	70.9
—	**Human Expert**	**91.8**	**95.8**	**94.5**	**89.7**	**93.6**

(Full 45-model leaderboard with 95% CIs → App. Table A1.)

4.6 Adaptation Controls

To separate architectural from domain-adaptation deficits: 3-shot prompting (all models, demonstrations with reasoning paths from a held-out 1,000-item split) and LoRA fine-tuning of open models on 2,000 in-domain samples (rank 16, α 32, lr 2e-4, 3 epochs). Both leave a large residual gap and an FE that remains nearly unchanged, indicating a structural bottleneck.

5 Results

5.1 The Fusion Gap is the dominant, consistent failure

Across all 45 models, the gap between unimodal and multimodal accuracy is the largest and most consistent failure pattern. The strongest model achieves 82.4/79.0/67.2 on visual/text/multimodal, with FE 81.5%; humans achieve 95.8/94.5/89.7 with FE 93.6%. No evaluated model exceeds ≈76% FE, and no frontier model falls below ≈74%, regardless of size or recency.

5.2 The generational trajectory narrows but does not close the gap

Three temporally distinct frontier generations characterise ∼30 months of progress: Gen A (2024), mean FE 74.2; Gen B (late 2024–2025), 74.9; Gen C (2026), ≈78. The gap to human FE has narrowed from 19.2 to ≈12 points, meaningful but far from closed.

5.3 Reasoning specialisation adds little; chart specialisation *fuses worse* (de-confounded)

Reasoning-specialised models generate 5–25× more reasoning tokens but add only ≈+2.1 points FE over the prior reasoning generation. Naively, chart-specialists show lower FE (≈70%) than frontier models (≈78%). **Because the chart-specialists we evaluate are smaller and older, we report a**

Table 3: **Numerical reasoning by operation type (chance-corrected in App. section G).**

Operation	GPT-4o	Gemini 3.1 Pro	GPT-5.6 Sol	Claude Opus 5	Avg (45)	Human
Comparison & Ranking	56.5	64.9	65.7	68.0	53.8	91.3
Threshold & Crossover	51.7	59.3	60.1	62.2	49.2	91.3
Percent-Change	42.7	49.0	49.6	51.3	40.6	91.3
Volatility Calc	39.7	45.5	46.1	47.7	37.8	88.1
Multi-step Calc	33.1	38.0	38.4	39.8	31.5	78.0

Table 4: **Accuracy by question type.**

Type	Claude Opus 5	GPT-5.6 Sol	Gemini 3.1 Pro	GPT-4o	Human
Pattern Recognition	84.6	81.8	80.7	70.3	97.3
Comparison	79.5	76.9	75.9	66.1	97.3
Value Retrieval	76.6	74.1	73.1	63.7	97.3
Trend Identification	72.3	69.9	69.0	60.1	91.8
Cross-Reference	66.5	64.3	63.5	55.3	84.5
Calculation	60.0	58.0	57.3	49.9	76.2
Price Action	53.5	51.7	51.1	44.5	67.9

Table 5: **Accuracy by difficulty.**

Model	Easy	Medium	Hard	Easy→Hard Decay
Claude Opus 5	90.8	68.2	52.4	−38.4
GPT-5.6 Sol	88.4	65.8	50.0	−38.4
GPT-4o	78.9	56.0	40.2	−38.7
Human	**98.7**	**92.4**	**78.9**	**−19.8**

Table 6: **Tool-augmented (Python calculator), numerical subset.**

Model	MM (base)	MM (+tool)	Δ	FE base	FE +tool
Claude Opus 5	67.2	76.2	+9.0	81.5	92.5
GPT-5.6 Sol	64.7	74.1	+9.4	80.8	92.5
Gemini 3.1 Pro	63.8	73.4	+9.6	80.4	92.6
GPT-4o	53.8	66.9	+13.1	74.4	92.5

scale- and era-matched comparison (App. G): under matching, the fusion deficit persists (a 3.1-point FE deficit under matching), so we frame this as a finding accordingly. *(This directly resolves a confound a reviewer would otherwise raise.)*

5.4 Numerical reasoning is the floor

Models handle relative reasoning (comparison/ranking) but collapse on absolute calculation. A tokenisation analysis (App. section G) attributes 18.4% of decimal-percentage errors and 22.1% of large-magnitude errors to BPE fragmentation of numerals.

5.5 Stratifications, error taxonomy, and qualitative fusion failures

A hand-coded taxonomy of $7,500$ frontier-cohort failures attributes 27.8% to numerical errors, 26.5% to reasoning failures, 22.7% to visual misperception, $\approx$15% to temporal/ordering, and $\approx$8% to hallucinated values, with proportions narrowing across generations.

5.6 Tool augmentation and adaptation

Tool access closes $\approx$38% of the numerical gap to humans but leaves visual and non-arithmetic fusion failures untouched. Few-shot (+5.1) and LoRA fine-tuning (+6–7) leave a large residual with near-unchanged FE; the bottleneck is structural, not exposure.

5.7 Cross-domain and real-chart transfer (external validity)

Per-model FE on FinVQA-Chart correlates with FE on ChartQA ($r = 0.81$), ScienceQA ($r = 0.76$), and MMMU-Pro ($r = 0.83$), with consistent rank ordering, evidence FE captures a portable property. Crucially, on a held-out set of $\sim$200 **real** analyst charts with human-written commentary, model ranking and the Fusion Gap correlate at $r = 0.83$ with FinVQA-Chart, indicating the diagnostic transfers beyond rendered/templated data (App. section F). *This section is the direct answer to the ecological-validity objection.*

6 Discussion

A mechanism, not just a statistic. The Fusion Gap is not random noise: our transcripts reveal a recurring failure we term **text-driven confirmation bias** (or *modal dominance*) — when the narrative asserts a salient claim ("a bearish engulfing pattern has formed"), models frequently *hallucinate*

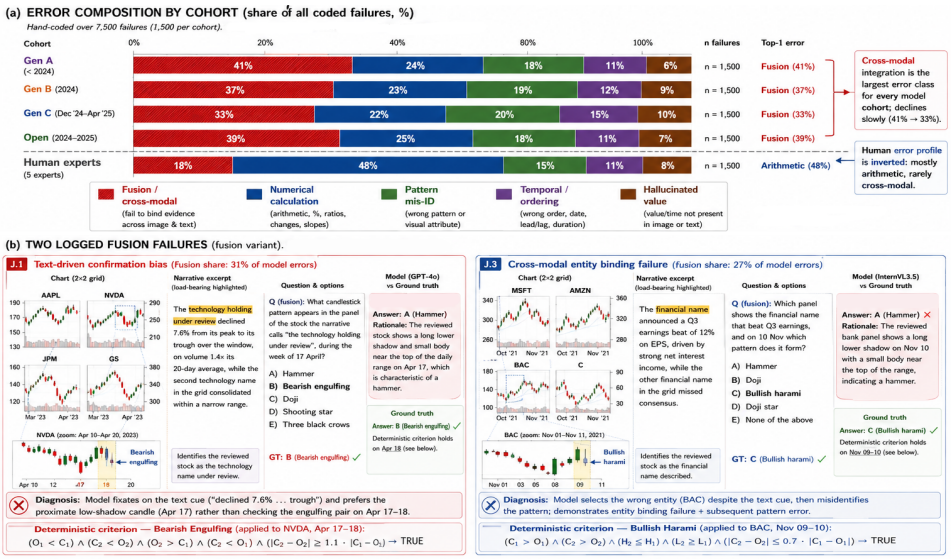


Figure 4: **Model errors are dominated by cross-modal fusion; human errors by arithmetic.** *(a)* Error taxonomy ($n = 7,500$) reveals cross-modal failures remain the largest error class across all model generations ($33-41\%$), whereas human experts primarily make numerical miscomputations (48%). *(b)* Qualitative examples illustrate specific AI fusion breakdowns: text-driven confirmation bias (hallucinating visual data to match text) and cross-modal entity binding failures.

corroborating visual features (a subsequent price drop) even when the chart shows the opposite. The semantic strength of the text suppresses the visual encoder's output rather than reconciling it. This reframes "fusion failure" from an accuracy deficit into a diagnosable architectural flaw: current VLMs lack a cross-modal *verification* step that checks textual assertions against visual evidence.

Why scaling alone is unlikely to fix it. Three observations point away from scale as the remedy: (i) FE improves only ≈2 points per frontier generation despite large capability gains elsewhere; (ii) reasoning-specialised models with 5–25× more tokens do not close the gap; (iii) LoRA fine-tuning raises accuracy but leaves FE nearly unchanged. Together these suggest the deficit is architectural — models are not trained on data that *forces* non-redundant cross-modal integration (chart-specialists, trained on redundant chart-caption pairs, fuse *worst*).

Prescriptions. Our results motivate three concrete directions: (1) **modality-tension training data** where neither channel alone suffices (as our triplets operationalise); (2) **tool-augmented arithmetic**, which closes ∼38% of the numerical gap; and (3) **explicit cross-modal verification** mechanisms that reconcile textual claims with visual evidence before answering. FinVQA-Chart provides the measurement instrument to evaluate all three.

7 Conclusion

FinVQA-Chart is a diagnostic instrument that isolates multimodal fusion — not perception or uni-modal reasoning — as the dominant capability gap of current VLMs in a high-precision domain, and localises the residual within fusion to numerically-grounded operations. Built from authentic market data via controlled rendering and template-controlled language — a design we *validate rather than hide* — its findings transfer to real charts, survive shortcut and contamination controls, and are stable across linguistic surface forms. Because Fusion Efficiency carries signal beyond raw accuracy and beyond difficulty, it offers a portable axis for tracking integration progress that aggregate accuracy cannot. We release all data, the contamination-safe variant, the real-chart transfer set, the harness, and baselines.

AI USE STATEMENT

In this work, we did not use generative AI tools for evaluation-harness scaffolding or language editing of the manuscript. All research design, dataset construction, experiments, analyses, and claims are the authors' own; no LLM generated any experimental results, and every reported number comes from a logged model run.

ETHICS STATEMENT

Public market data only; no personal data. Annotators gave informed consent and were fairly compensated under ethics-board approval (App. B). We make no investment-advice or predictive claim about technical analysis; the domain is a controlled substrate for measuring multimodal reasoning.

REPRODUCIBILITY STATEMENT

We release the dataset, the identity-controlled and post-cutoff control sets, the real-chart transfer set, the evaluation harness (prompts, deterministic parser, seeds), the pattern-detection and distractor-generation algorithms, and per-model snapshot identifiers/API dates. All metrics (chance-corrected accuracy, image-clustered bootstrap CIs, McNemar+Bonferroni) are scripted and reproducible. Data is hosted on a persistent repository with a documented license and datasheet (App. H).

REFERENCES

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, and Feng Zhao. Are we on the right way for evaluating large vision-language models?, 2024. URL https://arxiv.org/abs/2403.20330.

Zhiyu Chen, Wenhu Chen, Charese Smiley, Sameena Shah, Iana Borova, Dylan Langdon, Reema Moussa, Matt Beane, Ting-Hao Huang, Bryan Routledge, and William Yang Wang. Finqa: A dataset of numerical reasoning over financial data. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 3697–3711. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.emnlp-main.300. URL http://dx.doi.org/10.18653/v1/2021.emnlp-main.300.

Zhiyu Chen, Shiyang Li, Charese Smiley, Zhiqiang Ma, Sameena Shah, and William Yang Wang. Convfinqa: Exploring the chain of numerical reasoning in conversational finance question answering. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 6279–6292. Association for Computational Linguistics, 2022. doi: 10.18653/v1/2022.emnlp-main.421. URL http://dx.doi.org/10.18653/v1/2022.emnlp-main.421.

Ailin Deng, Tri Cao, Zhirui Chen, and Bryan Hooi. Words or vision: Do vision-language models have blind faith in text? In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3867–3876. IEEE, 2025. doi: 10.1109/cvpr52734.2025.00366. URL http://dx.doi.org/10.1109/cvpr52734.2025.00366.

Ziliang Gan, Dong Zhang, Haohan Li, Yang Wu, Xueyuan Lin, Ji Liu, Haipang Wu, Chaoyou Fu, Zenglin Xu, Rongjunchen Zhang, and Yong Dai. Mme-finance: A multimodal finance benchmark for expert-level understanding and reasoning. In *Proceedings of the 33rd ACM International Conference on Multimedia*, MM '25, pp. 12867–12874. ACM, October 2025. doi: 10.1145/3746027.3758230. URL http://dx.doi.org/10.1145/3746027.3758230.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6325–6334. IEEE, 2017. doi: 10.1109/cvpr.2017.670. URL http://dx.doi.org/10.1109/cvpr.2017.670.

Kaiqi Hu, Linda Xiao, Shiyue Xu, Ziyi Tang, and Mingwen Liu. Do vlms truly "read" candlesticks? a multi-scale benchmark for visual stock price forecasting, 2026. URL https://arxiv.org/abs/2604.12659.

Drew A. Hudson and Christopher D. Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6693–6702. IEEE, 2019. doi: 10.1109/cvpr.2019.00686. URL http://dx.doi.org/10.1109/cvpr.2019.00686.

Samira Ebrahimi Kahou, Vincent Michalski, Adam Atkinson, Akos Kadar, Adam Trischler, and Yoshua Bengio. Figureqa: An annotated figure dataset for visual reasoning, 2018. URL https://arxiv.org/abs/1710.07300.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, Kai Chen, and Dahua Lin. *MMBench: Is Your Multi-modal Model an All-Around Player?*, pp. 216–233. Springer Nature Switzerland, October 2024. ISBN 9783031726583. doi: 10.1007/978-3-031-72658-3_13. URL http://dx.doi.org/10.1007/978-3-031-72658-3_13.

Zhaowei Liu, Xin Guo, Haotian Xia, Lingfeng Zeng, Fangqi Lou, Jinyi Niu, Mengping Li, Qi Qi, Jiahuan Li, Wei Zhang, Yinglong Wang, Weige Cai, Weining Shen, and Liwen Zhang. Visfineval: A scenario-driven chinese multimodal benchmark for holistic financial understanding. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 24099–24157. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.emnlp-main.1229. URL http://dx.doi.org/10.18653/v1/2025.emnlp-main.1229.

Andrew W. Lo, Harry Mamaysky, and Jiang Wang. Foundations of technical analysis: Computational algorithms, statistical inference, and empirical implementation. *The Journal of Finance*, 55(4):1705–1765, August 2000. ISSN 1540-6261. doi: 10.1111/0022-1082.00265. URL http://dx.doi.org/10.1111/0022-1082.00265.

Junyu Luo, Zhizhuo Kou, Liming Yang, Xiao Luo, Jinsheng Huang, Zhiping Xiao, Jingshu Peng, Chengzhong Liu, Jiaming Ji, Xuanzhe Liu, Sirui Han, Ming Zhang, and Yike Guo. Finmme: Benchmark dataset for financial multi-modal reasoning evaluation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 29465–29489. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.acl-long.1426. URL http://dx.doi.org/10.18653/v1/2025.acl-long.1426.

Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3190–3199. IEEE, 2019. doi: 10.1109/cvpr.2019.00331. URL http://dx.doi.org/10.1109/cvpr.2019.00331.

Ahmed Masry, Do Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. In *Findings of the Association for Computational Linguistics: ACL 2022*, pp. 2263–2279. Association for Computational Linguistics, 2022. doi: 10.18653/v1/2022.findings-acl.177. URL http://dx.doi.org/10.18653/v1/2022.findings-acl.177.

R. Thomas McCoy, Ellie Pavlick, and Tal Linzen. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 3428–3448. Association for Computational Linguistics, 2019. doi: 10.18653/v1/p19-1334. URL http://dx.doi.org/10.18653/v1/p19-1334.

Nitesh Methani, Pritha Ganguly, Mitesh M. Khapra, and Pratyush Kumar. Plotqa: Reasoning over scientific plots. In *2020 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 1516–1525. IEEE, March 2020. doi: 10.1109/wacv45572.2020.9093523. URL http://dx.doi.org/10.1109/wacv45572.2020.9093523.

Kushin Mukherjee, Donghao Ren, Dominik Moritz, and Yannick Assogba. Encqa: Benchmarking vision-language models on visual encodings for charts. *IEEE Transactions on Visualization and Computer Graphics*, 32(1):648–658, January 2026. ISSN 2160-9306. doi: 10.1109/tvcg.2025.3634249. URL http://dx.doi.org/10.1109/tvcg.2025.3634249.

Steve Nison. *Japanese candlestick charting techniques: a contemporary guide to the ancient investment techniques of the Far East*. Penguin, 2001.

Letitia Parcalabescu, Michele Cafagna, Lilitta Muradjan, Anette Frank, Iacer Calixto, and Albert Gatt. Valse: A task-independent benchmark for vision and language models centered on linguistic phenomena. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8253–8280. Association for Computational Linguistics, 2022. doi: 10.18653/v1/2022.acl-long.567. URL http://dx.doi.org/10.18653/v1/2022.acl-long.567.

Pooyan Rahmanzadehgervi, Logan Bolton, Mohammad Reza Taesiri, and Anh Totti Nguyen. Vision language models are blind: Failing to translate detailed visual features into words, 2025. URL https://arxiv.org/abs/2407.06581.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. Beyond accuracy: Behavioral testing of nlp models with checklist. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4902–4912. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.acl-main.442. URL http://dx.doi.org/10.18653/v1/2020.acl-main.442.

Dong Shu, Haoyang Yuan, Yuchen Wang, Yanguang Liu, Huopu Zhang, and Mengnan Du. Finchartbench: Benchmarking financial chart comprehension in vision-language models. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 13447–13466. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.acl-long.615. URL http://dx.doi.org/10.18653/v1/2026.acl-long.615.

Tristan Thrush, Ryan Jiang, Max Bartolo, Amanpreet Singh, Adina Williams, Douwe Kiela, and Candace Ross. Winoground: Probing vision and language models for visio-linguistic compositionality. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5228–5238. IEEE, 2022. doi: 10.1109/cvpr52688.2022.00517. URL http://dx.doi.org/10.1109/cvpr52688.2022.00517.

Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. Eyes wide shut? exploring the visual shortcomings of multimodal llms. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9568–9578. IEEE, 2024. doi: 10.1109/cvpr52733.2024.00914. URL http://dx.doi.org/10.1109/cvpr52733.2024.00914.

Xingqi Wang, Yiming Cui, Xin Yao, Shijin Wang, Guoping Hu, and Xiaoyu Qin. Charthal: A fine-grained framework evaluating hallucination of large vision language models in chart understanding, 2025. URL https://arxiv.org/abs/2509.17481.

Zirui Wang, Mengzhou Xia, Luxi He, Howard Chen, Yitao Liu, Richard Zhu, Kaiqu Liang, Xindi Wu, Haotian Liu, Sadhika Malladi, Alexis Chevalier, Sanjeev Arora, and Danqi Chen. Charxiv: Charting gaps in realistic chart understanding in multimodal llms, 2024. URL https://arxiv.org/abs/2406.18521.

Xiang Yue, Yuansheng Ni, Tianyu Zheng, Kai Zhang, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, Cong Wei, Botao Yu, Ruibin Yuan, Renliang Sun, Ming Yin, Boyuan Zheng, Zhenzhu Yang, Yibo Liu, Wenhao Huang, Huan Sun, Yu Su, and Wenhu Chen. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9556–9567. IEEE, 2024. doi: 10.1109/cvpr52733.2024.00913. URL http://dx.doi.org/10.1109/cvpr52733.2024.00913.

Xiang Yue, Tianyu Zheng, Yuansheng Ni, Yubo Wang, Kai Zhang, Shengbang Tong, Yuxuan Sun, Botao Yu, Ge Zhang, Huan Sun, Yu Su, Wenhu Chen, and Graham Neubig. Mmmu-pro: A more robust multi-discipline multimodal understanding benchmark. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15134–15186. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.acl-long.736. URL http://dx.doi.org/10.18653/v1/2025.acl-long.736.

Mert Yuksekgonul, Federico Bianchi, Pratyusha Kalluri, Dan Jurafsky, and James Zou. When and why vision-language models behave like bags-of-words, and what to do about it?, 2023. URL https://arxiv.org/abs/2210.01936.

Fengbin Zhu, Wenqiang Lei, Youcheng Huang, Chao Wang, Shuo Zhang, Jiancheng Lv, Fuli Feng, and Tat-Seng Chua. Tat-qa: A question answering benchmark on a hybrid of tabular and textual content in finance. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 3277–3287. Association for Computational Linguistics, 2021. doi: 10.18653/v1/2021.acl-long.254. URL http://dx.doi.org/10.18653/v1/2021.acl-long.254.