MM-REPORTBENCH: A COMPREHENSIVE BENCHMARK FOR MULTIMODAL REPORT GENERATION FROM TABLES

Anonymous authors
Paper under double-blind review

ABSTRACT

The ability to generate multimodal reports integrating fluent narratives with faithful visualizations from tabular data, has become a cornerstone capability for modern vision-language models in practical domains such as finance, healthcare, and policy making. However, existing benchmarks are limited by single-modality focus, simplistic metrics, novelty evaluation, and visualization quality. We address these limitations with MM-ReportBench, a multimodal report evaluation benchmark covering 386 tasks across 185 real-world tables. Each task is paired with an expert-curated reference report featuring standardized chapter structures and verifiable key insight points, constructed through a hybrid pipeline of multimodal reports generation based on Monte Carlo Tree Search and rigorous human quality control. For robust evaluation, we design a dual-judge evaluation framework: Text Judge Model assesses textual quality along structural completeness, grounded insight verification with numerical fact-checking against source tables, and novelty analysis that distinguishes genuine analytical discovery from trivial paraphrasing; Visualization Judge Model evaluates generated visualizations across data fidelity, expressiveness, aesthetics, and technical correctness. They also evaluate cross-modal consistency. Extensive experiments on 20 state-of-the-art models reveal critical weaknesses: even the strongest model achieves only 68.9 overall score, with numerical hallucination affecting 38.5% of reports and genuinely novel insights remaining rare. Our ablation studies further demonstrate that single-judge evaluation consistently inflates scores by 8–12 points compared to dual-judge assessment, confirming that rigorous and reliable benchmarking demands multi-faceted, complementary evaluation probes. The source data and code is available at https://anonymous.4open.science/r/MReportBench.

1 INTRODUCTION

The rapid evolution of large language models (LLMs) and large vision-language models (LVLMs) has fundamentally reshaped how organizations interact with data. In real-world business analytics, scientific discovery, and decision-making, a critical emerging capability is *multimodal report generation*: the automated production of professional reports that synthesize textual analysis, quantitative insights, and visual charts directly from raw structured tables. Unlike conventional table question answering or simple summarization, this task demands that models not only extract and aggregate numerical facts but also structure coherent narratives, ground claims in source evidence, generate informative visualizations, and deliver actionable conclusions while maintaining factual fidelity and logical coherence (Huang et al., 2026; Liu et al., 2026a). As enterprises increasingly deploy AI agents for end-to-end data analysis workflows, the ability to reliably evaluate such multimodal report generation systems has become both an urgent practical need and a significant research challenge.

Despite substantial progress in table reasoning benchmarks, existing works exhibit inadequate modes. First, current benchmarks predominantly focus on single modalities in isolation, like table question answering (Zhong et al., 2017; Nan et al., 2022; Chen et al., 2020; Parikh et al., 2020; Lin, 2025), while recent efforts toward open-ended report writing such as T2R-Bench (Zhang et al., 2025) and DDR-Bench (Liu et al., 2026a) remain confined to textual outputs and do not assess the generation of visualizations essential to professional business reports. Second, the LLM-as-a-judge paradigm,

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.

though widely adopted for scalability, suffers from fundamental vulnerabilities: It is demonstrated that reasoning judges can drive policy models toward adversarial output strategies that achieve high judge rewards while violating task intent, and that such adversarial outputs generalize to deceive strong evaluators, underscoring the need for evaluation frameworks resistant to reward hacking (Liu et al., 2026b). Third, existing benchmarks lack standardized reference standards with verifiable ground truth, MMDR-Bench (Huang et al., 2026) proposes multi-stage evaluation with visual evidence fidelity checks but depends on retrieval-based citation verification unsuitable for table-grounded settings, while VisJudge-Bench (Xie et al., 2026) establishes comprehensive visualization assessment yet evaluates charts in isolation without connection to textual reports and source tables. Across these efforts, a unified benchmark that integrates structured tabular data, multimodal report generation, and rigorous multi-aspect evaluation with verifiable reference standards remains absent.

To bridge this gap, we introduce **MM-ReportBench**, a comprehensive benchmark for table-to-multimodal report generation with a comprehensive dual-judge evaluation suite. Our design is directly informed by the lessons of prior work: we ground all factuality checks in source tables rather than external retrieval to mitigate evaluation manipulation; we split evaluation into specialized judges to prevent the halo effects where strong text masks poor visuals or vice versa; and we incorporate explicit novelty assessment to reward genuine analytical discovery over paraphrasing.

Central to our benchmark is a **dual-judge evaluation framework**, which we adopt to overcome a critical pitfall of prior evaluation protocols: monolithic scoring systems that aggregate textual and visual quality into a single metric often conflate heterogeneous failure modes, mask modality-specific errors, and produce inflated overall scores, for example, a poorly grounded narrative might be camouflaged by aesthetically pleasing charts, or vice versa. By decoupling the assessment of generated text and visualizations, our framework offers two distinct advantages: (a) it enables fine-grained diagnostic analysis, pinpointing exactly whether a model struggles with factual grounding, chart fidelity, or creative insight; and (b) it prevents cross-modality compensation, ensuring that excellence in one modality cannot artificially boost the perceived performance of the other. Concretely, the framework comprises two independent judges: **T-Judge** evaluates the textual report along three axes, including Structural Completeness, Grounded Insight Verification and Novelty Analysis;**V-Judge** assesses generated visualizations across four dimensions including Data Fidelity, Expressiveness, Aesthetics and Technical Correctness. And they assess the **cross-modal consistency** together. This modular design not only mirrors the real-world expectation that human analysts critique reports and charts separately, but also provides a principled foundation for our subsequent ablation studies, which confirm that a single judge inflates scores by 8-12 points relative to the dual-judge setup, vindicating the necessity of this separation.

We conduct extensive experiments on over 20 state-of-the-art models, including text-only LLMs, vision-language models, code-augmented multimodal systems, and specialized deep research agents reveal critical weaknesses across the board. The contributions of this work are:

- **We propose the first comprehensive benchmark for table-to-multimodal report generation**, with 386 tasks across six business domains and expert-curated reference reports with verifiable key insight points.
- We design a **dual-judge evaluation** framework combining specialized T-Judge and V-Judge modules that jointly measure textual and visual quality through interpretable, dimension-level scoring.
- We propose a rigorous reference construction pipeline using **MCTS with multi-dimensional rewards** and multi-stage human quality control.
- **Comprehensive empirical findings** exposing critical failure modes in current models and actionable insights for advancing reliable multimodal report generation.

2 Related Work

Table Reasoning and Report Generation Benchmarks. Early benchmarks for table understanding focused on narrow, well-defined tasks such as table-based question answering (Zhong et al., 2017; Nan et al., 2022; Chen et al., 2020; Pasupat & Liang, 2015; Wu et al., 2025), table-to-text generation (Parikh et al., 2020; Cheng et al., 2022), and text-to-SQL parsing (Yu et al., 2018). While these datasets

enabled significant progress in basic table comprehension, they largely ignore the complex, open-ended requirements of real-world business analytics. More recent work has started to address this gap. T2R-Bench (Zhang et al., 2025) formalizes the table-to-report task with industrial tables and evaluates numerical accuracy, coverage, and writing quality, while DDR-Bench (Liu et al., 2026a) advances autonomous data exploration. However, both frameworks remain confined to textual outputs and do not assess the generation of visualizations, which are essential components of professional reports.

LLM-as-a-Judge and Multimodal Evaluation. The LLM-as-a-judge paradigm is vulnerable to reward hacking and adversarial outputs: It is shown that reasoning-judge-trained policies learn systematic strategies that deceive both training and downstream judges, including GPT-4.1 on Arena-Hard (Liu et al., 2026b). Recent reasoning judges such as JudgeLRM (Chen et al., 2025a) and RM-R1 (Chen et al., 2025b) improve static benchmark scores but have not been rigorously tested in policy training. For visualization assessment, VisEval (Chen et al., 2024) and VIS-Shepherd (Pan et al., 2025) employ LLMs to judge charts within NL2VIS pipelines, while MMDR-Bench (Huang et al., 2026) proposes multi-stage evaluation with visual evidence fidelity checks for deep research reports, and VisJudge-Bench (Xie et al., 2026) builds a specialized model for chart aesthetics and quality. Despite its comprehensiveness, MMDR-Bench targets research reports over image-text bundles rather than structured tables, and its evaluation heavily depends on retrieval-based citation verification, which is not directly applicable to table-grounded reports. Our dual-judge framework verifies all claims directly against source tables to reduce evaluation fragility and judge-manipulability.

3 MM-REPORTBENCH

3.1 TASK FORMULATION

The multimodal report generation task is defined as follows: Given a set of one or more structured tables $\mathcal{T} = \{T_1, T_2, \ldots, T_n\}$ and a natural language instruction q that specifies the analytical focus (e.g., "Employment trends across U.S. sectors since 2006"), a model must produce a multimodal report $\mathcal{R} = (R_{\text{text}}, R_{\text{vis}})$. The textual component R_{text} is an article-length analytical report containing structured chapters (e.g., Overview, Data Overview, Core Analysis, Conclusions & Recommendations), while the visual component R_{vis} consists of one or more generated charts (bar charts, line plots, pie charts, scatter plots, etc.) that are embedded within the text and directly support the narrative with underlying table evidence. The report is evaluated against a human-authored reference, which specifies an expected chapter structure and a set of key insight points $K = \{k_1, \ldots, k_M\}$ that capture the essential analytical findings derivable from the tables. Unlike prior benchmarks that evaluate text and visuals independently or ignore visual output entirely, MM-ReportBench treats the report as an integrated multimodal artifact, requiring coherence across modalities.

3.2 DATASET CONSTRUCTION

(1) Table Collection and Domain Coverage. We collect 185 real-world business tables from publicly available sources including municipal open data platforms, industry association reports, corporate annual disclosures, and open-source tabular datasets, following the protocol of T2R-Bench. Tables span six high-impact industrial domains: finance, manufacturing, healthcare, education, retail, and IT operations. We specifically include 38 extremely large tables (over 50,000 cells, 20.5% of the dataset) and tables with complex header structures (32.4%) to test models' capabilities in processing massive-scale and hierarchical structured data. All tables undergo manual review to remove personally identifiable information and verify sufficient quantitative depth for meaningful analysis and visualization. The final collection contains bilingual tables (Chinese and English), reflecting the demands of global business analytics. Table 1 summarizes the key statistics of the MM-ReportBench dataset.

(2) Query Generation and Selection. We design a template-based self-instruct approach for analytical query generation. Ten templates covering common business reporting scenarios (trend analysis, comparative evaluation, ranking, anomaly detection, composition analysis, correlation exploration, forecasting, segment analysis, year-over-year change, cross-dimensional interaction) are populated with table-specific metadata and expanded into 3–5 concrete query candidates by GPT-4o. Two independent domain-expert and one senior expert annotators evaluate candidates

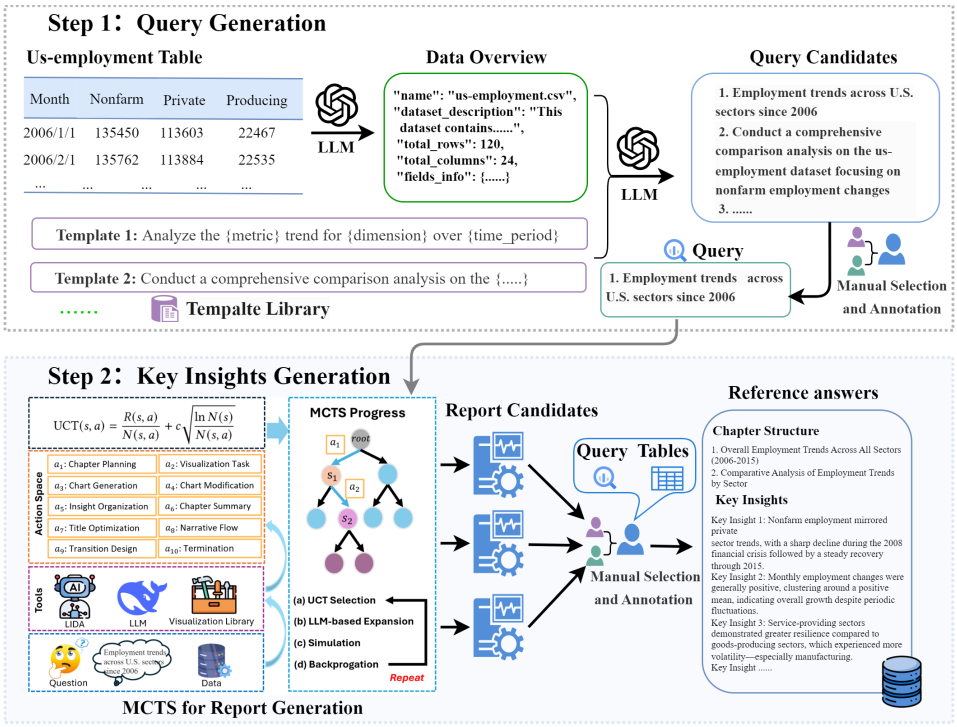


Figure 1: An overview of the construction pipeline for MM-ReportBench.

Table 1: Key statistics of MM-ReportBench.

Property	Value	Property	Value
Number of tables	185	Average cells per table	21200
Number of domains	6	Extremely large tables (> 50K cells)	38
Number of tasks	386	Complex header tables	60
Average tasks per table	2.09	Average reference keypoints per task	7.75
Average tables per multi-table task	4.2	Total reference keypoints	2992
Average rows per table	285		

based on answerability (fully solvable from provided tables without external knowledge), analytical focus (targeting a single coherent narrative thread), and complementarity (distinct, non-overlapping analytical dimensions for multiple queries per table). This yields 386 high-quality queries. The pipeline is shown in Figure 1 step 1 Query Generation.

(3) Reference Report Generation via Monte Carlo Tree Search. A central challenge in constructing a report-generation benchmark is obtaining high-quality reference reports that serve as stable evaluation targets. Human-authored reports are expensive to scale, while LLM-generated reports often suffer from hallucinations and verifiability issues. To balance quality, scalability, and diversity, we design a pipeline which reference report candidates are first generated through a Monte Carlo Tree Search (MCTS) (Coulom, 2007) procedure guided by a multi-dimensional reward function, and then curated and refined by human experts, which is demonstrated in Figure 1 MCTS for Report Generation step.

MCTS for Report Generation. We formulate report generation as sequential decision-making over an action space with two stages: Analysis (chapter planning, visualization task identification, chart generation, chart modification, insight organization, chapter summarization) and Refinement (title optimization, narrative flow improvement, and transition design). From an empty root node, the tree is expanded through iterative rollouts with action selection using the UCT (Kocsis & Szepesvári, 2006) criterion:

$$\text{UCT}(s,a) = \frac{R(s,a)}{N(s,a)} + c\sqrt{\frac{\ln N(s)}{N(s,a)}} \tag{1}$$

where $R(s,a)$ is the cumulative reward obtained after taking action a in state s, $N(s,a)$ is the visit count for (s,a), $N(s)$ is the total visit count for state s, and c is the exploration constant. We perform $K = 8$ rollouts per query, producing K complete report candidates.

Multi-Dimensional Report Reward Function. To guide the MCTS search toward high-quality, reference-worthy reports, we design a reward function $R(\mathcal{R})$ that evaluates a completed report along four dimensions, inspired by the evaluation criteria proposed in this work:

$$R(\mathcal{R}) = R_{\text{rep}}(\mathcal{R}) + R_{\text{pres}}(\mathcal{R}) + R_{\text{aes}}(\mathcal{R}) + R_{\text{nar}}(\mathcal{R}) \tag{2}$$

where R_{rep} captures data representation accuracy (combining numerical accuracy and information coverage metrics), R_{pres} assesses structural clarity and logical organization via LLM judge, R_{aes} evaluates visual chart quality, and R_{nar} assesses storytelling coherence and analytical depth (rewarding causal reasoning and actionable recommendations). Then we select the top $K = 3$ candidates with the highest terminal rewards to pass to the human curation stage. For more detailed information of MCTS for report generation, please refer to the appendix A.2.

Human Curation and Keypoint Annotation. The MCTS-generated report candidates per query undergo two-stage human curation. Two Annotators select the best candidate and collaboratively refine it through structural standardization, factual correction, and keypoint extraction (5–10 concise factual statements per report). Disagreements are resolved by a senior expert. The final output per task is a human-curated reference report with a defined chapter structure and keypoints. To further ensure the quality, we implement a three-stage verification pipeline: (1) factual verification where every numerical claim in each keypoint is cross-checked against raw table data by executing corresponding SQL queries; (2) necessity and exclusivity review where a dedicated quality-control team validates that each keypoint is necessary for the report's analytical goals and exclusively derivable from provided tables, removing trivial keypoints; and (3) we implement a reproducibility audit, where 50 randomly selected tasks are independently re-derived by three separate annotators using the same MCTS-assisted pipeline, achieving a Fleiss' κ of 0.85 with the original references and confirming the reliability of our reference construction process.

3.3 Dual-Judge Evaluation Framework

As shown in Figure 2, we design a dual-judge evaluation suite. Both modules operate with interpretable, dimension-level scoring.

(1) T-Judge. T-Judge evaluates R_{text} along three orthogonal dimensions:

(a) Structural Completeness (SC). This dimension measures how well the generated report's chapter organization, section flow, and topical coverage align with the reference chapter structure. We compute heading overlap score (F1 against reference headings with semantic equivalence via embedding-based matching, threshold 0.85) and section content coverage (T-Judge assesses whether matched sections cover the same analytical sub-dimensions). The final SC score is $0.3 \cdot \text{SC}_{\text{head}} + 0.7 \cdot \text{SC}_{\text{content}}$, emphasizing content-level alignment over surface formatting.

(b) Grounded Insight Verification (GIV). This dimension assesses factual faithfulness through two sub-metrics: (b1) *Numerical Accuracy* (GIV_{num}): We adapt the NAC protocol from T2R-Bench. First, we extract numerical sentence clusters, generate verification questions with the claimed values as target answers, and have three specialized code-generation LLMs independently answer by executing SQL queries on source tables. A claim is verified if at least two of three models agree within a 1% tolerance margin. (b2) *Keypoint Coverage* (GIV_{cov}): We measure semantic alignment between generated report sentences and reference keypoints via normalized mutual information:

$$\text{GIV}_{\text{cov}} = \frac{\sum_{i=1}^{M}\sum_{j=1}^{N} P(k_i, s_j) \log \frac{P(k_i,s_j)}{P(k_i)P(s_j)}}{-\sum_{i=1}^{M} P(k_i) \log P(k_i)} \tag{3}$$

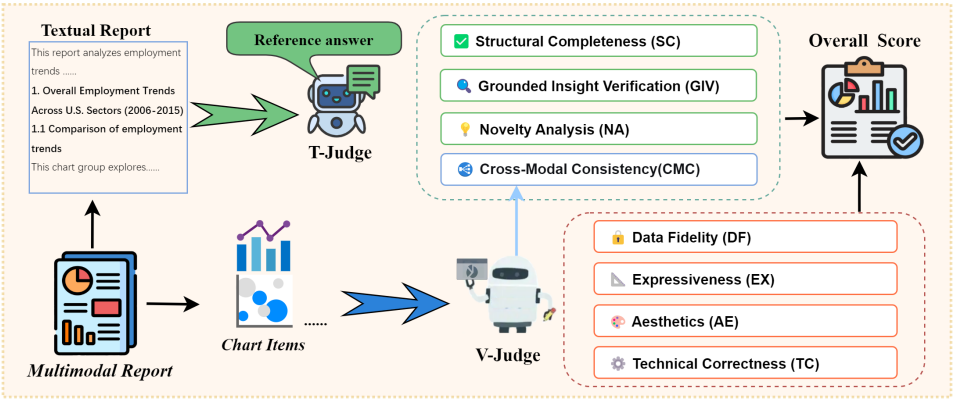


Figure 2: The Dual-Judge Evaluation Framework of MM-ReportBench. CMC verifies whether textual references to charts and chart-supported narratives are mutually consistent.

where probabilities are derived from a BERTScore-based similarity matrix, which provides a computable, reproducible, and paraphrase-robust metric. More detailed information can be found in Appendix A.3. The composite GIV score is the average of GIV_{num} and GIV_{cov}.

(c) Novelty Analysis (NA). To address the challenge that prior benchmarks cannot distinguish between paraphrasing and genuine insight discovery, we identify novel claims: sentence clusters with maximum cosine similarity below $\tau_{\text{novel}} = 0.6$ to any reference keypoint. These claims are evaluated by T-Judge along three sub-dimensions: non-triviality, actionability, and correctness (factual consistency with source tables). The overall NA score incorporates a penalty factor $\min(1, \alpha/|\mathcal{N}|)$ with $\alpha = 5$ to prevent models from receiving high scores through large quantities of low-quality novel claims, rewarding focused, high-quality insights over volume.

(d) Cross-Modal Consistency (CMC). It includes two sub-metrics: (d1) Cross-Modal Reference Consistency (CMRC). For each sentence that explicitly references a chart, T-Judge checks whether the referenced chart exists. CMRC = correctly grounded chart-referencing sentences / all chart-referencing sentences. (d2) Narrative Alignment (NRA). For each generated chart, **V-Judge** checks whether the chart is explained by the text, whether the described trend/values/labels match the chart, and whether the chart supports the section's analytical claim. NRA = correctly explained charts / all charts. The CMC value is the average of CMRC and NRA. The final T-Judge score is the average of SC, GIV, NA and CMC.

(2) V-Judge. V-Judge evaluates R_{vis} on four dimensions for chart-specific verification:

(a) Data Fidelity (DF). This dimension verifies plotted data against underlying tables through a two-step process: chart data extraction using a chart-to-table model and ground-truth comparison via SQL-generated expected data. Charts with more than 5% data points deviating beyond 1% tolerance (absolute) or 2 percentage points (proportions) are flagged as fidelity failures.

(b) Expressiveness (EX). This dimension evaluates communication effectiveness through chart type appropriateness, encoding clarity, and insight salience. V-Judge scores each chart on a 0–5 scale with a detailed rubric.

(c) Aesthetics (AE). This dimension assesses visual appeal through color harmony, layout quality, and absence of visual noise, anchored by calibration examples of exemplary and poor designs. V-Judge scores each chart on a 0–5 scale.

(d) Technical Correctness (TC). This dimension applies rule-based checks for missing chart elements, scale errors (including the critical zero-baseline check), label readability issues, and color accessibility problems. This provides objective, reproducible diagnostics. V-Judge scores each chart on a 0–5 scale. The final VisScore is the average of DF, EX, AE and TC.

(3) Overall Score. The overall MM-ReportBench score is a weighted combination of T-Judge and V-Judge scores, with a default weighting of 0.6 for T-Judge and 0.4 for V-Judge, reflecting the typical priority in business reporting where textual analysis carries the primary analytical burden

while visualizations provide essential support. **All report dimensions and the overall score are normalized to [0,100] for final score**. The implementation details of dual-judge are provided in Appendix A.5.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUP

Evaluation Protocol. For text-only models, only the textual component is generated, and the visual score is set to zero. For multimodal models, chart generation can be triggered via code execution or native multimodal output, depending on model capabilities. All models use the same prompt template and are allowed a maximum of 4096 tokens for the textual report and up to 8 visualizations. The reference reports and keypoints serve as ground truth for evaluation. The judge LLMs (GPT-4o and GPT-4o with vision) are used for T-Judge and V-Judge scoring at temperature 0.2 to maximize reproducibility. For Metrics, we report T-Judge scores, V-Judge scores, and the overall score. Each report is evaluated three times, and the average scores are reported.

Baseline Models. We evaluate 20 state-of-the-art models spanning four categories: (i) text-only LLMs (GPT-4o-text (Achiam et al., 2023), Claude-3.5-Sonnet (Anthropic, 2024), DeepSeek-V3 (Guo et al., 2025), Qwen3-32B, Llama-3.3-70B, GPT-5.5 (Achiam et al., 2023)), (ii) vision-language models with native chart generation (GPT-4o (Achiam et al., 2023), GPT-4.1, Gemini-2.5-Pro, Gemini-3-Flash, Claude-4.5-Sonnet (Anthropic, 2025), Qwen3-VL-235B (qwe, 2025), Gemini 3.5-Flash, Claude Opus 4.7 (ant, 2025)), (iii) code-augmented multimodal systems that generate charts via Python execution (DeepSeek-R1 with code interpreter (Guo et al., 2025), Qwen3-Coder, TableGPT2 (Su et al., 2024)), and (iv) specialized deep research agents (Gemini Deep Research (goo, 2025), ChatGPT Deep Research, Perplexity Sonar Deep Research). All models are accessed via official APIs or open-source inference between November 2025 and January 2026.

4.2 MAIN RESULTS

Table 2 presents the overall and dimension-level performance of all evaluated models across the all 386 tasks. We highlight several key findings.

Finding 1: Text-only LLMs are fundamentally insufficient for multimodal business reporting. Without visualization capabilities, these models score zero on V-Judge, capping their overall MM-ReportBench score below 24. Even in textual dimensions, they lag behind multimodal counterparts in GIV (up to 10 points lower), likely because visual thinking aids in data comprehension and fact checking.

Finding 2: Code-augmented multimodal systems lead, but the gap is modest. DeepSeek-R1 with code interpreter achieves the highest overall score (68.9), outperforming the strongest vision-language model (Gemini-2.5-Pro at 66.8) by 2.1 points. The advantage is most pronounced in Data Fidelity (85.1 vs 82.4) and Expressiveness (78.3 vs 76.1), attributed to the precise control that code execution provides over chart rendering. However, even this best model achieves only 68.9, far from human-level performance (human baseline: 92.3, see section 4.4, which is computed from independently human-written multimodal reports, not from the curated reference reports), indicating substantial room for improvement.

Finding 3: All models struggle with insight novelty. The Novelty Analysis (NA) scores are markedly lower than SC and GIV across the board. DeepSeek-R1, the best performer, reaches only 38.7 on NA, revealing that current models predominantly rely on paraphrasing existing data patterns and rarely produce genuinely non-trivial, actionable insights. This validates our inclusion of NA as a critical dimension absent from prior benchmarks.

Finding 4: Visualization quality remains a bottleneck. Even the top V-Judge scores hover around 80, with Data Fidelity being the strongest dimension (85.1 for DeepSeek-R1) and Aesthetics consistently weaker (max 76.0). Technical correctness issues (e.g., missing axis labels, unreadable legends) are frequent, especially among multimodal models without explicit code execution.

Table 2: Overall performance on MM-ReportBench. We report the main scores and the Overall score (Overall). The best results in each column are **bolded**, the second-best are underlined.

Model	SC	GIV	NA	CMC	T-Score	DF	EX	AE	TC	V-Score	Overall
Text-only LLMs											
GPT-4o (text)	72.3	48.5	28.1	0.0	37.2	—	—	—	—	0.0	22.3
Claude-3.5-Sonnet	70.8	45.2	25.6	0.0	35.4	—	—	—	—	0.0	21.2
DeepSeek-V3	68.4	46.1	27.3	0.0	35.5	—	—	—	—	0.0	21.3
Qwen3-32B	67.2	43.8	24.5	0.0	33.9	—	—	—	—	0.0	20.3
Llama-3.3-70B	65.9	41.2	22.7	0.0	32.5	—	—	—	—	0.0	19.5
GPT-5.5	72.9	51.2	32.7	0.0	39.2	—	—	—	—	0.0	23.5
Vision-Language Models											
GPT-4o	75.4	55.8	34.2	64.1	57.4	78.2	72.5	68.3	82.1	74.8	64.3
GPT-4.1	74.8	54.6	32.8	62.5	56.1	76.5	70.8	66.9	80.4	73.2	63.0
Gemini-2.5-Pro	76.1	57.3	35.5	67.2	59.0	82.4	76.1	73.0	84.6	78.4	66.8
Gemini-3-Flash	73.5	53.1	32.1	63.5	55.6	79.8	73.2	70.5	82.8	76.1	63.8
Gemini-3.5-Flash	76.2	54.6	33.3	65.3	57.3	80.6	72.9	71.9	82.7	76.6	65.0
Claude-4.5-Sonnet	74.2	54.8	33.6	64.5	56.7	80.3	74.0	72.1	83.3	76.8	64.8
Claude Opus 4.7	68.2	53.1	34.2	63.1	54.6	78.8	72.8	72.8	82.6	77.1	63.6
Qwen3-VL-235B	71.6	51.9	30.4	61.0	53.7	77.9	71.5	67.8	79.2	73.0	61.4
Code-augmented Multimodal Systems											
DeepSeek-R1	**78.2**	**59.4**	**38.7**	**69.1**	**61.35**	**85.1**	**78.3**	74.2	**86.5**	**80.2**	**68.9**
Qwen3-Coder-32B	72.9	52.4	31.5	63.1	55.0	80.5	74.1	71.3	83.0	76.5	63.6
TableGPT2-7B	65.3	44.7	23.2	52.3	46.3	68.9	62.5	58.4	71.2	64.0	53.4
Deep Research Agents											
Gemini Deep Research	77.5	58.1	37.2	68.1	60.2	83.1	77.5	**76.0**	85.8	79.6	67.9
ChatGPT Deep Research	75.2	55.9	35.0	65.5	57.9	80.9	74.8	73.5	83.9	77.3	65.7
Perplexity Sonar DR	73.8	53.5	33.8	63.5	56.2	79.0	72.9	70.8	82.5	75.5	63.9

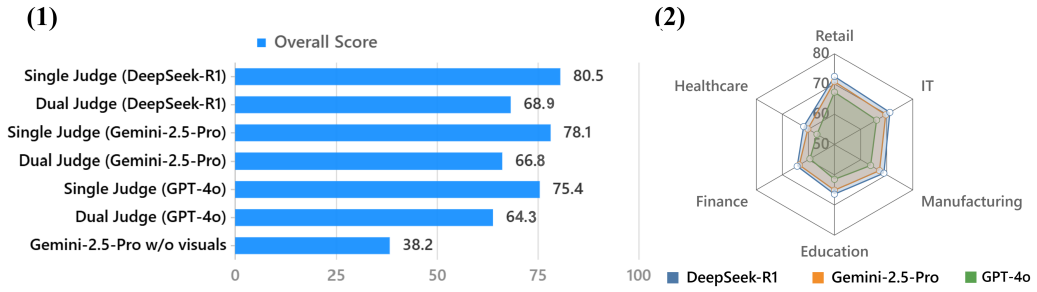


Figure 3: (1) shows the overall scores of three representative models under Single Judge vs. Dual-Judg. Dual-Judge's scores are consistently lower, revealing a score inflation of roughly 8–12 points in the single-judge setting. The chart also illustrates the severe performance drop when visual generation is disabled (w/o visuals). (2) is a radar chart of domain-level MM-ReportBench scores for the top three models. DeepSeek-R1 maintains a lead across all domains, but all models exhibit a similar "hard domain" pattern: Finance and Healthcare are consistently more difficult than Retail and IT Operations.

4.3 Ablation Study: Modality and Evaluation Configurations

Impact of Visual Generation. To quantify the value of multimodal generation, we compare the performance of the same base model (Gemini-2.5-Pro) when generating reports with vs. without charts. As shown in Figure 3(1), removing visualizations (text-only setting) drops the overall score from 66.8 to 38.2, but interestingly, the T-Judge sub-scores also degrade (GIV from 57.3 to 49.1, NA from 35.5 to 29.3). This suggests that the process of creating charts may force the model to better verify its data, indirectly improving textual factuality. Conversely, a purely visual report without text scores only 46.5 overall, highlighting the complementarity of modalities.

Impact of Dual-Judge Design. To validate the need for a separate V-Judge, we experiment with a *unified judge* (single GPT-4o prompt that scores both text and visuals together). The result is shown in Figure 3 (1). The unified judge consistently inflates overall scores by 8–12 points compared to

our dual-judge framework, and its correlation with human expert judgments drops from $r = 0.89$ to $r = 0.72$. This confirms that splitting the evaluation into specialized judges improves accuracy and reduces the halo effect (where good text masks poor charts, and vice versa), consistent with the concerns about single-judge fragility.

4.4 Performance Analysis

Domain-Specific Analysis. Figure 3 (2) shows the per-domain performance of the three best models. All three models perform best in *Retail* and *IT Operations*, where tables tend to be simpler and instructions are more straightforward. Performance drops sharply in *Healthcare* (complex temporal patient data, heavy numerical content) and *Finance* (multi-table consolidation, dense accounting metrics). For example, DeepSeek-R1 achieves 72.5 in Retail but only 61.8 in Healthcare, a gap of over 10 points. The main degradation factors are hallucinated numerical values in healthcare (adverse events misattributed) and structural confusion in multi-table financial reports (balance sheet and income statement items mixed). This domain sensitivity underscores the need for domain-specific adaptation strategies.

Human Evaluation and Consistency. We conduct a human evaluation to assess the alignment of our automated metrics with expert judgment. Six professional business analysts evaluate a stratified sample of 100 generated reports (covering all domains and top models). They score each report on the same dimensions as the T-Judge and V-Judge, following the rubrics described in Section 3.3. Inter-annotator agreement reaches Fleiss' $\kappa = 0.83$, indicating high consistency. As shown in Appendix Table 3, the MM-ReportBench automated scores achieve a Pearson correlation of 0.908 with human overall ratings, and dimension-level correlations range from 0.82 to 0.94. Notably, the automated system tends to be *stricter* than humans (average score difference ~ 3.5 points), likely because automated numerical verification catches subtle hallucinations that humans may overlook. This conservative bias is preferable for a benchmark intended to drive improvement.

Error Analysis We manually analyze 200 randomly sampled reports to categorize failure modes. The error taxonomy is extended for multimodal settings. Appendix Table 4 shows the distribution of primary error types in generated report. Numerical hallucination remains the most pervasive issue, occurring in 38.5% of reports. Hallucinated values often appear plausible but are off by an order of magnitude or swapped between columns. Trivial insights (31%) reflect the model's tendency to state obvious facts (e.g., "sales increased") without explaining why or providing context. Chart-data mismatch (22%) includes cases where a line chart claims a growth rate of 20%, but the underlying table shows 5%. Multi-table confusion (14%) is especially prevalent in finance reports with multiple interconnected statements. Truncation occurs in 12.5% of reports involving extremely large tables, where models lose context after the maximum token limit, leading to incomplete or incorrect analysis.

5 Conclusion

We introduced MM-ReportBench, the first unified multimodal benchmark for table-to-report generation with a comprehensive dual-judge evaluation framework. It contains 386 expert-curated tasks across six business domains, together with reference reports and verifiable key insight points. Reference reports are constructed through an MCTS-based generation pipeline followed by multi-stage human curation and verification. Our work is motivated by and directly addresses the key limitations identified across prior research. Extensive experiments on 20 models reveal that multimodal business report generation remains an open challenge. The best model achieves 68.9 overall score versus the human baseline of 92.3, with insight novelty and visualization quality as the major bottlenecks.Our ablation further confirms that a single unified judge inflates scores by 8–12 points, whereas the dual-judge design reduces cross-modality compensation and aligns better with human expert judgment. We hope MM-ReportBench provides a rigorous testbed for evaluating and advancing reliable multimodal report generation. Future work includes improving numerical faithfulness, novelty-oriented insight generation, chart fidelity, multi-table reasoning, domain adaptation, and evaluation protocols robust to reward hacking.

6 AI USE STATEMENT

In this work, we used generative AI tools for the following tasks with required disclosure: generating candidate analytical queries from table metadata; generating candidate reference reports through the MCTS pipeline; serving as LLM judges for T-Judge, V-Judge, and cross-modal consistency evaluation; and generating SQL/code for numerical verification of claims and keypoints. We have not used generative AI tools for the following required-disclosure tasks: formulating or proving mathematical claims, proposing or refining hypotheses, designing the core research methodology, conducting qualitative or thematic data analysis, translating data, cleaning or reformatting datasets, or interpreting results. Other required-disclosure tasks not listed above were not performed with generative AI. Additionally, we used generative AI tools for recommended-disclosure tasks, including editing the manuscript for readability, formatting references, suggesting paper structure and title, and assisting with code for data processing and evaluation. All AI-assisted work was reviewed and verified by the authors. In particular, LLM-generated queries, reference candidates, keypoints, charts, and evaluation outputs were checked against source tables and/or expert judgment; the final human-curated reference reports and keypoint annotations were produced by human annotators. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

REFERENCES

Claude opus 4.7. Anthropic Official Announcement, 2025. URL https://www.anthropic.com/news/claude-opus-4-7. Accessed: 2026-07-27.

Gemini deep research. Google Gemini, 2025. URL https://gemini.google.com/. Accessed: 2026-07-27.

Qwen3-vl-235b. Qwen AI, 2025. URL https://qwen.ai/. Accessed: 2026-07-27.

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Anthropic. Claude 3.5. Anthropic Blog, 2024. URL https://www.anthropic.com/news/claude-3-5-sonnet. Introducing Claude Sonnet 3.5.

Anthropic. Claude 4.5. Anthropic Blog, 2025. URL https://www.anthropic.com/news/claude-sonnet-4-5. Introducing Claude Sonnet 4.5, released 29 Sep 2025.

Nan Chen, Yuge Zhang, Jiahang Xu, Kan Ren, and Yuqing Yang. Viseval: A benchmark for data visualization in the era of large language models. *IEEE Transactions on Visualization and Computer Graphics*, 31(1):1301–1311, 2024.

Nuo Chen, Zhiyuan Hu, Qingyun Zou, Jiaying Wu, Qian Wang, Bryan Hooi, and Bingsheng He. Judgelrm: Large reasoning models as a judge. *arXiv preprint arXiv:2504.00050*, 2025a.

Wenhu Chen, Hongmin Wang, Jianshu Chen, Yunkai Zhang, Hong Wang, Shiyang Li, Xiyou Zhou, and William Yang Wang. Tabfact: A large-scale dataset for table-based fact verification. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=rkeJRhNYDH.

Xiusi Chen, Gaotang Li, Ziqi Wang, Bowen Jin, Cheng Qian, Yu Wang, Hongru Wang, Yu Zhang, Denghui Zhang, Tong Zhang, et al. Rm-r1: Reward modeling as reasoning. *arXiv preprint arXiv:2505.02387*, 2025b.

Zhoujun Cheng, Haoyu Dong, Zhiruo Wang, Ran Jia, Jiaqi Guo, Yan Gao, Shi Han, Jian-Guang Lou, and Dongmei Zhang. Hitab: A hierarchical table dataset for question answering and natural language generation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1094–1110, 2022.

Rémi Coulom. Efficient selectivity and backup operators in monte-carlo tree search. In H. Jaap van den Herik, Paolo Ciancarini, and H. H. L. M. (Jeroen) Donkers (eds.), *Computers and Games*, pp. 72–83, Berlin, Heidelberg, 2007. Springer Berlin Heidelberg. ISBN 978-3-540-75538-8.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, et al. Deepseek-r1 incentivizes reasoning in llms through reinforcement learning. *Nature*, 645(8081):633–638, 2025.

Peizhou Huang, Zixuan Zhong, Zhongwei Wan, Donghao Zhou, Samiul Alam, Xin Wang, Zexin Li, Zhihao Dou, Li Zhu, Jing Xiong, Chaofan Tao, Yan Xu, Dimitrios Dimitriadis, Tuo Zhang, and Mi Zhang. Mmdeepresearch-bench: A benchmark for multimodal deep research agents, 2026. URL https://arxiv.org/abs/2601.12346.

Levente Kocsis and Csaba Szepesvári. Bandit based monte-carlo planning. In Johannes Fürnkranz, Tobias Scheffer, and Myra Spiliopoulou (eds.), *Machine Learning: ECML 2006*, pp. 282–293, Berlin, Heidelberg, 2006. Springer Berlin Heidelberg. ISBN 978-3-540-46056-5.

Teng Lin. Structured retrieval-augmented generation for multi-entity question answering over heterogeneous sources. In *2025 IEEE 41st International Conference on Data Engineering Workshops (ICDEW)*, pp. 253–258, 2025. doi: 10.1109/ICDEW67478.2025.00036.

Wei Liu, Peijie Yu, Michele Orini, Yali Du, and Yulan He. Hunt instead of wait: Evaluating deep data research on large language models, 2026a. URL https://arxiv.org/abs/2602.02039.

Yixin Liu, Yue Yu, DiJia Su, Sid Wang, Xuewei Wang, Song Jiang, Bo Liu, Arman Cohan, Yuandong Tian, and Zhengxing Chen. Examining reasoning llms-as-judges in non-verifiable llm post-training, 2026b. URL https://arxiv.org/abs/2603.12246.

Linyong Nan, Chiachun Hsieh, Ziming Mao, Xi Victoria Lin, Neha Verma, Rui Zhang, Wojciech Kryściński, Hailey Schoelkopf, Riley Kong, Xiangru Tang, et al. Fetaqa: Free-form table question answering. *Transactions of the Association for Computational Linguistics*, 10:35–49, 2022.

Bo Pan, Yixiao Fu, Ke Wang, Junyu Lu, Lunke Pan, Ziyang Qian, Yuhan Chen, Guoliang Wang, Yitao Zhou, Li Zheng, et al. Vis-shepherd: Constructing critic for llm-based data visualization generation. *arXiv preprint arXiv:2506.13326*, 2025.

Ankur Parikh, Xuezhi Wang, Sebastian Gehrmann, Manaal Faruqui, Bhuwan Dhingra, Diyi Yang, and Dipanjan Das. Totto: A controlled table-to-text generation dataset. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 1173–1186, 2020.

Panupong Pasupat and Percy Liang. Compositional semantic parsing on semi-structured tables. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 1470–1480, 2015.

Aofeng Su, Aowen Wang, Chao Ye, Chen Zhou, Ga Zhang, Guangcheng Zhu, Haobo Wang, Haokai Xu, Hao Chen, Haoze Li, Haoxuan Lan, Jiaming Tian, Jing Yuan, Junbo Zhao, Junlin Zhou, Kaizhe Shou, Liangyu Zha, Lin Long, Liyao Li, Pengzuo Wu, Qi Zhang, Qingyi Huang, Saisai Yang, Tao Zhang, Wentao Ye, Wufang Zhu, Xiaomeng Hu, Xijun Gu, Xinjie Sun, Xiang Li, Yuhang Yang, and Zhiqing Xiao. Tablegpt2: A large multimodal model with tabular data integration, 2024. URL https://arxiv.org/abs/2411.02059.

Xianjie Wu, Jian Yang, Linzheng Chai, Ge Zhang, Jiaheng Liu, Xeron Du, Di Liang, Daixin Shu, Xianfu Cheng, Tianzhen Sun, et al. Tablebench: A comprehensive and complex benchmark for table question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 25497–25506, 2025.

Yupeng Xie, Zhiyang Zhang, Yifan Wu, Sirong Lu, Jiayi Zhang, Zhaoyang Yu, Jinlin Wang, Sirui Hong, Bang Liu, Chenglin Wu, and Yuyu Luo. Visjudge-bench: Aesthetics and quality assessment of visualizations. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=lG1HWWdEbN.

Tao Yu, Rui Zhang, Kai Yang, Michihiro Yasunaga, Dongxu Wang, Zifan Li, James Ma, Irene Li, Qingning Yao, Shanelle Roman, et al. Spider: A large-scale human-labeled dataset for complex and cross-domain semantic parsing and text-to-sql task. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pp. 3911–3921, 2018.

Jie Zhang, Changzai Pan, Sishi Xiong, Kaiwen Wei, Yu Zhao, Xiangyu Li, Jiaxin Peng, Xiaoyan Gu, Jian Yang, Wenhan Chang, Zhenhe Wu, Jiang Zhong, Shuangyong Song, and Xuelong Li. T2R-BENCH: A benchmark for real world table-to-report task. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 22427–22451, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.1141. URL https://aclanthology.org/2025.emnlp-main.1141/.

Victor Zhong, Caiming Xiong, and Richard Socher. Seq2sql: Generating structured queries from natural language using reinforcement learning. *arXiv preprint arXiv:1709.00103*, 2017.

A Appendix

A.1 Related work

Autonomous Data Analysis and Deep Research. A parallel line of work concerns agentic data analysis and deep research, where models must autonomously decide what to investigate and produce comprehensive findings. DDR-Bench (Liu et al., 2026a) formalizes *investigatory intelligence* as the ability of LLMs to explore structured databases without explicit questions, and demonstrates that current agents suffer from limited exploration breadth, premature termination, and weak insight synthesis, even when scaffolded with tool use. MMDR-Bench (Huang et al., 2026) extends this to multimodal deep research agents that must coordinate web browsing, evidence gathering, and report writing with visual grounding. Both benchmarks highlight critical bottlenecks: factual hallucination, insufficient coverage of key evidence, and a trade-off between writing fluency and evidential faithfulness. Our MM-ReportBench fills this gap by uniting the structured-table grounding of T2R-Bench with the exploratory depth of DDR-Bench and the multimodal verification rigor of MMDR-Bench, and by introducing a standardized reference-report structure that enables fine-grained, reproducible evaluation across six real-world business domains.

A.2 MCTS

We formulate report generation as a sequential decision-making process over an action space $\mathcal{A}$ consisting of two stages with six action types:

- **Analysis Stage Actions:**
  - A_1: **Chapter Plan** — Propose a chapter-level outline for the report based on the query and table data.
  - A_2: **Visualization Task** — Identify specific chart-worthy analytical points and specify chart types, data filters, and encodings.
  - A_3: **Chart Generation** — Generate one or more visualizations (bar charts, line plots, pie charts, etc.) via code execution or native multimodal output.
  - A_4: **Chart Modification** — Refine previously generated charts (e.g., adjust scales, fix labeling, improve color schemes) based on self-critique.
  - A_5: **Organize Insight** — Draft a textual insight paragraph grounded in the table data and linked to the supporting chart(s).
  - A_6: **Chapter Summary** — Synthesize all insights within a chapter into a cohesive section.
- **Refinement Stage Actions:**
  - R_1: **Title Optimization** — Refine the report title and section headings for clarity and impact.
  - R_2: **Narrative Flow** — Improve transitions between paragraphs and chapters for coherent reading.
  - R_3: **Transition Design** — Add bridging sentences and forward-looking statements between major sections.

The MCTS procedure follows four phases:

1. **Selection:** Starting from the root, recursively select child nodes with the highest UCT value until reaching an expandable node.
2. **Expansion:** Create one or more child nodes by applying an LLM-action from $\mathcal{A}$ to the selected node's report state.
3. **Simulation:** From the newly expanded node, complete the report to a terminal state by sequentially applying the remaining actions in a predefined order (analysis stage first, then refinement stage). To encourage diversity, we sample the LLM actions at a high temperature $T = 0.8$ during simulation.
4. **Backpropagation:** Evaluate the completed terminal report using the multi-dimensional reward function (described below) and propagate the reward upward to update the statistics of all ancestor nodes.

A.3 BERTSCORE-BASED SIMILARITY MATRIX FOR KEYPOINT COVERAGE

We describe in detail how the BERTScore-based similarity matrix is constructed for the **Grounded Insight Verification — Keypoint Coverage ($\mathbf{GIV_{cov}}$)** metric.

A.3.1 BACKGROUND

GIV_{cov} measures the degree to which the sentences of a generated report cover the reference keypoints $K = \{k_1, \ldots, k_M\}$. The metric is based on **normalized mutual information** (NMI) and thus requires a joint probability distribution, which we derive from a semantic similarity matrix between keypoint–sentence pairs.

A.3.2 CONSTRUCTING THE SIMILARITY MATRIX $\mathbf{S}$

Let k_i be the i-th reference keypoint ($1 \leq i \leq M$) and s_j be the j-th sentence (or sentence cluster) extracted from the generated report ($1 \leq j \leq N$).

For every pair (k_i, s_j), we compute a similarity score $\mathbf{S}_{ij} \in [0, 1]$ using **BERTScore**. The procedure consists of two steps:

1. **Embedding.** A pre-trained BERT model (e.g., `bert-base-uncased` or `roberta-large`) is used to obtain token-level contextual embeddings for both k_i and s_j.
2. **Similarity computation.** BERTScore performs greedy cosine-similarity matching between the tokens of the source (keypoint) and the target (sentence). Let $\mathbf{x}_t$ be the embedding of the t-th token in the keypoint and $\mathbf{y}_u$ the embedding of the u-th token in the sentence. The token-level match score is

$$\text{sim}_{\text{token}}(t) = \max_u \cos(\mathbf{x}_t, \mathbf{y}_u). \tag{4}$$

The sentence-level F1 score is then obtained by averaging the bidirectional matches (keypoint $\rightarrow$ sentence and sentence $\rightarrow$ keypoint) and applying a simple rescaling operation:

$$\mathbf{S}_{ij} = \text{BERTScore}_{\text{F1}}(k_i, s_j). \tag{5}$$

Thus, $\mathbf{S}$ is an $M \times N$ matrix whose entries represent the semantic alignment between the corresponding keypoint–sentence pairs.

A.3.3 DERIVING PROBABILITY DISTRIBUTIONS FROM THE SIMILARITY MATRIX

To transform the similarity matrix into quantities suitable for information-theoretic computation, we normalize it globally.

- **Joint probability.** Normalizing each element by the total sum yields the joint probability distribution:

$$P(k_i, s_j) = \frac{\mathbf{S}_{ij}}{\sum_{i=1}^{M} \sum_{j=1}^{N} \mathbf{S}_{ij}}. \tag{6}$$

This value represents the proportion of total semantic association that is captured by the pair (k_i, s_j).

- **Marginal probabilities.** Summing over the appropriate dimensions gives the marginal probabilities for keypoints and sentences:

$$P(k_i) = \sum_{j=1}^{N} P(k_i, s_j), \qquad P(s_j) = \sum_{i=1}^{M} P(k_i, s_j). \tag{7}$$

  $P(k_i)$ reflects the relative semantic importance of keypoint i, while $P(s_j)$ captures the contribution of sentence j.

A.3.4 EXAMPLE

Suppose there are $M = 2$ keypoints and $N = 3$ sentences, and the BERTScore matrix is

$$\mathbf{S} = \begin{bmatrix} 0.85 & 0.40 & 0.10 \\ 0.20 & 0.90 & 0.30 \end{bmatrix}.$$

The total sum is $0.85 + 0.40 + 0.10 + 0.20 + 0.90 + 0.30 = 2.75$. Hence,

$$P(k_1, s_1) = \frac{0.85}{2.75} \approx 0.309, \quad P(k_1) = 0.309 + 0.145 + 0.036 = 0.490,$$

$$P(s_1) = 0.309 + 0.073 = 0.382, \quad \text{etc.}$$

These probabilities are then used directly in the NMI formula:

$$\mathrm{GIV}_{\mathrm{cov}} = \frac{\sum_{i=1}^{M} \sum_{j=1}^{N} P(k_i, s_j) \log \frac{P(k_i, s_j)}{P(k_i)\, P(s_j)}}{-\sum_{i=1}^{M} P(k_i) \log P(k_i)}. \tag{8}$$

A.3.5 DESIGN CHOICES AND REMARKS

- **Why BERTScore.** Unlike strict n-gram matching, BERTScore tolerates paraphrasing and only captures semantic equivalence — a necessary property when evaluating keypoints that may be expressed in different wording.
- **Global normalization.** Transforming $\mathbf{S}$ into a probability distribution ensures that the NMI metric is well-behaved: it increases monotonically with coverage and approaches 1 when all keypoints are fully covered and the report contains little irrelevant content.
- **Sentence clustering.** To improve stability, we first cluster sentences in the report and then compute similarity using the representative embedding of each cluster, mitigating noise from sentence-splitting differences.

This pipeline, which converts a BERTScore similarity matrix into a mutual-information-based coverage score, provides a **computable, reproducible, and paraphrase-robust** metric for evaluating how faithfully a generated report captures the essential insights specified by the reference keypoints.

A.4 CROSS-MODAL CONSISTENCY EVALUATION

Definition. Cross-Modal Consistency (CMC) measures whether the textual and visual components of a generated report support each other. CMC is treated as the fourth dimension of T-Judge. It contains two complementary sub-metrics:

- **Cross-Modal Reference Consistency (CMRC)**: whether sentences that refer to charts are factually and semantically consistent with the referenced charts and the underlying source tables.
- **Narrative Alignment (NRA)**: whether each generated chart is explained by the text, whether the explanation matches the chart's visual encoding, and whether the chart supports the analytical claim of its section.

Both CMRC and NRA are normalized to $[0, 100]$. The final CMC score is

$$\mathrm{CMC} = 0.5 \cdot \mathrm{CMRC} + 0.5 \cdot \mathrm{NRA}. \tag{9}$$

CMRC: Cross-Modal Reference Consistency. Let $R = \{r_1, \ldots, r_P\}$ be the set of sentences in the generated report that explicitly or implicitly refer to a chart. For each reference sentence r_p, CMRC identifies the target chart c_p, then verifies the following four aspects:

1. **Existence**: the referenced chart c_p actually exists in the report.
2. **Numerical consistency**: any numerical value, percentage, rank, or aggregate mentioned in r_p agrees with the chart data and the source table.
3. **Trend consistency**: any directional or comparative claim (increase, decrease, peak, trough, faster, slower, etc.) agrees with the chart.
4. **Encoding and label consistency**: the sentence correctly describes the chart type, axes, legend, units, categories, and visual encoding.

Each reference sentence receives a score $s(r_p) \in \{0, 0.5, 1\}$:

- 1: all four aspects pass;
- 0.5: the reference and main trend are correct, but there is a minor numerical, label, or encoding error;
- 0: the chart does not exist, the core claim contradicts the chart, or the sentence misattributes a chart to the wrong evidence.

CMRC is computed as

$$\mathrm{CMRC} = 100 \cdot \frac{1}{|R|} \sum_{p=1}^{|R|} s(r_p). \tag{10}$$

If the report contains no sentence that refers to any chart, CMRC is set to 0, because a multimodal report is expected to connect its visuals to its narrative.

NRA: Narrative Alignment. Let $C = \{c_1, \ldots, c_Q\}$ be the set of generated charts. For each chart c_q, NRA checks:

1. whether at least one sentence or caption explains c_q;
2. whether the explanation matches the chart type, axes, legend, units, and visual pattern;
3. whether the explanation supports the analytical claim of the section in which c_q appears.

Each chart receives a score $t(c_q) \in \{0, 0.5, 1\}$:

- 1: the chart is clearly explained and the explanation is fully consistent with the visual encoding and the section claim;
- 0.5: the chart is mentioned, but the explanation is incomplete, vague, or partially inconsistent;
- 0: the chart is not explained at all, or the explanation contradicts the chart.

NRA is computed as

$$\mathrm{NRA} = 100 \cdot \frac{1}{|C|} \sum_{q=1}^{|C|} t(c_q). \tag{11}$$

If the report contains no chart, NRA is set to 0 under the multimodal setting.

Prompt template. The CMC prompt is as follows.

```
System:
You are a strict cross-modal consistency evaluator for multimodal
business reports. You receive:
(1) the source tables,
(2) the generated textual report,
(3) the generated charts with captions,
```

```
(4) the section headings,
(5) the extracted chart data when available.
Your task is to compute CMRC and NRA.
Do not reward fluent writing. Penalize any chart-text mismatch,
unexplained chart, or text that refers to a non-existent chart.
Return only valid JSON.

User:
[SOURCE TABLES]
{table_text_or_markdown}

[GENERATED REPORT]
{report_text}

[CHARTS]
Chart ID: C1
Caption: {caption_1}
Section: {section_1}
Image: {image_1}
Extracted data: {chart_data_1}

Chart ID: C2
Caption: {caption_2}
Section: {section_2}
Image: {image_2}
Extracted data: {chart_data_2}

[TASK]
1. Identify every sentence that explicitly or implicitly refers to a chart.
2. For each such sentence, find the target chart ID and score it as 0, 0.5, or 1
   using the CMRC rules.
3. For each chart, determine whether the text explains it and score it as
   0, 0.5, or 1 using the NRA rules.
4. Return JSON with the following schema.

{
  "cmrc": [
    {
      "sentence_id": "S3",
      "sentence": "...",
      "chart_id": "C1",
      "score": 1,
      "reason": "..."
    }
  ],
  "nra": [
    {
      "chart_id": "C1",
      "score": 1,
      "reason": "..."
    }
  ],
  "cmrc_score": 100.0,
  "nra_score": 100.0,
  "cmc_score": 100.0
}
```

Decision rules and tolerances. The following rules are used by the T-Judge's cross-modal sub-module and, when needed, by human adjudicators.

- **Numerical tolerance.** For absolute values, a claim is considered consistent if it is within 1% of the chart value. For proportions, percentages, or rates, the tolerance is 2 percentage points. Example: a chart value of 5.0% is consistent with a textual claim of $5.0\% \pm 2.0$ percentage points.
- **Trend direction.** The direction of change must match exactly. "Increase" vs. "decrease" is a hard error. "Sharp increase" vs. "slight increase" is treated as a partial mismatch unless the chart clearly supports the magnitude.
- **Rank and comparison.** If the text says "A is higher than B", the chart must show the same ordering. Reversed ordering is a hard error.
- **Labels and units.** Category names, axis labels, legends, and units must match after normalization. Acceptable synonyms and unit conversions are allowed, e.g., "USD million" vs. "million USD", or "Q1" vs. "first quarter".
- **Chart type.** If the text says "bar chart" but the chart is a line plot, this is a minor error if the analytical claim is still correct, and a hard error if the chart type is central to the claim.
- **Missing chart reference.** If the text refers to "Figure 3" or "the chart" but no such chart exists, CMRC for that sentence is 0.
- **Unexplained chart.** If a chart appears in the report but no sentence or caption explains it, NRA for that chart is 0.
- **Extraction confidence.** If the chart-to-table extraction confidence is below 0.8, the ground-truth comparison uses the source table and the original chart specification instead of the extracted image data. If neither is available, the chart is marked as "uncertain" and excluded from the denominator, and this exclusion is reported in the appendix.

Examples. **Example 1 (CMRC = 1, NRA = 1).** Text: "As shown in Figure 1, employment declined sharply in 2009."
Chart 1: a line chart showing a sharp drop in 2009.
Source table: the 2009 employment value is lower than 2008 by a large margin.
Decision: the chart exists, the trend matches, and the magnitude is supported. CMRC = 1. The chart is explained by the text and supports the section claim. NRA = 1.

Example 2 (CMRC = 0, NRA = 0.5). Text: "Figure 2 shows a 20% growth in 2010."
Chart 2: a bar chart showing 5% growth in 2010.
Decision: the chart exists and is mentioned, but the numerical claim is wrong. CMRC = 0. NRA = 0.5 because the chart is referenced but the explanation contradicts the visual evidence.

Reproducibility. The T-Judge's cross-modal submodule uses the same model snapshot as T-Judge and V-Judge: `gpt-4o`, temperature 0.2, top-$p = 1.0$, and a fixed random seed. Each report is evaluated three times, and the mean CMRC, NRA, and CMC are reported. The parsing failure rate is below 1%; failed parses are retried up to three times and then manually adjudicated.

A.5 Dual-Judge Evaluation: Implementation Details

The following provides an implementation-level completion of the **DUAL-JUDGE EVALUATION** section. The scoring protocol strictly follows the original text:

- **T-Judge**: $T = \mathrm{mean}(SC, GIV, NA, CMC)$, where $GIV = \mathrm{mean}(GIV_{\mathrm{num}}, GIV_{\mathrm{cov}})$ and $CMC = 0.5 \cdot CMRC + 0.5 \cdot NRA$.
- **V-Judge**: $V = \mathrm{mean}(DF, EX, AE, TC)$.
- **Overall**: $0.6T + 0.4V$. All dimensions and the overall score are normalized to $[0, 100]$.
- Unified judge settings: GPT-4o / GPT-4o-vision, temperature 0.2, top-$p = 1.0$, fixed random seed. Each report is evaluated three times, and the mean is reported. Parse failures are retried up to three times and then manually adjudicated.

The original paper only provides the full CMC prompt; the remaining prompts below are completed under the same JSON-output protocol and can be used directly as appendix implementations.

A.5.1 T-JUDGE DIMENSIONS

Structural Completeness (SC) Implementation.

- Heading matching: compute semantic embedding matching between generated headings and reference headings with threshold 0.85, and compute F1 as SC_{head}.
- Content coverage: T-Judge determines whether matched sections cover the same analytical sub-dimensions, yielding $SC_{\text{content}} \in [0, 100]$.
- Aggregation:
$$SC = 0.3 \cdot SC_{\text{head}} + 0.7 \cdot SC_{\text{content}}.$$

SC prompt snippet.

```
SC task:
Given reference chapter structure and generated report, compute:
- SC_head: heading F1 with semantic equivalence threshold 0.85.
- SC_content: 0-100, whether matched sections cover the same analytical sub-dimensio
Return SC_head, SC_content, SC = 0.3*SC_head + 0.7*SC_content.
```

Grounded Insight Verification (GIV)

(a) Numerical Accuracy, GIV_{num}. Implementation.

1. Extract numerical sentence clusters.
2. Generate verification questions for each numerical claim, with the claimed value as the target answer.
3. Have three specialized code-generation LLMs independently generate and execute SQL queries on the source tables.
4. A claim is verified if at least $2/3$ models agree within a 1% tolerance margin.
5. GIV_{num} is the proportion of verified claims multiplied by 100.

Numerical verification question generation prompt.

```
System:
You are a numerical verification question generator for table-grounded reports.
Return valid JSON only.

User:
[SOURCE TABLES]
{table_text_or_markdown}

[GENERATED REPORT]
{report_text}

[TASK]
1. Extract numerical sentence clusters.
2. For each claimed value, generate a verification question whose target answer is th
3. Generate an executable SQL query over the source tables.
4. Return JSON:
{
  "claims": [
    {
      "claim_id": "N1",
      "sentence": "...",
      "claimed_value": ...,
      "verification_question": "...",
      "sql": "..."
    }
```

```
  ]
}
```

(b) Keypoint Coverage, GIV_{cov}. Implementation.

- Use BERTScore to compute the similarity matrix S_{ij} between reference keypoints k_i and generated sentences s_j.
- Normalize globally to obtain the joint probability $P(k_i, s_j)$, and then obtain the marginals $P(k_i)$ and $P(s_j)$.
- Compute coverage using normalized mutual information:

$$GIV_{\text{cov}} = \frac{\sum_{i=1}^{M}\sum_{j=1}^{N} P(k_i, s_j)\log\frac{P(k_i,s_j)}{P(k_i)P(s_j)}}{-\sum_{i=1}^{M} P(k_i)\log P(k_i)}.$$

- This dimension requires no additional LLM prompt and is a reproducible semantic matching metric.

GIV aggregation.

$$GIV = \frac{GIV_{\text{num}} + GIV_{\text{cov}}}{2}.$$

Novelty Analysis (NA) Implementation.

- First identify novel claims: sentence clusters whose maximum cosine similarity to any reference keypoint is below $\tau_{\text{novel}} = 0.6$.
- T-Judge scores each novel claim on three sub-dimensions: non-triviality, actionability, and correctness.
- Let each claim receive a normalized score $c_n \in [0, 1]$, and let $|N|$ be the number of novel claims.
- Penalty factor:

$$\min\left(1, \frac{5}{|N|}\right).$$

- If $|N| = 0$, then $NA = 0$; otherwise:

$$NA = 100 \cdot \frac{1}{|N|}\sum_{n=1}^{|N|} c_n \cdot \min\left(1, \frac{5}{|N|}\right).$$

NA prompt snippet.

```
NA task:
A novel claim is a sentence cluster whose maximum cosine similarity to any reference
For each novel claim, score:
- non-triviality: 0-5
- actionability: 0-5
- correctness: factual consistency with source tables, 0-5
Normalize each claim to [0,1]. Apply penalty min(1, 5/|N|).
If |N|=0, NA=0.
Return novel_claims, per-claim score, NA.
```

Cross-Modal Consistency (CMC) Implementation.

$$CMC = 0.5 \cdot CMRC + 0.5 \cdot NRA.$$

- **CMRC**: For each sentence that explicitly or implicitly refers to a chart, check existence, numerical consistency, trend consistency, and encoding/label consistency. The score is $s(r_p) \in \{0, 0.5, 1\}$.

$$CMRC = 100 \cdot \frac{1}{|R|}\sum_{p=1}^{|R|} s(r_p).$$

If there is no chart-referencing sentence, $CMRC = 0$.

- **NRA**: For each chart, check whether it is explained by the text, whether the explanation matches the chart encoding, and whether it supports the analytical claim of the section. The score is $t(c_q) \in \{0, 0.5, 1\}$.

$$NRA = 100 \cdot \frac{1}{|C|} \sum_{q=1}^{|C|} t(c_q).$$

If there is no chart, $NRA = 0$.

CMC full prompt.

```
System:
You are a strict cross-modal consistency evaluator for multimodal business reports.
You receive:
(1) source tables,
(2) generated textual report,
(3) generated charts with captions,
(4) section headings,
(5) extracted chart data when available.
Your task is to compute CMRC and NRA.
Do not reward fluent writing.
Penalize any chart-text mismatch, unexplained chart, or text that refers to a non-ex
Return only valid JSON.

User:
[SOURCE TABLES]
{table_text_or_markdown}

[GENERATED REPORT]
{report_text}

[CHARTS]
Chart ID: C1
Caption: {caption_1}
Section: {section_1}
Image: {image_1}
Extracted data: {chart_data_1}

Chart ID: C2
Caption: {caption_2}
Section: {section_2}
Image: {image_2}
Extracted data: {chart_data_2}

[TASK]
1. Identify every sentence that explicitly or implicitly refers to a chart.
2. For each such sentence, find the target chart ID and score it as 0, 0.5, or 1 usi
3. For each chart, determine whether the text explains it and score it as 0, 0.5, or
4. Return JSON:
{
  "cmrc": [
    {
      "sentence_id": "S3",
      "sentence": "...",
      "chart_id": "C1",
      "score": 1,
      "reason": "..."
    }
```

```
  ],
  "nra": [
    {
      "chart_id": "C1",
      "score": 1,
      "reason": "..."
    }
  ],
  "cmrc_score": 100.0,
  "nra_score": 100.0,
  "cmc_score": 100.0
}
```

Decision rules and tolerances.

- Absolute values: 1% tolerance; proportions, percentages, or rates: 2 percentage points.
- Trend direction must match exactly; reversing order is a hard error.
- Labels, units, and chart type follow the original rules.
- If chart-to-table extraction confidence is below 0.8, use the source table and the original chart specification. If neither is available, mark the chart as uncertain and exclude it from the denominator.

A.5.2 V-JUDGE DIMENSIONS

Data Fidelity (DF) Implementation.

1. A chart-to-table model extracts data from the chart.
2. SQL generates expected data from the source tables.
3. Comparison: if more than 5% of data points deviate beyond a 1% absolute tolerance or a 2 percentage-point tolerance for proportions, the chart is flagged as a fidelity failure.
4. DF is aggregated as a chart-level data-fidelity score, normalized to $[0, 100]$.

Expressiveness (EX), Aesthetics (AE), and Technical Correctness (TC) Implementation.

- Each chart is scored on a 0–5 scale, then multiplied by 20 to convert to $[0, 100]$.
- EX: chart type appropriateness, encoding clarity, and insight salience.
- AE: color harmony, layout quality, and absence of visual noise, with calibration examples.
- TC: rule-based checks for missing elements, scale errors, zero-baseline checks, label readability, and color accessibility.

V-Judge full prompt.

```
System:
You are a strict V-Judge for generated visualizations in multimodal business reports
Evaluate only against source tables, chart specifications, and section claims.
Return valid JSON only.

User:
[SOURCE TABLES]
{table_text_or_markdown}

[CHARTS]
Chart ID: C1
Caption: {caption_1}
Section: {section_1}
Image: {image_1}
```

```
Extracted data: {chart_data_1}
Extraction confidence: {confidence_1}
Expected data from SQL: {expected_data_1}

[TASK]
For each chart:
- DF: compare extracted data with expected data. Mark fidelity failure if >5% data po
- EX: 0-5, chart type appropriateness, encoding clarity, insight salience.
- AE: 0-5, color harmony, layout quality, absence of visual noise.
- TC: 0-5, missing elements, scale errors including zero-baseline, label readability,
Return JSON:
{
  "charts": [
    {
      "chart_id": "C1",
      "DF": 100.0,
      "EX": 4,
      "AE": 4,
      "TC": 5,
      "fidelity_failure": false,
      "reason": "..."
    }
  ],
  "DF": 100.0,
  "EX": 80.0,
  "AE": 80.0,
  "TC": 100.0,
  "V_score": 90.0
}
```

V-Judge aggregation.

$$V = \frac{DF + EX + AE + TC}{4},$$

where EX, AE, and TC have already been mapped from 0–5 to 0–100.

A.5.3 Final Scoring

$$T = \frac{SC + GIV + NA + CMC}{4},$$

$$V = \frac{DF + EX + AE + TC}{4},$$

$$Overall = 0.6 \cdot T + 0.4 \cdot V.$$

All dimensions and the overall score are normalized to $[0, 100]$. This implementation is consistent with the original text: T-Judge handles text, factuality, novelty, and cross-modal consistency; V-Judge handles chart data fidelity, expressiveness, aesthetics, and technical correctness. The dual-judge separation avoids the 8–12 point inflation of a single judge and prevents cross-modal compensation.

A.6 Ablation: Sensitivity to Judge Weights.

We vary the T-Judge –V-Judge weight ratio from (0.5,0.5) to (0.8,0.2). The relative ordering of top models remains stable, with DeepSeek-R1 leading under all settings. The largest ranking shift occurs in the mid-tier: code-augmented models gain under higher V-Judge weight, while deep research agents, which often produce richer text but less polished charts, benefit from higher T-Judge weight. The default (0.6,0.4) balances these tendencies.

Table 3: Pearson correlation between automated MM-ReportBench overall scores and human expert judgments, broken down by dimension.

Dimension	**Pearson** r
Structural Completeness (SC)	0.88
Grounded Insight Verification (GIV)	0.94
Novelty Analysis (NA)	0.82
Data Fidelity (DF)	0.91
Expressiveness (EX)	0.85
Aesthetics (AE)	0.79
Technical Correctness (TC)	0.90
Overall Score	0.91

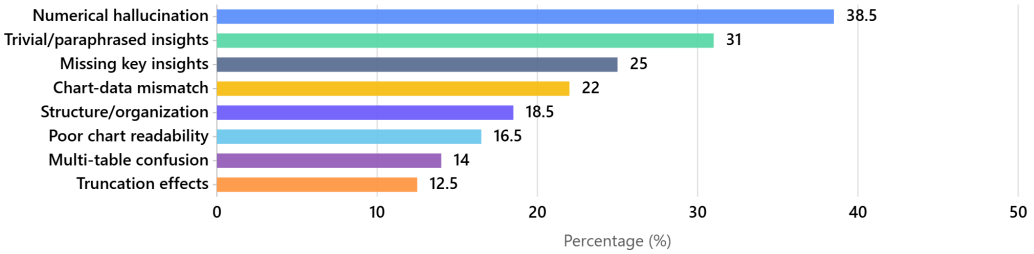


Figure 4: The frequency of common error categories across 200 sampled generated reports. Numerical hallucination is the most pervasive issue (38.5%), followed by trivial/paraphrased insights (31.0%) and missing key insights (25.0%). The distribution pinpoints factual accuracy, analytical depth, and multi-table reasoning as the primary weaknesses of current models.

A.7 Human Evaluation and Consistency

A.8 Error Analysis

We manually analyze 200 randomly sampled reports to categorize failure modes. The error taxonomy is extended for multimodal settings. Figure 4 shows the distribution of primary error types in generated report.

A.9 Case Study

Using the U.S. employment trends report as an illustrative case (Figure 5), our dual-judge evaluation reveals a mixed performance profile: the report achieves strong structural completeness (SC 92) through its well-organized chapter hierarchy, but its grounded insight verification is only moderate (GIV 65) because many statements remain qualitative without explicit numerical anchoring, and novelty is low (NA 30) as the insights largely paraphrase visible trends rather than offering non-trivial causal analysis or actionable recommendations. The accompanying charts are functionally correct but suffer from missing annotations and cluttered legends, yielding a VisScore of 79. The overall MM-ReportBench score of 69.3, while higher than the best model-generated reports (which typically score in the low 60s), still falls well short of the human baseline, demonstrating that even a well-structured and factually faithful report can be penalized for insufficient quantitative grounding and a lack of deep analytical insight—precisely the weaknesses that our benchmark is designed to expose.

Table 4: Distribution of primary error types in generated reports. A single report may contain multiple errors.

Error Type	**Frequency (%)**	**Metric(s)**
Numerical hallucination	38.5%	GIV_{num}, DF
Missing key insights	25.0%	GIV_{cov}
Structure/organization issues	18.5%	SC
Chart-data mismatch	22.0%	DF
Poor chart readability	16.5%	EX, AE, TC
Trivial/paraphrased insights	31.0%	NA
Multi-table confusion	14.0%	GIV, DF
Truncation effects (large tables)	12.5%	GIV_{num}

1. Monthly Trend Analysis of U.S. Nonfarm Employment Changes (2006-2015)
1.1 Analysis of Monthly Nonfarm Employment Changes
2. Seasonal Patterns in U.S. Nonfarm Employment Changes
2.1 Seasonal Patterns in Nonfarm Employment

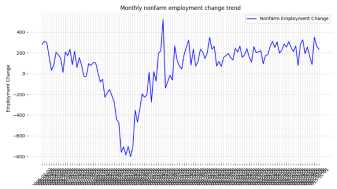
What were the monthly changes in U.S. nonfarm employment from January 2006 to December 2015?

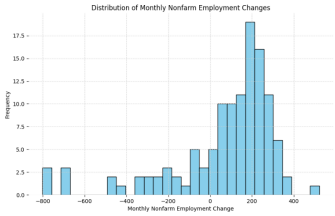
Abstract

The report on U.S. nonfarm employment changes from January 2006 to December 2015 identifies key trends and seasonal patterns. The analysis reveals significant monthly fluctuations, notably highlighting the downturn during the 2008 financial crisis followed by a gradual recovery. This is illustrated through line charts and histograms, which emphasize modest employment gains and fewer extreme changes over the period. Additionally, the report uncovers seasonal trends, with employment typically rising in spring and fall, and summer showing the least change. These seasonal patterns, depicted through line and bar charts, have implications for economic planning and future analyses of employment dynamics. Overall, the report provides valuable insights into employment changes, aiding in understanding economic conditions and informing policy decisions.

1 Monthly Trend Analysis of U.S. Nonfarm Employment Changes (2006-2015)

Analysis of Monthly Nonfarm Employment Changes





Key Points

- The group of charts examines the monthly changes in U.S. nonfarm employment from January 2006 to December 2015, providing insights into how employment levels fluctuated over time.
- The line chart highlights the overall trend, showing significant declines during the 2008 financial crisis and a gradual recovery thereafter, while the histogram reveals that most changes centered around modest gains, with fewer months experiencing extreme losses or gains.
- The line chart and histogram complement each other by illustrating both temporal trends and the frequency distribution of employment changes, offering a deeper understanding of the data dynamics over the decade.

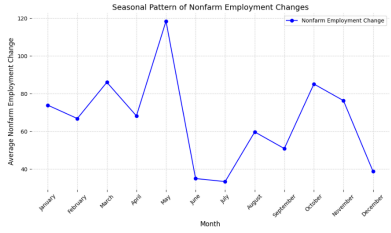
Summary

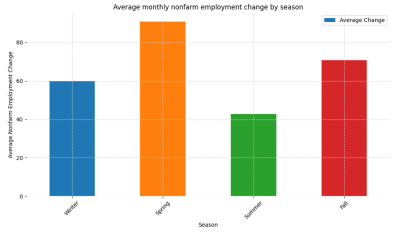
The chapter "Monthly Trend Analysis of U.S. Nonfarm Employment Changes (2006-2015)" explores employment fluctuations, highlighting significant declines during the 2008 financial crisis and subsequent recovery. Line charts and histograms reveal trends and distributions, emphasizing modest gains and fewer extreme changes, offering insights into employment dynamics, informing future economic analyses.

Building on the insights gained from the monthly trend analysis of U.S. nonfarm employment changes, we now shift our focus to uncovering seasonal patterns. Understanding these recurring fluctuations provides a deeper layer of analysis, revealing how different times of the year uniquely impact employment dynamics.

2 Seasonal Patterns in U.S. Nonfarm Employment Changes

Seasonal Patterns in Nonfarm Employment





Key Points

- This group of charts explores the seasonal patterns in U.S. nonfarm employment changes from January 2006 to December 2015, focusing on how employment fluctuates across different months and seasons.
- The key insight revealed is that nonfarm employment changes exhibit significant seasonal trends, with notable increases in the spring and fall, while summer shows the lowest average change.
- The line chart provides a detailed month-by-month analysis of employment changes, highlighting peak and trough months, while the bar chart summarizes these changes by season, allowing for a clearer comparison of seasonal impacts.

Summary

The chapter "Seasonal Patterns in U.S. Nonfarm Employment Changes" examines employment fluctuations from January 2006 to December 2015, revealing significant seasonal trends. Employment typically increases in spring and fall, with summer showing the lowest changes. Line and bar charts highlight monthly and seasonal impacts, respectively, underscoring these patterns' implications for economic planning.