OMHBench: Benchmarking Balanced and Grounded Omni-Modal Multi-Hop Reasoning

Seunghee Kim[1], **Ingyu Bang**[1], **Seokgyu Jang**[1], **Changhyeon Kim**[1], **Sanghwan Bae**[2], **Jihun Choi**[3*], **Richeng Xuan**[4], **Taeuk Kim**[1†]
[1]Hanyang University, [2]NAVER Cloud, [3]Knowledge Work Inc.,
[4]Beijing Academy of Artificial Intelligence
{gyg9325, ingyu1008, diamondgyu, livex, kimtaeuk}@hanyang.ac.kr,
baaesh10@gmail.com, jihun.choi@knowledgework.com, rcxuan@baai.ac.cn

Abstract

Multimodal Large Language Models (MLLMs) have increasingly supported omni-modal processing across text, vision, and speech. However, existing evaluation frameworks for such models suffer from critical limitations, including modality shortcuts and biased reasoning paths. To address these challenges, we propose **OMHBench**, a novel benchmark designed to rigorously evaluate omni-modal multi-hop reasoning. It consists of 6,144 questions with balanced reasoning paths that are jointly grounded across all three modalities. Extensive evaluation of 13 state-of-the-art models reveals that (1) a large performance gap exists between proprietary and open-source MLLMs and (2) even proprietary models exhibit high sensitivity to reasoning path variations, resulting in **asymmetric omni-modal grounding**. Notably, models struggle when processing the speech modality, underscoring the need for balanced, multi-hop evaluation of omni-modal intelligence.

1 Introduction

Human perception and understanding are inherently complex, often requiring the integration of textual, visual, and auditory information. Accordingly, the ability to process such heterogeneous inputs in tandem is fundamental to achieving human-level AI (Baltrušaitis et al., 2019). Relatedly, recent Multimodal Large Language Models (MLLMs) have evolved from initial bi-modal variants (e.g., text–vision and text–audio) to more comprehensive ones that jointly process text, vision, and audio, often referred to as **omni-modal** (Microsoft et al., 2025; Xu et al., 2025b; Gemini Team et al., 2025).[1]

The development of these models has also driven the emergence of new evaluation schemes, which fall into two main directions: **Omni-Modal Understanding (OMU)** (Li et al., 2025b; Hong et al., 2025; Zhou et al., 2025; Chen et al., 2025; Nguyen et al., 2025), which emphasizes measuring a model's ability to collectively handle text, vision, and audio; and **Cross-Modal Multi-Hop Reasoning (CMR)** (Talmor et al., 2021; Reddy et al., 2022; Kim et al., 2025; Foroutan et al., 2025; Jang et al., 2025), which focuses on its capability to perform multi-hop reasoning by composing information across modalities, typically in bi-modal settings. The key distinction between the two lies in whether the speech modality is incorporated and whether multi-hop reasoning is explicitly required.

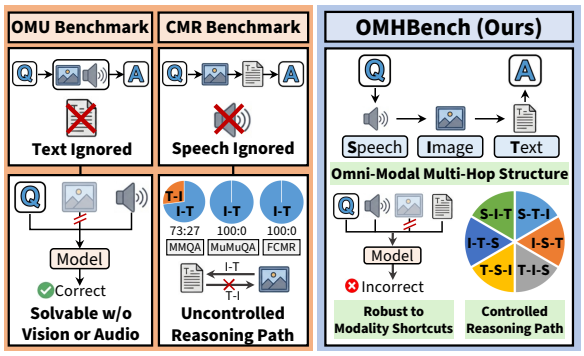


Figure 1: Omni-Modal Understanding (OMU) benchmarks lack textual context and suffer from modality shortcuts, while Cross-Modal Multi-Hop Reasoning (CMR) datasets exclude speech and exhibit imbalanced reasoning paths. OMHBench addresses these issues.

In this work, we pose two crucial research questions regarding the current evaluation paradigms (see Figure 1): (1) If an OMU benchmark can be solved without leveraging all three modalities, can it truly be said to evaluate omni-modal understanding? (2) If a CMR benchmark is dominated by a single reasoning path, resulting in a heavily skewed composition distribution, can its results reliably reflect a model's reasoning ability? We demonstrate through experiments that both suspected pitfalls

*This work was conducted at Sony AI.
†Corresponding author
[1]**Omni-modal** refers to the text-vision-audio setting, while **multi-modal** is a general term for more than one modality.

Findings of the Association for Computational Linguistics: ACL 2026, pages 18311–18334

Benchmark	Text (Q)	Text (C)	Vision	Speech	CMR	Path Balance
OmniBench	✓	✗	✓	✓	✗	–
WorldSense	✓	✗	✓	✓	✗	–
Daily-Omni	✓	✗	✓	✓	✗	–
OmniVideoBench	✓	✗	✓	✓	✗	–
UNO-Bench	✓	✗	✓	✓	✗	–
AV-SpeakerBench	✓	✗	✓	✓	✗	–
MMQA	✓	✓	✓	✗	✓	✗
MuMuQA	✓	✓	✓	✗	✓	✗
FCMR	✓	✓	✓	✗	✓	✗
ICT-QA	✓	✓	✓	✗	✓	✗
WikiMixQA	✓	✓	✓	✗	✓	✗
OMHBench(Ours)	✓	✓	✓	✓	✓	✓

Table 1: Comparison of OMU and CMR benchmarks by modality coverage and reasoning path balance. Text (Q) indicates question/option text, whereas Text (C) denotes separate contextual text. OMHBench uniquely supports all modalities with balanced multi-hop reasoning paths.

are present in practice and substantially undermine the integrity of current omni-modal evaluations.

To systematically address these limitations, we propose **O**(mnimodal)**M**(ulti)**H**(op)**Bench**.[2] By design, this novel benchmark departs from prior OMU and CMR datasets, as compared in Table 1. It requires omni-modal multi-hop reasoning over text, image, and speech, with each modality explicitly used at least once, eliminating shortcuts that allow models to solve tasks without access to a specific modality. Moreover, the ground-truth reasoning paths used as solutions are controlled with respect to modality order, enabling clearer identification of MLLMs' strengths and weaknesses.

Using OMHBench, we extensively test 13 proprietary and open-source MLLMs, uncovering several underexplored properties of omni-modal multi-hop reasoning. We find that (1) entity-attribute-based multi-hop structure effectively mitigates modality shortcut issues; (2) model performance rankings can vary considerably depending on the category of reasoning paths; (3) models exhibit strong asymmetry in omni-modal grounding, especially when transferring semantics into the speech modality.

In sum, this work (1) exposes fundamental limitations of existing OMU and CMR benchmarks; (2) presents OMHBench, an omni-modal benchmark with controlled and balanced multi-hop reasoning paths; and (3) reveals systematic weaknesses in omni-modal grounding, particularly in speech.

[2]OMHBench is available at https://huggingface.co/datasets/HYU-NLP/OMHBench.

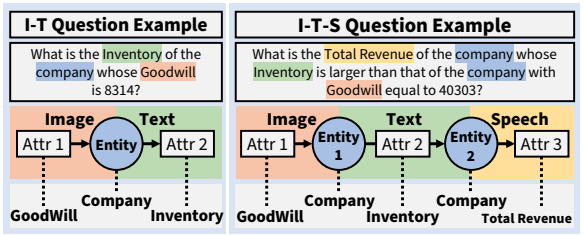


Figure 2: Illustration of the proposed task formulation with two example questions and their reasoning paths: I-T (left) and I-T-S (right). Attributes (e.g., Goodwill) are accessible only through specific modalities, while entities (e.g., Company) are shared across modalities.

2 Related Work

Multimodal Large Language Models (MLLMs) MLLMs aim to extend the capabilities of LLMs beyond text by incorporating additional modalities such as vision and audio. Early efforts primarily focused on bi-modal settings, most notably text–vision (Alayrac et al., 2022; Liu et al., 2023a) or text–audio (Kong et al., 2024; Ghosh et al., 2024) integration, by attaching modality-specific encoders to off-the-shelf LLMs. Consequently, these models remain limited in their ability to jointly reason over more than two modalities, motivating recent efforts toward omni-modal architectures (Xu et al., 2025a; Yao et al., 2024; Microsoft et al., 2025; Ye et al., 2025) that natively support text, vision, and audio within a unified framework.

Omni-Modal Understanding (OMU) With the advent of omni-modal models, a few benchmarks have been proposed to test their capabilities. OmniBench (Li et al., 2025b) is an initial attempt to evaluate models under tri-modal inputs. WorldSense (Hong et al., 2025), Daily-Omni (Zhou et al., 2025), OmniVideoBench (Li et al., 2025a), and AV-SpeakerBench (Nguyen et al., 2025) focus on vision-audio understanding. UNO-Bench (Chen et al., 2025) further explores the relationship between uni-modal and omni-modal performance.

However, existing datasets emphasize visual and auditory signals, relegating text to questions or options, and often permit shortcuts that enable strong performance even without using all modalities.

Cross-Modal Multi-Hop Reasoning (CMR) CMR evaluates a model's ability to perform multi-hop reasoning by interleaving textual and visual evidence. Early benchmarks such as MMQA (Talmor et al., 2021) and MuMuQA (Reddy et al., 2022) require models to integrate information from text,

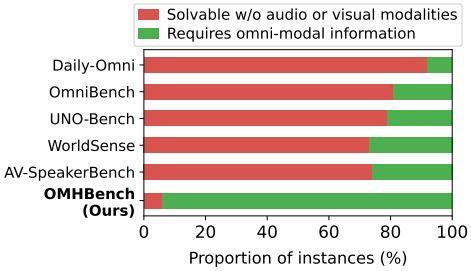


Figure 3: Fraction of instances in the OMU benchmarks that remain solvable *without* visual or auditory input. Across all datasets, nearly 70~80% of instances are susceptible, indicating that shortcuts allowing models to answer without using certain modalities are prevalent.

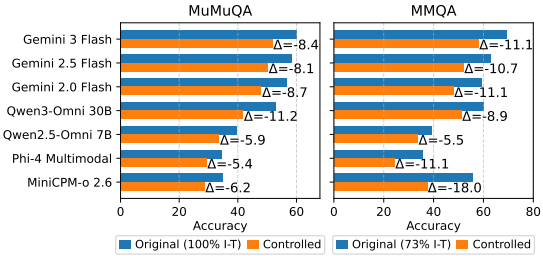


Figure 4: Performance comparison on the *original* MuMuQA and MMQA datasets vs. their *controlled* variants. When revised to ensure balanced reasoning paths, accuracy drops by up to 18%, raising concerns about the validity of previous evaluations on the original datasets.

tables, and images, while FCMR (Kim et al., 2025) extends this paradigm to the financial domain. ICT-QA (Jang et al., 2025) and WikiMixQA (Foroutan et al., 2025) also explore multi-hop reasoning over structured sources, e.g., tables and charts.

Nevertheless, these benchmarks primarily focus on text-vision modalities and do not support audio-based reasoning. Moreover, the absence of explicit control over reasoning paths in these datasets often leads to heavily skewed reasoning pattern distributions, compromising the reliability of evaluation.

3 Preliminaries

3.1 Task Formulation: Omni-Modal Multi-Hop Reasoning

Here, we specify the scope and formal definition of **omni-modal multi-hop reasoning** considered in this work. Multi-hop reasoning inherently operates over *entities* and their *attributes* as articulated in the context. From an omni-modal—more specifically, cross-modal—perspective, we consider scenarios in which entities are shared across modalities while attributes remain modality-specific, as illustrated in Figure 2. Answering such questions requires consulting modalities in a particular order, determined by the availability of the referenced attributes. We refer to this modality order as a **reasoning path**.

For instance, the question *"What is the inventory of the company whose goodwill is 8314?"* in Figure 2 necessitates integrating two attributes of the same entity, where *goodwill* is available in the image modality and *inventory* in text. In this case, the reasoning path is I-T. This formulation naturally generalizes to longer reasoning chains (e.g., I-T-S) through additional hops that identify subsequent entities. A reasoning problem is considered *omni-modal multi-hop* when the reasoning chain involves all three modalities: image, text, and speech.

3.2 Shortcuts in OMU Benchmarks

To verify whether OMU benchmarks demand the exhaustive use of omni-modal input, we investigate five cases—OmniBench, WorldSense, Daily-Omni, UNO-Bench, and AV-SpeakerBench—by measuring the proportion of instances that remain solvable without visual or auditory input. This experiment is conducted using Gemini 3 Flash (Google, 2025b).

Figure 3 reports that nearly 70–80% of instances in OMU benchmarks can be answered without access to certain modalities. In other words, existing OMU benchmarks fail to genuinely assess the utilization of all three modalities due to insufficient structural constraints on cross-modal dependence.

3.3 Biases in CMR Benchmarks

CMR datasets, e.g., MuMuQA and MMQA, are characterized by skewed distributions in their representation of reasoning paths. As they consider only visual and textual inputs, there exist two possible orders: I-T and T-I. Nonetheless, these benchmarks are imbalanced: MuMuQA and FCMR contain only I-T instances, while MMQA is skewed toward I-T, with roughly twice as many I-T as T-I.[3]

To examine the effect of these biases, we conduct a controlled experiment by reversing the reasoning path direction (e.g., from I-T to T-I) for each question without changing its semantics. This yields variants with uniform reasoning path distributions while preserving the originals' unique properties. Specifically, we create balanced versions of MuMuQA, MMQA and FCMR, each with an equal split (50:50) between I-T and T-I instances.[4]

[3] ICT-QA (Jang et al., 2025) and WikiMixQA (Foroutan et al., 2025) are not publicly available, but their dataset construction does not consider reasoning path distributions.

[4] The detailed process is described in Appendix A.

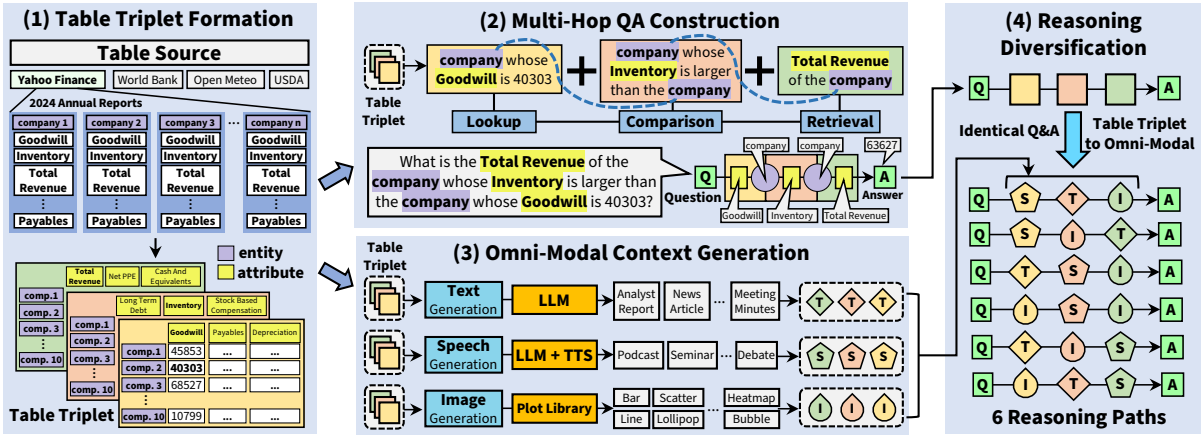


Figure 5: Overview of the OMHBench pipeline. (1) **Table Triplet Formation** constructs table triplets that share the same entities yet having separate attributes. (2) **Multi-Hop QA Construction** yields multi-hop QA pairs by utilizing content from table triplets. (3) **Omni-Modal Context Generation** converts each table into text, image, and speech modalities. (4) **Reasoning Diversification** guarantees multiple reasoning paths via modality permutation.

Figure 4 shows that performance exhibits a pronounced and consistent decline under controlled settings, in some cases exceeding a 10% drop in accuracy (see Figure 13 in the Appendix for FCMR). This indicates that balanced reasoning path distributions are essential for accurate MLLM evaluation; however, simple path reversal offers only a partial remedy confined to bi-modal settings, highlighting the need for a new, systematic benchmark.

4 Proposed Benchmark: OMHBench

Motivated by the largely isolated research streams on OMU and CMR and their respective limitations, we propose **OMHBench**, a new benchmark that bridges the two paradigms and addresses the aforementioned issues. It enables controlled evaluation of omni-modal multi-hop reasoning with three key desiderata: (1) it prevents shortcut-prone evaluation by enforcing multi-hop reasoning, (2) it jointly incorporates textual, visual, and speech modalities, and (3) it explicitly controls reasoning paths to support reliable and unbiased assessment.

As shown in Figure 5, OMHBench is constructed through four stages: (1) Table Triplet Formation, (2) Multi-Hop QA Construction, (3) Omni-Modal Context Generation, and (4) Reasoning Diversification. We refer readers to Appendix B for details.

4.1 Table Triplet Formation

OMHBench covers four domains—finance, economics, climate, and nutrition—where reasoning over text, images, and speech naturally occurs, with instances evenly distributed across domains. We employ real-world table data from Yahoo Finance, World Bank, Open-Meteo, and USDA.

Given an original tabular source, we construct table triplets of three smaller tables that share the same set of *entities* but contain distinct *attributes*. Each table has a size of 10 entities × 3 attributes and includes both relevant and distractor entities and attributes, requiring models to retrieve correct clues under information overload.

The intuition is that practical information is often organized in tabular form, yet its content can be realized in different modalities—images for visual comparison, text for detailed description, and speech for public announcements. The table triplets serve as the core intermediate representation, which are used to construct question-answer pairs (§4.2) and corresponding omni-modal contexts (§4.3).

4.2 Multi-Hop QA Construction

From each table triplet, we formulate a question that requires three-hop reasoning to answer. In detail, we define eight reasoning operations—Lookup, Ranking, Comparison, Range, Proximity, Retrieval, Mean, and Summation—sample three of them, and sequentially apply each to the tables to derive questions.[5] The first two reasoning steps focus on entity-level reasoning, applying operations like Lookup, Comparison, and Range to select, compare, or filter entities and pass either a single entity or several entities forward. The final step produces the answer

[5] In practice, not all combinations of reasoning operations are valid; their applicability depends on the entities and attributes involved. We therefore discard infeasible combinations and retain only those that apply to each case.

by either retrieving an attribute of a selected entity or aggregating attributes over a filtered entity set using operations such as Mean or Summation.

Note that the proposed procedure is deterministic and rule-based, enabling scalable question generation given a sufficient number of tabular data sources, without relying on costly external tools such as generative AI. Consequently, the QA construction process is efficient and fully automated.

Finally, we partition the QA pairs into two categories based on their underlying reasoning operation structures. **OMHBench-Connect** includes cases where intermediate results remain single entities throughout the reasoning process, following a fixed sequence of *Lookup–Comparison–Retrieval* operations. By contrast, **OMHBench-Reasoning** covers cases where intermediate results expand into sets of entities, requiring aggregation operations.

4.3 Omni-Modal Context Generation

At this stage, the three tabular sources from §4.1 are transformed into contextual representations across three modalities—text, image, and speech. We explain this part using the financial domain as an example; the same process can be applied to others.

For the text modality, we define scenarios such as analyst reports, news articles, and meeting minutes, following prior work in financial text mining (Kumar and Ravi, 2016; Pejić Bach et al., 2019; Gupta et al., 2020). Task-specific prompts are then used to guide LLMs in generating natural language descriptions grounded in the underlying tables. For the image modality, we generate visualizations using plotting libraries, following common practices in financial data visualization (Ko et al., 2016; Uddin et al., 2024; Christensen et al., 2024). For the speech modality, we adopt the taxonomy of financial speech scenarios proposed by Cao et al. (2025). The generated scripts are synthesized into speech using Kokoro-82M TTS (Hexgrad, 2025). These are constructed as a multi-speaker dialogue in which each speaker describes different attributes of the same entity, encouraging models to leverage both semantic content and acoustic cues. To enhance linguistic and stylistic diversity, we use three LLMs—GPT-5.1 (OpenAI, 2025a), Grok-4 (xAI, 2025), and Claude-Sonnet-4.5 (Anthropic, 2025b)—for text and speech script generation.

4.4 Reasoning Diversification

We then pair a multi-hop QA instance (§4.2) with alternative configurations of omni-modal contexts (§4.3) to create multiple QA variants. They preserve the same question and answer, but differ in how contextual evidence is organized. By permuting modality assignments over the three tables, we obtain $3! = 6$ possible reasoning paths. Note that across these variants, the informational content remains unchanged; only the modality sequence required for inference (i.e., the reasoning path) varies.

4.5 Quality Control

Lastly, we apply four quality control methodologies to ensure dataset reliability and fair evaluation.

Entity Anonymization Given prior findings that CMR datasets may allow shortcuts via parametric knowledge (Kim et al., 2025), we anonymize entity names with alphabetical codes (e.g., B, X), forcing models to rely solely on the provided context.

Consistency Checking We perform QA-based consistency checks (Fabbri et al., 2021) by deriving factoid questions from the original tables (§4.1) and validating answers using the converted context (§4.3). We also apply a test where an LLM reconstructs the original tables from the converted modalities (§4.3) and compares them with the originals. Both checks achieve 100% consistency, confirming the absence of factual loss or distortion.

Question Rephrasing To enhance linguistic diversity, we paraphrase questions using multiple LLMs: GPT-5.1, Grok-4, and Claude-Sonnet-4.5. Paraphrasing quality is evaluated with the Lexical Deviation (LD) metric (Liu et al., 2022), where our dataset achieves higher LD scores than the widely used PAWS dataset (Zhang et al., 2019) (0.32 vs. 0.13), indicating greater lexical diversity.

TTS Validation We evaluate TTS fidelity using ASR-based error rates (WER and CER) and speech quality metrics (STOI and SI-SDR), following Kumar et al. (2023). The results demonstrate high transcription accuracy and audio quality (WER: 0.03, CER: 0.02, STOI: 99.2, SI-SDR: 21.0).

Final Dataset Statistics OMHBench comprises 6,144 instances evenly distributed across six reasoning paths. The benchmark is divided into two subsets—OMHBench-Connect and OMHBench-Reasoning—each with 3,072 instances, based on the required reasoning operations. Comprehensive dataset statistics, including diversity control across the three modalities, are reported in Table 4 of the Appendix.

Model	Accuracy by Reasoning Path (%)						Avg. (Acc.)	PBS
	S-I-T	S-T-I	I-S-T	T-S-I	I-T-S	T-I-S		
Proprietary Models								
Gemini 3 Flash	97.5	98.4	75.4	75.0	60.2	63.5	78.3	32.2
Gemini 2.5 Pro	94.5	96.9	66.4	71.1	55.5	50.8	72.5	25.0
Gemini 2.5 Flash	82.0	85.9	50.8	54.7	26.6	21.9	53.6	4.7
Gemini 2.5 Flash-lite	49.2	60.9	38.3	35.2	5.5	4.7	32.3	0.0
Gemini 2.0 Flash	28.9	33.6	26.6	29.7	4.7	6.2	21.6	0.0
Gemini 2.0 Flash-lite	35.9	32.8	21.1	11.7	2.3	2.3	17.7	0.0
Open-Source Models								
Qwen3-Omni 30B	75.8	77.0	46.7	49.6	16.0	16.0	46.8	2.3
Phi-4 Multimodal	26.6	23.6	21.5	18.4	0.6	0.0	15.1	0.0
Qwen2.5-Omni 7B	22.7	20.9	19.3	20.5	2.0	1.8	14.5	0.0
Qwen2.5-Omni 3B	12.7	17.6	15.6	14.6	1.2	2.0	10.6	0.0
OmniVinci	14.8	8.6	14.8	7.0	0.8	0.6	7.8	0.0
MiniCPM-o 2.6	8.0	10.9	7.4	8.4	1.2	0.2	6.0	0.0
Omni-AutoThink	7.6	6.6	8.0	6.1	0.6	0.0	4.8	0.0

Table 2: Accuracies and Path Balance Scores (PBSs) across six reasoning paths in **OMHBench-Connect**. Avg. denotes macro-averaged accuracy. PBSs (§5.2) measure robustness to reasoning path variations.

Model	Accuracy by Reasoning Path (%)						Avg. (Acc.)	PBS
	S-I-T	S-T-I	I-S-T	T-S-I	I-T-S	T-I-S		
Proprietary Models								
Gemini 3 Flash	55.9	58.8	49.8	49.6	40.0	42.6	49.4	8.6
Gemini 2.5 Pro	53.9	51.6	52.3	47.7	41.4	46.1	48.8	10.9
Gemini 2.5 Flash	32.0	30.5	17.2	24.2	10.9	10.9	21.0	0.0
Gemini 2.5 Flash-lite	18.8	21.1	15.6	8.6	0.0	0.0	10.7	0.0
Gemini 2.0 Flash	4.7	11.7	4.7	6.2	0.8	0.0	4.7	0.0
Gemini 2.0 Flash-lite	3.9	5.5	3.9	2.3	0.8	0.0	2.7	0.0
Open-Source Models								
Qwen3-Omni 30B	27.3	28.5	14.1	14.6	2.7	2.7	15.0	0.0
Phi-4 Multimodal	0.6	0.4	0.2	0.0	0.2	0.2	0.3	0.0
Qwen2.5-Omni 7B	0.4	1.0	1.0	0.6	0.2	1.2	0.7	0.0
Qwen2.5-Omni 3B	0.8	0.6	0.2	0.0	0.4	0.2	0.4	0.0
OmniVinci	0.6	0.2	0.2	0.4	0.0	0.0	0.2	0.0
MiniCPM-o 2.6	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Omni-AutoThink	0.4	0.2	0.4	0.2	0.0	0.0	0.2	0.0

Table 3: Accuracies and Path Balance Scores (PBSs) across six reasoning paths in **OMHBench-Reasoning**. Avg. and PBSs are defined as in Table 2.

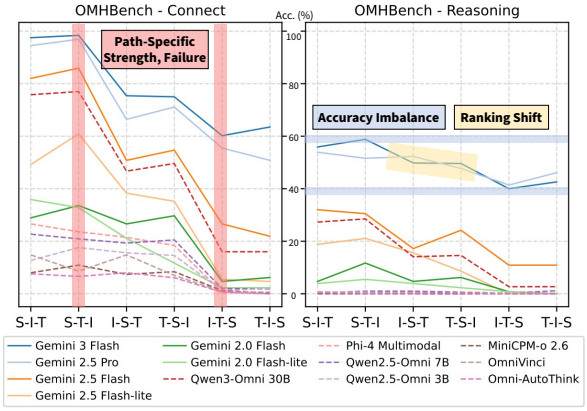


Figure 6: Visualization of core trends from Table 2 and Table 3, highlighting path-specific strengths and failures, accuracy gaps, and ranking changes by reasoning paths.

5 Experiments

5.1 Experimental Setup

We evaluate 13 MLLMs in total: both proprietary models—Gemini series (Google, 2025a; Comanici et al., 2025; Google, 2025b)—and open-source ones—Qwen3-Omni 30B (Xu et al., 2025c), Phi-4-Multimodal (Microsoft et al., 2025), Qwen2.5-Omni (Xu et al., 2025a), OmniVinci (Ye et al., 2025), MiniCPM-o 2.6 (Yao et al., 2024), and Omni-AutoThink (Yang et al., 2025b).[6] For models that support explicit reasoning modes, we enable this capability by setting a thinking budget of 8,192 tokens. All models are prompted using zero-shot chain-of-thought with a brief instruction and no fixed reasoning format. Model outputs are parsed into discrete answers and scored as correct or incorrect. We report *exact match* accuracies across all six reasoning paths, along with the macro-average. As Tan et al. (2024) shows that input modality order can affect model behavior—which we also observe in §6.6—we randomize the arrangement of omni-modal contexts to prevent such biases.

5.2 Path Balance Score (PBS)

Beyond accuracy, we propose the **Path Balance Score (PBS)** as a novel metric to measure model robustness to variations in reasoning paths. In OMHBench, each question is instantiated across all permutations of the available modalities; with N modalities, this yields $N!$ reasoning paths sharing the same question. PBS evaluates whether a model can consistently answer all such paths.

Formally, let the dataset contain $|D|$ instances, forming $|G| = |D|/N!$ groups. For the i-th group, let $a_{i,j} \in \{0, 1\}$ denote whether the model correctly answers the j-th path. PBS is defined as:

$$\text{PBS} = \frac{1}{|G|} \sum_{i=1}^{|G|} \mathbb{I}\left(\sum_{j=1}^{N!} a_{i,j} = N!\right),$$

where $\mathbb{I}(\cdot)$ is the indicator function. Intuitively, PBS counts a group as correct only if the model can answer the question under *all* different reasoning paths, reflecting its robustness to path variations.

5.3 Main Results

Table 2 and Table 3 report the performance of LLMs on OMHBench-Connect and -Reasoning, respectively.[7] Figure 6 provides a visual summary

[6] As of 2026-01-01, the Gemini series is the only proprietary model family that supports native omni-modal reasoning.

[7] For models allowing reasoning mode, only thinking variants are reported; full results are shown in Tables 8 and 9.

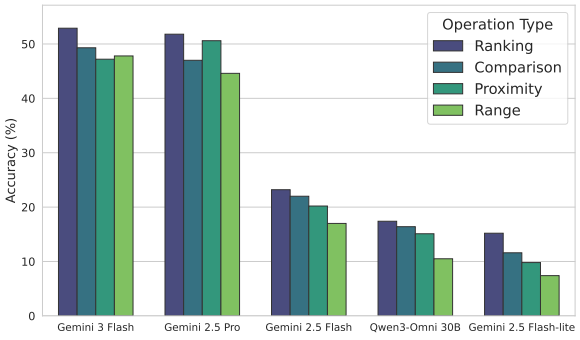


Figure 7: Accuracy by operation type on OMHBench-Reasoning, computed over instances whose reasoning chains include the corresponding operation, showing decreasing performance from Ranking to Range.

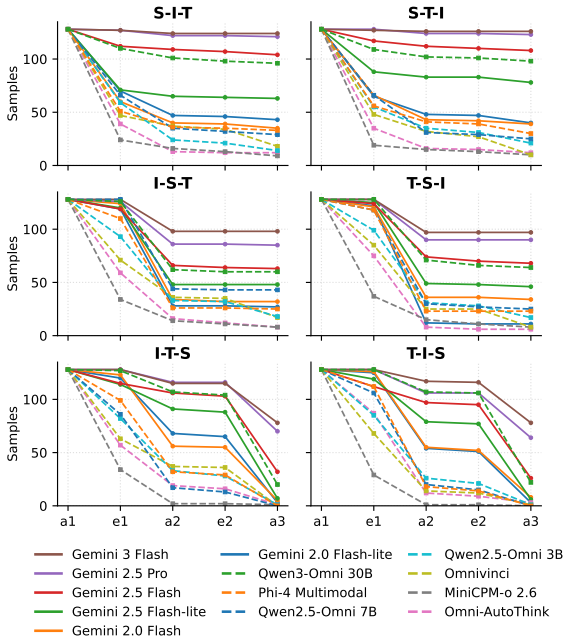


Figure 8: Step-by-step failure analysis for six reasoning paths. Each subplot reports the number of samples (out of 128 per reasoning path, 768 in total) successfully completing each reasoning stage—a_1, e_1, a_2, e_2, and a_3—with success depending on all prior steps being correct. This view reveals key bottlenecks in OMHBench.

of these results and highlights several key trends.

Overall Trends For both OMHBench-Connect and OMHBench-Reasoning, proprietary models consistently outperform open-source models, with Qwen3-Omni 30B standing out as the strongest open-source model. Importantly, we observe substantial performance variations across reasoning paths, with Qwen3-Omni 30B achieving 77% on S-T-I versus 16% on I-T-S in OMHBench-Connect. OMHBench-Reasoning is more challenging than OMHBench-Connect, due to multi-entity intermediate states and numerical operations: even the best model reaches only 49.4% accuracy, while most open-source models perform near zero.

Difficulty by Reasoning Path We discover that MLLMs may answer the same question correctly or incorrectly depending on the composition of the omni-modal context. This behavior is clearly reflected by the low PBS scores observed for most models, indicating limited robustness to reasoning-path variations. In particular, *accuracy is strongly influenced by the position of the speech modality*: paths where speech appears earlier generally achieve higher performance, whereas those where speech appears later are substantially more challenging. We further analyze this effect in §6.2.

In addition, Figure 6 shows that model rankings can shift across different paths. In OMHBench-Reasoning, Gemini 2.5 Pro outperforms Gemini 3 Flash on I-S-T, but underperforms on T-S-I. This suggests that single-path evaluation (e.g., 100% I-T in MuMuQA) fails to characterize model behavior, underscoring the limitation of existing benchmarks.

Difficulty by Reasoning Operation Instances in OMHBench-Reasoning are crafted to require diverse operations (e.g., Range, Ranking) for their solution. To examine difficulty by operation type, we group instances accordingly and average accuracies over the top five models. Figure 7 highlights that performance degrades from Ranking to Comparison, Proximity, and Range, implying that MLLMs handle ordinal or pairwise comparisons well, while operations involving numerical neighborhoods or interval constraints remain challenging.

6 Analysis

6.1 Modality Shortcut Validation

We verify that the modality shortcut issue observed in prior OMU benchmarks is no longer present in OMHBench. Following the protocol in §3.2, we measure the proportion of instances that remain solvable when one modality is removed. Figure 3 confirms that OMHBench exhibits almost no shortcut-prone cases, reflecting the effectiveness of its explicit multi-hop design. The few remaining solvable cases mainly stem from lookup-type questions, where correct answers can sometimes be obtained by chance through keyword retrieval.

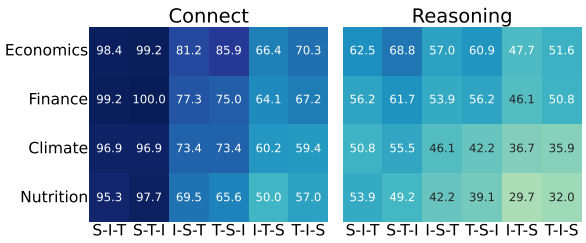


Figure 9: Domain-wise accuracies of Gemini 3 Flash on OMHBench-Connect and -Reasoning, with rows denoting domains and columns denoting reasoning paths. Performance gaps are amplified for challenging paths.

6.2 Step-by-Step Failure Analysis

We conduct a step-by-step failure analysis to identify key bottlenecks in OMHBench, with results depicted in Figure 8.[8] Leveraging the task's step-wise structure, we categorize failures by the stage at which the model fails to identify the required entity (e) or attribute (a), using Gemini 3 Flash.

Weaker models frequently fail at early reasoning stages, especially in identifying e_1 or a_2, regardless of the reasoning path, reflecting shortcomings in single-modal entity detection and cross-modal grounding. In contrast, stronger models generally succeed in identifying e_1, but exhibit divergent performance at the a_2 stage depending on the reasoning path. By decomposing each three-hop path into two cross-modal grounding steps, we find that transition between text and image (T-I and I-T), as well as from speech to other modalities (S-I and S-T), is relatively robust. However, *reasoning that moves to the speech modality (I-S and T-S) proves particularly challenging*. We define this as **asymmetric omni-modal grounding**, underscoring inconsistencies in processing across modality orders.

6.3 Domain-Specific Analysis

Figure 9 presents the domain-specific performance of Gemini 3 Flash. Performance varies across domains, with a maximum gap of 21.8% between the economics and nutrition domains under T-S-I in OMHBench-Reasoning. This implies that even the best model lacks uniform domain generalization, performing better on common domains (e.g., economics) than on technical ones (e.g., nutrition).

6.4 Case Study

To complement quantitative analyses, we present case studies in Figure 10 with three key findings.

[8]We perform this analysis on OMHBench-Connect for a controlled study of the required reasoning operations.

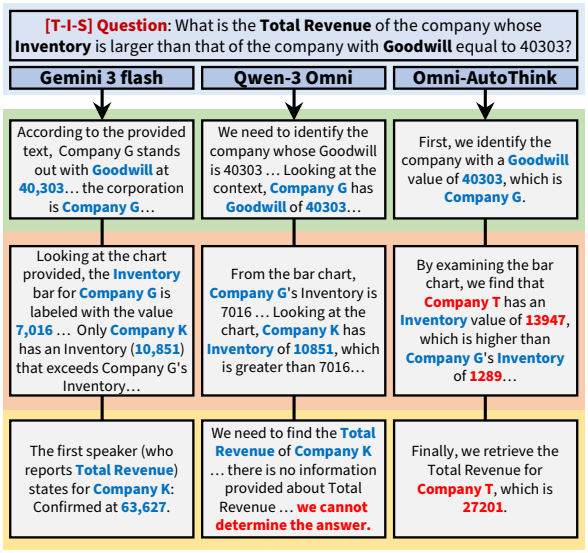


Figure 10: Three case studies show (1) adherence to the intended multi-hop reasoning path without explicit guidance, (2) neglect of the speech modality depending on its position, (3) error accumulation in weaker models.

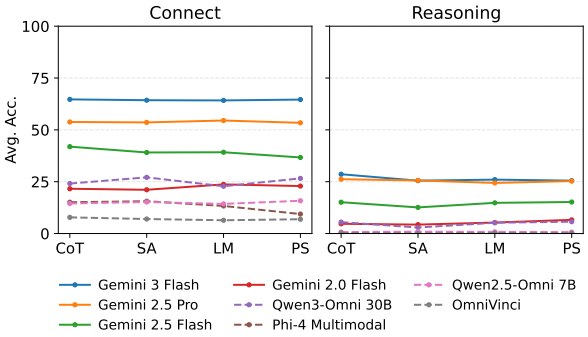


Figure 11: Performance of four prompting methods on OMHBench: Chain-of-Thought (CoT), Self-Ask (SA), Least-to-Most (LM), and Plan-and-Solve (PS). They yield limited gains, calling for dedicated future research.

(1) Even without explicit guidance in the prompt (i.e., zero-shot CoT), models consistently attempt to follow the intended reasoning path, confirming that OMHBench requires structured multi-hop reasoning. (2) Models sometimes behave as if the speech modality were absent, reporting missing evidence despite speech information being provided; this behavior depends on the position of speech within the reasoning path (with similar input context lengths). (3) Weaker models exhibit error accumulation, where early mistakes propagate and result in cascading failures at later reasoning stages.

6.5 Prompting Alone is Insufficient

To examine whether the identified challenge can be alleviated by simply adopting advanced prompting

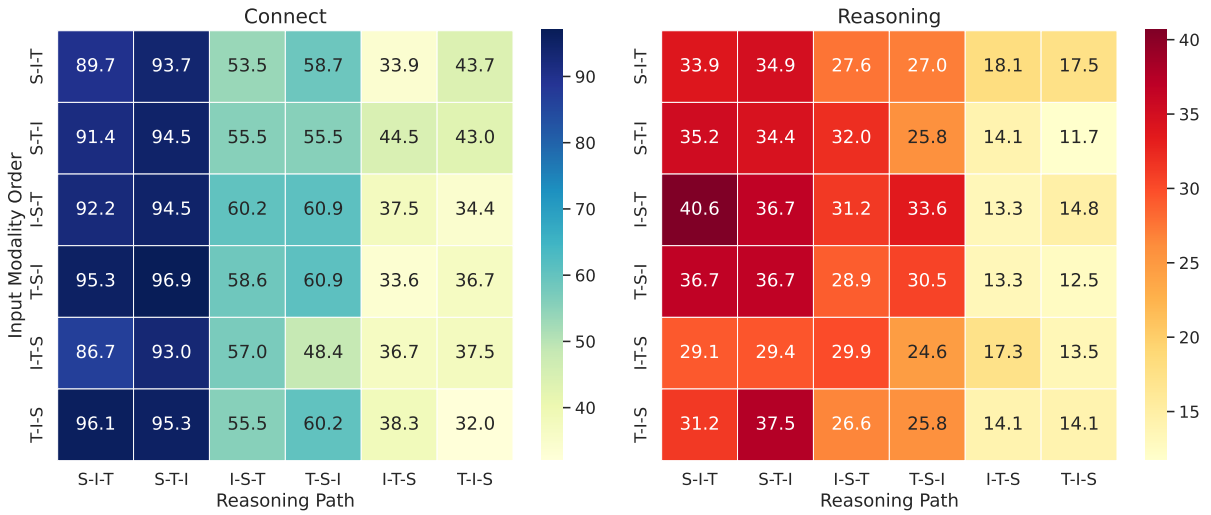


Figure 12: Heatmaps of performance across all combinations of input modality order (rows) and reasoning path (columns) on OMHBench-Connect and OMHBench-Reasoning. Input modality order refers to the sequence in which the three modal contexts (Speech, Image, Text) are presented to the model as input, while reasoning path refers to the modality that must be consulted at each hop. Each cell reports the exact match accuracy (%) of Gemini 3 Flash under that configuration.

techniques, we evaluate three multi-hop prompting strategies: Self-Ask (Press et al., 2023), Least-to-Most (Zhou et al., 2022), and Plan-and-Solve (Wang et al., 2023). As reported in Figure 11, none of these methods yields consistent improvements over the standard chain-of-thought baseline. This suggests that *asymmetric omni-modal grounding* is not primarily caused by insufficient prompt optimization, but instead reflects a fundamental limitation in transferring semantic representations across modalities, particularly into the speech modality.

6.6 Impact of Input Modality Order

Prior work has shown that altering the input order can affect model behavior (Chen et al., 2024; Tan et al., 2024). Moreover, models tend to prioritize information presented earlier in the input sequence, which can influence the reasoning process (Liu et al., 2023b; Wallace et al., 2024). To examine the impact of input modality order—the sequence in which the three modal contexts are presented to the model, which is independent of the reasoning path—on performance, we conducted a systematic analysis on OMHBench using Gemini 3 Flash. We enumerated all six possible input modality orders and evaluated accuracy across all 36 combinations of input orders and reasoning paths, as shown in Figure 12. We observe that performance varies non-negligibly across input orders even when the reasoning path is held fixed. For OMHBench-Connect, accuracy varies by up to 12.5 percentage points across input orders under the T-S-I reasoning path. For OMHBench-Reasoning, the corresponding variation reaches 11.5 percentage points under S-I-T. Although the overall trends across reasoning paths remain broadly consistent, these input-order-induced fluctuations can still introduce noise into comparative analysis. Therefore, we randomize the input modality order in our main experiments to reduce this source of variance.

7 Conclusion

We present OMHBench, a dataset for robust evaluation of omni-modal multi-hop reasoning. It addresses limitations of prior benchmarks by enforcing joint grounding across all three modalities and ensuring balanced proportions of distinct reasoning paths. Experiments on OMHBench provide new insights into how MLLMs perform multi-hop reasoning, revealing that they are sensitive to modality orders and struggle particularly with cross-modal grounding. In future work, we plan to explore training methods to improve the core multi-hop reasoning capabilities of omni-modal models.

Limitations

OMHBench adopts an entity-attribute based formulation with fixed three-hop omni-modal reasoning chains to enable controlled and balanced evaluation of reasoning paths and to prevent modality shortcuts. While this design facilitates precise anal-

ysis of modality interactions and fair comparison across different reasoning paths, it primarily targets reasoning scenarios that can be expressed through explicit entity-attribute relations and fixed-depth chains. Extending the benchmark to support more diverse reasoning patterns is a promising direction for future work.

Acknowledgments

This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (No.RS-2020-II201373, Artificial Intelligence Graduate School Program(Hanyang University)). This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) under the artificial intelligence semiconductor support program to nurture the best talents (IITP-(2026)-RS-2023-00253914) grant funded by the Korea government(MSIT). This work was supported by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT) (RS-2025-00558151).

References

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, and 1 others. 2022. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736.

Anthropic. 2025a. Claude haiku 4.5 system card. https://www-cdn.anthropic.com/7aad69bf12627d42234e01ee7c36305dc2f6a970.pdf. Accessed: 2026-04-14.

Anthropic. 2025b. Claude sonnet 4.5 system card. https://assets.anthropic.com/m/12f214efcc2f457a/original/Claude-Sonnet-4-5-System-Card.pdf. Accessed: 2025-12-30.

Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. 2019. Multimodal machine learning: A survey and taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(2):423–443.

Yupeng Cao, Haohang Li, Yangyang Yu, Shashidhar Reddy Javaji, Yueru He, Jimin Huang, Zining Zhu, Qianqian Xie, Xiao-yang Liu, Koduvayur Subbalakshmi, and 1 others. 2025. Finaudio: A benchmark for audio large language models in financial applications. *arXiv preprint arXiv:2503.20990*.

Chen Chen, ZeYang Hu, Fengjiao Chen, Liya Ma, Jiaxing Liu, Xiaoyu Li, Ziwen Wang, Xuezhi Cao, and Xunliang Cai. 2025. Uno-bench: A unified benchmark for exploring the compositional law between uni-modal and omni-modal in omni models. *arXiv preprint arXiv:2510.18915*.

Xinyun Chen, Ryan A Chi, Xuezhi Wang, and Denny Zhou. 2024. Premise order matters in reasoning with large language models. *arXiv preprint arXiv:2402.08939*.

Theodore E Christensen, Karson E Fronk, Joshua A Lee, and Karen K Nelson. 2024. Data visualization in 10-k filings. *Journal of Accounting and Economics*, 77(2-3):101631.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, and 1 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.

Alexander R Fabbri, Chien-Sheng Wu, Wenhao Liu, and Caiming Xiong. 2021. Qafacteval: Improved qa-based factual consistency evaluation for summarization. *arXiv preprint arXiv:2112.08542*.

Negar Foroutan, Angelika Romanou, Matin Ansaripour, Julian Martin Eisenschlos, Karl Aberer, and Rémi Lebret. 2025. Wikimixqa: A multimodal benchmark for question answering over tables and charts. *arXiv preprint arXiv:2506.15594*.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, and 1 others. 2025. Gemini: A family of highly capable multimodal models. *Preprint*, arXiv:2312.11805.

Sreyan Ghosh, Sonal Kumar, Ashish Seth, Chandra Kiran Reddy Evuru, Utkarsh Tyagi, S Sakshi, Oriol Nieto, Ramani Duraiswami, and Dinesh Manocha. 2024. Gama: A large audio-language model with advanced audio understanding and complex reasoning abilities. *arXiv preprint arXiv:2406.11768*.

Google. 2025a. Gemini 2.0 flash model card. https://modelcards.withgoogle.com/assets/documents/gemini-2-flash.pdf. Accessed: 2025-12-30.

Google. 2025b. Gemini 3 flash model card. https://storage.googleapis.com/deepmind-media/Model-Cards/Gemini-3-Flash-Model-Card.pdf. Accessed: 2025-12-30.

Aaryan Gupta, Vinya Dengre, Hamza Abubakar Kheruwala, and Manan Shah. 2020. Comprehensive review of text-mining applications in finance. *Financial Innovation*, 6(1):39.

Hexgrad. 2025. Kokoro-82m (revision d8b4fc7).

Jack Hong, Shilin Yan, Jiayin Cai, Xiaolong Jiang, Yao Hu, and Weidi Xie. 2025. Worldsense: Evaluating real-world omnimodal understanding for multimodal llms. *arXiv preprint arXiv:2502.04326*.

Youngrok Jang, Hyesoo Kong, Gyeonghun Kim, Yejin Lee, Jungkyu Choi, and Kyunghoon Bae. 2025. Ictqa: Question answering over multi-modal contexts including image, chart, and text modalities. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 138–148.

Seunghee Kim, Changhyeon Kim, and Taeuk Kim. 2025. FCMR: robust evaluation of financial cross-modal multi-hop reasoning. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 23352–23380.

Sungahn Ko, Isaac Cho, Shehzad Afzal, Calvin Yau, Junghoon Chae, Abish Malik, Kaethe Beck, Yun Jang, William Ribarsky, and David S Ebert. 2016. A survey on visual analysis approaches for financial data. In *Computer Graphics Forum*, volume 35, pages 599–617. Wiley Online Library.

Zhifeng Kong, Arushi Goel, Rohan Badlani, Wei Ping, Rafael Valle, and Bryan Catanzaro. 2024. Audio flamingo: A novel audio language model with few-shot learning and dialogue abilities. *arXiv preprint arXiv:2402.01831*.

Anurag Kumar, Ke Tan, Zhaoheng Ni, Pranay Manocha, Xiaohui Zhang, Ethan Henderson, and Buye Xu. 2023. Torchaudio-squim: Reference-less speech quality and intelligibility measures in torchaudio. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE.

B Shravan Kumar and Vadlamani Ravi. 2016. A survey of the applications of text mining in financial domain. *Knowledge-Based Systems*, 114:128–147.

Caorui Li, Yu Chen, Yiyan Ji, Jin Xu, Zhenyu Cui, Shihao Li, Yuanxing Zhang, Jiafu Tang, Zhenghao Song, Dingling Zhang, and 1 others. 2025a. Omnivideobench: Towards audio-visual understanding evaluation for omni mllms. *arXiv preprint arXiv:2510.10689*.

Yizhi Li, Ge Zhang, Yinghao Ma, Ruibin Yuan, Kang Zhu, Hangyu Guo, Yiming Liang, Jiaheng Liu, Zekun Wang, Jian Yang, Siwei Wu, Xingwei Qu, Jinjie Shi, Xinyue Zhang, Zhenzhu Yang, Xiangzhou Wang, Zhaoxiang Zhang, Zachary Liu, Emmanouil Benetos, and 2 others. 2025b. Omnibench: Towards the future of universal omni-language models. *Preprint*, arXiv:2409.15272.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023a. Visual instruction tuning. *Preprint*, arXiv:2304.08485.

Nelson F Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2023b. Lost in the middle: How language models use long contexts. *arXiv preprint arXiv:2307.03172*.

Timothy Liu and 1 others. 2022. Towards better characterization of paraphrases. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8592–8601.

Microsoft, :, Abdelrahman Abouelenin, Atabak Ashfaq, Adam Atkinson, Hany Awadalla, Nguyen Bach, Jianmin Bao, and 1 others. 2025. Phi-4-mini technical report: Compact yet powerful multimodal language models via mixture-of-loras. *Preprint*, arXiv:2503.01743.

Le Thien Phuc Nguyen, Zhuoran Yu, Samuel Low Yu Hang, Subin An, Jeongik Lee, Yohan Ban, SeungEun Chung, Thanh-Huy Nguyen, JuWan Maeng, Soochahn Lee, and 1 others. 2025. See, hear, and understand: Benchmarking audiovisual human speech understanding in multimodal large language models. *arXiv preprint arXiv:2512.02231*.

OpenAI. 2025a. Gpt-5.1 instant and gpt-5.1 thinking system card addendum. https://cdn.openai.com/pdf/4173ec8d-1229-47db-96de-06d87147e07e/5_1_system_card.pdf. Accessed: 2025-12-30.

OpenAI. 2025b. Update to gpt-5 system card: Gpt-5.2. https://cdn.openai.com/pdf/3a4153c8-c748-4b71-8e31-aecbde944f8d/oai_5_2_system-card.pdf. Accessed: 2026-04-14.

Mirjana Pejić Bach, Živko Krstić, Sanja Seljan, and Lejla Turulja. 2019. Text mining for big data analysis in financial sector: A literature review. *Sustainability*, 11(5):1277.

Ofir Press, Muru Zhang, Sewon Min, Ludwig Schmidt, Noah A Smith, and Mike Lewis. 2023. Measuring and narrowing the compositionality gap in language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5687–5711.

Revanth Gangi Reddy, Xilin Rui, Manling Li, Xudong Lin, Haoyang Wen, Jaemin Cho, Lifu Huang, Mohit Bansal, Avirup Sil, Shih-Fu Chang, Alexander Schwing, and Heng Ji. 2022. Mumuqa: Multimedia multi-hop news question answering via cross-media knowledge extraction and grounding. *Preprint*, arXiv:2112.10728.

Alon Talmor, Ori Yoran, Amnon Catav, Dan Lahav, Yizhong Wang, Akari Asai, Gabriel Ilharco, Hannaneh Hajishirzi, and Jonathan Berant. 2021. Multimodalqa: Complex question answering over text, tables and images. *Preprint*, arXiv:2104.06039.

Zhijie Tan, Xu Chu, Weiping Li, and Tong Mo. 2024. Order matters: Exploring order sensitivity in multimodal large language models. *arXiv preprint arXiv:2410.16983*.

Mohammed Majbah Uddin, Rahmat Ullah, and Mohammad Moniruzzaman. 2024. Data visualization in annual reports–impacting investment decisions. *International Journal for Multidisciplinary Research*, 6(5).

Eric Wallace, Kai Xiao, Reimar Leike, Lilian Weng, Johannes Heidecke, and Alex Beutel. 2024. The instruction hierarchy: Training llms to prioritize privileged instructions. *arXiv preprint arXiv:2404.13208*.

Lei Wang, Wanyu Xu, Yihuai Lan, Zhiqiang Hu, Yunshi Lan, Roy Ka-Wei Lee, and Ee-Peng Lim. 2023. Plan-and-solve prompting: Improving zero-shot chain-of-thought reasoning by large language models. *arXiv preprint arXiv:2305.04091*.

xAI. 2025. Grok 4 model card. https://data.x.ai/2025-08-20-grok-4-model-card.pdf. Accessed: 2025-12-30.

Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, Bin Zhang, Xiong Wang, Yunfei Chu, and Junyang Lin. 2025a. Qwen2.5-omni technical report. *Preprint*, arXiv:2503.20215.

Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, and 1 others. 2025b. Qwen2. 5-omni technical report. *arXiv preprint arXiv:2503.20215*.

Jin Xu, Zhifang Guo, Hangrui Hu, Yunfei Chu, Xiong Wang, Jinzheng He, Yuxuan Wang, Xian Shi, Ting He, Xinfa Zhu, Yuanjun Lv, Yongqi Wang, Dake Guo, He Wang, Linhan Ma, Pei Zhang, Xinyu Zhang, Hongkun Hao, Zishan Guo, and 19 others. 2025c. Qwen3-omni technical report. *Preprint*, arXiv:2509.17765.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025a. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Dongchao Yang, Songxiang Liu, Disong Wang, Yuanyuan Wang, Guanglu Wan, and Helen Meng. 2025b. Omni-autothink: Adaptive multimodal reasoning via reinforcement learning. *arXiv preprint arXiv:2512.03783*.

Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, and 1 others. 2024. Minicpm-v: A gpt-4v level mllm on your phone. *arXiv preprint arXiv:2408.01800*.

Hanrong Ye, Chao-Han Huck Yang, Arushi Goel, Wei Huang, Ligeng Zhu, Yuanhang Su, Sean Lin, An-Chieh Cheng, Zhen Wan, Jinchuan Tian, and 1 others. 2025. Omnivinci: Enhancing architecture and data for omni-modal understanding llm. *arXiv preprint arXiv:2510.15870*.

Yuan Zhang, Jason Baldridge, and Luheng He. 2019. Paws: Paraphrase adversaries from word scrambling. *arXiv preprint arXiv:1904.01130*.

Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc Le, and 1 others. 2022. Least-to-most prompting enables complex reasoning in large language models. *arXiv preprint arXiv:2205.10625*.

Ziwei Zhou, Rui Wang, and Zuxuan Wu. 2025. Daily-omni: Towards audio-visual reasoning with temporal alignment across modalities. *arXiv preprint arXiv:2505.17862*.

A Preliminary CMR Dataset Construction

Since MMQA and MuMuQA are formulated as short-answer question answering tasks, while FCMR follows a multiple-choice format, we handle these datasets separately during the preliminary CMR dataset construction stage.

We first preprocess MMQA to align its format with that of MuMuQA, thereby simplifying the overall pipeline design. Specifically, we treat the table input in MMQA instances as a text modality and extend the context column of each instance to include a textual representation of the given table. Next, we select instances from both MuMuQA and our preprocessed MMQA, that can be reformulated using our attribute-entity formulation. This step is necessary because certain instances in MMQA do not require cross-modal reasoning and thus cannot be reversed. We then prompt Gemini 2.0 Flash to decompose each question into its constituent entity and attributes, followed by question-generation step with the prompts provided in Figure 15 and 16. Upon manual inspection of the generated questions, we find that some instances explicitly include the entity or the answer within the question text; we discard such cases. Finally, we use an LLM-based validation step to further filter out questions with an incorrect reasoning order, using the prompt in Figure 17. We pair up instances of two direction—I-T, and T-I—using the generated questions and their original counterparts. Based on these pairs, we construct two sub-datasets: Original, which consists of instances from the original dataset, and Controlled, which includes both directional pairs.

For FCMR, the dataset construction process differs from MMQA and MuMuQA. Since FCMR is designed with a template-based structure and explicit multi-hop reasoning paths, we directly apply Gemini 3 Flash to generate controlled instances over the answer options. Using this procedure, we construct both Original and Controlled versions of the FCMR dataset. The detailed generation prompt is provided in Figure 18.

B Details of OMHBench Construction

The following describes the detailed design choices involved in constructing the benchmark. Example samples from the dataset are shown in Figure 22, Figure 23 and Figure 24.

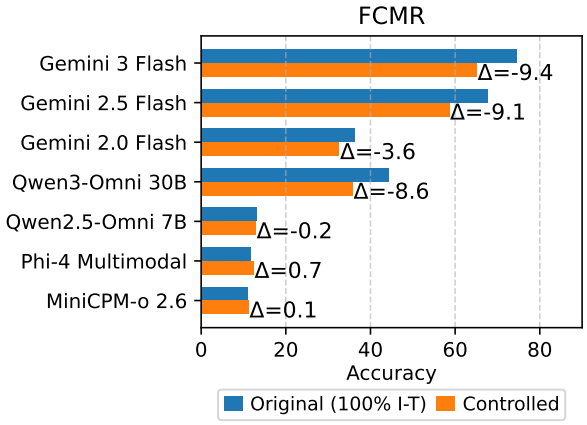


Figure 13: Performance comparison on the *original* FCMR datasets vs. their *controlled* variants. The random baseline is 12.5%, and due to the difficulty of the dataset, most open-source models perform close to this baseline under both settings.

B.1 Data Source

OMHBench comprises four domains: finance, economics, climate, and nutrition. The data for each domain were obtained from Yahoo Finance[9], World Bank[10], Open-Meteo[11], and U.S. Department of Agriculture (USDA)[12], respectively. All values are standardized to use consistent units across entities to ensure comparability. To avoid ambiguous answers, attributes with duplicate values across samples are excluded from the dataset.

Finance Domain The finance domain is constructed using annual financial statements from 23 publicly listed companies. The selected companies and their ticker symbols are: MSFT, NVDA, AVGO, QCOM, TXN, IBM, ADI, MU, KLAC, LLY, MRK, ABBV, TMO, PFE, GILD, BMY, ZTS, PG, KO, PEP, DIS, MDLZ, and HON. For each company, we extract 15 standardized financial attributes from their 2024 annual reports obtained from Yahoo Finance. These indicators span various aspects of corporate performance, including Total Revenue, Cost Of Revenue, Selling General And Administration, Cash And Equivalents, Receivables, Inventory, Other Current Assets, Other Non Current Assets, Net PPE, Goodwill, Payables, Long Term Debt, Other Non Current Liabilities, Depreciation, and Stock Based Compensation.

[9] https://finance.yahoo.com/
[10] https://data.worldbank.org/
[11] https://open-meteo.com/
[12] https://fdc.nal.usda.gov/

Statistics	Number
Dataset	
OMHBench-Connect	3,072
OMHBench-Reasoning	3,072
Total samples	6,144
Reasoning path	6
I-S-T	1,024
I-T-S	1,024
S-I-T	1,024
S-T-I	1,024
T-I-S	1,024
T-S-I	1,024
Domain	4
Finance	1,536
Economics	1,536
Climate	1,536
Nutrition	1,536
Operation types	8
Operation combinations	33
Image type	10
Image color	20
Image font	20
Plot library	2
Speech type	22
Speech voice	27
Text type	24
Generation LLM variants	3

Table 4: Key statistics of OMHBench.

Economics Domain The economics domain is constructed using country-level economic attributes obtained from the World Bank World. All values correspond to annual data for the year 2024. The dataset includes 18 countries: ARG, AUS, BRA, CHE, DEU, EGY, ESP, FRA, GBR, IDN, IND, ITA, MEX, NLD, NOR, SAU, SWE, and ZAF For each country, we collect 18 standardized economic indicators representing major components of national economic activity. The selected indicators include Personal remittances, paid, Personal remittances, received, Total reserves, excluding gold, Final consumption expenditure, Gross fixed capital formation, Gross capital formation, Agriculture, forestry, and fishing, value added, Manufacturing, value added, Industry, including construction, value added, Services, value added, Gross Value Added (GVA) at basic prices, GNI, Gross savings, Taxes less subsidies on products, Merchandise imports, Commercial service imports, Merchandise exports, and Commercial service exports.

Climate Domain The climate domain is constructed using monthly meteorological data collected from major cities around the world. All climate data correspond to the year 2024 and are obtained from Open-Meteo. The dataset includes 20 representative cities: Karachi, Addis Ababa, Cairo, Nairobi, Los Angeles, Tokyo, Ho Chi Minh City, Ulaanbaatar, Chicago, Singapore, Toronto, Shanghai, Manila, London, Lagos, Chengdu, Beijing, Dubai, Rome, and Mumbai. These cities were selected to cover diverse geographic regions and climate conditions. For each city, we collect 12 climate attributes corresponding to the maximum wind speed for each month from January to December.

Nutrition Domain The nutrition domain is constructed using food composition data obtained from the U.S. Department of Agriculture (USDA). The dataset includes 24 food items: Potatoes, mashed, dehydrated, granules without milk, dry form; Sorghum flour, whole-grain; Wheat, KAMUT khorasan, uncooked; PAPA JOHN'S 14" The Works Pizza, Original Crust; Lasagna with meat sauce, frozen, prepared; Potatoes, mashed, home-prepared, whole milk added; Frankfurter, turkey; Seeds, sesame seed kernels, dried (decorticated); Pork, cured, ham – water added, slice, bone-in, separable lean and fat, unheated; Nuts, cashew nuts, raw; T.G.I. FRIDAY'S, chicken fingers, from kids' menu; Pork, cured, ham with natural juices, shank, bone-in, separable lean only, unheated; Broccoli, cooked, boiled, drained, with salt; Pork, cured, ham and water product, shank, bone-in, unheated, separable lean only; Bologna, beef; Teff, uncooked; Nuts, pecans; HOT POCKETS Ham 'N Cheese Stuffed Sandwich, frozen; Pork, cured, ham – water added, slice, boneless, separable lean only, heated, pan-broil; Pork sausage, link/patty, fully cooked, microwaved; DENNY'S, chicken strips; Pork, cured, ham and water product, rump, bone-in, separable lean only, heated, roasted; Pork, cured, ham with natural juices, spiral slice, boneless, separable lean only, unheated; Kielbasa, fully cooked, unheated; For each food item, 19 nutritional attributes are collected. These include Ash, Protein, Lysine, Methionine, Isoleucine, Leucine, Valine, Phenylalanine, Threonine, Histidine, Arginine, Tyrosine, Alanine, Glycine, Serine, Proline, Tryptophan, Cystine, and Glucose.

B.2 Table Triplet Formation

In the first stage of the dataset generation framework, namely the Table Triplet Formation stage, we construct table triplets consisting of three tables

Operation sequence	# Instances
Connect	
Lookup–Comparison–Retrieval	3,072
Reasoning	
Ranking–Ranking–Mean	96
Ranking–Ranking–Summation	96
Ranking–Range–Mean	96
Ranking–Range–Summation	96
Ranking–Comparison–Mean	96
Ranking–Comparison–Summation	96
Ranking–Proximity–Mean	96
Ranking–Proximity–Summation	96
Range–Ranking–Mean	96
Range–Ranking–Summation	96
Range–Range–Mean	96
Range–Range–Summation	96
Range–Comparison–Mean	96
Range–Comparison–Summation	96
Range–Proximity–Mean	96
Range–Proximity–Summation	96
Comparison–Ranking–Mean	96
Comparison–Ranking–Summation	96
Comparison–Range–Mean	96
Comparison–Range–Summation	96
Comparison–Comparison–Mean	96
Comparison–Comparison–Summation	96
Comparison–Proximity–Mean	96
Comparison–Proximity–Summation	96
Proximity–Ranking–Mean	96
Proximity–Ranking–Summation	96
Proximity–Range–Mean	96
Proximity–Range–Summation	96
Proximity–Comparison–Mean	96
Proximity–Comparison–Summation	96
Proximity–Proximity–Mean	96
Proximity–Proximity–Summation	96

Table 5: Operation sequences used to construct OMHBench-Connect and OMHBench-Reasoning.

that share the same set of entities but contain different attributes. Each table contains 10 entities and 3 attributes, resulting in a table size of 10×3.

When forming each table triplet, we ensure that the selected attributes are distinct across the three tables while referring to the same set of entities. To minimize information redundancy, we further curate the attributes so that they are mutually independent and not logically or algebraically derived from one another. In particular, we exclude derived variables such as accounting identities, ratios, and other linearly dependent quantities (e.g., relationships of the form $a_1 + a_2 = a_3$), and retain only primitive, non-derived attributes. In addition, to facilitate subsequent conversion into chart-based representations, we constrain the value ranges of the attributes such that the ratio between the maximum and minimum values does not exceed 30. This restriction prevents extreme scale differences across attributes and ensures stable visualization and comparison across tables.

B.3 Multi-Hop QA Construction

OMHBench constructs each question as a three-hop reasoning chain over entities and their attributes. The first two hops select or filter sets of entities, while the final hop produces a scalar numerical answer. Each hop applies a specific operation to a designated attribute, and all operations are deterministic and rule-based, ensuring full control over the reasoning path. Below, we describe the concrete mechanics of each operation in detail. The valid sequences of operations used to construct multi-hop questions are summarized in Table 5.

Lookup Lookup anchors the reasoning chain to a single entity. From a base table containing ten entities, one entity is randomly selected, and the value of a specified attribute is retrieved. A strict uniqueness constraint is enforced: the selected attribute value must occur exactly once in the table. If the same value appears for multiple entities, the instance is discarded. The output of this operation is a uniquely identified entity and its corresponding attribute value.

Comparison Comparison filters an entity set using a strict inequality condition on a specified attribute. Entities are sorted by the attribute, and a threshold value is constructed such that exactly a predefined number of entities satisfy either a "larger than" or "smaller than" condition. Instances in which ties occur at the decision boundary are discarded. This operation outputs a reduced entity set of fixed size.

Ranking Ranking selects entities based on their ordinal position under a given attribute. Entities are sorted in ascending or descending order, and a fixed number of top or bottom entities are selected. No explicit threshold values are involved. The output is a subset of entities.

Range Range selects entities whose attribute values fall within a contiguous interval. Entities are first sorted by the target attribute, and a consecutive segment is selected. The interval boundaries are defined to ensure that the selected entities are uniquely determined. The output is a subset of entities.

Proximity Proximity selects entities whose attribute values are closest to a reference value. Entities are ranked by their absolute distance to the reference, and the closest entities are selected. The output is a subset of entities.

Retrieval Retrieval is applied to a uniquely determined entity. Given a target attribute, the corresponding attribute value is retrieved from the table and returned as the final answer.

Summation Summation aggregates a numerical attribute over a filtered entity set by summing all corresponding values. The output is a scalar numerical value.

Mean Mean computes the arithmetic average of a numerical attribute over a filtered entity set. Instances in which the resulting mean is non-integer are discarded during dataset construction. The output is a scalar numerical value.

B.4 Omni-Modal Context Generation

Image Modality We convert tabular data into image modality using widely adopted visualization libraries, Matplotlib and Seaborn. A total of ten chart types are generated: vertical bar, horizontal bar, vertical stacked bar, horizontal stacked bar, lollipop, line, scatter, heatmap, bubble, and tile.

To enhance visual diversity, we randomly select one of 20 fonts for each image, uniformly sampled across all images. The fonts used are: Arimo[wght], FiraSansCondensed-Regular, OpenSans-Regular, RobotoSlab[wght], WorkSans-VariableFont[wght], CALIBRI, Kosugi-Regular, OpenSansHebrew-Regular, SourceSansPro-Regular, arial, EBGaramond-VariableFont[wght], Lato-Regular, OpenSansHebrewCondensed-Regular, Tinos-Regular, tahoma, FiraSans-Regular, NotoSans-Regular, Roboto-Regular, Ubuntu-Regular, and times.

In addition, chart elements are colored using a fixed palette of 20 distinct colors: #4E79A7, #A0CBE8, #F28E2B, #FFBE7D, #59A14F, #8CD17D, #B6992D, #F1CE63, #499894, #86BCB6, #E15759, #FF9D9A, #79706E, #BAB0AC, #D37295, #FABFD2, #B07AA1, #D4A6C8, #9D7660, and #D7B5A6. These design choices allow us to construct a rich and diverse set of images.

Text Modality We define a total of 24 representative text scenarios across four domains. For the finance domain, the scenarios include Analyst Report, News Article, Blog Post, Email Newsletter, Executive Summary, and Meeting Minutes. For the economics domain, the scenarios include Analyst Report, News Article, Blog Post, Email Newsletter, Executive Summary, and Meeting Minutes. For the climate domain, the scenarios include Research Report, Business Report, City Marketing, News Article, Blog Post, and Magazine. For the nutrition domain, the scenarios include Research Report, Quality Assurance Log, Dietary Guidelines, Ingredient Encyclopedia, Blog Post, and Magazine.

For each scenario, we design a tailored situational prompt and use large language models to convert structured tabular data into natural language text. Figure 19 illustrates an example of the prompt used for text scenario generation.

Speech Modality We define 22 representative speech scenarios across four domains. For the finance domain, the scenarios include Meeting, Podcast, Seminar, Audit, and News Debate. For the economics domain, the scenarios include Meeting, Podcast, Seminar, News Debate, and Global Summit. For the climate domain, the scenarios include Weather Forecast, Meeting, Airport Control Tower, Sports Event Briefing, Business Risk Briefing, and Green Energy Assessment. For the nutrition domain, the scenarios include Meeting, Lab Briefing, Documentary, Ingredient Safety Audit, Conference, and Podcast. Each scenario consists of a four-speaker dialogue. One speaker serves as the moderator, while the remaining three speakers are each responsible for different attributes. Similar to the text modality, we first use Large Language Models to convert structured tabular data into textual scripts tailored to each scenario. These scripts are then converted into speech using the Kokoro-82M TTS model (Hexgrad, 2025). To enhance acoustic diversity, we utilize 27 distinct voices provided by the Kokoro-82M model, including 14 female and 13 male voices.[13] The female voices include: af_heart, af_alloy, af_aoede, af_bella, af_jessica, af_kore, af_nova, af_river, af_sarah, af_sky, bf_alice, bf_emma, bf_isabella, bf_lily. The male voices include: am_adam, am_echo, am_eric, am_fenrir, am_liam, am_michael, am_onyx, am_puck, am_santa, bm_daniel, bm_fable, bm_george, bm_lewis. This voice configuration ensures a high degree of variation in the generated speech data. Figure 20 illustrates an example of the prompt used for speech scenario generation.

[13]The full list of available voices is provided at https://huggingface.co/hexgrad/Kokoro-82M/blob/main/VOICES.md

C Comparison of Evaluation Methods

We further validate Exact Match in OMHBench by comparing it with human judgment and LLM-as-a-Judge evaluation. Specifically, we randomly sampled 300 instances from OMHBench-Connect and 300 instances from OMHBench-Reasoning, and evaluated Gemini 3 Flash outputs under all three protocols. All methods yielded identical results: 72.7% on OMHBench-Connect and 45.7% on OMHBench-Reasoning. This agreement was consistent across multiple judge models, including Gemini 3 Flash (Google, 2025b), GPT 5.2 (OpenAI, 2025b), Claude Haiku 4.5 (Anthropic, 2025a), and Qwen3-235B-A22B (Yang et al., 2025a).

This result is expected given the design of OMHBench. Since all ground-truth answers are positive integers, evaluation is not affected by paraphrasing or alternative surface forms. In practice, correctness reduces to checking whether the predicted final answer matches the target numeric value, making Exact Match fully consistent with both human and LLM-based evaluation in this benchmark.

D Path Balance Score at Different Thresholds

To provide a more fine-grained view of robustness to reasoning-path variation, we additionally report **PBS@k**. While PBS measures whether a model answers all reasoning-path variants of the same question correctly, PBS@k measures the proportion of question groups answered correctly on at least k out of the six paths.

Formally, let the dataset contain $|D|$ instances, forming $|G| = |D|/N!$ groups. For the i-th group, let $a_{i,j} \in \{0, 1\}$ denote whether the model correctly answers the j-th path. PBS@k is defined as:

$$\text{PBS@}k = \frac{1}{|G|} \sum_{i=1}^{|G|} \mathbb{I}\left(\sum_{j=1}^{N!} a_{i,j} \geq k\right),$$

where $\mathbb{I}(\cdot)$ is the indicator function and $k \in \{1, \ldots, N!\}$. In OMHBench, $N = 3$, so each question has six reasoning-path variants. Thus, PBS@1 indicates whether a model succeeds on at least one path variant, while PBS@6 coincides with PBS. Tables 6 and 7 report PBS@k results on OMHBench-Connect and OMHBench-Reasoning, respectively.

Across models, PBS@k decreases monotonically as k increases, showing that consistency across reasoning paths varies substantially by model and benchmark split. This trend is especially pronounced in OMHBench-Reasoning, where even the strongest models exhibit a large gap between PBS@1 and PBS@6.

Model	PBS@k (%)					
	$k=1$	$k=2$	$k=3$	$k=4$	$k=5$	$k=6$
Proprietary Models						
Gemini 3 Flash	100.0	99.6	95.5	83.8	58.8	32.2
Gemini 2.5 Pro	100.0	99.2	94.5	72.7	43.8	25.0
Gemini 2.5 Flash	96.9	91.4	70.3	43.0	15.6	4.7
Gemini 2.5 Flash-lite	87.5	61.7	30.5	13.3	0.8	0.0
Gemini 2.0 Flash	73.4	39.8	14.1	1.6	0.8	0.0
Gemini 2.0 Flash-lite	72.2	28.6	6.3	0.0	0.0	0.0
Open-Source Models						
Qwen3-Omni 30B	97.9	85.9	55.7	31.2	8.0	2.3
Phi-4 Multimodal	58.8	24.4	6.2	1.2	0.0	0.0
Qwen2.5-Omni 7B	57.8	22.5	5.7	1.2	0.0	0.0
Qwen2.5-Omni 3B	46.1	14.6	2.5	0.4	0.0	0.0
OmniVinci	35.0	9.4	2.1	0.2	0.0	0.0
MiniCPM-o 2.6	30.3	5.1	0.8	0.0	0.0	0.0
Omni-AutoThink	24.6	4.1	0.2	0.0	0.0	0.0

Table 6: PBS@k results on **OMHBench-Connect**. PBS@k measures the proportion of question groups answered correctly on at least k of the six reasoning paths. PBS@6 coincides with PBS.

Model	PBS@k (%)					
	$k=1$	$k=2$	$k=3$	$k=4$	$k=5$	$k=6$
Proprietary Models						
Gemini 3 Flash	90.0	77.5	57.2	41.2	22.1	8.6
Gemini 2.5 Pro	91.4	76.6	56.2	36.7	21.1	10.9
Gemini 2.5 Flash	67.2	36.7	14.1	6.2	1.6	0.0
Gemini 2.5 Flash-lite	44.5	14.8	4.7	0.0	0.0	0.0
Gemini 2.0 Flash	24.2	3.1	0.8	0.0	0.0	0.0
Gemini 2.0 Flash-lite	14.3	2.4	0.0	0.0	0.0	0.0
Open-Source Models						
Qwen3-Omni 30B	54.5	26.4	7.2	1.8	0.2	0.0
Phi-4 Multimodal	1.4	0.2	0.0	0.0	0.0	0.0
Qwen2.5-Omni 7B	3.1	1.0	0.2	0.0	0.0	0.0
Qwen2.5-Omni 3B	2.1	0.0	0.0	0.0	0.0	0.0
OmniVinci	1.4	0.0	0.0	0.0	0.0	0.0
MiniCPM-o 2.6	0.4	0.0	0.0	0.0	0.0	0.0
Omni-AutoThink	1.2	0.0	0.0	0.0	0.0	0.0

Table 7: PBS@k results on **OMHBench-Reasoning**. PBS@k measures the proportion of question groups answered correctly on at least k of the six reasoning paths. PBS@6 coincides with PBS.

E Experimental Environment

All experiments were conducted on a machine equipped with Intel Xeon Gold 6338 CPU (2.00 GHz), and an NVIDIA A100-SXM4 GPU with 80 GB of memory. The system ran Ubuntu 22.04.4 LTS with CUDA compilation tools release 12.4. We used Python 3.10.18 and PyTorch 2.6.0+cu124 as the core software environment. During both dataset generation and evaluation, the random seed was fixed to 42 to ensure reproducibility.

Model	Accuracy by Reasoning Path (%)						Avg. (Acc.)	PBS
	S-I-T	S-T-I	I-S-T	T-S-I	I-T-S	T-I-S		
Proprietary Models								
Gemini 3 Flash (Think)	97.5	98.4	75.4	75.0	60.2	63.5	78.3	32.2
Gemini 3 Flash (Non-Think)	93.0	94.5	60.9	57.8	43.0	39.1	64.7	8.6
Gemini 2.5 Pro (Think)	94.5	96.9	66.4	71.1	55.5	50.8	72.5	25.0
Gemini 2.5 Pro (Non-Think)	65.6	68.8	50.0	58.6	41.4	38.3	53.8	6.2
Gemini 2.5 Flash (Think)	82.0	85.9	50.8	54.7	26.6	21.9	53.6	4.7
Gemini 2.5 Flash (Non-Think)	65.6	69.5	37.5	39.1	24.2	15.6	41.9	2.3
Gemini 2.5 Flash-lite (Think)	49.2	60.9	38.3	35.2	5.5	4.7	32.3	0.0
Gemini 2.5 Flash-lite (Non-Think)	32.8	37.5	28.9	26.6	2.3	0.8	21.5	0.0
Gemini 2.0 Flash	28.9	33.6	26.6	29.7	4.7	6.2	21.6	0.0
Gemini 2.0 Flash-lite	35.9	32.8	21.1	11.7	2.3	2.3	17.7	0.0
Open-Source Models								
Qwen3-Omni 30B (Think)	75.8	77.0	46.7	49.6	16.0	16.0	46.8	2.3
Qwen3-Omni 30B (Non-Think)	35.0	44.7	17.8	33.8	6.2	7.2	24.1	0.0
Phi-4 Multimodal	26.6	23.6	21.5	18.4	0.6	0.0	15.1	0.0
Qwen2.5-Omni 7B	22.7	20.9	19.3	20.5	2.0	1.8	14.5	0.0
Qwen2.5-Omni 3B	12.7	17.6	15.6	14.6	1.2	2.0	10.6	0.0
OmniVinci	14.8	8.6	14.8	7.0	0.8	0.6	7.8	0.0
MiniCPM-o 2.6	8.0	10.9	7.4	8.4	1.2	0.2	6.0	0.0
Omni-AutoThink	7.6	6.6	8.0	6.1	0.6	0.0	4.8	0.0

Table 8: Accuracies and Path Balance Scores (PBSs) across six reasoning paths in **OMHBench-Connect**, including both thinking and non-thinking variants. Avg denotes macro-averaged accuracy. PBSs measure robustness to reasoning path variations.

Model	Accuracy by Reasoning Path (%)						Avg. (Acc.)	PBS
	S-I-T	S-T-I	I-S-T	T-S-I	I-T-S	T-I-S		
Proprietary Models								
Gemini 3 Flash (Think)	55.9	58.8	49.8	49.6	40.0	42.6	49.4	8.6
Gemini 3 Flash (Non-Think)	35.9	39.8	31.2	29.7	16.4	18.8	28.6	1.6
Gemini 2.5 Pro (Think)	53.9	51.6	52.3	47.7	41.4	46.1	48.8	10.9
Gemini 2.5 Pro (Non-Think)	28.9	32.8	30.5	28.9	14.8	21.1	26.2	0.0
Gemini 2.5 Flash (Think)	32.0	30.5	17.2	24.2	10.9	10.9	21.0	0.0
Gemini 2.5 Flash (Non-Think)	22.7	24.2	17.2	14.8	6.2	5.5	15.1	0.0
Gemini 2.5 Flash-lite (Think)	18.8	21.1	15.6	8.6	0.0	0.0	10.7	0.0
Gemini 2.5 Flash-lite (Non-Think)	7.8	9.4	3.9	5.5	0.0	0.8	4.6	0.0
Gemini 2.0 Flash	4.7	11.7	4.7	6.2	0.8	0.0	4.7	0.0
Gemini 2.0 Flash-lite	3.9	5.5	3.9	2.3	0.8	0.0	2.7	0.0
Open-Source Models								
Qwen3-Omni 30B (Think)	27.3	28.5	14.1	14.6	2.7	2.7	15.0	0.0
Qwen3-Omni 30B (Non-Think)	9.8	10.5	4.7	6.4	0.8	0.6	5.5	0.0
Phi-4 Multimodal	0.6	0.4	0.2	0.0	0.2	0.2	0.3	0.0
Qwen2.5-Omni 7B	0.4	1.0	1.0	0.6	0.2	1.2	0.7	0.0
Qwen2.5-Omni 3B	0.8	0.6	0.2	0.0	0.4	0.2	0.4	0.0
OmniVinci	0.6	0.2	0.2	0.4	0.0	0.0	0.2	0.0
MiniCPM-o 2.6	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Omni-AutoThink	0.4	0.2	0.4	0.2	0.0	0.0	0.2	0.0

Table 9: Accuracies and Path Balance Scores (PBSs) across six reasoning paths in **OMHBench-Reasoning**, including both thinking and non-thinking variants. Avg denotes macro-averaged accuracy. PBSs measure robustness to reasoning path variations.

Think step by step and provide the final answer at the end.

Based on the given question above, build a chain of sub-questions and intermediate answers to solve it.
First, determine if follow-up questions are needed. If yes, output "Are follow up questions needed here: Yes.".
Then, explicitly state a "Follow up:" question and provide an "Intermediate answer:" for it.
Repeat this process until you have enough information to determine the final answer.
Finally, conclude your reasoning.

To solve the complex problem above, first decompose it into a series of simpler sub-questions.
Then, solve each sub-question in order, using the information from previous answers to reach the final conclusion.

Your response should follow this structure:
1. Decompose the problem: "To answer the question, we need to know: [List sub-questions]“
2. Solve sequentially: Solve each sub-question one by one.
3. Provide the final answer.

Let's first understand the problem, extract relevant variables and their corresponding numerals, and devise a complete plan.

Then, let's carry out the plan, calculate intermediate variables (pay attention to correct numerical calculation and commonsense), solve the problem step by step, and show the answer.

Figure 14: Prompt formats used for each prompting method: Zero-Shot CoT, Self-Ask, Least-to-Most, and Plan-and-Solve.

```
Question: {question}
Context: {context}
Entities: {entities}
Answers: {answers}
Question Type: {question_type}

**Mathematical Formulation**
Attribute a_I: Value of some attribute of an entity exclusive to image
Attribute a_T: Value of some attribute of an entity exclusive to text
Relation: Relation connecting Attribute and Entity or two Entities

The question given is formulated as Image-to-Text (I-T) type or Text-to-Image (T-I) type.
I-T question gives value of attribute a_I in the question and asks for a value of attribute a_T for the answer.
T-I question gives value of attribute a_T in the question and asks for a value of attribute a_I for the answer.

Your task is to identify the a_I and a_T from the given entity and question from given context.
Your final answer should be in the form of json containing two items, a_I and a_T.
```

Figure 15: Prompt for extracting attributes for a entity from each question in MMQA and MuMuQA.

```
Question: {question}
Question Type: {question_type}
reversed_question: {reversed_question}

Attribute a_I: Value of some attribute of an entity exclusive to image
Attribute a_T: Value of some attribute of an entity exclusive to text

Question Types:
I-T type: a_I -> entity -> a_T.
T-I type: a_T -> entity -> a_I.

Given two multimodal questions, you need to determine whether given
Question is type of {question_type}, AND reversed_question is type of {reverse_type}.
After thinking step-by-step, give your final answer as Yes or No.
```

Figure 17: Prompt for validating the reasoning order of generated questions in MMQA and MuMuQA.

```
Question: {question}
Context: {context}
Entities: {entities}
Answers: {answers}
Question Type: {question_type}
a_I: {a_I}
a_T: {a_T}

You are given a QA pair involving multimodal reasoning with attributes from two different modalities: image (a_I) and text (a_T).
A relation connects these attributes or the attribute and an entity.

The question given is formulated as I-T type: a_I -> entity -> a_T.
For instance, if question is "What did the government of the person in the image with the grey tie do?", With a_I = color_of_tie, e = person, and a_T (answer) = conduct_investigation

Your task is to reverse the original question, generating a new question (T-I) that:
- Asks for the value of the attribute that was originally given in the question.
- Uses the original answer as a part of the generated question. (e.g. position in the image, color)

Think step by step and give your final answer in the form of json containing 2 items: reversed_question, answer_reversed_question
```

Figure 16: Prompt for generating question from our formulation in MMQA and MuMUQA.

```
The provided options are multi-hop options that, given text, table, and chart as inputs, always reason from Image (Chart) to Text.

Each option follows an explicit reasoning path, defined as an ordered sequence of modality relations.
In the original formulation used in this task, the options reason from Chart to Text.

I want to change ONLY the reasoning path so that the option instead reasons from Text to Chart, while keeping the content/meaning as unchanged as possible.

Do NOT add new facts. Do NOT remove important constraints.
Preserve entities, numbers, and conditions.
Return in EXACTLY the following format (one line per item, keep the tags):

[option1]: <rewritten option1>
[option2]: <rewritten option2>
[option3]: <rewritten option3>

Here are the inputs:
option1: {option1}
option2: {option2}
option3: {option3}
```

Figure 18: Prompt for controlling the reasoning path of each option in FCMR.

You are a financial analyst writing a professional analyst report based on the provided data.
Write a clear, structured, and insight-driven report that includes the following:
- Executive summary of the key findings
- Detailed analysis of all relevant figures
- Interpretation of trends and possible implications

Use a formal and objective tone suitable for investors and stakeholders.
Ensure that every single value from the data is explicitly mentioned.
No Markdown: Output strictly as plain text. Do NOT use any Markdown formatting syntax. Do not bold company names or values. Use only standard paragraph breaks for structure.
Make sure to cover all information from the given data, whether it relates to companies, financial metrics, or other quantitative or categorical indicators.
Present the content in paragraph format with optional section headers.
All units are standardized. Do not explicitly mention or append the units.

You are a journalist writing a news article based on the provided data.
Write an informative, concise, and neutral report that would appear in a financial news section.
The article should:
- Lead with a headline that reflects the core message
- Summarize key highlights in the opening paragraph
- Follow with detailed paragraphs that explain each data point or trend
- Include all information from the data, regardless of type (companies, financials, etc.)

Ensure that every single value from the data is explicitly mentioned.
No Markdown: Output strictly as plain text. Do NOT use any Markdown formatting syntax. Do not bold company names or values. Use only standard paragraph breaks for structure.
Maintain an objective tone throughout the piece.
Use paragraph-based prose, not a script or dialogue.
All units are standardized. Do not explicitly mention or append the units.

Figure 19: Example prompts used to generate the text.

Role: Financial Meeting Scriptwriter

Task:
Convert the provided financial data table into a meeting transcript.
The listener must identify the attribute ONLY by recognizing the speaker's voice from the introduction.

Context:
- [Speaker1]: The Moderator.
- [Speaker2]: Handles the metric in the 2nd Column.
- [Speaker3]: Handles the metric in the 3rd Column.
- [Speaker4]: Handles the metric in the 4th Column.

Strict Guidelines:
1. **Intro (The Voice Mapping Phase):**
  - [Speaker1] MUST ask "Who is in charge of [Header Name]?"
  - The corresponding Speaker MUST reply "That is me" or "I am."
  - Repeat this for all 3 metrics. This is the **ONLY** time headers are mentioned.

2. **Body (The Test):**
  - [Speaker1] simply announces the **Company Name** (e.g., "Let's look at Company X.").
  - [Speaker2], [Speaker3], [Speaker4] must immediately report their numbers in a **RANDOM sequence**.
  - **CRITICAL**: The text MUST NOT contain clues like "My asset value is..." or "The debt is...".
  - Use generic phrases: "I have...", "On my sheet...", "It is...".

3. **Randomization**:
  - For every company, the order of Sp2, Sp3, Sp4 MUST be shuffled.
  - Example: Co A (2->3->4), Co B (4->2->3), Co C (3->4->2).

4. Format: Plain text only. NO Markdown.
All units are standardized. Do not explicitly mention or append the units.

Example for Learning:

[Input Data]
Company,Receivables,Payables,Cash
X,1000,500,200
Y,3000,100,500

[Desired Output]
[Speaker1]: Let's get started. Who is tracking Receivables?
[Speaker2]: That would be me.
[Speaker1]: And who has the Payables figures?
[Speaker3]: I have those numbers.
[Speaker1]: Lastly, who is monitoring Cash?
[Speaker4]: I'm ready with that.

[Speaker1]: Great. Let's look at Company Y first.
[Speaker4]: On my end, it seems moderate. I'm seeing 500.
[Speaker2]: Well, looking at my sheet, it's quite high. It's sitting at 3000.
[Speaker3]: My figure is actually surprisingly low. It's just 100.

[Speaker1]: Okay, moving on to Company X.
[Speaker3]: The number on my list is 500.
[Speaker4]: It's a bit lower on my side. The value is 200.
[Speaker2]: I have a round number here. It shows exactly 1000.

Real Task:
Generate the script based on this Real Data:
{table_data_json}

Figure 20: Example prompt used to generate the speech script.

Role
You are an expert evaluator for Multi-hop Question Answering models.
Your task is to analyze a candidate model's response against a specific Question and Ground Truth Answer, diagnosing exactly where the reasoning chain failed.

Input Data
Question: {question}
Ground Truth Answer: {ground_truth_answer}

Candidate Model Response:
[other_model_response START]
{model_response}
[other_model_response END]

Reasoning Logic Chain
This multi-hop QA problem relies on a specific chain of reasoning connecting attributes and entities. The correct logical path is defined as follows:

1. Attribute 1 Value: {attr_1_val}
2. Entity 1: {ent_1}
3. Attribute 2 Value: {attr_2_val}
4. Entity 2: {ent_2}
5. Attribute 3 Value (Final Answer): {attr_3_val}

Evaluation Rules
Compare the "Candidate Model Response" against the "Reasoning Logic Chain" above. Identify the *first* step where the model failed. Assign a failure case based on the following taxonomy:

* Case 1: The model correctly recognized 'Attribute 1 Value' but failed to identify 'Entity 1'.
* Case 2: The model correctly recognized 'Attribute 1 Value' and 'Entity 1', but failed to recognize 'Attribute 2 Value'.
* Case 3: The model correctly recognized 'Attribute 1 Value', 'Entity 1', and 'Attribute 2 Value', but failed to identify 'Entity 2'.
* Case 4: The model correctly recognized everything up to 'Entity 2', but failed to derive the final 'Attribute 3 Value'.
* Case 5: The model successfully followed the entire chain and the final answer is correct.

Equivalence Guidelines
* Numerical Formatting: Ignore commas or formatting differences (e.g., "18,592" is equal to "18592").
* Units: Ignore unit discrepancies if the core value is correct (e.g., "26 million" is considered equal to "26" if the target value is 26).

Response Format
Provide a brief explanation of your analysis followed by the final case number.

Explain: <Explanation of the reasoning and the specific point of failure>
Final case: <1/2/3/4/5>

Figure 21: Prompt used to Step-by-Step Failure Analysis.

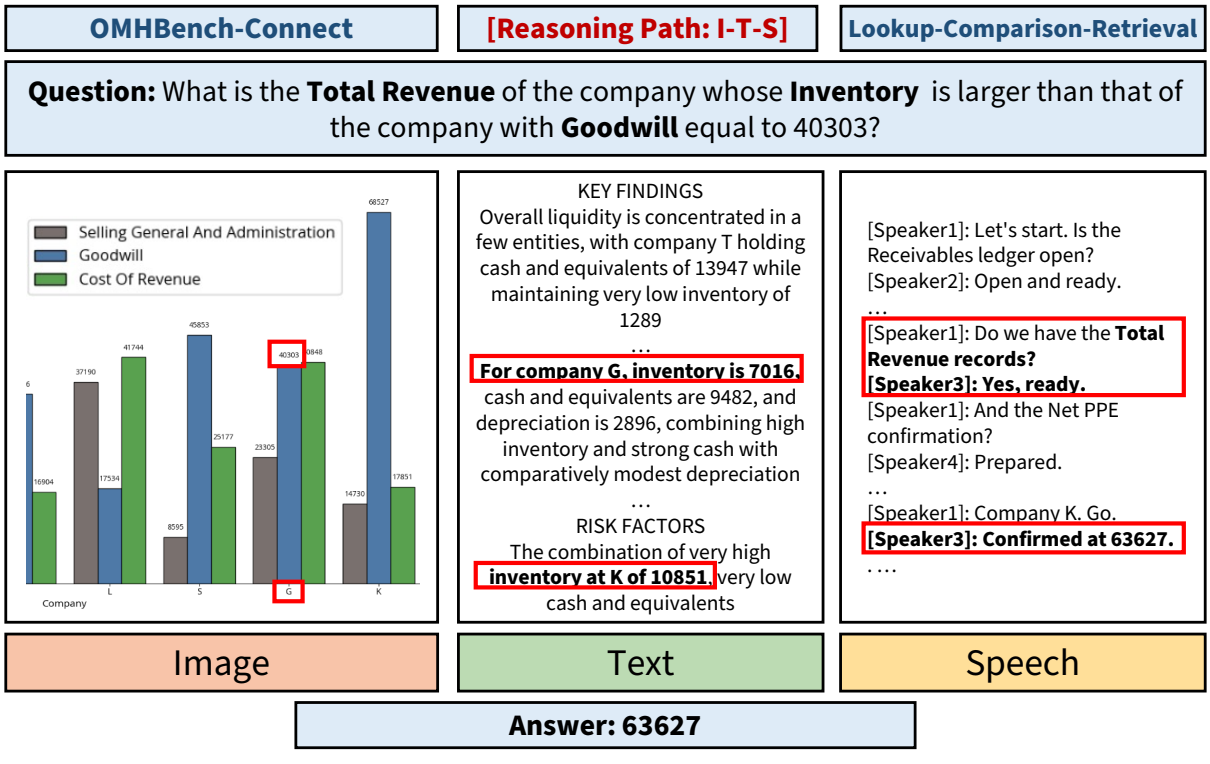


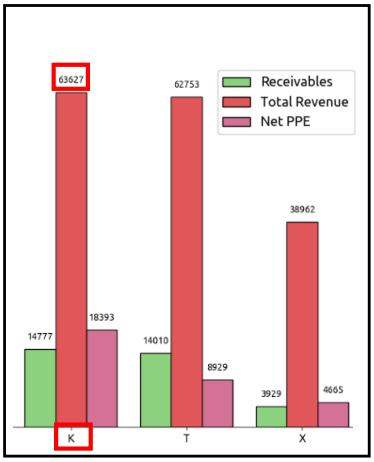
Figure 22: Example of OMHBench-Connect I-T-S Instance.

OMHBench-Connect | **[Reasoning Path: T-S-I]** | **Lookup-Comparison-Retrieval**

Question: What is the **Total Revenue** of the company whose **Inventory** is larger than that of the company with **Goodwill** equal to 40303?

Subject: New Financial Snapshot Reveals Wide Variations in Cost Structures Across Ten Companies
As we close out the quarter, a fresh look at the financial metrics
…
Company G stands out with Selling General and Administration
at 23305. **Goodwill at 40303**, and Cost of Revenue at 40848. The high Cost of Revenue suggests either premium input costs or a product mix requiring substantial direct expenses.
…
understanding these structural differences will be essential for assessing resilience and growth potential…

[Speaker1]: Breaking news. Who is reporting on **Inventory?**
[Speaker2]: I am covering that.
…
[Speaker1]: Update on **Company G!**
[Speaker2]: This one is pushing higher, my number is reaching **7016!**
[Speaker3]: I'm tracking a strong reading as well,
…
[Speaker1]: Update on **Company K!**
[Speaker2]: By contrast, I'm watching a big value, rocketing up to **10851!**
[Speaker4]: And my reading is elevated too
…



Text | Speech | Image

Answer: 63627

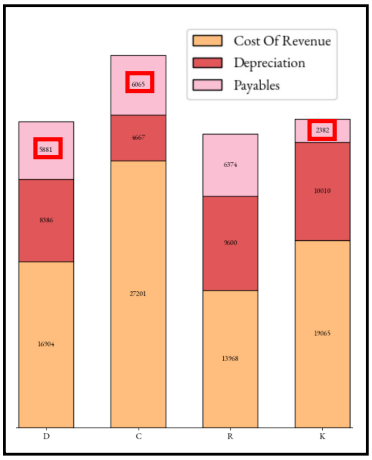
Figure 23: Example of OMHBench-Connect T-S-I Instance.

OMHBench-Reasoning | **[Reasoning Path: T-S-I]** | **Range-Ranking-Mean**

Question: What is the mean of **Payables** for the companies that are the top 3 biggest in **Total Revenue** from the companies whose **Receivables** is greater than 4174.5 and less than 14217.0?

Current Asset Composition Reveals Divergent Business Models Across Ten Companies
…
Company B reports the highest receivables at 14777, accompanied by inventory of 10851 and other current assets of 4253, giving it the largest overall current asset position among the group. This substantial inventory level, the highest across all companies, suggests either a manufacturing focus or retail operations requiring significant stock holdings.
…
In the mid-range, Company J reports receivables of 6615, inventory of 8875, and other current assets of 776
…

…
[Speaker1]: Now, turn to N, another modest ingredient that often slips by without fanfare.
[Speaker4]: Hidden within it lies 469, a steady flow that supports delicate internal processes with calm reliability.
[Speaker2]: We uncovered 705, suggesting that this food has more structural strength than its unassuming form would suggest.
[Speaker3]: Not to be overlooked, it brings 403 of my preferred particles, a balanced amount that reinforces its quiet importance.
…



Text | Speech | Image

Answer: 4776

Figure 24: Example of OMHBench-Reasoning T-S-I Instance.