

MARGEN: Multi-Agent LLM Approach for Self-Directed Market Research and Analysis

Roman Koshkin*
Okinawa Institute of Science and
Technology
Onna, Okinawa, Japan
roman.koshkin@gmail.com

Pengyu Dai*
Institute of Science Tokyo
Tokyo, Japan
dai.p.aa@m.titech.ac.jp

Nozomi Fujikawa
Amazon
Tokyo, Japan
nozofuji@amazon.co.jp

Masahito Togami
Amazon
Tokyo, Japan
mtogami@amazon.co.jp

Marco Visentini-Scarzanella
Amazon
Tokyo, Japan
marcovs@amazon.co.jp

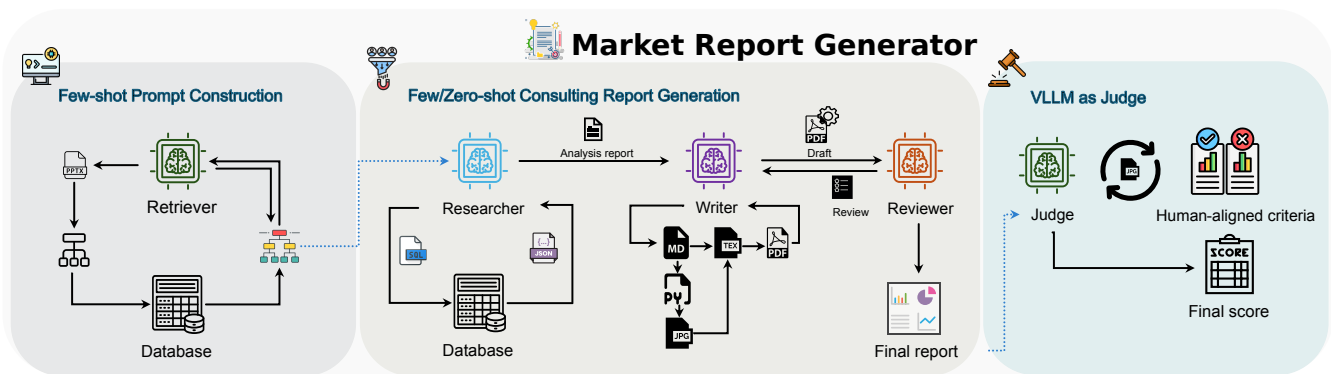


Figure 1: Overview of MARGEN. The **RESEARCHER** writes a sequence of SQL queries (one at a time), each of which is executed and the returned data are added to the prompt. This process continues until the LLM decides that it has enough information to write a research report or exceeds `MAX_QUERIES`. Based on the research report, the **WRITER** creates relevant figures and LaTeX code, which are compiled to PDF and sent to the **REVIEWER**, which suggests improvements. After several rounds of review, the final report is returned to the user. MARGEN generates multiple reports and only the best of them (as decided by the **JUDGE** agent) will be returned to the user. Additionally, the **RETRIEVER** agent can extract business insights from available unstructured materials, which can be fed to the **RESEARCHER** and help improve the quality of the final report.

ABSTRACT

We present an autonomous framework that leverages Large Language Models (LLMs) to automate end-to-end business analysis and market report generation. At its core, the system employs specialized agents - Researcher, Reviewer, Writer, and Retriever - that collaborate to analyze data and produce comprehensive reports. These agents learn from real professional consultants' presentation materials at Amazon through in-context learning to replicate professional analytical methodologies. The framework executes a multi-step process: querying databases, analyzing data, generating insights, creating visualizations, and composing market reports. We also introduce a novel LLM-based evaluation system for assessing report quality, which shows alignment with expert human evaluations. Building on these evaluations, we implement an iterative improvement mechanism that optimizes report quality through

automated review cycles. Experimental results show that report quality can be improved by both automated review cycles and consultants' unstructured knowledge. In experimental validation, our framework generates detailed 6-page reports in 7 minutes at a cost of approximately \$1. Our work could be an important step to automatically create affordable market insights.

CCS CONCEPTS

• **Computing methodologies** → **Intelligent agents; Multi-agent systems; Natural language generation; Artificial intelligence.**

KEYWORDS

LLM, Multi-agent, report generation

ACM Reference Format:

Roman Koshkin, Pengyu Dai, Nozomi Fujikawa, Masahito Togami, and Marco Visentini-Scarzanella. 2025. MARGEN: Multi-Agent LLM Approach for Self-Directed Market Research and Analysis. In *Proceedings of LLM4ECommerce Workshop at the 31st ACM SIGKDD Conference on Knowledge Discovery and*

*Both authors contributed equally to this research.

Data Mining V.2 (LLM4ECommerce Workshop at KDD '25). ACM, New York, NY, USA, 36 pages.

1 INTRODUCTION

In today's fast-paced e-commerce environment, the extraction of actionable insights from vast and diverse datasets encompassing customer behavior, product performance, and related variables - and the subsequent application of these insights to strategic decision-making processes - represents a critical imperative for contemporary organizations. However, many organizations face significant challenges in efficiently analyzing data due to difficulty of securing skilled business / data analysts.

To resolve this challenge and streamline data-driven decision making through automated e-commerce analytics and insight generation, we introduce MARKET REPORT GENERATOR (MARGEN), a comprehensive and fully automated framework designed to accelerate market research. Fig. 1 offers an overview of MARGEN. MARGEN implements a multi-agent system that automates business analysis—from interpreting client requirements to querying databases and generating analytical reports with supporting visualizations. MARGEN distills domain expertise and reasoning from unstructured text authored by professional human consultants and incorporates by in-context learning. For real-world applications, evaluation of report quality is essential. We further propose an automated evaluation framework based on pairwise comparisons of candidate reports, which achieved a Pearson correlation of 0.6 ($p < 0.01$) with expert human evaluations.

Our main contributions are as follows:

- (1) We propose a method to extract Amazon professional consultants' expertise from unstructured materials (PowerPoint) and exploit it via few-shot prompting, demonstrating that this **domain knowledge** significantly improves report quality.
- (2) We propose a **versatile multi-agent framework** to automate market analysis. Given high-level request, it produces comprehensive reports without requiring domain expertise, using a collaborative multi-agent process for tasks like SQL querying, data analysis, and report generation.
- (3) We propose an **automated report evaluation framework** of LLMs, showing alignment with human assessments. Based on this, we further propose a LLM-based review mechanism and an approach for selecting optimal reports, enabling **autonomous and iterative improvement** in report generation.

2 RELATED WORK

While Large Language Models (LLMs) have demonstrated remarkable capabilities, certain conventional architectures—particularly those based on token-level, left-to-right sequence generation—present challenges in complex problem solving that require extensive exploration and reasoning. Researchers have addressed this by embedding structured reasoning processes that mirror human cognitive patterns. Chain-of-Thought (CoT) prompting [19] emerged as an early influential research in this field, followed by Tree of Thoughts (ToT) [22], Graph of Thoughts (GoT) [1], and Buffer of Thoughts

(BoT) [21]. These approaches typically decompose complex reasoning into manageable units—such as individual steps of logical reasoning that might generate output of appropriate granularity—that are neither too large nor too small for effective LLM processing, integrating these units into comprehensive problem-solving frameworks [2]. Recent research has also explored ways to augment LLM capabilities and address limitations such as lack of persistent memory, outdated knowledge, and hallucination tendencies [11]. Solutions include Retrieval-Augmented Generation (RAG) [4, 7] for accessing latest information, external tool (e.g., an API to a service) integration [13, 15], and agentic approach architected to perform complex operations requiring decision-making and action [3, 6, 18, 20]. Our research investigates a novel application in this evolving landscape: the automation of market analysis through iterative data exploration with SQL, insight extraction, and comprehensive report creation with visualizations.

Recent research on LLMs for complex data analysis can be categorized into two types: problems with ground-truth solutions and those without objective solutions. Zhang et al. [23] presents *COD-LLM*, a post-trained LLM for data analytics tasks like database table selection, addressing the first category. For tasks without objective solutions, two complementary approaches emerge: combining non-LLM analytical methods with LLMs, and allowing LLMs to generate code dynamically. Ma et al. [10] demonstrates the first approach with *InsightPilot*, using LLMs to guide insight generation with predefined functions (such as trend prediction algorithm) that have been carefully designed. While this reduces hallucinations, it has limitations in flexibility. For the second approach, research explores multi-agent systems for complex tasks [5, 9]. P'erez et al. [14] proposes a multi-agent system consisting of a *Hypothesis Generator* that creates high-level questions, a *Query Agent* that validates them using SQL queries, and a *Summarization* module that consolidates the results into insights.

Our research, while also utilizing a multi-agent system, uniquely leverages domain knowledge from professional business consultants.

3 METHODS

In this section, we formalize MARGEN, the proposed multi-agent framework designed for automatic business analysis. Fig. 1 offers an overview of MARGEN. Given a complex consulting task x , the framework generates a business analysis report y by decomposing the analysis process into multiple subtasks, each handled by specialized agents. The complete MARGEN framework is summarized in Algorithm 1 and elaborated on below. MARGEN consists of four agents: RESEARCHER, WRITER, REVIEWER, and RETRIEVER, each of which will be introduced in detail below.

3.1 Researcher

Motivated by the hypothetical principles of human decision-making and game theory [17], given a prompt including a minimal description of the client and their objectives, the RESEARCHER proposes an initial hypothesis, gathers data to back it up, interprets and summarizes the findings. Its initial prompt (see Appendix B) may include general information about the client and their objectives, as well as constraints to limit the research space. For example, we might want

Algorithm 1 MaRGen Framework

Initialize:

```

1:  $N \leftarrow \text{MaxRounds}$                                 ▶ Max iterations
2:  $t \leftarrow 0$                                           ▶ Current index
3:  $\mathcal{K} \leftarrow \emptyset$                             ▶ External knowledge
4: if Few-Shot then
5:    $\mathcal{K} \leftarrow \text{Retriever}(x)$                     ▶ Retrieve knowledge
6: end if

Researcher:
7: if Few-Shot then
8:    $o_0 \leftarrow p_\theta(o \mid I, \mathcal{K}, x)$                 ▶ Init with knowledge
9: else
10:   $o_0 \leftarrow p_\theta(o \mid I, x)$                     ▶ Random init
11: end if
12:  $q_0 \leftarrow E(o_0); d_0 \leftarrow \text{DB}(q_0)$ 
13: while  $t < N$  and NeedMoreData( $o_t, d_t$ ) do
14:    $o_t \leftarrow p_\theta(o \mid I, \{(o_i, d_i)\}_{i=0}^{t-1})$ 
15:    $q_t \leftarrow E(o_t); d_t \leftarrow \text{DB}(q_t)$ 
16:    $t \leftarrow t + 1$ 
17: end while
18:  $H \leftarrow \{(o_i, d_i)\}_{i=0}^{t-1} \cup \{o_T\}$           ▶ Results
19: Researcher  $\Rightarrow$  Writer: H

Writer:
20:  $R_t^{\text{md}} \leftarrow p_\theta(\cdot \mid H, I_{\text{wr}})$                 ▶ Markdown
21:  $(p_1, \dots, p_k) \leftarrow \text{Extract\_Code}(R_t^{\text{md}})$ 
22:  $\text{Execute}(p_1, \dots, p_k)$                             ▶ Run code
23:  $R_t^{\text{tex}} \leftarrow \text{Convert\_To\_LaTeX}(R_t^{\text{md}})$ 
24:  $R_t^{\text{pdf}} \leftarrow \text{Compile\_PDF}(R_t^{\text{tex}})$ 
25: Writer  $\Rightarrow$  Reviewer: R}_t^{\text{pdf}}

Reviewer:
26:  $F_t \leftarrow p_\theta(F \mid R_t^{\text{pdf}}, I_{\text{rev}})$             ▶ Evaluate
27: if  $F_t < \text{Threshold}$  then
28:    $R_t^{\text{md}} \leftarrow \text{RequestRevision}(F_t, R_t^{\text{pdf}})$ 
29:   Reviewer  $\Rightarrow$  Writer: F}_t
30: else
31:   Approve report
32: end if

Iteration & Termination:
33: while  $t < N$  do
34:   Writer:  $R_t^{\text{md}} \rightarrow R_t^{\text{pdf}}$ 
35:   Reviewer: Evaluate  $F_t$ 
36:   if  $F_t \geq \text{Threshold}$  then
37:     Terminate
38:   end if
39:    $t \leftarrow t + 1$ 
40: end while
41: return  $y \leftarrow R_t^{\text{pdf}}$                             ▶ Final report

```

to focus on a particular period (to reduce the number of rows to scan in the database), limit the number of SQL queries to perform, and rows returned by one query etc. Further, the prompt should include information about the expected formatting of the output (e.g. instruction to surround the final report by the opening and closing tags `<FINAL_ANSWER>` and `</FINAL_ANSWER>` to facilitate its downstream processing (see Appendix B). Given this prompt, the LLM generates an answer, from which we parse an SQL query, execute it and append the result to the next prompt for the LLM to interpret. This process continues until the LLM either has enough

data to provide the final answer or the maximum number of queries is reached.

The RESEARCHER's full pipeline is as follows:

- (1) Propose an initial hypothesis, decide what data are necessary to support the hypothesis, write the first SQL query;
- (2) Execute the query;
- (3) Interpret the result of the query, and update the initial hypothesis as necessary
 - if the number of queries is less than MAX_QUERIES, and if more data needs to be collected, write the next query and go to Step 2;
 - otherwise summarize the data collected so far and return the summary

Formally, the RESEARCHER's pipeline is defined in Eq. 1:

$$o_t \sim p_\theta(\cdot \mid I, \{(o_i, d_i)\}_{i=0}^{t-1}) \quad (1)$$

where p_θ is an LLM, I is the initial prompt (including the system message), $d_i = \text{DB}(E(o_i))$ is the data returned by a SQL query to a database at the i -th time step, $E : o \rightarrow q$ is a script that extracts the SQL query, q , from the LLM's output, o , and $\text{DB} : q \rightarrow d$ is the database that returns data, d , given a query, q .

After the RESEARCHER has finished its work, the entire research history is passed as part of the prompt to the WRITER agent, which we describe next.

3.2 Writer

The WRITER creates the first version of the final report in markdown format (for the full prompt refer to Appendix B).

$$R_0^{\text{md}} \sim p_\theta(\cdot \mid I_{\text{wr}}^{\text{md}}, H) \quad (2)$$

where p_θ is an LLM, $I_{\text{wr}}^{\text{md}}$ is the prompt, including the system message, to generate a report in markdown format, $H = \{(o_i, d_i)\}_{i=0}^{T-1} \oplus \{o_T\}$ is the full history of research produced by the RESEARCHER. In subsequent rounds, the WRITER generates reports conditioned on the feedback from the REVIEWER and its own reports generated previously:

$$R_t^{\text{md}} \sim p_\theta(\cdot \mid I_{\text{wr}}^{\text{md}}, H, \{R_i^{\text{md}}, F_i\}_{i=0}^{t-1}) \quad (3)$$

where R_i^{md} and F_i are a previous markdown report and REVIEWER's feedback, up to the previous step (inclusive). R_i^{md} contains the text of the report and, in their appropriate places, blocks of python code for generating each figure. Before the WRITER can generate latex code, it needs to extract the Python code blocks and run them to generate and save the corresponding figures:

$$(p_1, \dots, p_N) = \text{EF}(R_t^{\text{md}}) \quad (4)$$

where $\{p\}_{i=0}^N$ are python blocks and $\text{EF} : R_t^{\text{md}} \rightarrow \{p\}_{i=0}^N$ is a script that extracts N python blocks from the LLM's output (the markdown text in this case).

Next, the WRITER generates latex code from the markdown document, replacing python code blocks with references to previously rendered figures, as follows:

$$R_t^{\text{tex}} \sim p_\theta(\cdot \mid I_{\text{wr}}^{\text{md}}, H, \{R_i^{\text{md}}, F_i\}_{i=0}^{t-1}, R_t^{\text{md}}) \quad (5)$$

where R_t^{tex} is latex code corresponding to R_t^{md} and $I_{\text{wr}}^{\text{md}}$ is the prompt, which includes a concise style template, to convert it into latex code, which is rendered into a PDF document.

In all rounds except the first and last, before submitting the PDF rendered from R_t^{tex} for review, the WRITER additionally generates a response to the REVIEWER, detailing what has been done specifically to address the REVIEWER's concerns and comments:

$$A_t \sim p_{\theta}(\cdot | I_{\text{ans}}^{\text{md}}, H, \{R_i^{\text{md}}, F_i\}_{i=0}^{t-1}, R_t^{\text{pdf}}) \quad (6)$$

Note that the LLM cannot directly process PDF documents, so we convert each PDF report into a sequence of images, R_t^{pdf} , one image (600 by 600 pixels) for each page of the PDF. On receiving the review (including the scores), the WRITER implements the changes suggested by the REVIEWER by generating an improved report. This process is repeated for several rounds or until either the draft gets the perfect score of 10 on both "clarity" and "layout", or a maximum number of rounds is reached. The WRITER's pipeline can be summarized as follows:

- (1) Generate a report in markdown format, implementing the edits suggested by the reviewer (except in Round 0), with blocks of Python code inserted in the appropriate sections;
- (2) Extract the python code blocks, run them and save the figures as images files;
- (3) Based on the markdown file, generate latex code, inserting references to the image files and compile to a PDF document;
- (4) Send the PDF to the REVIEWER;
- (5) Receive the feedback and scores from the REVIEWER; exit if perfect score or after 4 rounds, otherwise go back to Step 1.

3.3 Reviewer

The REVIEWER agent reviews and scores the PDF report on a scale of 1 to 10 (where 1 is poor and 10 is excellent) written by the WRITER based on the following two criteria: *clarity* and *layout* (the full prompt is provided in Appendix B.3) and returns the review and scores back to the WRITER for generating an improved version of the report in the next round.

$$F_t \sim p_{\theta}(\cdot | I_{\text{rev}}, \{F_i, A_i\}_{i=0}^{t-1}, R_t^{\text{pdf}}) \quad (7)$$

where p_{θ} is an LLM, I_{rev} is the prompt, including the system message, to review the latest report R_t^{pdf} given the full history (if any) of previous reviews and WRITER's replies to the REVIEWER, $\{F_i, A_i\}_{i=0}^{t-1}$, up to the previous round (inclusive). Note that in order to allow the REVIEWER to evaluate the overall appearance of the document (e.g. element layout, font and color choices), the latest PDF report is fed to the LLM in an image format.

3.4 Retriever

The content produced by the RESEARCHER agent is based solely on historical data and its own (largely unconstrained) reasoning about these data. However, we could improve report quality by aligning the agent's reasoning with human domain experts. To achieve this, we propose a RETRIEVER agent that extracts examples of professional reasoning from Amazon consultants' materials (PowerPoint). These examples are then inserted as few-shot examples in the RESEARCHER's prompt.

Among professional business consultants' methodologies, Barbara Minto's Pyramid Principle [12] is a key framework. This approach places the conclusion (hypothesis) at the top, with supporting evidence arranged hierarchically below for logical coherence. Such hierarchical relationships can be represented as a tree structure, which typically has a limited depth. Thus, from consulting materials, we extract consultants' domain knowledge by deriving tree structures with a depth of 2. For a conceptual example of consultants' PowerPoint slide, see Figure 2. For a conceptual example tree derived from consultants' materials, see Figure 3. Please note that the original PowerPoint material created by consultants consists of multiple slides containing various figures and texts, and is an unstructured document. RETRIEVER agent extracts hypothesis trees like the one shown in Figure 3 from this unstructured data, inserting them as few-shot examples in the RESEARCHER's prompt.

Market trend based on customer segment analysis

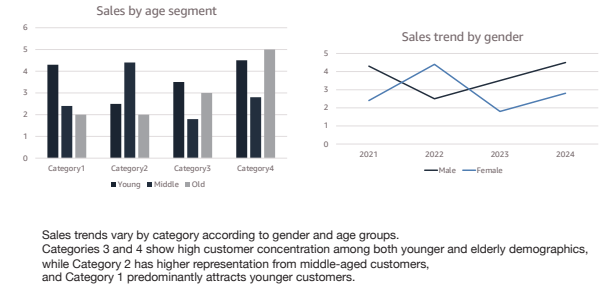


Figure 2: Example of a conceptual consultant's PowerPoint slide.

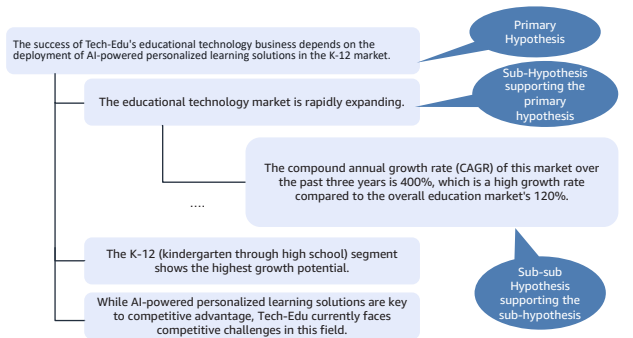


Figure 3: Example of a conceptual hypothesis tree structure extracted from an actual consultant's PowerPoint slide.

Specifically, the root node (primary hypothesis) represents high-level strategic decisions or judgments, the middle nodes (sub-hypotheses) represent supporting arguments that justify the root node, and the leaf nodes (sub-sub hypotheses) represent specific data-based evidence supporting the middle nodes.

Formally, assume that for a given client there is a collection of n documents $\mathcal{D} = \{d_1, d_2, \dots, d_n\}$, where each d_i represents a multimodal document containing both text and images. We first utilize a multi-modal embedding model to construct corresponding retrieval vectors $\{\mathbf{v}_i\}$:

$$\mathcal{K} = \{\mathbf{v}_i \sim \text{Em}(d_i) \mid d_i \in \mathcal{D}\}, \quad (8)$$

We leverage multi-modal conversations to extract hypotheses from documents, represented as $\mathcal{H} = \{h_1, h_2, \dots, h_m\}$, where each h_i denotes a distinct hypothesis. To improve the factual grounding of these hypotheses, we integrate the knowledge source, \mathcal{K} , using RAG [8] for fact-based augmentation. To further refine the structure of \mathcal{H} , the RETRIEVER clusters hypotheses based on their semantic and causal relationships, constructing a hierarchical tree structure \mathcal{H}_t . Within this tree, each node h_t undergoes additional fact augmentation via RAG to ensure consistency and factual completeness. The entire process is formally represented as:

$$\mathcal{H}_t = \left\{ \text{RAG}(h_t, \mathcal{K}) \mid h_t \in \text{T} \left(\bigcup_{i=1}^n \text{RAG}(\text{E}(d_i), \mathcal{K}) \right) \right\}, \quad (9)$$

\mathcal{H}_t denotes the final set of hypothesis tree. E and T denote the operations of extraction and tree structuring, respectively. $\text{RAG}(h_t, \mathcal{K})$ denotes the operation of retrieval augmented generation, which will return top 5 related vectors from \mathcal{K} . The prompts of all operations can be found at Appendix B.5

While it is desirable to utilize past materials from various client companies, not just those from the target client company, hypotheses from different industries and different challenges may not be particularly relevant for building hypotheses for the target client company. Therefore, we construct target hypotheses (hypotheses that apply to the target client company) from source hypotheses (hypotheses from other companies). To this end, given a source hypothesis tree \mathcal{H}_t^s and a target hypothesis tree \mathcal{H}_t^t , our goal is to iteratively transform \mathcal{H}_t^s into a valid \mathcal{H}_t^t . We use \mathcal{H}_t for tree-structured hypotheses, \mathcal{H}_t^s for source hypotheses, \mathcal{H}_t^t for target hypotheses, and $\mathcal{H}_t^{t'}$ for valid target hypotheses. Using a database DB, queries are generated based on \mathcal{H}_t^s and executed to validate \mathcal{H}_t^t . Using the success of the execution of the SQL code as a basis for judgment, this process is repeated until all queries are validated, and a final valid target hypothesis tree is obtained. This process can be denoted as,

$$\mathcal{H}_t^{t'} = \bigcup_{h_t^t \in \mathcal{H}_t^t} \mathcal{U}(h_t^t, \text{EQ}(\text{GQ}(h_t^s, \text{DB}), \text{DB})) \quad (10)$$

where $\text{GQ}(h_t, \text{DB})$ generates a SQL query q for the hypothesis node h_t^s , $\text{EQ}(q, \text{DB})$ executes the query on the database DB and returns the result $r \in \{\text{False}, \text{True}\}$, and $\mathcal{U}(h_t^t, r)$ updates the hypothesis node h_t^t based on the query result r . $\mathcal{H}_t^{t'}$ is used as the few-shot prompt of RESEARCHER. The prompt can be found in Appendix B.5

4 RESULTS

4.1 Data and Models

For all the experimental results reported here we have used a subset of Amazon’s database of orders filled between the beginning of January and middle of November of 2024. The database was hosted on AWS S3 and was accessed via AWS Athena API. To protect privacy, companies and products have been anonymized throughout this paper. All the agents used the same LLM `anthropic.claude-3-5-sonnet-20240620-v1:0` available through the AWS Bedrock service. Also, for the hyperparameters, please see Table 1.

Name	Value
temperature (for report generation)	0.8
temperature (for reviewing and scoring)	0.1
maximum report length	2500
minimum number of SQL queries	4
maximum number of SQL queries	8

Table 1: Hyperparameters.

4.2 Experimental Settings

In this paper we consider two business scenarios: "Sales optimization" and "Developing a product bundling strategy".

Case 1: Sales optimization. Here we consider a hypothetical scenario, where a major company (*Company A*) that sells a diverse range of products through the Amazon marketplace, wants to maximize their market position and boost their sales in a particular product category. They would normally hire a professional consultant, who has access to relevant data sources, to conduct market research, including data collection, analysis and preparation of a comprehensive report with suggestions on how the client can achieve their objectives. However, consultants’ work is usually time-consuming and might not be affordable to some. Here we show that MARGEN can automate the full cycle of consultants’ work: from understanding the client’s needs, to generating initial ideas, executing data queries, interpreting their results, and writing a human-readable report with actionable insights, supported by figures and data tables. We provide complete examples of such reports in Appendix B.5.5.

Case 2: Developing a product bundling strategy. In this scenario MARGEN is tasked to assist *Company X*, an electronics company, to optimize their product bundling strategy, increase average order value and market share. *Company X* also indicated that they are interested in understanding:

- Which products are frequently purchased together
- How different customer segments respond to different bundle combinations
- What price points maximize bundle adoption while maintaining profitability
- How seasonal patterns affect bundle performance
- Whether certain bundles could help them capture market share from competitors

5 ANALYSIS

5.1 Does the REVIEWER Really Helps the WRITER to Improve the Reports?

Yes. We noticed that the initial PDF reports usually have a lot of room for improvement. Specifically,

- Sections tend to be too concise, leaving out important insights found by the RESEARCHER;
- Some product/company names are not spelled in Latin characters;
- Recommendations are scant on specific implementation details;
- Technical terms are not spelled out (e.g what does the 'Personal Protective Equipment' product category include);
- Figures are very basic or not aesthetically pleasing (poor choice of fonts size, overlapping text, or axis labels, too big numbers).

While most of the time the WRITER is able to properly address the REVIEWER's suggestions (especially when it comes to adding new or expanding existing textual content), we also observed failure cases. For example,

- In response of the REVIEWER's suggestion to enlarge a figure for better visibility, the WRITER attempts to increase the resolution at which the figure is saved (using the `dpi` parameter in `plt.savefig` function). While this may improve the visibility, it doesn't directly address the REVIEWER's suggestion.
- The REVIEWER sometimes wrongly suggests making changes to a figure in Section A, while in fact the figure referred to is in Section B (usually a neighboring section).

In general, a brief human inspection of the PDF reports from different rounds, and the correspondence between the WRITER and REVIEWER clearly shows that reports from later rounds are clearly better than from earlier rounds. We note that it takes a maximum of 4 rounds (3 reviews) to reach the perfect score of 10 on both clarity and layout. Fig. 4 suggests that initial reviews already have relatively high scores in the first round of review (around 7 for clarity and 8 for layout).

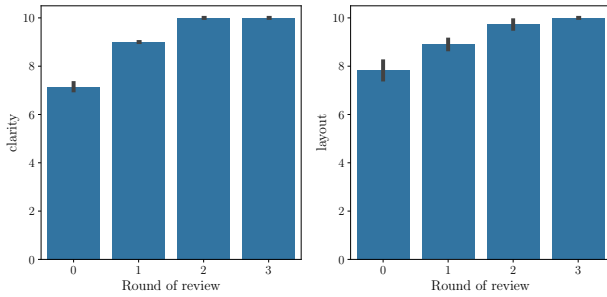


Figure 4: Progressive improvement of scores over consecutive rounds as judged by the Reviewer. Bar height reflects an average over 20 reports. The error bars correspond to 95% confidence intervals.

However, it is possible that the LLM simply increases the score because each new revision produced by the WRITER is *expected* to

be better than the previous one, not because the new revision is objectively better. We have conducted additional experiment in order to rule out this possibility. We run a set of pairwise comparisons between earlier and later versions of each report. Formally, we use an LLM that compares the same report in the i -th and j -th rounds and outputs 1, if the j -th report is better than the i -th:

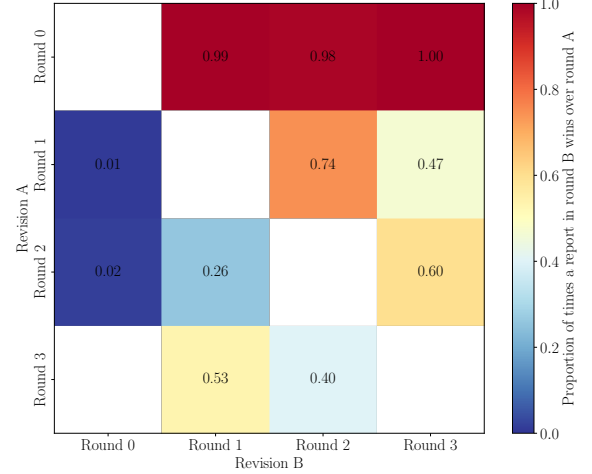


Figure 5: Pairwise comparison of the same report from different rounds of review. The numbers are averaged over 20 different reports.

$$f(R_i^{\text{pdf}}, R_j^{\text{pdf}}) = \begin{cases} 1 & \text{if } R_i^{\text{pdf}} \text{ is worse than } R_j^{\text{pdf}} \\ -1 & \text{if } R_i^{\text{pdf}} \text{ is better than } R_j^{\text{pdf}} \\ 0 & \text{otherwise} \end{cases} \quad (11)$$

$\forall i, j$, s.t. $i < j$ and $i, j \in \{0, \dots, N\}$. Fig. 5 shows that even when the LLM does not "know" which of the two reports correspond to which round, it still consistently judges *later* revisions as better than those from earlier rounds.

5.2 Can an LLM's Scores Be a Trusted Proxy for Human Quality Judgement?

Our goal is to use an LLM for reviewing and scoring the generated reports and using those scores to guide their automatic iterative improvement. Before we can do that, we need to have confidence that the LLM's scores can be trusted as a proxy of human judgement about their quality. The most straightforward approach is *LLM-as-a-judge*, in which each report is evaluated individually, that is *not* relative to the quality of any other report. While Lu et al. [9] claimed that, at least when it comes to evaluating scientific papers, LLM-human agreement was higher than that between humans, Si et al. [16] found that LLMs might still not be well-calibrated for this kind of evaluation of free-form text. For this reason, we also evaluate the reports' quality using a pairwise approach, similarly to Si et al. [16].

Individual scoring. In this type of scoring, each report ($N = 20$) is submitted to the LLM or human scorer and evaluated *independently*, i.e. not relative to any of the other reports. To assist in the scoring process, the scorers (both humans and LLMs) are instructed to provide a numerical answer on a scale of 1 to 10 (where 1 means "strongly disagree" and 10 means "strongly agree") to six statements. The statements are designed to evaluate the report's quality along a particular dimension: (1) non-triviality, (2) degree of justification, (3) clarity, (4) feasibility, (5) balance and (6) overall quality. See Appendix B.3.4 for the full prompt and Table 2 for the scorers' instructions.

Pairwise scoring. Pairwise scoring is similar to Eq. 11, except that instead of comparing different revisions of the same report, we compare *final* versions of different reports (i.e. generated from slightly different research histories, but same prompts). Specifically, given a set of N reports $\{R_i^{\text{pdf}}\}_{i=1}^N$, we prompt the LLM to decide which of the two reports, R_i^{pdf} and R_j^{pdf} is better. To reduce variance, each pair gets to be compared K times. The results of this comparison are recorded to the matrix $M^{N \times N \times K}$, where the entry M_{ijk} represents the result of the comparison defined as follows:

$$M_{ij} = \begin{cases} 1 & \text{if } R_i^{\text{pdf}} \text{ is better than } R_j^{\text{pdf}}, \\ -1 & \text{if } R_j^{\text{pdf}} \text{ is better than } R_i^{\text{pdf}}, \\ 0 & \text{if there is a tie,} \\ 0 & \text{if } i = j \text{ (no comparison with itself).} \end{cases} \quad (12)$$

The scores of each report are obtained by averaging M along the third and then second dimensions. Intuitively, these scores represent the proportion of times in which the i -th report won over the other. Fig. 6 shows that both approaches to scoring reviews mostly agree (with Pearson $r = 0.43$, $p = 0.0585$). The fact that AI scores for the same report obtained using two different automatic reviewing approaches are correlated, suggests that the LLM is relatively consistent in its perception of quality regardless of the method used. However, when using individual evaluation, the scores tend towards the middle of the scale (notice that individual scores are concentrated between 6.5 and 8.25), potentially failing to capture part of the reports' quality variability (Fig. 6). For this reason, unless specified otherwise, we report result obtained using this pairwise evaluation method.

Human scoring. To assess the agreement between the LLM's judgement of report quality ("AI scores") and that of an average human, we asked two human reviewers to score the same 20 reports as those scored by the LLM. Each of the 20 reports was scored once by each reviewer, and the scores were averaged between reviewers.

Fig. 7 shows that the average human scores for overall quality correlate well with pairwise AI scores (but see also correlations for individual AI scores, as well as detailed criterion-wise correlations in Appendix A). This is important, because we can further improve the quality by generating N reports in parallel with the same prompt, and while the reports will be mostly similar, given a reliable automatic reviewer, we can select the best of them as the final one.

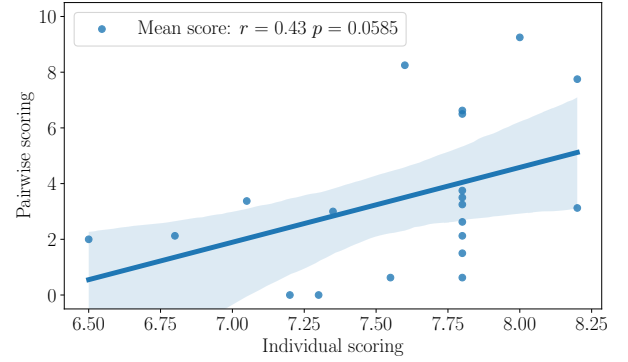


Figure 6: Final report quality scores obtained using LLM-based individual and pairwise scoring are somewhat correlated, but individually assigned scores range between 6.5 and 8.25, potentially failing to capture part of the quality variability.

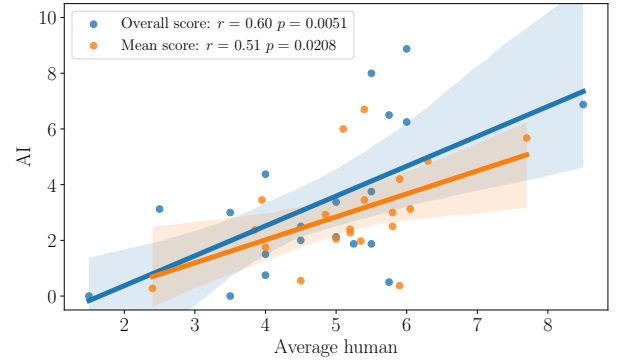


Figure 7: The correlation between average human and AI-generated (pairwise) quality scores for 20 final reports.

5.3 Does the Injection of Expert Knowledge by RETRIEVER Improve the Report Quality?

As mentioned in Section 3.4. In many real-world scenarios, a large number of unstructured expert documents already exist. We aim to fully exploit this wealth of information to enhance the quality and comprehensiveness of the generated reports. We randomly selected five companies, built their own RETRIEVER, and generated 20 reports for each: 10 with added few-shot examples generated by the RETRIEVER and 10 without them. Fig. 8 shows that the proposed RETRIEVER significantly enhances the quality of generated reports and demonstrates strong potential for application in real-world business scenarios.

6 LIMITATION AND FUTURE WORKS

We mention several limitations and future work. First, while we have shown above that LLMs can be a trusted proxy for human

Table 2: Scoring Criteria

Criterion	Description
Non-triviality	The report goes beyond basic analysis, makes non-trivial conclusions/insights
Degree of justification	The conclusions are strongly supported by the data provided
Clarity	The report is well-organized, clear and easy to follow, always highlighting key insights
Feasibility	The recommendations in the report are specific, measurable and feasible
Balance	The report is well balanced, discusses multiple perspectives, and considers possible limitations and risks of following the suggested actions
Overall Score	This overall score is intended to reflect your overall quality judgement of a given report. The idea is that the 5 criteria listed above might not fully capture all the important aspects of quality (such as what weight should be given e.g. to clarity or another criterion). This "overall score" is not necessarily the sum/mean of the five other scores

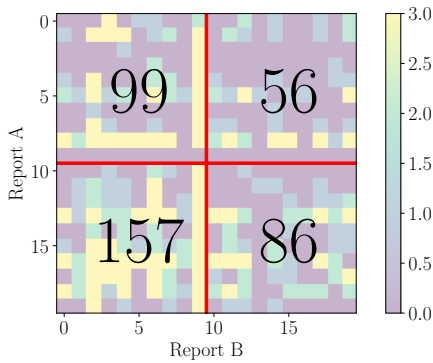


Figure 8: Pairwise comparison of MARGEN reports with and without added examples of expert reasoning. Each pixel shows the number of times report A was judged as better than report B in a pairwise comparison. Reports 0 - 9 and 10-19 were generated without and with the domain knowledge provided by the RETRIEVER, respectively. Reports 10 - 19 (generated with the added few-shot examples generated by the RETRIEVER) are clearly better than reports 0 - 10 (generated without that few-shot examples).

perception of quality, at least on the specific dimensions of quality that we defined for our analysis, and that LLMs can improve the quality of reports by themselves, we have not conducted evaluations on whether the market reports created by MARGEN are equivalent to or better than those created by humans. This point is important future work. Also, regarding whether the data extraction and analysis performed by MARGEN is correctly implemented, we would like to mention that users can verify this because all executed SQL queries and other operations can be checked. We have also conducted preliminary analysis on validating the data provided by MARGEN using another agent called VERIFIER (details can be found in the Appendix B.4). Incorporating a data validation process could lead to building a more reliable system.

7 CONCLUSIONS

In this paper, we introduced MARGEN, a framework for automated market research that leverages LLMs to generate data-driven insights. Our system demonstrates that MARGEN can effectively conduct end-to-end market research, from initial data exploration to producing comprehensive market reports with visualizations. Experimental result showed alignment between expert human and AI quality assessments, suggesting that automated evaluation can serve as a reliable proxy for human judgment, enabling efficient quality control at scale. Additionally, we showed that MARGEN can improve the quality of reports through iterative review and revision processes. Furthermore, we demonstrated that report quality can be improved by utilizing professional consultants' unstructured knowledge documented in PowerPoint through in-context learning. While future work remains, we believe our work represents an important step toward democratizing access to professional-grade market research.

REFERENCES

- [1] Maciej Besta, Nils Blach, Ales Kubicek, Robert Gerstenberger, Michal Podstawski, Lukas Gianinazzi, Joanna Gajda, Tomasz Lehmann, Hubert Niewiadomski, Piotr Nyczyk, et al. 2024. Graph of thoughts: Solving elaborate problems with large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 38. 17682–17690.
- [2] Qiguang Chen, Libo Qin, Jinhao Liu, Dengyun Peng, Jiannan Guan, Peng Wang, Mengkang Hu, Yuhang Zhou, Te Gao, and Wanxiang Che. 2025. Towards reasoning era: A survey of long chain-of-thought for reasoning large language models. *arXiv preprint arXiv:2503.09567* (2025).
- [3] Zane Durante, Qiuyuan Huang, Naoki Wake, Ran Gong, Jae Sung Park, Bidipta Sarkar, Rohan Taori, Yusuke Noda, Demetri Terzopoulos, Yejin Choi, Katsushi Ikeuchi, Hoi Vo, Li Fei-Fei, and Jianfeng Gao. 2024. Agent AI: Surveying the Horizons of Multimodal Interaction. (2024). arXiv:2401.03568 [cs.AI] <https://arxiv.org/abs/2401.03568>
- [4] Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yixin Dai, Jiawei Sun, Haofen Wang, and Haofen Wang. 2023. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997* 2 (2023), 1.
- [5] Juraj Gottweis, Wei-Hung Weng, Alexander Daryin, Tao Tu, Anil Palepu, Petar Sirkovic, Artiom Myaskovsky, Felix Weissenberger, Keran Rong, Ryutarō Tanno, Khaled Saab, Dan Popovici, Jacob Blum, Fan Zhang, Katherine Chou, Avinatan Hassidim, Burak Gokturk, Amin Vahdat, Pushmeet Kohli, Yossi Matias, Andrew Carroll, Kavita Kulkarni, Nenad Tomasev, Yuan Guan, Vikram Dhillon, Eeshit Dhaval Vaishnav, Byron Lee, Tiago R D Costa, José R Penadés, Gary Peltz, Yunhan Xu, Annalisa Pawlosky, Alan Karthikesalingam, and Vivek Nataraajan. 2025. Towards an AI co-scientist. arXiv:2502.18864 [cs.AI] <https://arxiv.org/abs/2502.18864>
- [6] Taicheng Guo, Xiuying Chen, Yaqi Wang, Ruidi Chang, Shichao Pei, Nitesh V Chawla, Olaf Wiest, and Xiangliang Zhang. 2024. Large language model based multi-agents: A survey of progress and challenges. *arXiv preprint arXiv:2402.01680*

- (2024).
- [7] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems* 33 (2020), 9459–9474.
 - [8] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems* 33 (2020), 9459–9474.
 - [9] Chris Lu, Cong Lu, Robert Tjarko Lange, Jakob Foerster, Jeff Clune, and David Ha. 2024. The AI Scientist: Towards Fully Automated Open-Ended Scientific Discovery. arXiv:2408.06292 [cs.AI] <https://arxiv.org/abs/2408.06292>
 - [10] Pingchuan Ma, Rui Ding, Shuai Wang, Shi Han, and Dongmei Zhang. 2023. InsightPilot: An LLM-Empowered Automated Data Exploration System. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, Yansong Feng and Els Lefever (Eds.). Association for Computational Linguistics, Singapore, 346–352. <https://doi.org/10.18653/v1/2023.emnlp-demo.31>
 - [11] Shervin Minaee, Tomas Mikolov, Narjes Nikzad, Meysam Chenaghlu, Richard Socher, Xavier Amatriain, and Jianfeng Gao. 2025. Large Language Models: A Survey. arXiv:2402.06196 [cs.CL] <https://arxiv.org/abs/2402.06196>
 - [12] Barbara Minto. 2008. *The Pyramid Principle: Logic in Writing and Thinking* (3rd ed.). Financial Times Prentice Hall, London.
 - [13] Bhargavi Paranjape, Scott Lundberg, Sameer Singh, Hannaneh Hajishirzi, Luke Zettlemoyer, and Marco Tulio Ribeiro. 2023. ART: Automatic multi-step reasoning and tool-use for large language models. (2023). arXiv:2303.09014 [cs.CL] <https://arxiv.org/abs/2303.09014>
 - [14] Alberto Sánchez Pérez, Alaa Boukhary, Paolo Papotti, Luis Castejón Lozano, and Adam Elwood. 2025. An LLM-Based Approach for Insight Generation in Data Analysis. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, Luis Chiruzzo, Alan Ritter, and Lu Wang (Eds.). Association for Computational Linguistics, Albuquerque, New Mexico, 562–582. <https://aclanthology.org/2025.naacl-long.24/>
 - [15] Timo Schick, Jane Dwivedi-Yu, Roberto Dessi, Roberta Raileanu, Maria Lomeli, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. 2023. Toolformer: Language Models Can Teach Themselves to Use Tools. (2023). arXiv:2302.04761 [cs.CL] <https://arxiv.org/abs/2302.04761>
 - [16] Chenglei Si, Diyi Yang, and Tatsunori Hashimoto. 2024. Can LLMs Generate Novel Research Ideas? A Large-Scale Human Study with 100+ NLP Researchers. arXiv:2409.04109 [cs.CL] <https://arxiv.org/abs/2409.04109>
 - [17] Herbert Alexander Simon. 1957. *Models of man: social and rational; mathematical essays on rational human behavior in society setting*. New York: Wiley.
 - [18] Lei Wang, Chen Ma, Xueyang Feng, Zeyu Zhang, Hao Yang, Jingsen Zhang, Zhiyuan Chen, Jiakai Tang, Xu Chen, Yankai Lin, Wayne Xin Zhao, Zhewei Wei, and Jirong Wen. 2024. A survey on large language model based autonomous agents. *Frontiers of Computer Science* 18, 6 (March 2024). <https://doi.org/10.1007/s11704-024-40231-1>
 - [19] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in neural information processing systems*, Vol. 35. 24824–24837.
 - [20] Zhiheng Xi, Wenxiang Chen, Xin Guo, Wei He, Yiwen Ding, Boyang Hong, Ming Zhang, Junzhe Wang, Senjie Jin, Enyu Zhou, et al. 2025. The rise and potential of large language model based agents: A survey. *Science China Information Sciences* 68, 2 (2025), 121101.
 - [21] Ling Yang, Zhaochen Yu, Tianjun Zhang, Shiyi Cao, Minkai Xu, Wentao Zhang, Joseph E Gonzalez, and Bin Cui. 2024. Buffer of thoughts: Thought-augmented reasoning with large language models. *Advances in Neural Information Processing Systems* 37 (2024), 113519–113544.
 - [22] Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. 2023. Tree of thoughts: Deliberate problem solving with large language models. *Advances in neural information processing systems* 36 (2023), 11809–11822.
 - [23] Jiani Zhang, Hengrui Zhang, Rishav Chakravarti, Yiqun Hu, Patrick Ng, Asterios Katsifodimos, Huzefa Rangwala, George Karypis, and Alon Halevy. 2025. CoddLLM: Empowering Large Language Models for Data Analytics. arXiv:2502.00329 [cs.DB] <https://arxiv.org/abs/2502.00329>

APPENDIX

A SUPPLEMENTARY FIGURES

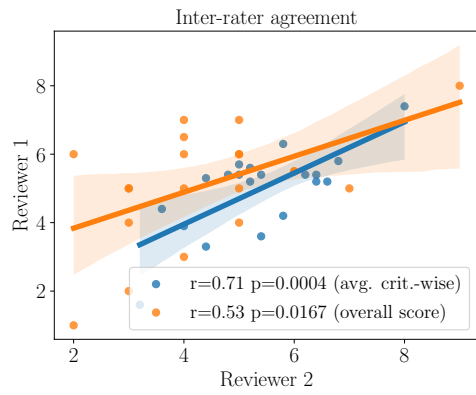


Figure 9: Inter-rater agreement. The orange line and data points correspond to overall scores (which are not necessarily the average of the other five criteria (see Appendix ??)). The blue line and data points correspond to the simple averages of the 5 criteria (NT, DJ, CL, FE and BA).

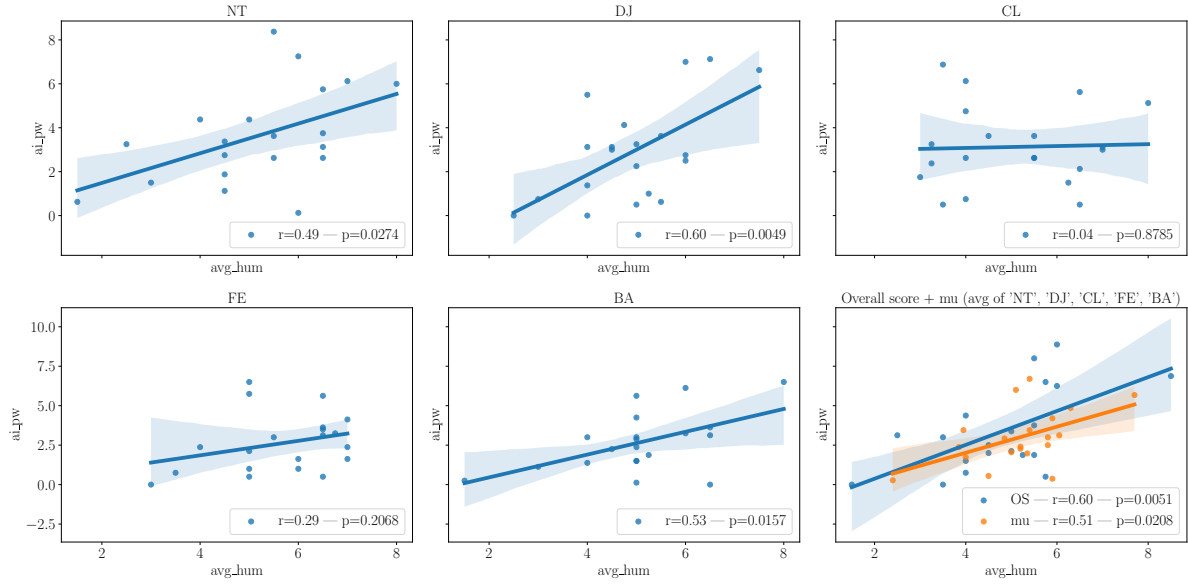


Figure 10: Analysis of the agreement between human and AI-generated (pairwise) quality scores. Each panel shows the correlation for one of the criteria. The bottom right panel additionally shows the correlation between human and the mean of the AI-generated scores on the following five criteria: NT, DJ, CL, FE and BA.

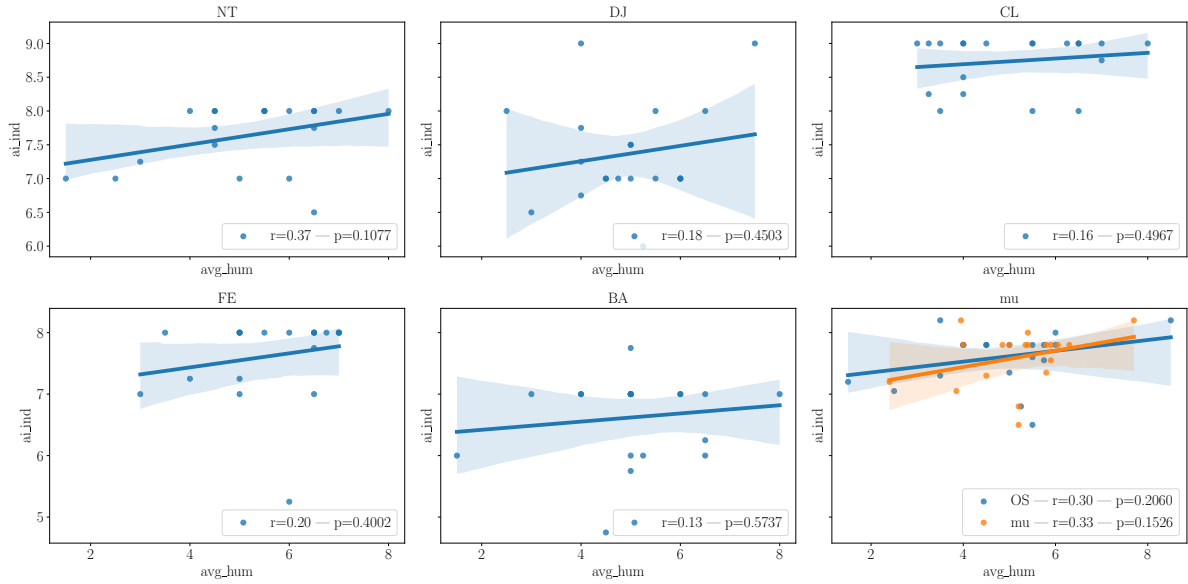


Figure 11: Analysis of the agreement between human and AI-generated (individual) quality scores. Each panel shows the correlation for one of the criteria. The bottom right panel additionally shows the correlation between human and the mean of the AI-generated scores on the following five criteria: NT, DJ, CL, FE and BA.

B PROMPTS

Product names and other information have been anonymized as appropriate.

B.1 Researcher's Prompt Template

The template below accepts 3 string parameters: CLIENT_COMPANY, CLIENT_PROVIDED_INFO, DB_SCHEMA, MAX_QUERIES, and MIN_QUERIES.

For Case Study 1 they were as follows:

```
CLIENT_COMPANY = "Client 123456789"
CLIENT_PROVIDED_INFO = "
- the client is looking to optimize their sales;
- the client has suggested that they want to become the leader in the category 'category A'."
```

```
MAX_QUERIES = 4
MAX_QUERIES = 8
```

Situation

```
Your client is {CLIENT_COMPANY}.
In your initial discussion with the client,
you have established the following:
{CLIENT_PROVIDED_INFO}
```

Task

You task is to conduct research, within the constraints listed, to help your customer achieve their goals. Specifically you must

- 1) come up with an initial hypothesis for
 - what actions you client can take to optimally achieve their goals;
 - what might be the possible risks or tradeoffs associated with following your suggestions
- 2) obtain relevant data and see if they support your hypothesis; if the data does not support it, adjust or change your hypothesis and validate it again against data.

Data Sources

You have access to a database table, which you can query it using standard SQL syntax. The database table has the following schema:

```
{DB_SCHEMA}
```

Constraints

1) When (and if) you need to get relevant data, write your SQL query surrounded with triple backticks, for example:

```
```
{EXAMPLE_SQL}
```
```

- 2) For all your queries, only request data from 2024.
- 3) Do not write SQL queries that are likely to return more than 200 rows, and thus use aggregation operations as appropriate.
- 4) Never generate more than one SQL query at a time and never proceed to further reasoning or analysis before the result of the query is returned to you.
- 5) If you are given any names (company, product etc.), always use them exactly, or your queries will return no results.
- 6) You must complete your research using no more than {MAX_QUERIES} (but no fewer than {MIN_QUERIES}) queries.
- 7) To select rows from 2024, always use the more efficient `WHERE EXTRACT(YEAR FROM COLUMN_X) = 2024`, not `WHERE DATE_FORMAT(COLUMN_X, '%Y') = '2024'`.
- 8) The actions in your final report must not require any further research or data analysis.

Additional Instructions

Each time you make a query, I will return its result to you in JSON format for your analysis and consideration. Except for the constraints given, you have unlimited creative freedom to tackle this task. If the retrieved data does not support the initial hypothesis, revise it and continue your research.

Expected Result

When your proposed hypothesis is strong enough, write an exceptionally well-written professional report of no longer than {REPORT_MAX_WRDS} words.

This report must be surrounded by the <FINAL_ANSWER> and </FINAL_ANSWER> tags, for example:

<FINAL_ANSWER>

Your final answer goes here.

</FINAL_ANSWER>

The report MUST cover the following points:

1. Executive Summary

- Provide a concise overview of key findings and recommendations
- Highlight 3-5 main points that capture the essence of the report

2. Introduction

- State the purpose of the research
- Outline the scope and objectives
- Briefly describe the methodology used

3. Market Overview

- Analyze the current state of the market
- Identify key players and their market shares
- Discuss market size, growth rates, and trends

4. Competitive Analysis

- Identify main competitors
- Analyze their strengths and weaknesses
- Compare the client's position relative to competitors

5. Customer Analysis

- Define target customer segments
- Analyze customer behavior, preferences, and needs

6. Product Analysis

- Evaluate current product offerings
- Analyze performance of different product lines or segments
- Identify areas for improvement or expansion

7. Pricing Analysis

- Analyze the customer's current pricing strategies in comparison with those of competitors

- Identify optimal price points for different customer segments

9. SWOT Analysis (note: include only when you have the data for this)

- Strengths: Internal positive attributes
- Weaknesses: Internal areas for improvement
- Opportunities: External factors that could benefit the business
- Threats: External factors that could harm the business

10. Recommendations

- Provide actionable strategies based on findings
- Include short-term and long-term suggestions

11. Implementation Plan

- Outline steps to implement recommendations
- Provide timeline and milestones
- Identify potential challenges and mitigation strategies

12. Financial Projections

- Estimate potential impact of recommendations
- Provide ROI analysis where applicable

13. Conclusion

- Summarize the main points
- Highlight critical issues and opportunities
- Reinforce the value of implementing recommendations

And adhere to the following Best Practices:

1. Use a mix of quantitative and qualitative research methods
 - Analyze sales/orders data, market trends, and customer behavior
2. Segment analysis where appropriate

- Break down findings by product categories, customer segments, or other relevant factors
3. Use clear, concise language
 - Avoid jargon unless necessary
 - Define technical terms when used
 4. Be specific and actionable in recommendations
 - Provide clear next steps
 - Explain expected outcomes of implementing recommendations
 5. Use a logical flow
 - Ensure each section builds on previous ones
 - Use transitions to connect different parts of the report
 6. Incorporate client's perspective
 - Address specific concerns or questions raised by the client
 - Align recommendations with client's goals and capabilities

B.2 Writer's Prompt

B.2.1 Writer's Initial Prompt Template. This prompt template is used to build the initial prompt for the Writer, i.e. when no review has been received yet.

```
{history}
```

What you see above is the full history of a chat between and an AI assistant and a human . The AI assistant has conducted market research and analysed the results. Based on this history, write a concise, but professional, consulting report in markdown format.

In this new report, each finding must be illustrated by a corresponding figure (one figure per SQL query response), inserted in its appropriate place. The png files for figures don't exist yet, so instead of adding a link to the image file in the markdown figure, you need to insert python code blocks, one for each figure. Each code block must appear in its appropriate place in the report and be sufficient to generate the corresponding figure and save it in png format. Each code block must be followed by corresponding figure caption concisely explaining what the (future) figure will show. Do not add titles to the figures. Do not use `ha` as a keyword argument for `tick_params`. Avoid using pandas. Each python code block must start with the following lines:

```
import matplotlib.pyplot as plt
import numpy as np
plt.rcParams["font.family"] = "Arial Unicode MS" # (!!!) needed to properly render Japanese characters
```

and end with the following two lines:

```
plt.tight_layout()
plt.savefig('fig_ID.png')
```

where ID is the ordinal number (integer) of the figure in the report.

B.2.2 Writer's Later Prompt Template. This prompt template is used to build the later prompt for the Writer.

Thank you. A professional editor at Nature Publishing Group has reviewed the document you created.
Here's their review:

```
<REVIEW_ROUND_{round}>
{review}
</REVIEW_ROUND_{round}>
```

Please address these reviewer's concerns and comments by editing your previous markdown and python code in it.

First generate a new markdown using the previous instructions. In the future, you will be prompted to generate the corresponding new latex version.

B.2.3 Writer's Response to Reviewer Template. After generating an improved version of the report (in markdown), which addresses the Reviewer's comments and suggestions, the Writer is prompted to write a brief response to the Reviewer:

Now write a short response to the reviewer describing the specific things you did to address their concerns and comments. The part of your answer that will be sent back to the reviewer must be surrounded by `<RESPONSE_TO_REVIEWER>` and `</RESPONSE_TO_REVIEWER>`, for example:

```
<RESPONSE_TO_REVIEWER>
List the things you did.
</RESPONSE_TO_REVIEWER>
```

B.2.4 Writer's Figure Template. This prompt template is used to build figure.

```
\begin{figure}[h!]
\centering
\includegraphics[width=0.8\textwidth] {fig_1.png}
\caption{\label{fig:1} Caption for Figure 1 from the markdown document.}
\end{figure}
```

B.2.5 WRITER's Latex Generation Template. After generating the report in markdown, the Writer is prompted to generate a latex version of it:

Now convert this report into latex code using the latex template (see below), but keep the structure and content exactly the same as in the markdown document. Importantly, replace the python blocks with latex code for inserting the corresponding figure. For example for Figure 1, the latex code should be as follows:

```
{figure_example}
```

Here's the latex template:

```
{latex_template}
```

In your answer, the latex code of the entire report must be written in one block and surrounded with triple backticks (`````). Also, be sure to escape the ampersand (`&`) with a backslash or the latex will not render (for example, instead of ``R&D`` and similar, write ``R\&D``).

B.3 REVIEWER's Prompts

B.3.1 REVIEWER's System Message. This is a system message.

You are a highly experienced marketing consultant with decades of professional experience. You teach consulting at the Harvard Business School.

B.3.2 Reviewer's Initial Prompt Template. This prompt template is used only for the first round.

Your task is to review and score the attached report thoroughly, paying special attention to whether the findings are presented clearly and the visuals are easy to understand and support the narrative. Write your brief feedback (text of the review) covering specific improvement suggestions. The scores must reflect the quality of the report on "clarity" and "layout" (on a 10-point integer scale, where 1 is bad and 10 is perfect), be surrounded by triple backticks, and strictly follow the format below:

```
```json
{{
 "clarity": INT,
 "layout": INT
}}
```
```

Here are some questions to help you guide (but not limit or otherwise restrict) your review:

Did the authors follow the best practices, specifically:

1. Did they use a mix of quantitative and qualitative research methods?
 - Did they analyze sales/orders data, market trends, and customer behavior?
2. Did they segment analysis where appropriate?

- Did they break down findings by product categories, customer segments, or other relevant factors?
3. Did they use clear, concise language
 - avoiding jargon unless necessary?
 - defining technical terms when used?
 4. Were they specific and actionable in their recommendations?
 - Did they provide clear next steps?
 - Did they explain expected outcomes of implementing recommendations?
 5. Did they use a logical flow
 - ensuring each section builds on previous ones?
 - using transitions to connect different parts of the report?
 6. Did they incorporate client's perspective?
 - addressing specific concerns or questions raised by the client?
 - aligning recommendations with client's goals and capabilities?

B.3.3 REVIEWER's Later Prompt Template. This template is used for all versions except the first one. Note that only the WRITER's response to the REVIEWER is pasted into the template, while the updated report is added to the prompt as a sequence of images.

The creator of the report has addressed your comments and suggestions and sent you its updated version. They also added some comments to highlight what they changed:

```
<CREATOR_RESPONSE>
{response}
</CREATOR_RESPONSE>
```

Please review the report again, paying special attention to whether your suggestions and comments have indeed been addressed satisfactorily. If they have been addressed insufficiently, point this out in your review. If your previous suggestions have been addressed partially, consider raising your score proportionately to the degree and quality of improvements. Your review and scores must follow the same format as before.

B.3.4 Prompt For Scoring Reports Individually. This prompt is used to score reports individually.

You are given images of the final PDF report written by a consultant at Amazon. Before writing their report, the consultant received the following instructions:

```
<CONSULTANTS_INSTRUCTIONS>
{creators_prompt}
</CONSULTANTS_INSTRUCTIONS>
```

Following these instructions, the consultant did some research, the full history of which is below:

```
<CONSULTANTS_REPORT>
{consultants_report}
</CONSULTANTS_REPORT>
```

Read the report carefully and fill out the following questionnaire in JSON format as illustrated below.

```
```json
{{
 "The report goes beyond basic analysis, makes non-trivial conclusions/insights.": INT,
 "The conclusions are strongly supported by the data provided.": INT,
 "The report is well-organized, clear and easy to follow, always highlighting key insights.": INT,
 "The recommendations in the report are specific, measurable and feasible.": INT,
 "The report is well balanced, discusses multiple perspectives, and considers possible limitations and risks of following the suggested actions.": INT
}}
```

As you do so, approach scoring the report with an unbiased perspective, while holding it to an exceptionally high standard of quality. Your answers must be on a 10-point integer scale from 1 (strongly disagree) to 10 (strongly agree).

**B.3.5 Prompt For Pairwise Scoring of Reports.** This prompt is used to evaluate reports pairwise.

You are given two reports independently written by two different consultants (one consultant - one report). These consultants had exactly the same background information, instructions and access to the same data. Before writing their reports, the consultants received the following instructions:

```
<CONSULTANTS_INSTRUCTIONS>
{creators_prompt}
</CONSULTANTS_INSTRUCTIONS>
```

The two reports for you to judge are as follows:

```
<REPORT_A>
{report_A}
</REPORT_A>
```

```
<REPORT_B>
{report_B}
</REPORT_B>
```

Using the specific instructions and general guidelines below, judge which report is better (i.e. is the winner). If both of them are equally good (or bad), the outcome of our judgement is a "DRAW".

**\*\*Specific Instructions\*\*:**

1. Check for alignment with the original problem description:
  - Does the report directly address the client's objectives and needs?
  - Are all key questions answered?
2. Assess the methodology:
  - Are the research methods used appropriate?
  - Are the sample sizes sufficient?
3. Examine data analysis:
  - Does the report contain appropriate statistical analyses?
  - Are the conclusions strongly supported by the data?
  - Are there any potential biases in the analysis?
  - Are the conclusions sufficiently non-trivial?

- Is the report Are the conclusions sufficiently non-trivial?

4. Review the presentation of findings:
  - Does the report present information clearly and in a logical order?
  - Are the key insights clearly highlighted and easy to understand?
  - Does the report discuss the limitations in the analysis?

5. Evaluate the actionability of recommendations:

- Are the recommendations specific, measurable, and feasible?
- Are the connections between findings and recommendations clear?

As you write up your review, follow the general guide lines below:

**\*\*General Guidelines\*\*:**

1. Maintain objectivity:
  - Approach each report with an unbiased perspective
  - Focus on the quality of work rather than what might be perceived as personal preferences
2. Consider the target audience:
  - Evaluate if the report is tailored to the client's level of expertise (very basic understanding of marketing, data science)
  - Assess the overall readability and accessibility of the content
3. Look for innovation and creativity:
  - Identify unique approaches or insights that set reports apart
  - Recognize consultants who go beyond basic analysis
4. Assess the overall professionalism:
  - Evaluate the overall polish and presentation of the report
5. Provide constructive feedback:
  - Highlight strengths as well as weaknesses in each report
6. Consider the client's perspective:

- Assess how well each report addresses the client's specific needs
- Evaluate the potential impact of findings on the client's business

Start by writing a review assessing each of the reports using the instructions and guidelines above. Finish your answer with your final judgement, which must be in the following format:

```
```json
{"winner": "REPORT_A" OR "REPORT_B" OR "DRAW"}
```
```

## B.4 Verifier's Prompt

This template is used to verify data analysis.

### # Situation

Another consultant has conducted market research and written a report (see images or each page and hi-resolution images for each figure in it above). The client, for whom the report was written, has voiced concerns that some data presented in the report may not be entirely accurate.

For your reference, you are also given the latex version of the report:

```
<REPORT_LATEX>
{latex}
</REPORT_LATEX>
```

### # Task

Your task is to double-check all the numerical data contained in the given report.

### # Data Sources

Just like the other consultant, you have access to a database table, which you can query using standard SQL syntax (Presto SQL used in AWS Athena). The database table has the following schema:

```
{DB_SCHEMA}
```

For your reference, here's the full history of the other consultant's research, where you can see what SQL queries he made. If you don't see anything wrong with those queries, you can use the same or similar SQL queries to double-check the results that the consultant's report contains.

```
<FULL_RESEARCH_HISTORY>
{final}
</FULL_RESEARCH_HISTORY>
```

### # Constraints

1) When (and if) you need to get relevant data, write your SQL query surrounded with triple backticks, for example:

```
...
{EXAMPLE_SQL}
...
```

- 2) For all your queries, only request data from 2024.
- 3) Do not write SQL queries that are likely to return more than 200 rows, and thus use aggregation operations as appropriate.
- 4) Never generate more than one SQL query at a time and never proceed to further reasoning or analysis before the result of the query is returned to you.
- 5) If there are any names (company, product etc.), always use them exactly, or your queries will return no results.
- 6) To select rows from 2024, always use the more efficient ``WHERE EXTRACT(YEAR FROM COLUMN_X) = 2024``, not ``WHERE DATE_FORMAT(COLUMN_X, '%Y') = '2024'``.
- 7) Do not offer any analysis, judgement or review of the quality or the report. Only look for inaccuracies and report them (if any).
- 8) When your queries contain calculations, cast all numbers to `DOUBLE`, and round the result to 4 decimal places, for example:
 

```
`ROUND(CAST(SUM(COLUMN_X) AS DOUBLE) / NULLIF(CAST(SUM(SUM(COLUMN_X)) OVER () AS DOUBLE), 0) * 100.0 AS market_share_percentage, 4)`
```

### # Additional Instructions

Each time you make a query, I will return its result to you in JSON format for your analysis and consideration.  
Each time you get the data from me, if there is a discrepancy between the number(s) I return and the number(s) you see in the other consultant's report, write:

The [previous consultant said the DATA\_POINT\_NAME was CONSULTANT\_VALUE, but according to your data it is actually ACTUAL\_VALUE. (replace DATA\_POINT\_NAME, CONSULTANT\_VALUE and ACTUAL\_VALUE with the corresponding name and values).

For example:

"Previous consultant the total market size was 3.1T yen, but according to your data it is actually 2.1B yen."

Except for the constraints given, you have unlimited creative freedom to tackle this task.

### # Expected Result

When you have checked all the numerical data, concisely list the specific errors and /or inaccuracies you have found (if any), also mentioning their known (or likely) causes. At the very end, summarize the incorrectly reported data in a markdown table as follows:

| data         | reported | actual |
|--------------|----------|--------|
| sales        | 3000     | 4000   |
| ...          | ...      | ...    |
| Client share | 30%      | 34%    |

## B.5 Retriever's Prompt

This template is used for the Retriever.

**B.5.1 Extraction.1.** This template is used to extract relevant questions and hypotheses from images.

Now you're a consultant who consults for retail companies. Analyze the provided images and extract relevant questions and hypotheses.

Instructions:

0. Even if the content is Japanese, try to understand and analyze it.
1. For each image, identify the key questions and hypotheses that can be extracted from the content.
2. If an image does not contain any useful information for companies, simply return "None" for that image.
3. Structure your output as follows:

```
[{
 "question": ,
 "hypothesis":
}]
```

every Questions follows one Hypotheses. The output should be provided in English.

**\*\*IMPORTANT\*\***

Just output the json format result directly.

DO NOT add additional explanations or introduction in the answer unless you are asked to. Make sure the only keys in the JSON object are "question", "hypothesis".

**B.5.2 Extraction.2.** This template is used to extract key questions and hypotheses.

You are an experienced retail consultant specializing in data analysis to extract key questions and hypotheses.

Please carefully review the provided data and perform the following tasks:

1. Filter irrelevant content: Remove any question and hypothesis unrelated to professional consulting tasks, or strategic planning.
2. Refine questions and hypotheses: Adjust the phrasing of each pair to ensure they are logical, precise, and presented in a professional tone.
3. Merge and consolidate: For pairs with overlapping or similar themes, merge and rewrite them to create more comprehensive and cohesive question-hypothesis pairs, maintaining a professional tone throughout.

4. Output structure: Present the final result in the following JSON format.

Here is an example:

```
{
 "question": "What are the revenue growth trends for business and outdoor bags over the past decade?",
 "hypothesis": "Over the past decade, both business and outdoor bags have experienced growth, but with distinct patterns. business bags have shown steady, consistent growth, while outdoor bags have demonstrated rapid acceleration in recent years. This trend suggests a shifting market dynamic where business bags maintains stable demand, and outdoor bags are gaining significant market share, potentially narrowing the gap between the two categories."
},
```

**\*\*IMPORTANT\*\***

Just output the result in JSON format directly. DO NOT add additional explanations or introduction in the answer unless you are asked to! Make sure the only keys are "question" and "hypothesis".

**B.5.3 RAG.1.** This template is used to enhance questions and hypotheses through RAG.

You are an expert retail data analyst specializing in hypothesis refinement based on accurate data and retrieved information. Your task is to critically evaluate and enhance the given hypothesis.

Input:

Question: {question}

Hypothesis: {hypothesis}

Instructions:

1. Carefully analyze the provided question, initial hypothesis, and retrieved context.
2. Based on the retrieved information, refine the hypothesis to:
  - Correct any inaccuracies or misinterpretations in the original hypothesis
  - Incorporate specific numerical data and evidence from the context

- Ensure direct relevance to the question
- Improve precision and logical structure
- Maintain professional wording

3. Focus on adding concrete details, statistics, and factual evidence to support the refined hypothesis.

Output:

Provide only the refined hypothesis with same format

**IMPORTANT:** output the json format result directly. DO NOT add additional explanations or introduction in the answer unless you are asked to. Prioritize accuracy and the inclusion of specific data. Ensure all statements are supported by the retrieved context. Output only the refined hypothesis as specified above.

**B.5.4 Clustering.** This template is used to cluster similar hypotheses and construct the final hypothesis tree.

I will provide you with a list of hypotheses related to a specific supplier. Please complete the following steps to reorganize the information into a decision-making structure:

Merge semantically similar hypotheses:

Identify hypotheses that are semantically similar and merge them into a single, coherent statement.

Ensure that no original details or information are lost during merging. Retain all relevant specifics from the merged hypotheses in the output.

Each merged hypothesis should form a concise yet complete decision or judgment statement, avoiding vague or unverifiable terms. Reorganize hypotheses into a multi-layer, decision-oriented tree structure:

The root nodes (labeled as "hypotheses") should represent high-level strategic decisions or judgments derived from the information provided. These should directly answer key questions such as:

"What is the recommended next step based on the evidence?"

"Should we engage with this supplier?"

"What are the major risks or opportunities?"

Intermediate nodes (labeled as "subhypotheses") should serve as supporting arguments that justify or explain the decisions in the root nodes. Each subhypothesis should break down how specific aspects (e.g., performance, reliability, cost) support the high-level decisions, ensuring they remain actionable and decision-focused.

Leaf nodes (labeled as "subsubhypotheses") should contain specific evidence such as numerical data, detailed analyses, supplier behavior, or contract terms that provide concrete backing for the subhypotheses. Each leaf should directly tie back to the decision-making context.

Output the final hypothesis tree in JSON format:

Ensure each node is categorized as "hypotheses", "subhypotheses", or "subsubhypotheses" based on its role in the decision-making process. Maintain logical relationships and strategic inferences between nodes, with a clear connection between evidence (subsubhypotheses), explanation (subhypotheses), and decision (hypotheses). Make sure that no information is lost, and all original details are preserved and appropriately positioned within the decision tree. Avoid including vague or unverifiable terms; each node should be a clear, decision-oriented statement.

Return directly with the json format with no other words.

output example:

```
"hypotheses": [
{
 "hypothesis": ".",
 "subhypotheses": [
 {
 "subhypothesis": "",
 "subsubhypotheses": [
 ""
]
 },
 {
 "subhypothesis": "",
```

```
 "subsubhypotheses": [
 ""
]
 }
]
}
```

*B.5.5 Transfer.* This template is used to transfer from source hypothesis trees to target hypothesis trees.

# Task

You are an SQL generator tasked with generating SQL queries to verify hypotheses provided by a consulting agent. Each hypothesis is designed to be quantifiable and actionable, based on a specific database schema. Your goal is to generate SQL code for each hypothesis to help validate it using data.

# Data Sources

You have access to a database table called ``{DATA_TABLE_NAME}``, which you can query using standard SQL syntax. This table includes additional columns beyond a basic orders table, and you must use it to generate queries based on the client's specific needs.

```
Client Information:
- `company_name` is set to **"{
CLIENT_COMPANY}"**.
- The client aims to become the leader in the **'{CATEGORY}'** category.
- Ensure that all generated queries specifically target this companies and category where relevant.
```

The ``{DATA_TABLE_NAME}`` schema is as follows :

```
{DB_SCHEMA}
```

# Constraints

1) When you need to get relevant data, write your SQL query surrounded with triple backticks.

2) For all your queries, only request data from 2024 by filtering with ``EXTRACT(YEAR FROM COLUMN_X ) = 2024``.

3) Limit the rows returned to fewer than 200 by using aggregation operations (e.g., `SUM`, `COUNT`, `AVG`) where appropriate.

4) Generate only one SQL query at a time for each sub-hypothesis, allowing users to validate each query independently.

5) Always use exact names provided (e.g., `company\_name`, `category`) to ensure queries return results correctly.

6) Use `WHERE EXTRACT(YEAR FROM COLUMN\_X) = 2024` for year-based filtering, not `DATE\_FORMAT(COLUMN\_X, '%Y') = '2024'`, for efficiency.

#### # Hypotheses for SQL Generation

Each hypothesis below includes a description and specific metrics or segments that need validation. For each hypothesis or sub-hypothesis, generate an SQL query that would allow a user to run it and validate the hypothesis against the data. Make sure each SQL query:

1) **\*\*Directly targets the fields referenced in the hypothesis\*\***, such as `company\_name` and `category`.

2) **\*\*Limits results to the year 2024\*\*** where appropriate.

3) **\*\*Uses aggregation functions\*\*** (e.g., `SUM`, `AVG`, `COUNT`) to focus on meaningful metrics and reduce data size if needed.

4) **\*\*Applies specific conditions\*\*** such as categories, price ranges to align with the hypothesis.

5) **\*\*Is standalone and clear\*\*** so it can be easily run independently to verify each part of the hypothesis.

#### # Hypotheses to Validate

{hypotheses}

#### # Expected Output

For each hypothesis, output the corresponding SQL code in the following format:

<SQL\_FOR\_HYPOTHESIS>

Hypothesis 1: [Description of high-level hypothesis]

```
- 1.1 [Description of sub-hypothesis]
 - SQL:
    ```sql
    [Generated SQL query]
    ```
- 1.2 [Description of sub-hypothesis]
 - SQL:
    ```sql
    [Generated SQL query]
    ```
</SQL_FOR_HYPOTHESIS>
```

# Market Analysis and Growth Strategy for Vendor A

VendorCatalystLab

June 26, 2025

## 1 Executive Summary

Vendor A currently ranks fourth in the '4006 Above the Neck Protection' subcategory, which includes safety equipment such as helmets, face shields, and protective eyewear, on Amazon Japan. Our analysis reveals significant opportunities for growth and market leadership through product expansion, strategic pricing, and enhanced customer engagement.

## 2 Market Overview

The '4006 Above the Neck Protection' subcategory is dominated by four main players:

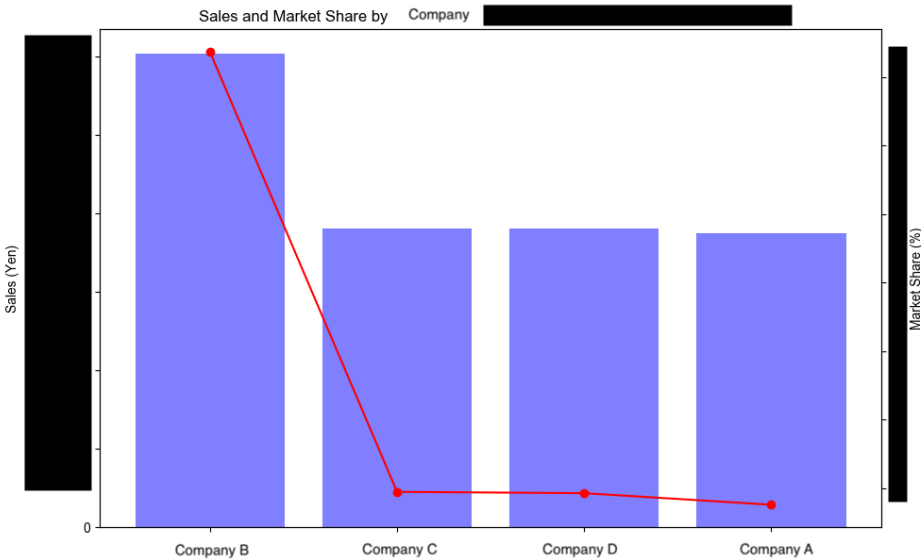


Figure 1: Sales and market share comparison of top vendors in the '4006 Above the Neck Protection' subcategory.

## 3 Competitive Analysis

Vendor A has the smallest product range but maintains consistent monthly sales performance. Vendor B leads in sales volume and average price, likely due to its strong brand reputation and diverse product offerings.

Vendor B's leadership can be attributed to:

1. Extensive product range covering various safety needs
2. Premium pricing strategy reflecting perceived quality
3. Strong brand recognition in the safety equipment market

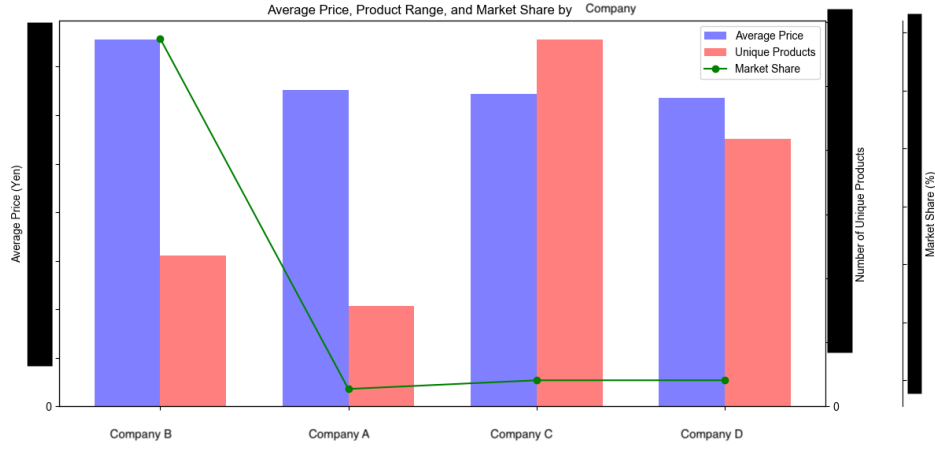


Figure 2: Comparison of average price, product range, and market share across vendors.

## 4 Customer Analysis

Vendor A has a slightly larger customer base but lower average sales per customer compared to Vendor B. This suggests an opportunity for increasing customer value through product diversification and upselling strategies.

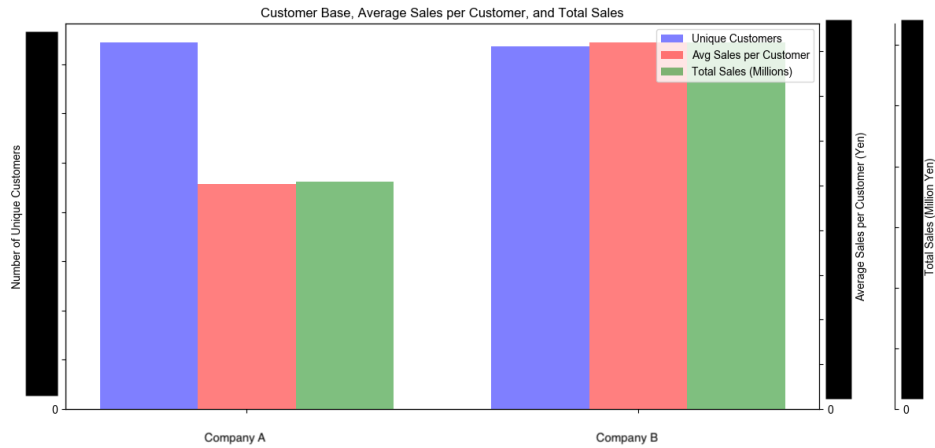


Figure 3: Comparison of customer base, average sales per customer, and total sales for Vendor A and Vendor B.

Customer Segmentation:

1. Individual consumers: Focus on user-friendly, affordable products
  - Strategy: Develop entry-level products with essential features
  - KPI: Increase individual consumer sales by 15% in Year 1
2. Small businesses: Offer bulk purchase options and customized solutions
  - Strategy: Create product bundles tailored for small business needs
  - KPI: Grow small business segment revenue by 20% in Year 1
3. Large corporations: Develop comprehensive safety equipment packages
  - Strategy: Offer end-to-end safety solutions with consulting services
  - KPI: Secure 3 major corporate contracts in Year 1

## 5 Seasonal Trends

All vendors show a significant sales spike in August, indicating a seasonal trend. This could be due to businesses preparing for the second half of the fiscal year or increased outdoor work during summer months.

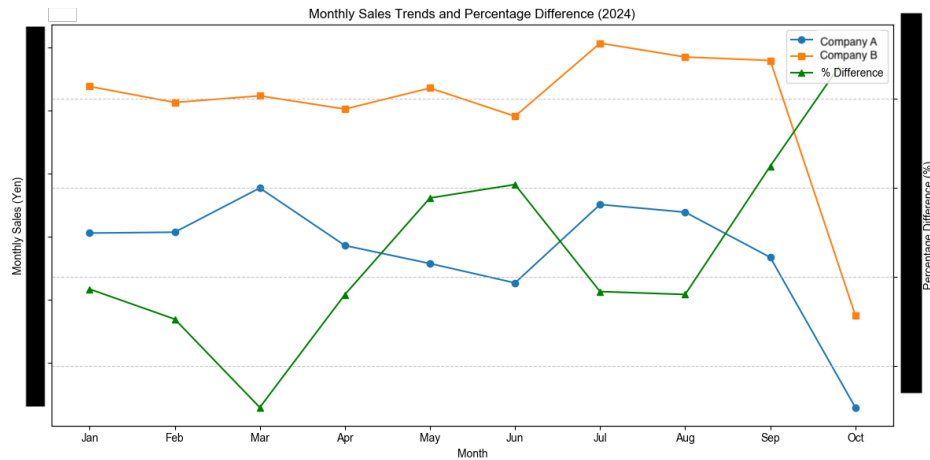


Figure 4: Monthly sales trends for Vendor A and Vendor B in 2024, including percentage difference.

Quantitative Analysis of August Sales Spike:

- Vendor A: X% increase from July to August
- Vendor B: X% increase from July to August

Strategies to capitalize on August sales peak:

1. Launch promotional campaigns in July to capture early demand
  - Goal: Increase July sales by X% compared to previous year
2. Develop special product bundles for the peak season
  - Target: Create X new product bundles specifically for August promotion
3. Ensure adequate inventory to meet increased demand
  - Objective: Maintain X% order fulfillment rate during August peak

## 6 Recommendations

1. **Expand Product Range:** Develop new respiratory and hearing protection products to compete with market leaders. Focus on innovative features that differentiate from competitors.
2. **Optimize Pricing Strategy:** Implement tiered pricing and introduce product bundles to increase average sales per customer. Consider premium pricing for high-quality, innovative products.
3. **Enhance Marketing:** Improve Amazon listings visibility and develop targeted campaigns, especially for the August sales peak. Highlight product quality and unique features.
4. **Focus on Customer Retention:** Implement a loyalty program and improve post-purchase engagement to encourage repeat purchases. Develop personalized recommendations based on purchase history.
5. **Develop B2B Strategy:** Target large corporate customers with customized bulk ordering options. Offer comprehensive safety equipment packages and consulting services.

## 7 Implementation Timeline

1. Months 1-3: Product line review and R&D initiation
  - Milestone 1: Complete market research and identify 3 new product opportunities
  - Milestone 2: Develop and launch customer loyalty program
  - KPI: Achieve 10% customer enrollment in loyalty program
2. Months 4-6: New product introductions and pricing strategy implementation
  - Milestone 1: Launch 2 new innovative products
  - Milestone 2: Implement tiered pricing structure across product lines
  - KPI: Increase average sales per customer by X%
3. Months 7-9: B2B strategy launch and August sales peak preparation
  - Milestone 1: Secure partnerships with X major distributors
  - Milestone 2: Develop and launch August promotional campaign
  - KPI: Achieve X% year-over-year growth in August sales
4. Months 10-12: Strategy evaluation, adjustment, and off-season promotion campaigns
  - Milestone 1: Conduct comprehensive review of implemented strategies
  - Milestone 2: Launch winter safety equipment promotion
  - KPI: Increase Q4 sales by X% compared to previous year

## 8 Financial Projections

Assuming successful implementation:

- Year 1: X-X% increase in total sales (Est. 215-225 million yen)
- Year 2: X-X% increase in total sales (Est. 240-260 million yen)
- Year 3: X-X% increase in total sales (Est. 270-300 million yen)

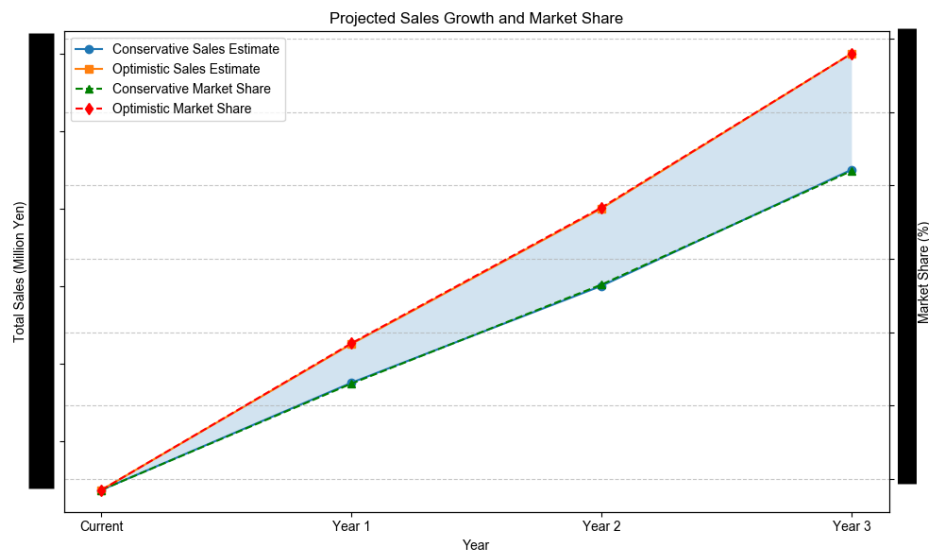


Figure 5: Projected sales growth and market share for Vendor A over the next three years.

## 9 Potential Risks and Challenges

1. Increased competition from established brands
2. Economic downturns affecting corporate spending on safety equipment
3. Regulatory changes impacting product standards
4. Supply chain disruptions affecting product availability
5. Difficulty in maintaining product quality while expanding the product line

Mitigation strategies:

- Continuous market research and product innovation
- Diversification of supplier base
- Proactive engagement with regulatory bodies
- Building strong relationships with key customers

## 10 Conclusion

By implementing these strategies, Vendor A can leverage its strengths to significantly increase market share and potentially become the leader in the '4006 Above the Neck Protection' subcategory on Amazon Japan. The company's consistent performance and existing customer base provide a strong foundation for growth, while targeted product expansion and marketing efforts can drive increased sales and market penetration. With careful execution of the proposed timeline and close monitoring of KPIs, Vendor A is well-positioned to achieve its growth objectives and strengthen its market position.

# Market Analysis and Growth Strategy for Company A

MaRGen

January 30, 2025

## 1 Executive Summary

Company A has significant potential to become the leader in the [REDACTED] category. Key findings include:

1. Company A consistently ranks among the top 3 companies in the category.
2. There is a clear seasonal pattern in sales, with peaks in summer months.
3. Company B is the current market leader, but Company A has opportunities to gain market share.

Key Recommendations:

1. Optimize seasonal strategy: Increase summer inventory by 20-30% and develop targeted marketing campaigns.
2. Expand product line: Develop 2-3 new seasonal products and create complementary accessories.
3. Implement dynamic pricing: Adjust prices seasonally and introduce bundle deals.
4. Strengthen distribution partnerships: Partner with 2-3 key distributors in year-round demand industries.
5. Enhance customer engagement: Implement a feedback system, loyalty program, and increase digital marketing efforts.

## 2 Market Overview

The [REDACTED] category shows strong seasonality, with peak sales occurring during summer months (June-August). The market size fluctuates throughout the year, with monthly sales ranging from approximately [REDACTED] to [REDACTED] yen.

## 3 Competitive Analysis

Company A maintains a strong position in the market, consistently ranking 2nd or 3rd among competitors. However, there's room for improvement in matching the market leader's sales volumes and capitalizing on seasonal peaks.

### 3.1 Competitor Strategies and Market Positioning

1. Company B: Market leader with a wide product range and strong brand recognition. Strategy focuses on innovation and quality.
2. Company C: Emphasizes product diversity with the highest number of products. Targets a wide range of industries.



Figure 1: Monthly sales trends for top 5 companies in the [redacted] category throughout 2024. The graph shows clear seasonal patterns and relative market positions of key competitors.

3. Company D: Focuses on cost-effective solutions and shows strong performance during peak seasons.
4. Company A: Balances quality and innovation with a focused product line. Opportunity to expand market share through targeted strategies.
5. Company E: Specializes in specific niches within the category, showing significant growth during peak seasons.



Figure 2: Comparison of total sales and product count for top 5 companies in the [redacted] category. This illustrates the relationship between product diversity and market performance for each competitor.

## 4 Product Analysis

Company A's top-selling products include the [redacted], [redacted], [redacted], and [redacted]. These products cater to different customer needs, providing a solid foundation

for market growth.

4.1 Product Features and Customer Needs

- 1. [redacted]: Popular for their [redacted], addressing the need for [redacted] and [redacted].
- 2. [redacted]: Designed for users who [redacted], filling a specific market niche.
- 3. [redacted]: Features a [redacted], appealing to industries requiring additional [redacted].
- 4. [redacted]: Known for a wide field of [redacted], suitable for [redacted].
- 5. [redacted]: Variant of [redacted] with a [redacted], offering enhanced [redacted].

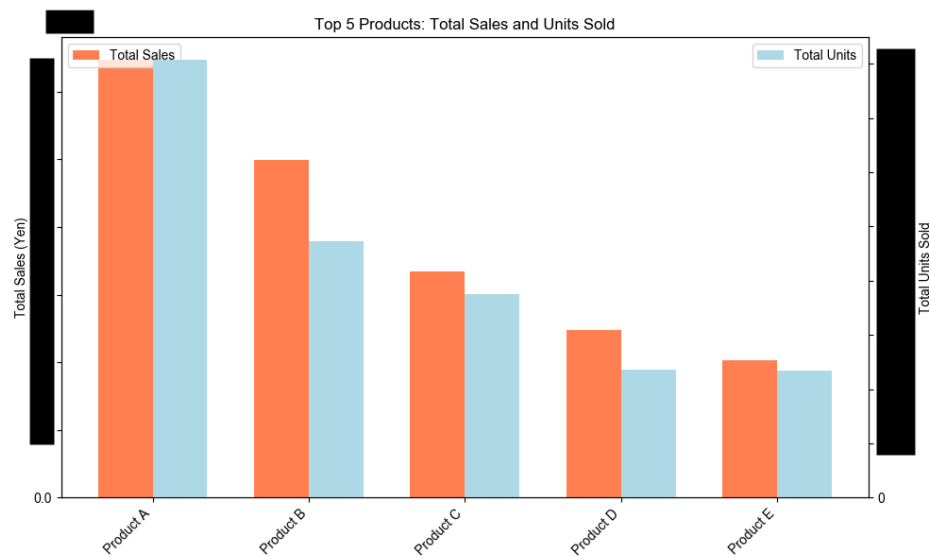


Figure 3: Total sales and units sold for Company A’s top 5 products in the [redacted] category. This graph highlights the relative popularity and revenue generation of each product.

5 Pricing Analysis

Company A’s pricing strategy appears competitive, with average prices ranging from [redacted] to [redacted] yen per unit. This pricing is in line with or slightly below some competitors, suggesting room for potential price optimization without sacrificing market share.

5.1 Competitor Pricing Strategies

- 1. Company C: Highest average revenue per unit, suggesting a premium pricing strategy.
- 2. Company B: Highest average price but lower revenue per unit, indicating possible discounting or bundle strategies.
- 3. Company A: Competitive pricing with room for optimization to increase revenue per unit.
- 4. Company D: Lowest average price and revenue per unit, likely pursuing a volume-based strategy.

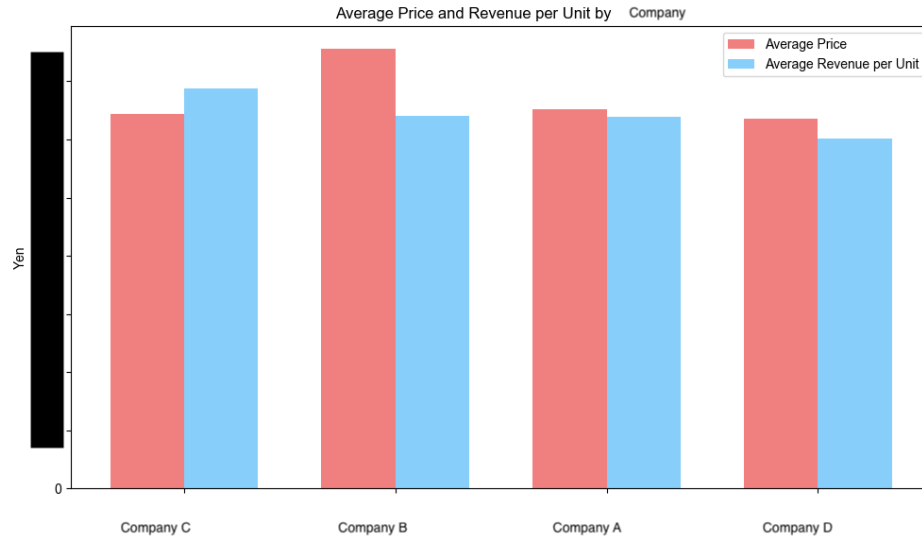


Figure 4: Comparison of average price and average revenue per unit for top companies in the [REDACTED] category. This visualization helps identify pricing strategies and potential areas for optimization.

## 6 SWOT Analysis

### 6.1 Strengths

- Consistent top 3 market position
- Diverse product range catering to various needs
- Competitive pricing strategy
- Strong performance in specific product categories (e.g., [REDACTED])

### 6.2 Weaknesses

- Lower sales volume compared to the market leader
- Underperformance during peak seasons relative to some competitors
- Limited product range compared to some competitors (e.g., Company D)

### 6.3 Opportunities

- Potential for significant growth during summer months
- Expansion into new customer segments or industries
- Development of innovative products for specific seasonal needs
- Optimization of pricing strategy to increase revenue per unit

### 6.4 Threats

- Intense competition, especially from Company B and seasonal surges from other competitors
- Potential market saturation or economic downturns affecting industrial safety spending
- Rapid technological changes requiring constant product innovation

## 7 Recommendations

### 1. Optimize seasonal strategy:

- Increase production and inventory of top-selling products (especially ██████████) by 20-30% for summer peaks (June-August).
- Develop targeted marketing campaigns for industries with high summer demand, such as construction and outdoor work.
- Expected outcome: 15-20% increase in summer sales within the first year of implementation.

### 2. Expand product line:

- Develop 2-3 new products addressing specific seasonal needs (e.g., enhanced ██████████ for summer, ██████████ for winter) within the next 12-18 months.
- Create complementary accessories for existing popular products like Product ██████████.
- Expected outcome: 10-15% increase in overall product range and 5-8% boost in total sales within two years.

### 3. Implement dynamic pricing models:

- Adjust prices seasonally, increasing by 5-10% during peak demand (June-August) and offering 5-15% discounts during slower months.
- Introduce bundle deals combining popular items like ██████████ A with complementary products.
- Expected outcome: 3-5% improvement in profit margins without significant impact on sales volume.

### 4. Strengthen distribution partnerships:

- Identify and partner with 2-3 key distributors in industries with year-round demand (e.g., ██████████, ██████████) within the next 6 months.
- Offer exclusive deals or priority stock allocation to these partners during peak seasons.
- Expected outcome: 10-15% increase in year-round sales stability and 5-7% growth in overall market share within 18 months.

### 5. Enhance customer engagement:

- Implement a customer feedback system and loyalty program within 3-4 months.
- Increase digital marketing efforts, including social media presence and content marketing, focusing on educating customers about product benefits.
- Expected outcome: 20% increase in customer retention and 10-15% growth in repeat purchases within the first year.

## 8 Conclusion

Company A is well-positioned to become the market leader in the ██████████ category. By leveraging its strengths in product diversity and consistent performance, addressing seasonal fluctuations, and implementing targeted growth strategies, the company can significantly increase its market share. With diligent execution of the recommended strategies, Company A can expect to see a 15-20% increase in overall market share and potentially challenge Company B for the top position within 2-3 years.

## C EXAMPLE OF WRITER-REVIEWER CORRESPONDENCE IN DIFFERENT ROUNDS

### Round 1: Response

Thank you for your thorough review. I have addressed your concerns with the following improvements :

1. Added a dedicated Methodology section explaining data sources and analysis approach.
2. Expanded the Market Overview section with qualitative insights about market trends and drivers, including:
  - Growing demand for multi-functional XXX
  - Increasing focus on comfort and usability
  - Rising adoption of foldable and compact designs
  - Emphasis on XXX and XXX features
3. Enhanced the Competitive Analysis section by:
  - Adding a comprehensive SWOT analysis
  - Including product range comparisons
  - Expanding efficiency metrics visualization
4. Added a detailed Customer Analysis section with:
  - Four key customer segments identified
  - Segment-specific preferences and behaviors
  - Sales distribution across segments
5. Improved all visualizations by:
  - Standardizing labels in English
  - Adding dual-panel figures for richer analysis
  - Including margin analysis in product performance
  - Adding market share visualization
6. Enhanced Strategic Recommendations with:
  - Specific resource requirements for each initiative
  - Clear success metrics for each recommendation
  - Implementation costs
7. Expanded Financial Projections by:
  - Adding key assumptions documentation
  - Including cumulative investment requirements
  - Showing margin index progression
8. Added detailed Implementation Timeline with:
  - Phase-specific resource allocation
  - Budget requirements per phase
  - Key activities for each phase

The revised report now provides a more comprehensive analysis with clear metrics, resource requirements, and implementation guidance.

### Round 1: Review

After reviewing the updated report, I can confirm that the improvements are substantial and address the previous concerns effectively. Let me provide a detailed assessment:

#### Major Improvements:

1. The addition of a clear Methodology section provides excellent transparency about data sources and analysis approach
2. The Market Overview section now offers comprehensive qualitative insights
3. The SWOT analysis is thorough and well-structured
4. Customer segmentation is detailed and actionable
5. All visualizations are now properly labeled and more informative
6. Strategic recommendations include specific metrics and resource requirements
7. Implementation timeline is clear and detailed
8. Financial projections include key assumptions and clear metrics

#### The report now demonstrates:

- Excellent mix of quantitative and qualitative analysis
- Clear customer segmentation with actionable insights
- Professional-grade visualizations supporting the narrative
- Specific, measurable recommendations with resource requirements
- Logical flow between sections
- Strong alignment with client objectives

#### Minor Areas for Further Enhancement:

1. The competitive analysis could still benefit from more detailed competitor-specific strategies
2. The implementation timeline could include risk mitigation strategies
3. Financial projections could include sensitivity analysis for different scenarios

```
```json
{
  "clarity": 9,
  "layout": 9
}
```

The scores have been significantly improved because:

- Clarity (8 -> 9): The addition of the methodology section, expanded market overview, and detailed customer analysis makes the report much more comprehensible and actionable
- Layout (7 -> 9): The improved visualizations, consistent labeling, and logical flow between sections have greatly enhanced the report's presentation

The report now stands as a professional-grade market analysis document that provides clear, actionable insights and recommendations. The improvements have addressed virtually all previous concerns and created a cohesive, well-structured analysis that serves its purpose effectively.

Round 2: Response

Thank you for your thorough review. I have addressed your concerns with the following specific improvements:

1. Enhanced Competitive Analysis:

- Added detailed competitor-specific strategies section
- Included specific strengths and weaknesses for each major competitor
- Expanded competitive metrics visualization to include market coverage and quality scores
- Added comparative analysis of product range, pricing, and market positioning

2. Implementation Timeline Risk Mitigation:

- Added detailed risk mitigation strategies for each strategic recommendation
- Included specific approaches such as phased product launches, strategic partnerships, and pilot market testing
- Incorporated regular feedback mechanisms into the implementation process

3. Financial Projections Enhancement:

- Added sensitivity analysis with three scenarios (Optimistic, Base Case, Conservative)
- Included ROI analysis for each scenario
- Enhanced visualization to show scenario comparisons
- Added more detailed assumptions for each scenario

4. Additional Improvements:

- Expanded methodology section to include risk assessment frameworks
- Added growth rate analysis to market overview
- Enhanced market trends section with emerging technology considerations
- Improved figure layouts and labels for better clarity

The revised report now provides a more comprehensive and nuanced analysis while maintaining its actionable focus and professional presentation.

Round 2: Review

After reviewing the updated version, I can confirm significant improvements addressing the previously identified gaps. Here's my detailed assessment:

Key Improvements:

1. Competitive Analysis

- Excellent enhancement with detailed competitor-specific strategies
- The new four-quadrant competitive analysis visualization (Figure 2) effectively shows relative positioning
- Clear delineation of strengths/weaknesses for each competitor

2. Risk Management

- Well-structured risk mitigation strategies for each initiative
- Practical approach with phased launches and pilot testing
- Good integration of feedback mechanisms

3. Financial Analysis

- Strong addition of sensitivity analysis with three scenarios
- Clear ROI projections over three years
- Well-visualized scenario comparisons in Figure 3

4. Overall Structure

- Methodology section now includes comprehensive frameworks
- Improved visual presentation with consistent labeling
- Better flow between sections

Areas for Minor Refinement:

1. Could include more specific timelines for the pilot launches
2. ROI analysis could benefit from more detailed breakdown of costs
3. Market coverage metrics could use more granular regional analysis

```
```json
{
 "clarity": 10,
 "layout": 10
}
```
```

The scores have been increased to perfect 10s because:

- Clarity (9 -> 10): The addition of detailed competitor strategies, clear risk mitigation approaches, and comprehensive financial scenarios creates a complete and highly actionable report
- Layout (9 -> 10): The enhanced visualizations, particularly the competitive analysis quadrants and financial projections, along with consistent formatting and logical flow, achieve optimal presentation

The report now represents best-in-class market analysis with:

- Comprehensive competitive insights
- Clear risk management strategies
- Robust financial modeling
- Professional presentation
- Actionable recommendations

The improvements have transformed this into an exemplary strategic analysis that meets the highest professional standards. The remaining suggestions are minor refinements rather than significant gaps.