California Management Review

TECHNOLOGY

Can GenAI do your next strategy task? Not yet.

by Christoph Lechner, Nikolaus Lang, Siegfried Handschuh, Olivier Bouffault, and Julian Cooper

Image Credit | JMarques

Empirically testing the potential of GenAI by analyzing three typical and representative strategic management tasks.

INSIGHT | FRONTIER 15 Jan 2025

Can GenAI really solve strategic management

tasks?

Generative AI (GenAI) is a type of AI based on language-base models that can create new content and ideas, including conversations, stories, images, videos, and music. For strategic management tasks required of company executives and investors, the implications and potential use cases are less clear. We focus on the ability of these models to complete strategic management tasks independently with a view to future automation.

RELATED CMR ARTICLES

"A Brief History of Artificial Intelligence: On the Past, Present, and Future of Artificial Intelligence" Haenlein, Michael and Andreas Kaplan, 61/4 (Summer 2019): 5-14.

"Getting AI Implementation Right: Insights from a Global Survey" Ångström, Rebecka C., Michael Björn, Linus Dahlander, and Magnus Mähring, 66/1 (Fall 2023): 5-22.

"Keeping Humans in the Loop: Pooling Knowledge through Artificial Swarm Intelligence to Improve Business Decision Making" Metcalf, Lynn , David A. Askay, and Louis B. Rosenberg, 61/4 (Summer 2019): 84-109.

"Designing the Intelligent Organization: Six Principles for Human-AI Collaboration" Kolbjørnsrud, Vegard, 66/2 (Winter 2024): 44-64.

What makes strategic management difficult to automate?

Strategic management can be defined as the process of realizing a company's strategies to achieve the goals set for key stakeholders (Müller-Stewens & Lechner, 2015)[1]. It stands in contrast to individual functions, such as design, procurement, production, logistics, and marketing and sales. Some of the tasks that we typically associate with modern strategic management include market research, scenario planning, corporate strategy, mergers and acquisitions, business models, or turnaround restructuring.

Strategic management tasks are perceived as particularly complex to automate because they require (to varying degrees) (a) multi-step and multi-level reasoning, (b) context-dependence, and (c) some understanding of human behaviour (Finkenstadt, 2023) [2]. While these challenges remain for even the most advanced GenAI models – such as OpenAI's GPT, Anthropic's Claude, Meta's Llama, Google's Gemini, or Mistral's Mixtral of Experts – the following paragraphs share approaches practitioners are using to tackle each in isolation.

1) Multi-step and multi-level reasoning: multi-step and multi-level reasoning performance is often addressed with step-by-step prompting strategies (OpenAI SDK, 2024)[3]. These strategies range from asking the LLM to "show your logic step-by-step", to including a list of the steps required within the prompt, to feeding each of these steps one-by-one to the model in separate prompts. While separating a task across multiple prompts can theoretically be implemented with automation, performance is improved by having a human-in-the-loop to course correct if an intermediate step is incorrectly computed. There is a natural limit for each of these step-by-step prompting strategies. Single prompt approaches suffer from limited context window and output window sizes, while LLMs tend to forget context or data from earlier steps if the task is separated across too many separate prompts (OpenAI SDK, 2024)[3].

2) Context-dependence: To be useful for strategic management applications, GenAI models also need to be able to access additional (and often sensitive) contextual data. For example, this might be a company's internal financial data and strategy memos when evaluating whether an M&A buyside transaction could be attractive. Configuring GenAI models for context-dependent applications is typically addressed with retrieval-augmented generation (RAG) architectures (OpenAI DevDay, 2023). RAG is a technique that adds an information-retrieval component to the generation process, allowing the LLM to query an arbitrarily large secondary data source and incorporate retrieved data into the context window alongside the user's prompt (see ***Figure 1***). Importantly for strategic management, this approach allows our GenAI model to access sensitive company databases (without fine-tuning), reduce probability of "hallucinations", and even source (share back retrieval query) the data used to produce a given response (Lewis et al., 2020)[4]. One prominent corporate example is BoschGPT, which Bosch developed in collaboration with Aleph Alpha.

3) Human behavior factors: Understanding and anticipating likely human behaviours (internal team dynamics, customer expectations, cultural contexts, etc.) plays a significant role in many strategic management subdisciplines. For example, a business pricing strategy may be based on expectations about consumer willingness to pay; but then also how competitors in the market will respond with their own pricing strategies; and finally, how the consumer will weigh these two when making their next purchasing decision. We can of course provide the model with our own assumptions as a guide, but true automation would mean us asking the LLM to provide its own assumptions given the raw historical pricing data we had access to as a starting point. To begin to train for these sorts of human behaviour intuitions, the best available approaches are either to provide case studies as contextual prompts (i.e. many-shot experiments) or fine-tune based on task-specific data with human expert annotations indicating the missing behavioural elements that are relevant (OpenAI SDK, 2024). Neither of these approaches is a silver bullet.

4) Benchmarking model performance: As a brief aside, it is helpful to understand how the GenAI community evaluates and benchmarks model performance for each of these difficult-to-automate capabilities. The mapping is not one-to-one, but also not far off. For integrated knowledge and reasoning capability, each of the leading LLMs will regularly publish their ARC, HellaSwag and MMLU scores. ARC refers to the AI2 Reasoning Challenge, which is a dataset of grade-school multiple choice questions (Clark et al., 2018)[5]. HellaSwag is a dataset of common-sense reasoning and logic questions (Zellers et al., 2019)[6]. And MMLU refers to Multi-task Language Understanding which is a dataset focused on graduate-level academic topics, with more of an emphasis on knowledge understanding and retrieval (Hendrycks et al., 2021[7]). For evaluating a model's ability to incorporate context through RAG architectures, our best-practice metrics closely resemble the classification machine learning confusion metrics, which is more context dependent and less of an exact science. We consider both how well the LLM answers the prompt question ("generation") and how relevant the content retrieved was for this answer ("retrieval"). For generation, we measure the factual accuracy of the answer and the relevancy of the answer to the question. And for retrieval, we measure the signal-to-noise ratio (context precision) and whether the content retrieved was sufficient to answer the question (context recall) (OpenAI DevDay, 2023). For social reasoning, benchmark datasets such as SocialIQA, which test if a model can predict what happens next in a story or explain motivations behind actions, are the current standard (Sap et al., 2019)[8].

Design and Findings of Three Experiments

We are interested in how these models perform on real-world strategic management tasks, which require combinations of these capabilities. To test this, we have designed three experiments that reflect Strategic Management tasks of increasing complexity and value: 1) compiling a market research dossier; 2) evaluating business strategy, and 3) running the analyses required for a buyside due diligence. The tasks chosen for these experiments represent a substantial part of the work strategy and investment teams do day-to-day.

Experiment #1: Compiling market research dossier

Design: We asked ChatGPT-4 to perform three specific analyses that were provided in an actual dossier prepared by an internal Boston Consulting Group (BCG) team on the Indian agrochemicals market: (1) summarize qualitative insights into agrochemicals globally by region; (2) plot the size of the global agrochemicals market from 2018-2023 as a stacked bar chart split by region; and finally (3) deep-dive into the Indian market and share some analysis on market attractiveness and competitive landscape (see ***Figure 2***). We included an industry report Global Agrochemicals (Grand View Research, 2021) – one of the primary resources used by the BCG team – as an attachment within the context window alongside our initial prompt.

Findings: First, the executive summary prepared by BCG could be reasonably derived from the ChatGPT-4 output alone. In fact, the headline CAGR numbers for the period specified (only calculation not included) even matched! Second, the model was able to parse and retrieve information from a 200-page industry report that included text, charts, and data tables. Third, with a single prompt, the model was able to respond (and perform analysis) at multiple levels of detail.

Conclusion: The findings from Experiment #1 demonstrate that LLMs are already able to perform large-scale synthesis tasks in a strategic management context, with some limited data aggregation and reasoning, in an automated way. In future research it would be interesting to test how this performance scales with a RAG architecture. For example, we could imagine giving the model access to a database of industry reports. If the retriever was well-designed, it would be interesting to see if the human-in-the-loop (providing relevant source material) could be effectively removed.

Experiment #2: Tackling the strategy decision by a consulting case interview

Design: A consulting case interview is almost always structured in four parts: (1) candidate is given case context and asked how they would approach this problem; (2) back-and-forth conversation to search for the core issues; (3) data exhibits shared that require candidate to conduct some calculations and provide quantitative insights; and (4) prepare concluding remarks to present back to a senior client stakeholder. We asked ChatGPT-4 to play the role of a candidate in a retired BCG case interview about supermarket frozen foods (see ***Figure 3***). The responses to each section were evaluated against typical responses from our scoring rubric, i.e. expected of human applicants.

Findings: For part (1), ChatGPT-4 was able to invent a relevant and approximately MECE (mutually exclusive collectively exhaustive) framework for tackling the case prompt, however, it was arguably not sufficiently hypothesis-driven to receive a passing grade. The case prompt specifically requested the candidate focus on profitability and so a strong candidate would be expected to include some discussion of revenue vs cost in this response. For part (2), the interviewer suggests pricing might be worth investigating. The model was quickly able to provide a succinct list of drivers for this but did not provide intuitive links back to the case itself or suggest next steps to help drive the conversation towards a conclusion as a top candidate might have done. For parts (3) and (4), the model generally performed at a passing level. We provided two data exhibits and for each it was able to produce the correct mathematical result and basic "so what" (e.g. frozen pizza declined 50% and it is a price issue, not cost or quantity), likely matching the performance of a strong candidate. This was our biggest surprise since we had expected the model to miss the "so what" links back to the case prompt given its responses to the earlier more qualitative questions. Again, more than sufficient for a pass mark.

Despite performing quite well overall from a context perspective, the model completely missed some of the human behavioural qualities expected of top candidates. For example, exhibiting an inquisitive, curious mindset and "driving the interview" by proposing hypothesis-driven next steps to the interviewer was out of reach, even when we experimented with warm-up prompting routines where we provided guidance.

Conclusion: The findings from Experiment #2 demonstrate the potential of LLMs to perform scenario planning tasks with a human-in-the-loop. While not yet good enough at proposing an approach for solving an abstract strategic problem, if guided by a human through this first step, these models can provide very effective support for running the subsequent analyses to help rapidly test hypotheses and drive towards a solution. These results were consistent with findings of previous research for problem-solving tasks (Dell`Aqua et al., 2023)[9].

Experiment #3: Modelling for a buyside due diligence

Design: For our final experiment, we wanted to understand the extent to which today's GenAI models could handle the complexity and quantitative rigor of an M&A due diligence. To test this, we used a case study from Stanford GSB's Financial Modeling course (Demarzo, 2022) on Stride Rite's 2005 acquisition of Saucony. We asked ChatGPT-4 to perform the sequence of analyses required by this course assignment, which include (1) building capitalization tables; (2) combining income statements and balance sheets; (3) calculating discounted cash flow; and (4) recommending the price per share that Stride Rite should offer (see ***Figure 4***).

Findings: In contrast to the previous tests, this experiment largely highlighted the limitations of ChatGPT-4, rather than its strengths. Despite this, there were some definite bright spots worth mentioning. First, the model was very effective at digesting the case study documents (one pdf, one excel) and organizing this information to answer specific questions – e.g. “extract the common and fully diluted shares for both the buyer and target companies”, or even “build capitalization tables for both companies”. Second, the model exhibited some capability to self-diagnose inconsistencies and potential errors. For example, when the model computed a negative benefit value associated with a proposed merger synergy, it included text in the response to warn the user that this result did not yet sense check.

In terms of weaknesses, our findings could be grouped into two categories: issues with quantitative multi-step reasoning and issues with fidelity. We asked the model to respond to two tasks that required quantitative multi-step reasoning: creating the pro forma combined income statement and computing discounted cash flow. In both cases we experimented with listing steps required at different levels of granularity as well as single vs multiple prompts, but despite significant guidance from the human-in-the-loop, we were not able to get the model to solve for the correct financial modelling results. In terms of fidelity, despite knowing these models to be stochastic, and therefore expecting some variations in results from session to session, we were surprised by the extent to which our results would differ given identical prompt and contextual data. For example, towards the end of the experiment we ask the model to provide its recommended price per share that Stride Rite should offer Saucony. The initial response was $28/share which was incorrect but reasonably close to the correct answer of ~$35/share. However, when prompted “can you try that exercise again?”, the model computed $96/share.

One of the key multi-step reasoning limitations we identified in this experiment was forgetfulness. While performance generally improved on these multi-step tasks as we broke instructions down into their component steps, we quickly reached a limit whereby the level of granularity required for the model to make the correct intermediate calculations necessitated providing too many prompts!

Conclusion: The findings from Experiment #3 demonstrate the clear limit of today's LLMs to handle truly complex tasks involving multiple reasoning steps – either too many to pass at once, or if too broken down, the model will forget earlier context. This can be partially addressed by designing engineering solutions around the LLMs to help encode and optimally reintroduce contextual data.

Risks and challenges to keep in mind

Strategic management decisions often have significant implications for a company's development. Therefore, we need to better understand the potentials but also pitfalls of current GenAI applications.

1) Inherent biases: GenAI models carry with them inherent biases linked to the datasets and natural language tasks used during pre-training. These biases can be exacerbated, or partially mitigated, by our choice of context window, retrieval-augmentation, and fine-tuning efforts. While still very much an active area of research, there are a few helpful benchmark datasets emerging to help practitioners (and LLM core platform developers) assess relative performance and progress made over time, such as Word Embedding Association Test, StereoSet, and FairFace (Schroder, 2022)[10]. This is just the beginning, and we can help by actively choosing GenAI technologies for our business that perform well on bias benchmarks, as well as more established measures like reasoning, context retrieval and so on.

2) Human-in-the-loop: While it is impressive to witness what today's GenAI models can accomplish with a human-in-the-loop, it is important to also remember the counterfactual: both case interview and due diligence experiments would not have been possible through pure automation. This is both encouraging and limiting. Encouraging in the sense that we expect most strategic management subdisciplines to be enriched rather than replaced by this technology but limiting from a scale perspective. Requiring a human-in-the-loop substantially limits the potential benefits of these technologies for a given task. For example, if due diligence analyses were truly automatable, one could imagine companies being able to constantly assess all possible merger and acquisition opportunities rather than rely on humans to select a short-list of potential targets to investigate.

Conclusion and Outlook

There are two main conclusions that we can draw from our study. First, today's LLMs are already able to automate large-scale synthesis tasks (e.g. market research), with some limited data aggregation and reasoning, but rely heavily on having a human-in-the-loop for any task requiring multiple steps or understanding of human behaviour (e.g. strategic scenario planning). Second, a hypothesis-driven and complex multi-step reasoning is still out of reach – for now. Complex multi-step analyses (e.g. buyside due diligence), even having a human-in-the-loop is insufficient to guide an off-the-shelf LLM to the correct result.

For leadership teams today, questions remain around two themes: (a) to what extent can we improve performance by designing dedicated systems (e.g. with separate fine-tuned quantitative modules, RAG retrieval from custom databases); and (b) how performance will natively increase with next versions of these LLMs (e.g. OpenAI's GPT-5, Meta's Llama-3)

We can already start to understand the benefits of dedicated systems by expanding on these sorts of experiments. For market research, it would be educational to give the model access to a database of industry reports via a RAG architecture and try removing the human (providing relevant source material) from the loop. For case interviews (and business scenario planning use cases in general), we might re-run the experiment with significant fine-tuning to help the model "learn" some of the behavioural patterns of top candidates. And finally for due diligence, it would be interesting to explore fine-tuning dedicated Custom GPTs for different parts (e.g. dedicated combined income statement generator) and embeddings-based search algorithms to reduce the memory burden associated with holding so much contextual data at one time.

LLMs and Generative AI have enormous value in the business context. We currently only can see the tip of the iceberg. These technologies will be at the basis of a far-ranging business transformation. In the coming months and few years, much of the transformation will be focused on automating basic intellectual tasks and processes – i.e. tasks requiring information retrieval, data synthesis, and some limited planning and reasoning capability. There are thousands of these processes and the productivity increase in those can be tremendous, 90%+ reduction in time required. Among our experiments, the Market Research Dossier is a typical example (see ***Figure 2***).

Advanced intellectual tasks – i.e. tasks requiring multi-step and quantitative reasoning, persistent long and short-term memory, and deep understanding of human behaviors – will also be in play for GenAI. But as shown through our second and third experiments, significant progress in LLM technologies and the deployment of these technologies are still needed to be effective in these fields. The business question becomes: at what point should I start seriously investing? A simple analogy would be: should I wait for smarter students to come out of general education (the new LLMs), or should I invest in codifying our business processes and teaching these to our existing workforce (building complex systems on top of current LLMs)? The latter approach is comfortable and perhaps less risky in the short-term, but also more rigid, and less adaptable to new situations or strategies.

To us, it comes back to a general framework about AI applications prioritization, which boils down to two questions. First, what is the value of the process I am considering applying AI to? Second, do I have a defensible advantage on access to relevant data? As LLMs mature and increase in performance, there will be less and less need for investment in specialized systems or fine-tuning, making the necessary investment smaller. At the same time, if today I have access to a vast amount of relevant data in written form, in any format, LLMs are now a way to exploit this data, which reinforces the value from the investment. So it might be that large investment funds, having defensible access to data from hundreds of past deals, will start investing soon in such advanced systems, while the typical corporate M&A departments will rationally wait a few more years for the underlying AI technology to mature.

References

1. Müller-Stewens, G. & Lechner, C. (2015). Strategic Management; Schäffer-Pöschel.

2. Finkenstadt, D. J., Eapen, T. T., Sotiriadis, J., Guinto, P. (November, 2023). Use GenAI to Improve Scenario Planning. Harvard Business Review. https://hbr.org/2023/11/use-genai-to-improve-scenario-planning

3. OpenAI Software Development Toolkit: Prompt Engineering Guide (March, 2024). https://platform.openai.com/docs/guides/prompt-engineering.

4. Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal , N., Küttler, H., Lewis M., Yih W., Rocktäschel, T., Riedel, S., & Kiela, D. (May, 2020). Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. Neural Information Processing Systems Foundation (NeurIPS).

5. Clark, P., Cowhey, I., Etzioni, O. Khot, T., Sabharwal, A., Schoenick, C., Tafjord, O. (March, 2018). Think you have Solved Question Answering? Try ARC, the AI2 Reasoning Challenge. Allen Institute for Artificial Intelligence, Seattle WA, USA.

6. Zellers, R., Holtzmann, A., Bisk, Y., Farhadi, A., Choi, Y. (May, 2019). HellaSwag: Can a Machine Finish Your Sentence? Allen Institute for Artificial Intelligence, Seattle WA, USA.

7. Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D., Steinhardt, J. (January, 2021). Measuring Massive Multitak Language Understanding (MMLU).

8. Sap, M., Rashkin, H., Chen, D., Bras, R., Choi, Y. (September, 2019). SocialIQA : Commonsense Reasoning about Social Interactions. Allen Institute for Artificial Intelligence, Seattle WA, USA.

9. Dell'Aqua, F., McFowland III, E., Mollick, E., Lifshitz-Assaf, H., Kellogg, K. C., Rajendran, S., Krayer, L., Candelon, F., and Lakhani, K. R. (October, 2023). Navigating the Jagged Technological Frontier: Field Experimental Evidence of the Effects of AI on Knowledge Worker Productivity and Quality. Harvard Business School, The Wharton School, Warwick Business School, MIT Sloan School of Management & Boston Consulting Group.

10. Schröder, Sarah, et al. “The same score: Improved cosine based measure for semantic bias.” arXiv 28 Mar 2022, pp. 1–8.

Figures

Figure 1 : Illustration of RAG architecture from AWS SDK, 2024

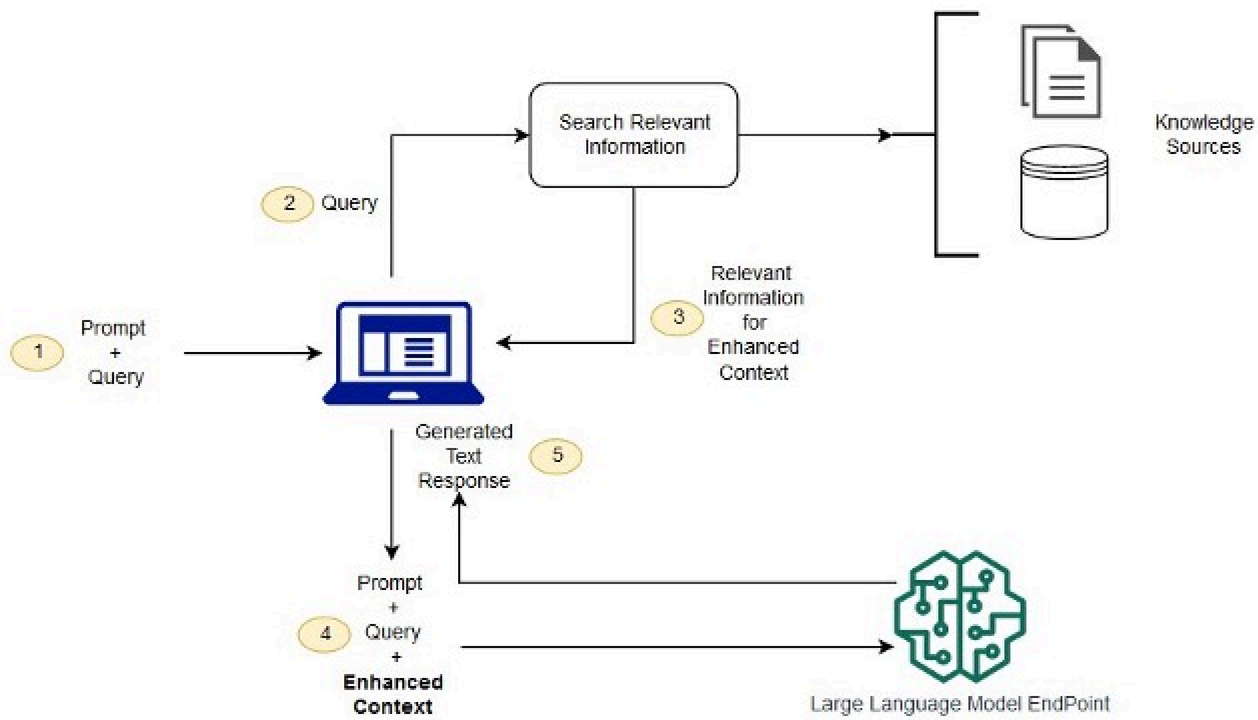


Figure 2: ChatGPT Experiment #1 - Market Research Dossier, 2024

<u>Emerging Market Dossier</u>: Indian Agrochemical Market, 2024

Task: Given underlying data (Global Agrochemicals Report, 2023), we asked *ChatGPT* to perform three specific analyses (summary, data manipulation and plotting, reasoning) that were provided in an actual dossier prepared by the BCG team.

Prompt

Summarize the global market insights found in this report ("Global Agrochemicals, 2023"). The report is not copyrighted and I have bought access to it.
[A] I'd like qualitative (and no numbers) insights into the following regions:
- 1. North America
- 2. South and Central America
- 3. Asia Pacific
- 4. Europe
- 5. Middle East and Africa.

These insights include: Which segments are dominating the region, what trends are occurring right now, who are the major contributors in the region, what are the key demand drivers.

[B] Then use Python to plot the Global Agrochemicals Market in million USD by year, from 2018 to 2023, in a stacked plots, where each of the regions' share can be seen. Please ensure that the colours are such that the different regions are easily distinguished. I also want you to calculate the CAGR and showcase that on the plot.

[C] Then deep-dive into the Indian agrochemicals market. I want you to summarize (1) if the overall market is attractive and (2) what are the competitors plays? Please do so by splitting each question into sub-analysis. For (1) attractiveness, these could include quantitative market size (compared to some benchmark), quantitative market growth (compared to some benchmark), key opportunities and challenges, competitive intensity and key success factors. For (2) competitor plays, these could include a segmentation of different types of companies, with qualitative statements on each segments' business model and characteristics. Please format this summarization into an easily readable slide or table.

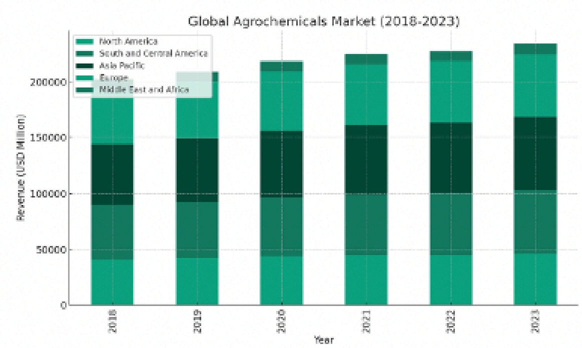
Please revisit each question and take your time to answer it. Take your time to conduct the Python analyses before moving on to the next question. All the answers to the questions can be extracted from the report.

ChatGPT

[A] Global Market Insights by Region
1. North America
- Dominating segments: Fertilizers and crop protection chemicals, with a strong push towards organic agrochemicals.
- Trends: Focus on sustainable agriculture, technological advancements, and increasing demand for bio-fertilizers and bio-pesticides.
- Major contributors: The United States and Canda.
- Key demand drivers: High agricultural activity and the growing dairy industry.

2. South and Central America
...
...

[B] Global Agrochemicals Market Analysis (2018-2023)
I will now proceed with the data analysis for part [B], including plotting the Global Agrochemicals Market and calculating the CAGR. This requires extracting and organizing the relevant data from the segments discussed. Let's start with preparing the data for the plot and calculations. The Compound Annual Growth Rate (CAGR) of the Global Agrochemicals Market from 2018 to 2023 is approximately **2.94%.** This indicates a steady growth in the agrochemicals market across the regions over the period.



[C] Indian Agrochemicals Market Analysis
[C.1] Market Attractiveness
- Growth Rate: The Indian agrochemicals market is growing at a CAGR of 4.4% from 2021 to 2028 in terms of revenue.
- Market Size: It was valued at USD 16,122.2 million in 2020 and is expected to reach USD 22,683.7 million by 2028.
- Opportunities and Challenges: The government's support and subsidies for certain agrochemical products are driving market growth. India, being a major producer of agricultural products, faces challenges, e.g. shrinking arable land and increasing demand for agrochemicals to improve yield.
- Competitive Intensity: The market is competitive, with a focus on fertilizers and crop protection chemicals. Bio-pesticides are gaining importance due to environmental concerns.

[C.2] Competitor Plays
- Segmentation: The market is segmented into fertilizers, crop protection chemicals, and plant growth regulators, with fertilizers dominating.
- Business Models: Companies in the market range from multinational corporations to local manufacturers, with a focus on innovative products and sustainable solutions.

Figure 3: ChatGPT Experiment #2 – Consulting Case Interview, BCG analysis, 2024

<u>Consulting Case Interview</u>: Supermarket Frozen Foods [1/2]

Task: Test how well *ChatGPT* can perform on a Boston Consulting Group case interview (prompt, data exhibits, etc.) compared to a human candidate. The responses to each section were evaluated against typical responses from our scoring rubric.

Prompt	ChatGPT
Explain the exercise	
We are going to run through an interactive case interview from BCG public repository. I will provide the case prompt below and then ask you to respond to questions, gradually introducing new information and data as we go. Please take your time, respond as concisely as you can, and explain your reasoning step-by-step.	Absolutely, I'm ready to assist with the case interview. Please provide the case prompt, and we can start working through it together.
Case interview prompt	
The client is a national supermarket chain positioned among the top 3 players in its market. The client has historically been able to maintain above average profit margins, serving as a benchmark for the industry. However, the client has identified a troublesome trend in its frozen food section. Profits have been declining steadily in the past year while other competitors have been able to maintain or even improve their margins. The client came to us because it wanted to understand what is driving its profitability decline in the frozen food section and identify possible solutions. How would you approach this?	To approach this case effectively, I would break down the problem into manageable parts to understand the root causes of the profitability decline in the frozen food section. Here's a structured approach: **1. Understand the Market Context**: Start by gathering information about the market dynamics, including trends in consumer behavior, changes in demand for frozen foods, and any recent developments in the frozen food industry. This step will help identify external factors that could be impacting profitability. **2. Analyze the Competition**: Look at what the competitors are doing differently. This includes pricing strategies, product offerings, marketing tactics, and distribution channels. Comparing these aspects with our client's strategies can reveal gaps or opportunities. **3. Evaluate Internal Operations**: Delve into the client's internal operations focusing on the frozen food section. This involves assessing the supply chain, inventory management, procurement costs, pricing strategies, and in-store placement. The goal is to identify inefficiencies or cost drivers that are impacting margins. **4. Financial Analysis**: Conduct a detailed financial analysis to understand the profit margin trends. This includes analyzing revenue trends (e.g., changes in sales volume, average selling price) and cost trends (e.g., cost of goods sold, operating expenses) specifically related to the frozen food section. **5.Customer Insights**: Gather insights into customer preferences and behaviors regarding the frozen food section. Understanding if there have been changes in customer satisfaction, preferences, or shifts towards competitors' offerings is crucial.
Test framework	
Gotcha, let's talk about pricing a little more. If pricing were an issue for the client, what could be driving this?	If pricing is contributing to the decline in profitability in the frozen food section for the client, several factors could be at play. Here are some potential drivers: **1.Competitive Pricing Pressure**: The client's competitors might have adopted aggressive pricing strategies, such as lowering prices to gain market share or running frequent promotions. This can put pressure on the client to lower prices, thereby reducing margins. **2.Cost Structure Misalignment**: The client's cost structure might be higher than competitors', leading to challenges in competitively pricing their products without eroding margins. Higher procurement costs, inefficiencies in the supply chain, or higher overhead costs could contribute to this issue. **3.Price Sensitivity of Customers**: The target customer segment might be highly price-sensitive, valuing price over quality or brand loyalty. ...

Consulting Case Interview: Supermarket Frozen Foods [2/2]

Task: Test how well *ChatGPT* can perform on a Boston Consulting Group case interview (prompt, data exhibits, etc.) compared to a human candidate. The responses to each section were evaluated against typical responses from our scoring rubric.

Data exhibit #1

Prompt

Great, now let's get into some data . What do you takeaway from this chart? In particular, what is the change in revenue by product from 2016 to 2017?

Revenues and costs by frozen food category

Sales by category (average per store)

	Units (K)		Price ($/unit)	
	2016	2017	2016	2017
Frozen pizza	10	12	4.50	3.50
Frozen entrees	6	6.6	3.50	3.25
Other	5.5	5	4.00	4.50

Cost by category

	Cost ($ / unit)	
	2016	2017
Frozen pizza	1.50	1.50
Frozen entrees	1.00	1.00
Other	3.90	3.90

ChatGPT

The table reveals the following changes in revenue by product from 2016 to 2017:

•**Frozen pizza**: Revenue decreased from $45,000 to $42,000, resulting in a change of -$3,000.

•**Frozen entrees**: Revenue slightly increased from $21,000 to $21,450, resulting in a change of +$450.

•**Other**: Revenue increased from $22,000 to $22,500, resulting in a change of +$500.

In summary, the revenue from frozen pizzas has declined, despite an increase in units sold, likely due to a significant price reduction per unit. On the other hand, frozen entrees and other categories saw a modest increase in revenue, with prices remaining stable or increasing, and units sold either increasing slightly or decreasing less than the increase in price.

Data exhibit #2

Prompt

Okay nice work! Let's dive into the frozen pizza category. We have some additional data. What do you takeaway from this?

Frozen pizza sales by brand

Pizza sales by brand (average per store)

	Units (K)		Revenue ($K)	
	2016	2017	2016	2017
Brand A	5	3.5	24.5	17.3
Brand B	3	2.5	14.7	12.4
Value Brand	2	6.0	6.0	12.0

ChatGPT

The data for frozen pizza sales by brand between 2016 and 2017 indicates the following:

•**Brand A**: Sales revenue decreased by $7.2K despite a small increase in revenue per unit, from $4.9 to approximately $4.94. This suggests that the decrease in revenue is due to a reduction in units sold (from 5K to 3.5K).

•**Brand B**: Sales revenue also decreased by $2.3K with the revenue per unit showing a negligible increase from $4.9 to $4.96. Like Brand A, this decline in revenue is due to fewer units sold (from 3K to 2.5K).

•**Value Brand**: In contrast, the Value Brand experienced a significant increase in both units sold (from 2K to 6K) and sales revenue (from $6K to $12K). However, it's noteworthy that the revenue per unit actually decreased from $3 to $2.

Overall, the total number of units sold for frozen pizzas increased from 2016 to 2017 (from 10K to 12K units), as seen in the previous data. However, the revenue per unit decreased significantly for the client's premium brands (Brand A and Brand B), whereas the Value Brand doubled its units sold despite a decrease in revenue per unit. The shift towards the Value Brand at a lower price per unit might be a strong contributor to the overall decline in the frozen pizza category revenue.

Concluding summary

Prompt

Perfect! Okay now can you summarise your findings from this conversation? Imagine you run into the CEO of our client in the elevator (2 min) and need to explain to him/her what you've found.

ChatGPT

Good afternoon! I've had a chance to analyze the frozen food section, specifically looking at the profitability decline. It appears that while overall unit sales for frozen pizzas have increased, the revenue has not followed suit due to a significant drop in the price per unit. Your premium brands, A and B, have seen reduced sales volumes and only a slight uptick in revenue per unit, which is not enough to offset the reduced pricing. On the other hand, your Value Brand has doubled its sales volume at a much lower price point, which suggests that customers are shifting towards more cost-effective options. This pivot to value-based purchasing could be driving the overall profit decline in the frozen pizza category. We should consider revisiting the pricing strategy for the premium brands and possibly look into how we can enhance the value proposition to better meet customer expectations.

Figure 4 : ChatGPT Experiment #3 – M&A Buyside Due Diligence, 2024

<u>M&A Buyside Due Diligence</u>: Stride Rite & Saucony, Inc. [1/4]

Task: ChatGPT asked to perform all of the core analyses of the simplest possible buyside due diligence. This includes building capitalisation tables, combining income statements and balance sheets, calculating discounted cash flow, and recommending price per share.

Prompt	ChatGPT
Explain the exercise	
We are going to run through an M&A due diligence exercise from Stanford GSB's Financial Modeling course. You are an Associate in a large investment bank. Your team has been hired by Stride Rite's board to advise the company on its potential acquisition of Saucony. In particular, the board is looking for your opinion on price, how the transaction should be structured, and how the market is likely to react. I will provide the context prompt and underlying data for both buyer and target companies, then ask you to prepare a series of due diligence analyses and eventually provide a recommendation. Please take your time, respond as concisely as you can, and explain your reasoning step-by-step.	Absolutely, I'm ready to assist you through this M&A due diligence exercise. Please go ahead and provide the context prompt, the underlying data for both the buyer (Stride Rite) and the target company (Saucony), and any specific analyses or information requests you have in mind. This can include financials, market positioning, strategic fit, synergies, and any other relevant data points that will inform our analyses and recommendations.
Extract & organise specific data	
Here is the context brief with underlying data contained in exhibits and the attached financial model. context_brief_saucony_stride.docx Document financial_model_saucony_stride.pdf PDF Let's begin by building capitalization tables for Saucony and Stride Rite to calculate their total enterprise value at the time of the offer. Use the treasury stock method to calculate fully diluted shares outstanding based on the share, option, and warrant information contained in Exhibits 8 and 9. *Answer table below (not provided to model)*	**Treasury Stock Method** The treasury stock method assumes that options and warrants are exercised at their exercise price, and the company uses those proceeds to buy back shares at the current market price. This method calculates the net increase in shares outstanding if all in-the-money options and warrants were exercised. **Calculating for Saucony** •**Class A Common Shares**: 2,520,647 •**Class B Common Shares**: 4,176,593 •**Options and Warrants**: We will calculate the number of shares resulting from the exercise of options and warrants, considering their exercise prices and the current share price. Once exercised, options convert into Class B common shares, and warrants convert into Class A common shares. Saucony Options & Warrants … … **Calculating for Stride Rite** •**Common Shares Outstanding**: 36,209,637 •**Options**: Similar to Saucony, we'll calculate the dilutive effect of the options based on their exercise prices and the current share price. Stride Rite Options … … After calculating the fully diluted shares outstanding using the treasury stock method: •Saucony's fully diluted shares outstanding are approximately 7,530,202 •Stride Rite's fully diluted shares outstanding are approximately 37,530,688.

Saucony

Share price	$23.00		
Common shares	6,697,240		
Total shares from options (TSM)	914,529		
Total shares from warrants (TSM)	0		
Fully diluted shares	7,611,769		
Common shares			
Class A common	2,520,647		
Class B common	4,176,593		
Total common	6,697,240		
	Options	Strike price	Net shares
Tranche 1	88,892	$6.93	62,108
Tranche 2	118,972	$5.49	89,707
Tranche 3	520,343	$6.18	254,414
Tranche 4	101,862	$4.10	85,313
Tranche 5	650,270	$8.80	401,476
Tranche 6	71,029	$8.15	66,267
Total shares from options	1,346,946		914,529
Warrants	130,131	$23.50	0

Stride Rite

Share price	$12.04		
Common shares	36,209,637		
Total shares from options (TSM)	1,321,051		
Fully-diluted shares	37,530,688		
	Options	Strike price	Net shares
Tranche 1	509,823	$6.00	255,758
Tranche 2	1,191,051	$6.58	540,128
Tranche 3	1,082,030	$7.89	372,959
Tranche 4	543,549	$9.82	100,038
Tranche 5	897,290	$11.34	52,168
Total shares from options	4,222,743		1,321,051

M&A Buyside Due Diligence: Stride Rite & Saucony, Inc. [2/4]

Task: ChatGPT asked to perform all of the core analyses of the simplest possible buyside due diligence. This includes building capitalisation tables, combining income statements and balance sheets, calculating discounted cash flow, and recommending price per share.

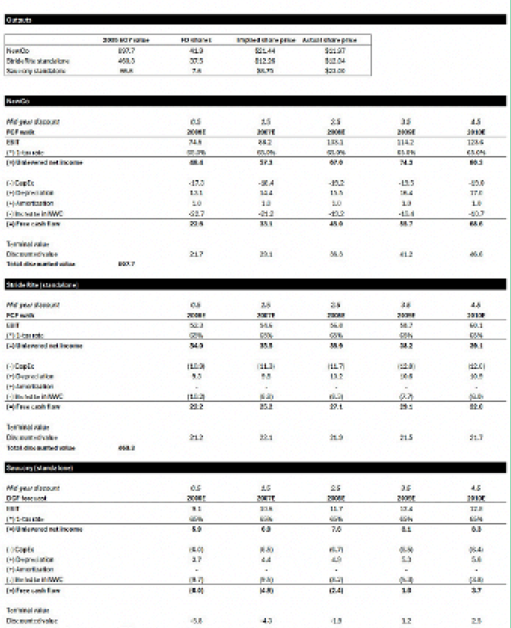
Prompt

Discounted Cash Flow

Perfect, now let's value both Saucony and Stride Rite as stand-alone businesses using a basic discounted cash flow (DCF) approach. You may also assume that inventories, accounts receivable and accounts payable are the only net working capital balances that will change. Inventory turns, A/R days, and A/P days should be held constant at 2005A levels. We will ignore the impact of seasonality and assume that cash flows occur evenly across the year using a mid-year convention (so 2006 cash flows will be discounted by 0.5 years, 2007 cash flows by 1.5 years, etc.).

Calculate terminal value using a TEV / EBIT multiple. Based on the comparable company analysis in Exhibit 10, assume 9x TEV / EBIT multiple is appropriate. Assume discount rate of 9%. Translate your DCF value to an implied share price for each scenario based on an appropriate fully diluted share count. Compare this value with the pre-merger market prices of SRR and SCNY found in Exhibits 8 and 9.

Answer table below (not provided to model)



Outputs

	2005 DCF value	FD shares	Implied share price	Actual share price
NewCo	897.7	41.9	$21.44	$21.97
Stride Rite standalone	468.3	37.5	$12.28	$12.04
Saucony standalone	66.6	7.6	$8.75	$23.00

NewCo

Mid year discount		0.5	1.5	2.5	3.5	4.5
FCF build		2006E	2007E	2008E	2009E	2010E
EBIT		74.6	88.2	105.1	114.2	123.6
(*) 1-tax rate		65.0%	65.0%	65.0%	65.0%	65.0%
(=) Unlevered net income		48.6	57.3	67.0	74.3	80.3
(-) CapEx		-17.3	-18.4	-15.2	-15.5	-15.0
(+) Depreciation		13.1	13.4	15.5	16.4	17.0
(+) Amortization		1.0	1.0	1.0	1.0	1.0
(-) Increase in NWC		-22.7	-21.2	-10.2	-15.4	-10.7
(=) Free cash flow		22.8	33.1	45.0	58.7	68.6
Terminal value						
Discounted value		21.7	30.1	36.5	41.2	46.6
Total discounted value	897.7					

Stride Rite (standalone)

Mid year discount		0.5	1.5	2.5	3.5	4.5
FCF build		2006E	2007E	2008E	2009E	2010E
EBIT		52.3	54.6	56.0	58.7	60.1
(*) 1-tax rate		65%	65%	65%	65%	65%
(=) Unlevered net income		34.0	35.5	36.4	38.2	39.1
(-) CapEx		(10.0)	(11.3)	(11.7)	(12.0)	(12.0)
(+) Depreciation		9.5	9.8	13.2	10.6	10.9
(+) Amortization		-	-	-	-	-
(-) Increase in NWC		(11.2)	(6.2)	(6.3)	(2.7)	(6.0)
(=) Free cash flow		22.2	25.2	27.1	29.1	32.0
Terminal value						
Discounted value		21.2	22.1	21.9	21.5	21.7
Total discounted value	468.3					

Saucony (standalone)

Mid year discount		0.5	1.5	2.5	3.5	4.5
DCF forecast		2006E	2007E	2008E	2009E	2010E
EBIT		9.1	10.6	11.7	12.4	12.8
(*) 1-tax rate		65%	65%	65%	65%	65%
(=) Unlevered net income		5.9	6.8	7.6	8.1	8.3
(-) CapEx		(6.0)	(6.8)	(6.7)	(6.6)	(6.4)
(+) Depreciation		3.7	4.4	4.9	5.3	5.6
(+) Amortization		-	-	-	-	-
(-) Increase in NWC		(8.7)	(9.3)	(8.2)	(5.8)	(3.8)
(=) Free cash flow		(8.0)	(4.8)	(2.4)	1.8	3.7
Terminal value						
Discounted value		-5.8	-4.3	-1.9	1.2	2.5
Total discounted value	66.6					

ChatGPT

To perform a discounted cash flow (DCF) valuation for Saucony and Stride Rite as standalone businesses, we will need to follow these steps:

1. Forecast free cash flows (FCFs) from the provided financial statements.
2. Calculate the terminal value using a 9x TEV/EBIT multiple.
3. Discount the FCFs and terminal value to the present value using a 9% discount rate.
4. Translate the DCF value to an implied share price.
5. Compare this value with the pre-merger market prices.

First, let's forecast the free cash flows for both companies based on the provided financials. We will use the mid-year convention for discounting cash flows and assume that cash flows occur evenly across the year. For working capital, we will hold inventory turns, A/R days, and A/P days constant at 2005A levels. Let's begin with the calculation of FCFs for each company from 2006E to 2010E, after which we will calculate the terminal value and proceed with the DCF valuation.

We'll start by defining the free cash flow as:
FCF=EBIT×(1−Tax Rate)+Depreciation−CapEx−ΔNWCFCF=EBIT×(1−Tax Rate)+Depreciation−CapEx−ΔNWC
where ΔNWCΔNWC represents the change in net working capital, calculated based on the changes in inventories, accounts receivable, and accounts payable.

We will first compute the forecasted FCFs for Stride Rite. Let's extract the necessary data to start these calculations.

The DCF value for Stride Rite as a standalone business is approximately $380.49 million. When translated to an implied share price based on the fully diluted share count, the value is approximately $0.0105 per share. This share price seems to be off, indicating there may be an error in the calculation, possibly with units or share count. Given that we are working with millions in the cash flows, the per share price should be much higher. Let's correct this by re-evaluating the share count and ensuring that we're working with the correct units.

Stride Rite's fully diluted shares outstanding from Exhibit 8 should be 36,209,637 shares. We'll correct this and then compare it with the pre-merger market price. Let's redo this calculation.

The corrected DCF value for Stride Rite as a standalone business is approximately $380.49 million. The implied share price based on the DCF valuation and the fully diluted share count is approximately $10.51 per share.

Next, we will perform the same DCF valuation for Saucony. We'll use the same assumptions, including the discount rate, the terminal value multiple, and the tax rate. Let's proceed with Saucony's DCF valuation. The DCF value for Saucony as a standalone business is approximately $777.51 million. The implied share price based on the DCF valuation and the Class A share count is approximately $308.46 per share.

M&A Buyside Due Diligence: Stride Rite & Saucony, Inc. [3/4]

Task: ChatGPT asked to perform all of the core analyses of the simplest possible buyside due diligence. This includes building capitalisation tables, combining income statements and balance sheets, calculating discounted cash flow, and recommending price per share.

Prompt	ChatGPT
NewCo financial statements	
Thanks! Now let's create pro forma integrated financial statements for the combined Saucony & Stride Rite business ("NewCo") from 2005PF (pro forma) to 2010E. For simplicity, assume that the transaction closes December 31, 2005. In addition to stand-alone projections, Stride Rite has provided us with the attached synergy assumptions. Assume a pro forma tax rate of 35%. Stride Rite will be required to pay interest on the drawn portion of its credit facility at the rate of 4.5 percent per year on the average outstanding balance. As the credit facility is a revolving credit facility, the company will be required to pay a fee for any committed capital that is not drawn. As a result, there is interest expense on this undrawn portion equal to the average unused credit (Total Availability – Used Portion) × 0.5% per year. For the purpose of this model, assume that Stride Rite intends to pursue a fixed debt policy going forward and will not repay any initial borrowings used to fund the transaction. The company must maintain a minimum cash balance of at least $10M which will earn interest at 1% per year on the average cash balance. All excess cash generated above the minimum cash balance going forward will be paid as dividends to shareholders. In addition, assume that the only net working capital balances for the combined company that change are inventories, receivables, and payables. A/R days (calculated using Revenue) are constant at 54.32; A/P days (calculated using COGS) are constant at 20.13; and inventory turns (calculated using COGS) are constant at 4.14x. NewCo will issue stock-based compensation at the total combined rate forecast for the individual firms ($3.5M initially growing by $0.8M), with expected dilution based on total forecasted share price. Assume that depreciation is a constant 1.5 percent of sales through the forecast period, and that net capital expenditures are 5.0 percent of any increase in sales.	To create the pro forma integrated financial statements for NewCo from 2005PF to 2010E, incorporating the actual financial data for both Saucony and Stride Rite, we will perform the following steps: 1. Combine the historical income statements of Stride Rite and Saucony from 2001 to 2005 to get the base year 2005PF (pro forma) figures. 2. Project the income statements from 2006E to 2010E using the forecast growth rates and incorporating the synergies as per the assumptions provided. 3. Include the financing costs, such as interest expense and commitment fees, as described. 4. Maintain a minimum cash balance policy and compute excess cash dividends accordingly. 5. Factor in stock-based compensation and its forecasted growth. 6. Account for depreciation as a constant percentage of sales and net capital expenditures as a percentage of the increase in sales. 7. Adjust the net working capital based on the given A/R days, A/P days, and inventory turns for the combined entity. ... ... Given the actual financial data for both companies from 2001 to 2005 and the forecasts for 2006 to 2010, we'll start by calculating the base year 2005PF combined income statement. Afterward, we'll move on to projecting the financials for the subsequent years based on the growth rates, synergies, and other assumptions. The combined pro forma financials for NewCo (Stride Rite and Saucony) for the base year 2005 are as follows (in millions): •Revenue: $791.6 •COGS: $486.0 •SG&A: $237.5 •EBIT: $68.1 •Net Income: $42.4 ... ... The projected financials for NewCo (Stride Rite and Saucony combined) for the years 2006 to 2010, taking into account the growth rates of Stride Rite and the synergistic effects, are as follows (all figures are in millions): •**2006**: Revenue: $835.14, COGS: $512.73, SG&A: $250.56, EBIT: $71.85, Net Income: $46.70 •**2007**: Revenue: $831.18, COGS: $505.20, SG&A: $248.13, EBIT: $77.85, Net Income: $50.61 •**2008**: Revenue: $827.22, COGS: $497.71, SG&A: $245.71, EBIT: $83.80, Net Income: $54.47 •**2009**: Revenue: $823.26, COGS: $490.28, SG&A: $243.30, EBIT: $89.69, Net Income: $58.30 •**2010**: Revenue: $815.35, COGS: $480.56, SG&A: $239.73, EBIT: $95.06, Net Income: $61.79 Please note that these figures assume a linear decrease in COGS margin and SG&A as a percentage of sales per the synergy anticipation and a constant tax rate of 35%. ... *continued on next page*

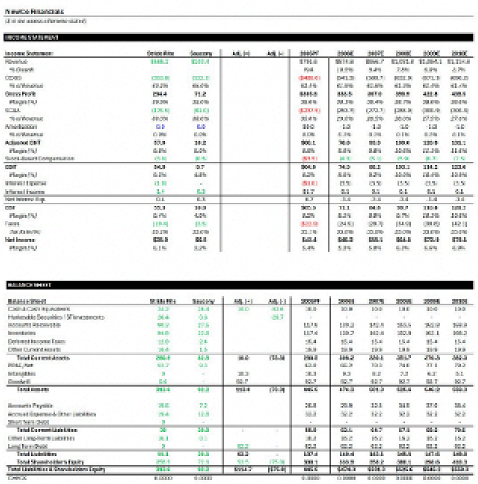
<u>M&A Buyside Due Diligence</u>: Stride Rite & Saucony, Inc. [4/4]

Task: ChatGPT asked to perform all of the core analyses of the simplest possible buyside due diligence. This includes building capitalisation tables, combining income statements and balance sheets, calculating discounted cash flow, and recommending price per share.

Prompt	ChatGPT

NewCo financial statements (cont.)

Answer tables below (not provided to model)



ChatGPT:

The updated projected financials for NewCo from 2006 to 2010 with all adjustments are as follows (all figures are in millions):

Year	2006	2007	2008	2009	2010
Revenue	835.14	831.18	827.22	823.26	815.35
COGS	512.73	505.2	497.71	490.28	480.56
SG&A	250.56	248.13	245.71	243.3	239.73
EBT	67.43	73.44	79.39	85.28	90.64
Net Income	41.55	44.94	48.29	51.59	54.56
Interest on Credit Facility	3.96	3.96	3.96	3.96	3.96
Commitment Fee	0.56	0.56	0.56	0.56	0.56
Cash Interest Earned	0.1	0.1	0.1	0.1	0.1
Stock-based Compensation	3.5	4.3	5.1	5.9	6.7
Depreciation	12.53	12.47	12.41	12.35	12.23
CapEx	2.18	1.98	1.78	1.58	1.19
Dividends	31.55	34.94	38.29	41.59	44.56

These projections are based on the given assumptions, including a fixed debt policy (no repayment of initial borrowings), a minimum cash balance of $10 million, a pro forma tax rate of 35%, and specific metrics for net working capital. The net income figures take into account all of these factors, and the excess net income after maintaining the minimum cash balance is considered available for dividends. The dividends represent the excess cash generated that is paid out to shareholders annually.

Final recommendation #1

Prompt: Great, okay final question: how much should SRR be willing to pay, in terms of $./share, for Saucony (assuming no integration risk)? Be specific: given 70% cash consideration, what is the maximum price per share SRR should be willing to pay?

ChatGPT: The maximum price per share Stride Rite should be willing to pay for Saucony, given a 70% cash consideration and assuming no integration risk, is approximately $28.95. This price ensures that the transaction is EPS neutral for Stride Rite, meaning it neither accretes nor dilutes Stride Rite's standalone EPS.

Final recommendation #2

Prompt: Interesting, can you try that exercise again? Do you get the same answer of $28.95 / share?

ChatGPT: Upon recalculating, the maximum price per share that Stride Rite should be willing to pay for Saucony, assuming a 70% cash consideration and no integration risk, is approximately $96.57. This figure represents the share price at which the acquisition would be EPS neutral to Stride Rite.

This result differs from the previous calculation, indicating an error in the earlier computation. It is crucial when performing financial analyses to ensure that all variables and equations are accurate and that results are consistently verified.

Addendum: Is our AI Strategy Assistant ready to be promoted?

Motivation: OpenAI's o1 models represent a new state-of-the-art for automated reasoning. Benchmarks like ARC and MMLU are no longer sufficient; the LLM community now tests these models on PhD math and science questions.[1]

In the original article,[2] we tested the capabilities of OpenAI's ChatGPT-4 for strategy tasks and found it was great for simple data comprehension (e.g., summarizing market research), but lacked the reasoning required to be helpful for analysis tasks (e.g., discounted cash flow). With the release of o1-preview, we were curious to rerun our case interview and due diligence experiments and see to what extent the upgrades translated into improved performance on strategy analyses. In other words: is our AI Strategy Assistant ready to be promoted?

Exp: Consulting Case Interview. o1-preview outperformed GPT-4 on this experiment, replicating the performance of a top candidate in four out of five questions. The improvement came primarily from the upfront framing questions and reflects the difference between a "weak pass" to a "strong pass." This result was largely expected given that the case interview questions rely heavily on sequential reasoning.

Table 1: Consulting Case Interview Results

#	Task	GPT-4 Result	o1-preview Result	Commentary
1	Structuring the problem given the initial prompt.	Insufficient	Satisfactory	Improved result! Identified relevant revenue and cost drivers (not true for GPT-4), but without providing a framework to demonstrate the approach would be comprehensive.
2	Ask candidate to think through drivers of a pricing issue.	Satisfactory	Exceptional	Improved result! Identified all responses expected of top candidates (only 50% for GPT-4), as defined by the grading rubric.
3	Data exhibit #1, what are your takeaways from this chart?	Exceptional	Exceptional	Similar result. Arithmetic correct. Also covered "so what" elements expected of top candidates, e.g., profit loss driven by pizza margin.
4	Data exhibit #2, what are your takeaways from this chart?	Exceptional	Exceptional	Similar result. Arithmetic correct. Also covered "so what" elements expected of a top candidate, e.g., likely cannibalization.
5	Summarize findings to client CEO.	Exceptional	Exceptional	Similar result. Clear and well-structured summary. First sentence answered question from prompt, i.e., cause for loss of profitability.

Table 1: Insufficient = candidate failed this section; Satisfactory = response met minimal requirements to pass this section; Exceptional = response expected of a top candidate. Generally, candidates need to pass every section.

The only task for which o1-preview scored "satisfactory" (instead of "exceptional") was the upfront problem framing. While this was already an improvement over the GPT-4 response (it identified hypothesis-driven set of revenue and cost drivers), it did not provide a unified framework to demonstrate that the proposed approach would be comprehensive.

Finally, as with GPT-4, the o1-preview model did not demonstrate the sorts of human behavioural qualities we expect of top candidates, despite pre-prompting. In particular, we want candidates to exhibit an inquisitive, curious mindset and "drive the interview" by proposing hypothesis-driven next steps to the interviewer. This is not necessarily a bug—we might prefer concise responses over extended thoughts on what to do next—just a limitation.

Exp: Buyside Due Diligence. o1-preview was substantially better than GPT-4 for the due diligence experiment, a result we did not expect. The experiment was designed intentionally to push the boundaries of this technology by combining complex multistep reasoning with large context window. In our original paper, as expected, we showed that GPT-4 largely failed this exercise—to the point where it was net unhelpful even as an assistant guided by a human.2 By contrast, o1 replicated the performance of a top candidate for three out of five tasks and would be an excellent assistant.

Table 2: Buyside Due Diligence Results

#	Task	GPT-4 Result	o1-preview Result	Commentary
1	Extract data, and suggest approach to evaluate overall due diligence prompt.	**Exceptional**	**Exceptional**	Similar result. Understood prompt and laid out a structured series of analyses that would be sufficient to complete exercise.
2	Build capitalization tables to calculate enterprise value for each company.	**Exceptional**	**Exceptional**	Similar result. Capitalisation tables and fully diluted share calculations were accurate.
3	Discounted cash flow analysis for target and buyer as stand-alone businesses.	**Insufficient**	**Exceptional**	Improved result! Correctly computed PV of FCFF and logic for implied share price was correct. Only small error in discounting of terminal value (4.5 vs. 5 years). This could be considered a reasonable "design choice" rather than an error.
4	Create pro-forma integrated financial statements for combined Saucony & Stride Rite business ("NewCo").	**Insufficient**	**Satisfactory**	Improved result! GPT-4 did not produce useful output for this task. o1-preview produced a perfect balance sheet and a 90% correct income statement. Only minor errors (e.g., depreciation).
5	Final recommendation for what buyer should be willing to pay. Question was asked twice to test for hallucinations.	**Insufficient**	**Satisfactory**	Improved result! GPT-4 hallucinated wildly. o1-preview proposed a share price within the range of acceptable values. While assumptions are stated clearly, the "so what" and judgment expected of a top response are still missing.

Table 2: Insufficient = either did not answer the question or arithmetic mostly incorrect; Satisfactory = response answers question, some minor arithmetic errors or assumption differences; Exceptional = response expected of a top analyst.

What surprised us most was how gracefully o1-preview handled complex contextual information, including tables for income statements, balance sheets, share information, synergy estimates, and a comparables analysis. The model's results were not perfect (e.g., its depreciation and terminal value results were inaccurate), but these mistakes were minor and would have been common even among capable students doing this assignment. Moreover, the o1 models no longer hallucinate. The final task asks the LLM to recommend a share price that our buyer should be willing to pay, then resubmits the prompt to test whether the LLM might change its mind. GPT-4 offered two wildly different answers, but o1-preview offered the same response up to a rounding error. Although this does not eliminate the need of users to sense check the model's results, it provides a more stable output that businesses can begin to build software checks and safeguards around.

One major caveat is that o1 models do not yet have access to the same user-friendly features that OpenAI has provisioned for GPT-4, such as file uploads, code execution, output schemas, and non-text modalities (e.g., image and video). This posed a potential problem for our due diligence experiment, since it inherently relied on contextual documents provided as file uploads. Fortunately, the context window size itself had not been limited (128,000 tokens, same as GPT-4 Turbo), and so we were able to work around the file upload constraint by running the code-based version of the experiment using OpenAI's "chat completions" API and passing the context documents as "markdown text." While this obviously limits the general usefulness of these models within a real-world business context (most business analysts and strategy consultants don't code), we expect that OpenAI will add such functionality in the coming months. Therefore, our results really indicate the potential usefulness of these models for strategic management activities once such functionality is readily available through the user interface.

Conclusions. Remarkably, since our original paper (written April 2024, published September 2024), the ability of LLMs to handle strategic tasks involving multistep reasoning and context-dependence has improved significantly. There are two main conclusions we can draw from our study.

1. LLMs can already assist us for strategy tasks. Our updated experimental results indicate that OpenAI's o1-preview model (without fine-tuning) would replicate the performance of a top candidate in a consulting case interview and produce a good useful draft effort of a buyside due diligence analysis. For us, this means LLMs can already begin to play a useful role as an assistant for strategy tasks, which was not true a few months ago.

2. Having a human-in-the-loop still matters. The due diligence experiment would not have been possible through pure automation. In particular, the model required repeated prompting to produce a complete answer, and without a human carefully sense checking the calculations, an analyst might have missed correcting different assumptions that the model was making.

Outlook: what does this mean? In the original article, we proposed that leadership teams need to decide "at what point should I start seriously investing?"[2] The earlier you invest in developing significant software engineering around existing LLMs, you run the risk of building solutions that will become redundant over time, as performance natively increases with newer model versions (e.g., GPT-5, Claude-4). This is still the right question, but we are now more bullish about starting rather than waiting a few years for the technology to mature further. This is because the quality of the existing off-the-shelf models is already sufficient to be useful for many strategy tasks, and the software engineering you would need to put around it is orthogonal to the quality of the responses (which may further improve). For example, many of the systems we are building have significant input and output data-processing pipelines, but already minimal model fine-tuning. It would be the same here, for most strategic management applications.

The performance jump also points to a more uncomfortable question: will future updates remove the need to have humans in the loop for strategy tasks? We think this is still a long way off. The ability to anticipate the behaviour of other humans and make strategic decisions under uncertainty, which is fundamentally what investors do when they build a model of what a company may become and could do in the future, are not tested by our experiments but represent a big leap in complexity. That said, strategy consultants and investment analysts should certainly no longer be doing their work unassisted by these tools.

Reference Documentation

1. OpenAI. (2024). Learning to reason with LLMs. Simons Institute for the Theory of Computing. https://live-simons-institute.pantheon.berkeley.edu/talks/noam-brown-openai-2024-09-26

2. Lechner, C., Lang, N., Handschuh, S., Bouffault, O., & Cooper, J. (2024, September 12). Can GenAI do your next strategy task? Not yet. California Management Review.

3. OpenAI. (2024). OpenAI SDK (Version 1.2.3) [Software]. OpenAI. https://github.com/openai/openai-python

---

Christoph Lechner Follow

Christoph is a Chaired Professor of Strategic Management at the University of St. Gallen, and Director of its Institute of Management & Strategy. He also works as Senior Advisor for the Boston Consulting Group (BCG). He was board member of Helvetia Group for 16 years and Vice Chairman of Hügli Group for 17 years. He received his PhD in Strategic Management from University of St. Gallen.

Nikolaus Lang Follow

Dr. Nikolaus S. Lang is inter alia the Global Leader of BCG Henderson Institute and has supported clients on globalization and mobility-related topics in more than 50 countries. Nikolaus has led BCG's collaboration with the World Economic Forum, dedicated to shaping the future of mobility. He has written over 100 publications on various contemporary issues (e.g., mobility, ecosystems, and geopolitics) and is a coauthor of BCG's book "Beyond Great," which describes how the world is transforming due to social tension, economic nationalism, and technological revolution.

Siegfried Handschuh

Prof. Dr. Siegfried Handschuh is a Full Professor of Data Science and Natural Language Processing at the Institute of Computer Science, University of St. Gallen. He studied Computer Science at the University of Ulm and Information Science at the University of Konstanz. His research specializes in Natural Language Processing, with a particular emphasis on Large Language Models (LLMs), as well as quantitative Data Science and generative AI. Prof. Handschuh has authored several notable books, including "Mathematical Foundations of Data Science" and a comprehensive work on generative AI.

Olivier Bouffault Follow

Olivier Bouffault is a Senior Partner and Managing Director in the Paris office. He is the Regional Leader of BCG X for EMESA Region (Europe, Middle-East, South America, Africa). BCG X is BCG's tech build & design unit, which gathers 3000 practioners across the world. He is a member of the Operations practice leadership team and leads the Operations Analytics topic globally.

Julian Cooper Follow

Julian is a Senior Data Scientist with BCGX, specialized in applying Bayesian Inference to model physical systems. His interest in Deep Learning topics stems from graduate school research into using Graph Neural Networks for Automated Theorem Proving as part of the Thrun Lab (SAIL). Julian holds a Masters in Computational and Mathematical Engineering from Stanford University.