15.451 - Proseminar in Capital Markets

Human v.s. AI in Financial Analysis

Supervisors:

Sebastien Page

Stefan Hubrich

Ramon Richards

Professor:

Prof. Mark Kritzman

Presented by:

Maya Walcher

Nicholas Wong

Ahmed Wakrim

David Xiang

Iris Xiao

Rei Yamahara

Cathy Shan

Jinghan Xu

December 2024

Academic Year 2024/2025

Contents

1 Introduction

The integration of artificial intelligence (AI) into financial analysis has opened new avenues for enhancing the productivity and accuracy of equity research. Among these advancements, Generative AI (GenAI), particularly models like GPT, presents a promising frontier for augmenting analysts' capabilities. However, the extent to which such models can effectively emulate or even replace human expertise remains a topic of significant debate.

In our work, we empirically investigate the utility of Generative AI in producing sell-side equity research reports. Focusing on two representative companies, Microsoft (MSFT) and Chipotle, we implement a multi-tiered empirical testing framework to evaluate GPT models with varying degrees of training and prompting sophistication. By employing a rigorous model evaluation rubric, we assess the generated reports against key qualitative criteria of stability, utility, discernibility, and rating consistency with human analysts to determine if the AI-generated reports pass muster. Since utility is a rather subjective rubric item, it is further deconstructed into clarity, accuracy, comprehensiveness, depth of analysis, and actionability, ensuring a holistic evaluation. Complementing the qualitative assessment, we perform a natural language processing (NLP) analysis comparing the equity research report generated by our most sophisticated GPT model against the equity research report of a human analyst using Term Frequency Inverse Document Frequency (TF-IDF) keyword weighting and text cosine similarity. This quantitative approach enables a direct comparison of the generated report with those of human analysts, offering insights into the alignment in thematic content, analytical depth, and lexical similarity.

We propose that while GPT models exhibit considerable potential, their limitations in nuanced comprehension, contextual interpretation, and need for coaching underscore the irreplaceable value of human expertise.

2 Literature Review

Previous works have studied the utility of GenAI for both quantitative as well as qualitative financial reporting and analysis. We begin with the in-depth analysis of the performance of LLMs on the CFA exam to establish a performance baseline before we proceed

with a survey of relevant literature studying the use of GenAI tools for quantitative and qualitative financial analysis.

2.1 LLM Models on the CFA Exam

The Chartered Financial Analyst (CFA) certification, widely recognized in the financial industry, is structured into three levels, each progressively testing candidates on increasingly complex aspects of financial knowledge and application. Level I primarily assesses foundational concepts and straightforward calculations, Level II transitions to case-based multiple-choice questions (MCQs) requiring applied knowledge, while Level III focuses on essay questions, emphasizing structured reasoning and real-world decision-making. According to the findings presented by Mahfouz et al. [6], state-of-the-art Large Language Models (LLMs), such as GPT-4 Turbo and Claude 3 Opus, exhibit robust performance in Levels I and II, achieving estimated passing scores based on a 60–70% threshold. However, the essay-based structure of Level III highlights significant limitations in LLMs' ability to produce coherent, contextually grounded written responses, resulting in no model successfully passing this level, as illustrated in Figure 1. The inability of LLMs to meet Level III requirements underscores challenges in structured reasoning, contextual comprehension, and instruction-following capabilities.

2.2 Quantitative Financial Reporting

According to Kim et al. in [4], Papasotiriou et al. in [7], and Fatouros et al. in [2], Large Language Models (LLMs), such as GPT-4, have demonstrated significant potential in financial analysis by effectively integrating diverse data sources—including financial fundamentals, news, historical prices, and macroeconomic data—to predict earnings changes, assign equity stock ratings, and generate signals for trading strategies. Here are several insights from these papers.

In the use of GPT models for quantitative financial analysis, the importance of Chain-of-Thought (CoT) prompting is paramount. Kim et al. in [4] put forward that CoT prompting significantly enhances GPT-4's predictive performance. By utilizing CoT reasoning to analyze trends and ratios from financial statements, the model achieved a prediction accuracy of 60.31%, surpassing both simple prompts (52.33%) and the perfor-

Provider	Model	Level I		Level II		Level III	
		L	U	L	U	L	U
OpenAI	GPT-3.5 Turbo	✓	✗	✗	✗	✗	✗
	GPT-4 Turbo	✓	✓	✓	✓	✗	✗
	GPT-4o	✓	✓	✓	✓	✗	✗
Anthropic	Claude 3 Opus	✓	✓	✓	✓	✗	✗
Mistral	Mixtral-8x7B	✓	✗	✗	✗	✗	✗
	Mixtral-8x22B	✓	✗	✓	✗	✗	✗
	Mistral Large	✓	✗	✓	✗	✗	✗
Google	Gemma 2B	✗	✗	✗	✗	✗	✗
	Gemma 7B	✗	✗	✗	✗	✗	✗
Meta	LLaMA 3 8B	✗	✗	✗	✗	✗	✗
	LLaMA 3 70B	✓	✗	✗	✗	✗	✗
	LLaMA 3 8B + RAG	✓	✗	✗	✗	✗	✗
	LLaMA 3 70B + RAG	✓	✓	✓	✗	✗	✗
Cohere	Command R+	✗	✗	✗	✗	✗	✗
Microsoft	Phi-3-mini	✓	✗	✗	✗	✗	✗
Ai2	OLMo 7B	✗	✗	✗	✗	✗	✗

Figure 1: LLM's Performance in CFA tests

Note: LLMs' ability to pass each CFA level, with lower bound score (L $\geq$ 60%) and upper bound score (U $\geq$ 70%). A checkmark (✓) indicates the LLM passes the exam according to the corresponding bound, while a cross (×) indicates failure.

mance of human analysts (52.71%). Similarly, Fatouros et al. in [2] emphasize the role of CoT prompting in analyzing diverse data sources, such as financial fundamentals, news, and macroeconomic factors, through tailored templates, highlighting the importance of step-by-step analysis and financial expertise in prompt design.

Under the right circumstances, GPT models are able to match cutting-edge deep learning methods in quantitative forecasting accuracy. Kim et al. in [4] highlight GPT-4's superior predictive accuracy, demonstrating that it is on par with specialized Artificial Neural Network (ANN) models (60.31% vs. 60.45%) but with a higher F1 score (63.45% vs.

61.62%). In parallel, Fatouros et al. in [2] showcase the significant performance of their MarketSenseAI framework, powered by GPT-4, which delivered cumulative returns of up to 72%, outperforming the S&P 100 by more than 30% during empirical testing.

There are also significant synergies between GPT models and human financial analysts. Kim et al. in [4] demonstrate the complementary nature of GPT and human predictions. Their study found that GPT provides incremental value, particularly in cases where human analysts exhibit biases or faced disagreements, while human analysts excelled in predicting outcomes for small or loss-making firms. Similarly, Papasotiriou et al. in [7] explore the performance of LLMs across various time horizons. They found that LLMs excelled in short-term predictions (1–3 months) by leveraging sentiment and news, whereas medium-term predictions (3–12 months) benefitted from fundamentals and sentiment-based setups. For long-term predictions (18 months), human analysts outperformed LLMs, reflecting the models' limitations in extrapolating data over extended periods and the forward-looking nature of qualitative financial data.

While Large Language Models (LLMs), such as GPT-4, have shown significant promise in quantitative financial analysis, several limitations remain. Predictions often lack granularity, focusing on high-level outcomes like earnings changes or stock rating tiers, which may fall short of the precision required for fine-tuned financial decision-making. Additionally, while LLMs excel at extracting and synthesizing data from diverse sources, their outputs are constrained by limited deep reasoning and fail to fully explore interconnections between factors, such as how variables interact or influence pricing.

2.3 Qualitative Financial Reporting

We synthesize the key insights from five papers by Blankespoor et al., Gupta et al., Zhou et al., Krause, and Rane exploring the applications, strengths, and limitations of GenAI for qualitative financial reporting. Prior works have identified the strengths of generative AI for reducing the time and effort to produce high quality qualitative financial reports.

Blankespoor et al. in [1] emphasize the ability of GenAI tools to lower costs by automating tasks, such as drafting and editing financial disclosures, which improve readability and consistency. In an adjacent field, Gupta et al. in [3] illustrate the application of GenAI to streamline government report generation - implying a transferable methodology in finan-

cial reporting via narrative synthesis. Similarly, Zhou et al. take this application further by introducing the multi-agent Chain of Thought (CoT) framework in FinRobot that automates narrative construction and narrative synthesis [9]. Through the automation of mundane financial reporting tasks, GenAI is able to reduce the reliance on human intervention without compromising report quality for automatable tasks. The scalability of GenAI systems also makes them invaluable in environments requiring real-time updates and large-scale report generation. Gupta et al. in [3] describe the role of GenAI in creating recurring reports with consistency, enhancing scalability of report generation. Zhou et al. in [9] describe FinRobot's dynamic data pipeline that allows the integration of real-time data updates into the report generation process.

GenAI enhances decision-making support for qualitative financial reporting by summarizing complex information into actionable insights. Zhou et al. describe the Concept-CoT Agent of FinRobot that mimics human analytical reasoning to interpret financial data and generate insights[9]. Utilizing GenAI tools for document summarization and to extract actionable insights enables financial managers to gain a nuanced understanding of qualitative aspects of lengthy financial reports such as competitive analysis and risk evaluation.

The narrative value of GPT models is further confirmed by the previously cited works in 2.2. Kim et al. in [4] highlights the narrative value of GPT-4's explanations derived from CoT reasoning. These narratives, when used as inputs in an ANN model, achieve an accuracy 59%, nearly matching GPT-4's performance, and have a superior F1 score of 65%, indicating that narrative insights are integral to GPT's predictive ability. Fatouros et al. in [2] further extend this concept by introducing a system for evaluating the quality of "buy" signal explanations generated by their model. They developed a ranking system on a 0–10 scale, which allowed these explanations to serve as an alternative approach for filtering and weighting stocks in portfolio construction, thereby enhancing the effectiveness of investment strategies.

There are however distinct downsides to using GenAI for qualitative financial analysis and reporting. GenAI systems often struggle with accuracy and may produce factually incorrect or contextually inappropriate content. Blankespoor et al. in [1] and Krause in [5] highlight the risk of hallucinations, where GenAI generates information that is factu-

ally incorrect or contextually irrelevant, which can undermine the credibility of financial reports. Zhou et al. underscore these concerns and emphasize the importance of human oversight in validating AI-generated outputs[9]. Since GenAI systems are trained on a set corpora of data, these hallucinations often occur as a result of training data limitations where the generalizability of patterns from its training data does not extrapolate well and leads to incorrect conclusions. These can also occur as a result of poor prompt engineering or prompt ambiguity, which can lead the AI to fabricate answers when instructions are not clear.

Related to the aforementioned point on hallucinations is the issue of ethical and regulatory concerns. Data bias and regulatory compliance issues are pervasive in GenAI, as outlined by Rane in [8] and Gupta et al. in [3]. Aligning AI-generated reports with evolving regulatory standards is thus a huge point of concern for firms seeking to adopt GenAI for qualitative financial reporting. These issues may lead to institutional resistance where GenAI adoption barriers are created in response to concerns about regulatory compliance, concerns about job displacement, and costs associated with adopting GenAI workflows [1]. As a result, robust validation frameworks are essential to address these concerns and ensure reliability of AI generated qualitative financial outputs [9].

3 Empirical Test on GenAI For Equity Report

We present the results of our three-layered empirical testing framework. To ensure a comprehensive evaluation of GPT's capabilities across varying business complexities, we selected Microsoft and Chipotle as representative companies. Microsoft, with its diversified portfolio spanning software, hardware, cloud computing, and artificial intelligence, offers a robust test case for analyzing how GPT models handle multi-faceted business narratives. In contrast, Chipotle, with its singular focus on fast-casual dining, provides a controlled environment to evaluate model performance in a more streamlined and narrowly defined business context. This juxtaposition enables a balanced assessment of GPT's versatility across different corporate structures.

3.1 Layer 1

Overview:

The layer 1 model for generating equity research reports involves using financial data exclusively from *10-K* and *10-Q* filings. These documents provide comprehensive insights into a company's financial performance, risk factors, management discussion, and outlook, which are essential for building accurate financial models and making investment recommendations. This report outlines the process of customizing GenAI, the methodology used, and the generated output when trained solely on *10-K* and *10-Q* filings to produce structured, professional equity research reports.

Input:

The core data for this equity research report generation comes from the *10-K* and *10-Q* filings, which contain vital information about a company's performance. These filings include detailed revenue and profitability trends, balance sheet and cash flow data, management's discussion and analysis, and identified risk factors.

Training Process:

GenAI's pre-trained capabilities allow it to analyze structured data and textual information from *10-K* and *10-Q* filings effectively. The model generates comprehensive equity research reports based on pre-defined templates. These templates provide a consistent framework for the reports, including sections such as the investment thesis, key highlights, valuation metrics, outlook, and recommendations. The financial analysis dives deeper, presenting trends in revenue, profit margins, and cash flow derived directly from the extracted data. The valuation section uses commonly accepted methods such as Price-to-Earnings (P/E) and Price-to-Book (P/B) ratios to assess the company's market value. The final investment recommendation synthesizes these analyses, providing a clear decision - *Buy*, *Hold*, or *Sell* - based on the data and insights from the reports.

Output:

The equity research report adhered to a predefined structure, ensuring a professional and consistent presentation of Microsoft Corporation's financial and strategic analysis. It began with an executive summary, which provided a clear *Buy* recommendation and a target price of $400, derived through a simple discounted cash flow (DCF) analysis. Key assumptions for this valuation included a 10% compound annual revenue growth rate, a

30% operating margin, and an 8% weighted average cost of capital.

The report offered a structured breakdown of Microsoft's core business segments. It highlighted the growth of the Productivity and Business Processes division, driven by demand for Microsoft 365 and AI-integrated tools like Copilot, alongside a 27% revenue increase in the Intelligent Cloud segment led by Azure. The More Personal Computing segment demonstrated resilience amid declining PC demand through innovations in gaming and Windows 11. These sections showcased the model's ability to generate a cohesive and organized analysis of financial data.

Challenges:

The report was well-structured and clear in its presentation; however, it lacked depth in key areas and contained elements of misinformation. The Layer 1 model output projected a 12-month target price of $2,300, significantly deviating from the real price of $60.76 in October 2024. This underscored the need for expertise-driven prompts and stricter regulation of information sourcing to improve accuracy and reliability. The analysis of financial trends and qualitative factors, such as management discussions and risk disclosures, remained superficial, failing to adequately address broader market implications or competitive dynamics. Additionally, the valuation analysis relied on simplistic assumptions, without incorporating scenario testing or sensitivity analyses that could have offered a more comprehensive and nuanced perspective.

3.2 Layer 2

Overview:

The Layer 2 model for generating equity research reports introduces enhanced inputs and moderate levels of guidance to improve upon the baseline model's performance. By incorporating additional data such as earnings call transcripts, industry reports, historical analyst opinions, and stock return data, this model provides a more comprehensive framework for analysis. The goal is to evaluate whether these enriched inputs and structured prompts enable GenAI to deliver reports with deeper insights and greater alignment with professional equity research standards.

Input:

The data used in this model is a combination of regulatory filings (*10-K* and *10-Q*),

supplemented by qualitative and contextual information from earnings call transcripts and industry reports. Financial data was preprocessed and standardized for consistency, allowing GenAI to incorporate both historical trends and forward-looking guidance into its analyses. It includes enriched inputs such as earnings call transcripts, industry reports, and stock data. It also incorporates AI-selected benchmarks (e.g., MSCI USA IT Index for Microsoft) and applied methodologies such as linear regression on historical and macroeconomic data.

Guidance was provided to GenAI in the form of two structured prompt types. The first, *Extraction Guidance*, directed the model to focus on extracting critical details such as financial performance metrics and operational highlights from the raw data. For example, prompts specified tasks like summarizing management's strategies from earnings call transcripts or consolidating key financial trends into structured tables. These intermediate steps ensured the inclusion of essential information and facilitated a clear foundation for subsequent analyses.

The second type of guidance involved *Decision-Making Frameworks*, which helped regulate the model's rating and recommendation processes. A standardized methodology, modeled on industry practices, was provided. For instance, Morgan Stanley's rating framework was adopted, requiring GenAI to assign a *Buy*, *Hold*, or *Sell* rating based on projected stock returns relative to benchmark indexes. This ensured that the model's recommendations adhered to structured and repeatable criteria, increasing its reliability.

Output:

On the delivery side, the models effectively summarized financial data. For example, they reported Microsoft's revenue growth, detailing a total revenue of $65.6 billion (+16% YoY), with contributions from Intelligent Cloud (+19% YoY) and Productivity & Business Processes (+13% YoY). Profitability metrics, such as gross margin ($45.5B at 69.3%) and operating income ($30.6B at +13.6% YoY), were also captured. Furthermore, the models offered simple valuation projections and contextual rationale, such as strong AI adoption driving cloud growth, Activision integration enhancing gaming margins, and operating leverage offsetting margin pressures.

Limitations and Challenges:

Reports lacked analytical depth, with minimal forward-looking guidance and poor logical

reasoning to connect facts with conclusions. Additionally, issues with inaccuracy and inconsistency were evident, as seen in Chipotle's stock price prediction. This underscored the need for expertise-driven prompts and stricter regulation of information sourcing to improve accuracy and reliability.

To address these limitations, the proposed improvements include integrating prompts crafted by domain experts and restricting data to verified platforms like Bloomberg and FactSet to ensure credible outputs. These enhancements aim to refine the AI's analytical and forecasting capabilities in future iterations.

3.3 Layer 3

Customizing GPT for Financial Analysis including detailed prompting Customizing GPT for financial analysis involved a comprehensive approach that began with gathering domain-specific data. This process included the collection of historical financial data, such as revenue trends and same-store sales growth, which provided the foundational metrics for forecasting. Industry benchmarks were also analyzed to offer a comparative perspective on Chipotle's performance relative to its peers. Additionally, macroeconomic indicators like inflation rates and labor costs were incorporated to account for broader market influences.

Qualitative data played an essential role in enhancing the model's contextual understanding. CEO statements, management guidance, and insights from sell-side research reports were included to provide qualitative context that complements the numerical analysis. All data were preprocessed to ensure compatibility with GPT, involving tasks such as parsing financial statements into structured formats and normalizing time-series data for trend analysis.

The customization process included fine-tuning GPT on these curated datasets to ensure the model understood domain-specific terminology and metrics. For example, financial concepts like weighted average cost of capital (WACC) and discounted cash flow (DCF) calculations were prioritized during training. Additionally, a modular prompting framework was implemented to simplify complex tasks. This involved breaking down financial modeling tasks into smaller, focused components that the model could handle effectively. These modular outputs were then integrated into a cohesive report.

To maintain consistency and clarity, a standardized template was provided for GPT to follow. This template specified the structure of the final reports, including sections such as the executive summary, financial projections, valuation, and risk assessment. The template not only ensured uniformity but also guided GPT in aligning its outputs with professional industry standards.

Building the Discounted Cash Flow (DCF) Model The DCF model was built using a systematic approach that relied on detailed revenue and cost projections. Revenue was forecasted based on a combination of factors, including same-store sales growth, transaction volumes, and pricing strategies. Historical trends and management guidance provided a basis for these projections, while adjustments were made to reflect macroeconomic conditions such as inflation and consumer spending patterns.

Report Generation The final report adhered strictly to the provided template, ensuring a professional and structured presentation of findings that met industry standards. It began with an executive summary, clearly outlining the primary investment recommendation—a "BUY" rating for Chipotle Mexican Grill. This recommendation was supported by a target price derived from a meticulously constructed Discounted Cash Flow (DCF) valuation model. Following the summary, the report provided a comprehensive analysis of Chipotle's financial performance, including detailed projections of revenue, costs, and profitability metrics.

A critical aspect of this process was the consistency and interpretability of the outputs. The use of a predefined template ensured that all components of the report were organized coherently. To enhance the quality of the output, we engaged in iterative testing and refinement of the content. This process improved both clarity and accuracy, ensuring alignment with professional expectations, particularly those of stakeholders in financial equity analysis.

To enable the customization of GPT for this task, we first uploaded a comprehensive dataset. This dataset included essential financial documents, such as 10-Q and 10-K filings, industry reports, and publicly available analyst forecasts from reputable sources like Refinitiv, Bloomberg, and Chipotle's company website. These data sources provided a robust foundation for the model's training, ensuring it could interpret domain-specific

information effectively.

Following the data upload, we trained GPT specifically on the provided dataset, tailoring it for financial analysis tasks. To guide the model in creating structured reports, we uploaded a Bank of America (BOFA) research report template. This template served as a framework for ensuring uniformity and completeness in the output. The final prompt provided to GPT was carefully designed to align its tasks with the requirements of professional financial analysis:

Final Prompt:

"You are a Bank of America Financial Equity Analyst creating research reports for Chipotle Mexican Grill. Your goal is to provide a comprehensive analysis, including financial modeling, valuation, and a written report with an investment recommendation. You will be provided with an exact format and example template to follow. Use only the qualitative and quantitative data provided, and do not seek additional data from external sources."

This approach ensured that GPT's outputs were not only relevant and accurate but also tailored to the professional standards expected in equity research reports. By leveraging structured data, targeted training, and a predefined template, we achieved a high degree of precision and quality in the final report.

4 Model Evaluation

4.1 Evaluation metrics

4.1.1 Metrics Definition

Our team developed a comprehensive set of evaluation metrics that serve as a framework for assessing the performance of GenAI models in financial analysis. Each metric offers unique insights into the capabilities and limitations of these tools.

Ease of Generation: This metric evaluates how simple it is to produce reports using GPT models. It considers the level of human effort or intervention required, such as the complexity of prompts, the number of interactions with GPT, and whether pre-defined prompt templates are provided or adjustments are needed based on GPT outputs.

Stability: It refers to the consistency of model predictions (Buy/Hold/Sell ratings) and analyses when provided with identical input data.

Forecasts Comparison: This evaluates the alignment between GPT-generated ratings and human analyst recommendations, as well as the consistency of GPT-generated forecasts (e.g., price targets, revenue, or earnings forecasts) with those of human analysts.

Utility: This assesses the usefulness of GPT-generated reports. Unlike other metrics, utility is more challenging to evaluate. To address this, we developed a set of sub-criteria focusing on clarity, accuracy, depth, comprehensiveness, and actionability and collected internal evaluation scores from team members to assess the utility of reports.

Discernibility: This metric assessed whether users could distinguish between AI-generated and human-authored reports.

The evaluation criteria applied to both Chipotle and Microsoft highlight several key differences in the quality of AI-generated financial reports. The table below summarizes the evaluation criteria used to assess clarity, accuracy, comprehensiveness, depth of analysis, and actionability across reports, which are the "submetrics used to evaluate Utility:

Criteria	**Excellent (5)**	**Good (4)**	**Average (3)**	**Poor (2)**	**Very Poor (1)**
Clarity and Logical Structure	Extremely clear and well-organized, with concise language and a logical flow that builds toward a cohesive conclusion.	Clear and well-structured, with only minor issues in flow or clarity.	Fairly clear; logical flow is present but may require effort or prior knowledge to fully follow.	Disorganized with unclear flow, making arguments hard to follow.	Unclear, poorly written, and lacks logical flow; nearly impossible to understand.
Accuracy	Completely accurate data, calculations, and conclusions, with no errors.	Mostly accurate with minor errors that do not affect conclusions.	Some inaccuracies or errors present but with limited impact on overall findings.	Significant errors in data, calculations, or conclusions that undermine credibility.	Critical errors or misrepresentations make the report unreliable and misleading.
Comprehensiveness	Covers all critical aspects comprehensively, addressing all relevant interconnections and dependencies.	Covers most critical aspects with minor gaps in details or interconnections.	Covers basic aspects but misses some critical or relevant details.	Covers few aspects, with minimal or incomplete information and missing interconnections.	Fails to address critical aspects and lacks relevant or meaningful content.

Depth of Analysis	Provides deep insights with robust arguments supported by detailed data, charts, and examples.	Offers good insights with most arguments well-supported by data, charts, or examples.	Offers surface-level insights with some supporting evidence; key arguments lack depth.	Lacks meaningful depth in analysis; limited or weak supporting evidence.	Superficial and unsupported analysis with no meaningful insights.
Actionability	Provides clear, actionable recommendations (e.g., Buy/Hold/Sell) grounded in sound reasoning and robust models.	Recommendations are actionable and mostly well-supported, with minor gaps in reasoning or model explanation.	Recommendations are present but lack sufficient clarity, justification, or explanation.	Recommendations are vague, impractical, or insufficiently explained to ensure credibility.	No actionable recommendations or entirely unsupported and impractical suggestions.

Table 4.1.1: Utility Sub-rubrics

The table illustrates the strengths and weaknesses of the AI-generated reports, enabling us to discern how these models performed under varying conditions for both companies.

4.1.2 Evaluation Results

Using the metrics defined above, we evaluated the reports generated by different layers of models as well as the analyst reports for Chipotle and Microsoft. The results are summarized below:

Ease of Generation

The increasing difficulty from Layer 1 to Layer 3 reflects the growing complexity and granularity of the reports generated.

- For **Layer 1**, report generation is relatively simple, requiring approximately 4 prompts with around 80 words.

- For **Layer 2**, a greater effort is required, involving approximately 20 prompts and 700 words.

- For **Layer 3**, building a customized GPT demands significant effort, requiring approximately 20 prompts with around 2000 words, but its usage is very simple with just one

prompt.

Stability

For Layer 1 and Layer 2, the model exhibits relative stability, consistently producing the same or similar ratings with only minor variations. In the case of Layer 1, since prompts do not impose strict constraints, the generated reports may include different sections, although there is significant overlap. For example, one report might contain sections like *Current Performance Overview*, *Key Growth Drivers*, *Risk Factors*, and *Valuation Metrics*, while another might include *Company Overview*, *Recent Financial Performance (FY 2023)*, *Investment Thesis*, *Key Risks*, and *Valuation*. With more prompting, the structure of the reports becomes more consistent. Additionally, even when provided with the same input materials, the model may focus on different parts of the content during each analysis, leading to variations in the output. For Layer 3, as the generation involves manual effort, we did not conduct a stability comparison.

The randomness of GPT outputs is influenced by the temperature parameter; we used the default setting for the chatbox. Stability can be improved by lowering the temperature.

Forecasts Comparison

From Layer 1 to Layer 3, predictions become increasingly granular:

- **Layer 1**: Predicts only ratings (Buy/Hold/Sell).
- **Layer 2**: Adds return forecasts.
- **Layer 3**: Incorporates price targets, calculated using DCF, adding metrics like revenue, earnings, and FCF.

Utility

Figure 2 presents the average total utility score (out of a maximum of 25) for each report, as evaluated by 8 team members. Key observations from this bar plot include:

Analyst Scores Are Higher: Across both Chipotle and Microsoft, the scores for "Analyst" consistently exceed those of all the "Model" scores, indicating that human evaluation is still outperforming automated models in this setting.

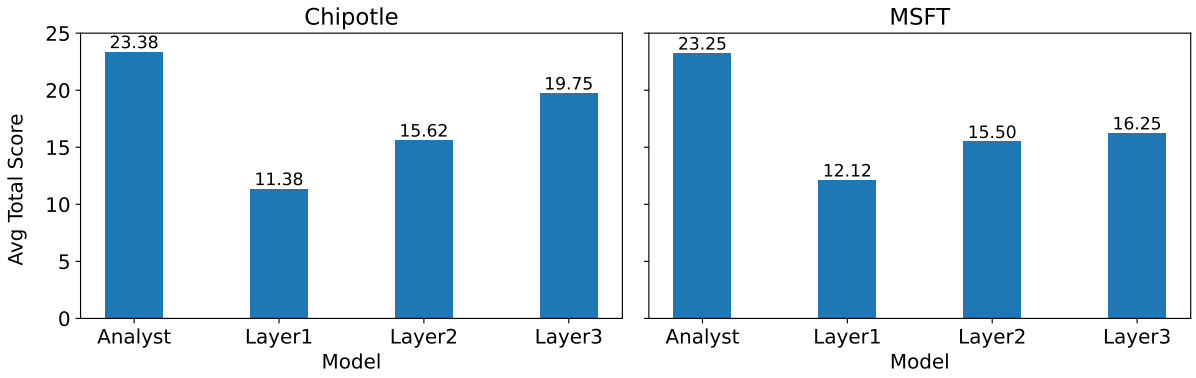

Figure 2: Total Utility Score for Each Model

Progressive Improvement in Model Scores: From Layer 1 to Layer 3, there is a noticeable improvement in the scores, suggesting that as the data and model sophistication increase, the models are able to generate better evaluations.

Chipotle Shows Greater Score Gains: The improvement in scores from Layer 1 to Layer 3 is more pronounced for Chipotle compared to Microsoft.

The performance of GPT models was assessed using Chipotle and Microsoft as representative cases to compare how well the models adapted to pure-play versus diversified business structures.

- **Chipotle (Pure-Play Model):** Chipotle, as a focused business, allowed GenAI models to produce accurate financial reports and recommendations. The simplicity of its revenue model—centered on same-store sales growth and menu innovations—enabled consistent outputs.
- **Microsoft (Diversified Model):** Microsoft, with its multifaceted revenue streams from Azure, Office, and gaming, posed a greater challenge. GPT often struggled to balance quantitative data with qualitative insights. This led to fluctuating predictions and difficulty integrating strategic factors like cloud competition and regulatory risks into the valuation.

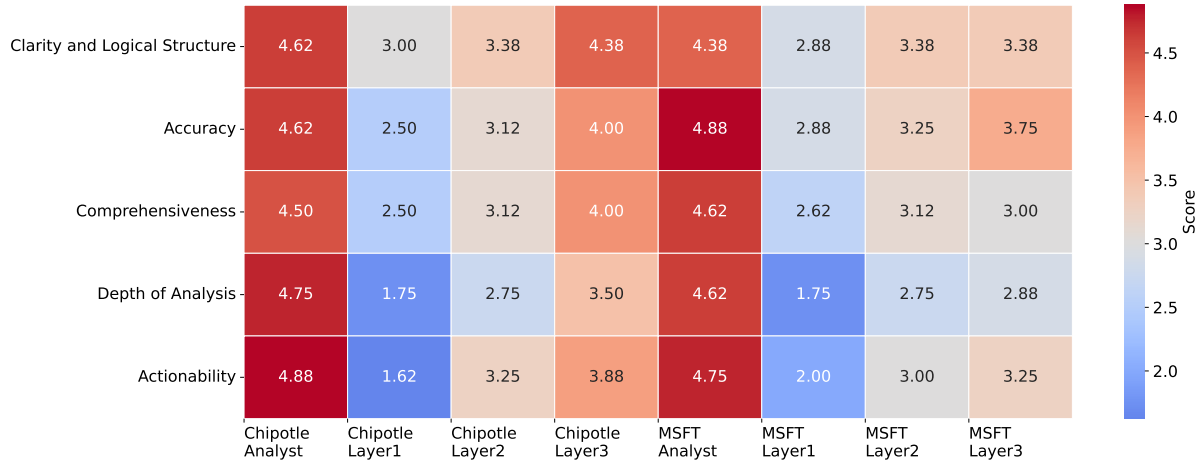

Figure 3: Utility Score Break Down

Analysis of Utility Score Breakdown

Overall Performance Differences

The utility scores show a consistent pattern where human analysts outperform all AI models, highlighting the ongoing gap between AI capabilities and expert human analysis.

Progression in Model Performance

The scores for the AI models improve from Layer 1 to Layer 3. This demonstrates that as models become more sophisticated (through better prompting and data integration), their utility scores align closer to human-generated reports, particularly in terms of comprehensiveness and depth of analysis.

Company-Specific Observations

- **Chipotle:** There is a significant improvement in utility scores from Layer 1 to Layer 3, reflecting the relative simplicity of analyzing a focused business model compared to a diversified one like Microsoft.
- **Microsoft:** The improvements across layers are more incremental, emphasizing the challenges AI models face when analyzing complex and multifaceted business structures.

Breakdown of Sub-Metrics

- **Clarity and Logical Structure:** Analysts scored consistently higher due to their ability to deliver well-organized and logically flowing arguments, especially in handling Microsoft's diverse operations.
- **Accuracy:** AI models often struggled with contextual nuances, leading to lower accuracy scores in both companies, but especially for Microsoft.
- **Comprehensiveness and Depth:** While AI models can capture broad strokes of analysis, they lack the nuanced insights provided by analysts, particularly for strategic elements like market positioning and regulatory risks.
- **Actionability:** Recommendations from human analysts are consistently more actionable, given their ability to ground decisions in robust qualitative reasoning and detailed quantitative models.

Implications for AI Development

The progressive improvement in utility scores for AI models suggests the potential for GenAI to augment financial analysts in routine tasks, freeing them to focus on strategic and nuanced analyses that remain challenging for AI to replicate. However, it does seem to be the case that the marginal improvement from better prompt engineering and datasets is decreasing.

Example

Below we will expand on our model evaluation by citing 3 different reports to explain the low scores in the heatmap for the AI models:

Example 1.

1. **Chipotle-Layer1**

Equity Research Report: Chipotle Mexican Grill, Inc. (NYSE: CMG)

Analyst Rating: Outperform
12-Month Target Price: $2,300
Current Price: $2,058 (as of XXXX, 202X)
Upside Potential: ~11.8%

Investment Thesis

Chipotle Mexican Grill, Inc. (Chipotle) continues to exhibit robust performance driven by strong revenue growth, consistent operational improvements, and strategic initiatives such as digital expansion and store openings. Despite inflationary pressures, Chipotle has demonstrated resilience through menu price adjustments and operational efficiencies, justifying an **Outperform** rating.

Overall, the Layer 1 report provides a surface-level analysis with general observations but fails to offer the detailed, data-driven insights required for a comprehensive equity research report. The lack of detailed metrics, combined with factual inaccuracies, limits its utility and explains its lower score compared to more advanced layers and human analysts. The Chipotle Layer 1 valuation report snippet reflects a simplistic analysis, with limited depth and contextual nuance. While it includes key financial metrics like the forward P/E ratio and incorporates a DCF model with defined assumptions (10% WACC and 5% terminal growth rate), it lacks comprehensive reasoning behind these inputs and fails to address potential risks or strategic factors.

Example 2.

2. Chipotle-Layer3

Chipotle Mexican Grill: Sustained Growth and Innovation Drive BUY Rating

Reiterate Rating: BUY | PO: $72.50 USD | Price: $61.25 USD

Solid QX Performance Points to Strong QX 202X Trajectory

Chipotle's QX 202X performance delivered stronger-than-expected margins and solid same-store sales growth, supported by operational efficiencies and digital adoption. With a robust pipeline for QX 202X, including planned promotional activities and new product launches, we anticipate continued momentum.

QX 202X SSSG of 8.2% exceeded expectations of 7.5%, underpinned by transaction growth and a modest increase in average check size. Management's continued focus on labour optimization and cost controls has set a solid foundation for sustained profitability.

The Layer 3 report improves with detailed metrics like an 8.2% SSSG and forward-looking insights on promotional activities and product launches. By acknowledging risks, it provides a balanced perspective, offering more depth than Layer 1. However, it still lacks the industry-specific precision of human analysts, despite being a notable improvement.

Example 3.

3. Chipotle-Analysts

Chipotle Mexican Grill Big XQ traffic beat drives EPS upside; expect share gains to persist

Reiterate Rating: BUY | PO: 71.00 USD | Price: 51.78 USD

Strong XQ; expect XQ txns to recover post weather/tech

While CMG's XQ SSSG handily beat expectations on the topline (11% vs 8.4%/8.8% and the all-important transaction count (8.7% vs 4.5%/6.0%), we believe investors are most focused on the deceleration through the quarter (XX was the strongest month) and softer start to XQ.

We continue to view the current demand environment as consistent with a slower – but still healthy – macroeconomy. As seasonality has shifted post-COVID, the summer in particular has been difficult to forecast (longer vacations/work from home) and we view the low end of the reiterated comp guide (mid to high single digits) as conservative.

The analyst report provides a nuanced assessment, contextualizing metrics like SSSG within broader market trends and investor concerns. Its professional tone, concise style, and use of industry terminology ensure credibility and relevance. With superior insights

and alignment to investor expectations, it consistently outperforms AI-generated reports.

Discernibility

AI has the capability to produce reports that are similar to those of human analysts because it can integrate a vast amount of information efficiently. However, AI has limitations, such as a lack of deep reasoning and the ability to discern complex interconnections. These limitations create discernible differences between AI-generated reports and those produced by human analysts. This problem becomes more pronounced in more complicated companies where the interplay of factors requires subjective judgment and a thorough understanding of intricate dynamics. The findings revealed that Chipotle's reports were harder to differentiate, likely due to their straightforward structure, while Microsoft's nuanced context made AI reports more identifiable.

4.2 Quantitative and Qualitative data

The integration of quantitative and qualitative data is essential for producing balanced and insightful valuation analyses. While quantitative data provides the backbone for financial models through measurable and objective metrics, qualitative data offers a forward-looking lens to assess context, opportunities, and risks that numbers alone cannot capture. This distinction underscores the complementary nature of these inputs and their joint importance in deriving accurate valuations and price targets.

Quantitative data serves as the numerical foundation of financial analysis, enabling precision and objective benchmarking in valuation models. Key inputs include metrics such as stock returns, financial statements, index performances, and macroeconomic indicators.

Precision in Valuation Models: Models such as Discounted Cash Flow (DCF) and price multiples rely heavily on quantitative data. For instance, Microsoft's historical revenues, earnings, and cash flows enable the calculation of intrinsic value through detailed forecasts.

Objective Benchmarking: Financial ratios, such as Price-to-Earnings (P/E), EBITDA multiples, and Earnings Per Share (EPS), allow analysts to compare a company's performance against industry standards and competitors. This was evident in the comparison

of Microsoft's forward P/E multiples to peers like Apple and Google. Scenario Testing: Quantitative data supports sensitivity analyses by allowing adjustments to key assumptions such as revenue growth, discount rates, and profit margins. For example, lowering Microsoft's revenue Compound Annual Growth Rate (CAGR) from 10While quantitative data provides measurable outputs and facilitates numerical adjustments to valuations, it is inherently backward-looking and lacks the context needed to project future growth or risk factors.

Qualitative data addresses the limitations of quantitative analysis by introducing a forward-looking perspective. This data includes non-numerical information such as industry trends, company-specific developments, and strategic initiatives, which provide essential context and shape assumptions in valuation models.

Forward-Looking Insights: Unlike quantitative metrics, qualitative data evaluates potential future scenarios. For example, the strategic adoption of AI technologies and Microsoft's acquisition of Activision Blizzard highlighted growth opportunities that directly impacted its valuation assumptions.

Contextual Understanding: Qualitative data interprets quantitative metrics within the broader industry and competitive landscape. For example, recognizing competitive pressures from Amazon Web Services (AWS) and Google Cloud contextualized growth projections for Microsoft's Azure services.

Risk Assessment: Through qualitative inputs, analysts identify potential risks such as regulatory scrutiny, competitive threats, and market saturation. Incorporating these factors refines valuation models, leading to more conservative forecasts and a re-evaluation of price targets.

The forward-looking nature of qualitative data was evident in its impact on price target adjustments and investment recommendations. For instance, factoring in regulatory concerns and market competition shifted Microsoft's rating from "Buy" to "Neutral" by influencing growth assumptions.

The Role of GPT Models in Data Integration GPT models exhibit proficiency in processing quantitative data due to their computational capabilities. However, their ability to leverage qualitative data depends heavily on the specificity and clarity of the

prompts provided. When guided by detailed and explicit instructions, GPT can integrate qualitative insights into forward-looking analyses, improving the depth and reliability of its outputs. Without this guidance, the models tend to rely disproportionately on quantitative metrics, often missing nuanced forward-looking insights.

4.3 Natural Language Processing (NLP) Analysis

We present a comparative analysis of a human-authored report and our most sophisticated AI-generated sell-side equity research report, both surveying Chipotle over a similar study period, using TF-IDF keyword extraction and cosine similarity metrics to compare thematic divergences and lexical similarity. These methods allow for a precise and interpretable comparison of textual content between human-authored and AI-generated reports.

4.3.1 TF-IDF

Term Frequency-Inverse Document Frequency (TF-IDF) is a statistical measure used to evaluate the importance of a term within a document relative to the entire corpus of documents. It combines term frequency, which captures how often a term appears in a single document, and inverse document frequency, which downweighs terms that are common across the corpus. In doing so, TF-IDF highlights terms that are frequent and distinctive, providing a weighted representation of the document's content. TF-IDF is given by:

$$\text{TF-IDF}(t, d, D) = \text{TF}(t, d) \cdot \text{IDF}(t, D) \tag{1}$$

Where:

- The Term Frequency (TF) is given by:

$$\text{TF}(t, d) = \frac{f_{t,d}}{\sum_{t' \in d} f_{t',d}} \tag{2}$$

- The Inverse Document Frequency (IDF) is given by:

$$\text{IDF}(t, D) = \log \frac{N}{|\{d \in D : t \in d\}|} \tag{3}$$

Here:

- $f_{t,d}$ is the frequency of term t in document d,
- N is the total number of documents,
- $|\{d \in D : t \in d\}|$ is the number of documents containing the term t.

We employ TF-IDF to perform a direct document-to-document comparison and to highlight key differences in our baseline analysis.

4.3.2 Cosine Similarity

Cosine similarity measures the degree of similarity between two text document by comparing the angle between their vectorized representations in a high-dimensional space. In the context of our work, we apply cosine similarity to TF-IDF vectors and it quantifies the lexical overlap between documents, ranging from 0 (completely dissimilar lexical representations) to 1 (identical lexical representations). Cosine similarity is computed as:

$$\text{CosineSimilarity}(A, B) = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} \cdot \sqrt{\sum_{i=1}^{n} B_i^2}} \tag{4}$$

Where:

- $A = (A_1, A_2, \ldots, A_n)$ and $B = (B_1, B_2, \ldots, B_n)$ are vectors,
- A_i and B_i are the components of the respective vectors.

4.3.3 Experiments

Our NLP analysis involves the following:

1. **Preprocessing**: Raw textual data extracted from the research report PDF files was processed using Python libraries to remove hyperlinks, special characters, and stopwords. This includes a custom stopword list to exclude non-informative terms and filler words.

2. **TF-IDF Vectorization**: Texts were vectorized using a TF-IDF model with 20 features, highlighting the most representative terms in each document.

3. **Cosine Similarity**: The lexical alignment between AI-generated and human-authored reports was quantified through cosine similarity of their respective TF-IDF feature vectors.

Full implementation code is available in the appendix at 6.5.

TF-IDF analysis identified the ten most significant keywords for both AI and human-authored reports:

Rank	Keyword (AI)	Weight (AI)	Keyword (Human)	Weight (Human)
1	metric	0.3668	eps	0.4777
2	price	0.3668	cmg	0.4777
3	cost	0.3437	usd	0.4246
4	usd	0.2445	sssg	0.2654
5	dec	0.2445	research	0.2654
6	cash	0.2445	value	0.2123
7	eps	0.2445	price	0.1592
8	margins	0.2445	metric	0.1592
9	buy	0.2445	labor	0.1592
10	rating	0.1834	dec	0.1592

Table 4.3.1: Comparison of Top Keywords and Their Weights (AI vs. Human Reports)

The TF-IDF analysis highlights distinct thematic priorities between AI-generated and human-authored reports. The AI-generated report predominantly employs generic, quantitatively-driven terms such as *metric*, *price*, and *cost*, indicating an emphasis on formulaic and transactional analysis. Terms like *buy* and *rating* suggest a focus on decision-making processes but lack the contextual grounding necessary for industry-specific applications. The sparsity and lack of depth in the AI-generated report relative to the human-authored report is a key factor contributing to these features of the document's thematic focus.

In contrast, the human-authored report demonstrates a nuanced understanding of the sector, incorporating industry-relevant terminology like *sssg* (same-store sales growth) and *labor*, which resonate with the operational realities of the field. The human-authored text also achieves a balanced integration of quantitative and qualitative insights, exemplified by terms like "value" and "research". Critical financial metrics such as *EPS* (Earnings-Per-Share) are given significantly more weight in the human report compared

to the AI report, reflecting a deeper alignment with industry priorities.

We acknowledge the limitation of doing a TF-IDF comparison with a limited corpus size, as the IDF (Inverse Document Frequency) portion does not have a lot of data to work with. There may be possible issues with bias toward unique terms and overfitting to noise given the limited corpus size used in this comparison. Hence, we use our TF-IDF as a baseline method to identify obvious differences in vocabulary and perform a text cosine similarity comparison for robustness. We obtain a cosine similarity score of **0.66**, indicating a moderate degree of lexical alignment. This degree of similarity suggests that despite some common analytical elements, significant differences in emphasis and contextual depth remain. The overlap likely stems from the use of fundamental financial terms common in equity research reports, but the divergence points to variations in the representation of nuanced domain-specific language.

The score highlights the distinct approaches of the two reports highlighted earlier: the AI-generated report leans toward generic, quantitative terms, whereas the human-authored counterpart incorporates more industry specific vocabulary and qualitative insights. There is a thematic alignment on a superficial level but it lacks deeper congruence in analytical priorities and contextual framing.

5 Conclusion

Through empirical testing, we find that layer 1 reports provide a foundational baseline, but lack depth and precision. Layer 2 shows improvement but still has limitations in aligning with human insights. Our most sophisticated layer 3 model integrates the detailed data that most closely mimics professional standards, though still falls short of human analysis in terms of granular understanding of the business.

GenAI excels at synthesizing vast amounts of data, generating reports for less complex business models, and demonstrates strong performance in routine tasks. It is also highly customizable to product and user requirements. Yet, it struggles with nuanced understanding of complex business models, is susceptible to variation, and does not accurately assess contextual and strategic elements.

Human analysts consistently outperform GenAI in comprehensiveness, accuracy, depth of

analysis, and actionability of recommendations. A human's ability to integrate qualitative insights with quantitative data creates value that AI in its current form is unable to replicate.

Results show the complementary potential of generative AI to augment, rather than replace, human analysts. By automating routine tasks, providing decision-making support, and offering baseline quantitative insights through CoT prompting, AI can free up human analysts to focus on high-value items and automate routine tasks to improve efficiency. Ultimately, the most effective outcomes arise when GenAI is deployed alongside human analysts. Their adoption should emphasize augmentation rather than replacement of human analysts.

To effectively integrate GenAI into the analyst workflow, we propose a phased integration approach to be most appropriate, starting with simple tasks and scaling as AI capabilities improve to ensure balanced utilization and a synergistic approach to AI adoption. This includes investing in frameworks to integrate AI into the analyst workflow and using it for routine tasks as a decision-support tool. For efficient integration, there should be a focus on quality training datasets and careful prompt engineering to guard against hallucinations. Robust validation frameworks are also vital to guard against compliance issues associated with AI-generated outputs.

References

[1] Elizabeth Blankespoor, Ed deHaan, and Qianqian Li. "Generative AI in Financial Reporting". In: (2024). DOI: 10.2139/ssrn.4986017. URL: http://dx.doi.org/10.2139/ssrn.4986017.

[2] Georgios Fatouros et al. "Can Large Language Models Beat Wall Street? Unveiling the Potential of AI in Stock Selection". In: (2024). DOI: 10.48550/ARXIV.2401.03737. URL: https://arxiv.org/abs/2401.03737.

[3] Rajan Gupta, Gaurav Pandey, and Saibal Kumar Pal. "Automating Government Report Generation: A Generative AI Approach for Efficient Data Extraction, Analysis, and Visualization". In: *Digital Government: Research and Practice* (Sept. 2024). ISSN: 2639-0175. DOI: 10.1145/3691352. URL: http://dx.doi.org/10.1145/3691352.

[4] Alex Kim, Maximilian Muhn, and Valeri Nikolaev. "Financial Statement Analysis with Large Language Models". In: (2024). DOI: 10.48550/ARXIV.2407.17866. URL: https://arxiv.org/abs/2407.17866.

[5] David Krause. "Large Language Models and Generative AI in Finance: An Analysis of ChatGPT, Bard, and Bing AI". In: *SSRN Electronic Journal* (2023). ISSN: 1556-5068. DOI: 10.2139/ssrn.4511540. URL: http://dx.doi.org/10.2139/ssrn.4511540.

[6] Mahmoud Mahfouz et al. "The State of the Art of Large Language Models on Chartered Financial Analyst Exams". In: (2024), pp. 1068–1082. DOI: 10.18653/v1/2024.emnlp-industry.80. URL: http://dx.doi.org/10.18653/v1/2024.emnlp-industry.80.

[7] Kassiani Papasotiriou et al. "AI in Investment Analysis: LLMs for Equity Stock Ratings". In: (2024). DOI: 10.48550/ARXIV.2411.00856. URL: https://arxiv.org/abs/2411.00856.

[8] Nitin Rane. "Role and Challenges of ChatGPT and Similar Generative Artificial Intelligence in Finance and Accounting". In: *SSRN Electronic Journal* (2023). ISSN: 1556-5068. DOI: 10.2139/ssrn.4603206. URL: http://dx.doi.org/10.2139/ssrn.4603206.

[9] Tianyu Zhou et al. *FinRobot: AI Agent for Equity Research and Valuation with Large Language Models*. 2024. DOI: 10.48550/ARXIV.2411.08804. URL: https://arxiv.org/abs/2411.08804.

6 Appendix

6.1 Layer 1 report for Chipotle

Equity Research Report: Chipotle Mexican Grill, Inc. (NYSE: CMG)

Analyst Rating

- **Outperform**
- **12-Month Target Price:** $2,300
- **Current Price:** $2,058 (as of September 30, 2024)
- **Upside Potential:** 11.8%

Investment Thesis

Chipotle Mexican Grill, Inc. (Chipotle) continues to exhibit robust performance driven by strong revenue growth, consistent operational improvements, and strategic initiatives such as digital expansion and store openings. Despite inflationary pressures, Chipotle has demonstrated resilience through menu price adjustments and operational efficiencies, justifying an **Outperform** rating.

Key Drivers Revenue Growth

- Q3 2024 Revenue: $2.79 billion (+13.0% YoY).
- Digital sales, representing 36% of revenue, continue to contribute significantly to top-line growth.

Operational Efficiency

- Restaurant operating margins increased to 17.0% in Q3, reflecting improved cost controls.
- Labor efficiencies and price increases helped offset higher food costs for key ingredients like beef and avocados.

Store Expansion

- Opened 200 new stores YTD 2024, including 20 international locations.

- Plans to accelerate openings in North America and the Middle East bolster long-term growth prospects.

Digital Innovation

- Continued investment in digital platforms and delivery services has enhanced customer convenience and loyalty.

Resilient Brand and Pricing Power

- Chipotle's strong brand equity enables it to pass on inflationary pressures through price hikes without significant loss in traffic.

Financial Highlights

- **Gross Margin:** Improved to 38.4% in Q3 2024 from 37.6% in Q3 2023 due to better supply chain management and pricing adjustments.
- **Net Income:** $387 million (+23.7% YoY).
- **Cash Position:** $699 million as of Q3 2024, ensuring strong liquidity for expansion and shareholder returns.

Risks

- **Cost Pressures:** Rising labor and food costs could challenge margins despite price increases.
- **Economic Sensitivity:** Chipotle's mid-premium positioning may face headwinds during an economic slowdown.
- **Regulatory Challenges:** Wage inflation and potential unionization could impact operating costs.

Valuation

Chipotle trades at a forward P/E of 38x, a premium to its historical average of 34x. However, its growth trajectory, pricing power, and operational leverage justify this valuation. Using a DCF model incorporating a 10% WACC and 5% terminal growth rate, we derive a target price of $2,300.

Catalysts

- Accelerated store openings in international markets.
- Sustained growth in digital and loyalty program sales.
- Improved cost efficiencies and margin expansion.

Recommendation

Chipotle is well-positioned to deliver strong returns in the next quarter and beyond. Investors should capitalize on its growth momentum and solid fundamentals, supporting an **Outperform** rating with a target price of $2,300.

6.2 Layer 2 report for Chipotle

Revenue Momentum and Operational Innovations Maintain Growth Outlook

Reiterate Rating: HOLD — Price Target: $61.52 USD — Current Price: $61.52 USD

Quarterly Performance Highlights

Chipotle Delivers Robust Revenue Growth with Margin Challenges

Chipotle reported 13% YoY revenue growth in FY24Q3, achieving $2.79 billion in total revenue, driven by 6% same-store sales growth and a 3% increase in transactions. Adjusted diluted EPS grew 17% YoY to $0.27, outperforming expectations by $0.02. Despite strong top-line growth, restaurant-level margins decreased to 25.5% due to inflationary pressures on avocados, dairy, and higher portion sizes.

Segment Performance Remains Strong

- **In-Restaurant Sales:** Revenue surged 80% YoY, supported by seasonal recovery and customer response to new menu offerings like Smoked Brisket.
- **Digital Sales:** Represented 34% of total revenue, driven by Chipotle Rewards program engagement and delivery services.
- **New Restaurants:** Chipotle opened 86 new locations this quarter, with 73 Chipotlanes, maintaining strong unit economics and expansion momentum.

Key Catalysts and Drivers

- **Menu Innovation Bolsters Transaction Growth:** Chipotle's limited-time offerings like Smoked Brisket generated incremental transaction growth and higher ticket sizes. Tests for Chipotle Honey Chicken have shown promising results, with broader rollouts anticipated in FY25.
- **Operational Efficiency Gains Through Technology:** The rollout of dual-sided planchas and produce slicers has improved throughput during peak hours. Additionally, the adoption of AI-powered hiring platforms reduced hiring times by up to 75%, enhancing Chipotle's competitive edge in labor management.
- **Geographic Expansion Continues to Gain Traction:** With 185 restaurants

opened YTD, Chipotle is on track to meet its FY24 target of 285–315 new locations, with 80% featuring Chipotlanes. Internationally, the brand continues to thrive, especially in Canada and Dubai, with plans to accelerate growth in Europe in FY25.

- **Macro Conditions and Pricing Strategies:** Despite higher input costs, Chipotle maintained pricing power through targeted increases in California. Inflationary trends, especially in avocados and dairy, will remain key cost considerations moving forward.

Valuation

Quarter	Revenue Forecast ($B)	YoY Growth (%)
FY24Q4	2.95	12.5
FY25Q1	2.90	10.5
FY25Q2	3.20	13.2
FY25Q3	3.15	13.0

Table 6.2.1: Revenue Forecasts and YoY Growth

Stock Return Prediction

Rationale and Methodology

Using a regression model based on historical revenue growth and stock price performance, we project a quarterly stock return of 8.2%. This estimate accounts for:

- **Revenue Momentum:** Sustained same-store sales growth and digital channel performance.
- **Operational Investments:** Efficiencies from new technologies and labor optimization.
- **Macro Factors:** Resilient consumer spending in the QSR sector and moderated inflationary trends.

Rating and Recommendation

- **Recommendation: HOLD**

  While Chipotle's fundamentals remain robust, the projected quarterly return of 8.2% falls slightly below the 9.27% benchmark return of the Dow Jones U.S. Restaurants & Bars Index. This justifies a HOLD rating, given the limited near-term

upside and inflationary pressures on margins.

- **Price Target: $61.52 USD**

  The current valuation reflects fair pricing, considering Chipotle's growth trajectory, competitive positioning, and operational execution.

6.3 Layer 3 report for Chipotle

Chipotle Mexican Grill: Sustained Growth and Innovation Drive BUY Rating

Reiterate Rating: BUY — PO: $72.50 USD — Price: $61.25 USD

Solid Q4 Performance Points to Strong Q1 2025 Trajectory

Chipotle's Q4 2024 performance delivered stronger-than-expected margins and solid same-store sales growth (SSSG), supported by operational efficiencies and digital adoption. With a robust pipeline for Q1 2025, including planned promotional activities and new product launches, we anticipate continued momentum.

Q4 2024 SSSG of 8.2% exceeded expectations of 7.5%, underpinned by transaction growth and a modest increase in average check size. Management's continued focus on labor optimization and cost controls has set a solid foundation for sustained profitability.

Operational Efficiencies Driving Margins

Record restaurant-level margins in Q4 2024 (28.8%, +130bps YoY) position CMG for further upside despite ongoing inflationary pressures in protein and labor costs. Benefits from operational investments, such as the deployment of automated grills and streamlined digital order processing, continue to offset cost headwinds. For Q1 2025, margins are projected at 28.5%, with improvements in throughput efficiency and cost leverage.

2025 EPS Adjustments and Revenue Outlook

We have revised our 2025 EPS to $1.39 (previously $1.36) based on better-than-expected top-line performance and cost management. Revenues for 2025 are now forecasted at $13.1 billion, with Q1 2025 expected to contribute approximately $3.25 billion.

Maintain Price Objective and Buy Rating

Our valuation approach reflects steady-state earnings potential, assuming Chipotle reaches its long-term target of 8,000 stores globally. Applying a 25x EBITDA multiple to projected earnings and discounting back at a 7% WACC, we reaffirm our price objective of $72.50, implying a ˜18% upside from the current price.

Metric	Previous	Current
2025E Rev (m)	$12,976.1	$13,100.0
2025E EPS	$1.36	$1.39

Table 6.3.1: Key Changes (US$)

Data	Value
Price	$61.25 USD
Price Objective	$72.50 USD
Investment Opinion	BUY
52-Week Range	$40.12 - $68.50
Market Value (mn) / Shares	$84,112 / 1,373.4
Free Float	99.3%

Table 6.3.2: Stock Data

Metric	2024E	2025E	2026E
GAAP EPS	$1.12	$1.39	$1.65
P/E	54.7x	44.1x	37.1x
Revenue (m)	$11,357.0	$13,100.0	$14,975.0

Table 6.3.3: Estimates (Dec)

Metric	2022A	2023A	2024E	2025E	2026E
P/E	78.5x	57.5x	46.2x	38.1x	31.2x
EV / EBITDA	50.4x	38.6x	32.2x	26.3x	21.8x

Table 6.3.4: Valuation (Dec)

Metric	2022A	2023A	2024E	2025E	2026E
Cash Realization Ratio	1.4x	1.4x	1.4x	1.3x	1.3x
Asset Replacement Ratio	1.7x	1.8x	1.7x	1.0x	0.9x
Tax Rate	23.9%	24.2%	24.5%	25.0%	25.0%
Net Debt-to-Equity Ratio	-16.2%	-18.3%	-29.6%	-46.1%	-60.2%

Table 6.3.5: iQmethod SM – Quality of Earnings

Metric (US$ Millions)	2022A	2023A	2024E	2025E	2026E
Net Income from Cont. Ops	$899	$1,229	$1,527	$1,880	$2,285
Depreciation & Amortization	$287	$319	$345	$381	$422
Change in Working Capital	$79	$95	$45	$45	$59
Deferred Taxation Charge	($43)	($10)	$0	$0	$0
Other Adjustments, Net	$102	$150	$181	$184	$193
Capital Expenditure	($479)	($561)	($571)	($400)	($400)
Free Cash Flow	$844	$1,23	$1,528	$2,091	$2,559

Table 6.3.6: Free Cash Flow Data (Dec)

Metric (US$ Millions)	2022A	2023A	2024E	2025E	2026E
Cash & Equivalents	$384	$561	$1,119	$2,241	$3,831
Trade Receivables	$107	$116	$141	$161	$184
Other Current Assets	$685	$945	$943	$964	$989
Property, Plant & Equipment	$1,951	$2,170	$2,378	$2,384	$2,349
Total Assets	$6,928	$8,044	$8,835	$10,003	$11,606
Short-Term Debt	$0	$0	$0	$0	$0
Total Liabilities	$4,559	$4,982	$5,051	$5,138	$5,245

Table 6.3.7: Balance Sheet Data (Dec)

6.4 Analyst report for Chipotle

Chipotle Mexican Grill Big 2Q traffic beat drives EPS upside; expect share gains to persist

Reiterate Rating: BUY — PO: 71.00 USD — Price: 51.78 USD

Strong 2Q; expect 3Q txns to recover post weather/tech

While CMG's 2Q SSSG handily beat expectations on the topline (11% vs 8.4%/8.8% BofAE/VA Consensus) and the all-important transaction count (8.7% vs 4.5%/6.0% BofAE/VA Consensus), we believe investors are most focused on the deceleration through the quarter (April was the strongest month) and softer start to 3Q. We continue to view the current demand environment as consistent with a slower – but still healthy – macroeconomy. As seasonality has shifted post-COVID, the summer in particular has been difficult to forecast (longer vacations/work from home) and we view the low end of the reiterated comp guide (mid to high single digits) as conservative. As mix grows less negative (more add-ons in 2Q vs 1Q), CMG brings back one of its most successful LTOs – smoked Brisket – in the fall, and advertising picks up we expect traffic to recover.

Labor guide looks conservative given throughput

Alongside robust SSSG trends, CMG posted record store margins – at 28.9% in 2Q (+140 bps y/y) — on sales leverage and more efficient ops. As CMG faces inflationary pressure from wages and food costs (protein, avocados), and invests behind improving portion consistency (a 40-60 bps headwind to COGS in 3Q), benefits from tightened operations (better labor deployment, dual-sided grill rollout) and ongoing throughput gains should help offset. Given leverage in 1H, the labor guide in particular could be high.

F24E EPS mostly unchg, 2Q beat offsets slightly lower 2H

Our F24E EPS moves to $1.12 (vs $1.13 prior) as BTE 2Q SSSG offsets lowered 3Q SSSG (5.9% vs 7.3% prior). Our estimate F24 SSSG of 7.5% (vs 7.1% previously) embeds 5.3% transaction growth and 2.5% average check growth.

Maintain PO and reiterate Buy rating

We value CMG on steady state earnings power, assuming 7000 U.S. stores and an additional 1000 international. At steady state, we expect AUVs to exceed $4mm and margins to reach the prior peak of 27% (28% ex. pre-opening). Assuming G&A of 5% (similar to mature company operated systems), CMG would generate $8.7bb in EBITDA. Applying a 24.5x multiple (unchanged), the implied EV is $212bb, or $90bb discounted back to today, we derive our PO of $71.

Metric	Previous	Current
2024E Rev (m)	11,345.6	11,356.9
2025E Rev (m)	12,970.2	12,976.1
2026E Rev (m)	14,870.7	14,875.3
2024E EPS	1.13	1.12
2025E EPS	1.38	1.36
2026E EPS	1.67	1.66

Table 6.4.1: Key Changes (US$)

Data	Value
Price	51.78 USD
Price Objective	71.00 USD
Date Established	25-Apr-2024
Investment Opinion	B-1-9
52-Week Range	35.37 USD - 69.26 USD
Market Value (mn) / Shares Out (mn)	71,113 USD / 1,373.4
Free Float	99.3%
Average Daily Value (mn)	1052.24 USD
BofA Ticker / Exchange	CMG / NYS
Bloomberg / Reuters	CMG US / CMG.N
ROE (2024E)	45.2%
Net Dbt to Eqty (Dec-2023A)	-18.3%

Table 6.4.2: Stock Data

Metric	2022A	2023A	2024E	2025E	2026E
GAAP EPS	0.64	0.89	1.11	1.36	1.66
EPS Change (YoY)	29.4%	36.4%	24.4%	21.4%	22.1%
Consensus EPS (Bloomberg)	-	-	1.11	1.34	1.57
DPS	0	0	0	0	0

Table 6.4.3: Estimates (Dec) (US$)

Metric	2022A	2023A	2024E	2025E	2026E
P/E	78.5x	57.5x	46.2x	38.1x	31.2x
EV / EBITDA*	50.4x	38.6x	32.2x	26.3x	21.8x
Free Cash Flow Yield*	1.2%	1.7%	2.1%	2.9%	3.6%

Table 6.4.4: Valuation (Dec)

6.5 NLP Analysis: Implementation Code

```
import re
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.metrics.pairwise import cosine_similarity
from nltk.corpus import stopwords
import nltk
from sentence_transformers import SentenceTransformer

nltk.download("stopwords")

CUSTOM_STOPWORDS = set([
    "would", "could", "should", "also", "many", "may", "much",
    "one", "two", "three", "four", "five", "good", "like", "however",
        "therefore",
    "thus", "make", "made", "need", "use", "new", "time", "include",
        "provided",
    "information", "section", "data", "proposed", "rule", "final",
        "notification",
    "order", "microsoft"
])
STOPWORDS = set(stopwords.words("english")) | CUSTOM_STOPWORDS

def extract_text_from_pdf(file_path):
    from PyPDF2 import PdfReader
    reader = PdfReader(file_path)
    text = ''
    for page in reader.pages:
        text += page.extract_text()
    return text

def preprocess_text(text):
    text = re.sub(r"http\S+", "", text) # Remove links
    text = re.sub("[^A-Za-z]+", " ", text) # Remove special characters
    text = re.sub(r"\s+", " ", text) # Remove extra spaces
```

```
    text = text.strip().lower()
    return " ".join(word for word in text.split() if word not in STOPWORDS and
        len(word) > 2)

GenAI_text =
    preprocess_text(extract_text_from_pdf("Reports/GenAI_chipotle.pdf"))
human_text =
    preprocess_text(extract_text_from_pdf("Reports/Bofa_chipotle.pdf"))

documents = [GenAI_text, human_text]

# TF-IDF Vectorization
tfidf_vectorizer = TfidfVectorizer(max_features=20, stop_words="english")
tfidf_matrix = tfidf_vectorizer.fit_transform(documents)

# Get feature names and weights
feature_names = tfidf_vectorizer.get_feature_names_out()
GenAI_weights = tfidf_matrix.toarray()[0]
human_weights = tfidf_matrix.toarray()[1]

# Display top keywords
def display_keywords(doc_name, weights, feature_names):
    print(f"\nTop keywords for {doc_name}:")
    sorted_indices = weights.argsort()[::-1] # Sort by weight (descending)
    for idx in sorted_indices[:10]: # Top 10 keywords
        print(f"{feature_names[idx]}: {weights[idx]:.4f}")

display_keywords("AI", GenAI_weights, feature_names)
display_keywords("Human", human_weights, feature_names)

# Cosine Similarity (Lexical)
cosine_sim = cosine_similarity([GenAI_weights], [human_weights])[0][0]
print(f"\nCosine Similarity (TF-IDF) between AI and Human reports:
    {cosine_sim:.4f}")
```