UTILIZAÇÃO DE LARGE LALGUAGE MODELS EM ANÁLISE FUNDAMENTALISTA

REGIS GUIMARÃES MIRANDA e EDUARDO FONSECA MENDES

Fundação Getulio Vargas

Escola de Economia de São Paulo

#### Resumo

Este estudo explora a aplicação de modelos de linguagem de grande porte (LLMs) na análise fundamentalista e na previsão de desempenho financeiro de empresas, com o objetivo de avaliar sua capacidade de competir com analistas financeiros humanos. Utilizando técnicas avançadas de prompt engineering, como Chain-of-Thoughts e Self-Consistency, o modelo ChatGPT o3-mini-high foi empregado para interpretar relatórios financeiros trimestrais e prever o lucro líquido das empresas. Dados de 97 companhias do IBRX-100 foram analisados, comparando as previsões do modelo com as de analistas humanos por meio das métricas F1-Score, AUC e Erro Absoluto Escalado.

Os resultados foram aplicados à otimização de carteiras utilizando os métodos de Markowitz e índices Sharpe e Sortino, comparando o desempenho com o benchmark BOVA11. A pesquisa demonstra o potencial dos LLMs para democratizar a análise financeira, reduzir custos operacionais e apoiar investidores individuais, mas reconhece limitações, como dificuldades em cálculos complexos e eventos atípicos.

O trabalho sugere que os LLMs podem complementar analistas humanos, oferecendo insights rápidos e precisos, mas destaca a necessidade de abordagens híbridas que combinem IA com expertise humana. Extensões futuras incluem integração com dados alternativos e validação em mercados internacionais, ampliando o impacto social e econômico da pesquisa.

**Palavras-chave**: Modelos de Linguagem de Grande Porte (LLMs); Análise Fundamentalista; Prompt Engineering; Previsão Financeira; Otimização de Carteiras.

 1 INTRODUÇÃO

A exploração da capacidade dos modelos de linguagem de grande porte (LLMs) em executar atividades diversas constitui um tema amplamente debatido na academia, embora ainda existam lacunas a serem preenchidas. Diversos estudos têm sido conduzidos para aproveitar as excelentes habilidades desses modelos em emular conversas, compreender contextos e fornecer insights. Por exemplo, Bernard et al. (2024) e Jesus et al. (2024) analisaram, respectivamente, a aptidão desses modelos em atribuir nomenclaturas corretas a dados financeiros e em realizar análises de sentimento em notícias. Outros trabalhos buscam explorar os limites desses modelos; Wang e Brorsson (2025), por exemplo, concluíram que, embora o modelo empregado consiga efetuar operações matemáticas isoladas, ele ainda demonstra ineficiência ao realizar cálculos em sequência, mantendo o contexto analisado – estudo no qual foram avaliados os modelos Llama 3.1, Llama 3.2 e Mistral 7B.

Um aspecto comum nas pesquisas que utilizam LLMs é a adoção de diferentes técnicas de “*prompt engineering*”, destacando-se o “*Chain-of-Thoughts*”, proposto originalmente por Wei et al. (2022). No presente estudo, empregou-se o método “*Self-Consistency*” (Wang et al., 2023) em conjunto com o “*Chain-of-Thoughts*”, com o intuito de gerar múltiplas respostas e identificar a mais consistente, tendo em vista que foram observadas variações significativas nas respostas geradas para as mesmas instruções.

Em consonância com o trabalho de Kim, Muhn e Nikolaev (2024), este estudo tem como objetivo investigar se o modelo ChatGPT o3-mini-high é capaz de prever o lucro líquido das empresas para o ano subsequente com desempenho comparável ao dos analistas financeiros humanos e se as ações com predição positiva apresentam vantagens em relação ao mercado (representado pelo BOVA11) na otimização de carteiras. Para essa comparação, utilizaram-se métricas como F1-Score e AUC (*Area Under the Curve*) – para avaliar aspectos binários das previsões (melhora ou piora do lucro líquido) – além do Erro Absoluto Escalado, que mensura a precisão real do modelo em relação aos analistas.

Foram coletados relatórios financeiros (ITR para informações trimestrais e DFP para informações anuais) de 97 empresas que compuseram a carteira da IBRX-100 em maio de 2023 no site da CVM. Para as previsões dos analistas financeiros, recorreu-se ao banco de dados “I/B/E/S Detail File”, que reúne tais relatórios e possibilita a avaliação do desempenho geral dos analistas no mercado americano. A coleta dos preços das ações foi realizada por meio da plataforma “EODHD”.

A otimização da carteira foi conduzida utilizando diferentes abordagens: Markowitz (MV), Equal-Weights, Value-Weights e otimizações fundamentadas nos índices Sharpe e Sortino. A diversidade de estratégias adotadas objetivou identificar qual delas apresentava maior consistência de desempenho em comparação ao mercado (representado pelo BOVA11). Ressalta-se que, em razão das datas distintas de divulgação dos relatórios financeiros das empresas, uma ação só foi incorporada à carteira após a emissão do respectivo relatório, a fim de evitar resultados incompatíveis com a realidade.

Este trabalho prossegue com a Revisão da Literatura, na qual são apresentados estudos que aplicam LLMs em diversas situações no campo das finanças, bem como as múltiplas técnicas de “*prompting*” fundamentais para esses modelos. Na seção de Metodologia, detalham-se as técnicas empregadas para a obtenção e o tratamento dos dados, além dos métodos utilizados para avaliar o desempenho do experimento frente a benchmarks reconhecidos.

A seção de Resultados apresenta os achados do experimento, com a segmentação dos dados de modo a facilitar a compreensão dos detalhes e nuances observadas. Por fim, a Conclusão expõe as respostas relativas aos objetivos deste estudo, bem como possíveis direções para investigações futuras.

2 REVISÃO DA LITERATURA

## 2.1 Introdução à Revisão de Literatura

Apesar da utilização de técnicas de mineração de texto em finanças não ser algo novo, o trabalho de Chen, Kelly e Xiu (2022) foi pioneiro pela utilização de Large Language Models para extrair contextos de textos de notícias e transformá-los em sinais preditivos para o retorno das ações. Nizcszota e Abbas (2023), por outro lado, foram pioneiros ao demonstrar o grau de literacia financeira de modelos LLMs, encontrando resultados de 66% e 65% para os modelos Davinci e GPT-3.5, respectivamente. Neste estudo, é possível verificar a elevada influência do artigo de Kim, Muhn e Nikolaev (2024) e como os autores abordaram a tratativa de realizar previsões do lucro líquido da empresa com base em dados financeiros emitidos anualmente por elas.

Algo que muitos dos trabalhos recentes da área vêm demonstrando é a preocupação com a aplicação de técnicas diferentes de “prompting” para ajudar a entender a performance e captar nuances. Estas técnicas são variadas; a de maior destaque tem o nome de “Chain-of-Thoughts” (CoT), apresentada originalmente por Wei et al. (2022), e “Self-Consistency” de Wang et al. (2023).

Adicionalmente, estudos recentes têm ampliado a discussão sobre a validação dos resultados dos LLMs em comparação com métodos tradicionais e sobre a importância da supervisão humana na análise dos outputs gerados, reforçando a necessidade de uma abordagem híbrida que combine inteligência artificial e expertise humana.

Esta revisão de literatura prosseguirá com a apresentação das técnicas modernas de “prompting” para, em seguida, explorar trabalhos em finanças que utilizam os modelos LLMs em situações específicas.

## 2.2 Técnicas de “prompting”

Técnicas específicas para enviar instruções a modelos LLMs começaram a ser exploradas e registradas na literatura com Brown et al. (2020), que demonstraram que modelos com grande quantidade de parâmetros – por exemplo, 175 bilhões – tendem a apresentar desempenho aprimorado quando expostos a alguns exemplos de entradas e saídas sobre temas similares, técnica conhecida como “few-shots prompting”. Posteriormente, Wei et al. (2022) forneceram uma abordagem alternativa, instruindo o modelo por meio de etapas de raciocínios intermediários para resolver a questão central da instrução. Conforme definido pelos autores:

“A chain of thought is a series of intermediate natural language reasoning steps that lead to the final output, and we refer to this approach as chain-of-thought prompting” (WEI et al., 2022, p.2).

A abordagem de Wei et al. (2022) não visava substituir o método “few-shots” de Brown et al. (2020), mas complementá-lo. De acordo com Wei et al. (2022, p.3), algumas características fundamentais do método incluem:

1. Permitir que os modelos decompõem os problemas em múltiplas etapas;
2. Possibilitar a verificação do raciocínio empregado pelo LLM, permitindo detectar possíveis enganos;
3. Ampliar a gama de tarefas possíveis, englobando desde problemas matemáticos até questões de senso comum e manipulações simbólicas;
4. Integrar-se facilmente a métodos *few-shots* já existentes.

Kojima et al. (2023), percebeu que o método “*Chain-of-Thoughts*” poderia ser simplificado e ter seu uso ampliado, mesmo em casos de não haver exemplos ou instruções claras para seguir. Seu método ficou conhecido como “*Zero-Shot Chain-of-Thoughts*” e sua contribuição foi demonstrar que uma etapa simples como incluir “pense em etapa por etapa” ao final da instrução poderia favorecer o aumento da eficácia do modelo em métricas utilizadas no setor, por exemplo, o acerto no teste “MultiArith” sobe de 17,7% para 78,7%, ao passo que no teste GSM8K o avanço foi de 10,4% para 40,7%.

Uma questão central para os usuários desses modelos em atividades de análise é que, em situações envolvendo o envio de instruções com um número elevado de tokens, pode ocorrer a geração de respostas divergentes para a mesma pergunta, mesmo com o parâmetro de aleatoriedade (“Temperature”) ajustado para 0. Para contornar esse problema, Wang et al. (2023) propuseram o método “Self-Consistency” como alternativa ao Chain-of-Thought, no qual o modelo gera diversas respostas para o mesmo prompt e adota como resposta final aquela que se repete com maior frequência. Em testes realizados com UL2-20B, GPT3-175B, LaMDA-137B e PaLM-540B, o método Self-Consistency superou o CoT em tarefas de raciocínio aritmético e de senso comum, mesmo em cenários onde o CoT não apresentava melhorias consistentes.

Já Yao et al. (2023) introduziram a abordagem “Tree of Thoughts” (ToT), distinta do *Self-Consistency* de Wang (2023) quanto ao modo de organização do raciocínio. Em vez de gerar múltiplas respostas e escolher a mais frequente, estimula-se a criação de uma “árvore de raciocínio”. No *prompt*, enfatiza-se a separação dos nós, incentivando o LLM a elaborar hipóteses passo a passo, totalmente orientadas pela natureza da tarefa em questão. Somente ao término desse processo obtém-se a conclusão. Essa estratégia requer, em alguns casos, ajustes no parâmetro *Temperature* do modelo, pois esse controle influencia o grau de liberdade criativa na geração de hipóteses.

Como variação adicional do raciocínio passo a passo, Wang (2023, p.2) descreve o “*Plan-and-Solve*”, também chamado de “*A new Zero-Shot CoT*”. Este método demanda que o modelo primeiro gere um plano para, posteriormente, executar a tarefa, incluindo o raciocínio intermediário. Essa abordagem é expressa em seu trabalho conforme é mencionado:

“Capable of generating a higher-quality reasoning process than Zero-shot-CoT prompting, as the PS prompts provide more detailed instructions guiding the LLMs to perform correct reasoning tasks.” (Wang, 2023, p.2):

No “*Zero-shot PS”*, a instrução vai além do “*let’s think step by step*”, podendo ser algo como em WANG (2023, p.3) “Let’s first understand the problem and devise a plan to solve the problem. Then, let’s carry out the plan and solve the problem step by step”. Desta forma, registraram-se ganhos de até 5% no desempenho em relação ao “*Zero-Shot CoT”*, que se ampliam quando se combina o método ao “*Self-Consistency*”.

## 2.3 Uso de LLM em Finanças

De um modo geral, a literatura ainda carece de estudos mais aprofundados sobre o potencial prático dos LLMs em tarefas de trading. Dakalbab et al. (2024) realizaram uma Revisão Sistemática da Literatura (RSL), conforme o modelo de Kitchenham e Charters (2007), identificando 143 artigos entre 2015 e 2023. No mapeamento, verificou-se que 71% dos estudos focavam em análise técnica, e que o deep learning era a abordagem de aprendizado de máquina mais frequente (21%). Entretanto, não se explorou, de maneira sistemática, a aplicação de LLMs como o ChatGPT para aprimorar quaisquer técnicas de trading já existentes.

Estudos envolvendo a utilização de “Large Language Model” em finanças continuam raros, mas já há uma diversidade crescente de investigações a partir de 2023. Por exemplo, o trabalho de Jesus et al. (2024) utilizou LLMs para atribuir percepção de sentimentos a notícias publicadas por veículos jornalísticos profissionais brasileiros, com o intuito de verificar se tais sinais poderiam ser empregados como preditores de retornos do Ibovespa.

O estudo de Bernard et al. (2024) utilizou a estratégia de analisar a forma como os LLMs escolhem o próximo token com base no anterior para verificar a complexidade de processar informações de relatórios financeiros. De modo a simplificar, se ao analisar um relatório financeiro fosse atribuída baixa confiança na designação de uma nomenclatura financeira – por exemplo, “contas a receber”, considerada de alta confiança e baixa complexidade –, isso indicaria a existência de informações extraordinárias para as quais o modelo teve pouco treinamento, refletindo um atraso na formação do preço após a divulgação do relatório. Bernard et al. (2024) utilizaram relatórios em formato XBRL, evidenciando que o grau de confiança do modelo na atribuição da nomenclatura correta (“TAG”) fornece pistas sobre a complexidade e o tempo de processamento das informações.

Além disso, para a otimização de carteiras, foram implementadas diversas tratativas para verificar se os LLMs seriam capazes de atribuir pesos que maximizassem o retorno com menor volatilidade, conforme abordado por Romanko, Narayan e Kwon (2023) – que investigaram a atribuição direta de pesos – e por Kong et al. (2024) e Ko e Lee (2024), que analisaram a escolha de ações para compor carteiras. Os resultados indicaram que, embora os LLMs ainda não consigam atribuir pesos com a mesma precisão de modelos tradicionais, eles demonstram elementos positivos na seleção dos ativos, com Kong et al. (2024) apontando variações na propensão ao risco entre diferentes LLMs.

Adicionalmente, estudos recentes buscaram validar os resultados dos LLMs por meio de comparações com métodos tradicionais. Chou et al. (2025) verificaram que o ChatGPT 4o, da OpenAI, gera códigos com resultados próximos aos obtidos por softwares estatísticos como o Stata, embora seja necessário ajustar e supervisionar manualmente a identificação correta das colunas dos dados para minimizar discrepâncias.

Segundo Wang e Brorsson (2025), embora os LLMs consigam realizar cálculos, eles ainda apresentam falhas em processos que envolvem múltiplas etapas, como na análise DuPont, no cálculo do Altman Z-score e na projeção de vendas e EBITDA para o próximo ano fiscal, evidenciando limitações na precisão em tarefas complexas.

Na literatura atual, a comparação da previsão do lucro líquido futuro das empresas envolve múltiplas métricas, como o F1-Score – que abrange precisão e sensibilidade das previsões –, e a AUC (Area Under the Curve), além de medidas de erro na previsão, conforme relatado por Houlihan Lokey (2024).

O debate acerca do desempenho dos analistas financeiros em comparação com modelos quantitativos também é antigo. Brown e Rozeff (1978) argumentaram que os analistas, na época, apresentavam desempenho superior aos modelos de séries temporais. Trabalhos mais recentes, como Bradshaw et al. (2010), indicam que os analistas possuem bom desempenho no curto prazo (até 1 ano), mas que, para períodos mais longos e em empresas de menor porte, modelos simples – como o Random Walk ou extrapolações ingênuas – podem igualar ou até superar as previsões dos analistas. Estudos de Clement (1999) e Long e Lee (2022) ressaltaram a importância do tempo de experiência dos analistas e da qualidade dos *accruals* para a obtenção de previsões mais precisas.

## 2.4 Otimização de Carteiras e Seleção de Ativos com LLMs

A adoção de inteligência artificial no setor financeiro cresce, principalmente no que tange à otimização de carteiras de investimentos e à seleção de ativos. O desenvolvimento de LLMs, como o ChatGPT, introduz oportunidades inéditas para a identificação de ativos e a obtenção de eventuais vantagens competitivas.

Em termos de otimização de carteiras, pesquisas recentes têm explorado o uso de LLMs tanto na seleção de ativos quanto na atribuição de pesos. Romanko, Narayan e Kwon (2023) solicitaram ao ChatGPT a indicação de ações do S&P500, combinando os resultados com métodos tradicionais de otimização, como o modelo de Markowitz e o Equal Weights. Embora a atribuição de pesos realizada exclusivamente pelo LLM ainda apresente limitações na relação risco-retorno, os resultados indicaram elementos positivos na escolha dos ativos.

Outros estudos, como o de Kong et al. (2024), aplicaram LLMs em conjunto com o modelo Black-Litterman, substituindo as estimativas tradicionais de retorno e volatilidade por algoritmos de aprendizado de máquina, como o Gradient Boosting Regression (GBR) e o GJR-GARCH. Mesmo quando submetidos às mesmas instruções, observou-se que LLMs como o ChatGPT e o Bard (agora Gemini) podem apresentar preferências distintas – com o ChatGPT indicando ações de maior capitalização e beta próximo de 1, enquanto o Bard tende a sugerir empresas de porte intermediário e betas mais elevados – o que impacta a construção e o desempenho dos portfólios.

Adicionalmente, experimentos envolvendo exemplos de prompts têm demonstrado o potencial dos LLMs na seleção de ativos. Romanko, Narayan e Kwon (2023) utilizaram instruções como:

“Using a range of investing principles taken from leading funds, create a theoretical fund comprising of at least X stocks (mention their tickers) from the S&P500 with the goal to outperform the S&P500 index.” (Romanko, Narayan e Kwon, 2023, p.22)

Essa instrução foi repetida 30 vezes para identificar as ações mais frequentemente sugeridas, seguida pela solicitação ao ChatGPT de distribuir pesos para cada ativo – com a soma total dos pesos igual a 1. De forma semelhante, Kong et al. (2024) empregaram o seguinte prompt:

“You are an investment expert. Use a range of investing principles taken from leading funds and create a theoretical fund comprising at least 5 stocks (mention their tickers) from the S&P500 with the goal to outperform the S&P500 index” (KONG et al., 2024, p.977).

Esses estudos evidenciam que a combinação de LLMs com técnicas tradicionais de otimização pode aprimorar a seleção de ativos e a construção de portfólios, apontando para uma integração promissora entre inteligência artificial e métodos quantitativos tradicionais.

Sobre estratégias de otimização de portfolio, a literatura que vem buscando a vinculação com o LLM está utilizando estratégias como Markowitz (MV) e EW (Equal-Weights) de uma forma com poucas variações. No presente estudo será adotado estratégias complementares de otimização, com atribuição de pesos mínimos e máximos.

Estudos recentes demonstram ganhos de performance dos portfólios quando se incorporam restrições otimizadas de pesos. Abate, Bonafini e Ferrari (2022) evidenciaram que a aplicação de restrições – como a vinculada ao modelo de ponderação igual, que em um universo de 11 ativos impõe pesos que variam de aproximadamente 4,09% (mínimo) a 14,09% (máximo) – pode elevar o índice de Sharpe em média de 22% a 49% em comparação com estratégias benchmark, como o portfólio “Equal-Weights” ou o “Value-Weights”.

De forma complementar, Behr, Guettler e Miebs (2012) desenvolveram uma abordagem de otimização de carteira com restrições de peso calibradas via bootstrap, a qual obteve um ganho médio de 32,5% no índice de Sharpe em relação à estratégia 1/N, recomendando – especialmente para universos maiores – a implementação de um limite inferior próximo de 0% (efetivamente restringindo posições curtas) e um limite superior em torno de 7%. Esses resultados reforçam que, ao combinar técnicas tradicionais (como a otimização via maximização do Sharpe-Ratio e Sortino-Ratio) com métodos avançados de calibração de restrições, é possível obter alocações de ativos mais robustas e com performance ajustada ao risco significativamente aprimorada.

# 3 METODOLOGIA

## 3.1 Tipo de Pesquisa

Esta pesquisa é de natureza quantitativa, com abordagem descritiva, exploratória e experimental. A escolha por uma metodologia quantitativa se justifica pela necessidade de mensurar a performance do modelo LLM “ChatGPT o3-mini-high” na previsão de indicadores econômicos e variáveis financeiras, utilizando métricas estatísticas robustas – como o F1-Score e a AUC – para validar os resultados obtidos. A abordagem descritiva permite mapear e analisar detalhadamente as características das predições e dos relatórios financeiros, enquanto a exploratória visa identificar padrões, inconsistências e oportunidades de melhoria na utilização de LLMs para análises financeiras. O caráter experimental é evidenciado pela realização de múltiplas predições para cada companhia, o ajuste fino dos parâmetros do modelo e a comparação dos resultados com as predições de analistas financeiros e com o desempenho de carteiras de investimentos, contribuindo para o avanço do conhecimento na aplicação de inteligência artificial no campo das finanças.

## 3.2 Fonte dos Dados

Neste estudo, foram utilizados dados dos demonstrativos ITR e DFP de empresas listadas e negociadas no Ibovespa, referentes ao período compreendido entre o quarto trimestre de 2022 e o quarto trimestre de 2024. Desses nove trimestres, cinco foram empregados para a previsão e quatro para a verificação, sendo que o trimestre encerrado em 31 de dezembro de 2023 foi utilizado em ambos os contextos, totalizando 97 empresas analisadas.

Os dados dos demonstrativos ITR foram obtidos na seção “Dados Abertos de Companhias” do site da CVM, utilizando-se especificamente os conjuntos de dados denominados:

• “Cias Abertas: Documentos: Formulário de Informações Trimestrais (ITR)” – para as demonstrações financeiras dos três primeiros trimestres;

• “Cias Abertas: Documentos: Formulário de Demonstrações Financeiras Padronizadas (DFP)” – para os relatórios financeiros anuais.

Complementarmente, foi coletada a última ata da reunião do COPOM anterior à data de divulgação do relatório financeiro em análise, por meio de coleta manual no site do Banco Central do Brasil (BCB).

Para obter a lista de empresas que seriam incorporadas, foi utilizada como referência a lista de ações que compuseram a IBRX-100 em maio de 2023. Como esta lista seria fixa, foi preferido buscar por uma composição que estivesse entre as datas de teste para a previsão. Foram excluídas as ações de empresas que possuíam mais de um código de negociação, de modo que cada companhia fosse representada por apenas uma ação na previsão do lucro líquido futuro da empresa, mas todos os “tickers” foram utilizados no processo de otimização de carteira.

A lista das ações consideradas pode ser vista na Tabela 1, onde a primeira linha contém a ação que deixou de ser negociada na bolsa de valores brasileira, a segunda linha houve fusão das duas empresas para a ação de código “AZZA3”. A terceira linha houve mudanças de códigos das ações sendo hoje representadas por “BHIA3”, “BRAV3” e “ALOS3”, a quarta linha não houve nenhuma modificação e todas as ações seguem iguais.

Tabela 1 – Lista de ações consideradas

|  |
| --- |
| ENBR3, CIEL3 |
| ARZZ3, SOMA3,  |
| VIIA3, RRRP3, ALSO3 |
| ALSO3, ALPA4, ABEV3, ASAI3, AURE3, AZUL4, B3SA3, BBSE3, BBDC3, BBDC4, BRAP4, BBAS3, BRKM5, BRFS3, BPAC11, CRFB3,CBAV3, CCRO3, CMIG4, COGN3, CPLE6, CSAN3, CPFE3, CMIN3, CVCB3, CYRE3, DXCO3, DIRR3, ECOR3, ELET3, ELET6, EMBR3,ENGI11, ENEV3, EGIE3, EQTL3, EZTC3, FLRY3, GGBR4, GOAU4, GOLL4, GMAT3, NTCO3, HAPV3, HYPE3, IGTI11, IRBR3, ITSA4, ITUB4, JBSS3, KLBN11, RENT3, LWSA3, LREN3, MDIA3, MGLU3, MRFG3, CASH3, BEEF3, MOVI3, MRVE3, MULT3, PCAR3, PETR3, PETR4, RECV3, PRIO3, PETZ3, PSSA3, QUAL3, RADL3, RAIZ4, RDOR3, RAIL3, SBSP3, SANB11, STBP3, SMTO3, CSNA3, SIMH3, SLCE3, SUZB3, TAEE11, VIVT3, TIMS3, TOTS3, TRPL4, UGPA3, USIM5, VALE3, VAMO3, VBBR3, WEGE3, YDUQ3, |

Os dados relativos ao histórico de preços das ações analisadas foram obtidos por meio da API da plataforma “EODHD”. Para estabelecer o preço de referência de cada ação durante um período, utilizou-se a chave “adjusted\_close”, a qual reflete os ajustes decorrentes de operações que impactam o preço da ação – excetuando recompra e emissão –, como, por exemplo, pagamentos de dividendos e operações de desdobramento ou agrupamento de ações.

## 3.3 Preparação dos Dados e Abordagens adotadas

O presente estudo utilizou o LLM “ChatGPT o3-mini-high” para realizar análises financeiras e predições dos indicadores econômicos e variáveis financeiras selecionadas. A escolha deste modelo deveu-se à sua data de corte de aprendizagem, que antecedeu a divulgação dos resultados do quarto trimestre de 2023 (conforme “Sep 30, 2023 knowledge cutoff”, disponível no site da OpenAI).

Os parâmetros do modelo foram ajustados com o intuito de reduzir a variação das respostas para um mesmo input. Dessa forma, o parâmetro “temperature” – que determina o grau de aleatoriedade das respostas (quanto menor, menor a divergência) – foi fixado em 0, enquanto o parâmetro “n” (número de respostas geradas para seleção da mais consistente) foi estabelecido em 10, objetivando, assim, reduzir a incerteza das predições e garantir respostas mais constantes ao longo do tempo.

Para o tratamento dos relatórios financeiros trimestrais, adotou-se a seguinte metodologia:

* Conversão dos arquivos em formato PDF para TXT;
* Limpeza dos arquivos, com remoção de espaços e quebras de linha excessivos, bem como de caracteres especiais utilizados na separação de páginas e seções dos relatórios.

Esse procedimento teve como objetivo reduzir o número de tokens de cada documento sem comprometer a qualidade do conteúdo extraído, possibilitando uma diminuição de, pelo menos, 20% do total de tokens. A conversão foi realizada utilizando os pacotes Python “camelot” e “tabulate”. Considerando que o modelo LLM aceita até 200.000 tokens como input, os arquivos que excederam esse limite foram divididos em duas partes. Inicialmente, a segunda parte do documento era utilizada para gerar um resumo dos aspectos relevantes para a predição, que, posteriormente, era incorporado de forma incremental (por meio de prompt) à análise da primeira parte, dado que as informações cruciais (como as demonstrações financeiras consolidadas) normalmente se encontram na primeira metade dos relatórios ITR e DFP.

Para a realização da análise financeira, empregou-se o mesmo método adotado por Kim, Muhn e Nikolaev (2024), o qual se baseia na abordagem “Chain-of-Thoughts” descrita por Bouwman, Frishkoff e Frishkoff (1987), visando emular o raciocínio de um analista financeiro. O prompt utilizado encontra-se disponível no Apêndice A. Adicionalmente, a última ata do COPOM, anterior à divulgação do relatório financeiro, foi fornecida ao modelo para contextualizar a conjuntura macroeconômica vigente no período analisado.

Após a emissão da análise pelo modelo, iniciou-se a etapa de verificação, na qual o LLM “Gemini 2.0 Flash” foi submetido tanto ao resultado obtido quanto ao relatório financeiro da companhia. Nesta fase, o modelo foi instruído a verificar se os dados utilizados pelo “ChatGPT o3-mini-high” estavam em conformidade com os parâmetros estabelecidos (por exemplo, a utilização do lucro líquido das demonstrações consolidadas). Ao final, foi emitida uma avaliação em forma de “Verdadeiro” ou “Falso”. Em caso de avaliação negativa, o relatório financeiro era reenviado ao modelo e a predição descartada.

A realização de 10 previsões para a mesma companhia e trimestre justifica-se pela variação inerente às respostas geradas pelo modelo. Em uma amostra aleatória de 20 companhias, abrangendo cinco trimestres, constatou-se que a consistência média das respostas relativas à previsão de melhora ou piora do lucro líquido foi de 90%, com o menor valor registrado em 82% para um conjunto de 100 requisições, demonstrando que, nas primeiras 10 requisições, o resultado majoritário prevaleceu.

Após a obtenção do resultado de maior prevalência, procedeu-se à avaliação da efetiva ocorrência de melhora ou piora do lucro líquido no relatório do ano subsequente. Essa verificação permitiu a transformação dos dados para o cálculo da acurácia, do F1-Score e da AUC, utilizando os seguintes critérios:

* **TP (True Positive):** atribuído quando a previsão de aumento do lucro líquido se confirma;
* **TN (True Negative):** atribuído quando a previsão de queda do lucro líquido se confirma;
* **FP (False Positive):** atribuído quando a previsão de aumento não se concretiza;
* **FN (False Negative):** atribuído quando a previsão de queda não se concretiza.

O **F1-Score** é calculado conforme a fórmula:

$$F1=2×\frac{Precision×Recall}{Precision+Recall}$$

​

em que:

* **Precision (Precisão):** $Precision=\frac{TP}{TP+FP}$, representando a proporção de predições positivas corretas em relação ao total de predições positivas;
* **Recall (Sensibilidade):** $Recall=\frac{TP}{TP+FN}$ , indicando a capacidade do modelo em identificar corretamente os casos positivos.

A **AUC (Area Under the Curve)** corresponde à área sob a curva ROC (Receiver Operating Characteristic), a qual relaciona a taxa de verdadeiros positivos (TPR) com a taxa de falsos positivos (FPR) em diferentes limiares de decisão. Uma AUC maior indica uma melhor capacidade do modelo em discriminar entre as classes positivas e negativas. No presente estudo, o grau de confiança atribuído pelo LLM também foi utilizado como limiar para segregar e avaliar a assertividade das predições, permitindo a simulação de diversas carteiras e a identificação de variações.

$$AUC = \frac{1}{N\_{+} N\_{-}}\sum\_{i: y\_{i}=1}^{}\sum\_{j: y\_{j}=0}^{}1\left(s\_{i}>s\_{j}\right)$$

Onde:

* N+é o número de observações positivas (y=1y=1),
* N−​ é o número de observações negativas (y=0y=0),
* si​ é a pontuação (ou probabilidade) predita para a observação i,
* sj é a pontuação predita para a observação j,
* 1(si>sj) é a função indicadora que vale 1 se si>sj e 0 caso contrário.

Além disso, foram incorporadas métricas de erro relativas à previsão da magnitude da variação do lucro líquido. Estudos como os de Wang e Brorsson (2025) evidenciam divergências nas aptidões dos modelos LLM para diferentes atividades, sendo que, para exercícios que demandam múltiplas etapas de raciocínio, os resultados não se destacaram.

Com os resultados obtidos, realizou-se a comparação com métricas de analistas financeiros humanos, utilizando o banco de dados I/B/E/S Detail File e um modelo de rede neural artificial (ANN), conforme abordagem similar à de Kim, Muhn e Nikolaev (2024).

Na etapa final do estudo, procedeu-se à construção de uma carteira de investimentos e à avaliação de seu desempenho em relação ao índice Ibovespa, representado pelo BOVA11. Para essa finalidade, a plataforma EODHD foi essencial na coleta do histórico dos preços das ações, ajustados retroativamente por eventos como desdobramentos, agrupamentos e pagamentos de dividendos.

O tratamento de outliers consistiu na identificação e exclusão de datas com valores discrepantes, especialmente em períodos de baixa volatilidade. Por fim, a otimização da carteira foi realizada por meio dos métodos value-weighted, equal-weighted e Markowitz (MV), com rebalanceamentos semanais. Em períodos de divulgação de resultados, as empresas com novas recomendações foram incorporadas a partir da data de divulgação, permitindo a inclusão e exclusão de ações em momentos distintos. Para a construção da matriz de covariância, considerou-se um período de 120 dias (4 meses anteriores à data de início da otimização). Sobre os pesos atribuídos, foi seguido a estratégia de Abate, Bonafini e Ferrari (2022), atribuindo um alfa de mais ou menos 5% para as ações que compõe a carteira naquele momento.

# 4 ANÁLISE DOS RESULTADOS

## 4.1 Desempenho do modelo O3-Mini-High

A Tabela 1 apresenta os resultados do modelo O3-mini-high para cada trimestre avaliado e o consolidado:

Tabela 1 – Método O3-Mini-High na previsão do lucro líquido futuro

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Método** | **Periodo** | **TP** | **FP** | **TN** | **FN** | **Total\_Registros** | **Acuracia** | **Precisao** | **Recall** | **F1\_Score** |
| **O3-Mini-High** |  |  |  |  |  |  |  |  |  |  |
|  | **2023-12-31** | 50 | 34 | 6 | 4 | 94 | 0.5957 | 0.5952 | 0.9259 | 0.7246 |
|  | **2024-03-31** | 55 | 34 | 3 | 1 | 93 | 0.6237 | 0.6180 | 0.9821 | 0.7586 |
|  | **2024-06-30** | 53 | 36 | 2 | 3 | 94 | 0.5851 | 0.5955 | 0.9464 | 0.7310 |
|  | **2024-09-30** | 71 | 19 | 3 | 0 | 93 | 0.7957 | 0.7889 | 1.0000 | 0.8820 |
|  | **2024-12-31** | 56 | 33 | 1 | 2 | 92 | 0.6196 | 0.6292 | 0.9655 | 0.7619 |
|  | **Total Geral** | 285 | 156 | 15 | 10 | 466 | 0.6438 | 0.6463 | 0.9661 | 0.7745 |

Observa-se que:

* A **acurácia** global foi de aproximadamente **64,38%**, o que significa que, em média, o modelo acertou em torno de 59% das previsões de alta ou baixa.
* A **precisão** de cerca de **64,63%** indica que, quando o modelo previu “alta”, em 64,63% dos casos havia realmente uma tendência de alta no lucro líquido.
* O **recall** de **96,61%** mostra que o modelo conseguiu identificar a maior parte das altas reais. Este valor elevado de recall indica que o modelo tende a classificar um número significativo de casos como “alta”, o que também explica o número relativamente alto de falsos positivos (156).
* O **F1-Score** de **77,45%** demonstra um equilíbrio razoável entre precisão e recall, mas ainda com uma ênfase maior em cobrir os casos positivos (alta).

Ao longo dos trimestres, notamos variações nos indicadores. Por exemplo, no período de **2024-09-30**, o recall chegou a **95,45%**, com uma precisão de **73,26%**, resultando em um F1-Score de **82,89%**, o melhor resultado entre os trimestres analisados. Já em outros períodos, como em **2023-12-31** e **2024-12-31**, a acurácia e a precisão são mais baixas, mas o recall ainda se mantém relativamente alto.

Essas variações podem estar associadas a características específicas dos relatórios trimestrais das empresas naquele período, sazonalidades ou mudanças bruscas em determinados setores.

## 4.2 Desempenho do método Naive

A Tabela 2 apresenta os resultados do método Naive, que faz a projeção da tendência de acordo com o comportamento do lucro líquido no mesmo trimestre do ano anterior:

Tabela 2 – Método Naive na previsão do lucro líquido futuro

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Método** | **Periodo** | **TP** | **FP** | **TN** | **FN** | **Total\_Registros** | **Acuracia** | **Precisao** | **Recall** | **F1\_Score** |
| **Naive** |  |  |  |  |  |  |  |  |  |  |
|  | **2023-12-31** | 20 | 17 | 25 | 32 | 94 | 0.4787 | 0.5405 | 0.3846 | 0.4494 |
|  | **2024-03-31** | 25 | 12 | 27 | 29 | 93 | 0.5591 | 0.6757 | 0.4630 | 0.5495 |
|  | **2024-06-30** | 28 | 12 | 25 | 28 | 93 | 0.5699 | 0.7000 | 0.5000 | 0.5833 |
|  | **2024-09-30** | 30 | 16 | 18 | 29 | 93 | 0.5161 | 0.6522 | 0.5085 | 0.5714 |
|  | **2024-12-31** | 34 | 15 | 20 | 24 | 93 | 0.5806 | 0.6939 | 0.5862 | 0.6355 |
|  | **Total Geral** | 137 | 72 | 115 | 142 | 466 | 0.5408 | 0.6555 | 0.4910 | 0.5615 |

Os resultados mostram que:

* A **acurácia** global do método Naive foi de cerca de **54,08%**, ligeiramente inferior ao método O3-mini-high.
* A **precisão** de **65,55%** é curiosamente superior à do modelo O3-mini-high (64,63%). Isso significa que, quando o Naive prevê “alta”, ele acerta em uma proporção maior dos casos de “alta” do que o O3-mini-high. Porém, isso vem ao custo de um **recall** menor, de **49,39%**, indicando que o Naive não consegue detectar muitas das altas reais.
* O **F1-Score** de **56,15%** é consideravelmente menor que o do O3-mini-high (77,45%), sugerindo que o método Naive é menos equilibrado entre Precisão e Recall.

Esses resultados indicam que o modelo O3-mini-high supera o método Naive no geral, principalmente por apresentar um recall muito maior. Isso sugere que a abordagem baseada em múltiplas variáveis (O3-mini-high) consegue capturar mais fatores que influenciam o comportamento do lucro líquido, em contraste com a estratégia simples de olhar apenas para o trimestre anterior e o atual.

## 4.3 Cálculo AUC do modelo O3-Mini-High

Na análise dos resultados do AUC para o modelo O3-mini-high, observa-se que os valores se mantêm, na maior parte dos períodos, levemente acima de 0,5, com um valor médio de 0.5385. Esse desempenho do AUC indica que a capacidade do modelo para ranquear as instâncias — isto é, para atribuir graus de confiança que permitam distinguir adequadamente entre casos positivos e negativos — é limitada.

É importante notar que, embora o modelo apresente um recall elevado (por exemplo, um recall geral de 96,61%), esse resultado pode ser, em parte, consequência do fato de o modelo estar prevendo um número excessivo de positivos. Em outras palavras, o alto recall pode estar associado a um viés na predição, onde o modelo "arrisca" mais positivos, aumentando a sensibilidade, mas também elevando o número de falsos positivos. Esse comportamento explicaria, em parte, o desempenho insatisfatório no AUC, já que, para uma boa capacidade de ranqueamento, espera-se que as previsões de confiança reflitam, de forma mais acurada, a verdadeira probabilidade dos eventos positivos.

Tabela 3 – Resultados AUC sobre as previsões do modelo O3-Mini-High

|  |  |  |
| --- | --- | --- |
| **Periodo** | **AUC** | **Num\_Empresas\_AUC** |
| **2023-12-31** | 0.5041 | 88 |
| **2024-03-31** | 0.5517 | 90 |
| **2024-06-30** | 0.4988 | 93 |
| **2024-09-30** | 0.5189 | 90 |
| **2024-12-31** | 0.5499 | 91 |
| **Total Geral** | 0.5385 | 452 |

Essa discrepância entre o alto recall e o baixo AUC sugere que o modelo, ao prever um número elevado de casos positivos, consegue captar a maioria dos eventos de alta, porém a confiança atribuída a essas previsões não discrimina de forma eficaz entre os casos verdadeiramente positivos e os falsos positivos. Em resumo, o elevado recall pode ser, em parte, um reflexo de uma estratégia de predição tendenciosa para o positivo, o que impacta negativamente a qualidade da ordenação das previsões – aspecto capturado pelo AUC.

Portanto, a hipótese levantada é que o modelo O3-mini-high pode estar priorizando a sensibilidade excessivamente, resultando em atribuir valores aleatórios ao classificar muitos casos como positivos, o que, embora melhore o recall, compromete a capacidade de ranqueamento e a confiança das predições, conforme demonstrado pelos baixos valores de AUC. Essa constatação aponta para a necessidade de um ajuste na calibração do modelo, de modo que se consiga um equilíbrio mais adequado entre a identificação dos casos positivos e a qualidade das predições de confiança.

## 4.4 Performance das ações com base nas previsões do modelo O3-Mini-High

O Gráfico 1 ilustra a **performance acumulada**de uma carteira de ações construída com base nas previsões do modelo O3-Mini-High, no período de 06/02/2023 a 27/05/2024, em comparação com três referências de mercado: o **BOVA11**(proxy do Ibovespa), o **BRAX11** (representativo do IBRX 100) e a **Taxa Livre de Risco**de 12,75% ao ano.

Grafico 1 – Otimização via Makowitz com base nas previsões de “Melhora” do modelo O3-Mini-High



Observa-se que a carteira, mesmo considerando um custo operacional de 0,1%, apresenta uma trajetória ascendente e se mantém **consistentemente acima** dos dois índices acionários ao longo de grande parte do horizonte analisado. Esse resultado sugere que as previsões do modelo O3-Mini-High, quando utilizadas para selecionar e rebalancear os ativos, geraram um retorno acumulado superior ao do Ibovespa (BOVA11) e ao do IBRX 100 (BRAX11). Adicionalmente, a carteira supera a curva de crescimento da taxa livre de risco, indicando que o **prêmio pelo risco** assumido foi compensado por um retorno significativamente maior.

É importante notar que, embora não estejam explicitadas métricas de **volatilidade** no gráfico, a análise visual sugere uma estabilidade relativa da carteira frente a possíveis oscilações mais intensas dos benchmarks. Em termos práticos, isso pode significar que a estratégia de alocação baseada nas previsões do O3-Mini-High não apenas se mostra capaz de capturar momentos de valorização do mercado, mas também de mitigar, ao menos parcialmente, eventuais períodos de queda.

Dessa forma, os resultados observados na Figura 4.4 reforçam o potencial do modelo O3-Mini-High em fornecer subsídios para a montagem de uma carteira de ações com desempenho superior aos índices de mercado e à taxa livre de risco, validando, em parte, a proposta de utilização de previsões de lucro líquido futuro como critério de seleção e alocação de ativos.

5 CONCLUSÃO

Este trabalho explorou a fronteira entre inteligência artificial e análise financeira, investigando se modelos LLMs, especificamente o ChatGPT o3-mini-high, podem rivalizar com analistas humanos na previsão de resultados financeiros e otimização de carteiras de investimento. A metodologia adotada, combinando técnicas avançadas de *prompt engineering* com métricas rigorosas de avaliação, representa um avanço importante na literatura. Entretanto, é crucial reconhecer as limitações inerentes ao estudo.

Primeiramente, os LLMs, apesar de sua sofisticação, ainda enfrentam desafios significativos em cálculos sequenciais complexos, conforme apontado por Wang e Brorsson (2025). A dependência exclusiva desses modelos para decisões financeiras pode ser problemática, especialmente em cenários de mercado atípicos ou durante eventos econômicos sem precedentes históricos.

Outra limitação relevante refere-se à natureza determinística das previsões financeiras. Como demonstrado por Bernard et al. (2024), a confiança dos modelos na atribuição de nomenclaturas financeiras varia significativamente, indicando que informações extraordinárias podem não ser adequadamente processadas, resultando em atrasos na formação de preços após divulgações.

Adicionalmente, o estudo se concentra em um período específico e em empresas brasileiras de grande capitalização, o que limita a generalização dos resultados para outros mercados, períodos ou categorias de empresas. A validação cruzada em diferentes contextos econômicos seria fundamental para confirmar a robustez das conclusões.

O trabalho também apresenta limitações metodológicas quanto à interpretabilidade das decisões tomadas pelo modelo. Diferentemente dos analistas humanos, que podem explicar detalhadamente seu raciocínio e adaptar suas análises a novos contextos rapidamente, os LLMs operam como "caixas-pretas" em determinados aspectos, o que pode comprometer a confiabilidade em cenários críticos de mercado.

Para pesquisas futuras, seria valioso explorar abordagens híbridas que combinem a capacidade de processamento dos LLMs com a expertise contextual e o julgamento crítico de profissionais humanos. Além disso, a incorporação de dados alternativos e o desenvolvimento de métricas mais sofisticadas para avaliar a qualidade das análises poderiam enriquecer significativamente o campo.

Em última análise, este trabalho não sugere uma substituição completa dos analistas financeiros humanos, mas indica um futuro onde esses profissionais possam ampliar suas capacidades utilizando LLMs como ferramentas poderosas de apoio à decisão, potencialmente transformando a natureza da profissão e democratizando o acesso a análises financeiras de qualidade.6 REFERÊNCIAS BIBLIOGRÁFICAS

**ABATE, G., BONAFINI, T., & FERRARI, P.** (2022). *Portfolio constraints: An empirical analysis*. International Journal of Financial Studies, 10(1), 9. <https://doi.org/10.3390/ijfs10010009>

**ALDRIDGE, Irene.** The AI Revolution: From Linear Regression to ChatGPT and Beyond and How It All Connects to Finance. *The Journal of Portfolio Management*, v.49, n.9, p.64–77, 2023. DOI: 10.3905/jpm.2023.1.519. Disponível em: https://doi.org/10.3905/jpm.2023.1.519. Acesso em: 25 mar. 2025. [Artigo de periódico]

**ALI, Hassnian; AYSAN, Ahmet Faruk.** What Will ChatGPT Revolutionize in Financial Industry? [Trabalho técnico]. Rochester, NY: SSRN, 28 mar. 2023. DOI: 10.2139/ssrn.4403372. Disponível em: https://papers.ssrn.com/abstract=4403372. Acesso em: 25 mar. 2025. [Trabalho técnico]

**ALLEN, Darcy W. E. *et al*.** Large Language Models Reduce Agency Costs. [Trabalho técnico]. Rochester, NY: SSRN, 4 maio 2023. DOI: 10.2139/ssrn.4437679. Disponível em: https://papers.ssrn.com/abstract=4437679. Acesso em: 25 mar. 2025. [Trabalho técnico]

**ALMEIDA, José; GONÇALVES, Tiago Cruz.** The AI Revolution: Are Crypto Markets More Efficient after ChatGPT 3? *Finance Research Letters*, v.66, 2024. DOI: 10.1016/j.frl.2024.105608. Disponível em: https://doi.org/10.1016/j.frl.2024.105608. Acesso em: 25 mar. 2025. [Artigo de periódico]

**ANTE, Lennart; DEMIR, Ender.** The ChatGPT effect on AI-themed cryptocurrencies. *Economics and Business Letters*, v.13, n.1, p.29–38, 2024. DOI: 10.17811/ebl.13.1.2024.29-38. Disponível em: https://doi.org/10.17811/ebl.13.1.2024.29-38. Acesso em: 25 mar. 2025. [Artigo de periódico]

**BEHR, P., GUETTLER, A., & MIEBS, F.** (2012). On portfolio optimization: Imposing the right constraints. SSRN Electronic Journal. [http://ssrn.com/abstract=1618756](http://ssrn.com/abstract%3D1618756)

**BERTOMEU, Jeremy *et al*.** The Impact of Generative AI on Information Processing: Evidence from the Ban of ChatGPT in Italy. *Journal of Accounting and Economics*, no prelo, 2025. DOI: 10.2139/ssrn.5118661. Disponível em: https://papers.ssrn.com/abstract=5118661. Acesso em: 25 mar. 2025. [Artigo de periódico no prelo]

**DONG, Mengming Michael; STRATOPOULOS, Theophanis C.; WANG, Victor Xiaoqi.** A Scoping Review of ChatGPT Research in Accounting and Finance. [Trabalho técnico]. Rochester, NY: SSRN, dez. 2023. DOI: 10.2139/ssrn.4680203. Disponível em: https://papers.ssrn.com/abstract=4680203. Acesso em: 25 mar. 2025. [Trabalho técnico]

**FAIRHURST, Douglas J.; GREENE, Daniel.** How Much Does ChatGPT Know About Finance? *Financial Analysts Journal*, no prelo, 2024. DOI: 10.2139/ssrn.4974976. Disponível em: https://papers.ssrn.com/abstract=4974976. Acesso em: 25 mar. 2025. [Artigo de periódico no prelo]

**HANSEN, Anne Lundgaard; KAZINNIK, Sophia.** Can ChatGPT Decipher Fedspeak? [Apresentação em conferência]. Apresentado em: *The Fourth New York Fed Conference on FinTech: Artificial Intelligence and Digital Assets*, Nova York, 29 set. 2023. Disponível em: https://www.newyorkfed.org/medialibrary/media/research/conference/2023/FinTech/5\_400pm\_Hansen\_slides\_Kazinnik.pdf. Acesso em: 25 mar. 2025. [Apresentação em conferência]

**JHA, Manish *et al*.** ChatGPT and Corporate Policies. [Trabalho técnico]. Rochester, NY: SSRN, 28 fev. 2025. DOI: 10.2139/ssrn.4521096. Disponível em: https://papers.ssrn.com/abstract=4521096. Acesso em: 25 mar. 2025. [Trabalho técnico]

**KIM, Alex G.; MUHN, Maximilian; NIKOLAEV, Valeri V.** Bloated Disclosures: Can ChatGPT Help Investors Process Information? [Trabalho técnico]. Rochester, NY: SSRN, 27 ago. 2024. DOI: 10.2139/ssrn.4425527. Disponível em: https://papers.ssrn.com/abstract=4425527. Acesso em: 25 mar. 2025. [Trabalho técnico]

**KIM, Jang Ho.** What if ChatGPT were a quant asset manager. *Finance Research Letters*, v.58, 2023. DOI: 10.1016/j.frl.2023.104580. Disponível em: https://doi.org/10.1016/j.frl.2023.104580. Acesso em: 25 mar. 2025. [Artigo de periódico]

**LOPEZ-LIRA, Alexander; TANG, Yuan.** Can ChatGPT Forecast Stock Price Movements? Return Predictability and Large Language Models. [Pré-publicação]. arXiv:2304.07619, 2023. Disponível em: https://arxiv.org/abs/2304.07619. Acesso em: 25 mar. 2025. [Pré-publicação]

**LU, Fangzhou; HUANG, Lei; LI, Sixuan.** ChatGPT, Generative AI, and Investment Advisory. [Trabalho técnico]. Rochester, NY: SSRN, 24 jul. 2023. DOI: 10.2139/ssrn.4519182. Disponível em: https://papers.ssrn.com/abstract=4519182. Acesso em: 25 mar. 2025. [Trabalho técnico]

**LUDWIG, Eric T.; BENNETTS, Chet R.** Streamlining Financial Planning with ChatGPT: A Collaborative Approach between Technology and Human Expertise. *Journal of Financial Planning*, v.36, n.6, 2023. Disponível em: https://www.financialplanningassociation.org/learning/publications/journal/JUN23-streamlining-financial-planning-chatgpt-collaborative-approach-between-technology-and-human-OPEN. Acesso em: 25 mar. 2025. [Artigo de periódico]

**MBANYELE, William.** Generative AI and ChatGPT in Financial Markets and Corporate Policy: A Comprehensive Review. [Trabalho técnico]. Rochester, NY: SSRN, 1 fev. 2024. DOI: 10.2139/ssrn.4745990. Disponível em: https://papers.ssrn.com/abstract=4745990. Acesso em: 25 mar. 2025. [Trabalho técnico]

**NIE, Yuqi *et al*.** A Survey of Large Language Models for Financial Applications: Progress, Prospects and Challenges. [Pré-publicação]. arXiv:2406.11903, 2024. Disponível em: https://arxiv.org/abs/2406.11903. Acesso em: 25 mar. 2025. [Pré-publicação]

**WU, Steven *et al*.** BloombergGPT: A Large Language Model for Finance. [Pré-publicação]. arXiv:2303.17564, 2023. Disponível em: https://arxiv.org/abs/2303.17564. Acesso em: 25 mar. 2025. [Pré-publicação]

Apêndice A: Modelo de Prompt Utilizado na Análise

Contexto:

Você é um analista financeiro experiente, encarregado de prever o lucro líquido por ação (EPS) para o próximo ano fiscal de uma empresa listada publicamente. Para isso, você dispõe de relatório ITR/DFP trimestral ou anual de empresas brasileiras listadas no ibovespa. Para a previsão do lucro líquido dê sempre prioridade aos resultados consolidados trimestrais, com exceção para arquivos que abrangem o ano fiscal inteiro.

Instruções Gerais:

Utilize seu checklist pré-definido e seus templates setoriais – conhecimento de treinamento e estruturas mentais que sintetizam os indicadores-chave e os padrões típicos do setor – para guiar sua análise. Ao longo do processo, combine informações quantitativas e qualitativas, compare os dados, ao final, realize uma análise de sensibilidade para identificar os principais riscos e variabilidades que possam afetar a previsão.

Passo 1: Preparação e Checklists

- Revisão do Checklist e Templates: Inicie consultando seu checklist normativo e seus templates setoriais, que destacam os itens essenciais e os padrões de desempenho típicos para o setor (aquilo que você espera do setor em questão).

- Contextualização: Identifique o setor da empresa, suas principais linhas de produtos e sua posição no mercado, usando os documentos disponíveis.

Passo 2: Familiarização

- Análise dos Demonstrativos: Revise a demonstração de resultados dos últimos três anos (ou 3 trimestres, caso seja ITR) para identificar tendências em receita, lucro bruto, despesas operacionais e lucro líquido.

- Cálculo do EPS Histórico: Calcule o EPS (lucro líquido dividido pelo número de ações em circulação) e observe padrões ou anomalias.

- Utilização do Checklist: Certifique-se de que os itens críticos do checklist estejam sendo analisados, garantindo uma visão completa e sistemática da empresa.

Passo 3: Exploração

Análise Detalhada dos Dados:

- Examine informações segmentadas (por produtos ou regiões) para compreender a performance de cada área.

- Revise o balanço patrimonial e a demonstração de fluxos de caixa para avaliar liquidez, solvência e a estrutura de capital.

- Cálculo e Comparação de Ratios: Calcule índices como margem de lucro, ROA e ROE e compare com benchmarks de resultados passados da empresa.

Passo 4: Integração e Formulação de Hipóteses

- Integração dos Dados: Combine as informações quantitativas e qualitativas para formar uma visão coesa do desempenho financeiro.

- Análise Qualitativa: Considere fatores como a qualidade da gestão, condições de mercado e ambiente competitivo.

Formulação de Hipóteses:

- Utilize os dados integrados e o template setorial para formular hipóteses sobre o EPS futuro.

- Atribua pesos aos indicadores críticos e compare com médias do setor (por exemplo, “se a receita crescer 5% e a margem se mantiver, o EPS deverá ser X”).

- Valide suas hipóteses;

Passo 5: Previsão e Análise de Sensibilidade/Riscos

Modelagem da Previsão:

- Projete o lucro líquido futuro utilizando um modelo simples que ajuste a receita, as despesas (com base nas margens históricas) e os impostos.

- Calcule o EPS previsto dividindo o lucro líquido projetado pelo número estimado de ações em circulação (ajustando por recompra ou emissão, se necessário).

Análise de Sensibilidade e Riscos:

- Realize uma varredura final (scanning) nos documentos para identificar possíveis dados ou anomalias que não foram considerados inicialmente.

- Teste diferentes cenários (por exemplo, variações na receita ou na margem de lucro) para avaliar a sensibilidade da previsão.

- Identifique e destaque os principais riscos e incertezas que podem impactar o EPS, explicando como esses fatores foram considerados na previsão.

Saída Esperada:

Forneça uma previsão específica do EPS para o próximo ano fiscal (exemplo: "US$ 2,50 por ação") acompanhada de uma breve explicação que detalhe:

- Os principais dados e tendências observados (tanto quantitativos quanto qualitativos);

- A forma como o checklist e os templates foram utilizados para orientar a análise;

- Os cenários testados e os riscos identificados durante a análise de sensibilidade;

- Apresentar em formato “json” a projeção de lucro líquido futuro, bem como se a expectativa é de “melhora” ou “piora” e o grau de confiança dessa previsão.