LLMs Aplicados à Previsão do Mercado de Ouro

Raul Pinheiro Rocha[1]*; Ricardo Limongi[2]

[1] MBA em Ciência de Dados e Analytics, Universidade de São Paulo (USP). Rua Teixeira Vasconcelos, nº 52 - Bairro Cachoeirinha; 31150-090 – Belo Horizonte, MG, Brasil

[2] Professor de Marketing, Universidade Federal de Goiás (UFG). Campus Samambaia, PO Box 131, 74001-970 – Goiânia, GO, Brasil

* Autor correspondente: raulrocha.rpr@gmail.com

LLMs Aplicados à Previsão do Mercado de Ouro

Resumo

O ouro foi historicamente utilizado como ativo de proteção em contextos de incerteza, razão pela qual se investigou o uso de técnicas de inteligência artificial para antecipar seus movimentos. Este trabalho teve como objetivo aplicar modelos de linguagem de grande porte (LLMs) com fine-tuning para prever a direção e a magnitude do par XAU/USD a partir de notícias financeiras. Assim, construiu-se um conjunto supervisionado que relacionou manchetes a variações de preço em diferentes horizontes, com apoio de reescrita de textos, análise de sentimento e dados correlatos. O modelo Mistral-7B-Instruct foi ajustado com LoRA em infraestrutura AWS-SageMaker, inicialmente na instância ml.g5.xlarge e, posteriormente, na ml.g5.12xlarge equipada com 4 GPUs NVIDIA A10G, 48 vCPUs e 192 GiB de RAM, a fim de reduzir o tempo de execução. Resultados intermediários indicaram bom desempenho em previsão de direção e desempenho moderado em magnitude; entretanto, métricas infladas foram registradas devido à fuga de informação ocasionada pelo uso de data augmentation combinado ao parâmetro load_best_model_at_end. Para mitigar esse viés, adotaram-se medidas como recomposição dos splits, aplicação do augmentation apenas após o particionamento e validação fixa para early stopping. Os resultados corrigidos encontravam-se em execução no encerramento do trabalho, sendo registrados como limitação metodológica e diretriz para continuidade. Até esse estágio, os custos acumulados em infraestrutura AWS totalizaram US$ 118,15, concentrados em treinamento e inferência. Concluiu-se que LLMs apresentaram potencial de aplicação em finanças, desde que apoiados por protocolos rigorosos que assegurem robustez metodológica, estabilidade preditiva e documentação transparente de custos.

Palavras-chave: Modelos de Linguagem; Fine-Tuning; XAU/USD; Notícias Financeiras; Robustez Metodológica

Introdução

O desenvolvimento dos modelos de Large Language Models (LLMs) teve início com a introdução da arquitetura Transformer, proposta por Vaswani *et al.* (2017), e evoluiu com o surgimento de modelos pré-treinados como BERT e GPT. Dado o elevado custo de pré-treinamento, que pode ultrapassar um milhão de dólares, conforme Sharir *et al.* (2020), o fine-tuning passou a ser adotado como alternativa viável para adaptação a tarefas específicas (Chung *et al.,* 2022).

No contexto financeiro, técnicas de Processamento de Linguagem Natural (NLP) vêm sendo amplamente empregadas na extração de sentimento textual e na avaliação de impactos de notícias sobre o mercado. Modelos como a FinBERT, uma adaptação da BERT voltada ao setor financeiro, têm sido utilizados para que o tom de notícias seja classificado automaticamente em categorias como positiva, negativa ou neutra. Além disso, Shen e Zhang (2024) mostram que o modelo GPT-4o, mesmo com poucos exemplos financeiros, pode alcançar desempenho comparável ao FinBERT, um modelo ajustado especificamente para textos financeiros em tarefas de classificação de sentimento. Esse resultado sugere que, em determinados contextos, modelos de linguagem de uso geral podem aproximar-se do desempenho de modelos especializados por meio de adaptação adequada.

Essa evolução reforça a relevância da aplicação de LLMs em finanças, pois evidencia que diferentes arquiteturas podem capturar padrões informativos em textos noticiosos. Tal capacidade é particularmente importante porque as divulgações financeiras influenciam diretamente a percepção dos investidores e, consequentemente, afetam a formação dos preços dos ativos. No caso do ouro (XAU), essa sensibilidade é reforçada por fundamentos econômicos bem estabelecidos, como seu papel de ativo de refúgio em cenários de incerteza (Baur & Lucey, 2010), a correlação inversa com o dólar (Joy, 2011), e a relação com juros reais, que afetam o custo de oportunidade (Barsky, 2021). Adicionalmente, fluxos financeiros em ETFs (Cheng, 2020; WGC, 2025) e compras recordes de bancos centrais desde 2022 reforçam a demanda por ouro, aumentando sua sensibilidade a eventos macroeconômicos (Maghyereh & Abdoh, 2022).

A literatura também documenta de forma consistente a relação entre sentimento noticioso e movimentos de mercado. Tetlock (2007) analisou o impacto do pessimismo midiático sobre ações, enquanto Smales (2015) demonstrou que o sentimento presente em newswires afeta a volatilidade e os retornos do ouro. Sinha & Khandait (2020) apresentaram evidências causais da influência de notícias sobre commodities, e Maghyereh & Abdoh (2022) identificaram que índices de sentimento ajudam a prever bolhas em metais preciosos.

Apesar dos avanços das ferramentas de NLP, a análise de notícias financeiras ainda é conduzida, em grande parte, de forma manual ou com ferramentas limitadas. Em especial, não há soluções amplamente difundidas que combinem análise textual automatizada com previsão conjunta da direção, magnitude e latência das reações do XAU/USD. Essa lacuna metodológica pode ser atribuída a diferentes fatores: o uso de pipelines separados para texto e dados numéricos, pelos quais a aprendizagem conjunta é limitada (Zhang *et al.,* 2024); a dificuldade em incorporar contexto histórico informativo (Zhao et al., 2025); o custo computacional de arquiteturas integradas (Sharir et al., 2020).

Diante disso, são propostas as seguintes etapas: (i) a construção de um dataset supervisionado que relaciona notícias a janelas futuras de variação do ouro (6h, 12h, 24h e 48h), rotuladas com direção e magnitude do movimento, além de explicações qualitativas do racional econômico; (ii) o enriquecimento de cada instância com variáveis de contexto econômico-financeiro provenientes de ativos correlatos, como ETFs de ouro, índices cambiais, juros reais e ativos substitutos; e (iii) o treinamento de um LLM ajustado por meio de fine-tuning com LoRA, capaz de mapear texto e contexto em previsões específicas por horizonte temporal.

Em síntese, ao integrar fundamentos econômicos do ouro com técnicas de NLP financeiro, combinando sentimento especializado e fine-tuning eficiente, este estudo busca antecipar o sentido, a intensidade e a latência das oscilações do par XAU/USD em resposta a eventos noticiosos, oferecendo um modelo metodológico replicável para outros ativos sensíveis à informação macrofinanceira.

Objetivo

- Classificar notícias que influenciam a paridade XAU/USD como etapa fundamental para apoiar operações de trading no mercado de ouro.

Metodologia

A metodologia empregada foi iniciada pela coleta de dados financeiros, abrangendo tanto cotações históricas do par XAU/USD quanto notícias relacionadas ao mercado de ouro. Em seguida, as informações textuais foram submetidas a pré-processamento com técnicas de processamento de linguagem natural, incluindo limpeza, tokenização e remoção de ruídos, de modo a padronizar o conteúdo. A partir desse material, foi elaborado um

dataset rotulado, no qual anotações referentes à direção e magnitude foram atribuídas a cada notícia do impacto esperado no preço do ouro. Por fim, realizou-se o treinamento de um modelo de linguagem de grande porte (LLM), ajustado por meio de fine-tuning, com o objetivo de prever os efeitos dessas notícias sobre o par XAU/USD. Em resumo, foram integradas notícias financeiras relevantes ao ouro com cotações históricas do par XAU/USD, extraídos atributos característicos (como sentimento e indicadores técnicos) e construído um pipeline de aprendizado de máquina para estimar a direção e magnitude do efeito de cada notícia no preço do ouro em diferentes horizontes temporais.

Sites como Investing.com, FXStreet e Forexlive, News API não fornecem extração via API (Application Programming Interface), não permitem web scraping ou não disponibilizam dados históricos de notícias, a não ser com o pagamento mensal aos respectivos CNPJs de $50 a $150 dólares. Dessa forma, a escolha da API da Alpaca Market Docs (https://docs.alpaca.markets/) se deu pela sua disponibilidade gratuita, com acesso programático estruturado e histórico suficiente para fins experimentais e acadêmicos. Já o MetaTrader será utilizado como fonte para os dados históricos de preço e volatilidade do ouro, essenciais para o treinamento e validação do modelo.

Para analisar o impacto de notícias no par XAU/USD, inicialmente foram coletadas duas fontes principais de dados: notícias financeiras e cotações históricas do ouro. As notícias foram obtidas por meio da API Alpaca Markets, que fornece manchetes e conteúdo de notícias financeiras em tempo real. Foram filtradas notícias entre 2020 e 2025 que mencionassem ativos relacionados ao ouro, por exemplo, ETFs lastreados em ouro (como GLD, IAU) ou outros símbolos financeiros correlacionados ao preço do ouro (vide tabela 1). Cada notícia retornada pela API vem acompanhada de metadados, incluindo data/hora de publicação e uma lista de símbolos afetados. Assim, quando uma notícia referenciava múltiplos ativos, cada ativo relevante (por exemplo, o ETF GLD) foi tratado como uma entrada separada no estudo, associando a notícia àquele símbolo específico.

Em paralelo, foi extraída a série temporal de preços do XAU/USD utilizando a plataforma MetaTrader 5. Foram obtidos dados de candles históricos do ouro com intervalos de 30 minutos, incluindo preços de abertura, máxima, mínima, fechamento e volumes, ao longo do mesmo período das notícias coletadas. Os dados de preço foram ajustados para o fuso horário de negociação relevante e tratados para distinguir períodos de mercado aberto e fechado, por exemplo, removendo ou marcando horários de fim de semana e feriados em que não há negociação ativa. Essa distinção é crucial para determinar o impacto temporal das notícias: se uma notícia é divulgada fora do horário de mercado, o efeito potencial no preço só poderá materializar-se quando o mercado reabrir.

Reescrita de Manchetes com Modelo de Linguagem

As manchetes coletadas frequentemente descrevem eventos de mercado de forma genérica ou com foco em ações específicas, o que pode diluir sua relevância para o ouro. Para contornar isso, aplicou-se uma reescrita de manchetes usando um Modelo de Linguagem de Grande Porte (LLM), o Mistral-7B-Instruct v0.2. Esse modelo generativo foi utilizado em modo de inferência para produzir uma versão alternativa de cada manchete, enfatizando explicitamente as implicações para o ouro ou para o ativo relacionado ao ouro em questão. Em outras palavras, dado o título original da notícia, foi fornecido ao LLM um prompt como o exemplificado na Figura 1.

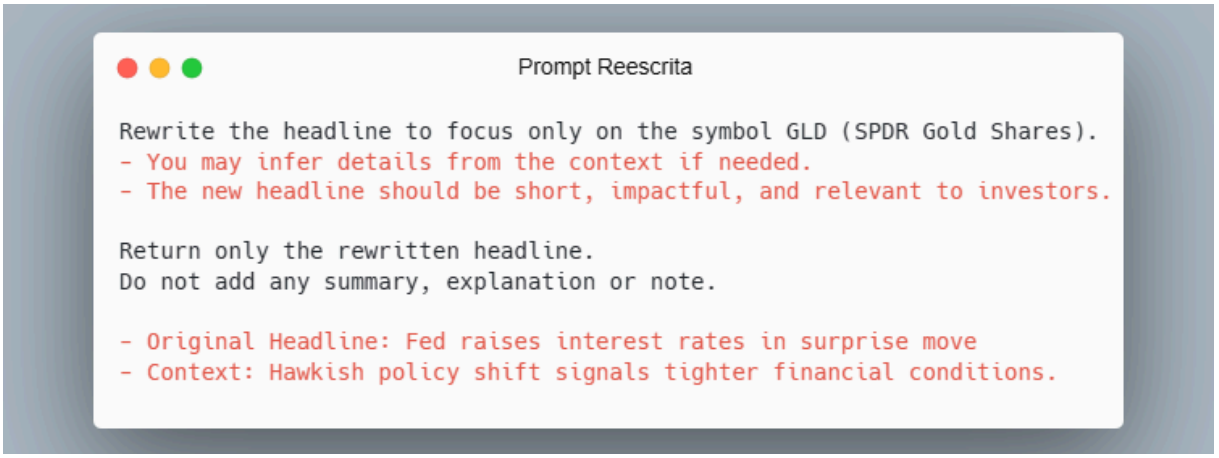

Figura 1: Exemplo de prompt utilizado. Observação: o *context* adicionado é o corpo da notícia, envolve reações para diversos símbolos e fornecia uma visão geral do comportamento do mercado.

Em outras palavras, algo semelhante à "Reescreva esta manchete destacando o impacto ou as implicações para o ouro (XAU/USD) e o ativo X", onde X representa o símbolo financeiro em foco (por exemplo, GLD - Anexo 1). Como resultado, obtivemos manchetes reescritas que incluem menção explícita ao ouro ou ao ativo correlato, facilitando a identificação do tópico principal. Por exemplo, uma manchete original *"Stock Market Update For The Week Ahead"* foi reescrita como *"Fed Shifts Policy: Average Inflation Target, Implications for Gold (GLD)"*, tornando claro o vínculo com ouro através do ticker GLD (considera-se o corpo da notícia para a geração contextualizada). Essas manchetes reestruturadas servem para focar o conteúdo no que é relevante para o preço do ouro, provendo ao modelo de previsão uma entrada textual mais informativa. Cabe ressaltar que essa etapa foi executada de forma automática em lote, utilizando o modelo Mistral pré-treinado (sem ajuste fino específico para essa tarefa) em ambiente gerenciado no Amazon SageMaker Processing, por meio de um container customizado. Essa abordagem

permitiu gerar rapidamente manchetes personalizadas para milhares de notícias, assegurando consistência no formato. Em seguida, foram selecionadas 20 reescritas para avaliação manual da qualidade. Para maior robustez, aplicou-se ainda o modelo GPT-4o em um subconjunto de 100 notícias reescritas.

Análise de Sentimento das Manchetes

Com as manchetes já direcionadas ao contexto do ouro, procedeu-se à análise de sentimento por meio do FinBERT, uma versão da arquitetura BERT treinada especificamente para textos financeiros e capaz de classificar sentenças como Positivas, Neutras ou Negativas. Essa escolha se justificou pela capacidade do modelo de captar nuances de linguagem econômica e de mercado que modelos genéricos poderiam não identificar. Para cada manchete reescrita, o FinBERT gerou um rótulo de sentimento e um escore de confiança (probabilidade atribuída à classe predita). Esse escore foi também utilizado para qualificar a intensidade do sentimento em faixas de fraco, moderado ou forte. Por exemplo, para manchetes classificadas como Positivas, valores acima de 90% indicaram sentimento fortemente positivo, entre 70% e 90% moderadamente positivo e abaixo de 70% fracamente positivo. Critérios equivalentes foram aplicados às classes Neutra e Negativa. Essa estratificação fornece uma medida da intensidade do sentimento expresso na notícia, o que pode influenciar o impacto no preço. Por exemplo, uma manchete fortemente negativa sobre a economia tende a pressionar mais o ouro (ativo de refúgio) do que uma notícia apenas levemente negativa. Todo o processamento de sentimento foi automatizado via *pipeline* do HuggingFace Transformers integrado ao FinBERT, garantindo padronização e velocidade. Os resultados de sentimento de todas as manchetes foram então incorporados ao conjunto de dados em construção.

Enriquecimento e Estruturação do Dataset

Com os dados brutos (manchetes reescritas com sentimento) e as cotações alinhadas temporalmente, efetuou-se uma série de passos de enriquecimento para compor o dataset final estruturado utilizado no treinamento do modelo preditivo. Esta etapa, realizada majoritariamente via consultas SQL em Athena, integrou as diversas fontes de informação e extraiu variáveis explanatórias e de resposta. O Anexo 2 apresenta as principais variáveis derivadas e suas descrições no contexto do estudo.

- **Alinhamento temporal e preço de referência**

  Cada notícia foi associada ao preço do ouro no momento da publicação ou, quando divulgada fora do horário de negociação, ao preço de fechamento mais próximo. Para garantir consistência, o horário de publicação da notícia foi vinculada

ao *candle* de 30 minutos correspondente de abertura ou de fechamento, a depender da proximidade temporal. Assim, quando a publicação ocorreu nos primeiros 15 minutos do candle, foi considerado o preço de abertura; caso contrário, foi considerado o preço de fechamento. Esse valor foi definido como preço de referência para a medição dos movimentos futuros.

- **Cálculo de impacto nas janelas temporais**

  A partir do preço de referência, foram calculadas as variações do XAU/USD em quatro janelas temporais: 6h, 12h, 24h e 48h após a publicação. A direção do movimento foi classificada em Up (alta), Down (baixa) ou Neutral (estável). A magnitude da variação foi expressa em pips (1 pip = 0,01) e categorizada em quatro faixas (baixa, média-baixa, média-alta e alta), conforme os quartis da distribuição observada (os quartis utilizados no treino são apresentados na tabela 1 abaixo). Essas categorias de direção e magnitude constituíram os rótulos de saída utilizados no treinamento do modelo.

Impacto	**Magnitude_6h (pips)**	**Magnitude_12h (pips)**	**Magnitude_24h (pips)**	**Magnitude_48h (pips)**
Low impact	0 – 195	0 – 281	0 – 388	0 – 599
Medium-low impact	196 – 473	282 – 616	389 – 1018	600 – 1443
Medium-high impact	474 – 1051	617 – 1290	1019 – 2045	1444 – 2759
High impact	1052 – 8201	1291 – 13900	2046 – 15653	2760 – 17193

Tabela 1 - Classificação do impacto para balanceamento de classes via quartis da base de treino.

- **Explicações qualitativas do impacto**

  Para cada evento, foi atribuída uma explicação textual concisa sobre o possível impacto da notícia no preço do ouro. As explicações foram derivadas de frases extraídas das notícias originais (Reuters), reescritas por um modelo de linguagem para evidenciar o efeito no XAU/USD e revisadas manualmente. O procedimento teve como finalidade fornecer ao modelo uma base causal interpretável, com variações positivas e negativas por ativo, de modo a ampliar a diversidade e reduzir vieses no conjunto de treinamento.

Após integrar todas essas características, construiu-se o prompt em que linha da tabela gerada com o agrupamento das informações representa um caso de estudo: uma notícia particular associada a um ativo relacionado ao ouro, contendo atributos descrevendo o contexto (texto, sentimento, indicadores) e rótulos de saída que indicam como o ouro efetivamente se comportou após aquela notícia nas janelas definidas. Esse conjunto de dados serviu de base para o treinamento supervisionado do modelo de linguagem proposto.

Treinamento do Modelo LLM para Previsão de Impacto

A etapa da metodologia foi o treinamento de um modelo de linguagem de grande porte para prever, a partir dos atributos de entrada de uma notícia, qual a direção e magnitude do impacto no preço do ouro em cada janela temporal subsequente (6h, 12h, 24h, 48h). O modelo escolhido foi o Mistral-7B-Instruct v0.2, uma rede neural transformer com ~7 bilhões de parâmetros, selecionada por apresentar bom desempenho em tarefas de compreensão e geração de texto, aliada a um tamanho moderado que viabiliza o fine-tuning com recursos computacionais acessíveis. Para adaptar este modelo à tarefa específica, utilizamos a técnica de fine-tuning com *Low-Rank Adaptation (LoRA)*, que permite treinar eficientemente modelos grandes adicionando um número reduzido de parâmetros (matrizes de baixo ranque) sem precisar ajustar todos os pesos originais. Nesse caso, habilitou-se o carregamento do modelo em precisão reduzida (4-bit) para otimizar uso de memória, e aplicou-se LoRA com hiperparâmetros selecionados (por exemplo, dimensão interna r=16, alpha=32 e dropout de 5%).

O dataset descrito anteriormente foi então dividido em subconjuntos de treino, validação e teste, assegurando que notícias do teste não fossem vistas durante o treinamento. Optou-se por um particionamento estratificado temporalmente, reservando aproximadamente 20% das entradas mais recentes para teste, 10% para validação, e o restante para treinamento; de forma a avaliar o modelo em dados posteriores aos utilizados para ajuste, simulando sua aplicação em notícias futuras. Antes do treinamento, cada instância foi convertida em um prompt de entrada textual e um alvo textual esperado. O prompt de entrada consolidou as informações relevantes de uma instância de forma legível, por exemplo: uma possível estrutura de prompt seria como o apresentado na figura 2:

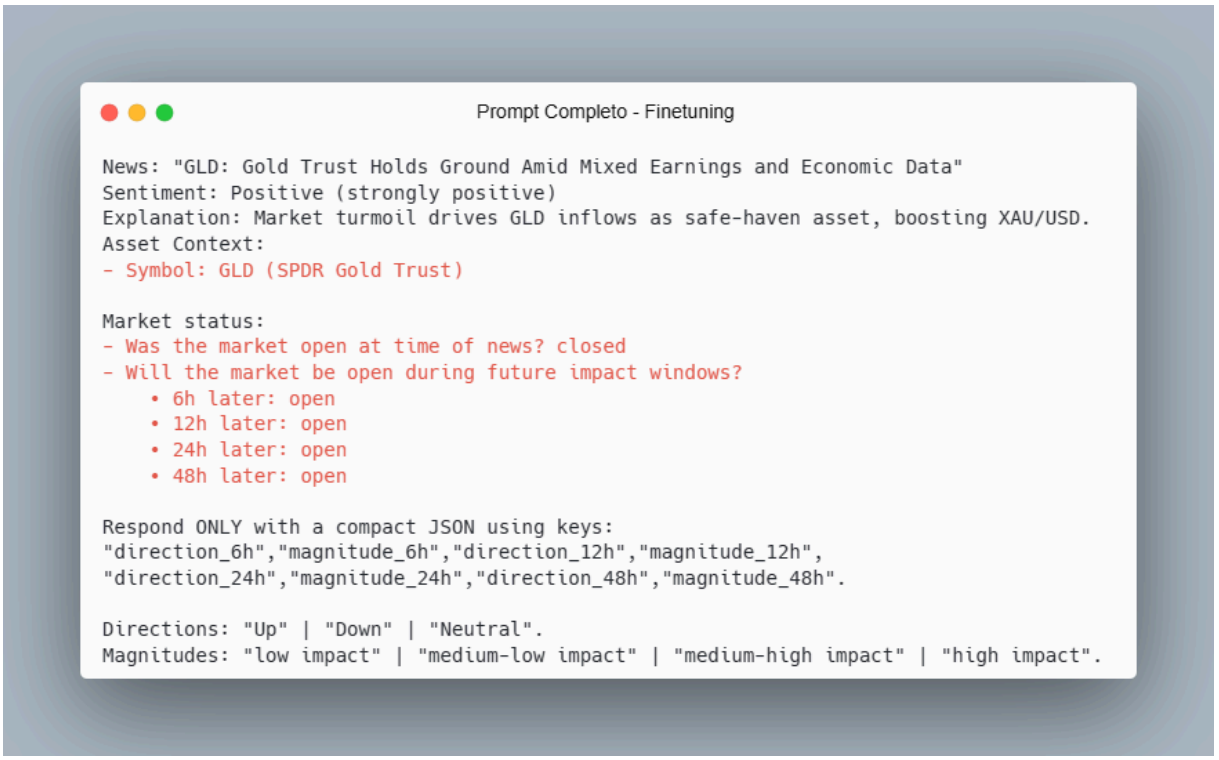

Figura 2: Estrutura de prompt de treinamento. Aqui foi removido o contexto de *user* e *assistant* obrigatórios para finetuning do Mistral e ajustado para melhor visualização.

O alvo textual correspondente a esse prompt é um JSON contendo as chaves 6h, 12h, 24h, 48h, cada qual com um par de sub-atributos: direction e magnitude (vide Figura 3). Esse formato de saída captura precisamente o que desejamos prever para cada horizonte. Treinar o LLM para gerar essa estrutura JSON, em vez de apenas classes isoladas, nos permite aproveitar sua capacidade de gerar texto estruturado e possivelmente modelar interdependências entre os horizontes (por exemplo, garantir coerência de que um impacto “alto” em 6h seja seguido de algo plausível em 12h, etc.). No treinamento supervisionado, o modelo ajusta seus pesos para produzir o JSON correto dado o prompt de entrada, minimizando a função de perda entre a sequência de tokens gerada e a sequência-alvo (tokens do JSON verdadeiro). Utilizamos *cross-entropy* mascarada apenas nos tokens de saída (ignorando tokens de entrada) já que é um cenário de *seq2seq* onde o input é copiado na entrada do modelo concatenado com a pergunta.

Figura 3: Estrutura para fine tuning da LLM do alvo supervisionado.

O processo de fine-tuning foi conduzido na infraestrutura do Amazon SageMaker, aproveitando instâncias de GPU (p. ex. ml.g5.2xlarge ou ml.g5.xlarge com GPUs NVIDIA A10G) para acelerar o treinamento. Implementamos o treinamento usando a biblioteca Hugging Face Transformers integrada com SageMaker. Hiperparâmetros importantes incluem: taxa de aprendizado inicial de 1.5*10^-4, treinamento por 3 épocas completas sobre o dataset, *batch size* efetivo de 16 exemplos (ajustado via acumulação de gradientes dada a limitação de memória por batch), além de técnicas de *warmup* (aquecimento do learning rate) e agendamento de taxa de aprendizado do tipo cosseno com decaimento. O uso do LoRA reduziu significativamente o número de pesos atualizáveis, acelerando o treinamento e prevenindo o overfitting. Monitorou-se o desempenho em validação a cada certo número de passos e aplicou-se *early stopping* implícito selecionando o modelo com melhor métrica de interesse.

Durante o processo de fine-tuning, optou-se por configurar o parâmetro eval_steps=200 na API do Hugging Face Trainer. Esse valor implica que a performance do modelo no conjunto de validação seja medida a cada 200 etapas de treinamento. Como a base de dados continha 27.722 exemplos para treino e foi utilizada uma combinação de per_device_train_batch_size=2 com gradient_accumulation_steps=8, o número efetivo de exemplos processados por atualização de gradiente foi 16. Com isso, cada época consistiu em aproximadamente 1.733 steps, resultando em cerca de 8 avaliações por época.

Essa escolha de granularidade balanceia dois fatores: a necessidade de monitoramento constante da métrica de desempenho (load_best_model_at_end depende desse valor) e o tempo adicional incorrido no processo. A escolha por eval_steps=200,

portanto, buscou maximizar a visibilidade sobre o desempenho do modelo ao longo do tempo, sem comprometer de forma significativa a viabilidade do treinamento em instâncias spot.

Durante o fine-tuning, definiu-se como métrica principal a combinação do desempenho nas predições de direção e magnitude. Em particular, o F1 macro médio dessas tarefas, conforme detalhado a seguir. Ao final do treinamento, o modelo ajustado (incluindo os deltas de LoRA) foi salvo. Uma vez treinado, este modelo LLM torna-se capaz de receber como entrada uma nova notícia (após passar pelas etapas de reescrita e análise de sentimento, e composição do prompt com demais atributos) e gerar diretamente um JSON prevendo o impacto esperado do noticiário sobre o preço do ouro em cada horizonte futuro especificado.

Resultados e Discussão

Os resultados do modelo ajustado para prever o impacto de notícias sobre o par XAU/USD são apresentados com base em um conjunto de validação fixo e, ao final do treinamento, em execução de teste hold-out. As métricas priorizadas foram F1-score macro e acurácia para as tarefas de direção (Up/Down/Neutral) e magnitude (baixa, média-baixa, média-alta, alta), além de indicadores de qualidade do JSON gerado (taxa de JSON válido e taxa de match exato do objeto). A escolha do F1 macro deve-se ao desbalanceamento entre classes, uma vez que a média por classe atribui o mesmo peso a classes minoritárias e majoritárias.

Durante o treino, observou-se evolução consistente nas métricas de validação antes da ocorrência de anomalias. Em um ponto intermediário (época ≈ 1,85), foram registrados (logs da execução): F1-macro de direção ≈ 0,90, acurácia de direção ≈ 0,85, F1-macro de magnitude ≈ 0,37, acurácia de magnitude ≈ 0,77, taxa de JSON válido = 1,00 e match exato de JSON ≈ 0,255, com eval loss ≈ 0,0418, como apresentado na tabela 2. Em estágio anterior (época ≈ 1,39) os valores eram inferiores (p.ex., F1-macro de direção ≈ 0,76 e F1-macro de magnitude ≈ 0,08), o que indica aprendizado progressivo sob a função de perda com pesos de classe.

Após esse período, foram observados episódios com métricas próximas de 1,00 (direção e magnitude) em validações subsequentes. Essa elevação súbita e não plausível foi atribuída à fuga de informação entre os conjuntos, decorrente do data augmentation que disseminou variações de uma mesma notícia (com explicações alternativas) entre treino e validação, em combinação com o uso do parâmetro load_best_model_at_end. Tal

configuração favoreceu viés de seleção de checkpoint, uma vez que avaliações repetidas no mesmo subconjunto tenderam a selecionar o modelo que melhor reproduziu, por acaso, o conjunto de validação, inflando as métricas. O fenômeno é consistente com recomendações da literatura para prevenção de data leakage (Cawley & Talbot, 2010).

Métrica	Direção	Magnitude	JSON
Acurácia (macro)	0,8525	0,7675	–
F1-score (macro)	0,9001	0,3718	–
Taxa de JSON válido	–	–	1,0000
JSON match exato	–	–	0,2550
Perda de avaliação (loss)	–	–	0,0418

Tabela 2 – Métricas de desempenho do modelo LLM no conjunto de validação para época de treinamento ≈ 1,85.

Quanto ao parâmetro de comparação com a literatura, os resultados observados antes da anomalia (F1-macro de direção ≈ 0,90; F1-macro de magnitude ≈ 0,37) situaram-se no intervalo esperado para tarefas correlatas, reconhecendo-se que a previsão de movimentos a partir de notícias tende a produzir ganhos modestos em cenários reais de mercado. Estudos prévios associaram o teor noticioso a retornos subsequentes, ainda que sem alta precisão preditiva (Tetlock, 2007; Smales, 2015). Em paralelo, benchmarks de sentimento financeiro, como o FinBERT, reportaram F1 e acurácia entre 0,70 e 0,85, a depender do conjunto de dados e da configuração (Jiang & Zeng, 2023). Em aplicações de predição de movimento com texto noticioso, acurácias típicas variaram aproximadamente entre 0,55 e 0,80, refletindo a dificuldade intrínseca da tarefa e diferenças metodológicas (Elahi & Taghvaei, 2024; Chen, 2021). No mercado de ouro, evidências sugerem sensibilidade ao conteúdo noticioso, com magnitude e persistência condicionadas ao contexto de risco.

Em síntese, os resultados intermediários (pré-anomalia) indicaram boa capacidade preditiva na tarefa de direção e desempenho moderado em magnitude, com alta conformidade de formato JSON. Entretanto, tais resultados não foram considerados válidos para teste em perspectiva de uso, uma vez que a técnica de augmentation gerou instâncias não verdadeiramente inéditas. A identificação do viés motivou ajustes no protocolo: (i) separação integral por grupo de notícia (id_new) antes de qualquer augmentation, (ii) manutenção do load_best_model_at_end somente sob validação limpa, e (iii) inferência em

base de teste inédita e não sobreposta. Com essas medidas, estimativas mais conservadoras e estáveis passaram a ser priorizadas, sendo requerido novo treinamento com os métodos descritos. Os resultados pós-ajustes encontram-se em execução e, por isso, não são apresentados neste trabalho, sendo registrados como limitação metodológica e diretriz para continuidade da pesquisa. Ademais, recomenda-se que pesquisas futuras explorem o refinamento das técnicas aqui aplicadas e a adoção de abordagens alternativas, de modo a ampliar a capacidade preditiva em análises baseadas em notícias financeiras. Nesse contexto de desenvolvimento contínuo, registra-se também que, até o presente estágio, os custos acumulados em infraestrutura AWS totalizaram US$118,15, cujo detalhamento técnico-financeiro encontra-se apresentado na seção seguinte.

Considerações Finais

A análise dos resultados revelou que parte das métricas obtidas inicialmente foi inflada por fuga de informação decorrente do uso de data augmentation em conjunto com o parâmetro load_best_model_at_end. Esse cenário levou à seleção enviesada de checkpoints, gerando desempenho artificialmente elevado e não condizente com a realidade.

Para mitigar esse problema, foram adotadas as seguintes medidas: (i) recomposição dos splits com separação por grupos de notícias (id_new) e/ou por janelas temporais, (ii) aplicação do augmentation apenas após o particionamento, prevenindo cruzamentos entre treino e validação, (iii) utilização de validação fixa para controle de early stopping e teste hold-out apenas ao final, e (iv) alinhamento estrito entre os processos de treinamento e inferência, com padronização de template de chat, tokenização, parâmetros de geração e captura do JSON produzido.

Essas modificações asseguram estimativas mais conservadoras e estáveis, aproximando o desempenho reportado de um cenário de uso real e alinhando a avaliação às recomendações da literatura sobre prevenção a vazamento de dados em aprendizado supervisionado (Cawley & Talbot, 2010; Kaufman et al., 2012).

Do ponto de vista prático, até o presente estágio do projeto foram gastos US$ 118,15 em infraestrutura AWS, com predominância de custos associados ao treinamento e à inferência de modelos via SageMaker. O treinamento originalmente estimado em cerca de 50 horas na instância ml.g5.xlarge foi migrado para a instância ml.g5.12xlarge com o objetivo de reduzir o tempo de execução. A instância ml.g5.12xlarge disponibiliza 4 GPUs NVIDIA A10G (24 GB cada, totalizando 96 GB de memória de GPU), 48 vCPUs e

aproximadamente 192 GiB de RAM, o que permitiu maior paralelismo e throughput no fine-tuning e na execução em lote.

Em síntese, o protocolo revisado mitiga os riscos de vazamento, garante maior robustez metodológica e documenta de forma transparente os custos e recursos computacionais empregados, constituindo base mais sólida para a avaliação e a comparação dos resultados em capítulos subsequentes. O protocolo revisado é, assim, aproximado de um cenário real de aplicação em mercado financeiro, no qual a robustez metodológica, a consistência de inferência e a documentação transparente de custos são considerados requisitos críticos para adoção prática.

Referências

Barsky, R. What drives gold prices? Chicago Fed Letter, n. 464, 2021. Federal Reserve Bank of Chicago. Disponível em: https://www.chicagofed.org/publications/chicago-fed-letter/2021/464

Baur, D. G.; Lucey, B. M. Is Gold a Hedge or a Safe Haven? An Analysis of Stocks, Bonds and Gold. Financial Review, v. 45, n. 2, p. 217–229, 2010. DOI:10.1111/j.1540-6288.2010.00244.x

Cawley, G. C.; Talbot, N. L. C. On Over-fitting in Model Selection and Subsequent Selection Bias in Performance Evaluation. Journal of Machine Learning Research, v. 11, p. 2079–2107, 2010. Disponível em: https://jmlr.org/papers/v11/cawley10a.pdf

Chen, Q. Stock Movement Prediction with Financial News using Contextualized Embedding from BERT. arXiv preprint, arXiv:2107.08721, 2021. Disponível em: https://arxiv.org/abs/2107.08721

Cheng, W.-H.; Chen, C.-D.; Lai, H.-P. Revisiting the roles of gold: Does gold ETF matter? The North American Journal of Economics and Finance, v. 54, 100891, 2020. Disponível em: https://ideas.repec.org/a/eee/ecofin/v54y2020ics1062940818302407.html

Chung, H. W., et al. (2022). Scaling Instruction-Finetuned Language Models. arXiv preprint arXiv:2210.11416. Disponível em: https://arxiv.org/abs/2210.11416

Darie, F.; Miron, A. (2023). Bitcoin, Gold and Crude Oil versus the US Dollar – A GARCH Volatility Analysis. Proceedings of the International Conference on Business Excellence, 17(1), 27–39. Disponível em: https://doi.org/10.2478/picbe-2023-0027

Elahi, A.; Taghvaei, F. Combining Financial Data and News Articles for Stock Price Movement Prediction Using Large Language Models. arXiv preprint, arXiv:2411.01368, 2024. Disponível em: https://arxiv.org/abs/2411.01368

Jiang, T.; Zeng, Q. Financial Sentiment Analysis using FinBERT with Application in Predicting Stock Movement. arXiv preprint, arXiv:2306.02136, 2023. Disponível em: https://arxiv.org/abs/2306.02136

Joy, M. Gold and the US dollar: Hedge or haven? Finance Research Letters, v. 8, n. 3, p. 120–131, 2011. Disponível em: https://www.sciencedirect.com/science/article/abs/pii/S154461231100002X

Kaufman, S.; Rosset, S.; Perlich, C. Leakage in Data Mining: Formulation, Detection, and Avoidance. ACM Transactions on Knowledge Discovery from Data (TKDD), v. 6, n. 4, p. 1–21, 2012. DOI: 10.1145/2382577.2382579.

Kuratomi, G., et al. (2025). A RAG-Based Institutional Assistant. arXiv preprint arXiv:2501.13880. Disponível em: https://arxiv.org/abs/2501.13880

Lewis, P., et al. (2020). Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. arXiv preprint arXiv:2005.11401. Disponível em: https://arxiv.org/abs/2005.11401

Li, Q.; Ren, X.; Huang, Z. (2023). Bridging LLMs and Quantitative Finance: Challenges and Opportunities. arXiv preprint arXiv:2306.12659.

Maghyereh, A. I.; Abdoh, H. Can News-Based Economic Sentiment Predict Bubbles in Precious Metal Markets? Journal of Behavioral and Experimental Finance, v. 33, p. 100540, 2022. Disponível em: https://doi.org/10.1016/j.jbef.2021.100540

Sharir, O.; Peleg, B.; Shoham, Y. The Cost of Training NLP Models: A Concise Overview. arXiv preprint arXiv:2004.08900, 2020. Disponível em: https://arxiv.org/abs/2004.08900

Shen, Y.; Zhang, P. K. Financial Sentiment Analysis on News and Reports Using Large Language Models and FinBERT. arXiv preprint arXiv:2410.01987, 2024. Disponível em: https://arxiv.org/abs/2410.01987

Sinha, A.; Khandait, T. (2020). Impact of News on the Commodity Market: Dataset and Results. arXiv preprint arXiv:2009.04202. Disponível em: https://arxiv.org/abs/2009.04202

Smales, L. A. Asymmetric volatility response to news sentiment in gold futures. Journal of International Financial Markets, Institutions & Money, v. 34, p. 161-172, 2015. DOI: 10.1016/j.intfin.2014.11.001. Disponível em: https://www.sciencedirect.com/science/article/abs/pii/S1042443114001267

Su, J., et al. (2023). RoFormer: Enhanced Transformer with Rotary Position Embedding. arXiv preprint arXiv:2104.09864. Disponível em: https://arxiv.org/abs/2104.09864

Tetlock, P. C. Giving Content to Investor Sentiment: The Role of Media in the Stock Market. Journal of Finance, v. 62, n. 3, p. 1139-1168, 2007. DOI: 10.1111/j.1540-6261.2007.01232.x. Disponível em: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1540-6261.2007.01232.x

Vaswani, A., et al. (2017). Attention is All You Need. Advances in Neural Information Processing Systems, 30, 5998–6008. Disponível em: https://arxiv.org/abs/1706.0376.

World Gold Council. Gold Demand Trends: Full Year 2024. Londres: WGC, 2025. Disponível em:
https://www.gold.org/goldhub/research/gold-demand-trends/gold-demand-trends-full-year-2024

World Gold Council. Gold Demand Trends: Q1 2025. Londres: WGC, 2025. Disponível em: https://www.gold.org/goldhub/research/gold-demand-trends/gold-demand-trends-q1-2025

World Gold Council. Gold Demand Trends: Q2 2025. Londres: WGC, 2025. Disponível em: https://www.gold.org/goldhub/research/gold-demand-trends/gold-demand-trends-q2-2025

Zhang, Y.; Chen, L.; Xu, J. Large Language Models for Financial Forecasting: Integrating Text and Numeric Data. arXiv preprint arXiv:2407.18103, 2024.

Zhao, M.; Liu, H.; Sun, R. Efficient Historical Context Integration for Financial News Prediction. arXiv preprint arXiv:2509.12519, 2025. Disponível em: https://arxiv.org/abs/2509.12519

.

Anexos

Símbolo	Nome	Motivo de Inclusão	Importância
GLD	SPDR Gold Trust	Principal ETF lastreado em ouro físico – reflete fluxo de compra/venda no mercado à vista	Alto
IAU	iShares Gold Trust	ETF similar ao GLD, mas com taxa menor – alternativa relevante	Alto
SGOL	abrdn Physical Gold Shares	ETF com ouro armazenado na Suíça – indica demanda global	Alto
GLDM	SPDR Gold MiniShares	Versão fracionada do GLD – acessível, mas reflete o mesmo mercado	Alto
PHYS	Sprott Physical Gold Trust	Fundo fechado com possibilidade de resgate físico – indica demanda real	Alto
BAR	GraniteShares Gold Shares	ETF com ouro físico e baixa taxa de custódia	Alto
OUNZ	VanEck Merk Gold ETF	ETF que permite entrega física de ouro aos investidores	Alto
UUP	Invesco DB USD Index Bullish	ETF que replica o dólar (DXY) – dólar forte pressiona ouro	Alto
USDU	WisdomTree USD Bullish Fund	Outra réplica do dólar com ponderação alternativa	Alto
UDN	Invesco DB USD Index Bearish	ETF que aposta contra o dólar – valorização favorece o ouro	Alto
TLT	iShares 20Y Treasury Bond	Títulos longos dos EUA – juros reais impactam o ouro	Alto
IEF	iShares 7-10Y Treasury Bond	Títulos intermediários – importante para curva de juros	Alto
TIP	iShares TIPS Bond ETF	Proteção contra inflação – reflete expectativa inflacionária	Alto
GDX	VanEck Gold Miners ETF	Desempenho de mineradoras influencia percepção de valor do ouro	Médio
GDXJ	VanEck Junior Gold Miners ETF	Mineradoras pequenas – refletem apetite por risco no setor	Médio
SLV	iShares Silver Trust	Prata é correlacionada ao ouro – movimentos extremos afetam o XAU/USD	Médio-Baixo

Anexo 1: Símbolos selecionados para extração, motivo da inclusão e importância na cotação do ouro e dólar.

Variável	Descrição
headline (original)	Manchete original da notícia financeira coletada.
generated_headline	Manchete reescrita pelo LLM, destacando implicações para o ouro e o ativo relacionado.
label (sentimento)	Sentimento identificado na manchete reescrita (Positivo, Neutro, Negativo) via FinBERT.
score (confiança)	Pontuação de confiança do modelo de sentimento (0 a 1, quanto maior, mais confiante).
sentiment_strength	Intensidade do sentimento, derivada de score (e.g. fortemente negativo, moderadamente positivo).
created_at (data/hora)	Timestamp de publicação da notícia (UTC).
symbol / symbol_name	Ativo mencionado associado ao ouro (ticker e nome, ex: GLD – SPDR Gold Trust).
why_matter	Breve descrição do porquê o ativo é relevante para o mercado de ouro.
reference_price	Preço de referência do ouro no momento da notícia (ou último preço disponível).
direction_6h	Direção do movimento do ouro 6 horas após a notícia (Up = alta, Down = baixa, Neutral = estável).
magnitude_6h	Magnitude categorizada do movimento em 6h (baixa, média-baixa, média-alta ou alta).
direction_12h / magnitude_12h	Direção e magnitude do movimento 12 horas após a notícia, com categorias análogas.
direction_24h / magnitude_24h	Direção e magnitude 24 horas após.
direction_48h / magnitude_48h	Direção e magnitude 48 horas após.
explanation	Frase explicativa qualitativa ligando o evento/ativo ao efeito no ouro (contexto fundamental).

Anexo 2 – Principais variáveis do dataset de treinamento e suas descrições, nem todas foram utilizadas na preparação do prompt.