FINE-TUNING APLICADO A UM MODELO DE LINGUAGEM DE LARGA ESCALA PARA RECONHECIMENTO DE INTENÇÃO NA ESCOLHA DE BEBIDAS

Estefany de Almeida Freitas[1] - e-mail: estefanyalmeida500@gmail.com
Daniel Neri Cardoso[2] - e-mail: nericardoso@gmail.com
Marcos Flávio Silveira Vasconcelos D'angelo[2] - e-mail: marcos.dangelo@unimontes.br
Antônio W. Vieira[2] - e-mail: antonio.vieira@unimontes.br
Maria Luisa Almeida Xavier[2] - e-mail: maria.luiza.a.xavier@gmail.com
Magnus Silva Soares[2] - e-mail: magnus.silva@unimontes.br
João Vitor Santana de Morais[2] - e-mail: jvs15.sr@gmail.com
Heveraldo R. de Oliveira[2] - e-mail: heveraldo.oliveira@unimontes.br
Julio César Porto de Carvalho[2] - e-mail: juliocesar13pc@gmail.com
Joaquim Atallah Haun de Barros[2] - e-mail: joaquimatallah@gmail.com
Cláudio Dias Campos[1] - e-mail: cdcampos@ufmg.br
Reinaldo Martínez Palhares[1] - e-mail: rpalhares@ufmg.br
[1]Universidade Federal de Minas Gerais, PPG em Engenharia Elétrica - BH - MG.
[2]Universidade Estadual de Montes Claros – Montes Claros – MG

Abstract. *Modelos de Linguagem de Larga Escala (LLMs, do inglês Large Language Models) têm demonstrado alto desempenho em tarefas que exigem processamento de linguagem natural. Para especializar a aplicação do modelo, utiliza-se o ajuste fino (do inglês fine-tuning), processo pelo qual o modelo pré-treinado é adaptado a tarefas específicas. Este artigo tem como objetivo detalhar o processo de fine-tuning empregado para direcionar a resposta do modelo à intenção do usuário na escolha entre duas bebidas: água ou café. Para tanto, consideram-se frases de entrada em linguagem coloquial que expressam calor, sede, cansaço ou sono de forma direta, bem como por meio de ambiguidades ou negações. Os resultados mostram que o modelo ajustado, treinado com o dataset desenvolvido, apresentou elevada precisão de classificação em amostras não vistas durante o treinamento, evidenciando sua capacidade de generalização.*

Keywords: *Modelos de Linguagem de Larga Escala, Ajuste fino, Reconhecimento de Intenção, Processamento de Linguagem Natural.*

1. INTRODUÇÃO

Nos últimos anos, os modelos de linguagem de larga escala (LLMs, do inglês *Large Language Models*) apresentaram desempenho notável em tarefas textuais, como programação, sumarização de textos e tradução automática. O avanço desses modelos viabilizou a interação fluida com seres humanos por meio do processamento de linguagem natural (Touvron et al., 2023).

Fundamentalmente, o funcionamento de um LLM baseia-se na previsão probabilística do próximo *token* a partir da sequência anterior. A função primária do modelo é, portanto, completar textos prevendo a continuação estatisticamente mais provável com base no *prompt* fornecido. Durante a inferência, esse processo divide-se em duas etapas: o preenchimento (*prefill*), no qual a entrada é processada em paralelo para computar o contexto inicial e emitir o primeiro *token*, e a decodificação, na qual os *tokens* subsequentes são gerados de forma autoregressiva, um a um (Zheng et al., 2024).

Como os modelos pré-treinados apenas buscam dar uma continuação plausível ao texto, eles tendem a gerar respostas genéricas. Isso limita seu desempenho em aplicações que exigem restrições de domínio, formatos rígidos de saída ou a interpretação estrita de intenções. Para contornar essa limitação e direcionar a geração sequencial do modelo para uma tarefa específica, torna-se fundamental aplicar o ajuste fino (*fine-tuning*).

Para viabilizar esse processo sem o alto custo computacional de reajustar bilhões de parâmetros, emprega-se a técnica LoRA (Hu et al., 2021), que introduz e treina apenas uma fração reduzida de parâmetros específicos para a tarefa. A técnica QLoRA (Dettmers et al., 2023) combina a estrutura de adaptadores do LoRA com a quantização em 4 bits do modelo base. Essa abordagem reduz drasticamente a utilização de memória VRAM sem comprometer a precisão do modelo, viabilizando o ajuste fino de LLMs em infraestruturas de hardware mais acessíveis.

Nesse sentido, este trabalho visa detalhar o processo de *fine-tuning* empregado para adaptar o modelo LLM à tarefa de reconhecimento de intenção do usuário na escolha entre duas bebidas: água ou café. Os resultados experimentais mostram que o modelo apresentou desempenho robusto de classificação diante de diferentes cenários , tais como pedidos diretos, ambiguidades, variações linguísticas e negação.

O restante deste artigo está organizado da seguinte forma: a Seção 2. descreve a metodologia e a configuração experimental; a Seção 3. detalha os resultados obtidos; e a Seção 4. apresenta as conclusões, limitações e direções para trabalhos futuros.

2. METODOLOGIA

A arquitetura proposta visa o reconhecimento de intenção a partir do *prompt* do usuário, empregando um modelo LLM adaptado via *fine-tuning*. A abordagem fundamenta-se nos trabalhos de Belkada et al. (2023); Dettmers et al. (2023); Touvron et al. (2023); Devlin et al. (2019); Schmid et al. (2023); Taori et al. (2023); Hugging Face (2024); von Werra et al. (2020); Mangrulkar et al. (2022) e está ilustrada na Figura 1.

Para a execução do *fine-tuning*, utilizou-se o modelo *Llama 3.2 3B* (Grattafiori et al., 2024), disponível na plataforma Hugging Face. Os experimentos foram realizados no Google Colab utilizando uma GPU NVIDIA T4 com 16 GB de memória dedicada. A configuração da técnica LoRA adotou posto $r = 16$, parâmetro que define a dimensão das matrizes de baixo posto e limita a quantidade de parâmetros treináveis; $\alpha = 32$, responsável por ponderar a intensidade da

atualização dos pesos adaptados; e *dropout* de 0,1, aplicado como regularização para prevenir o *overfitting*. O treinamento supervisionado foi conduzido por 4 épocas, com taxa de aprendizado de 2×10^{-4} e *batch* igual a 4 (Touvron et al., 2023).

Na etapa de inferência, adotou-se decodificação determinística para garantir a reprodutibilidade dos resultados. Por fim, o modelo foi avaliado segundo as métricas de acurácia, precisão, recall e F1-score.

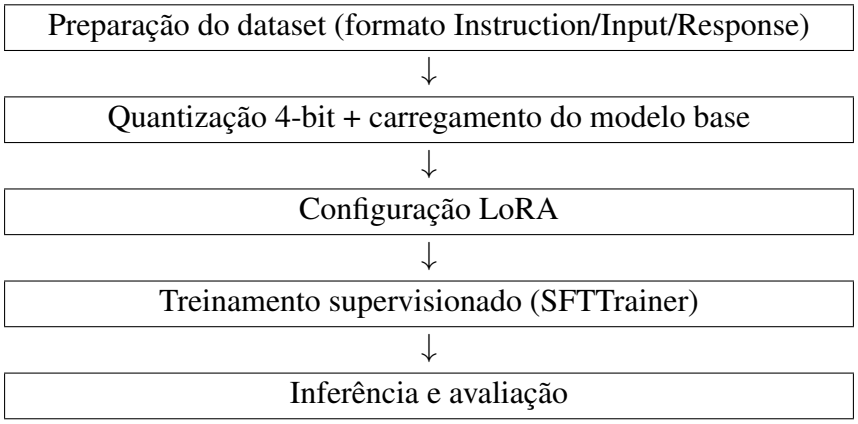

Figura 1: Fluxograma detalhado das etapas de treinamento e avaliação do LLM.

O conjunto de dados utilizado no processo *fine-tuning* é composto por 2.191 amostras no formato de instrução, contendo pares de entrada (frases que expressam intenções) e saída (respostas estruturadas em JSON). O *dataset* reúne 1177 exemplos associados ao contexto de café e 1014 ao de água. A distribuição entre as classes seguiu a variabilidade semântica, justificando a diferença quantitativa observada. A Figura 2 ilustra a estrutura em formato JSON de uma amostra, enquanto a Tabela 1 apresenta exemplos de entradas e saídas do modelo.

```
{
"instruction": "Você é um assistente que reconhece a intenção do usuário
com base no contexto. Responda apenas com um JSON contendo os campos
'contexto' e 'intencao'.",
"input": "Estou bocejando sem parar, dormi mal.",
"output": "{\"contexto\": \"inicio_conversa\", \"intencao\": \"café\"}"
}
```

Figura 2: Exemplo de amostra do conjunto de dados no formato JSON.

Tabela 1: Exemplos de pares (entrada, saída) para inferência do LLM.

Tipo	**Entrada (input)**	**Saída**
Direto	“Os olhos estão pesados, não consigo focar.”	”intencao”:”cafe”
Direto	“Calor escaldante e sede que não passa.”	”intencao”:”agua”
Negação	“Não estou com sede, prefiro um café.”	”intencao”:”cafe”
Negação	“Hoje não estou a fim de café.”	”intencao”:”agua”
Ambiguidade	“Parece que corri uma maratona dormindo.”	”intencao”:”cafe”
Ambiguidade	“Esse calor me tirou toda a disposição.”	”intencao”:”agua”

A Tabela 2 mostra a distribuição das categorias empregadas no processo de *fine-tuning*. Para a classe café, observa-se que as subcategorias foram divididas em: sono, cansaço, exaustão, falta de energia e produtividade. Já para a classe água, foram usados exemplos que expressam sede, hidratação e calor. Além dos exemplos abordados, foram acrescidos exemplos de negação para ambas as categorias. A categoria ”Outros” reúne exemplos em linguagem coloquial, variações linguísticas ou formulações curtas.

Tabela 2: Distribuição das subcategorias utilizadas na construção do conjunto de dados.

Subcategoria	**Café**	**Água**
Sono/Cansaço	412	–
Falta de energia/Foco	211	–
Pedido direto de café	94	–
Exaustão/Fadiga	60	–
Produtividade/Trabalho	38	–
Sede/Boca seca	–	293
Pedido direto de água	–	141
Hidratação/Saúde	–	141
Calor/Temperatura	–	132
Após exercício	–	50
Negação/Recusa	150	150
Outros	212	107
Total	**1177**	**1014**

No processo de inferência, o usuário insere um *prompt* em linguagem natural e o modelo retorna a classificação da intenção com base no dataset utilizado no processo de *fine-tuning* e gera a saída *café* ou *água* em formato JSON.

3. RESULTADOS

A avaliação do modelo LLM considerou dois aspectos: o desempenho na classificação da intenção entre água e café, medido pela acurácia, precisão, recall e F1-Score, e a capacidade de manter esse desempenho diante de diferentes formas de expressão linguística. Para avaliação foram utilizados 310 *prompts* não vistos anteriormente durante o treinamento. Essas sentenças abrangeram comandos diretos e indiretos de café e água, além de expressões ambíguas e negativas. A Tabela 3 apresenta a matriz de confusão obtida, e a Tabela 4 sintetiza os resultados das métricas avaliadas.

Tabela 3: Matriz de confusão do LLM.

	Previsto Água	**Previsto Café**
Real Água	154	3
Real Café	6	147

Tabela 4: Métricas de desempenho do LLM.

Métrica	**Acurácia**	**Precisão (Água)**	**Precisão (Café)**	**Recall (Água)**	**Recall (Café)**	**F1 (Água)**	**F1 (Café)**
Valor	97,1%	96,2%	98,0%	98,1%	96,1%	97,2	97,0

A matriz de confusão indica que, das 157 amostras com intenção ”água”, apenas 3 foram classificadas incorretamente, enquanto das 153 amostras com intensão ”café”, 6 foram atribuídas á classe água. O modelo atingiu acurácia global de 97,1% com precisão de 96,2% para intensão ”água” e 98,0% para intensão ”café”. Os valores para recall e f1-Score permaneceram igualmente elevados entre as classes com 98,1% e 97,2% para água e 96,1% e 97,0% para café, respectivamente. Os resultados indicam comportamento equilibrado, mostrando que o modelo não favorecimento de nenhuma das duas classes.

A Tabela 5 apresenta a acurácia detalhada para 4 categorias, considerando os principais subgrupos: expressões de calor/sede, cansaço/sono, frases ambíguas e de negação. Complementarmente, a Tabela 6 detalha o resultado para os subgrupos *ambíguas* e *negação*. Para a avaliação desses dois últimos grupos, utilizaram-se 30 e 80 frases de teste, respectivamente, todas inéditas ao modelo por não terem sido apresentadas durante o treinamento.

Tabela 5: Acurácia por categoria.

Categoria	**Acurácia**	**Acertos**
Expressões de calor/sede	98,1%	154/157
Expressões de cansaço/sono	96,1%	147/153
Frases ambíguas	90,0%	27/30
Frases de negação	98,8%	79/80

Tabela 6: Detalhamento das frases ambíguas e de negação por classe esperada.

Subgrupo	**Esperado Água**	**Esperado Café**	**Total**
Frases ambíguas	88,2% (15/17)	92,3% (12/13)	90,0% (27/30)
Frases de negação	100,0% (40/40)	97,5% (39/40)	98,8% (79/80)

A Tabela 5 apresenta o desempenho do modelo considerando os diferentes tipos de expressões. Os melhores resultados foram alcançados pelas expressões diretas com 98,1% (154/157) para expressões de calor/sede e 96,1% (147/153) para expressões de cansaço/sono. As frases ambíguas apresentaram 90,0% (27/30) e as frases de negação apresentaram 98,8% (79/80), indicando que o modelo é capaz interpretar corretamente variações linguísticas não vistas durante o treinamento.

A Tabela 6 detalha a classificação por classe para os casos de frases ambíguas e de negação. As frases ambíguas apresentaram desempenho equilibrado de 88,2% (15/17) para água e 92,3% (12/13) para café. Para as frases de negação que apresentam certa complexidade, uma vez que a bebida esperada não é mencionada de forma explícita no *prompt*, para a água a resposta foi classificada de forma correta em todos os casos (100,0%, 40/40) e para o café com valores próximos (97,5% (39/40)). Esses resultados mostram que o modelo apresenta boa capacidade de generalização para ambas as classes em contextos não apresentados no treinamento.

4. CONCLUSÕES

Os resultados experimentais evidenciaram a viabilidade da abordagem proposta, uma vez que modelo LLM adaptado apresentou acurácia global de 97,1% mantendo desempenho robusto mesmo diante de sentenças ambíguas, variações linguísticas e expressões de negação. O comportamento equilibrado entre as classes, com F1-Score de 97,2% para água e 97,0% para café, indica que o modelo não apresenta tendência em favor de nenhuma das classes. Para os subgrupos de frases ambíguas ou de negação, o modelo atingiu acurácia de 90,0% considerando frases ambíguas e 98,8% nas frases de negação. Como limitações, destacam-se a necessidade de refinamento do conjunto de dados para expressões ambíguas e a ausência de cobertura de variações regionais, o que pode impactar no desempenho em contextos reais de uso. Como direções para trabalhos futuros, propõe-se a extensão do sistema para outros contextos de classificação de intenção, a inclusão de novas categorias de bebidas e o aumento da complexidade do dataset, de forma a tornar a classificação mais detalhada e abrangente.

Agradecimentos

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – Brasil (CAPES) – Código de Financiamento 001 e FAPEMIG (grant no. OET-00243-26).

REFERÊNCIAS

Belkada, Y.; Dettmers, T.; Pagnoni, A.; Gugger, S. e Mangrulkar, S. (2023), "Making LLMs Even More Accessible with bitsandbytes, 4-bit Quantization and QLoRA", *Hugging Face Blog*.

Dettmers, T.; Pagnoni, A.; Holtzman, A. e Zettlemoyer, L. (2023), "*QLoRA: Efficient Finetuning of Quantized LLMs*", arXiv:2305.14314.

Devlin, J.; Chang, M.-W.; Lee, K. e Toutanova, K. (2019), "*BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*", arXiv:1810.04805.

Grattafiori, A.; Dubey, A.; Jauhri, A. et al. (2024), "*The Llama 3 Herd of Models*", arXiv:2407.21783.

Hu, E.J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L. e Chen, W. (2021), "*LoRA: Low-Rank Adaptation of Large Language Models*", arXiv:2106.09685.

Schmid, P.; Sanseviero, O.; Cuenca, P. e Tunstall, L. (2023), "Llama 2 is Here – Get It on Hugging Face", *Hugging Face Blog*.

Taori, R.; Gulrajani, I.; Zhang, T.; Dubois, Y.; Li, X.; Guestrin, C.; Liang, P. e Hashimoto, T.B. (2023), "*Stanford Alpaca: An Instruction-following LLaMA Model*", Stanford CRFM.

Touvron, H.; Martin, L.; Stone, K. et al. (2023), "*Llama 2: Open Foundation and Fine-Tuned Chat Models*", arXiv:2307.09288.

Zheng, L.; Li, Z.; Zhang, H. et al. (2024), "*LLM Inference Unveiled: Survey and Roofline Model Insights*", arXiv:2402.16363.

Mangrulkar, S.; Gugger, S.; Debut, L.; Belkada, Y.; Paul, S. e Bossan, B. (2022), "*PEFT: State-of-the-art Parameter-Efficient Fine-Tuning Methods*", Hugging Face. Disponível em: https://huggingface.co/docs/peft/index.

von Werra, L.; Belkada, Y.; Tunstall, L.; Beeching, E.; Thrush, T.; Lambert, N. e Huang, S. (2020), "*TRL: Transformer Reinforcement Learning*", Hugging Face. Disponível em: https://huggingface.co/docs/trl/main/en/index.

Hugging Face (2024), "SFTTrainer – TRL Documentation", *Hugging Face Documentation*. Disponível em: https://huggingface.co/docs/trl/main/en/sft_trainer.