Modelo de Linguagem Quantizados na Área da Saúde: Um Enfoque em Perguntas e Respostas com Base na Técnica DPO

Mario Pinto Freitas Filho, João Dallyson Sousa de Almeida, Anselmo C. Paiva

[1]Universidade Federal do Maranhão (UFMA)

[2]NCA - Nucleo de Computação Aplicada

Resumo. *A agilidade em diagnosticar os pacientes é um fator vital para o tratamento hábil de diversas enfermidades e muitas vezes é o parâmetro decisivo na recuperação dos pacientes. Ao levar em consideração que o tempo médio consumido por profissionais médicos em atividades de pesquisa, muitas vezes é de 4 horas de duração, e este se reduzido, de forma a não comprometer a qualidade dos resultados obtidos será valioso para o diagnóstico e tratamento principalmente em casos de maior urgência. A presente produção busca explorar a utilização de Modelos dos Large Language Models (LLMs) baseados na arquitetura Transformer para otimizar o tempo e a eficiência nas atividades de pesquisa de profissionais de saúde. Para tando, objetiva-se compreender o que são as LLM através do Transformer e suas funcionalidades além de apresentar o dataset-Medtext utilizado para treinar o modelo. Portanto, esta produção trata-se de uma pesquisa experimental na qual será aplicado o conhecimento teórico sobre LLMs e Transformers para resolver a problemática e otimização do tempo de pesquisa.*

Abstract. *The agility in diagnosing patients is a vital factor for the skillful treatment of various diseases and is often the decisive parameter in patient recovery. Considering that the average time spent by medical professionals on research activities is often 4 hours, reducing this time without compromising the quality of the results would be valuable for diagnosis and treatment, especially in more urgent cases. This work aims to explore the use of Large Language Models (LLMs) based on the Transformer architecture to optimize the time and efficiency of healthcare professionals' research activities. To this end, the objective is to understand what LLMs are through the Transformer and their functionalities, in addition to presenting the Medtext dataset used to train the model. Therefore, this work is an experimental study in which theoretical knowledge about LLMs and Transformers will be applied to address the problem and optimize research time.*

1. Introdução

O avanço acelerado das técnicas de Processamento de Linguagem Natural (PLN) nos últimos anos, resultou no surgimento de modelos de chat altamente sofisticados, como o GPT-4, LLAMA 2, e Falcon. Esses modelos exibem uma notável capacidade de compreender e gerar respostas semelhantes às humanas em diversos domínios, tornando-os cada vez mais populares em aplicações como suporte ao cliente, assistentes virtuais, moderação de mídia social, entretenimento e pesquisa.

No entanto, apesar do seu potencial, esses modelos são frequentemente acessíveis apenas por meio de APIs restritas, o que limita novas pesquisas e avanços na área de PLN. Além disso, eles são extremamente pesados e exigem grande poder computacional para treinamento e inferência.

Com isso em mente, o presente artigo busca apresentar abordagens para reduzir a demanda computacional desses modelos extensos, conhecidos como LLMs [Chen et al. 2023]. Serão exploradas técnicas como GPTQ [Frantar et al. 2023] para quantização de modelos e DPO (*Direct Preference Optimization*) [Rafailov et al. 2023] para otimização de pesos, as quais serão discutidas nas seções subsequentes.

Um estudo recente com profissionais de saúde revelou que eles gastam, em média, 60 minutos para formular uma estrategia de busca e durante suas pesquisas e dedicam cerca de 3 minutos para avaliar a relevância de cada documento, totalizando aproximadamente 4 horas de pesquisa [Russell-Rose T 2017].Uma solução viável para enfrentar esses desafios é a implementação de sistemas de perguntas e respostas dedicados .Esses modelos são capazes de compreender perguntas em linguagem natural e fornecer respostas baseadas em uma base de dados validada por especialistas.

A presente produção é composta pela fundamentação teórica em que são relatadas as técnicas utilizadas para a criação do modelo. A seguir são descritos o dataset utilizados, seu pre-processamento e experimentos realizados e por fim as considerações finais em que é sintetizado o trajeto percorrido até os resultados.

2. Fundamentação teórica

2.1. Arquitetura e Funcionamento de LLMs

Antes da arquitetura *Transformer*, modelos como LSTM (*Long Short-Term Memory*) e GRU(*Gated Recurrent Unit*) mitigavam, mas não resolviam completamente, a limitação dos RNNs (*Recurrent neural networks*) em lidar com dependências de longo alcance em sequências. O *Transformer*, apresentado em [Vaswani et al. 2023], introduziu uma abordagem inovadora com o mecanismo de atenção e o esquema codificador-decodificador, permitindo que o codificador transforme a sequência de entrada em uma representação vetorizada compreensível pela máquina.

Uma arquitetura *Transformer* em PLN, se configura por uma sequência de entradas passadas por uma camada de incorporação e codificação posicional antes de ser processada pelo codificador, que captura a semelhança entre as palavras e suas posições. O decodificador, então, usa esses vetores para produzir a saída de forma auto-regressiva, onde cada *tokens* de saída torna-se a entrada para o próximo passo.

O termo "auto-regressivo" refere-se ao processo de gerar saídas sequenciais, permitindo ao modelo criar frases de saída de comprimentos variáveis, adaptando-se a diferentes contextos e requisitos.

Llama-2 [Touvron et al. 2023] é uma coleção de quatro modelos baseados na arquitetura *Transformer*, variando em parâmetros: 7B, 13B, 34B e 70B. Todos compartilham a mesma função de ativação e método de normalização. Neste trabalho, foi utilizado o modelo LLAMA-2 7B.

O mesmo se diferencia por um novo método de ajuste fino chamado Ghost Atten-

tion (GAtt), projetado para que o modelo mantenha consistentemente o papel atribuído pelo usuário, como por exemplo ”Cardiologista”. O GAtt adiciona sinteticamente a instrução ”agir como”a todas as mensagens do usuário, permitindo que o modelo mantenha o contexto sem precisar da concatenação explícita durante o ajuste fino. Essa técnica melhora o controle do diálogo em múltiplos turnos, permitindo melhor adaptação às instruções dos usuários.

2.2. Quantização do modelo (GPTQ)

devido às características do modelo LLAMA 2 7B, com 7 bilhões de parâmetros, sua execução em GPUs convencionais ou no Google Colab não foi viável por causa da alta demanda de recursos. Para mitigar essa limitação, foram exploradas técnicas de quantização para reduzir o consumo de VRAM (*Video Random Access Memory*) sem comprometer o desempenho do modelo. Dois métodos principais de quantização são mencionados: a Quantização Pós-Treinamento (PTQ), que quantiza um modelo já pré-treinado, e o Treinamento com Reconhecimento de Quantização (QAT), que realiza a quantização antes ou durante o treinamento. O GPTQ (*Generative Pretrained Transformers Quantization*), uma técnica PTQ (*Post-training quantization*), é ideal para modelos muito grandes, onde o treinamento completo seria muito custoso.

O GPTQ utiliza um esquema misto de quantização int4/fp16 (*integer* de 4 bits e fp16 *float* de 16 bits), onde os pesos são quantizados como int4 e as ativações permanecem em float16. Durante a inferência, os pesos são desquantizados instantaneamente e o cálculo é feito em float16. Esse esquema oferece dois benefícios principais: economia de memória de até 4 vezes devido à quantização int4, e potencial aceleração da inferência e do treinamento, graças à menor largura de bits utilizada para os pesos, o que reduz o tempo de comunicação de dados [Frantar et al. 2023].

3. Materiais e Métodos

3.1. Dataset

MedText [Melamud and Shivade 2019] é um conjunto de dados para diagnósticos e tratamentos médicos, contendo 1.412 perguntas e respostas baseadas em casos de pacientes, vale ressaltar que este encontra-se na língua inglesa e aborda as 100 doenças e 30 lesões mais comuns nos hospitais. Cada condição possui perguntas e respostas variando entre leve, complicada e grave .

Ele foi desenvolvido a partir do MIMIC III [Johnson et al. 2018], um banco de dados que contém informações de prontuários médicos, incluindo diagnósticos, causas e tratamentos. A partir desses dados, foi criada uma estrutura artificial de perguntas e respostas (*question answering*) para o MedText, conforme descrito em [Melamud and Shivade 2019].

3.2. Pre-processamento do dataset e DPO

O pré-processamento do dataset é muito importante para o funcionamento do DPO (*Direct Preference Optimization*), que é uma alternativa ao aprendizado por reforço com feedback humano (RLHF). O DPO melhora o alinhamento da linguagem com as preferências humanas sem a necessidade de um modelo de recompensa. Ele utiliza dados compostos por triplas (*prompt*, resposta escolhida, resposta rejeitada), também usados no RLHF. Durante

	perplexidade
[Melamud and Shivade 2019]	12.5
Este artigo	1.98

Tabela 1. Metricas

o ajuste fino, o modelo de linguagem é duplicado, criando um modelo treinável (*policy model*) e outro congelado (*reference model*), ambos responsáveis por avaliar as respostas com base nas probabilidades dos *tokens*.

O DPO possui suas próprias métricas de avaliação, como a diferença de probabilidades logarítmicas entre as respostas escolhidas e rejeitadas, e a frequência de casos em que as respostas escolhidas superam as rejeitadas. Devido à estrutura de *prompt*, resposta escolhida e rejeitada, foi necessário adaptar o MedText, utilizando a tecla (espaço) a fim de preencher campo rejeitado sem adicionar quaisquer informação que possa interferir no processo. Zephyr [Tunstall et al. 2023], um *chatbot* que compete com grandes LLMs como mostrado em [Li et al. 2023] e [Zheng et al. 2023], adotou uma abordagem semelhante, mas a presença da teclada de (espaço) pode ser um risco visto que o modelo pode gerar respostas equivocadas.

3.3. Experimentos

A métrica usada na avaliação do modelo foi a perplexidade com isso em mente realizou-se a validação do modelo desenvolvido, comparando-o com o artigo original de onde o dataset foi extraído [Melamud and Shivade 2019]. Utilizando a técnica DPO, o modelo foi treinado por 10 épocas com uma taxa de aprendizado de 5e-07 e batch size de 2, permitindo uma análise detalhada das amostras. O otimizador AdamW foi escolhido para otimizar a eficiência na convergência. Os resultados visto na tabela 1 mostram que o modelo proposto teve uma perplexidade de **1.98**, Consideravelmente superior aos resultados do artigo do qual o conjunto de dados foi extraído [Melamud and Shivade 2019] **12.5**, indicando maior precisão na previsão de palavras e na compreensão das estruturas linguísticas.

4. Conclusão e Agradecimentos

O texto aborda estratégias para reduzir o custo computacional de grandes modelos de linguagem (LLMs), que frequentemente precisam ser executados em APIs externas. Uma das estratégias mencionadas é o GPTQ, que quantiza o modelo, reduzindo o uso de memória (VRAM), o tempo de treinamento e de inferência. O DPO, por sua vez, ao usar decodificadores causais, permite calcular recompensas em um único passo, otimizando o processo de treinamento.

Os autores agradecem Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES),Finance Code 001, Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), e a Fundação de Amparo à Pesquisa Desenvolvimento Científico e Tecnológico do Maranhão (FAPEMA), Empresa Brasileira de Serviços Hospitalares (Ebserh) Brasil (Grant number 409593/2021-4) pelo financiamento.

Referências

Chen, Z., Mao, H., Li, H., Jin, W., Wen, H., Wei, X., Wang, S., Yin, D., Fan, W., Liu, H., and Tang, J. (2023). Exploring the potential of large language models (llms) in learning on graphs.

Frantar, E., Ashkboos, S., Hoefler, T., and Alistarh, D. (2023). Gptq: Accurate post-training quantization for generative pre-trained transformers.

Johnson, A. E. W., Stone, D. J., Celi, L. A., and Pollard, T. J. (2018). The mimic code repository: enabling reproducibility in critical care research. *Journal of the American Medical Informatics Association*, 25(1):32–39.

Li, X., Zhang, T., Dubois, Y., Taori, R., Gulrajani, I., Guestrin, C., Liang, P., and Hashimoto, T. B. (2023). Length-controlled alpacaeval: A simple way to debias automatic evaluators. `https://github.com/tatsu-lab/alpaca_eval`.

Melamud, O. and Shivade, C. (2019). Towards automatic generation of shareable synthetic clinical notes using neural language models. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop at NAACL*, pages 35–45.

Rafailov, R., Sharma, A., Mitchell, E., Ermon, S., Manning, C. D., and Finn, C. (2023). Direct preference optimization: Your language model is secretly a reward model.

Russell-Rose T, C. J. (2017). Expert search strategies: The information retrieval practices of healthcare information professionals. *JMIR Med Inform 2017;5(4):e33*.

Touvron, H., Martin, L., and Al, E. (2023). Llama 2: Open foundation and fine-tuned chat models.

Tunstall, L., Beeching, E., Lambert, N., Rajani, N., Rasul, K., Belkada, Y., Huang, S., von Werra, L., Fourrier, C., Habib, N., Sarrazin, N., Sanseviero, O., Rush, A. M., and Wolf, T. (2023). Zephyr: Direct distillation of lm alignment.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2023). Attention is all you need.

Zheng, L., Chiang, W.-L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E. P., Zhang, H., Gonzalez, J. E., and Stoica, I. (2023). Judging llm-as-a-judge with mt-bench and chatbot arena.