IV Workshop de Matemática, Estatística e Computação Aplicadas à Indústria

Desempenho de Small Language Models na Sumarização de Diálogos Médico-Paciente no Formato SOAP

Aline Elí Gassenn[1], José F. Rodrigues-Jr[2]
ICMC-USP

1 Introdução

Modelos de linguagem têm sido empregados em diferentes tarefas de processamento de linguagem natural, incluindo classificação, sumarização e geração de texto. No contexto clínico, essas aplicações podem auxiliar na elaboração de registros estruturados e no apoio à documentação médica. Entretanto, o uso dessas ferramentas envolve restrições associadas à privacidade e à proteção de dados sensíveis.

Nesse cenário, modelos de menor porte (*Small Language Models* - SLMs) constituem uma alternativa aos *Large Language Models* (LLMs) em cenários que exigem execução local e controle sobre o processamento das informações. Por apresentarem menor número de parâmetros, esses modelos demandam menos recursos computacionais e podem ser empregados em sistemas embarcados ou ambientes corporativos com restrições de infraestrutura. A adoção de SLMs tem sido explorada em diferentes áreas, como saúde, manufatura e serviços, em tarefas que envolvem compreensão e geração de linguagem natural sob limitações de custo e privacidade.

Este trabalho avalia o desempenho dos modelos Qwen2.5-1.5B-Instruct e Phi-3-Mini-4K-Instruct na geração de resumos clínicos estruturados segundo o formato SOAP (*Subjective, Objective, Assessment, Plan*). Utiliza-se o conjunto de dados *Synthetic Medical Dialogues and SOAP Summaries*, composto por diálogos sintéticos entre paciente e profissional de saúde e seus respectivos resumos. As inferências são realizadas sobre o conjunto de validação, e o desempenho é mensurado por meio das métricas ROUGE-L, BLEU e BERTScore, bem como pelos indicadores de latência e uso de memória.

[1] aline.gassenn@usp.br
[2] junio@icmc.usp.br

2 Materiais e Métodos

2.1 Conjunto de Dados

O estudo utilizou o conjunto *Synthetic Medical Dialogues and SOAP Summaries* [4], disponível no repositório Hugging Face. O conjunto contém 10.000 pares de diálogos e resumos no formato SOAP (*Subjective, Objective, Assessment, Plan*), gerados a partir do modelo GPT-4 aplicado ao *NoteChat* [2], derivado de relatos clínicos do PubMed Central (PMC).

O corpus é dividido em 9.250 amostras de treinamento, 500 de validação e 250 de teste. Neste trabalho, foram utilizadas exclusivamente as amostras do conjunto de validação, compostas por diálogos sintéticos entre paciente e profissional de saúde e seus respectivos resumos em formato SOAP. Os dados não incluem informações provenientes de interações clínicas reais.

2.2 Modelos de Linguagem

Foram avaliados dois modelos de pequeno porte (*Small Language Models* - SLMs): Qwen2.5-1.5B-Instruct e Phi-3-Mini-4K-Instruct. Ambos foram utilizados exclusivamente em modo de inferência, sem ajuste de parâmetros. Os SLMs seguem a arquitetura *Transformer* do tipo decodificadora e destinam-se à execução em ambientes com restrições de recursos computacionais [5].

O Qwen2.5-1.5B-Instruct, desenvolvido pela equipe Qwen (Alibaba Group), possui 1,5 bilhão de parâmetros e foi pré-treinado em 18 trilhões de tokens. O modelo utiliza atenção *Group-Query Attention* (GQA) e normalização *RMSNorm*, com suporte a janelas de até 8 mil tokens. O ajuste supervisionado (*instruction tuning*) foi conduzido sobre aproximadamente 1 milhão de exemplos, abrangendo tarefas gerais de linguagem [1].

O Phi-3-Mini-4K-Instruct, desenvolvido pela Microsoft Research, contém 3,8 bilhões de parâmetros e foi treinado em 3,3 trilhões de tokens sob o regime *data-optimal*, que combina dados públicos filtrados e dados sintéticos. Após o pré-treinamento, o modelo passou por etapas de *Supervised Fine-Tuning* (SFT) e *Direct Preference Optimization* (DPO), com janelas de contexto de até 4 mil tokens [3].

A seleção dos modelos teve como objetivo contrastar duas abordagens representativas no desenvolvimento de SLMs. O Qwen2.5 foi incluído por empregar pré-treinamento em larga escala com diversidade de domínios, refletindo estratégias baseadas em volume de dados. O Phi-3, por sua vez, adota um processo de treinamento centrado na curadoria e na otimização de dados sintéticos. A comparação entre esses modelos permite avaliar o impacto de diferentes regimes de treinamento sobre a geração de resumos clínicos estruturados no formato SOAP.

2.3 Procedimento Experimental

Os experimentos foram realizados sobre o conjunto de validação do *Synthetic Medical Dialogues and SOAP Summaries*, com o objetivo de avaliar a capacidade dos modelos em gerar resumos clínicos estruturados no formato SOAP.

O processo envolveu três etapas principais: preparação dos dados, inferência e avaliação. Na primeira etapa, o conjunto de validação foi carregado em formato JSONL, com verificação de integridade e remoção de amostras inválidas. Na segunda, os modelos Qwen2.5-1.5B-Instruct e Phi-3-Mini-4K-Instruct foram executados em GPU por meio da biblioteca *Transformers*, utilizando

Tabela 1: Desempenho médio dos modelos no conjunto de validação.

Métrica	Qwen2.5-1.5B-Instruct	Phi-3-Mini-4K-Instruct
Parâmetros (B)	1.5	3.8
VRAM (GB)	4.5	5.0
Latência (ms/token)	6.47	10.17
Métricas Globais		
ROUGE-L	0.2225	0.2305
BLEU	9.88	11.13
BERTScore	0.8292	0.8292
S (Subjetivo)		
ROUGE-L	0.3490	0.4588
BLEU	16.31	21.54
BERTScore	0.8833	0.9144
O (Objetivo)		
ROUGE-L	0.3079	0.4731
BLEU	11.59	22.07
BERTScore	0.8642	0.9057
A (Avaliação)		
ROUGE-L	0.2344	0.3016
BLEU	5.94	8.36
BERTScore	0.8603	0.8806
P (Plano)		
ROUGE-L	0.1839	0.2199
BLEU	4.39	6.12
BERTScore	0.8543	0.8601

prompts padronizados para garantir consistência entre inferências. Cada diálogo foi processado individualmente, e o tempo médio de geração por token foi registrado para cálculo da latência.

Na etapa de avaliação, o desempenho foi mensurado por meio das métricas ROUGE-L, BLEU e BERTScore, aplicadas em dois níveis: global, considerando o resumo completo, e seccional, avaliando individualmente as partes S, O, A e P. Foram também registradas medidas de completude da estrutura SOAP, latência média e uso estimado de memória.

Todos os resultados foram exportados em formato JSON para posterior agregação e análise comparativa, servindo de base para a discussão apresentada na próxima seção.

3 Resultados e Discussão

Os experimentos avaliaram o desempenho dos modelos na geração de resumos clínicos estruturados no formato SOAP. A Tabela 1 apresenta as médias obtidas para cada métrica, considerando as avaliações globais e por seção.

O modelo *Phi-3-Mini-4K-Instruct* apresentou desempenho superior nas métricas de similaridade global, evidenciado pelos maiores valores de ROUGE-L e BLEU. Esses resultados indicam maior aderência lexical e estrutural em relação às referências. O BERTScore, voltado à similaridade semântica, manteve-se equivalente entre os modelos, sugerindo que ambos preservam coerência conceitual com o texto de referência. O *Qwen2.5-1.5B-Instruct*, embora apresente re-

sultados ligeiramente inferiores nas métricas de sobreposição, demonstrou menor latência e menor uso de memória, refletindo maior eficiência computacional.

A análise por seção mostra que as partes *Subjetiva (S)* e *Objetiva (O)* alcançaram as maiores pontuações em todas as métricas, o que indica melhor desempenho na reprodução de descrições e achados observacionais. As seções *Avaliação (A)* e *Plano (P)* apresentaram valores menores de ROUGE-L e BLEU, possivelmente em razão da natureza inferencial e prescritiva desses segmentos, que exigem maior capacidade de raciocínio clínico e síntese interpretativa.

4 Conclusão

Os resultados mostraram que ambos os modelos geraram resumos coerentes no formato SOAP, com o *Phi-3* apresentando melhor similaridade textual e o *Qwen2.5* maior eficiência computacional. De modo geral, ambos os modelos apresentaram desempenho consistente com o esperado para modelos de pequeno porte. A diferença observada sugere que abordagens baseadas em curadoria e otimização de dados podem favorecer a fidelidade lexical, enquanto arquiteturas otimizadas em escala reduzida tendem a apresentar vantagens operacionais em contextos de restrição de recursos.

No trabalho em desenvolvimento, o desempenho dos SLMs pode ser aprimorado por técnicas como *LoRA* (adaptação eficiente de parâmetros), quantização (redução de precisão para menor uso de memória), *pruning* (remoção de conexões redundantes) e distilação de conhecimento (transferência de competência de LLMs para modelos menores). Essas abordagens permitem melhorar a qualidade e a viabilidade de execução local dos modelos clínicos.

Referências

[1] A. Yang, B. Yang, B. Zhang, B. Hui, B. Zheng, B. Yu et al. Qwen2.5 Technical Report, *arXiv preprint arXiv:2412.15115 [cs.CL]*, 2024. DOI: 10.48550/arXiv.2412.15115.

[2] J. Wang, Z. Yao, Z. Yang, H. Zhou, R. Li, X. Wang, Y. Xu et al. NoteChat: A Dataset of Synthetic Patient–Physician Conversations Conditioned on Clinical Notes, *Findings of the Association for Computational Linguistics: ACL*, Bangkok, Tailândia, 2024. DOI: 10.18653/v1/2024.findings-acl.901.

[3] M. Abdin, J. Aneja, H. Awadalla, A. Awadallah, A. A. Awan, N. Bach et al. Phi-3 Technical Report: A Highly Capable Language Model Locally on Your Phone, *arXiv preprint arXiv:2404.14219 [cs.CL]*, 2024. DOI: 10.48550/arXiv.2404.14219.

[4] OMI Health. *Synthetic Medical Dialogues and SOAP Summaries Dataset*. Disponível em: `https://huggingface.co/datasets/omi-health/medical-dialogue-to-soap-summary`.

[5] Z. Lu, X. Li, D. Cai, R. Yi, F. Liu, X. Zhang et al. Small Language Models: Survey, Measurements, and Insights, *arXiv preprint arXiv:2409.15790 [cs.CL]*, 2025. DOI: 10.48550/arXiv.2409.15790.