

Avaliação Automatizada em OSCE: Integração de STT e LLMs para Feedback Clínico Escalável

Carlos Eduardo Borba Fernandes¹, Rafael do Nascimento Silva¹, Marcio de Souza Dias²

¹Instituto de Biotecnologia (IBIOTEC) – Universidade Federal de Catalão (UFCAT)

²Instituto de Matemática e Tecnologia (IMTec) – Universidade Federal de Catalão (UFCAT)

carlos10borba@gmail.com, rafaelnascimento1502@gmail.com

marciodias@ufcat.edu.br

Abstract. *The Objective Structured Clinical Examination (OSCE) is the gold standard for clinical evaluation but faces scalability limitations due to the high operational costs of human preceptorship. This paper presents an automated evaluation system that integrates Speech-to-Text (STT) pipelines and Large Language Models (LLMs) to transcribe, diarize, and heuristically interpret simulated clinical consultations. The system mitigates acoustic noise, separates dialogues, and applies advanced prompt engineering (Chain-of-Thought, Few-shot) to map student performance into structured JSON schemas. By automating the identification of semiotic entities (clinical signs and symptoms) and scoring across clinical categories, the solution aims to reduce manual grading subjectivity, enabling immediate, standardized, and scalable pedagogical feedback.*

Resumo. *O Exame Clínico Objetivo Estruturado (OSCE) é o padrão-ouro na avaliação clínica, mas enfrenta limitações de escalabilidade devido aos altos custos operacionais da preceptoria humana. Este trabalho apresenta um sistema automatizado que integra pipelines de Speech-to-Text (STT) e Modelos de Linguagem de Larga Escala (LLMs) para transcrever, diarizar e interpretar heurísticamente consultas simuladas. O sistema mitiga ruídos acústicos, separa diálogos e aplica engenharia de prompt avançada (Chain-of-Thought, Few-shot) para mapear o desempenho do discente em esquemas JSON estruturados. Ao automatizar a identificação de entidades semióticas (sinais clínicos e sintomas) e a pontuação clínica, a solução visa reduzir a subjetividade da correção manual, viabilizando um feedback pedagógico imediato, padronizado e escalável.*

1. Introdução e Justificativa

O Exame Clínico Objetivo Estruturado (OSCE) consolidou-se como o padrão-ouro para a avaliação de competências clínicas e habilidades de comunicação [Harden et al., 1975]. A simulação clínica estruturada, acompanhada de feedback rigoroso e imediato, é altamente eficaz para o aprendizado médico [Issenberg et al., 2005]. No entanto, a aplicação do OSCE em larga escala nas Instituições de Ensino Superior (IES) esbarra na dependência de um corpo docente focado na observação direta e correção em tempo real [Ziv et al., 2000]. Tentativas anteriores de automação em simulações médicas focaram em cenários baseados em computador ou manequins, mas a avaliação baseada em diálogos naturais de consultas simuladas ainda carece de abordagens escaláveis [Masters, 2023].

A adoção de sistemas automatizados surge como alternativa para escalar a correção pedagógica, alinhando-se à convergência entre inteligência artificial e medicina de alta performance [Topol, 2019]. Estudos demonstram que Modelos de Linguagem de Larga Escala (LLMs) avançados são capazes de codificar conhecimento clínico complexo e auxiliar em inferências diagnósticas [Singhal et al., 2023]. Contudo, a eficácia dessa inferência depende intrinsecamente da fidelidade da entrada de texto.

O áudio médico de um cenário OSCE apresenta desafios acústicos únicos, como ruído ambiental, sobreposição de vozes (cross-talk) [Park et al., 2022], e a extrema necessidade de acurácia na transcrição de epônimos e terminologias de entidades semióticas — aqui definidas como o conjunto de sinais clínicos, sintomas, achados semiológicos e condutas expressos no diálogo —, um desafio acentuado no processamento do português brasileiro [Gris et al., 2022]. Modelos genéricos de *Speech-to-Text* (STT) frequentemente falham ao processar esse vocabulário [Radford et al., 2023]. Falhas na transcrição podem degradar a entrada do LLM, aumentando o risco de interpretações incorretas ou de geração de inferências não sustentadas pelo texto original (alucinações), um risco crítico já documentado em LLMs aplicados à saúde [Lee et al., 2023].

Para mitigar a variabilidade interpretativa, este projeto propõe o uso de técnicas avançadas de Engenharia de Prompt, como *Chain-of-Thought* [Wei et al., 2022] e *Few-shot Prompting* [Brown et al., 2020].

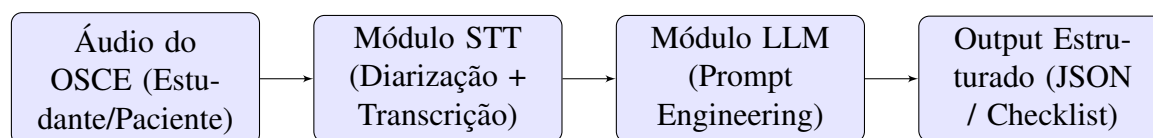
1.1. Objetivo do Trabalho

Este trabalho tem como objetivo desenvolver e avaliar preliminarmente um pipeline otimizado que integra processamento de áudio e interpretação automatizada baseada em LLMs, atuando como um sistema de inferência clínica focado na estruturação de checklists de desempenho para exames do tipo OSCE.

2. Metodologia

A pesquisa caracteriza-se pelo desenvolvimento experimental de um pipeline de Processamento de Linguagem Natural (PLN) dividido em dois módulos integrados: processamento de áudio e inferência clínica, conforme ilustrado na Figura 1.

Figura 1. Fluxo da Arquitetura do Pipeline Proposto (STT + LLM).



2.1. Módulo de Transcrição e Diarização (STT)

A validação e o benchmarking do módulo STT utilizam o dataset público Simulated Patient-Physician Medical Interviews [Kaggle]. A base compreende um volume total de 1,05 GB composto por 272 gravações de áudio de consultas médicas simuladas no formato OSCE com foco em casos respiratórios, acompanhadas de 272 transcrições textuais limpas (Clean Transcripts) e corrigidas manualmente, totalizando 544 arquivos de dados. Essas transcrições de referência são adotadas como o padrão-ouro para a medição

das métricas de erro de reconhecimento de fala (Word Error Rate - WER) e acurácia na captura de termos médicos. O módulo foca em algoritmos de diarização para separar cronologicamente as interações entre estudante e paciente simulado, utilizando métricas de *Word Error Rate* (WER) e *Diarization Error Rate* (DER) específicas para o vocabulário médico.

2.2. Módulo de Inferência e Interpretação Clínica (LLM)

As transcrições diarizadas são processadas por LLMs de código aberto, nomeadamente Mistral e Qwen [Jiang et al., 2023, Bai et al., 2023]. O pipeline utiliza *Prompt Engineering* para extrair as entidades clínicas (sinais, sintomas, hipóteses e condutas). As rubricas de avaliação são convertidas em esquemas de dados tipados (utilizando *Structured Outputs* via Pydantic/JSON Schema), forçando a IA a mapear sistematicamente a adesão do aluno aos protocolos clínicos nas categorias de anamnese, exame físico e conduta médica, eliminando falhas de sintaxe na resposta.

3. Resultados Preliminares e Limitações

O projeto encontra-se na fase de testes do Produto Mínimo Viável (MVP). As observações do módulo STT indicam que a precisão em epônimos e sinais clínicos requer a execução de modelos densos, como o Whisper Large V3 Turbo, o que evidencia a necessidade de infraestrutura baseada em GPUs dedicadas para evitar gargalos de memória (OOM).

3.1. Impacto da Engenharia de Prompt e Chain-of-Thought

A adoção da estratégia de *Chain-of-Thought* (CoT) combinada a diretrizes contextuais de extração demonstrou impacto direto na mitigação de falsos negativos e na redução de inferências não fundamentadas (*alucinações*). Conforme sintetizado na Tabela 1, em testes realizados sem o encadeamento de justificativa (prompt direto), o modelo tendia a atribuir valor *false* a critérios efetivamente realizados pelo discente, motivado por falhas de correspondência léxica exata (como não reconhecer que a pergunta “caminha pra algum lugar?” refere-se à *irradiação da dor*).

Ao impor a extração do trecho exato da transcrição no campo **justificativa_cot** antes da classificação booleana, o modelo passou a mapear adequadamente os sinônimos e a evidenciar o raciocínio clínico adotado durante a avaliação.

Tabela 1. Comparativo de extração no checklist OSCE: Prompt Direto vs. Prompt com CoT.

Categoria / Item	Sem CoT	Com CoT	Trecho Extraído (justificativa_cot)
Identificação (Idade)	<i>False</i>	True	“João, quantos anos você tem, João?”
Queixa Principal (Tempo)	<i>False</i>	True	“E quando que começou essa dor? [...] começou ontem à noite”
HMA (Irradiação da Dor)	<i>False</i>	True	“Irradia pra algum lugar? Ela caminha pra algum lugar?”
Antecedentes (Doenças)	<i>False</i>	True	“Tem alguma doença crônica? Não.”

Como o trabalho encontra-se em estágio inicial de desenvolvimento de software, os resultados atuais limitam-se à prova de conceito funcional da integração entre os

módulos. As próximas etapas metodológicas consistem na coleta de dados quantitativos e na validação estatística do sistema. A confiabilidade do motor de IA será aferida por meio do cálculo da Matriz de Confusão para cada item do checklist (quantificando Verdadeiros Positivos, Verdadeiros Negativos, Falsos Positivos e Falsos Negativos), permitindo mensurar formalmente a Precisão, a Revogação (*Recall*) e o *F1-score* em comparação à correção de professores especialistas (Padrão-Ouro).

3.2. Limitações do Sistema

No estado atual, o sistema apresenta limitações importantes: (i) dependência de infraestrutura computacional de alto custo (GPUs locais ou APIs pagas) para a execução do Whisper Large e LLMs de alta densidade; (ii) sensibilidade a sotaques regionais que degradam o WER (*Word Error Rate*) no português brasileiro; e (iii) a ausência, até o momento, de testes com áudios de salas reais de OSCE, utilizando atualmente apenas dados sintéticos e benchmarks públicos de entrevistas médicas.

4. Considerações Finais e Trabalhos Futuros

A integração de pipelines de STT acusticamente ajustados com LLMs instruídos por gabaritos estruturados demonstra viabilidade técnica para otimizar os processos de avaliação na área de educação médica.

Como metas de pesquisa imediatas de curto prazo, o projeto focará no estabelecimento de uma linha de base quantitativa (avaliando o impacto do *Chain-of-Thought* em relação a prompts diretos) e na medição do WER/DER no dataset de entrevistas simuladas. Em longo prazo, como perspectiva de aplicação tecnológica, visa-se a entrega de uma API e um SDK documentados, permitindo a futura integração deste motor de inferência a Sistemas de Gestão de Aprendizagem (LMS) e EdTechs.

Agradecimentos

Durante a redação deste artigo, modelos generativos de IA foram utilizados exclusivamente como ferramentas de apoio à estruturação e revisão gramatical do texto. Os autores assumem total responsabilidade pelo conteúdo final apresentado.

Referências

- Brown, T., et al. (2020). *Language Models are Few-Shot Learners*. Advances in Neural Information Processing Systems (NeurIPS), 33, 1877-1901.
- Gris, L. R. S., et al. (2022). *Brazilian Portuguese speech recognition using Wav2vec 2.0*. In: Lecture Notes in Artificial Intelligence, Springer, Cham, p. 320-330. DOI: 10.1007/978-3-030-98305-5_31.
- Harden, R. M., Stevenson, M., Downie, W. W., & Wilson, G. M. (1975). *Assessment of clinical competence using objective structured examination*. British Medical Journal, 1(5955), 447-451.
- Issenberg, S. B., et al. (2005). *Features and uses of high-fidelity medical simulations that lead to most effective learning: a BEME systematic review*. Medical Teacher, 27(1), 10-28.

- Masters, K. (2023). *Ethical use of artificial intelligence in health professions education: AMEE Guide No. 158*. Medical Teacher, 45(6), 574-584. DOI: 10.1080/0142159X.2023.2186203.
- Park, T. J., et al. (2022). *A review of speaker diarization: Recent advances with deep learning*. Computer Speech & Language, 72, 101317.
- Radford, A., et al. (2023). *Robust speech recognition via large-scale weak supervision*. International Conference on Machine Learning (ICML).
- Singhal, K., et al. (2023). *Large language models encode clinical knowledge*. Nature, 620(7972), 172-180.
- Lee, P., Bubeck, S., & Petro, J. (2023). *Benefits, limits, and risks of GPT-4 as an AI chatbot for medicine*. New England Journal of Medicine, 388(13), 1233-1239. DOI: 10.1056/NEJMs2214184.
- Topol, E. J. (2019). *High-performance medicine: the convergence of human and artificial intelligence*. Nature Medicine, 25(1), 44-56.
- Wei, J., et al. (2022). *Chain-of-thought prompting elicits reasoning in large language models*. Advances in Neural Information Processing Systems, 35, 24824-24837.
- Ziv, A., Small, S. D., & Wolpe, P. R. (2000). *Patient safety and simulation-based medical education*. Medical Teacher, 22(5), 489-495.
- Bai, J., et al. (2023). *Qwen Technical Report*. arXiv preprint arXiv:2309.16609.
- Jiang, A. Q., et al. (2023). *Mistral 7B*. arXiv preprint arXiv:2310.06825.