Módulo Integrado para Avaliação Contínua de Modelos de Linguagem em Plataforma de Saúde Digital

Ana Cláudia Machado[1], Elisa Tuler[1], Víctor Macul[3], Olívio Souza[3]
Jussara Almeida[2], Marcos André Gonçalves[2], Leonardo Rocha[1]

[1] Universidade Federal de São João del Rei, Brasil
[2] Universidade Federal de Minas Gerais Brasil
[3] Ana Health Brasil

anaclaudiamachado211@aluno.ufsj.edu.br, (etuler, lcrocha)@ufsj.edu.br
(jussara,mgoncalv@dcc.ufmg.br), (osouzaneto, victor)@anahealth.app

Resumo. *A atenção digital mediada por mensagens amplia acesso, mas intensifica a sobrecarga informacional e o risco de inconsistências em textos clínicos gerados automaticamente. Este artigo descreve um protótipo demonstrável integrado à plataforma Ana Health para estruturar um ciclo contínuo de avaliação de saídas de LLMs (sumarização, transcrição e resposta clínica). A ferramenta suporta avaliação* ***cega*** *(origens anonimizadas: humano vs LLM) e comparação A/B entre modelos, com critérios padronizados e score de 0–100 inspirado na SUS. Descrevemos a motivação, os stakeholders, o processo de desenvolvimento, a elicitação e os requisitos, além do desenho do módulo e do fluxo de avaliação. Por se tratar de trabalho em andamento, arquitetura detalhada, tecnologias finais e uma bateria completa de testes ainda estão em consolidação; contudo, o protótipo e um vídeo de instalação estão disponibilizados para demonstração.*

Abstract. *Digital primary care mediated by messaging platforms increases access but amplifies information overload and the risk of inconsistencies in automatically generated clinical texts. This paper presents a demonstrable prototype integrated into the Ana Health platform to operationalize a continuous evaluation cycle for LLM outputs (summarization, transcription, and patient-facing responses). The tool supports* ***blinded*** *assessment (anonymized origin: human vs LLM) and A/B model comparison, using standardized criteria and a SUS-inspired 0–100 score. We report the motivation, stakeholders, development process, requirements elicitation, key functional/non-functional requirements, and the module design. As work in progress, detailed architecture, final technologies, and a full testing suite are still being consolidated; nevertheless, the prototype and an installation/usage video are released for demonstration.*

1. Introdução

A expansão da atenção digital em saúde, especialmente em modelos de cuidado mediado por *mensagens instantâneas* (i.e., WhatsApp), vem alterando a dinâmica de comunicação entre pacientes e equipes clínicas. Embora esse formato amplie o acesso, ele intensifica um problema operacional recorrente: em atendimentos assíncronos, profissionais precisam revisar históricos longos, heterogêneos e pouco padronizados para recuperar informações clínicas, acompanhar a evolução de sintomas, compreender orientações

anteriores e responder com segurança. Em cenários de alta demanda, essa *sobrecarga informacional* pode aumentar o tempo de resposta, e elevar o risco de omissões, inconsistências e interpretações equivocadas ao longo da continuidade do cuidado.

Nesse contexto, soluções baseadas em Modelos de Linguagem (Grandes ou Pequenos – LLMs ou SLMs) têm sido investigadas para apoiar tarefas como sumarização de conversas, transcrição de áudios e geração de respostas ao paciente, com potencial de acelerar triagem e reduzir esforço cognitivo da equipe. Revisões e resultados recentes indicam que modelos de linguagem podem contribuir para síntese e estruturação de informação [Ferreira et al. 2025], porém ressaltam forte dependência do domínio e riscos associados a imprecisões, alucinações, omissões de alertas e inadequação clínica [Keszthelyi et al. 2023]. Assim, a adoção responsável desses modelos em saúde requer mecanismos **robustos e contínuos de avaliação**, capazes de produzir evidências rastreáveis sobre qualidade e segurança do texto gerado, apoiando governança, atualização controlada e tomada de decisão baseada em dados.

Avaliações de LLMs frequentemente dependem de processos *ad hoc* e de ferramentas externas (planilhas, formulários genéricos ou scripts), o que dificulta a recorrência, a padronização de critérios e a integração ao fluxo operacional das equipes clínicas. Além disso, avaliações comparativas entre modelos ou entre textos humanos e textos gerados por IA podem ser influenciadas por vieses do avaliador quando a origem do conteúdo é conhecida. Para mitigar esse problema, avaliações **cegas** (com anonimização da fonte) e procedimentos padronizados de coleta e agregação tornam-se desejáveis, sobretudo quando se pretende comparar modelos (A/B) e acompanhar evolução ao longo do tempo.

Este artigo descreve um **protótipo demonstrável** integrado à plataforma de atenção primária digital Ana Health, cujo objetivo é operacionalizar um *Ciclo de Avaliações* para textos produzidos em tarefas assistenciais (sumarização, transcrição e geração de respostas), viabilizando (i) a configuração de avaliações com **critérios e afirmações padronizadas**, (ii) a execução **cega** por avaliadores clínicos no mesmo ambiente e fluxo de trabalho, (iii) a comparação entre modelos em testes **A/B** e (iv) a consolidação de resultados em indicadores interpretáveis, incluindo um *score* global. Os principais stakeholders contemplados são: **Equipe de Saúde** (avaliadores e usuários do fluxo), **Gestor de Avaliações** (responsável por configurar ciclos, selecionar casos/datasets e acompanhar resultados), **Equipe de Pesquisa** (curadoria de critérios e análise) e **Desenvolvedores** (evolução do módulo). O desenvolvimento foi conduzido de forma incremental, com mapeamento do fluxo existente, entrevistas e reuniões com as equipes, pilotos externos para validação dos critérios e, por fim, implementação do protótipo integrado. Requisitos funcionais (por perfil Gestor/Avaliador) e não funcionais (recorrência, viabilidade operacional, integração ao fluxo, usabilidade e clareza dos resultados) foram consolidados para orientar o desenho dos fluxos e das interfaces.

2. Visão Geral da Ferramenta

A ferramenta implementa um *Ciclo de Avaliações* que permite: (i) configurar avaliações (critérios, casos/datasets e participantes); (ii) distribuir tarefas a avaliadores no fluxo operacional; (iii) coletar julgamentos de forma **cega** (anonimizando origem/modelo); (iv) consolidar resultados por critério e por ciclo, com *score* e visualizações para decisão. A Figura 1 ilustra a visão geral do módulo e de seus atores.

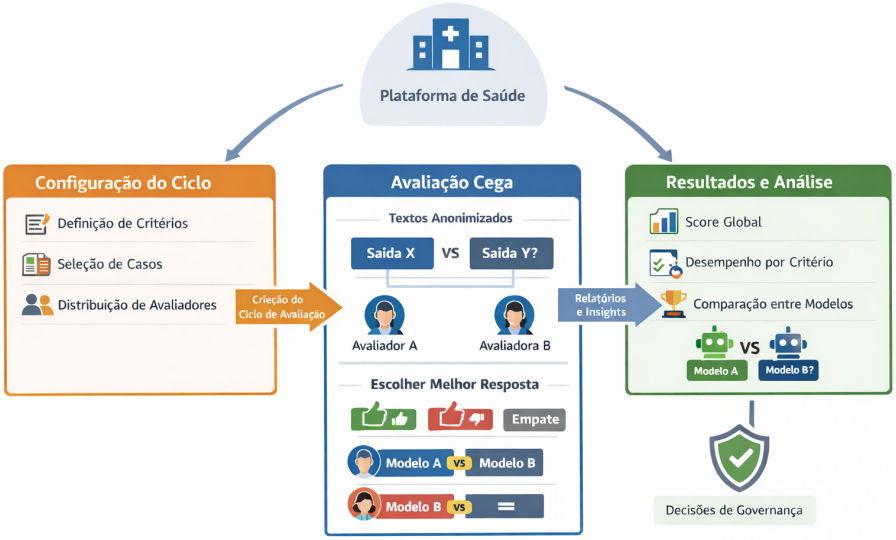

Figura 1. Visão geral do Ciclo de Avaliações (entradas, etapas e saídas).

2.1. Stakeholders e Perfis de Usuário

Identificamos como stakeholders centrais: **Associados** (usuários finais do serviço), **Equipe de Saúde** (acolhimento, equipe médica, psicoterapia) e **Gestor de Avaliações** (responsável pela governança do ciclo). Como atores de suporte: **Equipe de Pesquisa** (definição de critérios e análise) e **Desenvolvedores** (manutenção e evolução do módulo). O foco é viabilizar avaliação frequente com baixo atrito no contexto assistencial.

2.2. Processo de Desenvolvimento

O desenvolvimento foi incremental em quatro etapas: (i) mapeamento do fluxo atual e pontos de contato; (ii) elicitação e consolidação de requisitos; (iii) definição da metodologia de avaliação; e (iv) implementação do protótipo integrado. O desenho priorizou aderência ao fluxo existente, simplicidade para avaliadores e rastreabilidade de decisões.

3. Metodologia de Avaliação

A ferramenta suporta duas modalidades: **(1) Avaliação Tradicional** e **(2) Comparação de Modelos (A/B)**. Em ambas, avaliadores julgam saídas textuais (sumarização, transcrição, resposta) por critérios padronizados. Um **repositório de critérios** foi elaborado pela Equipe de Pesquisa e é composto por afirmações reutilizáveis por tipo de tarefa (e.g., precisão/ruído para transcrição; relevância/empatia/segurança para respostas).

Para a criação de um score, inspiramo-nos na System Usability Scale (SUS) [Brooke 1996] para estruturar uma escala com afirmações positivas e negativas alternadas em Likert (5 pontos). A pontuação é normalizada conforme a (SUS): nos itens positivos subtrai-se 1 e, nos negativos, subtrai-se a pontuação de 5; o somatório é multiplicado por 2,5, gerando *score* 0–100, o que facilita a interpretação e comparação entre modelos/ciclos. Para análise por critério, usamos $f(P, N) = P(6 - N)$, em que P é a média dos itens positivos e N a média dos itens negativos, resultando em valores de 1 a 25 por critério. Na comparação A/B, o avaliador escolhe qual saída atende melhor ao critério (ou empate), gerando taxas de vitória por critério e no total.

4. Elicitação e Requisitos

A elicitação de requisitos foi conduzida de forma incremental e centrada no contexto assistencial, combinando técnicas qualitativas com atividades práticas de validação. Inicialmente, realizamos **entrevistas semiestruturadas** e **reuniões de alinhamento** com representantes da equipe de saúde (acolhimento, médico e psicoterapia) e áreas de suporte, visando caracterizar o fluxo real de trabalho, identificar pontos de atrito na leitura de históricos e explicitar critérios de qualidade esperados para textos gerados (por humano ou por LLM). Em paralelo, conduzimos um **mapeamento de artefatos e rotinas existentes** (por exemplo, organização de tarefas, registros e acompanhamentos) para garantir a aderência ao processo já adotado e minimizar o custo de adoção.

Antes da implementação do módulo, executamos **avaliações piloto** com ferramentas externas (planilhas/formulários), testando a clareza das afirmações, a viabilidade de execução na rotina e a consistência das medidas coletadas. Desenvolvemos **scripts** para automatizar o cálculo do *score* e a agregação por critério, verificando a rastreabilidade e a reprodutibilidade dos resultados. Esses pilotos ajudaram a calibrar o tamanho do instrumento (número de critérios/afirmações) para manter a avaliação curta e frequente.

A especificação foi consolidada em **requisitos funcionais** (separados por perfis Gestor e Avaliador) e **requisitos não funcionais**, que orientaram diretamente o desenho do protótipo. Para manter comparabilidade entre ciclos e facilitar manutenção, priorizamos ainda a criação de um **repositório de critérios** reutilizável (afirmações positivas e negativas por tipo de tarefa), reduzindo retrabalho na configuração de novas avaliações e promovendo padronização ao longo do tempo.

4.1. Requisitos não funcionais

O protótipo foi guiado por cinco requisitos não funcionais centrais: **recorrência** (suportar ciclos periódicos), **viabilidade operacional** (baixo custo de execução por avaliação), **integração ao fluxo** (execução no processo já adotado, e.g., como tarefa no Kanban), **usabilidade** (interação simples e rápida) e **clareza dos resultados** (sínteses acionáveis para orientar decisões e melhorias do modelo).

4.2. Requisitos funcionais centrais

A ferramenta contempla dois perfis principais. O **Gestor de Avaliações** é responsável por criar, editar e encerrar ciclos de avaliação; configurar avaliações por meio da seleção de *critérios* e respectivas afirmações; selecionar casos de teste e/ou *datasets*; atribuir avaliadores (por perfil/equipe) e distribuir tarefas no fluxo operacional; e acompanhar o progresso e os resultados em painéis consolidados. O **Avaliador** (Equipe de Saúde) executa as avaliações atribuídas no contexto assistencial e sinaliza inconsistências, riscos e problemas identificados durante o julgamento das saídas textuais.

5. Protótipo e Demonstração

O módulo *Ciclo de Avaliações* foi implementado como um **protótipo integrado** à plataforma Ana Health para viabilizar avaliação contínua de textos gerados por LLMs (e/ou por profissionais) em tarefas assistenciais, como sumarização, transcrição e geração de respostas. O objetivo é tornar a avaliação **recorrente, rastreável e operacionalmente viável**, reduzindo a dependência de ferramentas externas e incorporando a governança do modelo ao fluxo de trabalho existente.

Figura 2. Exemplo de avaliação comparativa (A/B) com saídas anonimizadas no perfil Avaliador.

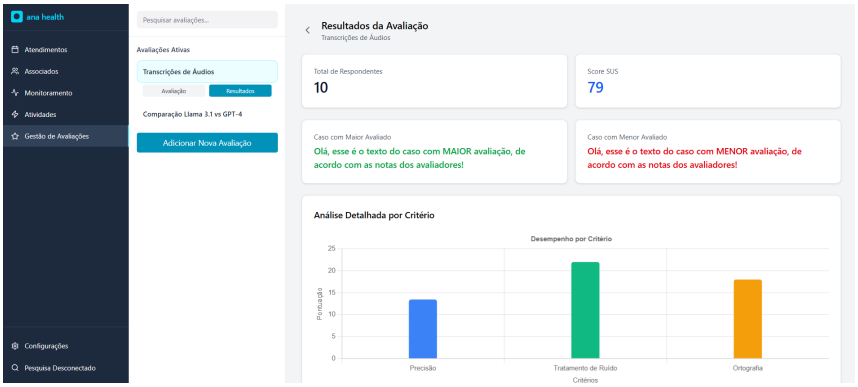

Figura 3. Exemplo de resultados consolidados no perfil Gestor: *score* global e análise por critério.

A Figura 2 ilustra o **momento de execução da avaliação** no perfil Avaliador, destacando o caráter **cego** e a modalidade de **comparação A/B**: para um mesmo caso, duas saídas anonimizadas são apresentadas, e o avaliador escolhe qual atende melhor a cada critério. Esse desenho reduz viés por conhecimento prévio da fonte/modelo e permite comparar alternativas de maneira padronizada. A Figura 3 exemplifica a **consolidação dos resultados** no perfil Gestor de Avaliações, reunindo um *score* global e a análise por critério. Essa síntese apoia decisões de governança (por exemplo, manutenção, ajuste ou substituição de modelos) e facilita a comunicação de evidências às equipes técnicas e clínicas, mantendo rastreabilidade por ciclo e por critério. A Figura 4 complementa o exemplo de execução A/B ao apresentar a **visão consolidada da comparação** no perfil Gestor. Nessa visualização, os votos dos avaliadores são agregados para indicar o desempenho relativo dos modelos, permitindo identificar vencedores globais e diferenças por critério. Esse resumo funciona como evidência prática para decisões de governança, como selecionar qual versão/modelo deve ser mantida em produção ou priorizada para ajustes.

O protótipo disponibilizado[1] [2], bem como o vídeo de demonstração[3] apresentam: (i) instalação/execução do protótipo; (ii) configuração de uma avaliação; (iii) execução

[1]Protótipo do perfil Gestor disponível em: Figma (Protótipo Gestor — AnaHealth).
[2]Protótipo do perfil Avaliador disponível em: Figma (Protótipo Avaliador — AnaHealth).
[3]Vídeo demonstrativo disponível em: Demonstração de Uso

Figura 4. Exemplo de resultados de uma avaliação comparativa: desempenho relativo entre modelos e votos por critério.

cega por avaliadores; e (iv) interpretação dos indicadores gerados. Por se tratar de um trabalho em andamento, a arquitetura detalhada, tecnologias finais e uma bateria formal de testes ainda estão em consolidação; ainda assim, a versão demonstrável já evidencia o ciclo ponta a ponta e sua integração ao fluxo assistencial.

6. Lições Aprendidas

(i) Avaliação contínua só é viável quando integrada ao fluxo existente; (ii) critérios precisam ser reusáveis e claros para reduzir esforço cognitivo; (iii) anonimização é crucial para reduzir viés em comparação humano vs LLM e entre modelos; (iv) síntese 0–100 facilita comunicação com operação e tomada de decisão.

7. Limitações

Este artigo descreve um **trabalho em andamento** focado em viabilizar avaliação contínua e cega no fluxo assistencial. Por se tratar de protótipo, a **arquitetura detalhada** (padrões e componentes), a **pilha tecnológica final** e uma **avaliação formal completa** (testes unitários/integração/sistema/usabilidade/satisfação) ainda estão em consolidação e serão relatadas em versão futura. Ainda assim, o sistema nosso protótipo é executável.

8. Trabalhos Futuros

Consolidar arquitetura e tecnologias; ampliar bateria de testes; conduzir estudo com avaliadores (usabilidade/satisfação) e avaliação longitudinal por ciclos; incorporar métricas automáticas complementares e mecanismos de auditoria; expandir repositório de critérios por tarefa clínica; e suportar análises estratificadas por perfil de caso/paciente e risco.

Agradecimentos Ana Health, INCT-TILDIAR, SEBRAE, EMBRAPII, FINEP, CAPES, CNPq, FAPEMIG, UFMG, UFSJ e UFOP.

Referências

Brooke, J. (1996). Sus-a quick and dirty usability scale. *Usability evaluation in industry*, 189(194):4–7.

Ferreira, A., Rocha, L., Cunha, W., Machado, A., Campos, J., Jallais, G., Viana, A., Tuler, E., Araújo, I., Macul, V., Souza Neto, O., Pereira de Souza Júnior, A., de Souza, G., Pallone, J., Aparecida, M., Santos, W., and Gonçalves, M. A. (2025). A comprehensive qualitative analysis of patient dialogue summarization using large language models applied to noisy, informal, non-english real-world data. *Scientific Reports*, 15:31660.

Keszthelyi, D., Gaudet-Blavignac, C., Bjelogrlic, M., and Lovis, C. (2023). Patient information summarization in clinical settings: Scoping review. *JMIR Medical Informatics*, 11:e44639.