

APRENDIZADO DE MÁQUINA PARA SAÚDE

LLMs e Multimodalidade: a nova fronteira da inteligência clínica

Da fluência estatística ao cuidado seguro, auditável e humano-no-loop

APOSTILA TÉCNICA · ENCONTRO 10

Prof. Flávio Luiz Seixas
Pós-graduação em Computação

Sumário

Sumário	2
Introdução	3
1. O que um LLM processa e produz	3
2. Modelos contextuais e geração	3
3. Sumarização, codificação e diálogo.....	4
4. Multimodalidade e recuperação	4
5. Alucinação, explicação e avaliação	4
6. Segurança humano-no-loop	4
Síntese do encontro.....	5
Referências	5

Introdução

Modelos de linguagem ampliam as possibilidades de geração, transformação e consulta de texto, mas fluência não é evidência de correção. Em saúde, qualquer uso deve separar suporte à organização da informação de decisão autônoma e deve tornar verificáveis dados de entrada, saídas e revisão humana. [1]

As referências atualizadas permitem aprofundar tokenização, geração autorregressiva, recuperação aumentada, alucinação, documentação ambiental, explicabilidade e segurança de LLMs clínicos. [1]

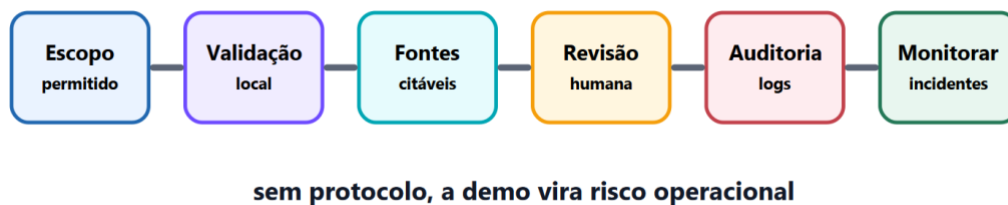


Figura 1 — Protocolo conceitual de segurança para uso clínico de LLMs. Fonte: elaboração própria com base em [1]–[5].

1. O que um LLM processa e produz

O texto é convertido em tokens, associados a embeddings, antes de atravessar blocos Transformer. Em modelos geradores, a saída é construída token a token: cada token escolhido é anexado ao contexto e alimenta a próxima passagem. Essa mecânica explica por que coerência linguística não equivale a consulta automática a uma base clínica [6].

Arquiteturas apenas com encoder, como BERT, são especialmente usadas para representar e classificar texto; arquiteturas geradoras usam decodificação para produzir sequências. A distinção ajuda a escolher entre extrair uma variável, recuperar um documento e redigir uma resposta aberta [6].

CONCEITO-CHAVE

O texto é convertido em tokens, associados a embeddings, antes de atravessar blocos Transformer.

2. Modelos contextuais e geração

Transformers produzem representações condicionadas pelo contexto por mecanismos de atenção. BERT é usado sobretudo para representar e compreender texto; modelos geradores estimam continuações e podem produzir sequências. Pré-treinamento seguido de adaptação permite reutilizar representações em tarefas biomédicas [1].

Em um sistema clínico, a entrada pode incluir notas, pergunta e contexto recuperado; a saída pode ser resumo, código, resposta ou rascunho. A avaliação precisa decompor fidelidade ao registro, omissões, adições não sustentadas e adequação ao usuário [1].

3. Sumarização, codificação e diálogo

Sumarização reduz documentos, mas precisa preservar fatos, incerteza e temporalidade. Codificação associa texto a terminologias; erros podem alterar pesquisa, faturamento e decisão. Sistemas de diálogo precisam gerenciar objetivos, turnos, confirmação e esclarecimento, não apenas responder a enunciados isolados [2].

Em documentação ambiental, um fluxo possível captura o áudio, transcreve e identifica interlocutores, extrai elementos clínicos, propõe uma nota estruturada e a entrega ao profissional para edição e assinatura. A saída permanece um rascunho: o sign-off clínico é parte explícita do processo [7].

4. Multimodalidade e recuperação

Dados clínicos incluem texto, imagem, sinais e medições. Modelos multimodais procuram combinar representações dessas fontes. O ganho potencial depende de alinhamento temporal, disponibilidade e significado de cada modalidade; ausência sistemática pode introduzir viés [3].

Recuperação de informação acrescenta uma etapa que localiza documentos ou trechos relevantes antes da resposta. Ela pode oferecer material verificável e reduzir dependência exclusiva dos parâmetros do modelo, mas não garante que a recuperação seja correta nem que a geração respeite o conteúdo [4].

Na recuperação aumentada por geração (RAG), consulta e documentos são representados para busca, os resultados podem ser reordenados e os trechos selecionados entram no contexto do gerador. A cadeia deve guardar consulta, coleção, versão, trechos e resposta para permitir auditoria; falhas de indexação ou recuperação continuam a contaminar a síntese [6] [8].

5. Alucinação, explicação e avaliação

Alucinação é a produção de conteúdo plausível, porém falso ou não verificável. Em saúde, um valor laboratorial inventado, uma interação medicamentosa incorreta ou uma referência inexistente pode alterar conduta. Barreiras mínimas incluem grounding em fontes, validação clínica e humano-no-loop [7].

Uma explicação útil precisa ser compreensível e também fiel ao comportamento do sistema. RAG pode expor evidências externas, mas não prova que a conclusão esteja correta; explicações narrativas podem soar convincentes sem refletir o mecanismo do modelo. Avaliação deve combinar fidelidade, robustez, compreensão e utilidade clínica [8].

O protocolo de teste deve separar recuperação, geração e efeito no trabalho. Além de métricas de linguagem, deve medir fatos omitidos ou adicionados, concordância com a fonte, estabilidade, calibração de confiança, desempenho por subgrupo, tempo de revisão, correções humanas e incidentes [7] [8].

6. Segurança humano-no-loop

Uma saída bem escrita pode conter informação ausente na fonte. Protocolos devem exigir ligação com evidências, apresentação de incerteza, verificação por profissional autorizado, registro de mudanças e possibilidade de recusa. Quanto maior o impacto da ação, maior deve ser a independência da verificação [5].

Privacidade exige controlar conteúdo enviado, destino, retenção e acesso. LLMs também ampliam a superfície de ataque com manipulação de prompt, envenenamento, inferência de pertencimento e vazamento de informação identificável; controles precisam acompanhar treinamento, adaptação, inferência e integração [5] [9].

Síntese do encontro

O que um LLM processa e produz: o texto é convertido em tokens, associados a embeddings, antes de atravessar blocos transformer.
Modelos contextuais e geração: transformers produzem representações condicionadas pelo contexto por mecanismos de atenção.
Sumarização, codificação e diálogo: sumarização reduz documentos, mas precisa preservar fatos, incerteza e temporalidade.
Multimodalidade e recuperação: dados clínicos incluem texto, imagem, sinais e medições.
Alucinação, explicação e avaliação: alucinação é a produção de conteúdo plausível, porém falso ou não verificável.
Segurança humano-no-loop: uma saída bem escrita pode conter informação ausente na fonte.

Referências

- [1] Xu, H.; Roberts, K. Natural Language Processing. In: Intelligent Systems in Medicine and Health. Springer, 2022.
- [2] Bickmore, T.; Wallace, B. Intelligent Agents and Dialog Systems. In: Intelligent Systems in Medicine and Health. Springer, 2022.
- [3] Sim, I.; Sirota, M. Data and Computation: A Contemporary Landscape. In: Intelligent Systems in Medicine and Health. Springer, 2022.
- [4] Hersh, W. Information Retrieval. In: Biomedical Informatics. Springer, 2021.
- [5] Goodman, K. W.; Miller, R. A. Ethics in Biomedical and Health Informatics: Users, Standards, and Outcomes. In: Biomedical Informatics. Springer, 2021.
- [6] Alammari, J.; Grootendorst, M. Hands-On Large Language Models: Language Understanding and Generation. O'Reilly Media, 2024.
- [7] Shaw, L.; Mahajan, S.; Upreti, K. (eds.). Natural Language Processing for Healthcare: The Rise of Intelligent Assistants. Academic Press, 2026.
- [8] Zamanifar, A.; Taherkordi, A.; Farhadi, A. (eds.). Explainable Large Language Models in Healthcare Applications. Springer, 2026.
- [9] Khalifa, N. E. M.; Taha, M. H. N. (eds.). Next-Gen Healthcare: AI-Powered Medical Innovations. Springer, 2026.