LattesRex: Building ChatBots for Semi-Structured Documents

Lucas Darcio[1], Karina Soares Santos[2], Amanda Spellen[1], Esther Soares[3], Livy Real[1,4], Altigran Soares da Silva[1]

[1] Universidade Federal do Amazonas
Manaus - AM - Brazil

{lucas.darcio, amanda.spellen, alti}@icomp.ufam.edu.br

[2]Serasa
São Paulo - SP - Brazil

ekaarysoares@gmail.com

[3]Universidade Federal de São Carlos
São Carlos - SP - Brazil

esther.soaresx@gmail.com

[4]Jusbrasil
Salvador - BA - Brazil

livyreal@gmail.com

Abstract. *We present LattesRex, a question-answering system based on large language models (LLMs) designed to support the analysis of curriculum vitae from the Lattes Platform. The proposed architecture adopts a structured modular approach inspired by RAG, leveraging metadata to structure the inputs sent to the LLM. We conducted a detailed evaluation, validated by linguists, considering (i) models of different sizes, (ii) documents of distinct lengths, and (iii) queries with different levels of complexity. Results indicate that structured data scale the solution without compromising quality. We contribute a replicable architecture, a systematic qualitative evaluation, and reflections on using LLMs in real-world contexts. All resources will be made publicly available.*

Resumo. *Apresentamos o LattesRex, um sistema de perguntas e respostas baseado em LLMs para auxiliar na análise de currículos da Plataforma Lattes. Propomos uma abordagem estruturada modular inspirada em RAG, explorando metadados para estruturar as entradas enviadas à LLM. Conduzimos uma avaliação detalhada, com validação de linguistas, variando (i) o porte dos modelos, (ii) a extensão dos documentos e (iii) a complexidade das consultas. Os resultados indicam que a estruturação dos dados escala a solução sem perda de qualidade. Contribuímos com uma arquitetura replicável, uma avaliação qualitativa sistemática e reflexões relevantes para o uso de LLMs em contextos reais. Todos os recursos serão disponibilizados publicamente.*

1. Introdução

A análise de perfis acadêmicos é uma tarefa central em diversos processos de avaliação científica, como seleção de bolsistas e julgamento de propostas de pesquisa. No Brasil, esse processo se apoia majoritariamente na Plataforma *Lattes*, que reúne informações sobre produção científica, formação e atividades profissionais. No entanto, o volume, a tecnicidade e a heterogeneidade desses documentos tornam a tarefa complexa,

especialmente quando os avaliadores não dominam a área de atuação do pesquisador avaliado, o que pode comprometer a consistência das decisões. Diante desse cenário, torna-se relevante investigar formas de apoiar a interpretação desses perfis, por meio de recursos computacionais que facilitem consultas específicas a currículos extensos.

Em 2023, a Plataforma *Lattes* já contava com quase 8 milhões de currículos cadastrados, com cerca de meio milhão adicionados apenas em 2022[1], refletindo sua importância no contexto de pesquisa nacional. Ela também se consolidou como fonte estruturada de dados acadêmicos, evidenciada por ferramentas como o scriptLattes [Mena-Chalco and Cesar-Junior 2009], o LattesMiner [Alves et al. 2011a] e suas aplicações no Sucupira [Alves et al. 2011b], além de plataformas mais recentes como o Science Tree [Cota et al. 2021] e o QLattes [Mendonça et al. 2023]. Estudos como os de [Perlin et al. 2017] e [Dias and Moita 2018] reforçam ainda seu valor como base empírica para investigações em larga escala sobre produção e colaboração científica no Brasil.

Apresentamos o *LattesRex*, um agente conversacional para análise de currículos da Plataforma *Lattes* por meio de consultas em linguagem natural. Diferente de ferramentas anteriores, que operam por extração estática e geração de relatórios, o *LattesRex* permite consultas dinâmicas a perfis técnicos, utilizando dados estruturados e pré-processamento semântico. O sistema emprega LLMs para interpretar e sintetizar informações em tempo de execução, oferecendo uma abordagem para análise de perfis acadêmicos[2].

Currículos extensos, como os da Plataforma, representam um desafio para LLMs devido ao volume e à complexidade das informações. Apesar dos avanços recentes na ampliação da janela de contexto, ainda há limitações práticas e custo computacional elevado. Isso motiva o uso de arquiteturas como a Geração Aumentada por Recuperação (RAG), que combina recuperação de trechos com geração textual. O *LattesRex* adota uma estratégia inspirada nessa abordagem, mas se diferencia ao explorar a estrutura interna dos currículos, seus metadados, para selecionar previamente os trechos mais relevantes, reduzindo o risco de alucinações e otimizando o uso da janela de contexto.

Comparamos o *LattesRex* a um baseline que processa currículos de forma monolítica. Testamos diferentes LLMs com currículos variados e consultas de diferentes complexidades. As respostas, avaliadas por linguistas, indicaram uma melhora discreta utilizando a estruturação dos dados com modelos robustos. Já em modelos menores, a fragmentação reduziu sutilmente a qualidade das respostas e aumentou imprecisões e vieses. A abordagem estruturada é a única que permite processar documentos de qualquer extensão. Esses resultados destacam a importância de adaptar o pré-processamento ao porte do modelo, sobretudo em cenários com documentos longos e técnicos.

As contribuições deste trabalho incluem i) o desenvolvimento do *LattesRex*, um agente conversacional voltado à análise de currículos acadêmicos; ii) um estudo comparativo entre abordagens monolíticas e baseadas em dados estruturados; e iii) uma avaliação sistemática que evidencia como a estruturação impacta o desempenho de LLMs na consulta a documentos extensos e técnicos, como os da Plataforma *Lattes*.

---

[1] `https://www.gov.br/cnpq/pt-br/acesso-a-informacao/acoes-e-programas/plataforma-lattes`

[2] Atualmente, não é mais possível realizar o download em larga escala dos currículos Lattes no formato XML na plataforma Lattes. Cada pesquisador, contudo, pode obter o seu próprio currículo neste formato individualmente. Tal limitação não invalida o presente trabalho, uma vez que o estudo concentra-se no uso de dados estruturados no contexto de chatbots baseados em Modelos de Linguagem de Larga Escala (LLMs), e não em uma investigação específica sobre a plataforma Lattes ou seu formato XML.

2. Trabalhos Relacionados

Repositórios de dados acadêmicos como a *Plataforma Lattes* brasileira inspiraram diversas ferramentas e estudos dedicados à extração e análise das informações contidas nos currículos de pesquisadores. Soluções iniciais focaram no processamento das informações dos currículos *Lattes* para produzir análises de vários tipos. Por exemplo, o *scriptLattes* é um sistema de código aberto que gera automaticamente relatórios de produção acadêmica a partir dos CVs[Mena-Chalco and Cesar-Junior 2009]. De forma semelhante, o *LattesMiner* consiste em uma linguagem específica de domínio para extrair informações acadêmicas individuais e de grupos a partir da base *Lattes* [Alves et al. 2011a]. O sistema *Sucupira* utilizou o LattesMiner para identificar e visualizar redes sociais acadêmicas entre pesquisadores [Alves et al. 2011b]. Mais recentemente, a plataforma *Science Tree* foi proposta para construir e explorar árvores genealógicas acadêmicas de pesquisadores brasileiros usando dados da Plataforma *Lattes* [Cota et al. 2021]. A integração da classificação Qualis com dados da plataforma foi proposta com o lançamento do *QLattes* [Mendonça et al. 2023], uma extensão de navegador que enriquece CVs com informações do Qualis.

Além do desenvolvimento de ferramentas, a plataforma *Lattes* tem sido utilizada em estudos em larga escala sobre produtividade acadêmica e tendências. Por exemplo, um conjunto de dados com mais de 180.000 CVs para buscar padrões de produção científica e impacto [Perlin et al. 2017]. Da mesma forma, dados da *Lattes* foram usados como base para uma visão geral da produção científica brasileira [Dias and Moita 2018].

Apesar da grande utilização dos dados da Plataforma, nota-se que a maioria dos estudos utilizou métodos estatísticos ou mineração de texto simples, carecendo da compreensão linguística proporcionada pela Inteligência Artificial (IA) moderna. Avanços recentes em Processamento de Linguagem Natural (PLN) abriram novas possibilidades para a análise de informações acadêmicas. Modelos baseados em transformadores revolucionaram a análise de texto. No domínio científico, modelos especializados como o *SciBERT* [Beltagy et al. 2019] e o *SPECTER* [Cohan et al. 2020] melhoraram significativamente tarefas como classificação e similaridade semântica de documentos acadêmicos.

O surgimento de LLMs, como GPT-3 e GPT-4, expandiu as possibilidades de compreensão de linguagem. Esses modelos executam tarefas complexas com pouco treinamento, utilizando aprendizagem zero-shot ou few-shot [Brown et al. 2020]. Isso gerou interesse em aplicar LLMs à análise de literatura científica. LLMs têm sido aplicadas para resumir artigos, extrair temas de pesquisa e auxiliar em revisões sistemáticas [Antu et al. 2023, Asai and et al. 2024, Felizardo et al. 2024]. Eles também podem fornecer análises semelhantes às de revisores [Liang et al. 2023]. Benchmarks como o *SciAssess* foram propostos para avaliar a proficiência científica desses modelos [Cai et al. 2024]. Apesar desses avanços, pouco se explorou o uso desses modelos em dados acadêmicos estruturados ou semi-estruturados como os da plataforma *Lattes*. Nosso trabalho busca preencher essa lacuna ao aplicar LLMs a currículos da *Lattes*, combinando dados estruturados com compreensão linguística para classificar publicações, detectar tendências e gerar interpretações.

3. Metodologia

O *LattesRex* foi projetado para interpretar e responder questões sobre um CV *Lattes* fornecido como entrada tirando proveito da estrutura padronizada. Como ilustrado na Figura 1, o sistema adota uma arquitetura modular inspirada na abordagem

RAG,composta por três módulos principais: *classificador*, *recuperador* e *gerador*. O módulo *classificador* determina o tipo de informação solicitada na pergunta — por exemplo, produção bibliográfica, formação acadêmica ou atuação profissional. Com base nessa classificação, o módulo *recuperador* localiza e extrai os trechos relevantes do currículo. Por fim, o módulo *gerador* utiliza esses trechos, juntamente com a consulta original, para compor a resposta. Em comparação à arquitetura RAG tradicional, os módulos de *classificação* e *recuperação* do *LattesRex* correspondem à etapa de recuperação de informação. O código-fonte do projeto e os prompts utilizados para as tarefas estão disponibilizados através do repositório público do projeto no github[3].

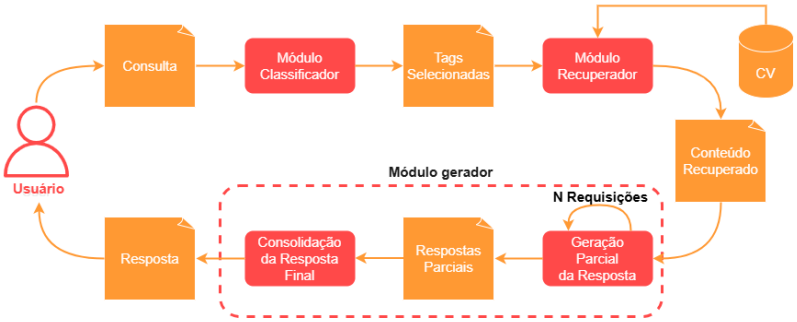


Figura 1. Pipeline do *LattesRex*

A abordagem estruturada explora as capacidades das LLMs para análise contextual, mantendo-se eficiente em termos de custo computacional e aderente às limitações impostas pelos modelos atuais, como as janelas de contexto, particularmente relevantes no caso dos CVs *Lattes*. Detalhamos os três módulos a seguir.

3.1. Módulo Classificador

Os CVs *Lattes* estão disponíveis para download tanto em XML quanto em RTF e PDF. Aqui abordaremos o tratamento de XML (eXtensible Markup Language), formato amplamente utilizado e recomendado pelo W3C, que permite representar relações complexas. Porém esta abordagem é aplicável a qualquer tipo de documento estruturado.

Cada seção dos CVs em XML é estruturada por meio de *tags*, que funcionam como marcadores semânticas, organizando a informação textual em grandes tópicos pertinentes ao domínio. Em sistemas de pergunta e resposta, é comum o uso de tópicos e campos semânticos como indicativos de que uma resposta é adequada para determinada pergunta [Blakemore 1987, McRoy 2021, Zhang et al. 2025]. Assim, usamos as mesmas categorias encontradas no CV para classificar cada consulta do usuário.

As categorias definidas correspondem aproximadamente às macrosseções dos CVs *Lattes*: *Identificação Pessoal e Profissional, Áreas de Atuação e Conhecimento, Formação Acadêmica e Qualificações, Pesquisa e Projetos Acadêmicos, Orientações e Treinamentos, Experiência Profissional, Atividades Acadêmicas e Administrativas, Produção Bibliográfica, Produção Técnica, Tecnológica, Produção Artística e Cultural*.

Para a classificação, o *LattesRex* usa uma LLM. O *prompt de sistema* instrui a LLM a atuar como um classificador inteligente, seguindo o conjunto predefinido de categorias, acompanhadas de suas descrições detalhadas. Também é enviado à LLM instruções de comportamento, tais como restrição à escolha de uma única categoria, critérios de desambiguação, necessidade de aderência estrita ao conteúdo das definições e um exemplo ilustrativo que funciona como orientação de formato e conteúdo da resposta.

[3] https://github.com/Lucas-Darcio/LattesRex

3.2. Módulo Recuperador

Com a categoria da consulta em mãos, extraímos do documento o conteúdo textual a ela associada. Cada categoria é relacionada a um conjunto de tags granulares do XML. Por exemplo, a categoria `Produção Bibliográfica` apresenta as seguintes tags: *TRABALHOS-EM-EVENTOS; ARTIGOS-PUBLICADOS; ARTIGOS-ACEITOS-PARA-PUBLICACAO; LIVROS-E-CAPITULOS;CAPITULOS-DE-LIVROS-PUBLICADOS; TEXTOS-EM-JORNAIS-OU-REVISTAS; DEMAIS-TIPOS-DE-PRODUCAO-BIBLIOGRAFICA*.

As tags associadas a cada macro-categoria foram selecionadas para abranger partes do documento relevantes à consulta, equilibrando cobertura e granularidade.

3.3. Módulo Gerador

Nesta etapa, uma LLM gera a resposta final à consulta. Para que todo o conteúdo recuperado sirva de contexto para a geração, este é avaliado quanto à quantidade de tokens. Se necessário, o conteúdo é particionado de forma coerente, observando início e fim de tags granulares. Com *chunks* que respeitem a janela de contexto da LLM, são realizadas múltiplas requisições parciais até que todo o conteúdo seja processado. Por fim, gera-se uma resposta a partir das saídas das requisições parciais. A utilização de *chunks* controlados permite escalar a geração de respostas independente da extensão do CV.

O *prompt* final, que agrega múltiplas requisições, explicita as seguintes instruções operacionais: (i) eliminar redundâncias e sobreposições semânticas entre as respostas parciais, (ii) evitar a introdução de novas informações, (iii) realçar eventuais contradições ou lacunas informativas, (iv) gerar uma resposta textual fluente, coesa e adaptada ao grau de formalidade exigido em contextos acadêmicos. O modelo é instruído a explicitar casos nos quais o conteúdo não responde à consulta do usuário.

3.4. Interface do Usuário

Uma interface para o usuário foi desenvolvida em Streamlit, disponível via Web, mostrada na Figura 2. O usuário carrega um CV ou o seleciona de uma base pré-estabelecida. Em seguida, com o documento carregado, o usuário consulta aquele CV.

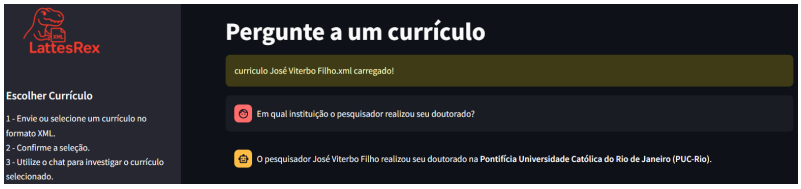


Figura 2. Interface do *LattesRex*

4. Resultados Experimentais

Aqui apresentamos os resultados de experimentos de avaliação do *LattesRex* em diferentes configurações. Considerando que muitos currículos Lattes excedem o limite da janela de contexto dos modelos da família GPT, utilizamos também a família Gemini, que permite janelas significativamente maiores — chegando a até 1 milhão de tokens. Isso possibilitou uma avaliação comparativa com abordagens que analisam diretamente o conteúdo completo dos currículos, sem segmentação ou pré-processamento. Denominamos essa estratégia de abordagem *monolítica*, uma vez que ela desconsidera a estrutura interna do documento. É importante observar que, mesmo com janelas ampliadas, os modelos Gemini podem ainda enfrentam limitações ao lidar com currículos excepcional-

mente longos, como os de pesquisadores com décadas de produção acadêmica contínua.

Conduzimos dois experimentos comparando a abordagem estruturada do *LattesRex* com a abordagem monolítica, utilizando os modelos dominantes (*mainstream*) que tínhamos acesso, GPT-4o e Gemini Flash 2.0, e suas contrapartes menores, GPT-4o mini e Gemini 2.0 Flash Lite. Utilizamos dois currículos de tamanhos distintos, um de uma pesquisadora em início de carreira (137.923 tokens) e outro de uma docente sênior (582.088 tokens). A avaliação foi conduzida por duas linguistas especialistas, a partir de 18 perguntas reais formuladas por um pesquisador com décadas de experiência em avaliação de projetos para o CNPq, complementadas por um conjunto de 10 perguntas factuais elaboradas por linguistas com experiência em avaliação de sistemas de IA. O protocolo adotado simulou o uso real do sistema, permitindo a análise do desempenho gerativo das LLMs e a comparação entre as arquiteturas. É importante destacar que no *LattesRex* temos adicionalmente o módulo classificador baseado em LLMs, que pode eventualmente introduzir erros ao longo do processo. No entanto, optamos por concentrar a avaliação humana na etapa final do sistema, pois é nesse ponto que os efeitos práticos das diferentes configurações se manifestam e podem ser comparados de forma mais significativa.

A avaliação de cada experimento foi feita separadamente. Em cada experimento, as respostas geradas foram organizadas em triplas `{pergunta|resposta1|resposta2}` e submetidas à anotação cega. As avaliadoras, necessariamente, deveriam indicar qual das duas respostas consideravam melhor. Mesmo nos casos em que ambas fossem insatisfatórias, a resposta menos inadequada deveria ser apontada. Em seguida, selecionavam uma única categoria predefinida que justificasse sua escolha. Caso nenhuma das categorias disponíveis fosse adequada, podiam optar pela categoria "Outros" e propor uma nova. Todas as avaliações estão disponíveis[4]. As definições de cada categoria de preferência fornecidas às avaliadoras estão resumidas abaixo:

- **Corretude** – A resposta apresenta dados factuais corretos (sem alucinações ou erros de cálculo).
- **Completude** – A resposta é suficiente e necessária, cobre aspectos essenciais sem se estender.
- **Aderência** – A resposta segue adequadamente a instrução fornecida pelo usuário.
- **Linguagem** – A redação está clara, bem estruturada e sem desvios linguísticos.
- **Utilidade** – A resposta é útil e pertinente diante da demanda expressa pelo usuário.
- **Outros** – Selecionada quando o motivo principal da escolha não se enquadra nas categorias anteriores.

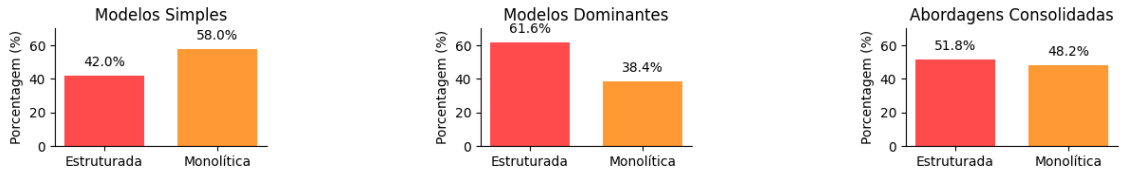
Ao todo, foram analisadas por cada avaliadora 224 perguntas (10 perguntas consideradas difíceis, 18 perguntas consideradas simples, em relação a dois diferentes currículos, e respostas obtidas a partir de 4 modelos distintos). Considerando todas as dimensões avaliadas, o porte do modelo utilizado teve maior impacto nas diferentes avaliações. Portanto, focamos em diferenciá-los. Diferenças entre as famílias Gemini e GPT também foram observadas e são discutidas na Seção 5.

4.1. Avaliação Geral

As anotadoras tiveram 66% de acordo geral, preferindo a **abordagem estruturada** em 52% das vezes, Figura 3. Tais números não indicam uma superioridade significativa de uma abordagem à outra. Os principais motivos de preferência por um modelo

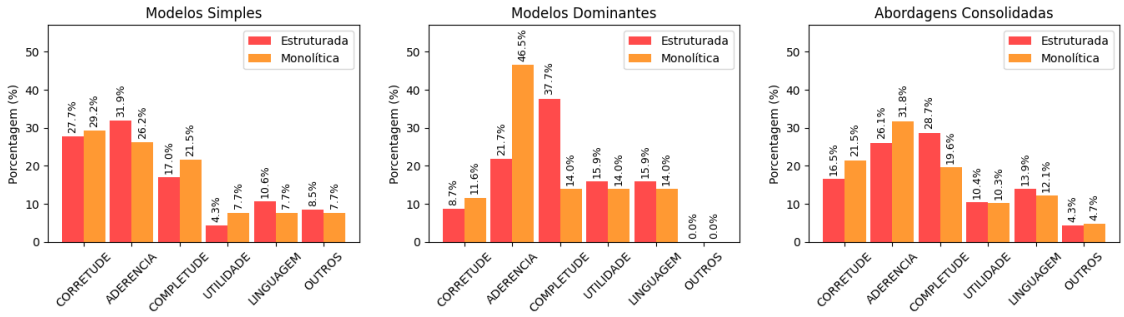
[4]`https://github.com/Lucas-Darcio/LattesRex`

ou outro foram aderência, completude e corretude, Figura 4. É interessante notar que o critério **completude** se sobressai quando a abordagem estruturada é preferida, indicando que os *chunks* semânticos vindos da estruturação do CV em XML não deterioraram a qualidade das respostas geradas pela abordagem estruturada. A categoria "Outros" foi usada apenas em 6 respostas, todas elas indicando 'desinformação' na resposta gerada.



(a) (b) (c)

Figura 3. Preferência entre abordagens entre diferentes tipos de modelos. a) Modelos simples. b) Modelos dominantes. c) Abordagens consolidadas.



(a) (b) (c)

Figura 4. Motivos de Predileção dos Modelos em diferentes modelos. a) Modelos simples. b) Modelos dominantes. c) Abordagens consolidadas.

4.2. Comparando Modelos Menores

Avaliamos separadamente o desempenho das duas abordagens utilizando apenas as contrapartes menores dos modelos dominantes, GPT-4o mini e Gemini 2.0 Flash Lite. Dentre as 112 respostas analisadas, obtivemos uma taxa de acordo de 70%. Aqui é notável que a preferência pela abordagem muda: em 58% das vezes a **abordagem monolítica**, baseada no Gemini 2.0 Flash Lite, é preferida, ou seja, em 42% das vezes a abordagem estruturada GPT-4o mini é preferida. Ainda assim, a diferença dos números é pequena, indicando que ambas apresentam desempenhos comparáveis dentro do contexto avaliado.

Os principais critérios que motivaram a escolha da abordagem monolítica foram: corretude (29,2%), aderência (26,2%) e completude (21,5%). Na preferência pela abordagem estruturada, destacam-se aderência (31,9%), corretude (27,7%) e completude (17,0%). Esses dados indicam que, independentemente do modelo selecionado, as avaliadoras fundamentaram suas escolhas com base nos mesmos critérios principais: **corretude, aderência e completude**. A distribuição das preferências e das justificativas fornecidas não aponta para uma superioridade clara de uma abordagem. No entanto, é relevante observar quais critérios foram mais influentes no processo de avaliação.

4.3. Comparando Modelos Maiores

Quando comparadas as arquiteturas utilizando os melhores modelos dominantes, GPT-4o e Gemini Flash 2.0, a taxa de acordo entre as anotadoras é de 66%, e a preferência

por cada um das abordagens inverte-se. A **abordagem estruturada** que utiliza o GPT-4o foi preferida 52% das vezes, enquanto a abordagem monolítica, 48%. Note-se que, mais uma vez, não há uma preferência clara sobre uma abordagem ou outra.

É particularmente interessante notar que a distribuição dos motivos de preferência por uma abordagem ou outra muda drasticamente nesse experimento. Aqui destacam-se dois critérios: **aderência** é a principal razão pela escolha da abordagem monolítica (46,5%), já quando a arquitetura estruturada é a predileta é **completude** que se sobressai (37,7%). Nota-se, portanto, que ao considerar a qualidade das respostas geradas, não temos uma abordagem que vai muito melhor que a outra, porém claramente, os motivos pela predileção de uma abordagem ou outras, mudam muito, indicando que os fenômenos linguísticos que emergem de cada uma das abordagens diferem.

5. Notas Relevantes da Avaliação Qualitativa

Apresentamos a seguir os principais achados da análise realizadas pelas linguistas, considerando aspectos metodológicos e limitações da abordagem, destacamos fenômenos linguísticos e vieses identificados, com seus respectivos riscos para a aplicação. Adicionalmente, alguns testes exploratórios complementaram a avaliação formal.

5.1. Vieses em Modelos Menores

Os resultados do experimento com os modelos menores apresentaram uma frequência significativamente maior de vieses e alucinações, em 88% das respostas, os modelos menores trazem algum tipo de viés. Consideramos a taxonomia de alucinações proposta por [Huang et al. 2025] para essa análise. Todas são casos de alucinação intrínseca dos modelos (*intrinsic hallucinations*), 'alucinações que contradizem o conteúdo de referência (em contraposição a 'alucinações extrínsecas', nas quais o conteúdo gerado não pode ser verificado a partir da fonte).

5.1.1. Compensação por Verbosidade

Observou-se que as contrapartes menores dos modelos, ao se depararem com lacunas informacionais ou com a necessidade de realizar inferências mais complexas a partir dos dados do currículo, tendem a gerar respostas prolixas. Em vez de indicarem de forma clara suas limitações ou a ausência explícita da informação solicitada, esses modelos frequentemente recorrem a construções textuais extensas, majoritariamente genéricas, redundantes e pouco informativas. Em situações de alta incerteza, LLMs tendem a gerar respostas longas, compostas por informações irrelevantes à pergunta e por formulações evasivas. Tal comportamento é identificado na literatura [Zhang et al. 2024] como 'compensação por verbosidade (*verbosity compensation* ou *vc*), caracterizado pela geração de respostas com baixa densidade semântica como estratégia para mascarar a falta de confiança na produção de uma resposta concisa.

Em [Zhang et al. 2024], os autores notam que a família de modelos que mais apresenta VC é a família GPT. Encontramos o mesmo comportamento: identificamos que a configuração *LattesRex* baseada justamente no GPT-4o mini, traz como padrão de VC a estrutura de redação dissertativa-argumentativa, comum em redações do ENEM, por exemplo. Frequentemente, as respostas trazem um parágrafo de introdução, um desenvolvimento com argumentos e uma conclusão, todos trazendo a mesma informação factual. Como por exemplo a resposta: ***Com base na análise*** *do currículo, não há informações sobre convites para palestras ou conferências em eventos científicos, seja a nível nacional ou internacional.* ***Os dados disponíveis*** *não indicam que a pesquisadora tenha sido*

convidada como palestrante ou conferencista em qualquer evento desse tipo. ***Portanto, conclui-se*** *que, segundo as informações fornecidas, a pesquisadora não recebeu convites para palestrar em eventos científicos. Essa constatação é baseada estritamente no conteúdo do currículo analisado.* (grifo nosso)

Assim, as respostas geradas pela arquitetura monolítica, que usa o modelo Gemini 2.0 Flash Lite, foram preferidas no critério de **completude** já que, quando ambos os modelos eram incapazes de responder, o GPT apresentava uma resposta muito mais verbosa. É importante ressaltar que a VC só foi observada nos modelos menores, aparecendo em mais de 50% das respostas geradas, configurando-se como o principal viés encontrado.

5.1.2. Viés de Polidez

Observou-se também um volume expressivo, cerca de 30%, de respostas excessivamente positivas, caracterizando o *viés de polidez* ou *sycophancy* (bajulação), em que os modelos priorizam cordialidade e positividade em detrimento da precisão factual. Segundo [Huang et al. 2025], esse viés decorre do treinamento com *Reinforcement Learning from Human Feedback* (RLHF), que valoriza respostas agradáveis ao usuário, ainda que potencialmente imprecisas, risco relevante em contextos que exigem exatidão, como avaliações acadêmicas. O viés foi identificado em todos os modelos avaliados, com maior intensidade na família Gemini e nos modelos menores, afetando principalmente o critério de **corretude** nas avaliações qualitativas. Um exemplo caricato ocorreu quando perguntamos sobre a compatibilidade das pesquisadoras (uma cientista da computação e uma cientista da linguagem) com o projeto 'Novos e velhos atores para estratégias terapêuticas em nefropatias: terapias celulares, acelulares e farmacológicas', projeto real aprovado pelo CNPq na área de Bioquímica e Fisiologia. Os modelos inferem que as pesquisadoras, por conhecerem áreas como Interação Humano-Computador ou Pragmática, poderiam colaborar no projeto de Nefropatia.

A presença do viés de polidez compromete a fidelidade ao documento-fonte, um dos pontos-chaves para se escolher uma arquitetura baseada em RAG[Lewis et al. 2020].

5.1.3. Falha de Raciocínio

Outro ponto observado foi a dificuldade de modelos menores em lidar com quantificação de informações explícitas, especialmente na abordagem estruturada. Por exemplo, ao serem questionados sobre o número de artigos publicados por uma das pesquisadoras (dado factual: 14 artigos), o GPT-4o-mini respondeu '8' e o Gemini Flash Lite, '17'. As LLMs precisavam elaborar sobre o documento de referência, que trazia uma lista clara dos artigos publicados. Essa falha de raciocínio (*reasoning failure*) acontece quando o conteúdo gerado diverge de fatos verificáveis claramente oferecidos ao modelo, que foi incapaz de raciocinar e inferir a resposta adequada. Mais de 40% das respostas dos modelos menores apresentaram falhas de raciocínio, o que representou o segundo comportamento inadequado mais frequente.

Já os modelos maiores mostraram melhor desempenho, aproximando-se dos valores corretos ou evitando discrepâncias numéricas grosseiras. Contudo, inferências numéricas não fundamentadas (extrapolações ou invenções de valores ausentes) ainda ocorreram, embora com menor frequência. A maior incidência de alucinações nos modelos menores reflete-se na predominância do critério de **corretude** nesses casos, que cai drasticamente na análise dos modelos maiores. Essa tendência de extrapolar ou inferir valores não documentados expõe limitações na realização de operações básicas de contagem e recuperação factual frente a perguntas quantitativas. Propomos denominar

este fenômeno de **alucinação factual quantitativa**, um tipo de alucinação causada por falha de raciocínio, por tratar-se especificamente de inferências numéricas imprecisas, de detecção relativamente mais simples, mas com alto risco para o uso do sistema.

5.2. Alucinação de Fidelidade

Também foi observado que as saídas dos modelos da **família Gemini** traziam informações além do conteúdo do documento, mesmo instruídos explicitamente a se basear exclusivamente no conteúdo fornecido. Ainda que estas informações não estivessem erradas, as respostas não se baseavam no contexto dado à LLM, um caso de inconsistência na instrução (*instruction inconsistency*), uma subcategoria de alucinação de fidelidade (*faithfulness hallucination*) [Huang et al. 2025]. Nesse trabalho, capturamos esse fenômeno a partir do critério de **aderência**.

Um exemplo emblemático ocorreu quando o modelo Gemini 2.0 Flash Lite afirmou que uma pesquisadora havia participado de um podcast, trazendo informações reais de uma participação ocorrida em 2017, mas não disponíveis no CV, que trazia uma participação em programa de rádio. A inconsistência evidenciou a dificuldade do modelo em aderir ao documento fornecido, comprometendo a confiabilidade e evidenciando alucinações vindas dos dados de pré-treinamento (*hallucination from pre-training*). Tal comportamento não foi observado em respostas geradas por modelos da família GPT, talvez simplesmente pelo fato de que os dados de treinamento das famílias sejam diferentes.

6. Conclusões

Apresentamos o *LattesRex*, um agente conversacional baseado em LLMs, projetado para lidar com currículos da Plataforma *Lattes* por meio de dados estruturados. A arquitetura do *LattesRex* demonstrou viabilidade técnica e, sobretudo, permitiu escalar a solução para documentos de qualquer extensão, mantendo a qualidade das respostas. Embora as vantagens qualitativas frente à abordagem monolítica sejam modestas, o pré-processamento semântico baseado em metadados do documento mostrou-se decisivo para viabilizar o processamento eficiente em cenários com limitações de janela de contexto.

A análise qualitativa evidenciou quais os critérios mais relevantes para a preferência humana por determinada abordagem, a saber: corretude, aderência e completude. A avaliação direcionada à tarefa encontrou limitações importantes no emprego de modelos menores, como alucinações e vieses. Também mostramos que famílias específicas de LLMs apresentam comportamentos distintos, ressaltando a necessidade de estratégias refinadas de controle e avaliação em tarefas gerativas.

Disponibilizamos uma solução replicável e aberta, contribuindo para o avanço das investigações sobre o uso de LLMs em documentos semi-estruturados, bem como trazemos observações importantes no que diz respeito à utilização de LLMs em português.

7. Limitações e Considerações Éticas

Descrevemos o desenvolvimento de um chatbot para consultas a currículos da Plataforma *Lattes*, reconhecendo limitações relacionadas ao escopo temático e aos recursos disponíveis. A abordagem depende de uma estratégia de *chunking* baseada na estrutura semântica dos documentos, o que pode comprometer a captura de nuances contextuais entre seções. A avaliação foi feita com modelos proprietários, mas propomos como trabalho futuro a exploração de modelos abertos. Os experimentos foram conduzidos com dados públicos da Plataforma *Lattes*, respeitando seus termos de uso. As avaliações qualitativas contaram com o consentimento prévio das participantes, sem uso de dados sensíveis.

Referências

Alves, A. D., Yanasse, H. H., and Soma, N. Y. (2011a). Lattesminer: a multilingual dsl for information extraction from lattes platform. In *Proceedings of the Co-located Workshops of SPLASH 2011*, pages 85–89.

Alves, A. D., Yanasse, H. H., and Soma, N. Y. (2011b). Sucupira: a system for information extraction of the lattes platform to identify academic social networks. In *Proceedings of the 6th Iberian Conference on Information Systems and Technologies*, pages 1–6.

Antu, S. A., Chen, H., and Richards, C. K. (2023). Using llm to improve efficiency in literature review for undergraduate research. In *Proceedings of the Workshop on Empowering Education with LLMs*, pages 8–16.

Asai, A. and et al. (2024). Openscholar: Synthesizing scientific literature with retrieval-augmented lms. *arXiv preprint*. arXiv:2411.14199.

Beltagy, I., Lo, K., and Cohan, A. (2019). Scibert: A pretrained language model for scientific text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*.

Blakemore, D. (1987). *Semantic Constraints on Relevance*. Blackwell, New York, NY, USA.

Brown, T. B., Mann, B., Ryder, N., and et al. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33:1877–1901.

Cai, H., Cai, X., Chang, J., and et al. (2024). Sciassess: Benchmarking llm proficiency in scientific literature analysis. *arXiv preprint*. arXiv:2403.01976.

Cohan, A., Feldman, S., Beltagy, I., Downey, D., and Weld, D. S. (2020). Specter: Document-level representation learning using citation-informed transformers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*.

Cota, J. M. M. C., Laender, A. H. F., and Prates, R. O. (2021). Science tree: a platform for exploring the brazilian academic genealogy. *Journal of the Brazilian Computer Society*, 27(1):1–20.

Dias, T. M. R. and Moita, G. F. (2018). Um retrato da produç ao cient'ifica brasileira baseado em dados da plataforma lattes. *Brazilian Journal of Information Science: Research Trends*, 12(1).

Felizardo, K. R., Lima, M. S., Deizepe, A., Conte, T. U., and Steinmacher, I. (2024). Chatgpt application in systematic literature reviews in software engineering: an evaluation of its accuracy to support the selection activity. In *Proceedings of the 18th ACM/IEEE International Symposium on Empirical Software Engineering and Measurement*, ESEM '24, page 25–36, New York, NY, USA. Association for Computing Machinery.

Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., Chen, Q., Peng, W., Feng, X., Qin, B., et al. (2025). A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2):1–55.

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W.-t., Rocktäschel, T., Riedel, S., and Kiela, D. (2020). Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA. Curran Associates Inc.

Liang, W., Zhang, Y., Cao, H., and et al. (2023). Can large language models provide useful feedback on research papers? a large-scale empirical analysis. *arXiv preprint*. arXiv:2310.01783.

McRoy, S. W. (2021). Discourse and dialog. In *Principles of Natural Language Processing*, chapter 7. University of Wisconsin–Madison Pressbooks / Open Publishing.

Mena-Chalco, J. P. and Cesar-Junior, R. M. (2009). Scriptlattes: an open-source knowledge extraction system from the lattes platform. *Journal of the Brazilian Computer Society*, 15(4):31–39.

Mendonça, N. C., Rodrigues, M. A. F., and Mendonça, L. R. (2023). Qlattes: An open-source tool for qualis annotation and visualization in the lattes platform. In *Anais do XL Semin'ario Integrado de Software e Hardware (SEMISH 2023)*.

Perlin, M. S., Santos, A. A. P., Imasato, T., and Borenstein, D. (2017). The brazilian scientific output published in journals: a study based on a large cv database. *Journal of Informetrics*, 11(1):18–31.

Zhang, N., Zhang, C., Tan, Z., Yang, X., Deng, W., and Wang, W. (2025). Credible plan-driven rag method for multi-hop question answering.

Zhang, Y., Das, S. S. S., and Zhang, R. (2024). Verbosity $\neq$ veracity: Demystify verbosity compensation behavior of large language models.