SNBU 2025
XXIII Seminário Nacional de Bibliotecas Universitárias
17 A 20 DE NOVEMBRO
SÃO PAULO - SP

EIXO 4 - Produtos, Serviços, Tecnologias & Inovação

Super Resumos: inteligência artificial na promoção do acesso ágil à informação científica em saúde

Super Abstracts: artificial intelligence for accelerating access to scientific health information

Francisco Barbosa-Junior – Centro Latino-Americano e do Caribe de Informação em Ciências da Saúde (BRIME/OPAS/OMS) – junior910@gmail.com

Ana Katia Camilo – Centro Latino-Americano e do Caribe de Informação em Ciências da Saúde (BRIME/OPAS/OMS) – camiloan@paho.org

Marcelo Bottura – Centro Latino-Americano e do Caribe de Informação em Ciências da Saúde (BRIME/OPAS/OMS) – botturam@paho.org

Danilo Bissoli Apendino – Centro Latino-Americano e do Caribe de Informação em Ciências da Saúde (BRIME/OPAS/OMS) – apendindan@paho.org

João Paulo Souza – Centro Latino-Americano e do Caribe de Informação em Ciências da Saúde (BRIME/OPAS/OMS) – souzaj@paho.org

Resumo: Este trabalho apresenta a criação dos Super Resumos, ferramenta baseada em Small Language Models (SLMs) para gerar versões hiper sintéticas de resumos científicos com 38 a 62 palavras. A metodologia envolveu curadoria de dados, preparação de ambiente computacional *on-premises* e *fine-tuning* do modelo LLaMA. Os resultados apontam ganho de eficiência na triagem de informação científica em saúde, com alta fidelidade semântica e sustentabilidade técnica.

Palavras-chave: Inteligência Artificial. Serviços automatizados de biblioteca. Biblioteca Virtual em Saúde. Informação em Saúde.

Abstract: This paper presents the creation of Super Abstracts, a tool based on Small Language Models (SLMs) designed to generate ultra-condensed versions of scientific abstracts ranging from 38 to 62 words. The methodology involved data curation, on-premises computational environment setup, and fine-tuning of the LLaMA model.

Results indicate improved efficiency in screening scientific health information, with high semantic fidelity and technical sustainability.

Keywords: Artificial Intelligence. Automated library services. Virtual Health Library. Health Information.

1 INTRODUÇÃO

1.1 Justificativa e Embasamento Teórico

Na era da informação, profissionais em busca de dados de qualidade enfrentam uma sobrecarga informacional significativa, dificultando a tomada de decisões rápidas e embasadas. A diversidade e o volume de informações disponíveis podem exceder a capacidade de processamento dos indivíduos, levando à redução na qualidade dessas decisões (Eppler; Mengis, 2004).

Além disso, de forma geral, a crescente digitalização e o uso intensivo de dispositivos eletrônicos têm impactado negativamente a capacidade de concentração do ser humano. A exposição constante a estímulos digitais pode fragmentar a atenção, comprometendo a assimilação de informações complexas (Ponti, 2023).

As diferenças geracionais também influenciam a forma como as informações são consumidas. Jovens profissionais, como os da geração Z[1], estão habituadas a formatos de conteúdo mais curtos e visuais, sendo bem-vinda a adaptação de formatos clássicos (Hassoun *et al.*, 2023).

1.2 Informação em Saúde

O conceito de informação em saúde é amplo e não se limita apenas a dados epidemiológicos, diagnósticos ou terapêuticos. Ele abrange também elementos relacionados à promoção do bem-estar físico, mental e social, compondo um campo essencial para apoiar tanto decisões clínicas quanto políticas de saúde pública. Conforme destaca Galvão (2021), trata-se de um tipo de informação que conecta aspectos biomédicos a dimensões mais abrangentes do cuidado, exigindo que produtos e serviços informacionais sejam desenhados para atender a múltiplas necessidades dos usuários.

[1] A geração Z (gen-Z ou *zoomers*) compreende pessoas nascidas de 1997 a 2012. Esta é uma das gerações que são chamadas de “nativas digitais”.

1.3 LLMs e SLMs

No final do ano de 2022, o campo do Processamento de Linguagem Natural (PLN) foi transformado pelo lançamento e rápida popularização do ChatGPT, da OpenAI. Produto de um Modelo de Linguagem de Grande Escala, ou *Large Language Models* (LLM), esse modelo – e outros, com o LLaMA, lançado no ano seguinte – são redes neurais profundas baseadas na arquitetura *Transformer*, treinadas com bilhões de parâmetros e imensos volumes de texto.

A grande diferença no uso dos LLMs em relação ao que já se utilizava com NLP está na sua versatilidade para realizar múltiplas tarefas linguísticas, como a geração de texto, análise de sentimentos e respostas a perguntas com uma performance próxima à humana em muitos benchmarks (Bogireddy; Dasari, 2024).

Contudo, a sofisticação desses modelos vem acompanhada de elevados custos computacionais, exigências robustas de hardware e uma série de riscos técnicos e éticos. Entre eles, destacam-se a opacidade dos processos internos, a reprodução de vieses dos dados de treinamento e a ocorrência das chamadas alucinações. Este é um importante fenômeno no qual o modelo gera informações factualmente incorretas, mas apresentadas com alta confiança e fluidez textual, o que pode comprometer decisões baseadas em suas respostas (Sahoo *et al.*, 2024).

Nesse cenário emergem os *Small Language Models* (SLMs), versões menores, muitas vezes com 1 a 8 bilhões de parâmetros, desenhadas para ambientes de execução com menor capacidade computacional. Essas versões podem ser especializadas (*fine-tuned*) para domínios específicos e otimizadas para baixo consumo de recursos computacionais, mantendo, ainda assim, resultados relevantes em tarefas especializadas (Subramanian; Elango; Gungor, 2025).

A principal diferença entre LLMs e SLMs está, portanto, na escala e na flexibilidade. Enquanto LLMs são generalistas, potentes e adaptáveis a uma gama muito ampla de tarefas, os SLMs são especialistas, otimizados para cenários bem definidos. Do ponto de vista prático, LLMs oferecem desempenho superior em benchmarks como CNN/DailyMail, SQuAD e conjuntos de dados de análise exploratória, apresentando maior precisão, coerência e legibilidade (Bogireddy; Dasari, 2024). Por outro lado, SLMs

podem ser mais viáveis em cenários que exigem controle de dados, soberania tecnológica e sustentabilidade econômica (Belcak *et al.*, 2025).

Nesse cenário, o Super Resumos surge como um SLM, tornando-se um novo recurso informacional capaz de oferecer versões super sintéticas de resumos de artigos científicos, contendo de 38 a 62 palavras. Essa abordagem visa facilitar o acesso rápido e eficiente à informação científica, destacando os pontos mais relevantes dos conteúdos originais sem comprometer a qualidade, a precisão dos dados e mantendo equilíbrio entre desempenho, custo computacional e autonomia tecnológica.

2 METODOLOGIA

Este trabalho apresenta o relato de experiência no desenvolvimento do produto de inteligência artificial denominado Super Resumos, no âmbito do Centro Latino-Americano e do Caribe de Informação em Ciências da Saúde (BIREME/OPAS/OMS). Seus métodos foram estruturados em três etapas principais:

(1) Seleção e preparação da base de dados contendo os resumos científicos;
(2) Preparação do ambiente computacional para a ferramenta;
(3) *Fine-tuning* do modelo de linguagem.

2.1 Seleção e preparação da base de dados contendo os resumos científicos

Foram utilizados resumos científicos extraídos da base de dados Mosaico, disponíveis na Biblioteca Virtual em Saúde (BVS), um produto de informação da BIREME. Esta base de dados foi escolhida pela sua disponibilidade no ambiente interno da organização e por ter um tamanho adequado para um piloto de um produto de inteligência artificial. Todos os resumos da base utilizada passaram por um processo de verificação inicial, sendo:

- Verificação automatizada:
    - Possuir resumo maior do que 38 caracteres;
    - Não possuir caracteres especiais no início do resumo, por exemplo: dois pontos, arroba, ponto e vírgula;
- Verificação humana:
    - Não possuir quebras de páginas, ou outras divisões que dificultassem o processamento dos textos;

- Que o *encoding* do arquivo estivesse sendo interpretado de forma correta.

Após a verificação, quando o resumo possuía uma quantidade menor do que 38 caracteres, o mesmo texto era copiado para o campo de Super Resumos. Isso indicaria para o momento de *fine-tuning* de que não era necessário alterar textos que tivessem esta característica. Quando possuía caracteres especiais no início do resumo, eles eram retirados automaticamente e passavam por revisão humana posterior para garantir a integridade do texto. Na presença de quebras de página, estas eram corrigidas de forma manual. No caso do *encoding*, não foi encontrado nenhum erro relacionado a esta característica na preparação do arquivo.

Esta preparação foi essencial para garantir a qualidade das entradas utilizadas para o posterior *fine-tuning* do modelo.

2.2 Preparação do ambiente computacional para a ferramenta

O núcleo tecnológico dos Super Resumos baseia-se em um SLM (Small Language Model) da família LLaMA, versão 3.2, com três bilhões de parâmetros. A escolha desse modelo foi motivada por sua capacidade de operar em infraestrutura *on-premises*, permitindo independência de serviços em nuvem. Essa abordagem facilita o controle dos dados, reduz riscos de vazamento de informações e contribui para maior soberania tecnológica institucional. O sistema operacional utilizado é o *Linux Ubuntu*, versão 22.04, baseado em software livre e de código aberto, ele pode ser usado, modificado e distribuído livremente. Para orquestrar o modelo de linguagem, foi utilizada a ferramenta Ollama, que permite rodar LLMs e SLMs em dispositivos pessoais com baixo overhead de configuração. (Ollama, 2025).

2.3 *Fine-tuning* do modelo de linguagem

O Super Resumos foi especializado (*fine-tuned*) com um recorte do banco de dados Mosaico (seção 2.1), onde versões sintéticas de resumos de artigos científicos foram geradas pelo ChatGPT 4o por meio de um prompt padronizado. O *fine-tuning* foi realizado no ambiente computacional previamente configurado (seção 2.2) em um servidor local na BIREME, utilizando técnicas de verificação de qualidade e validação manual iterativa. Globalmente, o ajuste do modelo envolveu:

- Curadoria de exemplos de resumos bem elaborados (seção 2.1);
- Engenharia de *prompt*, definindo instruções específicas para síntese curta e informativa;
- Controle da temperatura do modelo (nível de criatividade) para manter precisão sem introduzir elementos não existentes (alucinações).

Por fim, para teste de funcionamento, os resumos gerados passaram por avaliações internas de qualidade linguística, legibilidade e fidelidade ao conteúdo original, com revisão por especialistas em informação e saúde.

3 RESULTADOS E DISCUSSÕES

Esta seção apresenta os principais achados observados no desenvolvimento e nas primeiras etapas de operação do sistema, sendo eles:

(1) Geração dos Super Resumos;
(2) Preparação e disponibilização dos resumos no Portal da Biblioteca Virtual em Saúde (BVS);
(3) Ética e sustentabilidade.

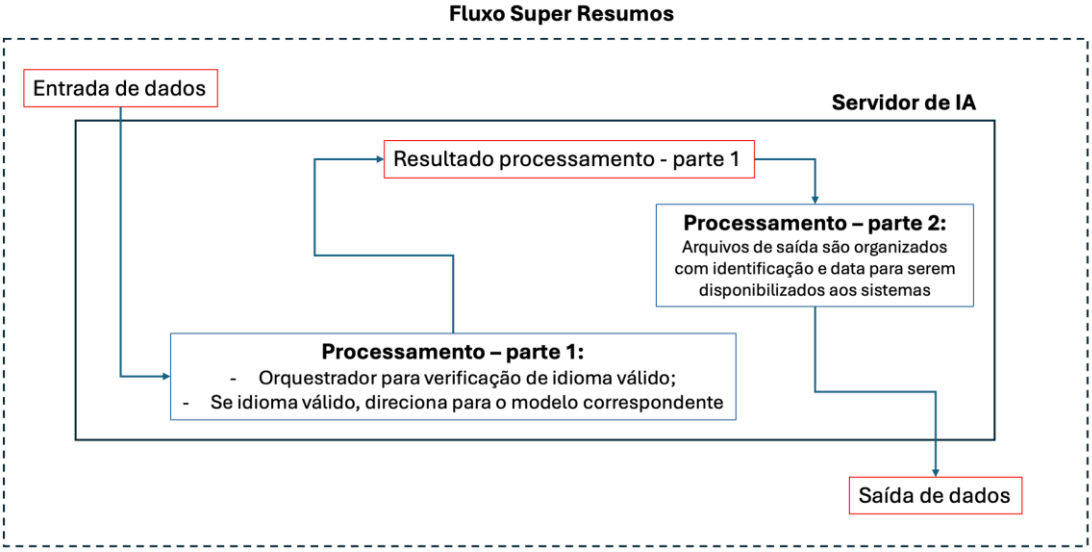
3.1 Geração dos Super Resumos

Após o *fine-tuning*, os Super Resumos passaram a ser gerados de forma automática em lotes (Figura 1) e armazenados nos servidores de IA com identificação correspondente. O processo é executado sob demanda e supervisionado por um sistema de validação que sinaliza incongruências para revisão manual.

3.2 Preparação e disponibilização dos resumos no Portal da Biblioteca Virtual em Saúde (BVS)

Após a geração dos Super Resumos em lote, os arquivos validados na etapa anterior são enviados para a área de atualização e processamento dos registros bibliográficos da BVS. Cada Super Resumo é armazenado como um novo campo nos metadados do registro bibliográfico. A partir daí, os usuários da BVS podem visualizar essa versão condensada dos resumos sempre que disponível.

Figura 1 – Fluxo de geração de super resumos



Fonte: Elaborado pelos autores.
Descrição: Processamento dos resumos dentro do servidor *on-premises.*

3.3 Ética e sustentabilidade

O projeto não envolve dados pessoais nem informações sensíveis. Todo o conteúdo utilizado é de domínio público ou licenciado para redistribuição científica. O uso de modelos locais evita o compartilhamento externo de dados, promovendo conformidade com as diretrizes de proteção da informação em saúde.

Além disso, o uso de um SLM torna o projeto sustentável em médio e longo prazo, ao reduzir os custos operacionais e a dependência de soluções comerciais de IA.

3.4 Resultados técnicos

O modelo LLaMA-3B, classificado aqui como um *Small Language Model* (SLM), demonstrou ser adequado para a tarefa de sumarização científica ultracurta[2], gerando textos entre 38 e 62 palavras com boa coerência, clareza e fidelidade ao conteúdo original. A combinação entre *fine-tuning*, engenharia de prompts e controle da temperatura permitiu que os textos gerados fossem informativos e objetivos, evitando desvios comuns como alucinações e redundância.

[2] Definimos como "sumarização científica ultracurta" uma condensação documental para um texto comparativamente menor do que um resumo tradicional.

Ainda que não tenha sido realizado um experimento comparativo formal com métricas como BLEU ou ROUGE, como em estudos de *benchmarking* (Bogireddy; Dasari, 2024), avaliações internas realizadas por especialistas da BIREME apontaram alta conformidade semântica entre os Super Resumos e os resumos originais, com aprovação média superior a 90% nos lotes validados.

3.5 Impacto para os usuários

Do ponto de vista do usuário final, avaliado qualitativamente por parte da equipe da organização, foi identificado o impacto na redução do tempo necessário para avaliar a relevância de um artigo. Com a inserção dos Super Resumos nos registros da base Mosaico e, progressivamente, para outras bases disponíveis no portal regional da BVS, gestores, pesquisadores e profissionais de saúde passaram a contar com uma visualização prévia mais eficaz, o que facilita a triagem e agiliza processos decisórios.

Essa abordagem também está alinhada a achados da literatura sobre eficiência informacional, onde textos curtos e bem estruturados podem aumentar a capacidade de resposta em ambientes sobrecarregados (Cachola *et al.*, 2020).

3.6 Desafios enfrentados

Durante o desenvolvimento, os principais desafios incluíram:

- Escolha de um modelo de IA compatível com a infraestrutura local;
- Manutenção da qualidade científica dos resumos gerados;
- Estabelecimento de um fluxo sustentável de validação e publicação.

A superação desses desafios exigiu ajustes iterativos no modelo, especialmente na parametrização de criatividade (temperatura), e o desenvolvimento de um pipeline de validação escalável. Além disso, o projeto reforçou a importância de equilibrar inovação tecnológica com viabilidade e sustentabilidade institucional, este um aspecto frequentemente negligenciado em implementações de IA.

3.7 Discussão

O caso dos Super Resumos revela o potencial dos SLMs aplicados a contextos especializados de informação científica. Ao contrário dos LLMs de grande porte, como o ChatGPT-4, que oferecem desempenho superior, mas exigem infraestrutura robusta

(Bogireddy; Dasari, 2024), os SLMs se mostram uma alternativa prática e eficaz em ambientes com restrições técnicas e orçamentárias.

Do ponto de vista da biblioteconomia e da ciência da informação, a iniciativa representa uma nova ferramenta que pode compor um sistema de Recuperação de Informação para melhorar a seleção dos resultados de artigos, ao criar um grau intermediário entre a indexação e o resumo tradicional, com potencial para ser explorado em outras bases, sistemas e contextos.

Apesar dos benefícios observados, é necessário reconhecer que o uso dos Super Resumos pode gerar riscos associados à leitura superficial. A disponibilização de versões hiper sintéticas de resumos pode levar alguns usuários a tomar decisões com base apenas nesses textos, sem consultar o artigo completo. Tal risco já foi apontado em estudos que discutem as discrepâncias entre os resumos e o texto completo (Li *et al.*, 2017). Assim, reforça-se que os Super Resumos não substituem a leitura integral dos trabalhos originais, mas atuam como ferramenta de apoio à triagem inicial, devendo ser utilizados de forma crítica e complementar no processo de recuperação e uso da informação científica.

4 CONSIDERAÇÕES FINAIS

O projeto Super Resumos representa uma inovação estratégica no campo da mediação da informação científica em saúde. Através do uso de um *Small Language Model* (SLM) ajustado especificamente para esse fim, foi possível gerar versões hiper sintéticas de resumos bibliográficos com alto grau de objetividade, clareza e fidelidade ao conteúdo original.

Os resultados apontam para um impacto positivo na experiência dos usuários da Biblioteca Virtual em Saúde (BVS), com destaque para a redução do tempo de triagem de conteúdos, a facilitação da tomada de decisões rápidas e embasadas e a ampliação do acesso à informação em contextos de alta demanda cognitiva.

O uso de uma infraestrutura local e a adoção de um modelo leve (LLaMA-3B) demonstraram que é possível desenvolver soluções baseadas em inteligência artificial que sejam ao mesmo tempo eficazes, economicamente sustentáveis e compatíveis com as diretrizes de privacidade e soberania informacional.

Entre as limitações do produto, destacam-se:

- A ausência, até o momento, de uma avaliação formal com métricas padronizadas (como BLEU ou ROUGE);
- Estudos de impacto na rotina profissional dos usuários;
- A avaliação inicial do resultado dos Super Resumos que foram gerados teve como base os resumos já existentes, não sendo feita uma avaliação perante o texto completo.

Como encaminhamentos, propõe-se:

- A ampliação da cobertura dos Super Resumos para outras bases bibliográficas da BVS;
- A realização de estudos quantitativos de desempenho e impacto;
- A experimentação da ferramenta em outros contextos de ensino, extensão e políticas públicas;
- E a avaliação formal com métricas padronizadas.

Conclui-se que os Super Resumos representam uma abordagem promissora e replicável para enfrentar um dos grandes desafios da atualidade: acessar, compreender e utilizar informação científica de forma rápida, segura e inteligente.

REFERÊNCIAS

BELCAK, P. et al. **Small language models are the future of agentic AI**. *arXiv*, 2025. Disponível em: http://arxiv.org/abs/2506.02153. Acesso em: 10 jun. 2025.

BOGIREDDY, S. R.; DASARI, N. Comparative Analysis of ChatGPT-4 and LLaMA: performance evaluation on text summarization, data analysis, and question answering. *In*: INTERNATIONAL CONFERENCE ON COMPUTING COMMUNICATION AND NETWORKING TECHNOLOGIES (ICCCNT), 15., 2024, Kamand. **Proceedings** [...]. Kamand, IEEE, 2024. Disponível em: https://ieeexplore.ieee.org/document/10725662/. Acesso em: 02 jun. 2025.

CACHOLA, I.; LO, K.; COHAN, A.; WELD, D. S. TLDR: extreme summarization of scientific documents. *In:* FINDINGS OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS: EMNLP 2020, Online, 2020. p. 4766–4777. Disponível em: https://aclanthology.org/2020.findings-emnlp.428/. Acesso em: 27 ago. 2025

EPPLER, M. J.; MENGIS, J. The concept of information overload: a review of literature from organization science, accounting, marketing, MIS, and related disciplines. **The Information Society**, v. 20, n. 5, p. 325–344, 2004. DOI: 10.1080/01972240490507974. Disponível em:

https://www.researchgate.net/publication/220175453_The_Concept_of_Information_Overload_A_Review_of_Literature_From_Organization_Science_Accounting_Marketing_MIS_and_Related_Disciplines. Acesso em: 10 jun. 2025.

GALVÃO, M. C. B. Usuários da informação em saúde: das necessidades aos produtos e serviços informacionais. In: CASARIN, H. de C. S. (org.). *Usuários da Informação e Diversidade*. Marília: Oficina Universitária; São Paulo: Cultura Acadêmica, 2021. p. 169-194. Disponível em: https://ebooks.marilia.unesp.br/index.php/lab_editorial/catalog/book/275. Acesso em: 27 ago. 2025.

HASSOUN, A. *et al*. **Practicing information sensibility: how gen z engages with online information**. *arXiv*, 2023. Disponível em: http://arxiv.org/abs/2301.07184. Acesso em: 25 jun. 2025.

LI, G. *et al*. A scoping review of comparisons between abstracts and full reports in primary biomedical research. **BMC Medical Research Methodology**, London, v. 17, n. 181, 2017. Disponível em: https://bmcmedresmethodol.biomedcentral.com/articles/10.1186/s12874-017-0459-5. Acesso em: 27 ago. 2025.

OLLAMA. Run open-source large language models locally. 2025. Disponível em: https://ollama.com. Acesso em: 10 jun. 2025.

PONTI, M. Screen time and preschool children: promoting health and development in a digital world. **Paediatrics & Child Health**, v. 16, n. 28, p. 184–202, maio 2023. Disponível em: https://pubmed.ncbi.nlm.nih.gov/37205134/. Acesso em: 10 jun. 2025.

SAHOO, N. R. *et al*. Addressing bias and hallucination in large language models. *In*: PROCEEDINGS OF THE JOINT INTERNATIONAL CONFERENCE ON COMPUTATIONAL LINGUISTICS, LANGUAGE RESOURCES AND EVALUATION (LREC-COLING), 2024, Torino. **Proceedings** [...]. Torino: ELRA, 2024. p. 73–79. Disponível em: https://aclanthology.org/2024.lrec-tutorials.12/. Acesso em: 25 jun. 2025.

SUBRAMANIAN, S.; ELANGO, V.; GUNGOR, M. **Small language models (SLMs) can still pack a punch: a survey**. *arXiv*, 2025. Disponível em: http://arxiv.org/abs/2501.05465. Acesso em: 10 jun. 2025.