# LLMs Aplicados à Gestão da Inovação: Extração e Estruturação de Dados do Lattes dos Docentes para Vitrine Tecnológica do PROFNIT.

Maria Sarajane Farias da Costa (IFPB, Campus Campina Grande), Katyusco de Farias Santos (IFPB, Campus Campina Grande).

**E-mails:** maria.sarajane@academico.ifpb.edu.br , katyusco.santos@ifpb.edu.br .

**Área de conhecimento (Tabela CNPq):** 12 - Pós-graduação stricto sensu PI&TT para Inovação - PROFNIT

**Palavras-chave**: Inteligência Artificial; *Large Language Models*; Currículo Lattes.

## 1. Introdução

Os Institutos de Ciência e Tecnologia - ICTs têm um papel fundamental na pesquisa e no desenvolvimento dos produtos tecnológicos, impactando positivamente a sociedade. Eles enfrentam uma dificuldade significativa em comunicar suas inovações tecnológicas para a sociedade e o setor produtivo, o que limita a visibilidade desses produtos e pode resultar em perda de receitas para os programas de pós-graduação.

Nesta perspectiva, o Programa de Pós-graduação em Propriedade Intelectual e Transferência de Tecnologia para a Inovação (PROFNIT) vem se destacando como uma iniciativa estratégica para a formação de profissionais voltados à inovação e à transferência tecnológica no Brasil. Nesse contexto, torna-se essencial fortalecer os mecanismos de gestão da informação científica e tecnológica gerada por seus docentes.

O PROFNIT teve início em 2016 e está em pleno funcionamento (CAPES, 2024). Avaliado com nota 4 pela Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), oferece o curso de Mestrado Profissional. O objetivo do programa é formar recursos humanos já engajados ou interessados em atuar nas competências dos Núcleos de Inovação Tecnológica (NITs) determinadas por lei, bem como nos Ambientes Promotores de Inovação em setores diversos, como o acadêmico, empresarial, governamental, e em órgãos (PROFNIT, 2024).

Com o intuito de ampliar a divulgação dos produtos tecnológicos gerado pelo programa, o projeto que está em andamento tem como objetivo aplicar *Large Language Models* (LLMs) para a extração de dados dos Currículos Lattes (CV) dos docentes a fim de estruturar os produtos tecnológicos do PROFNIT, ponto focal Instituto Federal da Paraíba (IFPB), visando subsidiar a criação de uma Vitrine Tecnológica (VT) e otimizar a divulgação da produção do programa. Como diferencial inovador, o modelo permite a construção de uma vitrine tecnológica a partir de recursos humanos sem formação específica em Tecnologia da Informação, área atualmente caracterizada por forte demanda e escassez de especialistas. Essa abordagem amplia a acessibilidade, otimiza recursos institucionais e contribui para a democratização do uso de inteligência artificial aplicada à gestão de dados acadêmicos e científicos.

Esta proposta é um desdobramento direto dos conhecimentos adquiridos na pesquisa anterior dos autores, que se concentrou no mapeamento temático das áreas do conhecimento dos Trabalhos de Conclusão de Curso (TCCs) do PROFNIT com o uso de LLMs. As competências adquiridas durante estudos anteriores foram fundamentais para a concepção deste projeto, que visa aplicar a inteligência artificial (IA) de forma prática na gestão acadêmica. Mesmo sem formação técnica em programação, a autora desenvolve este estudo com o suporte direto de *Large Language Models* (LLMs), evidenciando a facilidade de acesso e a versatilidade dessas ferramentas para pesquisadores de diversas áreas. Os LLMs são uma classe especial de modelos de linguagem pré-treinados (PLMs), desenvolvidos através do aumento do tamanho dos modelos, do uso de extensos corpus de pré-treinamento e do emprego de grande capacidade computacional (KALYAN, 2024).

A LLM utilizada foi o ChatGPT da OpenAI, versão 2, modelo GPT-4 versão paga. O ChatGPT avançou ao integrar entradas multimodais e dados visuais, marcando um progresso significativo no campo. Essa evolução está diretamente ligada ao desenvolvimento dos mecanismos de atenção nos LLMs, fundamentais para aprimorar a compreensão contextual. Inicialmente concebida como um conceito básico, a atenção evoluiu para estruturas mais complexas, cada uma enfrentando desafios próprios de otimização. Construindo sobre essas inovações, a arquitetura dos LMMs utiliza mecanismos de atenção avançados para processar e sintetizar informações provenientes de múltiplas modalidades (CHEN et al, 2024).

Esses avanços possibilitaram a criação de interfaces simples e acessíveis, permitindo que qualquer usuário, mesmo sem conhecimentos técnicos, interagisse com os sistemas de IA. Por meio de instruções em linguagem natural, como *prompts* em texto ou áudio, os usuários podem direcionar os sistemas para resolver problemas ou realizar tarefas técnicas e não técnicas de forma eficiente (VELA, 2024). Neste sentido, o Prompt é a linguagem de comunicação com a máquina (KASNECI et al., 2023).

Estudar como os prompts influenciam o fluxo de organização dessa arquitetura neural tornou-se fundamental para a prática do *prompting*, ou para uma engenharia de prompt mais eficaz, em qualquer área (LIU et al., 2023). A engenharia de prompt, ou prompt engineering, é o domínio da construção otimizada dessa comunicação, explicando como uma mesma tarefa atribuída à IA pode gerar resultados diferentes em função dos comandos aplicados

(KASNECI et al., 2023). Em outras palavras, os prompts desenvolvem o Deep Learning (DL) da IA para a forma e o foco que o usuário deseja, e há modelos e sequências de tarefas já estudados para isso (LIU et al., 2023).

Essas instruções, apresentadas como comandos estruturados na entrada de dados, ativam módulos nas redes neurais, promovendo a generalização de associações entre palavras, contextos e seus significados, o que constitui a definição aplicada de Machine Learning (LIU et al., 2023). A partir dessa capacidade, os LLMs podem realizar tarefas de extração com base em *prompts,* abrindo novas possibilidades para processar fontes complexas como os Currículos Lattes (BROWN et al, 2020).

**2. Materiais e métodos**

A metodologia aplicada para alcançar os objetivos deste projeto está dividida em sete etapas, que envolverá: **(1)** revisão bibliográfica sobre três eixos principais - (a) inteligência artificial e modelos de linguagem de grande escala (LLMs), com ênfase em suas aplicações para extração e estruturação de dados; (b) métodos de extração de informações de currículos acadêmicos, especialmente da Plataforma Lattes; e (c) gestão da informação sobre produção tecnológica em instituições de ensino e pesquisa; **(2)** busca de anterioridade de patentes de processo de extração de dados a partir do currículo lattes com o uso da plataforma *Orbit Intelligence*; **(3)** seleção e preparação dos dados; **(4)** desenvolvimento e refinamento de prompts para a extração dos dados; **(5)** aplicação do método de extração, validação dos dados; **(6)** análise crítica do processo e aplicação dos dados extraídos e **(7)** documentação do processo.

Como o projeto segue em andamento estão sendo realizados testes onde os Currículos Lattes nos formatos docx., HTML e PDF são submetidos ao ChatGPT através de técnicas de prompt engineering para instruir à identificação produções técnicas e tecnológicas como patentes, registros de softwares, protótipos, marcas, indicações geográficas e demais registros de proteções dentro do espectro de Propriedade Intelectual (PI).

Sob uma perspectiva doutrinária, a Propriedade Intelectual é comumente organizada em três categorias principais para fins de estudo e aplicação: o Direito Autoral, que protege as criações do intelecto humano como obras literárias, artísticas e programas de computador, incluindo os direitos do autor e direitos conexos; a Propriedade Industrial, que abrange elementos relacionados à atividade econômica e inovação, como marcas, patentes, desenhos industriais, indicações geográficas, segredos industriais e a repressão à concorrência desleal; e a Proteção Sui Generis, uma categoria específica para tutelar bens com características singulares que não se encaixam nas anteriores, como topografias de circuitos integrados, cultivares e conhecimentos tradicionais, justificando assim um regime jurídico próprio e adaptado às suas especificidades (JUNGMANN, 2010).

**3. Resultados e discussão**

Os resultados apresentados nesta seção são iniciais, uma vez que o projeto ainda se encontra em andamento. A etapa inicial focou em uma revisão bibliográfica sistemática, fundamental para embasar a pesquisa. Para esta atividade, as ferramentas de busca utilizadas incluíram o Portal de Periódicos CAPES, Scopus e Google Acadêmico, foram aplicados filtros como títulos e ano (2019 a 2025), abrangendo produções nacionais e internacionais. Os termos de busca incluíram "*Large Language Model*", "ChatGPT", "Inteligência Artificial", "Vitrine Tecnológica" e "Transferência de Tecnologia", combinados com operadores booleanos "AND" e "OR". A seleção dos artigos e dissertações priorizou títulos e resumos, resultando em 24 artigos nacionais, 8 artigos internacionais e 4 dissertações consideradas relevantes para esta fase. Todos os resultados foram agrupados no Mendeley Reference Manager.

O levantamento das produções tecnológicas se resumiu à localização do currículo lattes dos docentes e a testes que foram realizados, com apenas um currículo, para a extração das informações do currículo com o auxílio do LLM. Nesta fase já foram testados alguns *prompts*, pois o currículo lattes possui um vasto acervo de dados, sendo complexa a sua extração, e mesmo possuindo texto livre estruturado ele contém campos com descrições heterogêneas que demandam processamento adicional.

Existe, atualmente, o extrator lattes que é a ferramenta oficial do CNPq, regulamentada pela Resolução Normativa CNPq n. 01/2023, que permite que instituições realizem a extração de conjuntos de dados de currículos, entretanto, ele requer credenciamento institucional formal junto ao CNPq, e exige requisitos técnicos específicos - IP dedicado, desenvolvimento de solução para integração com o webservice. O extrator seria ideal para se alcançar esta fase, mas a ideia deste projeto é construir uma forma de extrair esses dados do lattes com o auxílio do *Large Language Model*, para que outras pessoas no futuro possam realizar o mesmo feito, principalmente, aquelas sem conhecimento de Programação possam pleitear tarefas inerentes de desenvolvedores de sistemas através podem do uso de LLM. Pois embora não existam registros na plataforma *Orbit Intelligence*, até o presente, de casos patenteados de aplicação direta de LLMs especificamente para extração de dados do lattes, há uma busca de exemplos análogos de uso de LLMs para extração de dados semi-estruturados que possam ser adaptados para este contexto.

Com o avanço da pesquisa, será avaliado o perfil das produções tecnológicas mapeadas, as possibilidades de colaboração institucional e a receptividade da comunidade acadêmica à proposta. Espera-se que o modelo desenvolvido possa ser replicado por outras instituições, promovendo o uso consciente e acessível da IA no meio acadêmico. A integração futura da vitrine tecnológica aos sites dos programas poderá ampliar sua funcionalidade como ferramenta permanente de divulgação e transferência de tecnologia.

## 4. Considerações finais

A origem desta proposta no estudo anterior de mapeamento temático dos TCCs do PROFNIT reforça a importância da continuidade metodológica e do aproveitamento incremental dos resultados de pesquisa. O apoio do GPT-4 foi essencial para operacionalizar a proposta, abrindo caminhos para novas aplicações de IA no ambiente acadêmico.

A expectativa é de que este método possa ser replicado em outras instituições e inspire iniciativas similares de sistematização e visualização da produção tecnológica, fomentando a transparência, a colaboração e a transferência de conhecimento no contexto da pós-graduação brasileira.

Neste sentido, este projeto busca também inspirar e encorajar pesquisadores, estudantes e profissionais que não atuam diretamente na área de programação a reconhecer o potencial transformador das ferramentas de Inteligência Artificial. Bem como, mostrar na prática, que é possível utilizar essas tecnologias avançadas de forma acessível e eficaz como instrumento de inovação, atualização e promoção do conhecimento, ampliando assim as perspectivas de desenvolvimento acadêmico e profissional de toda a comunidade envolvida.

## Agradecimentos

O apoio financeiro da Fundação de Apoio à Pesquisa do Estado da Paraíba (FAPESQ) está sendo crucial, pois viabiliza o acesso à ferramenta central da metodologia (GPT-4 em sua versão paga), garantindo a qualidade e a capacidade de processamento necessárias para a extração de dados.

## Referências

BROWN, T. B.; MANN, B.; RYDER, N.; SUBBIAH, M.; KAPLAN, J.; DHARIWAL, P.; ... & AMODEI, D. (2020). Language models are few-shot learners. Advances in neural information processing systems, 33, 1877-1901. Disponível em: https://papers.nips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf . Acesso em: 28 abr. 2025.

CABRAL, T. L. de O.; SILVA, F. C. da; PACHECO, A. S. V.; MELO, P. A. de . A CAPES E SUAS SETE DÉCADAS: trajetória da Pós-Graduação stricto sensu no Brasil. Revista Brasileira de Pós-Graduação, [S. l.], v. 16, n. 36, p. 1–22, 2020. DOI: 10.21713/rbpg.v16i36.1680. Disponível em: https://rbpg.capes.gov.br/rbpg/article/view/1680. Acesso em: 16 out. 2024.

CAPES. Plataforma Sucupira. Disponível em: https://sucupira.capes.gov.br/sucupira/. Acesso em: 16 out. 2024.

CHEN, Zheyi; XU, Liuchang; ZHENG, Hongting; CHEN, Luyao; TOLBA, Amr; ZHAO, Liang; YU, Keping; HAILIN, Feng . Evolution and Prospects of Foundation Models: From Large Language Models to Large Multimodal Models. Computers, Materials & Continua, v. 80, n. 2, 2024.Disponível em: https://www.sciencedirect.com/org/science/article/pii/S1546221824005472 . Acesso em: 10 mar.2025.DOI: https://doi.org/10.32604/cmc.2024.052618

JUNGMANN, D. M. Inovação e propriedade intelectual:guia para o docente. Brasília, DF: Senai, 2010. Disponível em: 93p.https://www.gov.br/inpi/pt-br/servicos/patentes/materiais-de-consulta-e-apoio/guia-para-o-docente.pdf. Acesso em: 13 jun. 2025.

KALYAN, Katikapalli Subramanyam. A survey of GPT-3 family large language models including ChatGPT and GPT-4. Natural Language Processing Journal, v. 6, p. 100048, 2024. Disponível em: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4593895 . Acesso em: 25 abr. 2025.

KASNECI, E,; SESSLER, K.; KÜCHEMANN, S.; BANNERT, M.; DEMENTIEVA, D.; FISCHER, F.; GASSER, U.; GROH, G.; GÜNNEMANN, S.; HÜLLERMEIER, E.; KRUSCHE, S.; KUTYNIOK, G.; MICHAELI, T.; NERDEL, C.; PFEFFER, J.; POQUET, O.; SAILER, M.; SCHMIDT, A.; SEIDEL, T.; STADLER, M.; & KASNECI, G.ChatGPT for good? On opportunities and challenges of large language models for education. Learning and Individual Differences, v. 103, p. 102274, 2023. Disponível em: https://www.sciencedirect.com/science/article/abs/pii/S1041608023000195. Acesso em: 13 jun. 2025.

LIU, Pengfei; WEIZHE, Yuan; Fu, Jinlan; JIANG, Zhengbao; HAYASHI, Hiroaki; NEUBIG, Graham. Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. ACM Computing Surveys, v. 55, n. 9, p. 1-35, 2023. Disponível em: https://dl.acm.org/doi/pdf/10.1145/3560815. Acesso em 13 jun.2025.

Programa de Pós-Graduação em Propriedade Intelectual e Transferência de Tecnologia para a Inovação. Disponível em: https://www.profnit.org.br. Acesso em: 24 ago. 2024

VELA, José Manuel Muñoz. Inteligência artificial generativa. Desafios para a propriedade intelectual. Revista de Direito da UNED (RDUNED) , n. 33, pág. 17-75, 2024. Disponível em: https://revistas.uned.es/index.php/RDUNED/article/view/41924/30456. Acesso em: 28 abr. 2025.