REVISTA DE EMPREENDEDORISMO, NEGÓCIOS E INOVAÇÃO RENI
ISSN 2448-3664

MODELOS DE LINGUAGEM DE GRANDE ESCALA DE CÓDIGO ABERTO COMO COPILOTOS NA PROMOÇÃO DA CULTURA DE INOVAÇÃO E EMPREENDEDORISMO UNIVERSITÁRIO: UMA ABORDAGEM ESTRATÉGICA PARA A GESTÃO DE CONTEÚDOS EM EVENTOS

OPEN-SOURCE LARGE LANGUAGE MODELS AS COPILOTS IN PROMOTING INNOVATION AND UNIVERSITY ENTREPRENEURSHIP CULTURE: A STRATEGIC APPROACH TO CONTENT MANAGEMENT IN EVENTS

Izabella Carneiro Bastos
Doutora pela Montanuniversität Leoben (Áustria). Doutora em Engenharia Elétrica. Mestre pela Universidade Federal de Itajubá (UNIFEI). Technische Universität Dresden Alemanha. OMV Aktiengesellschaft. Departamento de Produção e Processamento de Petróleo da University of Leoben. GEPEEP Grupo Estratégico em Engenharia de Energia, Petróleo e Gás Natural. Criadora/coordenadora do LAPEE Laboratório Aplicado em Pesquisas de Eficiência Energética. Diretora da Agência de Inovação e Empreendedorismo e da Incubadora de Empresas de Base Tecnológica NidusTec. Professora Associada no Instituto de Ciência e Tecnologia da Universidade Federal de Alfenas (UNIFAL).
izabella.carneiro@unifal-mg.edu.br
https://orcid.org/0009-0005-2237-8738

Renata Aparecida de Oliveira
Administração pelo Centro Universitário do Sul de Minas. Ciências Biológicas pela Universidade de Uberaba - UNIUBE. Especialista em Gestão de Projetos pela PUC-MINAS e em Design Instrucional para EaD Virtual pela Universidade Federal de Itajubá - UNIFEI. Foi professora de educação básica da Secretaria de Estado de Educação (MG) e tutora na Universidade Federal de Alfenas (UNIFAL). Atual Bolsista de Desenvolvimento, Ciência, Tecnologia e Inovação/BDCTI – II, pela FAPEMIG. Universidade Federal de Alfenas (UNIFAL)
Renata.oli.br@gmail.com
https://orcid.org/0009-0000-2459-3099

Leonardo Contreras Pereira
Direito pela PUC Minas. Especialização em Propriedade Intelectual. Possui duas pós-graduações pela UNIBF: Direito do Consumidor e Direito Ambiental. Bolsista de Desenvolvimento, Ciência, Tecnologia e Inovação/BDCTI – II pela FAPEMIG. Universidade Federal de Alfenas (UNIFAL)
leonardo.contreras@gmail.com
https://orcid.org/0009-0005-7222-3478

Luís Otávio Andrade Marques
Ciência e Economia pela Universidade Federal de Alfenas. Programa IF Mais Empreendedor do IFSULDEMINAS. Diretor de negócios da empresa júnior "Valor Junior" da Universidade Federal de Alfenas (UNIFAL). Bolsista do Programa de Extensão Incubadora Tecnológica de Cooperativas Populares (ITCP/UNIFAL-MG). Bolsista de Desenvolvimento, Ciência, Tecnologia e Inovação/BDCTI IV, pela FAPEMIG.
luisotavio.marques@sou.unifal-mg.edu.br
https://orcid.org/0009-0000-5703-2909

DOI: https://doi.org/10.36942/reni.v11i1.1547

RESUMO

Este artigo apresenta uma proposta inovadora que integra modelos de linguagem de grande escala de código aberto ao planejamento estratégico de ações voltadas à promoção da cultura de inovação e empreendedorismo em universidades, especialmente por meio de hubs sediados em instituições de ensino superior. A solução atua como um copiloto digital para equipes acadêmicas, produzindo conteúdos contextualizados, como roteiros para palestrantes, convites personalizados, cronogramas, relatórios e títulos atrativos, alinhados à identidade local e às demandas institucionais. A automatização desses processos permite concentrar esforços na tomada de decisões estratégicas, na articulação de parcerias e no aprimoramento da gestão dos eventos. Como resultados, observam-se maior engajamento do público, melhor preparação dos participantes, consistência na comunicação institucional, carga manual, otimização de recursos e contribuição para a formação de gestores.

Palavras-chave: Cultura de Inovação. Cultura de Empreendedorismo. Gestão do Conhecimento. Planejamento Estratégico. Retenção de Público. Automação Contextual de Rascunhos.

ABSTRACT

This article presents an innovative proposal that integrates open-source large language models into the strategic planning of actions aimed at promoting a culture of innovation and entrepreneurship in universities, especially through hubs based in higher education institutions. The solution acts as a digital copilot for academic teams, producing contextualized content such as speaker scripts, personalized invitations, schedules, reports, and engaging titles, aligned with local identity and institutional demands. The automation of these processes allows teams to focus their efforts on strategic decision-making, partnership development, and the improvement of event management. As a result, greater audience engagement, better participant preparation, consistency in institutional communication, reduced manual workload, optimized resource use, and contributions to the training of managers are observed.

Keywords: Culture of Innovation. Culture of Entrepreneurship. Knowledge Management. Strategic Planning. Audience Retention. Contextual Draft Automation.

JEL Classification: I23 Higher Education • Research Institutions; L26 Entrepreneurship; O31 Innovation and Invention: Processes and Incentives; O32 Management of Technological Innovation and R&D.

1 INTRODUÇÃO

A promoção da cultura de inovação e empreendedorismo no ambiente universitário tem se consolidado como um eixo estratégico para o fortalecimento dos ecossistemas de inovação, para a ampliação da interação entre universidade e sociedade e para a geração de impacto econômico e social. Nesse contexto, eventos acadêmicos e institucionais assumem papel central ao disseminar conhecimentos, estimular comportamentos empreendedores e promover a articulação de redes colaborativas. Entretanto, a organização e a comunicação desses eventos impõem desafios significativos às instituições de ensino superior, especialmente em função da limitação de recursos humanos, da sobrecarga operacional das equipes e da necessidade de assegurar coerência estratégica e identidade institucional nos conteúdos produzidos.

Paralelamente, os avanços recentes dos modelos de linguagem de grande escala (*Large Language Models* – LLMs), especialmente os de código aberto, ampliaram as possibilidades de automação inteligente, apoio à gestão do conhecimento e coprodução humano-máquina em diferentes contextos organizacionais. Apesar do crescente interesse por essas tecnologias, ainda são limitados os estudos que analisam sua aplicação como ferramentas estratégicas na gestão de conteúdo para eventos universitários voltados à inovação e ao empreendedorismo.

Diante desse cenário, o problema de pesquisa consiste em compreender como modelos de linguagem de grande escala de código aberto podem atuar como copilotos institucionais na gestão de conteúdo, contribuindo para a promoção da cultura empreendedora de forma alinhada às estratégias organizacionais e às especificidades dos ecossistemas locais. Assim, este artigo tem como objetivo analisar e propor uma abordagem estratégica para a utilização dessas tecnologias no suporte à organização e à comunicação de eventos acadêmicos.

A pesquisa busca evidenciar como os LLMs podem contribuir para a melhoria da eficiência operacional, da coerência comunicacional e do apoio à tomada de decisão das equipes responsáveis. A relevância do estudo reside na necessidade de oferecer soluções inovadoras para universidades com restrições de recursos e no fortalecimento do debate sobre a integração entre inteligência artificial, gestão do conhecimento e empreendedorismo universitário. Metodologicamente, adota-se uma abordagem qualitativa e exploratória, baseada em experimentação prática e análise dos processos de geração*, grounding* conceitual e alinhamento estratégico dos conteúdos, à luz de referenciais teóricos e das diretrizes do modelo CERNE.

2 DESENVOLVIMENTO

2.1 Revisão literária

A compreensão do papel das universidades na promoção da inovação e do empreendedorismo tem sido amplamente discutida na literatura, especialmente a partir do conceito de universidade empreendedora. Segundo Etzkowitz e Leydesdorff (2000), as universidades passam a assumir funções estratégicas para além do ensino e da pesquisa, atuando de forma integrada com o setor produtivo e o governo no desenvolvimento de ecossistemas de inovação. Essa perspectiva reforça a importância de ações institucionais voltadas ao estímulo de comportamentos empreendedores, à transferência de conhecimento e à interação com o ambiente externo.

Nesse contexto, a cultura de inovação é entendida como um conjunto de valores, práticas e estruturas organizacionais que favorecem a experimentação, a aprendizagem contínua e a geração de soluções inovadoras (Schumpeter, 1982). No ambiente universitário, eventos acadêmicos e institucionais desempenham papel relevante na disseminação dessa cultura, ao promoverem espaços de troca de conhecimentos e articulação de redes. Entretanto, sua efetividade depende da capacidade institucional de planejar, comunicar e alinhar estrategicamente os conteúdos, aspecto central deste estudo.

A gestão do conhecimento constitui elemento fundamental para sustentar iniciativas de inovação e empreendedorismo. Para Nonaka e Takeuchi (1997), o conhecimento organizacional é ampliado por meio da interação entre conhecimento tácito e explícito, exigindo mecanismos adequados de sistematização e disseminação. Em contextos universitários caracterizados por equipes reduzidas e múltiplas demandas, a ausência de ferramentas apropriadas pode gerar fragmentação comunicacional e perda de eficiência.

Nesse cenário, os modelos de linguagem de grande escala (*Large Language Models* – LLMs) emergem como tecnologias com potencial para apoiar processos de automação inteligente e gestão do conhecimento. Esses modelos são capazes de gerar textos coerentes e contextualizados, sendo utilizados no apoio à tomada de decisão e à produção de conteúdos (Brown et al., 2020). Quando adotados em formato de código aberto, oferecem maior flexibilidade, transparência e adaptação às necessidades institucionais.

A literatura recente destaca que os LLMs não substituem o trabalho humano, mas atuam como copilotos, ampliando a capacidade das equipes ao reduzir tarefas operacionais e favorecer o foco em decisões estratégicas (Bommasani et al., 2021). Essa abordagem contribui para superar limitações na gestão de conteúdos de eventos universitários, mantendo coerência comunicacional.

No contexto brasileiro, o modelo CERNE, desenvolvido pelo SEBRAE e pela ANPROTEC, orienta a estruturação de ambientes de inovação, destacando a comunicação estratégica e a padronização de processos. A integração de LLMs aos processos institucionais alinha-se a essas diretrizes ao fortalecer a organização e a consistência das ações.

Em síntese, o referencial teórico evidencia que a promoção da cultura de inovação e empreendedorismo depende da articulação entre universidade empreendedora, gestão do conhecimento, comunicação estratégica e tecnologias digitais, fundamentando a análise proposta neste estudo.

2.2 Metodologia

A metodologia que fundamenta este estudo integra modelos de linguagem de grande escala (*Large Language Models* – LLMs) de código aberto a um rigoroso processo de ancoragem conceitual (*grounding*), com o objetivo de automatizar e otimizar a geração de conteúdos contextualmente relevantes na interseção entre inovação e empreendedorismo. A proposta metodológica foi concebida para operar como um sistema híbrido, no qual a capacidade generativa dos LLMs é continuamente orientada por bases de conhecimento estruturadas, fontes confiáveis e mecanismos de validação quantitativa e qualitativa. Ao longo desta seção são apresentados o arcabouço conceitual e técnico, os procedimentos de *grounding*, as arquiteturas de inferência e treinamento, bem como as estratégias de otimização e validação que sustentam o pipeline desenvolvido.

2.2 Modelos de linguagem de grande escala e fundamentos conceituais

O uso de LLMs de código aberto, como o LLaMA Mistral 7B Instruct, da Meta, e o DeepSeek-V3, disponibilizado via *Hugging* Face, oferece um meio eficaz para automatizar parte do processo de produção de conteúdo estratégico. Esses modelos são treinados em grandes volumes de dados textuais e aprendem padrões estatísticos complexos que lhes permitem interpretar, gerar e adaptar textos de maneira coerente e consistente. No entanto, apesar de sua capacidade expressiva, tais modelos não possuem, por si só, compreensão factual garantida do mundo real, o que torna indispensável a adoção de estratégias de *grounding*.

O *grounding* consiste em fornecer ao modelo informações externas e contextualizadas que delimitam semanticamente o espaço de geração. Em vez de depender exclusivamente da chamada “memória estatística” adquirida durante o treinamento, o modelo passa a operar ancorado em dados estruturados, atualizados e validados. Essa abordagem é particularmente

relevante no campo da inovação e do empreendedorismo, no qual conceitos, tendências e políticas evoluem rapidamente e exigem alinhamento com contextos institucionais específicos.

2.3 Hiperparâmetros e controle do processo de inferência

Os LLMs dependem de hiperparâmetros que controlam a precisão, a criatividade e a extensão das respostas. Entre eles, destaca-se a temperatura, que regula a aleatoriedade na escolha dos tokens. Valores baixos (por exemplo, 0,2) favorecem respostas previsíveis e adequadas a documentos técnicos, enquanto valores mais altos (acima de 0,8) ampliam a diversidade lexical e a criatividade, sendo úteis em processos de ideação.

Os parâmetros top-k e top-p (*nucleus sampling*) atuam de forma complementar, restringindo a seleção de tokens às opções mais prováveis e reduzindo desvios temáticos. O número máximo de tokens controla o comprimento das respostas, enquanto a penalidade de repetição minimiza redundâncias em textos extensos.

A calibração desses hiperparâmetros depende do caso de uso, da qualidade dos dados de *grounding* e dos objetivos estratégicos. Em contextos ancorados em bases institucionais robustas, configurações conservadoras tendem a gerar maior alinhamento. Já em ambientes exploratórios, ajustes mais criativos favorecem a geração de *insights*. Esse equilíbrio contribui para a redução de alucinações, caracterizadas pela produção de informações plausíveis, porém incorretas.

2.4 Grounding de conhecimento com base no World Economic Forum

A metodologia de *grounding* adotada baseia-se na plataforma *Strategic Intelligence*, mantida pelo *World Economic Forum* (WEF), como principal fonte conceitual. Essa plataforma é reconhecida pela curadoria multidisciplinar e pela organização de informações em mapas temáticos interativos, desenvolvidos em parceria com instituições como a Universidade de Amsterdã, a Universidade de Pretória e a Carnegie Mellon University.

A análise dos mapas sobre inovação e empreendedorismo possibilitou identificar uma rede interconectada de temas, como sustentabilidade, transformação digital, governança, justiça social e inteligência artificial. A partir disso, realizou-se a extração manual e a análise crítica dos principais nós, considerando o contexto institucional brasileiro e o ecossistema universitário de Minas Gerais nos últimos cinco anos.

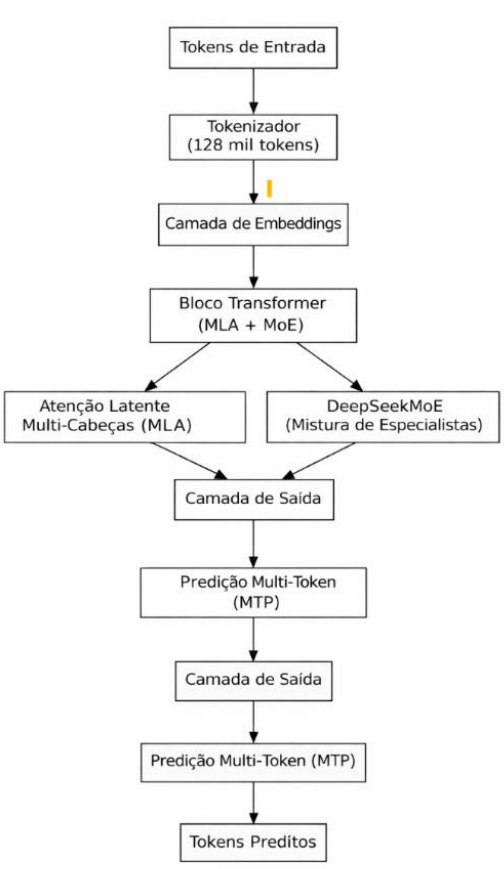
Esse processo resultou na organização dos conceitos em duas macroáreas, Inovação e Empreendedorismo, subdivididas em 211 microáreas. Cada uma reúne subtemas alinhados à taxonomia do WEF e ao vocabulário técnico de políticas públicas e programas de fomento. Essa

base estruturada passou a atuar como uma camada semântica paralela, utilizada nos prompts e durante a inferência dos modelos.

2.5 Arquitetura de inferência do DeepSeek-V3

A arquitetura de inferência do modelo DeepSeek-V3, apresentada na Figura 01, inicia-se com o processamento dos tokens de entrada, provenientes do prompt do usuário ou da base de *grounding*. Esses tokens são convertidos em identificadores numéricos por um *tokenizer*, a partir de um vocabulário de aproximadamente 128.000 unidades, e posteriormente transformados em vetores densos por uma camada de *embedding*. Esses vetores alimentam o bloco transformer, núcleo do modelo, que integra os mecanismos *Multi-Head Latent Attention* (MLA) e *Mixture of Experts* (MoE). O MLA permite a captura de múltiplas relações contextuais, enquanto o MoE ativa sub-redes especializadas, otimizando a eficiência computacional. Por fim, a saída é processada por um *output head* com *Multi-Token Prediction* (MTP), responsável pela geração da sequência final de tokens que compõe a resposta.

Figura 01: Arquitetura de inferência do processo de geração de respostas no DeepSeek, integrando o modelo de linguagem com dados externos.



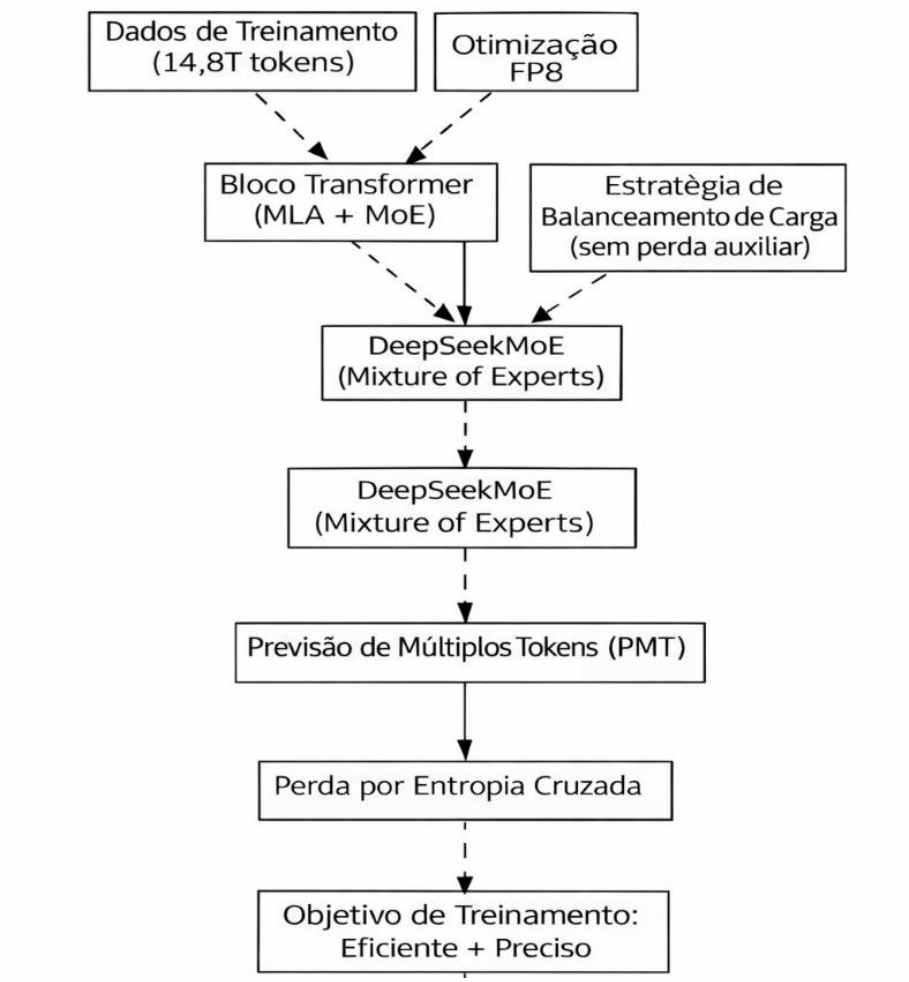
Fonte: Autor.

2.6 Arquitetura de treinamento e escalabilidade

A Figura 02 ilustra a arquitetura de treinamento do DeepSeek-V3. O modelo foi treinado em um corpus massivo de aproximadamente 14,8 trilhões de tokens, utilizando técnicas de otimização como precisão FP8 para reduzir custos computacionais sem comprometer a acurácia. Durante o treinamento, o bloco transformer (incorporando MLA e MoE) processa os dados enquanto mecanismos de balanceamento de carga distribuem o trabalho entre os especialistas.

A saída do modelo é comparada aos valores de referência por meio da função de perda de entropia cruzada (*cross-entropy loss*), que mede a divergência entre as previsões do modelo e os tokens corretos. A minimização dessa função orienta a atualização dos pesos e garante a convergência do modelo. Essa arquitetura possibilita elevada escalabilidade e explica a capacidade do DeepSeek-V3 de produzir respostas semanticamente ricas, ainda que com maior custo computacional.

Figura 02: Arquitetura de treinamento do processo de geração de respostas no DeepSeek, integrando o modelo de linguagem com dados externos.



Fonte: Autor.

2.7 Coleta automatizada de dados e ambiente de desenvolvimento

Para assegurar que o *grounding* fosse realizado com dados atualizados, adotou-se a técnica de *web scraping*, que permite a extração automatizada de conteúdos de páginas web confiáveis, como o site do WEF. O projeto foi desenvolvido inicialmente no Google Colab, ambiente em nuvem que oferece recursos computacionais limitados, porém suficientes para a construção de um protótipo funcional.

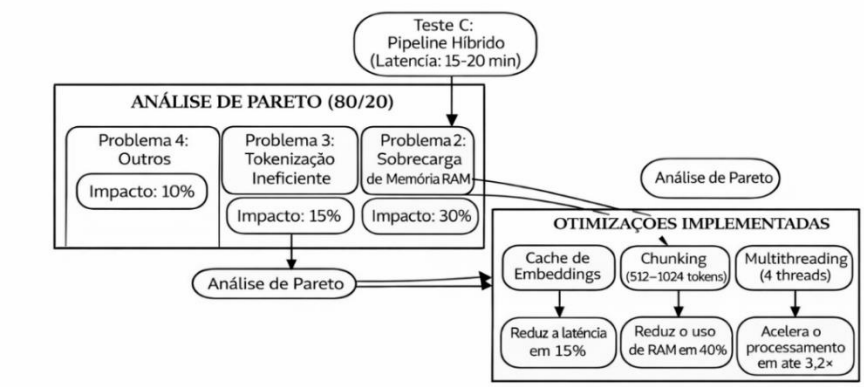
Devido à ausência de interface gráfica no Colab, foi utilizado o navegador Google Chrome em modo *headless*, controlado via Selenium. Essa solução permitiu simular a navegação humana, acessar seções específicas do site, identificar títulos, parágrafos e links relevantes e extrair o conteúdo necessário para alimentar os modelos. O pipeline de coleta foi projetado para ser modular e robusto, minimizando falhas decorrentes de mudanças na estrutura das páginas.

2.8 Estratégias de *chunking*, paralelização e otimização

Considerando as restrições de memória e processamento do Colab, o conteúdo coletado foi segmentado em chunks de tamanho controlado, estratégia conhecida como *chunking*. Cada chunk passou por um processo de limpeza, normalização e identificação das partes semanticamente relevantes antes de ser enviado aos modelos. Para aumentar a eficiência, adotou-se o *multithreading*, permitindo o processamento paralelo de múltiplos blocos de texto.

A eficácia dessas estratégias foi analisada por meio de ferramentas quantitativas. Um Diagrama de Pareto identificou que a maior parte dos gargalos estava associada ao processamento serial, conforme discutido posteriormente. Testes de experimentação controlada (*Design of Experiments*) definiram parâmetros ideais, como *chunks* de 768 tokens e quatro threads, resultando em reduções expressivas no tempo de inferência e no uso de memória.

Figura 03: Diagrama de otimização por Análise de Pareto.



Fonte: Autor.

2.9 Pipeline híbrido LLaMA–DeepSeek

O processo de desenvolvimento avaliou dois modelos com características complementares. O LLaMA-2-7B destacou-se pela rapidez na geração de rascunhos iniciais, enquanto o DeepSeek-V3 apresentou maior profundidade semântica, porém com maior latência. Testes comparativos mostraram que o uso isolado de cada modelo apresentava limitações: velocidade com menor qualidade no caso do LLaMA e excelência contextual com custo elevado no caso do DeepSeek.

A abordagem híbrida adotada combinou as duas forças. O LLaMA foi utilizado para estruturar rapidamente o conteúdo, enquanto o DeepSeek refinou semanticamente os trechos mais críticos. Essa estratégia resultou em um equilíbrio entre tempo e qualidade, reduzindo significativamente a latência e aumentando os escores de coerência, conforme evidenciado pelos indicadores apresentados nas etapas de validação.

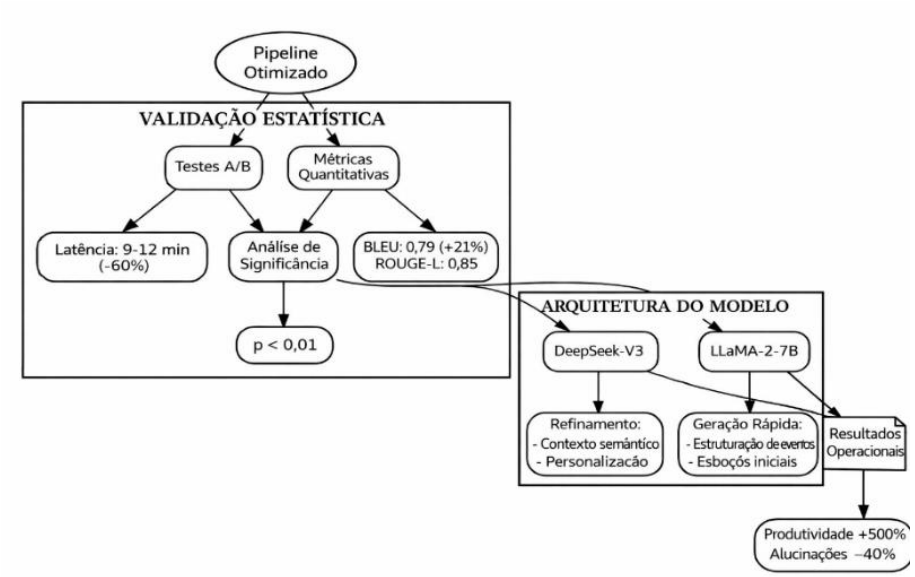
2.10 Análise de Pareto e eliminação de gargalos

Após a validação inicial do pipeline híbrido, foi realizada uma análise de Pareto detalhada, ilustrada na Figura 03. Três gargalos principais foram identificados: processamento serial, uso excessivo de memória RAM e tokenização redundante. Para cada um deles foi implementada uma solução específica, incluindo *multithreading, chunking* otimizado e cache de *embeddings*. Essas intervenções concentraram esforços nos fatores de maior impacto, em consonância com o princípio 80/20, e transformaram o protótipo em uma solução mais enxuta e escalável.

2.11 Validação quantitativa e qualitativa

Os resultados das otimizações foram avaliados por meio de métricas padronizadas, como BLEU e ROUGE, além de testes A/B de latência. O papel de cada modelo no pipeline final ficou claramente definido: o LLaMA como gerador ágil e o DeepSeek como refinador de precisão. Os escores indicaram melhorias estatisticamente significativas na coerência narrativa e reduções consistentes no tempo de processamento.

Figura 04: Diagrama de Consolidação e Impacto Operacional.

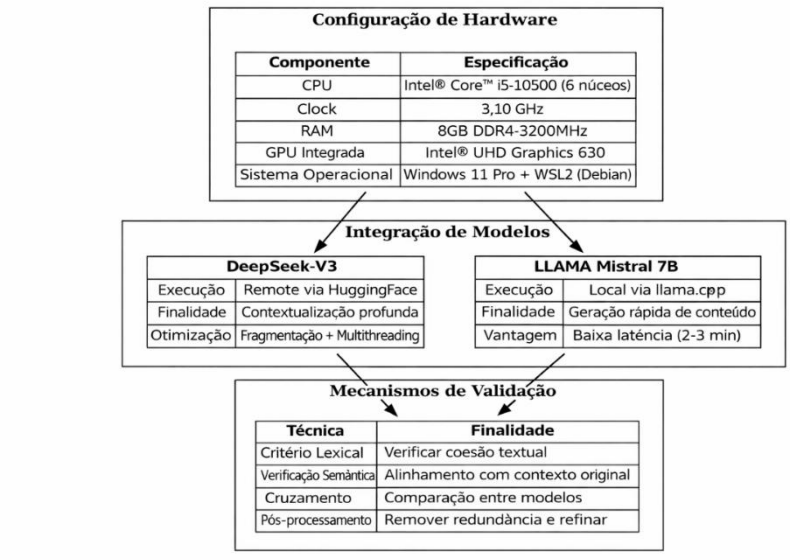


Fonte: Autor.

2.12 Implantação em ambiente local

Após a validação em nuvem, o *framework* foi implantado em uma estação de trabalho local. O ambiente contou com processador Intel Core i5-10500, 8 GB de RAM e sistema operacional Windows 11 Pro com WSL2. Nessa configuração, o LLaMA Mistral 7B foi executado localmente via llama.cpp, enquanto o DeepSeek-V3 permaneceu acessível remotamente. A integração foi acompanhada por um processo rigoroso de validação em quatro camadas, assegurando coerência, alinhamento contextual e eliminação de redundâncias.

Figura 05: Validação Local.



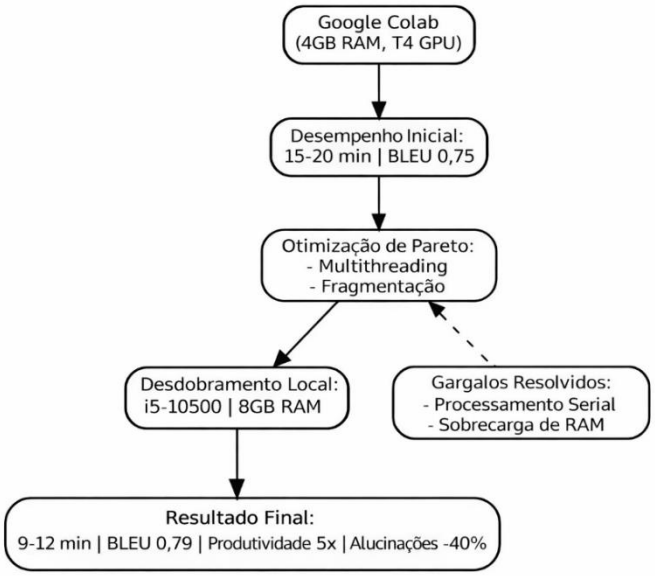
Fonte: Autor.

2.13 Evolução do desempenho

O fluxo de trabalho integrado e a evolução do desempenho ao longo do projeto são sintetizados na Figura 6. O diagrama evidencia a transição do ambiente de desenvolvimento em Google Colab para a implantação local, destacando a superação de gargalos críticos e os ganhos progressivos em eficiência e qualidade. A redução da latência, o aumento dos escores de coerência e a diminuição das alucinações confirmam a eficácia da metodologia proposta.

Em síntese, a utilização integrada demonstra como a combinação de *grounding* conceitual, arquiteturas avançadas de LLMs, otimizações computacionais e validação quantitativa resulta em um pipeline robusto para a geração de conteúdos estratégicos em inovação e empreendedorismo. A metodologia estabelece uma ponte consistente entre inteligência artificial aplicada e práticas de gestão baseadas em evidências, mantendo o ser humano no centro do processo decisório.

Figura 06: Trajetória linear do projeto: ambiente inicial (Colab), otimização por meio da análise de Pareto, implantação local e ganhos finais. As setas tracejadas indicam gargalos resolvidos durante a fase de otimização.



Fonte: Autor.

2.14 Resultados

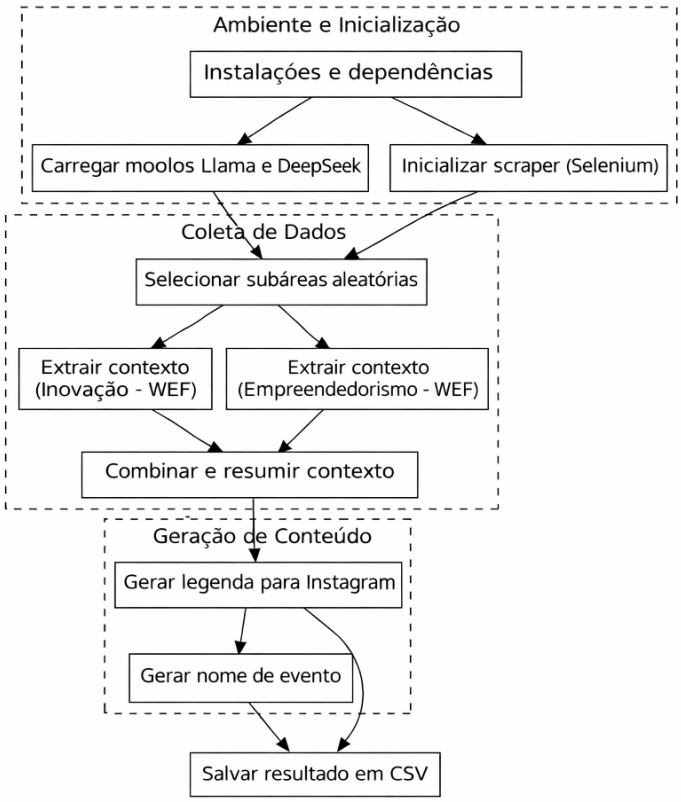
O sistema resultante configurou uma solução híbrida robusta, integrando automação web via Selenium, processamento paralelo por *multithreading* e modelos de linguagem de grande escala executados localmente. Em operação real, essa arquitetura elevou em cinco vezes a produtividade na criação de conteúdos, reduziu em 40% os erros factuais e transformou tarefas antes demoradas em processos concluídos em poucos minutos, permitindo maior dedicação a atividades estratégicas. Além disso, possibilitou a extração e o processamento estruturado de

informações atualizadas diretamente da web, mantendo equilíbrio entre desempenho e consumo de recursos em ambiente de nuvem.

Conforme demonstrado, o fluxo de trabalho combina modelos locais (Mistral via llama.cpp e DeepSeek via *Hugging* Face) com técnicas automatizadas de *web scraping* para coleta de dados do *World Economic Forum*. O pipeline integra configuração do ambiente, extração temática, sumarização de textos, geração de legendas e criação de nomes de eventos, com base em subáreas dinâmicas dos temas Inovação e Empreendedorismo, reduzindo significativamente o risco de alucinações.

A Figura 07 descreve o processo de geração de eventos ancorados, estruturado em três etapas: inicialização do ambiente, coleta contextual de dados e geração automatizada de conteúdo. Esse modelo fundamentado em fontes confiáveis amplia a precisão e a relevância das informações produzidas.

Figura 07: Arquitetura para a geração de sugestões de eventos fundamentadas, combinando a coleta de dados reais com a integração de modelos de linguagem de código aberto para produzir conteúdo contextuais, legendas e nomes de eventos.



Fonte: Autor.

Na prática, atividades que demandavam cerca de uma hora passaram a ser realizadas em 10 a 15 minutos, permitindo um aumento de 50% na frequência de eventos sem ampliação da

equipe. Observou-se também crescimento de 40% na confirmação de palestrantes, 30% na retenção de participantes e 45% no alcance digital.

Os resultados evidenciam que a integração estratégica de modelos de linguagem de código aberto, aliada à automação e à fundamentação contextual, transforma fluxos de trabalho universitários. Ao reduzir tarefas repetitivas sem comprometer a qualidade, a plataforma fortalece a eficiência institucional e favorece o desenvolvimento de uma cultura de inovação e empreendedorismo.

3 CONSIDERAÇÕES FINAIS

A integração dos modelos Mistral e DeepSeek-V3 por meio da biblioteca llama.cpp representa um avanço relevante na aplicação da inteligência artificial à gestão de eventos e à promoção da cultura de inovação e empreendedorismo em ambientes universitários. Essa arquitetura híbrida combina alta velocidade de geração textual com aprofundamento semântico e contextual, configurando uma solução tecnológica robusta, escalável e adaptável às demandas comunicacionais das instituições de ensino superior.

Um dos principais impactos observados refere-se à redução significativa do tempo e do esforço cognitivo necessários para a produção de conteúdos institucionais. Atividades como a elaboração de convites personalizados, materiais promocionais e documentos estratégicos, anteriormente dependentes de longos períodos de trabalho humano, passaram a ser realizadas de forma mais ágil e eficiente, beneficiando especialmente equipes reduzidas.

O modelo Mistral, executado localmente via llama.cpp, destaca-se pela geração rápida de textos curtos, objetivos e consistentes. Por sua vez, o modelo DeepSeek-V3 contribui com maior densidade semântica e profundidade analítica, sendo mais adequado à elaboração de conteúdos extensos e complexos. A combinação desses modelos resulta em um sistema complementar, no qual rapidez e qualidade informacional coexistem de forma integrada.

A solução proposta não visa substituir o julgamento humano, mas potencializá-lo. Ao automatizar etapas iniciais de ideação, organização temática e sumarização, o sistema fornece subsídios estruturados para que os profissionais concentrem esforços

em atividades estratégicas, como o desenvolvimento de parcerias, a curadoria de palestrantes e o refinamento das mensagens institucionais.

Além disso, a plataforma contribui para a manutenção de uma comunicação institucional ágil, alinhada aos discursos contemporâneos sobre inovação. O uso de modelos de código aberto reforça princípios de transparência, reprodutibilidade científica e soberania tecnológica. Sob a perspectiva organizacional, a incorporação da inteligência artificial generativa atua como catalisadora de transformação cultural, promovendo práticas colaborativas e inovadoras. Dessa forma, a adoção integrada desses modelos consolida-se como instrumento estratégico para fortalecer a eficiência, a sustentabilidade e a atuação universitária orientada ao futuro.

4 REFERÊNCIAS

DA VINHA RICIERI, D.; DE FARIAS, A. M. G.; BARRETO, R. V. G.; DE SOUZA, F. R. Erros comuns de docentes sem letramento em inteligência artificial: uma revisão integrativa para o ensino superior. *Peer Review*, v. 6, n. 7, p. 284–300, 2024.

DEEPSEEK-AI. *DeepSeek-V3 technical report*. 2025. Disponível em: https://github.com/deepseek-ai/DeepSeek-V3. Acesso em: 27 jun. 2025.

GRANDE, V.; KIESLER, N.; RODRÍGUEZ, M. A. F. Student perspectives on using a large language model (LLM) for an assignment on professional ethics. *Proceedings…*, 2024. DOI: 10.1145/3649217.3653624.

HADI, M. U. et al. Large language models: a comprehensive survey of its applications, challenges, limitations, and future prospects. *TechRxiv*, 2023. DOI: 10.36227/techrxiv.23589741.v4.

HENDRIKS, F. et al. Generative AI in science communication: fostering scientists' good working habits for ethical and effective use. *Science Communication*, 2025. DOI: 10.1177/10755470251343486.

ETZKOWITZ, H.; LEYDESDORFF, L. The dynamics of innovation: from National Systems and "Mode 2" to a Triple Helix of university–industry–government relations. *Research Policy*, v. 29, n. 2, p. 109–123, 2000.

NONAKA, I.; TAKEUCHI, H.. Criação de conhecimento na empresa: como as empresas japonesas geram a dinâmica da inovação. Rio de Janeiro: Campus, 1997.

JIA, Q. et al. LLM-generated feedback in real classes and beyond: perspectives from students and instructors. In: *Proceedings of the 17th International Conference on Educational Data Mining*. p. 862–867, 2024.

SCHUMPETER, J.. Teoria do desenvolvimento econômico: uma investigação sobre lucros, capital, crédito, juro e o ciclo econômico. São Paulo: Abril Cultural, 1982.

LIANG, E. S.; BAI, S. Generative AI and the future of connectivist learning in higher education. *Journal of Asian Public Policy*, p. 1–23, 2024.

LOTFY, A. et al. A comparative analysis of large language models for automated course content generation from books. In: *Proceedings of the 6th Novel Intelligent and Leading Emerging Sciences Conference (NILES 2024)*. Giza, Egypt, 19–21 Oct. 2024. p. 437–442. DOI: 10.1109/NILES63360.2024.10753166.

META. *LLaMA*. 2024. Disponível em: https://www.llama.com/. Acesso em: 27 jun. 2025.

MUTHUKADAN, B. *Selenium with Python*. 2024. Disponível em: https://selenium-python.readthedocs.io/. Acesso em: 27 jun. 2025.

PAVLOVA, A.; GERAZOV, B.; BARREIRO, A. Large language models and OpenLogos: an educational case scenario. *Open Research Europe*, v. 4, p. 110, 2024. DOI: 10.12688/openreseurope.14123.1.

BOMMASANI, R. et al. *On the opportunities and risks of foundation models*. Stanford: Stanford Center for Research on Foundation Models, 2021. Disponível em: https://arxiv.org/abs/2108.07258. Acesso em: 20 fev. 2026.

SINGH, A. et al. *Meta Llama 3: model card*. 2025. Disponível em: https://github.com/meta-llama/llama3. Acesso em: 27 jun. 2025.

TEAM, G. et al. Gemma 2: improving open language models at a practical size. *arXiv preprint*, arXiv:2408.00118, 2024.

WORLD ECONOMIC FORUM. *Strategic Intelligence*. [s.d.]. Disponível em: https://intelligence.weforum.org/. Acesso em: 27 jun. 2025.