Enriquecimento Semântico de Repositórios Acadêmicos de Computação Utilizando LLM

Mirelly Barabosa Fontes de Oliveira[1], **Valéria Silva Santos**[1], **Breno Santana Santos**[2], **Marianne Batista Diniz da Silva**[1]

[1]Universidade Federal de Alagoas (UFAL) – AL – Brasil

[2]Universidade Federal do Rio Grande do Norte (UFRN) – RN – Brasil

{mirelly.oliveira, valeria.silva}@arapiraca.ufal.br

breno1005@hotmail.com, marianne.silva@penedo.ufal.br

Abstract. *The expansion of higher education in Brazil has boosted academic production, generating increasing challenges for cataloging and retrieving information in institutional repositories. Undergraduate theses (TCCs) are deposited with degraded, incomplete, or generic metadata, drastically reducing the visibility of scientific output. Although manual correction is impractical due to the volume of documents, the use of Large Language Models (LLMs) emerges as an alternative for automated semantic processing. The objective of this work is to investigate the feasibility of LLMs in enriching digital collections through an automated pipeline based on the GPT-4o-mini model. The system was applied to a historical database of Computer Science TCCs from the Federal University of Alagoas (UFAL) and manually evaluated by a specialist on a sample of 50 works across three conceptual dimensions. The results indicate high performance, with an average of 4.80 in relevance and scores equal to or greater than 4.47 in the dimensions of specificity and search utility. It is concluded that the proposed approach constitutes a viable and low-cost alternative for the semantic standardization of institutional undergraduate collections, with the potential for direct integration into repository management systems.*
Keywords: *Semantic enrichment, Large Language Models, Institutional repositories, Metadata extraction, Natural Language Processing.*

Resumo. *A expansão do ensino superior no Brasil impulsionou a produção acadêmica, gerando desafios crescentes para a catalogação e recuperação de informações em repositórios institucionais. Os Trabalhos de Conclusão de Curso (TCCs) são depositados com metadados degradados, incompletos ou genéricos, o que reduz drasticamente a visibilidade da produção científica. Embora a correção manual seja inviável devido ao volume de documentos, o uso de Grandes Modelos de Linguagem (LLMs) surge como uma alternativa para o processamento semântico automatizado. O objetivo deste trabalho é investigar a viabilidade de LLMs no enriquecimento de acervos digitais por meio de um pipeline automatizado baseado no modelo gpt-4o-mini. O sistema foi aplicado em uma base histórica de TCCs de Computação da Universidade Federal de Alagoas (UFAL) e avaliado manualmente por um especialista em uma amostra de 50 trabalhos sob três dimensões conceituais. Os resultados indicam desempenho elevado, com média de 4,80 em relevância e escores iguais ou superiores*

a 4,47 nas dimensões de especificidade e utilidade de busca. Conclui-se que a abordagem proposta constitui uma alternativa viável e de baixo custo para a padronização semântica de acervos institucionais de graduação, com potencial de integração direta a sistemas de gerenciamento de repositórios.
Palavras-chave: *Enriquecimento semântico, Grandes Modelos de Linguagem, Repositórios institucionais, Extração de metadados, Processamento de Linguagem Natural.*

1. Introdução

A expansão do ensino superior brasileiro impulsionou o volume de Trabalhos de Conclusão de Curso (TCCs) depositados em repositórios institucionais, intensificando os desafios de catalogação e recuperação da informação [INEP 2024]. Frequentemente, esses documentos são indexados com metadados incompletos ou semanticamente imprecisos, comprometendo sua descoberta por sistemas de busca [Choudhury et al. 2023]. As palavras-chave exercem papel importante nos sistemas de busca, funcionando como termos de indexação que conectam documentos às consultas dos usuários, uma vez que palavras-chave mal definidas reduzem a precisão dos sistemas de recuperação por similaridade semântica. No entanto, a definição inadequada desses termos pelos próprios autores, que frequentemente escolhem palavras genéricas ou pouco representativas, reduz a precisão das buscas e compromete a visibilidade da produção científica [da Silva 2025, dos Santos e Eliseo Berni Reategui 2025].

Embora estudos indiquem que o refinamento de descritores e o mapeamento por redes de coocorrência melhorem a eficiência de recuperação [Lamba et al. 2025], o processo de catalogação e correção manual de grandes acervos é inviável dado o grande volume de documentos e à limitação de recursos humanos nas bibliotecas universitárias. Nesse contexto, modelos de linguagem de grande escala (LLMs) apresentam-se como alternativa ao processamento manual, dada sua capacidade de inferência contextual sobre texto não estruturado [Formánek 2024]. Modelos fundacionais demonstram desempenho superior aos métodos clássicos de PLN em tarefas de classificação e extração de informação [Zhao et al. 2023] justificando a investigação de pipelines automatizados para mitigar a degradação de metadados em repositórios digitais.

A resolução desse problema apresenta reflexos diretos no ecossistema de iniciação científica e na inserção de jovens pesquisadores no âmbito acadêmico [Braz 2024]. Ao prover um ambiente com metadados semanticamente ricos e padronizados, reduz-se o esforço rotineiro de levantamento bibliográfico e refinamento conceitual, permitindo que estudantes concentrem-se no desenvolvimento do pensamento crítico e na integridade da pesquisa [Braz 2024, dos Santos e Eliseo Berni Reategui 2025]. No entanto, a literatura ainda carece de investigações focadas no enriquecimento semântico de repositórios institucionais de graduação, sobretudo no domínio da Computação, lacuna que motiva a abordagem proposta neste estudo [Medina Nieto et al. 2020].

Deste modo, o objetivo deste trabalho é investigar como os LLMs podem apoiar a geração e a padronização de metadados semânticos em repositórios acadêmicos de computação, por meio do desenvolvimento e da avaliação de um pipeline automatizado baseado no modelo gpt-4o-mini [da Silva 2025]. O sistema foi aplicado para enriquecer semanticamente uma base histórica de trabalhos do repositório da Universidade Federal

de Alagoas (UFAL), gerando campos estruturados como áreas de conhecimento, subáreas e domínios de aplicação, cuja qualidade foi mensurada por meio de validação manual.

Para apresentar a pesquisa de forma estruturada, este artigo está organizado como: a Seção 2 contextualiza os trabalhos relacionados à extração de palavras-chave e ao uso de modelos *Transformers*. A Seção 3 detalha a metodologia empregada, descrevendo o *dataset* e o pipeline construído. Os resultados obtidos e a validação das métricas são discutidos na Seção 4. Por fim, a Seção 5 apresenta as conclusões e os direcionamentos para trabalhos futuros.

2. Trabalhos Relacionados

A extração automatizada de termos e o enriquecimento de metadados em acervos digitais têm sido abordados por diferentes estratégias tecnológicas. Tradicionalmente, técnicas baseadas em frequência estatística e centralidade de grafos, como TF-IDF e TextRank [Mihalcea and Tarau 2004], foram o padrão para indexação de documentos. Contudo, esses métodos puramente sintáticos são limitados pela incapacidade de mapear sinônimos e contornos contextuais complexos. O surgimento de modelos baseados em *Transformers*, como o BERT [Devlin et al. 2019], permitiu capturar representações dinâmicas e bidirecionais do texto, melhorando a precisão da extração em documentos científicos, mas ainda exigindo processos custosos de treinamento ou ajuste fino (*fine-tuning*) para domínios específicos.

Recentemente, trabalhos começaram a utilizar a capacidade de generalização e o capacidade de inferência contextual de LLMs generativos para contornar essas limitações de forma mais eficiente. Nessa direção, [D'souza et al. 2025] propuseram um método de enriquecimento de catálogos bibliográficos por meio de LLMs, demonstrando expansão da cobertura semântica e da recuperabilidade dos registros em relação aos métodos tradicionais. Todavia, a abordagem dos autores concentrou-se em grandes bases de dados consolidadas e metadados de publicações científicas de alto impacto, cenário distinto da realidade de repositórios institucionais de graduação.

Em linha com a proposta aqui apresentada, [da Silva 2025] investigou a extração de palavras-chave via prompting com LLMs, demonstrando superioridade sobre algoritmos clássicos de PLN em métricas de precisão e revocação. Contudo, esse estudo restringiu-se à geração de termos isolados a partir de resumos, sem avançar na categorização hierárquica dos documentos. Assim, o presente estudo diferencia-se e complementa os trabalhos descritos ao propor um pipeline automatizado que não apenas extrai palavras-chave, mas realiza o enriquecimento semântico multi-campo, compreendendo área, subárea, domínio de aplicação, palavras-chave e resumo padronizado do repositório da UFAL.

3. Metodologia

Esta seção descreve os procedimentos metodológicos adotados para o desenvolvimento e a avaliação do sistema de enriquecimento semântico. A abordagem compreende a caracterização do *dataset* experimental, a arquitetura detalhada do pipeline de automação construído e, por fim, o protocolo de validação manual utilizado para aferir a qualidade das saídas geradas pelo modelo de linguagem, conforme Figura 1.



Figura 1. Overview da metodologia proposta.

3.1. Dataset e Pré-processamento

O *dataset* utilizado neste estudo é composto por Trabalhos de Conclusão de Curso (TCCs) de graduação vinculados aos cursos da área de Computação da Universidade Federal de Alagoas (UFAL). A coleta compreendeu registros históricos defendidos entre os anos de 2005 e 2025.

A base de dados bruta passou por uma etapa de filtragem e saneamento automatizado utilizando a biblioteca *Pandas* em ambiente Python. Foram aplicados filtros para a remoção de registros duplicados e eliminação de inconsistências críticas. Como critério de inclusão, estabeleceu-se a obrigatoriedade da presença do título ou do resumo preenchidos, uma vez que estes campos constituem o contexto mínimo necessário para o processamento semântico do modelo. Após a etapa de limpeza, o acervo final totalizou 904 registros processados no experimento. Para cada documento, foram selecionadas cinco colunas estruturais: (i) título, (ii) resumo, (iii) palavras-chave originais, (iv) ano de defesa e (v) curso de origem.

3.2. Pipeline de Enriquecimento Automatizado

O pipeline de enriquecimento foi implementado de forma sequencial em Python, utilizando a biblioteca oficial da *OpenAI*. O fluxo operacional foi projetado para mitigar falhas comuns de interrupção em chamadas de API e estruturar os dados de forma estrita. Após o carregamento do *dataset*, o sistema itera sobre os registros construindo dinamicamente as requisições.

O núcleo do pipeline consiste no envio de um *prompt* de sistema estruturado para o modelo *gpt-4o-mini*. Para mitigar o comportamento estocástico inerente aos modelos generativos e assegurar respostas determinísticas e reproduzíveis, o parâmetro *temperature* foi fixado em zero. Por meio de *role prompting* de *prompts* o modelo recebeu instrução sistêmica para operar como bibliotecário digital especializado em Ciência da Computação, com saída obrigatória em formato JSON contendo os campos definidos a seguir:

- Palavras-chave sugeridas: cinco termos técnicos padronizados;
- Área principal e Subárea: classificações hierárquicas aderentes à taxonomia da área;
- Domínio de aplicação: cenário prático em que a solução se aplica (ex: Saúde, Educação);

- Resumo curto padronizado: uma síntese concisa do objetivo e resultado do trabalho;
- Avaliação e Justificativa: uma nota qualitativa atribuída pelo modelo às palavras-chave originais do autor, acompanhada de uma explicação textual sobre a adequação dos termos.

O gpt-4o-mini foi selecionado pelo equilíbrio entre custo operacional por token e desempenho em tarefas de extração de informação estruturada, representando uma alternativa de menor custo em relação a modelos de maior porte da mesma família [OpenAI 2024]. Para garantir a vastidão do pipeline, a saída da API foi submetida a um bloco de tratamento de exceções com validação de esquema (*schema validation*).

Caso o modelo gerasse uma *string* corrompida ou fora do padrão especificado, o sistema efetuava uma nova tentativa de requisição automática (*retry logic*), assegurando a consolidação final dos metadados estruturados no *dataset* enriquecido.

3.3. Protocolo de Validação Manual

Para aferir a confiabilidade das inferências realizadas pelo *gpt-4o-mini*, estabeleceu-se um protocolo de validação quantitativa por meio de inspeção humana. Uma amostra aleatória composta por 50 trabalhos foi extraída do *dataset* final, utilizando a função de amostragem pseudoaleatória com a semente fixa *random_state=42*, assegurando a reprodutibilidade do protocolo de amostragem.

A avaliação foi conduzida manualmente por um especialista da área de Computação. O avaliador analisou cegamente as saídas do modelo em relação ao texto original dos TCCs, atribuindo notas de 1 (desempenho insatisfatório) a 5 (desempenho excelente), dispostas em uma escala Likert ancorada sob três dimensões conceituais:

1. Relevância: avalia se as palavras-chave geradas possuem relação direta e coerente com a temática central abordada no conteúdo do trabalho;
2. Especificidade: mensura o grau de precisão técnica dos termos gerados, penalizando vocábulos excessivamente genéricos (como “software” ou “computador”) que não discriminam o trabalho;
3. Utilidade de Busca: afere a capacidade prática dos novos metadados em otimizar algoritmos de indexação e facilitar a recuperação do documento por pesquisadores em uma consulta real.

4. Resultados e Discussões

Esta seção apresenta os resultados obtidos a partir do pipeline de enriquecimento semântico aplicado ao acervo de Computação da UFAL, bem como a análise dos dados gerados. O processo de validação manual na amostra de 50 trabalhos indicou desempenho elevado do modelo. Na escala Likert de 1 a 5, a dimensão de Relevância obteve a maior pontuação média (4,80), evidenciando que o modelo produziu inferências semanticamente aderentes ao conteúdo técnico dos documentos, mesmo quando o texto de entrada era restrito ao título e a um resumo breve.

As dimensões de Especificidade e Utilidade de Busca mantiveram médias elevadas ($\geq$4,47), confirmando que os metadados gerados são precisos e capazes de mitigar o esforço de busca bibliográfica no ecossistema de iniciação científica, expandindo a

descoberta de produções correlatas [Mosha and Ngulube 2023]. A distribuição das áreas principais atribuídas automaticamente pelo LLM está apresentada na Figura 2. A área de maior prevalência no acervo foi Ciência da Computação (≈300 trabalhos), seguida por Inteligência Artificial (≈175) e Computação (≈105). Essa distribuição é consistente com o perfil curricular dos cursos de tecnologia da instituição, nos quais Ciência da Computação e Inteligência Artificial concentram o maior volume de trabalhos defendidos no período analisado.

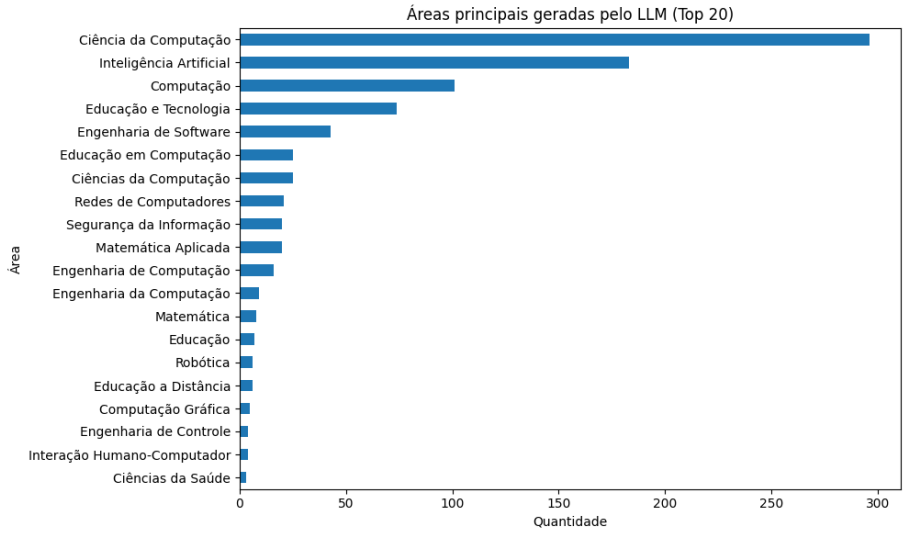


Figura 2. Distribuição das áreas principais geradas pelo LLM.

Para analisar a consistência histórica dessa classificação, a Figura 3 detalha a evolução temporal das 8 principais áreas mapeadas entre os anos de 2005 e 2025. Nota-se que Ciência da Computação mantém uma liderança consolidada ao longo das décadas, registrando um pico de produção em 2024 (≈36 trabalhos). Em paralelo, a área de Inteligência Artificial apresenta um crescimento acentuado e contínuo a partir do ano de 2020, atingindo ≈33 trabalhos em 2024. Esse comportamento corrobora com a expansão global e o interesse científico difuso em algoritmos de *machine learning* e LLMs observado na literatura recente [dos Santos e Eliseo Berni Reategui 2025], sinalizando que os repositórios institucionais demandam mecanismos de atualização semântica flexíveis para acompanhar essa rápida transição temática.

A avaliação automatizada acerca da qualidade das palavras-chave originais cadastradas pelos autores é exibida na Figura 4. A categoria *Boa* predominou amplamente no volume total de dados, abrangendo cerca de 890 trabalhos. A categoria *Regular* (≈20) representa uma parcela menor, porém significativa e a *Ruim* apresenta uma representação mínima, tornando-se estatisticamente desprezível no escopo desta análise. Esse diagnóstico inicial indica que, embora a maior parte dos discentes selecione termos minimamente adequados, subsiste uma margem crítica de degradação de metadados que justifica a intervenção de correções assistidas.

Complementarmente, a Figura 5 estabelece uma comparação direta entre a volumetria média de termos originais por documento (≈6,5) e a quantidade fixa sugerida pelo modelo generativo (≈5,0). Embora os dados históricos apresentem numericamente uma média superior, o comportamento determinístico do pipeline assegura um padrão de

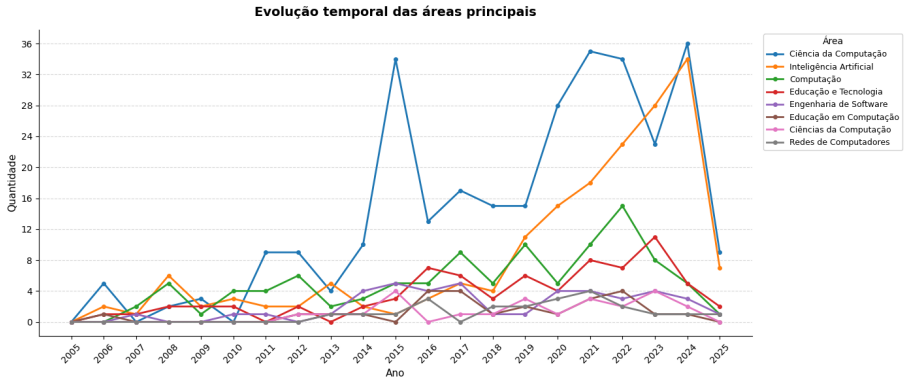


Figura 3. Evolução temporal das áreas principais mapeadas.

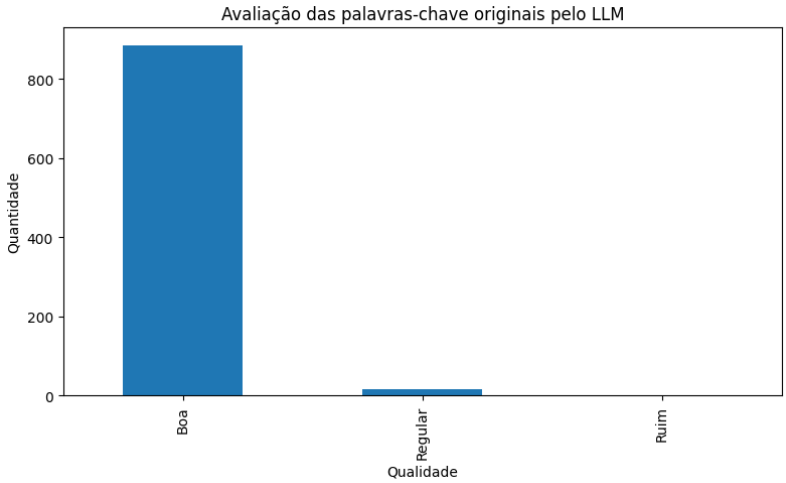


Figura 4. Avaliação da qualidade das palavras-chave originais realizada pelo LLM.

cobertura homogêneo. Essa padronização beneficia substancialmente os registros que sofriam com ausência crônica ou escassez de palavras-chave, equalizando o potencial de indexação da base de dados.

Ao analisar a frequência de vocábulos específicos, a Figura 6 ilustra as 20 palavras-chave mais recorrentes inseridas pelo pipeline. Destacam-se os termos *educação a distância* (24 ocorrências), *acessibilidade digital* (21) e *aprendizado de máquina* (16).

Essa prevalência terminológica é explicada diretamente pelos dados consolidados na Figura 7, que identifica os 20 domínios práticos de aplicação mais comuns nos projetos analisados. O segmento de Educação exerce uma liderança expressiva ($\approx$81 trabalhos), seguido por Saúde ($\approx$45) e Educação Superior ($\approx$39). A correspondência entre o domínio Educação e o termo *educação a distância*, o mais frequente entre as palavras-chave geradas, indica que o pipeline preserva a orientação aplicada dos trabalhos dos cursos analisados atribuindo metadados consistentes com o perfil temático dos cursos analisados.

5. Conclusão

Este trabalho demonstrou a viabilidade e a eficácia do uso de Grandes Modelos de Linguagem para o enriquecimento semântico automático de repositórios acadêmicos no domínio

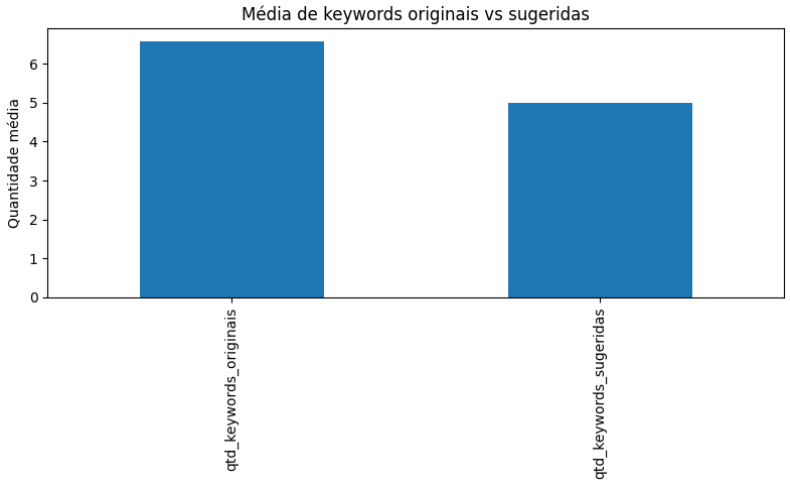


Figura 5. Média de termos por trabalho: Originais vs. Sugeridas pelo LLM.

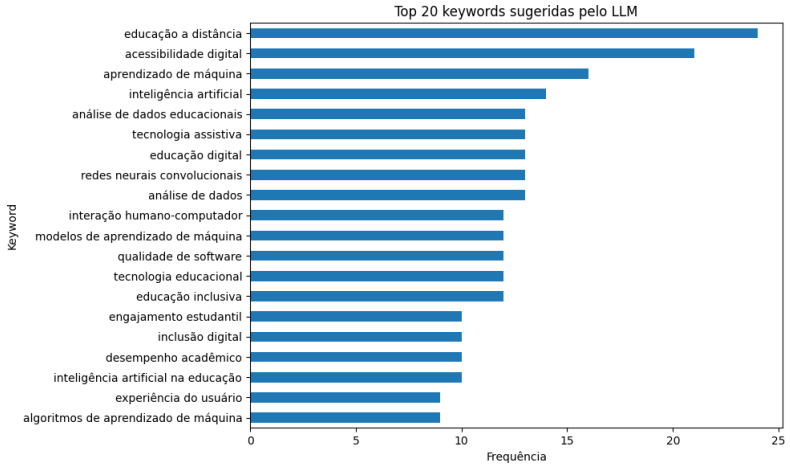


Figura 6. Top 20 palavras-chave sugeridas pelo LLM.

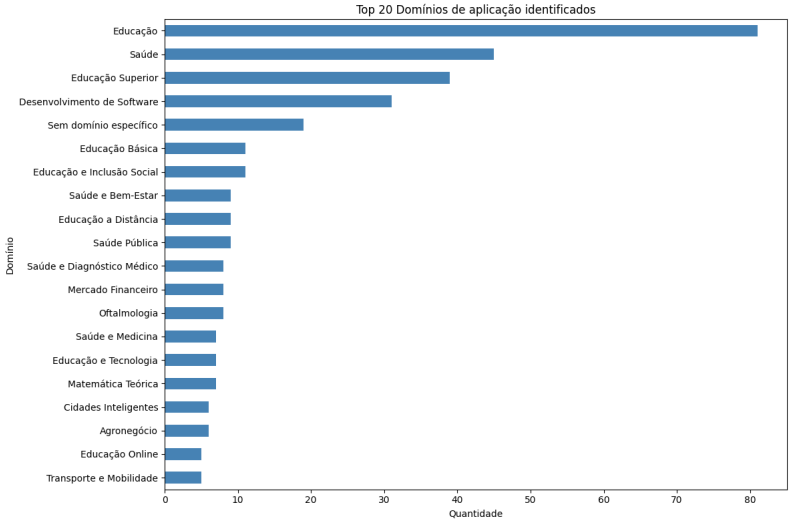


Figura 7. Top 20 domínios de aplicação identificados pelo LLM.

da Computação. O pipeline desenvolvido, baseado no modelo *gpt-4o-mini*, mostrou-se capaz de processar títulos e resumos históricos para gerar uma estrutura complexa de metadados padronizados em formato JSON, incluindo áreas, subáreas, domínios de aplicação e novas palavras-chave. A validação manual conduzida por um especialista técnico comprovou a alta qualidade das inferências, com médias superiores a 4,47 em uma escala de 1 a 5, destacando-se a relevância conceitual dos termos gerados (4,80). Os resultados demonstram que pipelines baseados em LLMs constituem uma abordagem viável para o enriquecimento semântico automatizado de repositórios institucionais, dispensando processos de ajuste fino *(fine-tuning)* específicos ao domínio, os quais demandam recursos computacionais e anotações manuais consideráveis.

Os achados trazem impactos diretos para o ecossistema de iniciação científica, uma vez que a padronização terminológica e a categorização hierárquica expandem a descoberta de produções correlatas e facilitam o levantamento bibliográfico por jovens pesquisadores. Além disso, o mapeamento temporal evidenciou as transições de interesse científico na instituição, como o crescimento acentuado da subárea de Inteligência Artificial nos últimos anos, corroborando a capacidade do pipeline de refletir a evolução temática da produção científica institucional ao longo do período analisado.

Este estudo apresenta, contudo, algumas limitações. A validação de qualidade foi conduzida por um único especialista, sem um segundo avaliador independente que permitisse o cálculo de métricas de concordância inter-avaliadores, o que introduz um componente de subjetividade na aferição dos resultados. Adicionalmente, a amostra de validação manual correspondeu a 50 dos 904 registros processados, podendo não capturar integralmente a variabilidade de qualidade presente no acervo completo. O pipeline também depende de um modelo proprietário acessado via API (gpt-4o-mini), o que implica custos recorrentes e sujeita a solução a eventuais descontinuações ou alterações de comportamento do provedor ao longo do tempo. Por fim, o estudo foi aplicado a um único repositório institucional, o que limita a generalização direta dos resultados para acervos de outras instituições com perfis curriculares distintos.

Como trabalhos futuros, planeja-se estender a avaliação quantitativa por meio de um protocolo multi-avaliador para o cálculo de concordância estatística, bem como replicar o pipeline utilizando modelos de código aberto locais, como a família LLAMA, visando a soberania dos dados e a redução de custos de API. Por fim, projeta-se o desenvolvimento de um plugin de integração direta do pipeline ao *software* DSpace, permitindo que o enriquecimento semântico ocorra de forma transparente no fluxo de depósito dos trabalhos institucionais.

Referências

Braz, M. P. (2024). Uso da inteligência artificial generativa na iniciação científica: um exercício de design especulativo. Dissertação (mestrado em tecnologias da inteligência e design digital), Pontifícia Universidade Católica de São Paulo (PUC-SP), São Paulo, SP.

Choudhury, M. H., Salsabil, L., Jayanetti, H. R., Wu, J., Ingram, W. A., and Fox, E. A. (2023). Metaenhance: Metadata quality improvement for electronic theses and dissertations of university libraries. In *Proceedings of the 2023 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*.

da Silva, L. F. M. (2025). *Análise e Visualização de Trabalhos de Conclusão de Curso utilizando Processamento de Linguagem Natural: Um Método Baseado em Grandes Modelos de Linguagem*. Trabalho de conclusão de curso (graduação em tecnologias da informação e comunicação), Universidade Federal de Santa Catarina (UFSC), Araranguá, SC.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 4171–4186.

dos Santos e Eliseo Berni Reategui, R. A. (2025). Uso de inteligência artificial generativa e análise de palavras-chave para apoiar o planejamento de projetos de pesquisa no ensino superior. *RELATEC: Revista Latinoamericana de Tecnología Educativa*, 24(2):87–104.

D'souza, J., Sadruddin, S., Israel, H., Begoin, M., and Slawig, D. (2025). Semeval-2025 task 5: Llms4subjects-llm-based automated subject tagging for a national technical library's open-access catalog. In *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*, pages 2570–2583.

Formánek, M. (2024). Exploring the potential of large language models and generative artificial intelligence (gpt): Applications in library and information science. *Journal of Librarianship and Information Science*.

INEP (2024). Censo da educação superior.

Lamba, M., Peng, Y., Nikolov, S., Layne-Worthey, G., and Downie, J. S. (2025). Metadata enrichment of long text documents using large language models. *arXiv preprint arXiv:2506.20918*.

Medina Nieto, M. A., Vázquez Mora, P. D., De la Calleja Mora, J., Tovar Vidal, M., López Domínguez, E., Arrieta Díaz, D., and Bárcenas Patiño, I. E. (2020). Web service to retrieve and semantically enrich datasets for theses from open educational repositories. *IEEE Access*, 8.

Mihalcea, R. and Tarau, P. (2004). TextRank: Bringing order into texts. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 404–411.

Mosha, N. and Ngulube, P. (2023). Metadata standard for continuous preservation, discovery, and reuse of research data in repositories by higher education institutions: A systematic review. *Information*, 14(8).

OpenAI (2024). GPT-4o mini: Advancing Cost-Efficient Intelligence. OpenAI Blog. Acesso em: jun. 2026.

Zhao, W. X., Zhou, K., Li, J., Tang, T., Wang, X., Hou, Y., Min, Y., Zhang, B., Zhang, J., Wang, Z., et al. (2023). A survey of large language models. *arXiv preprint arXiv:2303.18223*.