INFORMAÇÃO & INFORMAÇÃO
DOI: 10.543/1981-8920.2026v31n1p149
eISSN 1981-8920

Organização e arquitetura da informação em *large language models*

interseções na Ciência da Informação

Daiane Campos Procópio
Mestre em Gestão e Organização do Conhecimento pelo Programa de Pós-Graduação (PPGGOC) da Universidade Federal de Minas Gerais (UFMG). Belo Horizonte, Brasil.
campos-daiane@ufmg.br

Patrícia Nascimento Silva
Doutora em Gestão e Organização do Conhecimento pela Universidade Federal de Minas Gerais (UFMG). Docente na Escola de Ciência da Informação (ECI) e da Universidade Federal de Minas Gerais e no Programa de Pós-Graduação em Gestão e Organização do Conhecimento (PPGGOC). Belo Horizonte, Brasil.
patricians@ufmg.br

Resumo

Objetivo: identificar e analisar a organização e a arquitetura da informação em Large Language Models (LLMs), relacionando-as às técnicas de Inteligência Artificial (IA) e discutindo esses resultados no contexto da Ciência da Informação (CI). **Metodologia:** a pesquisa, de natureza bibliográfica, é uma revisão de literatura, com abordagem mista, que contou com um protocolo criterioso para investigar a temática nas bases ACM Digital Library, ScienceDirect, Scopus e Web of Science, realizada entre junho e agosto de 2025. Os dados foram analisados quantitativamente e qualitativamente, com aplicação da análise de conteúdo, na perspectiva de Bardin, por meio da análise categorial. **Resultados:** foram examinados 53 estudos publicados entre 2022 e 2025, oriundos de 26 países, identificando-se 16 categorias de técnicas de IA aplicadas em LLMs. Essas técnicas foram relacionadas aos quatro sistemas da Arquitetura da Informação propostos por Rosenfeld, Morville e Arango: organização, rotulagem, navegação e busca, demonstrando contribuições para o desenvolvimento e a aplicação de LLMs em diferentes domínios. **Conclusões:** a pesquisa contribui ao evidenciar como a organização e a arquitetura da informação dialogam com a inteligência artificial e com instrumentos tradicionais da área, como metadados, tesauros e taxonomias, podendo ser aplicadas para compreender de que maneira os LLMs organizam, representam e disponibilizam dados. Além disso, oferece uma base conceitual importante, subsidiando novos estudos e soluções que envolvem a curadoria de dados e o treinamento desses sistemas. O estudo também reforça a pertinência dos fundamentos da CI na análise de tecnologias emergentes.

Descritores: Arquitetura da Informação. Large Language Models. Revisões de literatura.

Recebido em: 10.09.2025 | **Aceito em:** 12.03.2026

1 Introdução

A ascensão dos Large Language Models (LLMs) tem transformado a maneira como as pessoas acessam e interagem com a informação. Esses modelos, baseados em Inteligência Artificial Generativa (IAG), são capazes de produzir textos coerentes, responder a perguntas complexas e realizar tarefas diversas em linguagem natural (Dhamani; Engler, 2024), o que tem ampliado seu uso em contextos pessoais, profissionais e acadêmicos. No entanto, apesar de sua popularidade crescente, há uma lacuna entre o uso cotidiano dessas ferramentas e a compreensão técnica de seu funcionamento. Compreender como a informação é organizada, tratada e estruturada nesses sistemas é fundamental para promover uma cultura digital ética e transparente, implicando em pessoas mais críticas e informadas.

Essa lacuna de compreensão pode gerar riscos como a desinformação, a exposição indevida de dados pessoais, a dependência acrítica de respostas automatizadas e a redução da capacidade analítica dos usuários (Chen *et al.,* 2025; Chow, 2025; Ferreira; Oliveira, 2021; Zhou; Geißler; Lukowicz, 2024). Além disso, o desempenho dos LLMs está diretamente relacionado à forma como os dados textuais são organizados e preparados durante seu treinamento.

Etapas como a curadoria, a limpeza e a rotulagem das informações influenciam diretamente a qualidade das respostas geradas, o nível de viés presente no modelo e a sua capacidade para fornecer informações confiáveis (Lukichev, 2024). Neste contexto, torna-se essencial investigar os elementos que compõem a organização e a arquitetura da informação em LLMs, sendo este o questionamento que norteia esta pesquisa, traduzido nas seguintes perguntas: Como a informação tem sido organizada, tratada e estruturada em Large Language Models? Quais técnicas de Inteligência Artificial são utilizadas nesse cenário?

Com isso, esta pesquisa objetiva identificar e analisar a organização e a arquitetura da informação em LLMs. Especificamente, buscou-se: 1) identificar e mapear estudos sobre a organização e a arquitetura da informação em LLMs e 2) relacionar esta arquitetura com as técnicas de Inteligência Artificial (IA), apresentando os conceitos principais que a compõem, no contexto da Ciência da Informação (CI).

A pesquisa, de natureza bibliográfica, é uma revisão de literatura de abordagem mista (qualitativa e quantitativa) que utilizou um protocolo criterioso para investigar a

temática em quatro bases de dados, executado em junho de 2025.

A justificativa para esta pesquisa reside na necessidade de ampliar o entendimento crítico sobre os modelos de linguagem e sua relação com os dados e as informações, apoiando novos projetos relacionados à governança desses dados e às formas de recuperação envolvendo fontes externas. Além disso, tornar compreensível o modo como esses modelos são construídos e operam, especialmente no que se refere à estruturação e ao tratamento da informação, é essencial para fortalecer a alfabetização digital, fomentar o uso responsável da tecnologia e mitigar riscos sociais relacionados à desinformação e à privacidade.

A pesquisa integra o projeto de pesquisa Dados abertos para potencializar a recuperação de informação em modelos de Inteligência Artificial, além de uma pesquisa de doutorado em andamento, e trata-se de uma iniciativa relevante do ponto de vista educacional e social, pois oferece subsídios para soluções nacionais e soberanas, utilizando dados e informações abertas para enriquecer sistemas e aplicações.

2 Fundamentação teórica

2.1 Large Language Models

De acordo com Kallens, Kristensen-McLachlan e Christiansen (2023, p. 1, tradução nossa), LLMs são “arquiteturas sofisticadas de aprendizagem profunda treinadas em grandes quantidades de dados de linguagem natural, permitindo-lhes realizar uma variedade de tarefas linguísticas”. Os LLMs possuem ampla versatilidade, podendo gerar, resumir e traduzir textos, produzir códigos computacionais, simular diálogos, corrigir erros gramaticais, entre outras tarefas.

Além disso, esses modelos realizam buscas semânticas, identificando não apenas termos exatos, mas também seu contexto, o que possibilita recuperar informações relevantes mesmo quando o usuário não utiliza as mesmas palavras na consulta (Wei; Huang; Wang, 2025).

O desenvolvimento de um LLM ocorre em duas etapas principais: o pré-treinamento (*pretraining*), que demanda grande volume de dados e recursos computacionais, no qual o modelo aprende padrões linguísticos gerais a partir de textos da *internet*, resultando nos chamados modelos base; e o refinamento ou ajuste

fino (*fine-tuning*), fase em que esse modelo é ajustado para tarefas ou comportamentos específicos, aproveitando o conhecimento já adquirido e reduzindo custos em relação ao pré-treinamento. Esse processo permite tanto a criação de modelos base generalistas quanto a sua adaptação para necessidades particulares, incluindo personalizações adicionais conforme as preferências dos usuários (Alammar; Grootendorst, 2024).

Enquanto o pré-treinamento e o ajuste fino descrevem as etapas de desenvolvimento de um LLM, o conceito de parâmetro refere-se ao quanto o modelo se internalizou durante o aprendizado e à sua capacidade de realizar tarefas complexas. Para exemplificar, o GPT-3.5, desenvolvido pela OpenAI, possui aproximadamente 175 bilhões de parâmetros (Dhamani; Engler, 2024).

Embora os LLMs tenham alcançado resultados importantes em diversas tarefas, ainda apresentam limitações relevantes, principalmente ligadas aos dados de treinamento, que podem absorver vieses sociais e estereótipos, refletindo desigualdades existentes em textos coletados da *internet*. Além disso, o volume de dados utilizados dificulta o controle de qualidade, aumentando o risco de incorporação de informações incorretas, ofensivas ou protegidas por direitos autorais. Outro desafio é a ocorrência de alucinações, que acontecem quando o modelo gera respostas incorretas, por considerar contextos diferentes, mas que parecem ser corretas, comprometendo a confiabilidade de seus resultados. Esses problemas reforçam a necessidade de cautela no uso e na interpretação das respostas geradas pelos LLMs, bem como de estratégias para mitigar esses problemas.

2.2 Arquitetura da informação

Com a popularização do uso da *internet*, a forma de acessar e utilizar informações mudou significativamente. Hoje, é possível consultar conteúdos variados diretamente do *smartphone*, acompanhar em tempo real indicadores de saúde por meio de dispositivos *wearebles* e realizar diversas atividades apoiadas em dados digitais. Esse cenário ampliou as possibilidades de acesso à informação, trazendo ganhos importantes para a qualidade de vida das pessoas. Porém, ao mesmo tempo em que trouxe benefícios, tal dinâmica também gerou desafios, pois quanto maior o volume de dados disponíveis, mais difícil se torna localizar, selecionar e,

principalmente, compreender a informação (Rosenfeld; Morville; Arango, 2015).

Nesse contexto, destaca-se a importância da Arquitetura da Informação, entendida como um campo que busca organizar, estruturar e facilitar a interação entre pessoas e sistemas digitais. Como ressalta Lima (2016, p. 48), sua principal função é "proporcionar uma estrutura lógica que possa ajudar o usuário a encontrar a informação de que necessita". Segundo Gabriel-Petit (2025) a Arquitetura da Informação corresponde a uma prática de *design* voltada à definição da estrutura de ambientes digitais, com foco na organização das informações e no apoio à localização e usabilidade por meio de sistemas de rotulagem, navegação e pesquisa bem elaborados.

De acordo com Rosenfeld, Morville e Arango (2015), a arquitetura da informação é estruturada em quatro sistemas fundamentais: organização, rotulação, navegação e busca. O sistema de organização diz respeito às formas de categorizar a informação; a rotulação trata da maneira como essa informação é representada; a navegação refere-se aos caminhos que permitem percorrê-la; e a busca está relacionada aos mecanismos utilizados para localizá-la.

Além desses sistemas, a Arquitetura da Informação também incorpora instrumentos amplamente estudados na CI, como tesauros, vocabulários controlados e metadados. Esses elementos funcionam como mediadores entre a linguagem utilizada pelos sistemas e aquela empregada pelos usuários, facilitando o acesso às informações (Rosenfeld; Morville; Arango, 2015).

Assim, a Arquitetura da Informação oferece suporte conceitual e metodológico para a construção de ambientes informacionais digitais mais compreensíveis aos usuários, buscando oferecer experiências de interação mais eficazes, orientadas tanto para públicos específicos quanto para potenciais usuários que ainda não têm clareza de suas necessidades informacionais (Macie; Nascimento; Madio, 2024; Torino *et al.*, 2022).

3 Metodologia

Neste estudo, buscou-se investigar a Arquitetura da Informação nos LLMs a partir da análise da literatura especializada. Para isso, adotou-se uma abordagem mista, ou qualiquantitativa, de caráter exploratório, fundamentada em uma revisão de

literatura.

Segundo Creswell e Creswell (2021, p. 4), a pesquisa de métodos mistos consiste em uma "[...] investigação que envolve a coleta de dados quantitativos e qualitativos, integrando os dois tipos de dados e usando desenhos distintos que refletem pressupostos filosóficos e estruturas teóricas". Quanto ao caráter exploratório, Corrêa (2008, p. 26) define esse tipo de pesquisa como aquela que "busca um conhecimento inicial sobre determinado tema ou objeto de estudo", visando à familiarização com o assunto. Por fim, a revisão de literatura foi adotada como principal procedimento metodológico. De acordo com Matias-Pereira (2019, p. 204), esse tipo de revisão consiste em "[...] uma análise comentada do que já foi escrito sobre o tema de sua pesquisa, procurando mostrar os pontos de vista convergentes e divergentes dos autores". No contexto deste estudo, a revisão foi utilizada para localizar e analisar publicações sobre a Arquitetura da Informação em LLMs.

4 Procedimentos metodológicos

4.1 Procedimentos metodológicos

Para alcançar os objetivos da pesquisa, esta foi dividida em três etapas metodológicas, que descrevem os critérios adotados para a realização da revisão de literatura. A primeira etapa consistiu na definição dos termos relacionados à temática investigada, que orientaram a construção da *string* de busca. Esses termos foram organizados nos idiomas português, inglês e espanhol, com o intuito de ampliar a cobertura e a recuperação de documentos relevantes em bases de dados multilíngues. A escolha dos termos considerou tanto as denominações mais comuns quanto suas variações, como apresentado no Quadro 1.

Quadro 1 - Definição dos termos de busca utilizados na pesquisa

Definição dos termos e seus sinônimos		
Termos em português	**Termos em inglês**	**Termos em espanhol**
Modelos de Linguagem de Larga Escala	*Large Language Models*	*Modelos de lenguaje a gran escala*
Grandes Modelos de Linguagem	*Large Language Models*	*Grandes modelos de lenguaje*
-	*LLM*	-
-	*LLMs*	-

Arquitetura da informação	*Information architecture*	*Arquitectura de la información*
Organização da informação	*Information organization*	*Organización de la información*
Tratamento de dados	*Data processing*	*Tratamiento de datos*
Termos retirados após teste na *string*		
Refinamento	*Fine-tuning*	*Refinamiento*
Arquitetura	*Architecture*	*Arquitectura*
Gestão da informação	*Information management*	*Gestión de la información*
Estruturação de dados	*Data structuring*	*Estructuración de datos*
Transformador	*Transformer*	*Transformador*
***String* geral**		
((("Modelos de Linguagem de Larga Escala" OR "Grandes Modelos de Linguagem") OR ("Large Language Models" OR LLM OR LLMs) OR ("Modelos de lenguaje a gran escala" OR "Grandes modelos de lenguaje")) AND (("Arquitetura da informação" OR "Information architecture" OR "Arquitectura de la información") OR ("Organização da informação" OR "Information organization" OR "Organización de la información") OR ("Tratamento de dados" OR "Data processing" OR "Tratamiento de datos")))		

Fonte: Elaborado pelas autoras (2025).

Na segunda etapa, foi elaborado um protocolo, indicando os objetivos e as questões da revisão, a seleção das bases e os parâmetros de elegibilidade, que englobam os critérios de inclusão e exclusão das publicações recuperadas nas buscas, visando garantir a relevância, a acessibilidade e a consistência do *corpus* analisado, bem como procedimentos para análise e tratamento, conforme apresentado no Quadro 02.

Quadro 02 - Protocolo da revisão da literatura

Critérios	Descrição
Objetivo geral	Identificar publicações sobre os principais elementos da Arquitetura da Informação em *Large Language Models* (LLMs).
Questões a serem resolvidas	Como a informação tem sido organizada, tratada e estruturada em LLMs?
	Quais técnicas de inteligência artificial são utilizadas?
Fontes de informação pesquisadas	Base de dados: ACM Digital Library, Science Direct, Scopus e Web Of Science. ● ACM Digital Library: por seu foco em Ciência da Computação e Tecnologia da Informação. ● ScienceDirect: por reunir periódicos da Elsevier nas áreas de ciência e tecnologia. ● Scopus: pela cobertura multidisciplinar e ampla indexação de periódicos revisados por pares. ● Web of Science: por sua curadoria rigorosa e cobertura internacional

	de pesquisas de alto impacto.
Critérios de elegibilidade	Sem delimitação de data.
	Tipologia documental: artigos de periódicos e trabalhos de eventos, pois possuem revisão por pares.
	Idiomas: português, inglês e espanhol, a fim de aumentar a abrangência dos estudos.
Critérios de inclusão e de exclusão	Inclusão de publicações de áreas que estudam Arquitetura da Informação e LLMs.
	Inclusão de publicações nos idiomas inglês, português e espanhol.
	Exclusão de publicações duplicadas.
	Exclusão de publicações que não foram redigidas em português, inglês ou espanhol.
	Publicações cujos títulos e resumos não abordam Arquitetura da Informação e LLMs.
	Publicações que não estejam disponíveis na íntegra via Portal de Periódicos da Capes.
Campos de busca	Título, resumo e palavras-chave.
Expressões de busca	Large Language Models e suas variações **E** arquitetura da informação e suas variações.
	As expressões foram utilizadas em português, inglês e espanhol.
String	((("Modelos de Linguagem de Larga Escala" OR "Grandes Modelos de Linguagem") OR ("Large Language Models" OR LLM OR LLMs) OR ("Modelos de lenguaje a gran escala" OR "Grandes modelos de lenguaje")) AND (("Arquitetura da informação" OR "Information architecture" OR "Arquitectura de la información") OR ("Organização da informação" OR "Information organization" OR "Organización de la información") OR ("Tratamento de dados" OR "Data processing" OR "Tratamiento de datos")))
Procedimentos de seleção dos documentos recuperados	Leitura dos títulos e resumos das publicações recuperadas com o intuito de verificar a pertinência do conteúdo aos objetivos da pesquisa.
Procedimentos de análise	Leitura completa das publicações, registrada em planilha estruturada, a fim de identificar elementos sobre a organização e a arquitetura da informação em LLMs.
Critério de exclusão após análise dos documentos	Pesquisas que não possuem abordagem conceitual, teórica ou metodológica sobre LLMs e Arquitetura da Informação.
Tratamento	*Software* Parsifal para a gestão do protocolo e das publicações, e Google Sheets para o compartilhamento de dados e desenvolvimento do relatório da pesquisa entre as pesquisadoras.

Fonte: Adaptado de Nascimento Silva (2023).

Na terceira etapa, foram executadas as buscas nas bases selecionadas, utilizando a *string* elaborada para cada base, conforme expressão geral do protocolo, apresentado no Quadro 3.

Quadro 3 - Strings utilizadas nas bases de dados

Base	*String*	Quant. de resultados	Data da busca	Filtros
ACM Digital Library (Title)	[[Title: "modelos de linguagem de larga escala"] OR [Title: "grandes modelos de linguagem"] OR [Title: "large language models"] OR [Title: llm] OR [Title: llms] OR [Title: "modelos de lenguaje a gran escala"] OR [Title: "grandes modelos de lenguaje"]] AND [[Title: "arquitetura da informação"] OR [Title: "information architecture"] OR [Title: "arquitectura de la información"] OR [Title: "organização da informação"] OR [Title: "information organization"] OR [Title: "organización de la información"] OR [Title: "tratamento de dados"] OR [Title: "data processing"] OR [Title: "tratamiento de datos"]]	2	25-06-25	Publicações de acesso aberto
ACM Digital Library (Abstract)	[[Abstract: "modelos de linguagem de larga escala"] OR [Abstract: "grandes modelos de linguagem"] OR [Abstract: "large language models"] OR [Abstract: llm] OR [Abstract: llms] OR [Abstract: "modelos de lenguaje a gran escala"] OR [Abstract: "grandes modelos de lenguaje"]] AND [[Abstract: "arquitetura da informação"] OR [Abstract: "information architecture"] OR [Abstract: "arquitectura de la información"] OR [Abstract: "organização da informação"] OR [Abstract: "information organization"] OR [Abstract: "organización de la información"] OR [Abstract: "tratamento de dados"] OR [Abstract: "data processing"] OR [Abstract: "tratamiento de datos"]]	29	25-06-25	Publicações de acesso aberto
ACM Digital Library (Author keyword)	[[Keywords: "modelos de linguagem de larga escala"] OR [Keywords: "grandes modelos de linguagem"] OR [Keywords: "large language models"] OR [Keywords: llm] OR [Keywords: llms] OR [Keywords: "modelos de lenguaje a gran escala"] OR [Keywords: "grandes modelos de lenguaje"]] AND [[Keywords: "arquitetura da informação"] OR	4	25-06-25	Publicações de acesso aberto

	[Keywords: "information architecture"] OR [Keywords: "arquitectura de la información"] OR [Keywords: "organização da informação"] OR [Keywords: "information organization"] OR [Keywords: "organización de la información"] OR [Keywords: "tratamento de dados"] OR [Keywords: "data processing"] OR [Keywords: "tratamiento de datos"]]			
ScienceDirect (português)	Title, abstract, keywords: (("Modelos de Linguagem de Larga Escala" OR "Grandes Modelos de Linguagem" OR LLM OR LLMs) AND ("Arquitetura da informação" OR "Organização da informação" OR "Tratamento de dados"))	0	25-06-25	Access type: Open access & Open archive
ScienceDirect (inglês)	Title, abstract, keywords: (("Large Language Models" OR LLM OR LLMs) AND ("Information architecture" OR "Information organization" OR "Data processing"))	47	25-06-25	Access type: Open access & Open archive
ScienceDirect (espanhol)	Title, abstract, keywords: (("Modelos de lenguaje a gran escala" OR "Grandes modelos de lenguaje" OR LLM OR LLMs) AND ("Arquitectura de la información" OR "Organización de la información" OR "Tratamiento de datos"))	0	25-06-25	Access type: Open access & Open archive
Scopus	TITLE-ABS-KEY (((("Modelos de Linguagem de Larga Escala" OR "Grandes Modelos de Linguagem") OR ("Large Language Models" OR LLM OR LLMs) OR ("Modelos de lenguaje a gran escala" OR "Grandes modelos de lenguaje")) AND (("Arquitetura da informação" OR "Information architecture" OR "Arquitectura de la información") OR ("Organização da informação" OR "Information organization" OR "Organización de la información") OR ("Tratamento de dados" OR "Data processing" OR "Tratamiento de datos")	654	25-06-25	Open access: All open access
Web of Science	TS=(("Modelos de Linguagem de Larga Escala" OR "Grandes Modelos de Linguagem" OR "Large Language Models" OR LLM OR LLMs OR "Modelos de lenguaje a gran escala" OR "Grandes modelos de lenguaje")	162	25-06-25	Open access: All open access

	AND ("Arquitetura da informação" OR "Information architecture" OR "Arquitectura de la información" OR "Organização da informação" OR "Information organization" OR "Organización de la información" OR "Tratamento de dados" OR "Data processing" OR "Tratamiento de datos"			

Fonte: Elaborado pelas autoras (2025).

Os documentos foram exportados das bases de dados e importados na ferramenta Parsifal, onde a revisão foi conduzida, aplicando-se os critérios de elegibilidade, inclusão e exclusão.

Em seguida, procedeu-se com os procedimentos de seleção e análise, iniciados pela leitura dos títulos e resumos das publicações para a seleção inicial dos documentos. As publicações elegíveis passaram, então, para a fase de leitura integral. A amostra de documentos selecionada foi analisada quantitativamente, com relação ao idioma, à vinculação e à área de conhecimento, e qualitativamente, por meio de uma categorização facetada, elaborada com base na análise de conteúdo proposta por Bardin (2016), especialmente em sua modalidade de análise categorial, o que permitiu identificar as técnicas utilizadas, bem como aspectos sobre a organização e a arquitetura da informação.

Para Bardin (2016), a análise de conteúdo consiste em um conjunto de instrumentos metodológicos em constante aprimoramento, aplicável a discursos diversos, considerando tanto seus conteúdos quanto seus suportes.

Essa análise compreende três fases principais: a pré-análise, a exploração do material e o tratamento dos resultados e interpretação. A pré-análise corresponde ao momento de organização e sistematização inicial do *corpus*, permitindo ao pesquisador delimitar o material e formular hipóteses preliminares. Na fase de exploração do material, acontece a codificação, entendida como a transformação dos dados brutos em unidades de análise significativas, que pode ocorrer de forma aberta ou fechada, conforme a existência ou não de categorias prévias. Por fim, o tratamento dos resultados e a interpretação consistem na etapa de inferência e atribuição de sentido aos dados, articulando-os ao referencial teórico da pesquisa (Bardin, 2016; Valle; Ferreira, 2025).

A revisão de literatura foi conduzida entre os meses de junho e agosto de 2025,

sendo o protocolo executado no dia 25/06/2025.

5 Análise e discussão dos resultados

5.1 Análise quantitativa

Após a aplicação do protocolo de revisão, foram inicialmente recuperadas 898 publicações, das quais 209 estavam duplicadas e foram removidas, restando 698. Após a aplicação de filtros de acesso aberto e de idioma (inglês, português e espanhol), a amostra foi reduzida para 101 publicações. A leitura de títulos e resumos eliminou trabalhos fora do escopo da pesquisa, conforme critérios definidos para a inclusão e a exclusão, resultando em 57 publicações. Por fim, a leitura completa levou à seleção de 53 publicações alinhadas aos objetivos da pesquisa. A relação dos estudos está disponível no Quadro 4.

Quadro 4 - Relação dos estudos selecionados na pesquisa

ID	Título	Autores	Ano
1	A Conceptual Framework for a Latest Information-Maintaining Method Using Retrieval-Augmented Generation and a Large Language Model in Smart Manufacturing: Theoretical Approach and Performance Analysis	Hangseo Choi e Jongpil Jeong	2025
2	A Dutch Financial Large Language Model	Sander Noels, Jorne De Blaere e Tijl De Bie	2024
3	A fine-tuning enhanced RAG system with quantized influence measure as AI judge	Keshav Rangan e Yiqiao Yin	2024
4	A framework enabling LLMs into regulatory environment for transparency and trustworthiness and its application to drug labeling document	Leihong Wu *et al*.	2024
5	A Framework for LLM-Assisted Smart Policing System	Paria Sarzaeim, Qusay H. Mahmoud e Akramul Azim	2024
6	A method to promote safe cycling powered by large language models and IA agents	Daniel G. Costa *et al*.	2024
7	A Multimodal Large Language Model Framework for Intelligent Perception and Decision-Making in Smart Manufacturing	Tianyu Wang *et al*.	2025
8	A New Pipeline for Generating Instruction Dataset via RAG and Self Fine-Tuning	Chih-Wei Sung, Yu-Kai Lee e Yin-Te Tsai	2024

9	A review on enhancing agricultural intelligence with large language models	Hongda Li *et al*.	2025
10	A scientific-article key-insight extraction system based on multiactor of fine-tuned open-source large language models	Zihan Song *et al*.	2025
11	A User-Centered Framework for Data Privacy Protection Using Large Language Models and Attention Mechanisms	Shutian Zhou *et al*.	2024
12	Advancing EHR analysis Predictive medication modeling using LLMs	Hanan Alghamdi e Abeer Mostafa	2025
13	Advancing Tinnitus TherapeuticsTherapeutics: GPT-2 Driven Clustering Analysis of Cognitive Behavioral Therapy Sessions and Google T5-Based Predictive Modeling for THI Score Assessment	Yongwoo Jeong *et al*.	2024
14	Aggregated Knowledge Model Enhancing Domain-Specific QA with Fine-Tuned and Retrieval-Augmented Generation Models	Fengchen Liu *et al*.	2024
15	Application of retrieval-augmented generation for interactive industrial knowledge management via a large language model	Lun-Chi Chen *et al*.	2025
16	Automating Systematic Literature Reviews with Retrieval-Augmented Generation: A Comprehensive Overview	Binglan Han, Teo Susnjak e Anuradha Mathrani	2024
17	Batch-ICL: Effective, Efficient, and Order-Agnostic In-Context Learning	Kaiyi Zhang *et al*.	2024
18	Building Lightweight Domain-Specific Consultation Systems via Inter-External Knowledge Fusion Contrastive Learning	Jiabin Zheng, Hanlin Wang e Jiahui Yao	2024
19	Can language models automate data wrangling?	Gonzalo Jaimovitch-López *et al*.	2022
20	CMLLM: A novel cross-modal large language model for wind power forecasting	Guopeng Zhu *et al*.	2025
21	Combining Financial Data and News Articles for Stock Price Movement Prediction Using Large Language Models	Ali Elahi e Fatemeh Taghvaei	2024
22	DataAgent: Evaluating Large Language Models' Ability to Answer Zero-Shot, Natural Language Queries	Manit Mishra *et al*.	2024
23	Deciphering Public Voices in the Digital Era: Benchmarking ChatGPT for Analyzing Citizen Feedback in Hamilton, New Zealand	Xinyu Fu *et al*.	2024
24	Deep Learning and Methods Based on Large Language Models Applied to Stellar Light Curve Classification	Yu-Yang Li *et al*.	2025
25	Development of a Natural Language Processing (NLP) model to automatically extract clinical data from electronic health records: results from an Italian comprehensive stroke center	Davide Badalotti *et al*.	2024
26	DocFinQA: A Long-Context Financial Reasoning Dataset	Varshini Reddy *et al*.	2024

27	Domain-Specific Manufacturing Analytics Framework: An Integrated Architecture with Retrieval-Augmented Generation and Ollama-Based Models for Manufacturing Execution Systems Environments	Hangseo Choi e Jongpil Jeong	2025
28	DRG-LLaMA: tuning LLaMA model to predict diagnosis-related group for hospitalized patients	Hanyin Wang *et al*.	2024
29	Enhancing E-Government Services through State-of-the-Art, Modular, and Reproducible Architecture over Large Language Models	George Papageorgiou *et al*.	2024
30	Enhancing Financial Domain Adaptation of Language Models via Model Augmentation	Kota Tanabe *et al*.	2024
31	Enriching building function classification using Large Language Model embeddings of OpenStreetMap Tags	Abdulkadir Memduhoğlu, Nir Fulman e Alexander Zipf	2024
32	Extracting lung cancer staging descriptors from pathology reportsreports: A generative language model approach	Hyeongmin Cho *et al*.	2024
33	Fisher Information-based Efficient Curriculum Federated Learning with Large Language Models	Ji Liu *et al*.	2024
34	Fishnet: Financial Intelligence from Sub-querying, Harmonizing, Neural-Conditioning, Expert Swarms, and Task Planning	Nicole Cho *et al*.	2024
35	JaFIn: Japanese Financial Instruction Dataset	Kota Tanabe *et al*.	2024
36	Large language models deconstruct the clinical intuition behind diagnosing autism	Jack Stanley *et al*.	2025
37	Large language models for preventing medication direction errors in online pharmacies	Cristobal Pais *et al*.	2024
38	LEGF-DST: LLMs-Enhanced Graph-Fusion Dual-Stream Transformer for Fine-Grained Chinese Malicious SMS Detection	Xin Tong *et al*.	2025
39	LLM-Assisted Multi-Teacher Continual Learning for Visual Question Answering in Robotic Surgery	Kexin Chen *et al*.	2024
40	LLM-Assisted Qualitative Data Analysis: Security and Privacy Concerns in Gamified Workforce Studies	Aisvarya Adeseyea, Jouni Isoahoa e Tahir Mohammad	2025
41	Military Equipment Entity Extraction Based on Large Language Model	Xuhong Liu *et al*.	2024
42	Model development for bespoke large language models for digital triage assistance in mental health care	Niall Taylor *et al*.	2024
43	Multimodal Medical Image Analysis: Integrating LLM and RAG Deep Learning Strategies	Hanrui Yan e Dan Shao	2025
44	Optimizing Large Language Models for Arabic Healthcare Communication: A Focus on Patient-Centered NLP Applications	Rasheed Mohammad, Omer S. Alkhnbashi, Mohammad	2024

		Hammoudeh	
45	Privacy-preserving large language models for structured medical information retrieval	Isabella Catharina Wiest *et al*.	2024
46	RDguru: A Conversational Intelligent Agent for Rare Diseases	Jian Yang *et al*.	2024
47	Research on Fine-Tuning Optimization Strategies for Large Language Models in Tabular Data Processing	Xiaoyong Zhao *et al*.	2024
48	TAT-LLM: A Specialized Language Model for Discrete Reasoning over Financial Tabular and Textual Data	Fengbin Zhu *et al*.	2024
49	Toward Regulatory Compliance: A few-shot Learning Approach to Extract Processing Activities	Pragyan K. C. *et al*.	2024
50	Visual-textual integration in LLMs for medical diagnosis: A preliminary quantitative analysis	Reem Agbareia *et al*.	2025
51	WaterGPT: Training a Large Language Model to Become a Hydrology Expert	Yi Ren *et al*.	2024
52	What did the occupant say? Fine-tuning and evaluating a large language model for efficient analysis of multi-domain indoor environmental	Abdul-Manan Sadick e Giorgia Chinazzo	2025
53	Zero Shot Classification of Art with Large Language Models	Tatsuya Tojima e Mitsuo Yoshida	2025

Fonte: Elaborado pelas autoras (2025).

Todas as 53 publicações analisadas foram redigidas em língua inglesa. Dentre elas, 18 disponibilizaram artefatos, sendo que, em 2 casos (ID 10 e ID 31), o acesso é concedido apenas mediante solicitação aos autores. Quanto ao período de publicação, foi identificado 1 estudo publicado em 2022 (ID 19), 36 em 2024 e 16 em 2025.

Com relação às instituições às quais os autores estão vinculados, estas estão distribuídas em 26 países. Os Estados Unidos e a China possuem 16 ocorrências cada, com 60,4% do total, seguidos pela Coreia do Sul (5 ocorrências) e por Japão e Reino Unido (3 ocorrências cada).

Foram identificadas 13 grandes áreas do conhecimento contemplando as temáticas abordadas nos estudos. A categorização da área de conhecimento foi realizada pelas autoras, com base na Tabela de Áreas do Conhecimento da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), adotando-se o primeiro nível, Grande Área, a partir da análise do conteúdo de cada estudo. A área de Ciência da Computação concentrou o maior número de publicações, com 20 estudos, seguida por Medicina (12), Economia (5), Administração (4), Desenho

Industrial (3) e Psicologia (2). As áreas de Ciências Biológicas, Planejamento Urbano e Regional, Agronomia, Engenharia Elétrica, Astronomia, Engenharia Civil e Artes foram representadas por apenas um estudo cada.

Com relação às técnicas utilizadas nos modelos, foram identificadas 16 categorias de técnicas de IA aplicadas aos experimentos com os modelos. Entre essas técnicas, *fine-tuning* foi utilizada em 18 estudos, Retrieval-Augmented Generation (RAG) em 12, engenharia de *prompt* em 10, *data augmentation* em 3 e *in-context learning* em 2. Outras técnicas apareceram em apenas um estudo cada: *pretraining*, arquitetura de LLMs, mecanismo de atenção, *contrastive learning*, aprendizado *cross-modal*, *active learning*, *model augmentation*, *embeddings*, *agentic architecture*, *instruction-tuning* e *continual learning*. Cabe destacar que alguns estudos combinaram mais de uma dessas categorias como técnica principal.

Para facilitar o entendimento e a análise, foi realizada a padronização dos nomes dos modelos, que unificou versões e quantidades de parâmetros quando pertencentes à mesma família. Observou-se que alguns estudos indicam explicitamente a versão e a quantidade de parâmetros utilizados, enquanto outros não fornecem essa informação. Considerando apenas os modelos, sem variações de parâmetros, identificou-se um total de 57 modelos distintos, dos quais 28 são de código aberto e 29 são proprietários. Parte desses modelos foi empregada diretamente nos experimentos, enquanto outros foram utilizados apenas em testes para comparação com o sistema desenvolvido. A família de modelos mais utilizada foi o GPT, com 33 ocorrências, seguida por LLaMA, com 31. O Quadro 5 apresenta a relação completa dos modelos e suas respectivas frequências de ocorrência.

Quadro 5 - Modelos utilizados nas pesquisas

Modelo	Quantidade
GPT	33
LLaMA	31
BERT	13
Mistral	10
Claude, Falcon, Flan-T5, GTE, Gemini, Gemma e Qwen	3
Ada, BLIP, ChatGLM, DaVinci, DeepSeek, InternLM2, Japanese-large-lm, MPT e Nekomata	2
AWS Titan, BART, Babbage, Baichuan, Bedrock, CLIP, Curie, GEITje, InstructBLIP, Llm-jp, Meditron, MiniMax, PaLM2 e Phi e Yi	1

Fonte: Elaborado pelas autoras (2025).

De modo geral, os resultados mostram um panorama de produção científica

concentrado em países com forte tradição em pesquisa e inovação, como Estados Unidos e China, que juntos respondem por grande parte das instituições e publicações identificadas. A diversidade temática é ampla, abrangendo desde áreas consolidadas, como Ciência da Computação e Medicina, até campos mais específicos, como Astronomia e Artes, o que demonstra o potencial de aplicação dos LLMs em diferentes contextos. Além disso, a variedade de áreas de aplicação indica que, embora haja predominância de alguns setores, como Tecnologia e Medicina, há espaço para exploração em domínios menos representados. Esses resultados reforçam a relevância do tema e o interesse de pesquisadores de diferentes campos na integração de LLMs em múltiplas áreas do conhecimento e da prática profissional.

5.2 Análise qualitativa

As pesquisas selecionadas nesta revisão de literatura, em sua maioria, apresentam estudos de caso que aplicam LLMs para solucionar problemas específicos ou promover melhorias em sistemas já existentes. Parte dos estudos precisou lidar com grandes volumes de dados nas etapas iniciais do desenvolvimento dos sistemas, como em ID 4 e 21. Em alguns casos, havia dados com diferentes formatos, como em ID 2 e 7, e, em outros, dados coletados em tempo real, como visto em ID 1 e 27. A utilização desses conjuntos de dados exigiu organização, rotulagem e limpeza, incluindo a remoção de ruídos, dados duplicados e desnecessários para os objetivos do sistema proposto.

Para enfrentar o desafio do grande volume e da heterogeneidade dos dados, os pesquisadores recorreram a diferentes técnicas de IA. Assim, neste estudo, essas abordagens foram classificadas em categorias facetadas, a fim de facilitar a análise comparativa, como mostrado na seção anterior. Para evitar ambiguidades terminológicas, optou-se por manter alguns termos em inglês, pois são mais conhecidos na área por seus nomes originais, evitando, assim, equívocos que poderiam surgir com traduções ainda não consolidadas na literatura. O Quadro 6 apresenta as categorias identificadas, sua descrição, baseada nas definições apresentadas nas pesquisas, e os estudos correspondentes a cada uma delas.

Quadro 6 - Categorias com suas respectivas descrições e estudos (ID)

Categoria	Descrição	Estudos (ID)
Fine-tuning (ajuste fino ou refinamento)	Processo de ajustar a base de conhecimento de um modelo pré-treinado utilizando um conjunto de dados adicionais de um domínio para que ele adquira maior especialização em determinado tema e/ou em uma tarefa em particular.	2, 3, 10, 12, 14, 24, 28, 32, 33, 36, 37, 41, 42, 44, 47, 48, 51 e 52
Retrieval-Augmented Generation (Geração Aumentada de Recuperação)	Método que aprimora as respostas de um modelo ao recuperar informações de fontes externas e incorporá-las como contexto nos *prompts*, permitindo que o modelo forneça respostas específicas sem a necessidade de um novo treinamento.	1, 3, 4, 5, 8, 14, 15, 16, 27, 29, 43 e 46
Engenharia de *prompt* (engenharia de comando ou instrução)	Técnica voltada para a formulação de instruções que orientam o comportamento do modelo, com o objetivo de gerar respostas mais alinhadas às necessidades do usuário e reduzir ocorrências de alucinação. Essa prática pode incluir, no *prompt*, o uso de exemplos e/ou cadeias de raciocínio.	6, 19, 21, 22, 23, 40, 45, 49, 50 e 53
Data augmentation (enriquecimento de dados)	Técnica para ampliar o conjunto de dados utilizados para o treinamento do modelo por meio de geração sintética, aumentando o conjunto sem a necessidade de coletar novos dados rotulados.	13, 38 e 41
In-context learning (aprendizado em contexto)	Capacidade do modelo de executar uma tarefa a partir de instruções e poucos exemplos fornecidos no próprio *prompt*, durante a inferência, sem atualizar os pesos do modelo.	17 e 26
Pretraining (pré-treinamento)	Etapa inicial de treinamento do modelo com grandes volumes de dados, na qual esse modelo aprende representações gerais antes de ser especializado.	7
Arquitetura de LLMs	Organização interna do modelo, normalmente em variantes *decoder-only* (apenas decodificação) ou *encoder-decoder* (codificador-decodificador), que define a sua capacidade, o número de camadas e a janela de contexto (espaço onde o usuário insere o *prompt*).	9
Mecanismo de atenção	Componente que identifica a importância de cada *token* em relação aos demais, permitindo ao modelo usar o contexto e compreender dependências mesmo em trechos longos.	11
Contrastive learning	Método de treino em que o modelo aprende a aproximar exemplos que têm relação entre si e a afastar os que	18

(aprendizado contrastivo)	não têm, ajudando a melhorar buscas, classificações e o alinhamento semântico entre itens.	
Aprendizado *cross-modal* (aprendizado intermodal)	Técnica que combina diferentes tipos de dados, como texto, imagem ou áudio, permitindo que o modelo realize tarefas entre variadas fontes, como buscar uma imagem a partir de uma descrição em texto.	20
Active learning (aprendizado ativo)	Estratégia em que o modelo escolhe os exemplos mais úteis ou incertos para serem revisados e categorizados por humanos, reduzindo a quantidade de dados necessários para treinar o modelo com qualidade.	25
Model augmentation (enriquecimento de modelo)	Uso de módulos ou ferramentas externas conectadas ao modelo para ampliar suas capacidades sem precisar treiná-lo novamente.	30
Embeddings (representações vetoriais)	Representações numéricas de textos, imagens ou outros dados em um espaço vetorial, permitindo medir a similaridade semântica entre esses elementos.	31
Agentic architecture (arquitetura agentiva)	Forma de organizar modelos para que funcionem como agentes capazes de agir sozinhos ou com pouca supervisão, interagir com ferramentas, dividir tarefas em etapas e tomar decisões com base em objetivos.	34
Instruction-tuning (ajuste ou refinamento por instruções)	Processo em que o modelo é treinado ou refinado com pares de instruções e respostas para alinhar as respostas geradas a comportamentos desejados pelo usuário em tarefas de linguagem natural.	35
Continual learning (aprendizado contínuo)	Abordagem que permite ao modelo aprender de forma incremental, incorporando novos dados e tarefas ao longo do tempo, sem esquecer conhecimentos previamente adquiridos.	39

Fonte: Elaborado pelas autoras (2025).

No que diz respeito às limitações, observa-se que a aplicação de LLMs em domínios específicos pode não refletir as características de outros contextos, o que compromete a generalização e não garante o mesmo nível de desempenho alcançado. Outro aspecto recorrente é a exigência de recursos computacionais substanciais para o treinamento e a utilização dos modelos, como em ID 2, somada às limitações de infraestrutura relatadas em alguns estudos, como em ID 4 e 22. Há também estudos de caráter teórico, que não chegaram a ser aplicados em ambientes reais, o que limita sua validação prática, como em ID 1 e 42.

Entre os demais desafios relatados estão: o risco de alucinações nas respostas

geradas, como em ID 4 e 6, potenciais vieses e problemas éticos, como apresentados em ID 5 e 43, a dependência de usuários em relação às soluções de IA, como mostrado em ID 6, e a carência de conjuntos de dados adequados para domínios específicos, como relatado em ID 7 e 42.

Apesar dessas limitações, os trabalhos apresentam contribuições significativas. Entre elas estão o desenvolvimento de metodologias replicáveis para análise e aplicação de modelos em diferentes domínios, como em ID 2 e 52, a automatização de processos em diferentes contextos, como apresentado em ID 5 e 53, e a criação de conjuntos de dados voltados a domínios especializados, como nos estudos de ID 8 e 9.

Após esta análise, foi possível estabelecer relações das técnicas de IA apresentadas nos estudos, com foco nas abordagens que utilizam LLMs para o tratamento de grandes volumes de dados, com os quatro sistemas de Arquitetura da Informação propostos por Rosenfeld, Morville e Arango (2015).

5.3 Análise na perspectiva da Arquitetura da informação em LLMs

Segundo Rosenfeld, Morville e Arango (2015), os **sistemas de organização** envolvem esquemas, que determinam as características comuns entre os itens para posterior agrupamento, e estruturas, que estabelecem as relações entre esses itens e seus grupos. Esses dois elementos influenciam a forma como a informação é localizada e compreendida pelos usuários. Em relação às técnicas de IA para organização, *fine-tuning, pretraining, data augmentation, embeddings, contrastive learning, continual learning* e *model augmentation* são aplicáveis, pois estruturam o conjunto de dados, refinam categorias, enriquecem exemplos, produzem representações semânticas e permitem a evolução contínua da organização da informação. Nos estudos de Noels, De Blaere e De Bie (2024), por exemplo, os dados foram coletados de diversas fontes financeiras, como notícias, relatórios e postagens em redes sociais. Esses dados foram traduzidos para o holandês e, então, passaram por um processo de tratamento. Após esta etapa, os dados, já organizados, foram utilizados para refinar o modelo para o domínio financeiro. Seguindo uma outra abordagem, Rangan e Yin (2024) utilizaram dados de uma única fonte, extraídos de documentos em PDF, e criaram pares de perguntas e respostas para refinar um

modelo para atuar como um *chatbot* que respondesse a perguntas sobre serviços sociais para pessoas em situação de vulnerabilidade.

Já os sistemas de **rotulação** envolvem a representação da informação e funcionam como a forma mais evidente de revelar os esquemas de organização adotados. Para serem eficazes, esses sistemas devem utilizar a mesma linguagem dos usuários, ao mesmo tempo em que refletem de maneira clara o conteúdo que representam. Entre as técnicas aplicadas neste tipo de sistema, destacam-se engenharia de *prompt, instruction-tuning, contrastive learning* e *in-context learning.* Nos estudos de Jaimovitch-López *et al*. (2022), a técnica de engenharia de *prompt* foi aplicada para automatizar tarefas de manipulação de dados que normalmente demandam um tempo significativo dos profissionais da área. Por meio da criação de *prompts*, elaborados com e sem exemplos de respostas, os autores ajustaram o comportamento do modelo para gerar saídas alinhadas às necessidades dos usuários. De forma semelhante, Elahi e Taghvaei (2024) utilizaram engenharia de *prompt* em dados financeiros, combinando informações rotuladas, como a descrição da empresa e de seu setor, para contextualizar e refinar as respostas do modelo.

Os **sistemas de navegação** têm como objetivo ajudar os usuários a se orientarem nos ambientes informacionais, oferecendo contexto, localização e caminhos de retorno. Observa-se que RAG, *agentic architectures*, *cross-modal learning*, *active learning* e arquiteturas próprias de LLMs se enquadram nesse sistema, pois ampliam os percursos de exploração e criam fluxos multimodais para conectar informações. Nos estudos de Choi e Jeong (2025a), foi proposto um modelo conceitual baseado em RAG para uso na manufatura inteligente. A arquitetura proposta possibilitou a coleta de dados em tempo real a partir de sensores industriais, a preservação do contexto temporal e semântico e a atualização constante do conhecimento, permitindo que os usuários navegassem pelos dados de forma contextualizada e dinâmica.

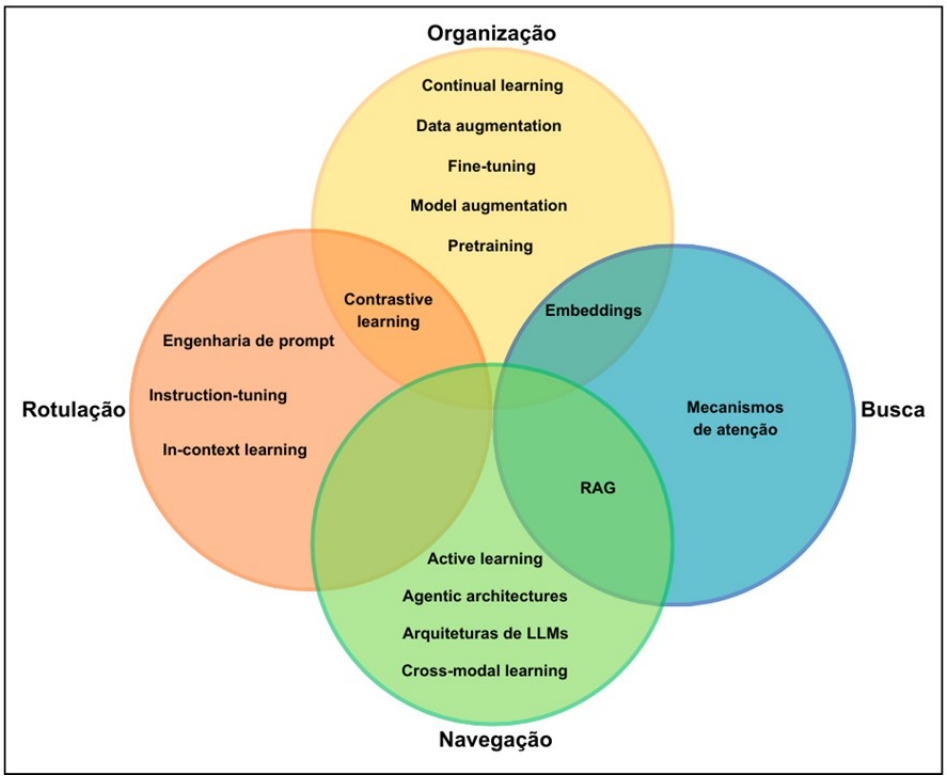
Adotando uma estratégia diferente, Wu *et al*. (2024) desenvolveram um sistema voltado à análise de documentos regulatórios da Food and Drugs Administration. O modelo, integrado a um mecanismo de RAG, permite que o usuário navegue pelos documentos por meio de fluxos de busca semântica, conectando trechos específicos a respostas geradas pelo modelo, garantindo precisão e confiabilidade.

Por fim, há os **sistemas de busca**, que são um recurso essencial para localizar

informações. A definição do que será indexado, a escolha dos algoritmos e as diferentes formas de apresentação dos resultados são fatores importantes para a eficácia da pesquisa, todos eles reunidos na interface que conecta o usuário ao sistema. *Embeddings*, mecanismos de atenção, e, novamente, RAG podem ser enquadrados nesse sistema. Nos estudos de Zhou *et al*., (2024), foi proposto um *framework* de proteção de privacidade que combina LLMs com mecanismos de atenção ajustáveis. Nessa abordagem, dados sensíveis de textos e imagens foram processados de forma a permitir que o sistema recuperasse informações relevantes sem comprometer a privacidade do usuário. A busca, nesse caso, foi estruturada para equilibrar precisão, *recall* e segurança, garantindo que apenas trechos adequados fossem retornados ao usuário. De forma distinta, Memduhoglu, Fulman e Zipf (2024) aplicaram *embeddings* de LLMs para enriquecer a classificação funcional de edificações urbanas. Com a integração de atributos físicos, espaciais e morfológicos, os autores demonstraram que os *embeddings* tornaram a busca e a recuperação mais precisas, superando técnicas tradicionais de PLN.

A Figura 1 apresenta um diagrama de Venn que ilustra a interseção entre os quatro sistemas da Arquitetura da Informação propostos por Rosenfeld, Morville e Arango (2015) - organização, rotulagem, navegação e busca - e as técnicas de IA identificadas neste estudo. Observa-se que três dessas técnicas são comuns a mais de um sistema, enquanto a maioria (13) apresenta associação específica a apenas um dos sistemas.

Figura 1 - Interseção entre os sistemas da Arquitetura da Informação e técnicas de IA



Fonte: Elaborado pelas autoras (2025).

Diante do exposto, é possível observar a relação entre os quatro sistemas de Arquitetura da Informação propostos por Rosenfeld, Morville e Arango (2015) e as técnicas de IA aplicadas ao tratamento de grandes volumes de dados descritas nos estudos, que podem atuar como embasamento teórico para compreender e orientar o desenvolvimento de soluções de IA que lidam com grandes volumes de dados.

Outro ponto relevante a ser destacado é que a CI precisa se apropriar das discussões relacionadas à informação, mesmo quando elas se manifestam em outros campos do conhecimento. Como afirma Araújo (2018, p. 7), “a CI que se faz hoje é muito diferente daquela de cinco décadas atrás”. Com o advento e a popularização das tecnologias, alguns problemas que motivaram a origem da área foram solucionados. Entretanto, novos desafios surgiram, relacionados às dimensões humanas da produção e do uso da informação. A CI contemporânea, conforme ressalta o autor, é mais atenta à complexidade dos fenômenos informacionais, considerando a relação entre registros do conhecimento, mediações tecnológicas e institucionais e saberes coletivos. Isso ressalta a necessidade de os profissionais da CI se inserirem ativamente em projetos e debates interdisciplinares, assumindo seu

papel estratégico na compreensão e mediação dos fluxos informacionais.

6 Considerações finais

Este estudo teve como objetivo identificar e analisar a organização e a arquitetura da informação em LLMs, a fim de compreender de que maneira a informação vem sendo organizada, tratada e estruturada em LLMs, identificando as principais técnicas de IA utilizadas nesse processo e analisando como essas técnicas se relacionam com os sistemas de Arquitetura da Informação. A partir de uma revisão de literatura, foi possível mapear um conjunto de categorias técnicas relevantes, como *fine-tuning*, RAG e engenharia de *prompt*, e relacioná-las aos quatro sistemas da Arquitetura da Informação propostos por Rosenfeld, Morville e Arango (2015), sendo eles a organização, a rotulagem, a navegação e a busca. A revisão permitiu observar que a Arquitetura da Informação fornece uma base conceitual para compreender como os LLMs estruturam, representam e disponibilizam informações de grandes volumes de dados.

Como contribuição, o trabalho indica a existência de técnicas variadas e aspectos relevantes da organização e a arquitetura da informação. Esses achados são essenciais para conduzir novos estudos e apoiar projetos em andamento, como no caso desta pesquisa, vinculada ao projeto Dados abertos para potencializar a recuperação de informação em modelos de Inteligência Artificial. Além disso, este estudo representa um passo relevante para a alfabetização digital, ao oferecer subsídios que ajudam os usuários a compreenderem o funcionamento de sistemas de IAG.

Para a CI, este estudo contribui ao evidenciar como a organização e a arquitetura da informação, campos interdisciplinares que dialogam com fundamentos da Organização e Representação da Informação e com instrumentos tradicionais da área, como metadados, tesauros e taxonomias, podem ser aplicados para que se compreenda de que maneira os LLMs organizam, representam e disponibilizam dados. Ao relacionar técnicas de IA com os sistemas de organização, rotulagem, navegação e busca, o trabalho oferece uma nova perspectiva para analisar esses modelos. Além disso, o estudo reforça a relevância da CI nos debates atuais sobre IA e destaca o papel do profissional da informação na mediação entre a tecnologia e a

sociedade.

Entretanto, algumas limitações devem ser destacadas: o estudo restringiu-se a uma revisão bibliográfica, sem a realização de análises em ambientes práticos, e foi influenciado pela rápida evolução tecnológica dos modelos de IA, que constantemente atualizam suas funcionalidades e possibilidades de aplicação. Nesse sentido, pesquisas futuras podem abordar o acompanhamento de projetos reais, observando como o tratamento e a organização dos dados se dão nessas situações. Além disso, torna-se relevante a realização de pesquisas que relacionem não apenas os aspectos técnicos, mas também as dimensões sociais e éticas envolvidas no uso dessas tecnologias.

Referências

ADESEYE, A.; ISOAHO, J.; MOHAMMAD, T. LLM-assisted qualitative data analysis: security and privacy concerns in gamified workforce studies. **Procedia Computer Science**, [S.l.], v. 257, p. 60-67, 2025. DOI: 10.1016/j.procs.2025.03.011.

AGBAREIA, R. *et al*. Visual-textual integration in LLMs for medical diagnosis: a preliminary quantitative analysis. **Computational and Structural Biotechnology Journal**, *[S.l.]*, v. 27, p. 184-189, 2025. DOI: 10.1016/j.csbj.2024.12.019.

ALAMMAR, J.; GROOENDORST, M. **Hands-On Large Language Models**. Sebastopol: O'Reilly, 2024.

ALGHAMDI, H.; MOSTAFA, A. Advancing EHR analysis: Predictive medication modeling using LLMs. **Information Systems**, [S*.l.*], v. 131, 102528, 2025. DOI: 10.1016/j.is.2025.102528.

ARAÚJO, C. A. A. **O que é Ciência da Informação?** Belo Horizonte: KMA, 2018. Disponível em: https://teste.eci.ufmg.br/wp-content/uploads/2024/03/O-QUE-E-CIENCIA-DA-INFORMACAO.pdf. Acesso em: 28 ago. 2025.

BADALOTTI, D. *et al*. Development of a natural language processing (NLP) model to automatically extract clinical data from electronic health records: results from an Italian comprehensive stroke center. **International Journal of Medical Informatics**, [S*.l.*], v. 192, 105626, 2024. DOI: 10.1016/j.ijmedinf.2024.105626.

BARDIN, L. **Análise de conteúdo**. São Paulo: Edições 70, 2016.

CHEN, K. *et al*. A Survey on Privacy Risks and Protection in Large Language Models. **arXiv**:2505.01976v1, 2025. DOI: 10.48550/arXiv.2505.01976.

CHEN, K. *et al.* LLM-assisted multi-teacher continual learning for visual question answering in robotic surgery. **arXiv**:2402.16664v3, 2024. DOI: 10.48550/arXiv.2402.16664.

CHEN, L. *et al*. Application of retrieval-augmented generation for interactive industrial knowledge management via a large language model. **Computer Standards & Interfaces**, [*S.l.*], v. 94, 103995, 2025. DOI: 1016/j.csi.2025.103995.

CHO, H. *et al*. Extracting lung cancer staging descriptors from pathology reports: a generative language model approach. **Journal of Biomedical Informatics**, *[S.l.]*, v. 157, 104720, 2024. DOI: 10.1016/j.jbi.2024.104720.

CHO, N. *et al*. FISHNET: financial intelligence from sub-querying, harmonizing, neural-conditioning, expert swarms, and task planning. *In*: ACM International Conference on AI in Finance (ICAIF '24), 5., 2024, Brooklyn. **Proceedings** [...]. *[S.l.]:* ACM, 2024. p. 1-9. DOI:1145/3677052.3698597.

CHOI, H.; JEONG, J. A Conceptual Framework for a Latest Information-Maintaining Method Using Retrieval-Augmented Generation and a Large Language Model in Smart Manufacturing: Theoretical Approach and Performance Analysis. **Machines**, Basel, v. 13, n. 94, 2025a. DOI: 10.3390/machines13020094.

CHOI, H.; JEONG, J. Domain-specific manufacturing analytics framework: an integrated architecture with retrieval-augmented generation and Ollama-based models for manufacturing execution systems environments. **Processes**, [S.*l.*], v. 13, n. 670, 2025b. DOI: 10.3390/pr13030670

CHOW, A. R. ChatGPT May Be Eroding Critical Thinking Skills, According to a New MIT Study. **Time**, 23 jun. 2025. Disponível em: https://time.com/7295195/ai-chatgpt-google-learning-school/. Acesso em: 30 jul. 2025.

CORRÊA, L. N. **Metodologia científica**: Para trabalhos acadêmicos e artigos. Florianopolis: Do Autor, 2008.

COSTA, D. G. *et al*. A method to promote safe cycling powered by large language models and AI agents. **MethodsX**, [*S.l.*], v. 13, 102880, 2024. DOI: 10.1016/j.mex.2024.102880.

CRESWELL, J. W.; CRESWELL, J. D. **Projeto de pesquisa**: métodos qualitativo, quantitativo e misto. 5. ed. Porto Alegre: Penso, 2021.

DHAMANI, N.; ENGLER, M. **Introduction to Generative AI**. Shelter Island: Manning, 2024.

ELAHI, A.; TAGHVAEI, F. Combining financial data and news articles for stock price movement prediction using large language models. **arXiv**:2411.01368v1, 2024. DOI: 10.48550/arXiv.2411.01368.

FERREIRA, S. A; OLIVEIRA, D. A. Desinformação, crise de confiança na ciência e necessidade de políticas para divulgação científica. **Ciência da Informação Express**, v. 2, n. 4, p. 1-6, 9 abr. 2021. DOI: 10.60144/v2i.2021.28.

FU, X. *et al*. Deciphering public voices in the digital era: benchmarking ChatGPT for analyzing citizen feedback in Hamilton, New Zealand. **Journal of the American**

Planning Association, *[S.l.*], v. 90, n. 4, p. 728-741, 2024. DOI: 10.1080/01944363.2024.2309259.

GABRIEL-PETIT, P. **Designing Information Architecture**: A practical guide to structuring digital content for findability and easy navigability. Birmingham: Packt Publishing, 2025.

HAN, B.; SUSNJAK, T.; MATHRANI, A. Automating systematic literature reviews with retrieval-augmented generation: A comprehensive overview. **Applied Sciences**, [*S.l.*], v. 14, n. 9103, 2024. DOI: 10.3390/app14199103.

JAIMOVITCH-LÓPEZ, G. *et al*. Can language models automate data wrangling? **Machine Learning**, [S.*l.*], v. 112, p. 2053-2082, 2022. DOI: 10.1007/s10994-022-06259-9.

JEONG, Y. *et al*. Advancing tinnitus therapeutics: GPT-2 driven clustering analysis of cognitive behavioral therapy sessions and Google T5-based predictive modeling for THI score assessment. **IEEE Access**, [*s.l.*], v. 12, p. 52414-52429, 2024. DOI: 10.1109/ACCESS.2024.3383020.

KALLENS, P. C.; KRISTENSEN-MCLACHLAN, R. D.; CHRISTIANSEN, M. H. Large Language Models Demonstrate the Potential of Statistical Learning in Language. **Cognitive Science**, *[S.l.],* v. 47, n. 3, 2023. DOI: 10.1111/cogs.13256.

LI, H. *et al.*, A review on enhancing agricultural intelligence with large language models. **Artificial Intelligence in Agriculture**, [S. l.], v. 15, p. 671-685, 2025. DOI: 10.1016/j.aiia.2025.05.006.

LI, Y. *et al*. Deep learning and methods based on large language models applied to stellar light curve classification. **Intelligent Computing**, [*s.l.*], v. 4, 0110, 2025. DOI: 10.34133/icomputing.0110.

LIU, F. *et al*. Aggregated knowledge model: Enhancing domain-specific QA with fine-tuned and retrieval-augmented generation models. *In*: INTERNATIONAL CONFERENCE ON AI-ML SYSTEMS, 4., 2024, Baton Rouge. **Proceedings** [...]. New York: ACM, 2024. p. 1-7. DOI: 10.1145/3703412.3703434.

LIU, J. *et al*. Fisher information-based efficient curriculum federated learning with large language models. **arXiv**:2410.00131v2, 2024. DOI: 10.48550/arXiv.2410.00131.

LIU, X. *et al*. Military equipment entity extraction based on large language model. **Applied Sciences**, *[s.l.]*, v. 14, n. 9063, 2024. DOI: 10.3390/app14199063.

LIMA, G. A. B. O. Arquitetura da Informação. *In*: Miranda, R. C. R. **Arquitetura da informação na Câmara dos Deputados**. Brasília: Câmara dos Deputados, 2016. Cap. 2, p. 45-63.

LUKICHEV, M. Understanding data quality's impact on Large Language Models. **Medium**, 18 jul. 2024. Disponível em: https://medium.com/telmai1/understanding-

data-qualitys-impact-on-large-language-models-01e9e54a5017. Acesso em: 30 jul. 2025.

NASCIMENTO SILVA, P. Recuperação de Informação na Ciência da Informação: produção acadêmico-científica brasileira (2012-2021). **Transinformação**, Campinas, v. 35, p. 1–17, 2023. Disponível em: https://periodicos.puc-campinas.edu.br/transinfo/article/view/7336. Acesso em: 19 fev. 2026.

MACIE, G. C.; NASCIMENTO, N. M.; MADIO, T. C. C. Arquitetura e recuperação da informação: uma abordagem do Sistema Integrado de Gestão Acadêmica (SIGA) da Universidade Eduardo Mondlane. **Em Questão**, Porto Alegre, v. 30, e-139451, 2024. DOI: 10.1590/1808-5245.30.139451.

MATIAS-PEREIRA, J. **Manual de metodologia da pesquisa científica**. 4. ed. São Paulo: Atlas, 2019.

MEMDUHOĞLU, A.; FULMAN, N.; ZIPF, A. Enriching building function classification using large language model embeddings of OpenStreetMap Tags. **Earth Science Informatics**, *[*S.l.], v. 17, p. 5403-5418, 2024. DOI: 10.1007/s12145-024-01463-8.

MISHRA, M. *et al*. DataAgent: evaluating large language models' ability to answer zero-shot, natural language queries. **arXiv**:2404.00188v1, 2024. DOI: 10.48550/arXiv.2404.00188.

MOHAMMAD, R.; ALKHNBASHI, O. S.; HAMMOUDEH, M. Optimizing large language models for Arabic healthcare communication: a focus on patient-centered NLP applications. **Big Data and Cognitive Computing**, *[*S.l.*]*, v. 8, n. 157, 2024. DOI: 10.3390/bdcc8110157.

NOELS, S; DE BLAERE, J; DE BIE, T. A Dutch Financial Large Language Model. *In*: ACM INTERNATIONAL CONFERENCE ON AI IN FINANCE, 5., 2024, Brooklyn, NY. **Proceedings** [...]. New York: ACM, 2024. p. 1-9. DOI: 10.1145/3677052.3698628.

PAPAGEORGIOU, G. *et al*. Enhancing e-government services through state-of-the-art, modular, and reproducible architecture over large language models. **Applied Sciences**, [S.*l.*], v. 14, n. 8259, 2024. DOI: 10.3390/app14188259.

PAIS, C. *et al*. Large language models for preventing medication direction errors in online pharmacies. **Nature Medicine**, *[s.l.]*, v. 30, n. 6, p. 1574-1582, 2024. DOI: 10.1038/s41591-024-02933-8.

PRAGYAN, K. C. *et al*. Toward regulatory compliance: a few-shot learning approach to extract processing activities. *In*: IEEE International Requirements Engineering Conference Workshops, 32., 2024. **Proceedings** [...]. [S.l.]: IEEE, 2024. p. 241-248. DOI: 10.1109/REW61692.2024.00038.

RANGAN, K.; YIN, Y. A fine-tuning enhanced RAG system with quantized influence measure as AI judge. **Scientific Reports**, [S. *l.*], v. 14, n. 27446, 2024. DOI: 10.1038/s41598-024-79110-x.

REDDY, V. *et al*. DocFinQA: a long-context financial reasoning dataset. **arXiv**:2401.06915v2, 2024. DOI: 10.48550/arXiv.2401.06915.

REN, Y. *et al*. WaterGPT: training a large language model to become a hydrology expert. **Water**, [S.l.], v. 16, n. 3075, 2024. DOI: 10.3390/w16213075.

ROSENFELD, L.; MORVILLE, P.; ARANGO, J. **Information Architecture**. 4th. ed. Sebastopol: O'Reilly, 2015.

SADICK, A. M.; CHINAZZO, G. What did the occupant say? Fine-tuning and evaluating a large language model for efficient analysis of multi-domain indoor environmental quality feedback. **Building and Environment**, *[*S.l.], v. 274, 112735, 2025. DOI: 10.1016/j.buildenv.2025.112735.

SARZAEIM, P.; MAHMOUD, Q. H.; AZIM, A. A Framework for LLM-Assisted Smart Policing System. **IEEE Access**, [S*.l.*], v. 12, p. 74915-74929, 2024. DOI: 10.1109/ACCESS.2024.3404862.

SONG, Z. *et al*. A scientific-article key-insight extraction system based on multi-actor of fine-tuned open-source large language models. **Scientific Reports**, [*s. l.*], v. 15, n. 1608, 2025. DOI: 10.1038/s41598-025-85715-7.

STANLEY, J. *et al*. Large language models deconstruct the clinical intuition behind diagnosing autism. **Cell**, *[S.l.]*, v. 188, p. 2235-2248, 2025. DOI: 10.1016/j.cell.2025.02.025.

SUNG, C.; LEE, Y.; TSAI, Y. A New Pipeline for Generating Instruction Dataset via RAG and Self Fine-Tuning. *In*: IEEE ANNUAL COMPUTERS, SOFTWARE, AND APPLICATIONS CONFERENCE, 48., 2024. **Proceedings** [...]. [S.l.]: IEEE, 2024. DOI: 10.1109/COMPSAC61105.2024.00371.

TANABE, K. *et al*. Enhancing financial domain adaptation of language models via model augmentation. **arXiv**:2411.09249v1, 2024. DOI: 10.48550/arXiv.2411.09249.

TANABE, K. *et al*. JaFIn: Japanese financial instruction dataset. *In*: IEEE Symposium on Computational Intelligence for Financial Engineering and Economics, 2024. **Proceedings** [...]. *[S.l.]*: IEEE, 2024. p. 1-8. DOI: 10.1109/CIFEr62890.2024.10772973.

TAYLOR, N. et al. Model development for bespoke large language models for digital triage assistance in mental health care. **Artificial Intelligence in Medicine**, [S.l.], v. 157, 102988, 2024. DOI: 10.1016/j.artmed.2024.102988.

TOJIMA, T.; YOSHIDA, M. Zero-shot classification of art with large language models. **IEEE Access**, *[S.l.],* v. 13, p. 17426-17435, 2025. DOI: 10.1109/ACCESS.2025.3532995.

TONG, X. *et al*. LEGF-DST: LLMs-enhanced graph-fusion dual-stream transformer for fine-grained Chinese malicious SMS detection. **Computers, Materials & Continua (CMC)**, [S.l.], v. 82, n. 2, p. 1902-1919, 2025. DOI: 10.32604/cmc.2024.059018.

TORINO, E. *et al*. A relação entre a arquitetura da informação e experiência do usuário sob a ótica dos pesquisadores da Ciência da Informação brasileira. **Biblos**: Revista do Instituto de Ciências Humanas e da Informação, Rio Grande, v. 36, n. 1, p. 219-237, 2022. DOI: 10.14295/biblos.v36i1.13769.

VALLE, P. R. D.; FERREIRA, J. L. Análise de conteúdo na perspectiva de Bardin: contribuições e limitações para a pesquisa qualitativa em Educação. **Educação em Revista**, Belo Horizonte, v. 41, e49377, 2025. DOI: http://dx.doi.org/10.1590/0102-469849377.

WANG, H. *et al*. DRG-LLaMA: tuning LLaMA model to predict diagnosis-related group for hospitalized patients. **npj Digital Medicine**, [S.l.], v. 7, n. 16, 2024. DOI: 10.1038/s41746-023-00989-3.

WANG, T. *et al*. A Multimodal Large Language Model Framework for Intelligent Perception and Decision-Making in Smart Manufacturing. **Sensors**, Basel, v. 25, n. 3072, 2025. Disponível em: https://www.mdpi.com/1424-8220/25/10/3072. Acesso em: 26 ago. 2025.

WEI, W. R.; HUANG, L.; WANG, J. J. **Retrieval-Augmented Generation for LLM Applications**: transforming search, recommendation, and AI assistants. Sebastopol, CA: O'Reilly, 2025.

WIEST, I. C. *et al*. Privacy-preserving large language models for structured medical information retrieval. **NPJ Digital Medicine**, [S.l.], v. 7, n. 257, 2024. DOI: 10.1038/s41746-024-01233-2.

WU, L. *et al*. A framework enabling LLMs into regulatory environment for transparency and trustworthiness and its application to drug labeling document. **Regulatory Toxicology and Pharmacology**, [*s l.*], v. 149, 105613, 2024. DOI: 10.1016/j.yrtph.2024.105613.

YAN, H.; SHAO, D. Multimodal medical image analysis: integrating LLM and RAG deep learning strategies. **Journal of Advances in Information Technology**, *[*S.l.], v. 16, n. 4, p. 568-581, 2025. DOI: 10.12720/jait.16.4.568-581.

YANG, J. *et al*. RDguru: a conversational intelligent agent for rare diseases. **IEEE Journal of Biomedical and Health Informatics**, *[*S.l.*]*, p. 1-12, 2024. DOI: 10.1109/JBHI.2024.3464555.

ZHAO, X. *et al*. Research on fine-tuning optimization strategies for large language models in tabular data processing. **Biomimetics**, *[*S.l.], v. 9, n. 708, 2024. DOI: 10.3390/biomimetics9110708.

ZHANG, K. *et al*. Batch-ICL: Effective, efficient, and order-agnostic in-context learning. **arXiv**:2401.06469v3, 2024. DOI: 10.48550/arXiv.2401.06469.

ZHENG, J.; WANG, H.; YAO, J. Building lightweight domain-specific consultation systems via inter-external knowledge fusion contrastive learning. **IEEE Access**, [S.*l.*], v. 12, p. 113244-113259, 2024. DOI: 10.1109/ACCESS.2024.3434648.

ZHOU, B.; GEIßLER, D.; LUKOWICZ, P. Misinforming LLMs: vulnerabilities, challenges and opportunities. **arXiv**:2408.01168v1, Aug. 2024. DOI: 10.48550/arXiv.2408.01168.

ZHOU, S. *et al*. A user-centered framework for data privacy protection using large language models and attention mechanisms. **Applied Sciences**, [*s.l.*], v. 14, n. 6824, 2024. DOI: 10.3390/app14156824.

ZHU, F. *et al*. TAT-LLM: a specialized language model for discrete reasoning over financial tabular and textual data. *In*: ACM International Conference on AI in Finance, 5., 2024, Brooklyn. **Proceedings** [...]. Brooklyn: ACM, 2024. p. 1-9. DOI: 10.1145/3677052.3698685.

ZHU, G. et al. CMLLM: A novel cross-modal large language model for wind power forecasting. **Energy Conversion and Management**, [S.l.], v. 330, 119673, 2025. DOI: 10.1016/j.enconman.2025.119673.

Agradecimentos

À Agência Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) pelo apoio financeiro concedido.

Ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) pelo apoio à pesquisa, processo 303721/2025-1.

Organization and information architecture in large language models

intersections in Information Science

Abstract

Objective: To identify and analyze the organization and information architecture in Large Language Models (LLMs), relating them to Artificial Intelligence (AI) techniques and discussing these findings within the context of Information Science (IS). **Methodology:** this bibliographic research consists of a literature review adopting a mixed-methods approach. A rigorous protocol was applied to investigate the topic in the ACM Digital Library, ScienceDirect, Scopus and Web of Science databases between June and August 2025. Data were analyzed both quantitatively and qualitatively through content analysis, following Bardin's framework, particularly through categorical analysis. **Results:** a total of 53 studies published between 2022 and 2025, from 26 countries, were examined, leading to the identification of 16 categories of AI techniques applied to LLMs. These techniques were related to the four Information Architecture systems proposed by Rosenfeld, Morville, and Arango: organization, labeling, navigation, and search, demonstrating contributions to the development and application of LLMs across different domains. **Conclusions:** this research contributes by highlighting how

information organization and architecture interact with artificial intelligence and traditional tools in the field, such as metadata, thesauri, and taxonomies, which can be applied to understand how LLMs organize, represent, and make data available. Furthermore, it provides an important conceptual foundation, supporting new studies and solutions involving data curation and the training of these systems. The study also reinforces the relevance of information science fundamentals in the analysis of emerging technologies.

Descriptors: Information Architecture. Large Language Models. Literature review.

Organización y arquitectura de la información en modelos de lenguaje a gran escala

intersecciones en la Ciencia de la Información

Resumen

Objetivo: Identificar y analizar la organización y la arquitectura de la información en los Large Language Models (LLMs), relacionándolas con técnicas de Inteligencia Artificial (IA) y discutiendo estos resultados en el contexto de la Ciencia de la Información (CI). **Metodología:** La investigación, de carácter bibliográfico, corresponde a una revisión de literatura con enfoque mixto. Se aplicó un protocolo riguroso para investigar la temática en las bases ACM Digital Library, ScienceDirect, Scopus y Web of Science, entre junio y agosto de 2025. Los datos fueron analizados cuantitativa y cualitativamente mediante el Análisis de Contenido, en la perspectiva de Bardin, especialmente a través del análisis categorial. **Resultados:** Se examinaron 53 estudios publicados entre 2022 y 2025, provenientes de 26 países, identificándose 16 categorías de técnicas de IA aplicadas a los LLMs. Estas técnicas fueron relacionadas con los cuatro sistemas de la Arquitectura de la Información propuestos por Rosenfeld, Morville y Arango: organización, rotulación, navegación y búsqueda, evidenciando contribuciones para el desarrollo y la aplicación de los LLMs en diferentes dominios. **Conclusiones:** La investigación contribuye a poner de manifiesto cómo la organización y la arquitectura de la información interactúan con la inteligencia artificial y con los instrumentos tradicionales del ámbito, como los metadatos, los tesauros y las taxonomías, y puede aplicarse para comprender de qué manera los LLM organizan, representan y ponen a disposición los datos. Además, ofrece una base conceptual importante, que sirve de apoyo a nuevos estudios y soluciones relacionados con la curación de datos y el entrenamiento de estos sistemas. El estudio también refuerza la relevancia de los fundamentos de la CI en el análisis de tecnologías emergentes.

Descriptores: Arquitectura de la Información. Modelos de Lenguaje a Gran Escala. Revisión de la literatura.