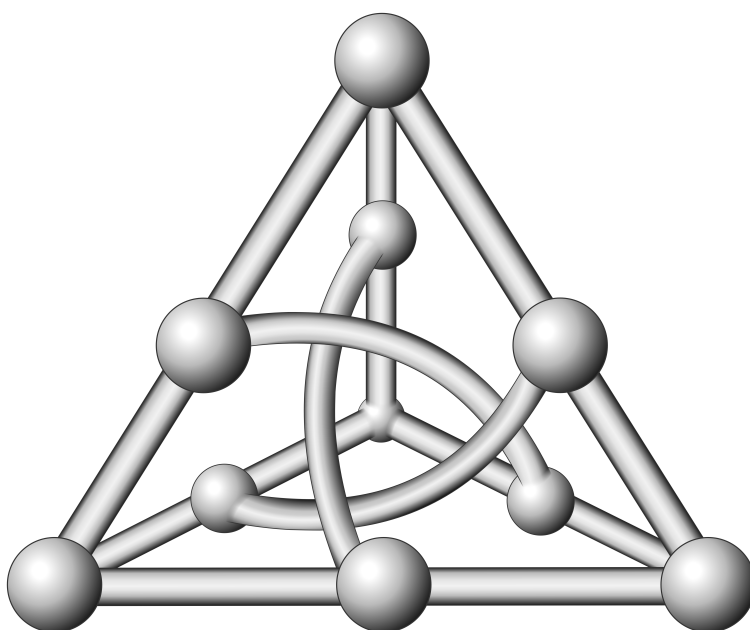


Como a Inteligência Artificial pode auxiliar a administração pública: Uma Revisão Sistemática da literatura

Guilherme Lube Machado de Melo

Universidade Federal do Mato Grosso do Sul - UFMS.



AGRADECIMENTOS

A Deus, pela minha vida, e por me permitir ultrapassar todos os obstáculos encontrados ao longo da realização deste trabalho.

Gostaria de expressar minha sincera gratidão ao professor Renato Porfirio Ishii, por sua orientação, paciência e valiosas contribuições durante o desenvolvimento deste trabalho. Seu conhecimento e incentivo foram essenciais para a conclusão desta pesquisa.

Agradeço à UFMS pela oportunidade de estudar e desenvolver meu projeto de conclusão de curso, bem como aos professores e funcionários que me ofereceram suporte ao longo desta jornada.

Agradeço também aos meus colegas de curso, em especial a Eduardo Ferreira Valim e Gustavo Felipe Gonçalves Reis, por compartilharem ideias, críticas construtivas e apoio ao longo desta trajetória.

Aos meus pais e familiares, meu agradecimento mais profundo pelo incentivo, amor e compreensão durante os momentos mais desafiadores. Sem o apoio de vocês, este trabalho não teria sido possível.

Contexto

Esta revisão sistemática explora o potencial da inteligência artificial (IA) para auxiliar o setor público em diversas funções, como a análise de grandes volumes de informação, a redação de documentos e o suporte à tomada de decisão. O estudo investiga as ferramentas e técnicas disponíveis, incluindo modelos de linguagem de grande escala (LLM) e processamento de linguagem natural (NLP). Por meio de uma análise rigorosa e da discussão das implicações, busca-se contribuir para o avanço do conhecimento nessa área crítica. O objetivo principal deste trabalho de conclusão de curso é analisar a integração da tecnologia de IA na administração pública, compreendendo seus impactos, desafios e benefícios. Entre as metas, destacam-se: posicionar as pesquisas identificadas em relação ao conhecimento existente sobre o tema; obter conhecimento útil sobre os métodos de pesquisa utilizados por outros estudiosos para estudar as implicações do uso de IA no governo e na administração pública; compreender as implicações do uso de IA no governo para a administração pública. Conclui-se que a tecnologia de inteligência artificial representa uma oportunidade promissora e desafiadora para a administração pública; neste estudo vão ser demonstrados o uso de RAG (Fleischer et al. 2024) (do inglês "geração aumentada por recuperação) e comandos para a IA, com o intuito de mostrar a sua capacidade para analisar grandes quantidades de dados e acelerar o processo de escrita de documentos. Sua implementação estratégica pode resultar em uma administração mais eficaz, em serviços públicos mais acessíveis e em uma sociedade mais conectada e informada.

This systematic review explores the potential of artificial intelligence (AI) to assist the public sector in various functions, such as analyzing large volumes of information, drafting documents, and supporting decision-making processes. The study investigates the available tools and techniques, including large language models (LLMs) and natural language processing (NLP). Through a rigorous analysis and discussion of implications, it aims to contribute to advancing knowledge in this critical area. The primary objective of this final undergraduate project is to analyze the integration of AI technology into public administration, understanding its impacts, challenges, and benefits. Among the goals are: positioning the identified research in relation to the existing knowledge on the subject; gaining useful insights into the research methods employed by other scholars to study the implications of AI use in government and public administration; and understanding the implications of AI use in government for public administration. The study concludes that artificial intelligence technology represents a promising and challenging opportunity for public administration. This research will demonstrate the use of RAG and prompts for AI to showcase its capability to analyze large datasets and accelerate the document-writing process. Its strategic implementation could result in more efficient administration, more accessible public services, and a more connected and informed society.

Problema

A produção de documentos públicos, como licitações, regulamentações e relatórios governamentais, é um processo essencial para a administração pública e o exercício da transparência governamental. No entanto, esse processo pode ser demorado, suscetível a erros humanos e requerer um conhecimento especializado significativo. A inteligência artificial (IA) surge como uma ferramenta promissora para auxiliar na produção desses documentos, automatizando tarefas repetitivas, aumentando a precisão e facilitando a análise de grandes volumes de dados (Zuiderwijk 2021).

Esta revisão sistemática da literatura visa examinar o estado atual da pesquisa sobre o uso de IA na produção de documentos públicos, destacando os principais benefícios, desafios e lacunas na literatura acadêmica. Com isso, procura-se entender de que forma a IA pode contribuir para a eficiência e qualidade na elaboração de documentos públicos, além de discutir as implicações e limites dessa tecnologia em termos de administração e transparência pública. No estudo do autor (Zuiderwijk 2021), são demonstrados os resultados positivos e em quais atividades a IA consegue auxiliar.

Para reduzir o tempo necessário na elaboração de documentos públicos, propõe-se o uso de LLM's (Large Language Models - LLMs). LLM's são uma classe de modelos de aprendizado de máquina baseados em redes neurais profundas que são projetados para entender, gerar e manipular linguagem natural de forma sofisticada. Esses modelos representam uma das inovações mais significativas no campo do processamento de linguagem natural (NLP), permitindo automatizar e otimizar a criação

de conteúdo textual com precisão e coerência ([Wikipedia](#)) ([Brown Benjamin Mann 2020](#)).

LLM's são usados para:

- Geração de Texto: criar conteúdo escrito, como artigos, histórias, posts de blog, e descrições de produtos. LLMs podem produzir texto coerente e envolvente com base em prompts dados pelo usuário.
- Assistentes Virtuais: assistentes como o ChatGPT da OpenAI são baseados em LLMs e podem realizar tarefas como responder perguntas, fornecer informações, agendar compromissos e até realizar conversas naturais.
- Tradução Automática: LLMs podem traduzir texto de um idioma para outro, mantendo o contexto e a nuance do texto original.
- Resposta a Perguntas: fornecer respostas precisas a perguntas com base em uma vasta base de conhecimento pré-treinada. LLMs podem ser usados em sistemas de perguntas e respostas para suporte ao cliente e outras aplicações.
- Análise de Sentimento: avaliar o tom e a emoção em textos, como comentários em redes sociais, avaliações de produtos, e feedback de clientes.

Todas essas funções estão explicadas no artigo ([Naveed et al. 2024](#)), que explica sobre LLM's, suas diferentes funções e a importância no campo de processamento de linguagem natural.

Referencial teórico

Machine learning

Machine Learning (ML) ([Carbonell, Michalski e Mitchell 1983](#)) é uma subárea da inteligência artificial focada na criação de algoritmos e modelos estatísticos que permitem aos computadores aprenderem a partir de dados e aprimorarem seu desempenho em tarefas específicas ao longo do tempo, sem serem explicitamente programados para cada uma dessas tarefas. No contexto da produção de documentos públicos, o ML tem demonstrado um potencial significativo para otimizar e automatizar diversos processos, tornando-os mais eficientes e precisos ([Goodfellow, Bengio e Courville 2016](#)). O ML possui uma subárea, o deep learning, que são redes neurais e técnicas de aprendizado profundo para analisar grandes volumes de dados e identificar padrões, com o objetivo de tomar decisões com pouca intervenção humana.

Machine learning tem uma aplicação muito importante na área da escrita e revisão de textos, os modelos de ML, especialmente aqueles baseados em técnicas de Processamento de Linguagem Natural (PLN), são amplamente utilizados para automatizar tarefas de escrita e revisão de documentos. Ferramentas como transformadores e modelos de linguagem de grande porte (Large Language Models - LLMs) são capazes de gerar texto que segue padrões específicos de estilo e formalidade, como exigido em documentos governamentais. Além disso, podem sugerir melhorias em termos de clareza, concisão e conformidade com normas legais ([Brown Benjamin Mann 2020](#)).

Deep-learning

Deep Learning é uma subárea do Machine Learning que utiliza redes neurais artificiais profundas para modelar padrões complexos em grandes volumes de dados. Ao contrário dos algoritmos de aprendizado de máquina tradicionais, que muitas vezes dependem de engenharia de características e abordagens estatísticas relativamente simples, os modelos de deep learning são capazes de aprender diretamente a partir dos dados brutos, o que permite a detecção de padrões intrincados e sutis ([Russell e Norvig 2010](#)).

Na produção de documentos públicos, o deep learning tem se mostrado uma ferramenta poderosa para diversas tarefas que envolvem o processamento de linguagem natural (PLN), como extração de informações, tradução automática e análise de grandes quantidades de texto. As redes neurais profundas são inspiradas no funcionamento do cérebro e consistem em neurônios artificiais interligados, denominados nós. Esses, por sua vez, são divididos em três camadas: a camada de entrada, que recebe

os dados brutos; as camadas internas (ocultas) responsáveis por processar e identificar padrões; e a camada de saída que gera o resultado final do modelo (Russell e Norvig 2010). A figura abaixo ilustra isso:

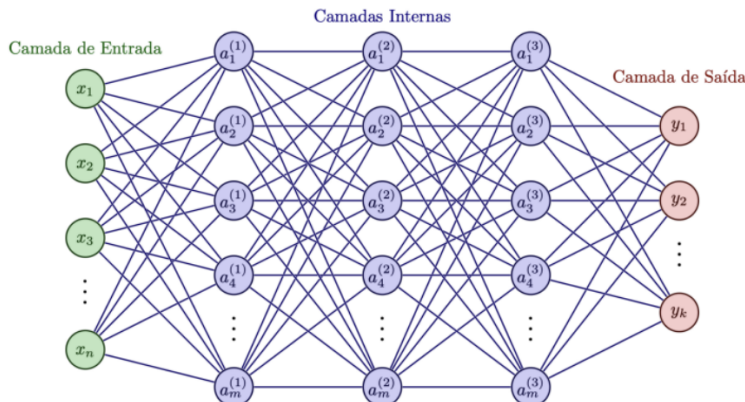


Figura 1: Representação de uma rede neural.

Para que a rede neural funcione, existem várias fórmulas que este usa. No caso mais básico, tem-se o perceptron (Gallant 1990), cuja fórmula é:

$$y = f \left(\sum_{i=1}^n w_i x_i + b \right)$$

onde x_i são as entradas, w_i são os pesos, b é o viés, e f é a função de ativação.

Depois é necessário uma função de ativação para introduzir a não-linearidade na rede neural, permitindo que ela aprenda e represente padrões mais complexos, que seriam impossíveis de captar com funções puramente lineares. Uma das funções de ativação mais utilizadas, especificamente em arquiteturas como transformadores, é a função de ativação softmax. A softmax é responsável por normalizar os valores de saída, convertendo-os em probabilidades que somam 1, o que é especialmente útil para tarefas de classificação em redes neurais (BRIDLE 1989). No contexto do uso de transformadores em redes neurais, a função softmax desempenha um papel crucial na etapa de atenção, atribuindo pesos às entradas com base em sua relevância para o contexto.

Dado um vetor de entrada \mathbf{z} com n elementos, a função softmax é definida como:

$$\text{softmax}(z_i) = \frac{e^{z_i}}{\sum_{j=1}^n e^{z_j}}$$

onde:

- z_i é o i -ésimo componente do vetor de entrada \mathbf{z} ,
- e^{z_i} é a função exponencial aplicada ao componente z_i ,
- $\sum_{j=1}^n e^{z_j}$ é a soma de todas as exponenciais dos componentes do vetor, garantindo que as probabilidades somem 1.

Após isso é realizado o backpropagation (Wythoff 1992) para o algoritmo calcular o gradiente da função de perda em relação aos pesos da rede. Este gradiente é utilizado para ajustar os pesos durante o treinamento, minimizando a função de perda.

Após a fase de cálculo das saídas da rede, é realizado o backpropagation (Wythoff 1992), que é o processo pelo qual o algoritmo calcula o gradiente da função de perda em relação aos pesos da rede. Esse gradiente é então utilizado para ajustar os pesos durante o treinamento, com o objetivo de minimizar a função de perda, melhorando o desempenho da rede.

O cálculo do gradiente é feito utilizando a regra da cadeia, um mecanismo matemático que permite calcular a derivada de funções compostas. Durante o treinamento, cada peso da rede é ajustado na direção oposta ao gradiente da função de perda com relação a esse peso. Isso significa que, ao identificar a inclinação que aumenta a perda, o algoritmo ajusta os pesos na direção contrária, reduzindo assim o erro da rede ao longo das iterações.

Para um peso $w_{jk}^{(l)}$ na camada l , a atualização é:

$$w_{jk}^{(l)} := w_{jk}^{(l)} - \eta \frac{\partial \mathcal{L}}{\partial w_{jk}^{(l)}} \quad (1)$$

E os pesos são atualizados usando um algoritmo de otimização, por exemplo, o Adam (Adaptive Moment Estimation) (Attrapadung Koki Hamada 2021) que ajusta a taxa de aprendizado de cada parâmetro individualmente, permitindo que diferentes parâmetros possam ter diferentes magnitudes de gradiente. O Adam combina as vantagens do método de gradiente estocástico com momentum e a adaptação da taxa de aprendizado, o que o torna eficiente em redes profundas. A fórmula do Adam é:

$$\begin{aligned} m_t &= \beta_1 m_{t-1} + (1 - \beta_1) g_t \\ v_t &= \beta_2 v_{t-1} + (1 - \beta_2) g_t^2 \\ \hat{m}_t &= \frac{m_t}{1 - \beta_1^t} \\ \hat{v}_t &= \frac{v_t}{1 - \beta_2^t} \\ \theta_t &= \theta_{t-1} - \frac{\eta}{\sqrt{\hat{v}_t} + \epsilon} \hat{m}_t \end{aligned}$$

onde:

- m_t é a estimativa do primeiro momento (média dos gradientes).
- v_t é a estimativa do segundo momento (média dos quadrados dos gradientes).
- β_1 e β_2 são os parâmetros de decaimento exponencial para as estimativas dos momentos.
- \hat{m}_t e \hat{v}_t são as estimativas corrigidas dos momentos.
- η é a taxa de aprendizado.
- ϵ é um pequeno valor para evitar divisão por zero.
- θ_t representa os parâmetros do modelo no passo t .

Após a atualização dos pesos e o término do treinamento, é necessário avaliar o desempenho da inteligência artificial na produção dos documentos. Para isso, são utilizadas métricas que avaliam a precisão e a qualidade da escrita gerada. Entre as principais métricas estão o F1-score, BLEU score, perplexidade e ROUGE score. Para este trabalho, será discutido o F1-score e o BLEU score.

O *F1-score* (Sasaki 2007) é uma métrica de avaliação usada principalmente em problemas de classificação para medir o equilíbrio entre a *precisão* (precision) e a *revocação* (recall). É particularmente útil em cenários onde as classes estão desbalanceadas, ou seja, quando uma classe é significativamente mais comum que a outra.

A **precisão** é a proporção de verdadeiros positivos (*True Positives*, *TP*) em relação ao total de previsões positivas, e a **revocação** é a proporção de verdadeiros positivos em relação ao total de reais positivos.

Matematicamente, a precisão (*Precision*) e a revocação (*Recall*) são definidas como:

$$\text{Precision} = \frac{TP}{TP + FP} \quad (2)$$

$$\text{Recall} = \frac{TP}{TP + FN} \quad (3)$$

onde:

- *TP* (True Positives) é o número de previsões positivas corretas,
- *FP* (False Positives) é o número de previsões incorretas onde o modelo previu positivo, mas o resultado real era negativo,
- *FN* (False Negatives) é o número de previsões incorretas onde o modelo previu negativo, mas o resultado real era positivo.

O *F1-score* é a média harmônica da precisão e da revocação, o que significa que o *F1-score* levará em consideração tanto os falsos positivos quanto os falsos negativos. A fórmula para calcular o *F1-score* é dada por:

$$F1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} \quad (4)$$

O *F1-score* varia de 0 a 1, onde 1 representa a melhor pontuação possível (perfeita precisão e revocação), e 0 representa a pior pontuação possível. Esta métrica é particularmente útil quando se deseja alcançar um equilíbrio entre a precisão e a revocação, sem favorecer demasiadamente uma em detrimento da outra.

O *BLEU score* (Bilingual Evaluation Understudy) é uma métrica que avalia a qualidade de textos gerados por sistemas de tradução automática comparando-os com traduções de referência feitas por humanos. Ele mede a similaridade com base na coincidência de n-gramas (sequências de n palavras) entre a tradução gerada e as traduções de referência.

O *BLEU score* calcula a precisão para diferentes tamanhos de n-gramas (como uni-gramas, bi-gramas, etc.) e combina essas precisões usando uma média geométrica, adicionando uma penalização para traduções que são significativamente mais curtas que as de referência. O resultado varia de 0 a 1, onde valores mais altos indicam maior similaridade com as traduções humanas.

Transformadores

Para que as LLM's funcionem de maneira adequada ao analisar textos, estas utilizam os transformadores, que são uma arquitetura inovadora no campo de deep learning, especialmente no processamento de linguagem natural (NLP), que têm revolucionado a forma como modelos de linguagem lidam com tarefas complexas. Esta arquitetura, que foi introduzida no artigo (Vaswani 2023) abandona as limitações das redes neurais recorrentes (RNNs) e convolucionais. Diferentemente dessas abordagens anteriores, os transformadores permitem o processamento paralelo de sequências de dados e são capazes de capturar relações de longo alcance entre elementos de uma sequência.

A arquitetura dos transformadores é composta principalmente por dois blocos: o codificador (encoder) e o decodificador (decoder). Ambos os blocos são empilhados em múltiplas camadas, onde cada camada é composta de subcamadas que incluem mecanismos de atenção e redes neurais feed-forward.

O codificador tem a função de transformar a sequência de entrada em uma representação abstrata que encapsula o contexto de cada token em relação à sequência completa. Esse processo começa com a conversão dos tokens de entrada em vetores de embeddings, seguida pela aplicação de uma codificação posicional para fornecer ao modelo a noção de ordem sequencial dos tokens. Em cada camada do codificador é aplicado um mecanismo de auto-atenção (self-attention), que permite ao modelo ponderar a importância de cada token em relação aos outros tokens da sequência. Em seguida, uma rede feed-forward refina a representação dos dados, proporcionando uma melhor compreensão do contexto global da sequência.

O decodificador, por sua vez, recebe a representação produzida pelo codificador e gera a sequência de saída, que pode ser, por exemplo, uma tradução de texto, uma resposta a uma pergunta ou outro tipo de texto. Similar ao codificador, o decodificador utiliza camadas de auto-atenção, mas também

inclui um mecanismo de atenção cruzada (cross-attention), que permite ao modelo focar em diferentes partes da representação de entrada ao gerar cada token de saída, garantindo que a saída seja coerente com a entrada.

O mecanismo de atenção é o núcleo dos transformadores, permitindo que o modelo se concentre em diferentes partes da entrada de forma simultânea. A atenção escalonada é calculada comparando vetores de consulta (queries), chaves (keys) e valores (values), e gerando pesos que determinam quais partes da entrada são mais relevantes para a tarefa em questão.

A inovação do transformador está na atenção multi-head, onde múltiplas "cabeças" de atenção são aplicadas em paralelo. Cada "cabeça" de atenção captura diferentes aspectos da entrada, permitindo ao modelo aprender uma representação mais rica e robusta dos dados. Isso aumenta significativamente a capacidade do modelo de lidar com tarefas complexas e longas sequências de texto.

O mecanismo de atenção é calculado usando a fórmula:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

Onde:

- Q (Queries), K (Keys) e V (Values) são matrizes derivadas das entradas.
- d_k é a dimensão dos vetores de keys.

Essa fórmula descreve como o modelo calcula a "atenção" de cada palavra em relação às outras, determinando como as informações fluem através da rede.

Os transformadores usam uma arquitetura composta de camadas de atenção multi-head e feed-forward. A atenção multi-head permite que o modelo aprenda diferentes representações ao olhar para diferentes partes do texto. A fórmula para a atenção multi-head é:

$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, \dots, \text{head}_h)W^O$$

Onde:

- Cada $\text{head}_i = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V)$, e W^O é a matriz de projeção final.
- h é o número de cabeças de atenção.

Além disso, as camadas de feed-forward aplicam transformações não lineares adicionais. Cada camada é definida como:

$$\text{FFN}(x) = \max(0, xW_1 + b_1)W_2 + b_2$$

No qual W_1 , W_2 , b_1 , e b_2 são parâmetros aprendidos.

No contexto dos LLM's, os transformadores são treinados em grandes volumes de texto, usando uma abordagem de pré-treinamento seguida de ajuste fino. O pré-treinamento é, geralmente, um processo não supervisionado, onde o modelo aprende representações ricas de linguagem, capturando padrões gerais e dependências de longo alcance nas sequências de texto. Esse processo permite que o modelo desenvolva uma compreensão ampla e geral da linguagem.

Após o pré-treinamento, o modelo passa por uma etapa de ajuste fino supervisionado, onde este é adaptado para tarefas específicas, como tradução automática, geração de textos ou resumo de documentos. Durante essa fase, o modelo é refinado utilizando conjuntos de dados rotulados para melhorar seu desempenho em tarefas concretas como resumo de textos.

Esses LLMs, como GPT, são capazes de realizar tarefas complexas de NLP, como tradução, resumo, e geração de texto, graças à sua capacidade de capturar dependências de longo alcance e complexas nas sequências de texto. Permitindo que compreendam o contexto e gerem respostas coerentes e relevantes em uma ampla variedade de tarefas.

Large Language Models(LLM)

LLM's são redes neurais profundas muito grandes que são pré-treinadas em grandes quantidades de dados, mais precisamente elas se encaixam na categoria de modelos generativos que usam aprendizado supervisionado e não supervisionado para processar e gerar linguagem (Carbonell, Michalski e Mitchell 1983).

Os modelos de linguagem de grande escala representam um avanço significativo no campo de processamento de linguagem natural (PLN), sendo construídos com base na arquitetura Transformer explicado anteriormente e possuírem milhões, ou até trilhões de parâmetros, o que lhes permite gerar texto com coerência e relevância.

Com o surgimento dos Transformers (Vaswani 2023), houve uma mudança significativa no desenvolvimento de modelos de processamento de linguagem natural (PLN). O mecanismo de auto-atenção (self-attention) e as camadas de feedforward dos Transformers permitiram grandes avanços. O mecanismo de auto-atenção possibilita que o modelo se concentre em partes relevantes da sequência de entrada, capturando dependências de longo alcance entre palavras, o que é essencial para uma compreensão contextual mais profunda.

Uma das características que define os LLMs é sua escala. O aumento no número de parâmetros e o uso de grandes volumes de dados de treinamento permitem que esses modelos capturem uma gama muito mais ampla de nuances linguísticas e padrões contextuais do que os modelos menores. Isso resultou em algumas capacidades como:

- **Aprendizado em contexto:** A capacidade de resolver novas tarefas simplesmente fornecendo exemplos no prompt, sem a necessidade de ajuste fino adicional.
- **Raciocínio e planejamento:** Modelos de grande escala demonstraram habilidades em resolver problemas de raciocínio lógico ou tomar decisões com base em múltiplos passos de pensamento.

Os LLM's podem ser categorizados em dois tipos principais:

- **Modelos autoregressivos:** Como o GPT, que geram texto de maneira sequencial, prevendo a próxima palavra com base nas palavras anteriores. Esse tipo de modelo é eficaz para tarefas de geração de texto, como criação de diálogos, redação de conteúdo e autocompletar (Dalal, Li e Taori 2019).
- **Modelos bidirecionais:** Como o BERT, que processa simultaneamente palavras anteriores e subsequentes para entender o contexto completo de uma palavra em uma frase. Isso os torna altamente eficientes para tarefas como perguntas e respostas, classificação de sentimentos e reconhecimento de entidades nomeadas (Rogers, Kovaleva e Rumshisky 2021).

Em resumo, os modelos de linguagem de grande escala representam o ápice atual do aprendizado profundo aplicado à linguagem, combinando escalabilidade com compreensão contextual.

Abordagem metodológica

Esta seção descreve a abordagem utilizada para realizar uma revisão extensa da literatura. Foi adotada a abordagem de revisão sistemática da literatura, conforme definida por (Kitchenham 2004), com o objetivo de identificar, avaliar e sintetizar as evidências disponíveis sobre o uso de Large Language Models (LLM's) na administração pública.

Nas subseções a seguir, detalhamos as seguintes etapas da revisão sistemática: (1) identificação dos estudos, (2) seleção dos estudos, (3) avaliação da relevância e qualidade dos estudos, (4) extração de dados e (5) síntese dos dados.

Passo 1: Identificação dos estudos

Na primeira etapa, serão definidos os objetivos e as perguntas que moldaram a revisão de literatura. Como reconhecido por (Kitchenham 2004), revisões de literatura podem ser realizadas para diversos propósitos. Para este trabalho, os objetivos da revisão são: (1) posicionar as pesquisas identificadas

em relação ao conhecimento existente; (2) obter insights úteis sobre os métodos de pesquisa que outros pesquisadores utilizaram para estudar as implicações do uso de IA no governo para a administração pública.

Para atingir o primeiro objetivo da revisão de literatura, ou seja, posicionar as pesquisas identificadas em relação ao conhecimento existente, serão feitas as seguintes perguntas:

- a) Em quais contextos o tema das implicações da governança pública em relação ao uso de IA no governo foi investigado por pesquisas anteriores (por exemplo, disciplinas de pesquisa, regiões, países)?
- b) Quais são os objetivos e as contribuições das pesquisas anteriores sobre as implicações do uso de IA no governo para a administração pública?
- c) Quais teorias e modelos teóricos foram desenvolvidos, usados, testados ou aplicados em estudos sobre as implicações do uso de IA no governo para a administração pública?

O segundo objetivo da revisão, que é obter insights úteis sobre os métodos de pesquisa que outros estudiosos usaram para estudar as implicações do uso de IA no governo para a administração pública, levou às seguintes perguntas:

- d) Quais abordagens e métodos de pesquisa foram empregados em estudos que abordam o uso de IA no governo para a administração pública?
- e) Quais são os principais elementos da administração pública afetados pelo uso de IA?
- f) Quais são os benefícios e desafios potenciais do uso de IA para a administração pública?

Será usado seis fontes complementares para identificar estudos científicos sobre as implicações do uso de IA na administração pública: Web of Science, Scopus, IEEE(Institute of Electrical and Electronic Engineers), ArXiv, Google scholar e Papers with Code. Essas bases de dados, em conjunto, cobrem mais de 200 editores que publicam materiais intimamente relacionados ao tema em estudo.

Seleção de estudos

No segundo passo, a seleção de estudos, serão definidos os critérios de busca, exclusão e inclusão. Usando os termos de busca especificados na tabela abaixo, serão limitados a artigos de periódicos e conferências publicados entre 1996 e 2024. Para as buscas em bases de dados acadêmicos como o Scopus, os resultados serão restringidos a disciplinas de pesquisa específicas, com o objetivo de identificar os estudos mais relevantes para a temática.

Banco de dados	Web of Science, Scopus, ArXiv, IEEE, Google Scholar e Papers with code
Termos de busca/keywords	(Large Language Model* OR LQA OR RAG OR fine-tuning) AND ("legal document*" OR "judicial Record*" OR ((public OR government) AND procurement* OR (write legal document*))) (public procurement OR public bidding OR bidding) AND (NLP OR LLM OR fine-tuning) (NLP OR transformers OR LLM) AND (bidding))

Passo 3: Relevância dos estudos e qualidade da avaliação

O terceiro passo da revisão sistemática da literatura é garantir a relevância e qualidade dos estudos selecionados. Essa fase consiste em dois passos principais. Primeiro, entre os mais de 50 estudos

selecionados, o título e o resumo de cada artigo serão analisados para determinar sua relevância, com base em três critérios:

1) A IA deve desempenhar um papel substancial ou principal no estudo(o objetivo, a pergunta da pesquisa, etc). Estudos em que o uso de IA era marginal ou secundário foram excluídos nesta fase.

2) Uso de IA em administração governamental deve ser central no estudo. Se o estudo ou artigo não abordou o uso de IA (pelo menos parcialmente) no contexto da administração pública, ele foi excluído nessa fase.

3) As implicações do uso de IA no setor público devem ser o foco principal do artigo. Em outras palavras, se IA for apenas mencionado superficialmente ou listado sem ser o foco principal do assunto, o estudo será excluído. Por exemplo, se a IA for discutida apenas como uma ferramenta, sem ser fortemente ligada a uma ou mais implicações na administração pública, o artigo será excluído.

4) Como avaliação de qualidade, foi utilizado o sistema QUALIS (Capes), sendo selecionadas apenas as publicações com estrato maior ou igual a B2. As publicações que não estavam no QUALIS, foi feita uma comparação com o índice H5 do Google Scholar e com o percentual do Journal Impact Factor(JIF) que se compara à qualidade de um periódico de nota B2 ou superior.

Com base na relevância de cada estudo, 28 artigos foram selecionados para servir como base para esta revisão sistemática. Cada um desses estudos foi avaliado de acordo com as seguintes dimensões de qualidade derivadas do (Batini et al. 2009) e (Bano e Zowghi 2015):

- **Precisão:** Os objetivos do estudo estão claramente definidos, os dados e métodos de coleta estão adequadamente descritos. Referências de apoio são essenciais para as afirmações feitas no artigo.
- **Consistência:** O design do estudo está adequado para os objetivos da pesquisa. As perguntas da pesquisa são respondidas ou o objetivo da pesquisa é atingido.
- **Integridade:** A abordagem de pesquisa do artigo é descrita com detalhes suficientes.

De acordo com os critérios de avaliação dos artigos, quinze artigos foram excluídos da lista pelas seguintes razões: 1) Os artigos não atenderam aos critérios de qualidade mencionados acima; 2) Os estudos continham apenas a opinião do autor sem descrever qualquer abordagem ou design de pesquisa específico; 3) Os estudos tinham um foco insuficiente nas implicações do uso da IA para a administração pública; 4) Os artigos foram publicados como descrições breves de pôsteres incluídas em conferências.

Isso resultou em uma seleção final de 28 estudos que abordam diretamente questões de administração pública em relação à IA. De acordo com os critérios de inclusão e exclusão, o número limitado de estudos restantes, apesar da extensa busca por literatura relevante, é um achado importante por si só. Esse pequeno número destaca a escassez de pesquisas que examinam o impacto da IA na administração pública.

Passo 4: Extração de dados

Para a extração de dados nesta revisão, será utilizada uma planilha para registrar os metadados de cada um dos estudos selecionados. A Tabela 1 apresenta os metadados coletados dos 28 estudos incluídos, abrangendo informações descritivas, detalhes sobre a abordagem metodológica, qualidade do estudo, aspectos da administração pública e dados relacionados ao uso da IA. Para garantir coerência na coleta de informações, essas categorias de metadados foram derivadas das perguntas de pesquisa mencionadas na seção de relevância dos estudos e qualidade da avaliação.

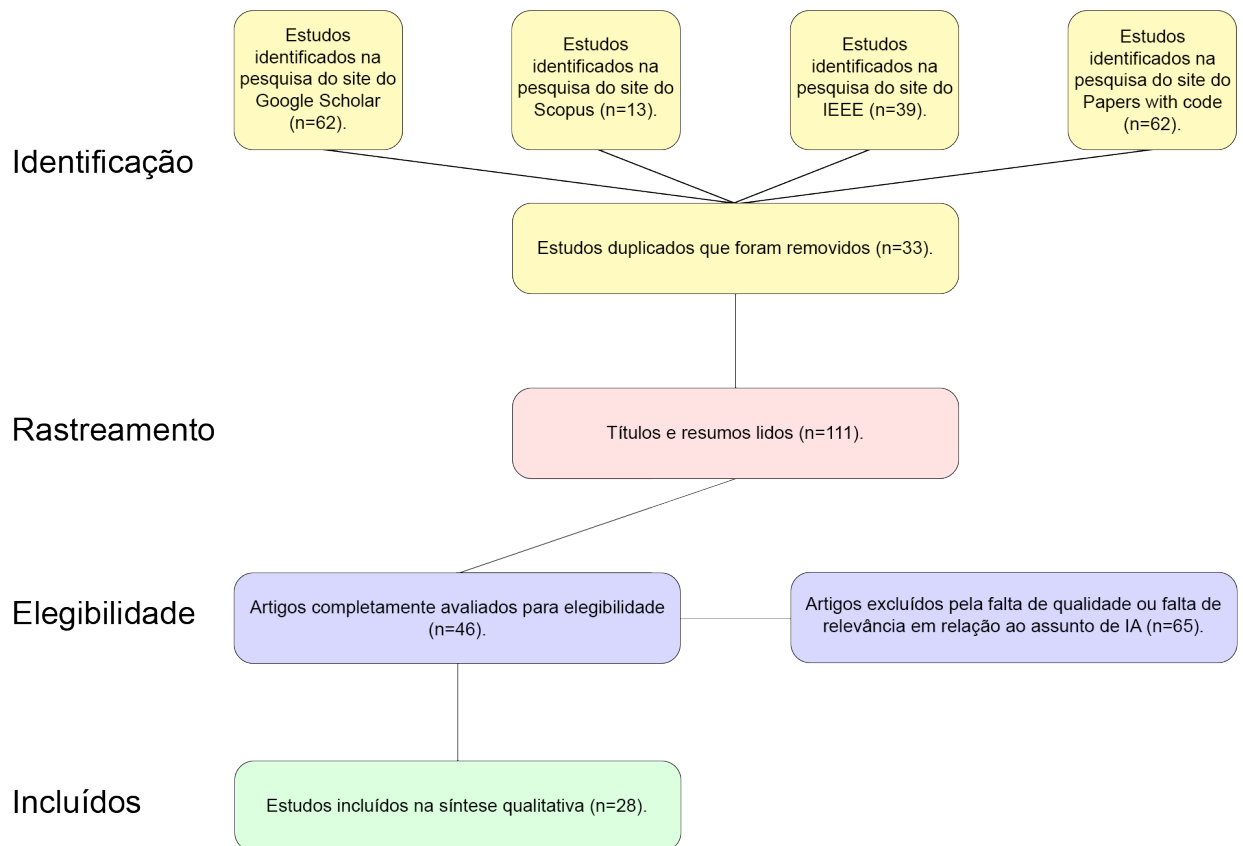


Figura 2: Extração de dados da revisão sistemática.

Síntese de dados

A etapa final do estudo ficou focada na síntese dos dados. Onde foram analisados sistematicamente os dados brutos derivados do procedimento da revisão da literatura mencionada na figura acima. Para os mais de 80 artigos que foram analisados, a grande maioria foi removida pela sua falta de relevância em relação ao assunto ou falta de qualidade. Nas tabelas abaixo é citado o nome do autor do artigo e sobre o que é o artigo que este escreveu, e em alguns, a precisão da IA.

Tabela 1: Visão geral dos estudos incluídos na revisão sistemática da literatura.

Referências	Objetivo do estudo
1 (Vaswani 2023)	Os modelos dominantes de transdução de sequência são baseados em redes neurais recorrentes ou convolucionais complexas que incluem um codificador e um decodificador. Os modelos de melhor desempenho também conectam o codificador e o decodificador por meio de um mecanismo de atenção. Nós propomos uma nova arquitetura de rede simples, o Transformer, baseada exclusivamente em mecanismos de atenção, eliminando completamente a recorrência e as convoluções.
2 (Goodfellow, Bengio e Courville 2016)	Uma introdução a uma ampla gama de tópicos em deep learning, cobrindo fundamentos matemáticos e conceituais, técnicas de deep learning utilizadas na indústria e perspectivas de pesquisa.
3 (Brown Benjamin Mann 2020)	O artigo apresenta o GPT-3, um modelo de linguagem com 175 bilhões de parâmetros, que realiza várias tarefas de linguagem sem ajuste fino. Ele se destaca por aprender com poucos exemplos (few-shot learning), alcançando resultados competitivos em tradução, perguntas e geração de texto. No entanto, ainda enfrenta limitações, como dificuldades em tarefas de raciocínio, viés nos dados e alto custo computacional.
4 (Russell e Norvig 2010)	O livro Inteligência Artificial: Uma Abordagem Moderna apresenta a IA como o estudo de agentes inteligentes que percebem e agem no ambiente para alcançar seus objetivos. Ele aborda diversos subcampos da IA e destaca os avanços e desafios na tomada de decisões, aprendizado de máquina e robótica em cenários complexos.
5 (Gallant 1990)	O artigo discute os algoritmos de aprendizado baseados no perceptron, analisando suas limitações e propondo melhorias para aumentar o desempenho em tarefas de classificação e reconhecimento de padrões.
6 (BRIDLE 1989)	O artigo propõe modificações em redes neurais MLPs para tratar saídas como probabilidades. Ele utiliza a função softmax e pontuação de probabilidade, simplificando os cálculos. O uso de unidades radiais melhora a discriminação de classes, superando métodos tradicionais, mesmo com suposições imprecisas sobre distribuições.

Tabela 2: Visão geral dos estudos incluídos na revisão sistemática da literatura.

Referências	Objetivo do estudo
7 (Wythoff 1992)	O artigo menciona as aplicações amplas das redes neurais, incluindo pesquisas químicas. O foco principal é fornecer uma introdução teórica detalhada sobre redes neurais de retropropagação e discutir questões práticas para desenvolvedores.
8 (Attrapadung Koki Hamada 2021)	O artigo ajuda em LLMs ao melhorar as funções de ativação, como as sigmóides, tornando a rede mais eficiente e precisa na modelagem de padrões complexos e tarefas de linguagem.
9 (Sasaki 2007)	O artigo explica o F-measure, uma métrica que combina precisão e recall como uma média harmônica, amplamente usada na avaliação de sistemas de extração de informação.
10 (Papineni Salim Roukos 2002)	O artigo propõe o BLEU, um método automático de avaliação de traduções que compara n-gramas da tradução gerada com traduções humanas, oferecendo uma alternativa rápida e barata à avaliação humana, com boa correlação aos julgamentos humanos.
11 (Zuiderwijk 2021)	O artigo realiza uma revisão sistemática sobre o uso da IA na governança pública. Ele destaca as oportunidades e desafios da IA, propondo uma agenda de pesquisa com recomendações para estudos futuros.
12 (Bano e Zowghi 2015)	O artigo revisa 87 estudos e conclui que o envolvimento de usuários geralmente contribui para o sucesso dos sistemas, mas pode causar problemas se mal gerido, destacando fatores e estratégias para uma gestão eficaz durante o desenvolvimento de software.
13 (Shao et al. 2024)	O artigo propõe o sistema STORM, que usa modelos de linguagem para gerar artigos longos, como os da Wikipedia, a partir do zero, focando na pesquisa e criação de esboços antes da escrita. Ele também avalia a qualidade dos artigos gerados e destaca desafios de manter a veracidade das informações.
14 (Li 2022)	O artigo propõe um modelo PSO-SVM para prever preços de licitação em projetos de construção, melhorando a precisão em problemas não lineares. Testado em projetos reais, o modelo teve uma taxa de erro de 1,96%, mostrando-se eficaz e superior a métodos tradicionais.

Tabela 3: Visão geral dos estudos incluídos na revisão sistemática da literatura.

Referências	Objetivo do estudo
15 (Tang e Yu 2024)	O artigo propõe o modelo CPMARL-AFL, que agrupa usuários de modelos (MUs) em clusters para otimizar lances em leilões de aprendizado federado. Isso melhora a eficiência e precisão dos lances. O modelo superou outras abordagens em termos de utilidade e precisão, com ganhos de até 1,40%.
16 (Wei et al. 2023)	O artigo mostra que o ajuste fino do modelo DistilBERT melhora a precisão e o F1-score na classificação de textos jurídicos, especialmente em trechos curtos. O ajuste fino é eficaz para revisão de documentos legais, mas exige mais recursos computacionais.
17 (Kamoi et al. 2024)	O artigo avalia a capacidade dos LLMs de detectar erros em suas respostas usando o benchmark ReaLMistake. Ele mostra que modelos como GPT-4 e Llama 2 têm baixo desempenho na detecção de erros, superados pelos humanos, e que técnicas comuns não melhoram essa habilidade.
18 (Liu et al. 2023)	O artigo avalia o uso de grandes modelos de linguagem em tarefas de recomendação. LLMs têm bom desempenho em explicabilidade, mas são menos eficazes em tarefas de precisão. O ajuste supervisionado melhora seu desempenho em recomendações.
19 (Bai et al. 2024)	O artigo aborda como os modelos de linguagem (LLMs) têm dificuldade em gerar textos com mais de 2.000 palavras. A solução proposta é o AgentWrite, que divide grandes tarefas de escrita em partes menores, permitindo gerar textos de até 20.000 palavras. Usando o conjunto de dados LongWriter-6k, os autores aumentaram a capacidade de saída dos modelos sem comprometer a qualidade. Eles também criaram o LongBenchWrite para avaliar essa habilidade, e o modelo treinado superou até modelos maiores.
20 (Dettmers et al. 2023)	O QLORA é uma técnica que permite ajustar grandes modelos de linguagem de forma eficiente, usando menos memória (apenas uma GPU de 48 GB) sem perder desempenho. Utiliza quantização de 4 bits e adaptadores de baixa rank (LoRA). O modelo Guanaco 65B, ajustado com QLORA, alcançou 99,3% da performance do ChatGPT em 24 horas. A técnica torna o ajuste fino acessível mesmo com recursos limitados.
21 (Fleischer et al. 2024)	O artigo apresenta um framework open-source para facilitar a implementação de Retrieval-Augmented Generation (RAG) em Large Language Models (LLMs). O RAG Foundry é modular e permite a criação, treinamento, inferência e avaliação de modelos de linguagem que utilizam informações externas para melhorar a precisão e relevância das respostas.

Tabela 4: Visão geral dos estudos incluídos na revisão sistemática da literatura.

Referências	Objetivo do estudo
22 (Lewis et al. 2021)	O artigo apresenta o Retrieval-Augmented Generation (RAG), uma abordagem que combina memória paramétrica (modelos seq2seq) e não-paramétrica (recuperação de documentos) para melhorar tarefas intensivas em conhecimento, como resposta a perguntas e verificação de fatos.
23 (Yang, Ji e Jing 2021)	O artigo propõe um método de aprendizado seletivo para otimizar lances de produtores de energia em mercados de eletricidade de preço uniforme. Usando vários algoritmos de aprendizado de máquina, o sistema faz previsões de preços e escolhe os melhores lances. As simulações mostram que esse método aumenta os lucros dos produtores ao melhorar a precisão das previsões e se adaptar ao mercado.
24 (Harikrishnakumar et al. 2019)	O artigo propõe o uso de algoritmos de aprendizado supervisionado para classificar fornecedores em quatro categorias (excelente, bom, satisfatório e insatisfatório), substituindo métodos tradicionais de tomada de decisão que são enviesados e subjetivos. Utilizando um conjunto de dados de 350 fornecedores da indústria petroquímica, o estudo avalia diversos algoritmos, como SVM, regressão logística e KNN. O SVM com kernel linear foi o mais preciso, com 87,12% de acurácia.
25 (Valim 2024)	O artigo investiga métodos de machine learning para identificar produtos idênticos em fontes distintas. Abordagens tradicionais baseadas em regras têm limitações, e técnicas modernas, como modelos de linguagem pré-treinados (BERT e variantes), superam esses desafios, atingindo escores F1 superiores a 95%.
26 (Rhythm et al. 2023)	O artigo propõe um sistema que gera automaticamente perguntas e respostas a partir de textos usando modelos de linguagem como BengaliBERT. Avaliado com métricas de precisão (EM Score) e F1 Score, o sistema mostrou bom desempenho em perguntas factuais, com melhorias necessárias em questões mais complexas.
27 (Howard e Ruder 2018)	O artigo propõe o ULMFiT, um método de aprendizado por transferência para PLN que melhora a classificação de textos, reduzindo erros em até 24% em tarefas amplamente usadas. Ele utiliza técnicas de ajuste fino para evitar o esquecimento catastrófico e é altamente eficiente, alcançando bom desempenho com poucos dados rotulados. O modelo e o código foram disponibilizados para facilitar a adoção.
28 (Eken 2022)	A tese investiga o uso de PLN e aprendizado de máquina para automatizar a revisão de contratos de construção, alcançando 89% de precisão na categorização de tipos de sentenças e 83% na identificação das partes envolvidas. O estudo demonstra como essas técnicas podem agilizar a análise contratual e reduzir erros, especialmente na fase de licitação.

Resultados

Esta seção irá examinar o potencial da IA para a administração pública de um modo mais prático, discutindo métodos como RAG(Fleischer et al. 2024), prompts para o ChatGPT(Naveed et al. 2024) e métodos para se usar LLM's de maneira gratuita (Naveed et al. 2024) como o Llama, e como usar a biblioteca LangChain(Jiang et al. 2024) para usar LLM's, vetorização de textos e banco de dados de forma gratuita com o intuito de cortar custos.

Em que a IA não só consegue resolver problemas de NLP(do inglês, processamento de linguagem natural), mas também consegue fazer análise de dados, analisando grandes quantidades de informações, com isso a IA consegue resumir documentos, calcular o preço total de uma licitação, escrever documentos, etc.

Usando API do *ChatGPT*

É possível usar as APIs do ChatGPT para automatizar a leitura e escrita do documento de licitação, onde o funcionário público irá escrever prompts ao chatbot(Brown Benjamin Mann 2020), tais como:

Você é um servidor público do(a) [nome do órgão/entidade] e necessita escrever o trecho de um estudo técnico preliminar de acordo com a Lei 14.133/2021.

Redija um texto com a descrição da necessidade da contratação, considerado o problema a ser resolvido sob a perspectiva do interesse público, justificando a necessidade do(a) [objeto da contratação] para [unidade demandante].

Deste jeito o chatbot vai começar a escrever na formalidade e os textos necessários para uma licitação. O *ChatGPT* já é treinado para escrever na formalidade e estrutura de licitação, basta usar o prompt correto.

E para o seguinte prompt:

Redija novamente o texto considerando que o nome da escola é 'escola Tio Patinhas' e a cidade já informada.

E o chatbot deu a seguinte resposta:

Estudo Técnico Preliminar (ETP) para Aquisição de Cadeiras Universitárias para a Escola Tio Patinhas

****1. Descrição da Necessidade e Justificativa da Contratação****

A Prefeitura de Embu das Artes, compreendendo a essencialidade da educação como vetor de desenvolvimento social e econômico, e alinhada ao compromisso de propiciar um ensino público de qualidade, identificou uma demanda imperativa associada à recém-inaugurada Escola Tio Patinhas, localizada no bairro [Nome do Bairro]. Este educandário foi estrategicamente edificado para atender a uma crescente demanda estudantil da região, garantindo um acesso mais equânime e qualitativo à educação.

****Problema Identificado:****

A Escola Tio Patinhas, construída para acolher [Número de Estudantes] estudantes, distribuídos em [Número de Turmas] turmas, nos turnos matutino e vespertino, confronta-se, no entanto, com um desafio premente: a ausência de um mobiliário escolar adequado para assegurar uma experiência educacional propícia e ergonômica aos alunos, uma vez que atualmente não dispõe de assentos apropriados para os discentes nas salas de aula.

****Justificativa do Interesse Público:****

É indiscutível que a falta de cadeiras universitárias apropriadas pode incidir negativamente na experiência educacional dos alunos e, conseqüentemente, em seu desempenho acadêmico, além de implicar riscos ergonômicos. Portanto, assegurar um ambiente escolar estruturado, que viabilize condições físicas adequadas para o aprendizado, não é apenas uma questão de conforto, mas um requisito para promover educação de qualidade.

****Importância da Contratação de 100 Cadeiras Universitárias:****

Diante da lacuna identificada e cientes da imperativa necessidade de garantir que cada estudante da Escola Tio Patinhas disponha de um assento adequado e confortável durante suas atividades acadêmicas, propomos a aquisição de 100 cadeiras universitárias. Esta quantidade visa não apenas atender todas as turmas da escola mas também assegurar que as instalações estejam equipadas de forma a proporcionar ambientes de aprendizagem plenamente funcionais.

A contratação se justifica por:

- ****Garantir Acessibilidade e Conforto:****

Oferecer assentos adequados e ergonômicos para todos os estudantes, sem distinção. - ****Fomentar o Aprendizado:****

Facilitar a criação de um ambiente propício para o desenvolvimento integral dos estudantes, em aspectos acadêmicos, sociais e pessoais.

- ****Promover Saúde e Bem-Estar:****

Ao adotar cadeiras que atendem critérios ergonômicos, estamos prevenindo problemas de saúde e promovendo um bem-estar físico aos estudantes. - ****Maximizar a Utilização da Escola Tio Patinhas:****

Garantir que todos os espaços sejam utilizados de forma plena e eficaz, permitindo que todas as salas de aula estejam operacionais e aptas a acolher os alunos. Assim, a aquisição das cadeiras universitárias para a Escola Tio Patinhas não apenas colmata uma necessidade imediata mas também se alinha de forma intrínseca ao comprometimento desta administração pública com a educação de alta qualidade, garantindo que os espaços educativos sejam verdadeiramente ambientes de crescimento, aprendizagem e bem-estar para todos os alunos de Embu das Artes.

— Este exemplo, ainda que adaptado para o contexto local, permanece um modelo que deve ser refinado e ajustado conforme as especificidades do projeto, garantindo que todas as análises e etapas preconizadas pela Lei nº 14.133/2021 sejam rigorosamente cumpridas.

Porém, este método pode sair caro, já que para usar a API do *ChatGPT* é necessário pagar a inscrição mensal, e para cada mil caracteres que este escreve, é cobrado um preço em dólar. Para um bloco do governo, isso seria barato; porém, se a *IA* pode ser usada para um tipo de documento governamental, ele pode ser usado para a maioria, senão todos. Se todos os blocos do governo tentarem automatizar todas as escritas de documentos, isso demandaria muito dinheiro dos cofres públicos. O preço do uso das APIs está ilustrado abaixo:

Model	Input	Output
gpt-4	\$30.00 / 1M tokens	\$60.00 / 1M tokens
gpt-4-32k	\$60.00 / 1M tokens	\$120.00 / 1M tokens

Figura 3: Preço das APIs do GPT-4

Onde cada 1000 tokens equivale a 750 caracteres.

Pensando em reduzir o preço de uma IA, são usados os modelos LLM(modelo de linguagem ampla) open-source(código aberto), que não têm custo para serem usados, porém deve ser realizado todo o treinamento da IA usando fine-tuning, já que a IA deve aprender como escrever na formalidade uma licitação. Na próxima sessão discutiremos como será realizado esse treinamento.

Existem vários modelos de LLM open-source disponibilizados na Internet, e para conseguir usar esses modelos será usado o Ollama, que é um projeto de código aberto que serve como uma plataforma poderosa e amigável para executar modelos de linguagem grande (LLMs) na sua máquina local. Ele atua como uma ponte entre as complexidades da tecnologia LLM e o desejo por uma experiência de IA acessível e personalizável.

No site deles é possível fazer o download de várias LLM's como o Llama2, Llama3, Mistral, GEMMA 2, no site do [Ollama](#) e cada LLM pode ter diferentes respostas e parâmetros, por exemplo, o Llama2 possui 7 bilhões de parâmetros. Em LLM's parâmetros são componentes ajustáveis que permitem ao modelo aprender a partir dos dados. Eles incluem pesos e vieses, que são ajustados durante o treinamento para minimizar o erro e melhorar o desempenho do modelo em tarefas específicas. O

número de parâmetros é uma medida chave da capacidade do modelo, influenciando sua capacidade de aprender e generalizar a partir de grandes volumes de dados.

Para usar todas as funções necessárias para o bom funcionamento da LLM, como coleta de dados, transformação dos documentos em algo mais leve e legível para a LLM, embarcar os dados no banco de dados e trocar os diferentes modelos de LLM, será usado o LangChain.

Langchain

LangChain é uma estrutura que facilita a criação de aplicativos usando grandes modelos de linguagem (LLMs)(Pandya e Holia 2023). Ele ajuda a conectar esses modelos a fontes de dados externas, permite que eles mantenham o contexto em conversas longas, e possibilita a criação de fluxos de trabalho complexos com várias etapas de processamento. Em resumo, LangChain torna mais fácil e eficiente usar LLMs em diversas aplicações, como assistentes virtuais e sistemas de recomendação.

Ele possui as seguintes funcionalidades:

1. Cadeia de Chamadas: permite a construção de cadeias de chamadas de API ou funções que podem ser orquestradas para realizar tarefas complexas.

2. Memória: suporte a diferentes tipos de memória para os LLMs, permitindo que eles mantenham o contexto entre chamadas, o que é essencial para diálogos contínuos e aplicações que precisam de memória de longo prazo.

3. Integração com dados: possui suporte para integrar LLMs com diversas fontes de dados, como bancos de dados, APIs web e sistemas de arquivos, permitindo que os modelos de linguagem acessem e usem informações dinâmicas e atualizadas.

4. Execução de código: suporte para executar código, permitindo a criação de aplicações onde os LLMs podem gerar e executar código para responder perguntas, transformar dados ou realizar outras tarefas programáticas.

5. Agentes: LangChain permite a criação de agentes que podem tomar decisões baseadas em entradas, chamadas de API e respostas de modelos de linguagem, permitindo um comportamento mais dinâmico e adaptável.

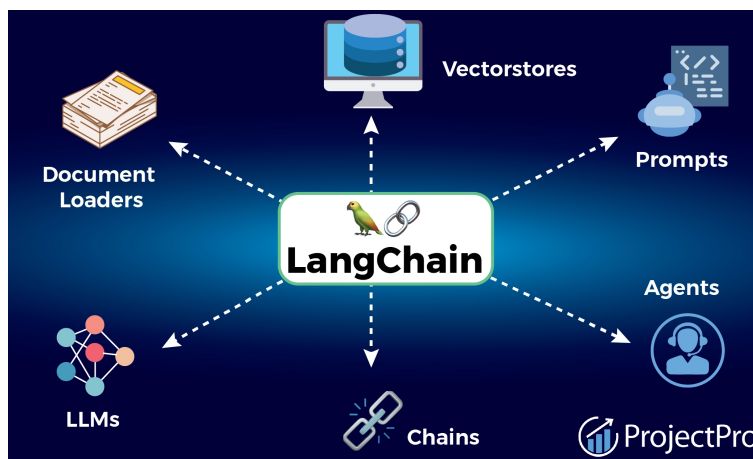


Figura 4: Figura ilustrativa do Langchain e suas utilidades

Usando RAG em um modelo open source de IA

Em Large Language Models (LLMs), como o GPT-4, é possível a realização de uma RAG(Fleischer et al. 2024). Este conceito combina a geração de texto com a recuperação de informações, utilizando a capacidade de grandes modelos de linguagem para melhorar a precisão e relevância das respostas.

Uma RAG é dividida em duas partes, a retrieval(recuperação), onde um componente de recuperação busca informações relevantes em uma base de dados ou conjunto de documentos para responder a uma

consulta específica. A segunda é generation(geração) um modelo de linguagem como o Llama2 gera uma resposta utilizando as informações recuperadas, integrando-as de forma coerente e contextualmente apropriada.

Este método é útil para fazer pesquisa de preços de acordo com outros documentos, resolver problema de decisão sobre o que deve ser feito de acordo com o orçamento e contexto do problema, por exemplo em uma licitação de veículos, qual veículo deve ser comprado, se deve ser feito leasing, compra ou aluguel dos veículos.

Uma RAG possui várias etapas, como é possível ver na imagem abaixo:

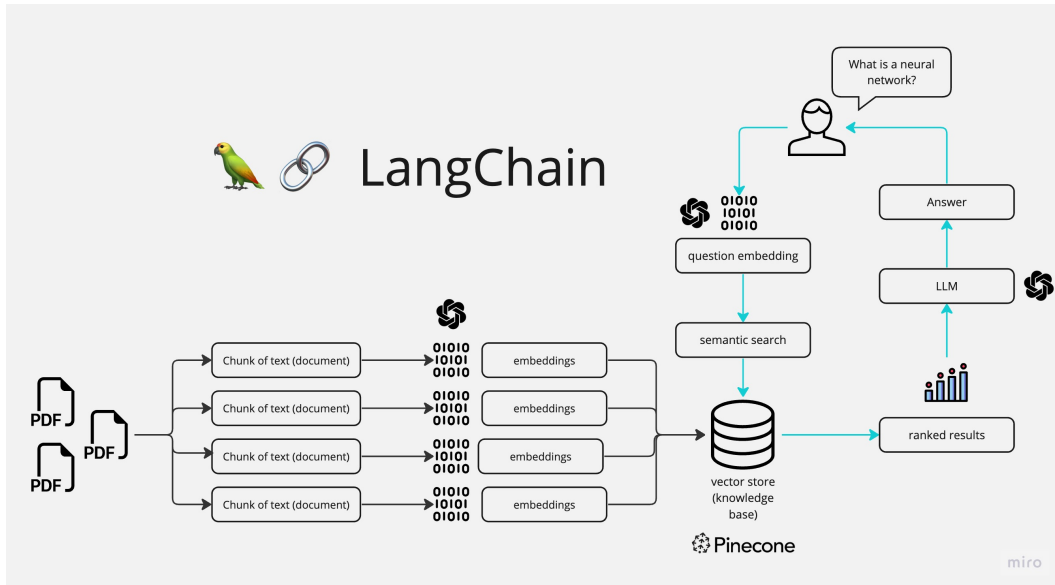


Figura 5: Estrutura de uma RAG usando Langchain.

No qual os documentos são separados em pedaços de texto para carregar melhor os dados para a IA, depois esses pedaços são transformados em modelo de texto embarcado que é uma técnica de processamento de linguagem natural (NLP) que transforma palavras, frases ou documentos inteiros em vetores numéricos de dimensão fixa. Esses vetores, conhecidos como embeddings, capturam o significado semântico dos textos de uma forma que facilita o processamento por algoritmos de aprendizado de máquina. Depois esse modelo é armazenado em um banco de dados de código aberto, foi escolhido o Chroma DB por sua facilidade de uso e por ser um banco de dados open source. Para transformar o texto em modelo de texto embarcado será usado o modelo do hugging face ([Hugging Face](#)), que é um modelo aberto para grandes quantidades de texto, pode ser menos rápido que o modelo da OpenAI ([Open AI](#)), porém para este projeto o importante é cortar custos.

Após o texto ser transformado em um modelo embarcado, este é armazenado no banco de dados para futuro estudo da LLM. Após isso, o usuário fará uma série de perguntas para a LLM e as próprias perguntas vão ser transformadas em modelo de texto embarcado e usadas como estudo pela LLM, para que esta melhore sua compreensão sobre o assunto e seja feita a validação de suas respostas ([Kamoi et al. 2024](#)). As respostas da LLM serão muito importantes nesse artigo para testar quão precisas são as respostas da IA.

Abaixo têm as respostas do Llama 2 de 7 bilhões de parâmetros em relação às perguntas:

Pergunta : "Como funciona um edital de aquisição de veículo?"

Resposta: An edital de aquisição de veículo é um documento que descreve o processo de comprimento de um veículo ou motocicleta para uma entidade pública ou privada, geralmente utilizado em licitações. O edital contém informações sobre as condições da aquisição, como o preço do veículo, o prazo de entrega, as obrigações da contratante e as características do objeto. Aqui estão algumas das funções básicas de um edital de aquisição de veículo: 1. Descrever as condições da aquisição: O edital deve descrever as condições da aquisição, incluindo o preço do veículo, o prazo de entrega e as obrigações da contratante. 2. Fornecer informações sobre o veículo: O edital deve fornecer informações detalhadas sobre o veículo ou motor.

Pergunta: O que é mais barato, alugar, comprar ou leasing de um veículo?

Resposta: The cost of renting, buying or leasing a vehicle varies depending on several factors, including the type of vehicle, the location, and the duration of the rental or purchase agreement. Here are some general estimates for each option: Renting a vehicle:

-The average daily rate for renting a car in Brazil is around BRL 50-70 (around USD 15-20).

-For a short-term rental (less than 3 months), the total cost can range from BRL 500 to BRL 1,000 (around USD 150 to USD 400).

* For a long-term rental (more than 3 months), the total cost can range from BRL 2,000 to BRL 5,000 (around USD 600 to USD 1,500).

Buying a vehicle:

The average price of a new car in Brazil is around BRL 30,000-40,000 (around USD 8,000-16,000).

Used cars can be purchased for lower prices, with an average cost of around BRL 15,000-25,000 (around USD 4,000-7,000).

Leasing a vehicle:

The average monthly payment for leasing a car in Brazil is around BRL 1,000-2,000 (around USD 300-600).

The total cost of leasing a vehicle can range from BRL 10,000 to BRL 50,000 (around USD 3,000 to USD 12,000), depending on the length of the lease agreement.

It's worth noting that these estimates are based on average prices and can vary depending on several factors, such as the make and model of the vehicle, the location, and the duration of the rental of the location, as to BRL 5/50,3250,2')

Ao analisar o Llama para analisar documentos em uma RAG, ela se mostrou menos precisa que o GPT-4, não só alucinando em suas respostas, mas respondendo as perguntas que lhe eram feitas em português em inglês. Analisando isso, foi preferida uma inteligência artificial que respondesse seus comandos em português e não tivesse um preço tão exorbitante, foi usada a IA sabiá.

Uma alternativa nacional

Como o GPT vem de uma empresa americana, logo o preço de seus serviços é em dólar e as LLM's gratuitas não têm uma performance muito boa em relação à uma paga. Pensando nisso, foi usada também uma LLM nacional para realizar os experimentos de RAG, o sabiá, que é uma LLM desenvolvida pela empresa Maritaca AI.

O sabiá-3 tem um preço acessível, como mostra a imagem abaixo:



Figura 6: Preço da API do sabiá-3

Pergunta: Como funciona um edital de aquisição de veículo?

Resposta: Um edital de aquisição de veículo é um documento oficial que detalha o processo pelo qual uma entidade pública ou privada adquire veículos para sua frota ou uso específico. O edital estabelece as regras, especificações técnicas, critérios de avaliação e demais condições que os licitantes devem seguir para participar do processo licitatório. Vou explicar como funciona um edital de aquisição de veículo com base nos documentos fornecidos:

1. **Justificativa e Objetivo**: O edital começa com uma justificativa para a necessidade da aquisição e o objetivo do processo licitatório. Isso inclui a descrição dos requisitos da contratação, como no caso do Estudo Técnico Preliminar (ETP) para a aquisição de viaturas de Representação, de Comando e Administrativas.
2. **Descrição do Objeto**: O edital deve descrever detalhadamente o objeto da licitação, ou seja, o tipo de veículo que está sendo adquirido. Isso inclui especificações técnicas, quantidades e quaisquer outros requisitos relevantes.
3. **Critérios de Avaliação**: O edital estabelece os critérios pelos quais as propostas serão avaliadas. No caso de um pregão eletrônico, como o mencionado no documento, o tipo de critério é geralmente o de menor preço por item.
4. **Documentação Necessária**: O edital detalha toda a documentação que os licitantes devem apresentar, incluindo propostas de preços, declarações unificadas e quaisquer outros documentos necessários para a habilitação e qualificação.
5. **Condições de Entrega**: O edital especifica as condições de entrega dos veículos, incluindo locais de entrega, prazos e condições de pagamento.
6. **Garantias e Manutenção**: O edital pode exigir que o fornecedor ofereça garantias e manutenção para os veículos adquiridos, bem como a disponibilidade de assistência técnica.
7. **Sanções e Penalidades**: O edital deve conter informações sobre as sanções e penalidades aplicáveis em caso de descumprimento das obrigações contratuais.
8. **Validade das Propostas**: O edital define a validade das propostas apresentadas pelos licitantes, que geralmente é de 90 dias corridos a partir da data de abertura da sessão pública.
9. **Assinatura e Autenticidade**: O edital deve ser assinado por autoridades competentes e conter informações sobre a autenticidade do documento, como códigos de verificação e locais de validação.

Ao participar de um processo licitatório para aquisição de veículos, os interessados devem seguir rigorosamente as instruções e requisitos estabelecidos no edital para garantir a conformidade e a competitividade de suas propostas.

Também para perguntas para resolver problemas de decisão, o *sabiá-3* teve respostas concisas de acordo com o contexto dado:

Pergunta: Qual o preço de um carro? Baseado nos documentos disponibilizados.

Resposta: O preço de um carro pode variar dependendo do modelo e das características específicas mencionadas nos documentos fornecidos. Aqui estão alguns exemplos de preços para diferentes tipos de veículos:

- Automóvel (Mínimo 1.3 L, 5 portas, Álcool/Gasolina, 90 CV): R\$ 99.314,63 - Veículo Transporte Pessoal (Aplicação Militar, Tipo Automóvel, Tração 4x2, Gasolina): R\$ 133.692,54
- Veículo Transporte Pessoal (Aplicação Civil, Tipo Automóvel, Tração 4x2, Gasolina): R\$ 105.066,80
- SUV (Mínimo 2.8l, 4 portas, Diesel, 170 CV, Cor Preta): R\$ 534.360,00

Esses são apenas alguns exemplos dos preços mencionados nos documentos. O preço específico de um carro pode ser encontrado nas descrições detalhadas para cada item listado.

Pergunta: O que é mais barato, alugar, comprar ou leasing de um veículo?

Resposta: Para determinar se é mais barato alugar, comprar ou optar por um leasing de um veículo, é necessário considerar vários fatores, como os custos iniciais, custos de manutenção, desvalorização do veículo, seguro, impostos, e o tempo de uso previsto para o veículo. Cada opção tem suas vantagens e desvantagens específicas.

1. **Aluguel**: Geralmente, o aluguel de veículos é mais adequado para necessidades temporárias ou esporádicas. Os custos são normalmente calculados com base no uso (quilômetros rodados e dias de uso), e geralmente incluem manutenção e seguro. Pode ser mais caro a longo prazo, mas oferece flexibilidade e não requer um investimento inicial significativo.
2. **Compra**: A compra de um veículo envolve um investimento inicial alto, mas pode ser mais econômica a longo prazo se o veículo for usado por um período prolongado. A compra também permite personalização e não há limitações de quilometragem. No entanto, há custos adicionais como seguro, manutenção e desvalorização.
3. **Leasing (Arrendamento Mercantil)**: O leasing é uma espécie de aluguel de longo prazo, onde o usuário paga uma quantia mensal pelo uso do veículo. Ao final do contrato, o usuário pode optar por comprar o veículo pelo valor residual. O leasing pode ser mais barato que a compra à vista, pois os pagamentos são geralmente menores do que os de um financiamento. No entanto, pode haver limitações de quilometragem e custos adicionais caso essas limitações sejam excedidas.

Para uma resposta mais precisa, seria necessário analisar o contexto específico, como a frequência de uso, a finalidade do veículo, o orçamento disponível e os termos de cada contrato de aluguel, compra ou leasing. Portanto, eu não sei qual é a opção mais barata sem mais detalhes sobre o uso pretendido e as condições financeiras do interessado.

E também tem uma precisão muito boa para responder e analisar problemas, como demonstrado no artigo (Abonizio et al. 2024) que mostra que a LLM tem uma taxa de 86,6% de acerto ao responder questões no ENEM, mostrando uma boa compreensão de língua portuguesa e de resoluções de problemas.

Ao fazer perguntas de licitação ao *sabiá-3* como:

Discussão

Pelos trabalhos analisados de outros artigos é possível concluir que as LLM's têm uma boa avaliação para busca de dados, escrita de documentos e auxílio em problemas de decisão. Concluindo que é possível usar LLM's no setor público e trazer vários benefícios. Abaixo tem alguns exemplos de artigos citando a precisão de IA ao auxiliar a análise de preços e detecção de fraudes, como um exemplo de como uma LLM pode ser útil para um órgão público ao escrever uma licitação. Depois, uma tabela mostrando a precisão das LLM's utilizadas neste artigo em alguns testes de inteligência artificial e no vestibular ENEM, com o intuito de provar a sua eficácia em interpretar e escrever documentos.

No artigo (Hong et al. 2024) o modelo Data Interpreter conseguiu uma avaliação de 96% de respostas corretas na previsão de preços de casas de acordo com os dados do terreno e área, demonstrando potencial para prever preços e custos em projetos de licitações.

No artigo (Modrušan, Rabuzin e Mrsic 2020) o algoritmo Logistic Regression obteve 76% de acurácia ao detectar fraudes em processos de licitação pública.

Na tabela abaixo tem a métrica das LLM's Llama 2 7B, sabiá-3 e GPT-4o para as seguintes métricas:

- WSC: O WSC foi introduzido tanto como uma alternativa ao Teste de Turing quanto como um teste da capacidade de um sistema de realizar raciocínio de senso comum.
- BLUEX: O BLUEX é um conjunto de dados projetado para avaliar modelos de linguagem em português.
- ASSIN 2 RTE(F1): Avaliação de Similaridade Semântica e Inferência Textual é um teste que se refere à tarefa de medir o grau de semelhança em significado entre dois textos.
- BoolQ: BoolQ é um conjunto de dados para tarefas de resposta a perguntas do tipo sim ou não, contendo 15.942 exemplos. Essas perguntas surgem de forma natural, sendo geradas em contextos não induzidos e sem restrições.
- FaQuAD(F1): O conjunto de dados FaQuAD é um conjunto de leitura e compreensão projetado para avaliar modelos de resposta a perguntas.
- IMDB(Acc): O conjunto de dados IMDb Movie Reviews é um dataset de análise de sentimento binária, composto por 50.000 avaliações do Internet Movie Database (IMDb) classificadas como positivas ou negativas.
- ENEM: Apesar de ser um vestibular e não um dataset este ainda é um bom parâmetro para analisar a interpretação de texto e conhecimento geral da IA.

LLM	GPT-4o	Llama2	Sabia-3
ENEM(Acc)	84,9%	23,07%	86,6%
BLUEX(Acc)	86,87%	32,02%	86,2%
ASSIN 2 RTE(F1)	90,96%	56,82%	88,7%
BoolQ(Acc)	86,50%	57,37%	75,96%
WSC(Acc)	92,63%	47,72%	74,74%
FaQuAD(F1)	84,74%	77,38%	88,47%
IMDB(Acc)	97,00%	86,92%	92,76%

Essas métricas são muito importantes para julgar a compreensão de analisar textos e analisar sua precisão em responder perguntas de um determinado documento e extrair informações valiosas.

Conclusão

Neste trabalho de conclusão de curso, foi feito o estudo de como a inteligência artificial pode auxiliar o setor público. Sob a metodologia da revisão sistemática da literatura, foi selecionada uma gama extensa de artigos científicos relacionados ao tema. Pela análise dos trabalhos investigados, é possível concluir que a RAG tem potencial para auxiliar servidores públicos na hora de analisar documentos e resolver problemas de decisão.

Como o número de LLM's no mercado é extenso, foi escolhida três para serem analisadas nesse estudo: O GPT-4o, o Llama2 e o sabiá-3. Esses modelos foram escolhidos pela sua precisão em resolver NLP, porém eles diferem bastante em seu custo.

O GPT-4o tem as melhores respostas, porém com sua cotação em dólar fica inviável sua implementação em um setor governamental. O Llama2 é completamente gratuito, mas alucina em suas respostas, respondendo às perguntas em inglês. O que fica como o melhor custo-benefício seria o sabiá-3, que possui boa interpretação de texto pela sua precisão em provas como o ENEM e possui um

custo acessível. De acordo com a análise das LLM's a que mais se encaixa no projeto pelo seu preço e escalabilidade seria o sabiá-3.

Referências

- Abonizio et al. 2024 ABONIZIO, H. et al. *Sabiá-3 Technical Report*. 2024. Disponível em: <https://arxiv.org/abs/2410.12049>.
- Attrapadung Koki Hamada 2021 ATTRAPADUNG KOKI HAMADA, D. I. N. *Adam in Private: Secure and Fast Training of Deep Neural Networks with Adaptive Moment Estimation*. [S.l.]: Arxiv, 2021. <https://arxiv.org/pdf/2106.02203>.
- Bai et al. 2024 BAI, Y. et al. *LongWriter: Unleashing 10,000+ Word Generation from Long Context LLMs*. 2024. Disponível em: <https://arxiv.org/abs/2408.07055>.
- Bano e Zowghi 2015 BANO, M.; ZOWGHI, D. A systematic review on the relationship between user involvement and system success. *Information and Software Technology*, 2015. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0950584914001505>.
- Batini et al. 2009 BATINI, C. et al. Methodologies for data quality assessment and improvement. *ACM Comput. Surv.*, Association for Computing Machinery, 2009. Disponível em: <https://doi.org/10.1145/1541880.1541883>.
- Bridle 1989 BRIDLE, J. S. *Probabilistic Interpretation of Feedforward Classification Network Outputs, with Relationships to Statistical Pattern Recognition*. [S.l.]: Springer Link, 1989. https://link.springer.com/chapter/10.1007/978-3-642-76153-9_28.
- Brown Benjamin Mann 2020 BROWN BENJAMIN MANN, N. R. e. M. S. T. B. *Language Models are Few-Shot Learners*. [S.l.]: arXiv, 2020. <https://arxiv.org/pdf/2005.14165>.
- Carbonell, Michalski e Mitchell 1983 CARBONELL, J. G.; MICHALSKI, R. S.; MITCHELL, T. M. 1 - an overview of machine learning. In: MICHALSKI, R. S.; CARBONELL, J. G.; MITCHELL, T. M. (Ed.). *Machine Learning*. San Francisco (CA): Morgan Kaufmann, 1983. p. 3–23. ISBN 978-0-08-051054-5. Disponível em: <https://www.sciencedirect.com/science/article/pii/B9780080510545500054>.
- Dalal, Li e Taori 2019 DALAL, M.; LI, A. C.; TAORI, R. *Autoregressive Models: What Are They Good For?* 2019. Disponível em: <https://arxiv.org/abs/1910.07737>.
- Dettmers et al. 2023 DETTMERS, T. et al. *QLoRA: Efficient Finetuning of Quantized LLMs*. 2023. Disponível em: <https://arxiv.org/abs/2305.14314>.
- Eken 2022 EKEN, G. *Using natural language processing for automated construction contract review during risk assessment at the bidding stage*. Tese (Doutorado) — Middle East Technical University, 2022.
- Fleischer et al. 2024 FLEISCHER, D. et al. *RAG Foundry: A Framework for Enhancing LLMs for Retrieval Augmented Generation*. 2024. Disponível em: <https://arxiv.org/abs/2408.02545>.
- Gallant 1990 GALLANT, S. I. *Perceptron-based learn algoritms*. [S.l.]: IEEE, 1990. <https://www.ling.upenn.edu/courses/cogs501/Gallant1990.pdf>.
- Goodfellow, Bengio e Courville 2016 GOODFELLOW, I.; BENGIO, Y.; COURVILLE, A. *Deep Learning*. [S.l.]: MIT Press, 2016. <http://www.deeplearningbook.org>.
- Harikrishnakumar et al. 2019 HARIKRISHNAKUMAR, R. et al. Supervised machine learning approach for effective supplier classification. In: *2019 18th IEEE International Conference On Machine Learning And Applications (ICMLA)*. [S.l.: s.n.], 2019. p. 240–245.
- Hong et al. 2024 HONG, S. et al. MetaGPT: Meta programming for a multi-agent collaborative framework. In: *The Twelfth International Conference on Learning Representations*. [s.n.], 2024. Disponível em: <https://openreview.net/forum?id=VtmBAGCN7o>.
- Howard e Ruder 2018 HOWARD, J.; RUDER, S. *Universal Language Model Fine-tuning for Text Classification*. 2018. Disponível em: <https://arxiv.org/abs/1801.06146>.
- Hugging Face Hugging Face. *Hugging Face transformer*. <https://huggingface.co/docs/transformers/index>.

- Jiang et al. 2024 JIANG, A. Q. et al. *Mixtral of Experts*. 2024. Disponível em: <https://arxiv.org/abs/2401.04088>).
- Kamoi et al. 2024 KAMOI, R. et al. *Evaluating LLMs at Detecting Errors in LLM Responses*. 2024. Disponível em: <https://arxiv.org/abs/2404.03602>).
- Kitchenham 2004 KITCHENHAM, B. *Procedures for Performing Systematic Reviews*. [S.l.]: Citeseer, 2004. <https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=29890a936639862f45cb9a987dd599dce9759bf5>).
- Lewis et al. 2021 LEWIS, P. et al. *Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks*. 2021. Disponível em: <https://arxiv.org/abs/2005.11401>).
- Li 2022 LI, H. Bidding price forecast of construction projects based on machine learning algorithm. In: *2022 International Conference on Computation, Big-Data and Engineering (ICCBDE)*. [S.l.: s.n.], 2022. p. 54–57.
- Liu et al. 2023 LIU, J. et al. *LLMRec: Benchmarking Large Language Models on Recommendation Task*. 2023. Disponível em: <https://arxiv.org/abs/2308.12241>).
- Modrušan, Rabuzin e Mrcic 2020 MODRUŠAN, N.; RABUZIN, K.; MRSIC, L. Improving public sector efficiency using advanced text mining in the procurement process. In: . [S.l.: s.n.], 2020. p. 200–206.
- Naveed et al. 2024 NAVEED, H. et al. *A Comprehensive Overview of Large Language Models*. 2024. Disponível em: <https://arxiv.org/abs/2307.06435>).
- Open AI Open AI. *Open AI transformer*. [\(\(https://platform.openai.com/docs/guides/embeddings/what-are-embeddings\)\)](https://platform.openai.com/docs/guides/embeddings/what-are-embeddings).
- Pandya e Holia 2023 PANDYA, K.; HOLIA, M. *Automating Customer Service using LangChain: Building custom open-source GPT Chatbot for organizations*. 2023. Disponível em: <https://arxiv.org/abs/2310.05421>).
- Papineni Salim Roukos 2002 PAPINENI SALIM ROUKOS, T. W. W.-J. Z. K. *BLEU: a Method for Automatic Evaluation of Machine Translation*. [S.l.]: ACL Anthology, 2002. <https://aclanthology.org/P02-1040.pdf>).
- Rhythm et al. 2023 RHYTHM, E. R. et al. *Text-based QA: Automated Question Generation and Answering for Enhanced Data Processing*. 2023. 1-5 p.
- Rogers, Kovaleva e Rumshisky 2021 ROGERS, A.; KOVALEVA, O.; RUMSHISKY, A. A Primer in BERTology: What We Know About How BERT Works. *Transactions of the Association for Computational Linguistics*, v. 8, p. 842–866, 01 2021. ISSN 2307-387X. Disponível em: https://doi.org/10.1162/tacl_a_00349).
- Russell e Norvig 2010 RUSSELL, S. J.; NORVIG, P. *Artificial intelligence : a modern approach, Third Edition*. [S.l.]: Prentice-Hall, 2010. <https://people.engr.tamu.edu/guni/csce421/files/AI.Russell.Norvig.pdf>).
- Sasaki 2007 SASAKI, R. F. Y. *The truth of the F-measure*. [S.l.]: School of Computer Science, University of Manchester, 2007. https://nicolasshu.com/assets/pdf/Sasaki_2007_The%20Truth%20of%20the%20F-measure.pdf).
- Shao et al. 2024 SHAO, Y. et al. *Assisting in Writing Wikipedia-like Articles From Scratch with Large Language Models*. 2024. Disponível em: <https://arxiv.org/abs/2402.14207>).
- Tang e Yu 2024 TANG, X.; YU, H. Efficient large-scale personalizable bidding for multiagent auction-based federated learning. *IEEE Internet of Things Journal*, v. 11, n. 15, p. 26518–26530, 2024.
- Valim 2024 VALIM, E. F. Análise e avaliação do desempenho de técnicas de machine learning para product matching. *Repositorio UFMS*, 2024. <https://repositorio.ufms.br/handle/123456789/9103>).
- Vaswani 2023 VASWANI, A. Attention is all you need. *Google scholar*, 2023. <https://arxiv.org/abs/1706.03762>).

[//arxiv.org/abs/1706.03762](https://arxiv.org/abs/1706.03762)).

Wei et al. 2023 WEI, F. et al. Empirical study of llm fine-tuning for text classification in legal document review. In: *2023 IEEE International Conference on Big Data (BigData)*. [S.l.: s.n.], 2023. p. 2786–2792.

Wikipedia Wikipedia. *Large Language Model*. (https://en.wikipedia.org/wiki/Large_language_model).

Wythoff 1992 WYTHOFF, B. J. *Backpropagation neural networks A tutorial*. [S.l.]: Science direct, 1992. (<https://www.sciencedirect.com/science/article/pii/016974399380052J>).

Yang, Ji e Jing 2021 YANG, Y.; JI, T.; JING, Z. Selective learning for strategic bidding in uniform pricing electricity spot market. *CSEE Journal of Power and Energy Systems*, v. 7, n. 6, p. 1334–1344, 2021.

Zuiderwijk 2021 ZUIDERWIJK, A. Implications of the use of artificial intelligence in public governance: A systematic literature review and a research agenda. *Google scholar*, 2021.

Zuiderwijk 2021 ZUIDERWIJK, Y.-C. C. F. S. A. *Implications of the use of artificial intelligence in public governance: A systematic literature review and a research agenda*. [S.l.]: Elsevier, 2021. (<https://www.sciencedirect.com/science/article/pii/S0740624X21000137>).