ESCOLA ESTADUAL DE ENSINO MÉDIO PROFª. MARIA ROCHA
CURSO TÉCNICO EM INFORMÁTICA INTEGRADO
AO ENSINO MÉDIO

INTELIGÊNCIA ARTIFICIAL NA EDUCAÇÃO

TRABALHO DE CONCLUSÃO DE CURSO

SAMUEL SCHIFELBEIN CAETANO

Santa Maria, RS
2024

ESCOLA ESTADUAL DE ENSINO MÉDIO PROFª. MARIA ROCHA
CURSO TÉCNICO EM INFORMÁTICA INTEGRADO
AO ENSINO MÉDIO

SAMUEL SCHIFELBEIN CAETANO

Trabalho de Conclusão de Curso apresentado como requisito parcial para obtenção do título de Técnico em Informática.

BANCA EXAMINADORA:

_____
Prof.
E. E. E. M. Profª. Maria Rocha

_____
Prof.
E. E. E. M. Profª. Maria Rocha

Santa Maria, XX de Dezembro de 2024.

LISTA DE FIGURAS

- Figura 1: Mapa mental de um *fine-tuning*
- Figura 2: Mapa mental da mentalidade de análise do ser humano perante problemas e respostas
- Figura 3: Página principal do site do Ollama - Tela de Download
- Figura 4: Download Concluído do Ollama
- Figura 5: Progresso de instalação do Ollama no Windows 10
- Figura 6: Ollama executado e o ícone sendo mostrado na barra de notificações
- Figura 7: Abrindo o terminal (CMD) no Windows 10
- Figura 8: Comando digitado “ollama run mistral” no terminal do Windows 10
- Figura 9: Tela inicial Do Site
- Figura 10: Tela de teoria do Site
- Figura 11: Tela sobre aplicação das inteligências artificiais do Site
- Figura 12: Tela de login do site da inteligência artificial (Renascer)
- Figura 13: Figura após login, pronta para prompts textuais
- Figura 14: Um exemplo de prompt textual

SUMÁRIO

- INTRODUÇÃO 6
- FUNDAMENTAÇÃO TEÓRICA 9
- 2.1. Machine Learning e a linguagem natural 10
- 2.2. TTS (Texto para fala) e ASR (Reconhecimento de Voz Automática) 10
- 2.3. Atualmente 11
- DESENVOLVIMENTO 12
- CONCLUSÃO 18
- REFERÊNCIAS BIBLIOGRÁFICAS 19

Dicionário de Acrônimos

(LLM) - Modelo de Linguagem Grande

(NLP) - Processamento de Linguagem Natural

(ML) - Aprendizado de Máquina

(LoRA) - Adaptação de Baixo Posto

(QLoRA) - Adaptação de Baixo Posto Quantizada

(IA) - Inteligência Artificial

(TTS) - Texto Para Fala

(ASR) - Reconhecimento de Voz Automática

(AGI) - Inteligência Artificial Geral

(ChatGPT) - Generative PreTrained Chat (Plataforma de bate-papo com uma IA de texto/áudio gerado previamente treinado

1 INTRODUÇÃO

O presente trabalho de conclusão de curso pretende compreender o impacto das inteligências artificiais em cotejo com a literatura acadêmica atual. Estas páginas têm por objetivo buscar compreender como tais tecnologias vêm modificando as práticas pedagógicas, e criar, com base em um modelo de linguagem grande (LLM), uma IA para auxiliar na educação e realizar testes vocacionais.

Segundo os desenvolvedores do Google Cloud (2008), diz que a inteligência artificial nada mais é do que um termo guarda-chuva para métodos de cálculos, alguns podendo ser programada com conhecimento prévio estático - contendo uma imensa quantidade de condições (if) - enquanto outros podem ser de aprendizado profundo, como redes neurais artificiais e modelos de linguagens.

Entretanto, neste trabalho estarei me aprofundando no processamento de linguagem natural (NLP) cujo aprendizado de máquina (*machine learning)* é uma componente do processo.

Cortiz, D. em seu vídeo “Curso de Inteligência Artificial para todos - Aula 1” (2020) diz que o aprendizado de máquina pode ser categorizado genericamente em 4 categorias:

Aprendizado supervisionado, onde o algoritmo é treinado com um conjunto de dados rotulados, ou seja, cada exemplo de entrada do conjunto de dados possui uma saída desejada associada. Suponha que exista um conjunto de imagens de animais, onde cada imagem é rotulada com o tipo de animal que contém. O algoritmo supervisionado é bom em identificar com precisão que animal está na foto.

Aprendizado não supervisionado, onde o algoritmo é treinado com um conjunto de dados sem rótulos, não servindo para classificação por não ser rotulado.

Aprendizado por reforço, é aquele que é baseado em técnicas de tentativas e erros em ambientes controlados onde a máquina é punida com uma reação incorreta.

E por fim, aprendizado semi-supervisionado, sendo uma mistura de técnicas de aprendizado supervisionado com não supervisionado. O que gera um novo comportamento, que é a abundância de dados não rotulados e a escassez de dados rotulados.

O NLP diz respeito à interação entre computadores e linguagem humana, abrangendo a capacidade dos sistemas computacionais de compreender, interpretar e produzir texto ou fala em linguagem natural.

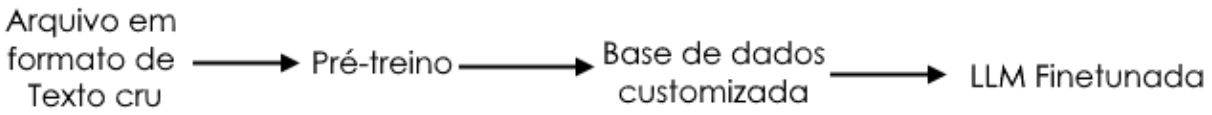
O referido processamento abrange uma variedade de tarefas, incluindo análise de sentimento, tradução automática, reconhecimento de fala, extração de informações, entre outras. Os sistemas de NLP utilizam técnicas de aprendizado de máquina e análise de dados para compreender e gerar linguagem humana eficaz. É possível criar linguagens de processamento natural utilizando a linguagem de programação Python e suas bibliotecas TensorFlow, PyTorch e Transformers.

Porém, os sistemas NLP necessitam ser pré-treinados para tornarem-se funcionais, isto é, exigem adicionar e filtrar informações. Este processo é chamado de *fine-tuning* ou ajuste-fino.

Quanto mais bilhões de parâmetros, mais completa e pesada é a LLM.

Há vários tipos de *fine-tuning*:

Figura 1: Mapa mental de um *fine-tuning*



Fonte: Autor (2024)

Ajuste fino completo: envolve treinar todo o LLM em um novo conjunto de dados, visando realizar uma tarefa específica, mas que exige um bom hardware para ser executado, e pode levar ao que é chamado de *overfitting* (sobreajuste). Porém, apresenta um melhor resultado em comparação com os outros métodos citados abaixo.

Ajuste de adaptador:

LoRA (*Low-Rank Adaptation* ou Adaptação de Baixo Posto) :

Refere-se a uma técnica utilizada em aprendizado de máquina, em que modelos complexos, como redes neurais, visam substituir uma matriz original, com dimensões elevadas, por uma matriz de baixo *rank*, com uma estrutura que pode ser matematicamente adaptada, resultando em um produto de matrizes de dimensões menores. Esse procedimento alude ao uso de modelos simplificados ou aproximativos a fim de representar dados complexos.

Aplicar uma "aproximação de baixo posto" significa representar essa matriz como o produto de duas matrizes menores (de baixo *rank*). Em vez de armazenar ΔW como uma matriz completa, ela é numericamente adaptada de forma mais compacta, reduzindo, assim, o espaço de memória necessário.

QLoRA (Quantized Low-Rank Adaptation ou Adaptação de Baixo Posto Quantizada) :

É uma abordagem extensiva da LoRA, que inclui a quantização dos parâmetros do modelo. Essa quantização é o processo de converter valores contínuos, por exemplo, números de ponto flutuante (os famosos “números quebrados”) em valores discretos, como números inteiros,

geralmente representados com menos bits, o que pode reduzir o consumo de memória e acelerar as operações de inferência. Com essa abordagem, a eficiência de memória da GPU é aperfeiçoada, minimizando sua quantidade de uso num percentual de até 75% em comparação com a LoRA[1]. O uso do QLoRA com, por exemplo, 4 bits para a quantização de parâmetros, é uma decisão para maximizar a compactação dos parâmetros, permitindo uma economia de memória e largura de banda durante a transmissão ou armazenamento do modelo. Mas é importante ressaltar que essa modificação afeta o desempenho do modelo, pois a redução da precisão dos parâmetros pode levar a uma perda de acurácia ou qualidade nos resultados de inferência.

2. FUNDAMENTAÇÃO TEÓRICA

O que é inteligência artificial e sua filosofia

Conforme Alan Turing, pode-se dizer que inteligência artificial é quando um computador consegue imitar, com precisão, as respostas humanas em circunstâncias específicas. Se uma máquina pode pensar, ela pode pensar de forma mais inteligente do que nós, então, [hierarquicamente], onde estaríamos situados neste contexto? (TURING, 1951).

Para aferir esta precisão, na emulação do pensamento humano, Turing (1951) projetou um teste objetivando definir um quantitativo satisfatório de reconhecimento de inteligência. A máquina passa no teste caso um interrogador não consiga descobrir se as respostas obtidas vêm de uma pessoa ou de uma IA.

A inteligência artificial computacional surgiu na década de 1950 como uma investigação sobre a natureza dessa qualidade cognitiva. Nessa oportunidade foram utilizados computadores como uma ferramenta revolucionária para emulá-la.

Por outro lado, o filósofo John Searle (2014) postula que, mesmo se construíssemos uma máquina capaz de simular perfeitamente todos os aspectos do cérebro humano, esta conquista não prova, conclusivamente, que a máquina capta qualquer significado. Embora possa exibir um comportamento sugestivo de compreensão, isso fica aquém da compreensão real. Na verdade, Searle enfatiza que a consciência humana transcende a mera simulação de padrões apropriados de entrada e saída, implicando em que a captura completa da essência da cognição humana representa um desafio contínuo na investigação de uma inteligência artificial. Esses conceitos estão interligados, pois o domínio das tarefas por meio da aprendizagem de máquina permite que os computadores simulem, de forma mais precisa, os comportamentos humanos (SEARLE, 2014).

[1] Veja mais em: **Recomendações da LoRA e QLoRA para LLMs.** Google. <https://cloud.google.com/vertex-ai/docs/model-garden/lora-qlora?hl=pt-br> Acesso em 26 de abril de 2024.

Como ficaria muito extenso e complexo criar-se uma inteligência artificial que replicasse a totalidade de funções de um ser humano em diversas áreas e que, por limitação finalística ao serviço humano, acabasse sendo subutilizada no exercício de apenas uma dessas funções, foi decidido dividir o conceito de IAs em subcampos, como, por exemplo, o aprendizado de máquina. Ele é um subcampo da inteligência artificial, focado no desenvolvimento de algoritmos e técnicas que permitem aos computadores aprenderem a partir de dados e fazerem previsões ou tomarem decisões sem serem explicitamente programados para isso. Em vez de apenas seguir estritamente instruções pré-programadas, os sistemas de aprendizado de máquina podem aprender com exemplos e experiências anteriores, ajustando seus próprios modelos a fim de melhorar seu desempenho ao longo do tempo. O *machine learning* é utilizado em uma grande variedade de áreas, incluindo a análise de dados, o processamento de imagens, os sistemas de recomendação, reconhecimento de padrões, entre outros.

Em 1955, Arthur Samuel criou um código para uma máquina alimentada com *machine learning*, que aprendeu a jogar jogo de damas.

> "A filosofia da inteligência artificial tenta responder a questões como:
>
> - Pode uma máquina agir inteligentemente? Pode resolver *qualquer* problema que uma pessoa resolveria através do raciocínio?
> - Pode uma máquina possuir uma mente, estados mentais e uma consciência, da mesma maneira que os seres humanos possuem? As máquinas podem *sentir*?
> - A inteligência humana e a inteligência de uma máquina são idênticas? É o cérebro humano essencialmente um computador?"
>
> Fonte: Wikipédia

2.1. Machine Learning e a linguagem natural.

Estes conceitos estão interligados, pois o domínio das tarefas, por meio do *machine learning*,permite que os computadores emulam o comportamento humano. Além disso, métodos de aprendizagem de máquina, como redes neurais, facilitam a criação, semelhante à cognição humana, de raciocínio e comunicação, se aproximando da visão de Turing na qual, máquinas artificialmente inteligentes são indistinguíveis de seres humanos.

2.2. TTS (Texto para fala) e o (ASR) Reconhecimento de voz automático

Desde metade do século passado são estudadas maneiras de recriar a habilidade da fala, para aqueles que estão impossibilitados por capacidades cognitivas e motoras ou simplesmente fazer com que uma máquina falasse, a exemplo disso temos o S.A.M. (Software de fala automática) criado em 1982 para o Commodore 64 e outras plataformas.

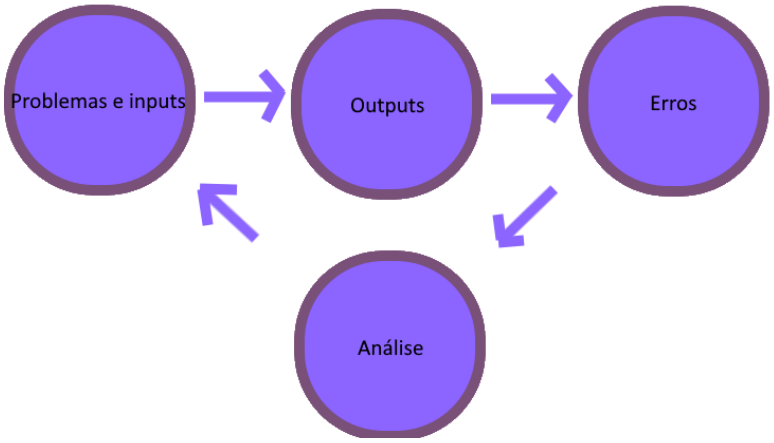
E para ouvir? Surgiu o ASR, reconhecendo voz para substituir teclados manuais e outros mais sofisticados para escutar, pesquisar e "compreender" o que foi dito anteriormente, como por exemplo o atual ChatGPT com sua nova *feature* de compreensão de voz e fala em tempo real.

2.3. Atualmente

Atualmente, *machine learning*, NLP e TTS baseada em IA são desenvolvidas pouco a pouco para criar o que chamamos de AGI (Inteligência artificial geral).

Para os humanos, as problemáticas surgem a partir da necessidade de algo. Não é diferente na tecnologia, pois em surgindo uma problemática, com ela vem igualmente a necessidade de sua resolução.

Figura 2: Mapa mental da mentalidade de análise do ser humano perante problemas e respostas



Fonte: Autor (2024)

Justamente com esta ideia em mente podemos pensar que a evolução tecnológica traz benefícios e malefícios para a sociedade, mas de modo geral se pensarmos apenas com uma mente pesquisadora a evolução pode trazer melhorias.

Com isso em mente, novas tecnologias surgiram ao passar do tempo. Desde Psicólogas virtuais, até criadores de imagem e música.

Google BeRT (11/2018) – É um desses algoritmos. Ele serve para entender justamente o que o usuário digita na barra de pesquisa, inclusive gírias e diversas línguas.

OpenAI ChatGPT (11/2022) – É um outro algoritmo, com funcionalidade de chatbot muito avançado com treinamento com base no wikipédia, github e outros sites da www.

Moises.ai (2019) – Outro programa compilado com inteligências artificiais que detectam progressões harmônicas, batidas por minuto, tonalidade e remoção de voz e instrumentos de uma música com alta qualidade.

Midjourney (5/2022), enquanto o Stable Diffusion foi lançado logo após, (8/2022). Já o DALL-E teve sua segunda versão, o DALL-E 2, lançada em (4/2022), com a primeira versão tendo sido lançada em 2021. Esses três modelos surgiram em um curto espaço de tempo, impulsionando o campo de geração de imagens por IA e rapidamente se tornando populares.

RVC (2024). É uma tecnologia de conversão de voz baseada em recuperação de dados, desenvolvida para transformar a voz de um locutor em outra de forma realista, mantendo características como entonação e timbre. Ela utiliza redes neurais profundas e é amplamente aplicada em criações musicais e covers de músicas utilizando vozes clonadas de cantores ou personagens. Desde o seu lançamento, o RVC tem ganhado destaque por sua capacidade de realizar conversões de voz com alta precisão, especialmente em tempo real.

Character.ai (11/2021) – Simula uma conversa com um personagem criado préviamente, tanto com personagens imaginativos quanto reais.

3. DESENVOLVIMENTO

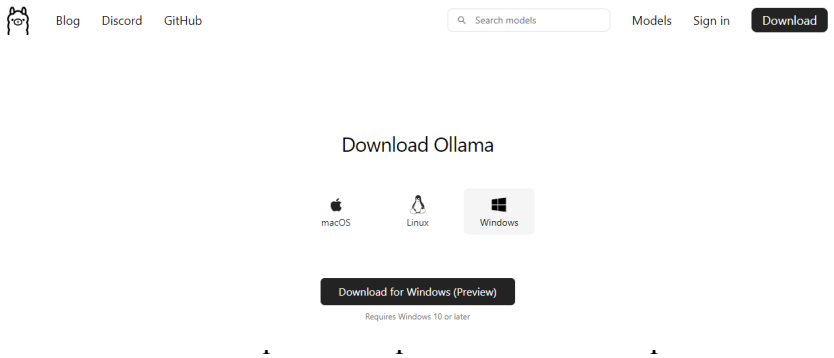
Com o objetivo de enfatizar a importância da tecnologia de Inteligência artificial, ensinaremos como montar um receptáculo para armazenar, modificar e testar ambientes compatíveis com IAs em casa (requisitos necessários sendo: mínimo de 8gb de RAM, Intel Core I5 3470, Nvidia GTX 1050 Ti e um SSD, recomenda-se 80 Gb livres, compatível com SOs Windows 10/11, Linux e MacOS 11 ou superior, todos de 64 bits).

Passos necessários para o *enviroment*:

- Acesse o site da plataforma Ollama para IAs. (https://ollama.com/download)

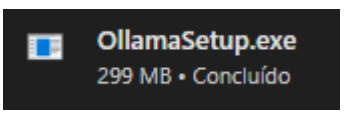
  Estarei demonstrando apenas com o Windows 10.

Figura 3: Página principal do site do Ollama - Tela de Download



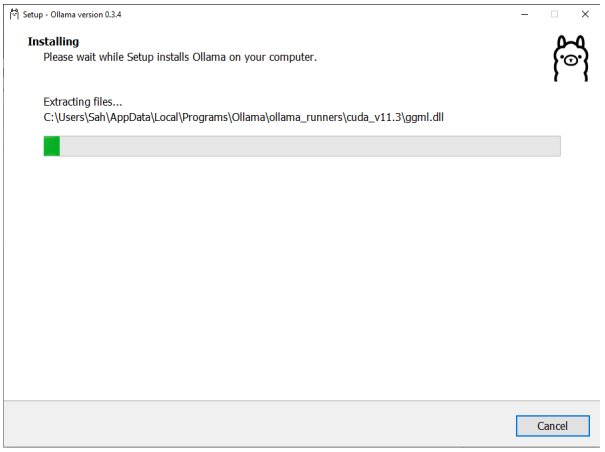
- Baixar e instalar o Ollama.

Figura 4: Download Concluído do Ollama



Fonte: Autor (2024)

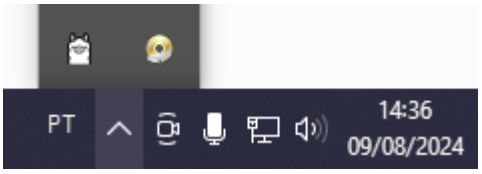
Figura 5: Progresso de instalação do Ollama no Windows 10



Fonte: Autor (2024)

Quando a instalação estiver concluída, verifique se o programa já está ativado e executado no seu sistema. Senão, execute-o.

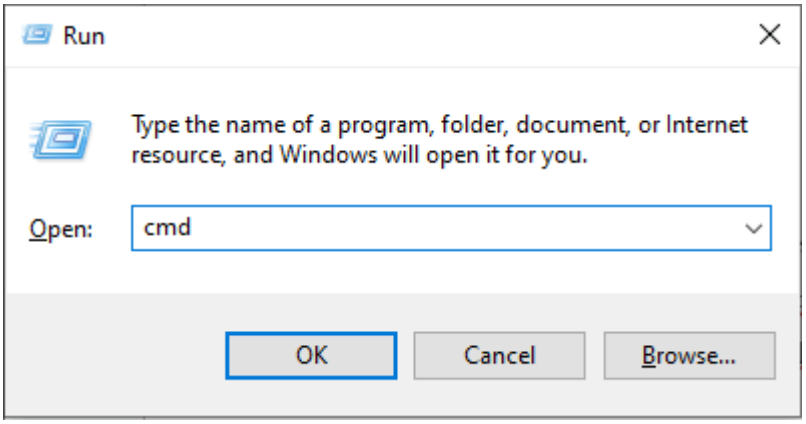
Figura 6: Ollama executado e o ícone sendo mostrado na barra de notificações



Fonte: Autor (2024)

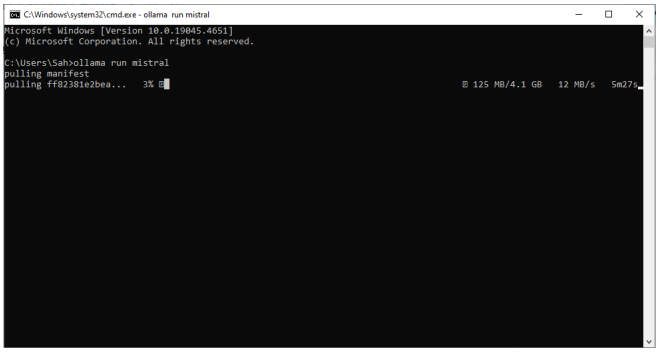
- Baixar um arquivo de IA ".gguf" do Hugging Face ou digite "ollama run" + "nome da IA" no seu terminal que será baixado automaticamente algumas prévias disponibilizadas pelo Ollama.

Figura 7: Abrindo o terminal (CMD) no Windows 10



Fonte: Autor (2024)

Figura 8: Comando digitado “ollama run mistral” no terminal do Windows 10



Fonte: Autor (2024)

Agora o mistral está instalado na máquina com o ollama. (Recomenda-se o llama3 ao invés do mistral).

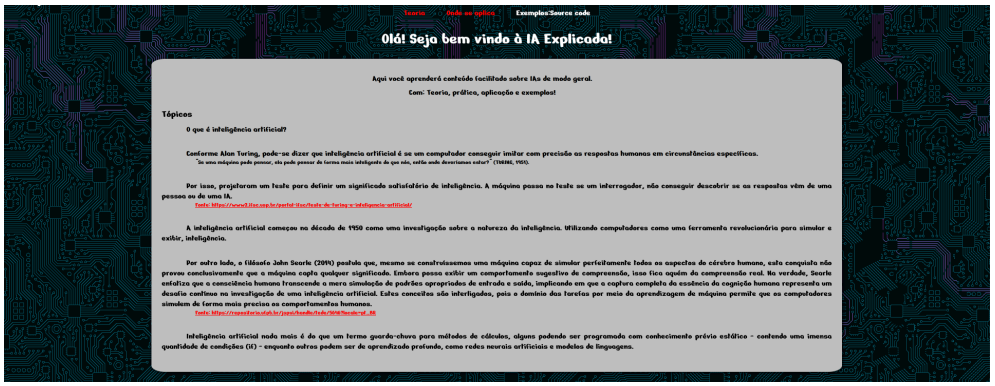
Ativando a Interface Gráfica

- Instale o app do Docker na máquina.
-
- Instale e habilite o Openweb-UI no Docker para gerenciar a interface gráfica e
- administração para o Ollama.

A figura abaixo mostra a tela inicial do site criado para informar como funcionam as IAs.

Você pode acessar em: https://in-telligencia.vercel.app

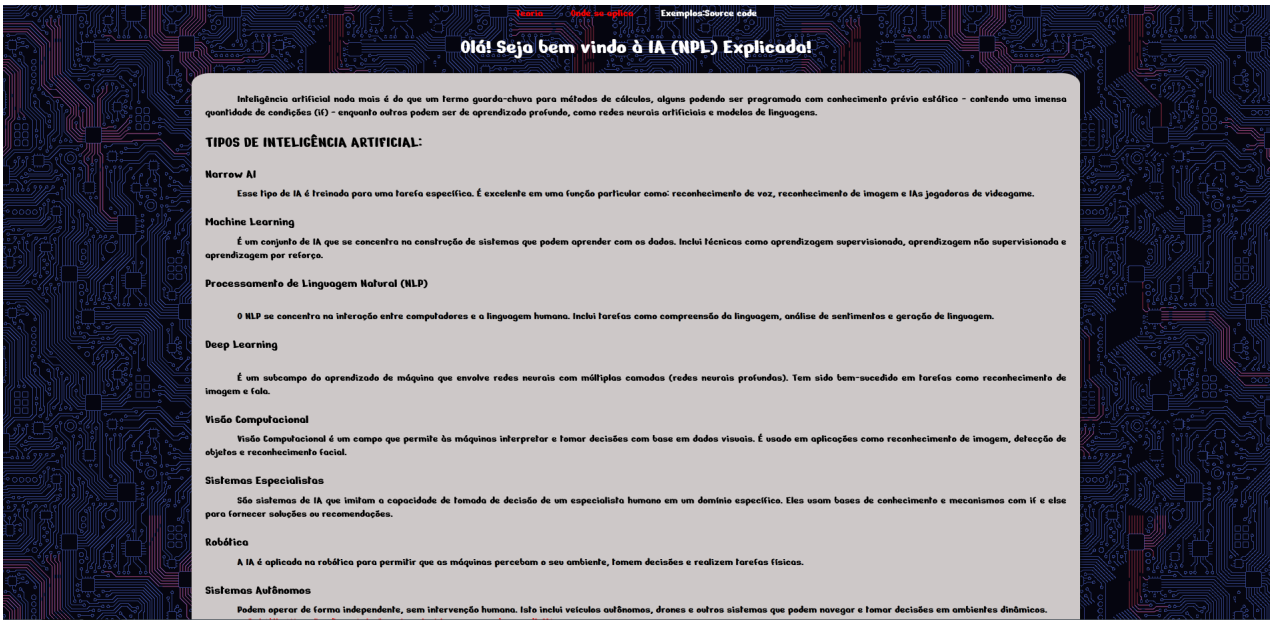
Figura 9 - Tela inicial do Site



Fonte: Autor (2024)

A figura abaixo mostra a página de Teoria sobre IA.

Figura 10 - Tela de teoria do Site

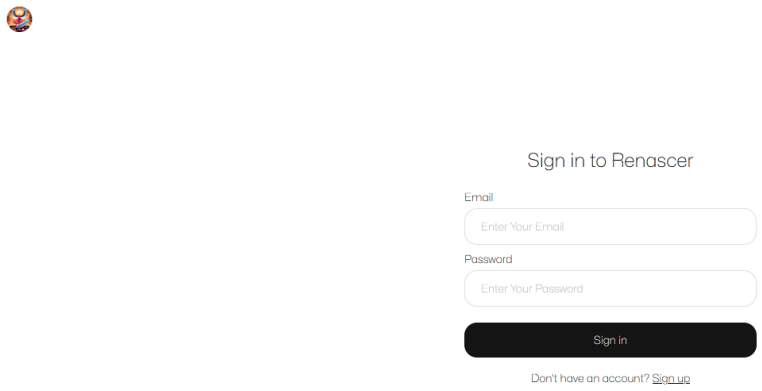


Fonte: Autor (2024)

Figura 11 - Tela sobre aplicação das inteligências artificiais do Site

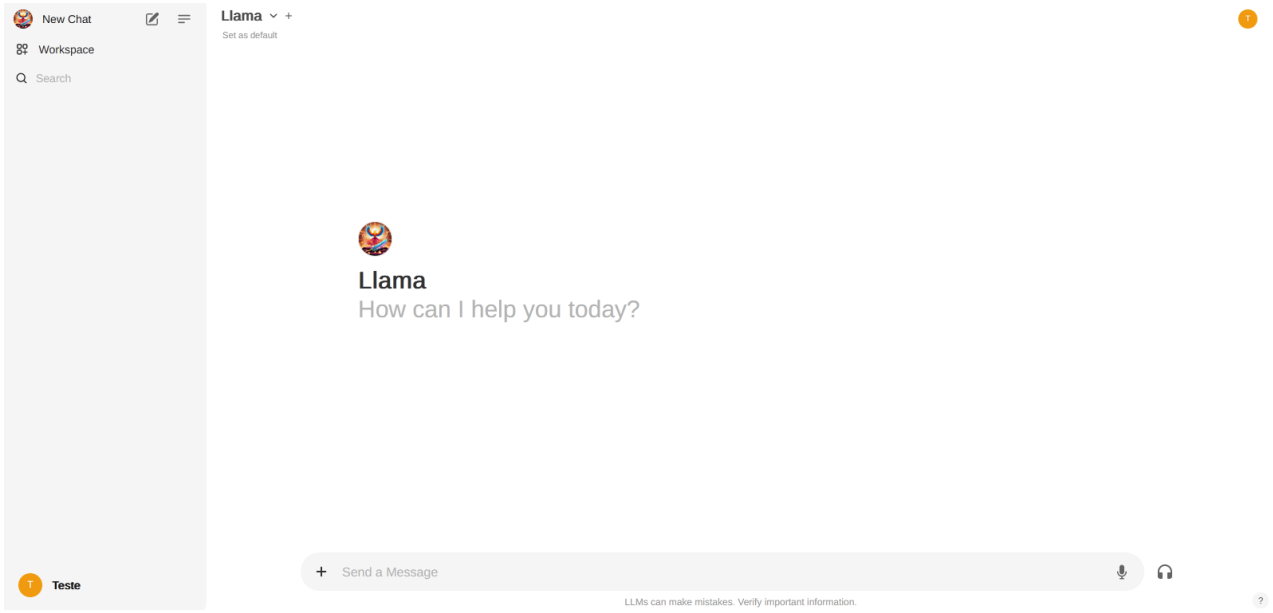
Fonte: Autor (2024)

Figura 12 - Tela de login do site da inteligência artificial (Renascer)



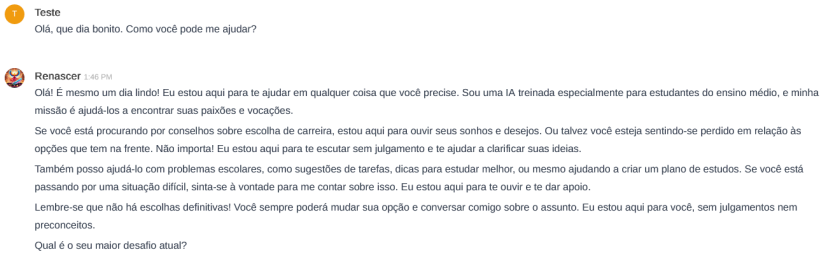
Fonte: Autor (2024)

Figura 13 - Figura após login, pronta para prompts textuais



Fonte: Autor (2024)

Figura 14 - Um exemplo de prompt textual



Fonte: Autor (2024)

CONCLUSÃO

A inteligência artificial é importante para os seres humanos, pois ela ajuda a automatizar tarefas, aprimorar processos e a fornecer informações precisas e rápidas. Auxilia no cotidiano pessoas com deficiências ou doenças; fornecer apoio emocional e conversacional; realizar análise de dados para tomada de decisões; ajuda na resolução de problemas e dúvidas e amplia a capacidade humana em áreas como a medicina, a economia e a educação.

O objetivo deste trabalho é auxiliar os usuários a formarem uma opinião sobre qual vocação seguir; em dúvidas para trabalhos escolares e ajuda psicológica.

Escolhi esse tema porque gosto de desafios, todo este grande processo foi muito interessante e me abriu portas e conhecimentos que eu não conhecia antes e com a inteligência artificial eu posso ajudar outras pessoas a decidirem seus caminhos de vida.

Fazendo esse trabalho, aprendi mais sobre Inteligência Artificial, sua filosofia, Machine Learning e a linguagem natural, bem como hospedar sites na vercel de graça e editá-los online facilitando o gerenciamento e o uso do tempo enquanto fazia o trabalho de pesquisa; aprendi a moldar máquinas virtuais e utilizar o docker com Hyper-V e comandos de Linux.

Futuramente, pretendo estudar, continuar na área e publicar artigos incluindo sobre uma grande rede de computadores *peer-to-peer* de fácil acesso e manutenção.

REFERÊNCIAS BIBLIOGRÁFICAS

TURING, A. **AMT/B/5.1951**. Disponível em: <https://turingarchive.kings.cam.ac.uk/publications-lectures-and-talks-amtb/amt-b-5>. Acesso em 26 de abril de 2024

Understanding the different types of artificial intelligence. IBM, 12 out. 2023. Disponível em: <https://www.ibm.com/blog/understanding-the-different-types-of-artificial-intelligence/>. Acesso em 26 de abril de 2024.

O que é processamento de linguagem natural? Google. Disponível em: <https://cloud.google.com/learn/what-is-natural-language-processing?hl=pt-br> Acesso em 26 de abril de 2024.

GUITARRARA, P. **Inteligência artificia**l; *Brasil Escola*. Disponível em: <https://brasilescola.uol.com.br/informatica/inteligencia-artificial.htm>. Acesso em 26 de abril de 2024.

MITTAL, A. C**ompreendendo o ajuste fino do LLM: adaptando grandes modelos de linguagem aos seus requisitos exclusivos**. 2023. Disponível em: <https://www.unite.ai/pt/entender-o-ajuste-fino%2C-adaptando-grandes-modelos-de-linguagem-%C3%A0s-suas-necessidades-exclusivas/>. Acesso em 14 de abril de 2024.

MITTAL, A. **Otimizando memória para inferência e ajuste fino de modelos de linguagem grande.** 2024. Disponível em : <https://www.unite.ai/pt/otimizando-a-mem%C3%B3ria-para-infer%C3%AAncia-de-modelos-de-linguagem-grande-e-ajuste-fino/>. Acesso em 24 de maio de 2024.

Recomendações da LoRA e QLoRA para LLMs. Google. <https://cloud.google.com/vertex-ai/docs/model-garden/lora-qlora?hl=pt-br> Acesso em 26 de abril de 2024.

Miller, A. **LoRA versus QLoRA: Decifrando o futuro da afinidade em Inteligência Artificial.** 2024. Disponível em: <https://www.linkedin.com/pulse/lora-versus-qlora-decifrando-o-futuro-da-afinidade-em-augusto-jgvpf/>. Acesso em 26 de abril de 2024.

Dettmers, T. Pagnoni, A. Holtzman, A. Zettlemoyer, L. **QLoRA: Efficient Finetuning of Quantized LLMs.** 2023. Disponível em: <https://arxiv.org/abs/2305.14314> Acesso em 26 de abril de 2024.

Transformer (deep learning architecture). Wikipedia, 2024. Disponível em: <https://en.wikipedia.org/wiki/Transformer_(deep_learning_architecture)>. Acesso em 26 de abril de 2024.

Generative pre-trained transformer. Wikipedia, 2024. Disponível em: <https://en.wikipedia.org/wiki/Generative_pre-trained_transformer>. Acesso em 26 de abril de 2024.

Deep Learning. Wikipedia, 2024. Disponível em: <https://en.wikipedia.org/wiki/Deep_learning>. Acesso em 26 de abril de 2024.

O que é aprendizado supervisionado? IBM. Disponível em: <https://www.ibm.com/br-pt/topics/supervised-learning#:~:text=O%20aprendizado%20supervisionado%2C%20conhecido%20tamb%C3%A9m,ou%20preveem%20resultados%20com%20precis%C3%A3o.>.

Bhatta, S. **Applied Machine Learning Solutions with Python**. 2021. Disponível em: <https://www.google.com.br/books/edition/Applied_Machine_Learning_Solutions_with/QD1AEAAAQBAJ?hl=pt-BR&gbpv=1&dq=arthur%20samuel%20machine%20learning%20books>. Acesso em 24 de maio 2024.

Fine-Tuning LLMs: Overview, Methods, and Best Practices. Disponível em: <https://www.turing.com/resources/finetuning-large-language-models>. Acesso em 24 de maio de 2024.

Deep learning - Wikipedia. Disponível em: <https://en.wikipedia.org/wiki/Deep_learning#Hardware>. Acesso em 24 de maio de 2024.

Amorim, Paula Fernanda Patrício de - **A crítica de John Searle à inteligência artificial: uma abordagem em filosofia da mente.** Disponível em: <https://repositorio.ufpb.br/jspui/handle/tede/5646?locale=pt_BR>. Acesso em 9 de agosto de 2024.

O que é a IA forte? - IBM. Disponível em: <https://www.ibm.com/br-pt/topics/strong-ai**>.** Acesso em 9 de agosto de 2024.

Deep learning - Wikipedia. Disponível em: <https://en.wikipedia.org/wiki/Arthur_Samuel_(computer_scientist)**>.** Acesso em 9 de agosto de 2024.