

UNIVERSIDADE FEDERAL DE PELOTAS
Centro de Desenvolvimento Tecnológico
Curso de Bacharelado em Ciência da Computação



Trabalho de Conclusão de Curso

LLM-Powered Applications: Tecnologia, Questões e estudo de caso.

Marilton Sanchotene de Aguiar

Pelotas, 2024

Marilton Sanchotene de Aguiar

LLM-Powered Applications: Tecnologia, Questões e estudo de caso.

Trabalho de Conclusão de Curso apresentado ao Centro de Desenvolvimento Tecnológico da Universidade Federal de Pelotas, como requisito parcial à obtenção do título de Bacharel em Ciência da Computação.

Orientador: Prof. Dr. Marilton Sanchotene de Aguiar
Coorientador: Prof. Dr. Marilton Sanchotene de Aguiar
Colaborador: Prof. Dr. Marilton Sanchotene de Aguiar

Pelotas, 2024

Insira AQUI a ficha catalográfica
Quando finalizado o trabalho, deve ser
solicitada através do Sistema Cobalto
Biblioteca – Cadastro – Ficha catalográfica.

Marilton Sanchotene de Aguiar

LLM-Powered Applications: Tecnologia, Questões e estudo de caso.

Trabalho de Conclusão de Curso aprovado, como requisito parcial, para obtenção do grau de Bacharel em Ciência da Computação, Centro de Desenvolvimento Tecnológico, Universidade Federal de Pelotas.

Data da Defesa: 30 de fevereiro de 2019

Banca Examinadora:

Prof. Dr. Marilton Sanchotene de Aguiar (orientador)
Doutor em Computação pela Universidade Federal do Rio Grande do Sul.

Prof. Dr. Paulo Roberto Ferreira Jr.
Doutor em Computação pela Universidade Federal do Rio Grande do Sul.

Prof. Dr. Ricardo Matsumura Araujo
Doutor em Computação pela Universidade Federal do Rio Grande do Sul.

Prof. Dr. Luciano da Silva Pinto
Doutor em Biotecnologia pela Universidade Federal de Pelotas.

Dedico...

AGRADECIMENTOS

Agradeço...

Só sei que nada sei.

— SÓCRATES

RESUMO

AGUIAR, Marilton Sanchotene de. **LLM-Powered Applications: Tecnologia, Questões e estudo de caso..** Orientador: Marilton Sanchotene de Aguiar. 2024. 28 f. Trabalho de Conclusão de Curso (Ciência da Computação) – Centro de Desenvolvimento Tecnológico, Universidade Federal de Pelotas, Pelotas, 2024.

Bla blabla blablabla bla. Bla blabla blablabla bla. Bla blabla blablabla bla. Bla
blabla blablabla bla. Bla blabla blablabla bla. Bla blabla blablabla bla. Bla blabla
blablabla bla. Bla blabla blablabla bla. Bla blabla blablabla bla. Bla blabla blablabla bla.
Bla blabla blablabla bla. Bla blabla blablabla bla. Bla blabla blablabla bla. Bla blabla
blablabla bla. Bla blabla blablabla bla. Bla blabla blablabla bla. Bla blabla blablabla bla.
Bla blabla blablabla bla. Bla blabla blablabla bla. Bla blabla blablabla bla. Bla blabla
blablabla bla.

Palavras-chave: palavrachave-um; palavrachave-dois; palavrachave-tres; palavrachave-quatro.

RESUMO

AGUIAR, Marilton Sanchotene de. **Titulo do Trabalho em Ingles**. Orientador: Marilton Sanchotene de Aguiar. 2024. 28 f. Trabalho de Conclusão de Curso (Ciência da Computação) – Centro de Desenvolvimento Tecnológico, Universidade Federal de Pelotas, Pelotas, 2024.

Bla blabla blablalba bla. Bla blabla blablalba bla. Bla blabla blablalba bla. Bla
blabla blablalba bla. Bla blabla blablalba bla. Bla blabla blablalba bla. Bla blabla
blablalba bla. Bla blabla blablalba bla. Bla blabla blablalba bla. Bla blabla blablalba bla.
Bla blabla blablalba bla. Bla blabla blablalba bla. Bla blabla blablalba bla. Bla blabla
blablalba bla. Bla blabla blablalba bla. Bla blabla blablalba bla. Bla blabla blablalba bla.
Bla blabla blablalba bla. Bla blabla blablalba bla. Bla blabla blablalba bla. Bla blabla
blablalba bla.

Palavras-chave: keyword-one; keyword-two; keyword-three; keyword-four.

LISTA DE FIGURAS

Figura 1	Nome da figura	16
----------	--------------------------	----

LISTA DE TABELAS

Tabela 1	Nome da Tabela	15
Tabela 2	Nome da Tabela	17

LISTA DE ABREVIATURAS E SIGLAS

ABNT	Associação Brasileira de Normas Técnicas
NUMA	Non-Uniform Memory Access
SIMD	Single Instruction Multiple Data
SMP	Symmetric Multi-Processor
SPMD	Single Program Multiple Data

SUMÁRIO

1	INTRODUÇÃO	14
1.1	Grandes Modelos de Linguagem	14
1.2	Outra seção	14
1.2.1	Uma subseção	15
2	LLM POWERED APPLICATIONS	16
2.1	Fine tuning	17
2.2	Retrieval Augmented Generation	17
2.3	Engenharia de prompt	17
2.4	Function Calling e Output Parsers	19
2.5	Agentes	19
3	DESENVOLVIMENTO DE UMA LLM POWERED APPLICATION PARA FACILITAR O CONSUMO E APRENDIZAGEM ATRAVÉS DE VÍDEOS	20
3.1	Motivação	20
3.2	Features	21
3.2.1	Improved Readability	21
3.2.2	Transcrição/Tradução com Whisper	21
3.2.3	Auto Chapter	21
3.2.4	Geração de perguntas por capítulo	21
4	CONCLUSÃO	22
	REFERÊNCIAS	23
	APÊNDICE A UM APÊNDICE	25
	ANEXO A UM ANEXO	27
	ANEXO B OUTRO ANEXO	28

1 INTRODUÇÃO

1.1 Grandes Modelos de Linguagem

Bla blabla blablabla bla. Bla blabla blablabla bla. Bla blabla blablabla bla. Bla blabla
blablabla bla. Bla blabla blablabla bla. Bla blabla blablabla bla. Bla blabla blablabla bla.
Bla blabla blablabla bla. Bla blabla blablabla bla. Bla blabla blablabla bla. Bla blabla
blablabla bla.

Bla blabla blablabla bla. Bla blabla blablabla bla. Bla blabla blablabla
bla. Bla blabla blablabla bla. Bla blabla blablabla bla. Bla blabla
blablabla bla. Bla blabla blablabla bla. Bla blabla blablabla bla. Bla
blabla blablabla bla. Bla blabla blablabla bla Moore (1979); Aguiar; Mar-
ilton (2005).

Bla blabla blablabla bla. Bla blabla blablabla bla. Bla blabla blablabla bla. Bla blabla
blablabla bla. Bla blabla blablabla bla. Bla blabla blablabla bla. Bla blabla blablabla bla.
Bla blabla blablabla bla. Bla blabla blablabla bla. Bla blabla blablabla bla. Bla blabla
blablabla bla. Bla blabla blablabla bla. Bla blabla blablabla bla. Bla blabla blablabla bla.
Bla blabla blablabla bla. Bla blabla blablabla bla. Bla blabla blablabla bla. Bla blabla
blablabla bla. Bla blabla blablabla bla. Bla blabla blablabla bla. Bla blabla blablabla bla.

Bla blabla blablabla bla. Bla blabla blablabla bla. Bla blabla blablabla bla. Bla blabla
blablabla bla. Bla blabla blablabla bla. Bla blabla blablabla bla. Bla blabla blablabla bla.
Bla blabla blablabla bla. Bla blabla blablabla bla. Bla blabla blablabla bla. Bla blabla
blablabla bla. Bla blabla blablabla bla. Bla blabla blablabla bla. Bla blabla blablabla bla.
Bla blabla blablabla bla. Bla blabla blablabla bla. Bla blabla blablabla bla. Bla blabla
blablabla bla. Bla blabla blablabla bla. Bla blabla blablabla bla. Bla blabla blablabla
bla (Neumann; Aguiar, 1966).

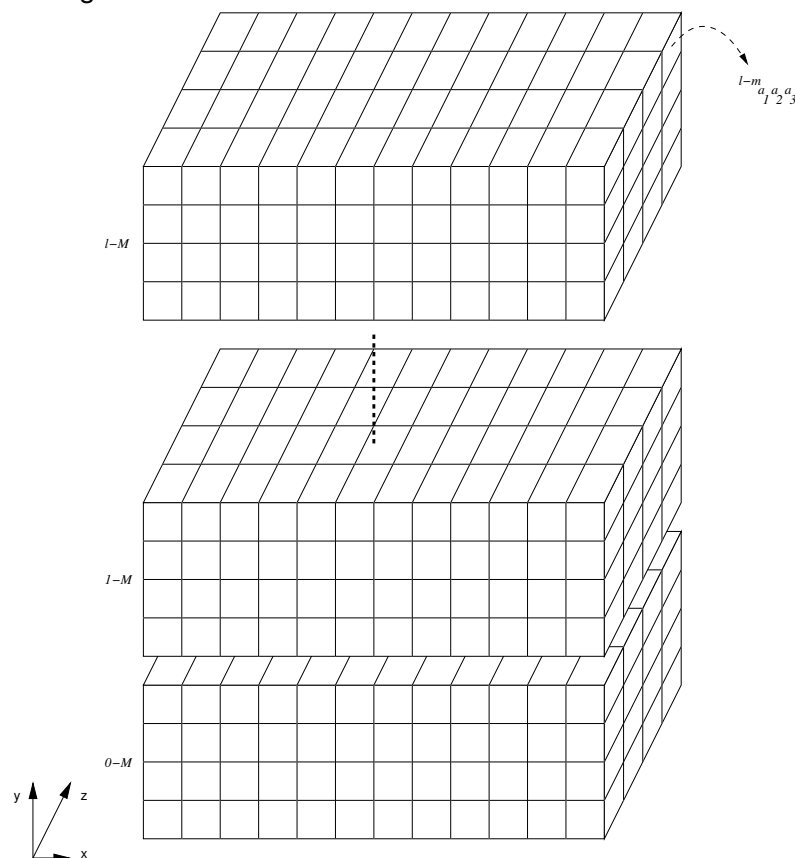
1.2 Outra seção

Bla blabla blablalba bla. Bla blabla blablalba bla. Bla blabla blablalba bla. Bla blabla
blablalba bla. Bla blabla blablalba bla. Bla blabla blablalba bla. Bla blabla blablalba bla.
Bla blabla blablalba bla. Bla blabla blablalba bla. Bla blabla blablalba bla. Bla blabla

2 LLM POWERED APPLICATIONS

Bla blabla blablabla bla. Bla blabla blablabla bla. Bla blabla blablabla bla. Bla blabla
blablabla bla. Bla blabla blablabla bla. Bla blabla blablabla bla. Bla blabla blablabla bla.
Bla blabla blablabla bla. Bla blabla blablabla bla. Bla blabla blablabla bla. Bla blabla
blablabla bla. Bla blabla blablabla bla. Bla blabla blablabla bla. Bla blabla blablabla bla.
Bla blabla blablabla bla. Bla blabla blablabla bla. Bla blabla blablabla bla. Bla blabla
blablabla bla. Bla blabla blablabla bla. Bla blabla blablabla bla. Bla blabla blablabla
bla 2.

Figura 1 – Nome da figura



Fonte: Elaborada pelo autor.

Tabela 2 – Nome da Tabela

Blabla	Blabla	Blablabla
Bla	Blabla	<i>Bla blabla blablabla blabla blablabla blabla blablabla.</i>
Bla	Blabla	<i>Bla blabla blablabla blabla blablabla blabla blablabla.</i>
Bla	Blabla	<i>Bla blabla blablabla blabla blablabla blabla blablabla.</i>
Bla	Blabla	<i>Bla blabla blablabla blabla blablabla blabla blablabla.</i>
Bla	Blabla	<i>Bla blabla blablabla blabla blablabla blabla blablabla.</i>
Bla	Blabla	<i>Bla blabla blablabla blabla blablabla blabla blablabla. Conforme a figura 1</i>

- 2.1 Fine tuning
- 2.2 Retrieval Augmented Generation
- 2.3 Engenharia de prompt

A engenharia de prompt é uma disciplina emergente que desempenha um papel fundamental na utilização e eficiência de modelos de linguagem grandes (LLMs), especialmente dentro do domínio das tarefas de processamento de linguagem natural. A medida que a IA continua a permear diversos setores, a habilidade de comunicar efetivamente com esses sistemas torna-se crucial. Este capítulo tem como objetivo dissecar o conceito de engenharia de prompt, suas técnicas, aplicações e o profundo impacto que ela tem em diferentes campos.

Técnicas de Engenharia de Prompt

Essencialmente, a engenharia de prompt é a prática de formular e refinar prompts para maximizar o desempenho de modelos de linguagem. Um prompt bem elaborado pode alterar significativamente a saída de uma IA, fazendo a diferença entre uma resposta útil e uma irrelevante.

A engenharia de prompt eficaz depende de vários princípios centrais que garantem clareza, especificidade e relevância na interação com sistemas de IA. Esses princípios incluem escrever instruções claras e descritivas, usar delimitadores, fornecer exemplos, atribuir papéis, adicionar informações de contexto, dividir tarefas complexas e solicitar múltiplas soluções. Técnicas avançadas também envolvem a sinergia de raciocínio e ação em modelos de linguagem, como modelos auxiliados por programas, raciocínio automático e uso de ferramentas, e ajuste de prompt para adaptar as

respostas do modelo a necessidades específicas (DaveAI).

Técnicas Avançadas de Prompt

As técnicas avançadas de prompt englobam uma variedade de estratégias projetadas para obter respostas mais sofisticadas de LLMs. Estas incluem prompt de encadeamento de pensamento, que encoraja os modelos a exibir seu processo de raciocínio, e prompt de menor para maior, que guia os modelos por meio de uma série progressiva de complexidade em suas respostas. Além disso, o prompt de geração de conhecimento e o uso de modelos de árvore de pensamento representam abordagens inovadoras para a resolução de problemas com IA (Arxiv).

Aplicações da Engenharia de Prompt

As aplicações da engenharia de prompt são vastas e variadas. Na escrita acadêmica, os pesquisadores podem aproveitar a engenharia de prompt para simplificar revisões de literatura, sintetizar informações complexas e até mesmo gerar rascunhos de artigos. Além da academia, a engenharia de prompt é instrumental em áreas como atendimento ao cliente, onde a IA pode fornecer respostas personalizadas a perguntas, e em indústrias criativas, onde a IA pode auxiliar na criação de conteúdo que varia da escrita à geração de imagens (Intellyverse).

Impacto da Engenharia de Prompt

O impacto da engenharia de prompt é substancial, pois influencia diretamente a eficácia das interações de IA. Ao otimizar prompts, os usuários podem obter saídas mais precisas e contextualmente relevantes dos sistemas de IA, aumentando assim a produtividade e reduzindo o potencial de mal-entendidos. O campo também levanta importantes considerações sobre o uso ético da IA, já que os prompts em si podem moldar a natureza das informações fornecidas pela IA (Giray).

Conclusão

A engenharia de prompt não é apenas uma habilidade técnica, mas uma forma de arte que requer um profundo entendimento tanto da linguagem quanto da tecnologia. À medida que a IA se torna cada vez mais sofisticada, o papel do engenheiro de prompt se tornará ainda mais crítico na moldagem das interações entre humanos e máquinas. A disciplina está no cruzamento da comunicação, tecnologia e criatividade, oferecendo uma nova fronteira para exploração e inovação na era da IA.

Lista de Referências

- Giray, Louie. "Engenharia de Prompt com ChatGPT: Um Guia para Escritores Acadêmicos." *Springer Nature*, 2023, <https://link.springer.com/article/10.1007/s10439-023-03272-4>.
- "Como se Tornar um Engenheiro de Prompt." *DataCamp*, <https://www.datacamp.com/blog/how-to-become-a-prompt-engineer>.
- "Guia Completo de Engenharia de Prompt." *DaveAI*, <https://daveai.substack.com/p/prompt-engineering-full-guide>.
- "Sinergizando Raciocínio e Ação em Modelos de Linguagem." *Arxiv*, <https://arxiv.org/abs/2210.03629>.
- "Engenharia de Prompt." *Intellyverse*, <https://intellyverse.com/blog/prompt-engineering>.

2.4 Function Calling e Output Parsers

2.5 Agentes

3 DESENVOLVIMENTO DE UMA LLM POWERED APPLICATION PARA FACILITAR O CONSUMO E APRENDIZAGEM ATRAVÉS DE VÍDEOS

3.1 Motivação

A aprendizagem ativa tem sido objeto de vários artigos acadêmicos importantes, destacando seu impacto positivo nos resultados dos alunos. A pesquisa de Freeman et al. (2014) demonstrou que a aprendizagem ativa pode aumentar significativamente as notas dos alunos em relação aos métodos didáticos, com alunos em cursos sem aprendizagem ativa sendo 1,5 vezes mais propensos a reprovar do que aqueles com aprendizagem ativa. Além disso, a aprendizagem ativa tem sido mostrada positiva para melhorar a motivação dos alunos, habilidades de pensamento crítico, retenção de informações e habilidades interpessoais. Por outro lado, a aprendizagem passiva, como assistir a vídeos, tem sido associada a desvantagens, como níveis mais baixos de engajamento e compreensão superficial de conceitos-chave.

Além disso, dado o contexto de excesso de dados e nossa crescente dificuldade em navegar em meio a um mar de informações, cada vez mais se faz necessário tecnologias que possam ajudar com esses problemas contemporâneos.

É importante ressaltar também que muitas vezes o conteúdo de interesse está disponível em outra língua, dificultando a aprendizagem. Apesar do Youtube possibilitar a geração de legendas automáticas, a legenda gerada por esse método não possui tanta qualidade quanto por exemplo com os modelos mais avançados de Reconhecimento de Voz Automático (ASR).

A criação de uma aplicação potencializada por LLMs e ASR, como o Whisper, tem o poder de transformar vídeos educacionais em experiências de aprendizagem mais ricas e interativas. Ao incorporar funcionalidades como a geração automática de capítulos, facilitação de perguntas e respostas por capítulos, e o uso de Retrieval Augmented Generation (RAG) para melhoria do QA em vídeos, esta aplicação não apenas melhora a acessibilidade e a personalização do conteúdo, mas também promove a aprendizagem ativa e engajada. Ademais, a transcrição e geração automática

de legendas tornam o conteúdo acessível para um público mais amplo, independentemente do idioma nativo do espectador, ampliando assim o alcance e a eficácia da educação por meio de vídeos.

3.2 Features

3.2.1 Improved Readability

3.2.2 Transcrição/Tradução com Whisper

3.2.3 Auto Chapter

3.2.4 Geração de perguntas por capítulo

4 CONCLUSÃO

REFERÊNCIAS

AGUIAR, M.; MARILTON, A. **Título da Monografia**. 2005. 85p. Trabalho de Conclusão (Curso de Ciência da Computação) — Centro de Desenvolvimento Tecnológico, Universidade Federal de Pelotas, Pelotas.

MOORE, R. E. **Methods and Applications of Interval Analysis**. Philadelphia, PA, USA: Society for Industrial and Applied Mathematics, 1979. xi + 190p.

NEUMANN, J. von; AGUIAR, M. **Theory of Self-Reproducing Automata**. [S.l.: s.n.], 1966. xix + 388p.

Apêndices

APÊNDICE A – Um Apêndice

Anexos

