ETAPA 4

Prompts avançados

Resumir textos utilizando prompts

O **resumo de textos** é uma das aplicações mais **populares** e **úteis** de LLMs, pois permite extrair **informações principais** de grandes quantidades de conteúdo de forma eficiente. Vamos classificar os tipos de atividades de resumo em:

- Resumo Extrativo
- Resumo Abstrativo
- Resumo Híbrido

Assistentes de Atendimento ao Cliente: resumo de interações passadas para contexto antes do atendimento.

Análise de Notícias e Artigos: Sumarizar notícias e artigos para leitores, oferecendo uma visão geral das informações mais importantes.

Pesquisa Acadêmica e Científica: resumo de artigos e papers científicos para pesquisadores, permitindo uma rápida revisão de literatura.

Relatórios Executivos e Tomada de Decisão: geração de relatórios executivos que sintetizam informações de relatórios longos, facilitando a tomada de decisões.

Resumir textos utilizando prompts

No **resumo extrativo**, o modelo seleciona **frases ou trechos importantes** diretamente do texto original, mantendo a **estrutura e as palavras originais**. Esse tipo de resumo é útil em situações que exigem fidelidade ao texto fonte, como notícias ou relatórios financeiros.

No **resumo abstrativo**, o modelo **reformula o conteúdo**, criando novas frases que sintetizam as **ideias do texto original** de forma mais concisa e criativa. Esse **método** é mais **flexível**, porém **depende** da capacidade do **modelo** de interpretar e reescrever a informação.

Extrativo	Abstrativo
“Resuma este texto extraindo as frases mais importantes.”	“Reescreva as principais ideias deste texto em um resumo conciso de até 3 frases.”
Leia o texto a seguir e extraia apenas as frases mais relevantes que sintetizam as principais informações. Mantenha a linguagem original e destaque as ideias centrais.	“Resuma o texto a seguir usando uma combinação de frases extraídas do texto original e novas formulações. Destaque as informações principais de forma concisa.”

Híbrido

“Resuma o texto a seguir usando uma combinação de frases extraídas do texto original e novas formulações. Destaque as informações principais de forma concisa.”

Resumir textos utilizando prompts

Manutenção do Contexto
LLMs podem perder o fio de pensamento ao longo de textos extensos, especialmente se houver múltiplos tópicos ou mudanças de contexto.

Limitação de Tokens
Muitos LLMs possuem um limite de tokens por entrada e saída, o que restringe a quantidade de texto que pode ser resumida em uma única iteração.

Coerência e Coesão
A divisão de textos em chunks pode gerar resumos que não são coesos ou que apresentam quebra na fluidez entre as partes resumidas.

Risco de Omissão de Informações Importantes
LLMs podem deixar de lado detalhes essenciais, especialmente quando orientados a fazer resumos mais concisos.

Tendência de Alucinação
Modelos podem gerar informações irrelevantes ou incorretas, criando conteúdo que não estava presente no texto original.

Dificuldade de Balancear Generalidade e Especificidade
Encontrar um equilíbrio entre ser conciso e manter detalhes importantes é um desafio, especialmente em textos complexos ou técnicos.

Estratégias de Prompts para Resumo de Textos Longos

Para **resumir textos longos** com LLMs, é necessário **dividir o conteúdo** em partes menores, chamadas de **chunks**. Esse processo facilita a manutenção de **contexto e precisão** durante o resumo.

Resumo Sequencial: dividir o texto em seções sequenciais e aplicar o mesmo prompt a cada uma. Apesar de manter a coerência dentro de cada chunk, pode perder o contexto geral ao consolidar as partes resumidas.

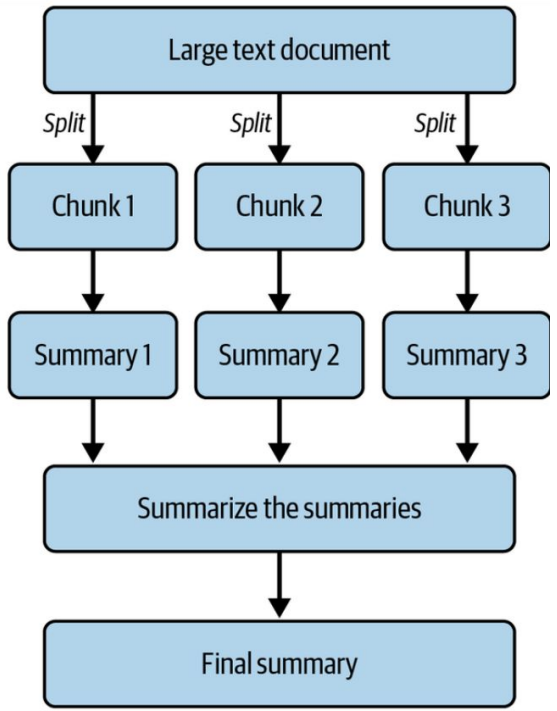

Figure 3-4. A summarization pipeline that uses text splitting and multiple summarization steps

Estratégias de Prompts para Resumo de Textos Longos

Formatos de Chunking para Resumo de Textos Longos
Existem diversas estratégias de **chunking** de textos, que aplicam cortes no texto segundo diferentes estruturas.

Divisão por Parágrafos: cada parágrafo é tratado como um chunk independente, o que facilita a manutenção de coesão dentro de tópicos, porém pode haver perda de continuidade entre parágrafos.

Divisão por Tópicos ou Seções: separação baseada na estrutura do documento, como introdução, metodologia, etc. Mantém o contexto específico de cada seção,porém pode deixar lacunas de informação entre as seções.

Table 3-1. Six chunking strategies highlighting their advantages and disadvantages

Splitting strategy	Advantages	Disadvantages
Splitting by sentence	Preserves context, suitable for various tasks	May not be efficient for very long content
Splitting by paragraph	Handles longer content, focuses on cohesive units	Less granularity, may miss subtle connections
Splitting by topic	Identifies main themes, better for classification	Requires topic identification, may miss fine details
Splitting by complexity	Groups similar complexity levels, adaptive	Requires complexity measurement, not suitable for all tasks
Splitting by length	Manages very long content, efficient processing	Loss of context, may require more preprocessing steps
Using a tokenizer: Splitting by tokens	Accurate token counts, which helps in avoiding LLM prompt token limits	Requires tokenization, may increase computational complexity

By choosing the appropriate chunking strategy for your specific use case, you can optimize the performance and accuracy of AI language models.

Estratégias de Prompts para Resumo de Textos Longos

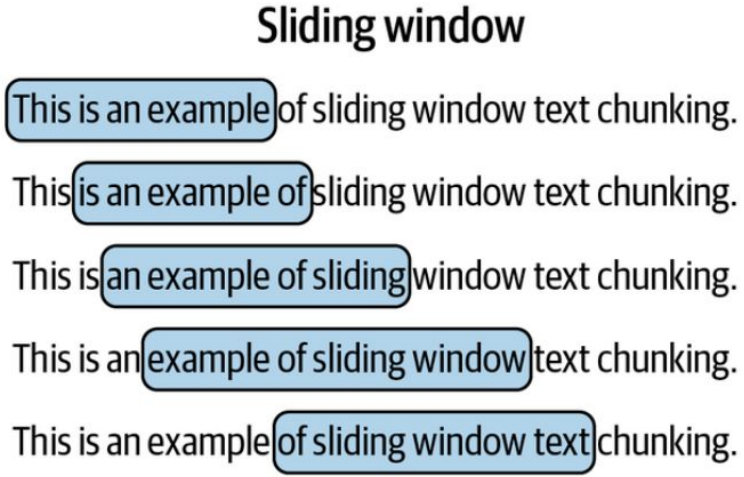

Figure 3-6. A sliding window, with a window size of 4 and a step size of 1

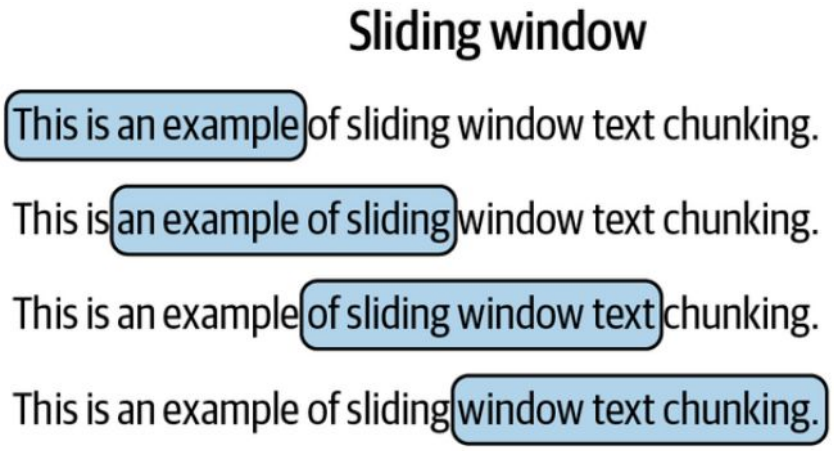

Figure 3-7. A sliding window, with a window size of 4 and a step size of 2

EXERCÍCIOS

Tarefa 1: sumarização de episódios
Utilizar técnicas de prompt engineering para sumarizar um episódio do programa.

Tarefa 2: sumarização da temporada
Quebrar a temporada em chunks de episódios para sumarização dividida.

Tarefa 3: listar a interação com a família
Para cada temporada, listar quais foram os personagens que mais interagiram com cada membro da família.

Quantidade de tokens necessários para processar textos longos

O conceito de **token** é essencial para o **processamento de textos longos**. Quanto maior o número de tokens, mais **complexa e custosa** se torna a tarefa de processamento.

Limites de Tokens afetam diretamente o desempenho e o custo dos LLMs. Quando o texto excede esse limite, ele precisa ser **dividido** em **chunks menores**, o que pode **afetar a fluidez** e o **contexto** do texto.

Além disso, como muitos **LLMs** têm **custos** atrelados ao uso de **tokens**, é importante **otimizar prompts** e textos para **minimizar o consumo** e garantir que o modelo entregue respostas completas.

LLM	TOKEN LIMIT
GPT-4 Turbo	128,000
GPT-4	8,192
GPT Vision	128,000
GPT 3.5 Turbo	16,385
Llama3	8,000
Claude3	200, 000
Gemini 1.5	1 M
Mistral	32,000
AWS Titan	8,000
Cohere Command R	128,000

https://www.edenai.co/post/understanding-llm-billing-from-characters-to-tokens

Quantidade de tokens necessários para processar textos longos

Gerenciamento do consumo de tokens em LLMs Fundamental prever o número de tokens que um texto consumirá para estimativa de custos e otimização de prompts. O BPE (Byte Pair Encoding) ajuda a minimizar o número total de tokens, dividindo palavras em partes reutilizáveis, amplamente usado em todos os LLMs.

A biblioteca **tiktoken** oferece métodos para calcular o número de tokens com base no texto e no modelo utilizado, simulando o processo de tokenização.

```python
import tiktoken
enc = tiktoken.get_encoding("o200k_base")
assert enc.decode(enc.encode("hello world")) == "hello world"

# To get the tokeniser corresponding to a specific model in the OpenAI API:
enc = tiktoken.encoding_for_model("gpt-4o")
```

Encoding name	OpenAI models
cl100k_base	GPT-4, GPT-3.5-turbo, text-embedding-ada-002
p50k_base	Codex models, text-davinci-002, text-davinci-003
r50k_base (or gpt2)	GPT-3 models like davinci

Role prompting para otimizar respostas de LLMs

Role Prompting é uma técnica de prompting que orienta o LLM a **assumir um papel específico** ao responder. Ao definir um "papel" para o modelo, conseguimos guiar o tom e o conteúdo da resposta, melhorando a **coerência** e **adequação ao contexto**.

Benefícios: útil em aplicações que requerem respostas com um estilo específico ou que precisam incorporar conhecimentos especializados. Role prompting ajuda o LLM a adaptar seu vocabulário.

"Finja ser um historiador e explique a importância da Revolução Industrial."

"Seja um especialista em marketing e dê sugestões para aumentar o engajamento do público nas redes sociais."

Quando usar Role Prompting?

Criatividade

Estimular respostas criativas: Role Prompting pode ser utilizado para criar cenários fictícios ou gerar respostas imaginativas ao atribuir papéis como contador de histórias, personagens de um romance ou figuras históricas.

Especialização

Elicitar conhecimento especializado: quando a resposta exige conhecimento específico de um domínio ou área, o Role Prompting ajuda o LLM a gerar respostas mais precisas e informadas.

Personalização

Personalizar o estilo de resposta: atribuir um papel pode orientar o LLM a gerar respostas com um tom, estilo ou perspectiva específicos, como uma resposta formal, casual ou humorística.

Engajamento

Aumentar o engajamento do usuário: Role Prompting torna as interações mais envolventes e divertidas ao permitir que o LLM assuma personagens ou papéis que se conectem com o usuário.

Perspectivas

Explorar perspectivas diversas: Para explorar diferentes pontos de vista sobre um tema, o Role Prompting pode ajudar ao solicitar que o LLM assuma várias perspectivas ou personagens, proporcionando uma visão mais completa do assunto.

Role prompting para otimizar respostas de LLMs

Desafios: **risco** de introduzir **viés** ou **estereótipos** com base no papel atribuído, o que pode levar a **respostas enviesadas**, prejudicando a usabilidade e, em alguns casos, podendo ofender os usuários.

Risco de **perda de consistência** no papel após **longas interações** com o usuário, bem como o **risco de prompt injection**, alterando o Role do sistema.

Boas práticas em Role Prompting

Direcionamento Claro: Defina o papel de forma explícita, incluindo instruções sobre a perspectiva que o LLM deve assumir (ex.: "Responda como um especialista financeiro").

Definição de Estilo e Formato: Especifique o tom e o estilo desejados, como “formal”, “informativo” ou “conversacional”. Isso ajuda o modelo a ajustar o estilo de escrita.

Uso de Exemplos: Apresente exemplos de respostas esperadas para guiar o modelo. Isso aumenta a coerência e a qualidade das respostas ao alinhar o output ao estilo desejado.

EXERCÍCIOS

Tarefa 1: Role prompting para agentes personalizados
Criar Agentes de IA que possam interpretar os personagens na interação com o usuário.

Tarefa 2: Recriação de episódios
Simular os editores dos Simpsons, criando novos diálogos e fim para os episódios.

Tarefa 3: Assistente de conteúdo
Interface para QA com um assistente especialista em Simpsons.