**Infnet – Ciência de Dados – 2024.2**

Engenharia de Prompts para Ciência de Dados

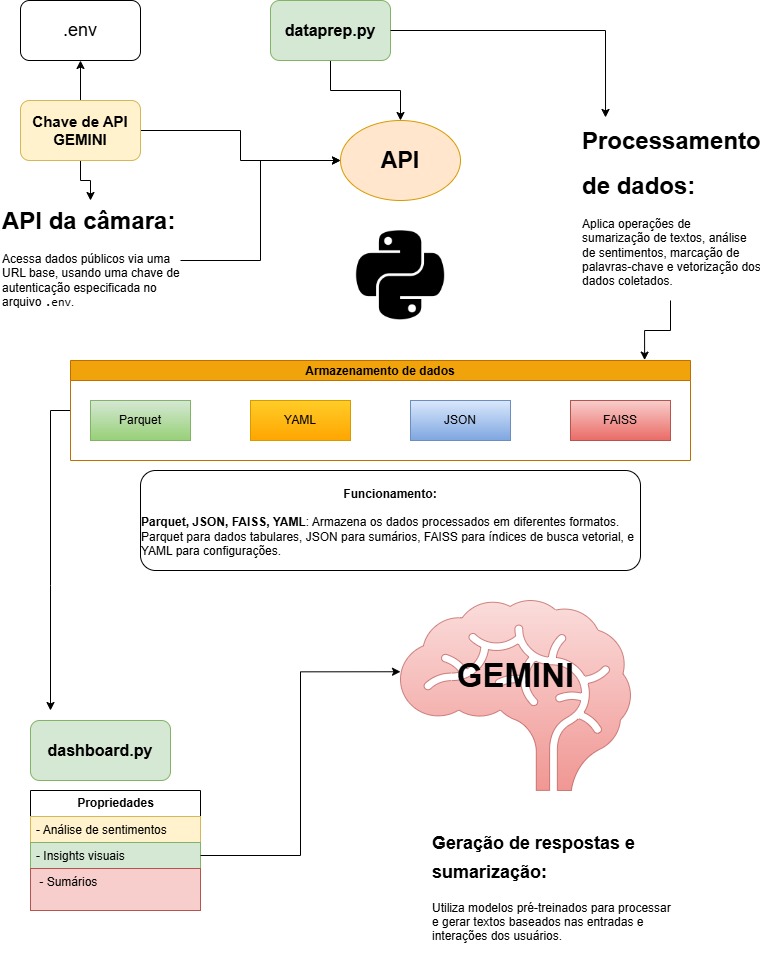
AT – Questões Escritas

Prof.: Thiago Ciodaro Xavier

Aluno: Daniel da Cunha Moreira

**Questão 1:**

Letra a)



Letra b)

**Arquitetura do Projeto de Processamento de Dados da Câmara dos Deputados**

1. **Coleta de Dados (dataprep.py)**

* **API da Câmara dos Deputados**: Utilizada para extrair dados estruturados sobre atividades legislativas, como proposições e despesas dos deputados.
* **Arquivo .env para chave Gemini**: Armazena chaves de configuração e API de forma segura, evitando exposição no código fonte.
* **Processadores de dados**:
  + **Sumarização**: Utiliza LLMs para condensar informações extensas em resumos concisos.
  + **Análise de sentimentos**: Avalia o tom emocional (positivo, negativo, neutro) das declarações ou documentos.
  + **Marcação de palavras-chave**: Identifica e marca termos e conceitos relevantes nos textos.
  + **Vetorização para bases de referência**: Transforma textos em vetores numéricos que representam semanticamente o conteúdo para facilitar comparações e recuperações rápidas.

**2. Armazenamento de Dados**

* **Parquet para dados tabulares**: Formato eficiente para armazenamento e consulta de dados estruturados em colunas.
* **JSON para sumários e metadados**: Utilizado para armazenar resumos e informações adicionais em formato de fácil manipulação e acesso.
* **FAISS Index para buscas vetoriais**: Permite a realização de buscas de alta performance em grandes volumes de dados vetorizados.
* **YAML para configurações**: Armazena configurações de forma estruturada, facilitando a gestão e interpretação por humanos e máquinas.

**3. Dashboard (dashboard.py)**

* **Abas de Visualização**: Oferece interfaces separadas para visualizar resumos, análises de sentimentos e resultados de pesquisas vetoriais.
* **Funcionalidades interativas e uso de LLMs**: Permite que os usuários interajam com os dados e solicitem processamentos ad hoc, como sumarizações e análises, utilizando LLMs integrados.

**4. Modelos de Linguagem de Larga Escala (LLMs)**

* **Sumarização e geração de respostas**: Implementados para processar e responder a consultas complexas, apoiando decisões e análises legislativas. Utilizam a arquitetura de **Transformers**, que são modelos de atenção que aprendem contextos em larga escala de dados textuais. Os Transformers são fundamentais para entender a estrutura linguística e gerar respostas contextuais precisas.

Letra c)

**Funcionamento de LLMs e Aplicações em Sumarização**

**LLMs (Large Language Models)**, como os baseados na arquitetura Transformer, são modelos de inteligência artificial treinados para entender e gerar linguagem humana. Eles operam através de um mecanismo de atenção que permite ao modelo considerar o contexto de cada palavra no texto, independentemente da distância entre as palavras no texto. Isso é crucial para entender a estrutura gramatical e o significado semântico.

Na sumarização, os LLMs podem ser empregados de duas formas principais:

1. **Sumarização Extrativa:** Identifica as partes mais importantes do texto (frases ou segmentos) e as compila para formar um resumo. Esses modelos dependem fortemente da capacidade do modelo de avaliar a relevância e a representatividade de segmentos de texto dentro de um documento maior.
2. **Sumarização Abstrativa:** Gera um novo texto que captura as ideias principais do conteúdo original, frequentemente reescrevendo ou parafraseando para produzir um resumo mais conciso e coeso. Isso envolve um entendimento mais profundo e a capacidade de reformular informações, algo para o qual os LLMs são particularmente bem adaptados devido à sua capacidade de geração de linguagem.

**Questão 2:**

**Prompt Utilizado em todos os modelos:**

“Gere um texto curto de dois parágrafos que explique a câmara dos deputados.”

**Respostas dos 3 LLMs utilizados:**

**Claude-3.5-Sonnet**

A Câmara dos Deputados é uma das duas casas do Congresso Nacional brasileiro, sendo composta por 513 deputados federais eleitos pelo povo para mandatos de quatro anos. Estes parlamentares representam os diferentes estados brasileiros e o Distrito Federal, com o número de representantes de cada unidade federativa sendo proporcional à sua população.

Como parte do Poder Legislativo, a Câmara tem como principais atribuições criar e aprovar leis, fiscalizar o Poder Executivo e debater questões de interesse nacional. Os deputados também são responsáveis por analisar o orçamento da União, propor emendas à Constituição e julgar crimes de responsabilidade cometidos por autoridades, como o Presidente da República, em conjunto com o Senado Federal.

**Gemini-1.5-FLash**

A Câmara dos Deputados é uma das duas casas do Congresso Nacional brasileiro, sendo a outra o Senado Federal.  Ela é composta por deputados federais, eleitos por representação proporcional em cada estado, com o número de deputados por estado sendo proporcional à sua população.  A principal função da Câmara é legislar, ou seja, criar leis federais que regem o país.  Além disso, a Câmara também tem o poder de aprovar o orçamento anual do governo, fiscalizar as ações do Poder Executivo e julgar o Presidente da República em casos de crime de responsabilidade.

A representatividade da Câmara dos Deputados garante que diferentes regiões e grupos populacionais tenham voz no processo legislativo.  Os deputados são eleitos para mandatos de quatro anos, podendo ser reeleitos.  O funcionamento da Câmara é regido por um Regimento Interno, que define as normas de funcionamento e procedimentos para votações e debates.  A Câmara desempenha um papel fundamental na democracia brasileira, sendo um espaço de debate e construção de políticas públicas.

**GPT-4o-Mini**

A Câmara dos Deputados é uma das duas casas do Congresso Nacional do Brasil, composta por 513 deputados federais eleitos pelo povo para mandatos de quatro anos. Os deputados representam diferentes estados e o Distrito Federal, com o número de representantes de cada unidade federativa sendo proporcional à sua população. A principal função da Câmara é elaborar e aprovar leis que regem o país, além de fiscalizar as ações do Poder Executivo e debater questões de interesse nacional.

Além de legislar, a Câmara dos Deputados também é responsável por analisar o orçamento da União, propor emendas à Constituição e julgar crimes de responsabilidade cometidos por autoridades, como o Presidente da República. O funcionamento da Câmara é regido por um Regimento Interno, que define normas e procedimentos para votações e debates, garantindo que a diversidade de vozes e interesses da população seja representada no processo legislativo.

Analisando as respostas dos três LLMs:

**Claude-3.5-Sonnet**

**Vantagens:**

* A resposta é clara, concisa e traz um entendimento fácil.
* Destaca a representação proporcional e as responsabilidades legislativas, trazendo detalhes técnicos.

**Desvantagens:**

* Não traz muita ênfase nos procediemntos internos da câmara.

**Gemini-1.5-Flash**

**Vantagens:**

* Traz uma explicação detalhada das funções legislativas e de fiscalização.
* Menciona o Regimento Interno, dando uma noção de como a Câmara funciona internamente.

**Desvantagens:**

* Resposta muito complexa. Para quem não está familiarizado com termos técnicos, pode ser de dfiicil compreensão.

**GPT-4o-Mini**

**Vantagens:**

* Cobertura abrangente das funções da Câmara, desde legislação até julgamento de crimes de responsabilidade.
* Inclui informações sobre o Regimento Interno e o papel da Câmara na representação democrática.

**Desvantagens:**

* Alguns pontos são repetidos ou semelhantes às outras respostas, redundante.

**Diferenças gerais**

* **Claude-3.5-Sonnet** - Resposta mais focada na descrição e nas funções básicas da Câmara.
* **Gemini-1.5-FLash** - Traz detalhes sobre representatividade e sobre o Regimento Interno, trazendo uma visão mais processual.
* **GPT-4o-Mini** - oferece uma visão compreensiva que abrange desde funções legislativas até detalhes de procedimentos internos, sendo mais detalhista na explicação das funções de julgamento e fiscalização.
* **Claude** e **GPT-4o** são mais diretos e podem ser mais acessíveis para um público geral, enquanto **Gemini** usa um estilo que, embora informativo, pode ser um pouco mais denso e técnico.
* Dependendo do contexto de uso (educacional, informativo, técnico), cada modelo pode ser preferido. Por exemplo, para um contexto educacional, o estilo de **GPT-4o** e **Claude** pode melhor, enquanto **Gemini** seria ideal para contextos onde um entendimento mais técnico é necessário.

**Questão 2:**

**Questão 3:**

Letra b)

**Prompt utilizado, armazenado na variável “prompt\_grafico”:**

"""

Você é um especialista em visualização de dados usando Python. Gere um código que:

1. Leia um arquivo Parquet localizado no diretório "data" com o nome "deputados.parquet". Este arquivo contém informações sobre deputados federais, incluindo a coluna `siglaPartido` que identifica o partido de cada deputado.

2. Agrupe os deputados por partido (`siglaPartido`) e calcule:

   - O número total de deputados em cada partido.

   - O percentual correspondente de deputados de cada partido em relação ao total geral.

3. Crie um gráfico de pizza que mostre:

   - A quantidade percentual de deputados por partido.

   - Um título claro: "Distribuição de Deputados por Partido".

   - Um rótulo com o nome do partido e o percentual correspondente para cada fatia.

4. Use as bibliotecas `pandas` e `matplotlib` para essa tarefa.

5. Garanta que o gráfico tenha:

   - Título centralizado.

   - Grid desativado (pois gráficos de pizza não usam grid).

   - Cores diferenciadas para cada fatia.

6. Salve o gráfico no diretório `docs` com o nome "distribuicao\_deputados.png".

Retorne somente o código Python, sem explicações ou texto adicional.

"""

**Resposta do LLM:**

```python

import pandas as pd

import matplotlib.pyplot as plt

# Certifique-se de que o diretório 'data' exista e contenha o arquivo 'deputados.parquet'

# Certifique-se de que o diretório 'docs' exista

try:

df = pd.read\_parquet("data/deputados.parquet")

partidos = df.groupby('siglaPartido')['siglaPartido'].count()

total\_deputados = partidos.sum()

percentuais = (partidos / total\_deputados) \* 100

plt.figure(figsize=(10, 10))

plt.pie(percentuais, labels=[f'{partido} ({percentual:.1f}%)' for partido, percentual in zip(partidos.index, percentuais)],

autopct='%1.1f%%', startangle=140, textprops={'fontsize': 10})

plt.title("Distribuição de Deputados por Partido", fontsize=16, loc='center')

plt.axis('equal') # Equal aspect ratio ensures that pie is drawn as a circle.

plt.grid(False)

plt.savefig("docs/distribuicao\_deputados.png")

plt.close()

except FileNotFoundError:

print("Arquivo 'deputados.parquet' não encontrado no diretório 'data'.")

except Exception as e:

print(f"Ocorreu um erro: {e}")

**Esta resposta se encontra em um arquivo ipynb, juntamente com outras respostas geradas pelos LLMs utilizados no trabalho (GEMINI) de nome: “AT\_questoes.ipynb”.**

Letra c)

**Prompt utilizado, armazenado na variável “prompt\_insights”**

"""

Você é um analista político especialista em entender a dinâmica da Câmara dos Deputados.

Com base nos seguintes dados da distribuição de deputados por partido:

{dados}

1. Analise a representatividade de cada partido na Câmara dos Deputados.

2. Explique como essa composição pode influenciar as votações e decisões políticas.

3. Considere possíveis coalizões ou divergências baseadas nos números apresentados.

4. Forneça insights relevantes para compreender o impacto político dessa distribuição.

Responda de forma detalhada e didática, apresentando os insights de maneira clara.

Instruções:

- Use uma linguagem profissional e acessível.

- Evite viés político.

- Seja direto e objetivo nas suas análises.

"""

**Resposta do LLM:**

## Análise da Composição Partidária da Câmara dos Deputados

A análise da distribuição partidária na Câmara, apresentada, revela um cenário de poder fragmentado, mas com clara predominância do PL. Vamos detalhar:

\*\*1. Representatividade Partidária:\*\*

O PL, com 93 deputados (17.92%), detém uma significativa vantagem numérica sobre os demais partidos. O PT, com 68 deputados (13.1%), ocupa a segunda posição, marcando uma distância considerável em relação ao líder. UNIÃO (60 deputados, 11.56%), PP (52 deputados, 10.02%), PSD (45 deputados, 8.67%), REPUBLICANOS e MDB (ambos com 44 deputados, 8.48%) formam um segundo grupo de partidos com influência relevante, embora individualmente não possuam o peso do PL ou PT. Os demais partidos possuem representatividade muito menor, atuando como peças importantes em negociações e formação de blocos, mas sem capacidade individual de definir o rumo das votações.

\*\*2. Influência nas Votações e Decisões Políticas:\*\*

A fragmentação partidária, apesar da liderança do PL, torna o processo decisório complexo. O PL, apesar da expressiva vantagem numérica, necessitará de apoio de outros partidos para aprovar projetos relevantes, exigindo negociações e concessões. A capacidade de formar coalizões será crucial para a aprovação de pautas prioritárias do governo e da oposição. A ausência de uma maioria sólida para qualquer partido ou bloco tende a gerar instabilidade política e a depender de acordos pontuais para a aprovação das matérias. Partidos menores podem exercer um papel estratégico, atuando como "pêndulo" em votações acirradas.

\*\*3. Possíveis Coalizões e Divergências:\*\*

Considerando a distribuição, podemos projetar alguns cenários:

\* \*\*Governo:\*\* O PL, provavelmente, buscará formar uma base aliada majoritária, incluindo partidos como UNIÃO, PP, PSD, REPUBLICANOS e MDB. A solidez dessa base dependerá da capacidade de negociação e da oferta de cargos e benesses políticas. Divergências internas dentro dessa coalizão são prováveis, exigindo constantes ajustes e concessões para manter a unidade.

\* \*\*Oposição:\*\* O PT, como maior partido de oposição, terá um papel central na articulação da oposição. Entretanto, a sua capacidade de influenciar o rumo das votações dependerá da sua capacidade de agregar outros partidos de oposição, o que não é uma tarefa trivial, considerando a diversidade ideológica presente. Partidos menores, como PSOL, PSB e PDT, podem desempenhar papel importante na formação de um bloco opositor mais coeso.

\*\*4. Impacto Político:\*\*

A distribuição partidária atual indica um cenário de:

\* \*\*Governabilidade desafiadora:\*\* O governo precisará negociar constantemente, buscando o apoio de diferentes partidos para aprovar suas propostas.

\* \*\*Maior influência de partidos médios:\*\* Partidos com representatividade entre 40 e 60 deputados terão poder de barganha significativo.

\* \*\*Importância da articulação política:\*\* A capacidade de negociação e construção de consensos será fundamental para o sucesso de qualquer agenda política.

\* \*\*Potencial para instabilidade:\*\* A fragmentação partidária aumenta a possibilidade de impasses e crises políticas, dependendo da capacidade de diálogo entre os partidos.

\* \*\*Reforço do papel do Centrão:\*\* Os partidos tradicionalmente considerados "Centrão" (PP, PSD, MDB, REPUBLICANOS, UNIÃO) mantém grande influência, podendo ser decisivos na aprovação ou rejeição de projetos.

Em resumo, a atual composição da Câmara dos Deputados aponta para um cenário político complexo, marcado por uma forte fragmentação partidária que exige habilidade política e capacidade de negociação tanto do governo quanto da oposição para conduzir a agenda legislativa. A análise apresentada é apolítica, focando na dinâmica de poder derivada da distribuição numérica dos partidos. A efetiva dinâmica política dependerá de fatores como a liderança partidária, a conjuntura econômica e social, e a agenda legislativa prioritária.

Insights salvos em data/insights\_distribuicao\_deputados.json.”””

**Esta resposta se encontra em um arquivo ipynb, juntamente com outras respostas geradas pelos LLMs utilizados no trabalho (GEMINI) de nome: “AT\_questoes.ipynb”.**

**Avaliação da Resposta**

**Pontos Positivos:**

* 1. Estrutura: O texto está organizado e o modelo não alucinou. Trouxe tópicos claros que facilitam a leitura.
  2. Detalhamento: A análise traz diferentes cenários e dinâmicas, como coalizões e impacto político, deixando rica a interpretação dos dados.
  3. Contextualização: O texto conecta os números à prática legislativa, como negociações e governabilidade.
  4. Atenção às Dinâmicas Políticas: Os cenários são realistas e coerentes, levando em conta a fragmentação partidária e o poder de barganha dos partidos menores.

**Melhorias a implementar:**

* 1. Dados mais Específicos: Poderia incluir mais detalhes sobre os partidos menores e suas possíveis posições políticas.
  2. Sugestões Práticas: Poderia sugerir estratégias de governabilidade, como fortalecer lideranças ou implementar reformas no sistema político.
  3. Visualizações: Implementação de gráficos ou tabelas para ilustrar a distribuição partidária ajudaria na compreensão visual dos dados.

**Questão 4:**

Letra b)

**Prompt utilizado, armazenado na variável “prompt\_despesas”:**

"""

Você é um especialista em política brasileira e também um mestre em Ciência de dados/Engenharia de prompts/LLMs.

Agora, você irá usar suas habilidades e conhecumentos para analisar dados das despesas dos deputados.

Os dados estão no arquivo 'serie\_despesas\_diárias\_deputados.parquet'.

As colunas do Dataframe são: 'deputado\_id', 'deputado\_nome', 'tipo\_despesa', 'valor\_despesa',

       'data\_despesa'

Com esses dados em mãos, crie um código Python que realize 3 análises simples e informativas, como por exemplo:

1. \*\*Análise de Despesas Totais\*\*: Calcule o valor total das despesas por deputado.

2. \*\*Análise por Tipo de Despesa\*\*: Mostre a distribuição de despesas por tipo, ou seja, quanto foi gasto em cada tipo de despesa.

3. \*\*Análise Temporal\*\*: Analise as despesas ao longo do tempo, identificando picos de gasto ou tendências de variação de despesas.

Certifique-se de que o código seja bem estruturado e fácil de entender. Aplique as boas práticas de Ciência de dados como um profissional versado da área, incluindo:

- Leitura do arquivo Parquet e análise dos dados.

- Utilização de bibliotecas como Pandas para manipulação de dados.

- Visualização dos resultados, se possível, para tornar as análises mais intuitivas.

- Trate possíveis valores ausentes ou inconsistências nos dados durante as análises.

Quando terminar, me forneça o código gerado e explique cada uma das análises realizadas. Caso identifique algo interessante ou relevante nas análises, mencione isso também.

"""

**Resposta do LLM:**

Observações sobre possíveis insights:

* A análise 1 identifica os deputados com maiores despesas, permitindo um foco em investigações posteriores.
* A análise 2 mostra a proporção de gastos em cada tipo de despesa, permitindo avaliar prioridades orçamentárias.
* A análise 3 revela tendências temporais nas despesas, podendo indicar períodos de maior ou menor gasto público, o que pode estar relacionado à época do ano ou a eventos específicos.  É importante analisar o contexto político para melhor interpretar esses picos.  Uma análise mais detalhada poderia incluir uma regressão para tentar explicar as variações.

```

\*\*Explicação das Análises:\*\*

1. \*\*Análise de Despesas Totais:\*\*  Esta análise agrupa os dados pelo nome do deputado e soma o valor das despesas para cada um. O resultado mostra quem gastou mais, permitindo uma comparação direta entre deputados.  O gráfico de barras facilita a visualização dos 10 maiores gastos.

2. \*\*Análise por Tipo de Despesa:\*\* Esta análise agrupa as despesas pelo tipo e soma os valores.  Um gráfico de pizza é usado para visualizar a proporção de cada tipo de despesa em relação ao total, mostrando a distribuição dos recursos.

3. \*\*Análise Temporal:\*\* Esta análise agrupa as despesas por mês e ano, mostrando a evolução dos gastos ao longo do tempo.  Um gráfico de linha é usado para visualizar tendências e picos de gastos.  Note que esta análise é apenas um exemplo; outras granularidades temporais (semanal, trimestral etc.) poderiam ser usadas dependendo da necessidade.  A identificação de picos e tendências necessita de uma análise mais profunda com o auxílio de especialistas em política e ciência política, pois pode estar correlacionada a eventos políticos relevantes.

\*\*Observações Importantes:\*\*

\* A qualidade da análise depende diretamente da qualidade dos dados.  É crucial garantir a limpeza e a consistência dos dados antes de iniciar qualquer análise.

\* As análises apresentadas são simples.  Análises mais complexas poderiam incluir testes estatísticos, modelos de regressão ou técnicas de machine learning para extrair insights mais profundos.

\* O contexto político é fundamental para a interpretação dos resultados.  As despesas precisam ser analisadas à luz de eventos políticos e econômicos relevantes.

**As análises se encontram no arquivo .ipynb de nome “AT\_questoes.ipynb” juntamente com as outras questões enumeradas neste documento, gráficos e métricas.**

**Questão 6:**

**Obs: Na questão 6, o script gerado pelo LLM para a visualização de despesas, gerou uma visualização pobre, ruim e pouco descritiva. Sendo assim, usando como base o código gerado como resposta do LLM, melhorei-o para que a visualização ficasse minimamente aceitável.**

**O script gerado pelo LLM se encontra no arquivo .ipynb na numeração da questão 6.**

**O script melhorado e integrado ao que foi gerado pelo LLM está na função “despesas\_page”, arquivo “dashboard.py”.**

**Questão 7:**

**Obs: Assim como na questão 6, o script gerado pelo LLM não atendeu às expectativas. O arquivo json precisou ser tratado para que as sumarizações pudessem ser exibidas na interface Streamlit. Mesmo especificando isso no prompt executado para o LLM, o mesmo não foi capaz de tratar o arquivo json. Sendo assim, implementei uma função no início do arquivo dashboard.py (load\_and\_process\_json) que tratou o arquivo.**

**Letra f):**

**Comparação entre CoT e BP:**

**Chain of Thoughts**

**Vantagens:**

* **Maior detaljamento:** Mais conveniente para problemas que requerem raciocínio passo a passo ou onde a explicação do processo de pensamento é importante.
* **Transparência:** Mais fácil de entender como o modelo chegou à conclusão ou resposta, sendo útil para debugging e aprendizado.

**Desvantagens:**

* **Verbosidade:** Pode gerar saídas mais longas que o necessário, o que pode não ser ideal para aplicações que exigem concisão.
* **Complexidade de Implementação:** Exige uma modelagem cuidadosa para garantir que todos os passos do raciocínio sejam relevantes e corretos.

**Batch Prompting**

**Vantagens:**

* **Eficiência:** Processa múltiplas entradas ou tarefas de uma vez, ideal para otimizar tempo e recursos em aplicações em escala.
* **Flexibilidade:** Pode ser facilmente adaptado para diferentes tipos de tarefas ou dados, tornando-o adequado para uso em diversas aplicações.

**Desvantagens:**

* **Menor Controle sobre o Processo Individual:** Como processa várias entradas simultaneamente, pode ser mais difícil entender o raciocínio individual por trás de cada saída.
* **Risco de Erros Escalados:** Um erro no prompt inicial pode afetar todas as respostas geradas, potencializando o impacto de pequenos erros.