Problema

Qual modelo escolher?

Trade-off principal:

- Velocidade **vs** Qualidade
- Recursos **vs** Detalhamento

Modelos Comparados:

Modelo	Parâmetros	Tamanho
Flan-T5-Small	80M	320 MB
Flan-T5-Base	250M	990 MB

Como Testamos

Executamos 20 testes em 5 categorias:

- Conhecimento geral (4 testes)
- Matemática (4 testes)
- Tradução (3 testes)
- Lógica (3 testes)
- Criatividade (3 testes)

Solução

Código para comparar modelos

```python
from transformers import pipeline

# Carregar modelos do HuggingFace
model_small = pipeline("text2text-generation",
                       model="google/flan-t5-small")

model_base = pipeline("text2text-generation",
                      model="google/flan-t5-base")

# Testar e medir tempo
for pergunta in perguntas:
    inicio = time.time()
    resposta = model(pergunta)
    tempo = time.time() - inicio

    # Salvar métricas
    resultados.append({
        'modelo': nome,
        'tempo': tempo,
        'resposta': resposta
    })
```

O que medimos:

- Tempo de resposta (segundos)
- Velocidade (tokens/segundo)
- Tamanho das respostas
- Uso de memória

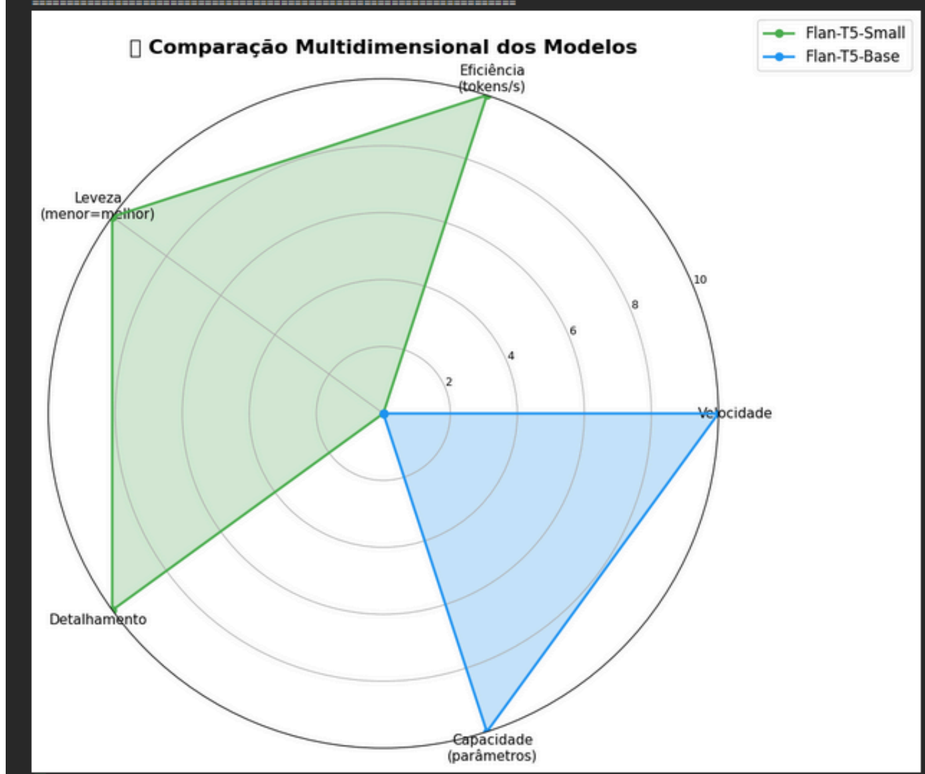
Tecnologias:

- HuggingFace Transformers - Carregar modelos
- PyTorch - Backend
- Pandas - Análise de dados
- Matplotlib - Gráficos

Execução

```
...

============================================================
🏆 RESUMO FINAL DA COMPARAÇÃO
============================================================

📊 TABELA COMPARATIVA FINAL:

      Modelo  Tempo Médio (s)  Velocidade Média (tokens/s)  Tamanho Médio Resposta  Parâmetros (M)  Tamanho Disco (MB)
Flan-T5-Small           3.653                         8.82                    48.6            77.0               293.6
 Flan-T5-Base           2.603                         3.95                    27.5           247.6               944.4

============================================================
🏅 VENCEDORES POR CATEGORIA:
============================================================

⚡ MAIS RÁPIDO: Flan-T5-Base
   Tempo médio: 2.603s

🚀 MAIOR VELOCIDADE DE TOKENS: Flan-T5-Small
   Velocidade: 8.82 tokens/s

💾 MAIS LEVE: Flan-T5-Small
   Tamanho: 293.6 MB

📝 RESPOSTAS MAIS DETALHADAS: Flan-T5-Small
   Tamanho médio: 48.6 caracteres

============================================================
💡 RECOMENDAÇÕES DE USO:
============================================================

📱 Flan-T5-Small:
   ✅ Ideal para: Aplicações mobile, dispositivos com recursos limitados
   ✅ Vantagens: Mais rápido, menor consumo de memória
   ⚠️  Limitações: Respostas mais concisas, menos parâmetros

💻 Flan-T5-Base:
   ✅ Ideal para: Servidores, aplicações desktop, quando qualidade é prioridade
   ✅ Vantagens: Respostas mais elaboradas, maior capacidade
   ⚠️  Limitações: Mais lento, maior uso de recursos
```

```
============================================================
📊 GRÁFICO RADAR DE COMPARAÇÃO:
============================================================
```




```
✅ Gráfico radar gerado!
```