Tech Discovers

Gabriel Cassol Bach, Bruno Emanuel Zenatti

UniversidadequeTransforma
UTFPR
UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ

Contexto Geral

- **Objetivo**: Identificar e descrever padrões, tendências e inovações tecnológicas a partir das informações recuperadas de documentos de patentes por meio de um sistema RAG.

- **Dataset**: Solicitações de patentes feitas no EUA entre 2011 e 2016 (~4.5 milhões de pedidos de patentes).
  - Utilizado somente o ano de 2016 (~373 mil patentes).

Perguntas de Pesquisa

- **Primeira pergunta:** Quais estratégias de recuperação são mais eficazes para encontrar patentes relevantes?

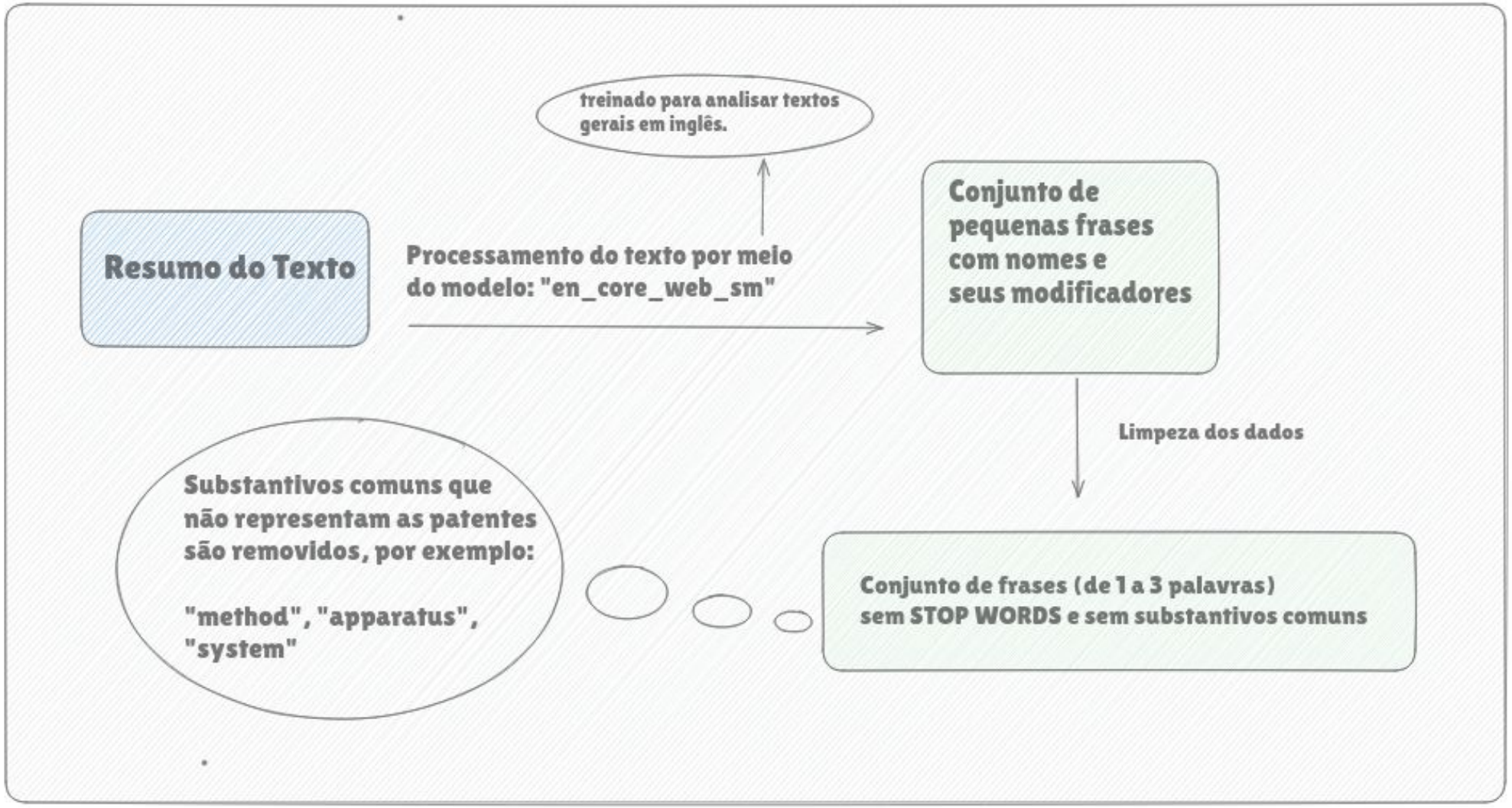
- **Segunda Pergunta:** Como o RAG facilita tarefas de análise de patentes, como identificação de antecedentes e inovações?

- **Terceira Pergunta:** Quais são as limitações das buscas utilizando o RAG?

Pré-Processamento

- Geração de palavras chave que representam o texto.
- Vetorização de campos específicos das patentes:
  - Título;
  - Palavras chave;
  - Resumo da patente.

Extração de Palavras-Chave

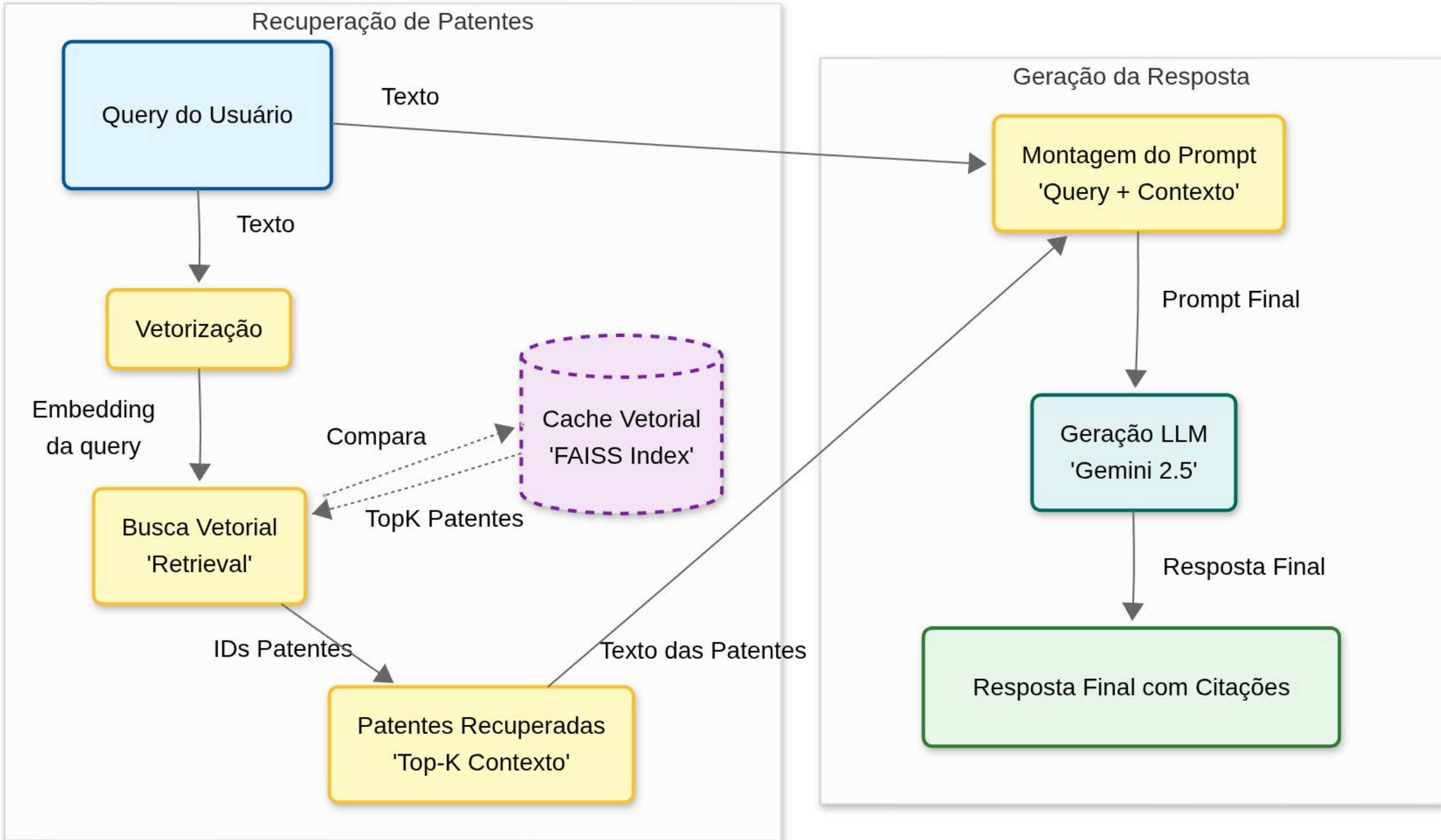


RAG (Retrieval-Augmented Generation)

- **Conceito Chave:** Mecanismo que combina recuperação de documentos relevantes + geração por LLM usando esses documentos como contexto.
- **Palavras-chave:** Auxiliam no Retrieval do RAG.
- **Aplicação no Projeto:** O sistema fornece um contexto preciso, verificável e alinhado ao domínio técnico analisado para o modelo.

Por que RAG?

- **Problema de LLMs Puros:** Falta de contexto específico, o que impede que as respostas sejam precisas para um certo domínio.
- **Contexto de Patentes:** Necessidade de precisão factual.
- **Rastreabilidade:** Cita exatamente qual documento embasou a resposta.
- **Janela de Contexto:** O RAG seleciona um pequeno contexto, tendo em vista que é impossível passar 373 mil patentes no prompt.


Recuperação de Patentes
Query do Usuário
Texto
Texto
Vetorização
Embedding
da query
Compara
Cache Vetorial
'FAISS Index'
TopK Patentes
Busca Vetorial
'Retrieval'
IDs Patentes
Patentes Recuperadas
'Top-K Contexto'
Texto das Patentes
Geração da Resposta
Montagem do Prompt
'Query + Contexto'
Prompt Final
Geração LLM
'Gemini 2.5'
Resposta Final
Resposta Final com Citações

Detalhes Técnicos

- **Geração dos *Embeddings*:** Testes com all-MiniLM-L6-v2 e BAAI/bge-large-en-v1.5.
- **Biblioteca de Busca:** FAISS (Facebook AI Similarity Search).
- **Métrica de Similaridade:**
  - Normalização L2 dos vetores.
  - IndexFlatIP: Produto interno com vetores normalizados, equivale à Similaridade de Cosseno.
- **Cache**:
  - Os vetores das 373 mil patentes são pré-calculados e salvos (.npy).
  - As palavras chaves, e os seus textos, são salvos e mapeados com o respectivo vetor.

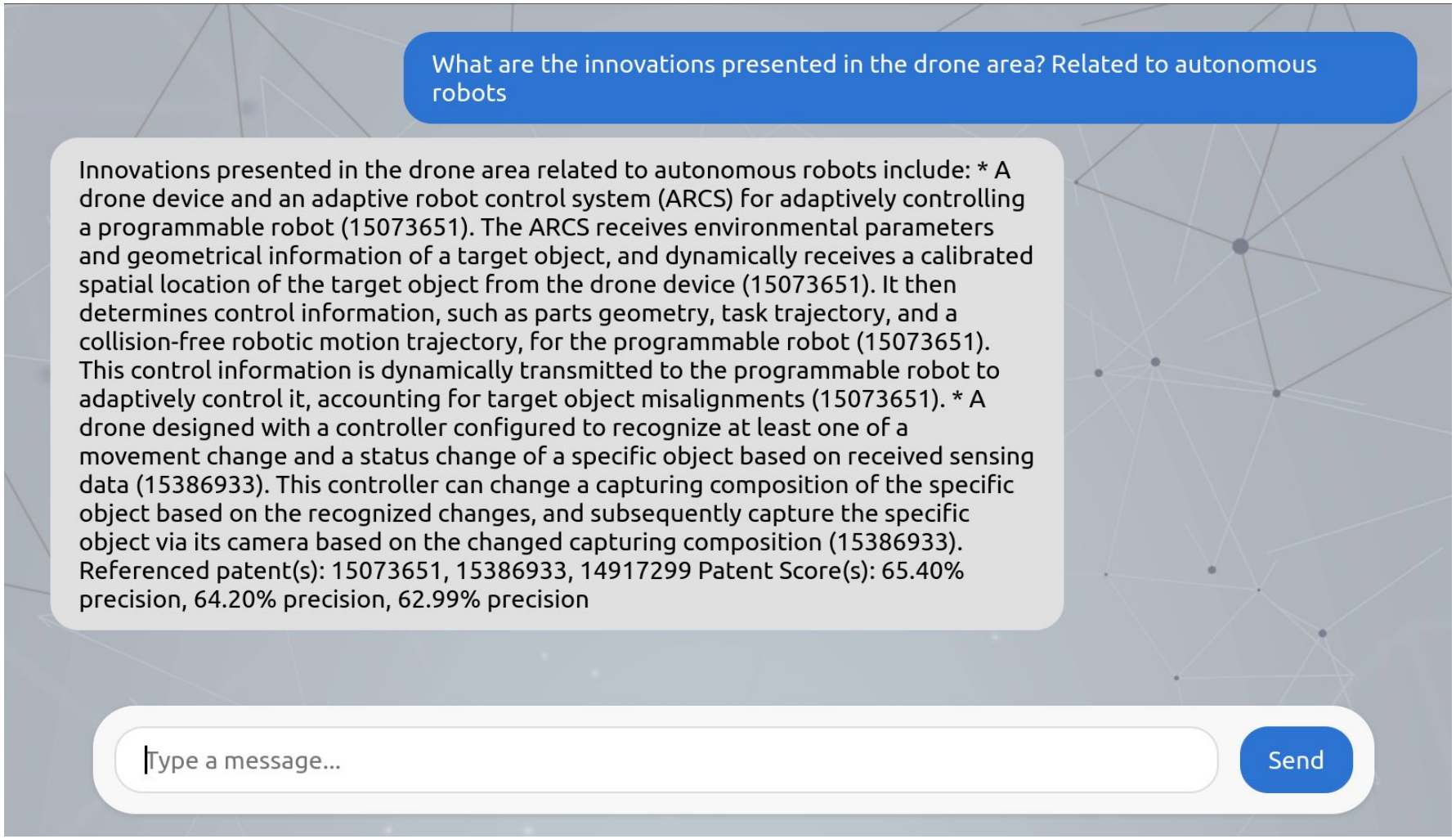
Engenharia de Prompt e Resposta

- **Seleção de Contexto:** Top-K (ex: 3) patentes mais relevantes acima de um score threshold (0.1).
- **Estrutura do Prompt:**
  - Regra 1: Usar apenas informações do contexto fornecido.
  - Regra 2: Responder "insufficient evidence" se não houver dados.
  - Regra 3: Citar o ID da patente para cada afirmação factual.
- **Output Estruturado:** Resposta explicativa seguida de metadados (IDs das patentes e scores de confiança).

Avaliação e Resultados

- **Metodologia:** Avaliação qualitativa manual.
- **Amostra:** 10 perguntas diversificadas sobre o dataset.
- **Critérios de Sucesso:**
  - **Capacidade de Recuperação (Retrieval)**: O sistema encontrou patentes relevantes?
  - **Fidelidade**: A resposta obedeceu ao prompt e citou as fontes?
  - **Alucinação**: O modelo inventou informações fora do contexto?

Interface Para Consultas



Análise dos Parâmetros

***Score Threshold* do *Retrieval* (RAG):**

- SC < 50 %: Patentes recuperadas com pouca relação com a pergunta.
- SC >= 50%: Ponto ótimo, conteúdo relevante não é perdido.
- SC >= 60%: Melhora a precisão, mas diversas patentes úteis são perdidas.

Quantidade de patentes recuperadas (*Top-K*):

- Top-K = 3: Retorna uma resposta mais concisa, porém, às vezes, faltam informações.
- Top-K = 5: Fornece mais contexto para a geração, melhorando a qualidade da resposta. A LLM, no geral, não se confunde com mais contexto.

Análise dos Parâmetros

Modelo para geração de *Embedding:*

- *all-MiniLM-L6-v2*: modelo genérico para gerar *embeddings.* Funciona de forma satisfatória, mas não é tão efetivo.

- *BAAI/bge-large-en-v1.5:* modelo similar ao all-MiniLM-L6-v2 com mais parâmetros. Além disso, possui como ponto forte o retrieval (MTEB).

Resultados

(Top-3; All-Mini) X (Top-5; Bge-large)

Pergunta	Tema	Nota 1	Nota 2
1	Diabetes	9/10	7/10
2	Redes	7/10	10/10
3	Biodegradáveis	9/10	8/10
4	Drones	7/10	9/10
5	Máquinas elétricas	10/10	7.5/10
6	ML/IR/NLP	5/10	9/10
7	Wearables	10/10	9/10
8	E-cigarettes	7/10	10/10
9	Antecedentes de lentes de câmera	6/10	8/10
10	Lentes câmera	10/10	8/10

Trabalhos Futuros

Possíveis Melhorias:

- Tanto a recuperação quanto a geração das respostas dependem de boas perguntas.
- Encontrar uma forma de tratar as perguntas que o usuário insere na aplicação:
  - Ideias: Gerar modelos de pergunta para o usuário na interface ou criar um algoritmo que consegue adequar diferentes perguntas a um modelo.
- Remover patentes duplicadas.

Obrigado!

UTFPR
UNIVERSIDADE TECNOLÓGICA FEDERAL DO PARANÁ