Evaluación y mejora del contexto en Large Language Models

Universidad Internacional de Valencia

Álvaro Beltrán Camacho

De:
Planeta Formación y Universidades

Indice

1. Large Language Models
2. Retrieval Augmented Generation
3. Caso Práctico
4. Demo

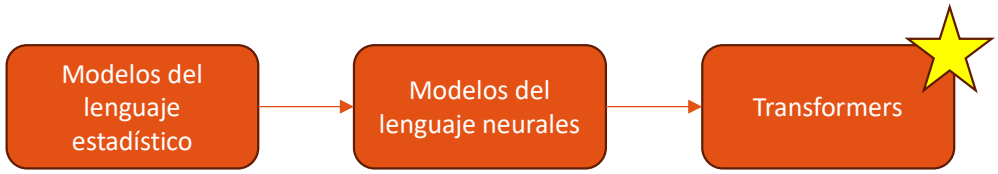
Large Language Models

Problema: Generación de texto

Predicción del siguiente token

Formalmente, si tenemos una secuencia de tokens $t_1, t_2, ..., t_n$, el modelo aprende a predecir la probabilidad del token t_{n+1} basándose en los tokens $t_1, t_2, ..., t_n$. Este proceso se puede expresar como:

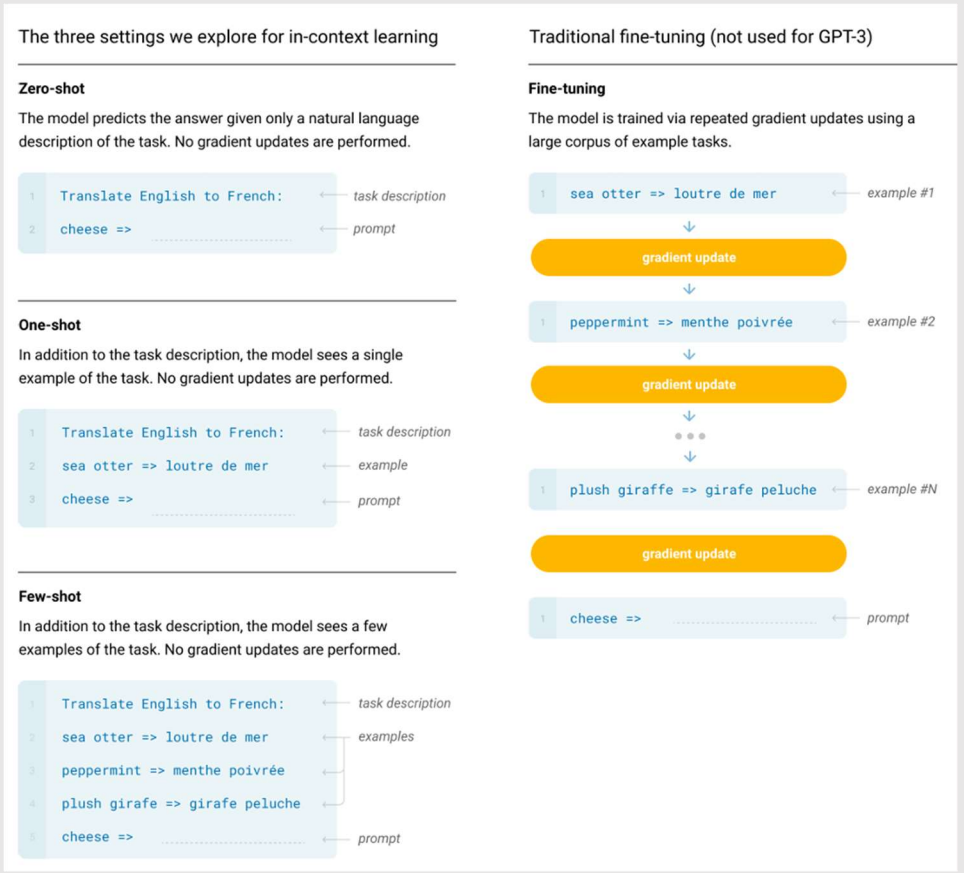
$$P(t_{n+1}|t_1, t_2, ..., t_n)$$



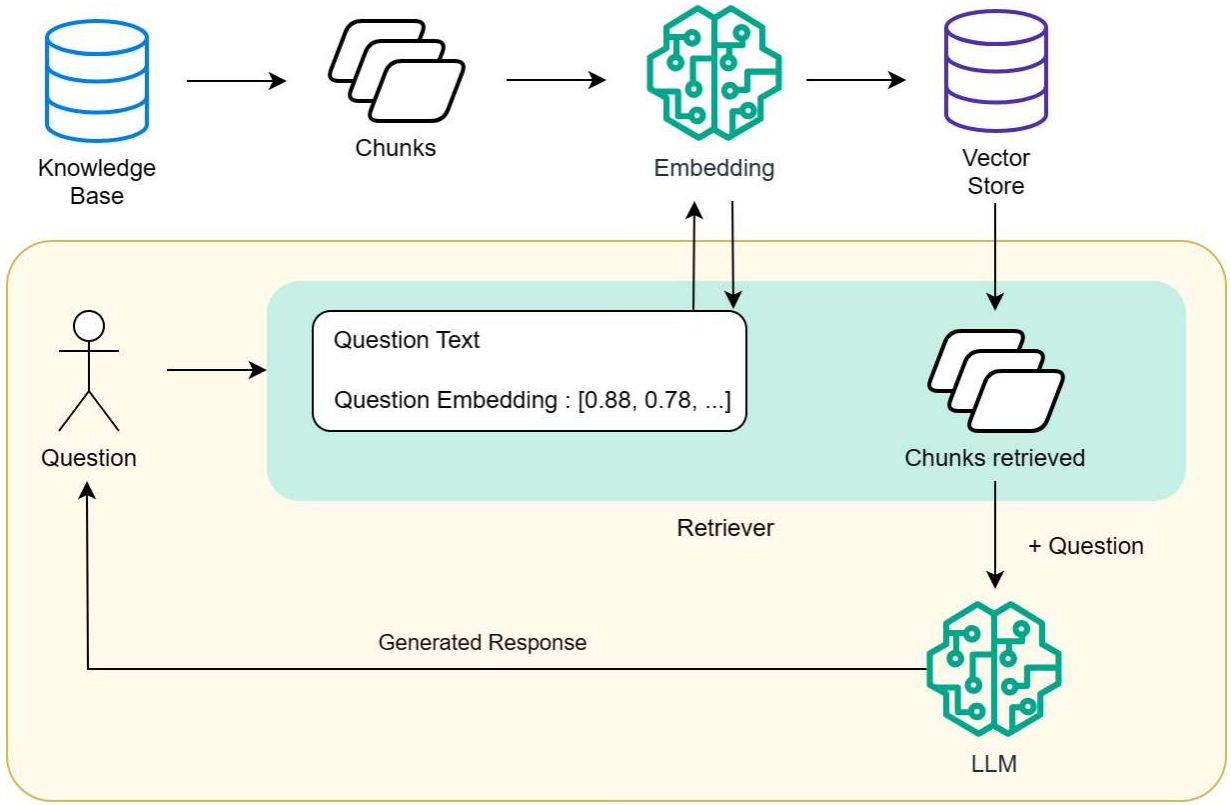
Subtítulo presentación

Universidad Internacional de Valencia

Cambio de paradigma: zero-shot, one-shot y few-shot



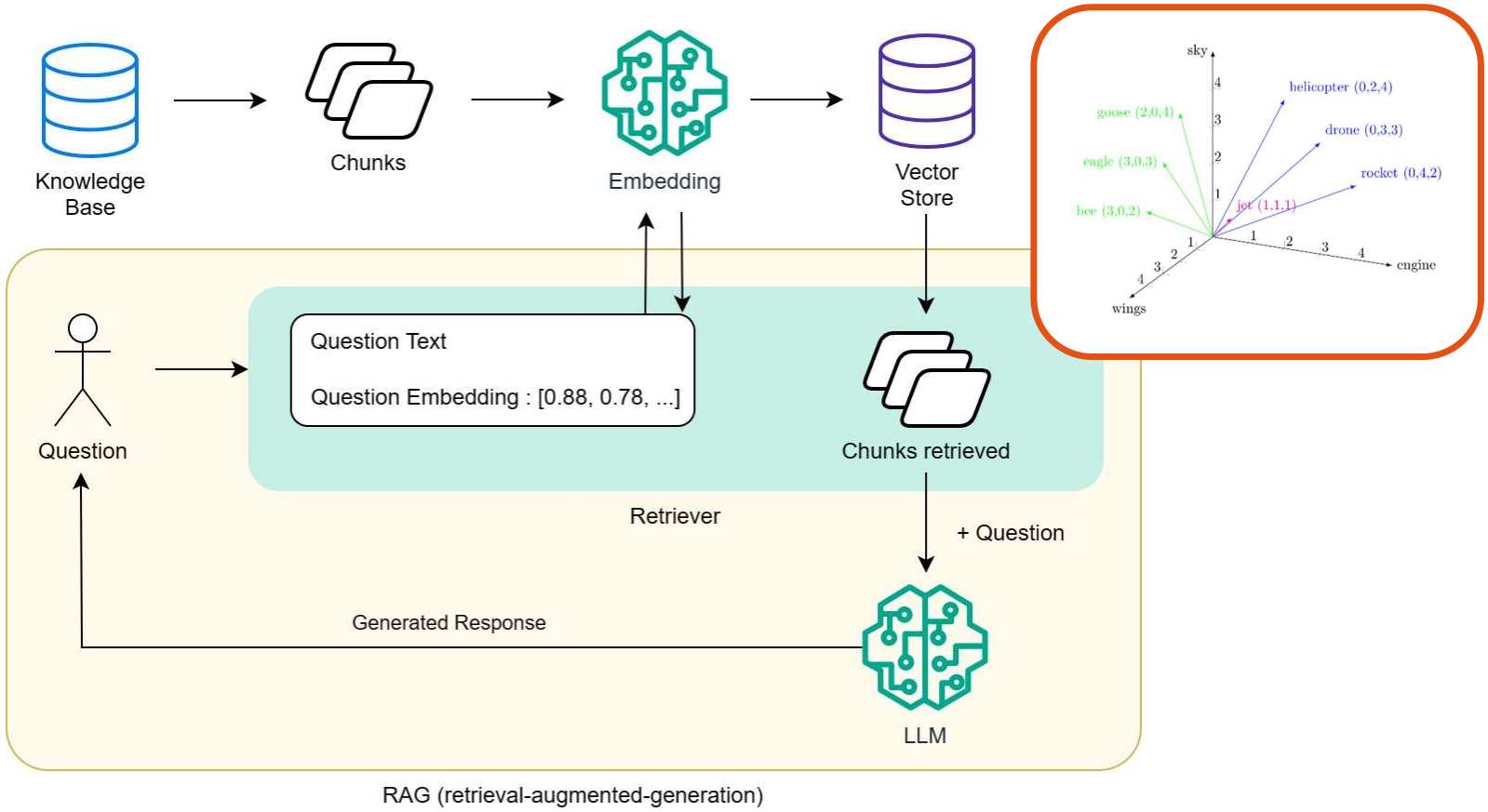
Retrieval Augmented Generation



RAG (retrieval-augmented-generation)

Subtítulo presentación

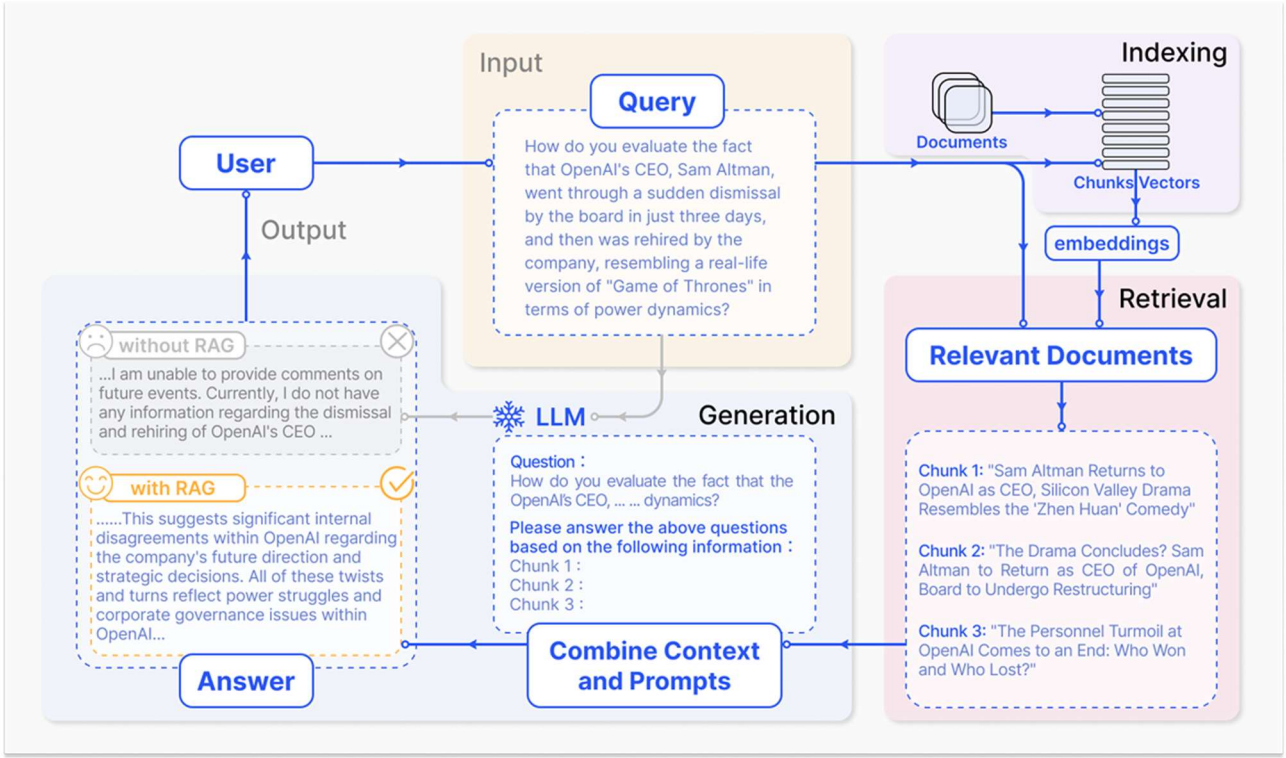
Retrieval Augmented Generation



RAG (retrieval-augmented-generation)

Subtítulo presentación

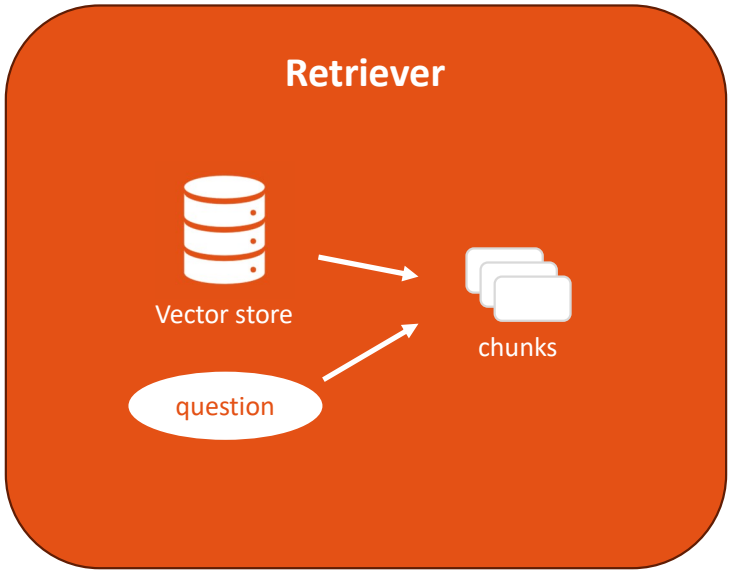
Retrieval Augmented Generation

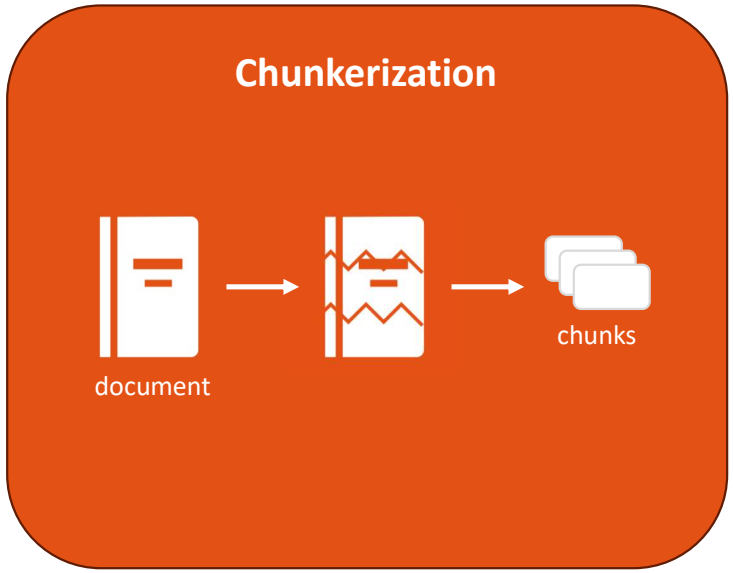


Subtítulo presentación

Retrieval Augmented Generation

Puntos relevantes en el sistema

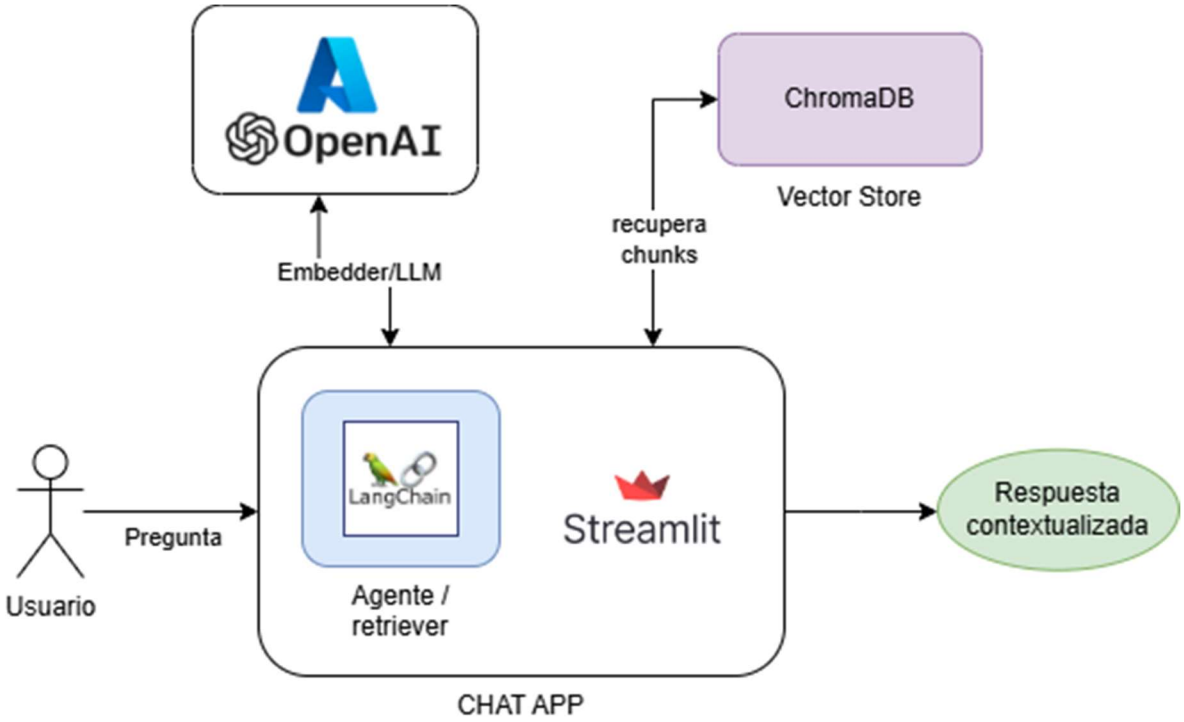




Caso Práctico:

RAG sobre constitución española

Diagrama Arquitectura



Subtítulo presentación

Caso Práctico:

RAG sobre constitución española

Modelos usados

Modelo Juez
-
GPT4o

Modelo del Agente
-
GPT 3.5 Turbo

Embedder
-
Text-embedding-3-small

Caso Práctico:

RAG sobre constitución española

Pasos para la creación

Chunkerization

7 formas de chunkerizar con
3 metodologías diferntes

RecursiveCharacterSplitter

CharacterSplitter

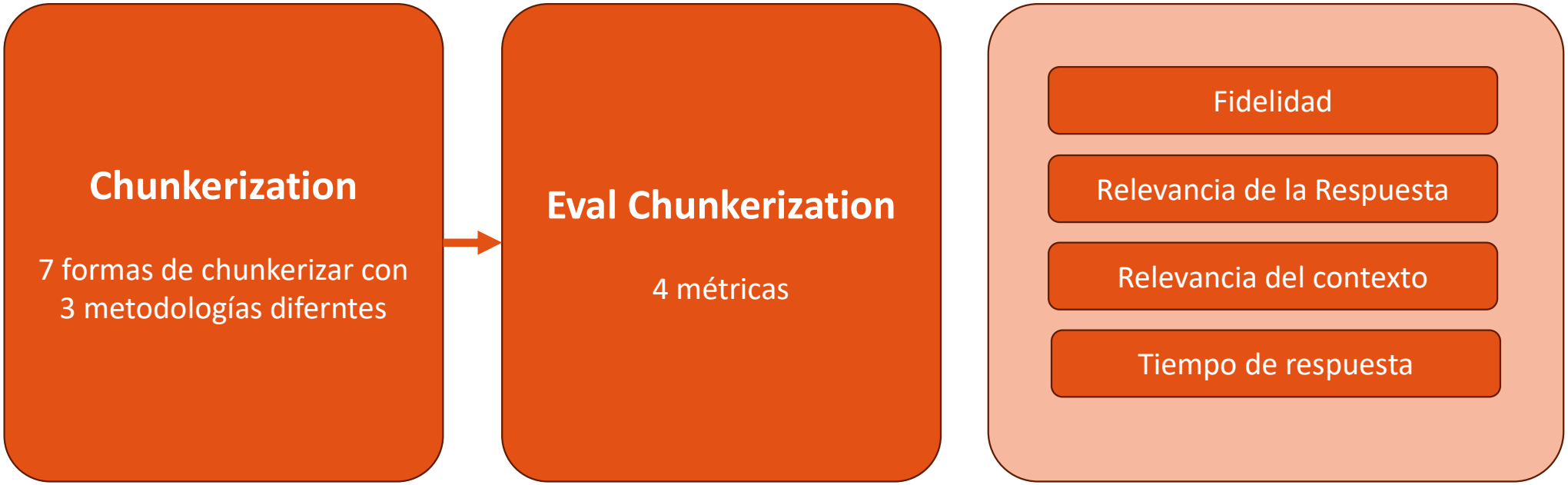
CustomSplitter

```
<h5 class="articulo">Artículo 3</h5>
<p class="parrafo">1. El castellano es la lengua e
Todos los españoles tienen el deber de conocerla y
<p class="parrafo">2. Las demás lenguas españolas
respectivas Comunidades Autónomas de acuerdo con s
<p class="parrafo">3. La riqueza de las distintas
España es un patrimonio cultural que será objeto d
</div>
```

Caso Práctico:

RAG sobre constitución española

Pasos para la creación



Caso Práctico:

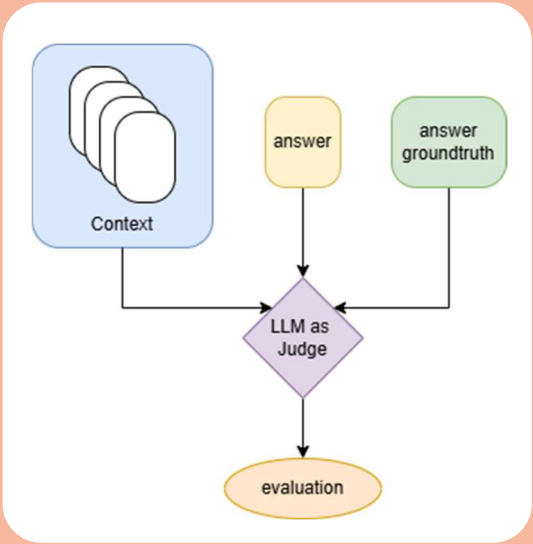
RAG sobre constitución española

Pasos para la creación

Eval Chunkerization

4 métricas

- Fidelidad
- Relevancia de la Respuesta
- Relevancia del contexto
- Tiempo de respuesta



Caso Práctico:

RAG sobre constitución española

Pasos para la creación

Demo

Subtítulo presentación