

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

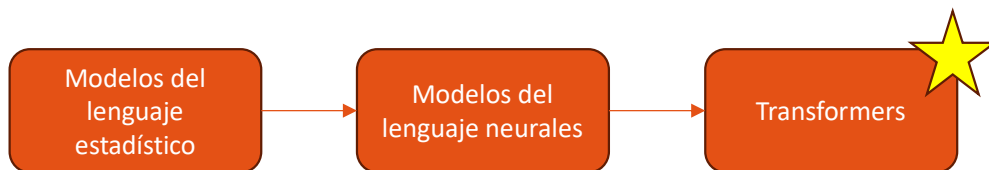
Large Language Models

Problema: Generación de texto

Predicción del siguiente token

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

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



Subtítulo presentación

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

The three settings we explore for in-context learning

Zero-shot

The model predicts the answer given only a natural language description of the task. No gradient updates are performed.

```

1 Translate English to French: <-- task description
2 cheese => ..... <-- prompt
  
```

One-shot

In addition to the task description, the model sees a single example of the task. No gradient updates are performed.

```

1 Translate English to French: <-- task description
2 sea otter => loutre de mer <-- example
3 cheese => ..... <-- prompt
  
```

Few-shot

In addition to the task description, the model sees a few examples of the task. No gradient updates are performed.

```

1 Translate English to French: <-- task description
2 sea otter => loutre de mer <-- examples
3 peppermint => menthe poivrée <--
4 plush giraffe => girafe peluche <--
5 cheese => ..... <-- prompt
  
```

Traditional fine-tuning (not used for GPT-3)

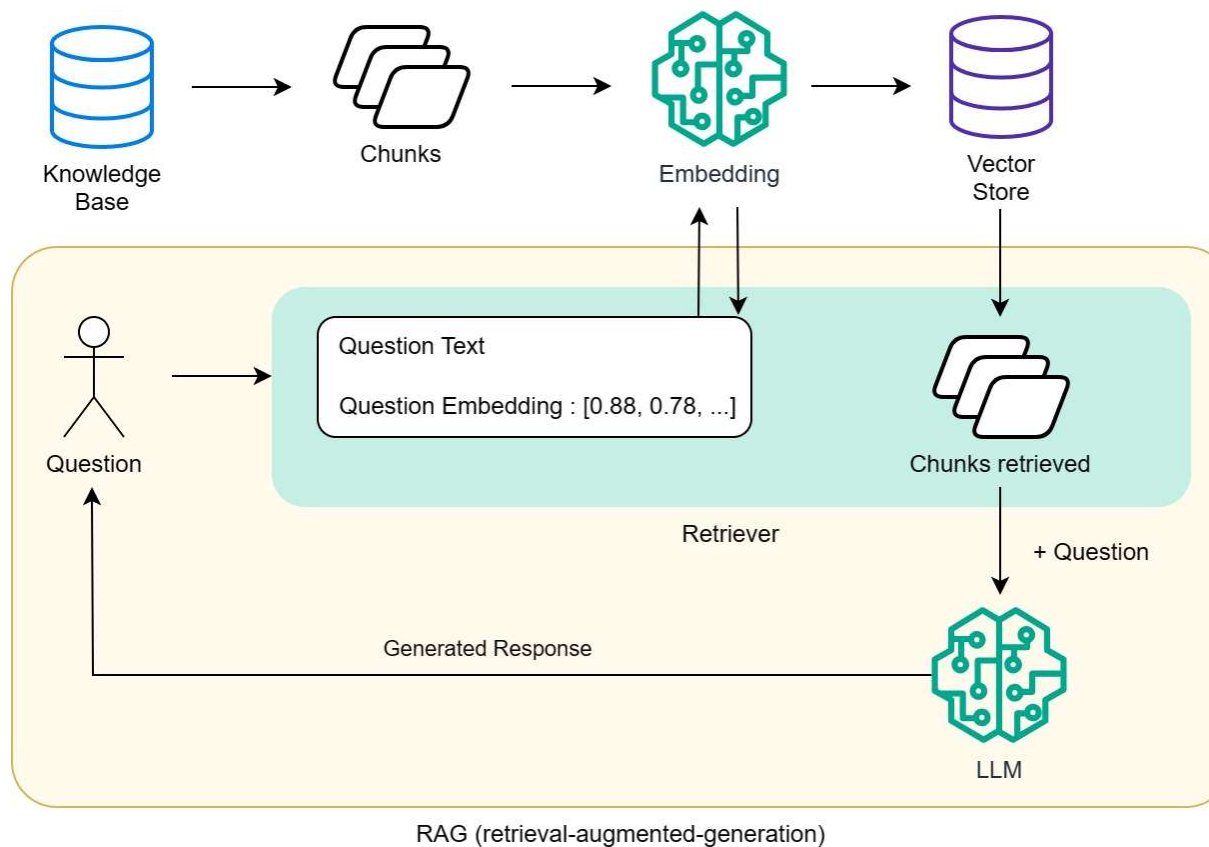
Fine-tuning

The model is trained via repeated gradient updates using a large corpus of example tasks.

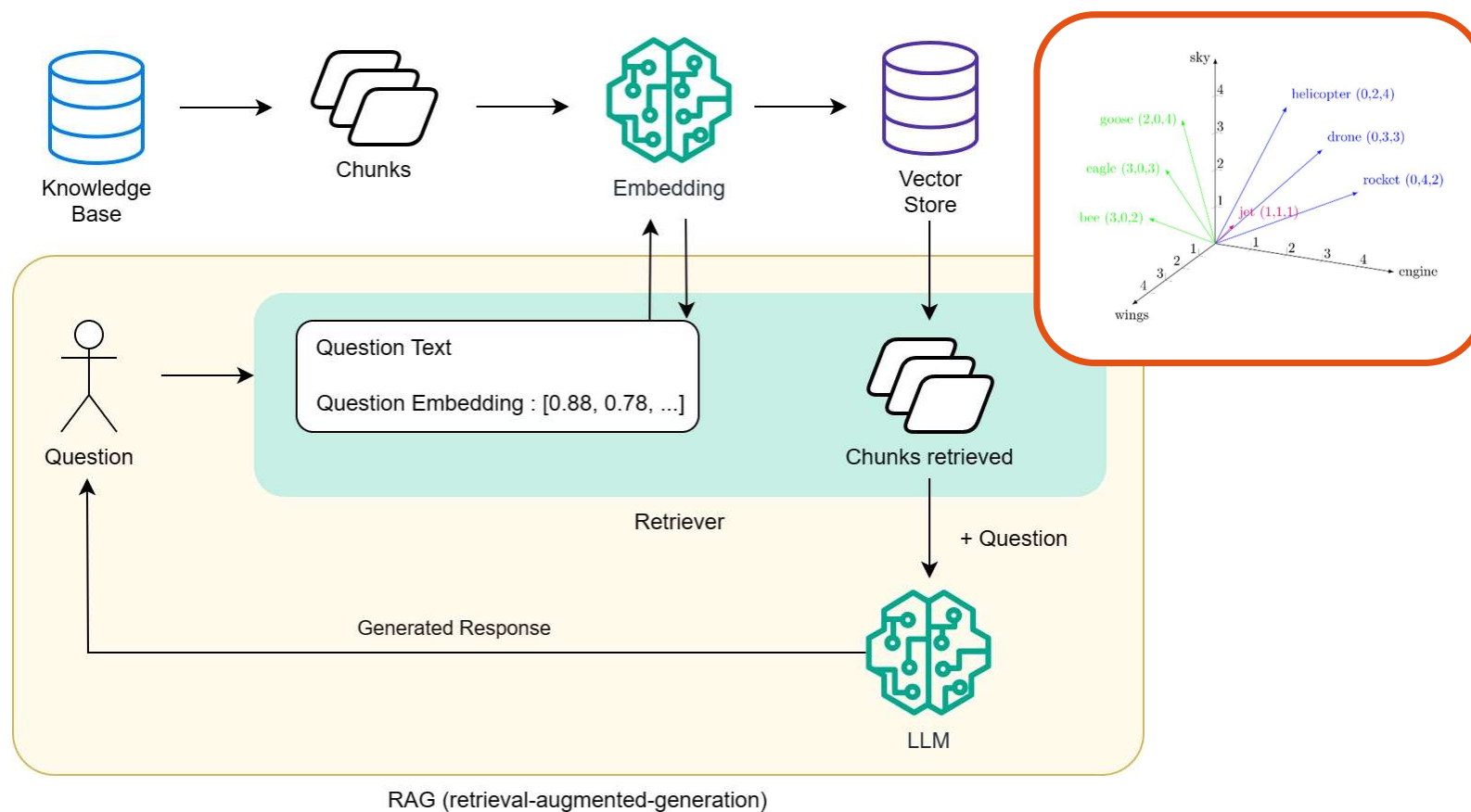
```

1 sea otter => loutre de mer <-- example #1
  |
  v
gradient update
  |
  v
1 peppermint => menthe poivrée <-- example #2
  |
  v
gradient update
  |
  v
...
  |
  v
1 plush giraffe => girafe peluche <-- example #N
  |
  v
gradient update
  |
  v
1 cheese => ..... <-- prompt
  
```

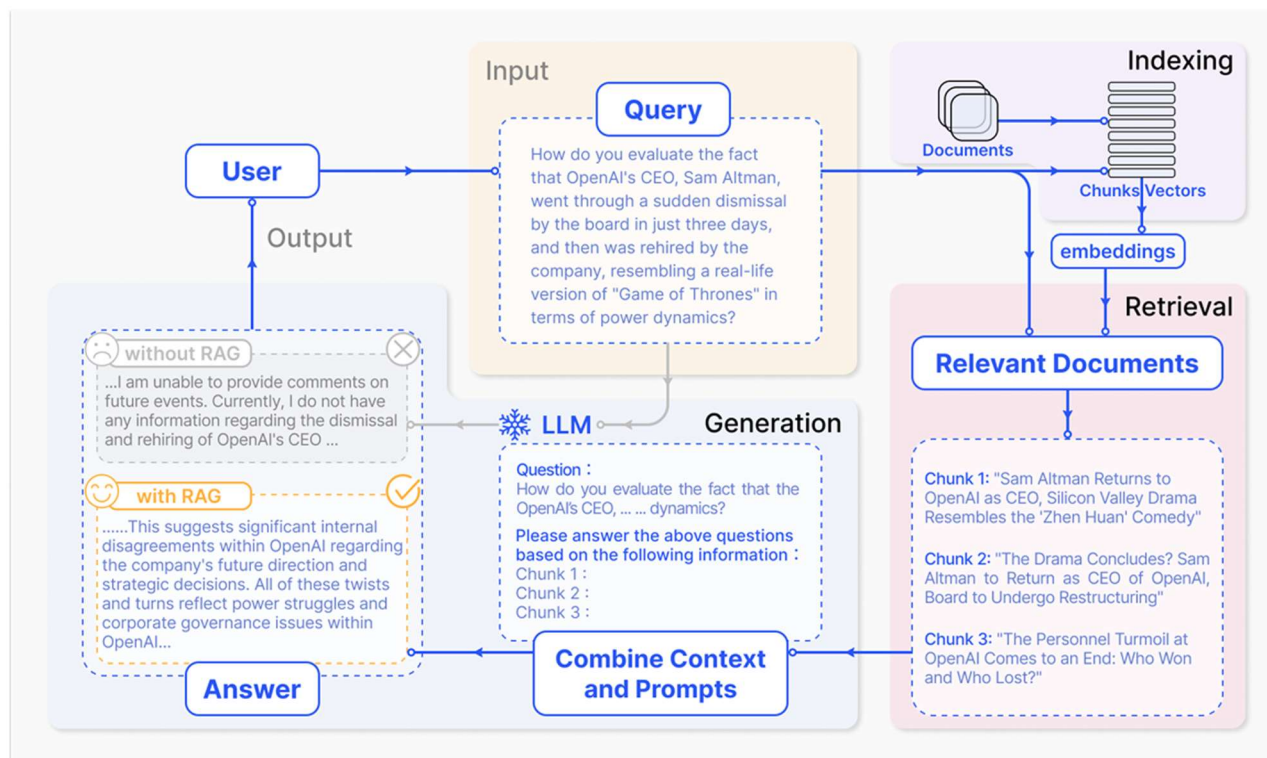
Retrieval Augmented Generation



Retrieval Augmented Generation

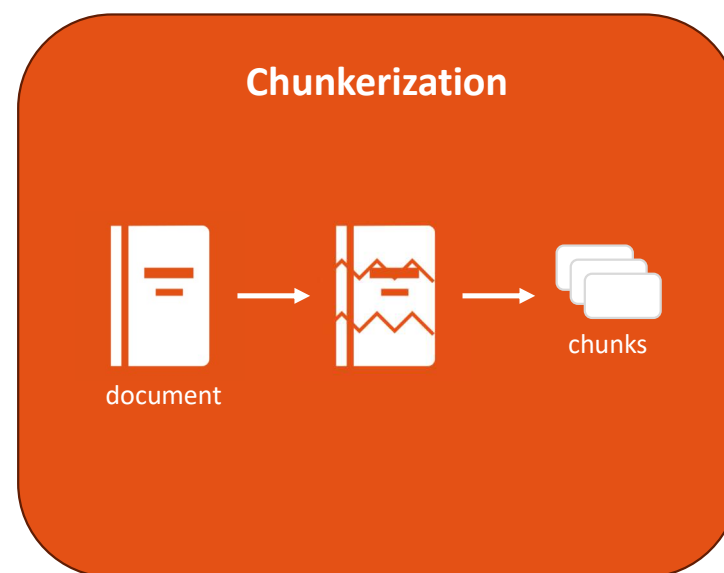
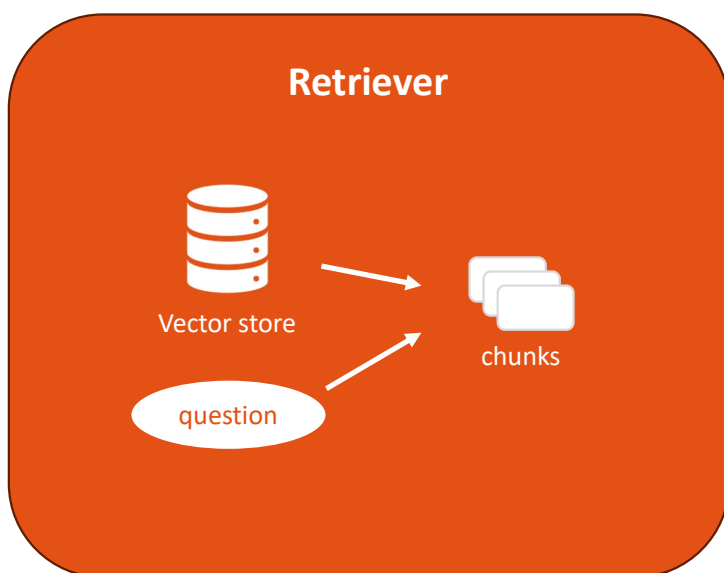


Retrieval Augmented Generation



Retrieval Augmented Generation

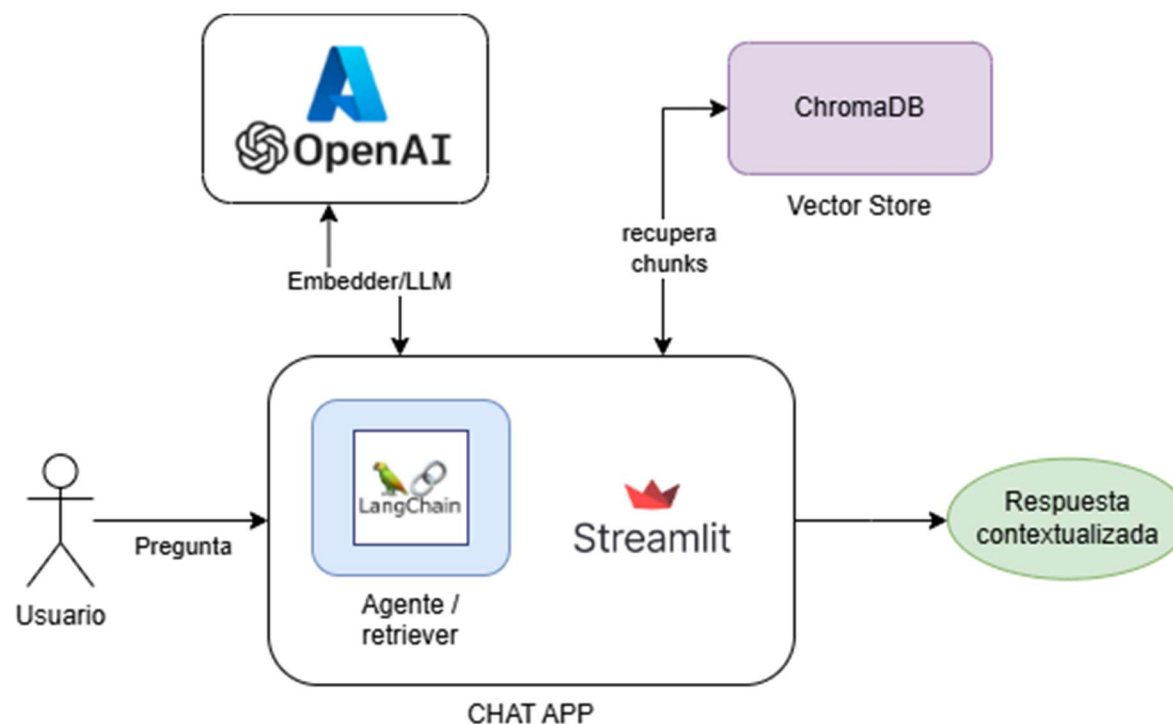
Puntos relevantes en el sistema



Caso Práctico:

RAG sobre constitución española

Diagrama Arquitectura



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 diferentes

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 c
</div>
```

Caso Práctico:

RAG sobre constitución española

Pasos para la creación



Caso Práctico:

RAG sobre constitución española

Pasos para la creación

Eval Chunkerization

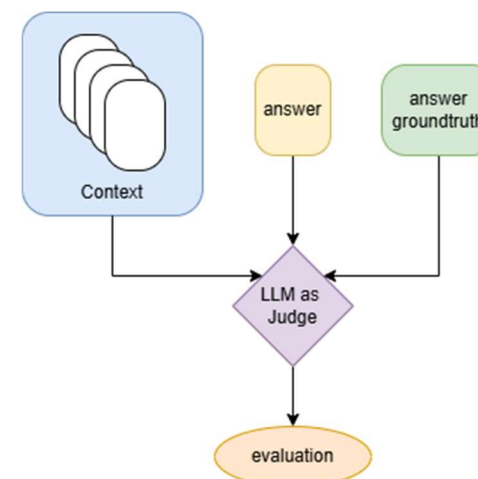
4 métricas

Fidelidad

Relevancia de la Respuesta

Relevancia del contexto

Tiempo de respuesta



Caso Práctico:

RAG sobre constitución española

Pasos para la creación



Demo