Grans Models de Llenguatge (com ChatGPT) per entorns locals i tancats

Dr. Isaac Lera

III Jornada
Networking per a la Transferència d'Aplicacions de la IA a les Empreses i la Societat de les Illes Balears (JAIA2024+)

Fundació Universitat Empresa de les Illes Balears

laia.uib.cat

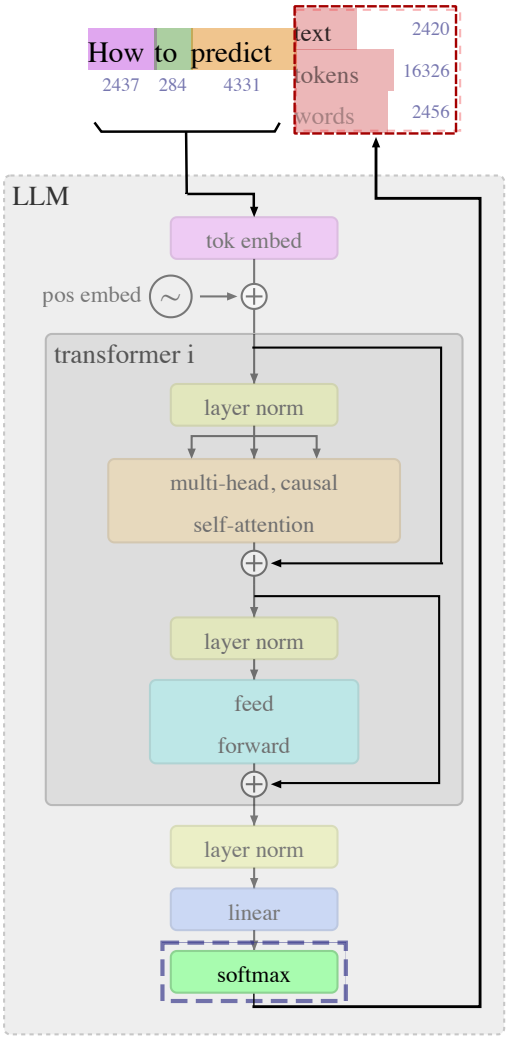
Roadmap

Basic Concepts on LLM

How to use LLM

Our Support: LAIA@UIB

Basic concepts on LLM (I)



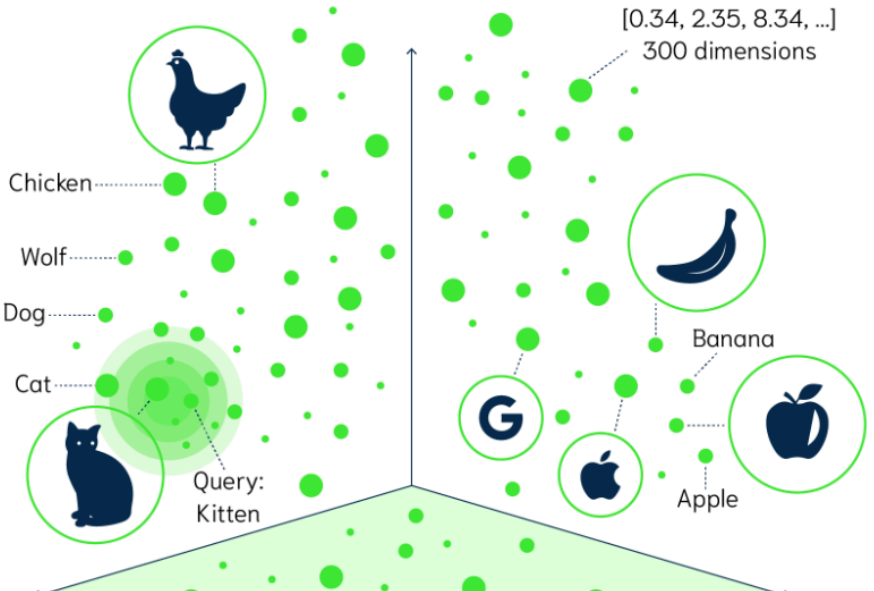
Tokens

Aixòpotrepresentaruntokenen2024
(GPT-3)

Embeddings

[-0.09144739806652069, 0.06914552301168442, 0.009324240498244762, ...]

Search Space



Basic concepts on LLM (II)

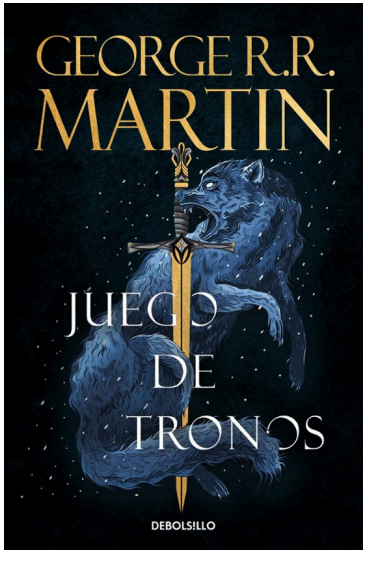
OpenAI o1

Frontier reasoning model that supports tools, Structured Outputs, and vision | 200k context length

Price Input:
$15.00 / 1M tokens

Cached input:
$7.50 / 1M tokens

Output:
$60.00 / 1M tokens



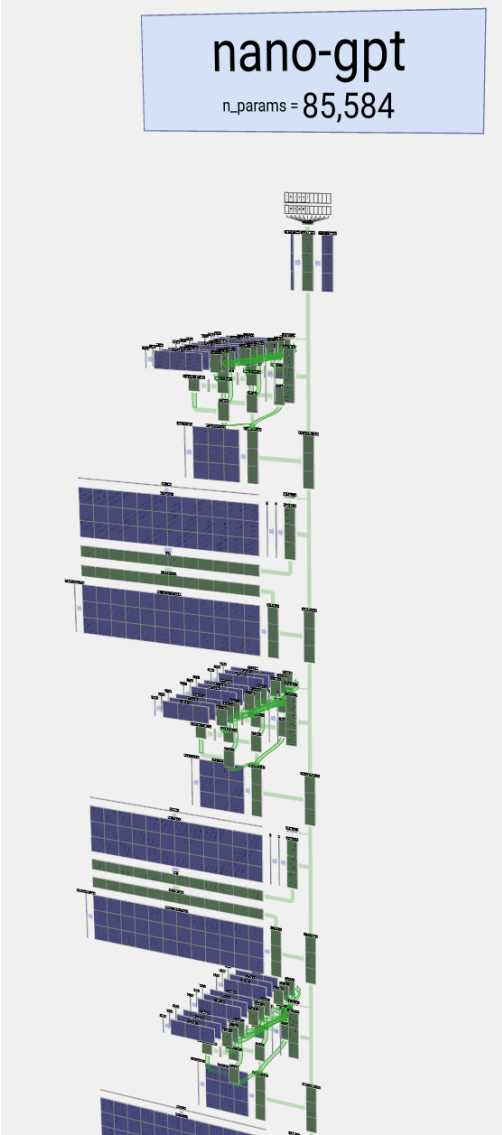
~ 1.7 M Words | ~ 2.4 M tokens

Ratio:

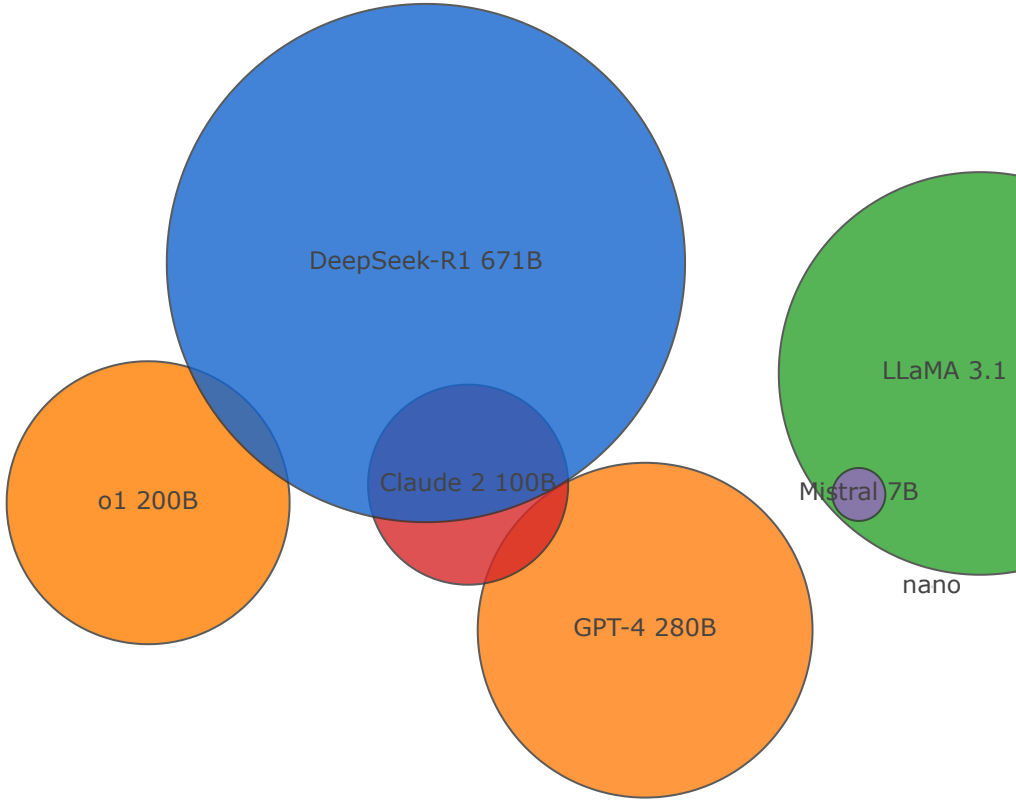
- 4 tokens for 3 words on average, so 0.75 word per token 🇬🇧
- 1 page are ~1k tokens

Windows size: 1k, 8k, 128k, 200k tokens

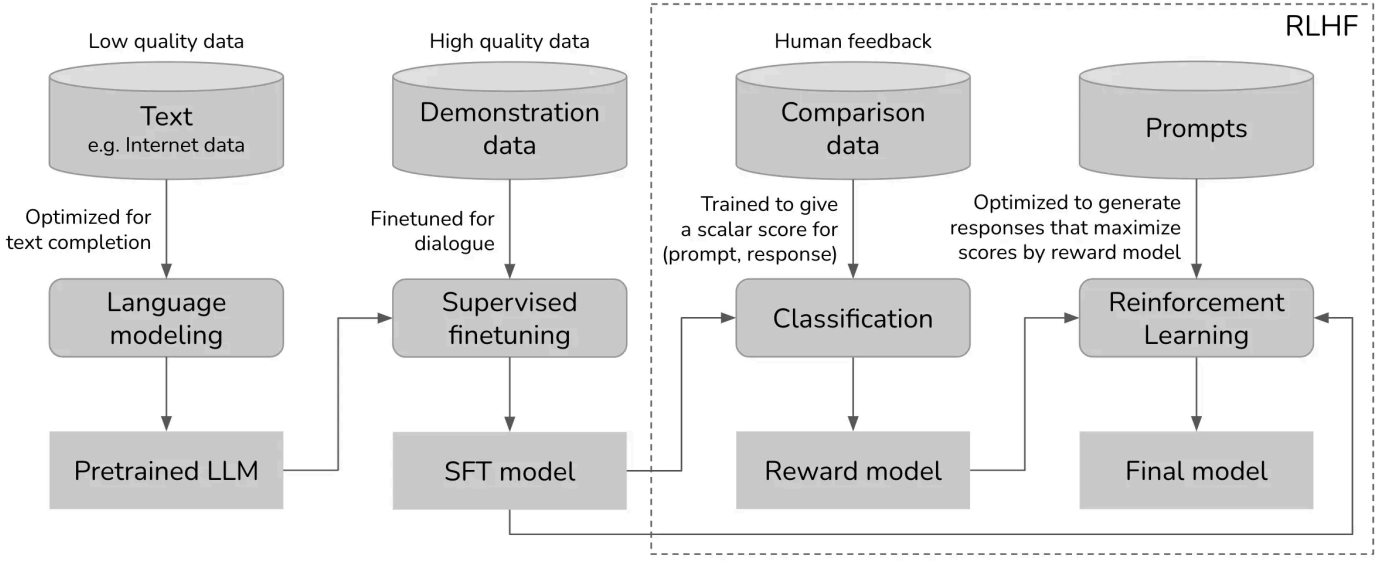
Basic concepts on LLM (III)



B = 10^9 parameters



Great Diversity of Models



Scale May '23	**>1 trillion** tokens	**10K - 100K** (prompt, response)	**100K - 1M** comparisons (prompt, winning_response, losing_response)	**10K - 100K** prompts
Examples **Bolded**: open sourced	GPT-x, Gopher, **Falcon**, LLaMa, **Pythia**, **Bloom**, **StableLM**	**Dolly-v2, Falcon-Instruct**		InstructGPT, ChatGPT, Claude, **StableVicuna**

Reference

Training Models

Memory: each parameter FP16 requieres 2 bytes. RAM = $\frac{405 \cdot 10^9 * 2}{10^9} \approx 810GB$

Quantification techniques: 8-bit, 4-bit, 2-bits, ... but at a Quality Cost

Total training of **Llama 3** (100 epochs) are:

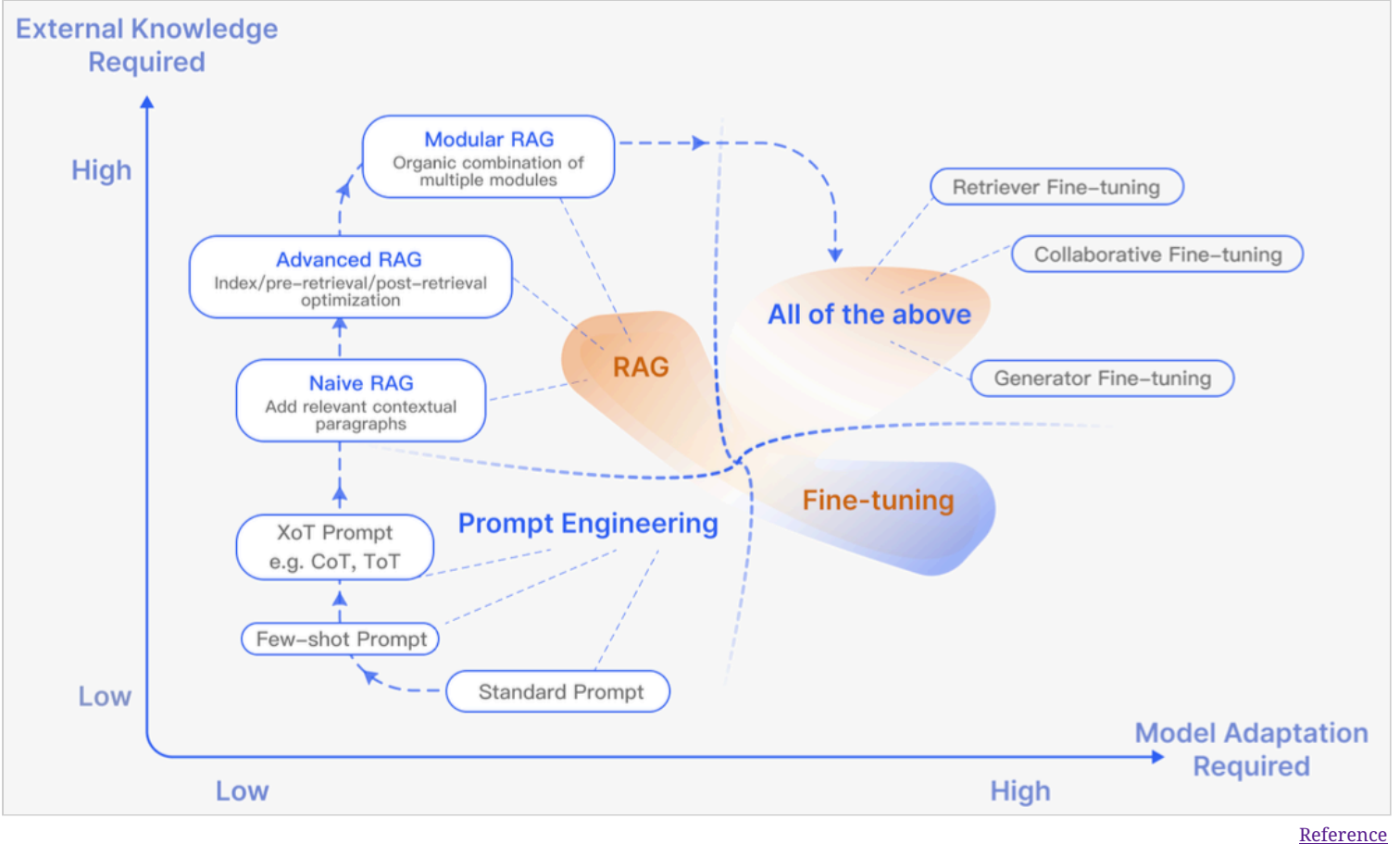
- 3.43 yotta FLOPs (3.43×10^{24} FLOPs)
- 3 months to train on 350.000 * AMD MI300X GPUs
- 350 MegaWatts of power; Central Nuclear Nacional: ~1000MW
- Carbon footprint 62.000 x higher than the avearage US household

The **Llama 3.1** models were trained on over 15 trillion tokens (10^{12}) on a custom-built GPU (H100, 30k€) cluster with a total of 39.3M GPU hours (~4486 years)

Cost: 2€/hour GPU on Cloud with 39.3M hours: 80M€

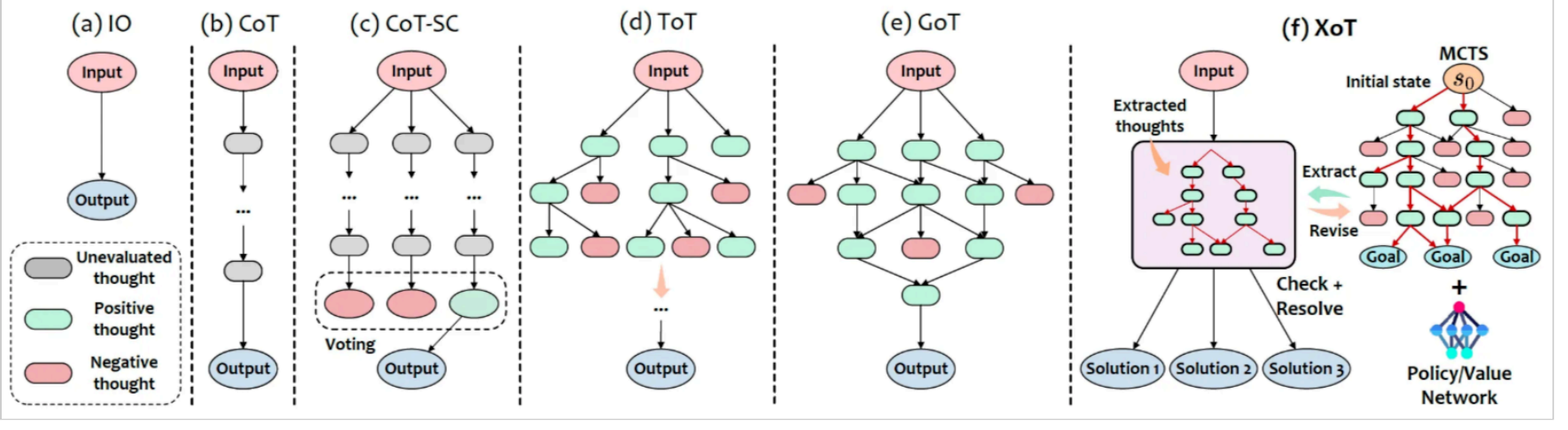
Strategies for Leveraging Local LLMs

LLM with our data



Reference

Prompting Techniques I



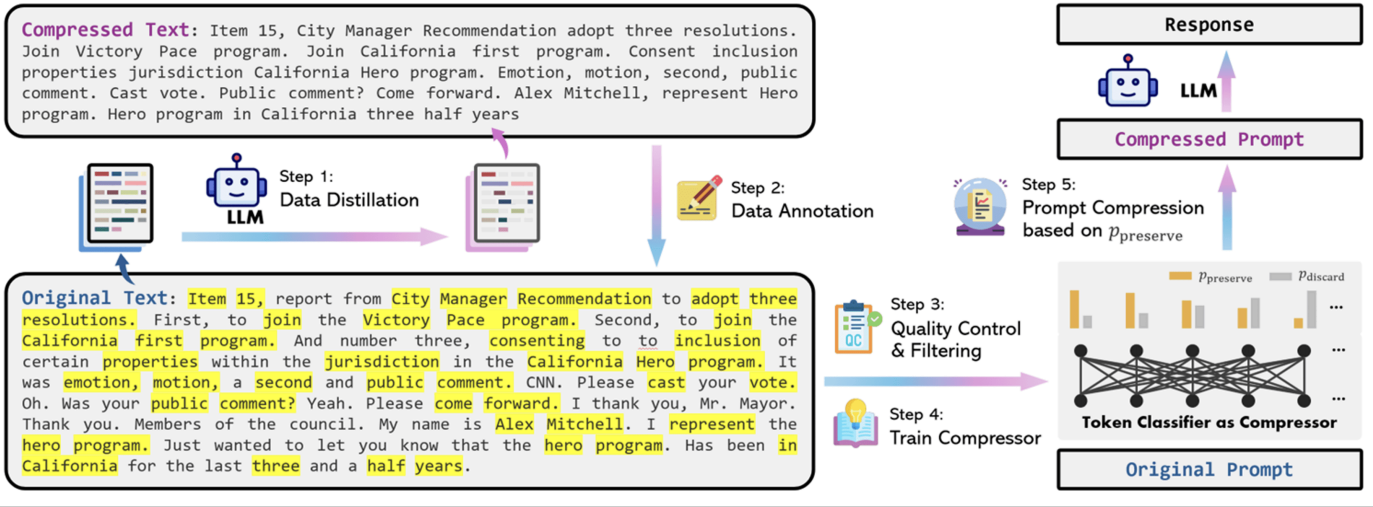
Reference

Prompting Techniques II

Microsoft ***LLMLingua-2***: Data Distillation for Efficient and Faithful Task-Agnostic Prompt Compression

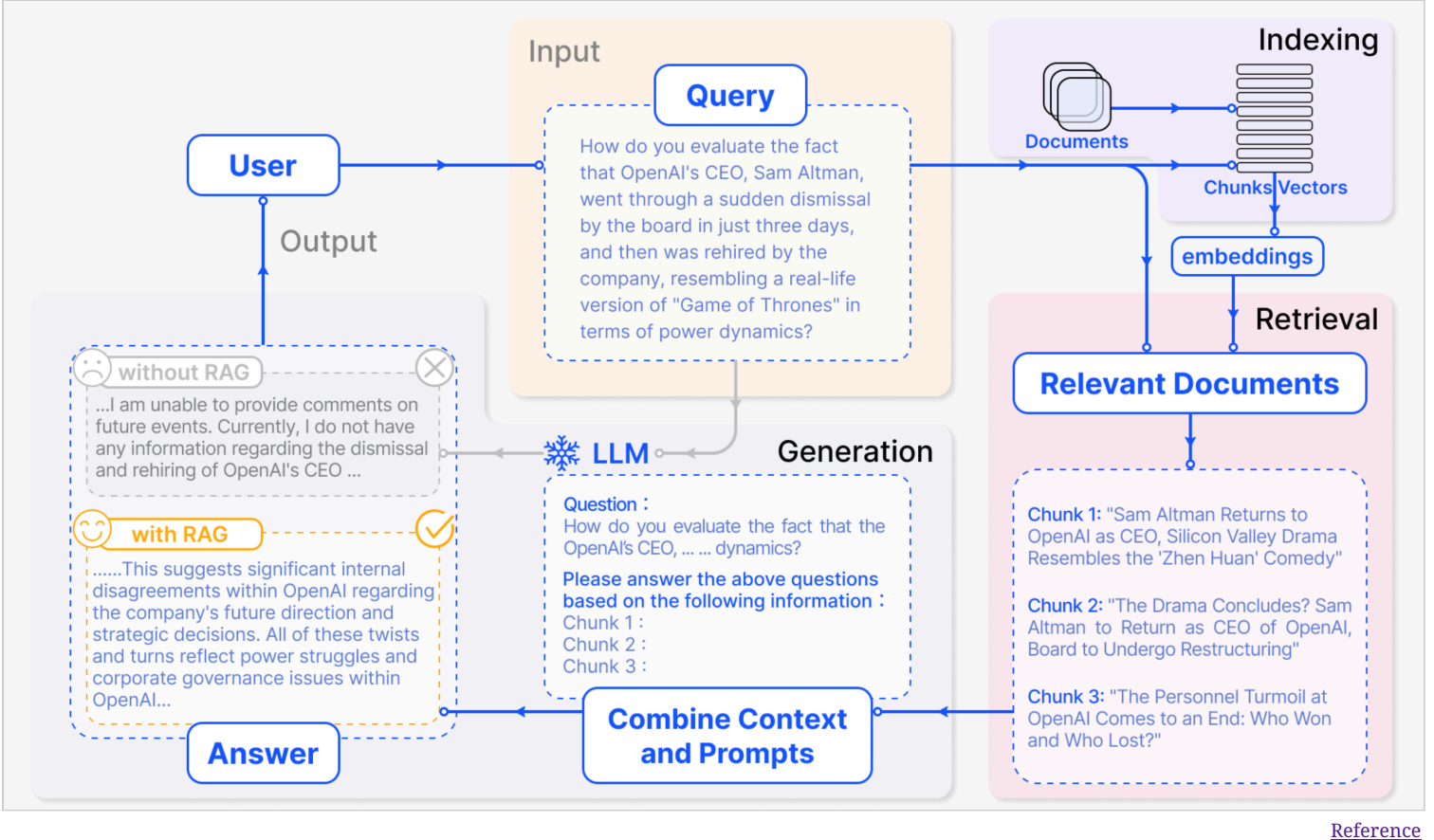
An efficient option (at bert-base scale) with good performance and generalizability across different scenarios.

- Propose a data distillation procedure to derive knowledge from an LLM of compressing prompts without losing crucial information.
- Approach prompt compression as a token-classification task to capture all essential information needed for compression from bidirectional context while ensuring the faithfulness.



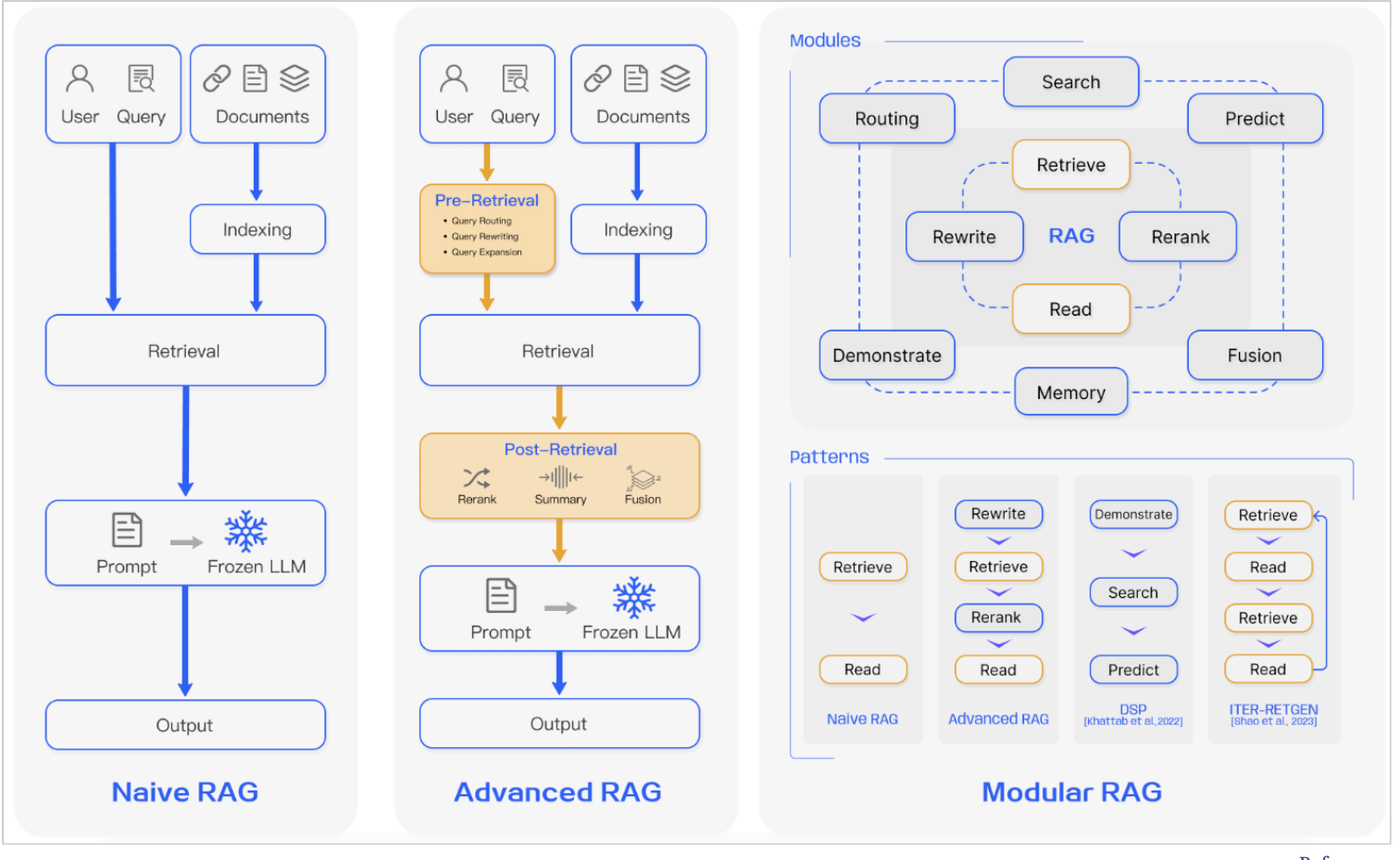
Reference

Retrieval Augmented Generation Workflow



Reference

RAG Techniques



Reference

Fine-tuning

?

When

- Specific domain knowledge
- Specialization: tasks and style

El nostre acompanyament

Laboratori d'Aplicacions de la Intel·ligència Artificial

Identificació del Problema / Solució (I)

Un problema de classificació?

- Similiarity by embeddings
- A specific model: *re-ranker*

```
BAAI/bge-reranker-base | 278M, f32, ~1Gb | Output: {"relevance_score": 0.99853}
```

- Prompting Design

```
template = ""
        Tarea: Clasifica el texto acorde a su departamento ... Proveedores, ...
        Ejemplo:
        Texto: En XX trabajamos con empresas de transporte de primer nivel...
        Departamento: Proveedores
        ...
        Por favor, clasifica el departamento del siguiente texto:
        Texto: {input_text}
        Departamento:
""
```

- Fine-Tuning Design 🎯💥

Identificació del Problema / Solució (II)

Integració de consultes SQL | NoSQL?

- A specific model: *text-to-sql*
- Fine-Tuning with the current DB Schema

Sortida SQL	**Descripció**	**Context**
SELECT COUNT(*) FROM head WHERE age > 56	¿Cuántos jefes de departamento tienen más de 56 años?	CREATE TABLE head (age INTEGER)
SELECT name, born_state, age FROM head ORDER BY age	Lista el nombre, estado de nacimiento y edad de los jefes de departamento ordenados por edad.	CREATE TABLE head (name VARCHAR, born_state VARCHAR, age VARCHAR)

- Agent Architecture: Sensitive data

Identificació del Problema / Solució (III)

Invocació de funcions?

- Prompts ?
- Specific Models: **Function Calling**

```
@tool(parse_docstring=True)
def foo(bar: str, baz: int) -> str:
    """The foo.

    Args:
        bar: The bar.
        baz: The baz.
    """
return bar
```

- Router Agents

Support in Data preparation (I)

¿Qué cantidad puede tomar un adulto?

Posología

	Vía oral o rectal (paracetamol)	Vía parenteral (propacetamol)
Adultos	325-650mg/4-6h ó 500-1000mg/6-8h, máximo 4 g/día	1-2g/4-6h, máximo 8g/día
Niños	10-15 mg/kg/4-6h, máximo 50mg/kg/día	1-2g/4-6h, máximo 8g/día, sólo en mayores de 13 años ó 50 kg

Vía parenteral: el paracetamol se absorbe rápida y casi completamente por vía oral, por lo que la vía parenteral debería reservarse para aquellos pacientes en los que la vía oral no resulta posible. Se puede administrar vía intramuscular profunda, pero es preferible la vía intravenosa en infusión de 100 ml de suero fisiológico.

Support in Data preparation (II)

Contenido A ❌

```
Posología
Vía oral o rectal (paracetamol) Vía parenteral (propacetamol)
325–650mg/4–6h o 500–1000mg/6–8h,
Adultos 1–2g/4–6h, máximo 8g/día
máximo 4 g/día
1–2g/4–6h, máximo 8g/día, sólo en mayores de 13
Niños 10–15 mg/kg/4–6h, máximo 60mg/kg/día
años o 50 kg
```

Contenido B ✅

```
Tratamiento sintomático del dolor leve o moderado y de la fiebre.
 |  | Vía oral o rectal (paracetamol) | Vía parenteral (propacetamol
 | --- | --- | ---
 | Adultos | 325–650mg/4–6h o 500–1000mg/6–8h, máximo 4 g/día | 1–2g/
 | Niños | 10–15 mg/kg/4–6h, máximo 60mg/kg/día | 1–2g/4–6h, máximo 8
```

Support RAG design

Chunk and Indexing Data

- **Chunking techniques**:
  - Tokenizer: unexpected
  - Word and Sentence: [un,expect, ed]
  - Sentence Window
  - Semantic (Embedding Model ⚙️)
  - Semantic Double-Pass Merging
  - ...

```
The blue, the red. Football and Tennis. Gray and white.

Semantic Chunking
Chunk #1: The blue, the red.
Chunk #2: Football and Tennis.
Chunk #3: Gray and white.

Semantic Double-Pass Merging
Chunk #1: The blue, the red. Gray and white.
Chunk #2: Football and Tennis
```

- **Metadata Attachments**: pages, sections, timestamps, ...

Support RAG design

Information Retrieval

- Pre-retrieval:
  - Query expansion:
    - Semantic Expansions: GPT-2, BERT, ...
    - Stemming & Lemmatization: Porter Stemmer, KoKiwi, or KoOKT.
    - Synonims or Ontologies: WordNet, ConceptNet,...
  - Decomposition:
    - Intent improvement: LLMs or Semantic
- Retrieval
  - Format: Text, Tabular, Knowledge Graph, Code
  - Method: Iterative, Recursive and Adaptive
- Post-retrieval:
  - Passage Augmenter (with LLMs)
  - Reranker
  - Filter
  - Compressor

Support RAG design

Evaluation

- **Aspects**: Context Relevance, Faithfulness Answer Relevance, Noise Robustness, Negative Rejection, Information Integration, Counterfactual Robustness,...
- **Test data preparation** under specific task: dialogue generation, code search, summarization, ...
- Metric interpretations:
  - Retrieval:
    - **QA**: EM and F1 metrics,
    - **Quality**: Hit Rate, MRR (Mean Reciprocal Rank), NDCG (Normalized Discounted Cumulative Gain)
  - N-Gram Based: BLEU and ROUGE, Meteor,
  - Sem Score (Semantic Similarity)
  - G-Eval (Coherence, Consistency, Fluency, Relevance)
  - ...

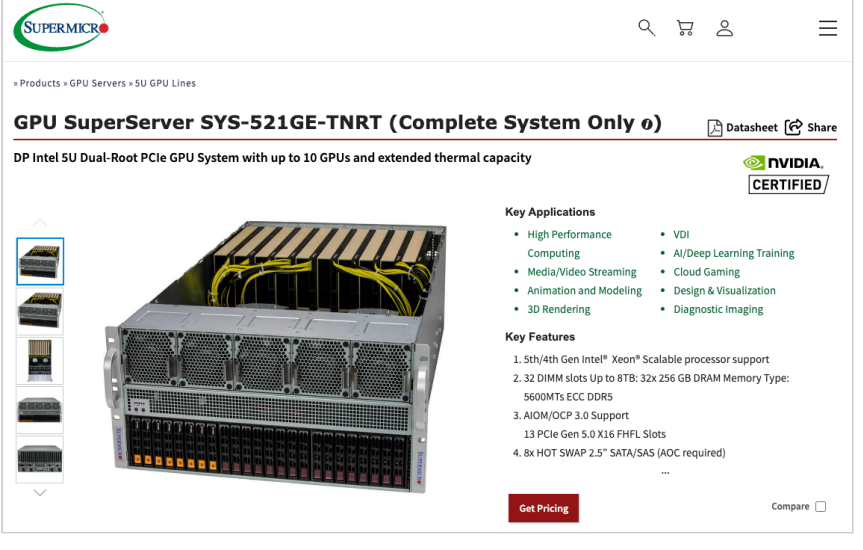
Fine-tuning

Preparation and Evaluation

- Data preparation
- Fine-tunning the model
- Evaluation

Architecture Stack and Infrastructure Recommendation

- On cloud: hosting the model
- On-premise



Gràcies

III Jornada
Networking per a la Transferència d'Aplicacions de la IA a les Empreses i la Societat de les Illes Balears (JAIA2024+)

Fundació Universitat Empresa de les Illes Balears

laia.uib.cat

🥂