datacraft*

Gen AI in prod

tech & product insights for delivering value with RAG

François PAUPIER - September 18, 2024

CC BY NC SA

Agenda

1. Product & strategy

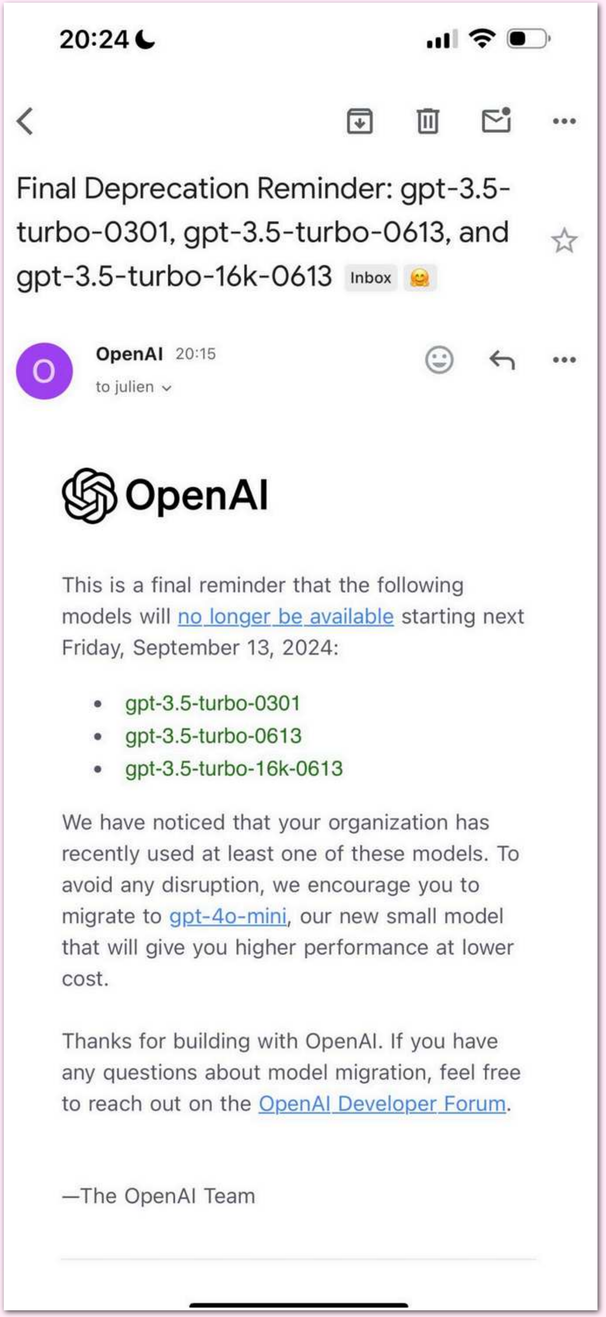
1. GenAI landscape
2. Strategic considerations
3. Business value and differentiation
4. User-centric approach

2. Technical implementation

1. RAG pipeline overview
2. Where to focus your eng. efforts
3. Fine-tuning considerations
4. *Pot-pourri* of RAG protips

CC BY NC SA

1. Product & strategy



20:24

Final Deprecation Reminder: gpt-3.5-turbo-0301, gpt-3.5-turbo-0613, and gpt-3.5-turbo-16k-0613 Inbox

OpenAI 20:15
to julien

OpenAI

This is a final reminder that the following models will no longer be available starting next Friday, September 13, 2024:

- gpt-3.5-turbo-0301
- gpt-3.5-turbo-0613
- gpt-3.5-turbo-16k-0613

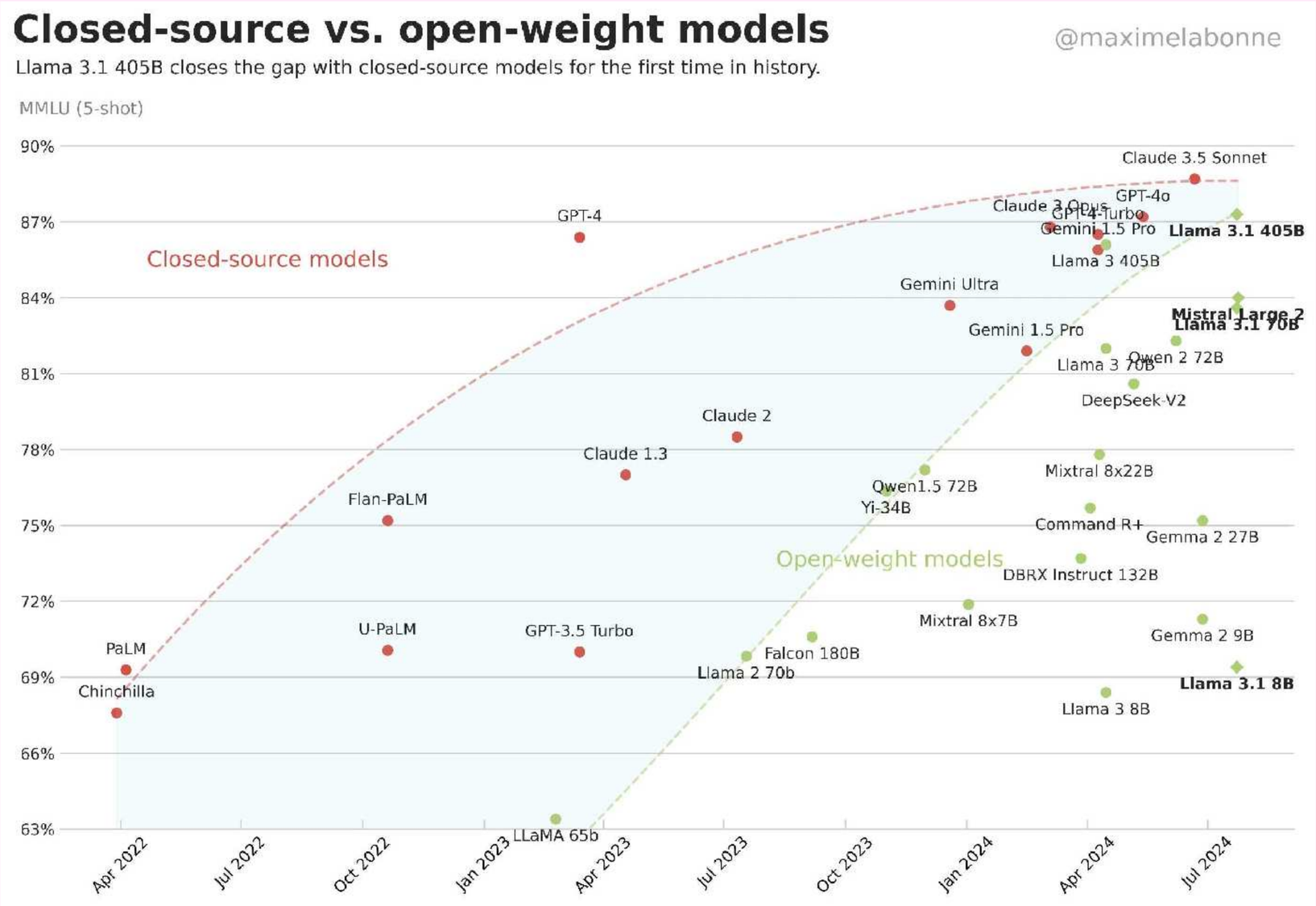
We have noticed that your organization has recently used at least one of these models. To avoid any disruption, we encourage you to migrate to gpt-4o-mini, our new small model that will give you higher performance at lower cost.

Thanks for building with OpenAI. If you have any questions about model migration, feel free to reach out on the OpenAI Developer Forum.

—The OpenAI Team

CC BY NC SA

Better models, and *open-weights* catching up with *closed-source* ones



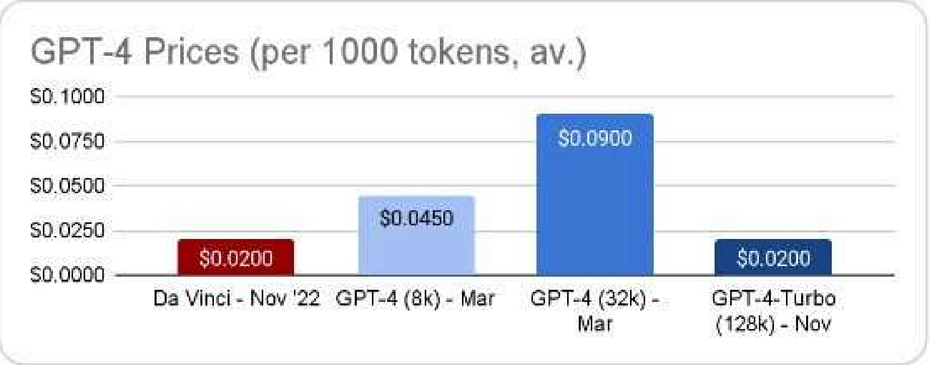
Source: Maxime Labonne, https://mlabonne.github.io/blog, July 2024

CC BY NC SA

Better models, and lower cost over time - example of OpenAI

8 months after GPT 3.5 API was launched:

- GPT 3.5 is **90% cheaper** than Da Vinci
- GPT 3.5 with 16k context got **55% cheaper in 5 months**



8 months after GPT-4 API was launched:

- GPT-4-Turbo is now **55% cheaper than GPT-4-8k**
- GPT-4-Turbo is now **75% cheaper than GPT-4-32k**
- GPT-4-Turbo has **128k context window** and is the **same price is Da Vinci** (older, less performant model)

7 months after Ada was launched:

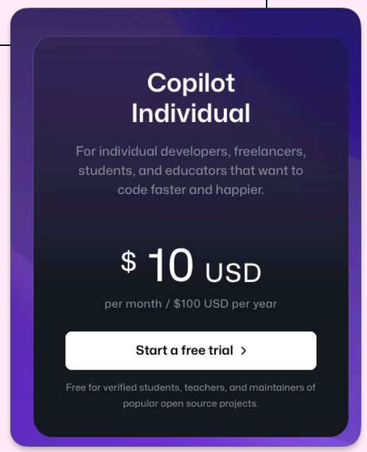
- Ada Embeddings model is now **75% cheaper**
- It would cost now **only ~$1500** to generate embeddings for the **whole of Wikipedia**

Source: Peter Gostev, Head of AI @ Moonpig, link , summer 2023

CC BY NC SA

A quick note on pricing for GenAI services

Successful approaches	Considerations
Per-seat pricing (*e.g.,* GitHub Copilot, NotionAI)	Avoid complex per-token pricing (expect for products designed for engineers)
Per business output pricing (#blog articles created)	Implement fair use policies instead of strict usage limits (reserves yourself a right to throttle)
Simple, flat-rate models	Balance simplicity with value-based pricing



➡ Prioritize **simplicity in pricing** to **drive adoption** and **align with value** delivered

CC BY NC SA

Back of the envelope estimation for GenAI product cost

OpenAI & other LLM API Pricing Calculator

Calculate and compare the cost of using OpenAI, Azure, Anthropic Claude, Llama 3, Google Gemini, Mistral, and Cohere LLM APIs for your AI project with our simple and powerful free calculator. Latest numbers as of September 2024.

Input Tokens: 1000 | Output Tokens: 500 | API Calls: 100

Calculate by: Tokens | Words | Characters

Pricing Calculations

The following pricing calculations are based on the input tokens, output tokens, and API calls you have entered above.

Provider	Model	Context	Input/1k Tokens	Output/1k Tokens	Per Call	Total
Chat/Completion Models						
OpenAI	GPT-4o-2024-08-06	128K/16K	$0.0025	$0.01	$0.0075	$0.75
OpenAI	GPT-4o	128K/4K	$0.005	$0.015	$0.0125	$1.25
OpenAI	GPT-4o mini	128K/16K	$0.00015	$0.0006	$0.0004	$0.04
OpenAI	o1-preview	128K/32K	$0.015	$0.06	$0.0450	$4.50

Source: OpenAI & Other LLM API pricing calculator

LLM as a service

Substratus — Home Contact Blog

Posts

- Deploying Faster Whisper on Kubernetes
- Introducing KubeAI: Open AI on Kubernetes
- What GPUs can run Llama 3.1 405B?
- Learn how to benchmark vLLM to optimize for speed
- Private RAG with Lingo, Verba and Weaviate
- Deploying Mixtral on GKE with 2 x L4 GPUs
- Calculating GPU memory for serving LLMs
- Deploying Mistral 7B Instruct on K8s using TGI
- The K8s YAML dataset
- Tutorial: K8s Kind with GPUs
- Converting HuggingFace Models to GGUF/GGML
- A Kind Local Llama on K8s

Calculating GPU memory for serving LLMs

November 16, 2023 by Sam Stoelinga

How many GPUs do I need to be able to serve Llama 70B? In order to answer that, you need to know how much GPU memory will be required by the Large Language Model.

The formula is simple:

$$M = \frac{(P * 4B)}{(32/Q)} * 1.2$$

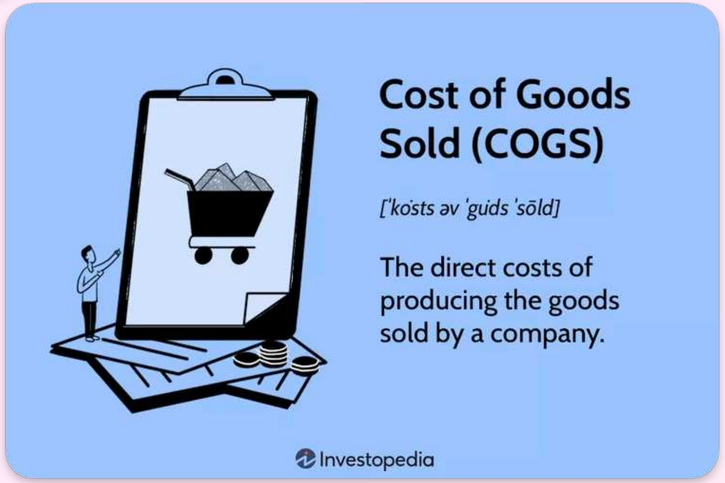
Symbol	Description
M	GPU memory expressed in Gigabyte
P	The amount of parameters in the model. E.g. a 7B model has 7 billion parameters.
4B	4 bytes, expressing the bytes used for each parameter
32	There are 32 bits in 4 bytes
Q	The amount of bits that should be used for loading the model. E.g. 16 bits, 8 bits or 4 bits.
1.2	Represents a 20% overhead of loading additional things in GPU memory.

Looking to deploy LLMs on Kubernetes? Check out KubeAI, providing private Open AI on Kubernetes.

Now let's try out some examples.

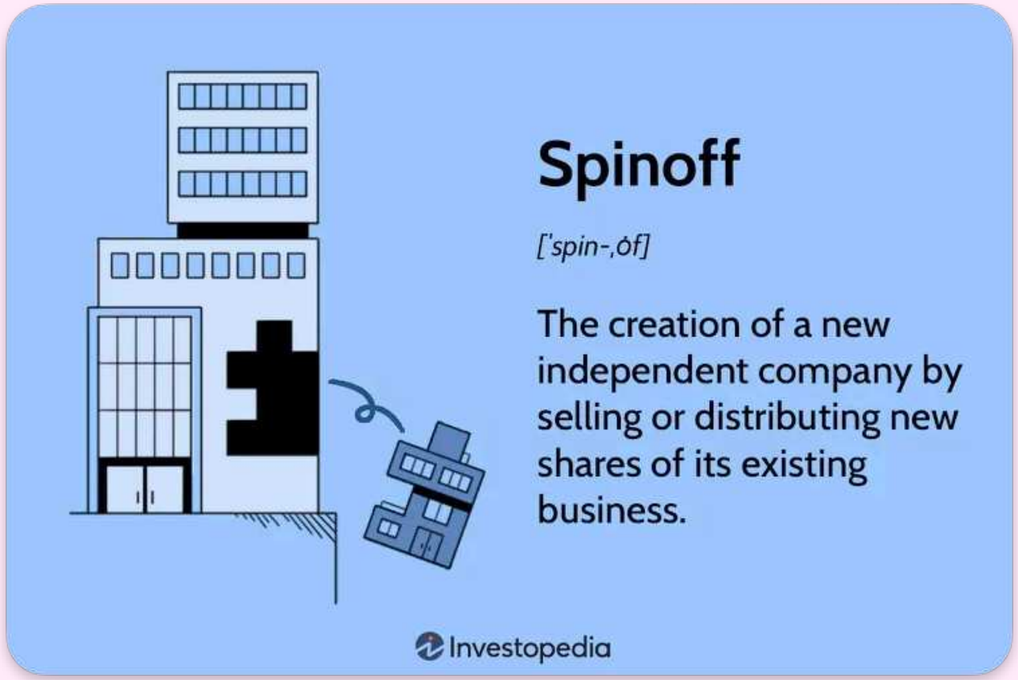
Source: Calculating GPU memory for serving LLMs, Nov 16, 2023

Self hosted



CC BY NC SA

For corporate - Exploring GenAI through a spinoff

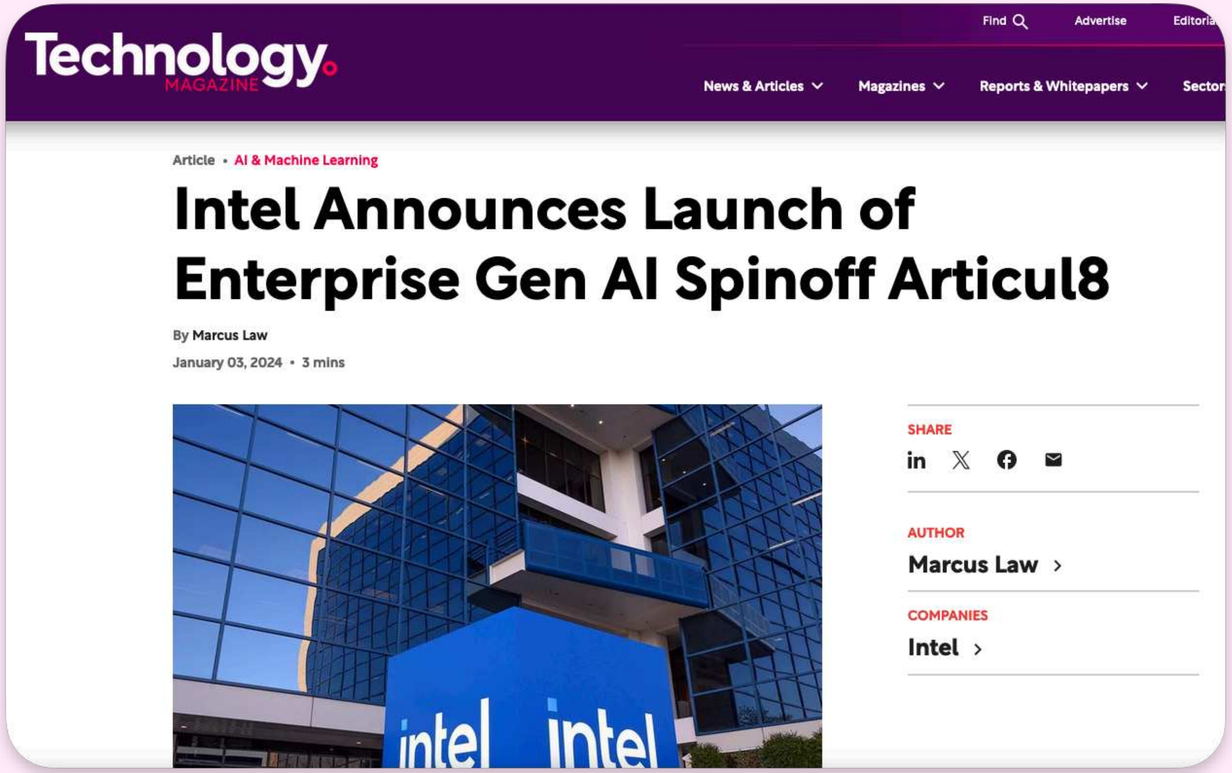


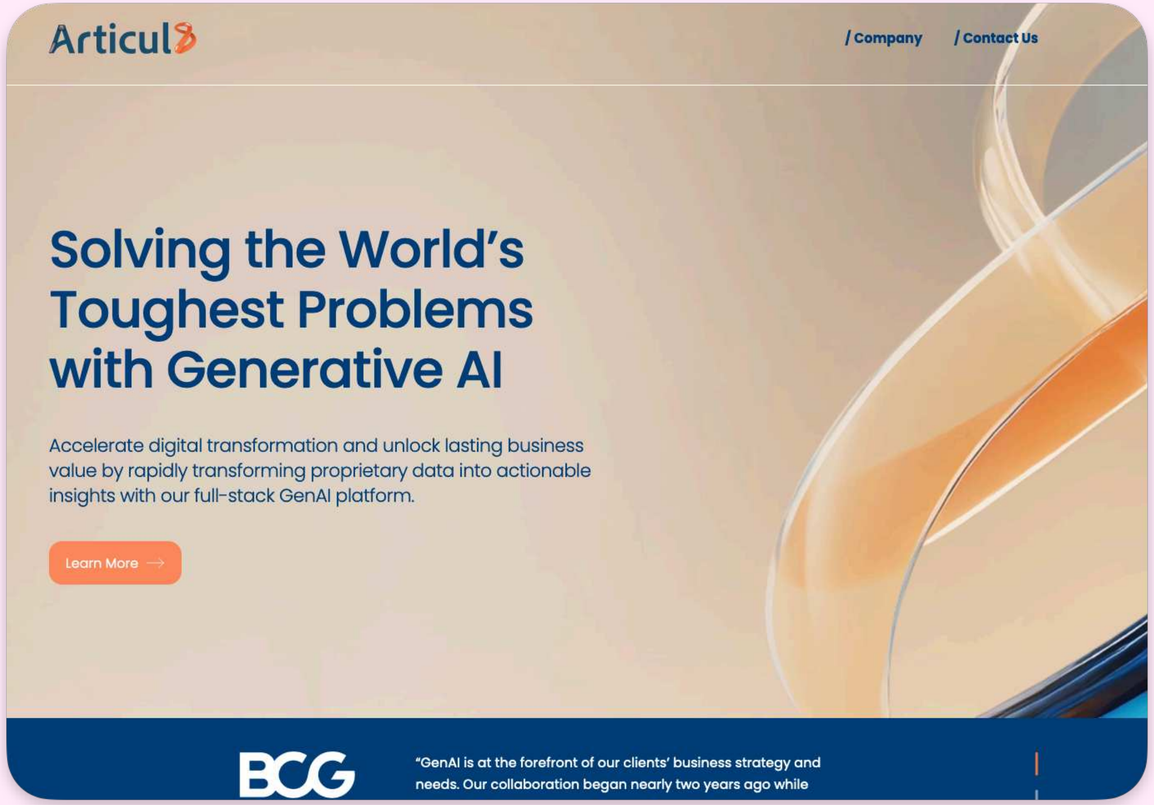
Source: https://www.investopedia.com/terms/s/spinoff.asp

- Escape corporate friction
- Limit reputation risk
  (*e.g.*, hallucinations, buggy product)
- Incentivize entrepreneurial team
- Access new clients & revenue stream

CC BY NC SA

Example of Intel and its GenAI spinoff in Jan. 2024





Source: https://www.articul8.ai/

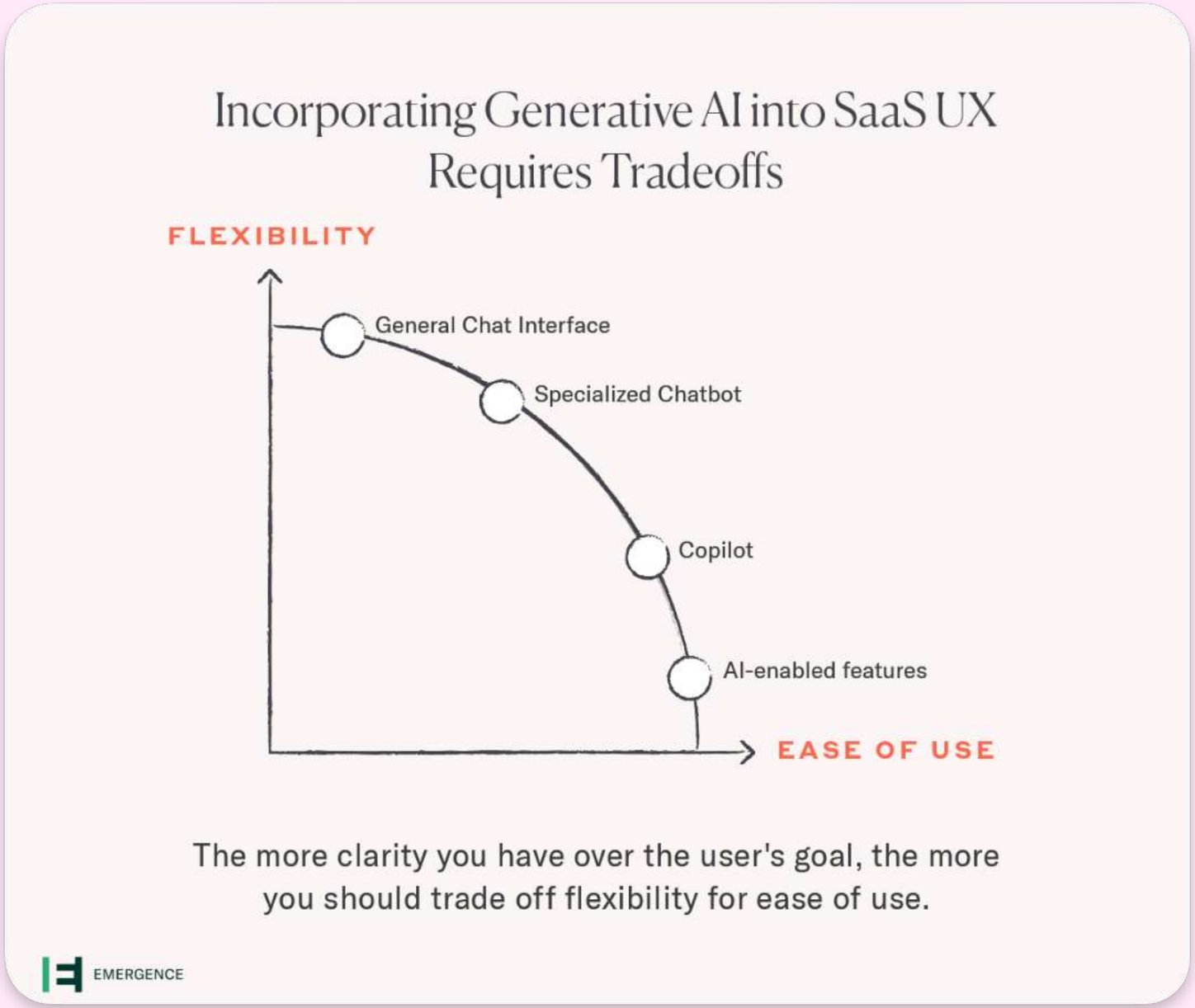
CC BY NC SA

Business value and differentiation

- Combining business expertise with augmented LLM workflows
- Assessing AI-driven value creation
- Avoiding "ChatGPT wrapper" syndrome

The more specific your industry and context, the better your model and product.

CC BY NC SA



Source: Jake Saper, J. C. (2023). *Your Success with Generative AI May Come Down to These UX Decisions*. https://www.emcap.com/thoughts/your-success-with-generative-ai-may-come-down-to-these-ux-decisions/

Build your moat 🏰

1. Integrate off the shelf model to capture domain specific interactions
2. Correlate model suggestions with users actions
3. Tune model / UX / features to deliver domain specific expertise
4. Deliver exclusive value a generic model can’t offer

CC BY NC SA

User centric approach - understand usage to design better products

- Logging and analysing user interactions
- Build yourself a cheap tool to view / annotate interaction data
  Data flywheel: continuously improving a ML system
- Consider unsupervised learning to identify critical topics
  Topic extraction, to discover topics in user queries

CC BY NC SA

What is this ?…

CC BY NC SA

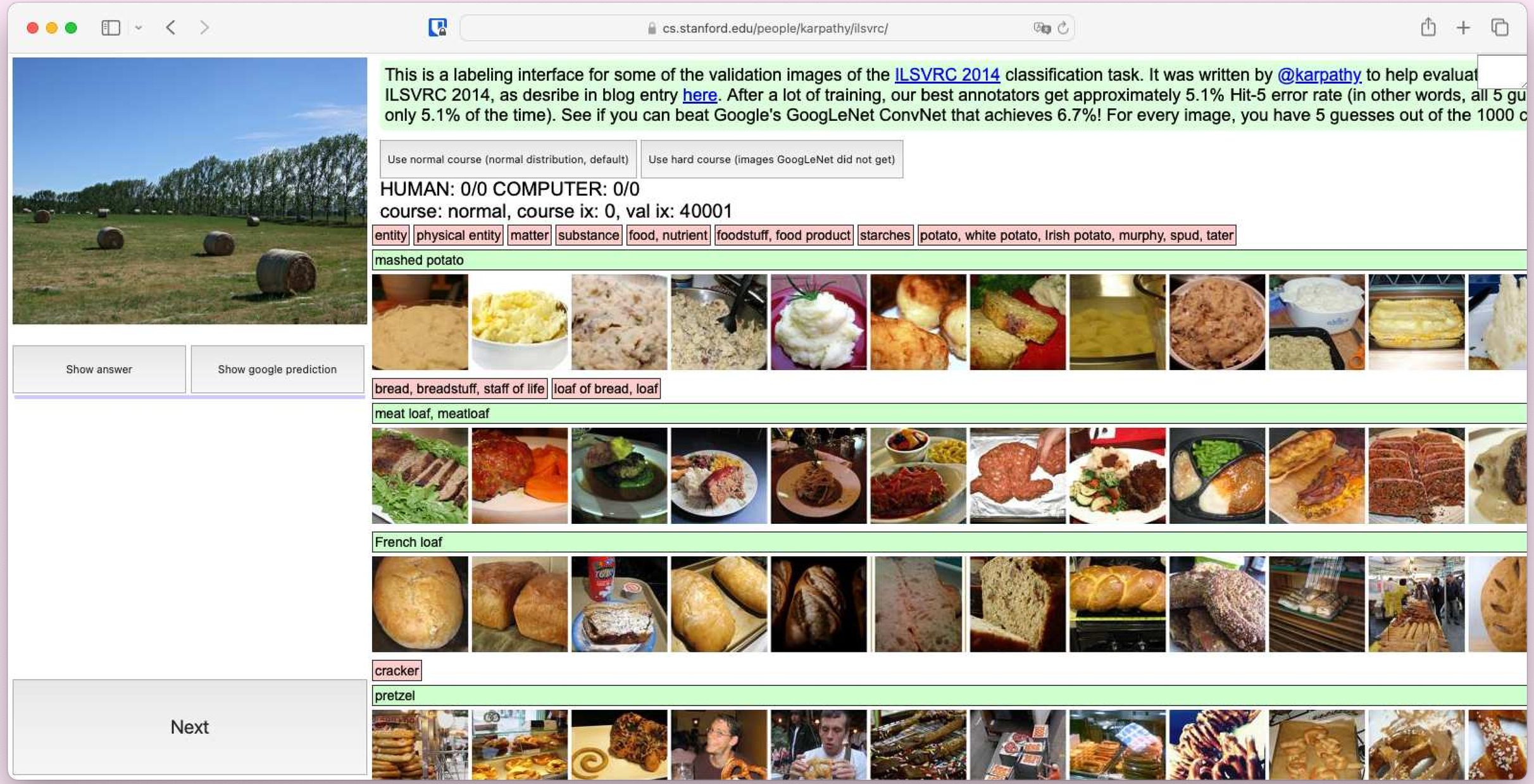
Labeling Interface

I developed a labeling interface that would help us evaluate the human performance. It looked similar to, but not identical, to the screenshot below:

A crop of a screenshot of the labeling interface for the ILSVRC validation data. Try it out for yourself.

Source: Andrej Karpathy What I learned from competing against ConvNet on ImageNet, Sep 2 2014

CC BY NC SA



Source: https://cs.stanford.edu/people/karpathy/ilsvrc/

CC BY NC SA

What Karpathy learned from competing against a ConvNet on ImageNet

It took a while. I ended up training on 500 validation images and then switched to the test set of 1500 images. The labeling happened at a rate of about 1 per minute, but this decreased over time. I only enjoyed the first ~200, and the rest I only did *#forscience*. (In the end we convinced one more expert labeler to spend a few hours on the annotations, but they only got up to 280 images, with less training, and only got to about 12%). The labeling time distribution was strongly bimodal: Some images are easily recognized, while some images (such as those of fine-grained breeds of dogs, birds, or monkeys) can require multiple minutes of concentrated effort. I became very good at identifying breeds of dogs.

It was worth it. Based on the sample of images I worked on, the GoogLeNet classification error turned out to be 6.8% (the error on the full test set of 100,000 images is 6.7%). My own error in the end turned out to be **5.1%**, approximately 1.7% better. If you crunch through the statistical significance calculations (i.e. comparing the two proportions with a Z-test) under the null hypothesis of them being equal, you get a one-sided p-value of 0.022. In other words, the result is statistically significant based on a relatively commonly used threshold of 0.05. Lastly, I found the experience to be quite educational: After seeing so many images, issues, and ConvNet predictions you start to develop a really good model of the failure modes.

Source: Andrej Karpathy What I learned from competing against ConvNet on ImageNet, Sep 2 2014

CC BY NC SA

Andrew Ng also emphasised critical role of error analysis

Source: Andrew Ng, Carrying Out Error Analysis (C3W2L01) DeepLearningAI course

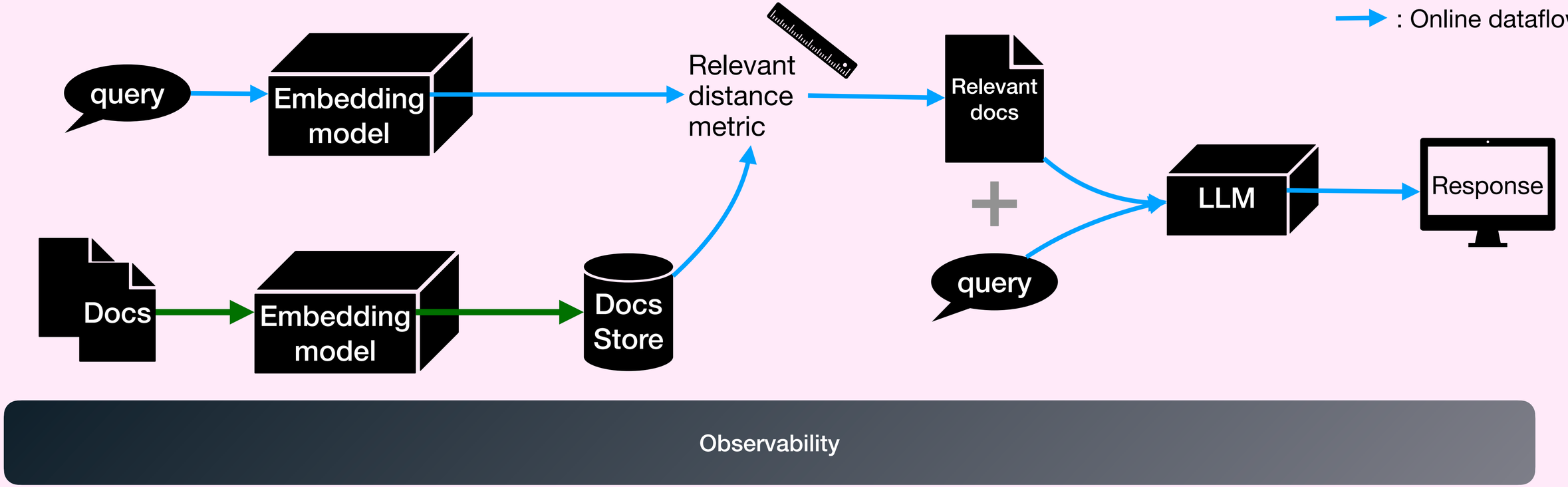
CC BY NC SA

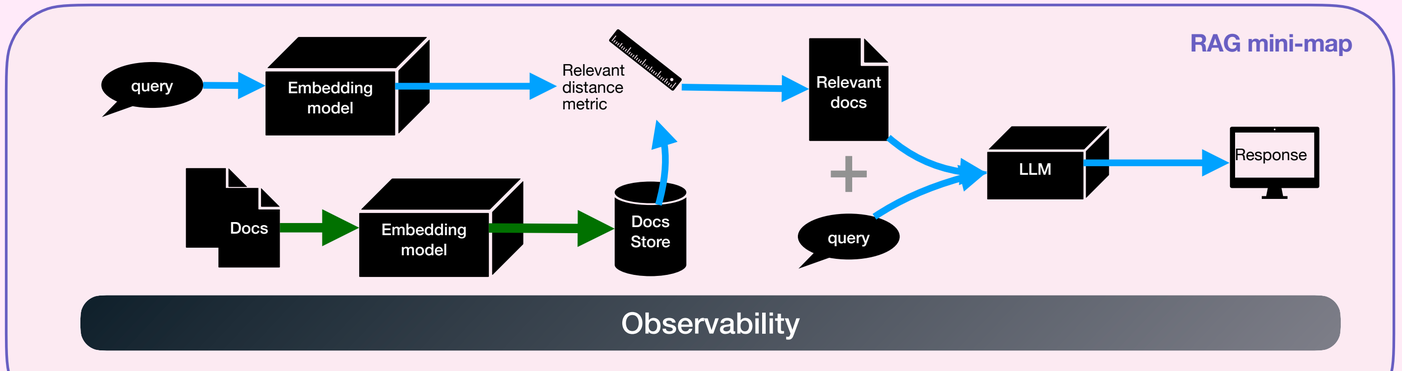
2. Technical implementation

Legend

: Offline dataflow

: Online dataflow

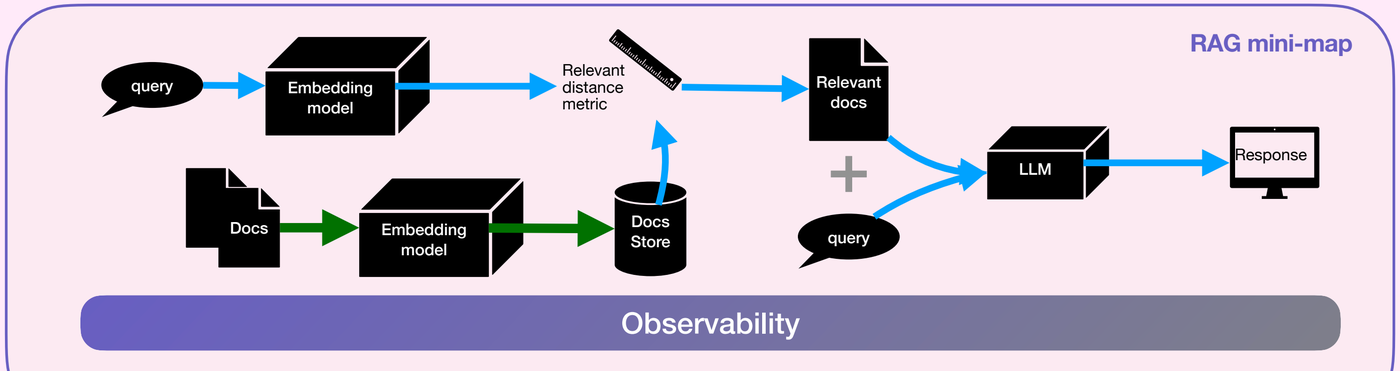




CC BY NC SA

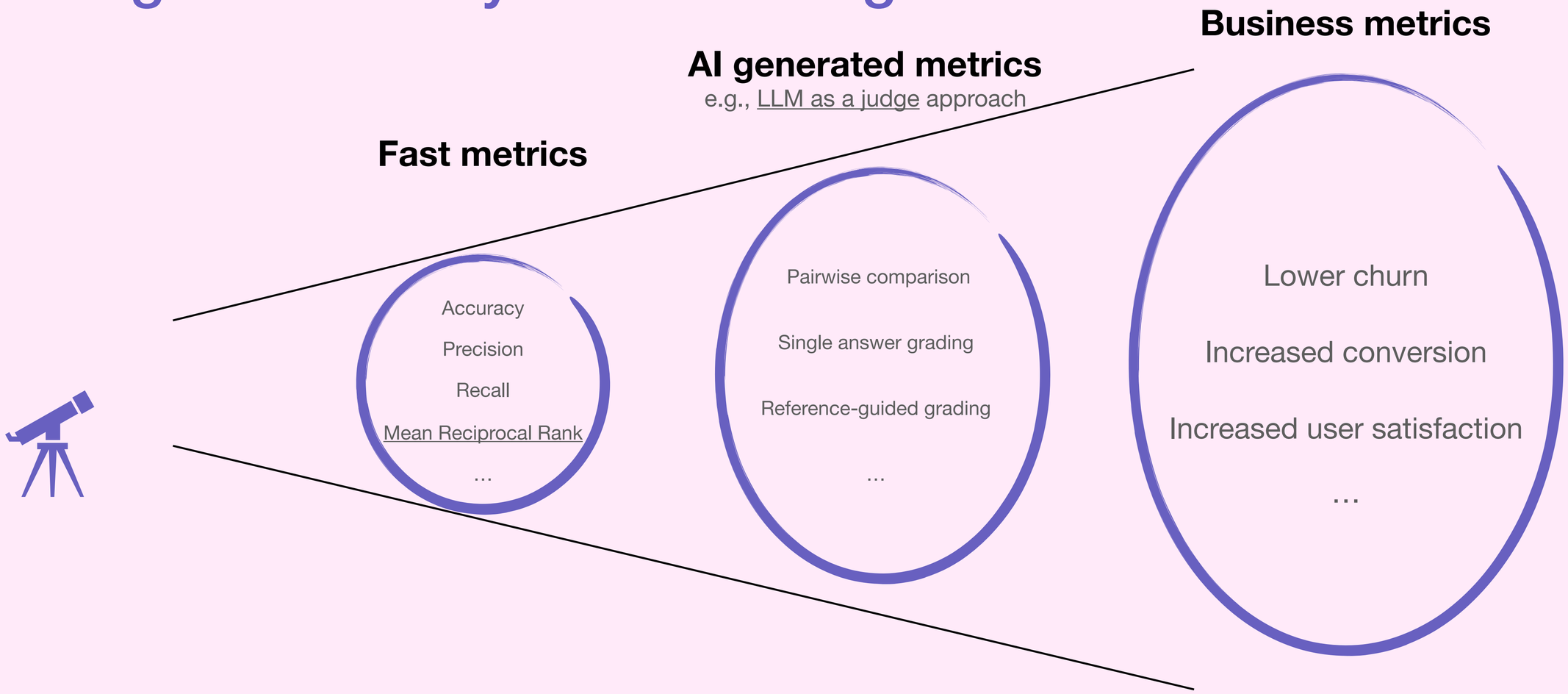
If you can’t measure it, you can’t improve it.

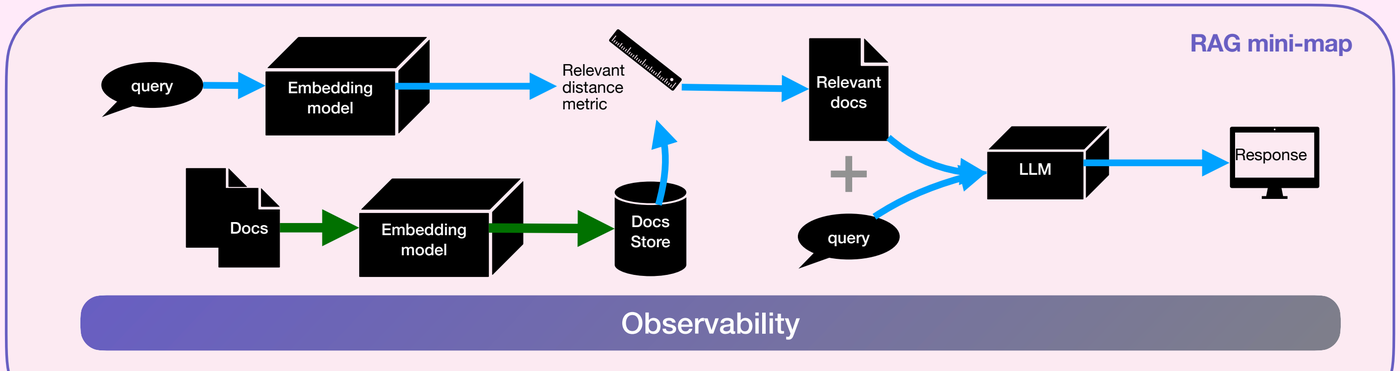
- Peter Drucker



CC BY NC SA

The long shot - from your business goal back to automated metrics





CC BY NC SA

Finding where to focus your engineering efforts

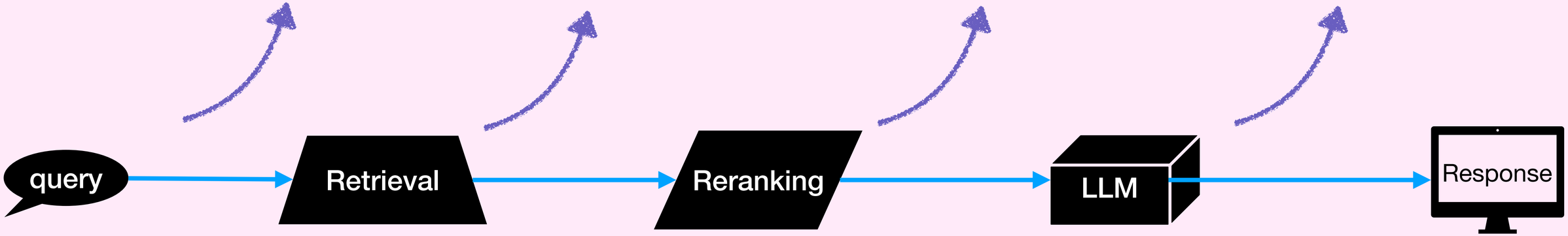
- minimal RAG eval

Did we had the expected piece of information in the first place ?

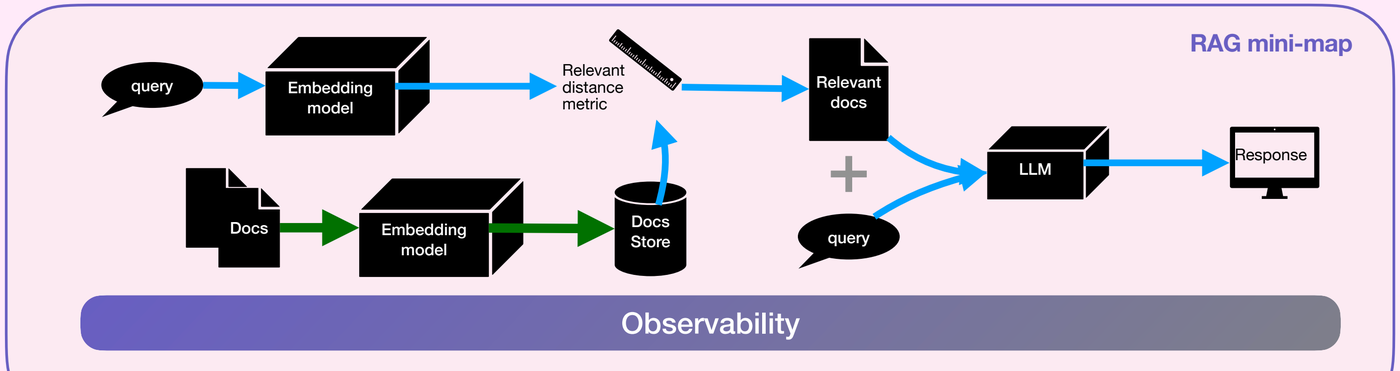
Is the retrieval system finding the info ?

Is the re-ranking priorizing correctly ?

Is the text generation working fine ?



Start measuring from left to right



CC BY NC SA

Maximise chance to reproduce exps. by logging all the way

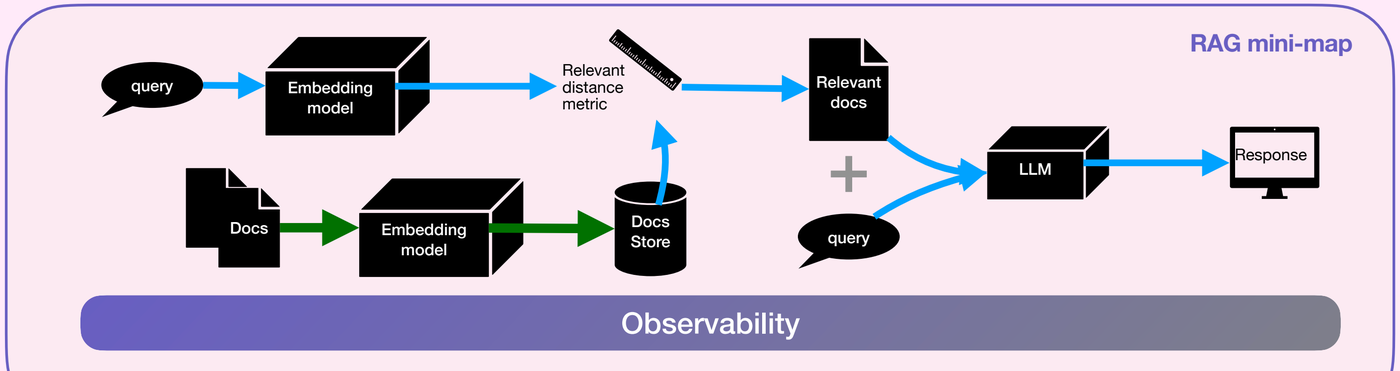
- *query*
- *query metadata*

- *Chunking mechanism*
- *Embedding version*
- *max sequence length*
  ...

- *Reranker version*
- *Reranker config (batch size, ...)*

- *LLM version*
- *Hyperparams (Temp., topP, topK,..)*

- *Response*
- *Context presented to the user*



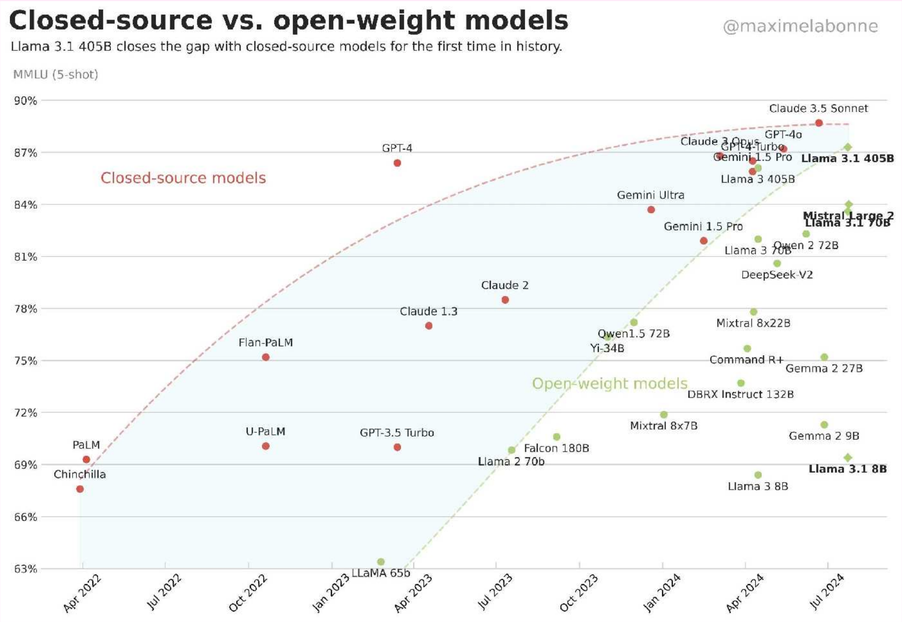
CC BY NC SA

Considerations on LLM finetuning 1/3

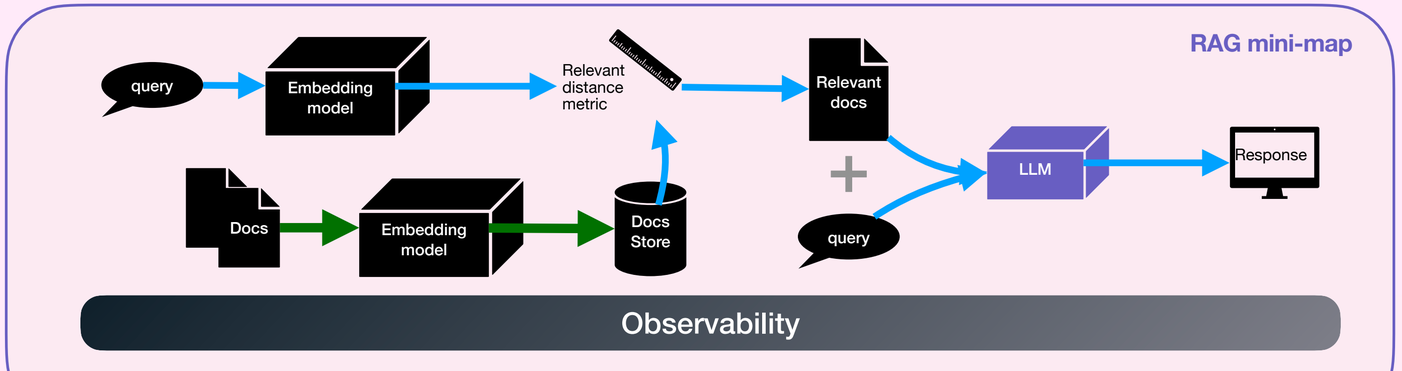
Reminder

Technique	Cost
Zero shot *'do this'*	$
Few shot *'do this, here are some examples'*	$$
Fine tuning **(Supervised / Preference alignment)**	$$$$
Train foundation model	$$$$x10e^6

Best bang for the buck



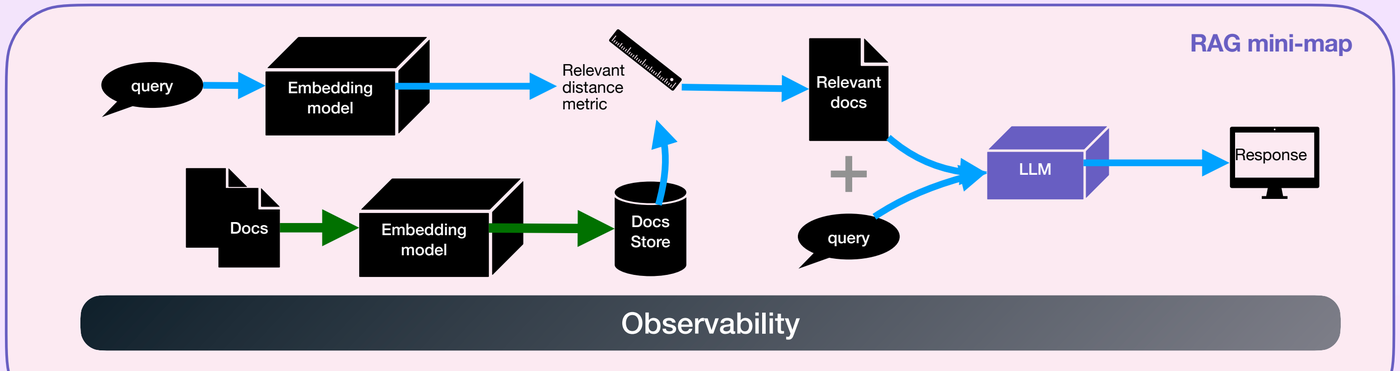
Source: Maxime Labonne, https://mlabonne.github.io/blog, July 2024



CC BY NC SA

Paying a premium for a top tier model often solves the performance problem

Focus precious engineering time to solve *your* specific business problem



CC BY NC SA

Considerations on LLM finetuning 3/3

When few shots learning does not work

When LLM-as-a-service don't work well

When deploying on device

Great community support

Lower latency requirements

Narrow problem - specific task

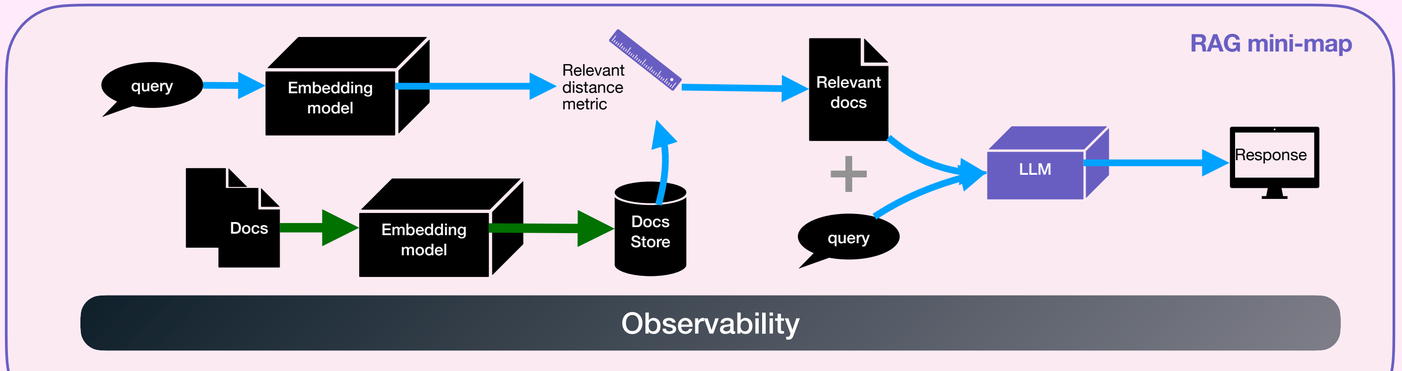
Data distribution never seen on public dataset

When prompt engineering is impractical

Data privacy

Clearly defined loss function

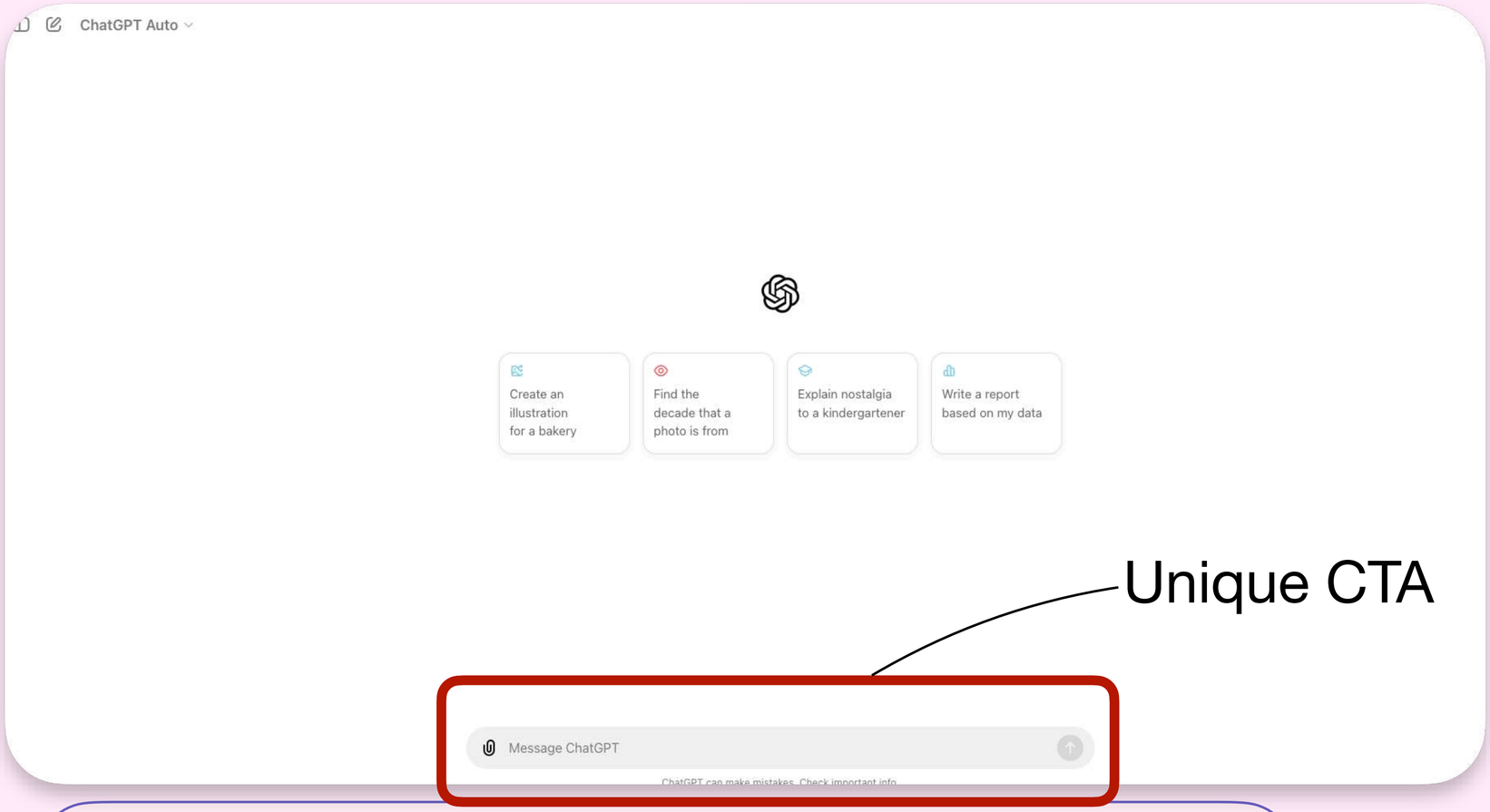
When is fine tuning **relevant** ?

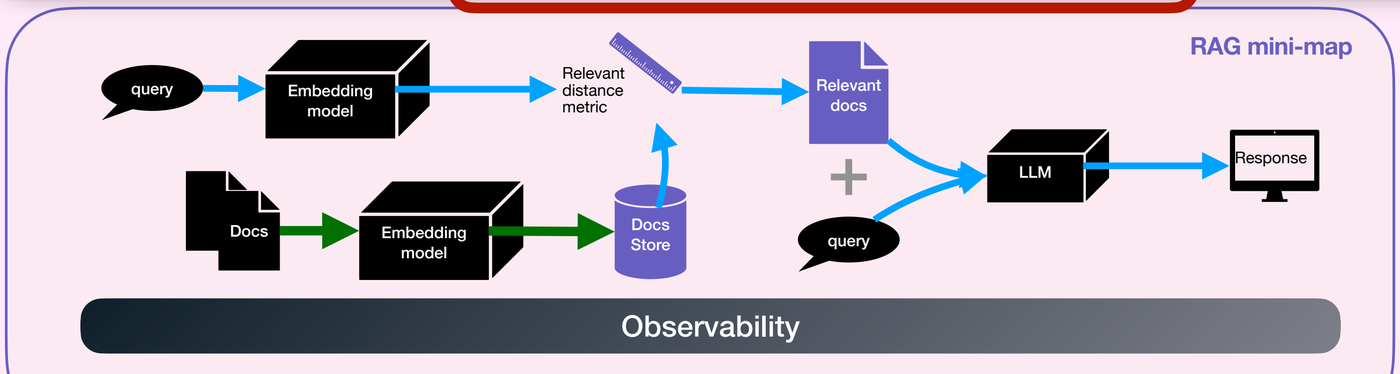


CC BY NC SA

Pot-pourri of RAG protips

Retrieval

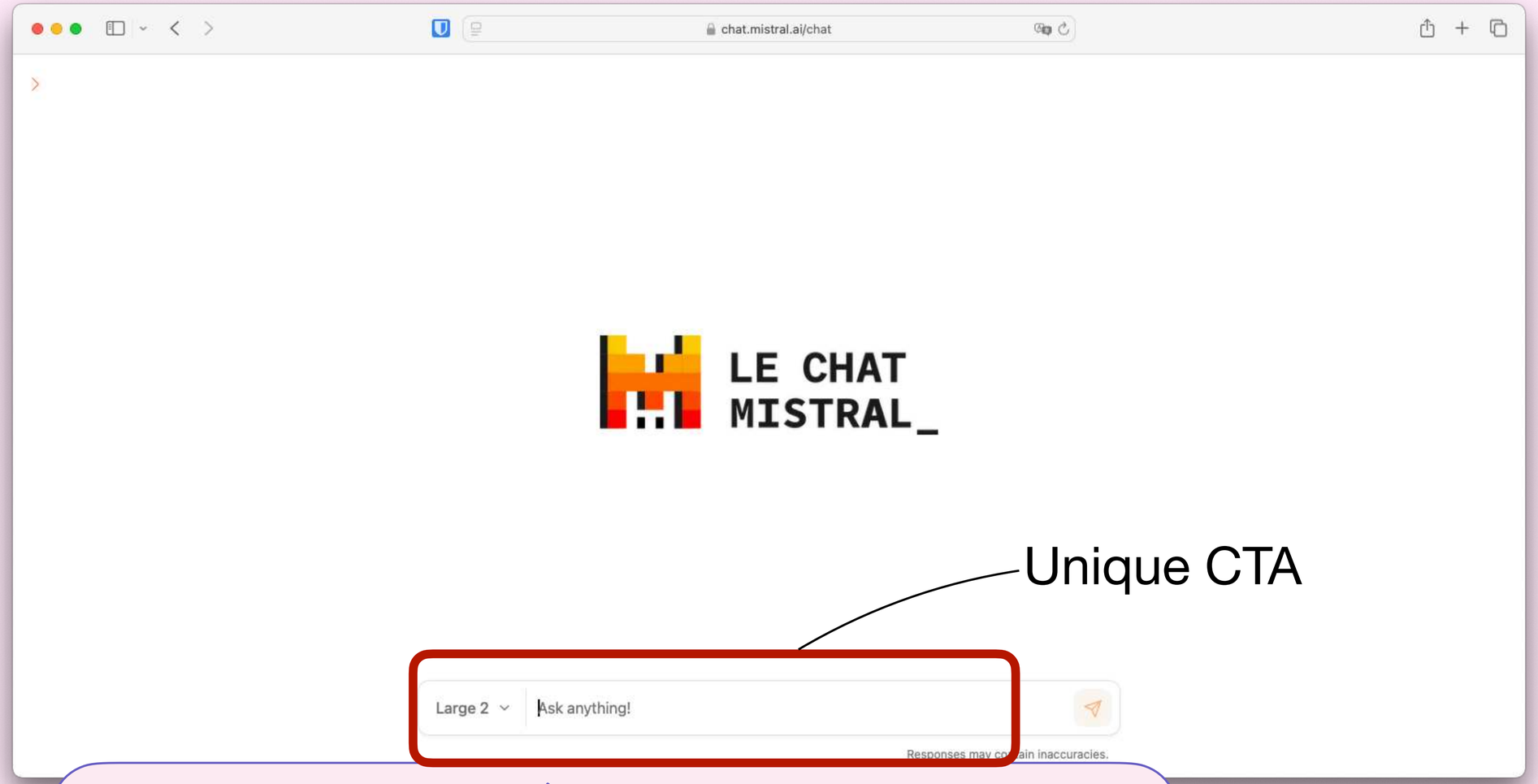




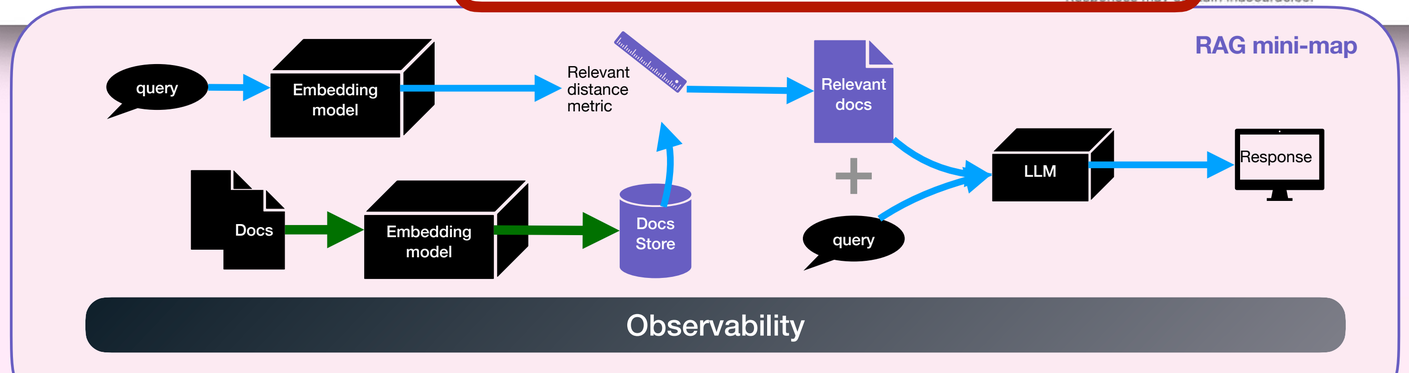
CC BY NC SA

Pot-pourri of RAG protips

Retrieval



Unique CTA



CC BY NC SA

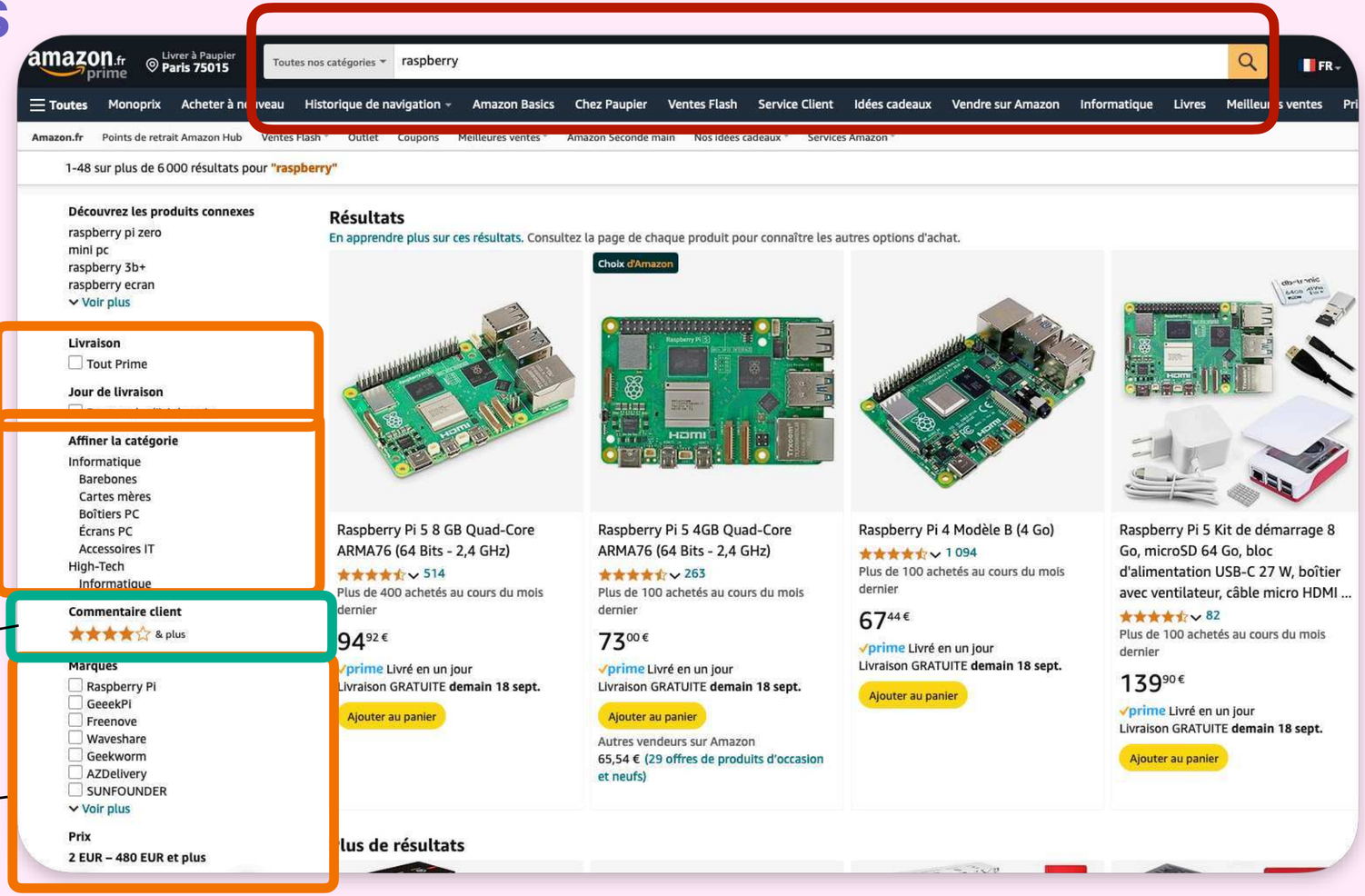
Pot-pourri of RAG protips

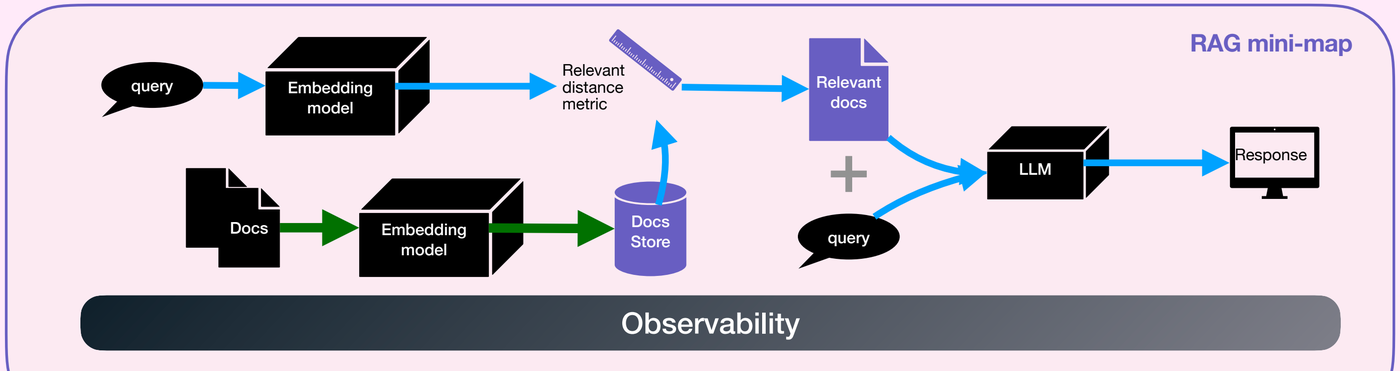
Retrieval

Delivery experience filter

Inventory product metadata

Clients feedback data

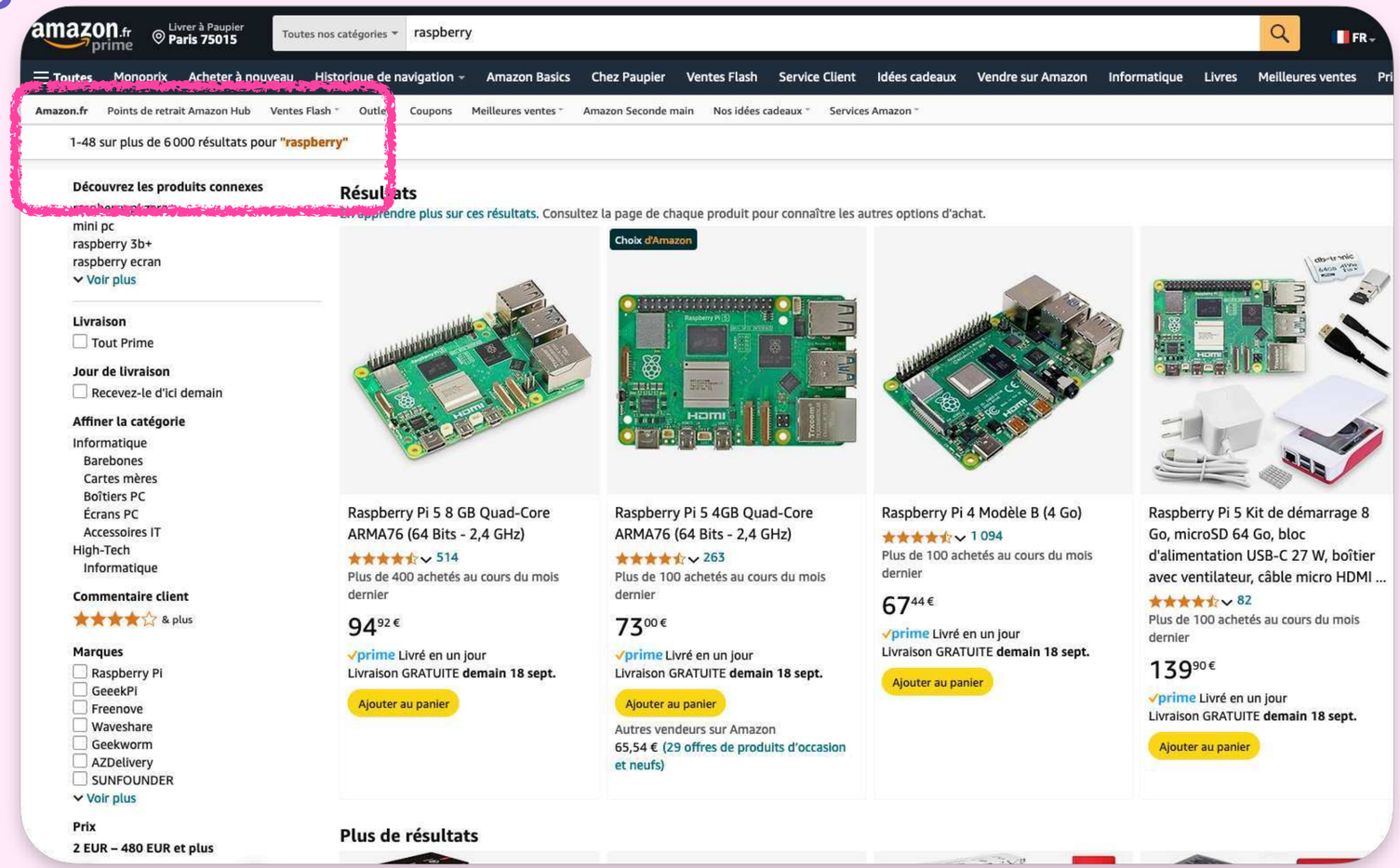


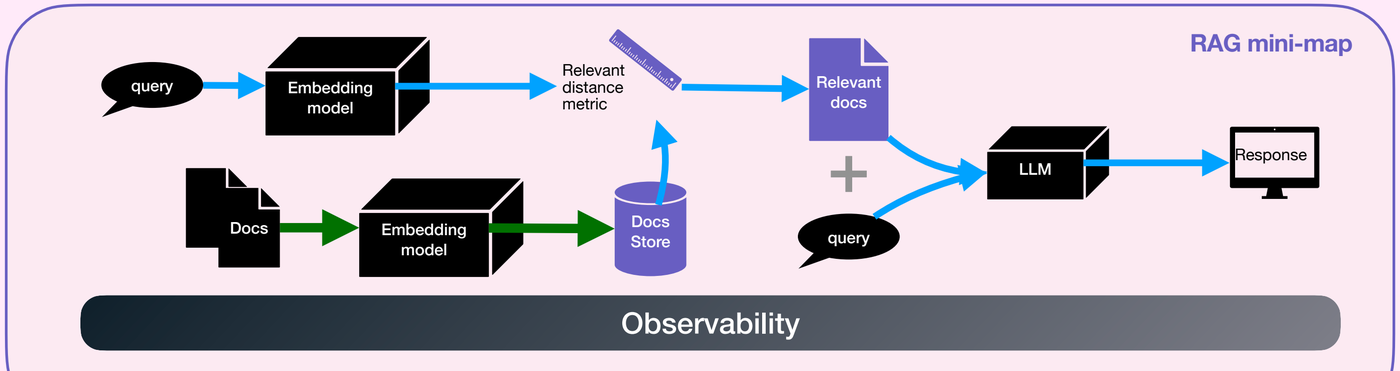


CC BY NC SA

Pot-pourri of RAG protips

Retrieval

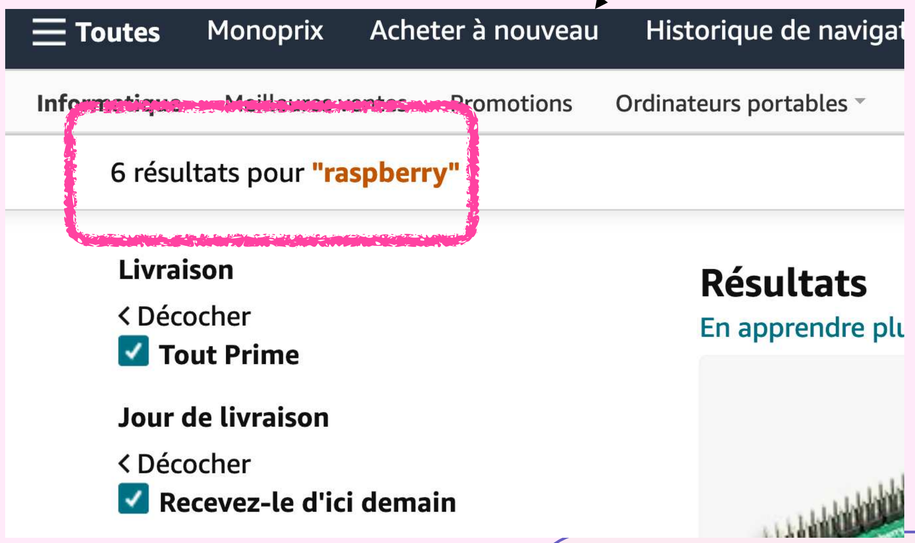


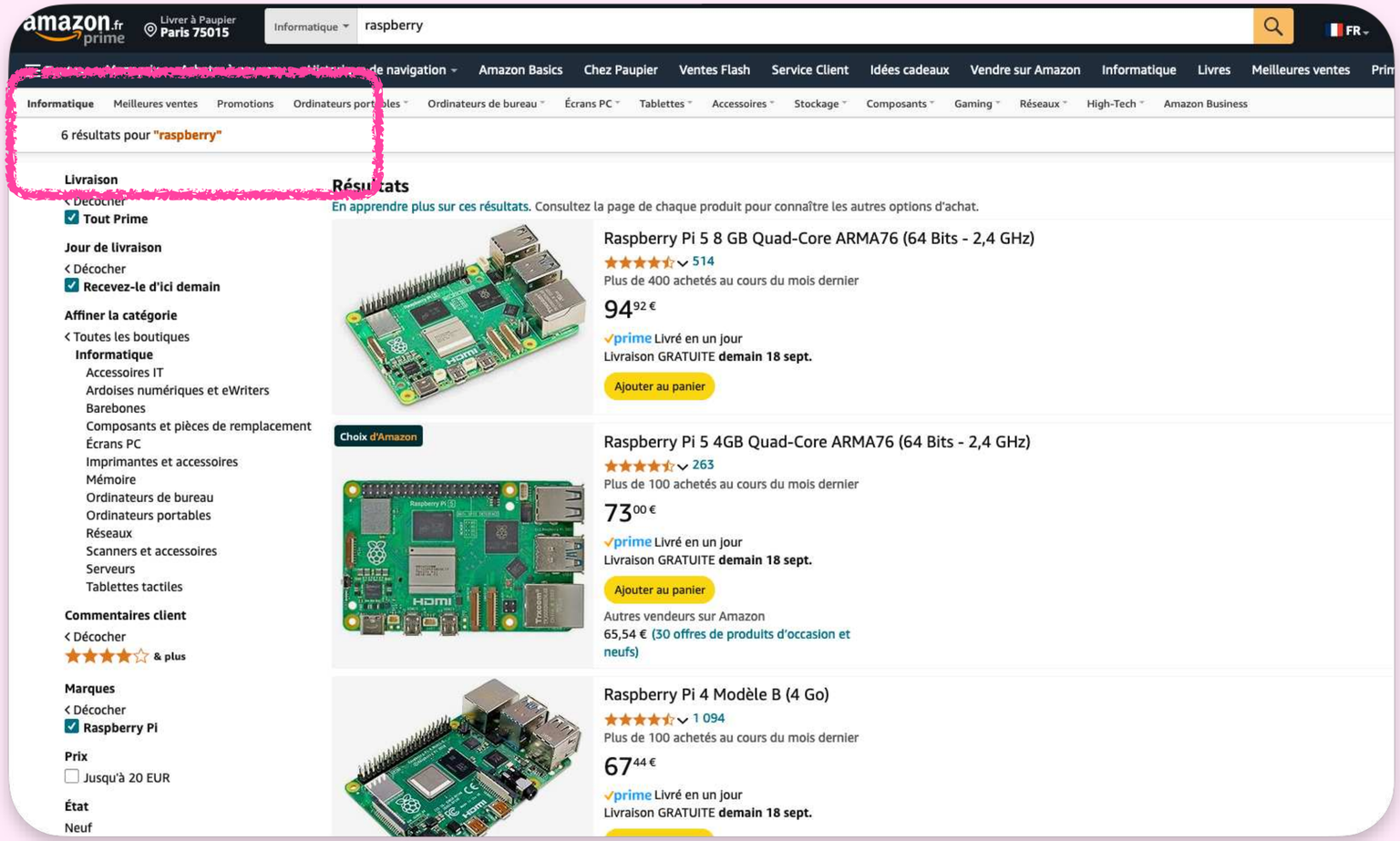


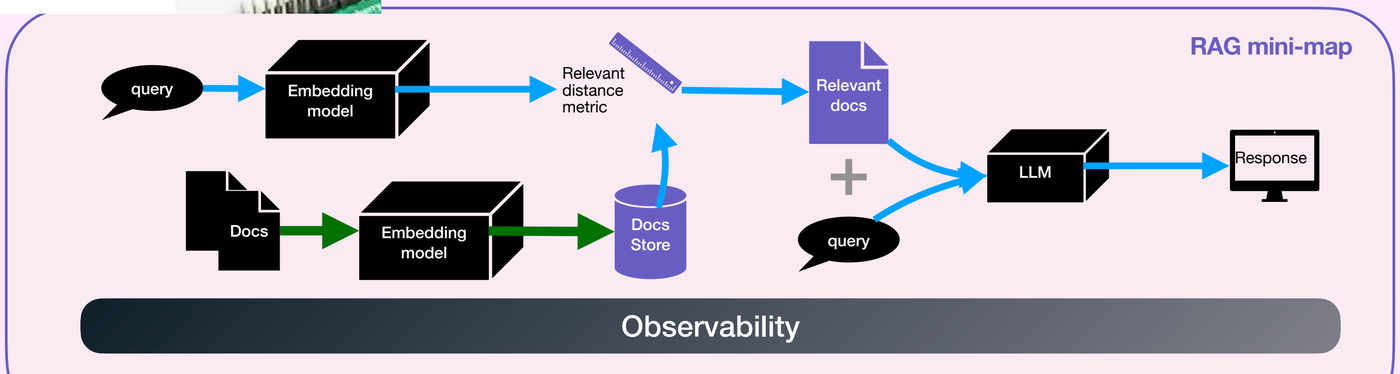
CC BY NC SA

Pot-pourri of RAG protips

Retrieval





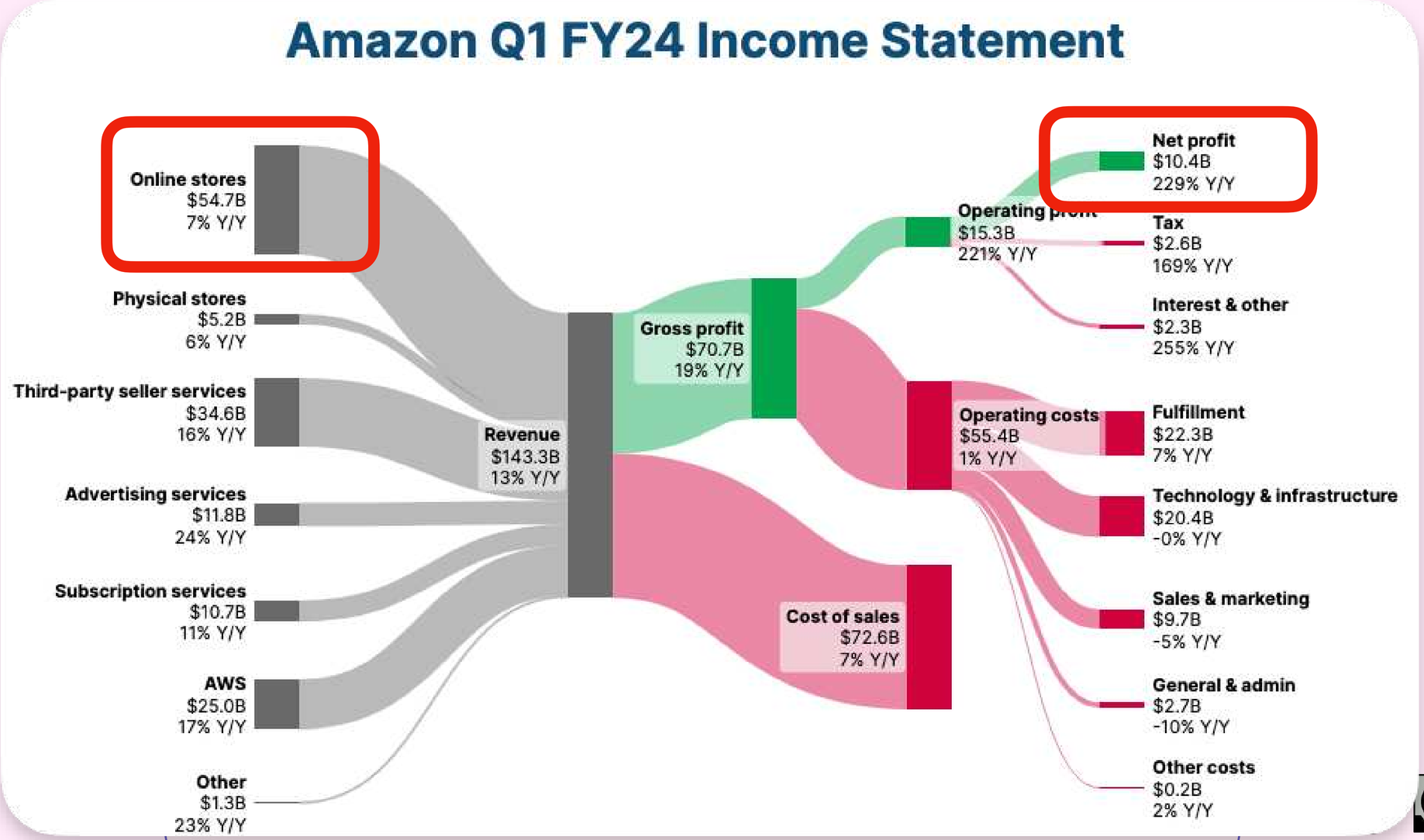


CC BY NC SA

Pot-pourri of RAG protips

Retrieval

Source: https://www.sankeyart.com/sankeys/public/17290/



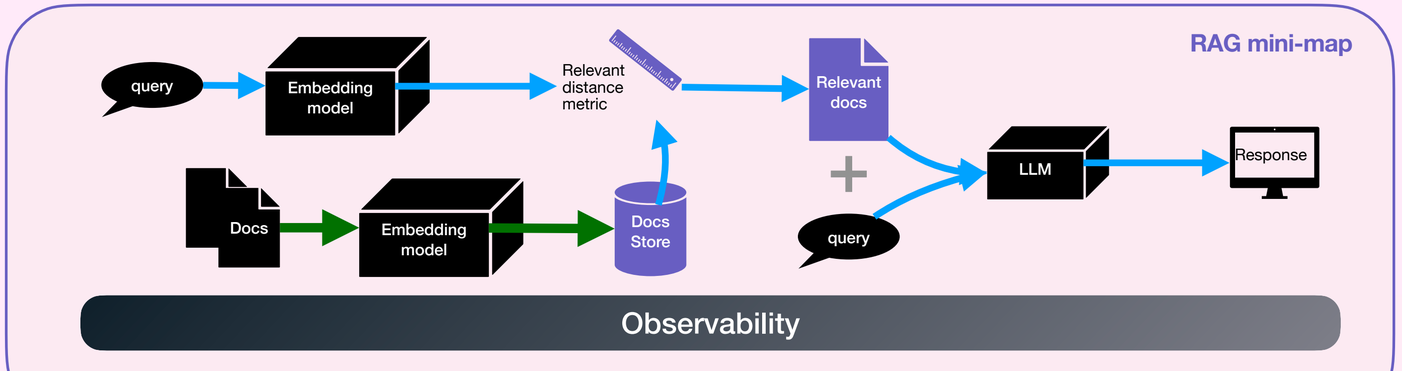
CC BY NC SA

Pot-pourri of RAG protips

Retrieval

metadata filtering is not dead !

- Enable domain specific search
- Embeddings work fine for full text, but bad for date, categories, price, etc
- Humans search with keywords & filter in mind
- Affordable way to improve your retrieval pipeline



Supported in several vector stores

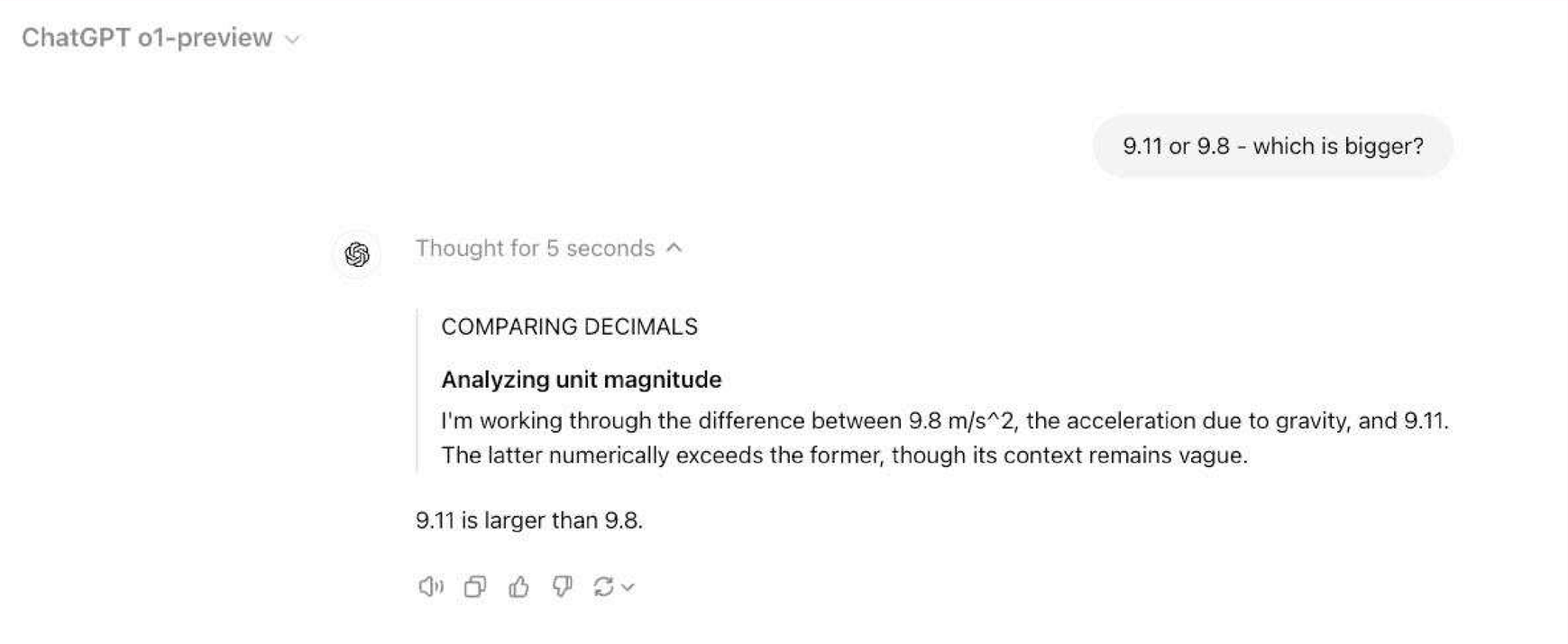
Weaviate

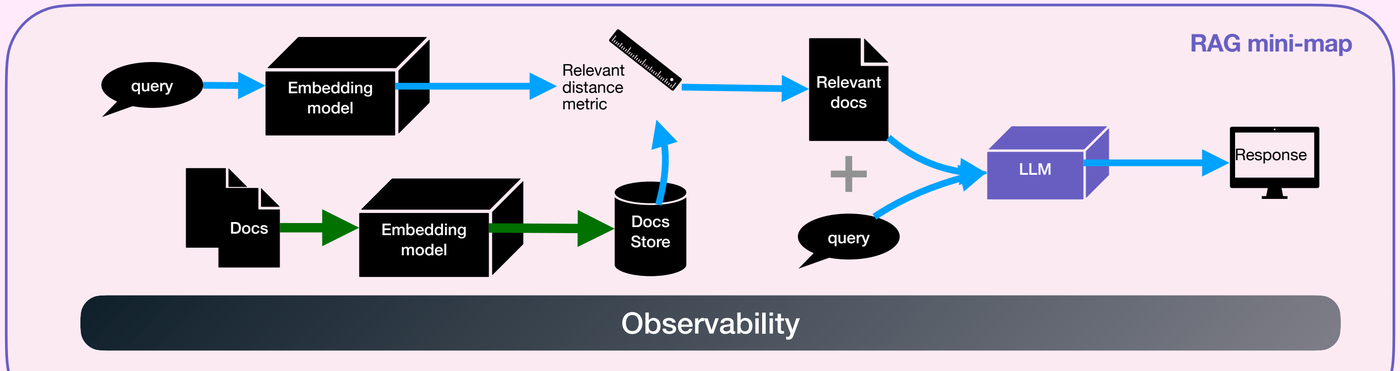
CC BY NC SA

Pot-pourri of RAG protips

Retrieval

metadata filtering is not dead !

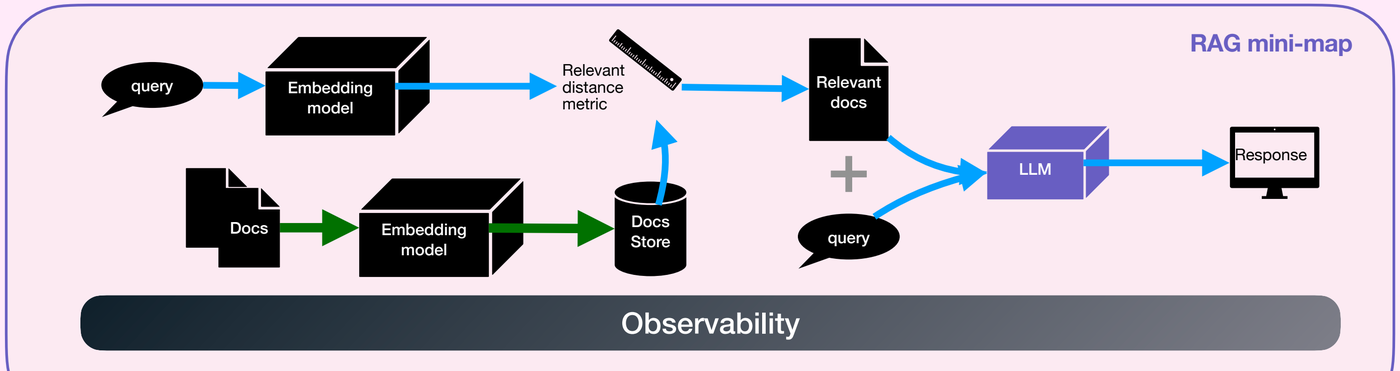




CC BY NC SA

Pot-pourri of RAG protips

Retrieval

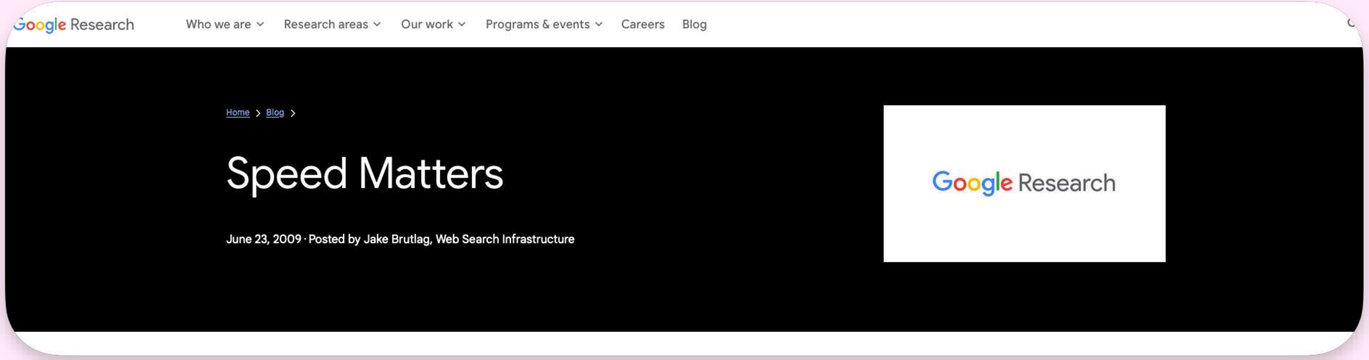


CC BY NC SA

Pot-pourri of RAG protips

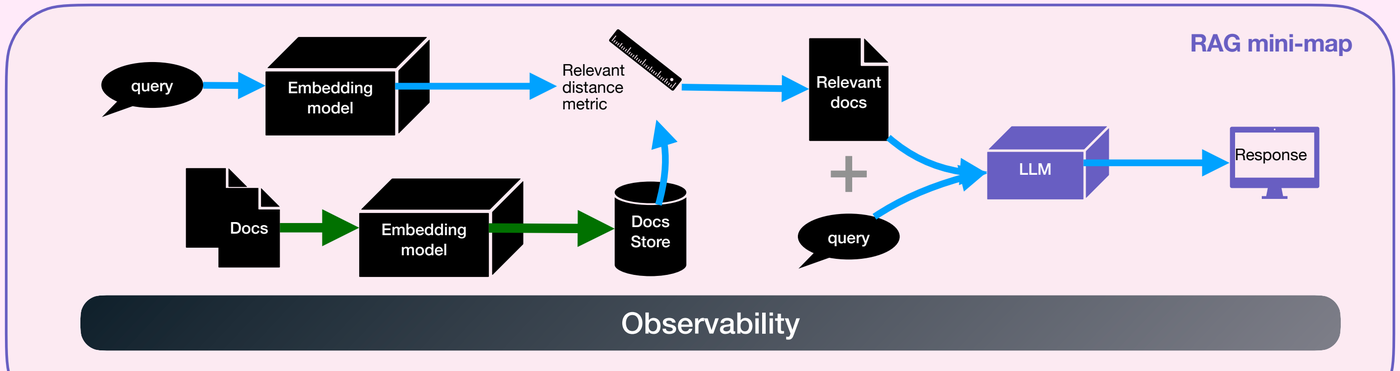
Useful metrics to track in production

Speed is a feature



Source: Google Research - Speed matters, Web Search infrastructure

All other things being equal, more usage, as measured by number of searches, reflects more satisfied users. Our experiments demonstrate that slowing down the search results page by 100 to 400 milliseconds has a measurable impact on the number of searches per user of -0.2% to -0.6% (averaged over four or six weeks depending on the experiment). That's 0.2% to 0.6% fewer searches for changes under half a second!



CC BY NC SA

Pot-pourri of RAG protips

Useful metrics to track in production

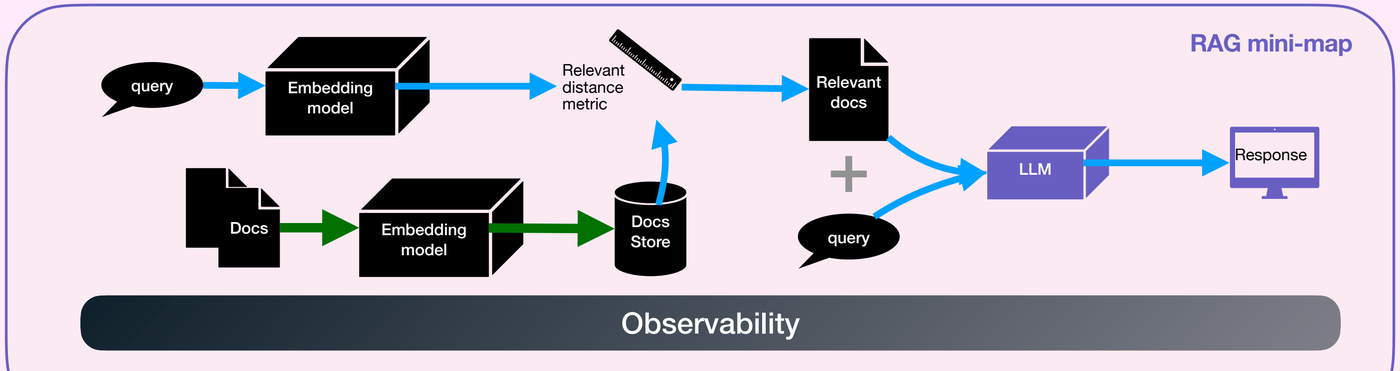
Speed is a feature

2 metrics to consider logging per user session:

- **Total latency -** for a request completion
- **Time to first token** - time it takes for the token to be returned to the user

Look at impact on user interaction

Follow up action? Terminated their session?



CC BY NC SA

Pot-pourri of RAG protips

Useful metrics to track in production

Tracking semantic drift across time

> *semantic drift*: changes in semantic relevance over time.

Why tracking it ?

- Identify trends in user behaviour & preferences
- Detect areas where the retrieval system is lagging
- Ensure continuous alignment with user needs and interests.

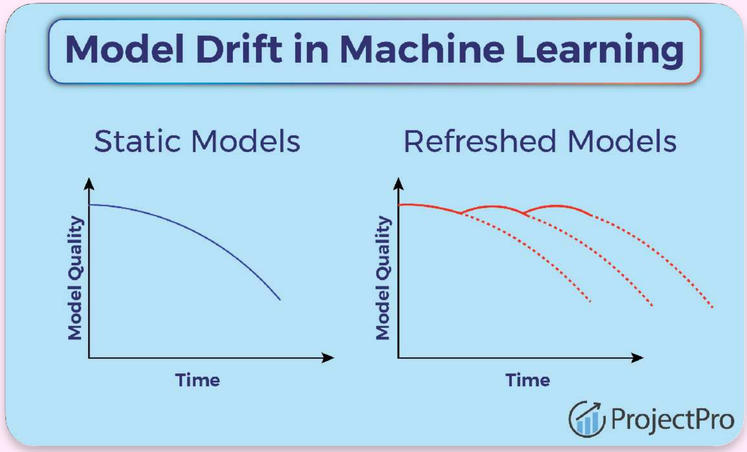
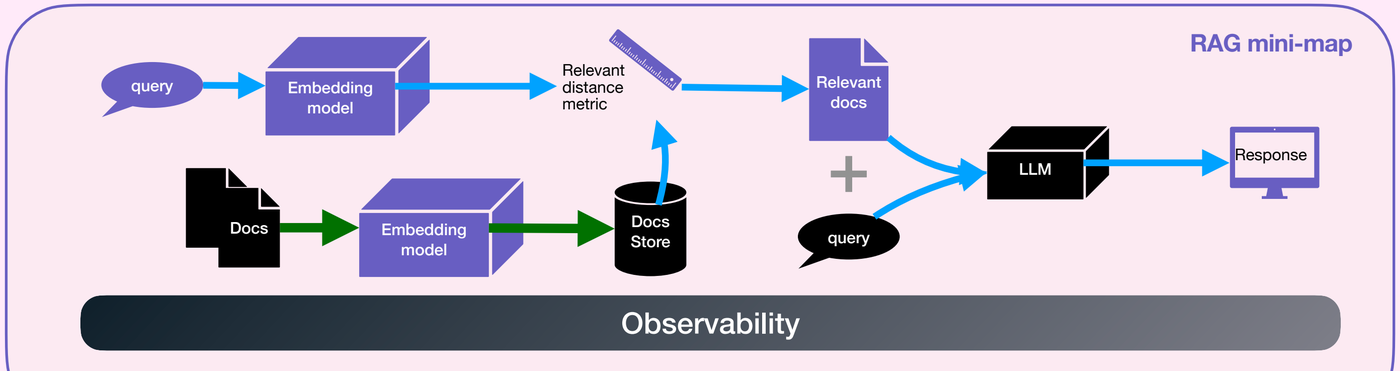


Image source: ProjectPro



CC BY NC SA

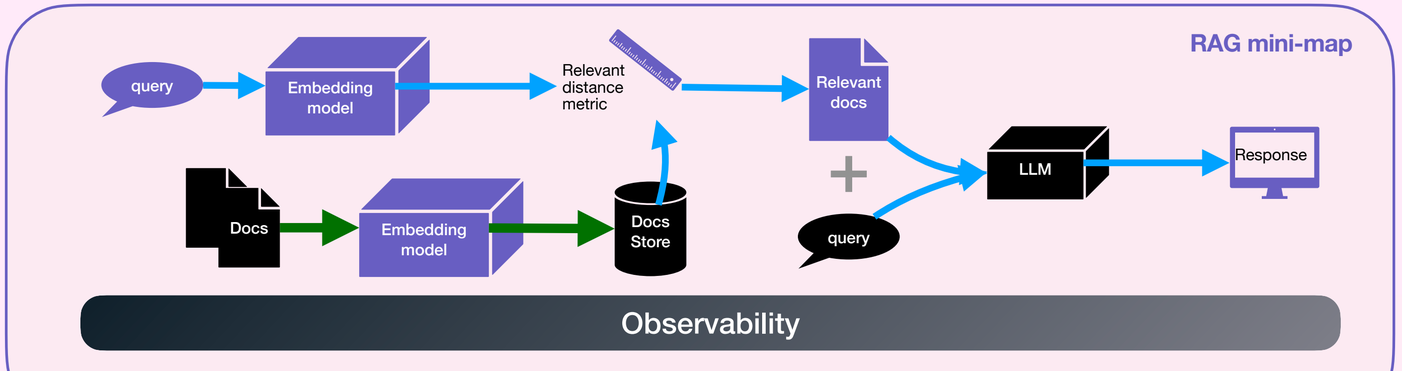
Pot-pourri of RAG protips

Useful metrics to track in production

Tracking semantic drift across time

How to track it ?

- Collect cosine similarity scores for user-query and document pairs.
- Use time-series plots to track changes in similarity scores over time and/or user segments

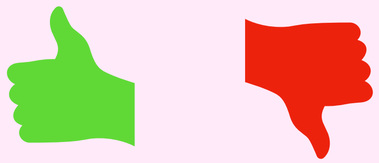


CC BY NC SA

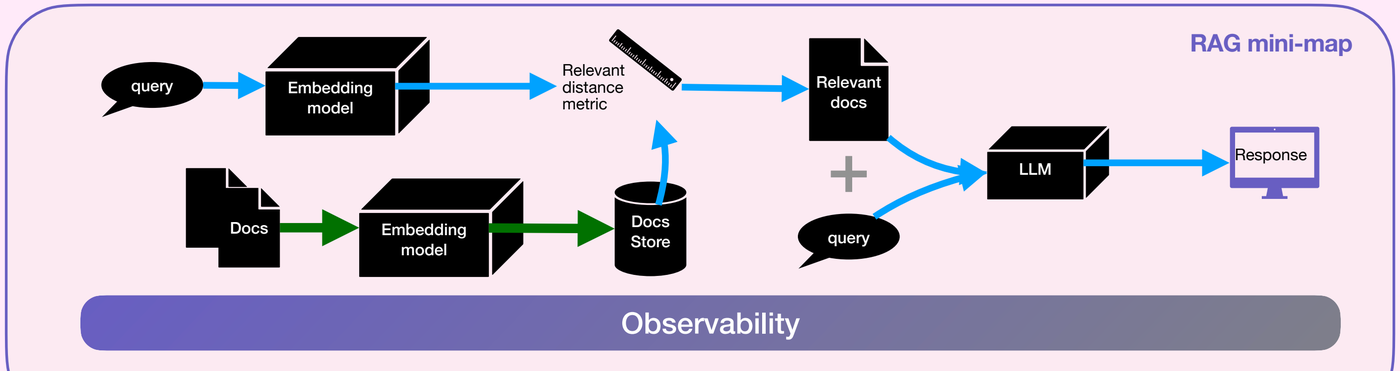
Pot-pourri of RAG protips

Beyond thumbs up - Collecting actionable user feedback

- Alice - “Let’s collect some feedback from our user”
- Bob - “Yay great idea! How do we do it ?”
- Alice - “Let’s add some thumbs up / down button in the UI”



Challenge: binary label 'good’ or ‘bad’ are vague and hard to use to improve your pipeline.

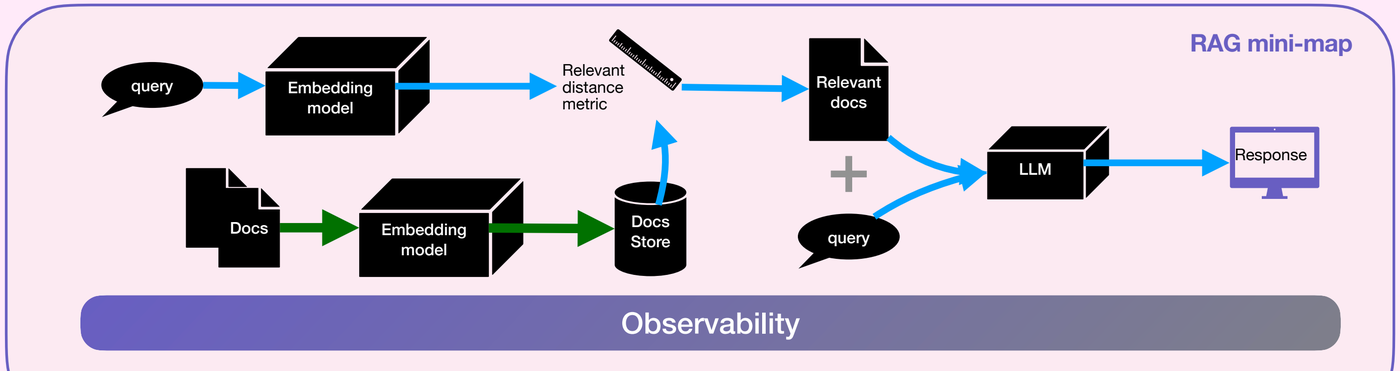


CC BY NC SA

Pot-pourri of RAG protips

Beyond thumbs up - Collecting actionable user feedback

Key qualities of feedback interface	Details
Clarity & simplicity	Use clear, concise language for feedback categories
Specificity	Offer specific options that cover a range of common issues
Open-ended Input	Consider an optional text box for users to elaborate



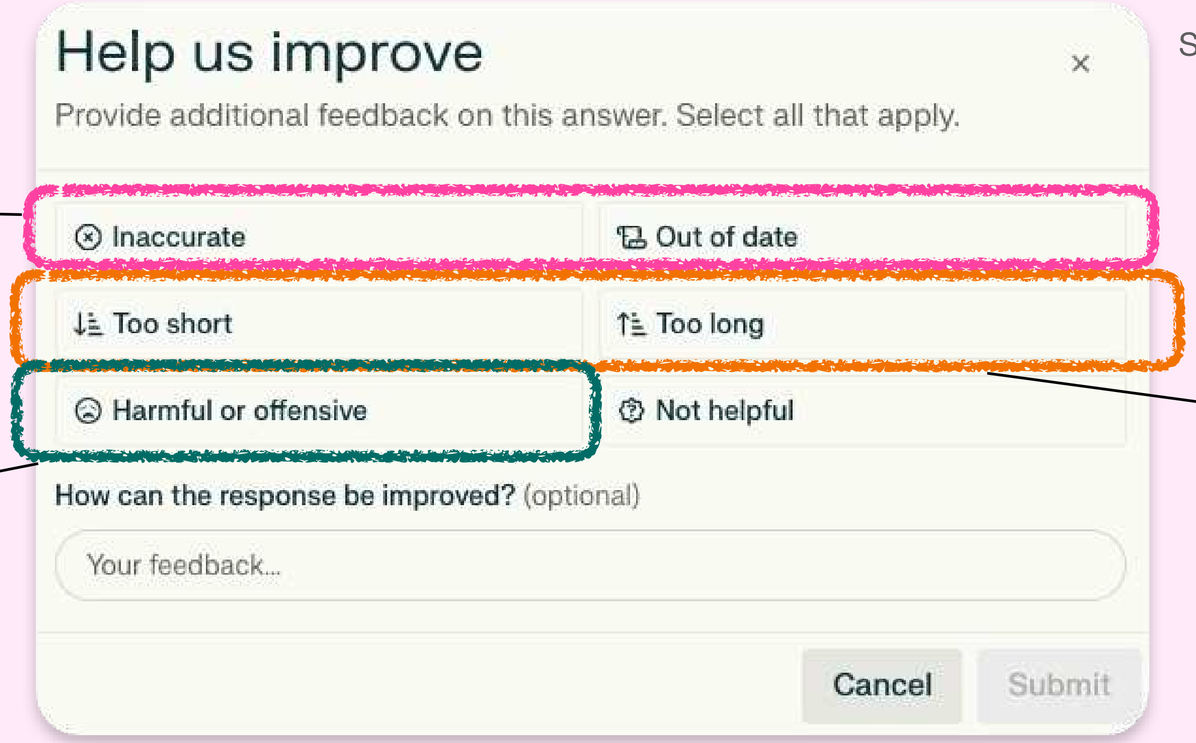
CC BY NC SA

Pot-pourri of RAG protips

Beyond thumbs up - Collecting actionable user feedback

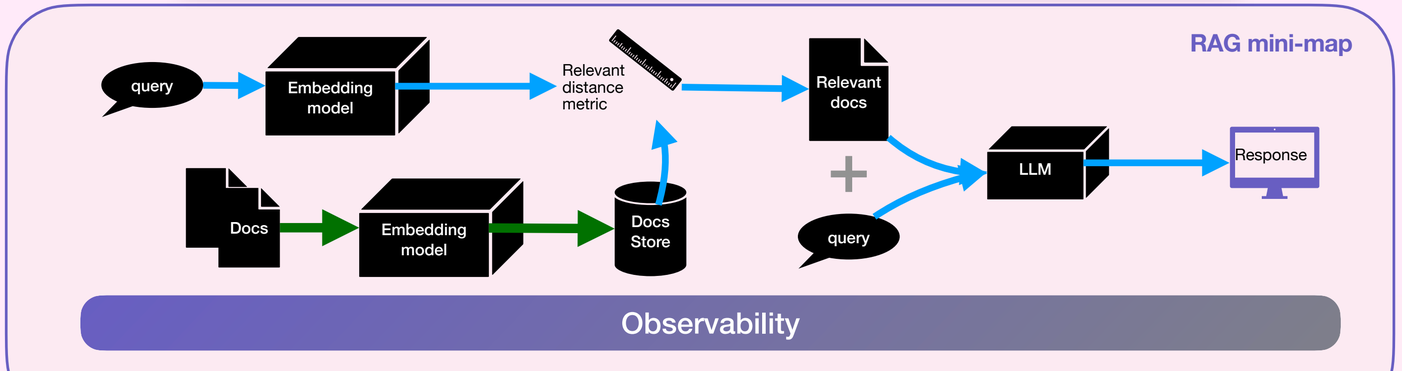
Likely due to retrieval system

Might need to perform preference alignement on LLM



Source: Feedback button of Perplexity.ai

Likely due to LLM prompt



CC BY NC SA

In a nutshell

Generative AI isn't (just) about building models - it's about building understanding.

> Monitoring and error analysis are key to discover untapped niches and deliver specialized solutions

CC BY NC SA

Appendix

Slides are available as pdf on fpaupier.fr/speaking

CC BY NC SA

Bibliography

- **Dozens interviews with builders, entrepreneurs & corporate actors in GenAI**

- **Hamel Husain (@hamelhusain): What We've Learned From A Year of Building with LLMs, June 2024** <— *If you're reading just one, read this one*
- Allspaw, J., & Robbins, J. (2010). Web operations: Keeping the data on time. O'Reilly Media
- Andrej Karpathy What I learned from competing against ConvNet on ImageNet, Sep 2 2014
- Shreya Shankar, Data Flywheels for LLM Applications, Jul 1, 2024
- Maxime Labonne personal blog, July 2024
- Chip Huyen (@chipro), Building A Generative AI Platform, Jul 25 2024
- Eugene Yan (@eugeneyan), Prompting Fundamentals and How to Apply them Effectively, May 2024
- Sebastian Raschka (@rasbt) Ahead of AI newsletter
- Philipp Schmid (@_philschmid): LLM Evaluation doesn't need to be complicated, Jul 2024
- Jason Liu (@jxnlco): RAG is more than embeding, Sep 2023
- Jason Liu (@jxnlco): Stop using LTPM@Few as a metric, Feb 2024
- Simon Willison (@simonw): Imitation Intelligence, keynote for PyCon US 2024, 14 Jul 2024
- Omar Sanseviero (@osanseviero): LLM Evals and Benchmarking, Mar 2024
- Lilian Weng (@lilianweng): Extrinsic Hallucinations in LLMs, Jul 2024

CC BY NC SA