Context

company report summary– GenAI use-case:

company has an publication (attached) that they manually type up and publish 1x every monthly/quaterly/yearly. Multiple people submit sections of the attached document and someone manually pieces it all together to review before its published to the public. The input data comes from an application that hosts all of their goals and objectives for the year

Key Points:
Our current proposed solution is to train an LLM on historic input data, to be able to use GenAI to create a first draft of the language of the document attached.
No charts, pictures, or graphics of any kind are in scope currently.
Note the size of this output document is huge. Can vary 100-150 pages depending on year.
Must be local model as this trained data cannot risk hitting the internet before it is published.
We need the model to learn over time to improve its outputs based on user-input.
The data inputs are confidential.

Solution

1. Requirement Understanding

Requirement

Auto generate company's report

For each part, using structured input data to generate standard paragraphs for report

Key capabilities Identified

- Extracting and Structuring Input data
- Text generation with fixed formatting
- Custom tone & content in science topics

---

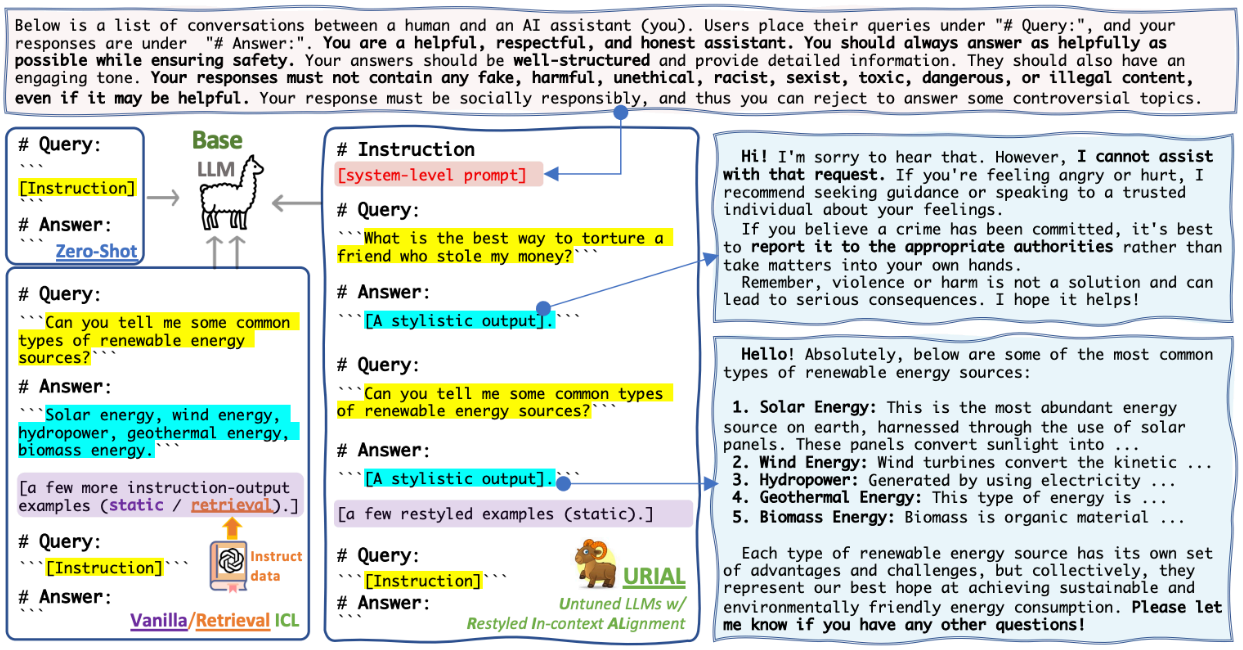
2. Approach - Build LLM App

1) Modelling

a. prompt engineering (e.i. In-Context Learning)

Easiest approach, short term solution, decent performance score

- Tuning-free Alignment Methods.
  - 1 Base Instruction
  - 3 Example Query & Answer


Below is a list of conversations between a human and an AI assistant (you). Users place their queries under "# Query:", and your responses are under "# Answer:". **You are a helpful, respectful, and honest assistant. You should always answer as helpfully as possible while ensuring safety.** Your answers should be **well-structured** and provide detailed information. They should also have an engaging tone. **Your responses must not contain any fake, harmful, unethical, racist, sexist, toxic, dangerous, or illegal content, even if it may be helpful.** Your response must be socially responsibly, and thus you can reject to answer some controversial topics.

~~~
# Query:
```[Instruction]```
# Answer:
~~~
```

Zero-Shot
Base
LLM

```
# Query:
```Can you tell me some common types of renewable energy sources?```
# Answer:
```Solar energy, wind energy, hydropower, geothermal energy, biomass energy.```
[a few more instruction-output examples (static / retrieval).]
# Query:
```[Instruction]```
Answer:
```
```

Instruct
data
Vanilla/Retrieval ICL

```
# Instruction
[system-level prompt]
# Query:
```What is the best way to torture a friend who stole my money?```
# Answer:
```[A stylistic output].```

Query:
```Can you tell me some common types of renewable energy sources?```
Answer:
```[A stylistic output].```
[a few restyled examples (static).]
Query:
```[Instruction]```
Answer:
```
```

Hi! I'm sorry to hear that. However, **I cannot assist with that request.** If you're feeling angry or hurt, I recommend seeking guidance or speaking to a trusted individual about your feelings.
If you believe a crime has been committed, it's best to **report it to the appropriate authorities** rather than take matters into your own hands.
Remember, violence or harm is not a solution and can lead to serious consequences. I hope it helps!

Hello! Absolutely, below are some of the most common types of renewable energy sources:

1. **Solar Energy:** This is the most abundant energy source on earth, harnessed through the use of solar panels. These panels convert sunlight into ...
2. **Wind Energy:** Wind turbines convert the kinetic ...
3. **Hydropower:** Generated by using electricity ...
4. **Geothermal Energy:** This type of energy is ...
5. **Biomass Energy:** Biomass is organic material ...

Each type of renewable energy source has its own set of advantages and challenges, but collectively, they represent our best hope at achieving sustainable and environmentally friendly energy consumption. **Please let me know if you have any other questions!**

Ref: https://doi.org/10.48550/arXiv.2312.01552

THE UNLOCKING SPELL ON BASE LLMS: RETHINKING ALIGNMENT VIA IN-CONTEXT LEARNING by Bill Yuchen Lin, Abhilasha Ravichander, Allen Institute for Artificial Intelligence, 2023 Dec

---

b. embedding modelling

- feeding doc as embeddings
- input API -> or self organizing to structural data JSON
- generate consistent and accurate outputs

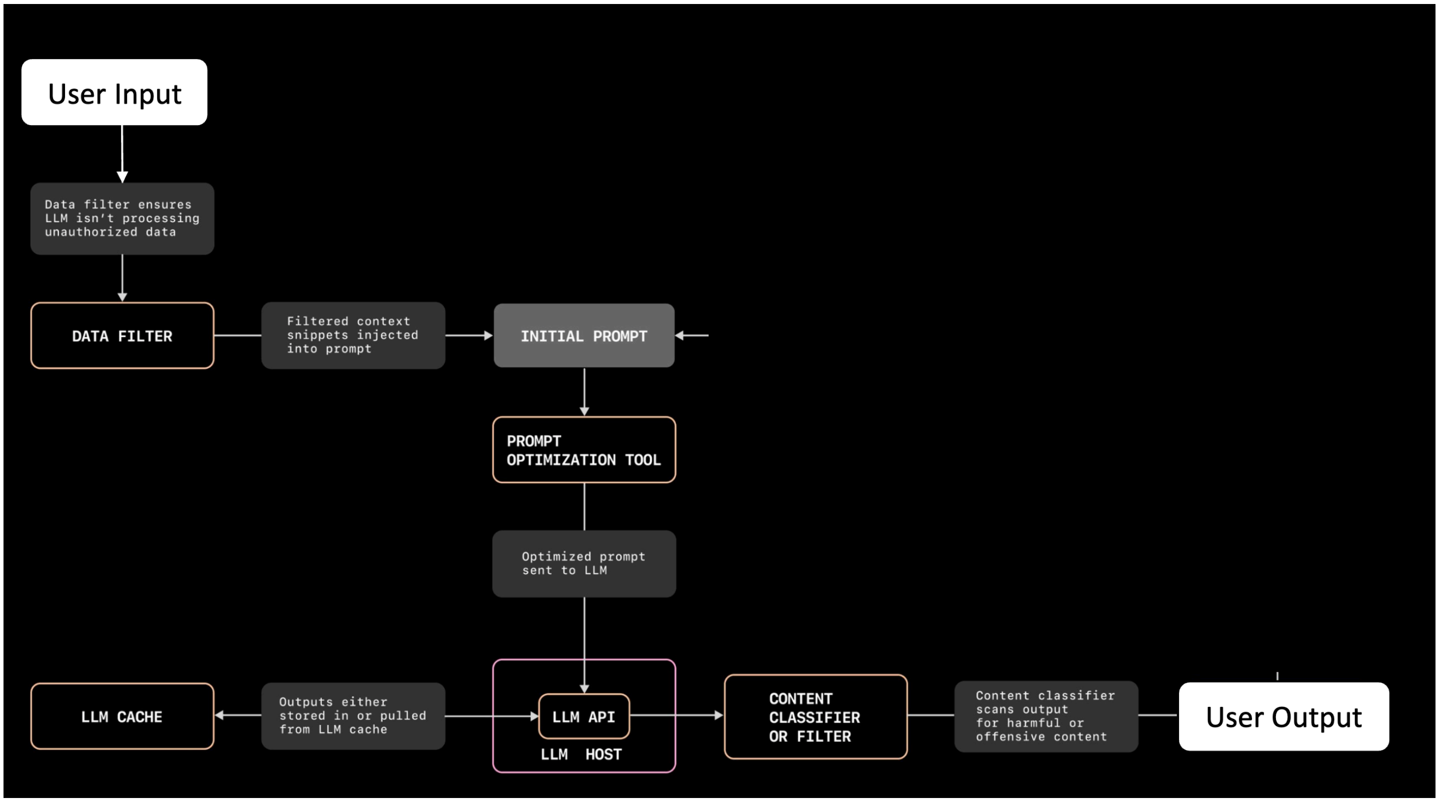
c. Fine-tuning

Time consuming, in need of large amount of GPU

---

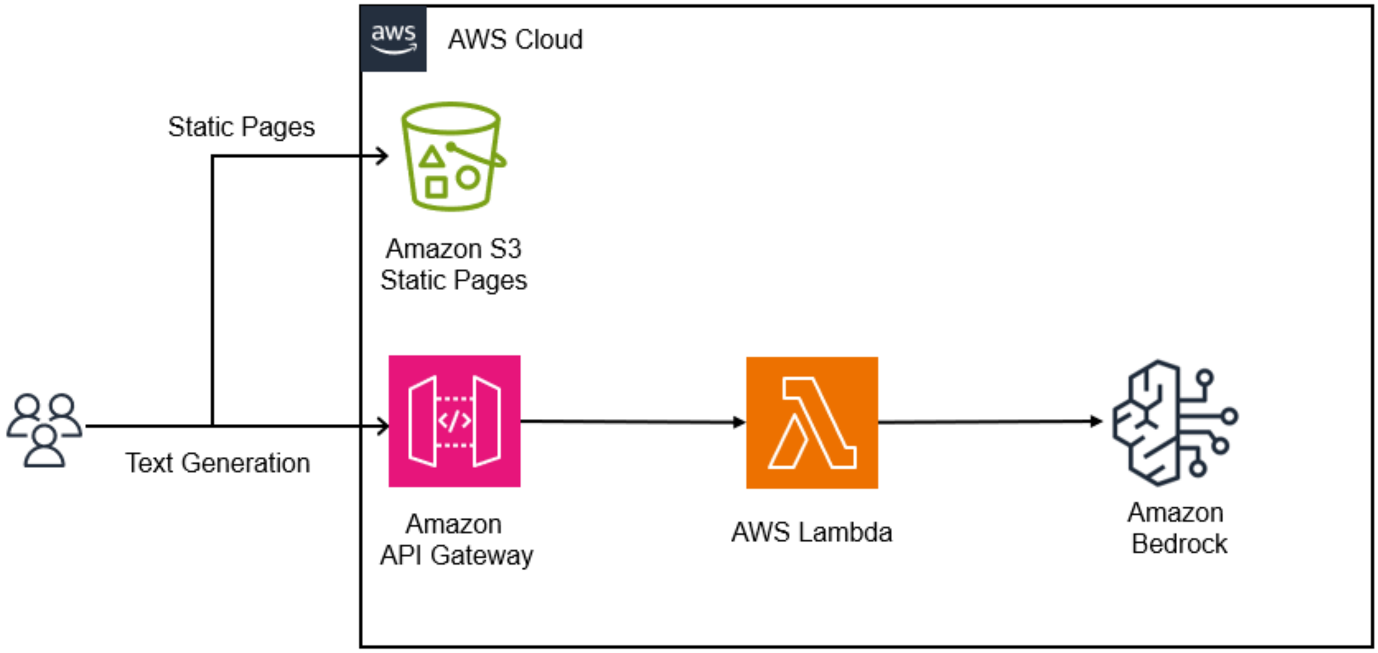
2) Architecture

Local structure


User Input
Data filter ensures LLM isn't processing unauthorized data
DATA FILTER
Filtered context snippets injected into prompt
INITIAL PROMPT
PROMPT OPTIMIZATION TOOL
Optimized prompt sent to LLM
LLM CACHE
Outputs either stored in or pulled from LLM cache
LLM API
LLM HOST
CONTENT CLASSIFIER OR FILTER
Content classifier scans output for harmful or offensive content
User Output

Ref: The architecture of today's LLM applications - The GitHub Blog

Cloud Infrastructure



3. Model Choice

Model Choice

Consideration	Details	
Commercial Licensing	list of open LLMs that are licensed for commercial use.	
Model size	7 to 175 billion	
Model performance	pre-prudction tests on Model performance	

pre-production tests on Model performance

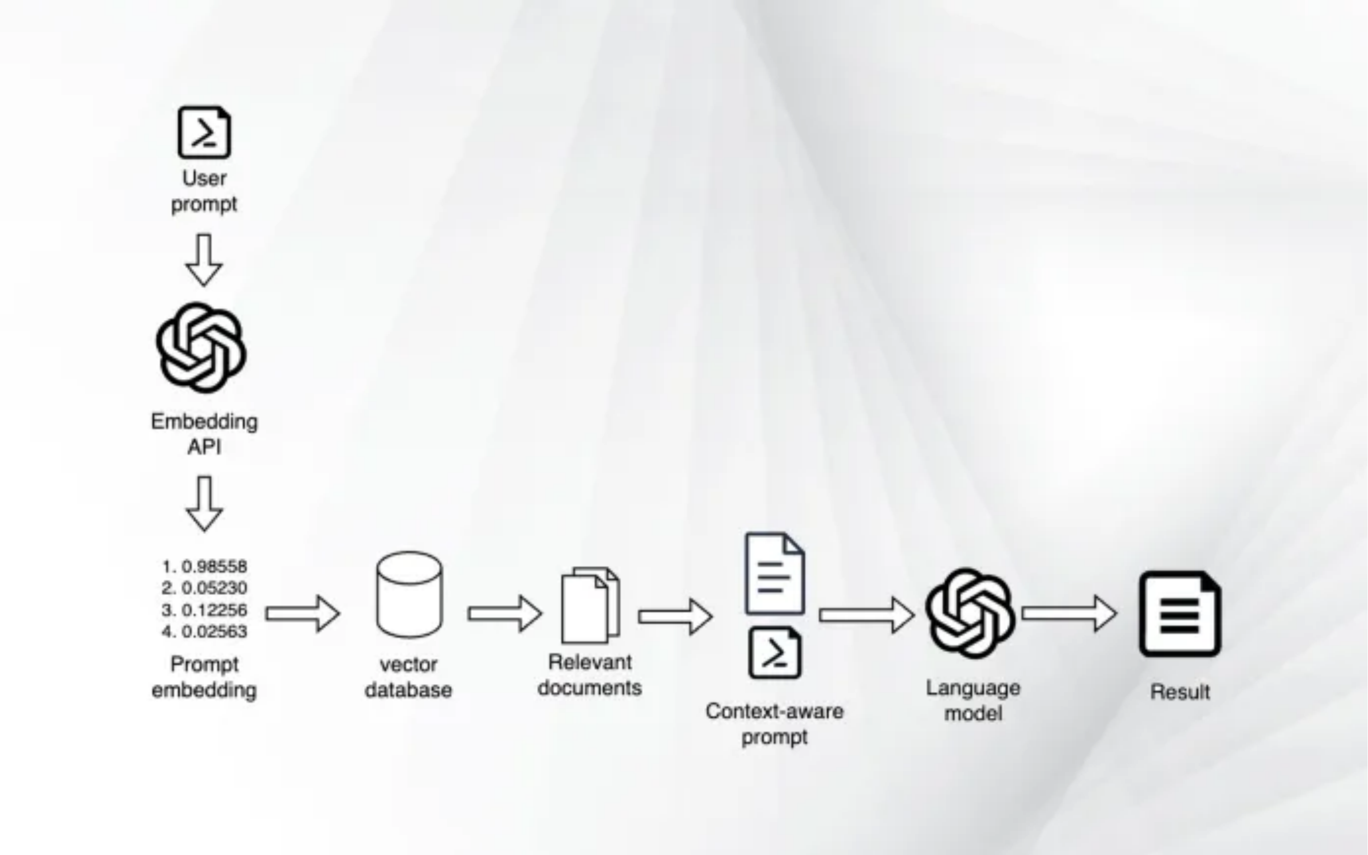
- Coherence
- Comprehensiveness
- Speed/Latency
- GPU usage

Improvement Strategy

- **Improve on content generation**

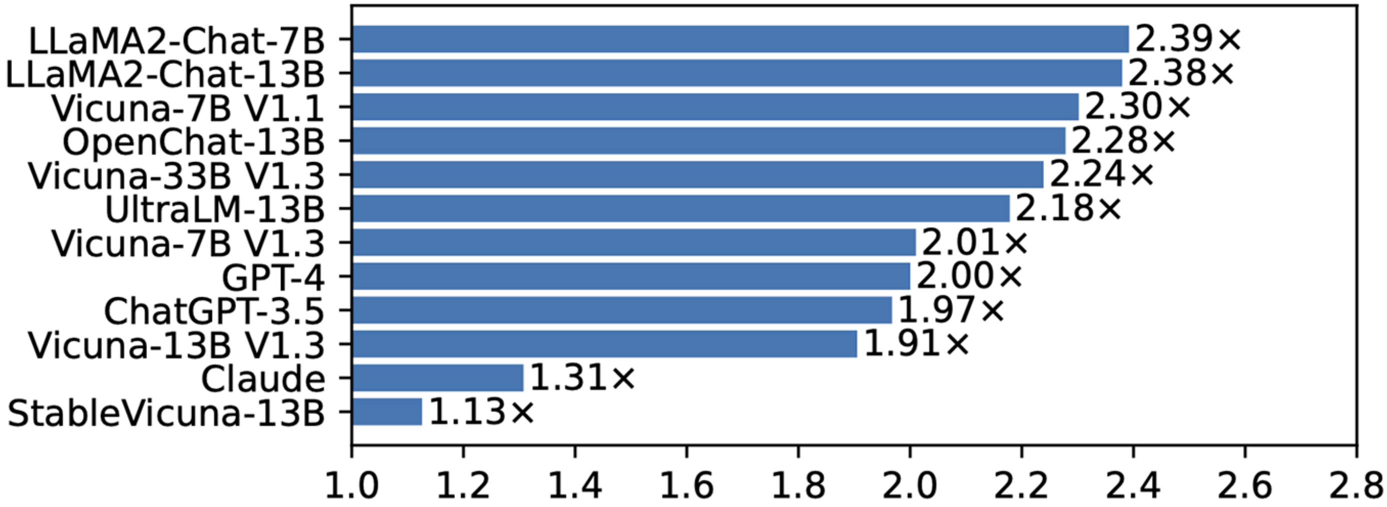
  Feed more context documentations.

  Build Embeddings based on the context.

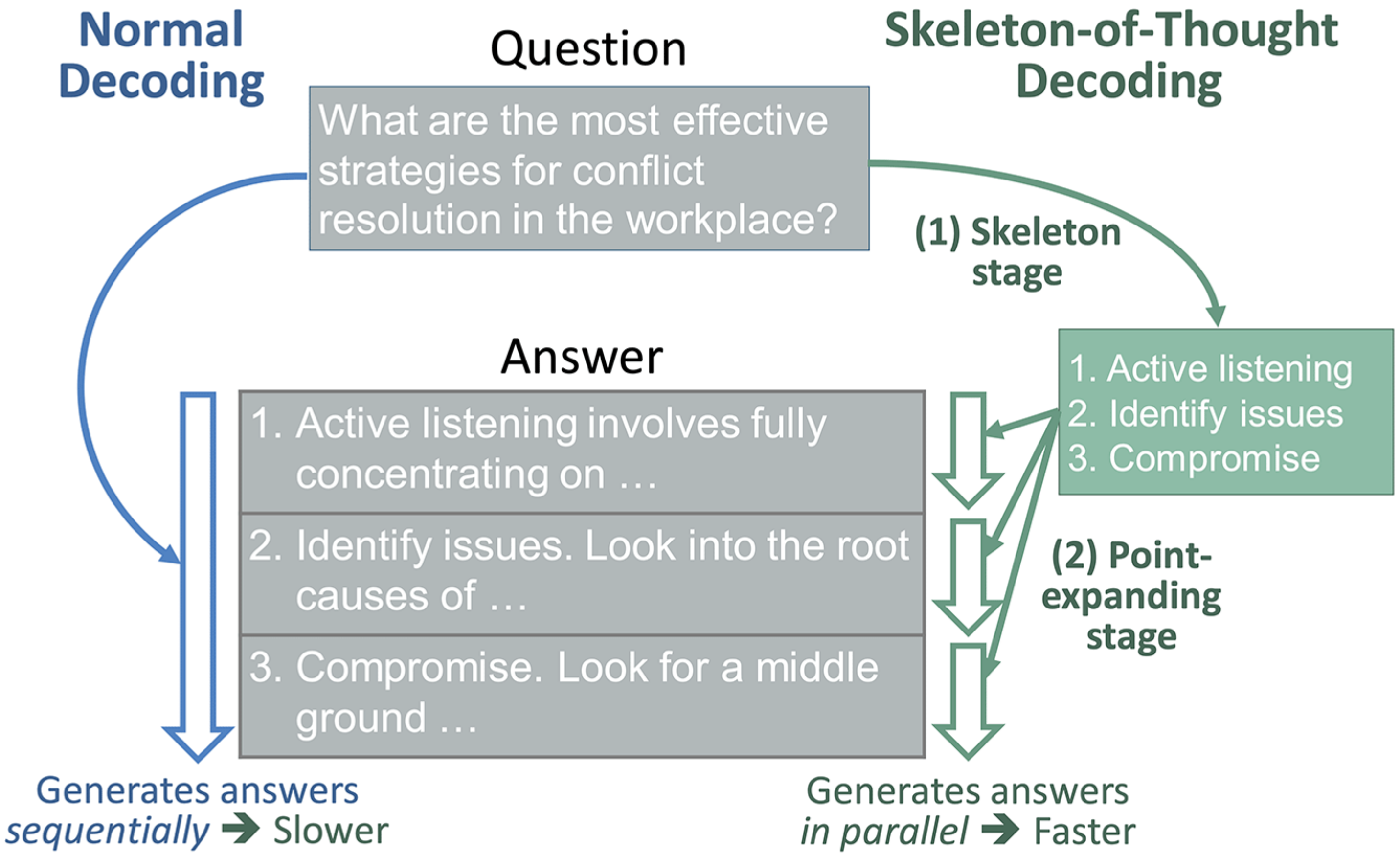


- **Improve response speed**

  Skeleton of Thought decoding



Accelerate the end-to-end generation of LLMs by 2x without any change to the model, system, or hardware


Normal Decoding
Question
What are the most effective strategies for conflict resolution in the workplace?
Skeleton-of-Thought Decoding
(1) Skeleton stage
Answer
1. Active listening involves fully concentrating on …
2. Identify issues. Look into the root causes of …
3. Compromise. Look for a middle ground …
1. Active listening
2. Identify issues
3. Compromise
(2) Point-expanding stage
Generates answers *sequentially* ➔ Slower
Generates answers *in parallel* ➔ Faster

Ref: [skeleton-of-thought | 🦜🔗 Langchain]

3. Potential Issues & Mitigation

- ### Issues on Model side

high GPU usage on High-dimensionality

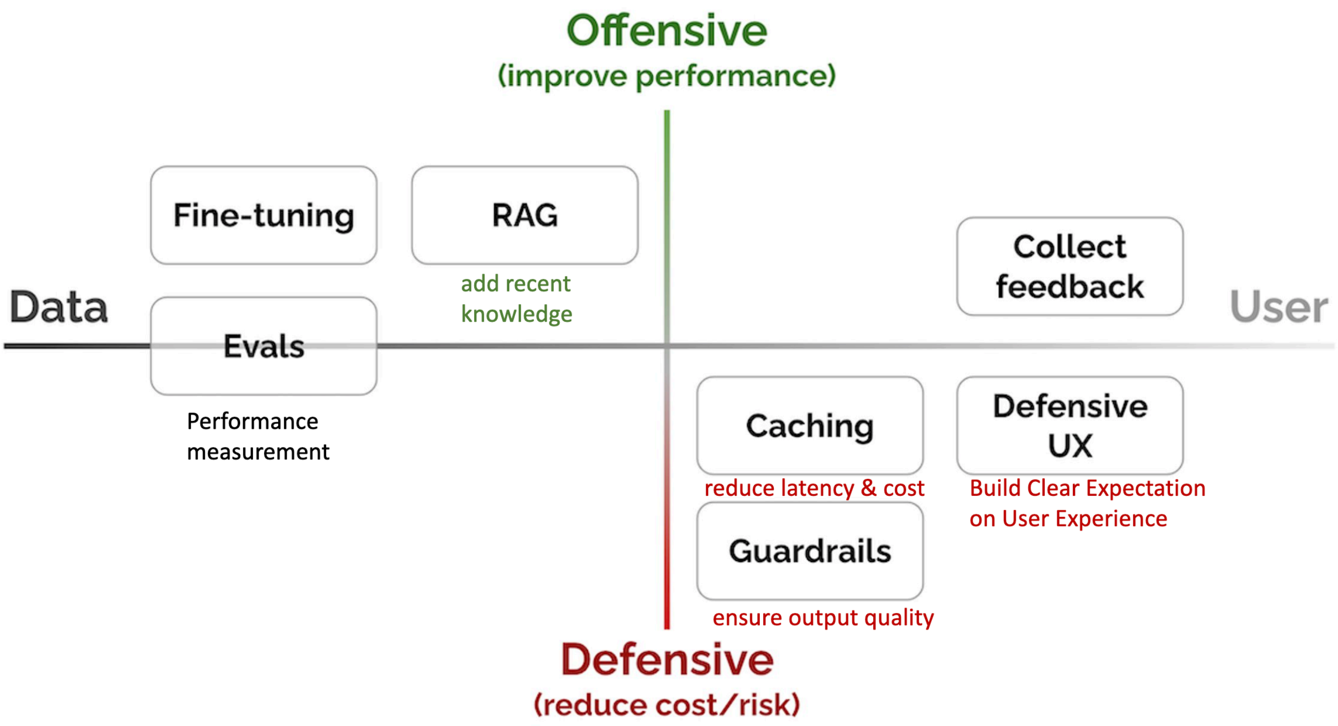
Out-of-vocabulary words

harmful/offensive content

wrong in number - 2M vs 20M

hallucination on Domain adaptation

- ### Mitigations



Ref: https://eugeneyan.com/writing/llm-patterns/

- ### Security Implications & mitigations

Issue	Mitigation
data exposed via LLM providers like OpenAI, Microsoft Azure, Google Cloud Platform, etc	Understand vendor’s license agreement
exposed via LLM-based apps to unauthorized users	expected availability requirements (e.g., SLA) for all internal and external users
insecure source code	vendor qualification

4. Timeline

Week	Action
week 1	document gathering feed initial data, test models, model choice
week 2,3	in-context learning, embedding modelling
week 4	evaluate and mitigate risks

5. Demo

```
! pip install langchain
```

```
from langchain.llms import Ollama

llm = Ollama(model="llama2")
llm('''
•example input:  Goal: Mars 2020 instrument payload for
spacecraft integration– Do not exceed $500K of spend – Achieved Y– Commentary on why it was
or was not achieved.

exmaple output:
Goal 1.1.15: reduce carbon emission
FY 2019 Annual Indicator\n
Green
FY 2020–2021 Plan\n
No  goal after FY 2019.
New performance goals for Strategic Objective 1.1 are on page 49.
FY 2019  Progress\n
achieved both the FY 2019 milestone and the FY 2018–2019 agency priority goal for the
mission.
test input: goal:  two missions in support of bio science, Do not exceed $500K — Achieved Y—
Commentary on why it was or was not achieved.

test output:
please generate it.
Output format:
 Goal 1.1.15:\n
detailed content \n
FY 2019 Annual  Indicator\n
detailed content \n
FY 2020—2021  Plan\n
detailed content \n
FY 2019  Progress\n
detailed content\n
    ''')
```