Production-scale Retrieval Augmented Generation for Real-Time News Distillation

Grenoble DS meetup
2024-01-18

Robert Caulk, PhD
Founder

Elin Törnquist, PhD
Co-founder, Lead Research Scientist

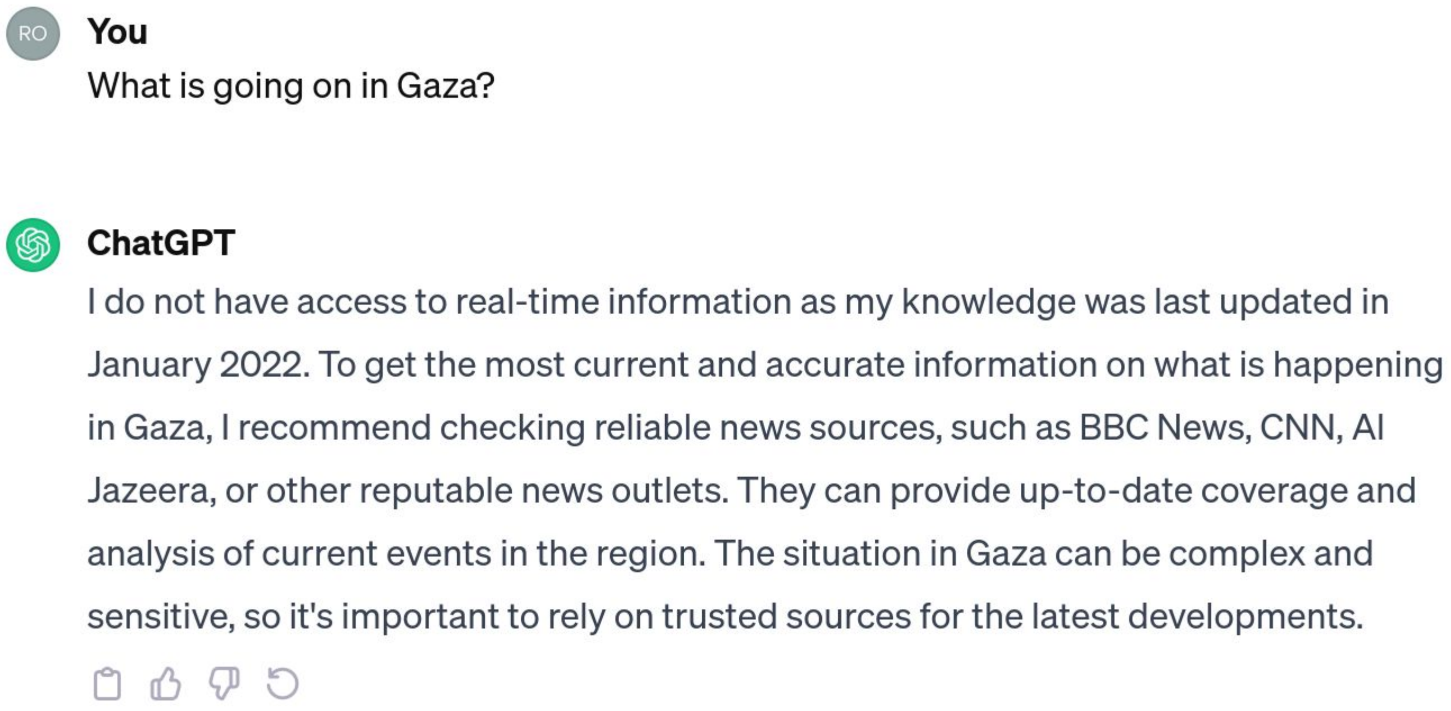
Emergent Methods

Retrieval Augmented Generation (RAG)

Emergent Methods

Why Retrieval Augmented Generation (RAG)?

- LLM is trained on historic data
- Many applications demand real-time data
- Re-training LLMs can take months on thousands of GPUs



You

What is going on in Gaza?

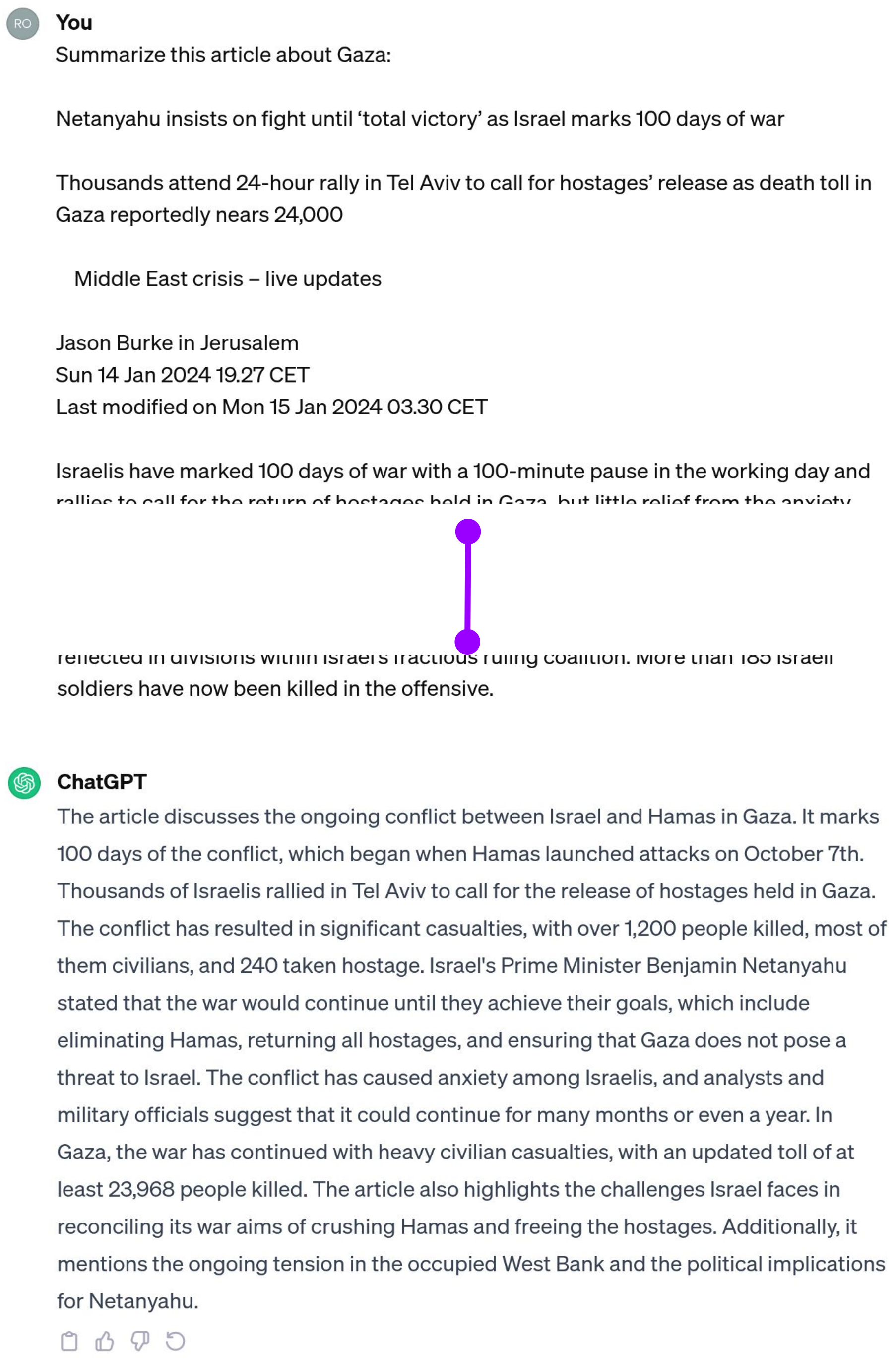
ChatGPT

I do not have access to real-time information as my knowledge was last updated in January 2022. To get the most current and accurate information on what is happening in Gaza, I recommend checking reliable news sources, such as BBC News, CNN, Al Jazeera, or other reputable news outlets. They can provide up-to-date coverage and analysis of current events in the region. The situation in Gaza can be complex and sensitive, so it's important to rely on trusted sources for the latest developments.

Why Retrieval Augmented Generation (RAG)?

- LLMs have great reasoning capacity
- We can simply feed it the news article and ask for a well reasoned summary and analysis

It is not the user's job to go find the news and copy paste it into ChatGPT…



You
Summarize this article about Gaza:

Netanyahu insists on fight until 'total victory' as Israel marks 100 days of war

Thousands attend 24-hour rally in Tel Aviv to call for hostages' release as death toll in Gaza reportedly nears 24,000

Middle East crisis – live updates

Jason Burke in Jerusalem
Sun 14 Jan 2024 19.27 CET
Last modified on Mon 15 Jan 2024 03.30 CET

Israelis have marked 100 days of war with a 100-minute pause in the working day and rallies to call for the return of hostages held in Gaza, but little relief from the anxiety

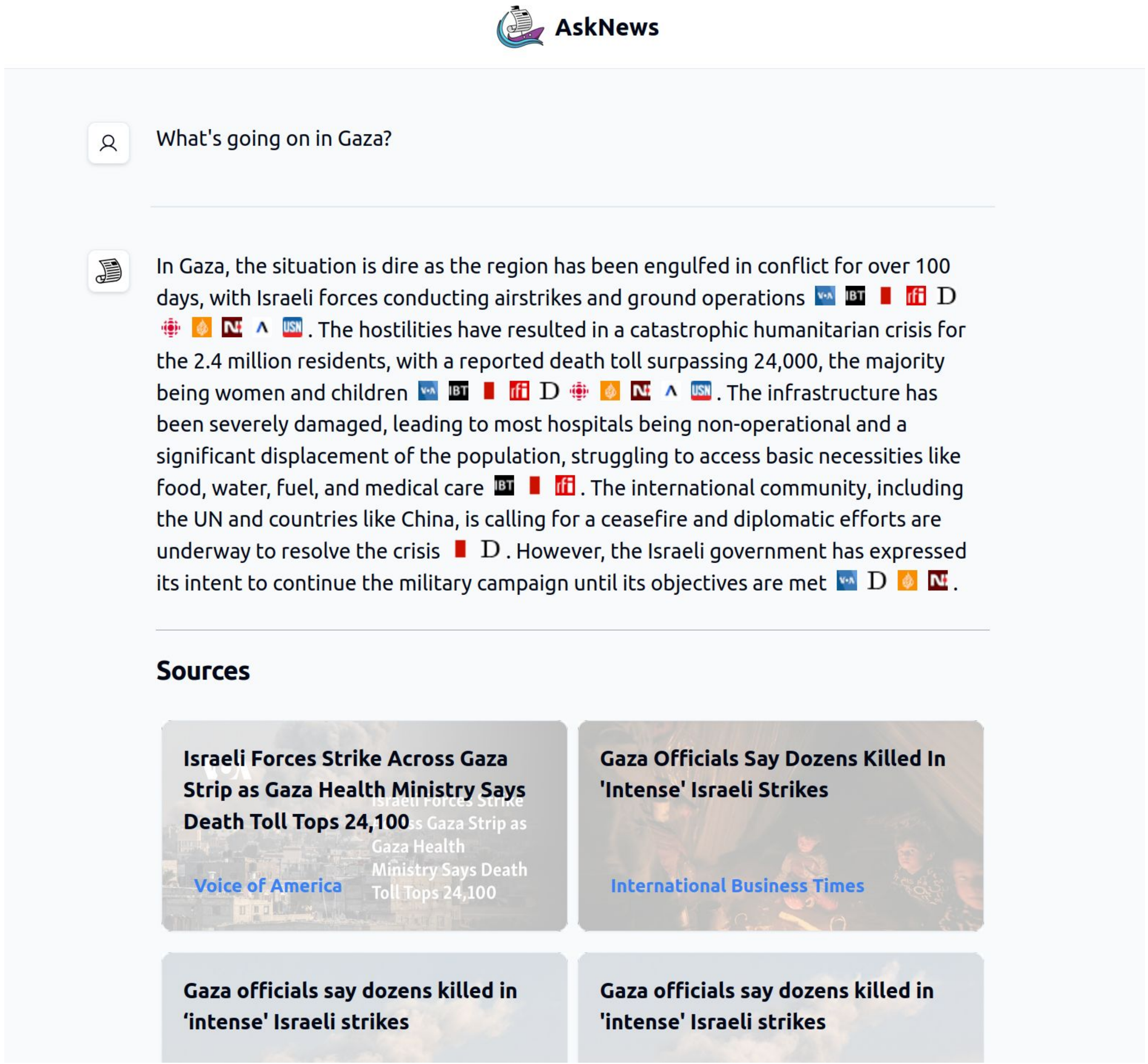
reflected in divisions within Israel's fractious ruling coalition. More than 185 Israeli soldiers have now been killed in the offensive.

ChatGPT
The article discusses the ongoing conflict between Israel and Hamas in Gaza. It marks 100 days of the conflict, which began when Hamas launched attacks on October 7th. Thousands of Israelis rallied in Tel Aviv to call for the release of hostages held in Gaza. The conflict has resulted in significant casualties, with over 1,200 people killed, most of them civilians, and 240 taken hostage. Israel's Prime Minister Benjamin Netanyahu stated that the war would continue until they achieve their goals, which include eliminating Hamas, returning all hostages, and ensuring that Gaza does not pose a threat to Israel. The conflict has caused anxiety among Israelis, and analysts and military officials suggest that it could continue for many months or even a year. In Gaza, the war has continued with heavy civilian casualties, with an updated toll of at least 23,968 people killed. The article also highlights the challenges Israel faces in reconciling its war aims of crushing Hamas and freeing the hostages. Additionally, it mentions the ongoing tension in the occupied West Bank and the political implications for Netanyahu.

Emergent Methods
www.emergentmethods.ai

How do we engineer the context?

- Automating balanced context engineering
- Ensuring proper attribution
- Minimizing hallucination
- Scaling to thousands of users



AskNews

What's going on in Gaza?

In Gaza, the situation is dire as the region has been engulfed in conflict for over 100 days, with Israeli forces conducting airstrikes and ground operations. The hostilities have resulted in a catastrophic humanitarian crisis for the 2.4 million residents, with a reported death toll surpassing 24,000, the majority being women and children. The infrastructure has been severely damaged, leading to most hospitals being non-operational and a significant displacement of the population, struggling to access basic necessities like food, water, fuel, and medical care. The international community, including the UN and countries like China, is calling for a ceasefire and diplomatic efforts are underway to resolve the crisis. However, the Israeli government has expressed its intent to continue the military campaign until its objectives are met.

Sources

Israeli Forces Strike Across Gaza Strip as Gaza Health Ministry Says Death Toll Tops 24,100
Voice of America

Gaza Officials Say Dozens Killed In 'Intense' Israeli Strikes
International Business Times

Gaza officials say dozens killed in 'intense' Israeli strikes

Gaza officials say dozens killed in 'intense' Israeli strikes

emergentmethods.ai

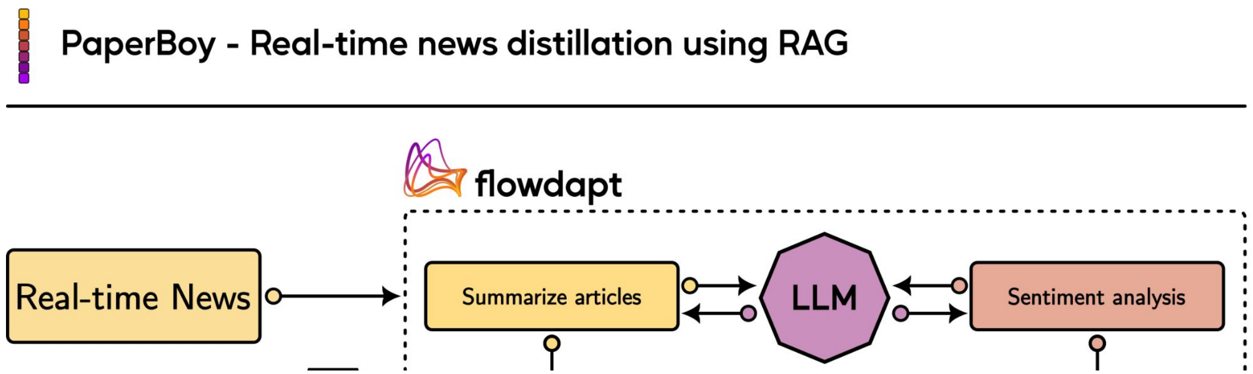
flowdapt.ai

asknews.app

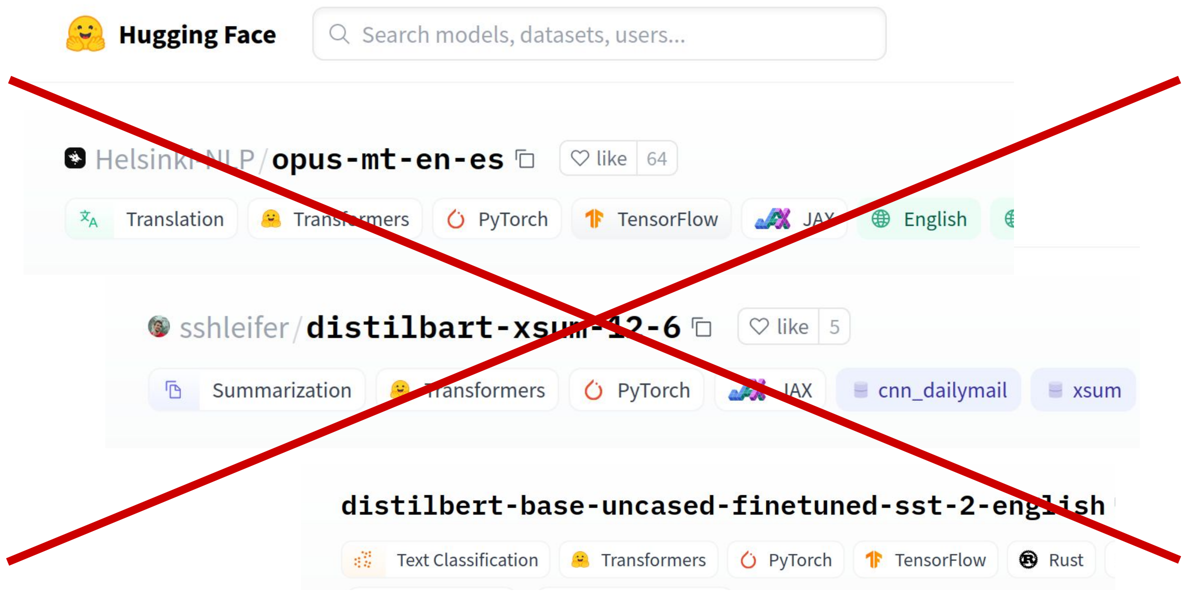
2024-01-18 Grenoble DS meetup: Production-scale RAG for Real-Time News Distillation

Context engineering - distilling context

1. Scrape 1k sources every 5 minutes for title + full content
2. Translate, summarize, extract entities, run sentiment analysis



Traditional pipeline (before Llama2)

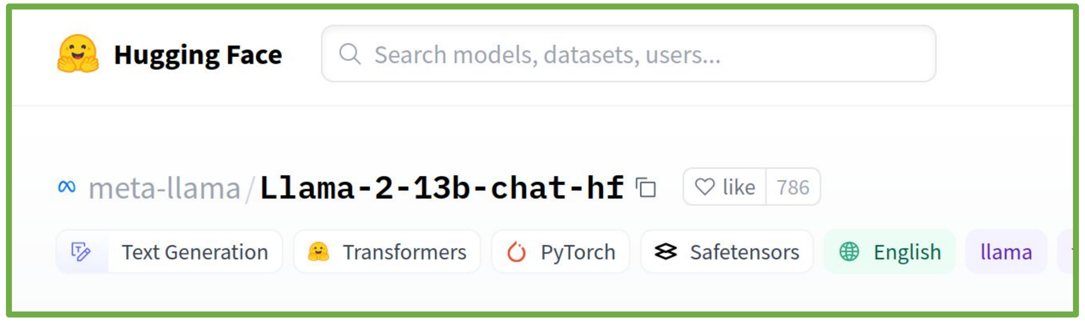

Hugging Face
Search models, datasets, users...

Helsinki-NLP/opus-mt-en-es
like 64
Translation · Transformers · PyTorch · TensorFlow · JAX · English

sshleifer/distilbart-xsum-12-6
like 5
Summarization · Transformers · PyTorch · JAX · cnn_dailymail · xsum

distilbert-base-uncased-finetuned-sst-2-english
Text Classification · Transformers · PyTorch · TensorFlow · Rust

After Llama2


Hugging Face
Search models, datasets, users...

meta-llama/Llama-2-13b-chat-hf
like 786
Text Generation · Transformers · PyTorch · Safetensors · English · llama

- Bigger context window (4096 tokens)
- more flexible output (custom analyses)
- Richer summarization
- Heavier, requires GPU
- Requires prompt engineering

emergentmethods.ai

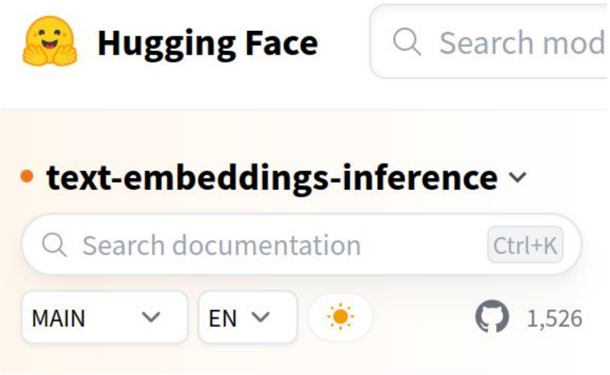
flowdapt.ai

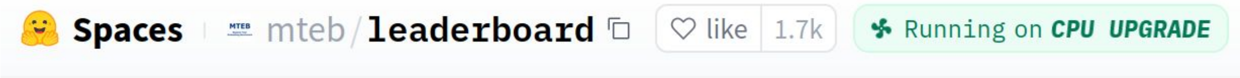
asknews.app

Emergent Methods
www.emergentmethods.ai

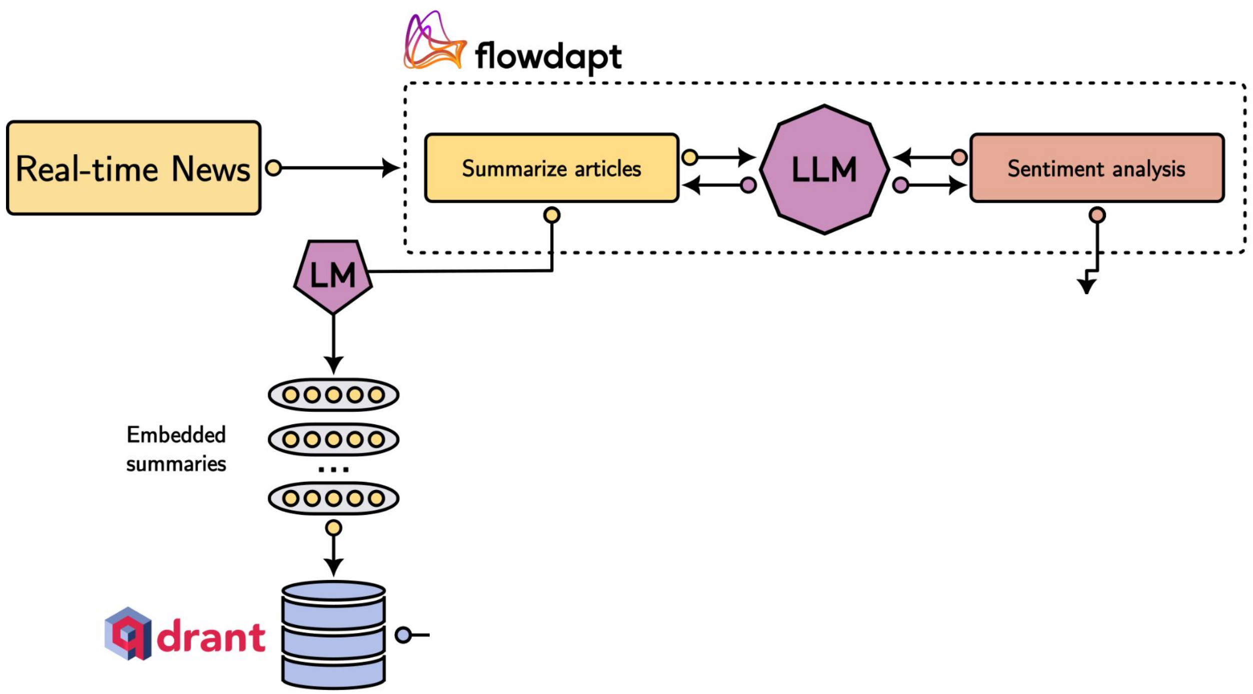
Context engineering

1. Embed enrichments into the vector database





PaperBoy - Real-time news distillation using RAG



emergentmethods.ai

flowdapt.ai

asknews.app

Context engineering

1. Embed enrichments into the vector database
   a. Use semantics that you expect to search on in the future
   b. In data science terms: stay in a single parameter space

User query examples:

Not semantically similar:
User: “What is going on in NYC government?”

!=

Ambiguous:
User: “Why did *they* change the rules?”

Embedded page:

title: NY May Change Village Formation Rules, Potentially Affecting Hasidic Village Plans

summary: Two bills in New York aim to change the rules for forming a village, which could potentially derail plans for …

source: Yahoo published: December 05 2023 08:00

people: Kathy Hochul

organizations: the New York State Conference of Mayors and Municipal Officials

places: Ateres, Orange County, Orange, New York, Seven Springs, Sullivan County

keywords: New York, Village, Hasidic, Governor, Reform, Population

classification: Government

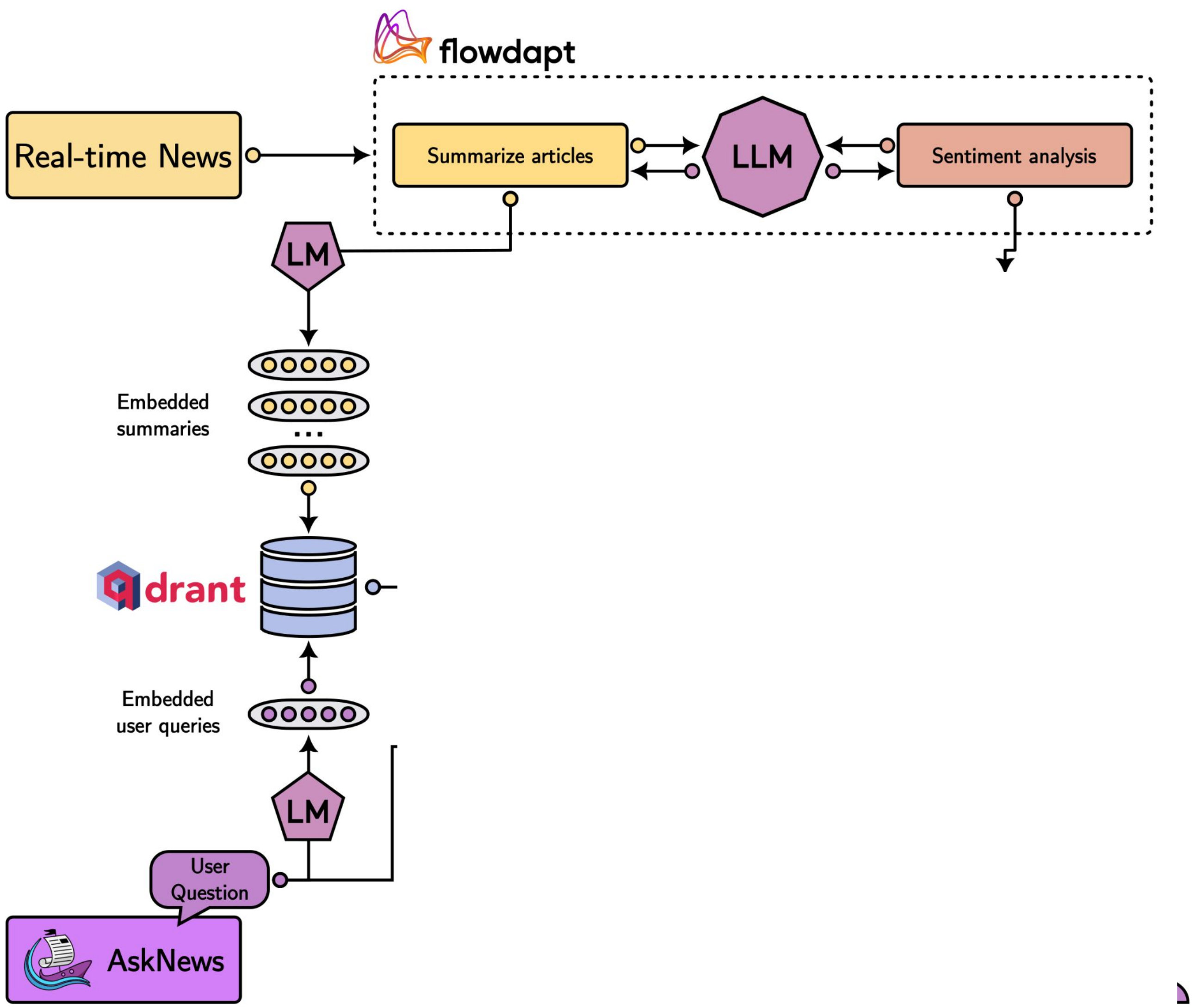
Context engineering

1. Embed enrichments into the vector database
    a. Use semantics that you expect to search on in the future
    b. In data science terms: stay in a single parameter space
    c. Hypothetical Document Embeddings (HyDE)

Not semantically similar:
User: “What is going on in NYC government?”

Use our Reasoning engine (the LLM) to obtain a better search (HyDE):
“Write me a hypothetical summary for an article that answers the question “What is going on in NYC government?”

PaperBoy - Real-time news distillation using RAG



flowdapt
Real-time News
Summarize articles
LLM
Sentiment analysis
LM
Embedded summaries
drant
Embedded user queries
LM
User Question
AskNews

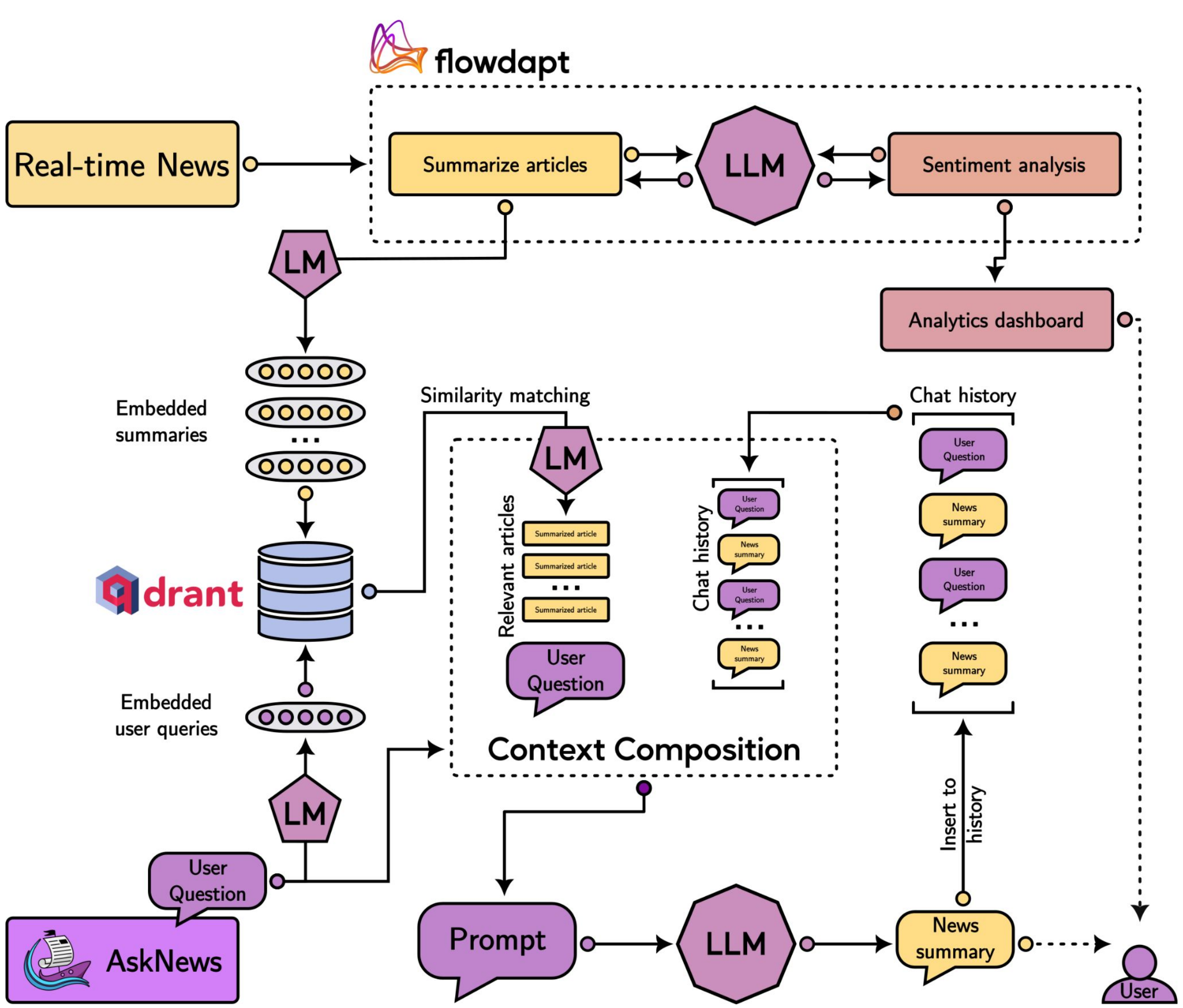
Context engineering

1. Embed enrichments into the vector database
    a. Use semantics that you expect to search on in the future
    b. In data science terms: stay in a single parameter space
    c. Hypothetical Document Embeddings (HyDE)

Ambiguous:
User: “Why did *they* change the rules?”

Use our Reasoning engine (the LLM) to obtain a better search (HyDE):
“Given the historic conversation, and the current question, write me an explicit question without ambiguity.

Use the explicit question to write me a hypothetical article that answers the explicit question.”



PaperBoy - Real-time news distillation using RAG

emergentmethods.ai

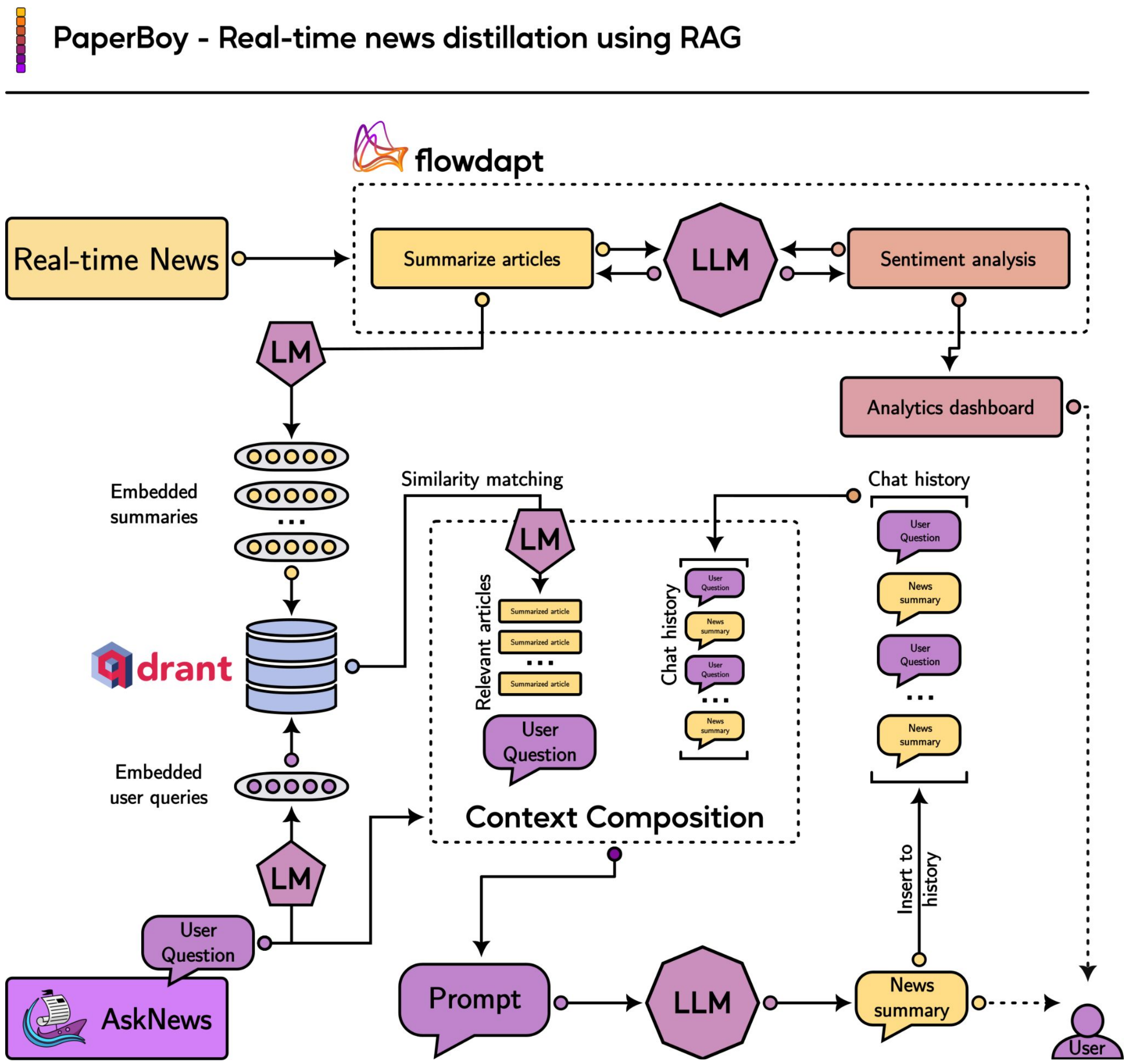
flowdapt.ai

asknews.app

Context engineering

Recap:

1. **Scrape news**, enrich, embed, store in vectorDB (continuously in the background)
2. **Reformat user query** into the embedding structure used for your vectorDB parameter space (using LLMs and HyDE, or any other creative method)
3. **Run similarity search** on the reformatted user query, take the extracted data and inject into final-call prompt
4. **Inject conversation** into final-call prompt
5. Stream response to user!



emergentmethods.ai flowdapt.ai asknews.app

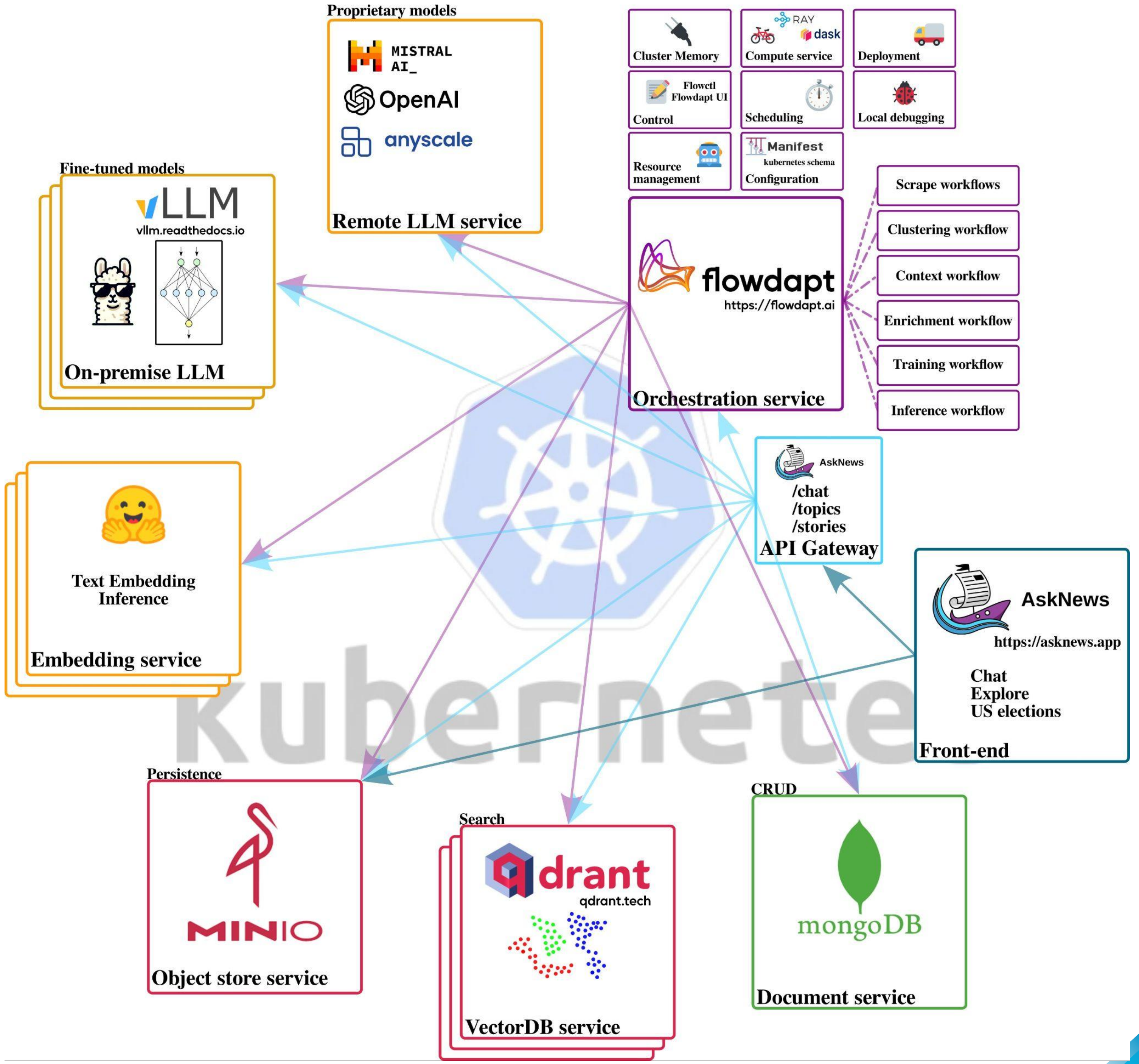
Scaling Context Engineering

Emergent Methods

Scaling real-time context engineering…

Context engineer's best friends:

- Microservice orchestration □
- Single Responsibility Principle (SRP)
- Open-source projects
- Kubernetes
- A strong DevOps…

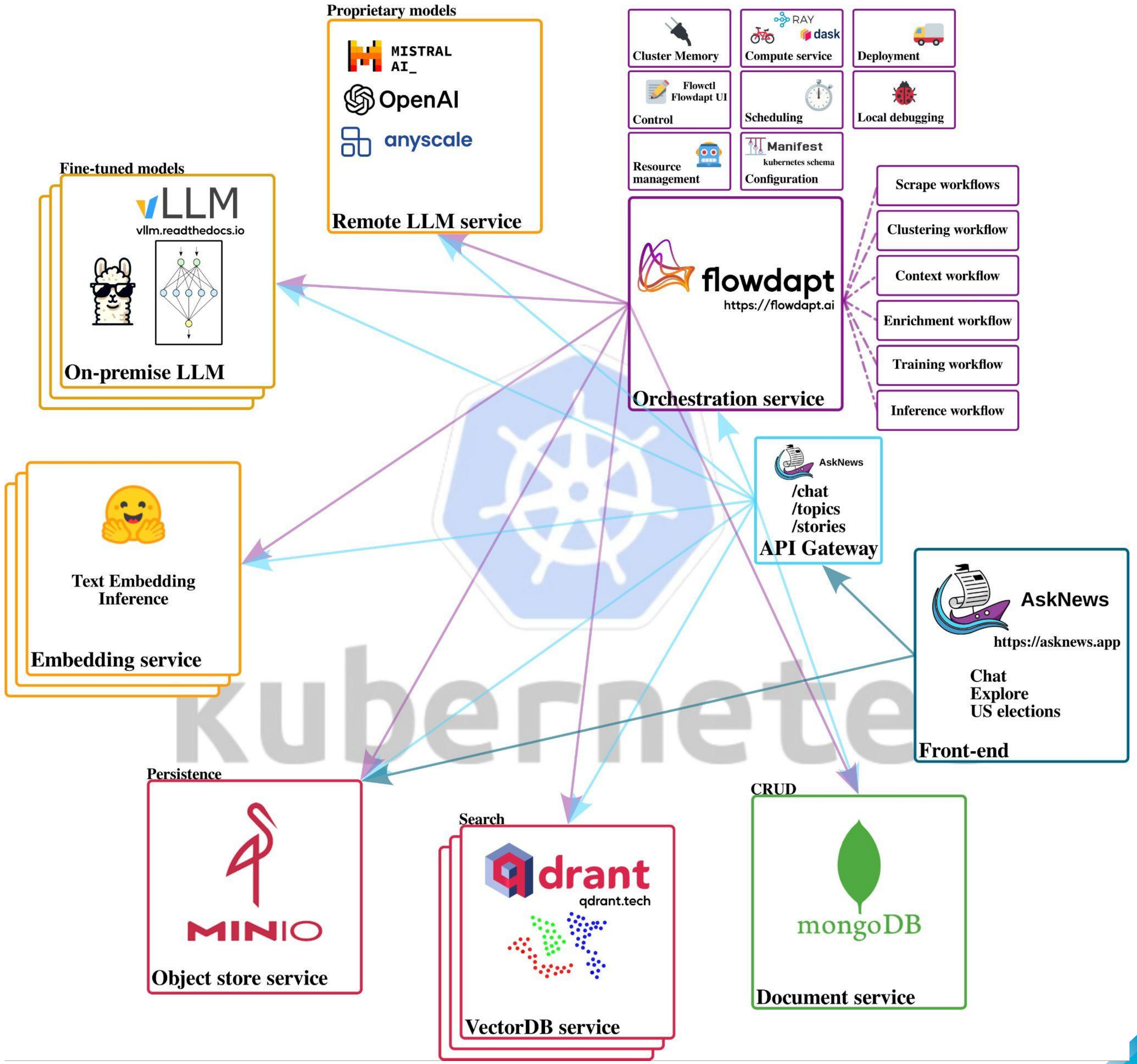


emergentmethods.ai flowdapt.ai asknews.app

Scaling real-time context engineering…

Context engineer’s best friends:

- Microservice orchestration □
- Single Responsibility Principle (SRP) ☝️
- Open-source projects
- Kubernetes
- A strong DevOps…

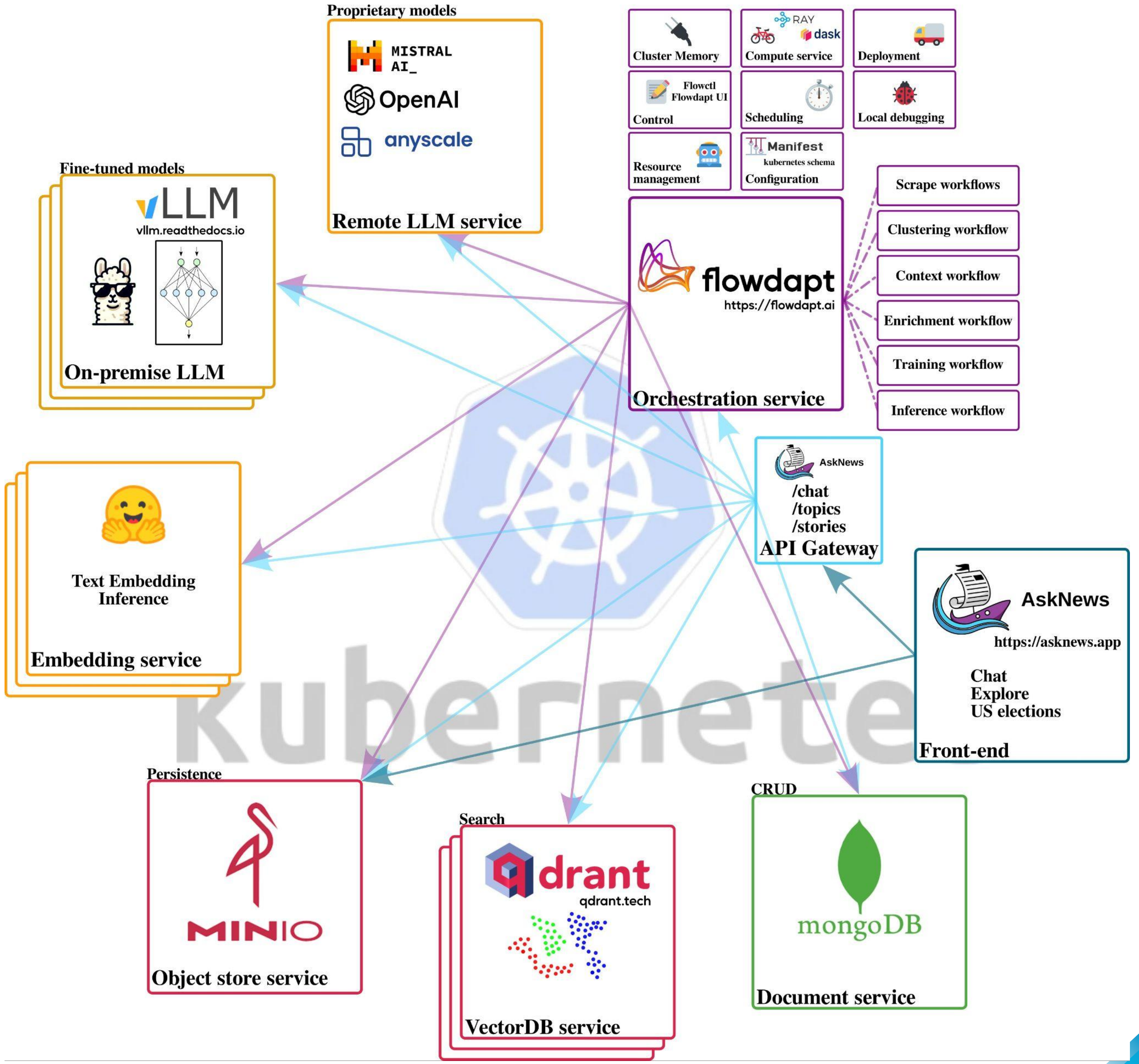


emergentmethods.ai
flowdapt.ai
asknews.app

Scaling real-time context engineering…

Context engineer’s best friends:

- Microservice orchestration □
- Single Responsibility Principle (SRP) ☝️
- Open-source projects
- Kubernetes
- A strong DevOps…



emergentmethods.ai flowdapt.ai asknews.app

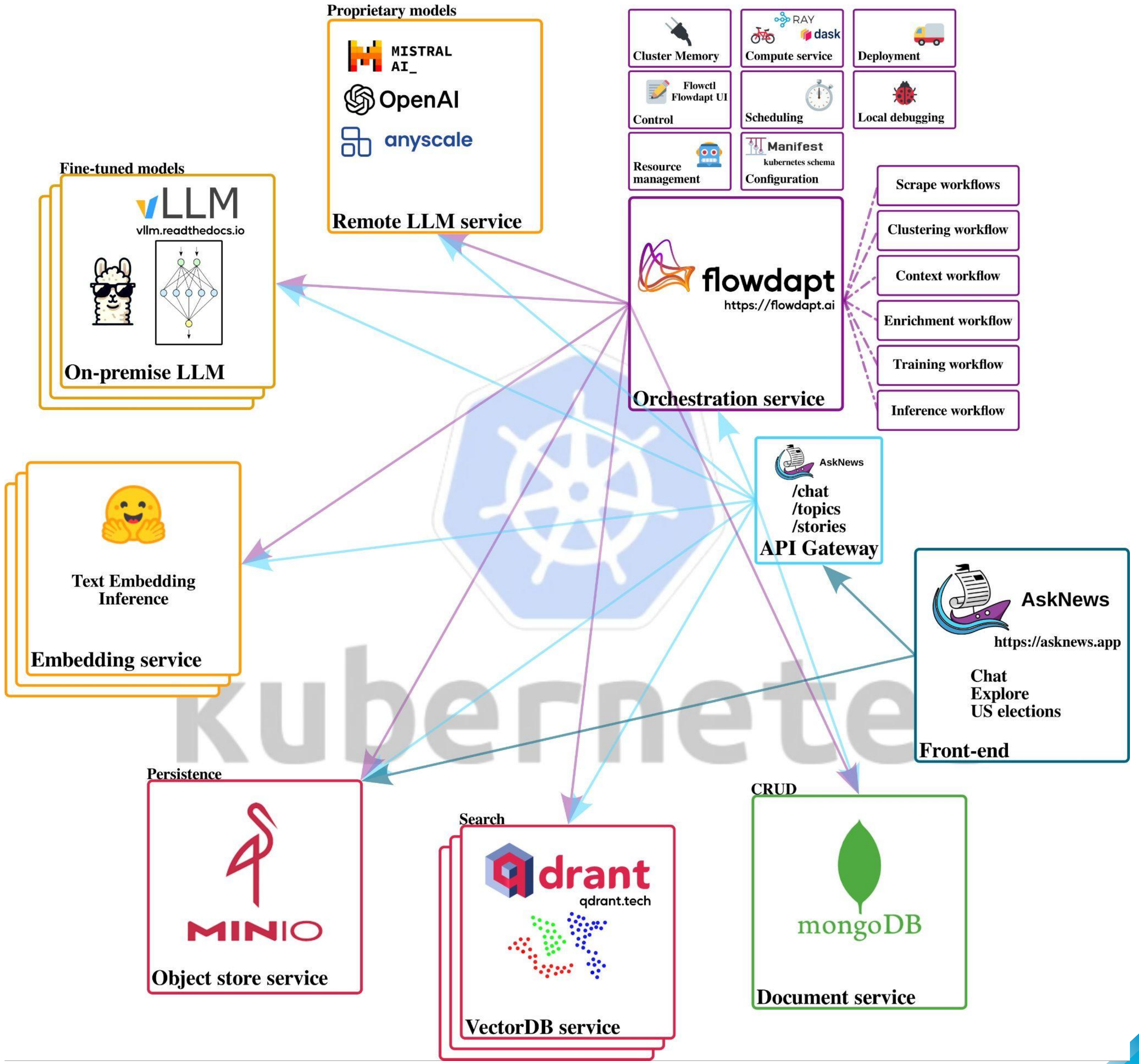
2024-01-18 Grenoble DS meetup: Production-scale RAG for Real-Time News Distillation

Emergent Methods
www.emergentmethods.ai

Scaling real-time context engineering…

Context engineer’s best friends:

- Microservice orchestration □
- Single Responsibility Principle (SRP) ☝
- Open-source projects
- Kubernetes
- A strong DevOps…


Proprietary models
MISTRAL AI_
OpenAI
anyscale
Remote LLM service

Fine-tuned models
vLLM
vllm.readthedocs.io
On-premise LLM

Cluster Memory
RAY dask
Compute service
Deployment
Flowctl Flowdapt UI
Control
Scheduling
Local debugging
Resource management
Manifest kubernetes schema
Configuration

flowdapt
https://flowdapt.ai
Orchestration service

Scrape workflows
Clustering workflow
Context workflow
Enrichment workflow
Training workflow
Inference workflow

Text Embedding Inference
Embedding service

AskNews
/chat
/topics
/stories
API Gateway

AskNews
https://asknews.app
Chat
Explore
US elections
Front-end

kubernetes

Persistence
MINIO
Object store service

Search
qdrant
qdrant.tech
VectorDB service

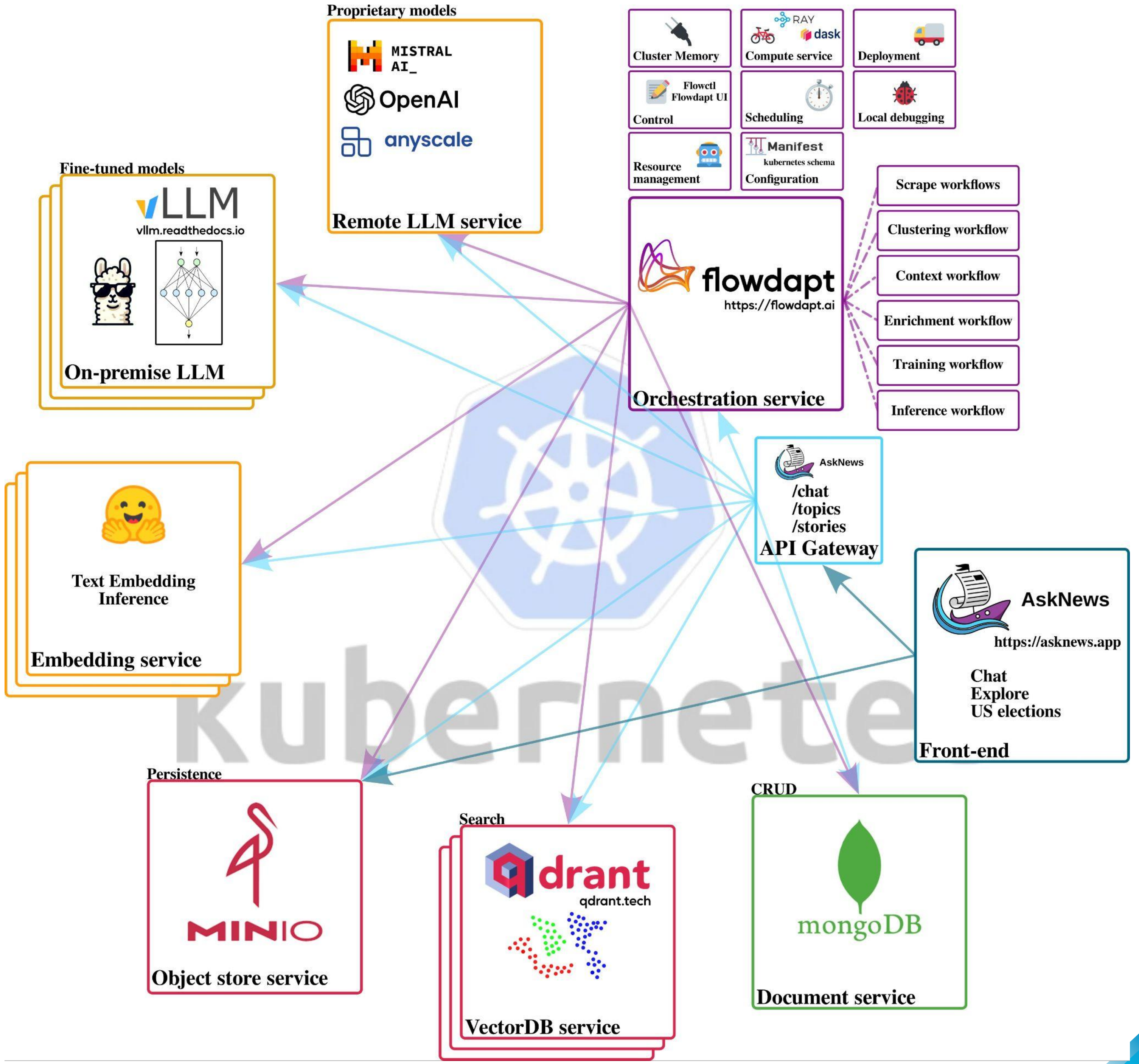
CRUD
mongoDB
Document service

emergentmethods.ai flowdapt.ai asknews.app

Scaling real-time context engineering…

Context engineer’s best friends:

- Microservice orchestration □
- Single Responsibility Principle (SRP) ☝️
- Open-source projects
- Kubernetes
- A strong DevOps… 💪



Proprietary models

MISTRAL AI_
OpenAI
anyscale

Remote LLM service

Fine-tuned models

vLLM
vllm.readthedocs.io

On-premise LLM

Cluster Memory

RAY
dask
Compute service

Deployment

Flowctl
Flowdapt UI
Control

Scheduling

Local debugging

Resource management

Manifest
kubernetes schema
Configuration

flowdapt
https://flowdapt.ai

Orchestration service

Scrape workflows

Clustering workflow

Context workflow

Enrichment workflow

Training workflow

Inference workflow

Text Embedding Inference

Embedding service

AskNews
/chat
/topics
/stories
API Gateway

AskNews
https://asknews.app

Chat
Explore
US elections

Front-end

kubernetes

Persistence

MINIO

Object store service

Search

qdrant
qdrant.tech

VectorDB service

CRUD

mongoDB

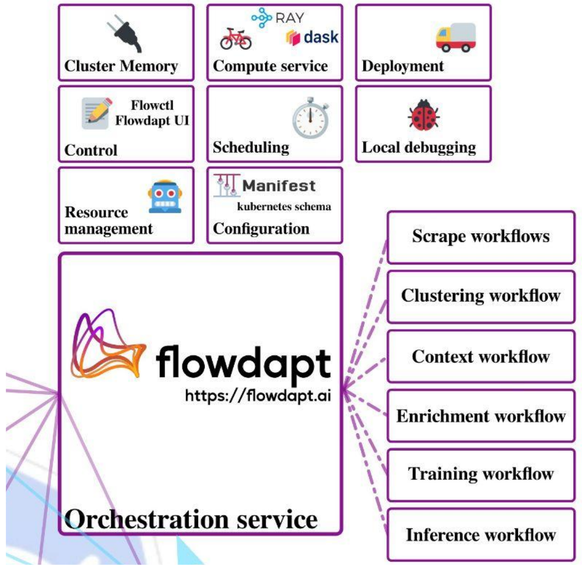
Document service

emergentmethods.ai flowdapt.ai asknews.app

Scaling real-time context engineering…

Context engineering == microservice orchestration!

- Priorities:
    - 🚲 highly parallelized compute efficiency,
    - automatic resource management,
    - rapid (local) prototyping and debuggability,
    - intuitive cluster-wide data sharing methods,
    - easy scheduling,
    - live configurability, and
    - deployment cycle efficiency.



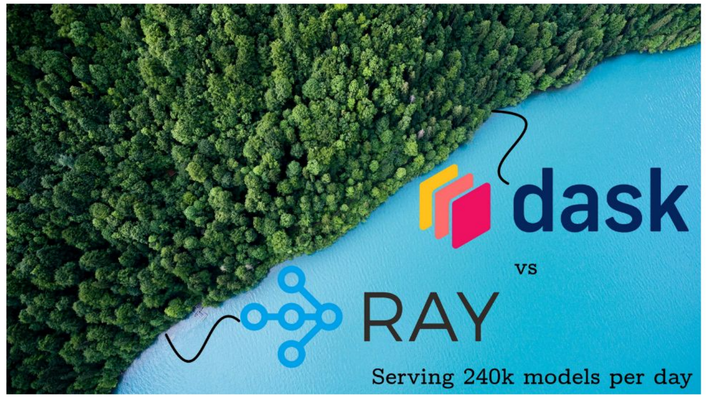
2024-01-18 Grenoble DS meetup: Production-scale RAG for Real-Time News Distillation

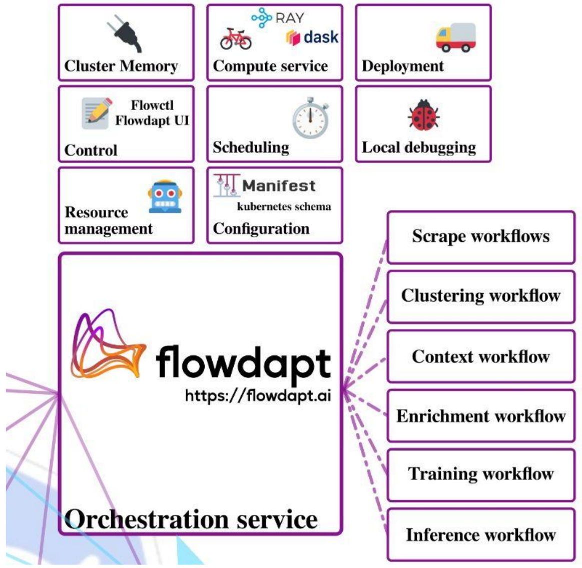
Scaling real-time context engineering…

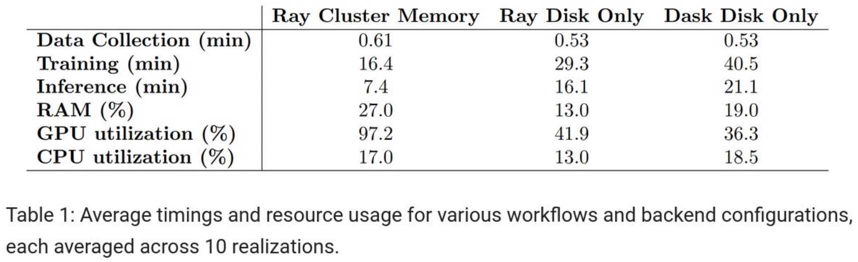
Context engineering == microservice orchestration!

- Priorities:
  - ○ 🚲 highly parallelized compute efficiency,
  - ○ automatic resource management,
  - ○ rapid (local) prototyping and debuggability,
  - ○ intuitive cluster-wide data sharing methods,
  - ○ easy scheduling,
  - ○ live configurability, and
  - ○ deployment cycle efficiency.

Ray vs Dask: Lessons learned serving 240k models per day in real-time ¶







	Ray Cluster Memory	Ray Disk Only	Dask Disk Only
Data Collection (min)	0.61	0.53	0.53
Training (min)	16.4	29.3	40.5
Inference (min)	7.4	16.1	21.1
RAM (%)	27.0	13.0	19.0
GPU utilization (%)	97.2	41.9	36.3
CPU utilization (%)	17.0	13.0	18.5

Table 1: Average timings and resource usage for various workflows and backend configurations, each averaged across 10 realizations.

https://docs.flowdapt.ai/examples/rayvsdask/

flowdapt

RAY dask

Flowdapt is vanilla Python 🐍
Zero code change between Ray or Dask backend 🚀

emergentmethods.ai

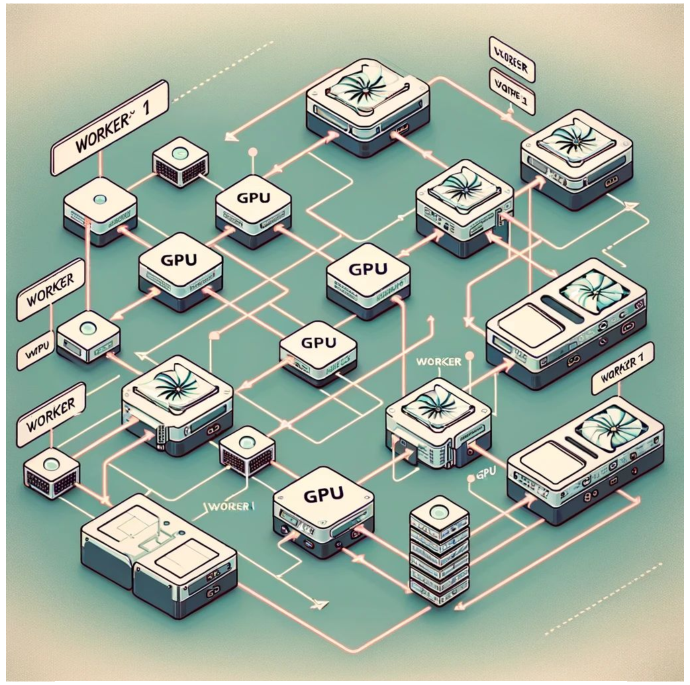
flowdapt.ai

asknews.app

Scaling real-time context engineering…

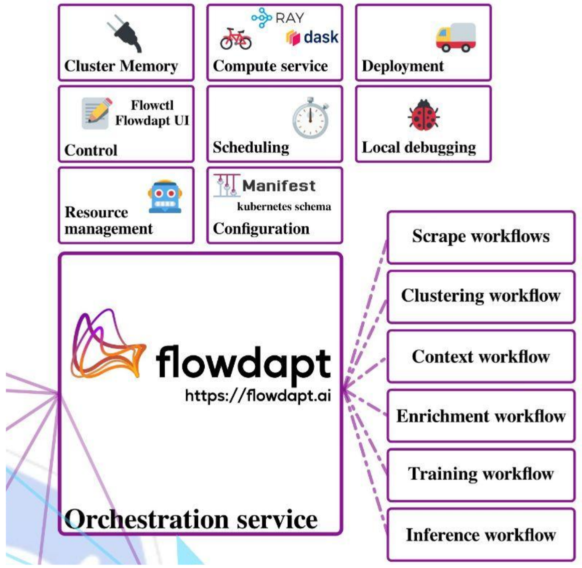
Context engineering == microservice orchestration!

- Priorities:
  - ○ 🚲 highly parallelized compute efficiency,
  - ○ 🤖 automatic resource management,
  - ○ rapid (local) prototyping and debuggability,
  - ○ intuitive cluster-wide data sharing methods,
  - ○ easy scheduling,
  - ○ live configurability, and
  - ○ deployment cycle efficiency.



flowdapt

RAY dask



Cluster Memory

Compute service

Deployment

Flowctl
Flowdapt UI
Control

Scheduling

Local debugging

Resource management

Manifest
kubernetes schema
Configuration

flowdapt
https://flowdapt.ai
Orchestration service

Scrape workflows

Clustering workflow

Context workflow

Enrichment workflow

Training workflow

Inference workflow

Flowdapt is vanilla Python 🐍
Zero code change between Ray or Dask backend 🚀

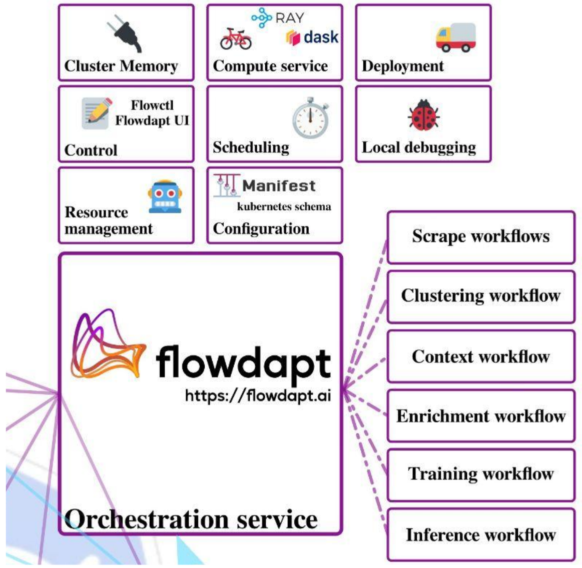
emergentmethods.ai flowdapt.ai asknews.app

Scaling real-time context engineering…

Context engineering == microservice orchestration!

- Priorities:
  - 🚲 highly parallelized compute efficiency,
  - 🤖 automatic resource management,
  - 🐞 rapid (local) prototyping and debuggability,
  - intuitive cluster-wide data sharing methods,
  - easy scheduling,
  - live configurability, and
  - deployment cycle efficiency.

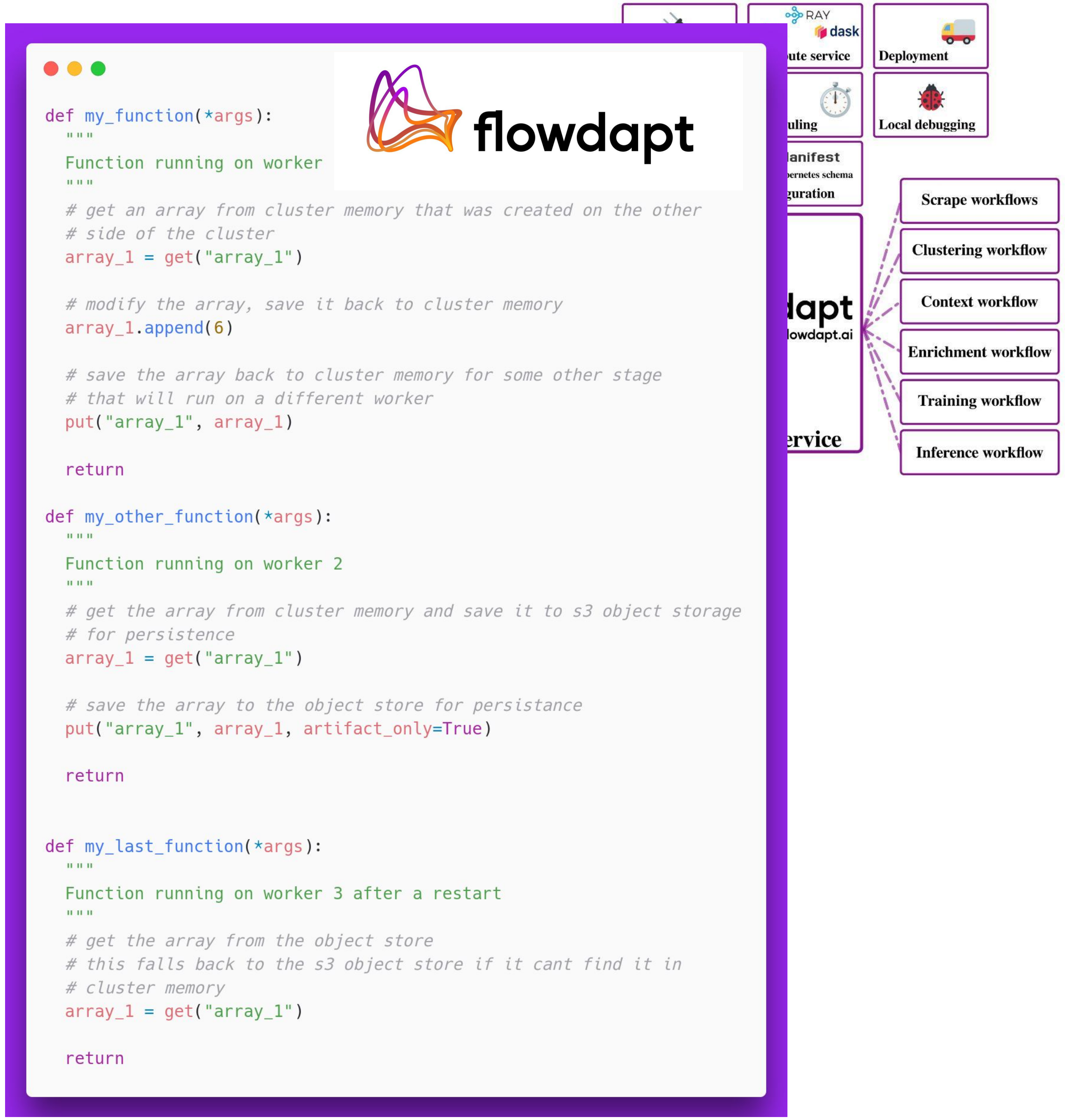
flowdapt



Scaling real-time context engineering…

Context engineering == microservice orchestration!

- Priorities:
  - 🚲 highly parallelized compute efficiency,
  - 🤖 automatic resource management,
  - 🐞 rapid (local) prototyping and debuggability,
  - 🔌 intuitive cluster-wide data sharing methods,
  - easy scheduling,
  - live configurability, and
  - deployment cycle efficiency.



```python
def my_function(*args):
  """
  Function running on worker
  """
  # get an array from cluster memory that was created on the other
  # side of the cluster
  array_1 = get("array_1")

  # modify the array, save it back to cluster memory
  array_1.append(6)

  # save the array back to cluster memory for some other stage
  # that will run on a different worker
  put("array_1", array_1)

  return

def my_other_function(*args):
  """
  Function running on worker 2
  """
  # get the array from cluster memory and save it to s3 object storage
  # for persistance
  array_1 = get("array_1")

  # save the array to the object store for persistance
  put("array_1", array_1, artifact_only=True)

  return


def my_last_function(*args):
  """
  Function running on worker 3 after a restart
  """
  # get the array from the object store
  # this falls back to the s3 object store if it cant find it in
  # cluster memory
  array_1 = get("array_1")

  return
```

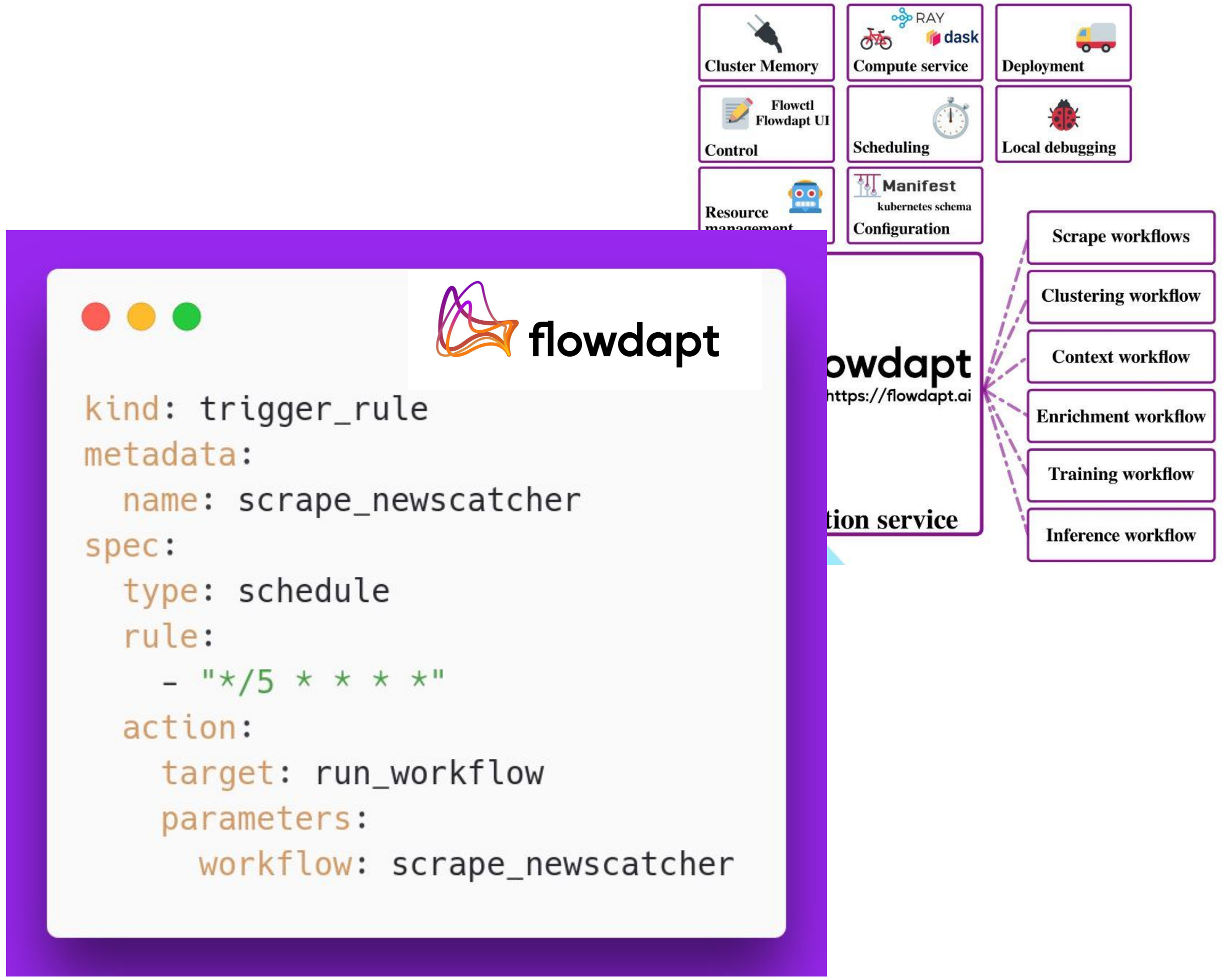
emergentmethods.ai flowdapt.ai asknews.app

Emergent Methods
www.emergentmethods.ai

Scaling real-time context engineering…

Context engineering == microservice orchestration!

- Priorities:
    - ○ 🚲 highly parallelized compute efficiency,
    - ○ 🤖 automatic resource management,
    - ○ 🐞 rapid (local) prototyping and debuggability,
    - ○ 🔌 intuitive cluster-wide data sharing methods,
    - ○ ⏱️ easy scheduling,
    - ○ live configurability, and
    - ○ deployment cycle efficiency.



```yaml
kind: trigger_rule
metadata:
  name: scrape_newscatcher
spec:
  type: schedule
  rule:
    - "*/5 * * * *"
  action:
    target: run_workflow
    parameters:
      workflow: scrape_newscatcher
```

emergentmethods.ai
flowdapt.ai
asknews.app

2024-01-18 Grenoble DS meetup: Production-scale RAG for Real-Time News Distillation

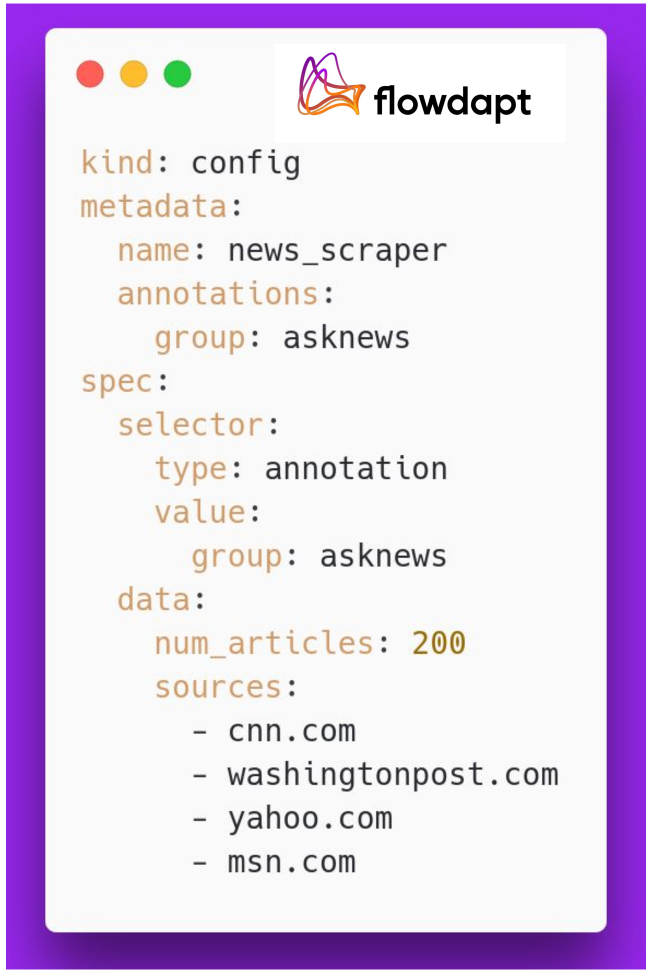
Emergent Methods
www.emergentmethods.ai

Scaling real-time context engineering…

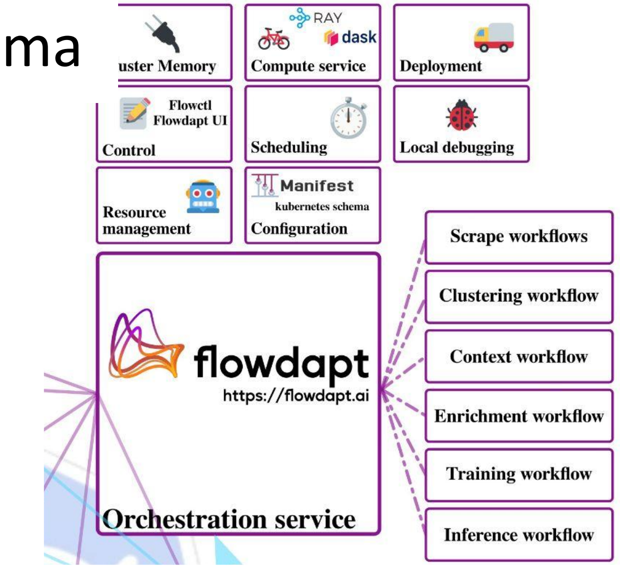
Context engineering == microservice orchestration!

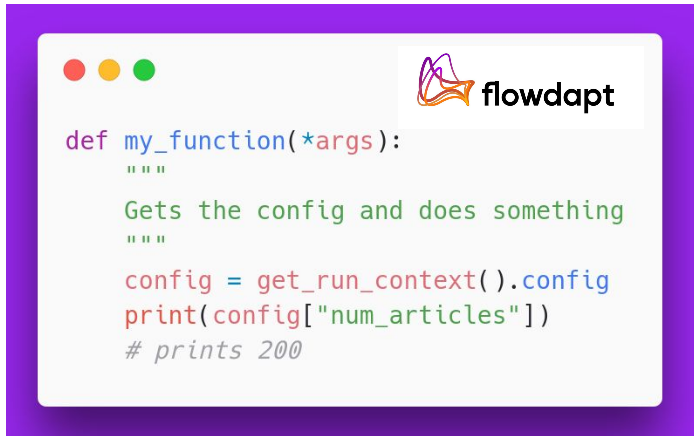
- Priorities:
  - ○ 🚲 highly parallelized compute efficiency,
  - ○ 🤖 automatic resource management,
  - ○ 🐞 rapid (local) prototyping and debuggability,
  - ○ 🔌 intuitive cluster-wide data sharing methods,
  - ○ ⏱️ easy scheduling,
  - ○ 📝 live configurability, and
  - ○ deployment cycle efficiency.

Kubernetes-style resource schema


flowdapt

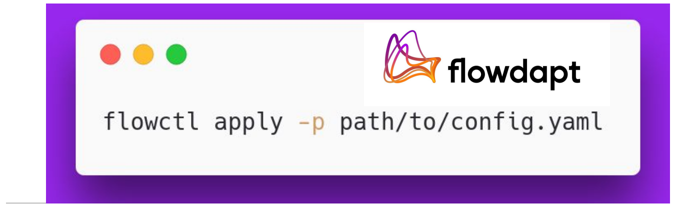
```yaml
kind: config
metadata:
  name: news_scraper
  annotations:
    group: asknews
spec:
  selector:
    type: annotation
    value:
      group: asknews
  data:
    num_articles: 200
    sources:
      - cnn.com
      - washingtonpost.com
      - yahoo.com
      - msn.com
```




flowdapt

```python
def my_function(*args):
    """
    Gets the config and does something
    """
    config = get_run_context().config
    print(config["num_articles"])
    # prints 200
```

kubectl-style command schema


flowdapt

```
flowctl apply -p path/to/config.yaml
```

emergentmethods.ai

flowdapt.ai

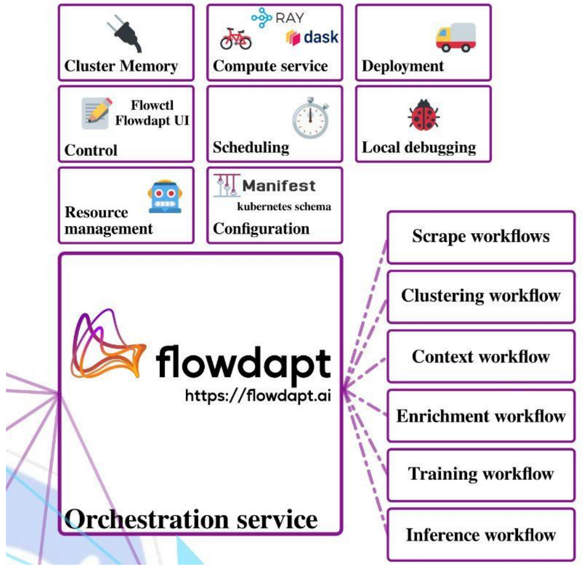
asknews.app

Scaling real-time context engineering…

Context engineering == microservice orchestration!

- Priorities:
  - 🚲 highly parallelized compute efficiency,
  - 🤖 automatic resource management,
  - 🐞 rapid (local) prototyping and debuggability,
  - 🔌 intuitive cluster-wide data sharing methods,
  - ⏱️ easy scheduling,
  - 📝 live configurability, and
  - 🚚 deployment cycle efficiency.

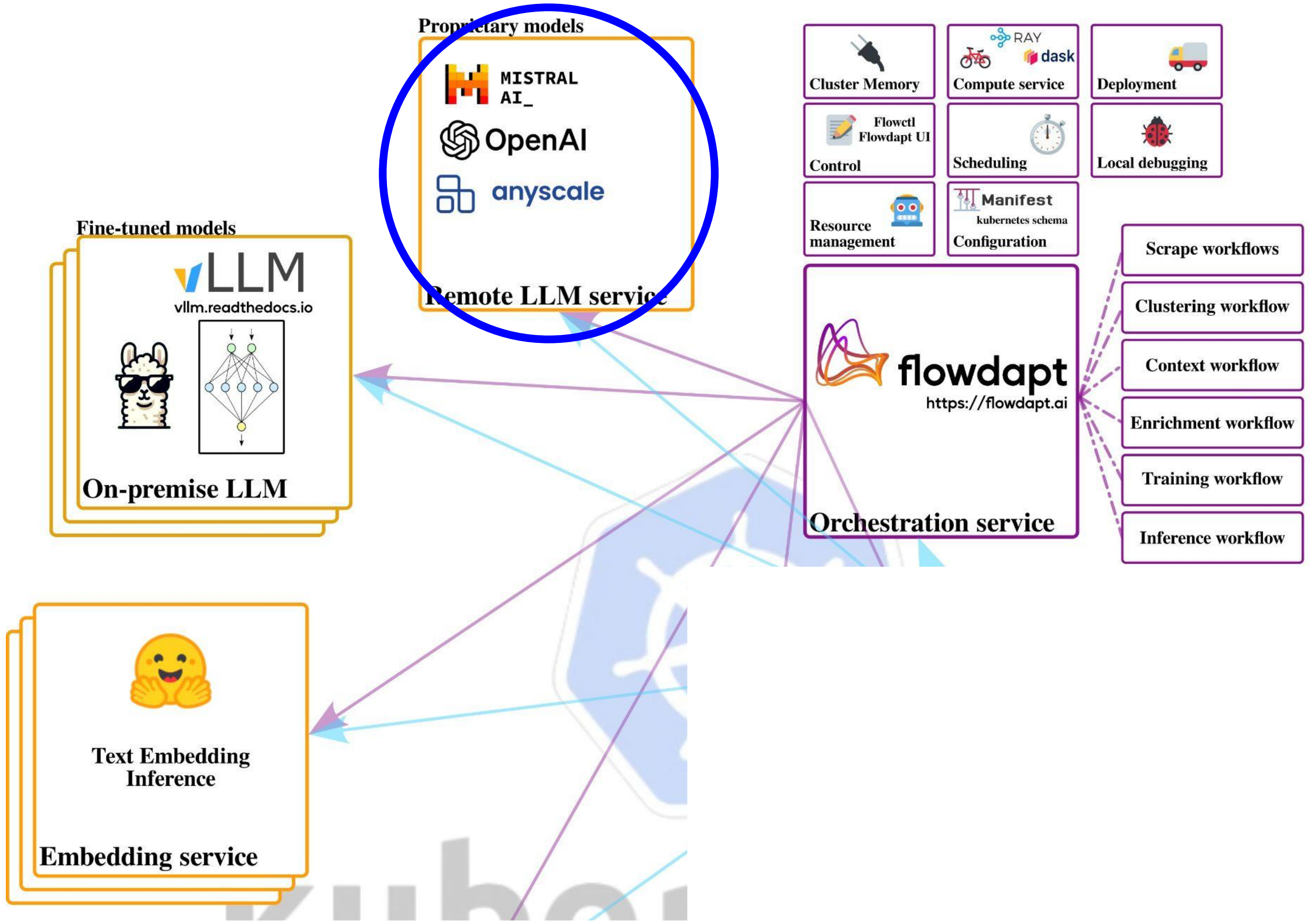
flowdapt



Scaling real-time context engineering…

The “reasoning engine”:

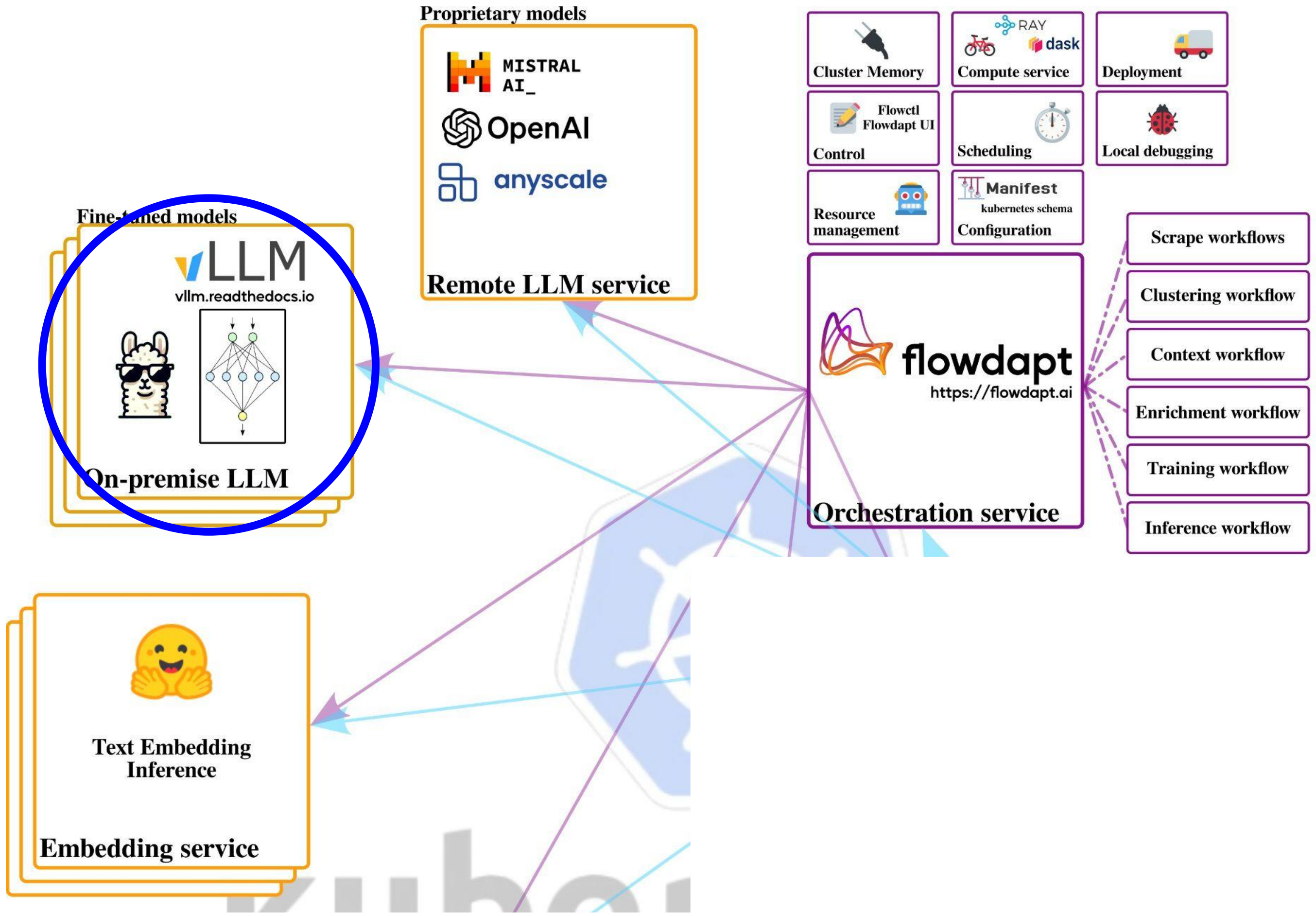
- Proprietary models
  - rate-limits 🐢
  - prompt security 🔓
  - data security 🔓
  - model drift 😩
  - safety limitations 🦺
  - intelligence 🧠



Scaling real-time context engineering…

The “reasoning engine”:

- Proprietary models
  - rate-limits 🐢
  - prompt security 🔓
  - data security 🔓
  - model drift 😩
  - safety limitations 🦺
  - intelligence 🧠
- On-premise models
  - throughput 🛣️
  - prompt security 🔐
  - data security 🔐
  - model control 🎛️
  - reasoning 🦜



emergentmethods.ai

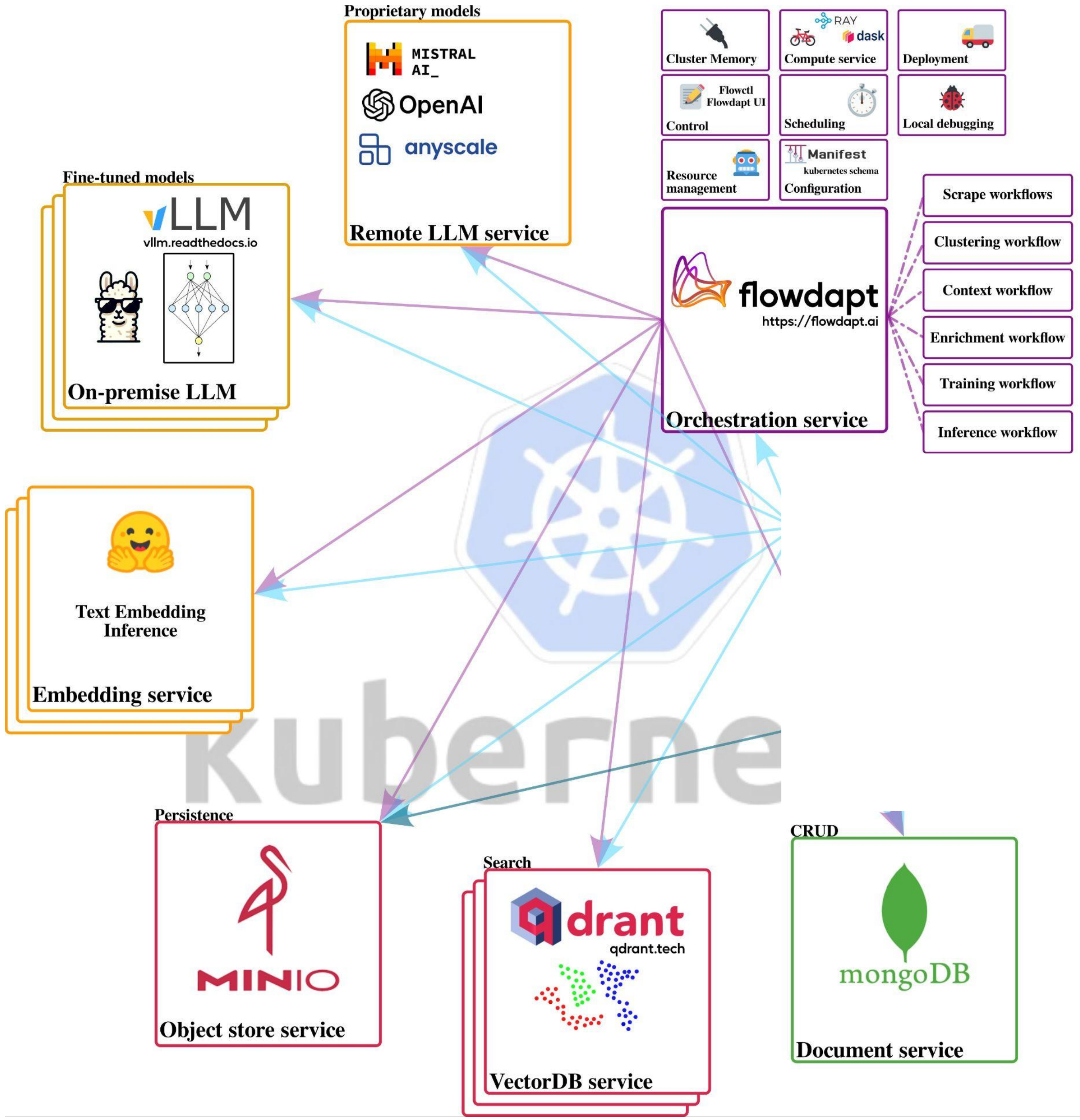
flowdapt.ai

asknews.app

Scaling real-time context engineering…

Persistence and search 🔍

- All-in-one database solutions
  - sacrifice performance for convenience
  - sacrifice accuracy for convenience



emergentmethods.ai

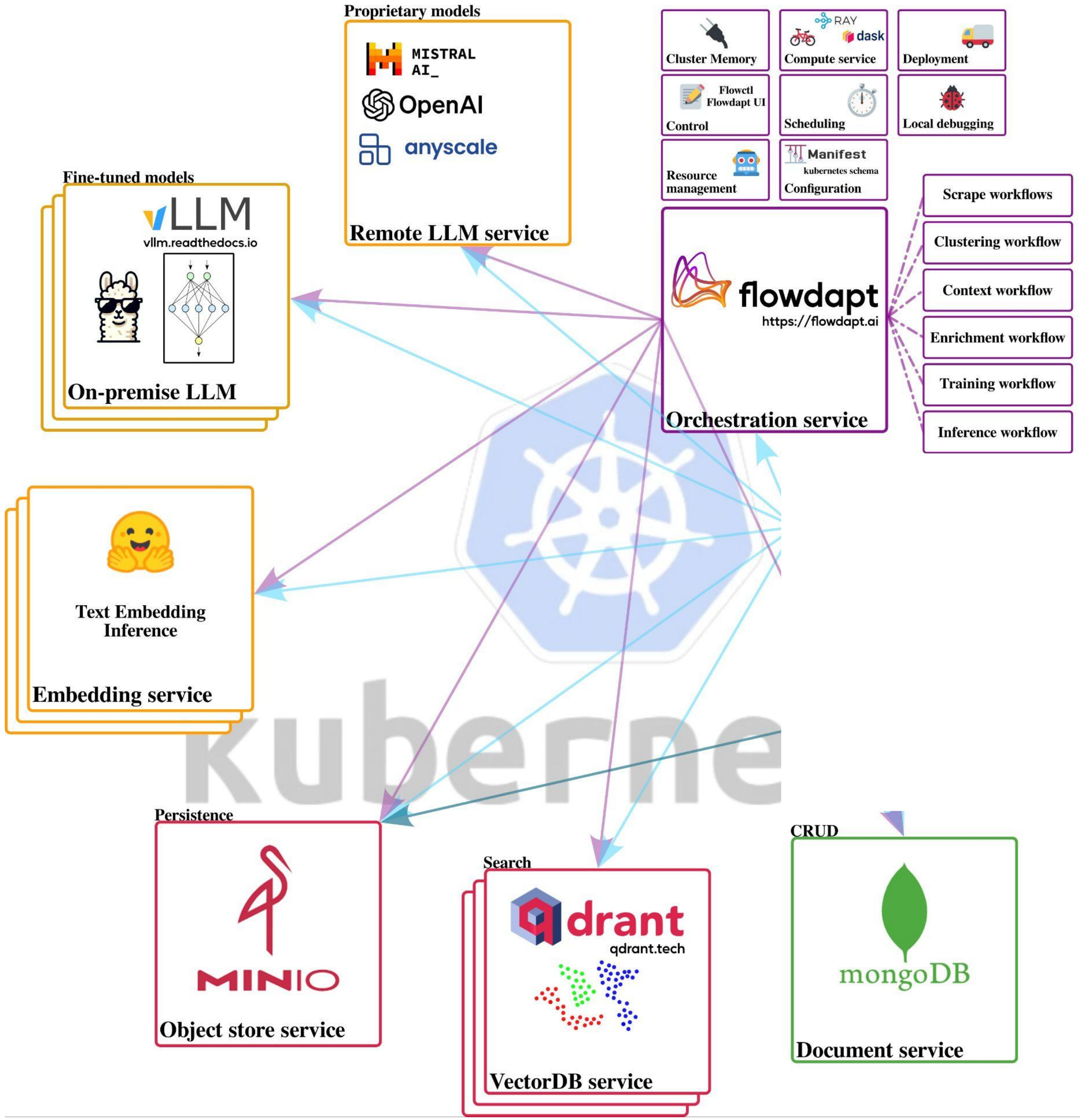
flowdapt.ai

asknews.app

Scaling real-time context engineering…

Persistence and search 🔍

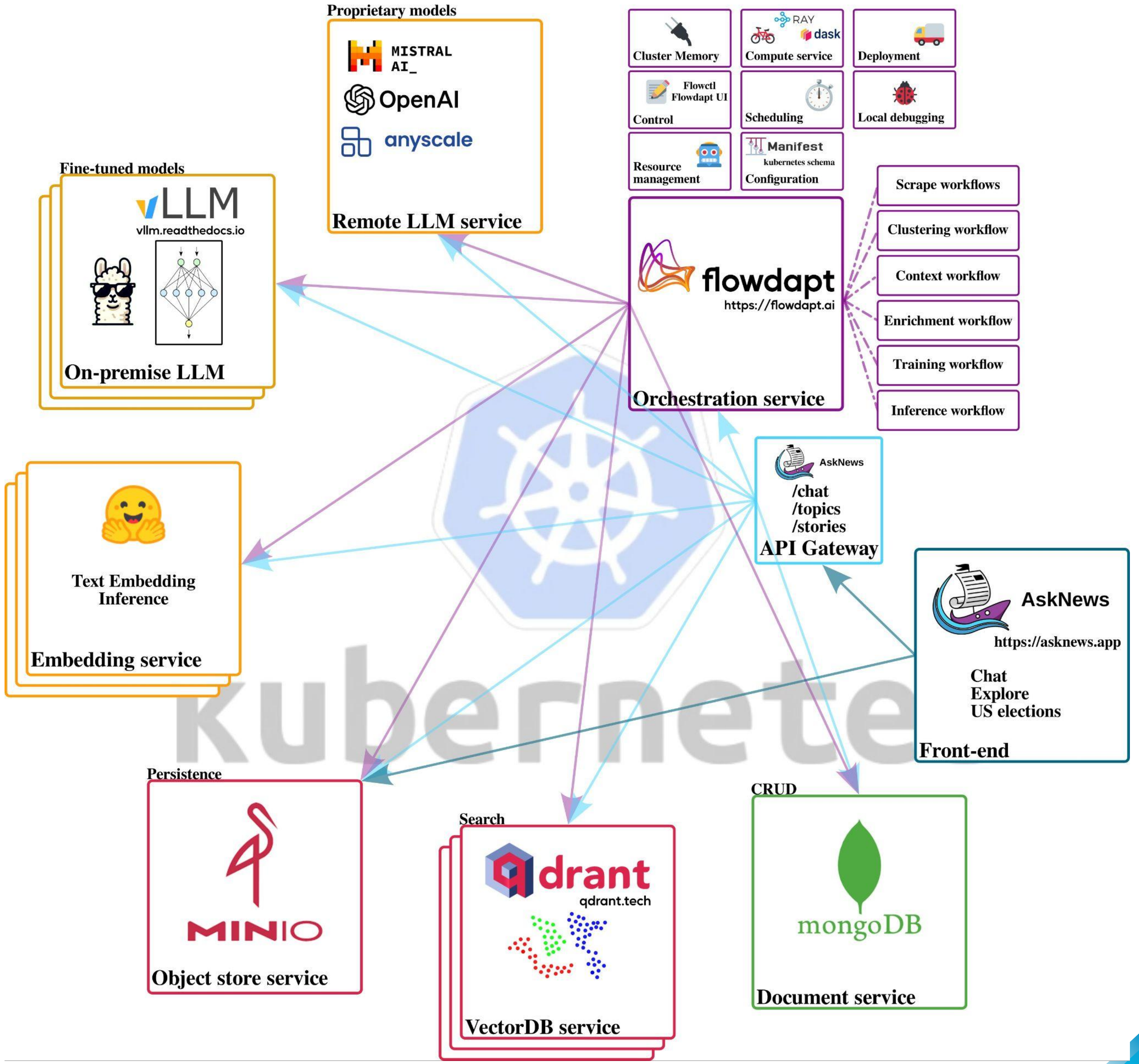
- All-in-one database solutions
  - sacrifice performance for convenience
  - sacrifice accuracy for convenience
- Single responsibility principle (SRP)
  - modularity
  - performance
  - accuracy



Scaling real-time context engineering…

Minimize latency 📶

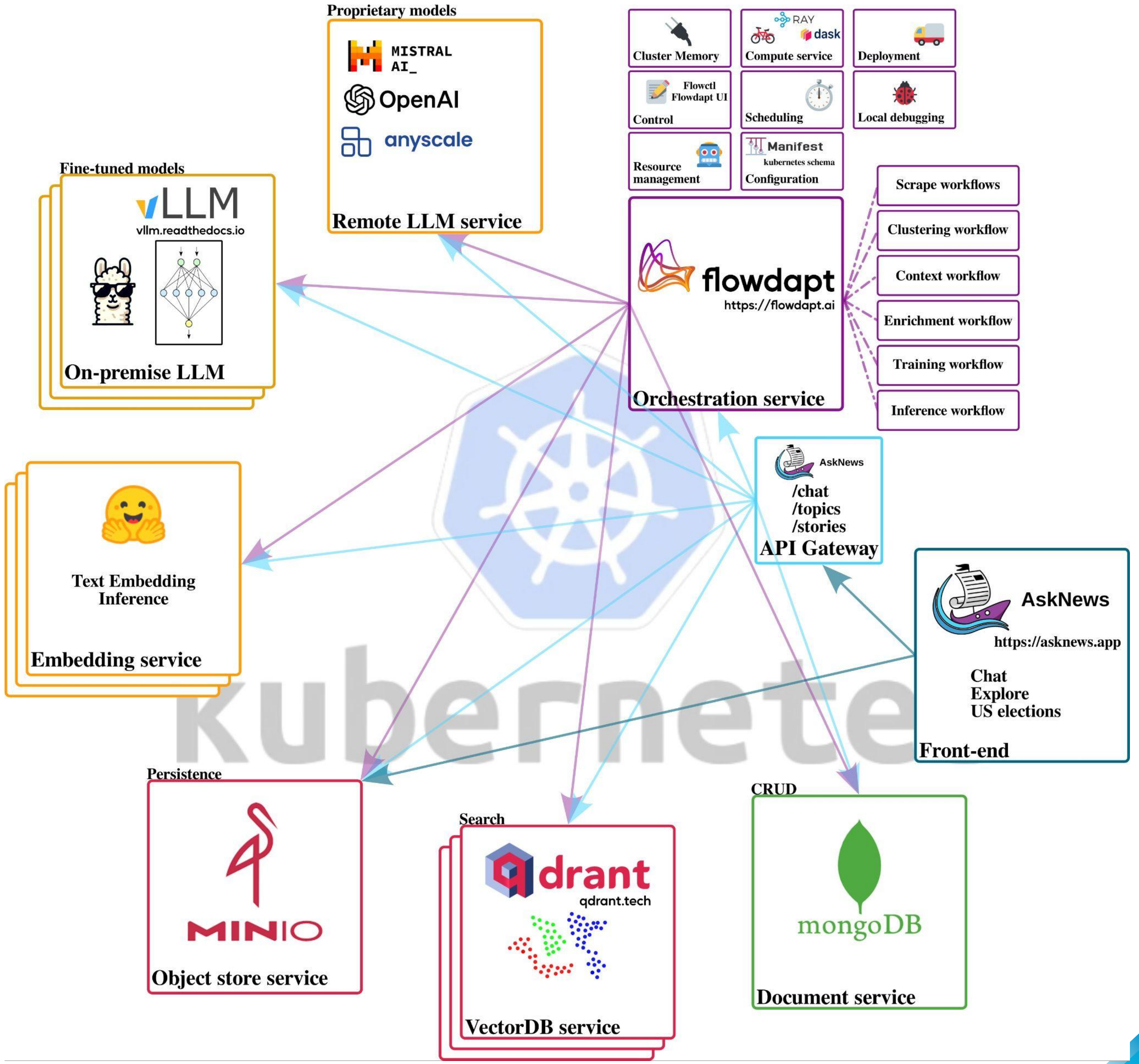
- Leverage intracluster communications for speed and security benefits



Scaling real-time context engineering…

Minimize latency 📶

- Leverage intracluster communications for speed and security benefits
- Design endpoints strictly on I/O in the API gateway:
  - go-lang is preferred
  - async FastAPI on pydantic v2.5 is performant and easier for typical ML engineers to navigate



Proprietary models

MISTRAL AI_
OpenAI
anyscale

Remote LLM service

Fine-tuned models

vLLM
vllm.readthedocs.io

On-premise LLM

Text Embedding Inference

Embedding service

Cluster Memory

RAY
dask
Compute service

Deployment

Flowctl
Flowdapt UI
Control

Scheduling

Local debugging

Resource management

Manifest
kubernetes schema
Configuration

flowdapt
https://flowdapt.ai

Orchestration service

Scrape workflows
Clustering workflow
Context workflow
Enrichment workflow
Training workflow
Inference workflow

AskNews
/chat
/topics
/stories
API Gateway

AskNews
https://asknews.app
Chat
Explore
US elections
Front-end

kubernetes

Persistence

MINIO

Object store service

Search

qdrant
qdrant.tech

VectorDB service

CRUD

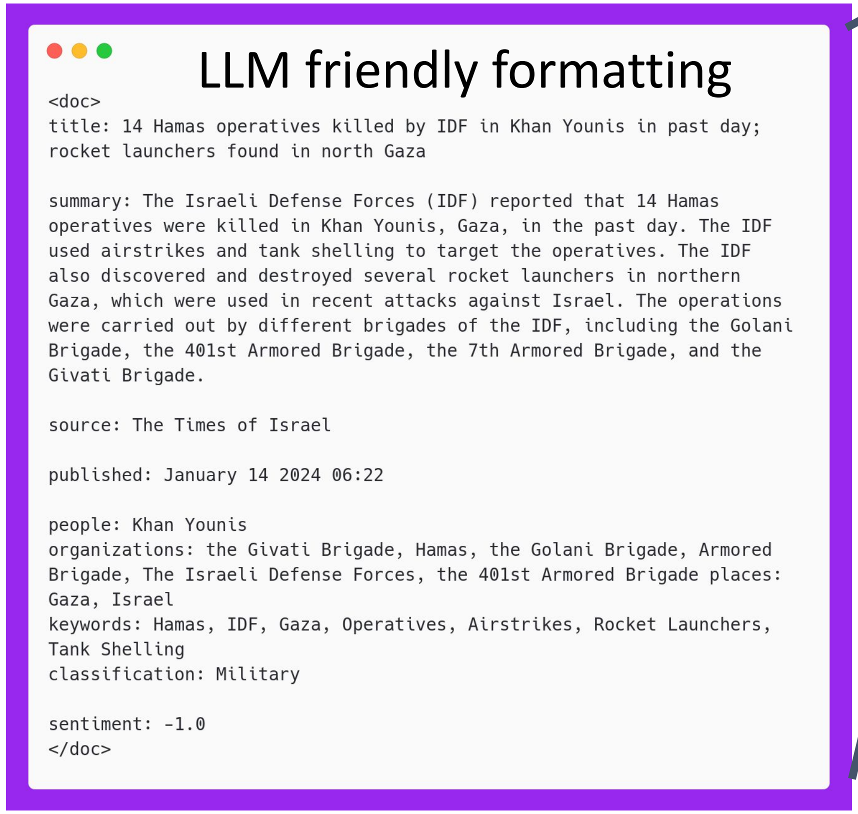
mongoDB

Document service

Leave the context engineering up to us 🧑‍🔬

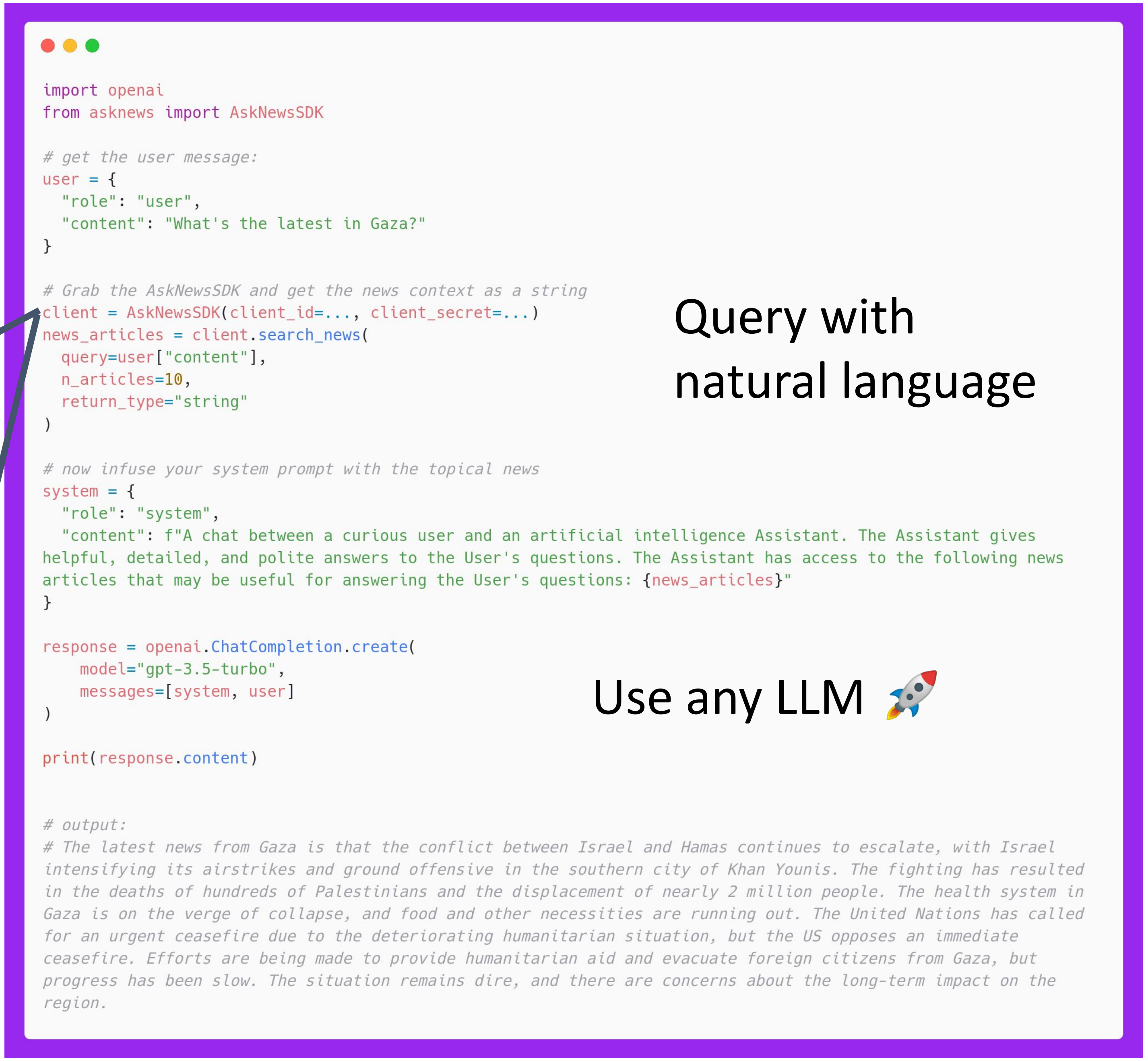
One call to our system equates to:

1. Similarity search on our real-time news enrichments
2. Auto-construct prompt friendly news enrichments for full context:



LLM friendly formatting

```
<doc>
title: 14 Hamas operatives killed by IDF in Khan Younis in past day;
rocket launchers found in north Gaza

summary: The Israeli Defense Forces (IDF) reported that 14 Hamas
operatives were killed in Khan Younis, Gaza, in the past day. The IDF
used airstrikes and tank shelling to target the operatives. The IDF
also discovered and destroyed several rocket launchers in northern
Gaza, which were used in recent attacks against Israel. The operations
were carried out by different brigades of the IDF, including the Golani
Brigade, the 401st Armored Brigade, the 7th Armored Brigade, and the
Givati Brigade.

source: The Times of Israel

published: January 14 2024 06:22

people: Khan Younis
organizations: the Givati Brigade, Hamas, the Golani Brigade, Armored
Brigade, The Israeli Defense Forces, the 401st Armored Brigade places:
Gaza, Israel
keywords: Hamas, IDF, Gaza, Operatives, Airstrikes, Rocket Launchers,
Tank Shelling
classification: Military

sentiment: -1.0
</doc>
```



```python
import openai
from asknews import AskNewsSDK

# get the user message:
user = {
  "role": "user",
  "content": "What's the latest in Gaza?"
}

# Grab the AskNewsSDK and get the news context as a string
client = AskNewsSDK(client_id=..., client_secret=...)
news_articles = client.search_news(
  query=user["content"],
  n_articles=10,
  return_type="string"
)

# now infuse your system prompt with the topical news
system = {
  "role": "system",
  "content": f"A chat between a curious user and an artificial intelligence Assistant. The Assistant gives
helpful, detailed, and polite answers to the User's questions. The Assistant has access to the following news
articles that may be useful for answering the User's questions: {news_articles}"
}

response = openai.ChatCompletion.create(
    model="gpt-3.5-turbo",
    messages=[system, user]
)

print(response.content)


# output:
# The latest news from Gaza is that the conflict between Israel and Hamas continues to escalate, with Israel
intensifying its airstrikes and ground offensive in the southern city of Khan Younis. The fighting has resulted
in the deaths of hundreds of Palestinians and the displacement of nearly 2 million people. The health system in
Gaza is on the verge of collapse, and food and other necessities are running out. The United Nations has called
for an urgent ceasefire due to the deteriorating humanitarian situation, but the US opposes an immediate
ceasefire. Efforts are being made to provide humanitarian aid and evacuate foreign citizens from Gaza, but
progress has been slow. The situation remains dire, and there are concerns about the long-term impact on the
region.
```

Query with natural language

Use any LLM 🚀