Breaking Black Box - Evals

Craig West

https://craig-west.netlify.app/

https://evaluating-ai-agents.com/

- Contains links for slides and eval-framework
- Everything in this talk is included

My desired outcome

Not for you to fully understand Evals

But…

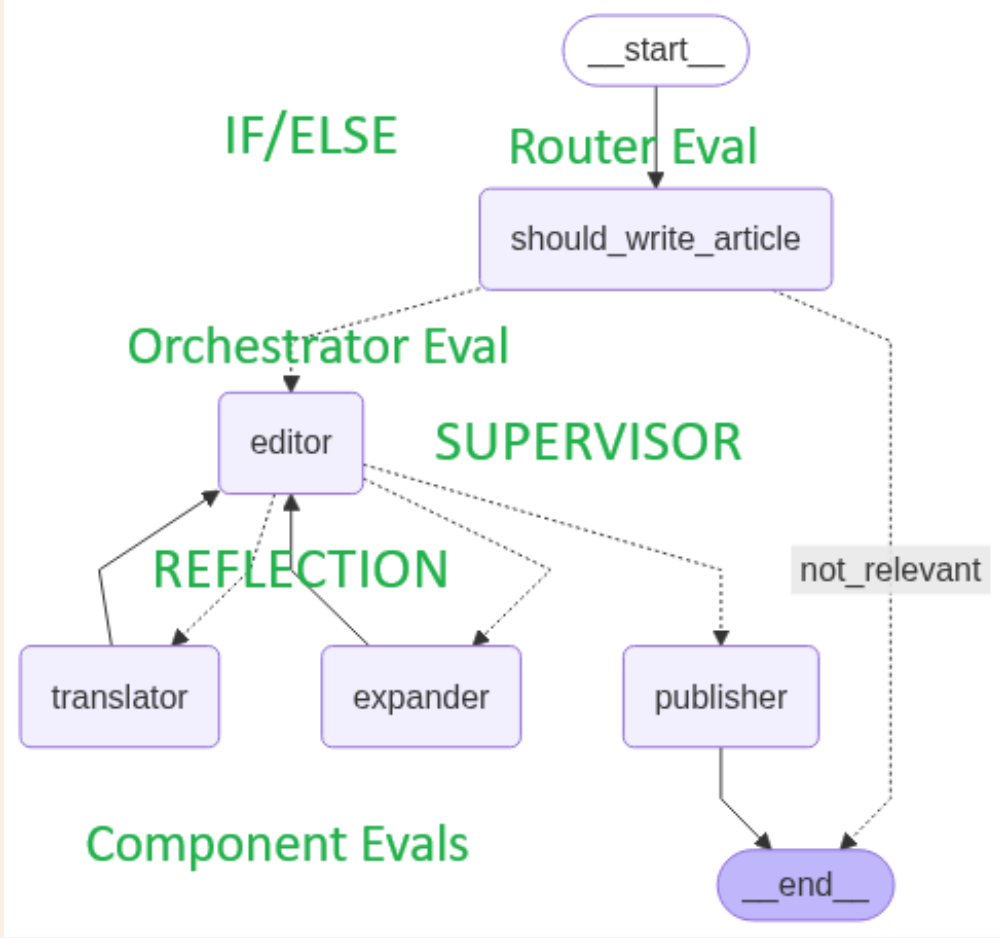
That Evals are doable

There are useful resources available for you

Demystify and simplify what an Agent is

Our product

To provide some context we will use this toy app.
It contains many common patterns.
It can also be a ‘leaf’ app for a much larger app.
1. Given a subject title, it **classifies** it as AI/NOT_AI
and if so, continues to the ‘editor’ which **orchestrates**
two agents:
2. One translates article into a selected foreign language,
the other expands content to a selected number of words.
3. The editor also judges the article regarding being
sensational or not.



It requires a **YES** to **correct language**, **correct length** and **not_sensational** before passing it to the publishing node.

Why Evals?

Ensure TRUST and CAPABILITIES of our AGENTS so that somebody will pay for them.

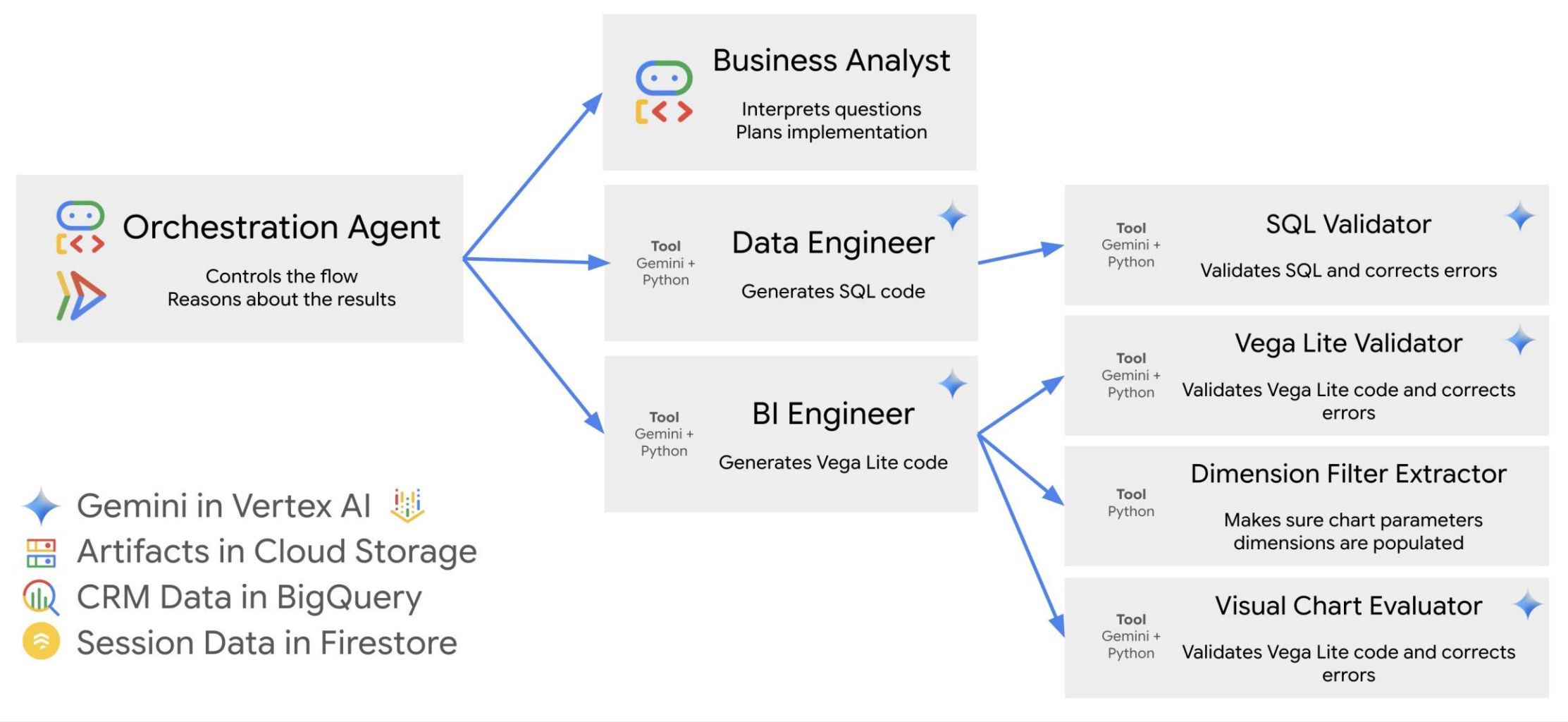
Would you pay for this substandard product?

It is also very useful during dev and debugging.

But I don’t use Agents!

- Perhaps not now, but it might be that your app is required to use Natural Language, (text, voice) from the user.
- You might use or incorporate ready made 3rf party software that is Agentic as in the next slide…Google’s CRM Agent…
- or use Model Context Protocol/Agent2Agent in your ‘app’
- It may be helpful to understand what Agents are

Google CRM Agent

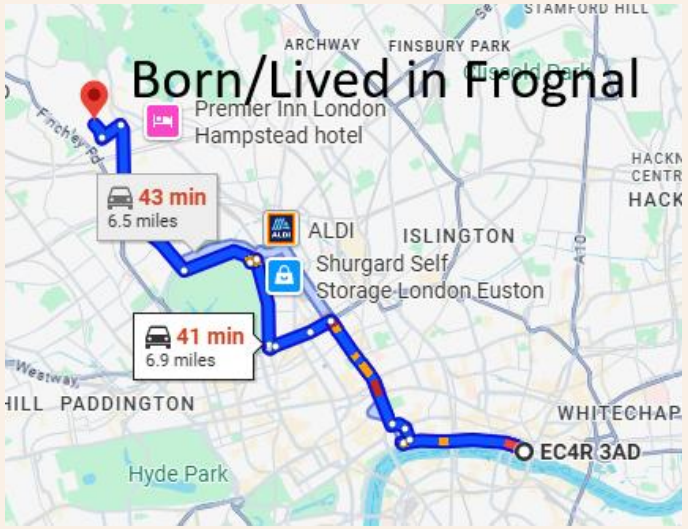


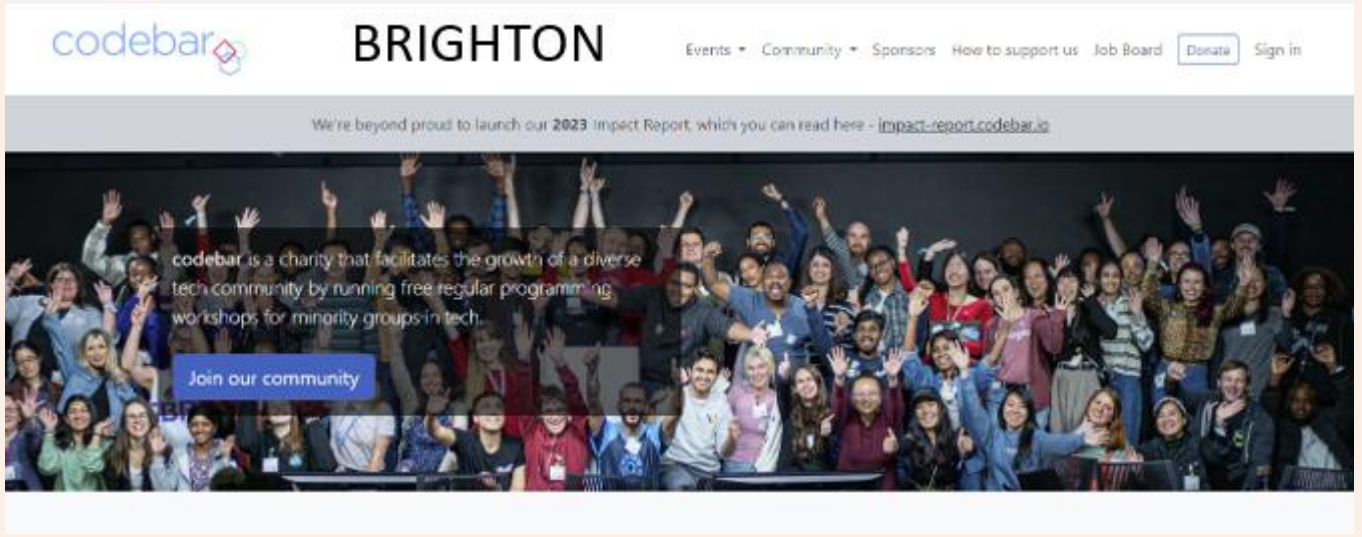
https://github.com/vladkol/crm-data-agent/blob/main/src/agents/data_agent/prompts/crm_business_analyst.py

https://www.youtube.com/watch?v=2-AJszogj7Y

About me


Brighton
I live ~ one mile


Born/Lived in Frognal
Premier Inn London Hampstead hotel
43 min
6.5 miles
ALDI
Shurgard Self Storage London Euston
41 min
6.9 miles
ARCHWAY
FINSBURY PARK
ISLINGTON
PADDINGTON
Hyde Park
EC4R 3AD


codebar
BRIGHTON
Events
Community
Sponsors
How to support us
Job Board
Donate
Sign in
We're beyond proud to launch our 2023 Impact Report, which you can read here - impact-report.codebar.io
codebar is a charity that facilitates the growth of a diverse tech community by running free regular programming workshops for minority groups in tech.
Join our community

What is an AI Agent?

Many definitions and people opt for ‘Agentic Apps’

Anthropic

- **Workflows** are systems where LLMs and tools are orchestrated through predefined code paths.
- **Agents**, on the other hand, are systems where LLMs dynamically direct their own processes and tool usage, maintaining control over how they accomplish tasks.

API

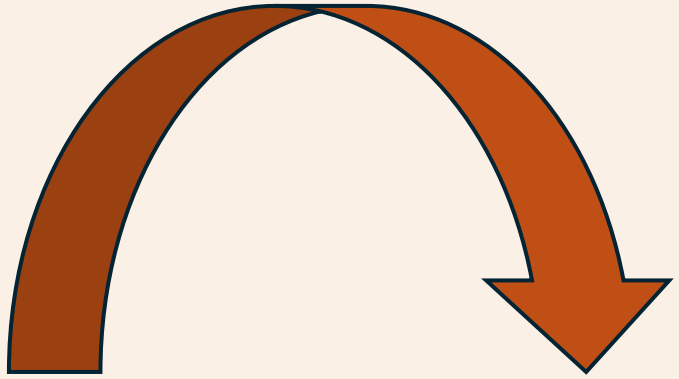
```
model = "gpt-4o-mini"  # Model
model_endpoint = "https://api.openai.com/v1/chat/completions"  # just one endpoint
headers = {
    "Content-Type": "application/json",  # Authorisation
    "Authorization": f"Bearer {api_key}",
}
# payload structure may vary from LLM Organisation but it is a text string.
payload = {
    "model": model,
    "messages": [
        {"role": "system", "content": system_prompt},
        {"role": "user", "content": user_prompt},
    ],
    # additional parameters
    "stream": False,
    "temperature": temperature,
}
# Use HTTP POST method
response = requests.post(
    url=model_endpoint,  # The endpoint we are sending the request to. Low Temperature:
    headers=headers,  # Headers for authentication etc
    data=json.dumps(payload),  # Inputs, Context, Instructions, Additonal parameters
).json()
```

These messages contain Instructions and Context

What is Prompt/Context Engineering?

- Prompt Engineering is going out of fashion and people opt for *Context Engineering* and *In Context Learning*.
- We create a set of instructions and add **additional information** to a request we send to the LLM.
- We can let LLM know that more context is available if it tells us which of the **tools/functions** we have told it about it wants us to call with **arguments** it has provided – we will run it on our box and send the extra context/information.

180 degree shift



No more difficult than what we do currently – just 180 degrees different

“It doesn’t get any easier – just different” - Anon

Model costs

Cost of a unit of ‘intelligence has dropped ~ 99% in last two years and is 10x a year (anecdotal)

Where to start? What to do?

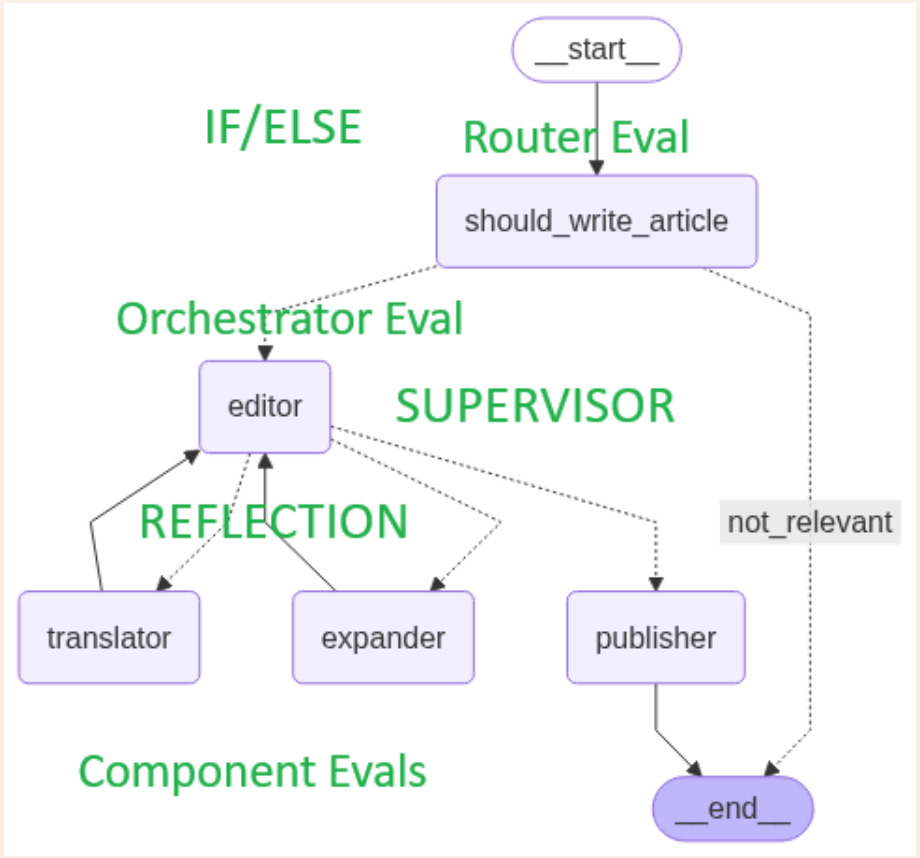
- We will look at Case Study 1 in the manual and framework.

Given an article title, determine if it is about AI

1. Take title and create article of N words (EXPANDER)
2. In a specified language (TRANSLATOR)
3. Ensure article is not sensational (EDITOR)

PUBLISH = YES only if all three are YES

Let's go to the guide and code…



Summary Key Points

- We need to look at the data to determine what are EVAL criteria are
- We need to determine what success will look like
- We will need humans at certain points
- We need a Domain Expert
- LLM Judges are for scaling
- We need to do EVALS on LLM Judges
- Ongoing monitoring
- It can seem to be both art and science

In closing

Craig West

https://craig-west.netlify.app/

https://evaluating-ai-agents.com/