Tantus Data

that uncovers wisdom.

AI chats: the conundrums of business integration

Agenda

01. RAG - Retrieval Augmented Generation

02. Hallucinations

03. Cost

04. Privacy

About me

Marcin Szymaniuk

Senior Data Engineer / CEO

Tantus Data

Time flies.

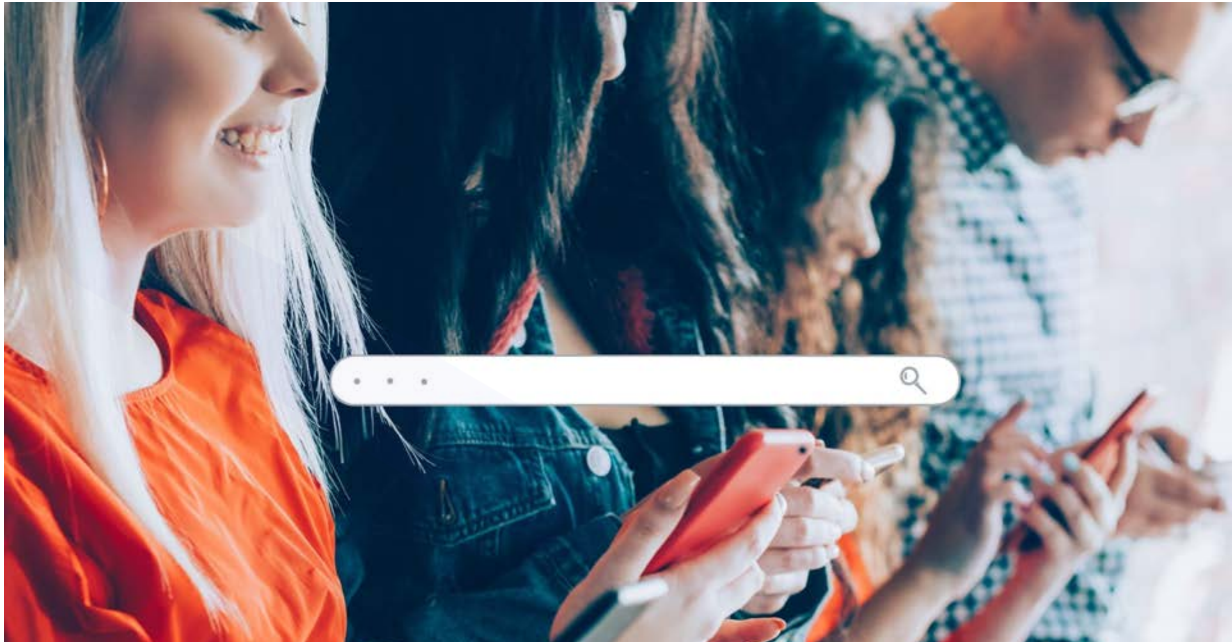
LLMs improve

Libraries improve

Libraries die

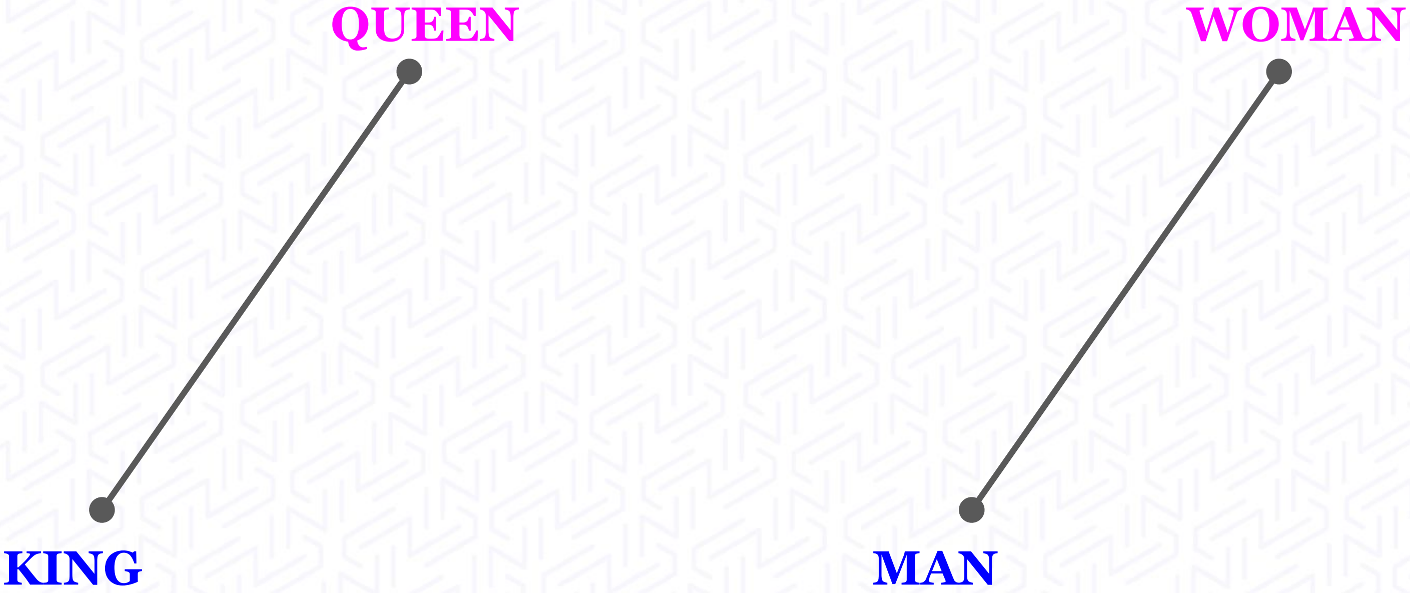
Search Query

```
I need an apartment with elevator in London
```



Vector Embeddings

- Text => Vector
- Semantic meaning



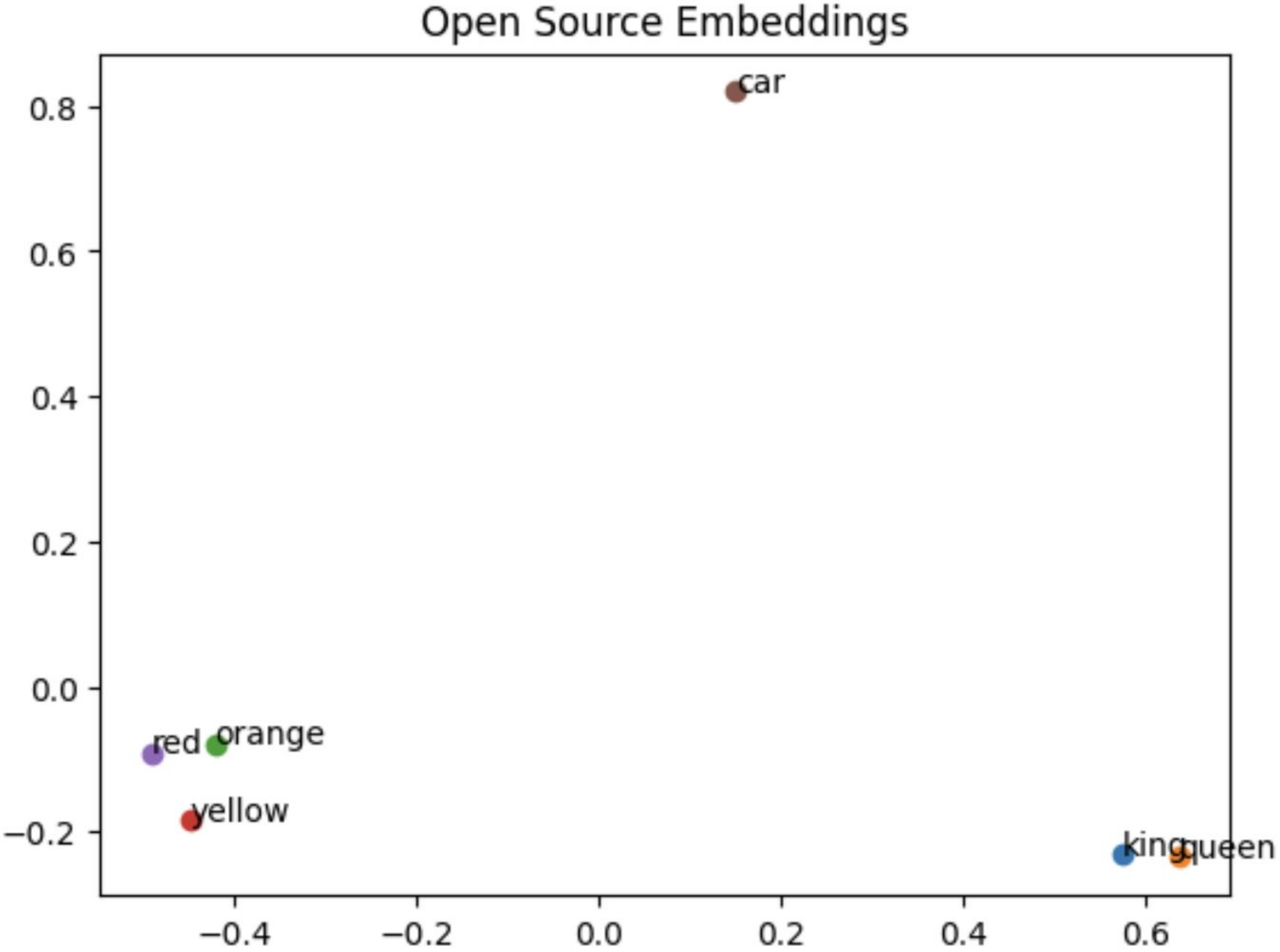
QUEEN

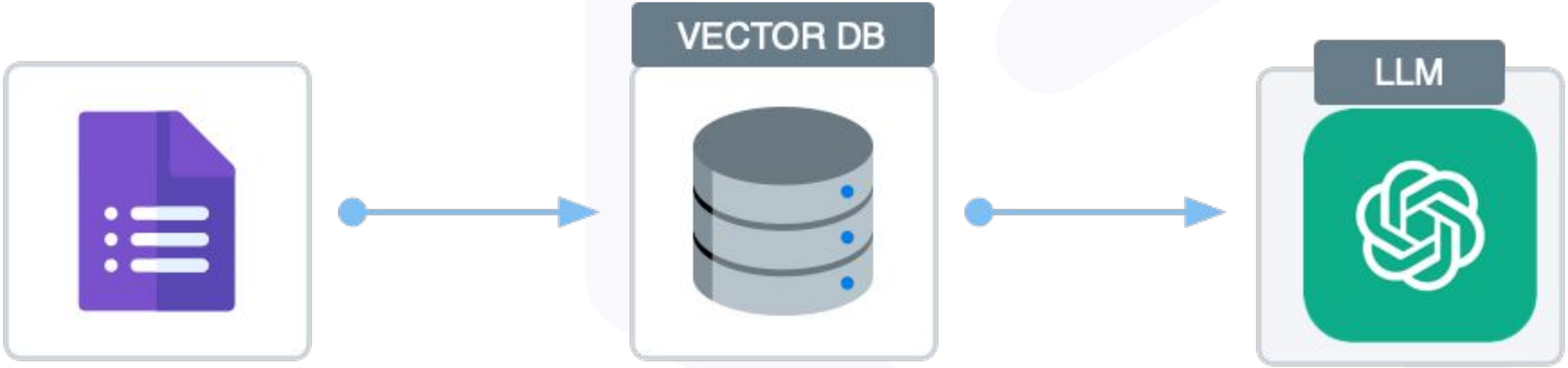
WOMAN

KING

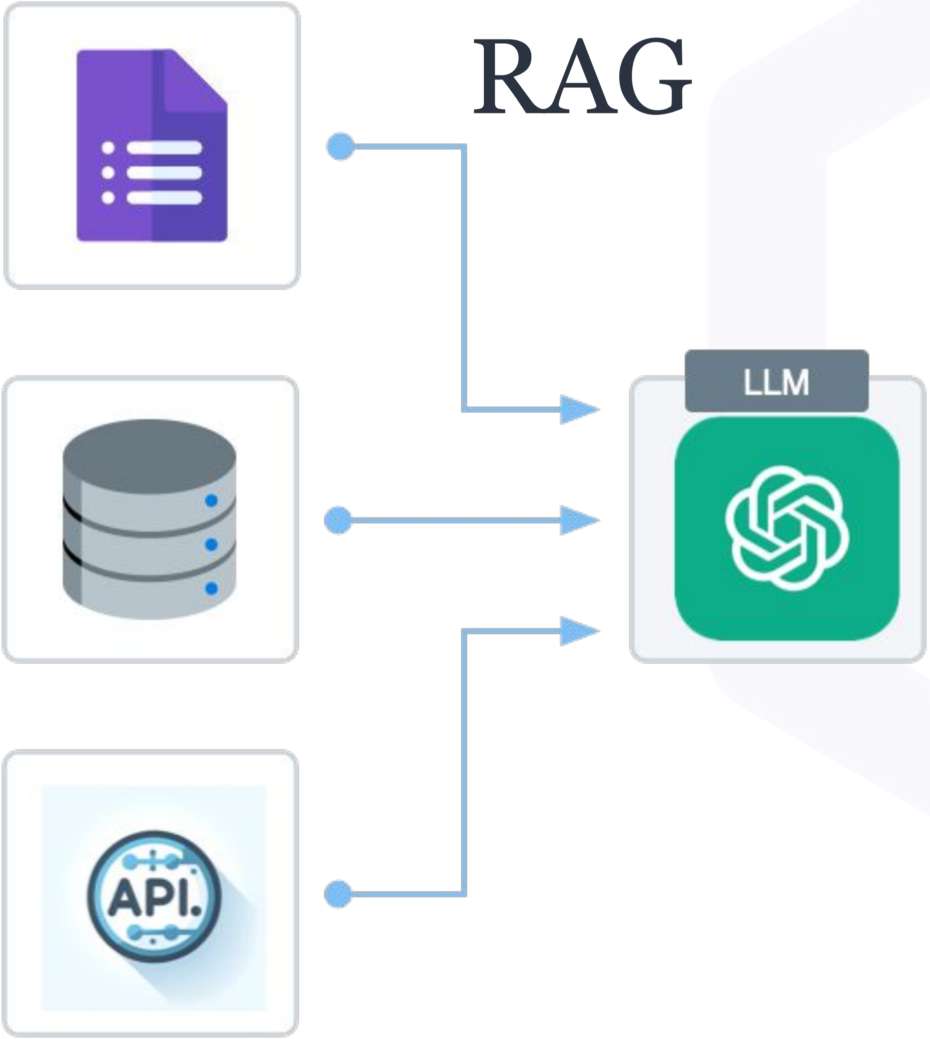
MAN

QUEEN = KING - MAN + WOMAN


Open Source Embeddings
car
red
orange
yellow
king
queen
0.8
0.6
0.4
0.2
0.0
−0.2
−0.4
−0.2
0.0
0.2
0.4
0.6

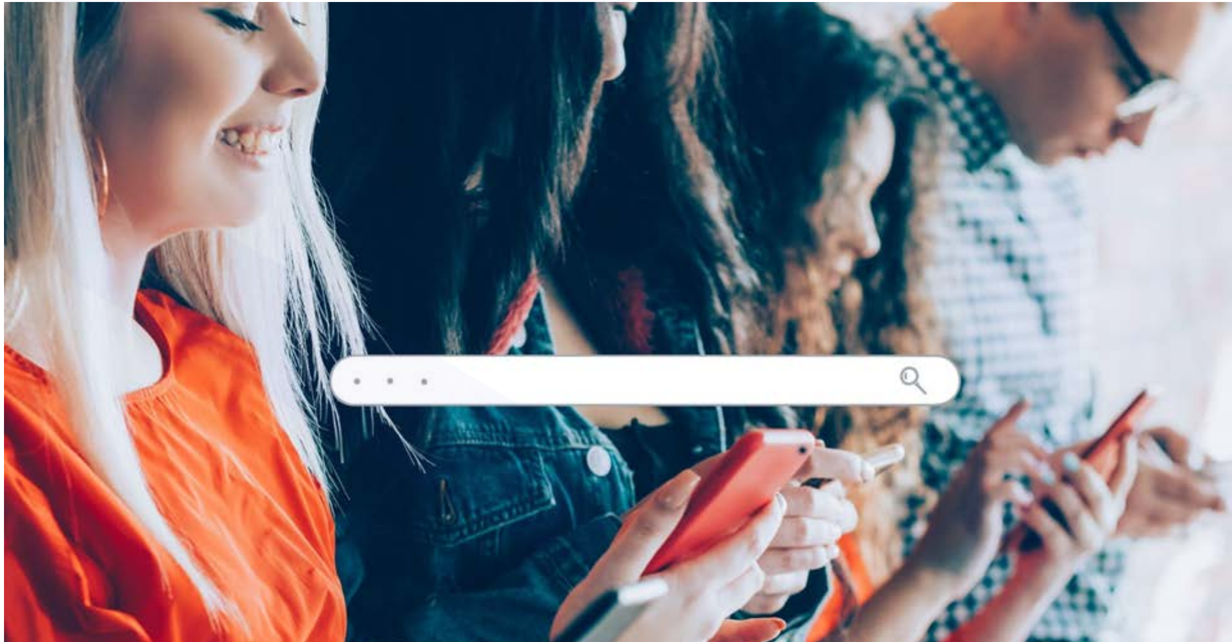

VECTOR DB
LLM

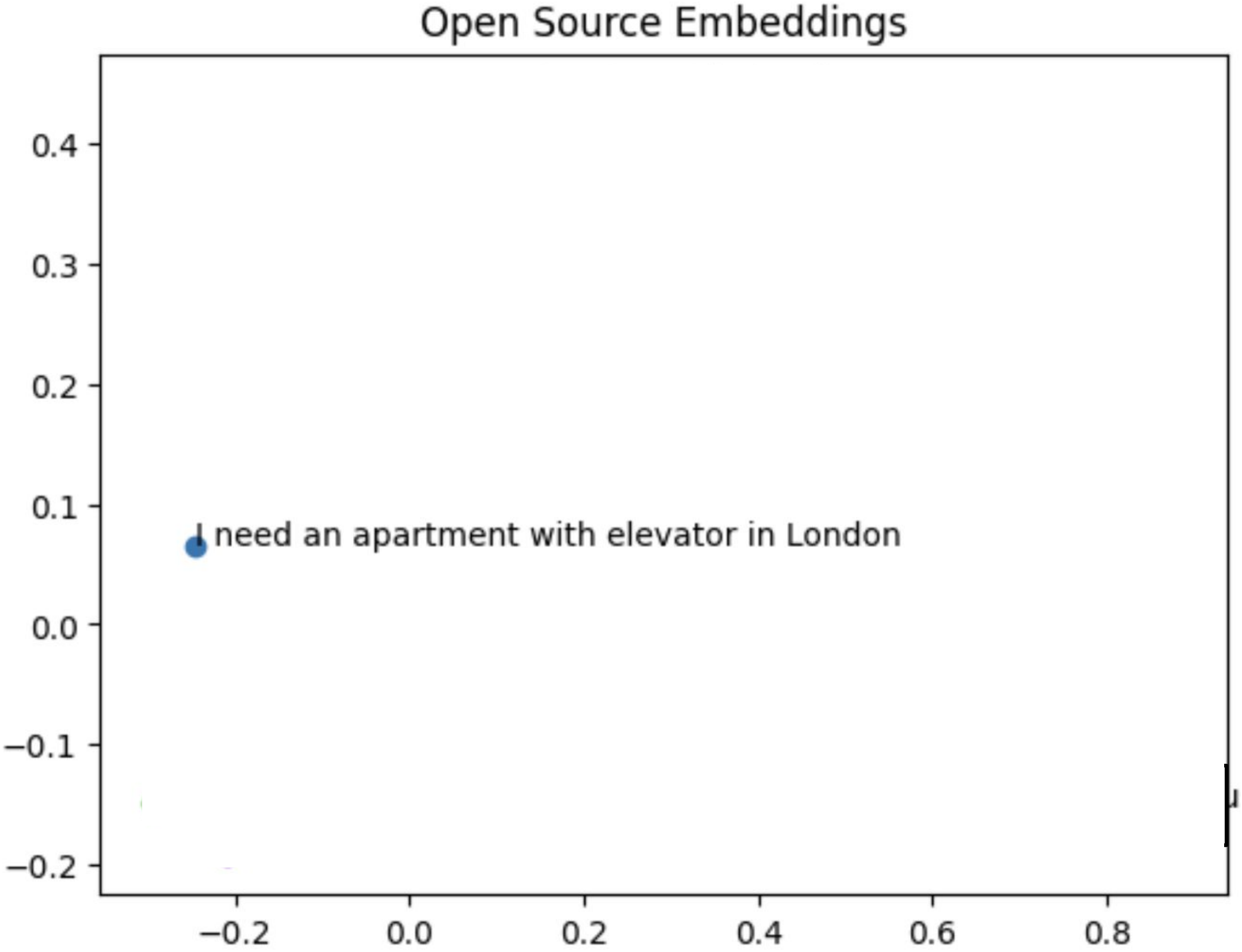
RAG

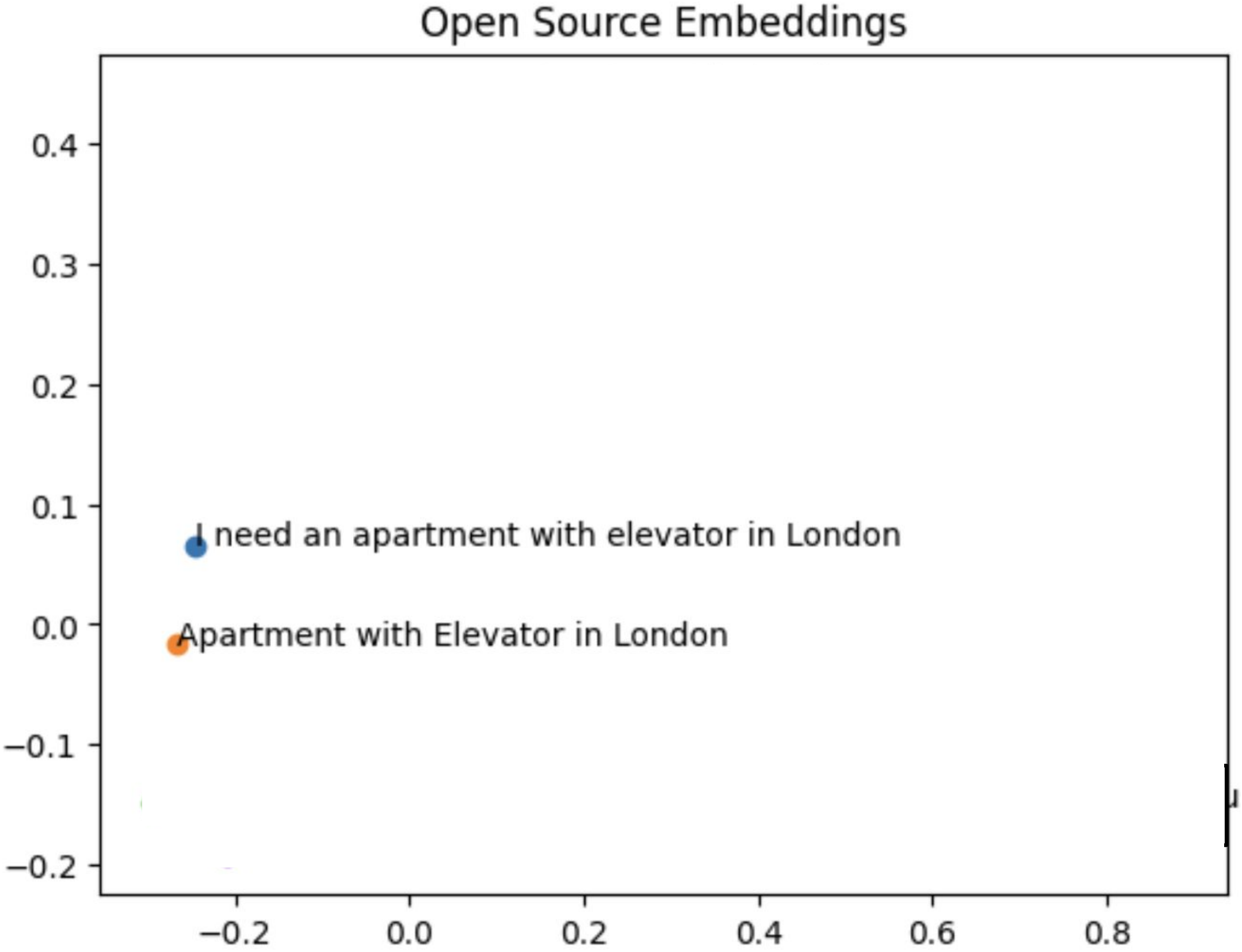


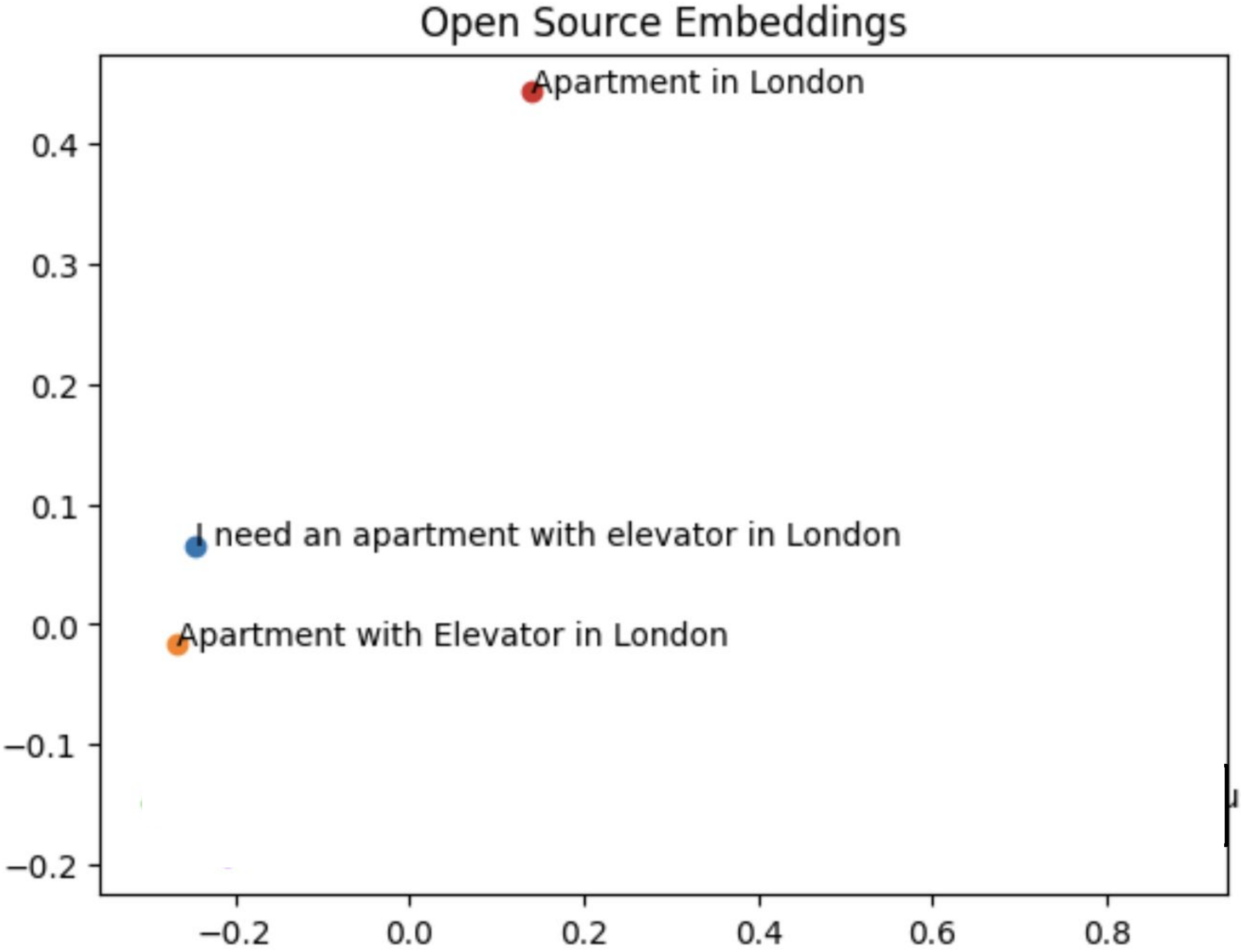
Search Query

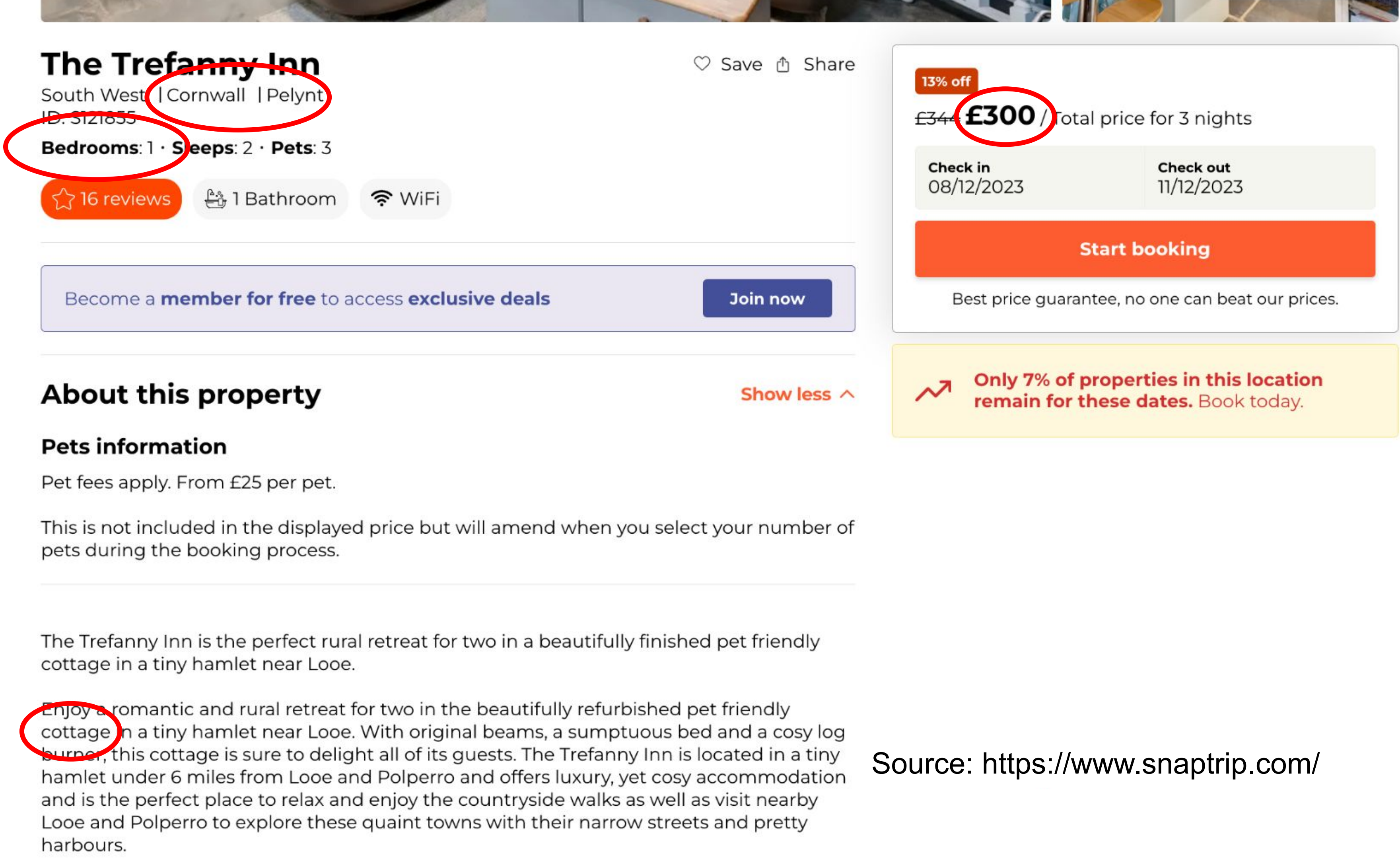
```
I need an apartment with elevator in London
```




Open Source Embeddings
I need an apartment with elevator in London
0.4
0.3
0.2
0.1
0.0
−0.1
−0.2
−0.2
0.0
0.2
0.4
0.6
0.8


Open Source Embeddings
I need an apartment with elevator in London
Apartment with Elevator in London
0.4
0.3
0.2
0.1
0.0
−0.1
−0.2
−0.2
0.0
0.2
0.4
0.6
0.8


Open Source Embeddings
Apartment in London
I need an apartment with elevator in London
Apartment with Elevator in London
0.4
0.3
0.2
0.1
0.0
−0.1
−0.2
−0.2
0.0
0.2
0.4
0.6
0.8



The Trefanny Inn

South West | Cornwall | Pelynt

ID: S121855

Bedrooms: 1 · **Sleeps**: 2 · **Pets**: 3

13% off

~~£344~~ **£300** / Total price for 3 nights

Check in	Check out
08/12/2023	11/12/2023

Best price guarantee, no one can beat our prices.

Become a **member for free** to access **exclusive deals**

Only 7% of properties in this location remain for these dates. Book today.

About this property

Pets information

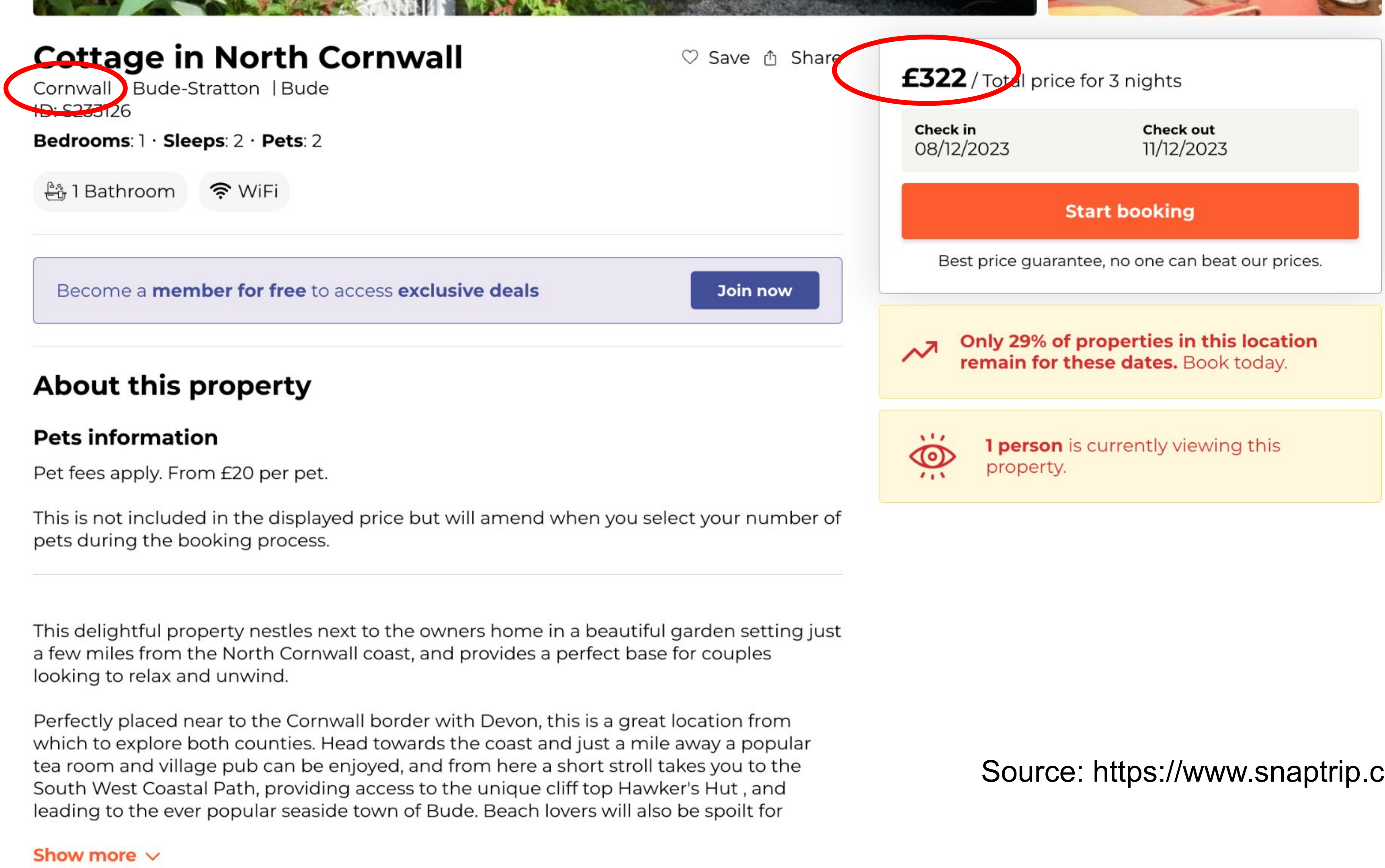
Pet fees apply. From £25 per pet.

This is not included in the displayed price but will amend when you select your number of pets during the booking process.

The Trefanny Inn is the perfect rural retreat for two in a beautifully finished pet friendly cottage in a tiny hamlet near Looe.

Enjoy a romantic and rural retreat for two in the beautifully refurbished pet friendly cottage in a tiny hamlet near Looe. With original beams, a sumptuous bed and a cosy log burner, this cottage is sure to delight all of its guests. The Trefanny Inn is located in a tiny hamlet under 6 miles from Looe and Polperro and offers luxury, yet cosy accommodation and is the perfect place to relax and enjoy the countryside walks as well as visit nearby Looe and Polperro to explore these quaint towns with their narrow streets and pretty harbours.

Source: https://www.snaptrip.com/



Cottage in North Cornwall

Cornwall | Bude-Stratton | Bude

ID: S233126

Bedrooms: 1 · **Sleeps**: 2 · **Pets**: 2

Become a **member for free** to access **exclusive deals**

About this property

Pets information

Pet fees apply. From £20 per pet.

This is not included in the displayed price but will amend when you select your number of pets during the booking process.

This delightful property nestles next to the owners home in a beautiful garden setting just a few miles from the North Cornwall coast, and provides a perfect base for couples looking to relax and unwind.

Perfectly placed near to the Cornwall border with Devon, this is a great location from which to explore both counties. Head towards the coast and just a mile away a popular tea room and village pub can be enjoyed, and from here a short stroll takes you to the South West Coastal Path, providing access to the unique cliff top Hawker's Hut , and leading to the ever popular seaside town of Bude. Beach lovers will also be spoilt for

£322 / Total price for 3 nights

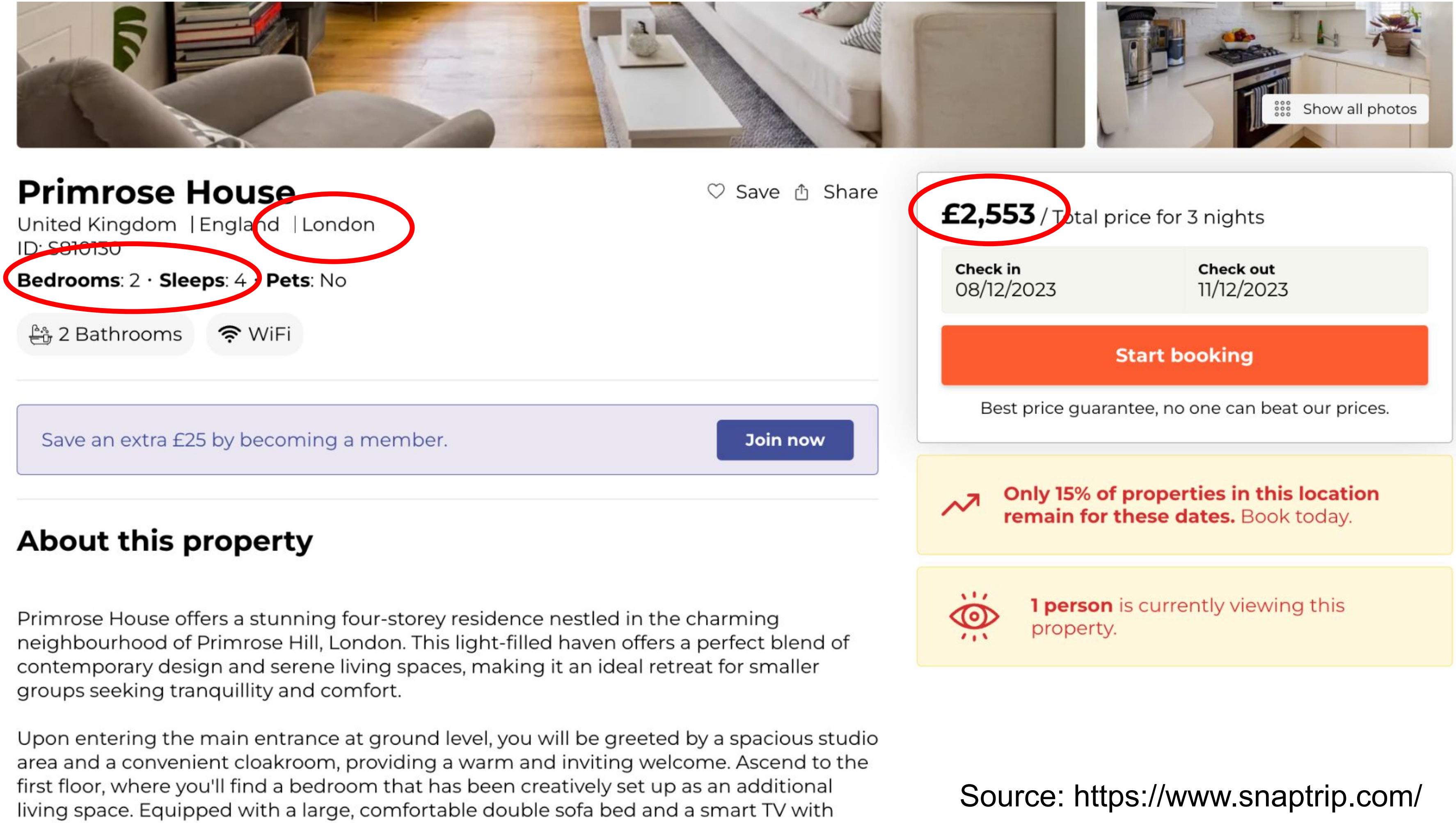
Check in	Check out
08/12/2023	11/12/2023

Best price guarantee, no one can beat our prices.

Only 29% of properties in this location remain for these dates. Book today.

1 person is currently viewing this property.

Source: https://www.snaptrip.com/



Primrose House

United Kingdom | England | London

ID: S810130

Bedrooms: 2 · **Sleeps**: 4 · **Pets**: No

£2,553 / Total price for 3 nights

Check in	Check out
08/12/2023	11/12/2023

Best price guarantee, no one can beat our prices.

Save an extra £25 by becoming a member.

Only 15% of properties in this location remain for these dates. Book today.

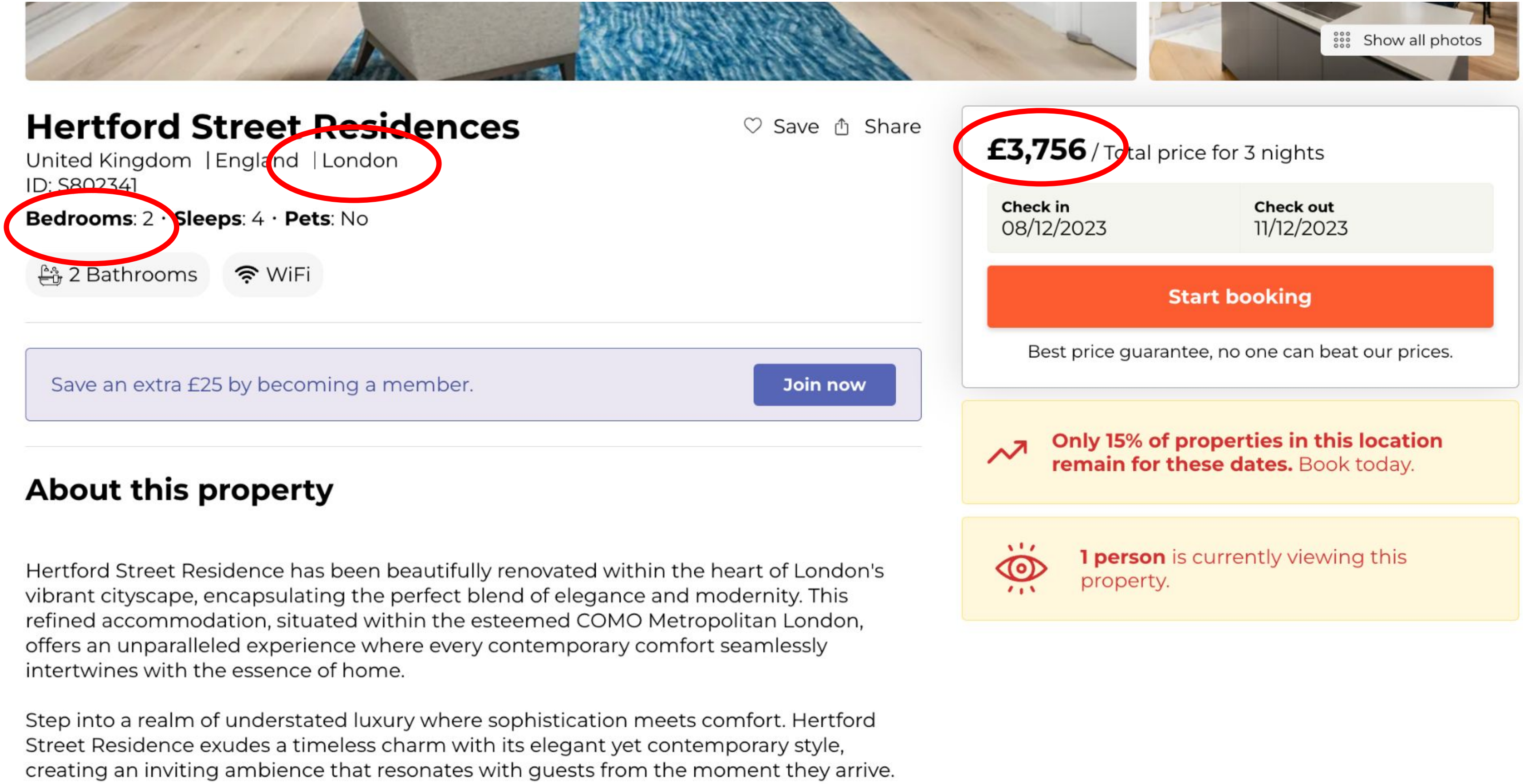
1 person is currently viewing this property.

About this property

Primrose House offers a stunning four-storey residence nestled in the charming neighbourhood of Primrose Hill, London. This light-filled haven offers a perfect blend of contemporary design and serene living spaces, making it an ideal retreat for smaller groups seeking tranquillity and comfort.

Upon entering the main entrance at ground level, you will be greeted by a spacious studio area and a convenient cloakroom, providing a warm and inviting welcome. Ascend to the first floor, where you'll find a bedroom that has been creatively set up as an additional living space. Equipped with a large, comfortable double sofa bed and a smart TV with

Source: https://www.snaptrip.com/



Hertford Street Residences

United Kingdom | England | London

ID: S802341

Bedrooms: 2 · **Sleeps**: 4 · **Pets**: No

2 Bathrooms WiFi

£3,756 / Total price for 3 nights

Check in 08/12/2023

Check out 11/12/2023

Best price guarantee, no one can beat our prices.

Save an extra £25 by becoming a member.

Only 15% of properties in this location remain for these dates. Book today.

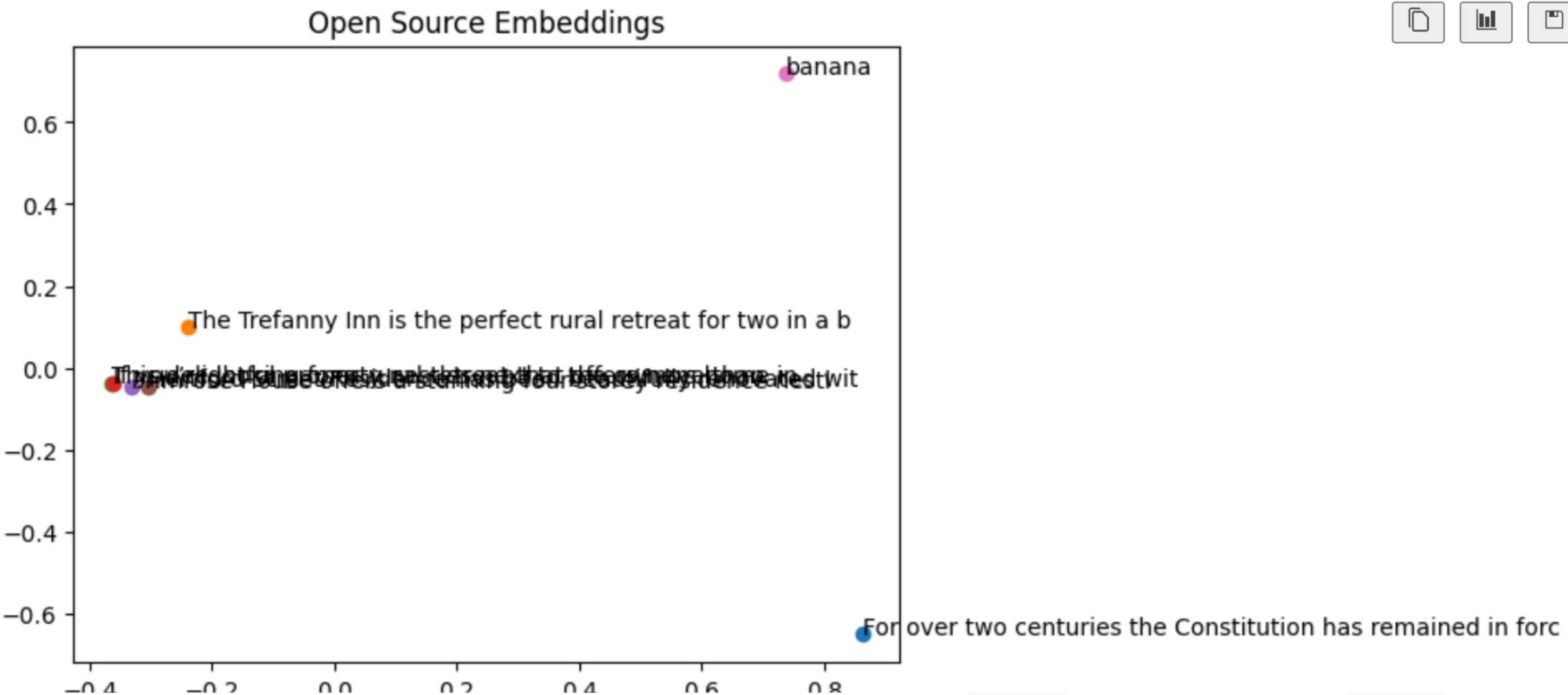
1 person is currently viewing this property.

About this property

Hertford Street Residence has been beautifully renovated within the heart of London's vibrant cityscape, encapsulating the perfect blend of elegance and modernity. This refined accommodation, situated within the esteemed COMO Metropolitan London, offers an unparalleled experience where every contemporary comfort seamlessly intertwines with the essence of home.

Step into a realm of understated luxury where sophistication meets comfort. Hertford Street Residence exudes a timeless charm with its elegant yet contemporary style, creating an inviting ambience that resonates with guests from the moment they arrive.

Source: https://www.snaptrip.com/


Open Source Embeddings
banana
The Trefanny Inn is the perfect rural retreat for two in a b
For over two centuries the Constitution has remained in forc
0.6
0.4
0.2
0.0
−0.2
−0.4
−0.6
−0.4
−0.2
0.0
0.2
0.4
0.6
0.8

Magic does not exist

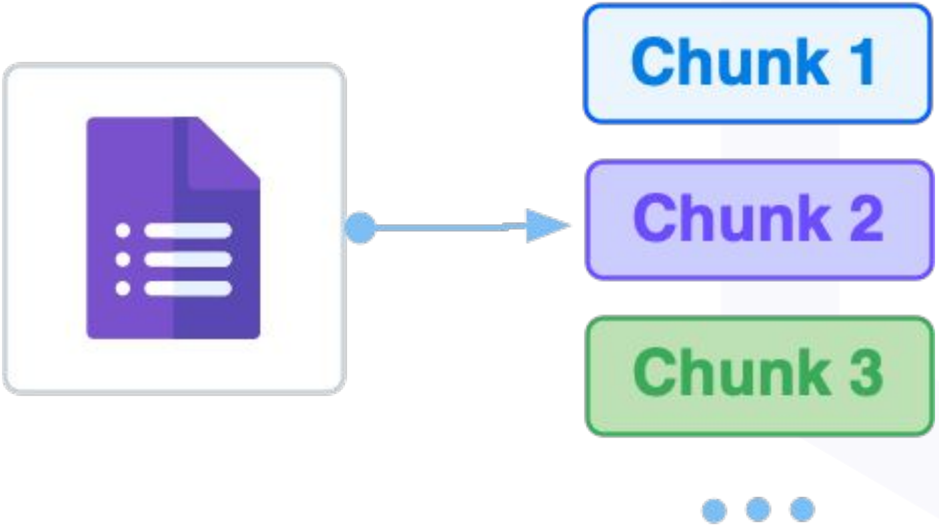
Specialisation

Context length

Longer is not better

Secret ingredients

Splitting documents

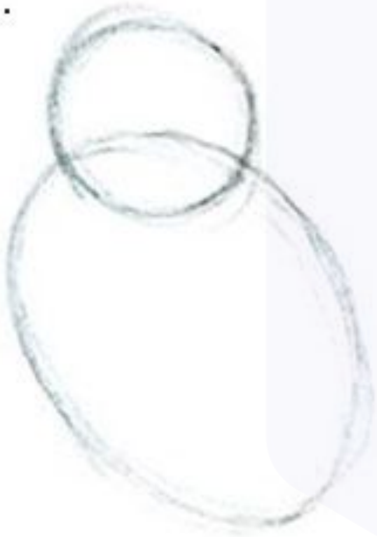


Chunk 1

Chunk 2

Chunk 3

1.



1.

2.

Splitting documents

Size

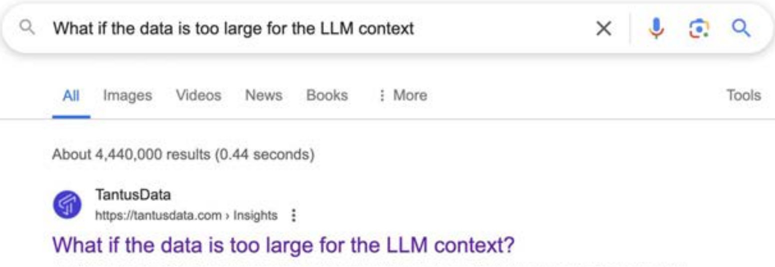
Context

Splitting documents

Size

Context

Self-Query



What if the data is too large for the LLM context

About 4,440,000 results (0.44 seconds)

TantusData
https://tantusdata.com › Insights

What if the data is too large for the LLM context?

Splitting documents

Size

Context

Self-Query

Tricky formats

Secret ingredients

Splitting documents

Hybrid search

Search options

Backend database

Data lake

Data Warehouse

Internal API

External API

Secret ingredients

Splitting documents

Hybrid search

Re-rank

Secret ingredients

Splitting documents

Hybrid search

Re-rank

Preprocessing using LLM

Hallucinations

Provide accurate info

Provide relevant info

Demo

Hallucinations check

Consequence

Cost of false positive

Other costs

Back to square one

Data quality

Careful engineering

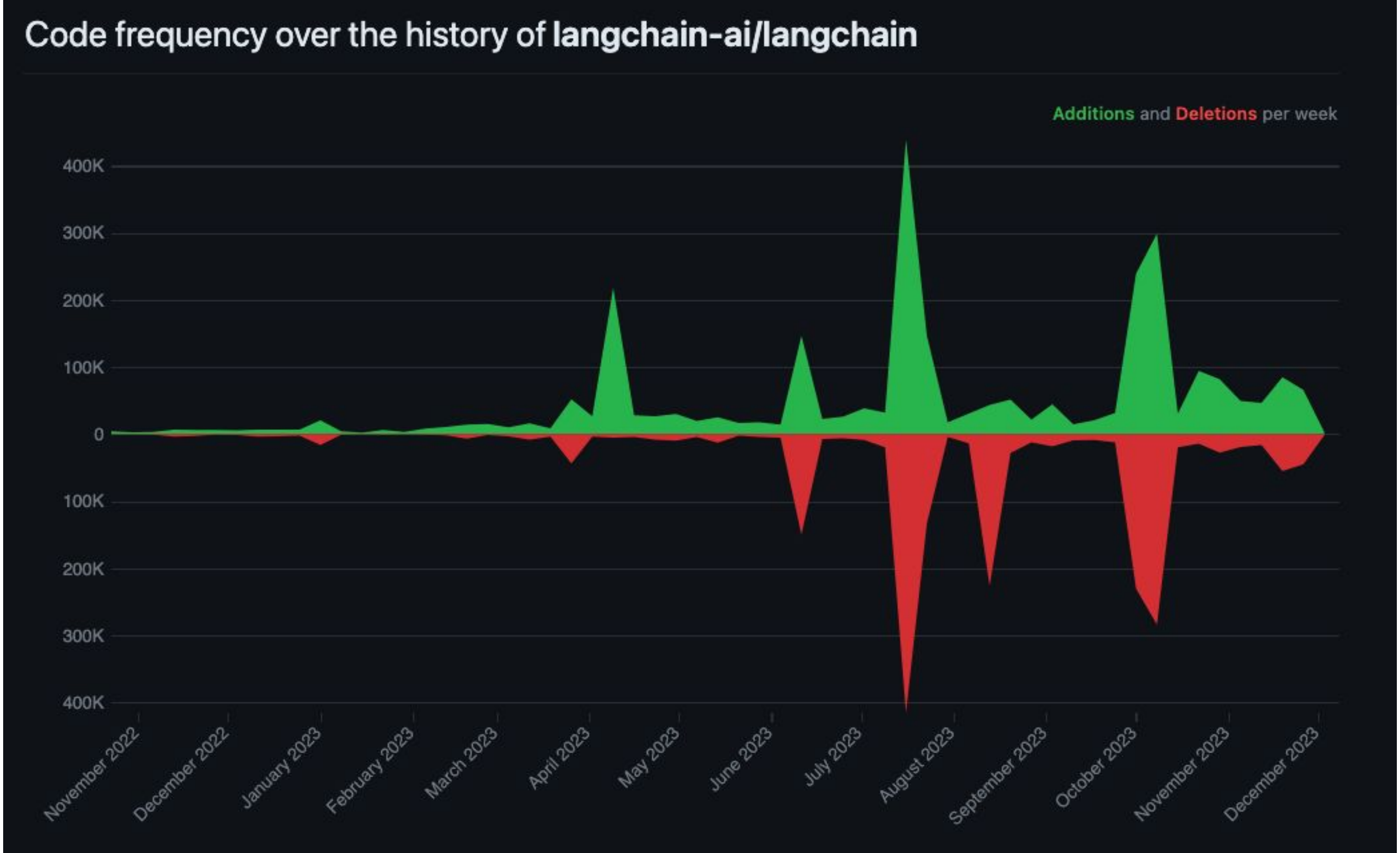
Tools

Nemo Guardrails

MemGPT

Weaviate

Tools



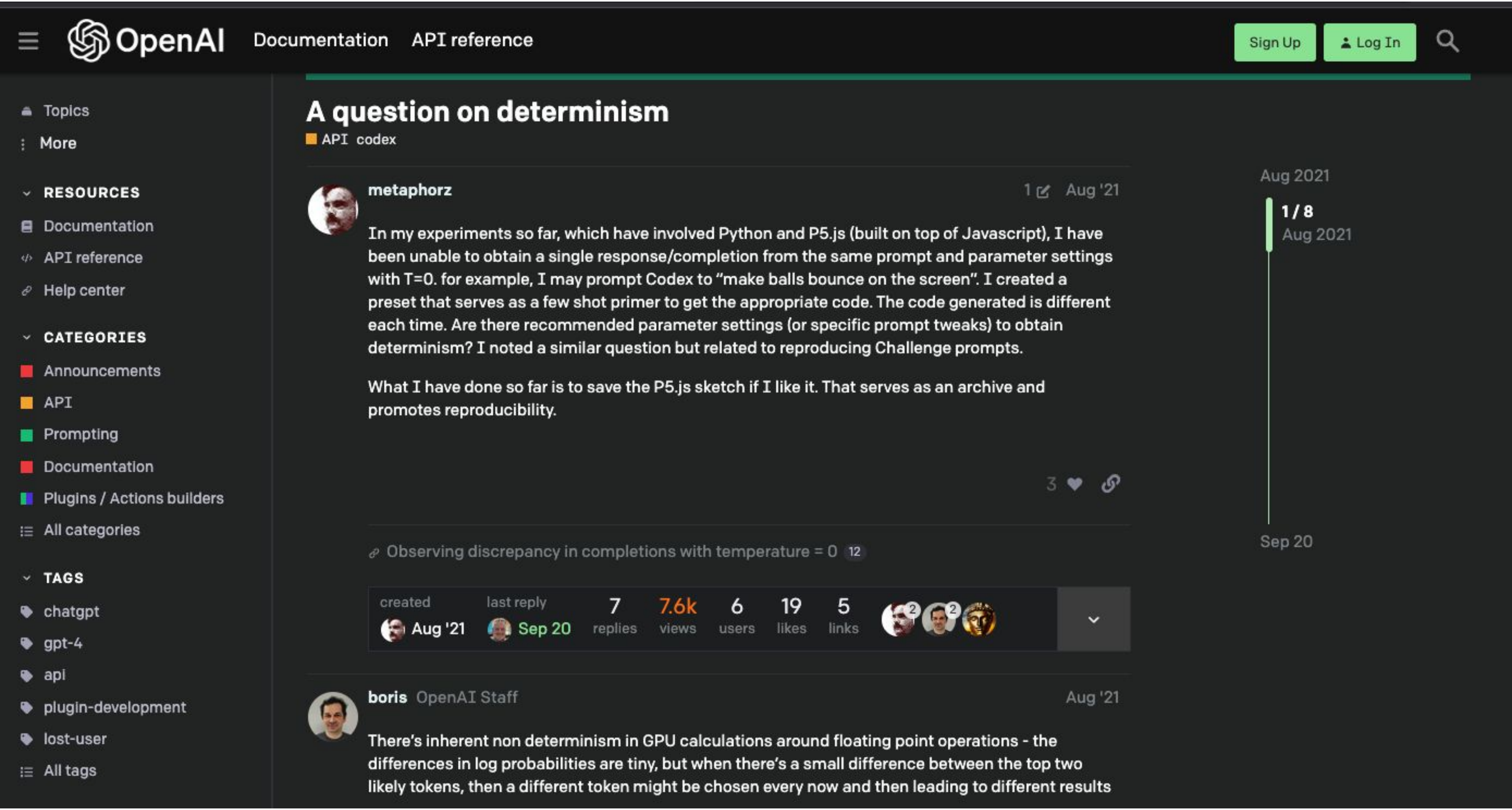
Code frequency over the history of **langchain-ai/langchain**

Testing

Test the retrieval!

Test the LLM actions… wherever you can..

Collect the data



A question on determinism

API codex

metaphorz 1 Aug '21

In my experiments so far, which have involved Python and P5.js (built on top of Javascript), I have been unable to obtain a single response/completion from the same prompt and parameter settings with T=0. for example, I may prompt Codex to "make balls bounce on the screen". I created a preset that serves as a few shot primer to get the appropriate code. The code generated is different each time. Are there recommended parameter settings (or specific prompt tweaks) to obtain determinism? I noted a similar question but related to reproducing Challenge prompts.

What I have done so far is to save the P5.js sketch if I like it. That serves as an archive and promotes reproducibility.

3

Observing discrepancy in completions with temperature = 0 12

created	last reply	7	7.6k	6	19	5
Aug '21	Sep 20	replies	views	users	likes	links

boris OpenAI Staff Aug '21

There's inherent non determinism in GPU calculations around floating point operations - the differences in log probabilities are tiny, but when there's a small difference between the top two likely tokens, then a different token might be chosen every now and then leading to different results

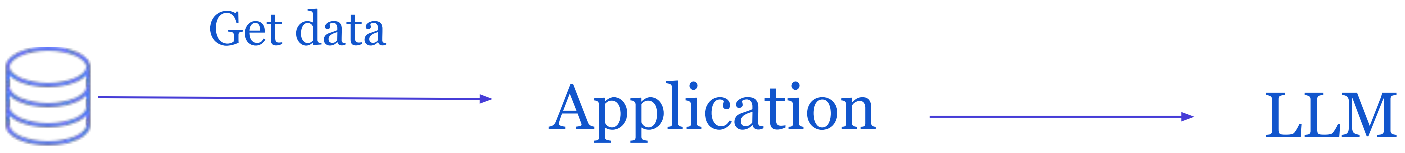
Testing

Test the retrieval!

Test the LLM actions… wherever you can..

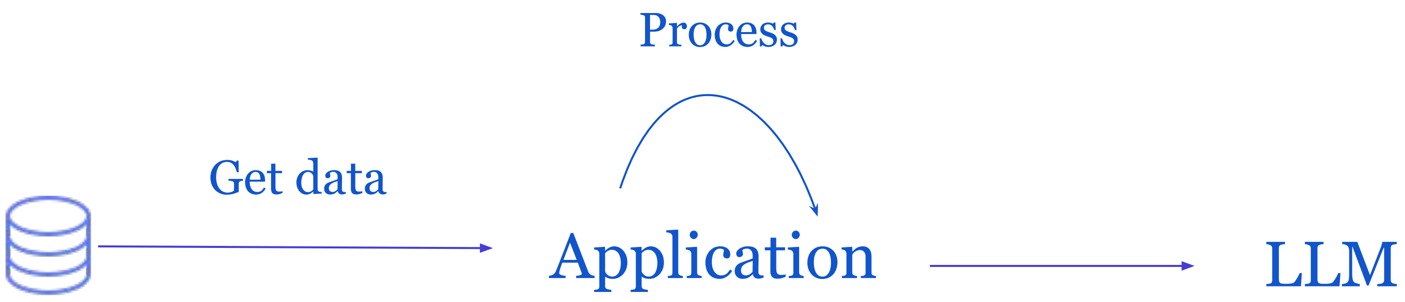
Collect the data

Privacy



Icons courtesy of Icons8

Privacy



Process

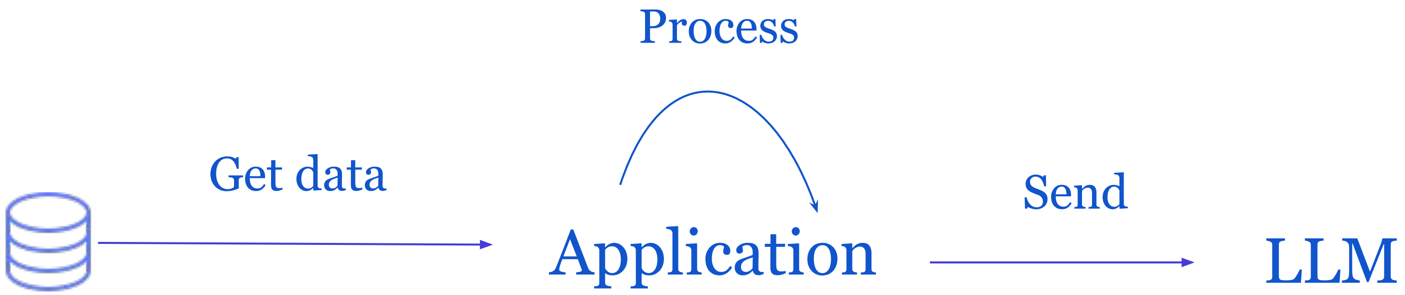
Get data

Application

LLM

Icons courtesy of Icons8

Privacy



Icons courtesy of Icons8

Privacy

API is not alway possible

Private LLM

Private Embedding model

Private Vector DB

Legal Aspects

Licence

Data For fine-tuning

Synthetic data

Copyrights

…

Cost

OS is cheaper…

Or not?

Cost

Chat GPT 3.5

In: 0.5$ / 1 mln tokens

Out: 1.5$ / 1 mln tokens

Chat GPT 4

In: 30$ / 1 mln tokens

Out: 60$ / 1 mln tokens

Claude 3 Opus

In: 15$ / 1 mln tokens

Out: 75$ / 1 mln tokens

Gemini

In: 0.125$ / 1 mln characters

Out: 0.375$ / 1 mln characters

Cost : API

Input vs output

Token vs token

Token vs character

Cost : OpenSource

Hardware Cost

Traffic

Not static cost per request

Cost

Conversion rate vs cost

Development time

Maintenance cost

Open source

Out of the box LLMs are behind

Fine tuning is possible

Data is the key!

Retrieval to start with

Open source - Training

P-Tuning

LoRA

IA3

SFT

Summary

Experiment

Business goal first

Be pragmatic

Thank you!

Do you have any questions?

marcin@tantusdata.com

www.tantusdata.com