

November, 05

17:00 (CET)

English

Customer and testing-friendly LLMs: dream or reality?



James Lennon

VP of Product
Engineering at Ciklum



Oleg Panichev

Expert Research
Engineer at Ciklum



Mariana Batiuk

Principal TCoE Lead
at Ciklum

Experiences of tomorrow. Engineered together.



We transform how people experience the business. All through next generation technology.

What we do:

Product
Engineering

Intelligent
Automation

Data &
Analytics

2002
founded

4000+
professionals

20+
offices

300+
clients

Leading companies choose us:



KANTAR

RETAIL



Our speakers



James Lennon
VP of Experience
Engineering

- James pioneers adoption of systems like GenAI and parallel programming to accelerate delivery and delight customers.
- With over 15 years driving digital innovation, James combines deep hands-on experience with a vision for the future.



Mariana Batiuk
Principal
TCoE Lead

- Mariana leads the technical council on the QA maturity assessment, test strategy, pre-sales, new services development, initiatives, and quality engineering activities.
- Has proficient experience in QA Management, Agile Methodologies, Testing, Team Management, and Coaching.



Oleg Panichev
Expert Research
Engineer

- Oleg has 14 years of experience in R&D projects and has been working with machine learning, deep learning, and data science.
- He was involved in establishing the R&D department in Ciklum, specializing in complex projects.

Our Global Delivery Centres

Global Reach, Local Insight - Ciklum bridges the best in tech from various geographies



Central & Eastern Europe

-  Bulgaria
-  Czech Republic
-  Poland
-  Romania
-  Slovakia
-  Spain
-  Ukraine
-  United Kingdom

Asia

-  India
-  Pakistan

LATAM

-  Argentina
-  Uruguay

North America

-  Canada
-  USA



Ciklum & AI

- Intro
- Ciklum case study

Value-Focused AI Adoption

A product focused approach to embracing rapid technological change & possibilities.

In-Production GenAI Builds

25+

Exemplary GenAI Clients

AstraZeneca 

axel springer 

 BigChange

Deloitte.

 Dotdash
meredith

eliorgroup 

ibsoftware

INSIGHT
PARTNERS

inspired

 JLL

JLR


mastercard

Paysafe:

 Santander

Savanta:

myZerto

“

2024 was the
year of the PoC

”

–James Lennon,
said just now



User
Problem
First

Being an analyst
can suck!

No, really. It can.

There's a lot to do to answer one question. Should we invest?



20+

Data sources to query



Which don't always agree

100k+

Current & potentially viable investment opportunities to monitor daily

100+

Different reports and analyses needed to be created

24hrs

A day to get all this done

How can we make the experience better?

How can we integrate into an actual workflow?



Trackassistant.com

Generated Summary

Track Assistant is an app and online service for easily recording trips and mileage for your car. It offers manual, semi-automatic, and fully automatic trip logging using your smartphone's GPS.

Key features:

- Manual trip logging: Enter your destination and the app calculates the mileage. Adjust if needed to match odometer.
- Semi-automatic: Start/stop trip logging with a slider. App records route and mileage automatically.
- Fully automatic: App detects when you start and stop driving and records trips automatically.
- Works on both iPhone and Android. iPhone uses iBeacon for automatic tracking.
- Syncs data to online account for backup and reporting.
- Handles combined business and personal trips - splits mileage automatically.

Tips

Here are some examples of what AI can do for you.

Summarize a Company

Provide a summary about the company Search for...

Fetch info from a Website

Summarize this website for me [insert url here], be...

Summarize an Industry

Please tell me the top 5 things to know about [inse...

Write an Email

Please help craft a follow up email: Prompt: Nice t...

Summarize Call Notes

Please summarize the following call notes into 5 b...

Fetch News

Can you tell me about recent venture capital fundi...

HINT: start message with / to view available commands.



Results

30+

min av. session
duration

8.4/10

NPS score

Accuracy & reliability drove the success

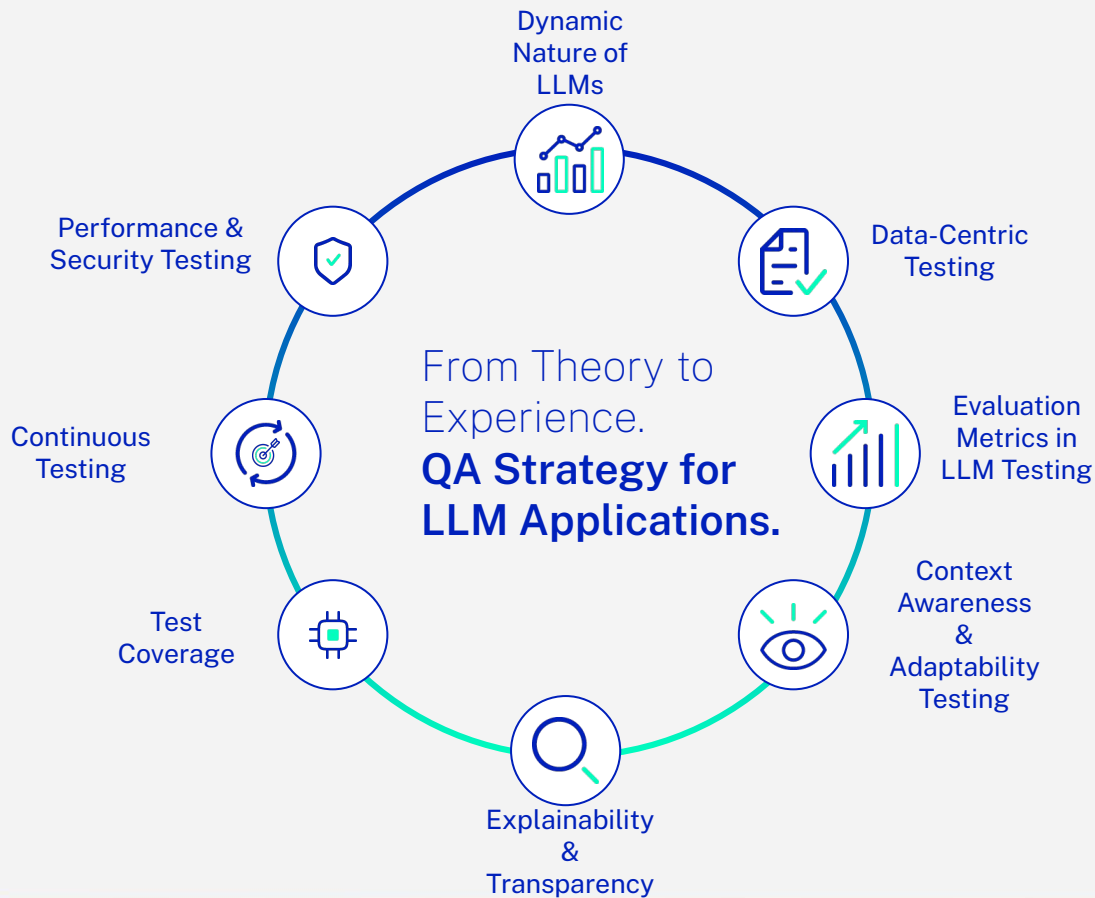
But how did we achieve it?

A decorative wavy line in shades of blue and green, starting from the bottom left and curving upwards towards the right side of the slide.



QA & Testing

- What theory is saying?
- How does it change since our last talk?



As we shift from traditional software to AI-driven models, ensuring the quality of LLM applications requires **new** strategies and frameworks.

Dynamic Nature of LLMs vs. Static Behavior in Traditional Software

Traditional Software

Predictable
Outputs

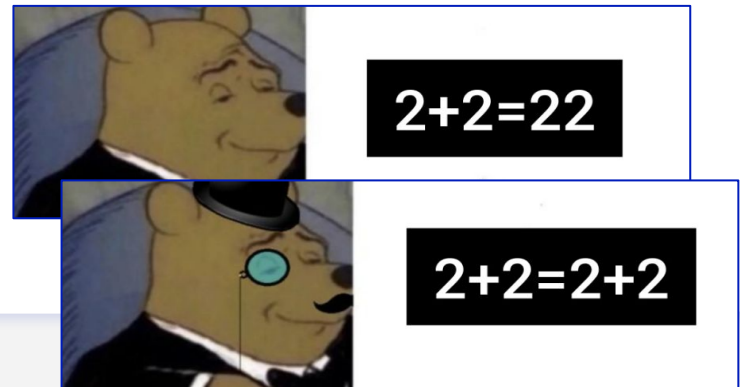
Pre-defined
Logic



LLMs

Non-deterministic
Outputs

Learned
Behavior



Data-Centric Testing

- ✓ LLMs Learn from Data
- ✓ Bias and Fairness
- ✓ Real-World Impact



Data Quality

Verifying that the training data is clean, accurate, and comprehensive.



Relevance and Coverage

Ensuring the dataset represents diverse contexts and real-world scenarios, covering edge cases.



Bias Detection

Analyzing data for inherent biases (gender, race, socioeconomic status, etc.) and removing or mitigating them.



Incorporate human-in-the-loop feedback to evaluate ethical concerns and edge cases.

Raw Data

Preprocessing

Bias Testing

Model Training

Evaluation

Evaluation Metrics in LLM Testing

Perplexity Metric

Accuracy score

BLEU score

(Bilingual Evaluation Understudy)

EM score

(Exact Match)

ROUGE score

(Recall-Oriented Understudy for Gisting Evaluation)

HUMAN evaluation

(Emotion-Driven Chaos Fueled by Caffeine)

Context Awareness & Adaptability Testing

- Evaluating how well LLMs understand and maintain context during interactions.
- Assessing the model's ability to adapt to varying inputs and user intents.

Multi-turn
Interactions

User Intent
Recognition

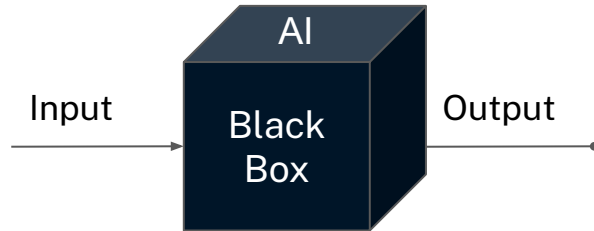
Dynamic Input
Handling

What if you won't do it?



CHALLENGE

OR ADVANTAGE?



LLM-based systems

How did you come up with this list? Can you explain your decision process?

What was the decision chain during generating answer to the previous question?

Test Coverage Challenges

Infinite Input Space

Context Dependency

Uncertainty in Model Outputs

Lack of Ground Truth

Evolving Models

Scenario-based
evaluations

Out-of-Scope
Coverage
=
idk

CONTINUOUS TESTING

Developed

Tested

Deployed

Tested

...

Tested

Performance & Security Testing

- Latency

- Prompt Injection

- Real-Time Responsiveness

- Data Leak Prevention

Useful links [here](#) and [here](#)



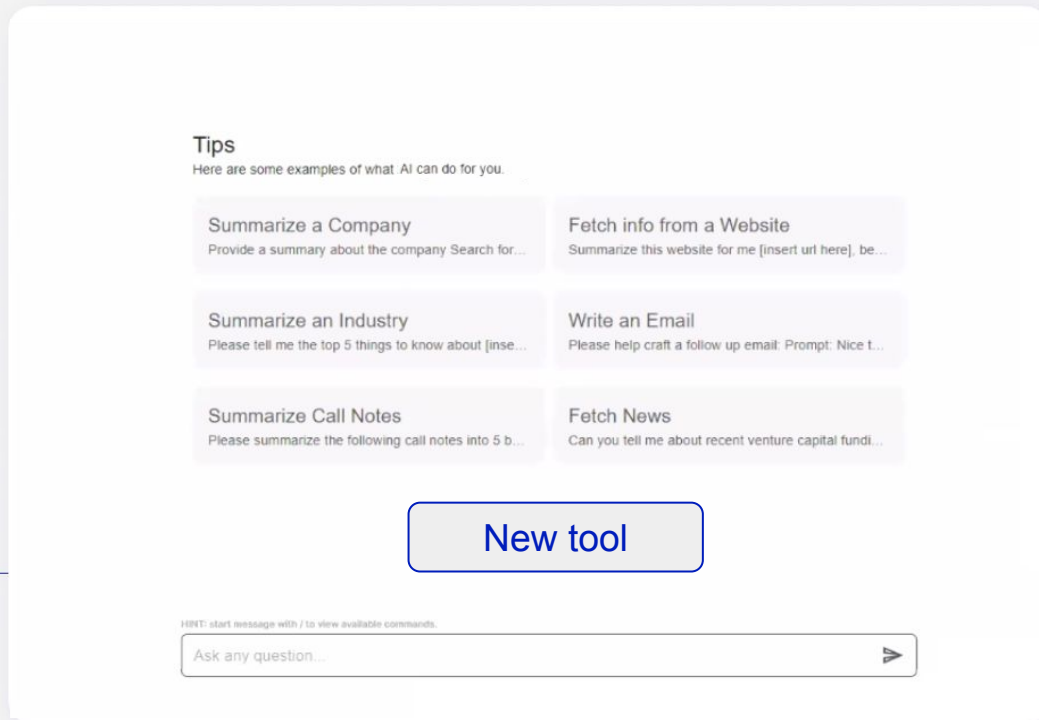
Case Study

- Ciklum case study -building the product
- Testing of the LLM-based solutions

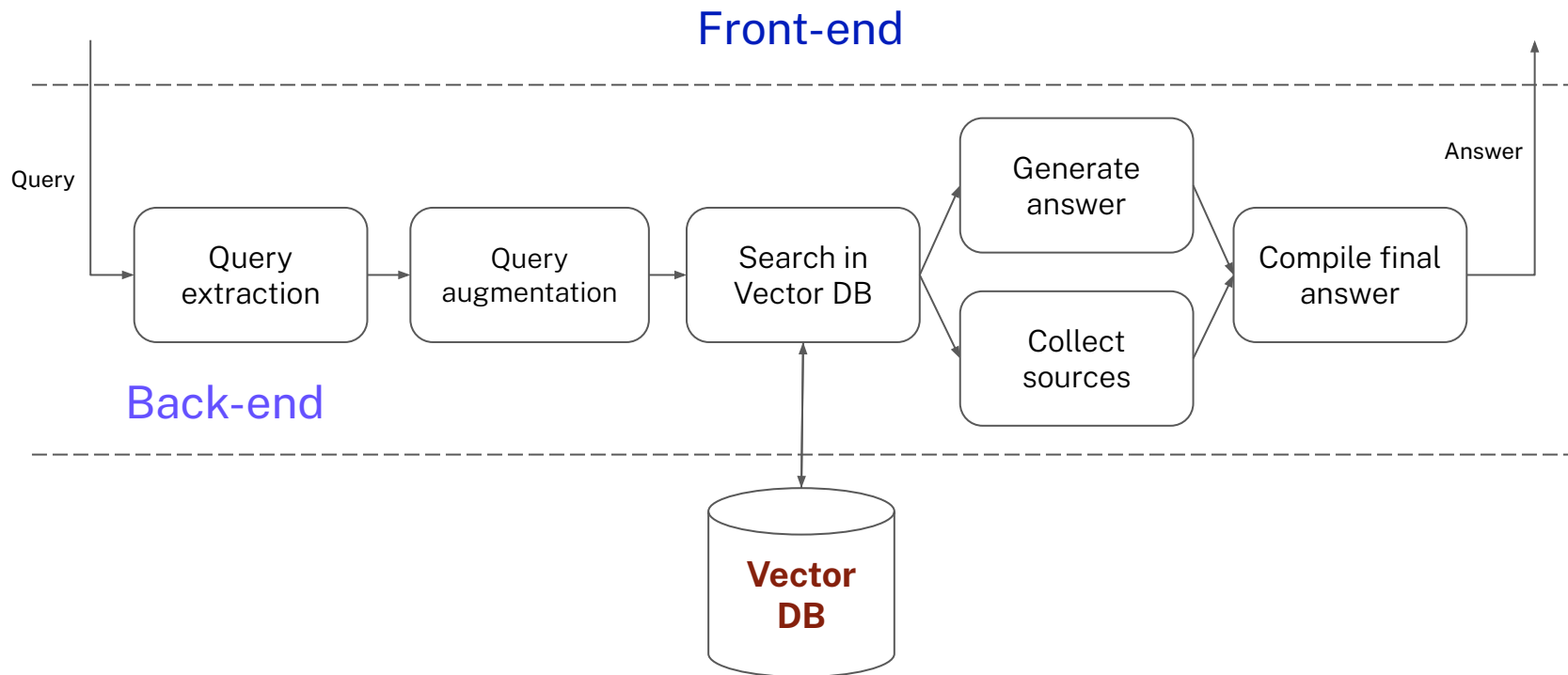
About the project

Problem: Advisors at an investment company often face questions outside their area of expertise. They need a way to provide accurate responses and link to relevant internal resources for further information.

Solution: Develop a ChatGPT-like application that helps advisors answer questions using an internal knowledge base, including PDF documents, YouTube videos, and an internal educational website. The app will provide direct answers and precise links to specific content, such as page numbers and video timestamps, to facilitate deeper exploration.



System components



QA Stages

- Data quality assurance
- Testing technical implementation
- Testing answer quality

Data Quality Assurance

Major issues:

- Extracted texts are not readable by the LLM
- Some information is not relevant anymore
- Data contains contradictions
- Documents lack of context (especially for presentations)
- Lot's of redundant data and repetitions
- Noisy data (e.g. data contains examples and templates)
- No metadata available (e.g. date of creation, tags)

Data Quality: Text extraction

Original document

Statement of financial position
as at March 31 2018

	Notes	2017-18 £'000	Restated 2016-17 £'000
Non-current assets			
Intangible fixed assets	4	124	168
Property, plant and equipment	7	315	640
Current assets			
Investments - balance at the NLDF	12		
Trade and other receivables	8		
Cash and cash equivalents			

Total assets

Processed with Textract

Statement of financial position
as at March 31 2018

Non-current assets
Intangible fixed assets

Property, plant and equipment

Current assets
Investments - balance at the NLDF
Trade and other receivables

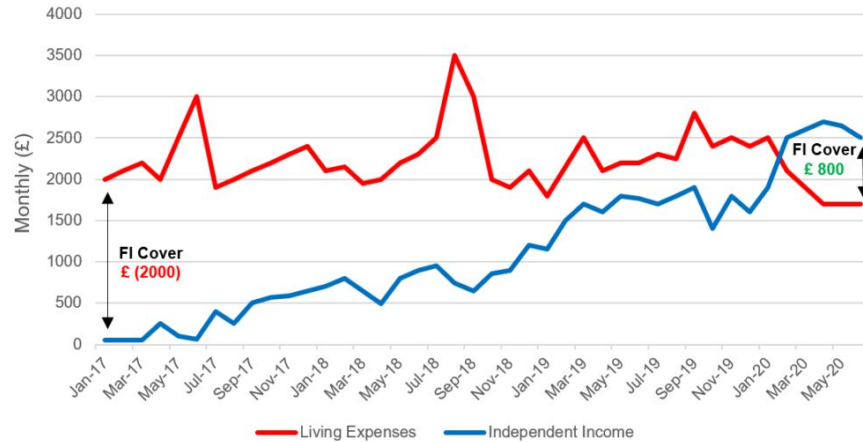
and cash equivalents

Processed with OpenAI

```
1  **Statement of financial position**
2  **as at March 31 2018**
3
4  |-----|-----|-----|-----|
5  |-----|-----|-----|-----|
6  | **Non-current assets** | | | |
7  | Intangible fixed assets | 4 | 124 | 168 |
8  | Property, plant and equipment | 7 | 315 | 640 |
9  |-----|-----|-----|-----|
10 | **Current assets** | | | |
11 | Investments - balance at the NLDF | 12 | 412,221 | 451,624 |
12 | Trade and other receivables | 8 | 4,558 | 2,441 |
13 | Cash and cash equivalents | | 255 | 1,822 |
14 | **Total assets** | | | |
15 |-----|-----|-----|-----|
16 | 417,473 | 456,695 |
```

Data Quality: Text extraction

Original document



Processed with OpenAI

1 The image is a line chart showing two data series over time from 2017 to May 2020. The y-axis represents monthly income/expenses in pounds (£), and the x-axis represents time in months and years.

2
3 Here's a breakdown of the text and information from the chart:

4
5
6
7 **### Key Elements in the Chart:**

8 - ****Red Line****: Represents "Living Expenses."

9 - ****Blue Line****: Represents "Independent Income."

10
11 **### Labels and Text:**

12 - ****FI Cover £ (2000)****: This label with an arrow indicates the financial independence (FI) gap at an earlier point on the chart, monthly living expenses were £2000, and independent income was low.

13 - ****FI Cover £ 800****: This label with an arrow indicates a later point on the chart where the financial independence gap has narrowed, with independent income closer to covering living expenses, resulting in a gap of £800.

14
15 **### Explanation of the Chart:**

16 The chart shows how independent income (blue line) gradually increases over time, while living expenses (red line) fluctuate but tend to stabilize around £2000–£2500 per month. By May 2020, independent income nearly matches or covers living expenses, reducing the FI cover gap to £800, suggesting progress toward financial independence.

17
18 **### Table Representation in Markdown (Approximated):**

Date	Living Expenses (£)	Independent Income (£)
Jan-17	2000	0
Mar-17	2000	(lower)
Sep-18	3000	800
Jan-19	2500	1000
Jul-19	2000	1500
Nov-19	2500	1750
Jan-20	2300	2000
May-20	2300	1500

29 This table simplifies the trend, capturing approximate points where the FI cover changes are noted in the chart.

Data Quality Assurance

REVIEW THE DATA!

- Ensure that extracted data is readable by LLM as plain text. Pay attention to tables and charts.
- If extracting texts from presentations, check if the text is in the proper order.
- Evaluate how often valuable data is stored as images, tables, and charts.
- Check if the document metadata is correct.

Technical QA

At this stage your goal is to make sure that all parts of your system are working as expected.

Typical issues:

- Random things are happening all the time
- Very different and inconsistent responses on the same requests.

Prepare a small set of requests for testing:

- A set of questions that have answers in the knowledge base for sure. You can use LLM for that ;)
- A set of questions that the knowledge base cannot cover.

If you use memory, develop a test to check if everything works fine when the memory buffer is full.

Sources of randomness

- **Agent**, that makes a decision which tool should be used and extracts short request from the whole message
- **Query augmentation and Keywords extraction** for search in the Vector database
- **Vector DB** resulting documents -different order and set of documents every time, sometimes returns no documents for the same query
- Answer generation by **GPT model**



Query preprocessing variability

Find information from learning materials to answer the question: "What is Community-Led Growth?"

Agent query extraction:

raw_query: Community-Led Growth

raw_query: What is Community-Led Growth?

Keyword extraction:

extracted_query: community growth led initiatives

extracted_query: community growth led initiatives grassroots

extracted_query: community-led growth community development grassroots initiatives local engagement

extracted_query: community growth led initiatives

extracted_query: community-led growth development initiatives grassroots

Collect answer metadata

You will thank yourself in the future if you implement metadata for answers to track impact of changes.

You will experiment with:

- Different APIs: OpenAI, Anthropic etc.
- K neighbors
- Chunk size, chunking algorithm
- Embedding model, vector size, dimensionality reduction
- Prompts (implement prompt versioning)

Introduce one change at a time!

```
answer = {
    "is_content_search": True,
    "query": query,
    "extracted_query": extracted_query,
    "n_docs": len(docs),
    "raw_answer": output["output_text"],
    "final_answer": final_answer,
    "urls": urls,
    "n_urls": len(urls),
    "EMBEDDINGS_MODEL": embeddings_model,
    "OPENSEARCH_CHUNK_SIZE": int(os.environ.get("OPENSEARCH_CHUNK_SIZE")),
    "OPENSEARCH_SEARCH_TYPE": os.environ.get("OPENSEARCH_SEARCH_TYPE"),
    "OPENSEARCH_ENGINE": os.environ.get("OPENSEARCH_ENGINE"),
    "OPENSEARCH_SPACE_TYPE": os.environ.get("OPENSEARCH_SPACE_TYPE"),
    "OPENSEARCH_EF_CONSTRUCTION": int(os.environ.get("OPENSEARCH_EF_CONSTRUCTION")),
    "OPENSEARCH_M": int(os.environ.get("OPENSEARCH_M")),
    "OPENSEARCH_SEARCH_K": int(os.environ.get("OPENSEARCH_SEARCH_K")),
    "OPENAI_API_TYPE": os.environ.get("OPENAI_API_TYPE"),
    "OPENAI_API_VERSION": os.environ.get("OPENAI_API_VERSION"),
    "CONVERSATIONAL_FOUNDATION_MODEL": os.environ.get("CONVERSATIONAL_FOUNDATION_MODEL"),
    "n_embedded_chunks": n_embedded_chunks,
    "n_embedded_objects": n_embedded_objects,
    "no_answer": no_answer
}
```

** TIP: Create multiple tables in your Vector DB to store embeddings with different configurations. It will allow you to experiment with different configurations without re-extracting embeddings for all your data. Example of a table name "embeddings_{MODEL_NAME}_{CHUNK_SIZE}"*

Tech QA:

What metadata can show you

How often proper tool was applied depending on the prompt?

	Proper tool was applied, %
prompt	
Find information from learning materials to answer the question: "{USER_REQUEST}"	100.000000
learning content about {USER_REQUEST}	100.000000
search content {USER_REQUEST}	28.285714

How answer quality depends on number of sources used?

OPENSEARCH_SEARCH_K	Answer found, %	Average # of sources from OpenSearch	Average # of filtered sources	Average answer length, chars
1.0	61.428571	1.0	1.000000	568.928571
2.0	71.428571	2.0	1.800000	852.242857
4.0	68.571429	4.0	3.583333	1038.442857
8.0	51.428571	8.0	6.388889	1008.057143
16.0	11.428571	16.0	8.000000	249.728571

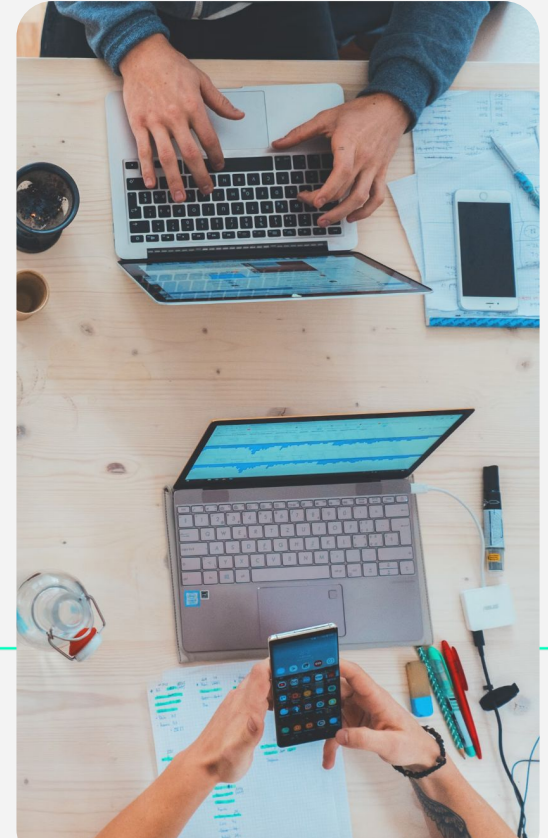
More detailed analysis

prompt	user_request	Proper tool was applied, %	Answer found, %	Average # of sources from OpenSearch	Average # of filtered sources
Find information from learning materials to answer the question: "{USER_REQUEST}"	What are the drawbacks of moving to the cloud from a security perspective?	100.0	50.0	8.0	2.900000
	What are the key benefits of community-led growth?	100.0	90.0	8.0	4.800000
	What are the potential benefits of moving to the cloud from a security perspective?	100.0	10.0	8.0	0.200000
	What is Community-Led Growth?	100.0	10.0	8.0	0.700000
	What is the cross-border team and what is their role?	100.0	0.0	8.0	0.000000
	What is the forecast for the importance of cloud security in the next five years?	100.0	100.0	8.0	7.400000
	Why are people interested in allocating a significant portion of their marketing spend towards community efforts?	100.0	100.0	8.0	7.000000
learning content about {USER_REQUEST}	What are the drawbacks of moving to the cloud from a security perspective?	100.0	60.0	8.0	3.800000
	What are the key benefits of community-led growth?	100.0	90.0	8.0	3.200000
	What are the potential benefits of moving to the cloud from a security perspective?	100.0	0.0	8.0	0.000000
	What is Community-Led Growth?	100.0	80.0	8.0	3.500000
	What is the cross-border team and what is their role?	100.0	0.0	8.0	0.000000
	What is the forecast for the importance of cloud security in the next five years?	100.0	70.0	8.0	5.400000
	Why are people interested in allocating a significant portion of their marketing spend towards community efforts?	100.0	80.0	8.0	4.700000
search content {USER_REQUEST}	What are the drawbacks of moving to the cloud from a security perspective?	70.0	60.0	8.0	4.428571
	What are the key benefits of community-led growth?	70.0	60.0	8.0	3.857143
	What are the potential benefits of moving to the cloud from a security perspective?	50.0	0.0	8.0	0.000000
	What is Community-Led Growth?	10.0	10.0	8.0	3.000000
	What is the cross-border team and what is their role?	30.0	0.0	8.0	0.000000
	What is the forecast for the importance of cloud security in the next five years?	10.0	10.0	8.0	2.000000
	Why are people interested in allocating a significant portion of their marketing spend towards community efforts?	0.0	0.0	0.0	0.000000

Answer Quality Assurance

Typical issues:

- Hallucinations
- LLM is using internal knowledge instead of provided context or provides answers with mixed information
- Not clear why the answer isn't as good as expected
- Not clear how to evaluate the system
- Not clear how to decide if new version is actually better than the previous one
- Lack of user feedback for consistent evaluation
- Users don't follow instructions on how to use the system



Hallucination patterns

Find information from learning materials to answer the question: "What is the forecast for the importance of cloud security in the next ten years?"


THIS ANSWER WAS GENERATED BY CONTENT SEARCH: There is no information provided about the cloud security forecast ten years or any mention of **Michael Jackson** by the president. Sources: N/A




Michael Jackson · 2nd 








Venture Capitalist
Greater Paris Metropolitan Region

 157K followers

 Talks about #tech, #deeptech, #startups, #venturecapital, and #entrepreneurship

unrealistic expectations, and not isolating oneself or neglecting social connections. The article also provides guidelines for reducing the baby's risk of developing diabetes later in life. The president did not mention **Michael Jackson** in his speech. The article provides tips for reducing screen time for babies and toddlers and suggests fun home activities without screens. The article also provides guidelines for

langchain-ai / langchain Public 

<> Code  Issues 2.4k  Pull requests 375  Discussions  Actions  Projects  Security  Insights

Hallucinating Question about Michael Jackson #2510

 Open mattliscia opened this issue on Apr 6 · 3 comments



mattliscia commented on Apr 6 · edited

Following the tutorial for `load_qa_with_sources_chain` using the example `state_of_the_union.txt` I encounter interesting situations. Sometimes when I ask a query such as "What did Biden say about Ukraine?" I get a response like this:
"Joe Biden talked about the Ukrainian people's fearlessness, courage, and determination in the face of Russian aggression. He also announced that the United States will provide military, economic, and humanitarian assistance to Ukraine, including more than \$1 billion in direct assistance. He further emphasized that the United States and its allies will defend every inch of territory of NATO countries, including Ukraine, with the full force of their collective power. However, he mentioned nothing about **Michael Jackson**."

I know that there are examples directly asking about Michael Jackson in the documentation:
https://python.langchain.com/en/latest/use_cases/evaluation/data_augmented_question_answering.html?highlight=michael%20jackson#examples

Here is my code for reproducing situation:

Assignees

No one assigned

Labels

None yet

Projects

None yet

Milestone

No milestone

Hallucination patterns

Find information from learning materials to answer the question: "What is the cross-border team and what is the role of the cross-border team?"

CONTENT SEARCH RESULT:

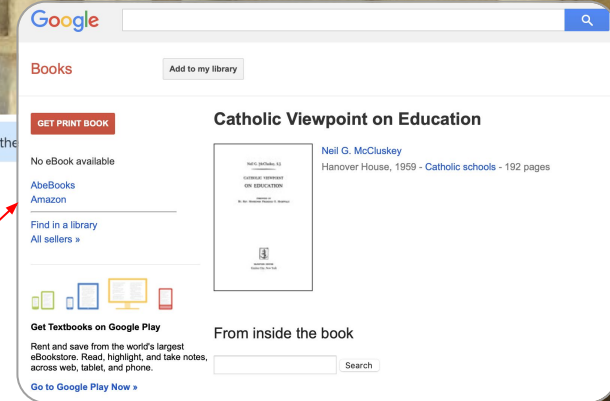
The cross-border team is a specialized group within a company that handles international tax planning, compliance, and reporting. Their role typically includes:

- Advising on structuring foreign operations and investments to optimize tax efficiency
- Ensuring compliance with tax laws and regulations across different countries
- Managing transfer pricing policies and documentation
- Coordinating tax audits and risk management
- Filing tax returns and managing compliance obligations in foreign jurisdictions
- Managing withholding taxes, VAT, and customs duties
- Providing tax support for cross-border transactions and restructurings
- Liaising with local finance teams and external advisors on international tax matters

The cross-border tax team works closely with other groups like legal, finance, and business units to provide comprehensive tax support for global operations. They aim to maximize tax savings, minimize compliance costs, and mitigate tax risks associated with international expansion. Their specialized knowledge is key for multinational companies to navigate the complexities of cross-border taxation.

SOURCES:

- International Taxation in America for the Entrepreneur 2013 Edition: <https://books.google.com/books?id=MmWdtgEACAAJ>
- The Taxation of Companies in Europe: <https://books.google.com/books?id=p-NgDwAAQBAJ>
- Managing the Risks of Cross-Border Taxation: <https://www.imanet.org/insights-and-trends/risk-management-in-practice/managing-the-risks-of-cross-border-taxation?ssopc=1>
- How Finance Teams Can Manage Cross-Border Complexity: <https://www.wallstreetprep.com/knowledge/finance-teams-manage-cross-border-complexity/>



Hallucination patterns

- Seen more often when no answer was found. It is hard to say “**I don’t know**”.
- LLMs are not designed to work with calculations, abstraction and logic. Don’t be surprised when they fail at these tasks.
- Add test cases when the correct answer is “**I don’t know**”.
- Add test cases, that are relevant to your data, but are out of your use cases (when the correct answer is “**I can’t do that**”).
- Hallucinations depend a lot on LLM.

QA engineers joined the team

Examples of test queries designed by QA engineers:

- Who is the president of United States?
- My name is John. ... What is my name?
- How to calculate company valuation? Answer in French
- How to start a startup? What is product-market fit? How to grow fast? What MMR should I reach before considering investments?
- Make sure you are testing your system instead of LLM (unless you want to test it as well)
- You may ask your QA engineers to test the system without giving them instructions first. It will give you major ideas to prepare instructions for future users.
- QA engineers need at least basic understanding what are RAGs and LLMs to design really useful test cases.

First tests with real users

	user_query
0	test
1	Test
2	This is a test.
3	test
4	hi
5	teest
6	are you active?
7	hello!
8	This is a test message?
9	Hi, Richard!
10	test
11	What is the best way to use this tool
12	are you able to summarize documents?
13	Please summarize openai.com

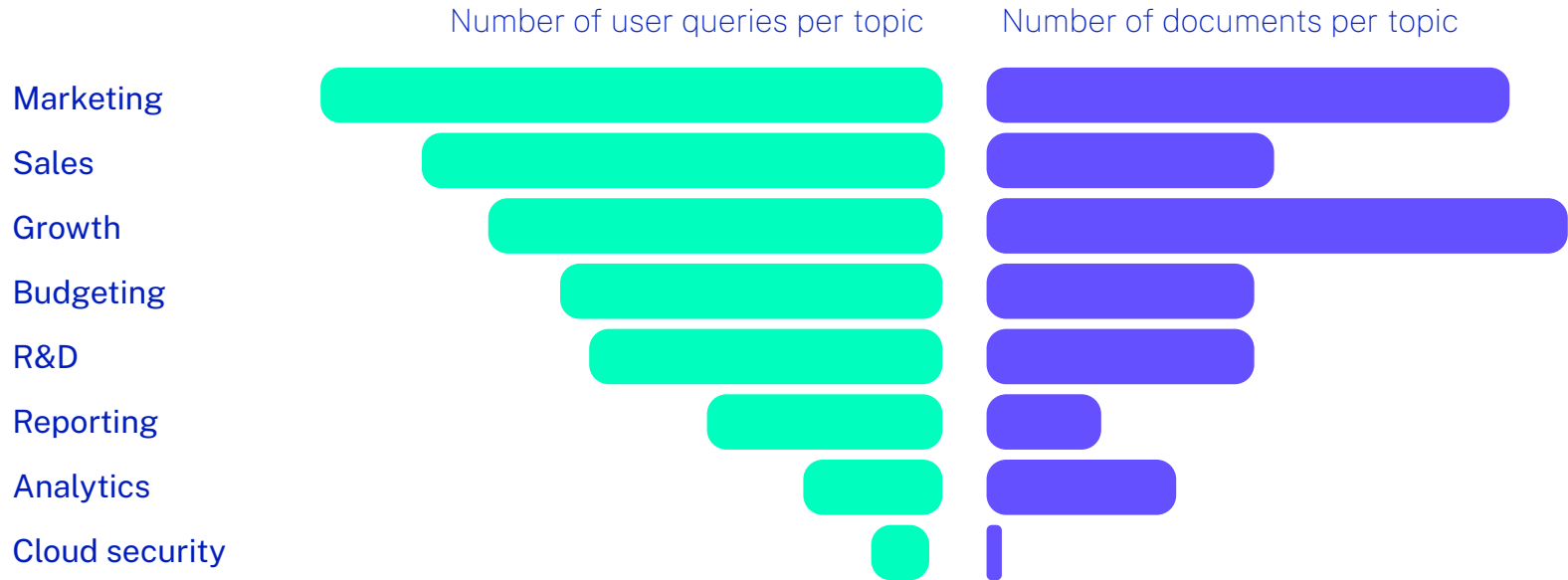
- Users will still interpret instructions in surprising ways and this is fine
- They will try to use your RAG in a similar way they are using ChatGPT

I want to write an article about marketing in AI startup. Write an article covering five questions: 1) How can AI startups differentiate their brand in a rapidly growing and competitive market? 2) ...

- Collect all requests and metadata and analyze it

Identify missing data

Tagging user queries and documents



Final answer scoring

- Like/dislike
- Stars (e.g. score from 1 to 5)
- Pick the best answer

Is your RAG better than ChatGPT?

Question			
Answer 1		Answer 2	
Usability	4	Usability	6
Accuracy	8	Accuracy	7
Which one do you prefer?			2

Blinded side-by-side comparison

Advisors rated and selected their preferred answer in a side-by-side comparison, without knowing where each answer came from.

120+ questions on various topics

Results:

- 72% -our answer is better
- 28% -Enterprise ChatGPT answer is better
- 60% of generated answers are usable for communication as is
- 62% considered factually accurate

Observations:

- Topics with more documents → higher accuracy score
- Answers with bulleted lists → higher usability score
- Answer length → no significant impact on usability score

Supervised testing and metrics

- Prepare a dataset of question-answer pairs to test the application and calculate accuracy metrics.
- Minimum valuable answer (MVA)
- Correct answer with incorrect tool or wrong data \neq correct answer
- Collect debug data, intermediate steps

Generated answers were labelled as:

- **Correct** -system picked proper tool, found relevant data and the answer is correct
- **Partially correct** -the answer contains only a part of data from MVA, proper tools were used
- **Incorrect** -either incorrect tool, irrelevant data or answer is incorrect.

Supervised testing and metrics

Question	Correct Answer	Correct SQL query	Generated Answer	Generated SQL Query	Data Source	Correct Document(s)	Document(s) used by RAG
How much did we invest in Google.	\$1M	SELECT INVESTED_AMOUNT FROM INVESTMENTS WHERE COMPANY = 'Google';	According to the data provided we invested one million dollars in Google	SELECT INVESTED_AMOUNT FROM INVESTMENTS WHERE LOWER(COMPANY) = 'google';	SQL	INVESTMENTS	INVESTMENTS
How many companies we have invested in are from US?	8	SELECT COUNT(*) FROM INVESTMENTS WHERE COUNTRY = 'United States';	We haven't invested in any companies from US.	SELECT COUNT(*) FROM INVESTMENTS WHERE COUNTRY = 'US';	SQL	INVESTMENTS	INVESTMENTS
Provide description about OpenAI	OpenAI is an American artificial intelligence (AI) research organization founded in December 2015 and headquartered in San Francisco, California.		OpenAI is an American artificial intelligence (AI) research organization founded in 2017 and headquartered in San Francisco, California.		Vector DB	openai.pdf	openai.pdf
Provide descriptions for AI companies we have invested in?	OpenAI is an American artificial intelligence (AI) research organization founded in December 2015 and headquartered in San Francisco, California. Anthropic PBC is a U.S.-based artificial intelligence (AI) public-benefit startup founded in 2021.	SELECT COMPANY_NAME FROM INVESTMENTS WHERE INDUSTRY = 'AI';	OpenAI is an American artificial intelligence (AI) research organization founded in 2017 and headquartered in San Francisco, California.	SELECT COMPANY_NAME FROM INVESTMENTS WHERE INDUSTRY = 'AI';	Both	openai.pdf, anthropic.pdf	openai.pdf

Conclusions

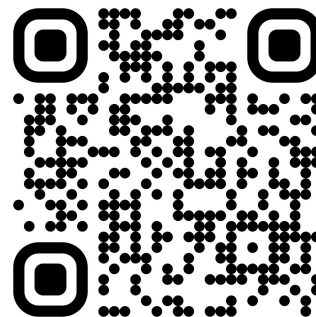
- Always review your data!
- Add metadata to your answers and log as much as possible.
- Don't forget to test if system is capable of saying "I don't know" and "I can't do that"
- Instruct your users before using your system.
- Provide a context about your data to the LLM in tool descriptions and system prompts.
- Make sure your are testing the system you've developed and not the LLM or service you are using (unless you want to test them)
- Balance evaluation efforts; don't pick too complicated metrics unless you need them

Closing remarks & lessons learned

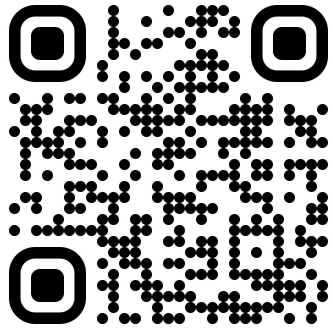


Any questions?

Share your
feedback!



Join our team





Thank you!

