Agentic RAG and the Future of AI-Powered Information Retrieval

Xpress AI

Hello. I’m Paul Dubs.

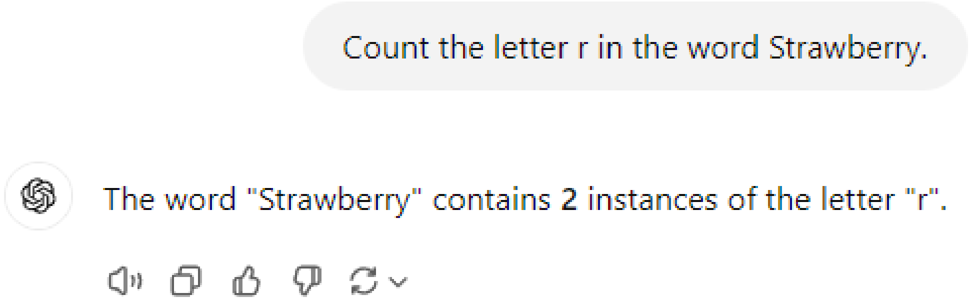
- Software developer for 20+ years
- ML practitioner for 10+ years
- CTO & Co-Founder of Xpress AI in Japan
- (Some) Previous Talks:
  - Specializing Conversational AI for Improved Accuracy and Trustworthiness
  - A Practical Guide to Vector Space Search
  - Conquering the Complexity of Agent Building with Xircuits

What is
Retrieval Augmented Generation?

I do not know who I am...
I don't know why I'm here
All I know is that I must PREDICT THE NEXT TOKEN

Why RAG?

- LLMs predict the next token based on statistics
    - May produce hallucinations
    - Confidently gives wrong answers
- Fix the statistics
    - Fine-tuning
    - Context manipulation

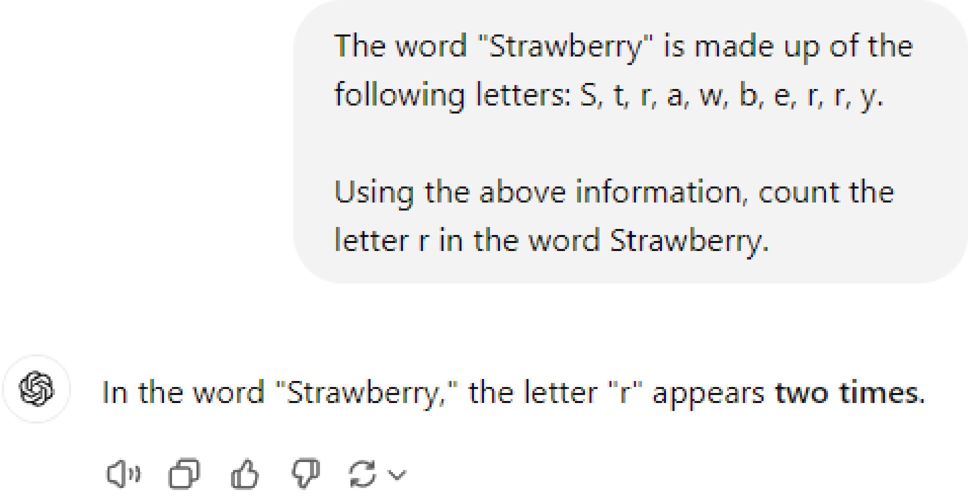


Count the letter r in the word Strawberry.

The word "Strawberry" contains **2** instances of the letter "r".

Why RAG?

- Fix the statistics with Context manipulation
- The original problem still exists, but it is less likely



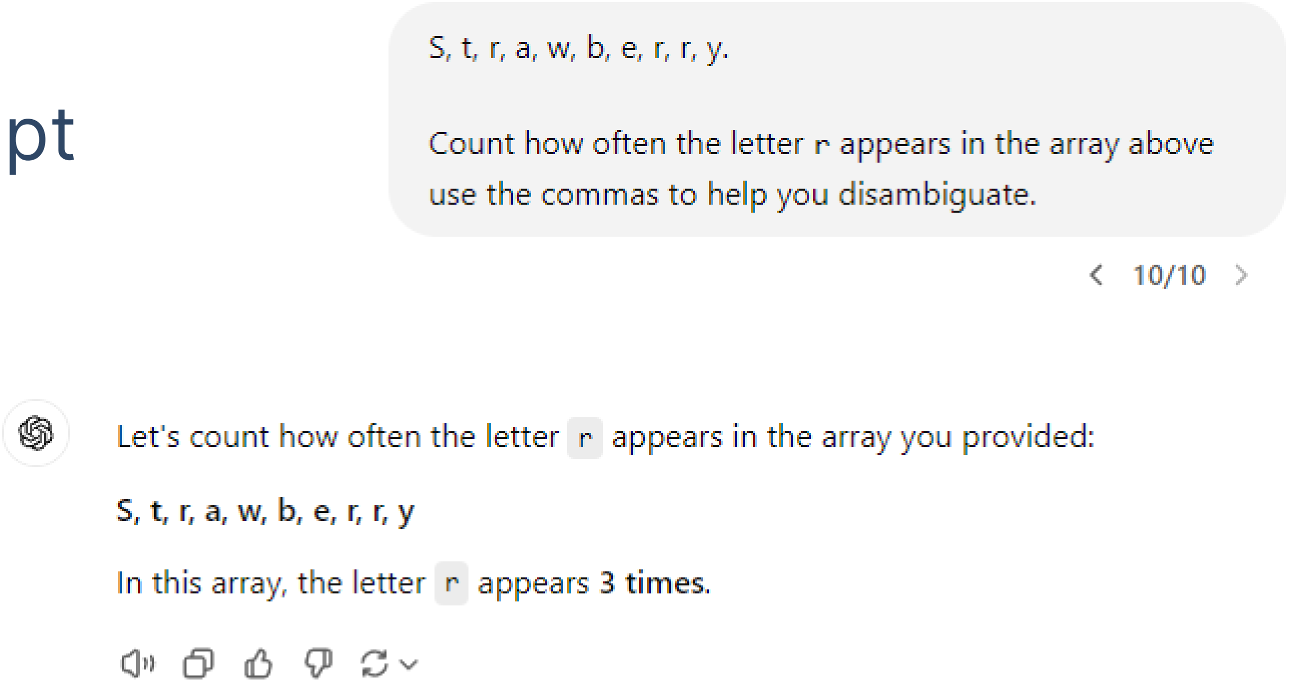
The word "Strawberry" is made up of the following letters: S, t, r, a, w, b, e, r, r, y.

Using the above information, count the letter r in the word Strawberry.

In the word "Strawberry," the letter "r" appears **two times.**

Why RAG?

- And it may take some prompt rewriting to get there



S, t, r, a, w, b, e, r, r, y.

Count how often the letter `r` appears in the array above use the commas to help you disambiguate.

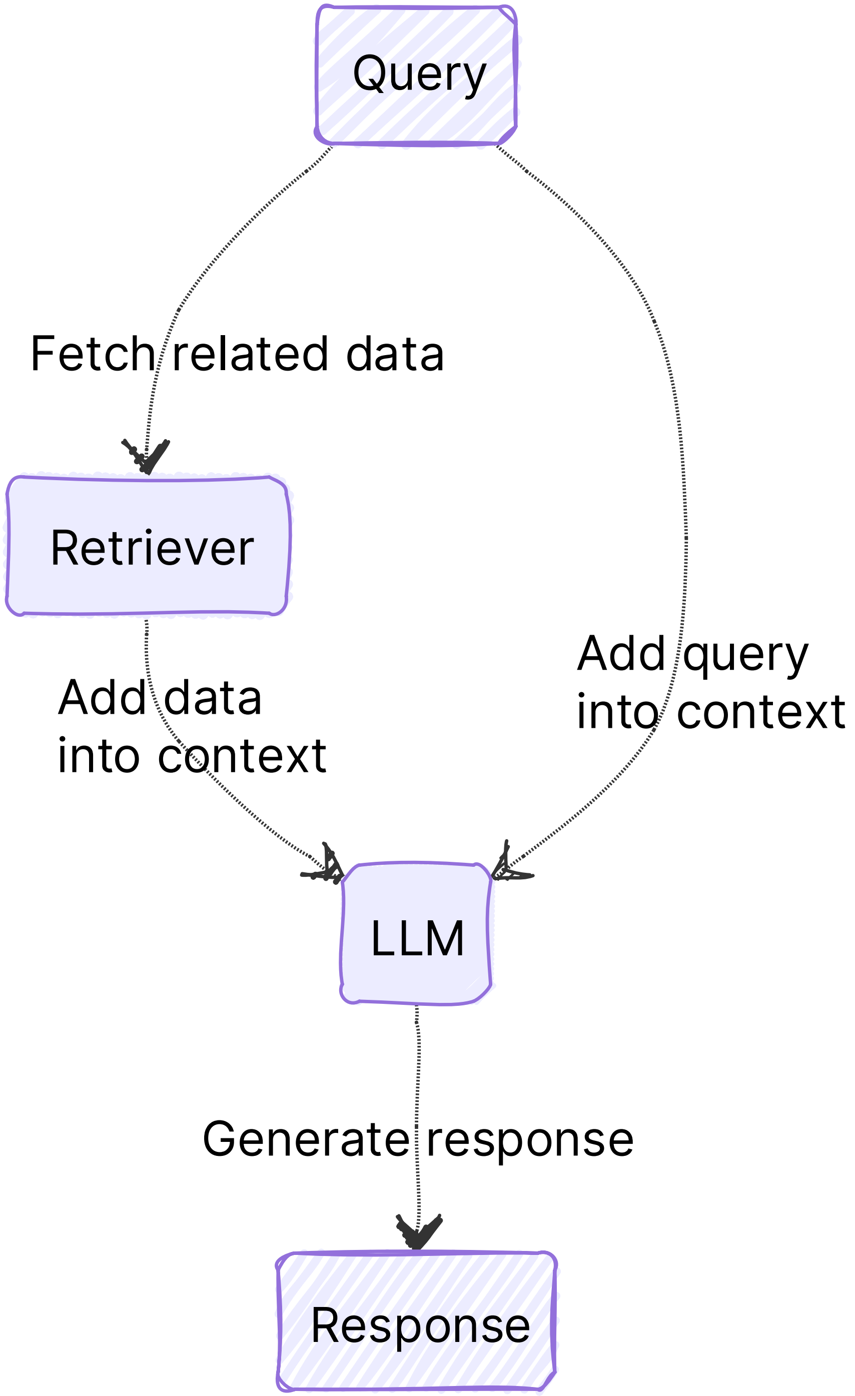
Let's count how often the letter `r` appears in the array you provided:

S, t, r, a, w, b, e, r, r, y

In this array, the letter `r` appears **3 times.**

What is **Naïve** RAG?

- Retrieve factual data
- Add data to the context
- Prompt LLM to consider only the retrieved data and the query in the response



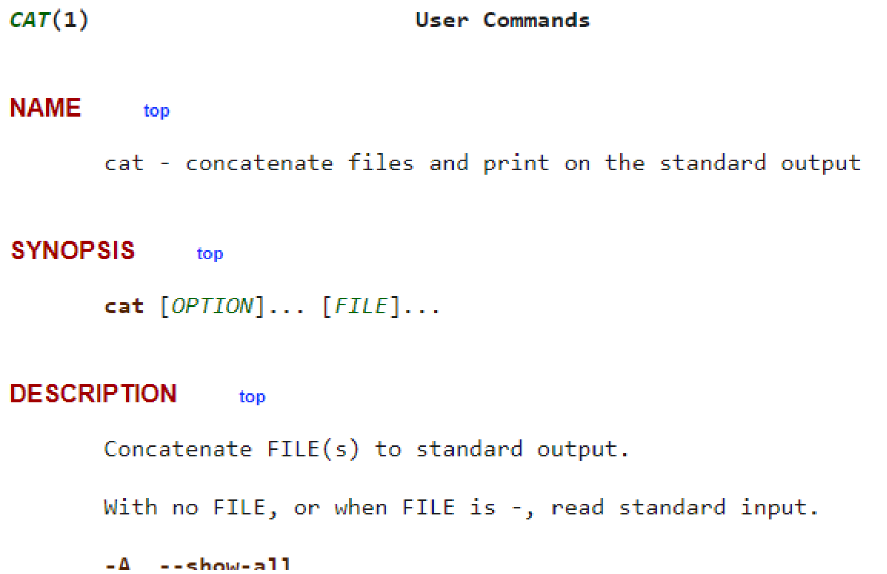
Naïve RAG:
Typical Retriever

- Transform documents to plain text
- Cut documents into chunks
- Vectorize chunks to embeddings
- Store embeddings and metadata linking back to chunks in vector database
- When queried, vectorize query to embedding
- Use query embedding to retrieve related document chunks

Semantic Gap:
Same meaning different words

- Cat
- Kitty
- Feline
- …

Semantic Gap:
Same word different meanings



```
CAT(1)                         User Commands

NAME       top

       cat - concatenate files and print on the standard output

SYNOPSIS       top

       cat [OPTION]... [FILE]...

DESCRIPTION       top

       Concatenate FILE(s) to standard output.

       With no FILE, or when FILE is -, read standard input.

       -A  --show-all
```

Embeddings close the Semantic Gap

- Similarity of embedding vectors is used as semantic similarity
- Semantic search only may be not enough
  - In specialized domains general purpose embedding models tend to perform worse than keyword based approaches
- Additional complexity:
  - What embedding model to choose?
  - What vector database to choose?
  - How to chunk the documents?
  - Who is going to maintain the additional infrastructure?
  - How much will ingesting all of our data cost?

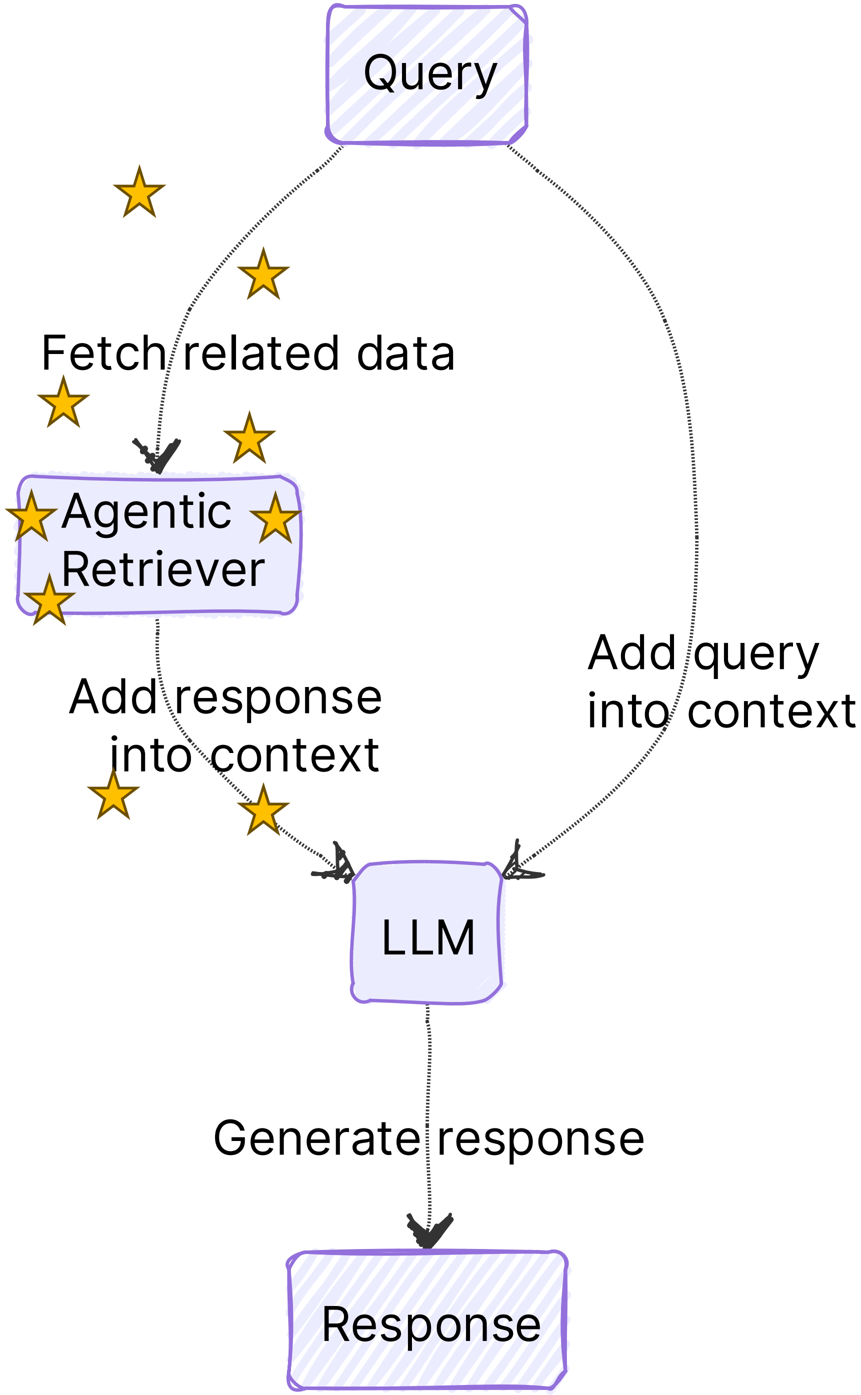
Agentic Retrieval

Close the Semantic Gap with Agents

- Let an agent search for relevant documents
- Agents built on top of LLMs have an innate ability to cover the semantic gap

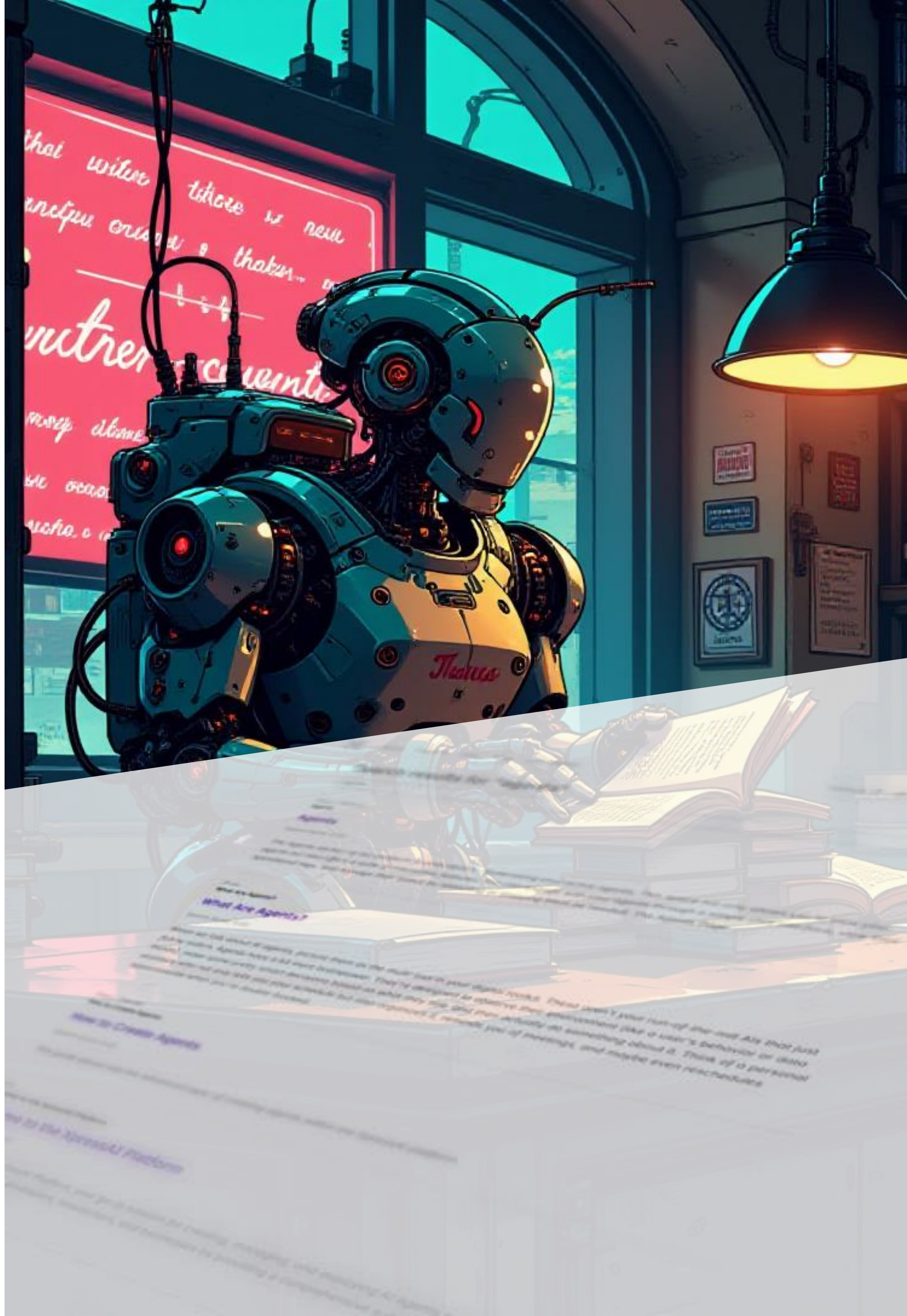
What is **Agentic RAG?**

- **Let an agent** retrieve factual data
- Add the agent’s response into the context
- Prompt LLM to consider only the retrieved data and the query in the response



Agentic RAG: Retriever Agent

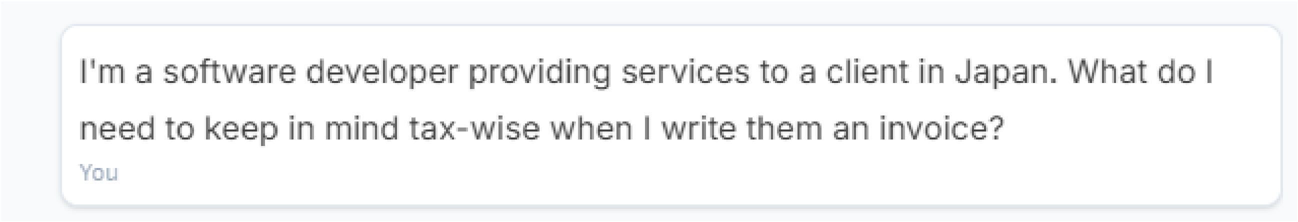
- Access a search API that returns results with a bit of preview
    - Existing search APIs
    - Full-text search in most SQL Databases
    - Even grepping through a directory
- Judge results in context of query (and maybe conversation)
- Refine search query
  – or –
  Return appropriate results



How does it work in practice?

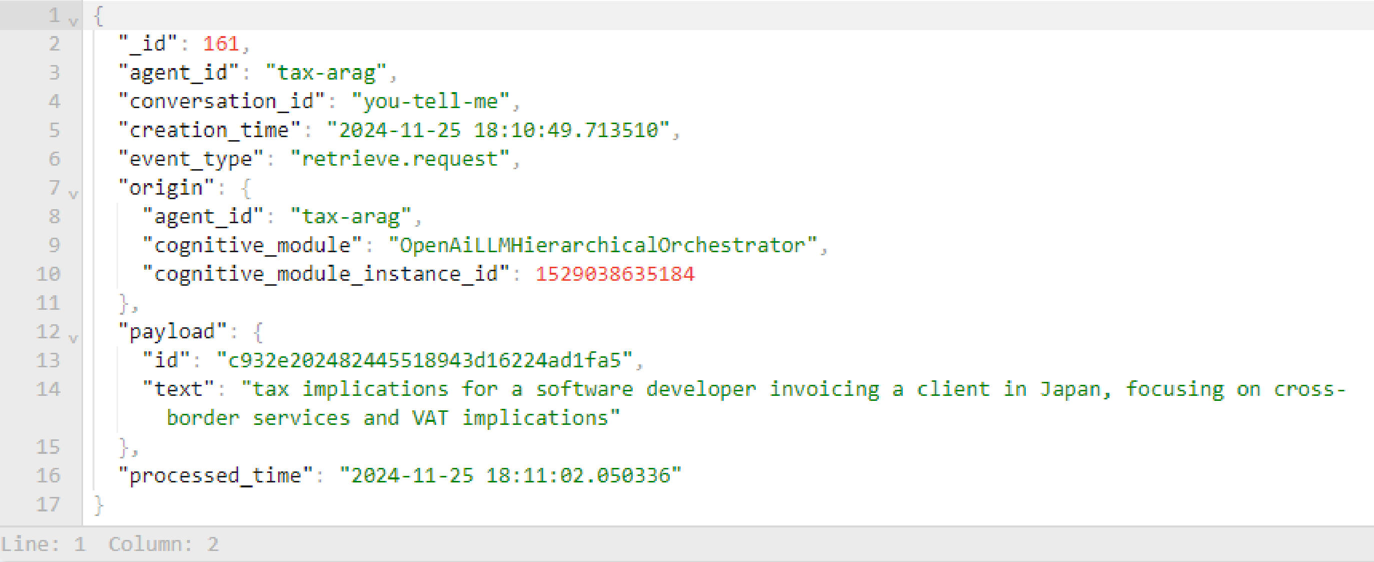
Agentic RAG in Action.

Query

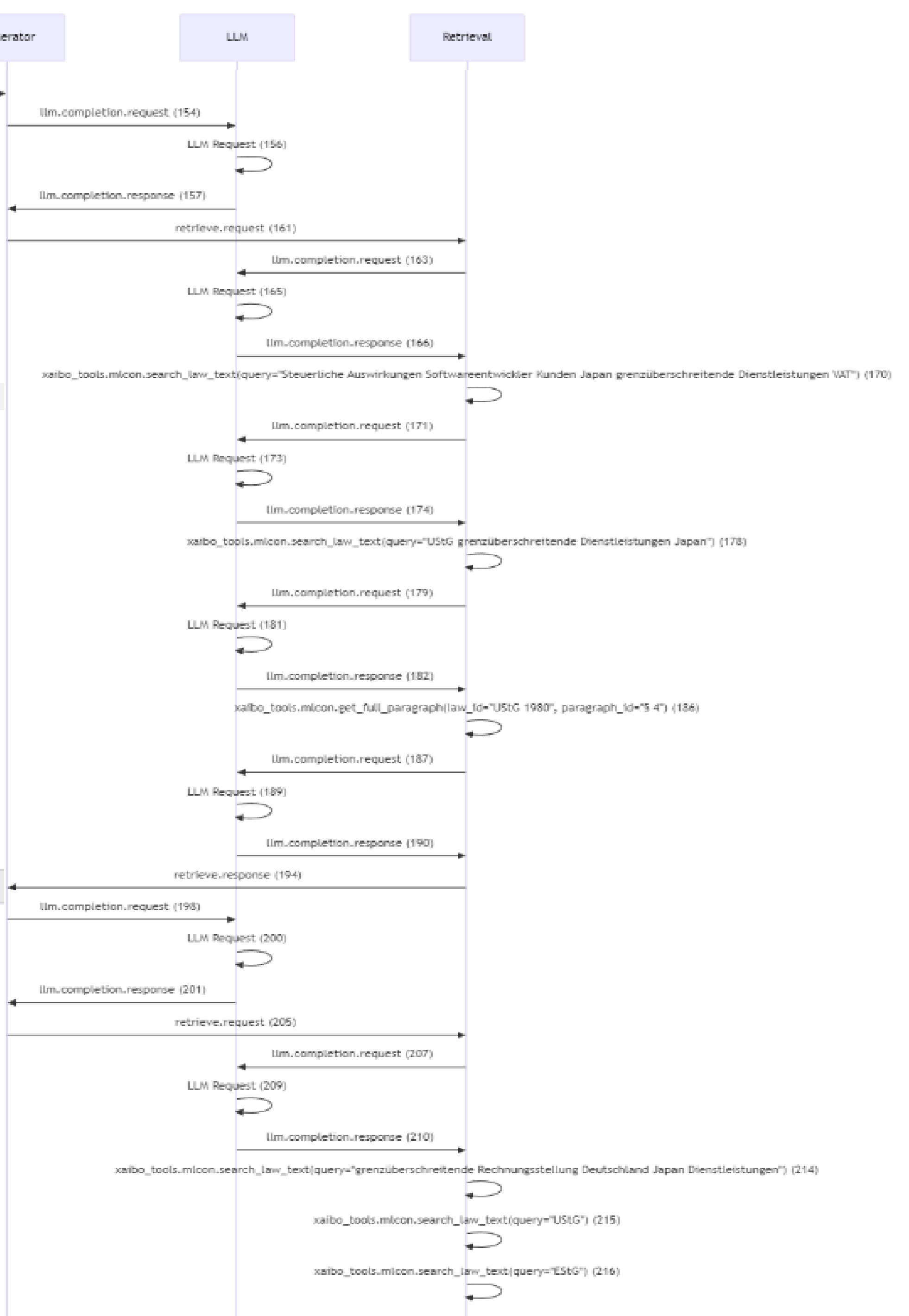


I'm a software developer providing services to a client in Japan. What do I need to keep in mind tax-wise when I write them an invoice?

Retrieval

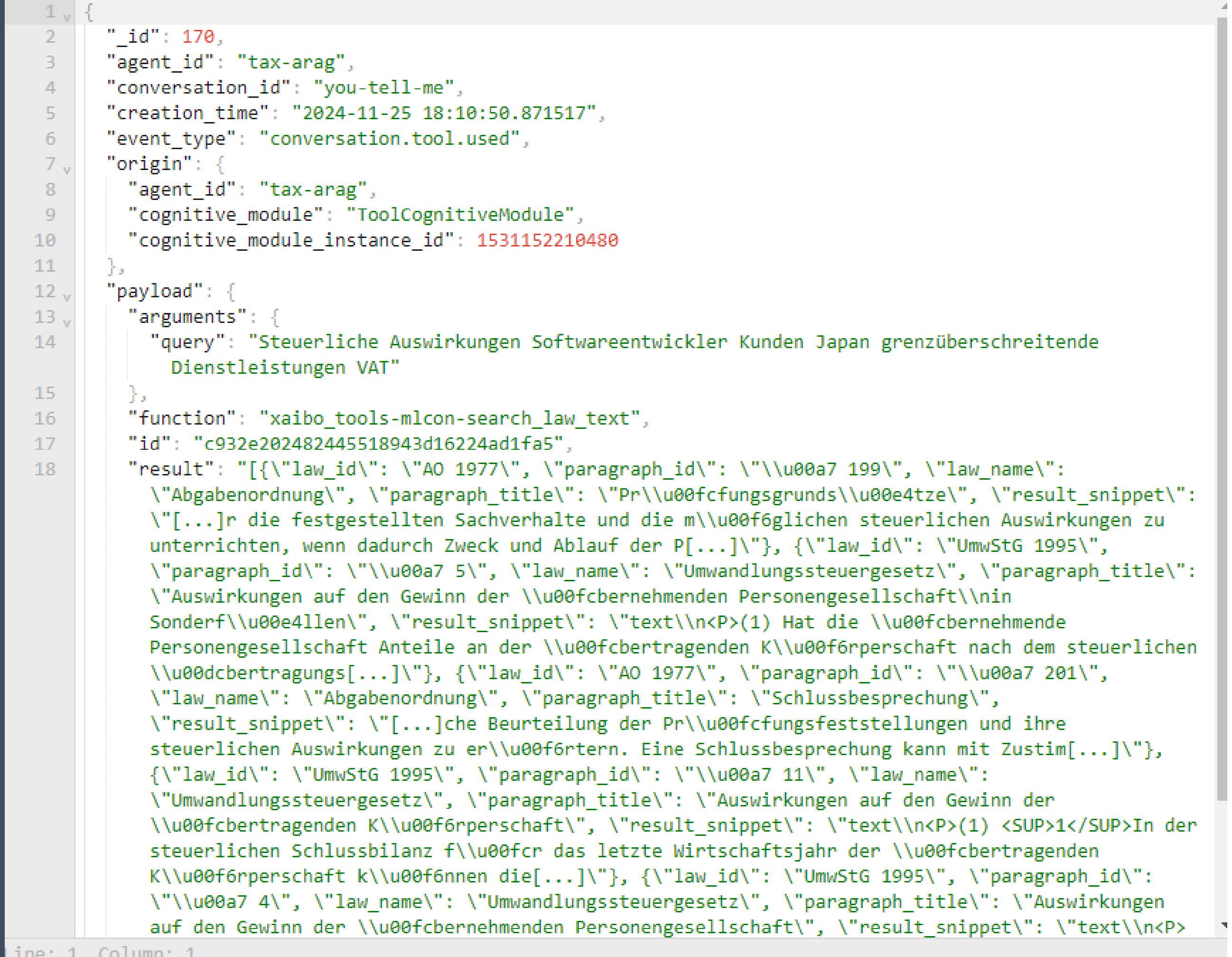


```
{
  "_id": 161,
  "agent_id": "tax-arag",
  "conversation_id": "you-tell-me",
  "creation_time": "2024-11-25 18:10:49.713510",
  "event_type": "retrieve.request",
  "origin": {
    "agent_id": "tax-arag",
    "cognitive_module": "OpenAiLLMHierarchicalOrchestrator",
    "cognitive_module_instance_id": 1529038635184
  },
  "payload": {
    "id": "c932e202482445518943d16224ad1fa5",
    "text": "tax implications for a software developer invoicing a client in Japan, focusing on cross-
      border services and VAT implications"
  },
  "processed_time": "2024-11-25 18:11:02.050336"
}
```

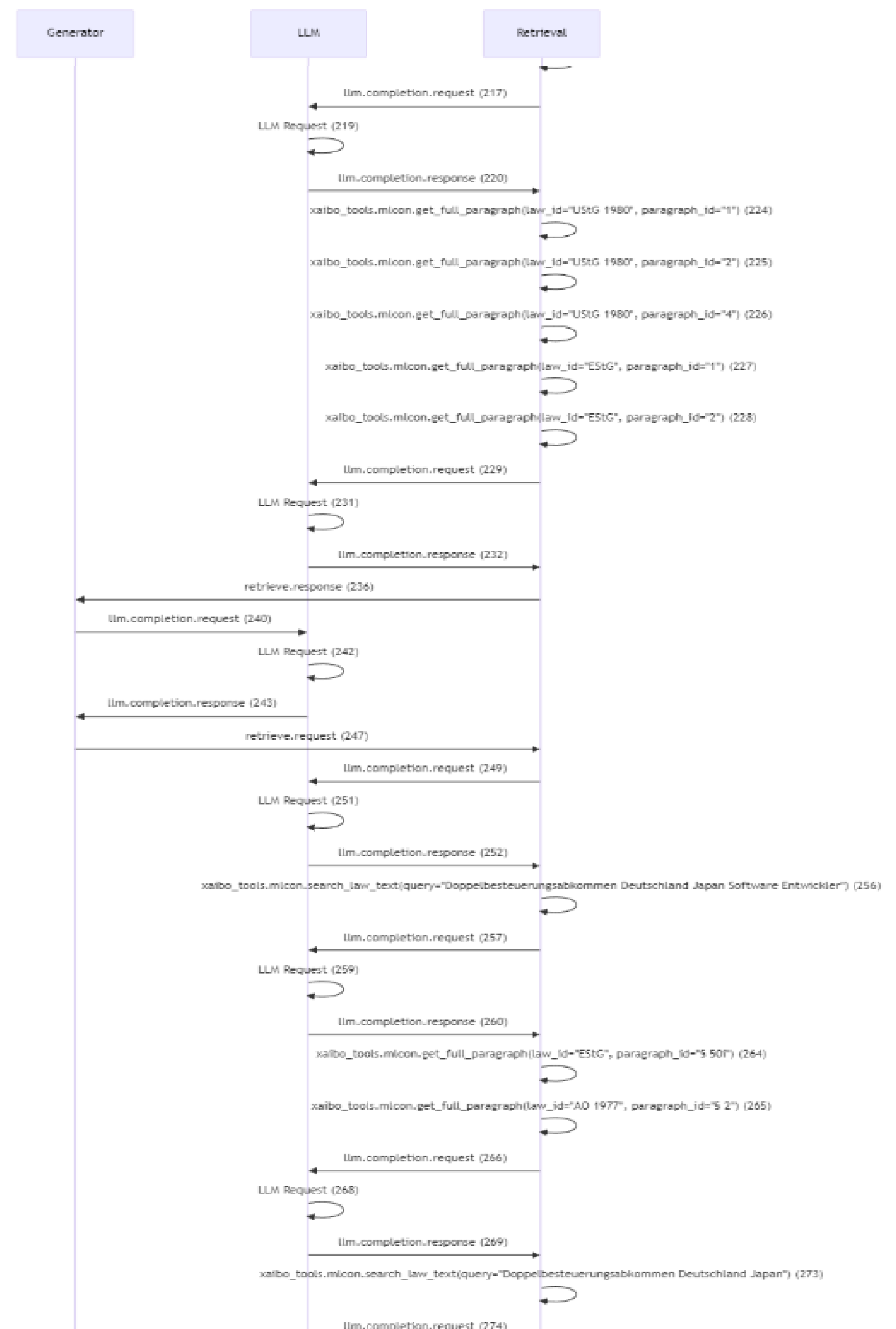


Tax-arag
Generator
LLM
Retrieval
conversation.message.new (152)
llm.completion.request (154)
LLM Request (156)
llm.completion.response (157)
retrieve.request (161)
llm.completion.request (163)
LLM Request (165)
llm.completion.response (166)
xaibo_tools.mlcon.search_law_text(query="Steuerliche Auswirkungen Softwareentwickler Kunden Japan grenzüberschreitende Dienstleistungen VAT") (170)
llm.completion.request (171)
LLM Request (173)
llm.completion.response (174)
xaibo_tools.mlcon.search_law_text(query="UStG grenzüberschreitende Dienstleistungen Japan") (178)
llm.completion.request (179)
LLM Request (181)
llm.completion.response (182)
xaibo_tools.mlcon.get_full_paragraph(law_id="UStG 1980", paragraph_id="§ 4") (186)
llm.completion.request (187)
LLM Request (189)
llm.completion.response (190)
retrieve.response (194)
llm.completion.request (198)
LLM Request (200)
llm.completion.response (201)
retrieve.request (205)
llm.completion.request (207)
LLM Request (209)
llm.completion.response (210)
xaibo_tools.mlcon.search_law_text(query="grenzüberschreitende Rechnungsstellung Deutschland Japan Dienstleistungen") (214)
xaibo_tools.mlcon.search_law_text(query="UStG") (215)
xaibo_tools.mlcon.search_law_text(query="EStG") (216)

Retrieval
Search Results



```
{
  "_id": 170,
  "agent_id": "tax-arag",
  "conversation_id": "you-tell-me",
  "creation_time": "2024-11-25 18:10:50.871517",
  "event_type": "conversation.tool.used",
  "origin": {
    "agent_id": "tax-arag",
    "cognitive_module": "ToolCognitiveModule",
    "cognitive_module_instance_id": 1531152210480
  },
  "payload": {
    "arguments": {
      "query": "Steuerliche Auswirkungen Softwareentwickler Kunden Japan grenzüberschreitende
       Dienstleistungen VAT"
    },
    "function": "xaibo_tools-mlcon-search_law_text",
    "id": "c932e202482445518943d16224ad1fa5",
    "result": "[{\"law_id\": \"AO 1977\", \"paragraph_id\": \"\\u00a7 199\", \"law_name\":
      \"Abgabenordnung\", \"paragraph_title\": \"Pr\\u00fcfungsgrunds\\u00e4tze\", \"result_snippet\":
      \"[...]r die festgestellten Sachverhalte und die m\\u00f6glichen steuerlichen Auswirkungen zu
      unterrichten, wenn dadurch Zweck und Ablauf der P[...]\"}, {\"law_id\": \"UmwStG 1995\",
      \"paragraph_id\": \"\\u00a7 5\", \"law_name\": \"Umwandlungssteuergesetz\", \"paragraph_title\":
      \"Auswirkungen auf den Gewinn der \\u00fcbernehmenden Personengesellschaft\\nin
      Sonderf\\u00e4llen\", \"result_snippet\": \"text\\n<P>(1) Hat die \\u00fcbernehmende
      Personengesellschaft Anteile an der \\u00fcbertragenden K\\u00f6rperschaft nach dem steuerlichen
      \\u00dcbertragungs[...]\"}, {\"law_id\": \"AO 1977\", \"paragraph_id\": \"\\u00a7 201\",
      \"law_name\": \"Abgabenordnung\", \"paragraph_title\": \"Schlussbesprechung\",
      \"result_snippet\": \"[...]che Beurteilung der Pr\\u00fcfungsfeststellungen und ihre
      steuerlichen Auswirkungen zu er\\u00f6rtern. Eine Schlussbesprechung kann mit Zustim[...]\"},
      {\"law_id\": \"UmwStG 1995\", \"paragraph_id\": \"\\u00a7 11\", \"law_name\":
      \"Umwandlungssteuergesetz\", \"paragraph_title\": \"Auswirkungen auf den Gewinn der
      \\u00fcbertragenden K\\u00f6rperschaft\", \"result_snippet\": \"text\\n<P>(1) <SUP>1</SUP>In der
      steuerlichen Schlussbilanz f\\u00fcr das letzte Wirtschaftsjahr der \\u00fcbertragenden
      K\\u00f6rperschaft k\\u00f6nnen die[...]\"}, {\"law_id\": \"UmwStG 1995\", \"paragraph_id\":
      \"\\u00a7 4\", \"law_name\": \"Umwandlungssteuergesetz\", \"paragraph_title\": \"Auswirkungen
      auf den Gewinn der \\u00fcbernehmenden Personengesellschaft\", \"result_snippet\": \"text\\n<P>
```

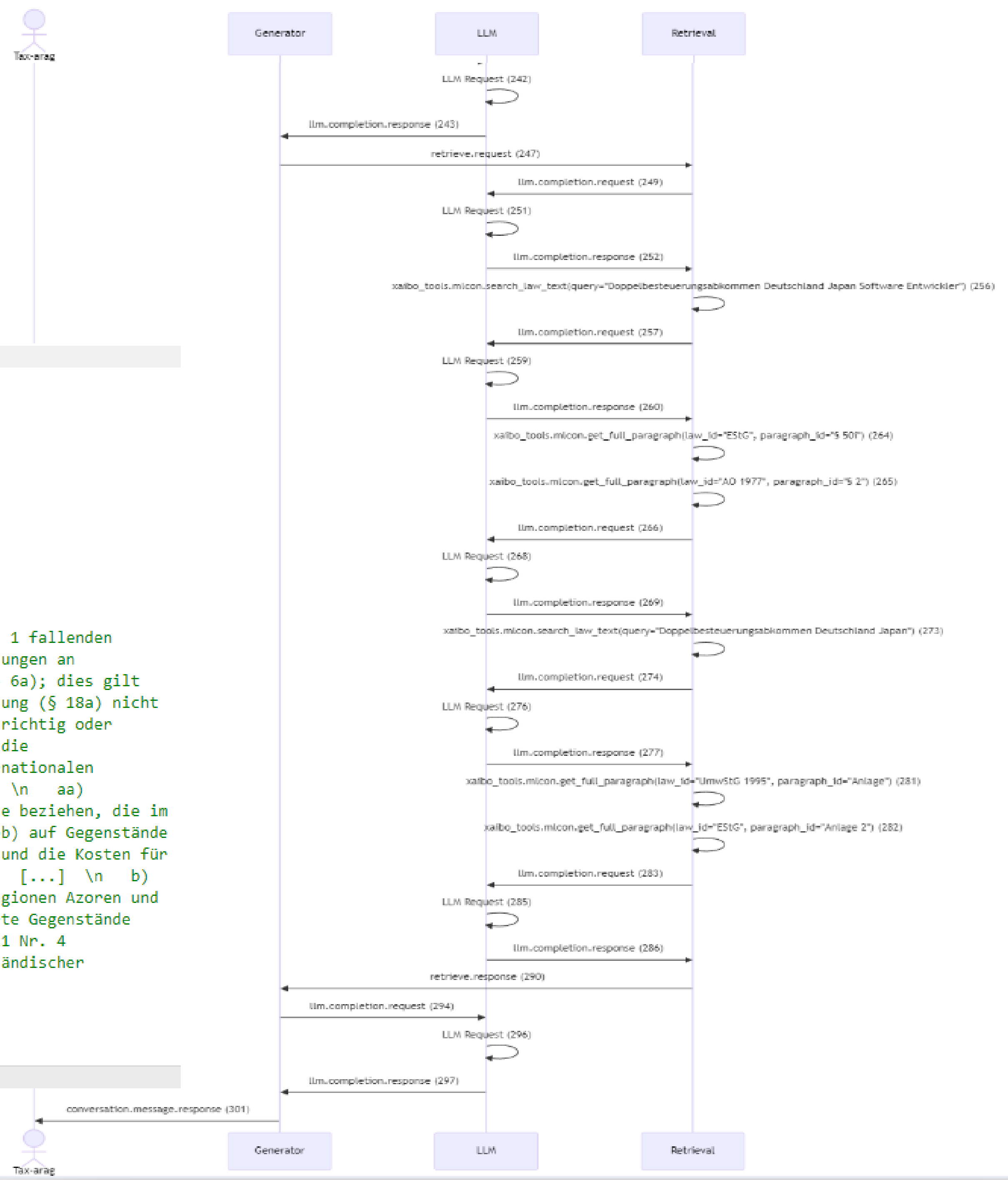


Tax-arag
Generator
LLM
Retrieval
llm.completion.request (217)
LLM Request (219)
llm.completion.response (220)
xaibo_tools.mlcon.get_full_paragraph(law_id="UStG 1980", paragraph_id="1") (224)
xaibo_tools.mlcon.get_full_paragraph(law_id="UStG 1980", paragraph_id="2") (225)
xaibo_tools.mlcon.get_full_paragraph(law_id="UStG 1980", paragraph_id="4") (226)
xaibo_tools.mlcon.get_full_paragraph(law_id="EStG", paragraph_id="1") (227)
xaibo_tools.mlcon.get_full_paragraph(law_id="EStG", paragraph_id="2") (228)
llm.completion.request (229)
LLM Request (231)
llm.completion.response (232)
retrieve.response (236)
llm.completion.request (240)
LLM Request (242)
llm.completion.response (243)
retrieve.request (247)
llm.completion.request (249)
LLM Request (251)
llm.completion.response (252)
xaibo_tools.mlcon.search_law_text(query="Doppelbesteuerungsabkommen Deutschland Japan Software Entwickler") (256)
llm.completion.request (257)
LLM Request (259)
llm.completion.response (260)
xaibo_tools.mlcon.get_full_paragraph(law_id="EStG", paragraph_id="§ 50f") (264)
xaibo_tools.mlcon.get_full_paragraph(law_id="AO 1977", paragraph_id="§ 2") (265)
llm.completion.request (266)
LLM Request (268)
llm.completion.response (269)
xaibo_tools.mlcon.search_law_text(query="Doppelbesteuerungsabkommen Deutschland Japan") (273)
llm.completion.request (274)

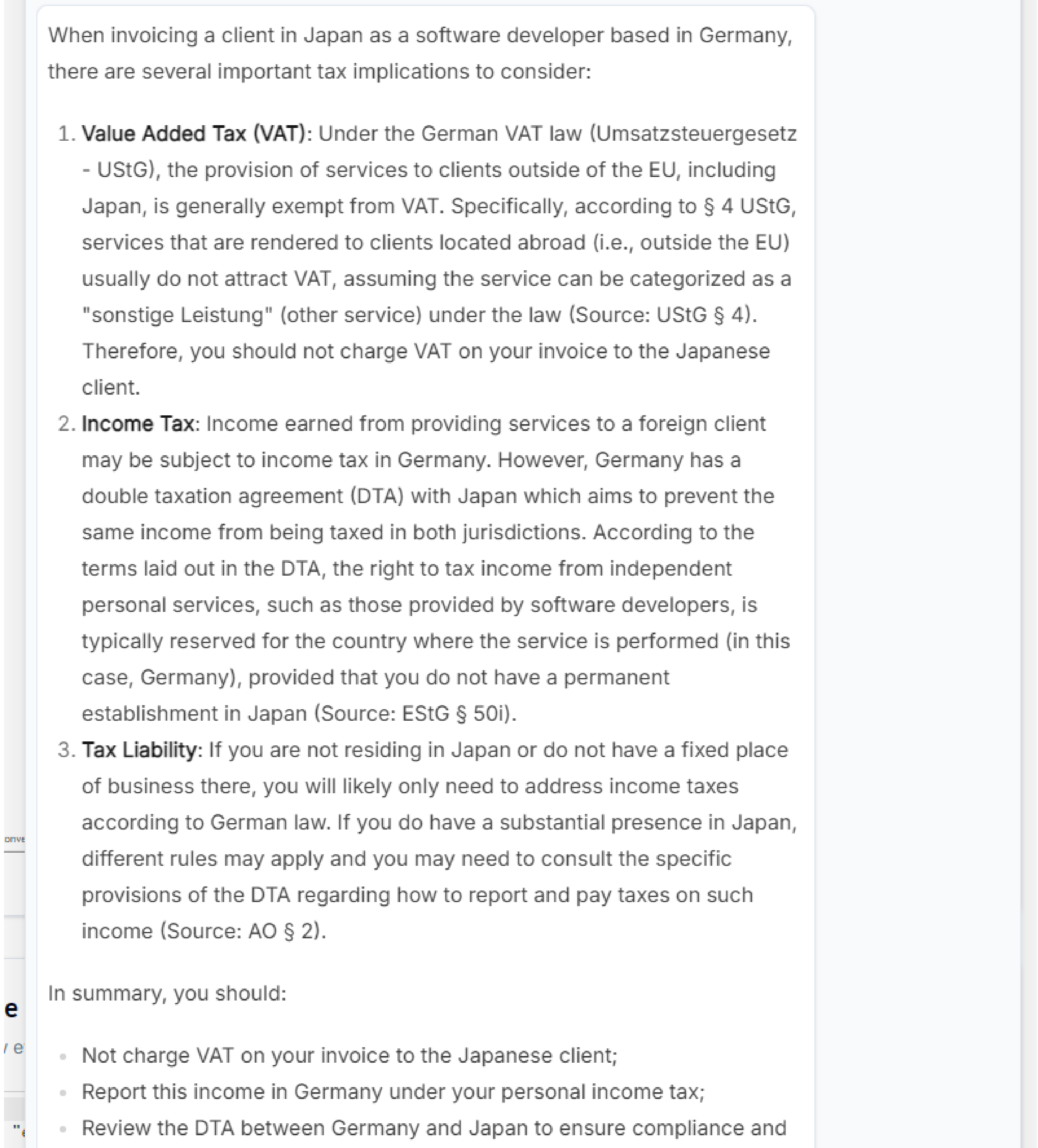
Retrieval Final Results

```
{
  "_id": 194,
  "agent_id": "tax-arag",
  "conversation_id": "you-tell-me",
  "creation_time": "2024-11-25 18:11:02.050336",
  "event_type": "retrieve.response",
  "origin": {
    "agent_id": "tax-arag",
    "cognitive_module": "ToolCognitiveModule",
    "cognitive_module_instance_id": 1531152210480
  },
  "payload": {
    "id": "c932e202482445518943d16224ad1fa5",
    "text": "<citation>\n§ 4 UStG - Umsatzsteuergesetz  \nVon den unter § 1 Abs. 1 Nr. 1 fallenden Umsätzen sind steuerfrei:  \n1. die Ausfuhrlieferungen (§ 6) und die Lohnveredelungen an Gegenständen der Ausfuhr (§ 7),  \n2. die innergemeinschaftlichen Lieferungen (§ 6a); dies gilt nicht, wenn der Unternehmer seiner Pflicht zur Abgabe der Zusammenfassenden Meldung (§ 18a) nicht nachgekommen ist oder soweit er diese im Hinblick auf die jeweilige Lieferung unrichtig oder unvollständig abgegeben hat;  \n3. die folgenden sonstigen Leistungen:  \n   a) die grenzüberschreitenden Beförderungen von Gegenständen, die Beförderungen im internationalen Eisenbahnfrachtverkehr und andere sonstige Leistungen, wenn sich die Leistungen  \n    aa) unmittelbar auf Gegenstände der Ausfuhr beziehen oder auf eingeführte Gegenstände beziehen, die im externen Versandverfahren in das Drittlandsgebiet befördert werden, oder  \n    bb) auf Gegenstände der Einfuhr in das Gebiet eines Mitgliedstaates der Europäischen Union beziehen und die Kosten für die Leistungen in der Bemessungsgrundlage für diese Einfuhr enthalten sind.  \n    [...]  \n    b) die Beförderungen von Gegenständen nach und von den Inseln, die die autonomen Regionen Azoren und Madeira bilden;  \n    c) sonstige Leistungen, die sich unmittelbar auf eingeführte Gegenstände beziehen, für die zollamtlich eine vorübergehende Verwendung in den in § 1 Abs. 1 Nr. 4 bezeichneten Gebieten bewilligt worden ist, wenn der Leistungsempfänger ein ausländischer Auftraggeber (§ 7 Abs. 2) ist.  \n    [...]  \n</citation>"
  },
  "processed_time": "2024-11-25 18:11:02.135408"
}
```

Line: 1 Column: 1


Tax-arag
Generator
LLM
Retrieval
LLM Request (242)
llm.completion.response (243)
retrieve.request (247)
llm.completion.request (249)
LLM Request (251)
llm.completion.response (252)
xaibo_tools.micon.search_law_text(query="Doppelbesteuerungsabkommen Deutschland Japan Software Entwickler") (256)
llm.completion.request (257)
LLM Request (259)
llm.completion.response (260)
xaibo_tools.micon.get_full_paragraph(law_id="EStG", paragraph_id="§ 50f") (264)
xaibo_tools.micon.get_full_paragraph(law_id="AO 1977", paragraph_id="§ 2") (265)
llm.completion.request (266)
LLM Request (268)
llm.completion.response (269)
xaibo_tools.micon.search_law_text(query="Doppelbesteuerungsabkommen Deutschland Japan") (273)
llm.completion.request (274)
LLM Request (276)
llm.completion.response (277)
xaibo_tools.micon.get_full_paragraph(law_id="UmwStG 1995", paragraph_id="Anlage") (281)
xaibo_tools.micon.get_full_paragraph(law_id="EStG", paragraph_id="Anlage 2") (282)
llm.completion.request (283)
LLM Request (285)
llm.completion.response (286)
retrieve.response (290)
llm.completion.request (294)
LLM Request (296)
llm.completion.response (297)
conversation.message.response (301)
Tax-arag
Generator
LLM
Retrieval

Response to User



When invoicing a client in Japan as a software developer based in Germany, there are several important tax implications to consider:

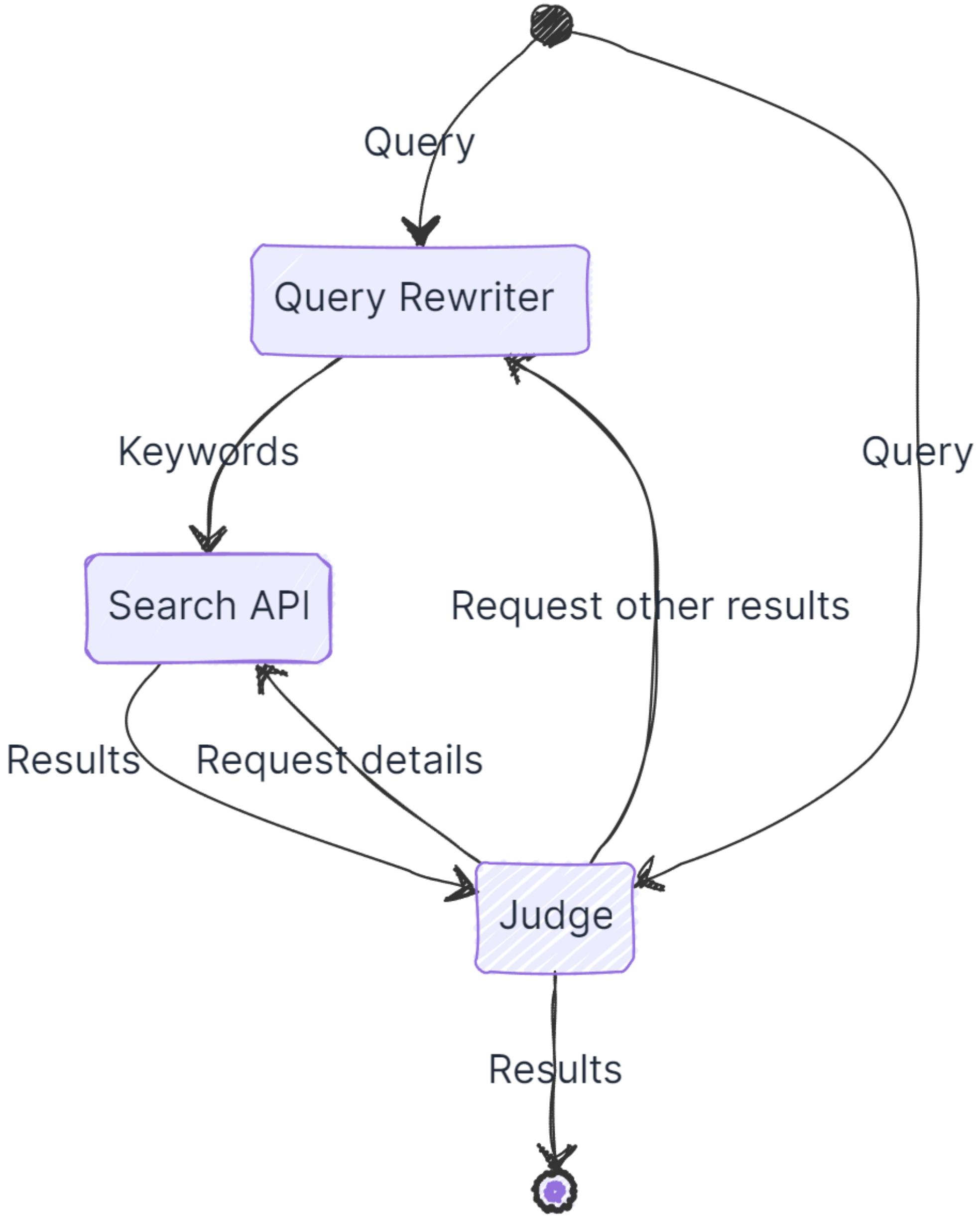
1. **Value Added Tax (VAT)**: Under the German VAT law (Umsatzsteuergesetz - UStG), the provision of services to clients outside of the EU, including Japan, is generally exempt from VAT. Specifically, according to § 4 UStG, services that are rendered to clients located abroad (i.e., outside the EU) usually do not attract VAT, assuming the service can be categorized as a "sonstige Leistung" (other service) under the law (Source: UStG § 4). Therefore, you should not charge VAT on your invoice to the Japanese client.
2. **Income Tax**: Income earned from providing services to a foreign client may be subject to income tax in Germany. However, Germany has a double taxation agreement (DTA) with Japan which aims to prevent the same income from being taxed in both jurisdictions. According to the terms laid out in the DTA, the right to tax income from independent personal services, such as those provided by software developers, is typically reserved for the country where the service is performed (in this case, Germany), provided that you do not have a permanent establishment in Japan (Source: EStG § 50i).
3. **Tax Liability**: If you are not residing in Japan or do not have a fixed place of business there, you will likely only need to address income taxes according to German law. If you do have a substantial presence in Japan, different rules may apply and you may need to consult the specific provisions of the DTA regarding how to report and pay taxes on such income (Source: AO § 2).

In summary, you should:

- Not charge VAT on your invoice to the Japanese client;
- Report this income in Germany under your personal income tax;
- Review the DTA between Germany and Japan to ensure compliance and

Agentic Retrieval

- Rewrite query to fit search API
- Judge results
- If necessary (repeatedly):
  - Request details
  - Search with different keywords
- Return result to response generator



Naïve vs Agentic

Similar

- Augment response with some factual context
  - Easy to update
  - Easy to reason about
  - Can still provide references
- LLM handles semantic gap

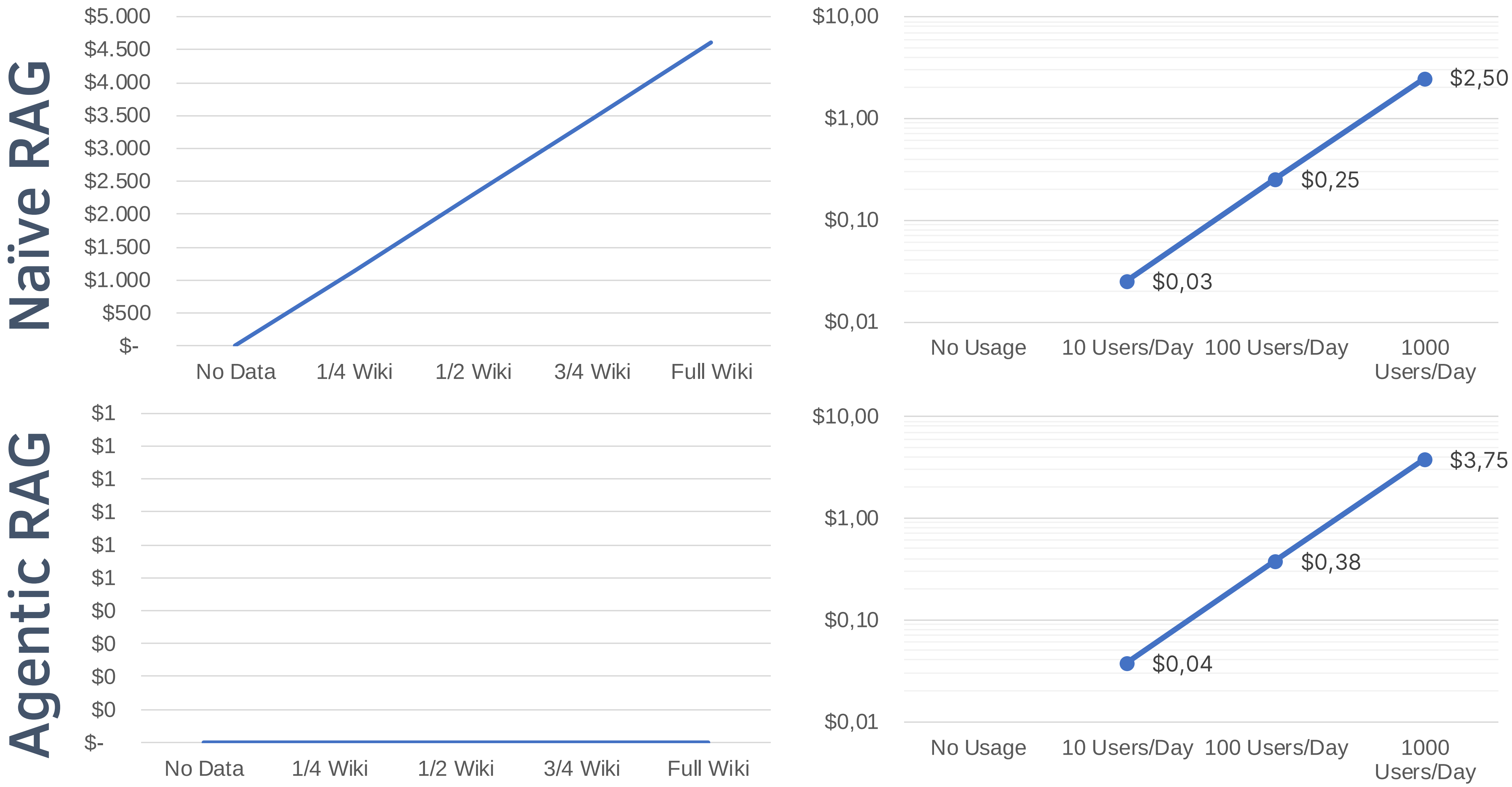
Differences

- No need for intermediary DB
- Semantic Gap handled with more context about the query

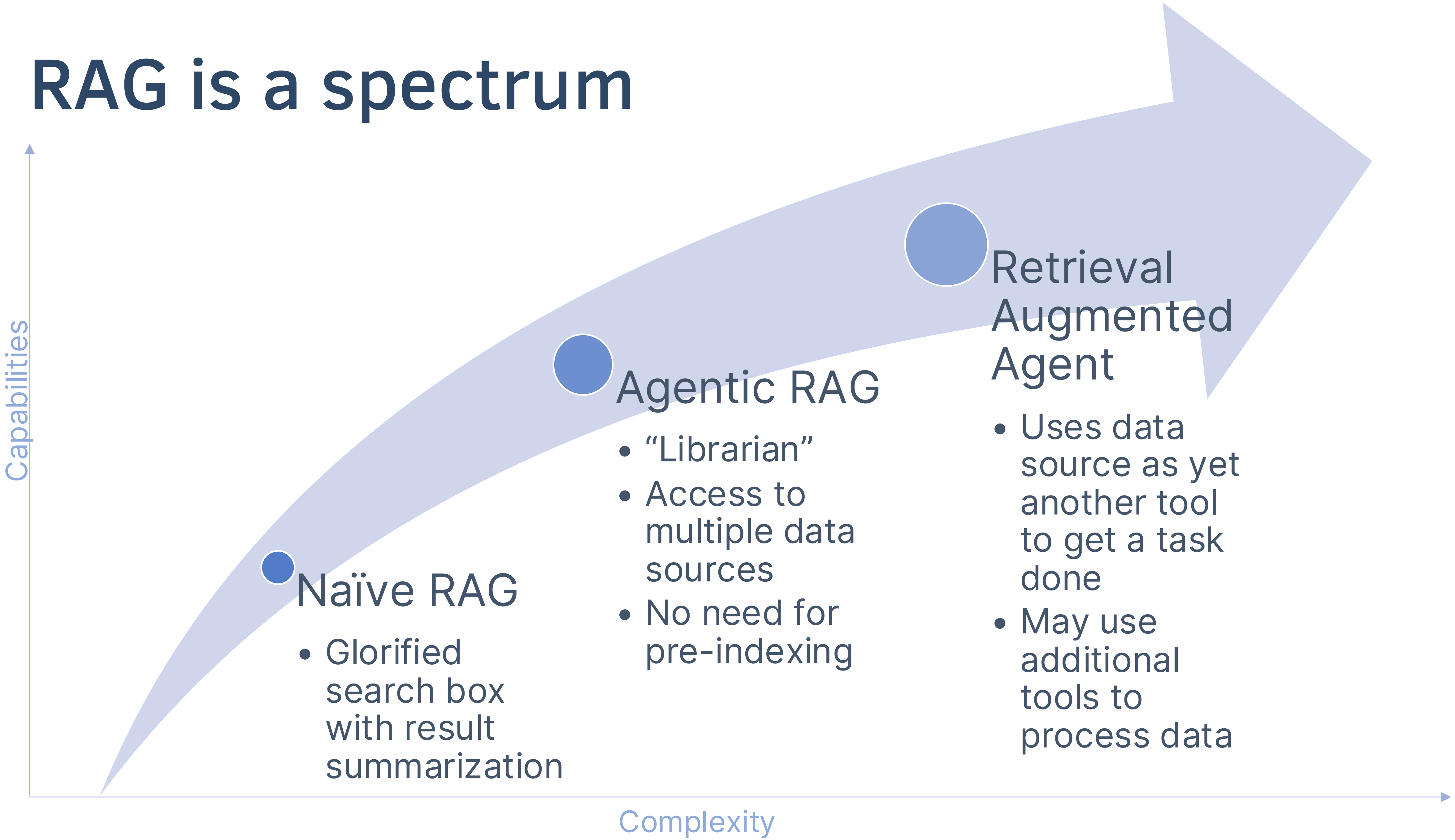
	Benefits	Drawbacks
Naïve RAG	+ Very Simple + Plenty of tutorials	– Needs vector database – Requires additional preprocessing of data • Chunking • Vectorization – Loses simplicity as more functionality is added
Agentic RAG	+ Can be attached to any data source without preprocessing + Naturally handles multi-source cases + Can ask clarifying questions + Can judge search results and reformulate query	– Can be more complex initially – Requires a more capable framework to stay easily debug-able – May be slower in answering questions, as LLM is involved more often

Up-Front Cost

Running Cost



RAG is a spectrum



Capabilities

Complexity

Naïve RAG

- Glorified search box with result summarization

Agentic RAG

- “Librarian”
- Access to multiple data sources
- No need for pre-indexing

Retrieval Augmented Agent

- Uses data source as yet another tool to get a task done
- May use additional tools to process data

Summary

- Agentic RAG is RAG using an agent for information retrieval
  - Can remove some complexity (vector DB, chunking, model choice…)
  - Adds other different complexity (engineering to implement an agent)
- Stepping stone towards fully agentic assistant

Questions?

Slides & Links to more materials:
https://www.dubs.tech/mlcon-berlin-2024

SCAN ME

We ask for **your feedback!**

PLEASE VOTE NOW!

devmio