



Using NLP Methods to Improve SQL Data Accessibility

SDMX + AI Workshop

27.03.2024

Who We Are – A Brief Introduction



1989 established

160+ employees

~18 Mio. € revenue in 2023

HQ in Heidelberg

Offices in Germany

Focus on regulated industries

Worldwide Network and Clients

Speakers

Julian Kurz



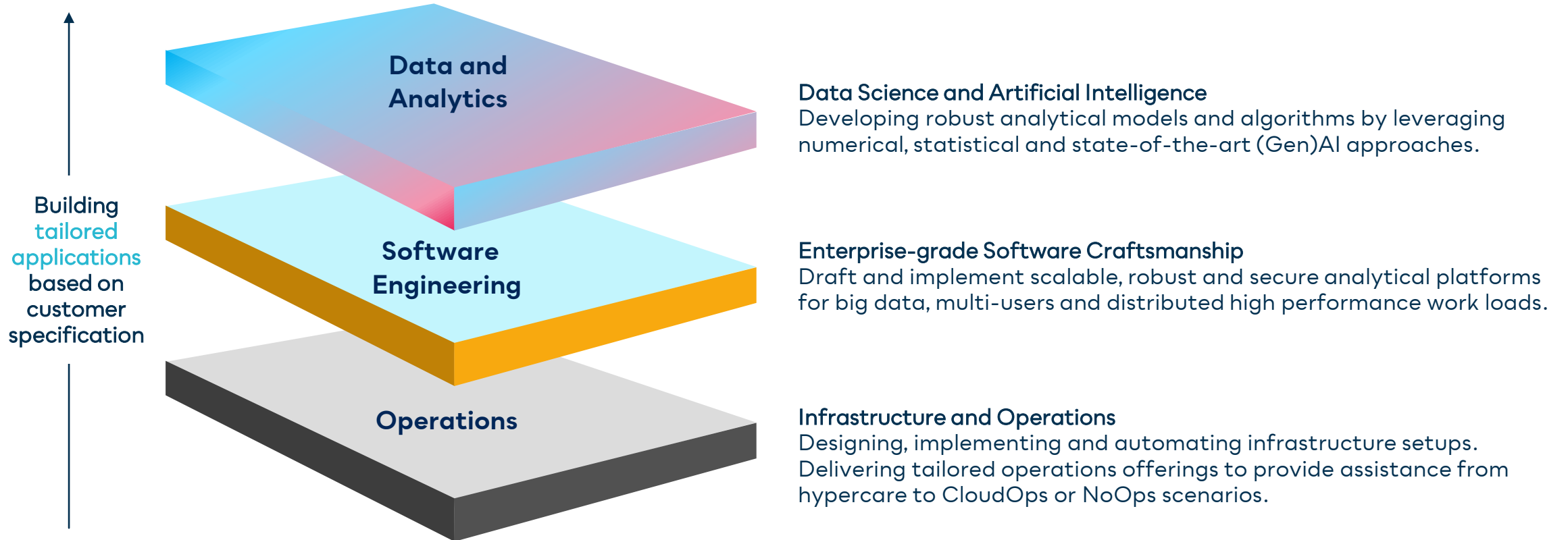
Senior ML Engineer,
Lead Expert
Generative AI

Christoph Bergen



Senior Data Scientist,
Team Lead
Data Science

Turning Data into Value – Our End-to-End Offering



Things You Can Do with Data and Generative AI



Find Data

- Finding the right dataset(s) to answer your question
- Join datasets
- Examples:
 - Utilizing RAG to search for relevant data and documents across a large database.
 - Employing LLM capabilities to transform non-technical search queries into technical queries.



Explore Data

- Understand the content of the data
- Explore patterns, anomalies



Analyze Data

- Apply calculations, statistics or derive KPIs
- Answer hypothesis and questions
- Visualize and communicate insights

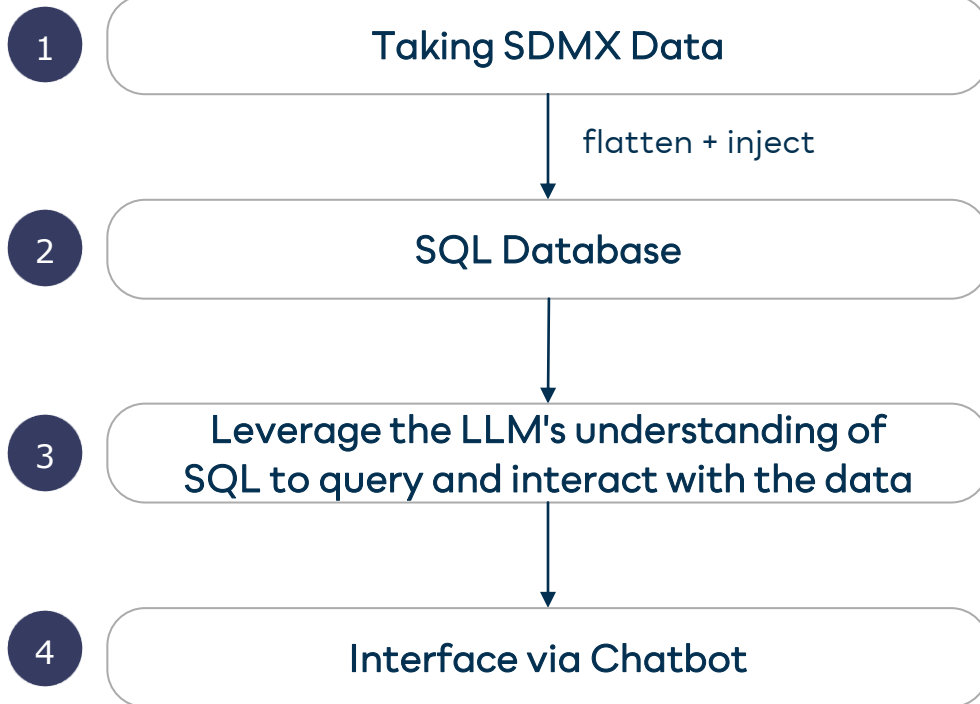
Focus of
the talk



Via natural language...
...using large language models.

Setting the Scene – Our Approach

Our Use Case



Target Audience

Our Generative AI system enables natural language-based data querying, interaction, and transformation.

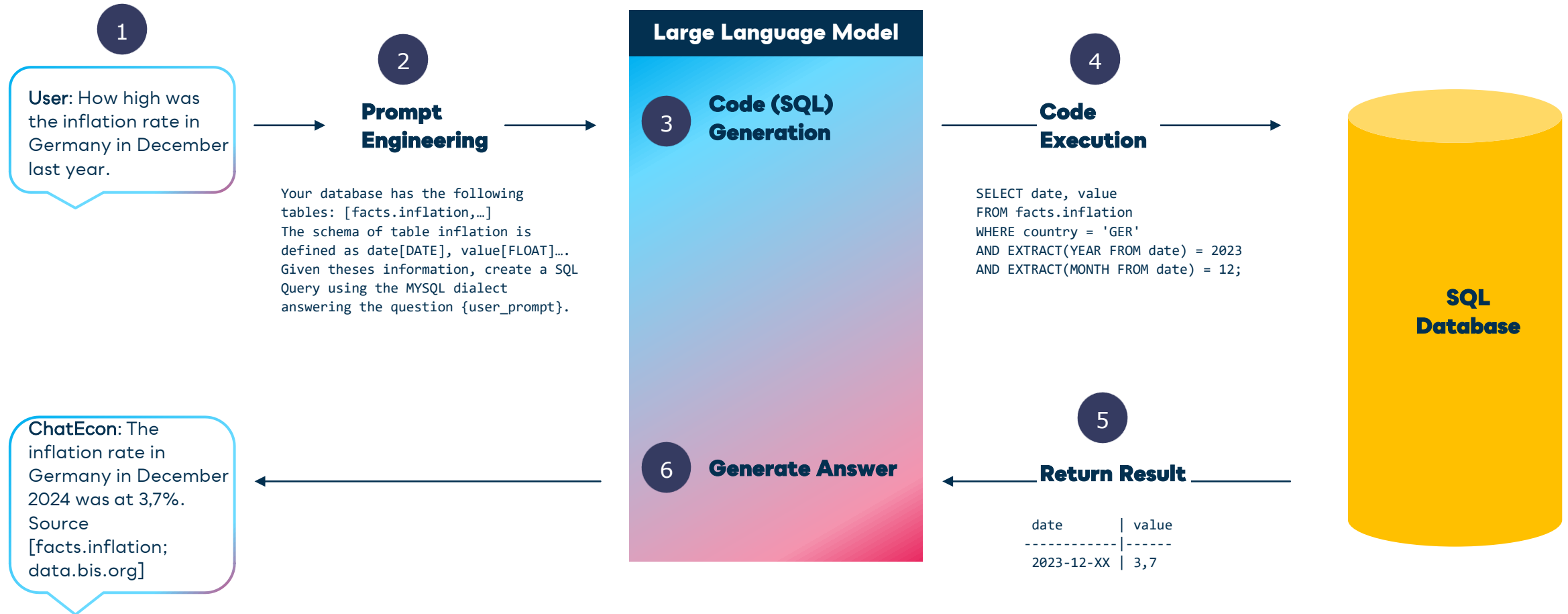


- No specialized training or specific tools required.
- Lowers barriers to data access and utilization.
- Complements, but does not replace, traditional data processing and visualization tools (yet).

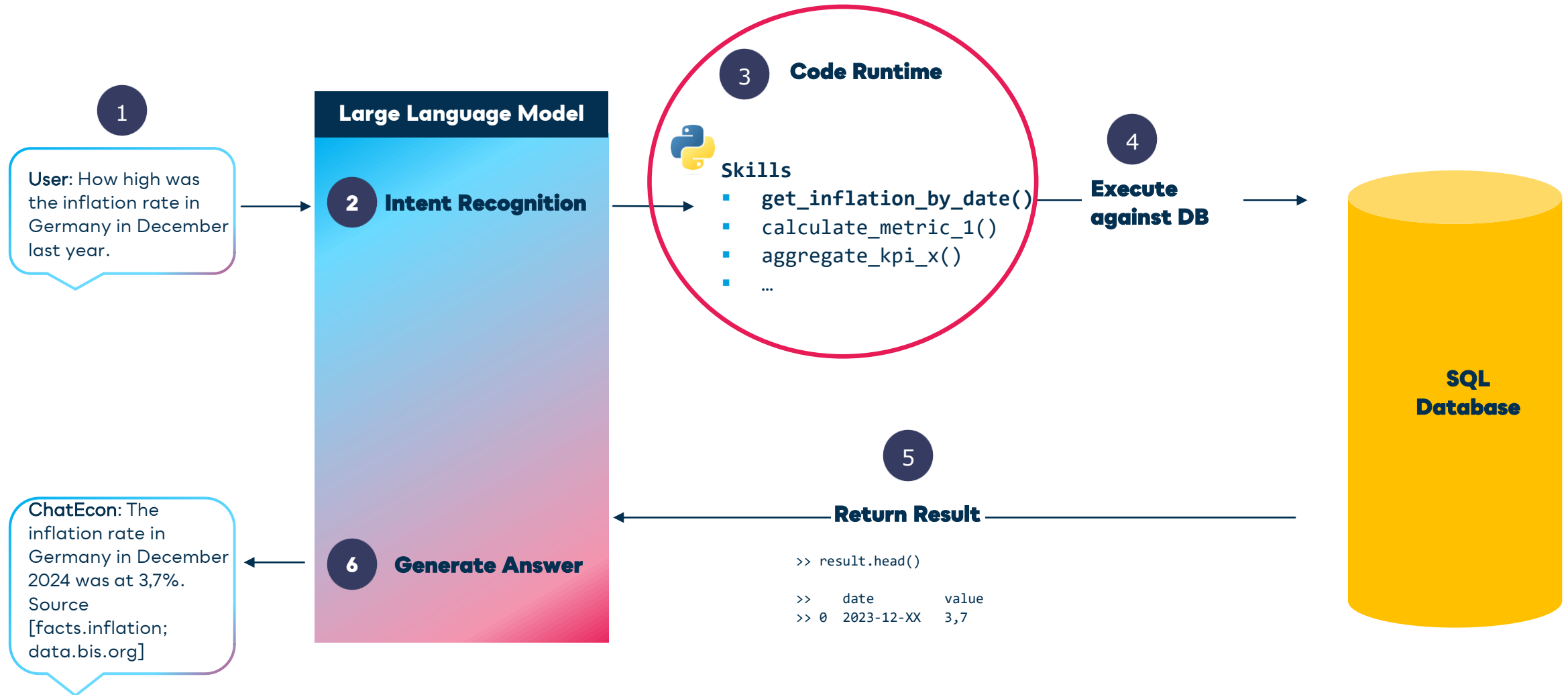


- Example users:
 - Journalists for source verification and dataset comparison.
 - General public for independent data interpretation and visualization.
 - And potentially many more...

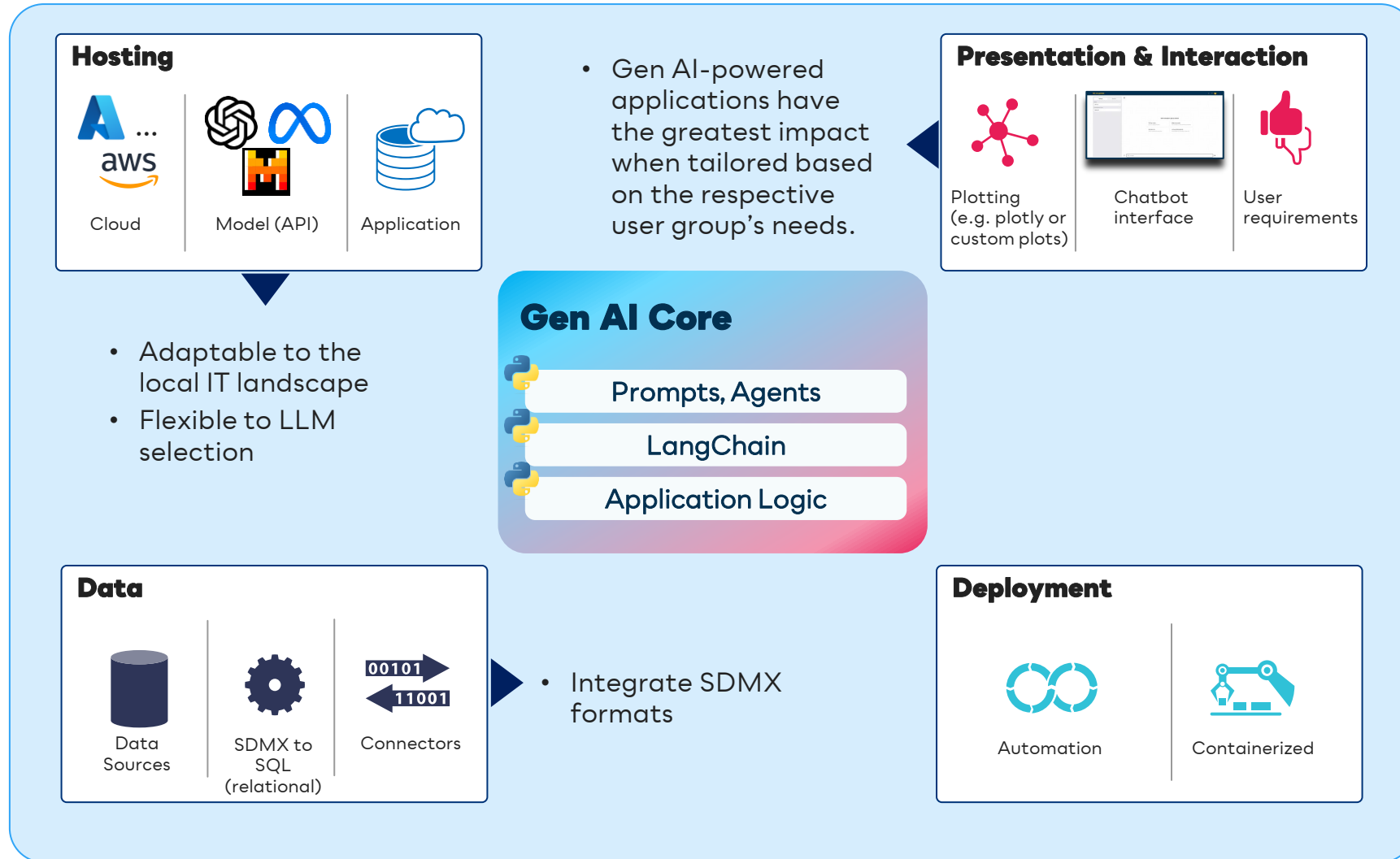
Two Ways for LLMs to Interact with Tabular Data 1/2



Two Ways for LLMs to Interact with Tabular Data 2/2



Our Solution is based on a Modular Software Architecture around a Gen AI Core



- Typical application modules can be realized based on a standardized **Generative AI Core Module**
- The modular approach provides the **flexibility to integrate** into heterogeneous IT landscapes...
- ... and also allows for the specific solutions to be **tailored to user needs**.

Currently, we are witnessing Rapid Advancements in Gen AI Models, with Expectations for Continued Progress.

The performance leap from GPT-3.5 to GPT-4 has enabled us to achieve satisfactory quality in the underlying use case.

Inter-model updates (e.g. new training data cutoffs)

Arena Elo

Full Leaderboard

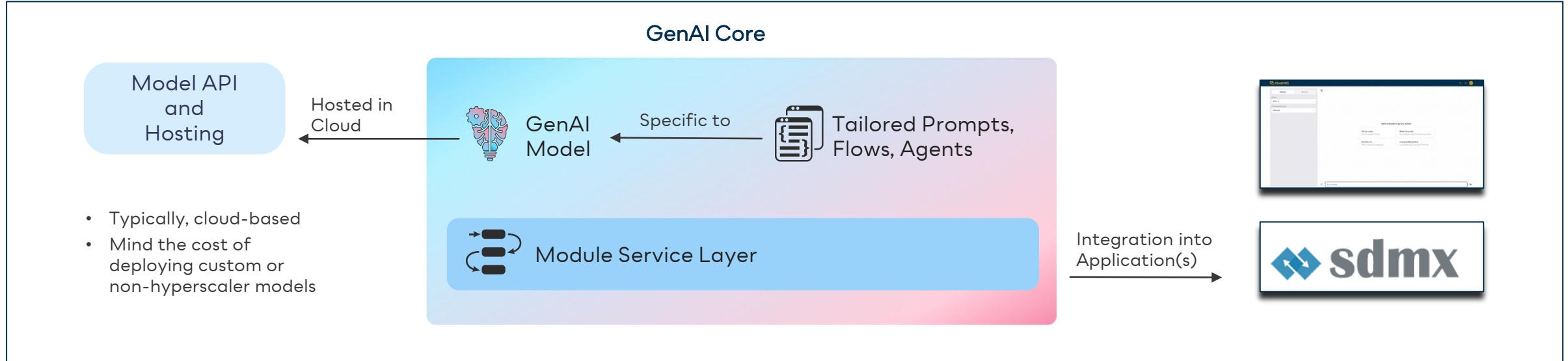
Total #models: 73. Total #votes: 408144. Last updated: March 13, 2024.

Contribute your vote 🗳 at chat.lmsys.org! Find more analysis in the [notebook](#).

Rank	Model	Arena Elo	95% CI	Votes	Organization	License
1	GPT-4-1106-preview	1251	+5/-4	48226	OpenAI	Proprietary
1	GPT-4-0125-preview	1249	+5/-6	22282	OpenAI	Proprietary
1	Claude 3 Opus	1247	+6/-6	14854	Anthropic	Proprietary
4	Bard (Gemini Pro)	1202	+6/-7	12623	Google	Proprietary
4	Claude 3 Sonnet	1190	+6/-6	14845	Anthropic	Proprietary
5	GPT-4-0314	1185	+4/-6	27245	OpenAI	Proprietary
7	GPT-4-0613	1159	+4/-5	43783	OpenAI	Proprietary
7	Mistral-Large-2402	1155	+5/-6	18959	Mistral	Proprietary
8	Qwen1.5-72B-Chat	1146	+4/-5	16729	Alibaba	Qianwen LICENSE
8	Claude-1	1145	+5/-6	21929	Anthropic	Proprietary
8	Mistral Medium	1145	+5/-4	23931	Mistral	Proprietary
12	Claude-2.0	1126	+7/-5	13679	Anthropic	Proprietary
12	Mistral-Next	1123	+5/-5	11875	Mistral	Proprietary
12	Gemini Pro (Dev API)	1118	+5/-6	12146	Google	Proprietary
13	Claude-2.1	1115	+4/-6	32431	Anthropic	Proprietary
13	Mixtral-8x7b-Instruct-v0.1	1114	+5/-5	27892	Mistral	Apache 2.0
13	GPT-3.5-Turbo-0613	1113	+5/-4	41354	OpenAI	Proprietary

Source: <https://chat.lmsys.org/> - LMSYS Chatbot Arena Leaderboard

Rapid Developments in the GenAI Field require Flexible Implementations



- The lifespan of a foundation model (e.g. GPT4, Gemini, Claude 3) being considered state-of-the-art is a **maximum of 2 years**
- Better model capabilities, as well as factors such as API costs, almost always justify switching to newer models
- Tailored components (i.e. capabilities currently enabled by prompting-techniques) might be in the “base” capabilities of future foundation model iterations

➤ Hence, we advice to **integrate GenAI components in a microservice fashion**, which enables easier upgrading to newer foundation model in the future.

Open Source Framework Approach

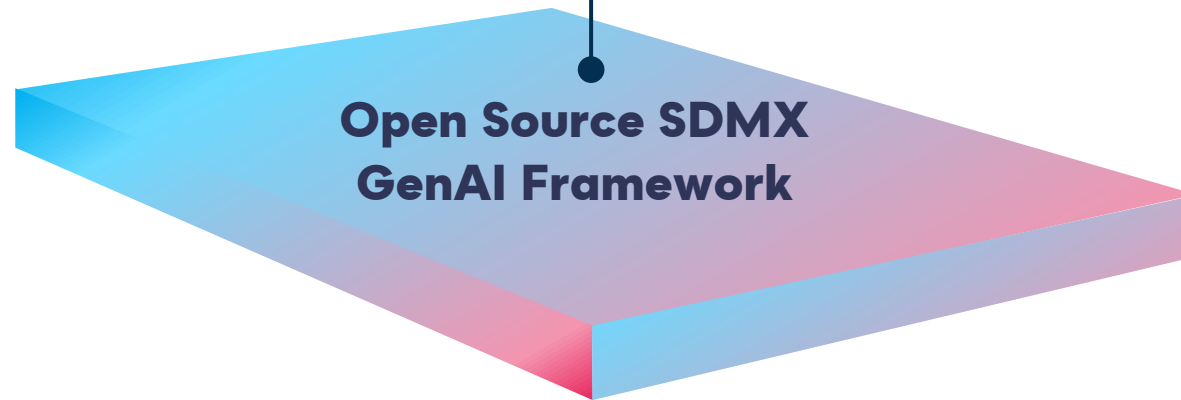


develops

- Open source implementation (Python, Javascript)
- Implements common use cases (e.g. NLP-based SQL querying)
- Provides core components, e.g. data connectors or SDMX domain specific prompts

- Open for both commercial and open source GenAI models
- Cloud agnostic

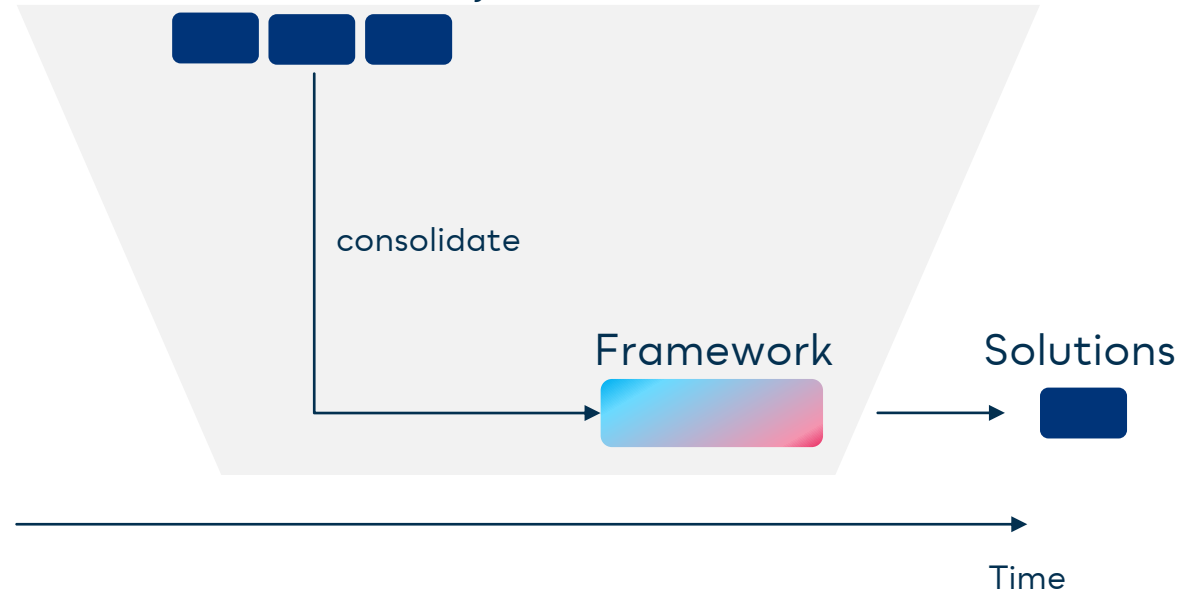
available to



How to Build such a Framework – Possible Roadmaps

Option 1

Start with POCs / Projects

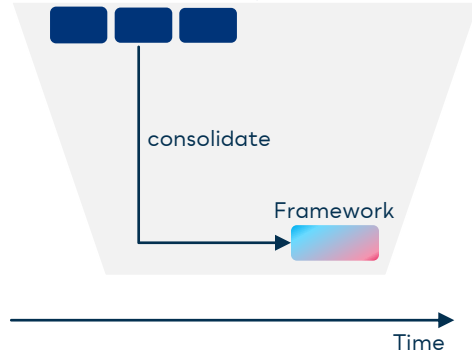


- + Gain experience with common components upfront
- Risk of neglecting backward integration to initial POCs
- Development time until the framework consolidation can begin

How to Build such a Framework – Possible Roadmaps

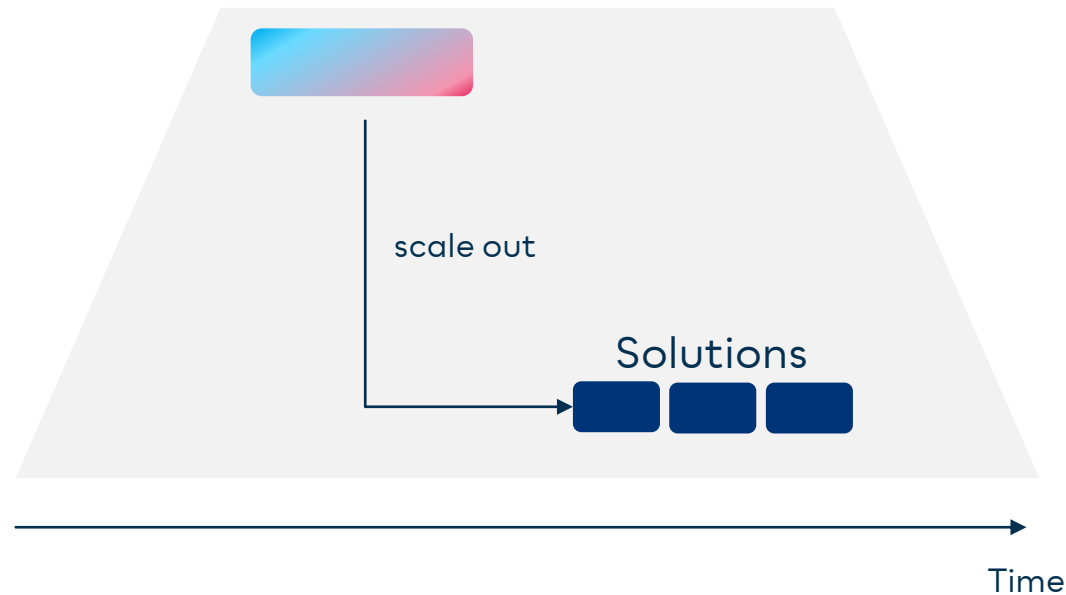
1

Start with POCs / Projects



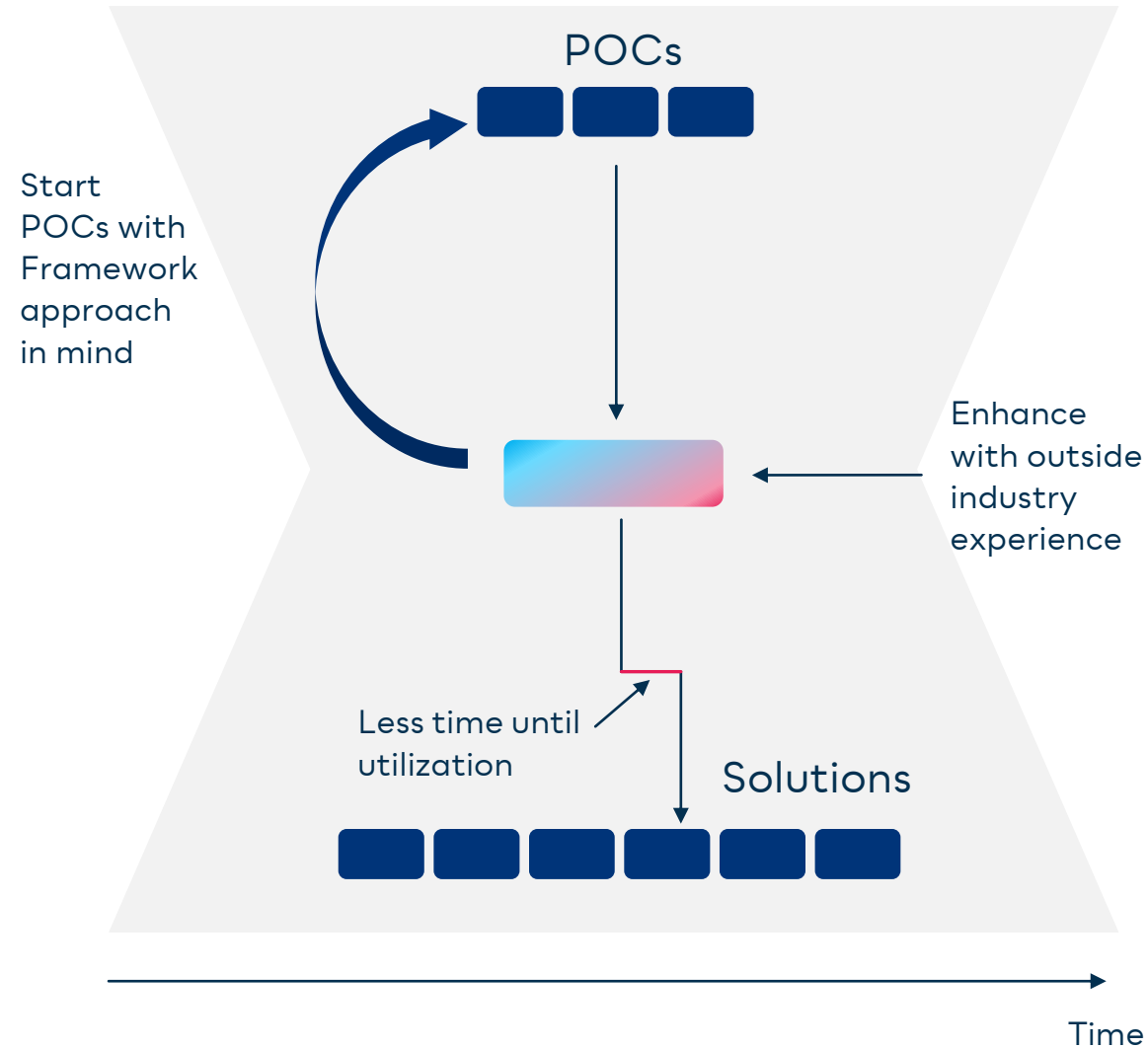
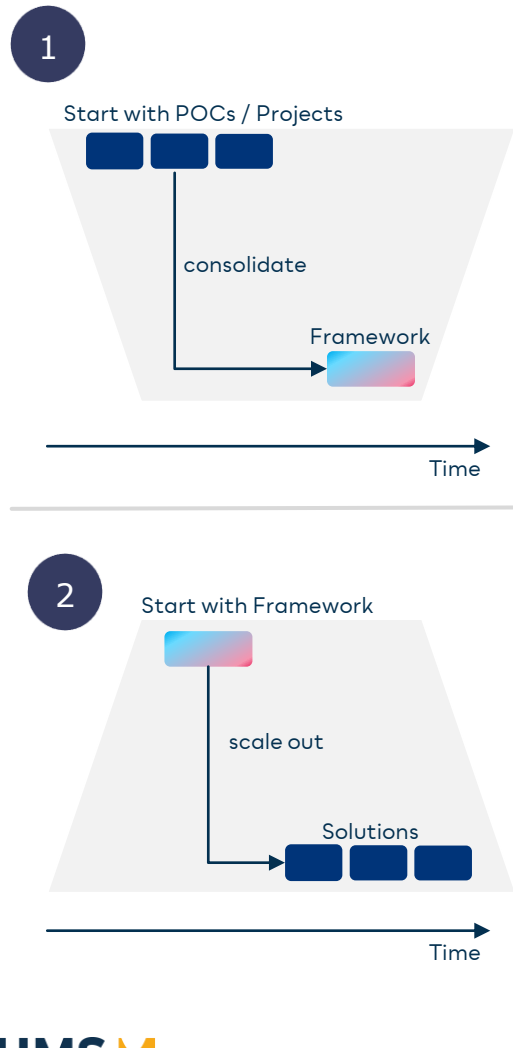
Option 2

Start with Framework

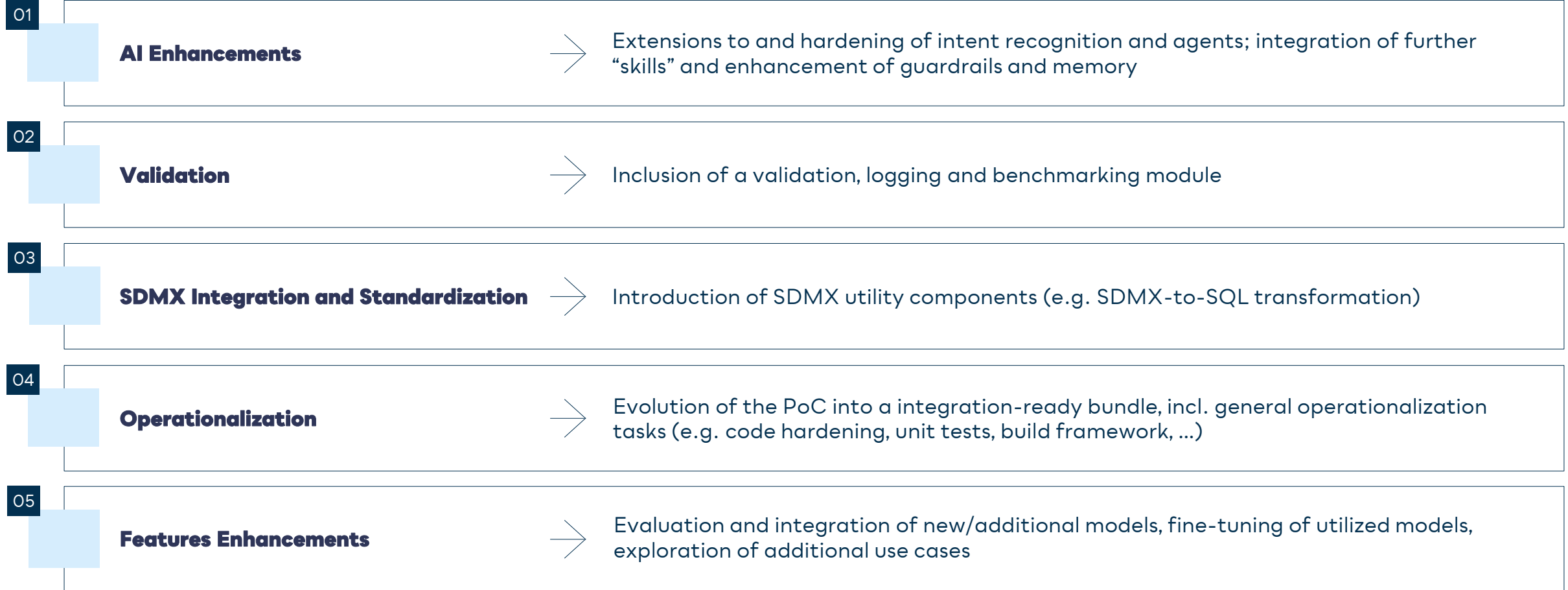


- + Each solution can benefit from the framework
- Risk of overengineering the framework
- Longer lead time until Solutions can be rolled out

Best of Both Worlds: Combining Both Approaches



Potential Next Steps



DEMO

NLP-BASED SQL QUERYING

Settings

Sources

Model

GPT-3.5



Speech-to-Text Model

Web Speech API



Template

Assistant



Fully customizable and flexible generative AI framework, allowing to experiment with, implement and integrate various use cases.

Some examples to get you started

Tell me a story
set in a magical forest

Share a piece of wisdom
about building strong relationships

Share a random thought
about the quirks of human behavior

Let's get philosophical
and debate the existence of extraterrestrial life



Your message





Using GPT-4 (Assistant)

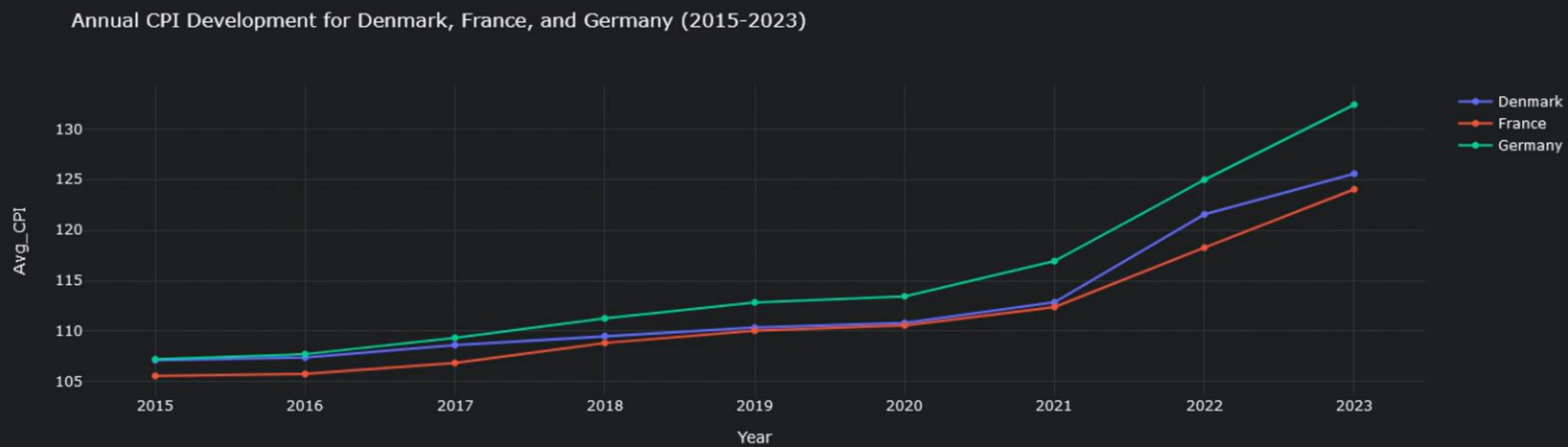
Please visualize the annual CPI development for Denmark, France and Germany from 2015-2023. ⋮

⋮ Here is the annual average CPI development for Denmark, France, and Germany from 2015 to 2023:

Country	Year	Avg_CPI
Denmark	2015	107.13
Denmark	2016	107.40
Denmark	2017	108.63
Denmark	2018	109.52
Denmark	2019	110.35
Denmark	2020	110.81
Denmark	2021	112.86

Interact with visualization frameworks (such as plotly) using natural language.

Now, let's visualize the data:



Please note that the consumer price index has the year 2010 as the index year with a value of 100.



Your message



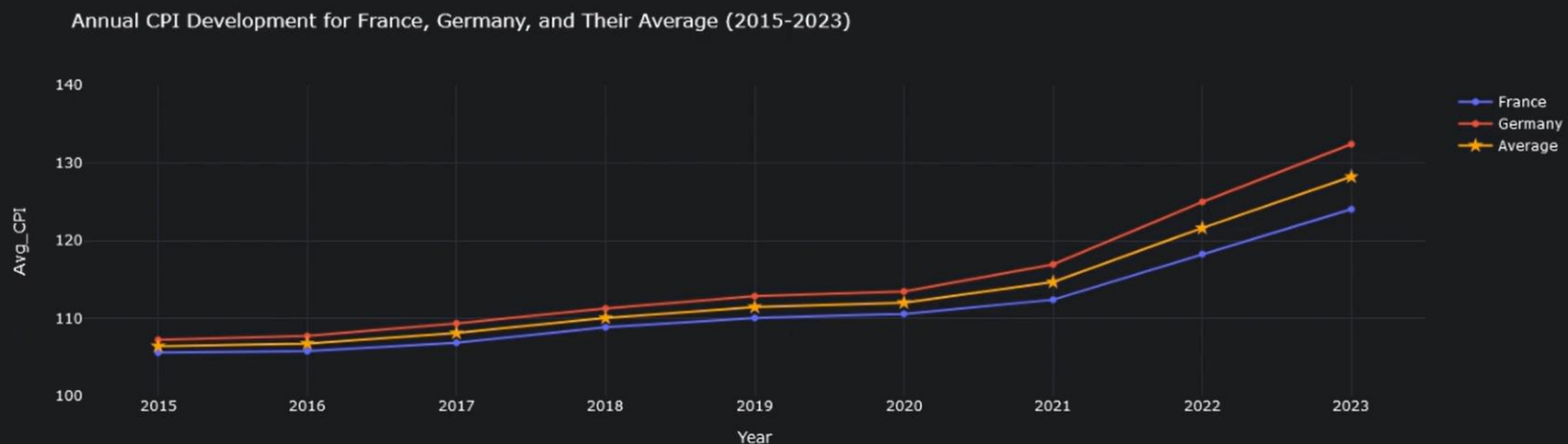


This graph shows the annual average CPI development for France and Germany from 2015 to 2023. Remember that the consumer price index (CPI) is based on the year 2010, with an index value of 100.

Calculate new/additional time series based on the underlying standardized data sets.

Now add another line showing the average between these countries. The line should be in bright orange color with big star-shaped markers. Also, please start the y-axis at 100. ⋮

⋮ Here is the updated graph with the average CPI line between France and Germany in bright orange color with big star-shaped markers, and the y-axis starting at 100:



This graph shows the annual average CPI development for France, Germany, and their average from 2015 to 2023. The average line is in bright orange color with big star-shaped markers. The y-axis starts at 100, which is the index value for the year 2010.



Your message





Using GPT-4 (Assistant)

Hello - What kind of data do you know? ⋮

⋮ I have access to the following data:

consumer_price_index: Contains the consumer price index with the year 2010 being the index year with a value of 100. The data is monthly (columns "year", "month", "cpi").

otc_exchange_turnover_in_2022: Contains data on over-the-counter (OTC) exchange turnover in 2022.

residential_property_prices: Contains data on residential property prices with the index year 2010. The data is in years and quarters (columns "year", "quarter", "rpp").

Please let me know if you have any questions or requests.

Detailed logging view (audit trail) for transparency and control purposes.

_Exception

_Exception

sql_db_query

sql_db_schema

sql_db_query

Input

```
SELECT
  (SELECT AVG(`rpp`) FROM residential_property_prices
   WHERE `year` = 2022 AND `country` = 'Germany') AS
  `rpp_2022`,
  (SELECT AVG(`rpp`) FROM residential_property_prices
   WHERE `year` = 2015 AND `country` = 'Germany') AS
  `rpp_2015`
```



Output

[[190.875, 117.125]]

Close

⋮ The inflation rate in Germany for 2015 was approximately 24.50% in January. Keep in mind that the consumer price index has the year 2010 as the index year with a value of 100.

What was the inflation rate of Germany in 2015? ⋮

...ing of the

Goods in Germany in 2023 compared to 2015? ⋮

⋮ Goods in Germany were approximately 24.50% more expensive at the end of 2015 (December). Keep in mind that the consumer price index has the year 2010 as the index year with a value of 100.

...the index at

Now that's quite a lot. And what about buildings? How much more expensive were residential properties at that time in Germany? ⋮

⋮ I am sorry, but I don't have data for residential property prices in Germany for 2023. The available data in the **residential_property_prices** table is up to 2022. Please let me know if you would like to know the change in residential property prices between 2015 and 2022.

Yes, please compare those prices then. ⋮

⋮ The average residential property price index in Germany for 2015 was **117.125**, and for 2022, it was **190.875**. This means that residential properties in Germany were approximately **62.91%** more expensive in 2022 compared to 2015. Keep in mind that the residential property price index has the year 2010 as the index year.

Your message



Thank You For Your Attention

Presenters:

Julian.Kurz@analytical-software.de

Christoph.Bergen@analytical-software.de

Managing contact:

Daniel.Goldfuss@analytical-software.de