



Generative AI at SAP

## Unit 3: Adapting Generative AI to Business Context

## Objectives of Unit 3

In Unit 3, you will learn about

- The limitations of generative AI models
- Methods to address limitations by grounding and adapting generative AI to business context
- Best practices when adapting generative AI



## Limitations of generative AI

Generative AI's ability to answer broad, general knowledge problems is remarkable!

**Forbes**

### GPT-4 Beats 90% Of Lawyers Trying To Pass The Bar

John Koetsier Senior Contributor   
*Journalist, analyst, author, and speaker.*

Follow

Mar 14, 2023, 06:40pm EDT

*The Atlantic*

### Is This the Singularity for Standardized Tests?

GPT-4's mastery of the SAT will re-entrench the power and influence of rote exams.

By Ian Bogost

**MEDPAGETODAY**<sup>®</sup>

### AI Passes U.S. Medical Licensing Exam

— Two papers show that large language models, including ChatGPT, can pass the USMLE

by [Michael DePeau-Wilson](#), Enterprise & Investigative Writer, MedPage Today January 19, 2023

Such successes may give the impression that generative AI can tackle any problem out of the box.

Users must be aware of current limitations, keep expectations in check, and understand how to make generative AI enterprise-ready to handle business problems.

## Adapting generative AI to business context




### Example limitations of generative AI

**Hallucination.** Large language models can generate plausible-sounding yet false answers.  
Related: LLMs can provide different “facts” depending on the language in which the prompt was given.

Model: GPT-4

JE Which review papers discuss challenges and applications of large language models?

As of my knowledge cutoff in September 2021, there were several influential papers discussing the challenges and applications of large language models. Here are some key resources that were available at that time:

1. "Language Models are Few-Shot Learners" by Tom B. Brown, Benjamin Mann, Nick Ryder, et al., 2020. This is not exactly a review paper, but it's a seminal work by OpenAI that discusses the potential and challenges of large-scale language models.  **Correct!**
2. "The Power and Limitations of Language Models" by Jacob Devlin, 2021. Although this is a blog post, it provides a good summary of the current understanding of the power and limitations of large language models, especially the Transformer-based models like BERT and GPT-3.  **Does not exist!**
3. "A Survey of Transformers" by Tianyu Liu, Xiang Li, Zhengyang Liu, et al., 2021. This is a survey paper that reviews the Transformer architecture, which is the basis for large language models such as GPT-3. It discusses various applications and challenges of these models.  **Wrong authors!**

**AP**

U.S. NEWS

**Lawyers submitted bogus case law created by ChatGPT. A judge fined them \$5,000**

“... six legal cases listed to support their March 1 written arguments did not exist.”

“... we made a good faith mistake in failing to believe that a piece of technology could be making up cases out of whole cloth.”

Source: arXiv 2307.10169. Challenges and Applications of Large Language Models.



# Adapting generative AI to business context

## Example limitations of generative AI

**Up-to-date and specific knowledge.** The knowledge of a generative AI model is frozen in time from when it was trained. In contrast to world general knowledge, business information changes quickly.



Prompt

Who won the NBA preseason game last night between the LA Lakers and Golden State Warriors?



I'm sorry for any inconvenience, but as an AI last trained on data up to September 2021, I don't have the ability to access real-time information or updates, including recent sports scores.



Output  
(model 1)



The Golden State Warriors defeated the Los Angeles Lakers 114-101 in their preseason game last night.



Output  
(model 2 –  
hallucination)



Prompt

What libraries and SDKs are available with SAP AI Core?



As of my training data until September 2021, SAP AI Core does not provide any specific libraries or SDKs. Instead, it integrates with various Python libraries and ML frameworks used for machine learning and data science, such as TensorFlow, PyTorch, Scikit-Learn, etc.



Output

11 Libraries and SDKs		
Explore additional SDKs and Libraries for use with SAP AI Core.		
SDKs Available with SAP AI Core		
Resource	Description	More Information
SAP AI Core SDK	The SAP AI Core SDK is a Python-based SDK that lets you access SAP AI Core using Python methods and code structures. It can be used to interact with SAP AI Core, and provides access to all public (library and administrative) APIs.	<a href="#">SAP AI Core SDK API</a> <a href="#">Business Accelerator Hub</a>
AI API Client SDK	The AI API Client SDK is a Python-based SDK that enables you to access the AI API using Python methods and data structures.	<a href="#">AI API Client SDK API</a> <a href="#">AI API</a>
Libraries Available with SAP AI Core		
Resource	Description	More Information
Metaflow Python Library for SAP AI Core	You can download the software development SDK for SAP AI Core from the Python Package Index (PyPI). This Python library is a plugin that lets you generate workflow templates that can be executed in SAP AI Core.	<a href="#">PyPI Metadata</a> <a href="#">Metaflow</a> <a href="#">How to get started</a>
Metaflow Python Library for SAP AI Core	Metaflow is an open-source workflow tool that lets you manage the productivity of data science. The Metaflow Python library for SAP AI Core extends Metaflow's capabilities so that machine learning pipelines can be run on SAP AI Core (which are also used in SAP AI Cloud).	

### Ground truth:

SAP AI Core, **launched October 12<sup>th</sup> 2021**, provides SAP AI Core SDK, AP API Client SDK, and Metaflow Python Library (SAP AI Core Product Documentation, 2023-09-01)

## Adapting generative AI to business context

### Example limitations of generative AI

**Inconsistent math abilities and limited notion of time.** Although improving, large language models are not calculators. Other kinds of foundation models could conceivably complete forecasts and math.



Prompt

Jane was elected class president in 1973 when she was 12 years old.

Her daughter, Jill, was elected class president in 2012 when she was 13 years old.

Who was older when she was elected class president, Jane or Jill?



*Correct: When elected, Jill was 13 and Jane was 12. Therefore, **Jill was older**.*

*\* Experiment conducted October 2023 using three different leading LLMs. Same result when the years were removed.*

*Note: Various prompt engineering techniques could guide the model to provide the correct answer.*



**Jane was older** when she was elected class president. In 1973, Jane was 12 years old, while in 2012, Jill was 13 years old.



Output  
(model 1)



Jane was elected class president in 1973 when she was 12 years old, and her daughter, Jill, was elected class president in 2012 when she was 13 years old. Therefore, **Jane was older** when she was elected class president.



Output  
(model 2)



**Jane was older** when she was elected class president, because she was 12 years old in 1973, and Jill was 13 years old in 2012.



Output  
(model 3)

## Adapting generative AI to business context

### Example limitations of generative AI

... even though the capabilities of generative AI to solve and explain some math problems are impressive! Techniques like calling functions and chain-of-thought prompting can increase performance on math problems.

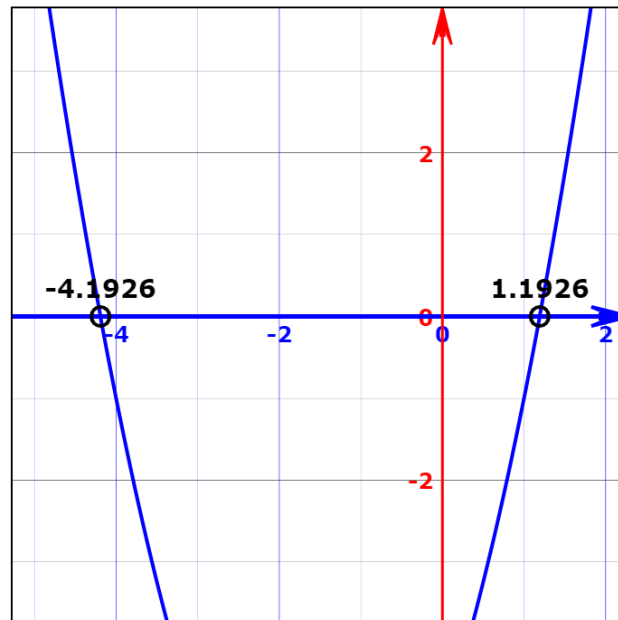


**Prompt**

Solve for X, show your work, and list the answer:

$$X^2 + 3x - 5 = 0$$

*Note:  
Checking the  
output, the LLM got  
the correct answer!*



This is a quadratic equation, and we can solve it by using the quadratic formula which is  $x = [-b \pm \sqrt{b^2 - 4ac}] / (2a)$ .

Here,  $a = 1$ ,  $b = 3$ , and  $c = -5$ .

Let's plug these values into the formula:

$$x = [-3 \pm \sqrt{(3)^2 - 4 \cdot 1 \cdot (-5)}] / (2 \cdot 1)$$

$$x = [-3 \pm \sqrt{9 + 20}] / 2$$

$$x = [-3 \pm \sqrt{29}] / 2$$

So, the solutions are:

$$x = (-3 + \sqrt{29}) / 2$$

$$x = (-3 - \sqrt{29}) / 2$$

These are the exact solutions. If you want to approximate to decimal places:

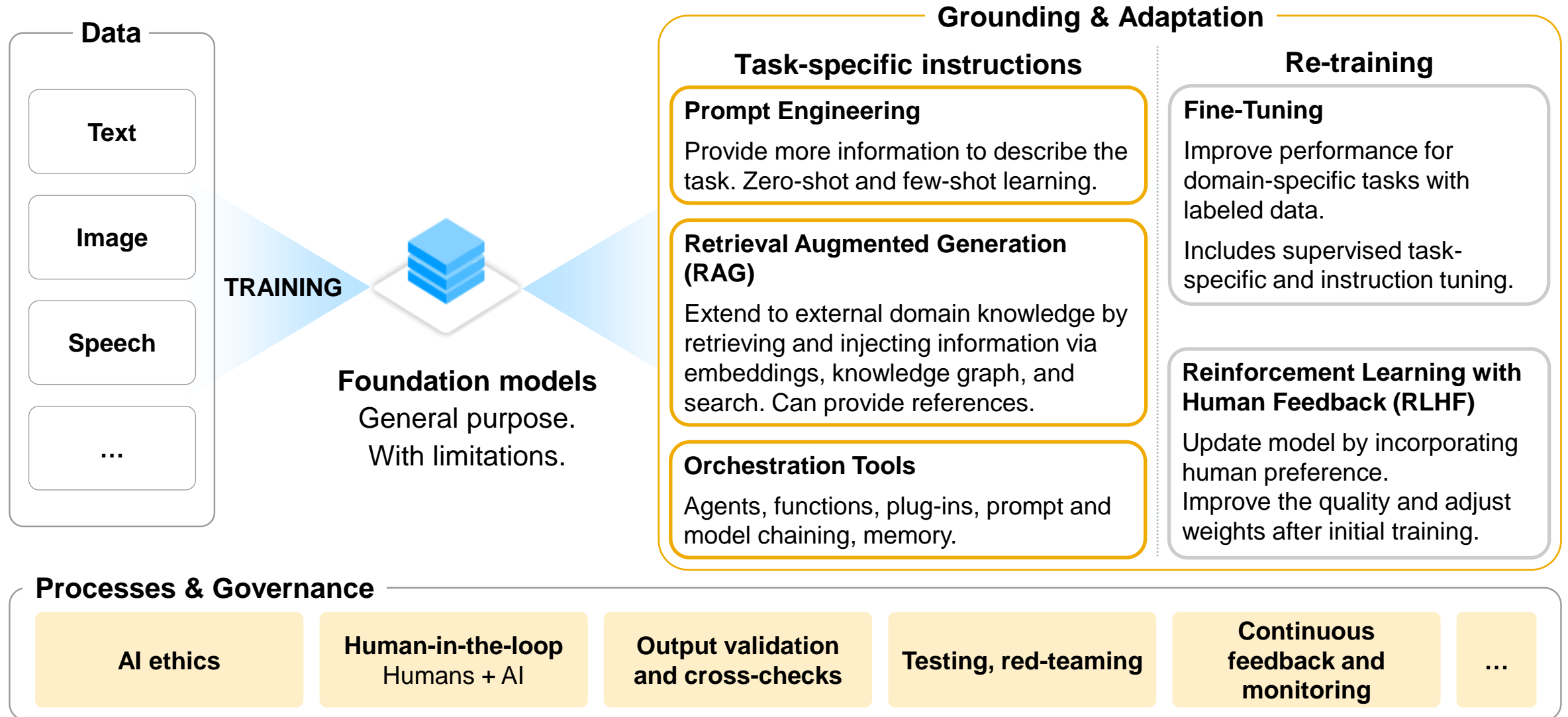
$$x \approx 1.1925824$$

$$x \approx -4.1925824$$



**Output  
(GPT-4)**

## Methods to ground and adapt generative AI to business context





## Prompt engineering example to create a job description

➤ Prompting: Give the model detailed information to reliably produce the desired output.

### Zero-shot learning

Prompt 1:

Write a job description for a support engineer.

### Instruction following

Prompt 2:

Write a job description for a support engineer. Keep it less than 300 words.  
  
Locate the job in Chicago.  
  
Include the following skills:  
1) ...

### In-context learning including examples

Prompt 3:

Create engaging job descriptions for company ACME. Location is Chicago and start date is January.

```
{ "job_title": "support engineer",  
  "required_skills": "..."} 
```

Ensure diversity and fairness.  
Use subtitles for required skills and for competencies.  
List skills and competencies as bulleted lists.

Use the same style based on examples of ACME job postings:  
...

User input

System prompt with content

System prompt with instructions

System prompt with examples



Number of tokens in the **context window** sent to the LLM increases (and along with it cost and response time).

More information and consistent input result in more precise output.

# Adapting generative AI to business context

## Prompt engineering example with SAP business documents

➤ Extract information from business documents with large language models

### Original input image



### Text with preserved 2D structure

Lieferschein		Donnan Schwelb Technik GmbH	
During Schwelbtechnik GmbH Postfach 1188 86329 Königsbrunn		Guldenstr. 1 1 75671 Eggenstein Deutschland Telefon: Fax.n.r. E-Mail:	
888 AG Abt. 4000 41002 Str. Kirzinger MT-144 Albert-Einstein-Str. 5 76938 Eggenstein Deutschland		Seite 1	
Ihre Kunden-Nr.:	38951	Lieferdatum:	7. März 2023
Lieferanten-Nr.:	778578-38	Lieferungs-Nr.:	78588
Auftragsnummer:	5838865	Ihre Ust-ID:	
Auftragsdatum:	10.02.2023		
Externe Belegnr.:	05BPLGK		
Pos. Nr.	FD838385	Beschreibung	Menge Einheit
2		PASSSCHRAUBE 2508 1878 001 007 Zeichnungsnummer: Ref-Nr. Einzelgewicht (netto) in kg: Gesamtgewicht (netto) in kg:	0005 SAP-Nr. 7303500 0,81 0,84
3	FD838600	GEWINDEBOLENZEN 2508 841 011007 Zeichnungsnummer: Ref-Nr.	1 Stück 0003 SAP-Nr. 7303506
Lieferbedingung: FREI HAUS Frei Haus			
Rech. an Adresse Rech. an Deb.-Nr. 380078 888 AG Kreditorenbuchhaltung, Rechnungsprüfung Heidenmaierstraße 170 80785 München Deutschland			
TUV SUD			



### Structured result

```
{
  "DeliveryNoteNo": "70500",
  "DocumentDate": "2023-03-07",
  "PurchaseDoc": "G5BPLGK",
  "SupplierID": "770570-30",
  "Items": [
    {
      "ItemNo": "2",
      "ProductId": "7303500",
      "Quantity": "4",
      "UnitOfMeasure": "Stück"
    },
    {
      "ItemNo": "3",
      "ProductId": "7303506",
      "Quantity": "1",
      "UnitOfMeasure": "Stück"
    }
  ]
}
```



### Output validation

Supplier ID XXXXXX-XX incorrect.  
Review proposed supplier ID match  
from master data: XXXXXY-XX

+ Append instructions/examples to the prompt (“meta prompt”)

Extract the following fields as JSON:

- **DeliveryNoteNo**: ID for the delivery note
- **DocumentDate**: Date of the document
- **PurchaseDoc**: Reference to purchase doc
- **SupplierID**: Supplier's unique ID
- **Items**: List of products
- **ItemNo**: Number for the item in the list
- ...

## Retrieval augmented generation (RAG) and embeddings example

### ➤ First: what are embeddings?

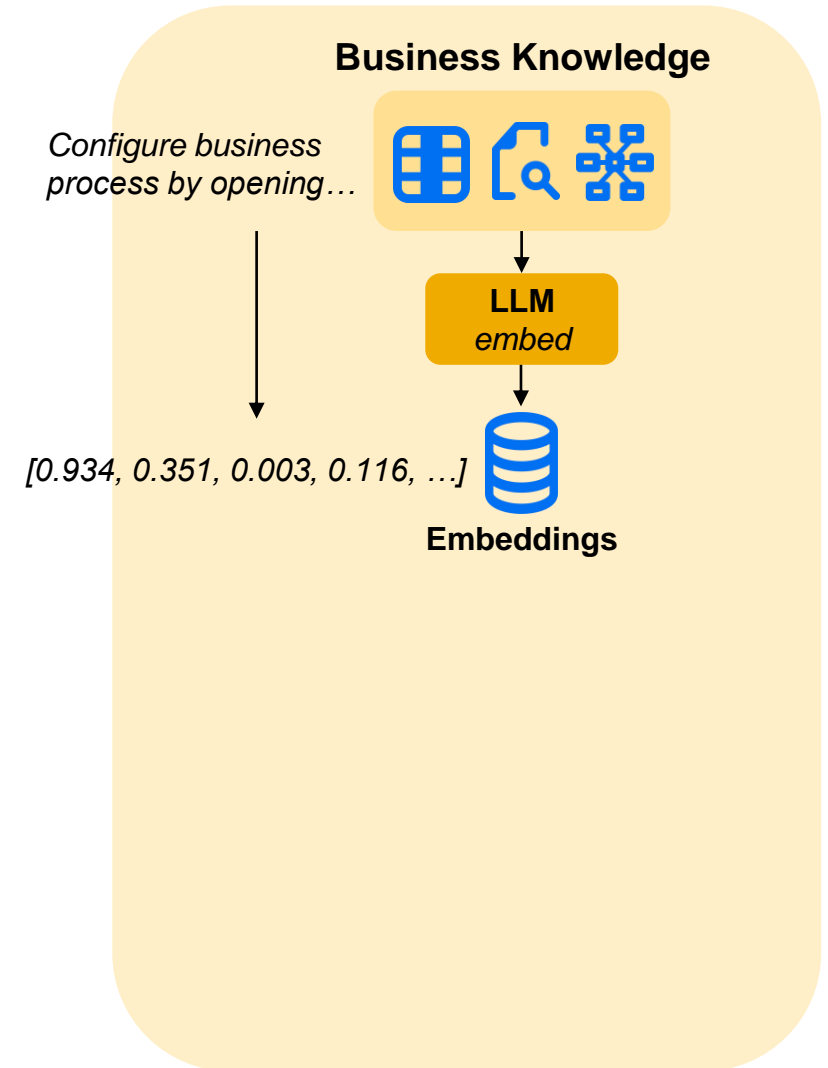
Embeddings: numerical representations of information that retain semantic, contextual meaning (text, images, etc.)

Example:


- Encode product documentation as embeddings using machine learning (Word2Vec, SBERT...). Store in vector database.
- “*Configure business process by opening...*” is represented as a vector [0.934, 0.351, 0.003, 0.116, ...]

The vector captures the semantic meaning in the text, e.g. “*Apple*” has a different vector in “*Apple* makes phones” vs. “I have an *Apple* iPhone 15” vs. “*Apple* is an ingredient in pies”.

Business data represented as embeddings can be easily searched and retrieved using techniques like *vector similarity scoring*.



## Retrieval augmented generation (RAG) and embeddings example

 You have a set of business documents that should be used to answer user questions

**Answering business-specific questions requires two steps:**

1. Embed your business knowledge to retrieve relevant items given a question, using LLMs and other foundation models.
2. Generate the answer from the best results using LLMs. Use prompt engineering as needed.

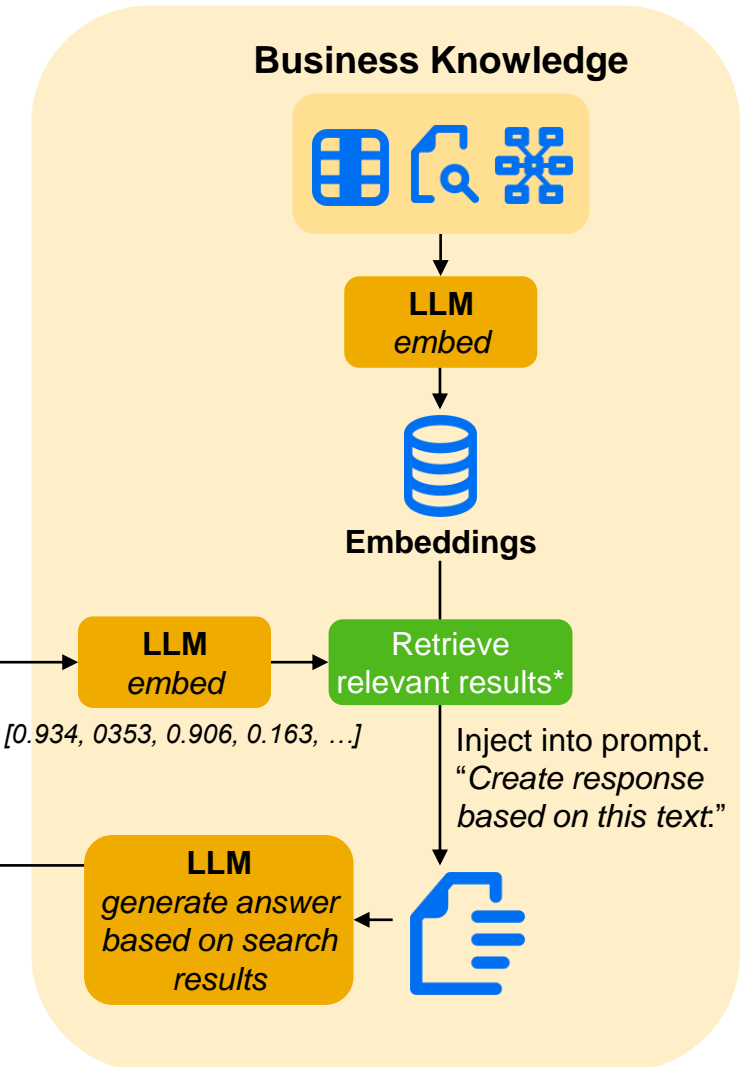


“How can I adapt my business process for VAT calculation?”

“You need to configure your business process and add a step by...”



Using techniques like embeddings 1) grounds the prompt with relevant information and 2) references source information to generate the response.

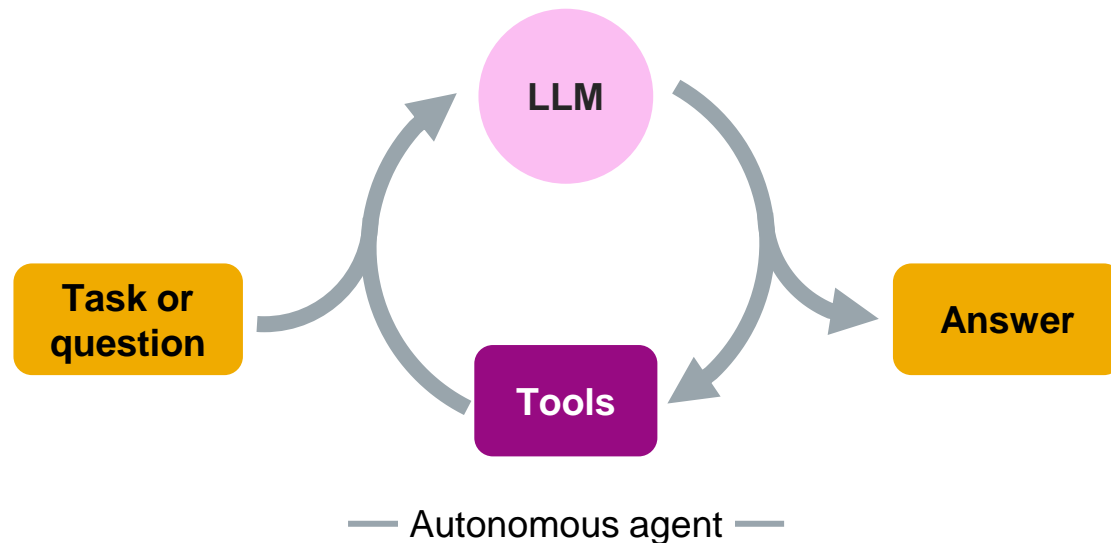


\* Possible to use classical keyword-based search as an alternative

## Adapting generative AI to business context

### Orchestration tools (“Agents”)

**Orchestration example (experimental):** Give an LLM access to tools. An agent retrieves API specifications from SAP systems, develops plan, and extracts data via API calls to formulate a response.



#### Example

**Task/Question:** User Inputs a Task

- Autonomous Agent
1. Retrieve  
SAP systems and their API specifications are retrieved based on the input (powered by **knowledge graphs**)
  2. Plan  
The **LLM** generates a plan based on the API retrieved specifications
  3. Execute  
The **LLM** follows the plan and generates API calls and reacts to the system responses

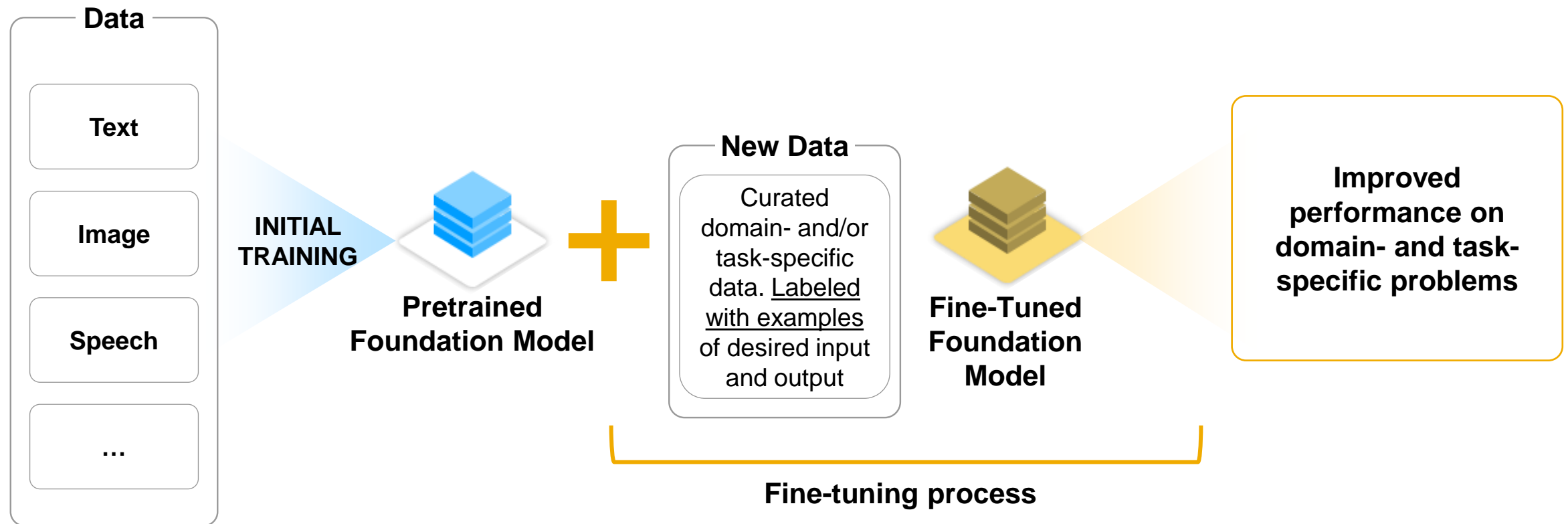
**Answer:** The final result is returned

- Use existing systems via API calls.
- LLM has no prior knowledge of the system. Knowledge is provided as part of the prompt.



## Fine-tuning

➤ Fine-tuning: adjusting an existing foundation model's parameters (i.e. weights) to perform better at a specific task by retraining it on a new data set.



## Fine-tuning



### Considerations for fine-tuning foundation models

#### Opportunities

- Strong potential to create **task-specific** models with significant boost in accuracy.
- Continued training on domain- or industry-specific data improves quality for downstream tasks in an area.
- Smaller to medium-sized (~5-70 billion parameter) LLMs can reach or exceed performance of more costly massive models.

#### Challenges

- May reduce but will not eliminate problems like hallucination.
- May not “forget” existing biases and information.
- Does not reliably learn new factual information.
- Requires sufficient (curated) training data.
- Can be computationally expensive and costly.
- Currently difficult to generalize the volume of training data needed to achieve meaningful results.



Consider and test task-specific instructions (prompt engineering) first before investing in fine-tuning models.

## Recommendations for adaptation of foundation models

### DO NOT

- ✗ Trust generic generative AI models to answer factual questions correctly, esp. in a business context.
- ✗ Start with fine-tuning.
- ✗ Use the biggest model by default.
- ✗ Productize a technically functional generative AI use case without established supporting processes.

### DO

- ✓ Be aware of limitations. Ground and adapt generative AI models on business data to ensure output is relevant and reliable.
- ✓ Start with task-specific instructions like prompt engineering and RAG.
- ✓ Test and adapt different generative AI models and optimize based on price and performance.
- ✓ Ensure proper governance and design for generative AI (e.g. human in the loop, output validation, monitoring...).



# Thank you.

**Contact information:**

**open@sap.com**

Follow all of SAP



[www.sap.com/contactsap](http://www.sap.com/contactsap)

© 2023 SAP SE or an SAP affiliate company. All rights reserved.

No part of this publication may be reproduced or transmitted in any form or for any purpose without the express permission of SAP SE or an SAP affiliate company.

The information contained herein may be changed without prior notice. Some software products marketed by SAP SE and its distributors contain proprietary software components of other software vendors. National product specifications may vary.

These materials are provided by SAP SE or an SAP affiliate company for informational purposes only, without representation or warranty of any kind, and SAP or its affiliated companies shall not be liable for errors or omissions with respect to the materials. The only warranties for SAP or SAP affiliate company products and services are those that are set forth in the express warranty statements accompanying such products and services, if any. Nothing herein should be construed as constituting an additional warranty.

In particular, SAP SE or its affiliated companies have no obligation to pursue any course of business outlined in this document or any related presentation, or to develop or release any functionality mentioned therein. This document, or any related presentation, and SAP SE's or its affiliated companies' strategy and possible future developments, products, and/or platforms, directions, and functionality are all subject to change and may be changed by SAP SE or its affiliated companies at any time for any reason without notice. The information in this document is not a commitment, promise, or legal obligation to deliver any material, code, or functionality. All forward-looking statements are subject to various risks and uncertainties that could cause actual results to differ materially from expectations. Readers are cautioned not to place undue reliance on these forward-looking statements, and they should not be relied upon in making purchasing decisions.

SAP and other SAP products and services mentioned herein as well as their respective logos are trademarks or registered trademarks of SAP SE (or an SAP affiliate company) in Germany and other countries. All other product and service names mentioned are the trademarks of their respective companies.

See [www.sap.com/trademark](http://www.sap.com/trademark) for additional trademark information and notices.