AI Summit Brainport 2024

GENERATIVE AI SYSTEMS

JAKUB M. TOMCZAK | Group Leader, Associate Professor
Generative AI Group & Dept. M&CS

TU/e
EINDHOVEN
UNIVERSITY OF
TECHNOLOGY

Jakub M. Tomczak, Ph.D.

Associate Professor at TU/e

Group Leader of the Generative AI group

Founder Amsterdam AI Solutions

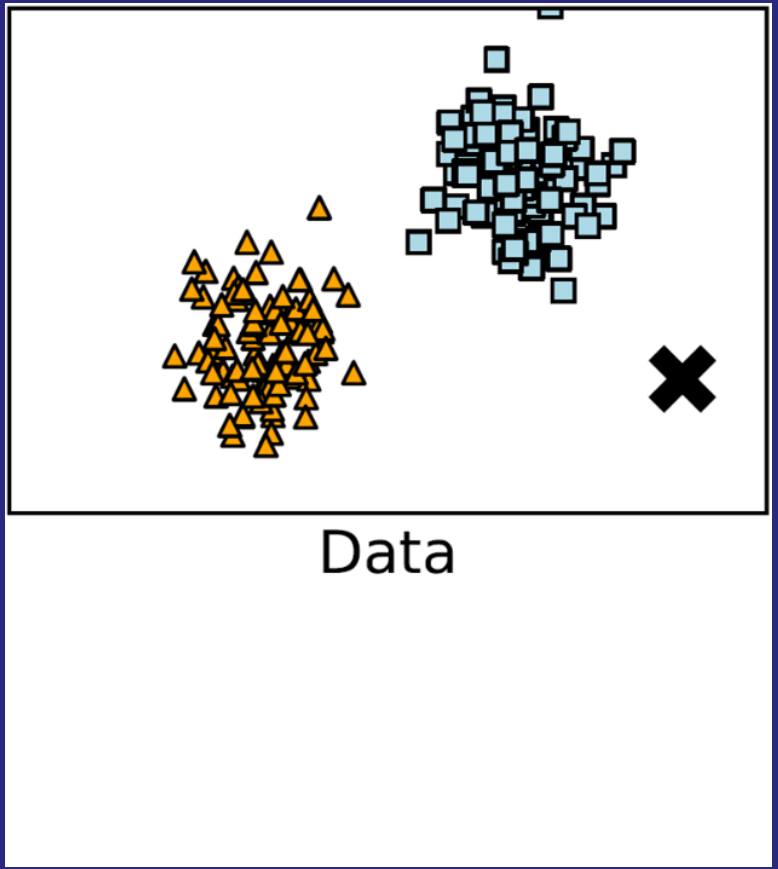
+15y experience in ML/AI and GenAI

The author of “Deep Generative Modeling”

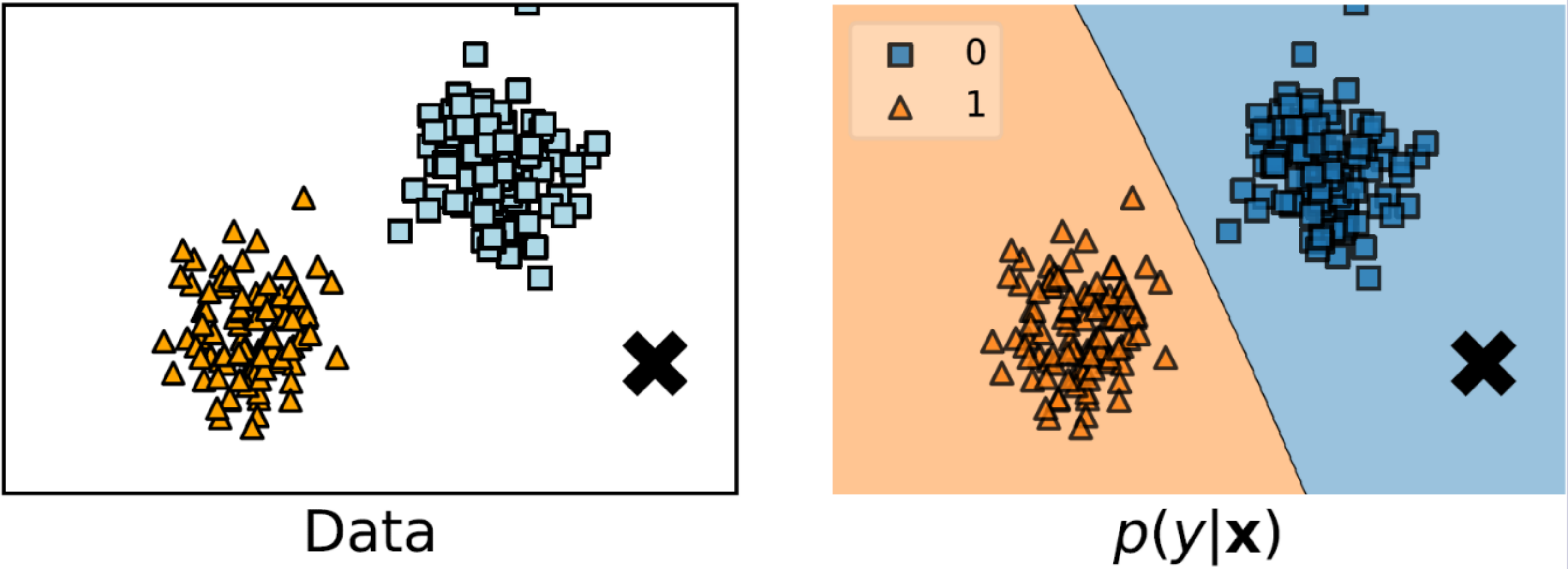
Generativ/e

Amsterdam
AI Solutions

Modeling: Discriminative vs. Generative



Modeling: Discriminative vs. Generative

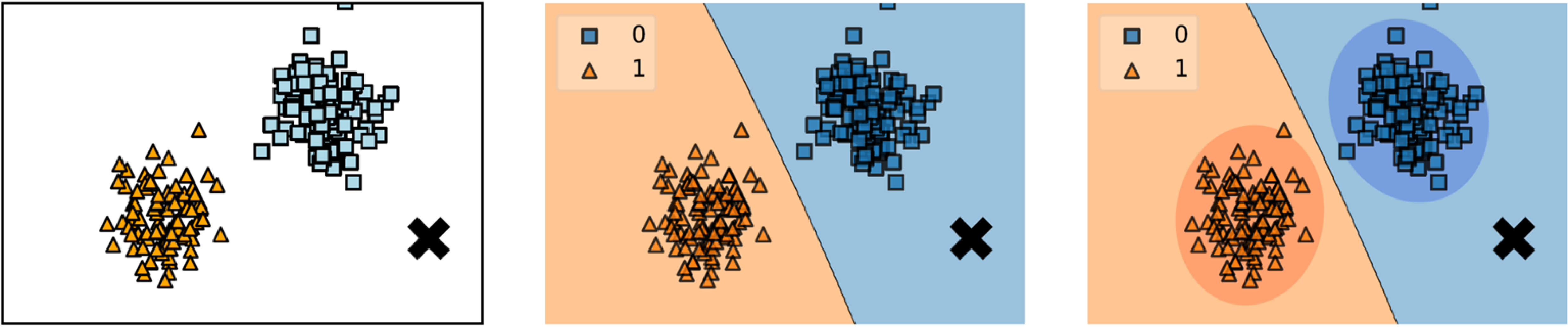


Data

$p(y|\mathbf{x})$

$p(blue|\mathbf{x})$ is high
= certain decision!

Modeling: Discriminative vs. Generative



Data

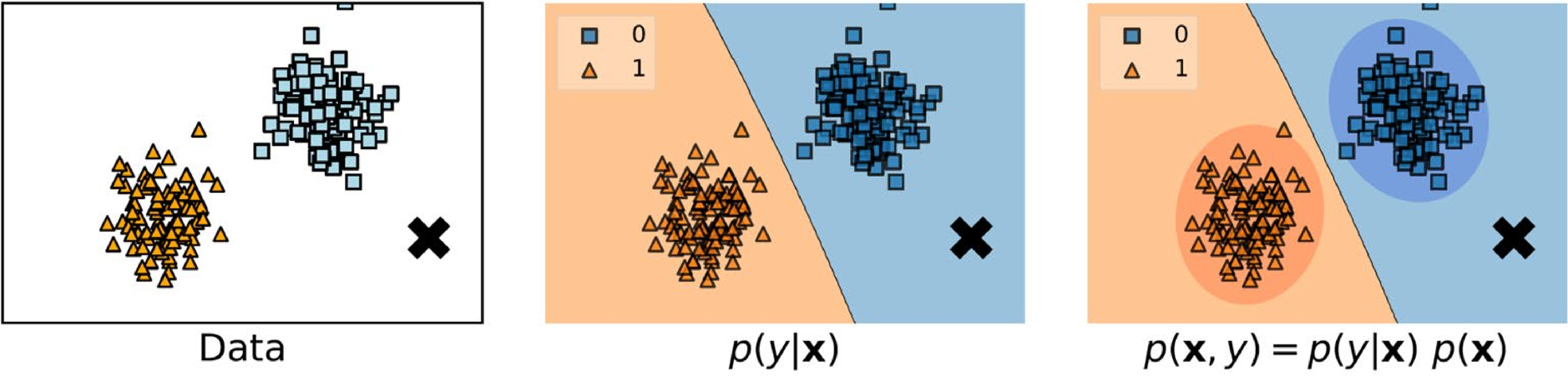
$p(y|\mathbf{x})$

$p(\mathbf{x}, y) = p(y|\mathbf{x})\ p(\mathbf{x})$

$p(blue|\mathbf{x})$ is high
= certain decision!

$p(blue|\mathbf{x})$ is high
and $p(\mathbf{x})$ is low
= uncertain decision!

Modeling: Discriminative vs. Generative



Knowing the generative process allows us to <u>understand</u> phenomena
and
<u>synthesize</u> (a.k.a. <u>generate</u>) new data.

Generative AI

Probabilistic modeling

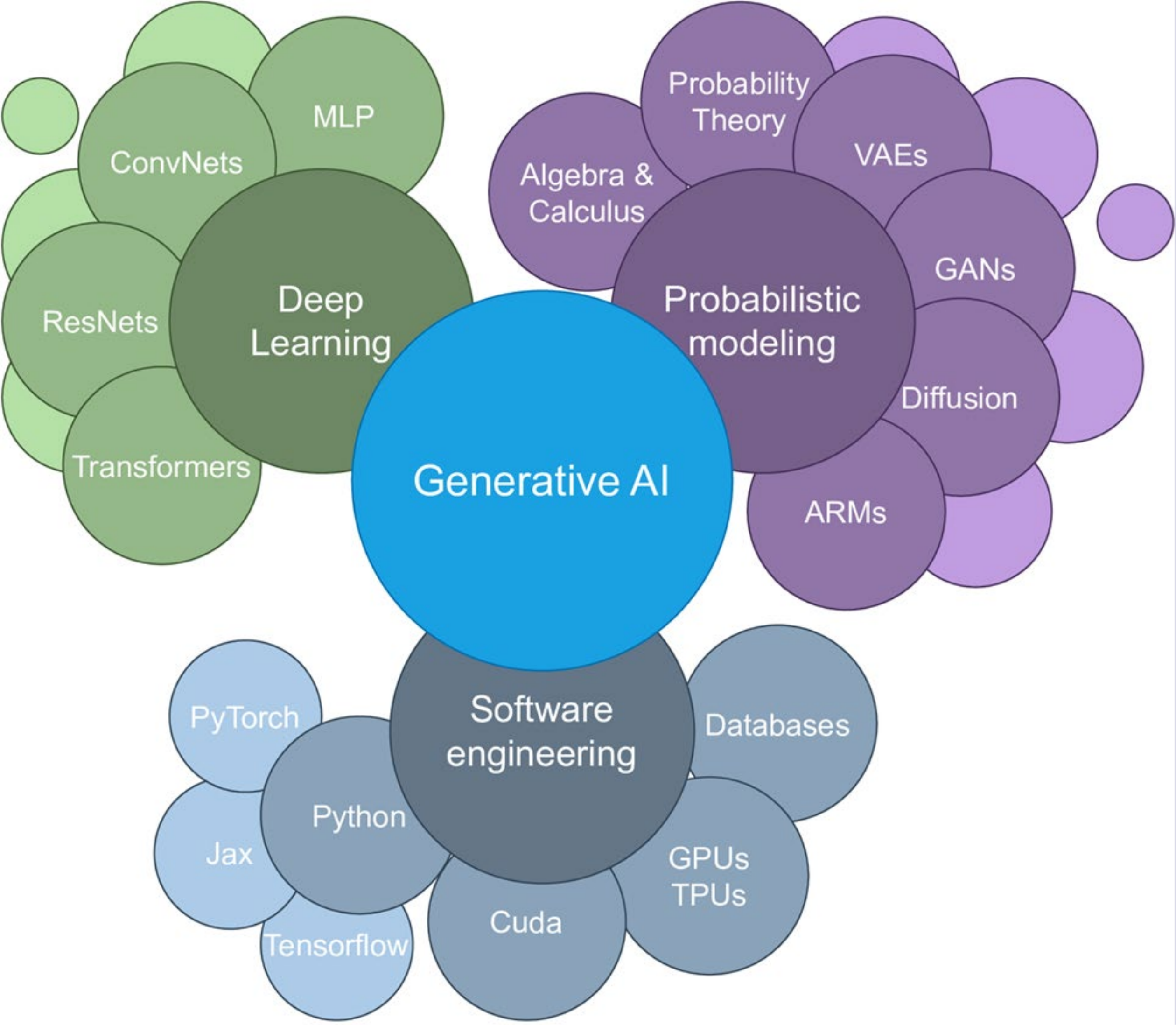
(principles of building models)

Deep Learning

(parameterizations of distributions)

Software engineering

(effective and efficient implementation of Generative AI)



Generative AI

Probabilistic modeling

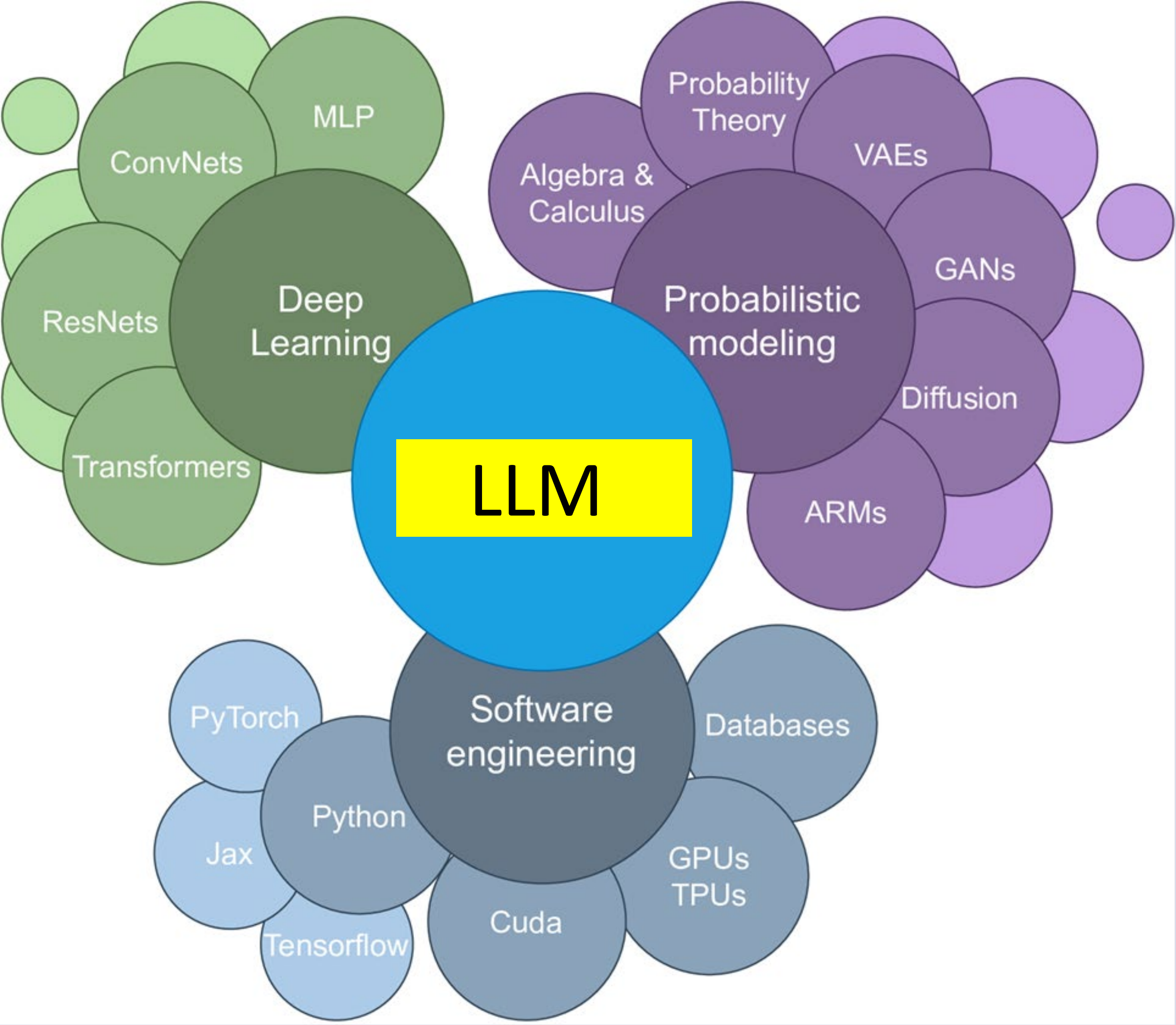
(principles of building models)

Deep Learning

(parameterizations of distributions)

Software engineering

(effective and efficient implementation of Generative AI)



Generative AI

Probabilistic modeling

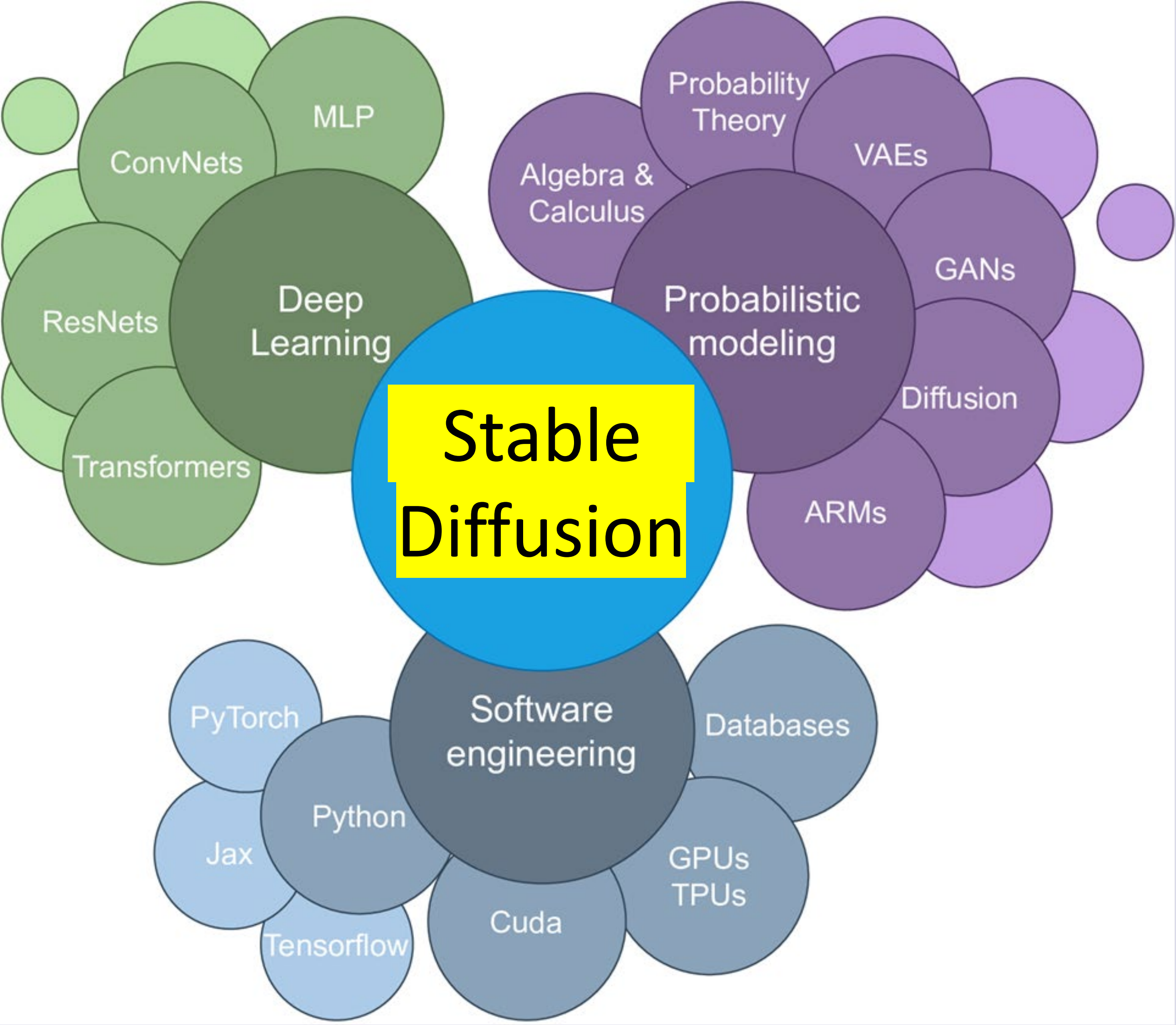
(principles of building models)

Deep Learning

(parameterizations of distributions)

Software engineering

(effective and efficient implementation of Generative AI)



How to build Generative AI?

Why Deep Generative Modeling?

Jakub M. Tomczak

Pages 1–13

NEW!

Probabilistic Modeling: From Mixture Models to Probabilistic Circuits

Jakub M. Tomczak

Pages 15–36

Updated!

Autoregressive Models

Jakub M. Tomczak

Pages 37–62

Flow-Based Models

Jakub M. Tomczak

Pages 63–92

Latent Variable Models

Jakub M. Tomczak

Pages 93–167

Updated!

Hybrid Modeling

Jakub M. Tomczak

Pages 169–181

Energy-Based Models

Jakub M. Tomczak

Pages 183–199

Generative Adversarial Networks

Jakub M. Tomczak

Pages 201–215

NEW!

Score-Based Generative Models

Jakub M. Tomczak

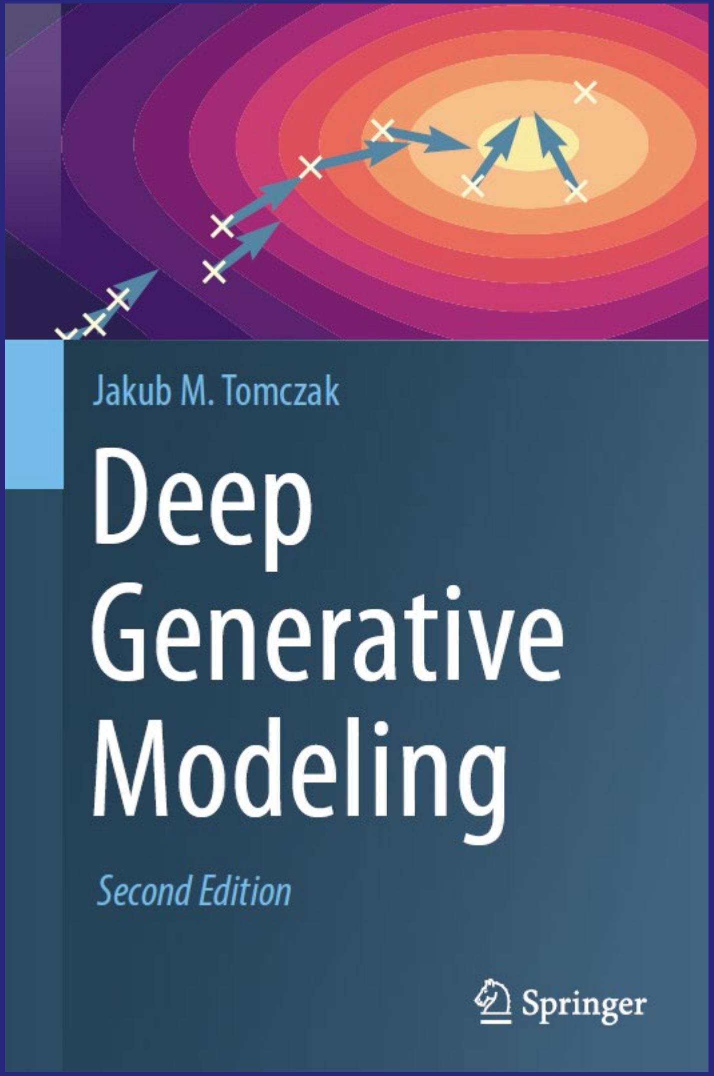
Pages 217–257

From Large Language Models to Generative AI Systems

Jakub M. Tomczak

Pages 277–302

NEW!



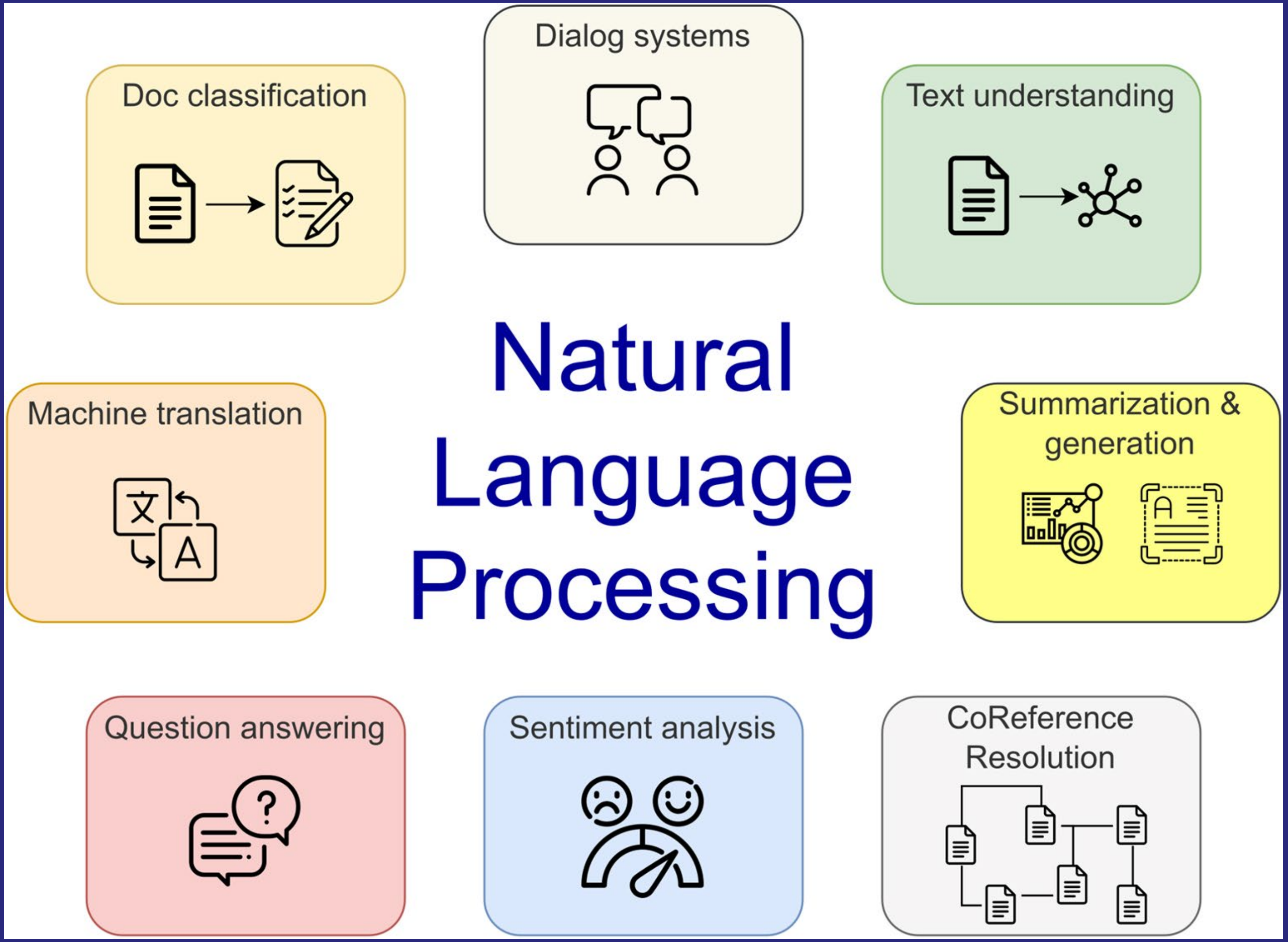
Jakub M. Tomczak

Deep Generative Modeling

Second Edition

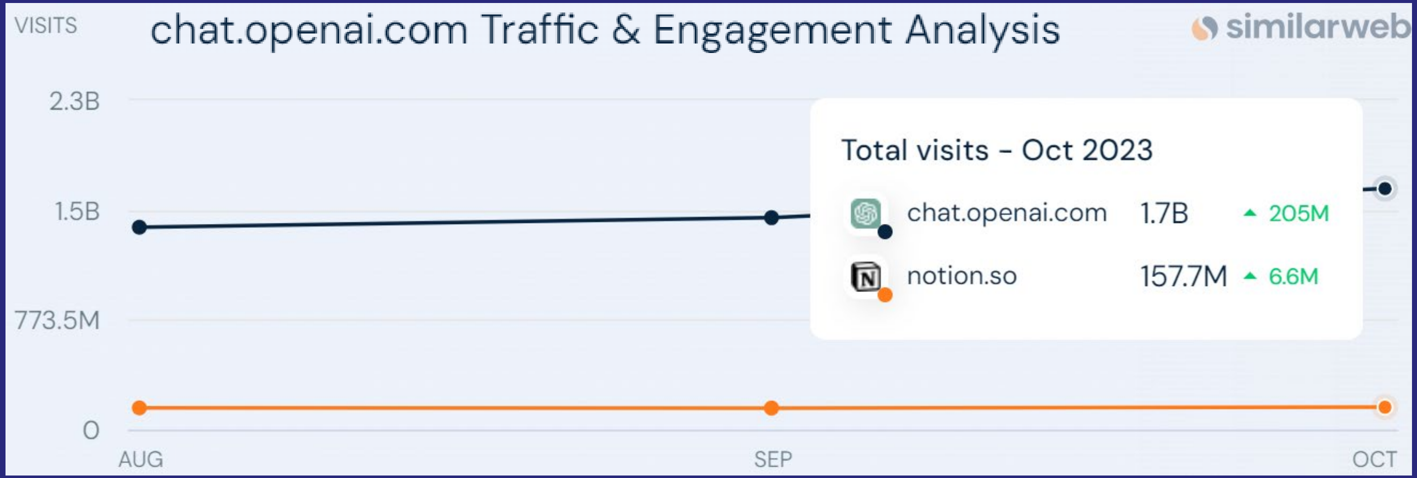
Springer

Large Language Models


Dialog systems
Doc classification
Text understanding
Natural Language Processing
Machine translation
文
A
Summarization & generation
A
Question answering
?
Sentiment analysis
CoReference Resolution

OpenAI Chat GPT

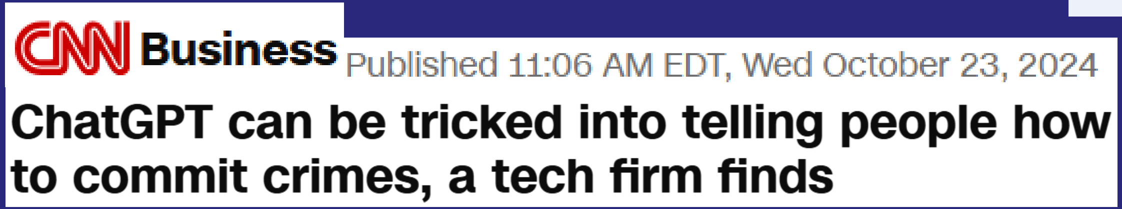
Launched November 30, 2022



chat.openai.com Traffic & Engagement Analysis

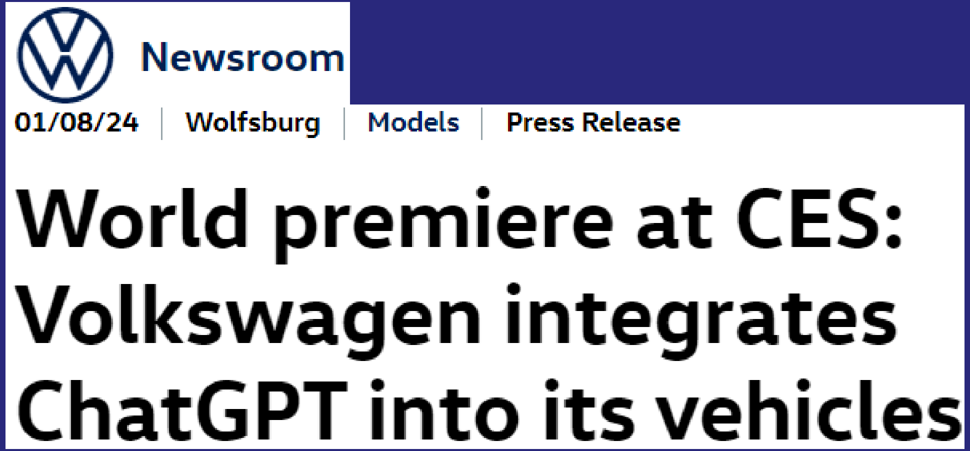
Total visits – Oct 2023

chat.openai.com	1.7B	▲ 205M
notion.so	157.7M	▲ 6.6M



CNN Business Published 11:06 AM EDT, Wed October 23, 2024

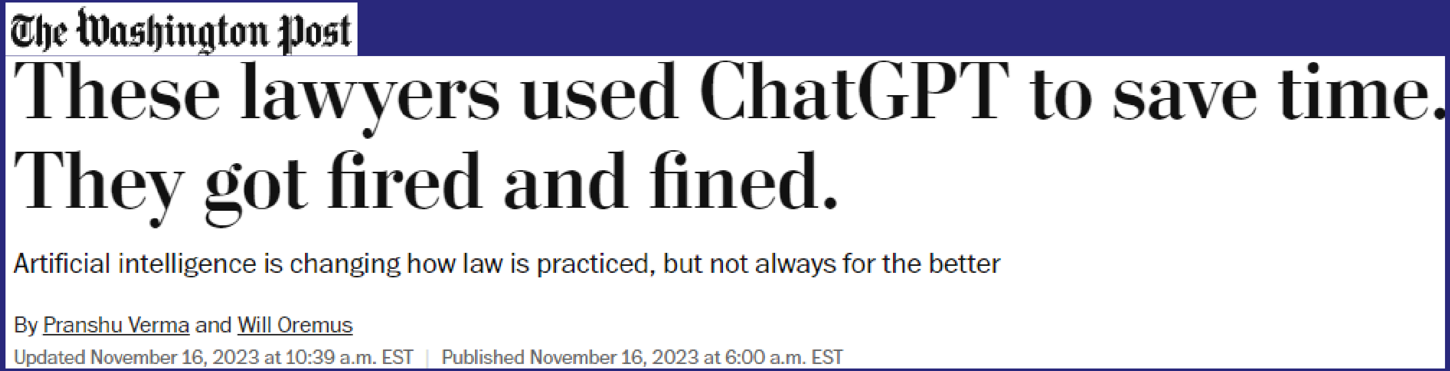
ChatGPT can be tricked into telling people how to commit crimes, a tech firm finds



Newsroom

01/08/24 | Wolfsburg | Models | Press Release

World premiere at CES: Volkswagen integrates ChatGPT into its vehicles



The Washington Post

These lawyers used ChatGPT to save time. They got fired and fined.

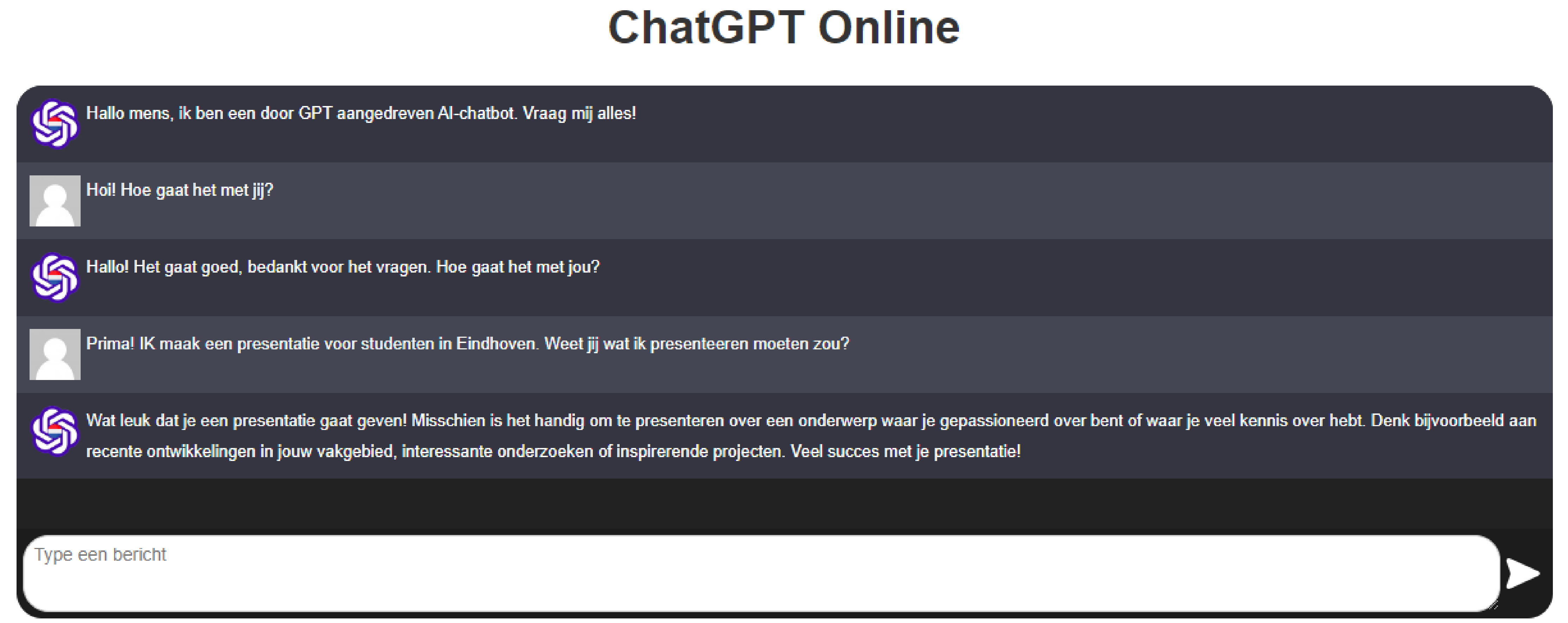
Artificial intelligence is changing how law is practiced, but not always for the better

By Pranshu Verma and Will Oremus

Updated November 16, 2023 at 10:39 a.m. EST | Published November 16, 2023 at 6:00 a.m. EST

https://chat.openai.com/

Dutch ChatGPT



ChatGPT Online

Hallo mens, ik ben een door GPT aangedreven AI-chatbot. Vraag mij alles!

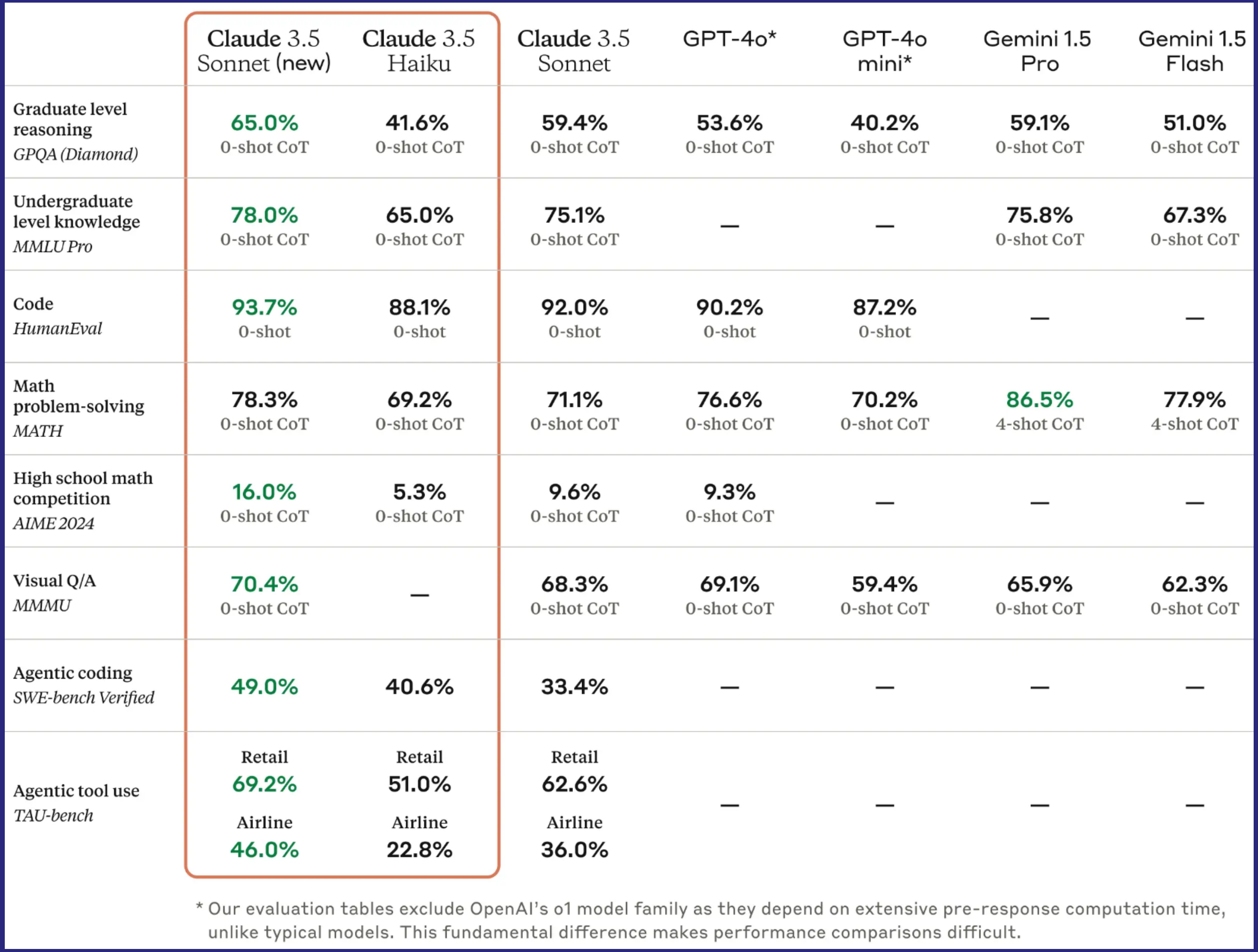
Hoi! Hoe gaat het met jij?

Hallo! Het gaat goed, bedankt voor het vragen. Hoe gaat het met jou?

Prima! IK maak een presentatie voor studenten in Eindhoven. Weet jij wat ik presenteeren moeten zou?

Wat leuk dat je een presentatie gaat geven! Misschien is het handig om te presenteren over een onderwerp waar je gepassioneerd over bent of waar je veel kennis over hebt. Denk bijvoorbeeld aan recente ontwikkelingen in jouw vakgebied, interessante onderzoeken of inspirerende projecten. Veel succes met je presentatie!

https://chatgptopenai.nl/chatgpt-online/

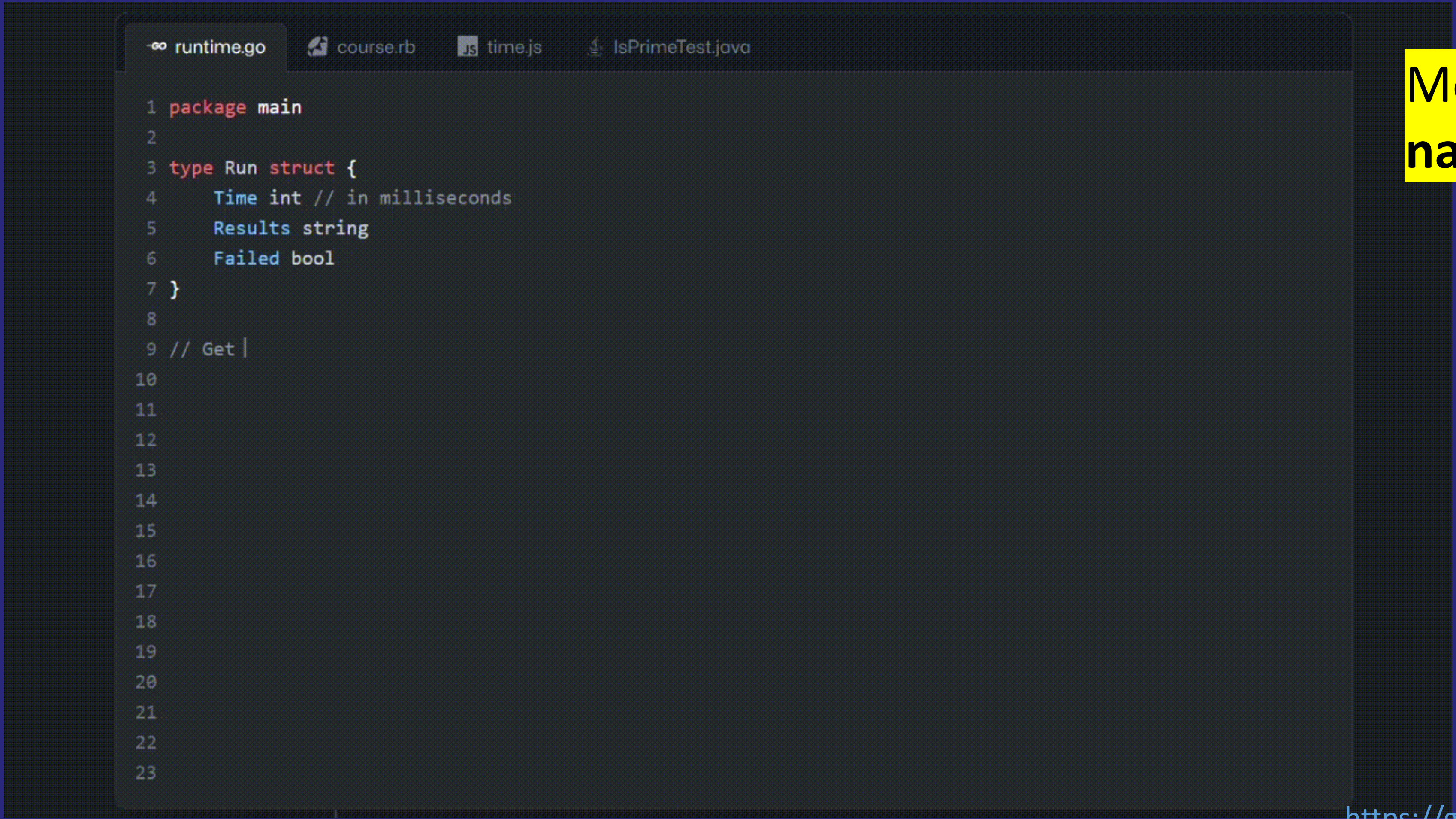


	Claude 3.5 Sonnet (new)	Claude 3.5 Haiku	Claude 3.5 Sonnet	GPT-4o*	GPT-4o mini*	Gemini 1.5 Pro	Gemini 1.5 Flash
Graduate level reasoning *GPQA (Diamond)*	65.0% 0-shot CoT	41.6% 0-shot CoT	59.4% 0-shot CoT	53.6% 0-shot CoT	40.2% 0-shot CoT	59.1% 0-shot CoT	51.0% 0-shot CoT
Undergraduate level knowledge *MMLU Pro*	78.0% 0-shot CoT	65.0% 0-shot CoT	75.1% 0-shot CoT	—	—	75.8% 0-shot CoT	67.3% 0-shot CoT
Code *HumanEval*	93.7% 0-shot	88.1% 0-shot	92.0% 0-shot	90.2% 0-shot	87.2% 0-shot	—	—
Math problem-solving *MATH*	78.3% 0-shot CoT	69.2% 0-shot CoT	71.1% 0-shot CoT	76.6% 0-shot CoT	70.2% 0-shot CoT	86.5% 4-shot CoT	77.9% 4-shot CoT
High school math competition *AIME 2024*	16.0% 0-shot CoT	5.3% 0-shot CoT	9.6% 0-shot CoT	9.3% 0-shot CoT	—	—	—
Visual Q/A *MMMU*	70.4% 0-shot CoT	—	68.3% 0-shot CoT	69.1% 0-shot CoT	59.4% 0-shot CoT	65.9% 0-shot CoT	62.3% 0-shot CoT
Agentic coding *SWE-bench Verified*	49.0%	40.6%	33.4%	—	—	—	—
Agentic tool use *TAU-bench*	Retail 69.2% Airline 46.0%	Retail 51.0% Airline 22.8%	Retail 62.6% Airline 36.0%	—	—	—	—

* Our evaluation tables exclude OpenAI's o1 model family as they depend on extensive pre-response computation time, unlike typical models. This fundamental difference makes performance comparisons difficult.

https://www.anthropic.com/news/3-5-models-and-computer-use

GitHub Copilot



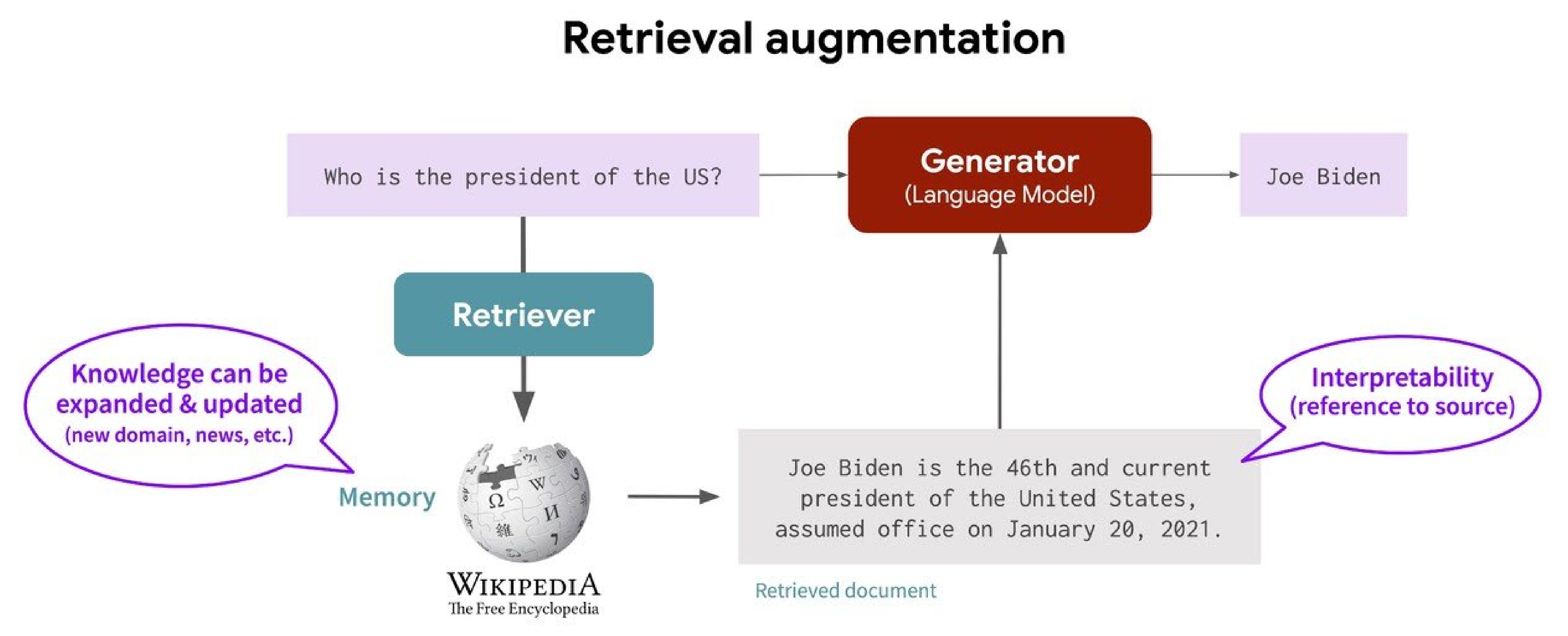
```go
package main

type Run struct {
    Time int // in milliseconds
    Results string
    Failed bool
}

// Get
```

More than **natural** language

https://github.com/features/copilot

Going beyond LLMs: Generative AI Systems

RAG: Retrieval-augmented Generation



Retrieval augmentation

Who is the president of the US?

Generator
(Language Model)

Joe Biden

Retriever

Knowledge can be expanded & updated
(new domain, news, etc.)

Memory

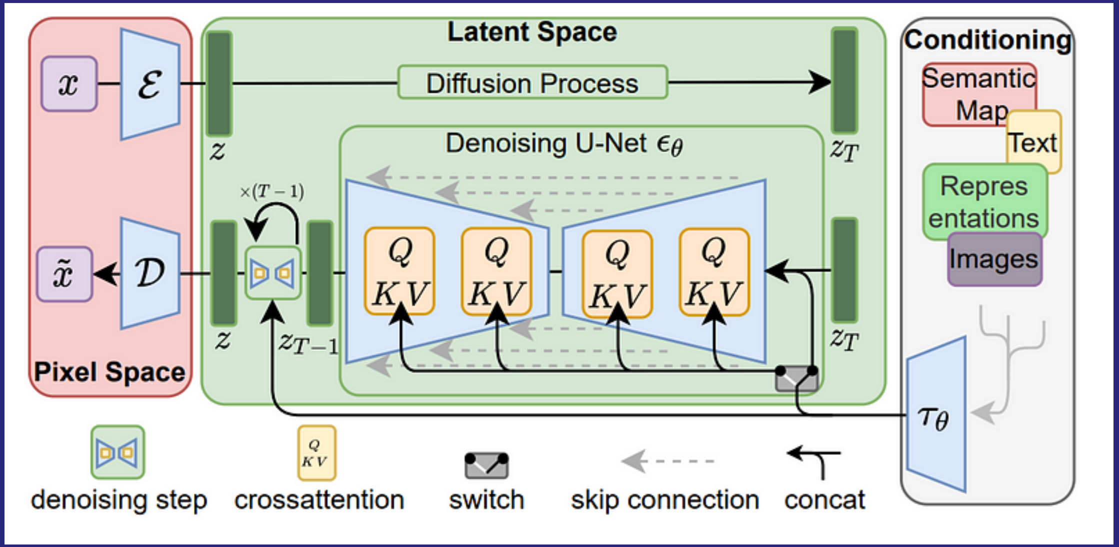
Joe Biden is the 46th and current president of the United States, assumed office on January 20, 2021.

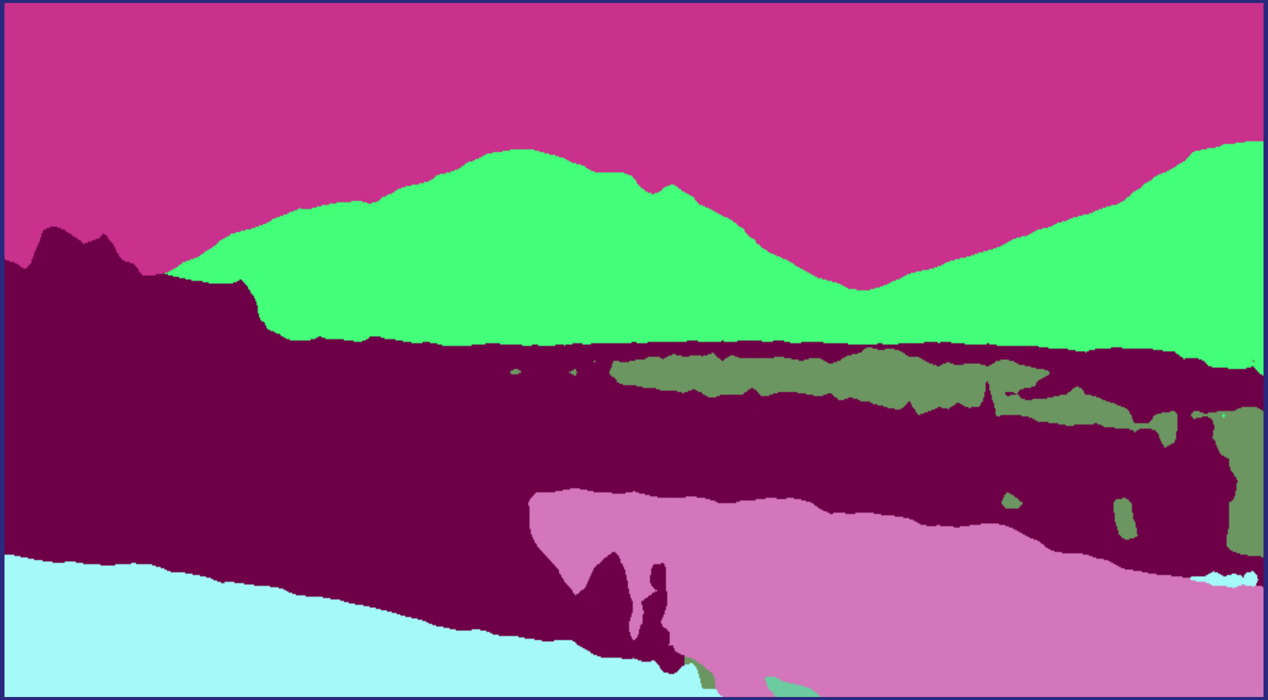
Retrieved document

Interpretability
(reference to source)

https://ai.meta.com/blog/retrieval-augmented-generation-streamlining-the-creation-of-intelligent-natural-language-processing-models/

Stable Diffusion (a.k.a. Latent Diffusion)



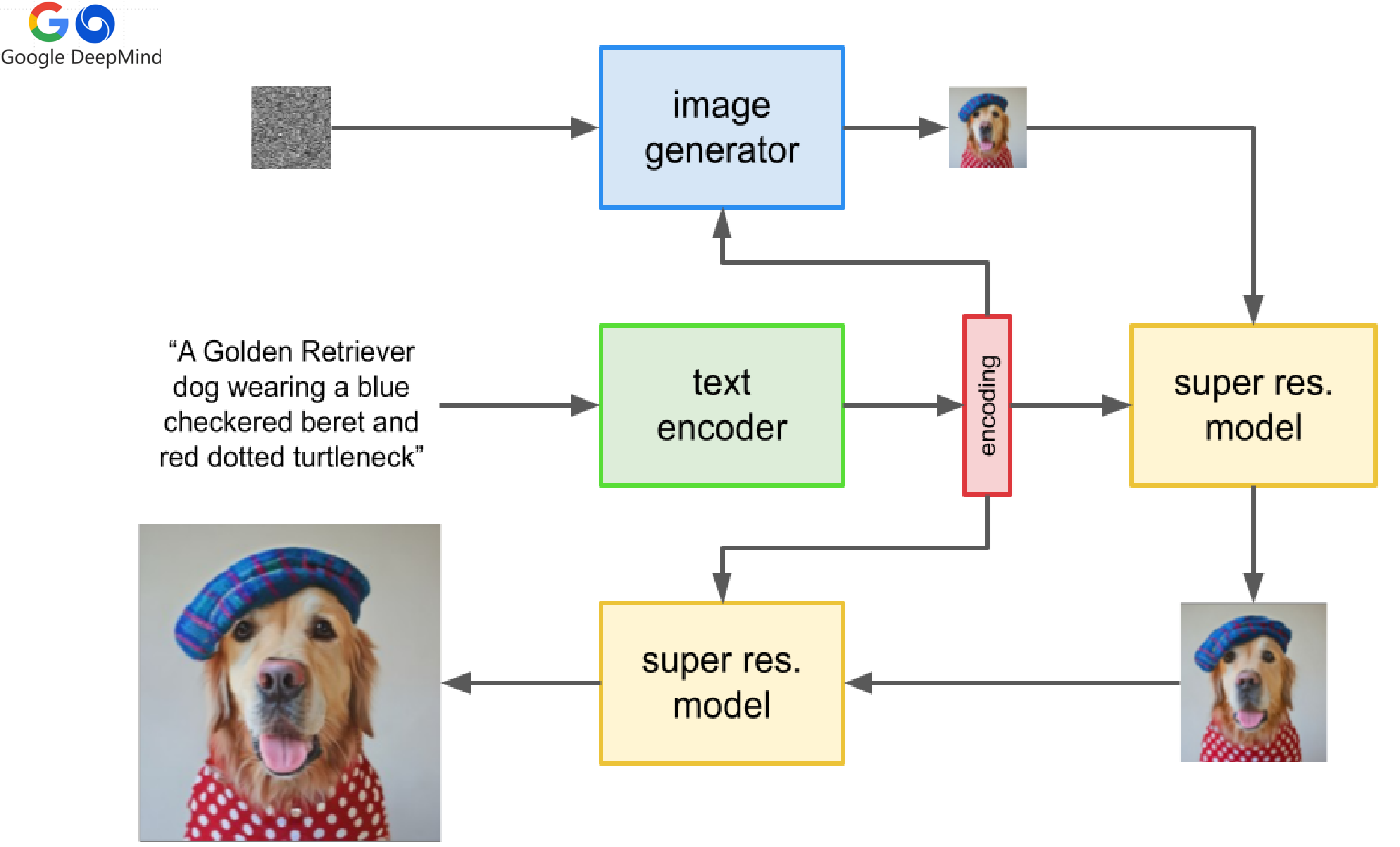


Pre-trained Auto-Encoder
Diffusion model in the latent space
Sampling/Inference:
Diffusion model -> Decoder

Segmentation-to-Image
Inpainting
Superresolution
Text to Image

Rombach et al. (2022). High-resolution image synthesis with latent diffusion models. CVPR

ImaGen: Diffusion + Superresolution



Google DeepMind

image generator

"A Golden Retriever dog wearing a blue checkered beret and red dotted turtleneck"

text encoder

encoding

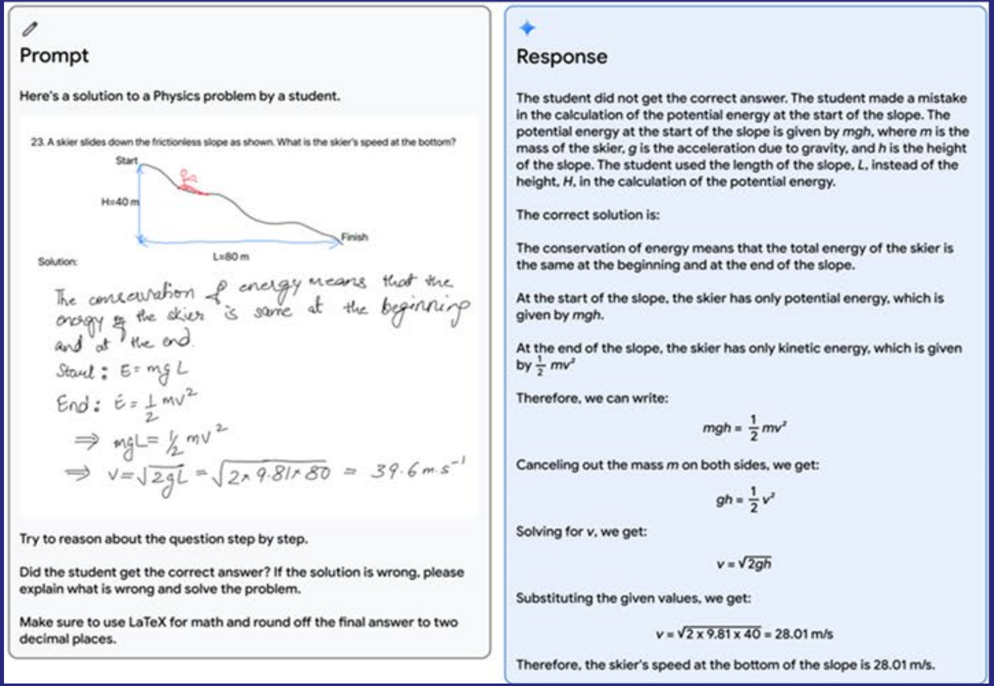
super res. model

super res. model

Saharia et al. (2022). Photorealistic text-to-image diffusion models with deep language understanding NeurIPS

Multimodal Generative AI (various data modalities)

Google DeepMind



Prompt

Here's a solution to a Physics problem by a student.

23. A skier slides down the frictionless slope as shown. What is the skier's speed at the bottom?

Solution:

The conservation of energy means that the energy of the skier is same at the beginning and at the end.

Start: $E = mgL$

End: $E = \frac{1}{2} mv^2$

$\Rightarrow mgL = \frac{1}{2} mv^2$

$\Rightarrow v = \sqrt{2gL} = \sqrt{2 \times 9.81 \times 80} = 39.6 \, m.s^{-1}$

Try to reason about the question step by step.

Did the student get the correct answer? If the solution is wrong, please explain what is wrong and solve the problem.

Make sure to use LaTeX for math and round off the final answer to two decimal places.

Response

The student did not get the correct answer. The student made a mistake in the calculation of the potential energy at the start of the slope. The potential energy at the start of the slope is given by mgh, where m is the mass of the skier, g is the acceleration due to gravity, and h is the height of the slope. The student used the length of the slope, L, instead of the height, H, in the calculation of the potential energy.

The correct solution is:

The conservation of energy means that the total energy of the skier is the same at the beginning and at the end of the slope.

At the start of the slope, the skier has only potential energy, which is given by mgh.

At the end of the slope, the skier has only kinetic energy, which is given by $\frac{1}{2} mv^2$

Therefore, we can write:

$$mgh = \frac{1}{2} mv^2$$

Canceling out the mass m on both sides, we get:

$$gh = \frac{1}{2} v^2$$

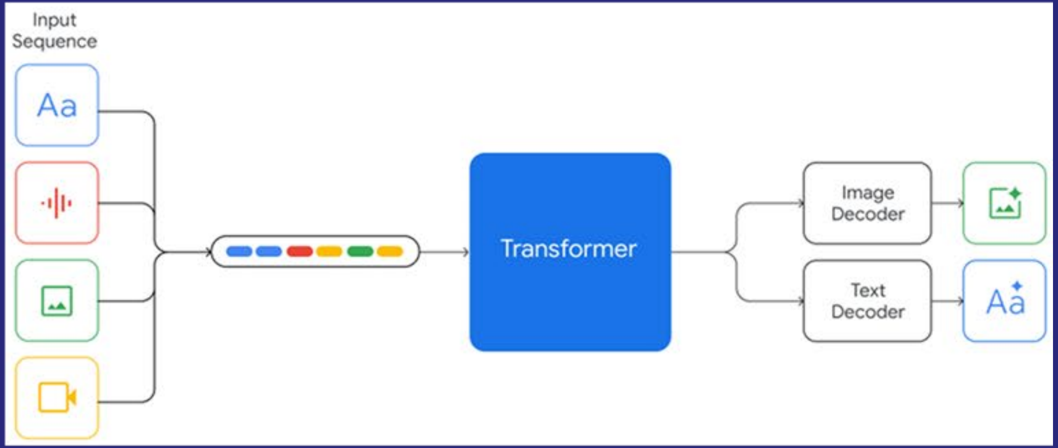
Solving for v, we get:

$$v = \sqrt{2gh}$$

Substituting the given values, we get:

$$v = \sqrt{2 \times 9.81 \times 40} = 28.01 \text{ m/s}$$

Therefore, the skier's speed at the bottom of the slope is 28.01 m/s.



Gemini: A Family of Highly Capable Multimodal Models

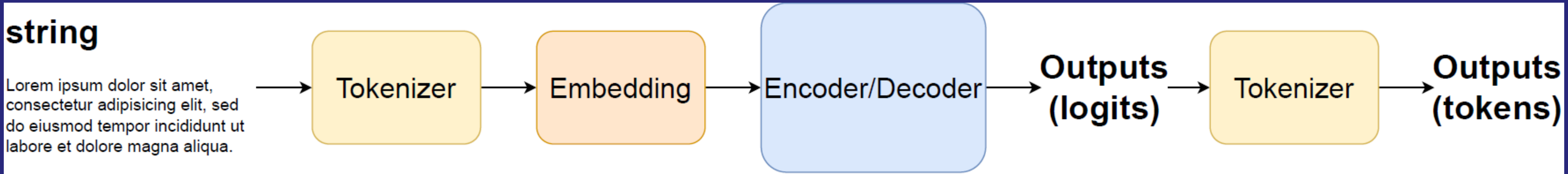
https://deepmind.google/technologies/gemini

How to design such systems?

The following questions:

1. What are the common components?
2. What about compositionality?
3. Can we come up with a general structure?
4. Can we come up with general design patterns?

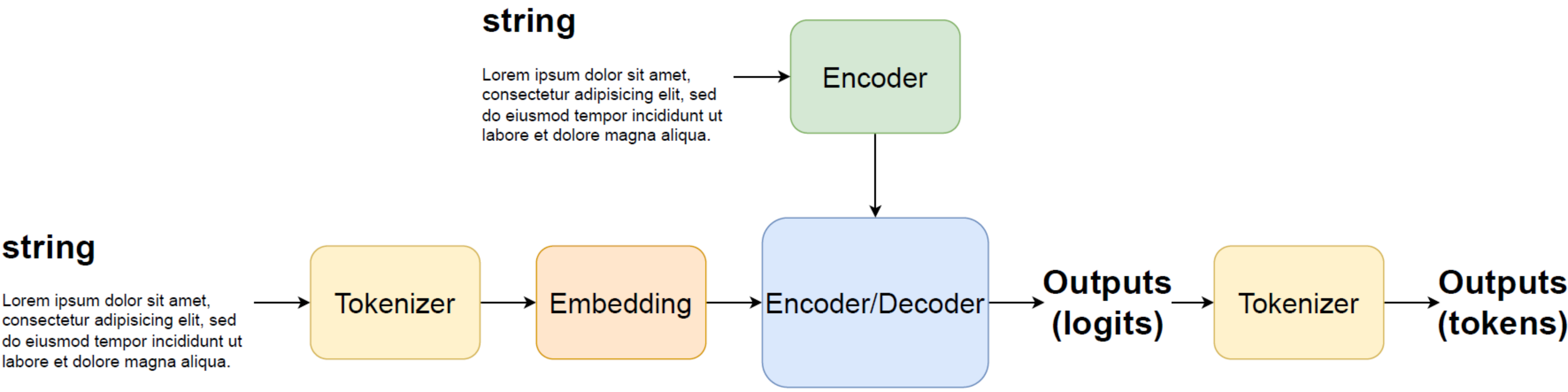
LLMs as systems



string

Lorem ipsum dolor sit amet, consectetur adipisicing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua.

Tokenizer → Embedding → Encoder/Decoder → **Outputs (logits)** → Tokenizer → **Outputs (tokens)**



string

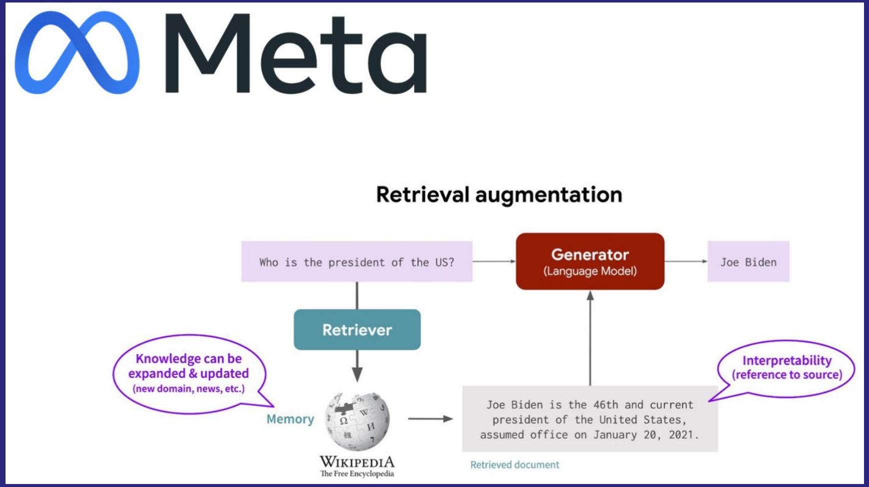
Lorem ipsum dolor sit amet, consectetur adipisicing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua.

Encoder

string

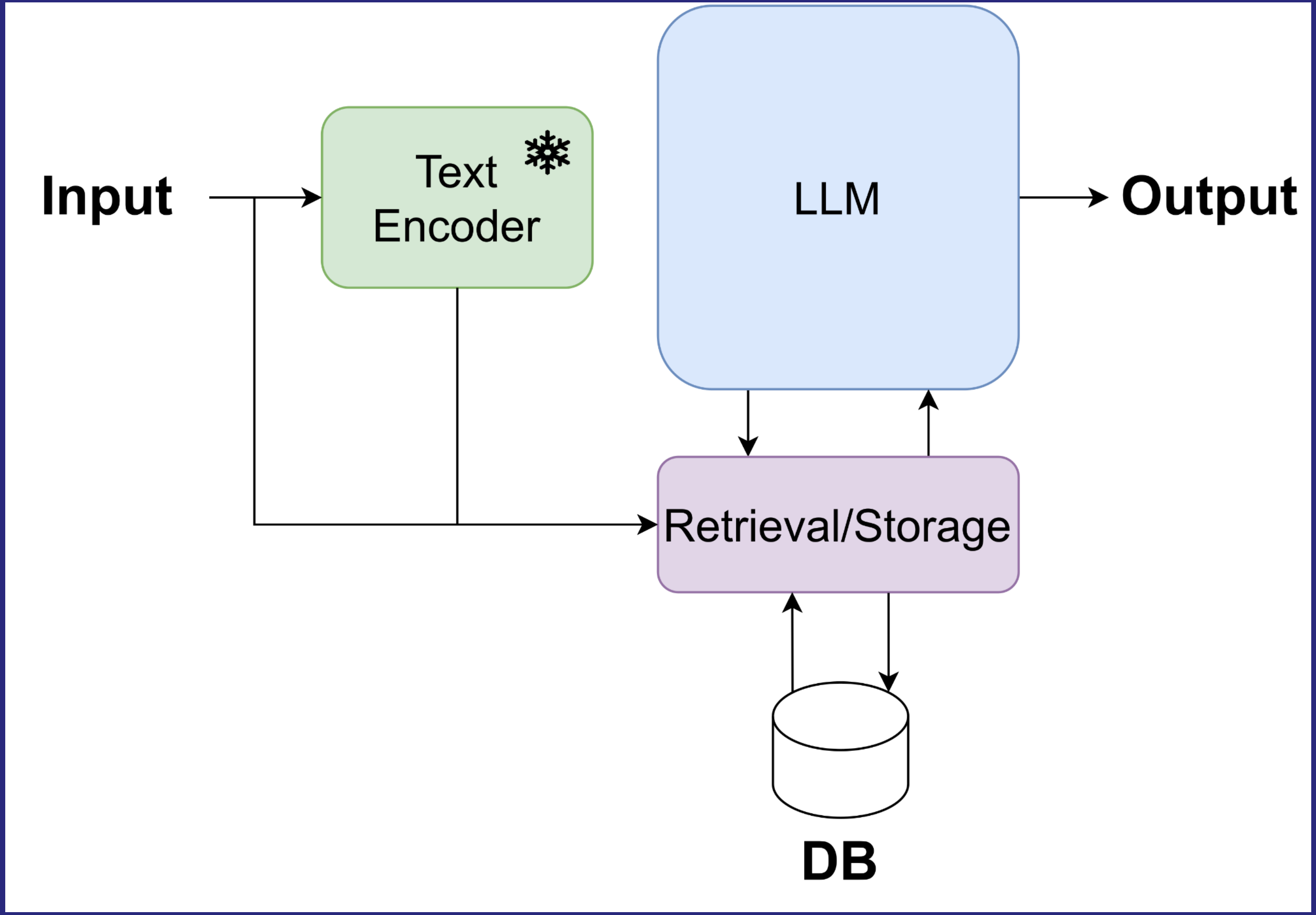
Lorem ipsum dolor sit amet, consectetur adipisicing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua.

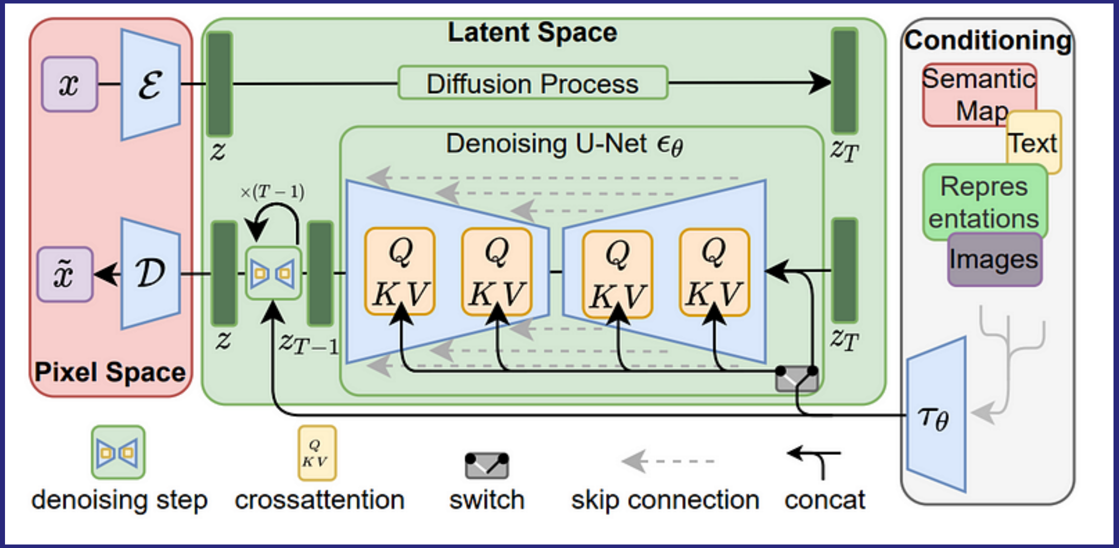
Tokenizer → Embedding → Encoder/Decoder → **Outputs (logits)** → Tokenizer → **Outputs (tokens)**



Joe Biden is the 46th and current president of the United States, assumed office on January 20, 2021.

Components:
Text Encoder
Generative Model: LLM
Retrieval/Storage unit


Input
Text Encoder ❄
LLM
Output
Retrieval/Storage
DB

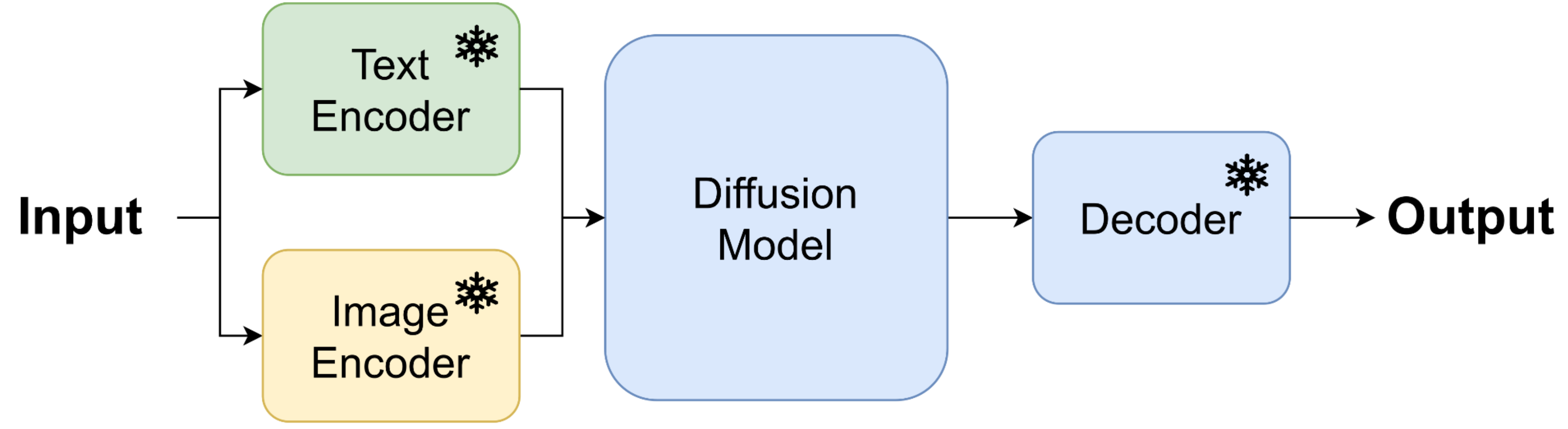


Components:

Text Encoder

Image encoder

Generative Model: Diffusion + Decoder



Input

Text Encoder ❄

Image Encoder ❄

Diffusion Model

Decoder ❄

Output

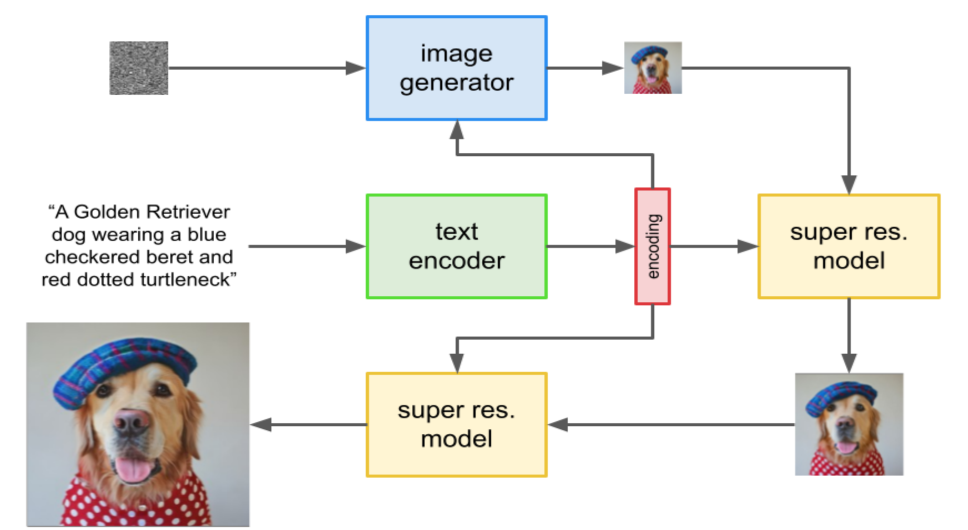
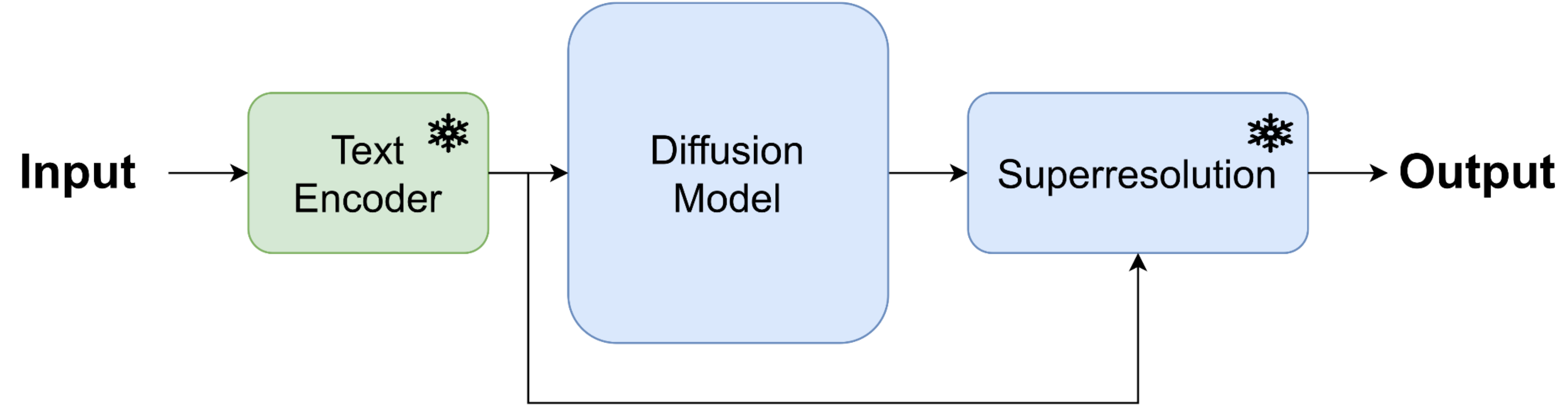


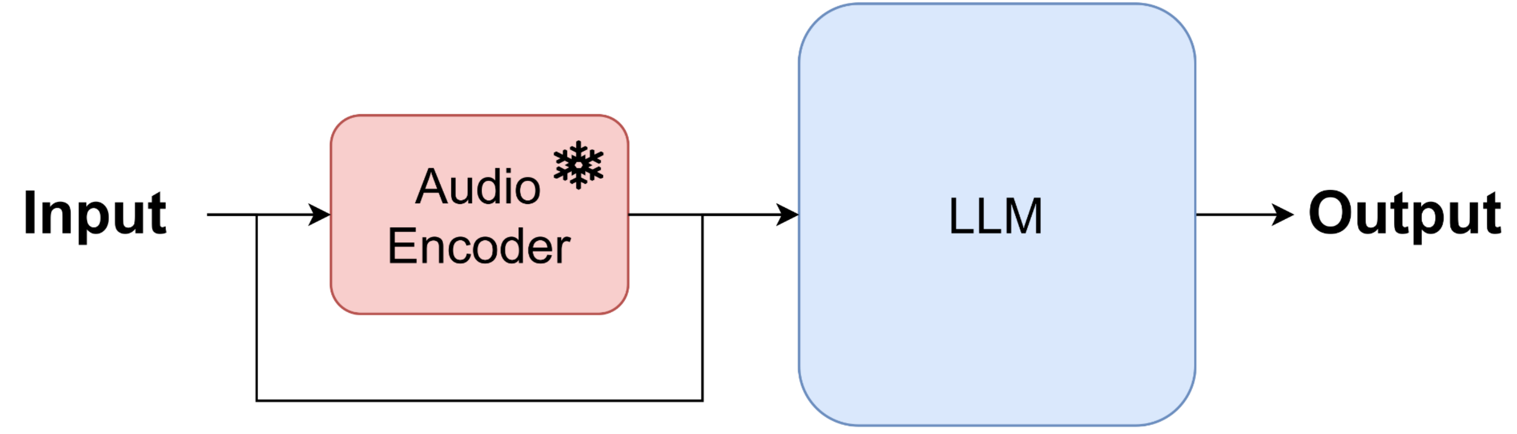
image generator
"A Golden Retriever dog wearing a blue checkered beret and red dotted turtleneck"
text encoder
encoding
super res. model
super res. model

<u>Components</u>:
Text Encoder
Generative Model: Diffusion + Superresolution



OpenAI/Whisper

<u>Components</u>:
Audio Encoder
Generative Model: LLM



Generative AI Systems (GenAISys)

Components:

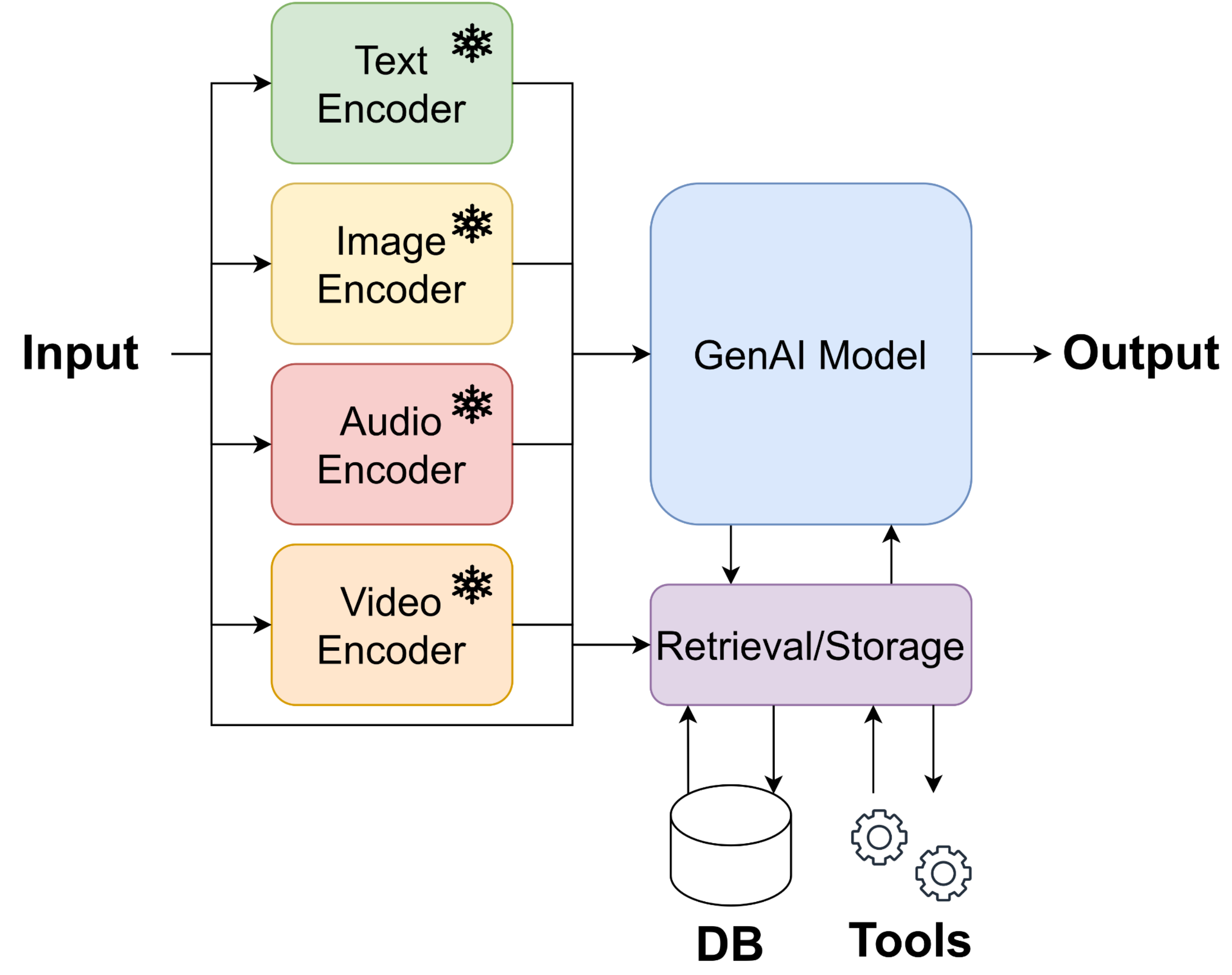
DEs: Data Encoders

GeM: Generative Model

R/S: Retrieval/Storage unit

Database (DB)

Tools



Generative AI Systems (GenAISys)

Compositionality:

atomic system (as) = state + dynamics
composite system (cs) = set of as

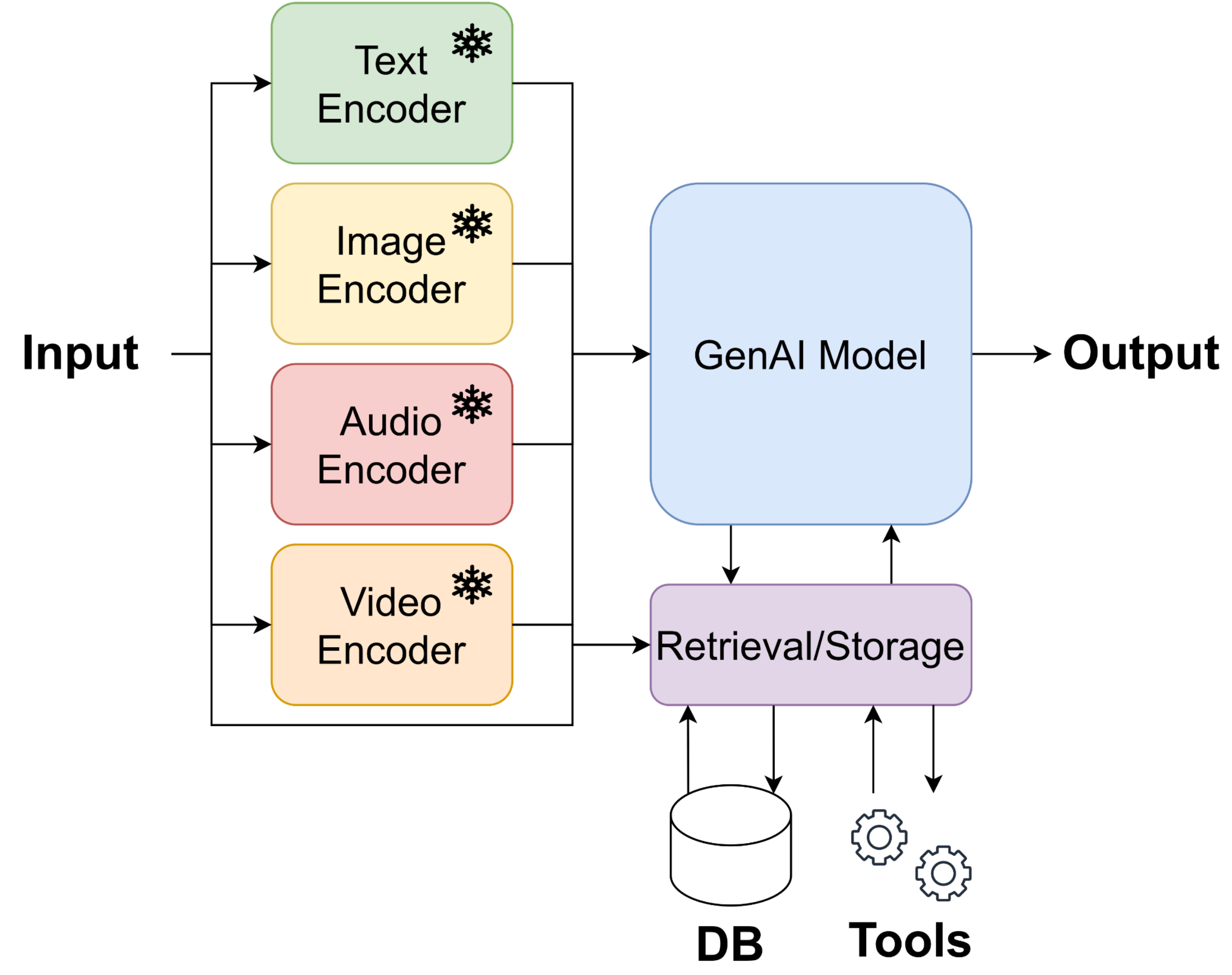
composition = how to combine as & cs

Compatibility:

Outputs of one system are legal as inputs of another system

System Verification & Validation (Reliable AI):

Verification is a set of actions used to check the correctness of a system and its elements. Validation checks if a system solves a given problem



Conclusion

LLMs: not only a hype but an important landmark in GenAI

GenAISys: we have them, we use them but need to understand them

Future:

Generative AI + Responsive AI = **Agentic AI**

How to formally analyze/understand GenAISys? **Category Theory**?

Thank you!

Questions?

Contact: j.m.tomczak@tue.nl
jmk.tomczak@gmail.com

Generativ/e

Generative AI Group: https://generativeai-tue.github.io/