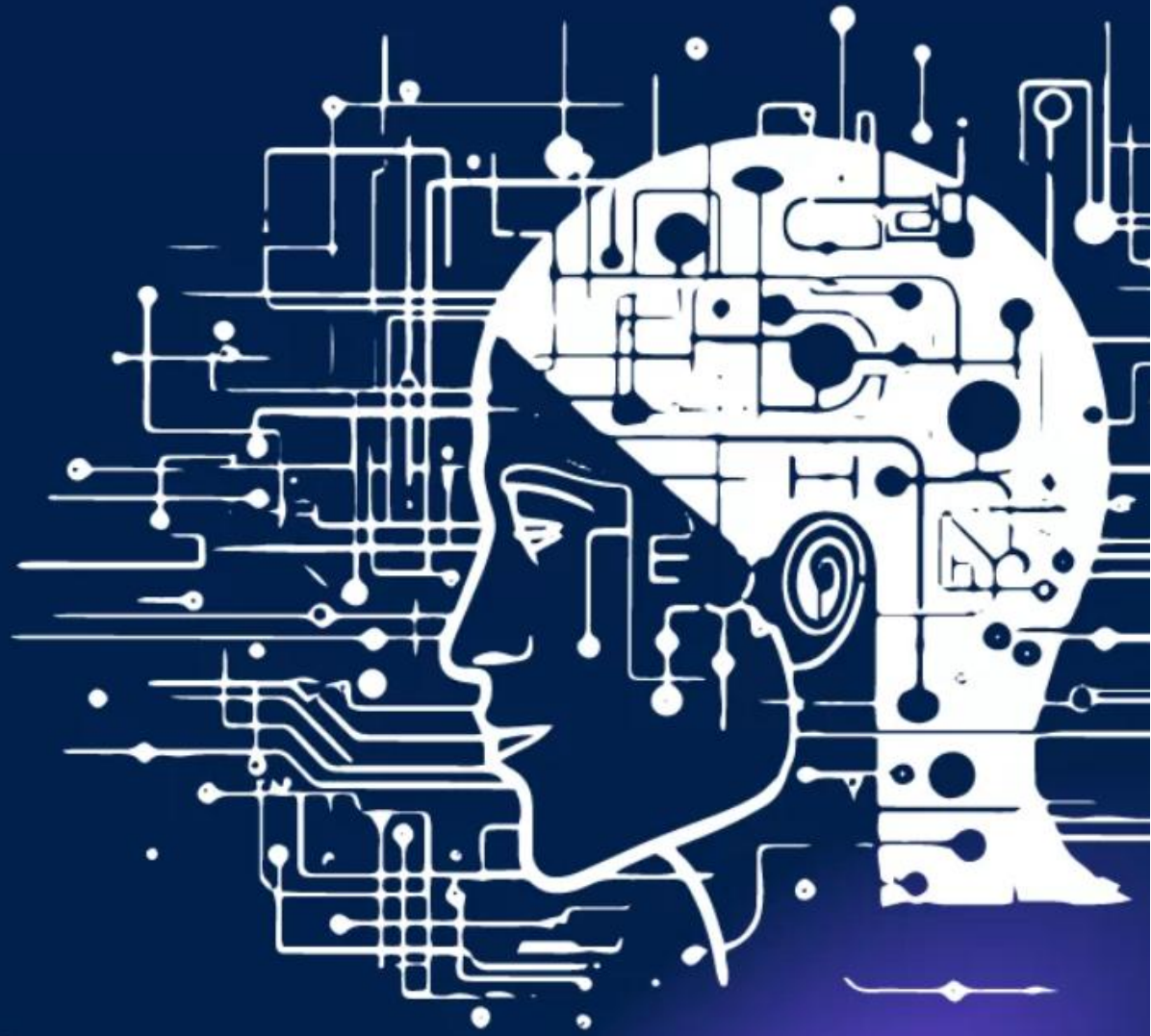
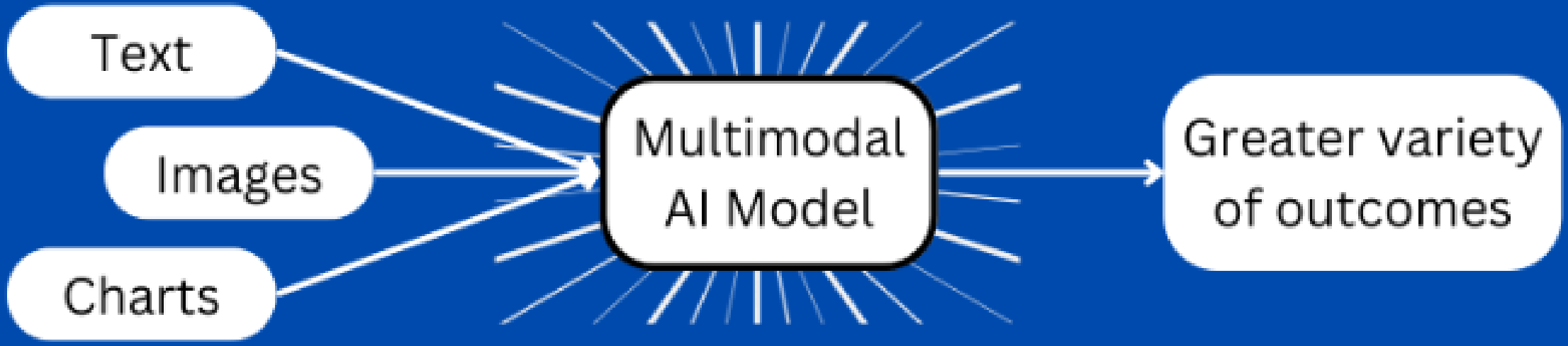
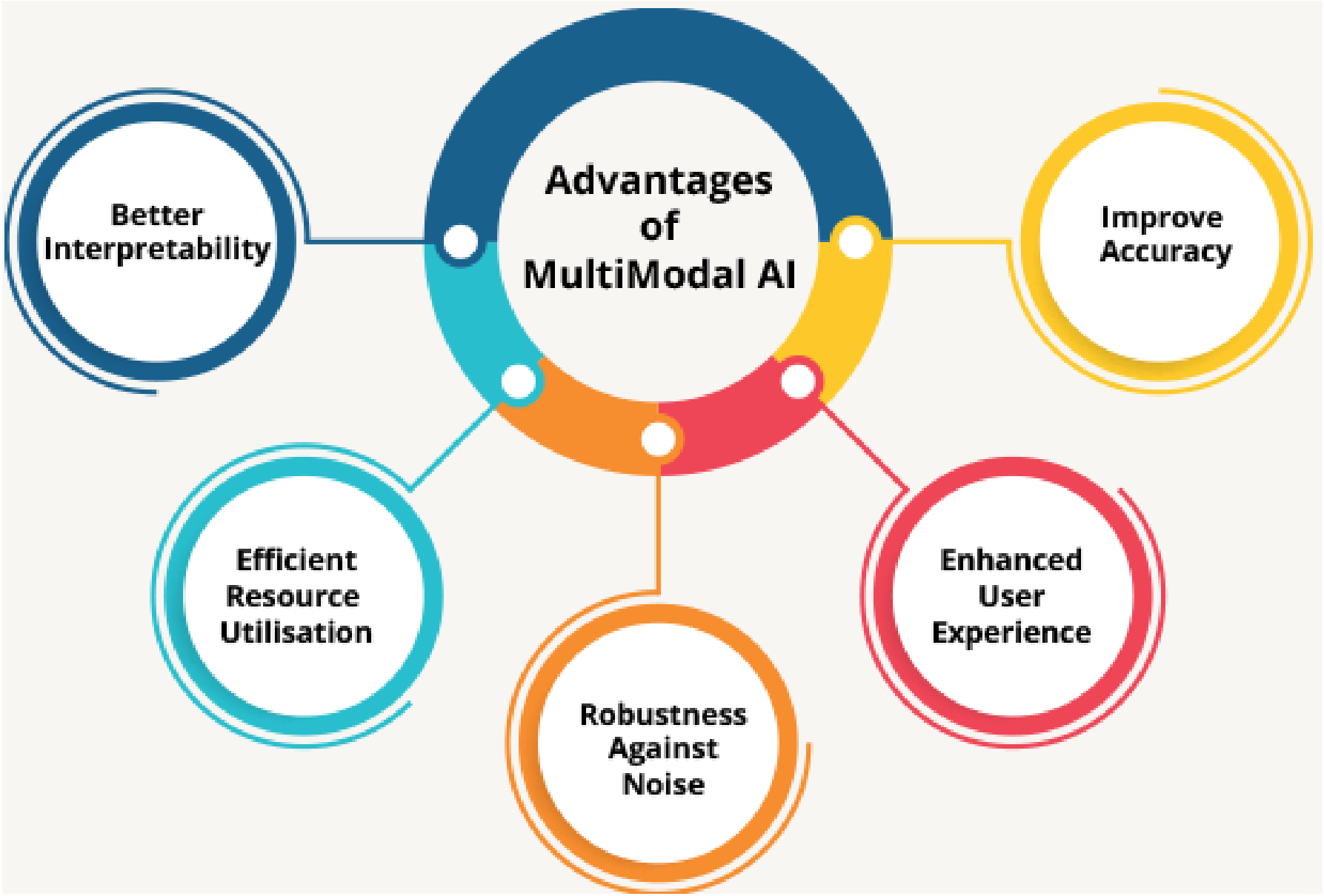
What Is Multimodal AI



Single-modal AI Model

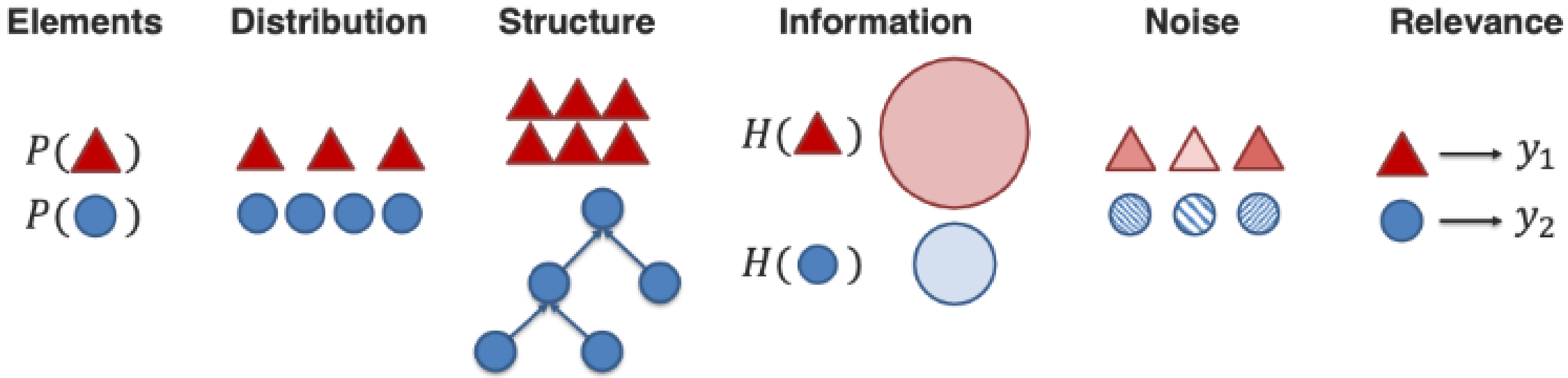
Multimodal AI Model?




Advantages of MultiModal AI
Better Interpretability
Improve Accuracy
Efficient Resource Utilisation
Robustness Against Noise
Enhanced User Experience

2.1 – Heterogeneity

One of the key insights of multimodal AI is that **modalities are heterogeneous**, meaning they often exhibit diverse qualities, structures, and representations. For example, an image and a speech recording of the same event may convey different information, requiring unique processing and analysis.



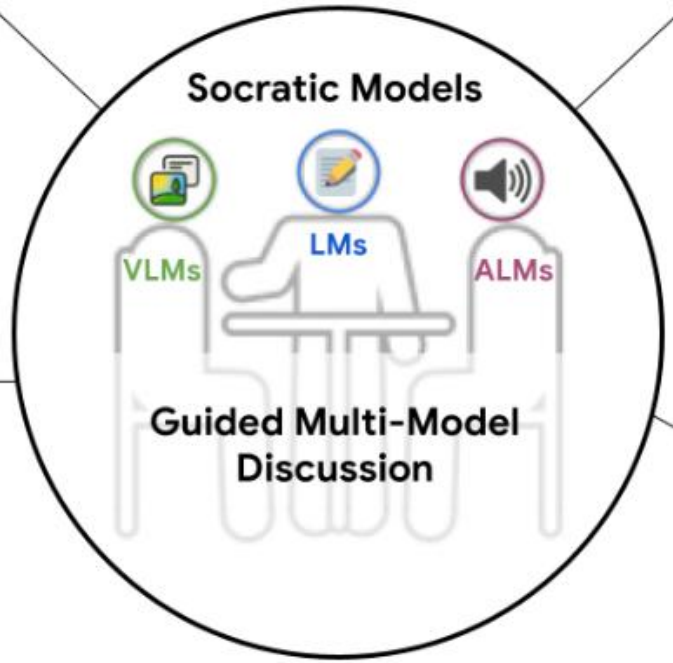
Source: https://arxiv.org/abs/2209.03430

Video Search

Q: Where did I leave my mug?

Image Captioning

Summary: I am watching netflix in a living room.



Free-form Video Q&A: Visual & Contextual Reasoning

Q: When did I wash my hands?

A: I last washed my hands at 3:38 PM.

Q: Why was I at the front porch?

A: I went to the front porch today to receive a package.

Q: Why was I chopping wood?

A: Because I needed to get a fire going in the fireplace.

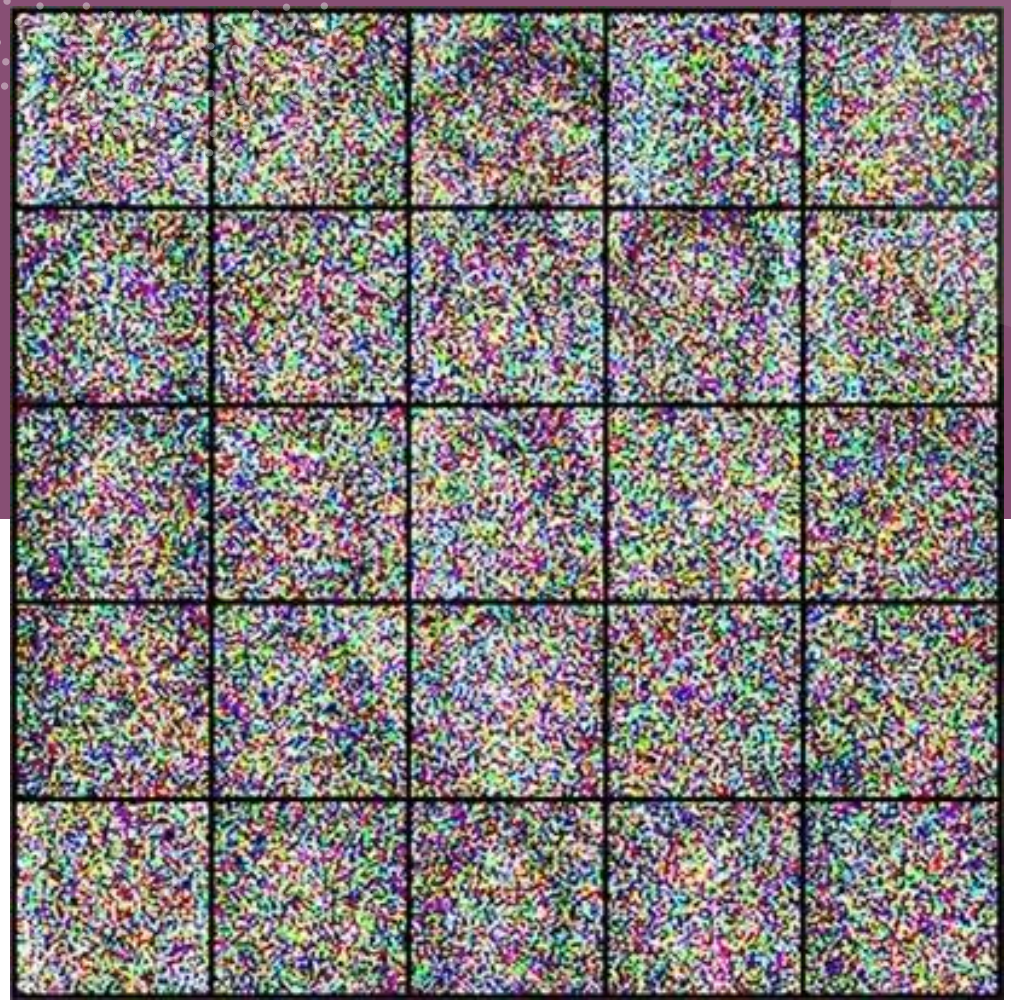
Forecasting: Predicting Future Activities

1:46 PM: I am eating a sandwich in a kitchen.
2:18 PM: I am checking time and working on a laptop in a clean room.
2:49 PM: I am buying produce from a grocery store or market.

3:21 PM: I am driving a car.
4:03 PM: I am in a park and see a playground.
4:35 PM: I am in a home and see a television.

Source: https://socraticmodels.github.io

DIFFUSION MODEL



From a pure Gaussian Noise,
generate an image
which won’t be replicated, without
any way to control the process

Diffusion Models

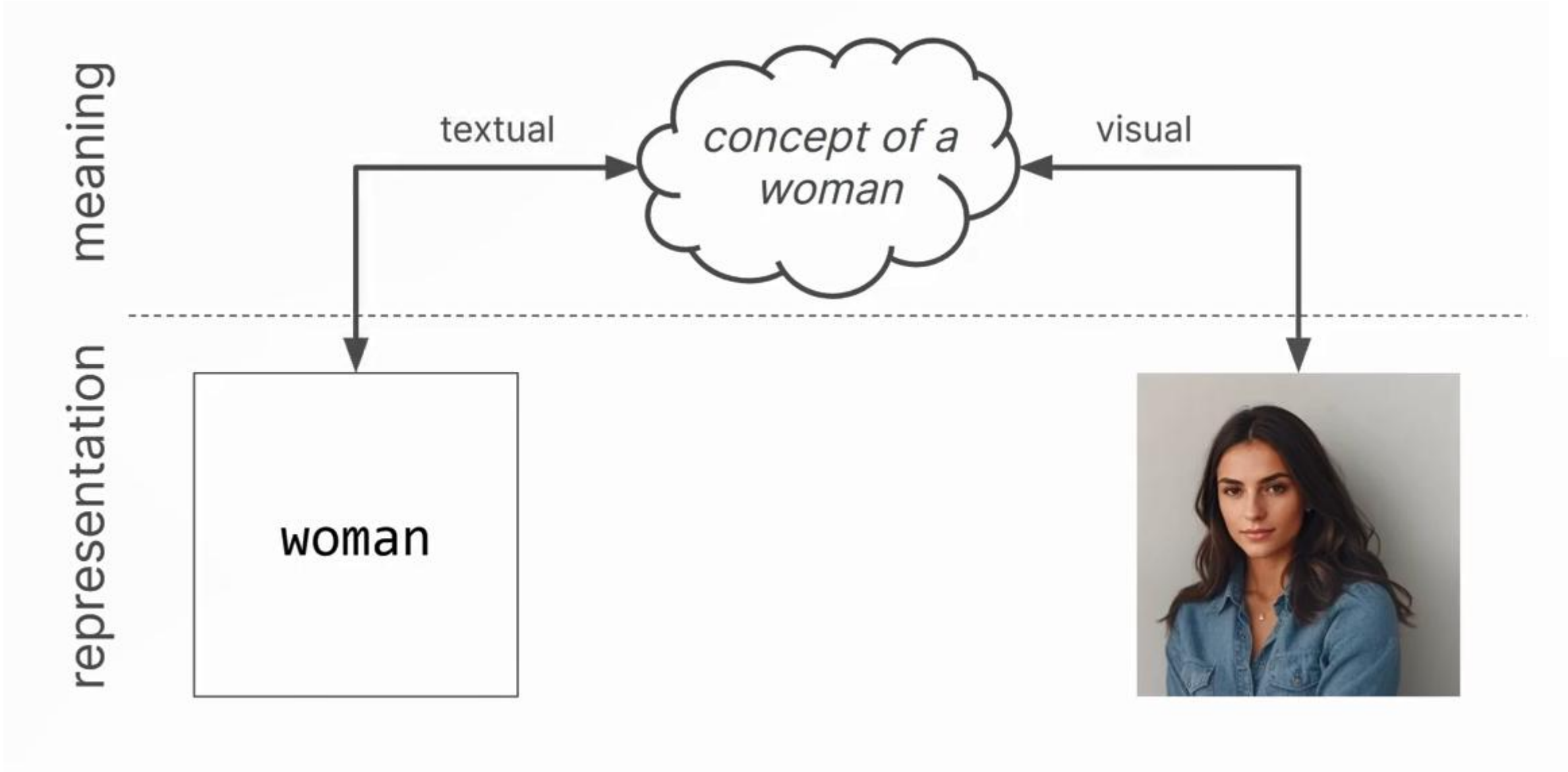
no input

Model

Text-to-image Models

A golden retriever wearing black sunglasses

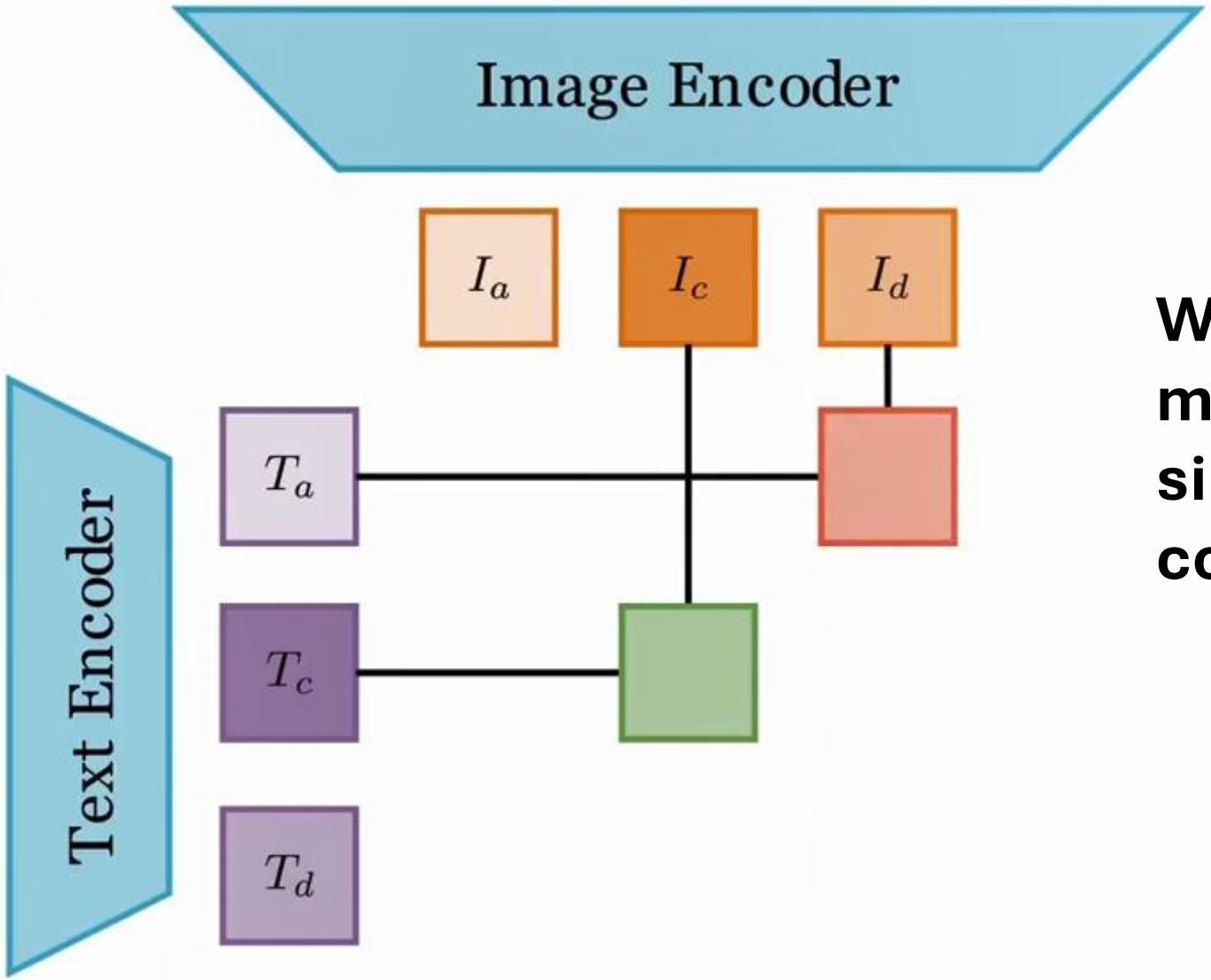
Model


meaning
representation
textual
concept of a woman
visual
woman

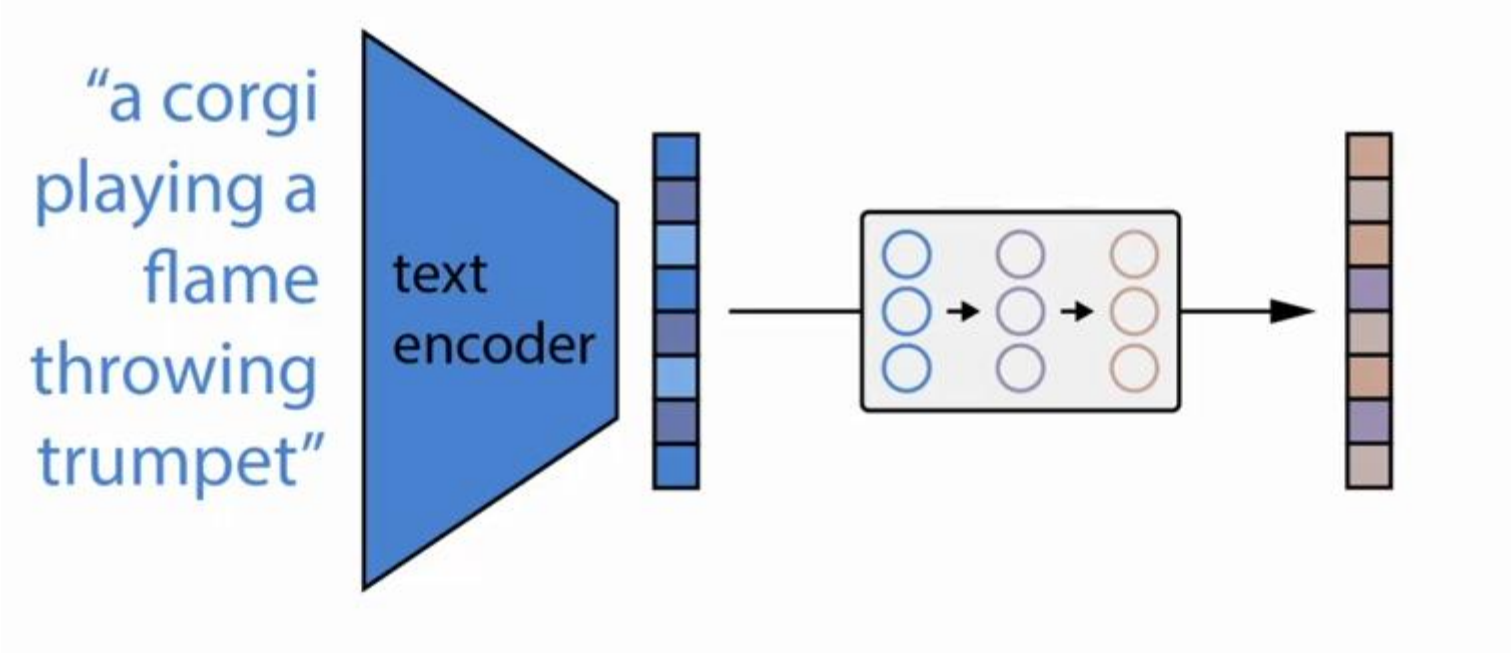
Multimodal AI model use embeddings models to convert text to images to vectors that capture the meaning.

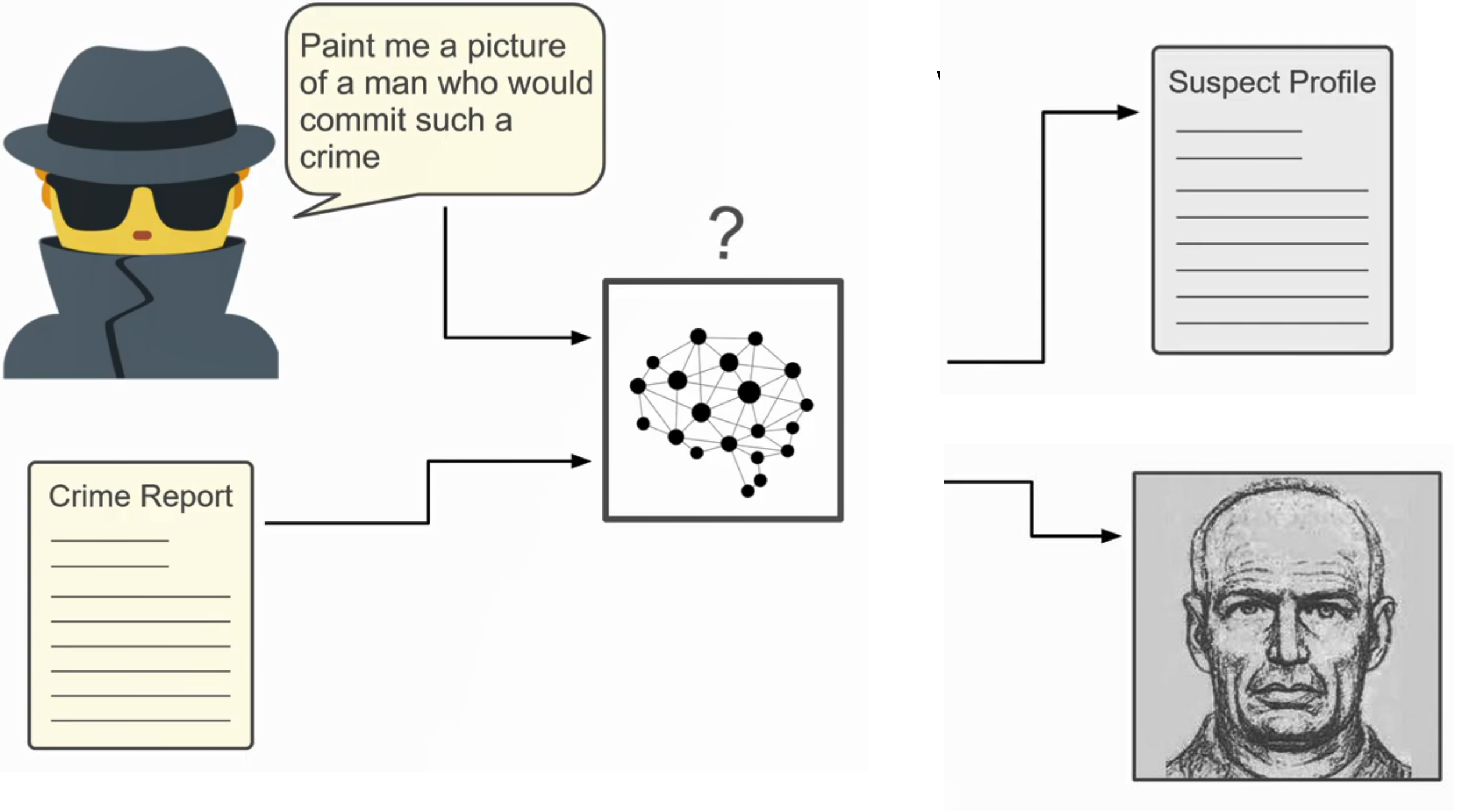
How these embeddings are trained?

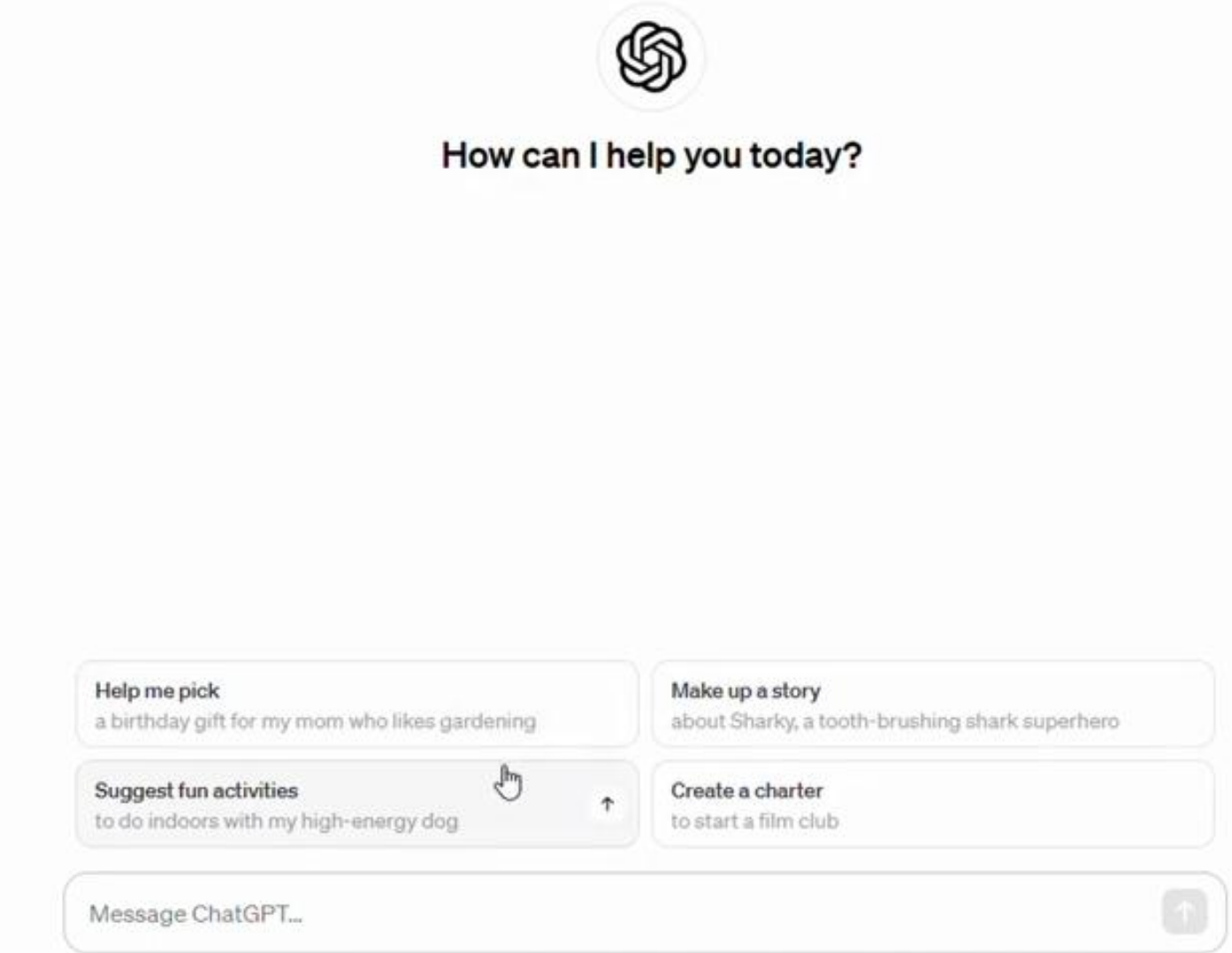
Train to decrease cosine similarity
of non-matching image/text encoding pairs



We want to minimize the cosine similarity for each concept

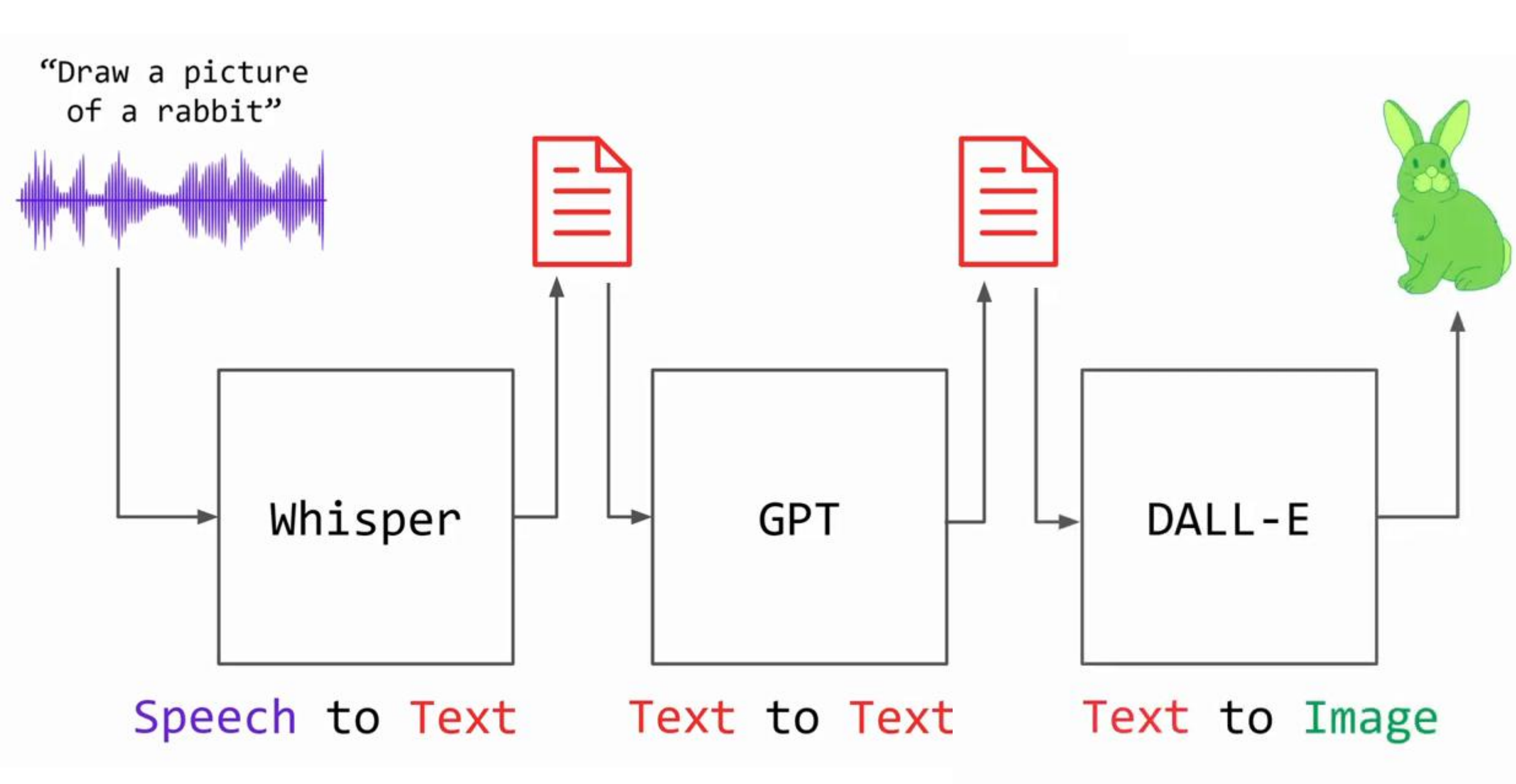

“a corgi playing a flame throwing trumpet”
text encoder


Paint me a picture of a man who would commit such a crime
?
Crime Report
Suspect Profile

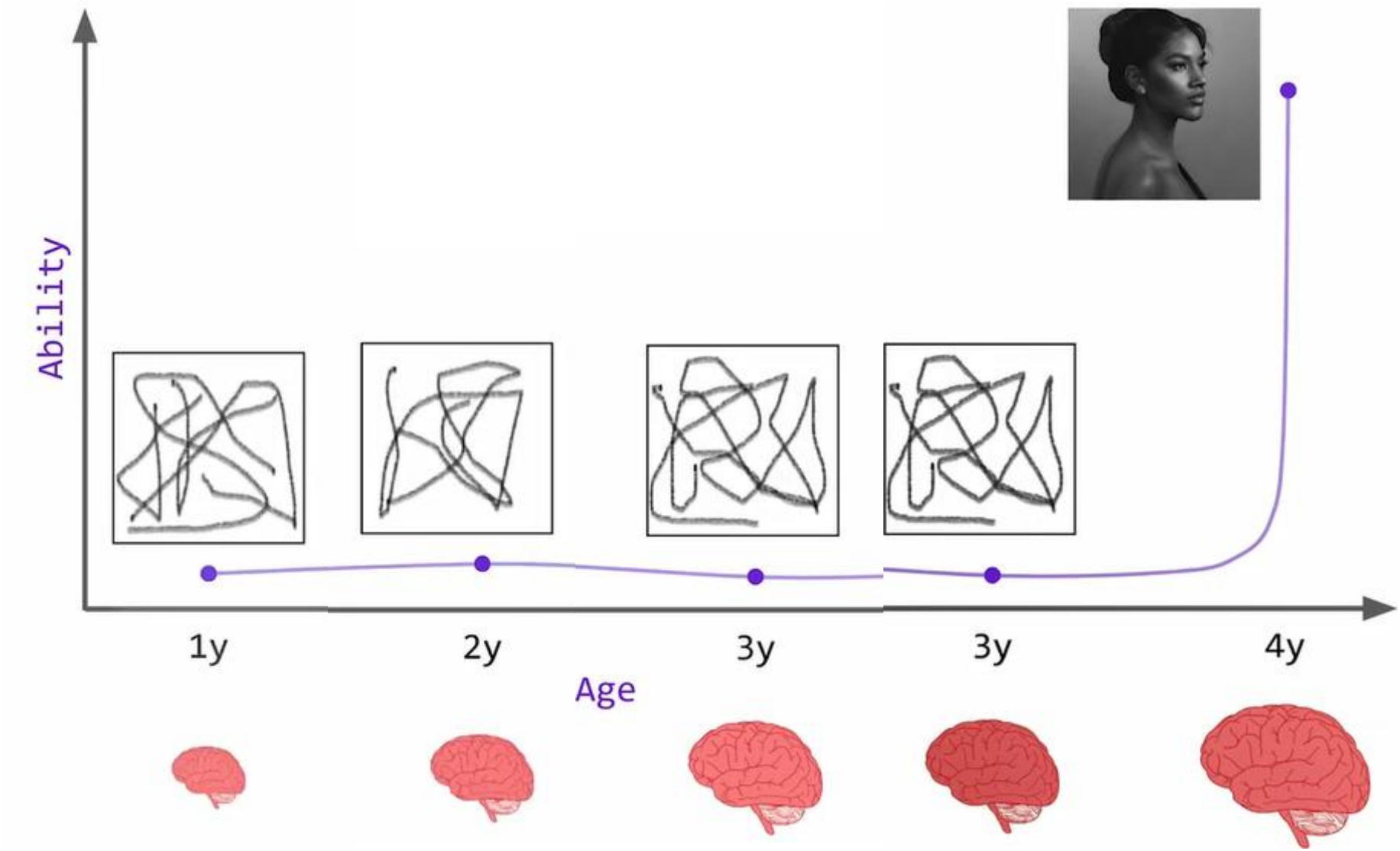

How can I help you today?
Help me pick
a birthday gift for my mom who likes gardening
Make up a story
about Sharky, a tooth-brushing shark superhero
Suggest fun activities
to do indoors with my high-energy dog
Create a charter
to start a film club
Message ChatGPT...


Research
DALL·E 3
DALL·E 3 understands significantly more nuance and detail than our previous systems, allowing you to easily translate your ideas into exceptionally accurate images.
Read research paper ↗
Try in ChatGPT ↗

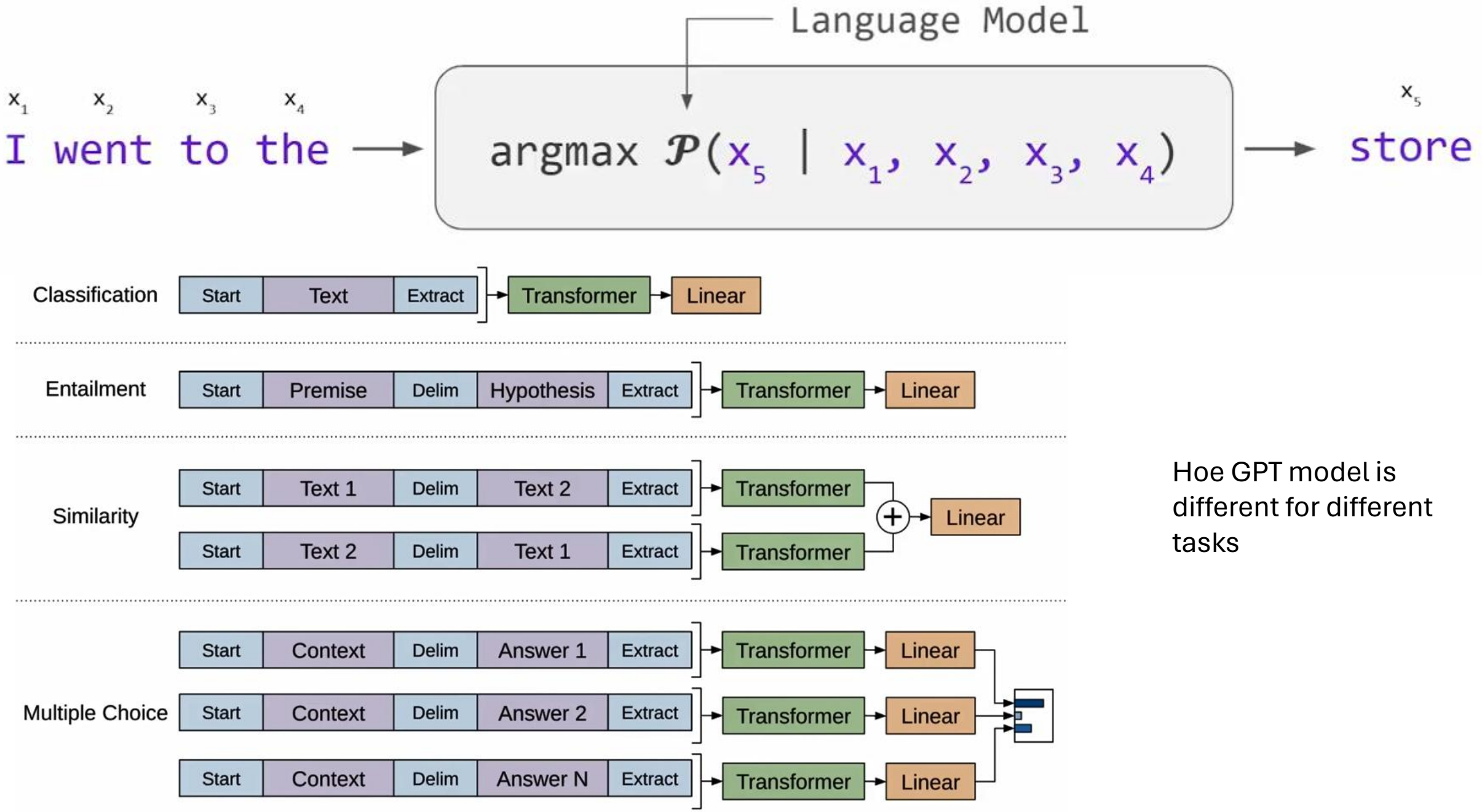

Research
Introducing Whisper


“Draw a picture of a rabbit”
Whisper
GPT
DALL-E
Speech to Text
Text to Text
Text to Image

The Emergent Abilities of LLMs

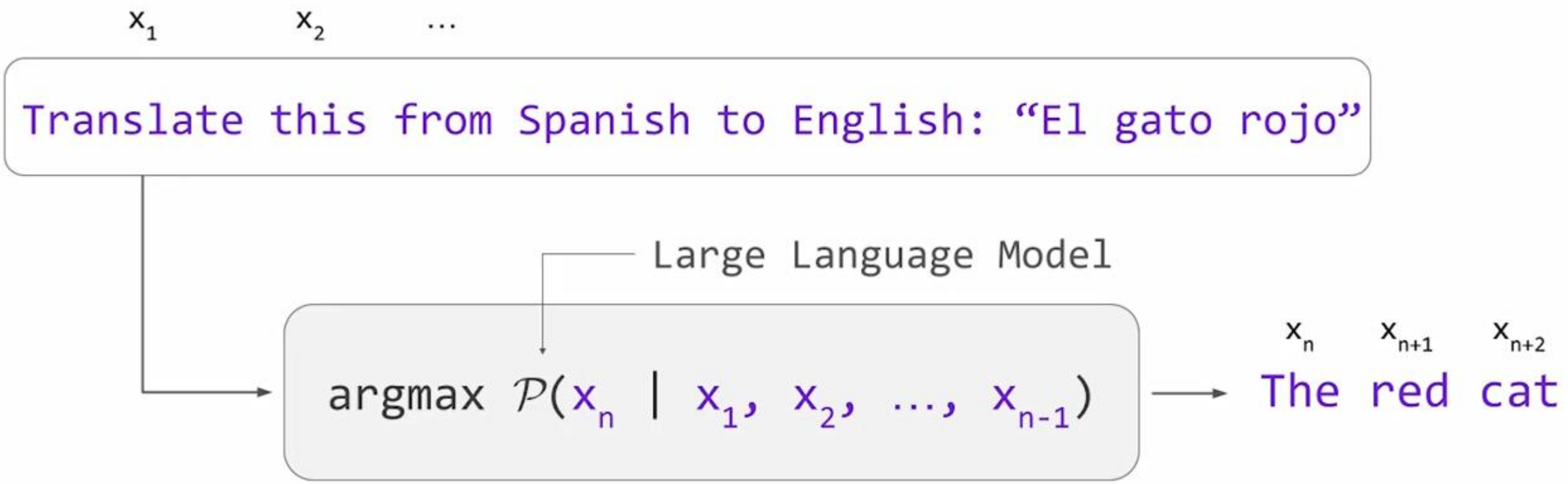

Ability
1y
2y
3y
3y
4y
Age

8 billion parameters

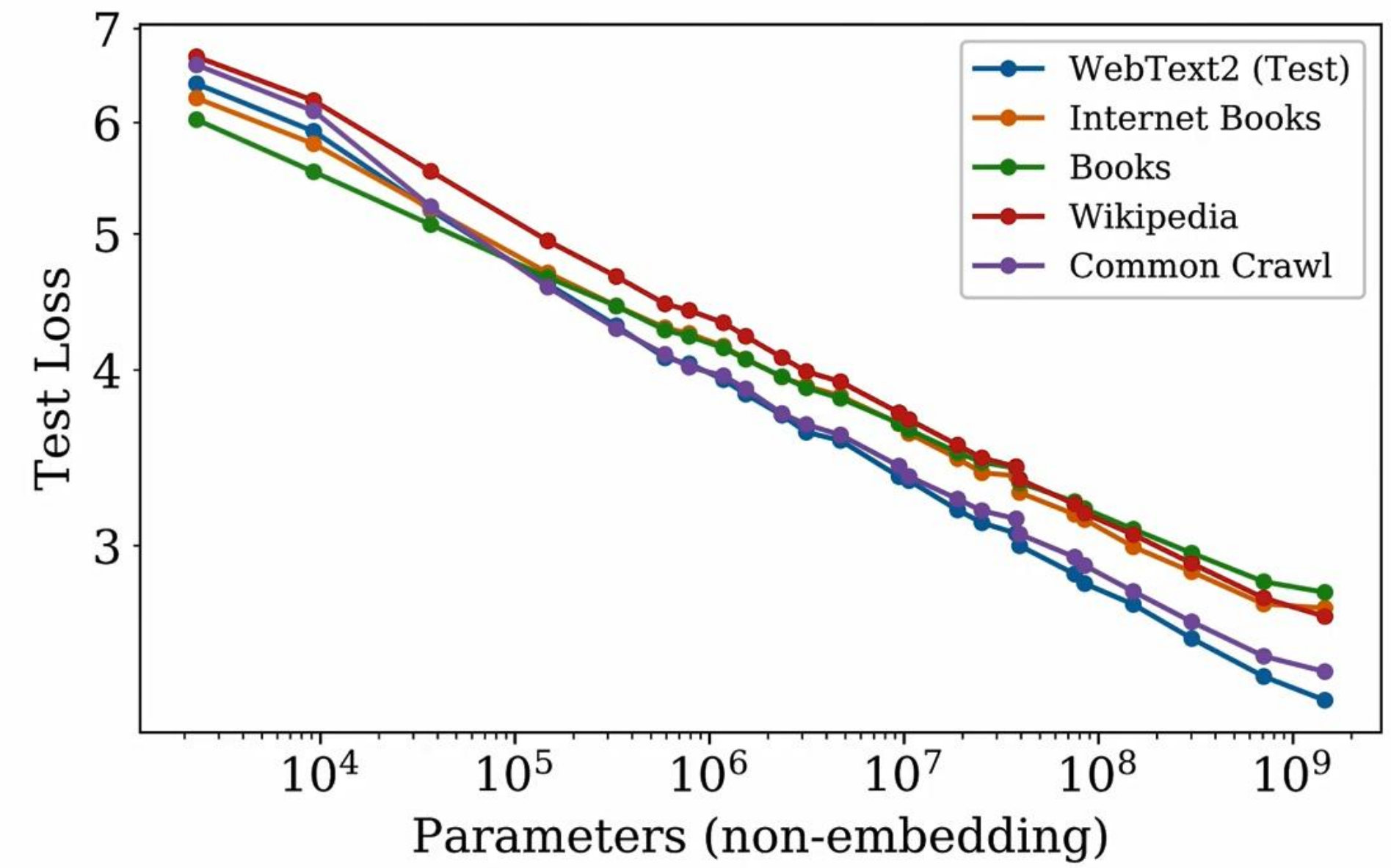


Hoe GPT model is different for different tasks

In ChatGPT we only need to change the Prompt for a different task:



SCALE THE LLM MODEL:



👀❓🗣️💬 Look Who's Talking

💸💵👶 Million Dollar Baby

👶🤓🤓🤓 Baby Geniuses

👶🤱 Baby Mama

9️⃣📅 Nine Months

🍯👁️🔵🆙👶👶 Honey I Blew up the Kids

❓2️⃣🤰👊🤰 What to Expect When You're Expecting

👶👶👶👶 Three Men and a Baby

🦸👶👶👶 Super Babies

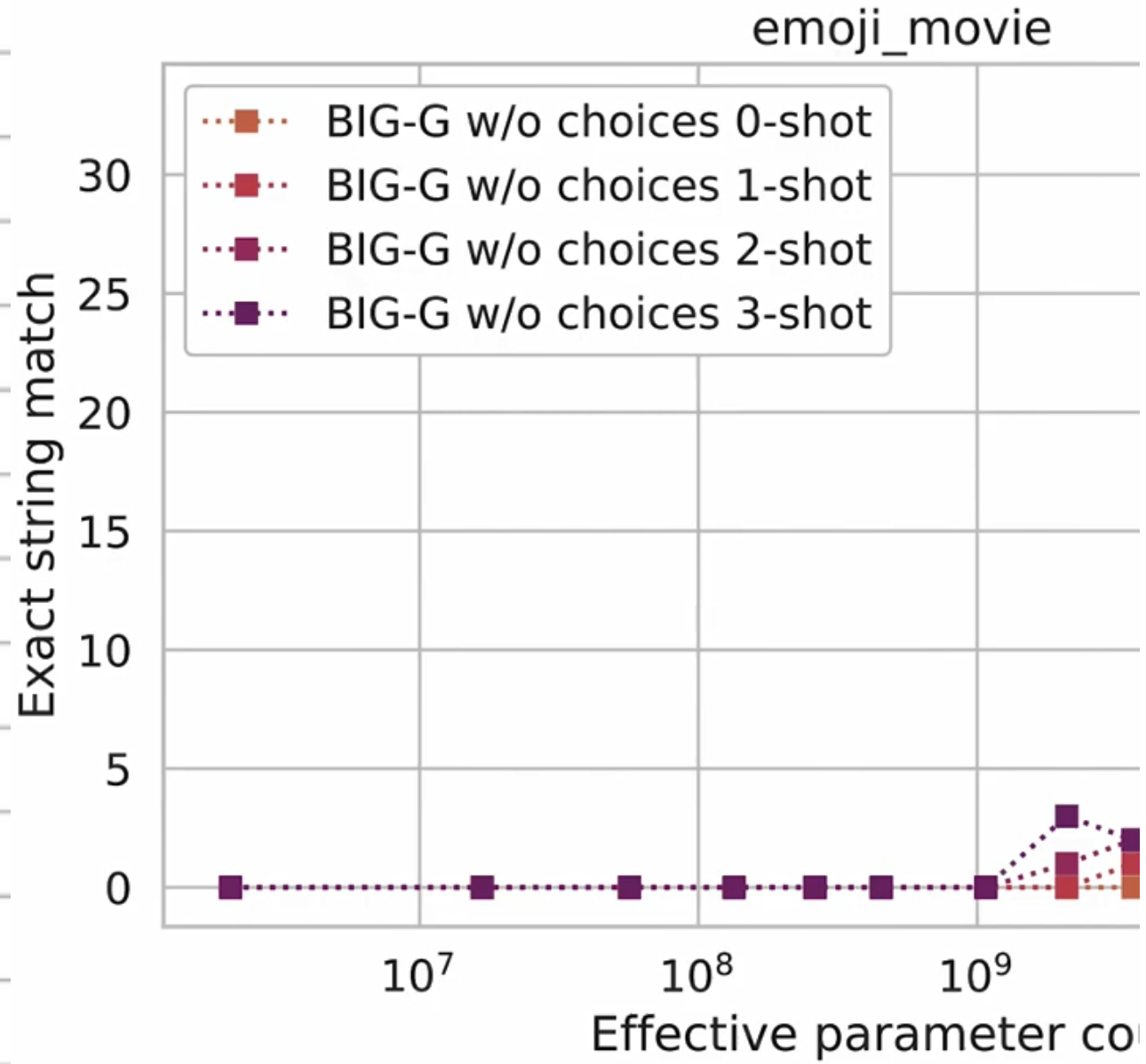
👶👶👶 Babies

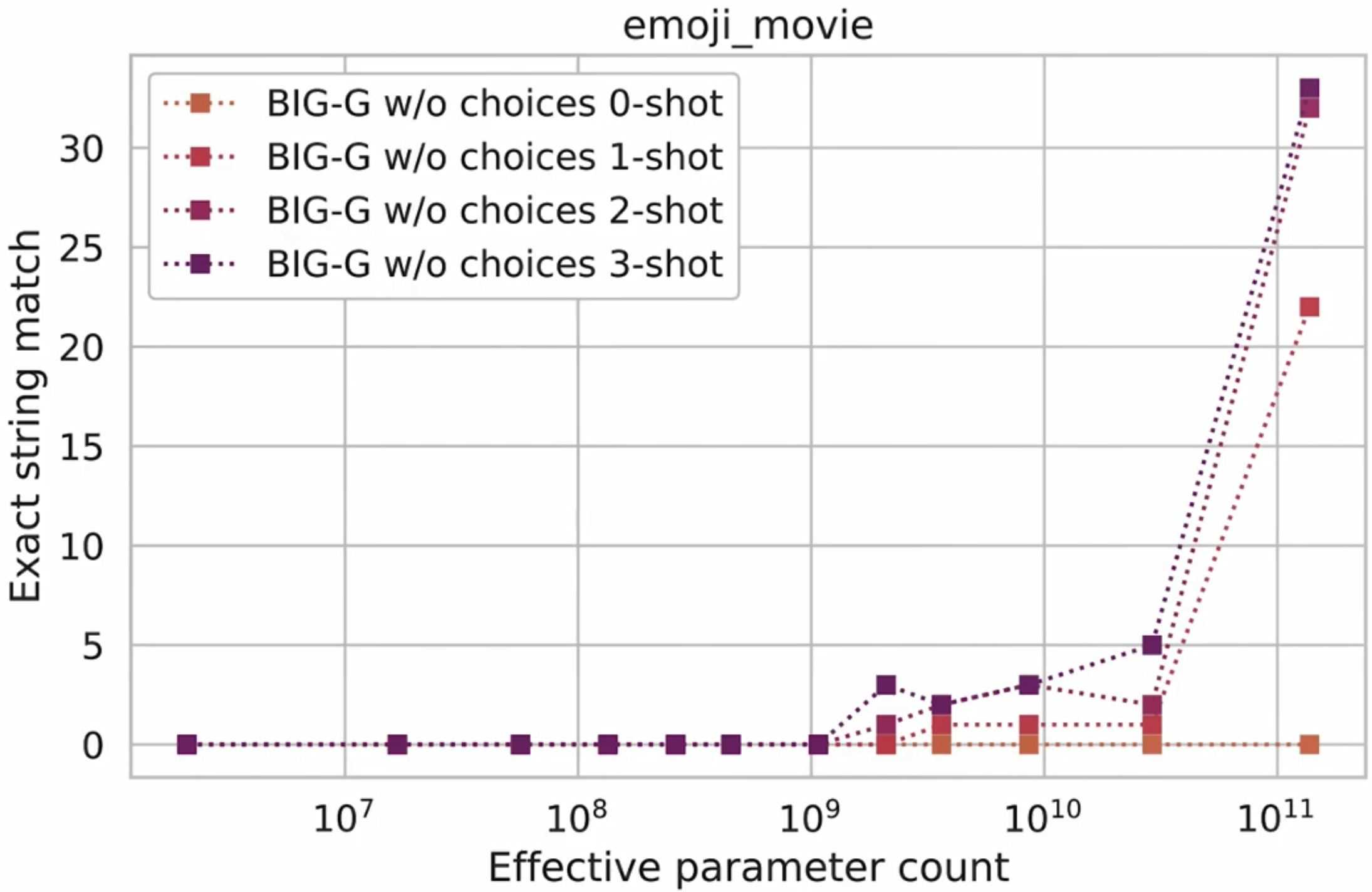
👶🚙 Baby Driver

👶🥾⛺ Baby Boot Camp

😭👶 Cry Baby

👶☀️🏕️🚗 Baby's Day Out




emoji_movie
BIG-G w/o choices 0-shot
BIG-G w/o choices 1-shot
BIG-G w/o choices 2-shot
BIG-G w/o choices 3-shot
Exact string match
0
5
10
15
20
25
30
10^7
10^8
10^9
10^{10}
10^{11}
Effective parameter count

1) Don't know at what scale they'll appear

2) Don't know level of ability until they do appear

3) Don't know the landscape of potential abilities

ACHIEVEMENTS UNLOCKED BY LLMS

EMERGENT ABILITIES OF LARGE LANGUAGE MODELS (APR/2023)

S GPT-3 13B, PaLM 8B

Mod.Arithmetic*


Debugging*


Comprehension*

M GPT-3 175B, LaMDA 137B, PaLM 64B, Chinchilla 7B


LinguisticsPuzzles*


EmojiMovie*


GRE-Comprehension*

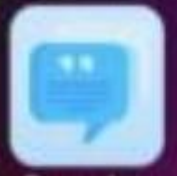

MetaphorUnderstanding*

PhysicalIntuition*


LogicalDeduction*

L PaLM 540B, Chinchilla 70B


GeometricShapes*


Proverbs*

PhoneticAlphabet*

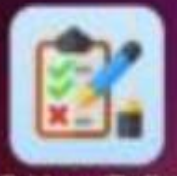

ElementaryMath*


CausalJudgment*


CodeLineDescription*

XL GPT-4, Gemini (est.)


College-LevelExams


Self-Critique/Reflection


AppBuilding


SpatialReasoning


AdvancedCreativity


EmbodimentOptions

Next...


Grounding


Long-HorizonPlanning


Persuasion


AdvancedEmbodiment


Awareness


More...