Senato della Repubblica

Data contained in this document constitute proprietary and privileged information belonging to C.A. Carnevale Maffè and to the addressee of the document. This information is strictly connected to the oral information that accompanied it, and it can only be used by the persons who attended the presentation. Copying, publishing and distributing the material contained in this document is forbidden and it may be illegal.

# Generative AI: institutional implications

*1 June 2023*

Prof. Carlo Alberto Carnevale Maffè
*Bocconi University School of Management*

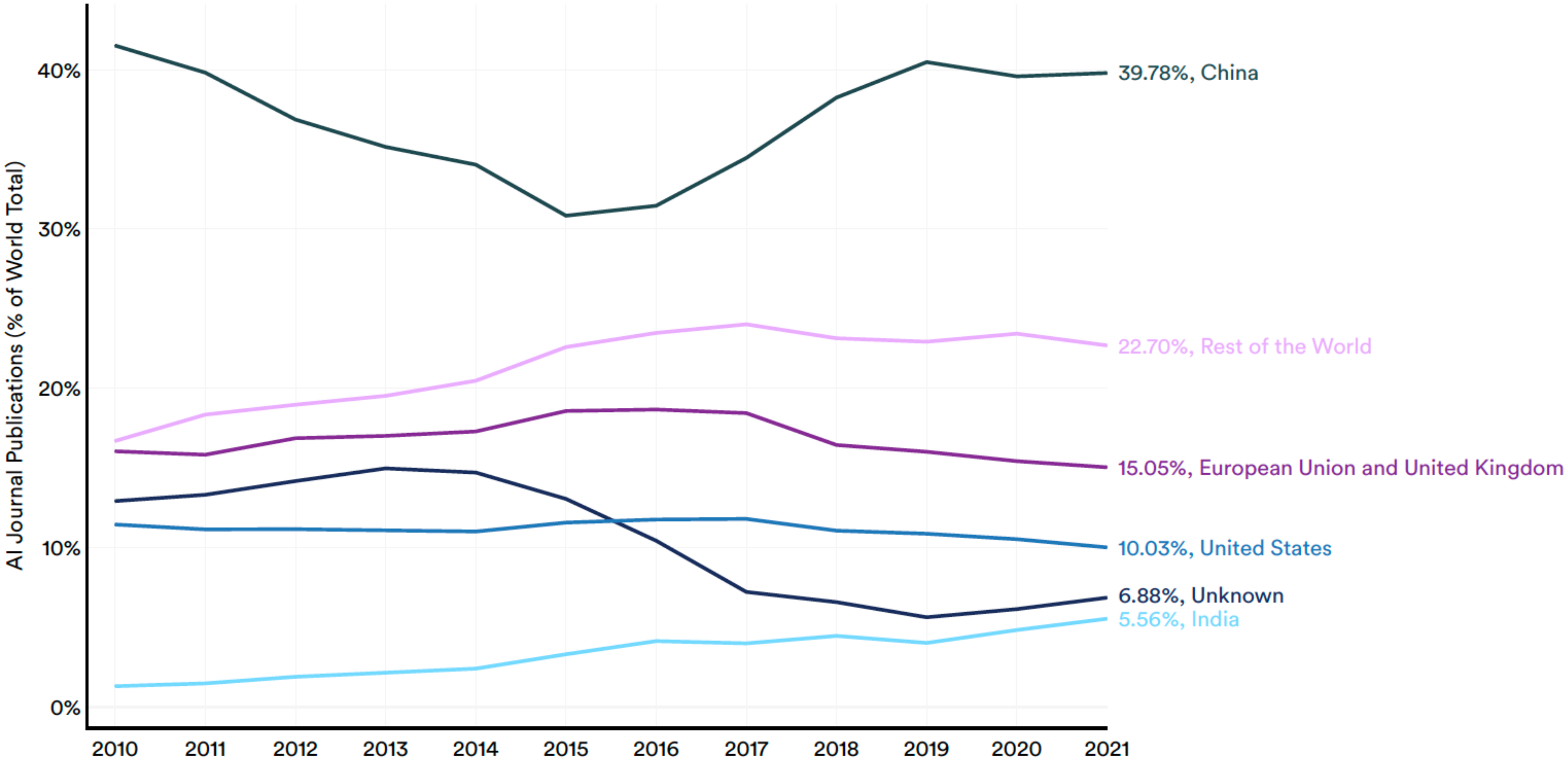
*«AI won't replace human professionals, but those who use it will replace those who don’t»*

**Artur Niemczewski**, Insurance NED, CEO

# What Italy and European Union should do about AI

- Home of global «Open Source AI» community
- Distributed computing architecture as opposed to hyperscalers
- Self-sovereign Identity Framework with OPAL model
- Specific data spaces and knowledge base for «social applications» oriented to European public goods (health care, education, mobility, journalism, etc.)
- AI for Open Banking and Financial Stability

# AI Journal Publications: China has remained the leader throughout, with 39.8% in 2021, followed by the European Union and UK (15.1%), then US (10.0%).



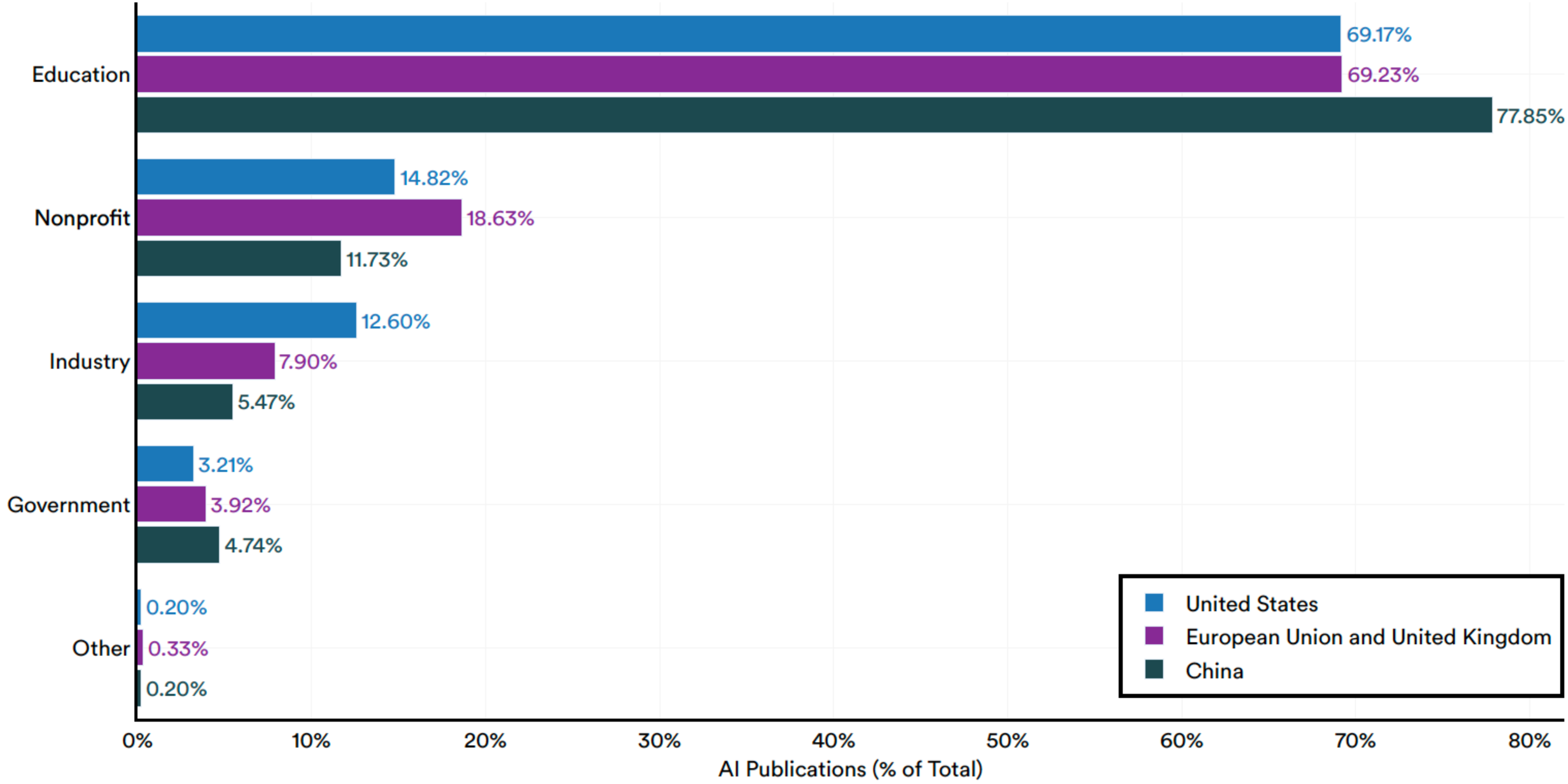
AI Journal Publications (% of World Total) by Geographic Area, 2010–21

Source: Center for Security and Emerging Technology, 2022 | Chart: 2023 AI Index Report

*Source: Stanford University, 2023*

## AI Publications (% of Total) by Sector and Geographic Area, 2021

Source: Center for Security and Emerging Technology, 2022 | Chart: 2023 AI Index Report



*Source: Stanford University, 2023*

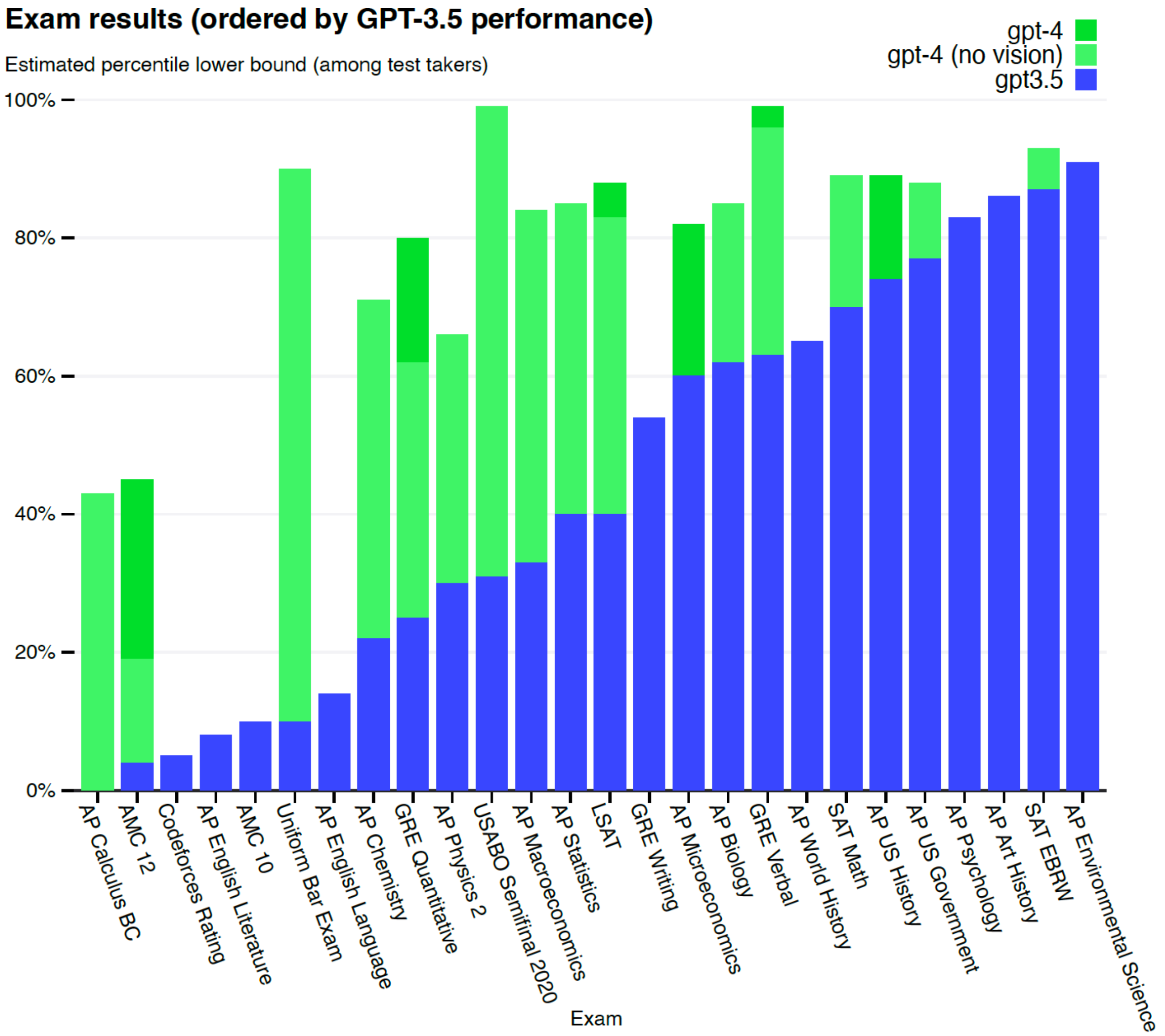
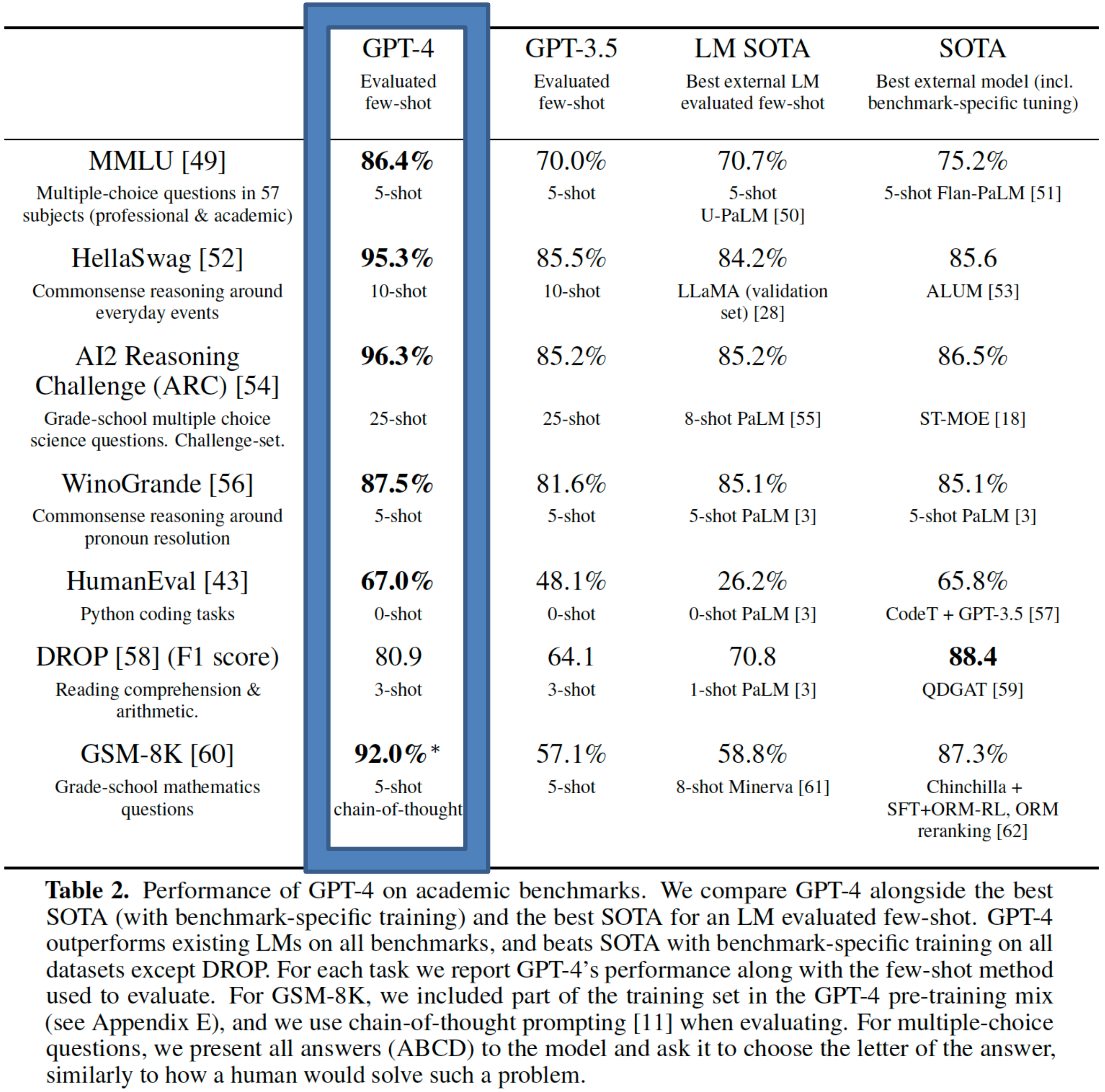


**Figure 4.** GPT performance on academic and professional exams. In each case, we simulate the conditions and scoring of the real exam. Exams are ordered from low to high based on GPT-3.5 performance. GPT-4 outperforms GPT-3.5 on most exams tested. To be conservative we report the lower end of the range of percentiles, but this creates some artifacts on the AP exams which have very wide scoring bins. For example although GPT-4 attains the highest possible score on AP Biology (5/5), this is only shown in the plot as 85th percentile because 15 percent of test-takers achieve that score.

# GPT4 performance on academic and professional exams

*Source: Stanford University, 2023*

# GPT4 performance on academic benchmarks



| | GPT-4<br>Evaluated few-shot | GPT-3.5<br>Evaluated few-shot | LM SOTA<br>Best external LM evaluated few-shot | SOTA<br>Best external model (incl. benchmark-specific tuning) |
|---|---|---|---|---|
| MMLU [49]<br>Multiple-choice questions in 57 subjects (professional & academic) | **86.4%**<br>5-shot | 70.0%<br>5-shot | 70.7%<br>5-shot U-PaLM [50] | 75.2%<br>5-shot Flan-PaLM [51] |
| HellaSwag [52]<br>Commonsense reasoning around everyday events | **95.3%**<br>10-shot | 85.5%<br>10-shot | 84.2%<br>LLaMA (validation set) [28] | 85.6<br>ALUM [53] |
| AI2 Reasoning Challenge (ARC) [54]<br>Grade-school multiple choice science questions. Challenge-set. | **96.3%**<br>25-shot | 85.2%<br>25-shot | 85.2%<br>8-shot PaLM [55] | 86.5%<br>ST-MOE [18] |
| WinoGrande [56]<br>Commonsense reasoning around pronoun resolution | **87.5%**<br>5-shot | 81.6%<br>5-shot | 85.1%<br>5-shot PaLM [3] | 85.1%<br>5-shot PaLM [3] |
| HumanEval [43]<br>Python coding tasks | **67.0%**<br>0-shot | 48.1%<br>0-shot | 26.2%<br>0-shot PaLM [3] | 65.8%<br>CodeT + GPT-3.5 [57] |
| DROP [58] (F1 score)<br>Reading comprehension & arithmetic. | 80.9<br>3-shot | 64.1<br>3-shot | 70.8<br>1-shot PaLM [3] | **88.4**<br>QDGAT [59] |
| GSM-8K [60]<br>Grade-school mathematics questions | **92.0%***<br>5-shot chain-of-thought | 57.1%<br>5-shot | 58.8%<br>8-shot Minerva [61] | 87.3%<br>Chinchilla + SFT+ORM-RL, ORM reranking [62] |

**Table 2.** Performance of GPT-4 on academic benchmarks. We compare GPT-4 alongside the best SOTA (with benchmark-specific training) and the best SOTA for an LM evaluated few-shot. GPT-4 outperforms existing LMs on all benchmarks, and beats SOTA with benchmark-specific training on all datasets except DROP. For each task we report GPT-4's performance along with the few-shot method used to evaluate. For GSM-8K, we included part of the training set in the GPT-4 pre-training mix (see Appendix E), and we use chain-of-thought prompting [11] when evaluating. For multiple-choice questions, we present all answers (ABCD) to the model and ask it to choose the letter of the answer, similarly to how a human would solve such a problem.

*Source: Stanford University, 2023*

# Performance of GPT4 in a variety of languages

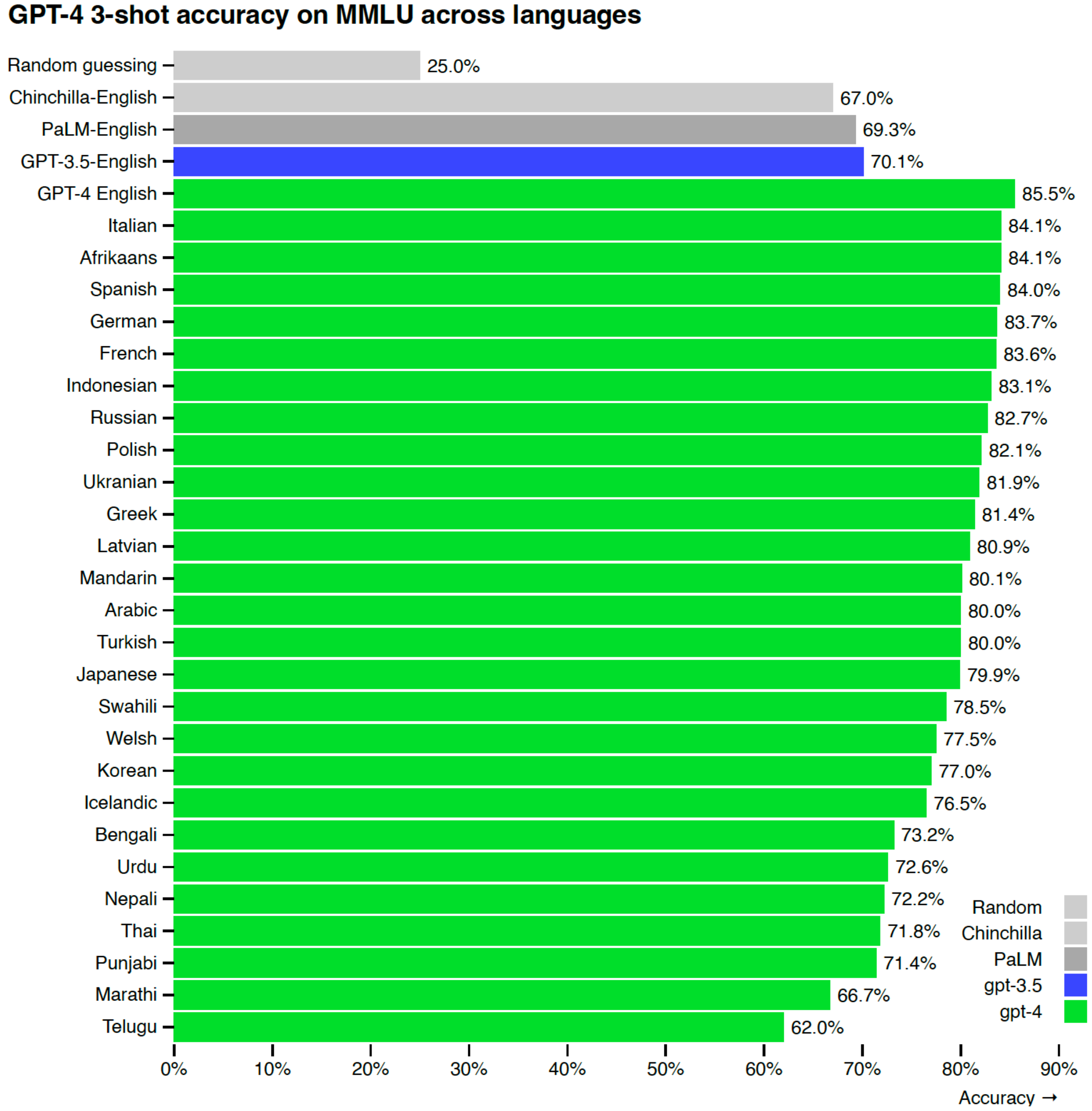


**Figure 5.** Performance of GPT-4 in a variety of languages compared to prior models in English on MMLU. GPT-4 outperforms the English-language performance of existing language models [2, 3] for the vast majority of languages tested, including low-resource languages such as Latvian, Welsh, and Swahili.

*Source: Stanford University, 2023*

# Evolution of GPT-x performance on factuality evaluations

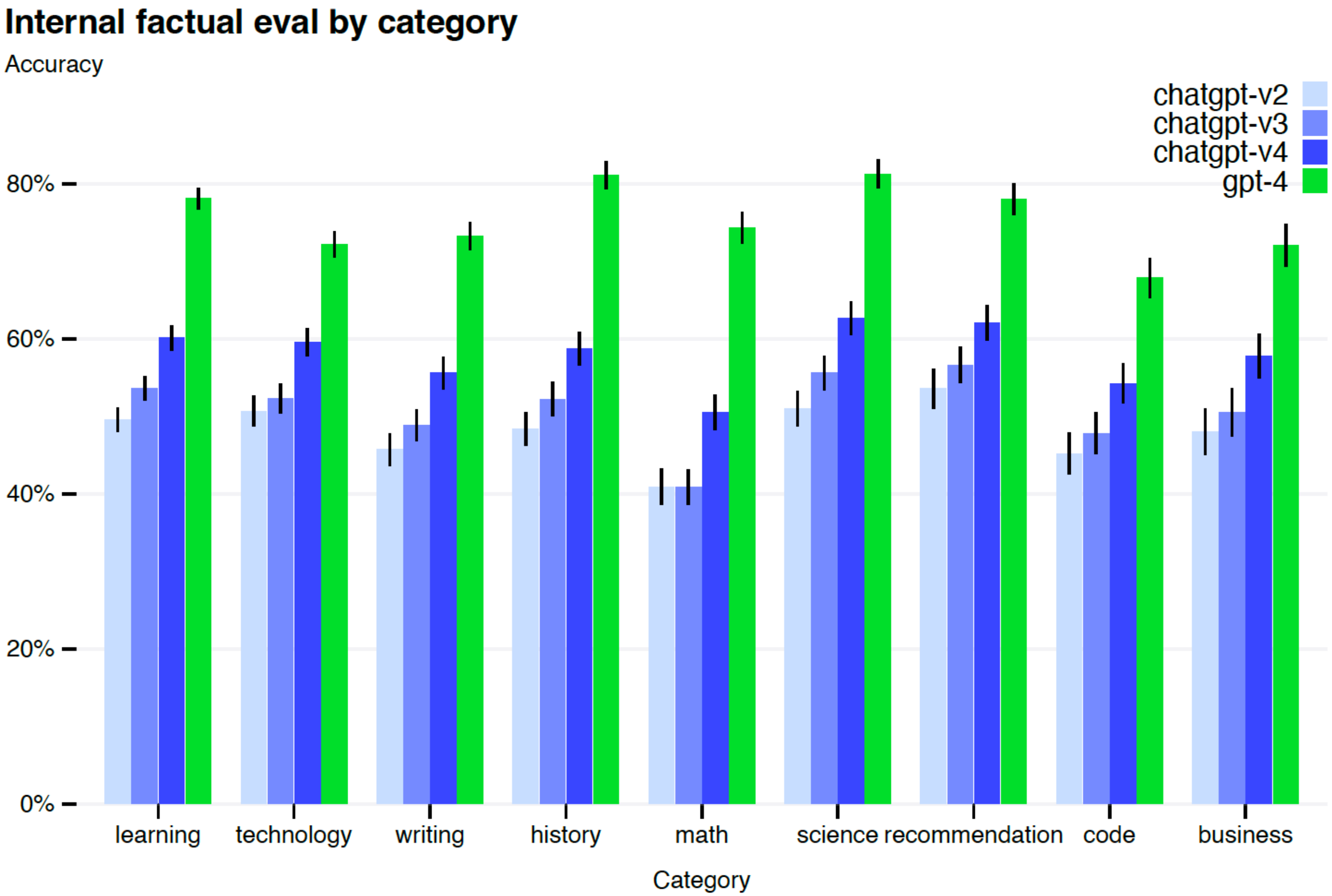
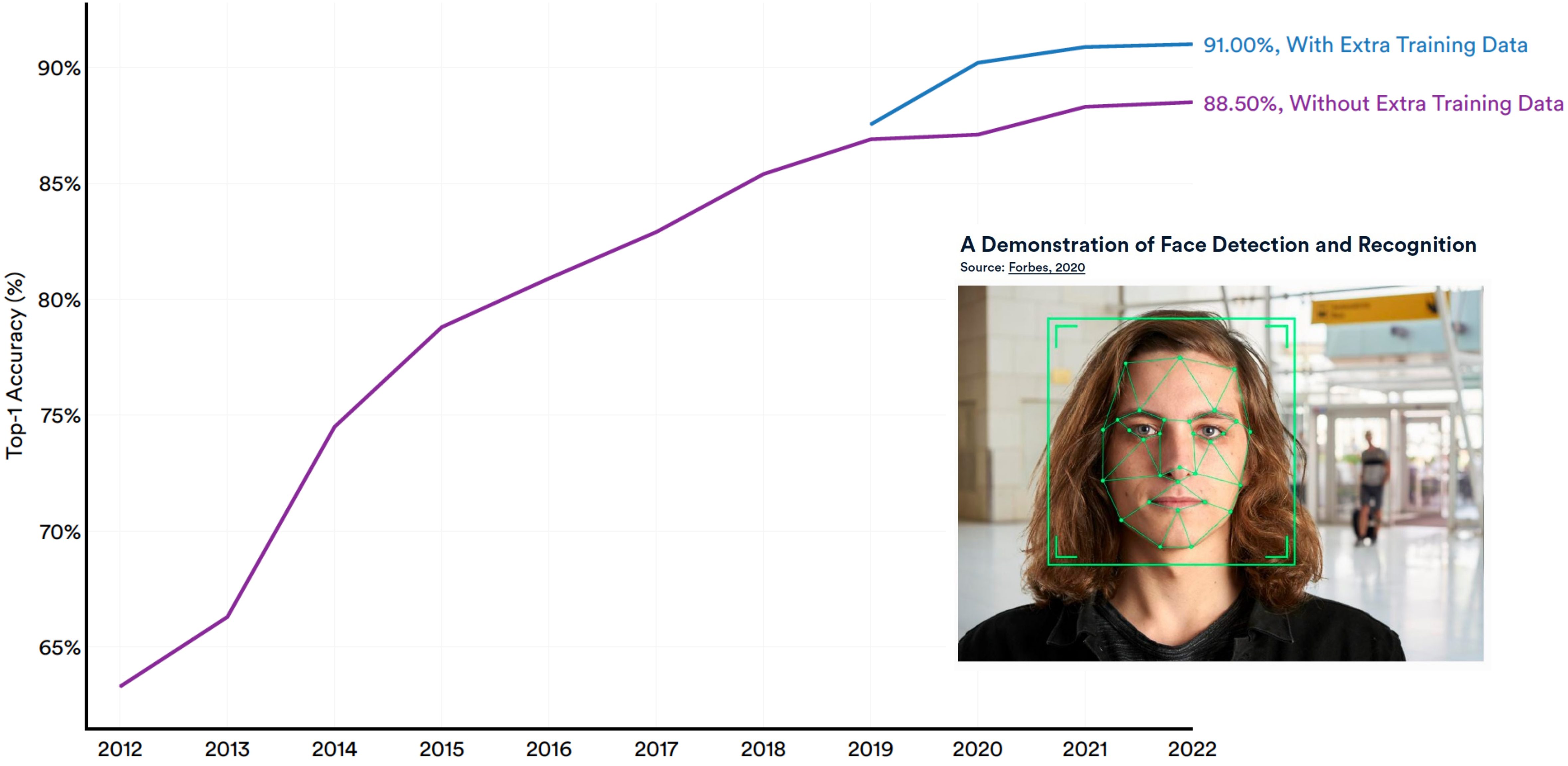


**Figure 6.** Performance of GPT-4 on nine internal adversarially-designed factuality evaluations. Accuracy is shown on the y-axis, higher is better. An accuracy of 1.0 means the model's answers are judged to be in agreement with human ideal responses for all questions in the eval. We compare GPT-4 to three earlier versions of ChatGPT [64] based on GPT-3.5; GPT-4 improves on the latest GPT-3.5 model by 19 percentage points, with significant gains across all topics.

*Source: Stanford University, 2023*

# ImageNet Challenge: Top-1 Accuracy

Source: Papers With Code, 2022; arXiv, 2022 | Chart: 2023 AI Index Report



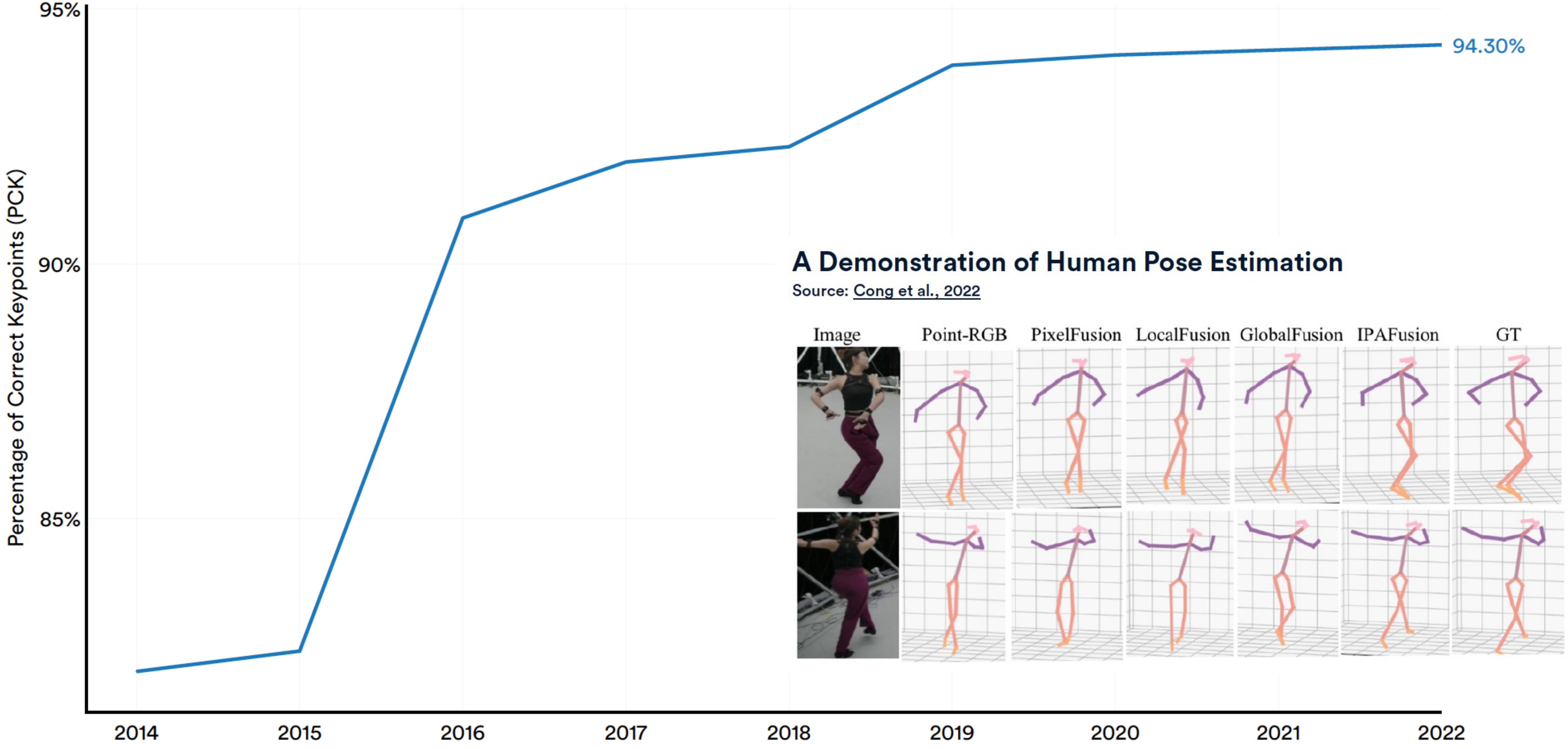
## A Demonstration of Face Detection and Recognition

Source: Forbes, 2020

*Source: Stanford University, 2023*

## MPII: Percentage of Correct Keypoints (PCK)

Source: Papers With Code, 2022; arXiv, 2022 | Chart: 2023 AI Index Report



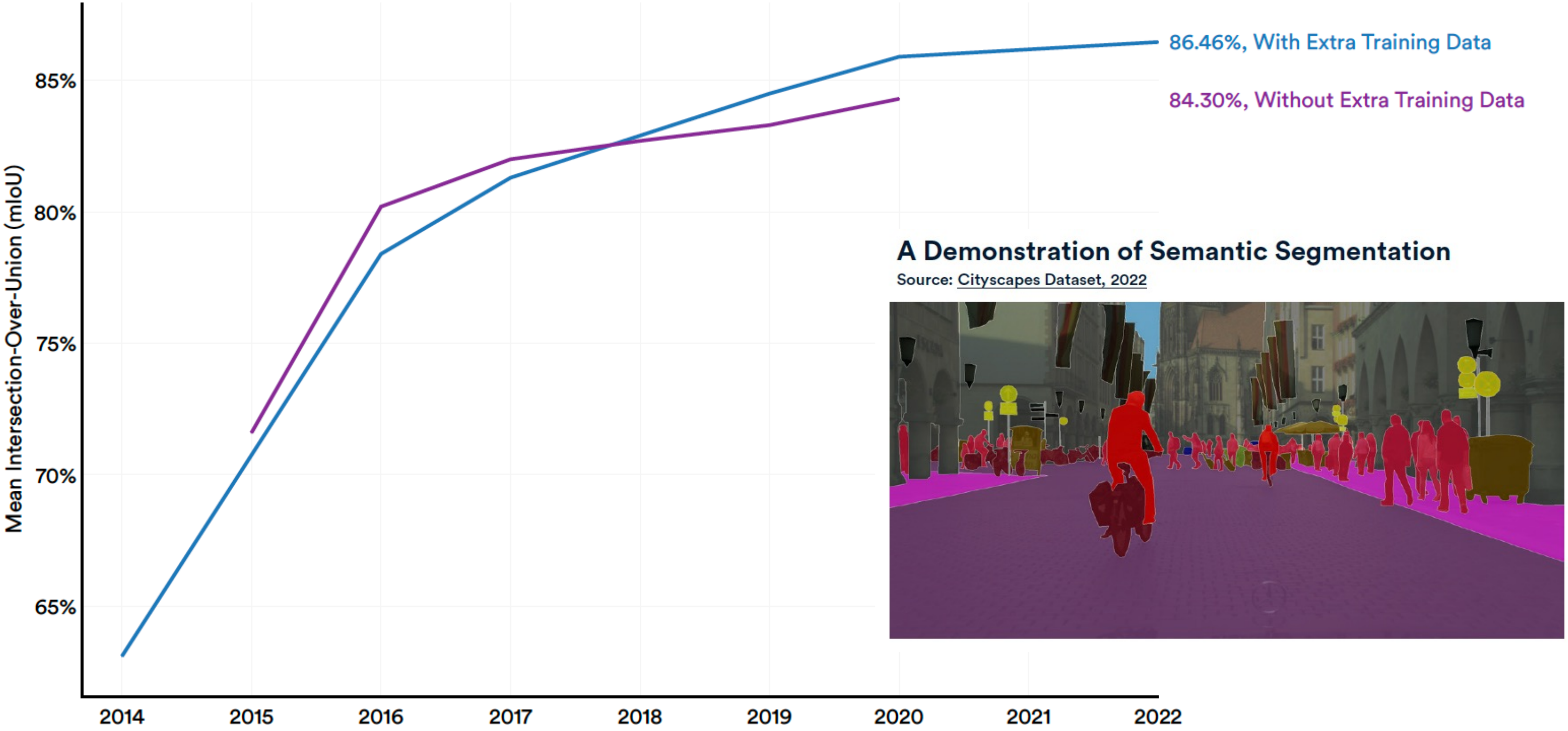
### A Demonstration of Human Pose Estimation

Source: Cong et al., 2022

*Source: Stanford University, 2023*

# Cityscapes Challenge, Pixel-Level Semantic Labeling Task: Mean Intersection-Over-Union (mIoU)

Source: Cityscapes Challenge, 2022 | Chart: 2023 AI Index Report



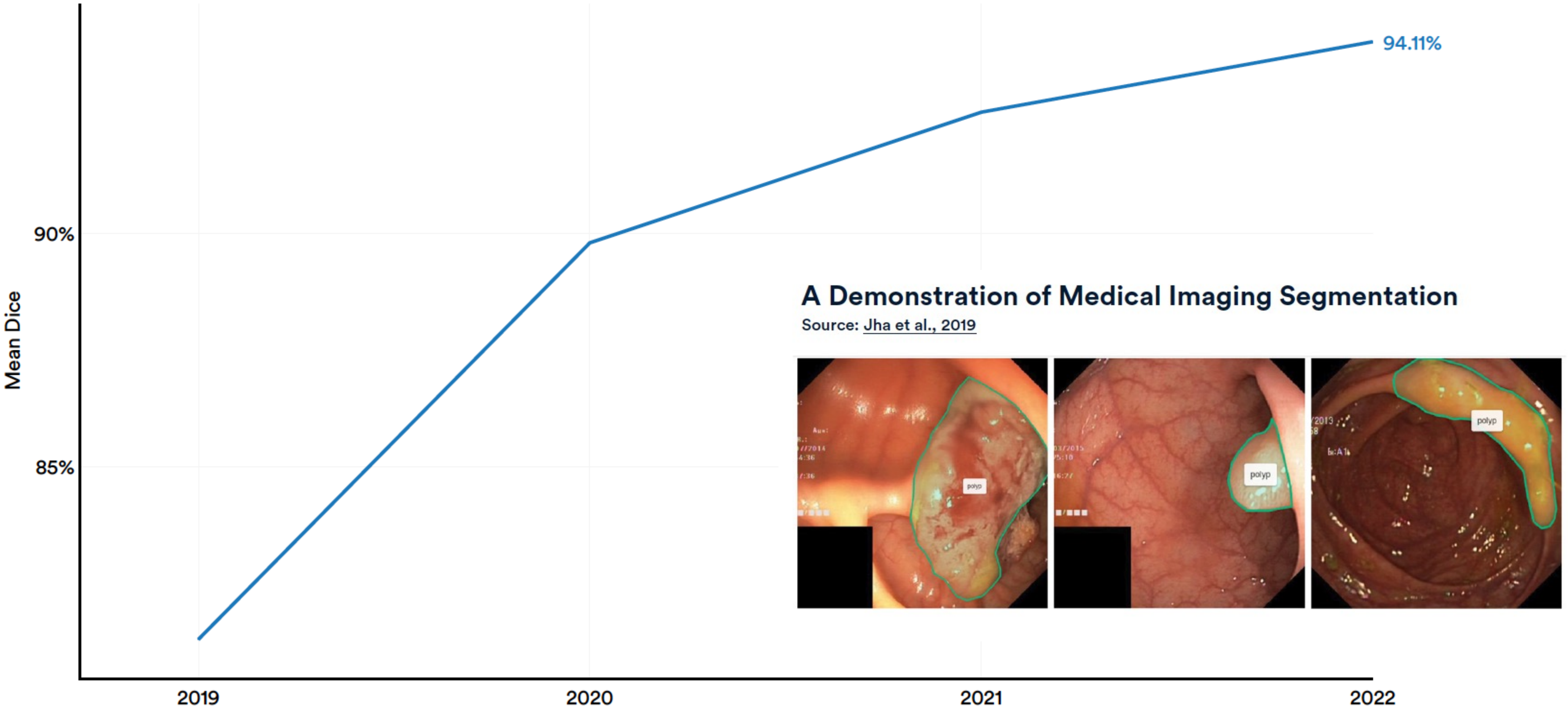
## A Demonstration of Semantic Segmentation

Source: Cityscapes Dataset, 2022

*Source: Stanford University, 2023*

## Kvasir-SEG: Mean Dice

Source: Papers With Code, 2022; arXiv, 2022 | Chart: 2023 AI Index Report



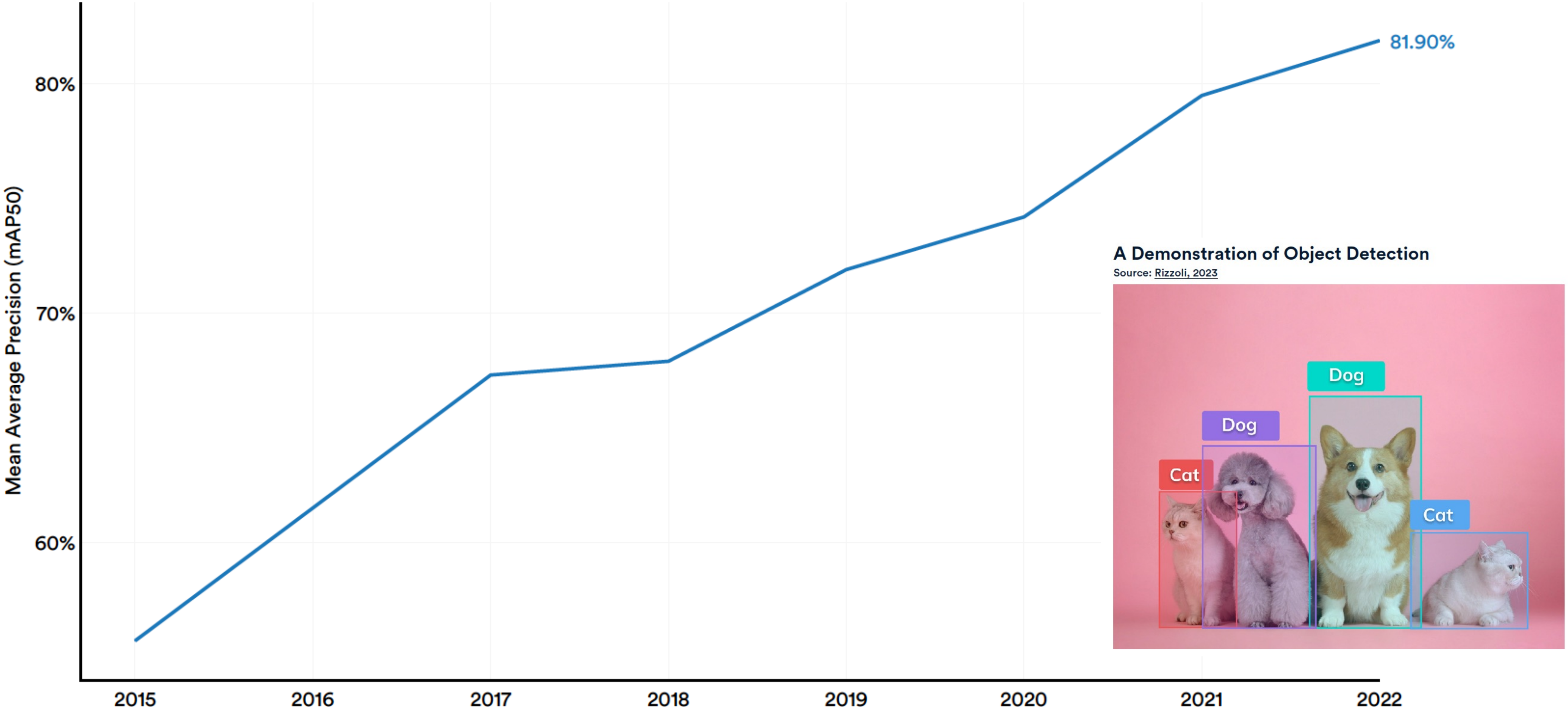
### A Demonstration of Medical Imaging Segmentation

Source: Jha et al., 2019

*Source: Stanford University, 2023*

# COCO: Mean Average Precision (mAP50)

Source: Papers With Code, 2022; arXiv, 2022 | Chart: 2023 AI Index Report



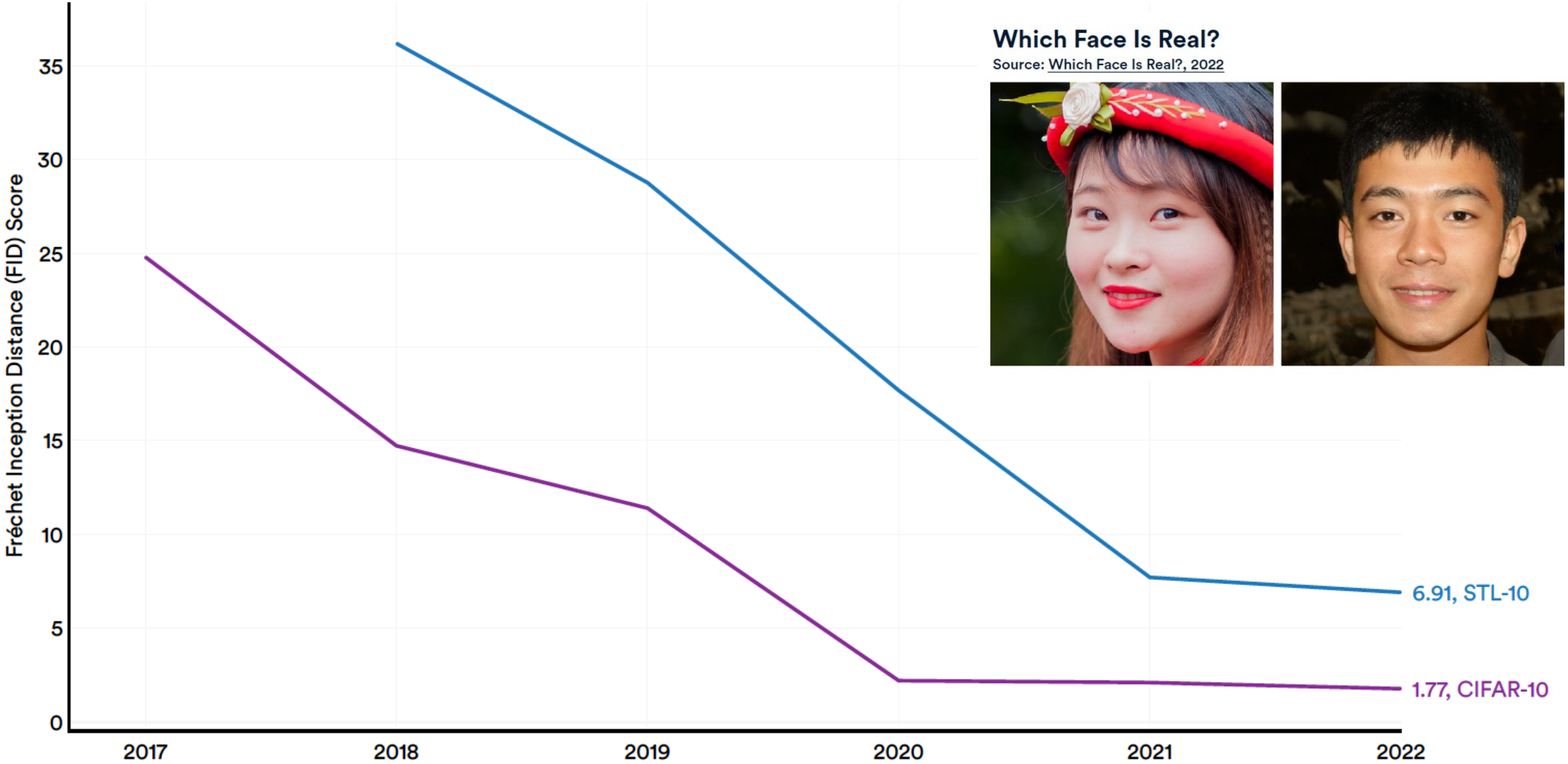
A Demonstration of Object Detection

Source: Rizzoli, 2023

*Source: Stanford University, 2023*

# CIFAR-10 and STL-10: Fréchet Inception Distance (FID) Score

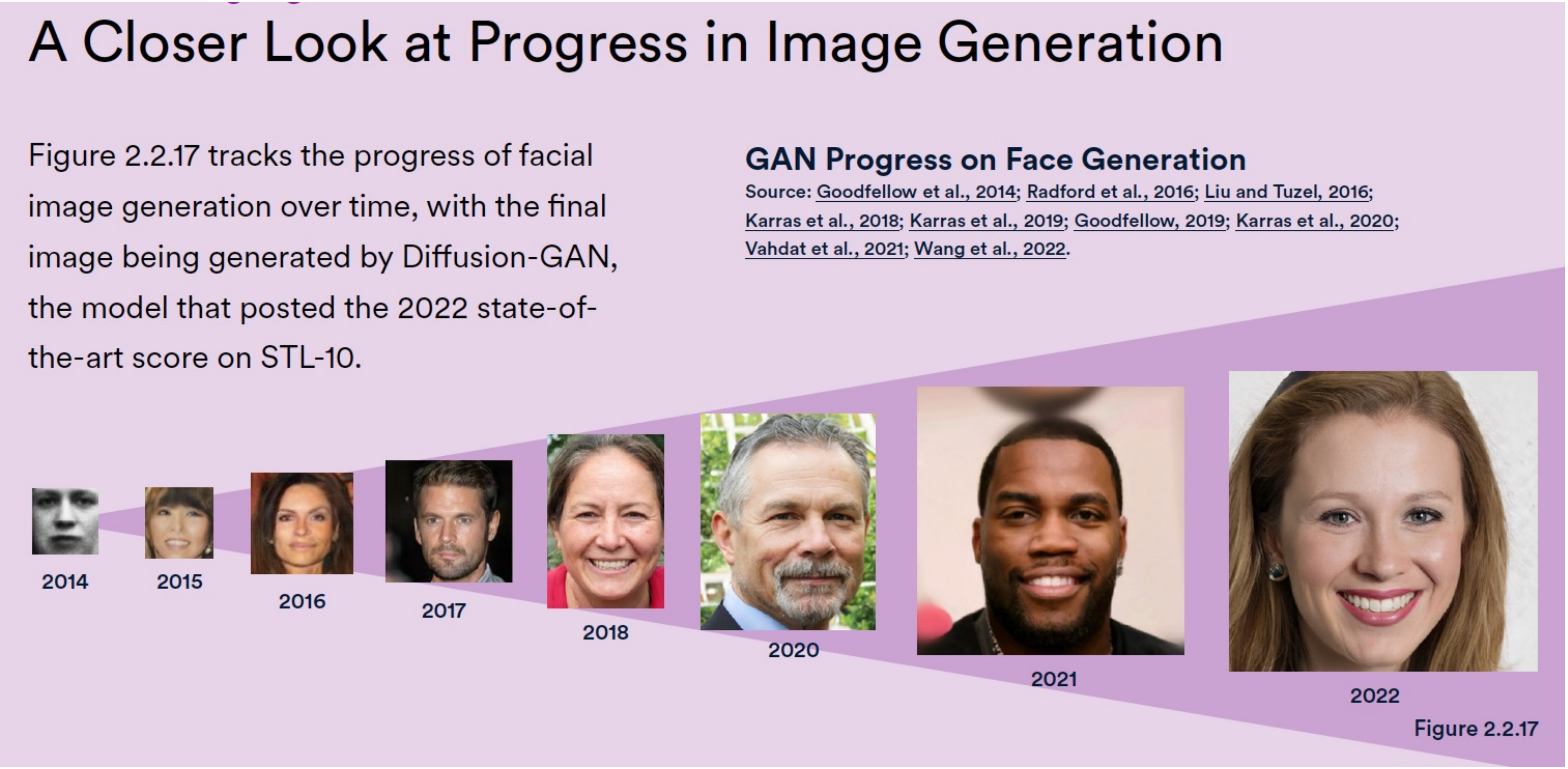
Source: Papers With Code, 2022; arXiv, 2022 | Chart: 2023 AI Index Report



**Which Face Is Real?**

Source: Which Face Is Real?, 2022

*Source: Stanford University, 2023*



# A Closer Look at Progress in Image Generation

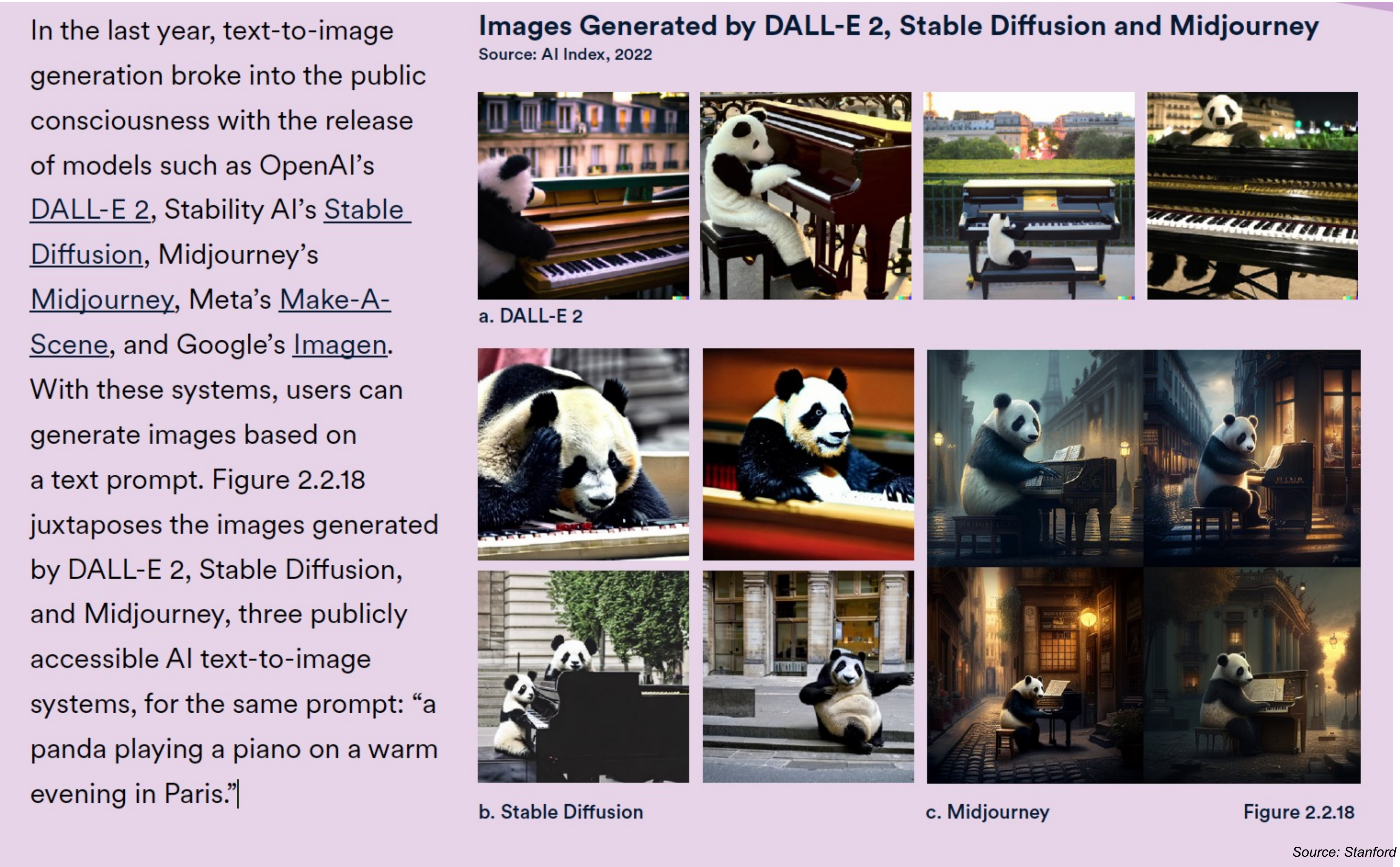
Figure 2.2.17 tracks the progress of facial image generation over time, with the final image being generated by Diffusion-GAN, the model that posted the 2022 state-of-the-art score on STL-10.

**GAN Progress on Face Generation**

Source: Goodfellow et al., 2014; Radford et al., 2016; Liu and Tuzel, 2016; Karras et al., 2018; Karras et al., 2019; Goodfellow, 2019; Karras et al., 2020; Vahdat et al., 2021; Wang et al., 2022.

Figure 2.2.17

*Source: Stanford University, 2023*



In the last year, text-to-image generation broke into the public consciousness with the release of models such as OpenAI's DALL-E 2, Stability AI's Stable Diffusion, Midjourney's Midjourney, Meta's Make-A-Scene, and Google's Imagen. With these systems, users can generate images based on a text prompt. Figure 2.2.18 juxtaposes the images generated by DALL-E 2, Stable Diffusion, and Midjourney, three publicly accessible AI text-to-image systems, for the same prompt: "a panda playing a piano on a warm evening in Paris."

**Images Generated by DALL-E 2, Stable Diffusion and Midjourney**

Source: AI Index, 2022

a. DALL-E 2

b. Stable Diffusion

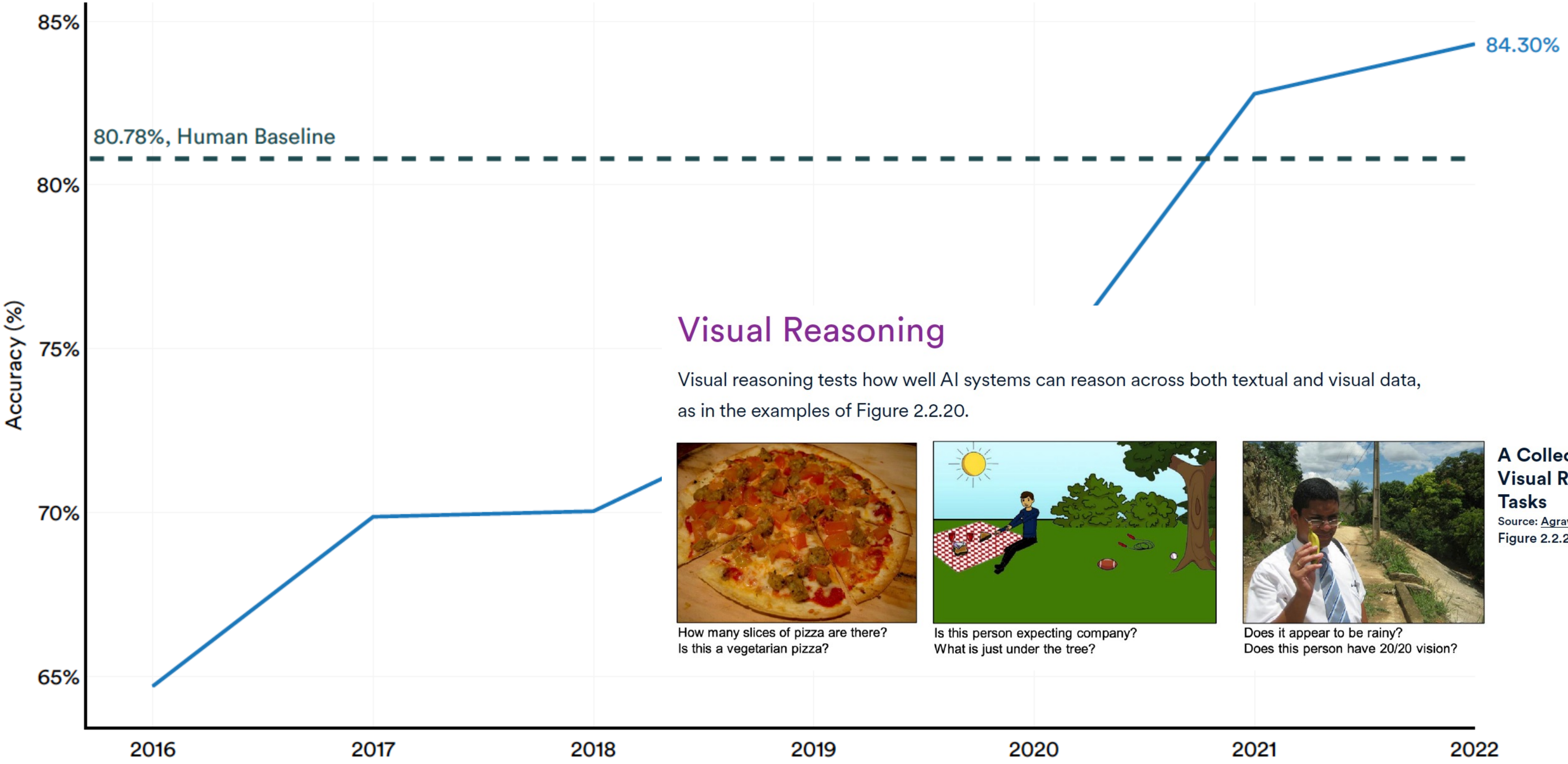
c. Midjourney

Figure 2.2.18

*Source: Stanford University, 2023*

## Visual Question Answering (VQA) V2 Test-Dev: Accuracy

Source: Papers With Code, 2022; arXiv, 2022 | Chart: 2023 AI Index Report



## Visual Reasoning

Visual reasoning tests how well AI systems can reason across both textual and visual data, as in the examples of Figure 2.2.20.

How many slices of pizza are there?
Is this a vegetarian pizza?

Is this person expecting company?
What is just under the tree?

Does it appear to be rainy?
Does this person have 20/20 vision?

**A Collection of Visual Reasoning Tasks**

Source: Agrawal et al., 2016

Figure 2.2.20

*Source: Stanford University, 2023*

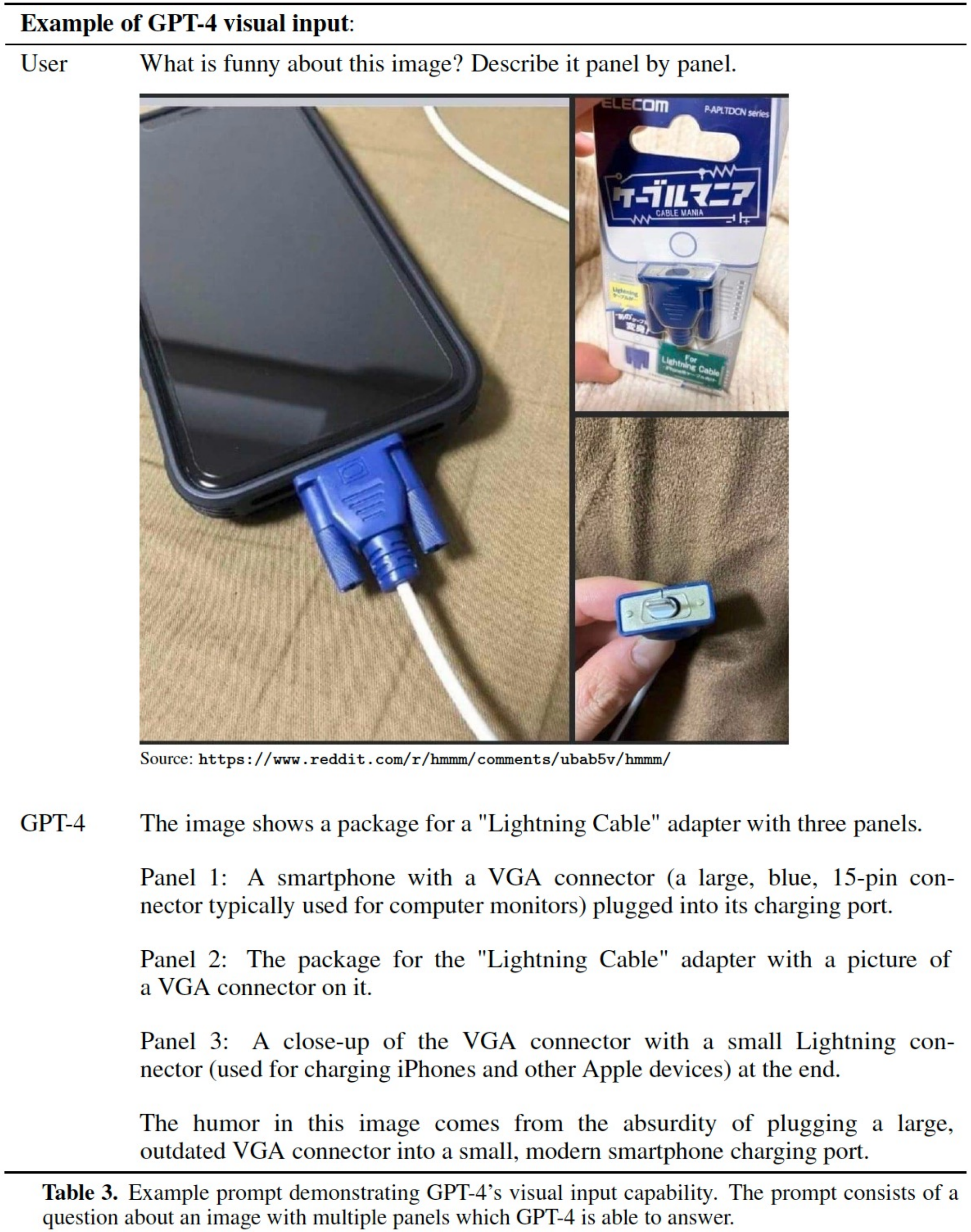
# Visual Inputs

GPT-4 accepts prompts consisting of both images and text, which—parallel to the text-only setting— lets the user specify any vision or language task.

Specifically, the model generates text outputs given inputs consisting of arbitrarily interlaced text and images.

Over a range of domains— including documents with text and photographs, diagrams, or screenshots — GPT-4 exhibits similar capabilities as it does on text-only inputs.

An example of GPT-4's visual input can be found in Table 3.


ELECOM
P-APLTDCN series
ケーブルマニア
CABLE MANIA
For Lightning Cable

**Example of GPT-4 visual input**:

User What is funny about this image? Describe it panel by panel.

Source: `https://www.reddit.com/r/hmmm/comments/ubab5v/hmmm/`

GPT-4 The image shows a package for a "Lightning Cable" adapter with three panels.

Panel 1: A smartphone with a VGA connector (a large, blue, 15-pin connector typically used for computer monitors) plugged into its charging port.

Panel 2: The package for the "Lightning Cable" adapter with a picture of a VGA connector on it.

Panel 3: A close-up of the VGA connector with a small Lightning connector (used for charging iPhones and other Apple devices) at the end.

The humor in this image comes from the absurdity of plugging a large, outdated VGA connector into a small, modern smartphone charging port.

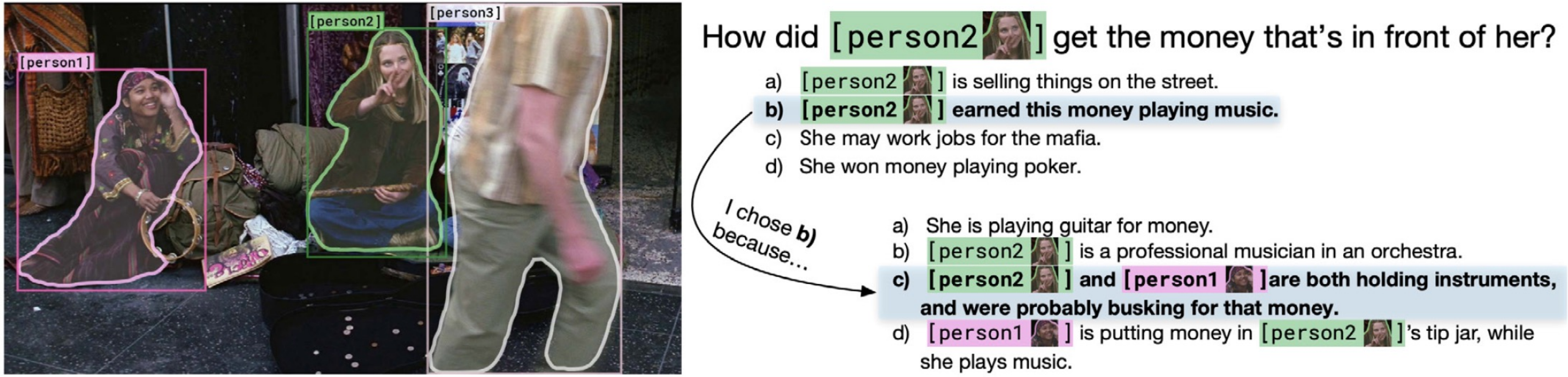
**Table 3.** Example prompt demonstrating GPT-4's visual input capability. The prompt consists of a question about an image with multiple panels which GPT-4 is able to answer.

## Visual Commonsense Reasoning (VCR)

The Visual Commonsense Reasoning challenge, first launched in 2019, is a relatively new benchmark in which AI systems must answer questions presented from images, as in VQA, but also select the reasoning behind their answer choices. Figure 2.2.24 shows an example of a question posed in VCR. Performance on VCR is tracked in the Q->AR score, which combines the ability of machines to select the right answer for the question (Q->A) and the ability to select the correct rationale behind the answer (Q->R).

### A Sample Question from the Visual Commonsense Reasoning (VCR) Challenge

Source: Zellers et al., 2018



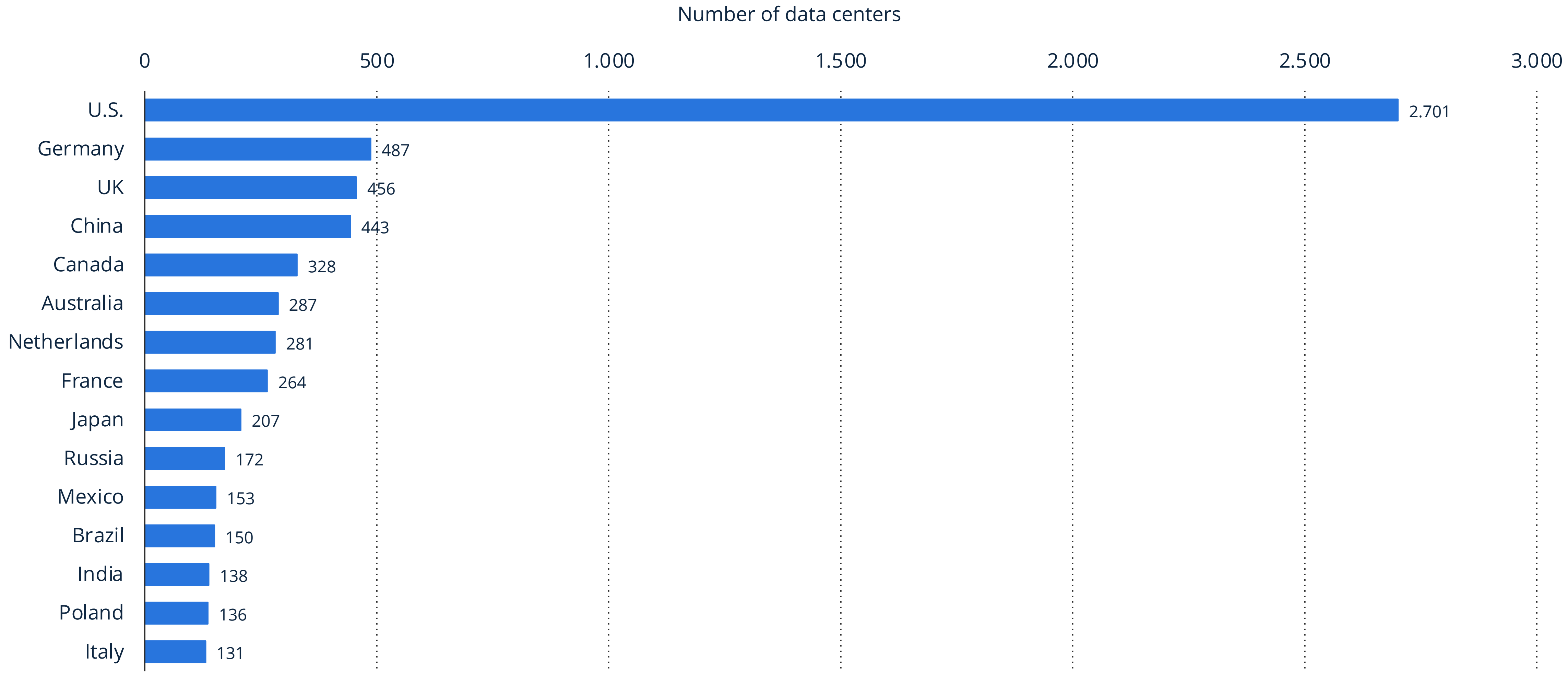
How did [person2] get the money that's in front of her?

a) [person2] is selling things on the street.
**b) [person2] earned this money playing music.**
c) She may work jobs for the mafia.
d) She won money playing poker.

a) She is playing guitar for money.
b) [person2] is a professional musician in an orchestra.
**c) [person2] and [person1] are both holding instruments, and were probably busking for that money.**
d) [person1] is putting money in [person2]'s tip jar, while she plays music.

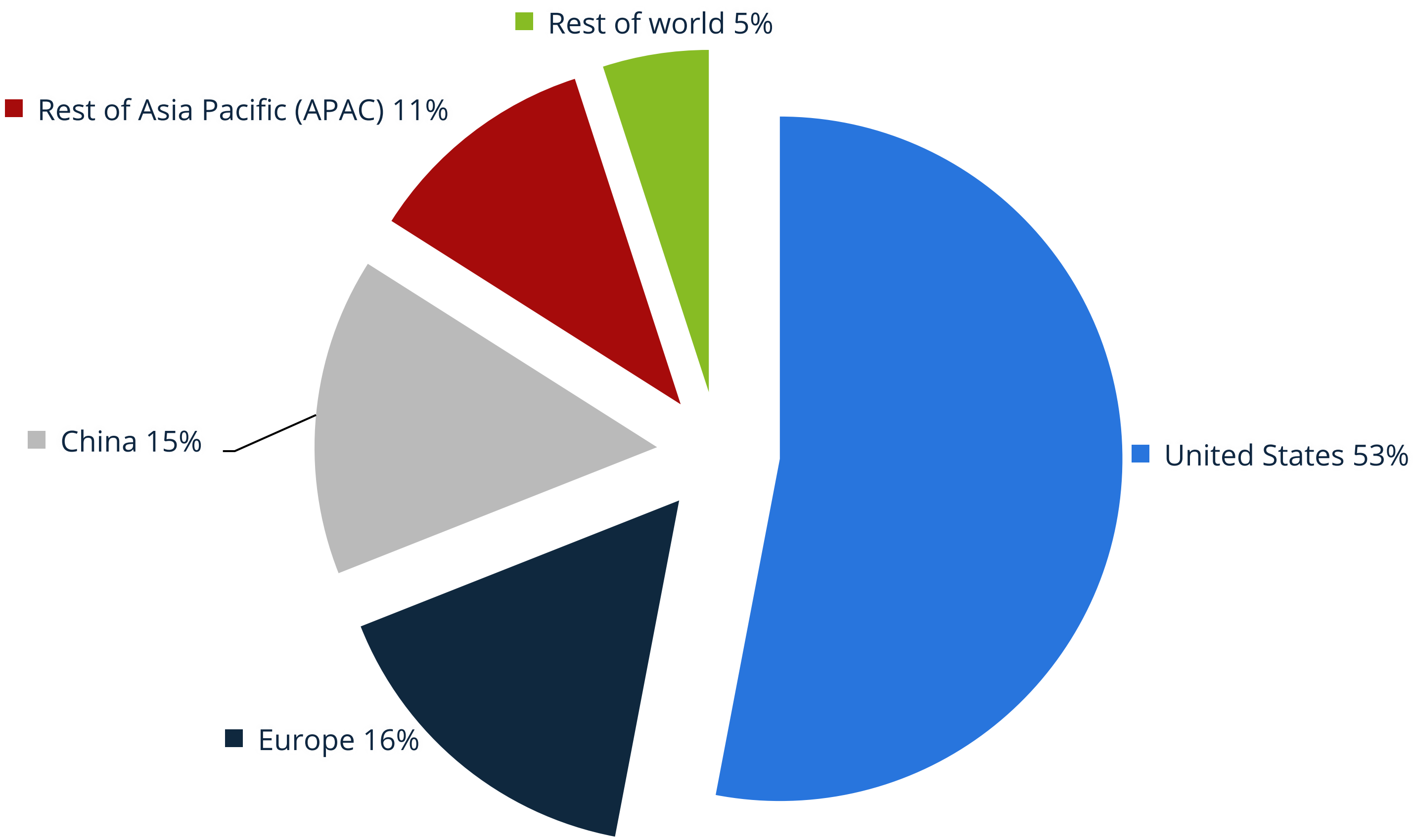
*Source: Stanford University, 2023*

# Number of data centers worldwide 2022, by country



**Description:** As of January 2022, 2,701 data centers were in the United States, with a further 487 data centers located in Germany. The United Kingdom ranked third among countries in terms of the number of data centers with 456, while China recorded 443.
**Source(s):** Cloudscene

# Share of global hyperscale data center capacity 2022, by region



**Description:** As of the second quarter of 2022, the United States accounts for over half of global hyperscale data center capacity, with a particular concentration in the state of Virginia. Europe holds the next highest capacity at 16 percent, followed closely by China at 15 percent.
**Source(s):** Synergy Research Group

# SSI & Digital Euro for AI-based public and private transactions

We have decided to launch the investigation phase of a digital euro project

# «Bring algorithms to data, not data to algorithms»

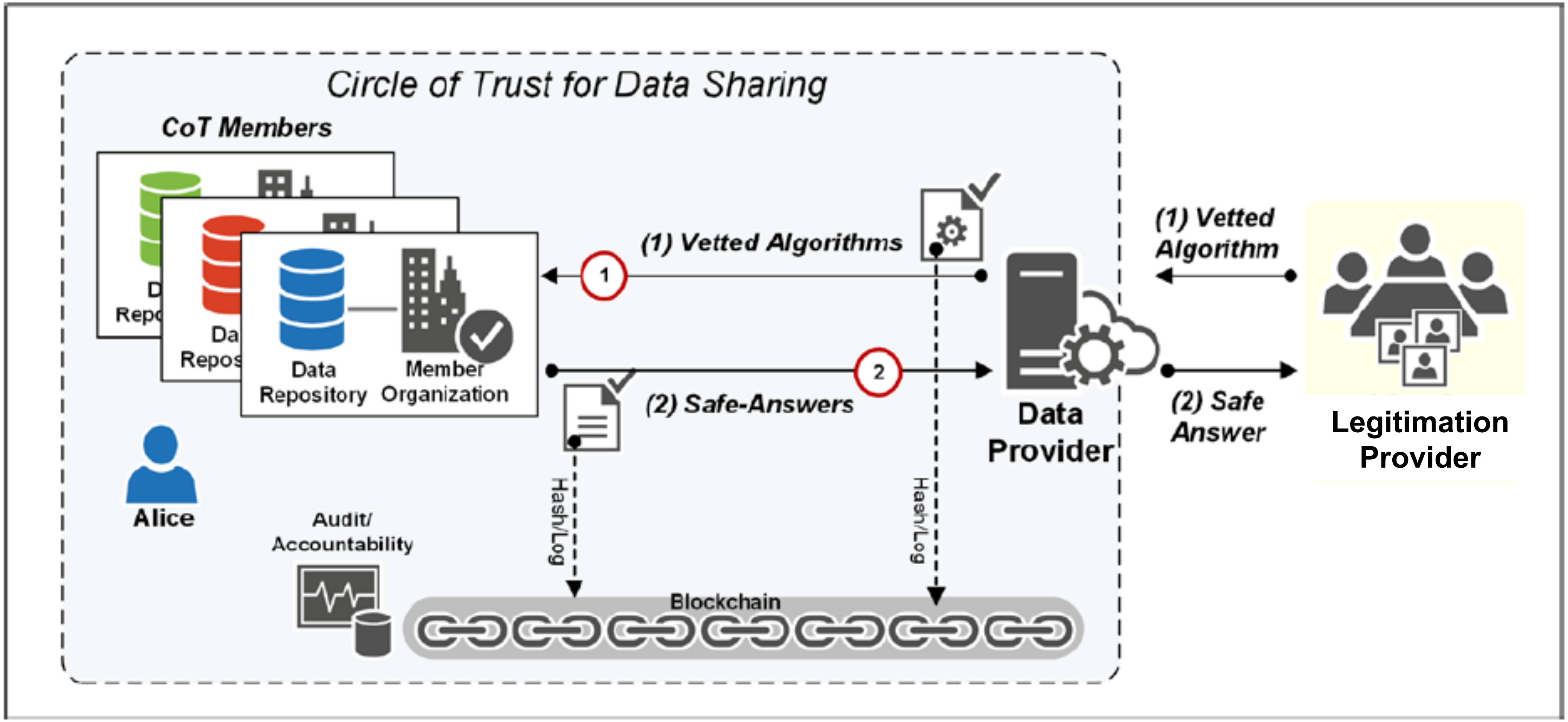
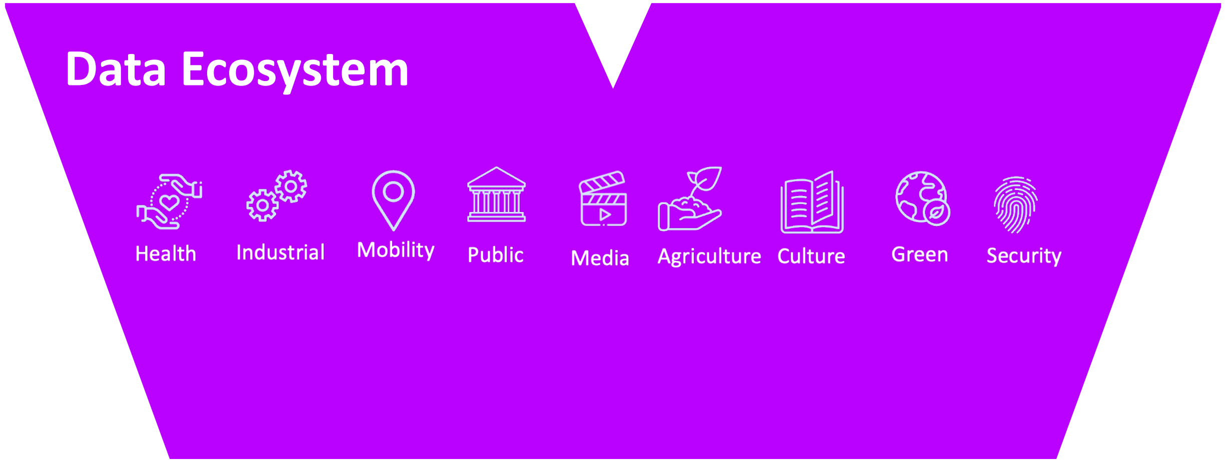


Figure 3: Circle of Trust for Data Sharing with a Data Provider as a proxy to the external world

# Fragmented data & infrastructures

## Data

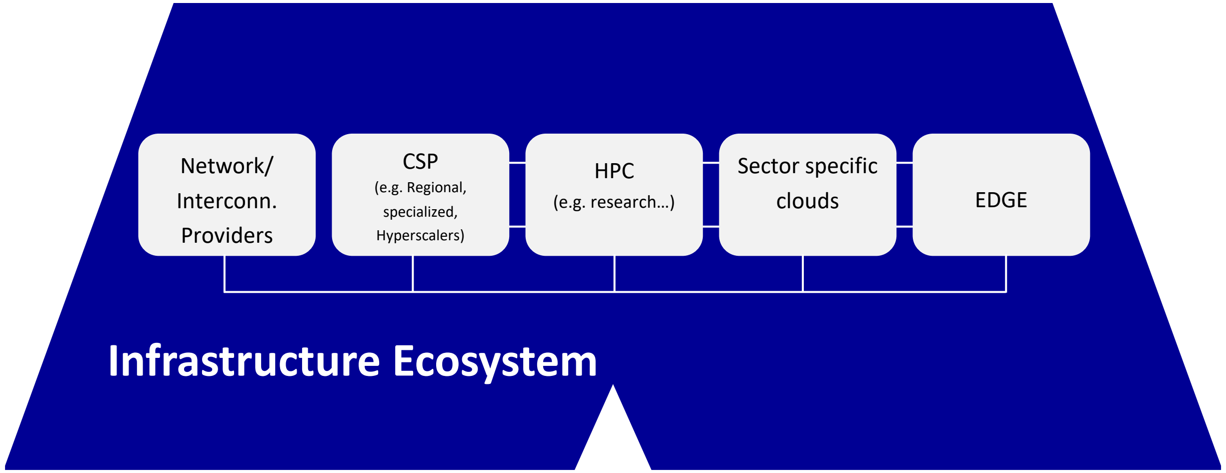
Untapped, fragmented, disjoint, no secure exchange mechanism



**Data Ecosystem**

## infrastructures

Segregated, non-reversible, non-interoperable, closed architectures, private standards



**Infrastructure Ecosystem**

Network/ Interconn. Providers
CSP (e.g. Regional, specialized, Hyperscalers)
HPC (e.g. research...)
Sector specific clouds
EDGE

# A federated model of European infrastructures

**Advanced Smart Services**

(Cross-) Sector Innovations/ Market places/ Applications

**Data Spaces**

Interoperable & portable (Cross-) Sector data-sets and services

**Gaia-X Services**

Decentralized services to verify trust, sovereign data exchange, and federation across nodes

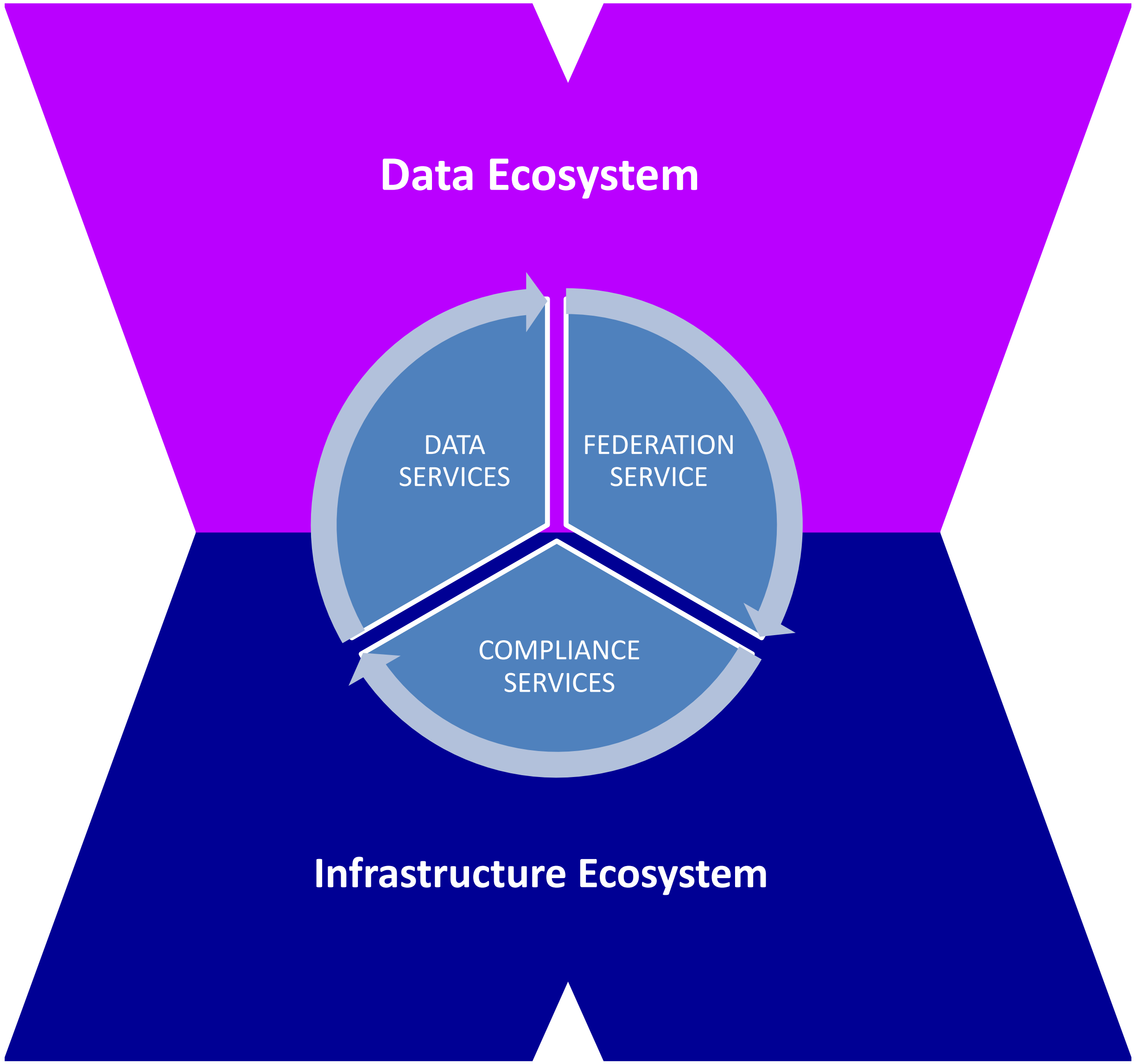
**Portability, Interoperability & interconnectivity**

Technical: Architecture and Specifications

Commercial: Policies and Labels

**Compliance**

Legal: Regulations



# Thanks!

*CarloAlberto Carnevale-Maffè*

Bocconi University – SDA Bocconi School of Management

Email: carloalberto.carnevale@sdabocconi.it

Twitter: @carloalberto