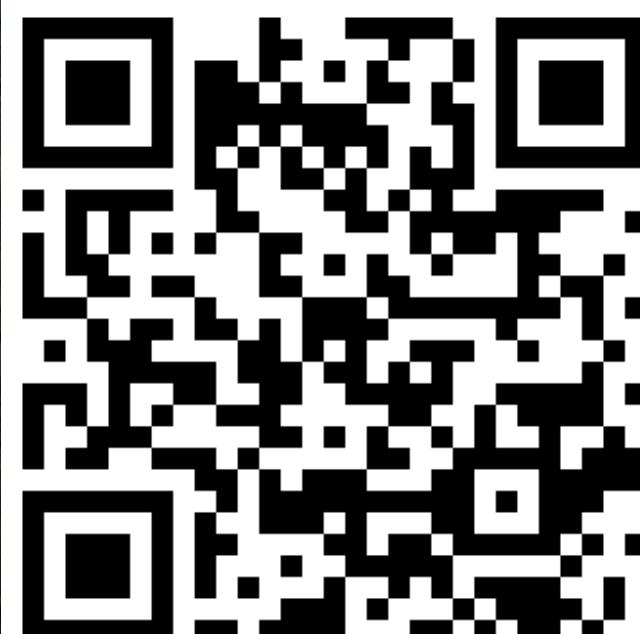
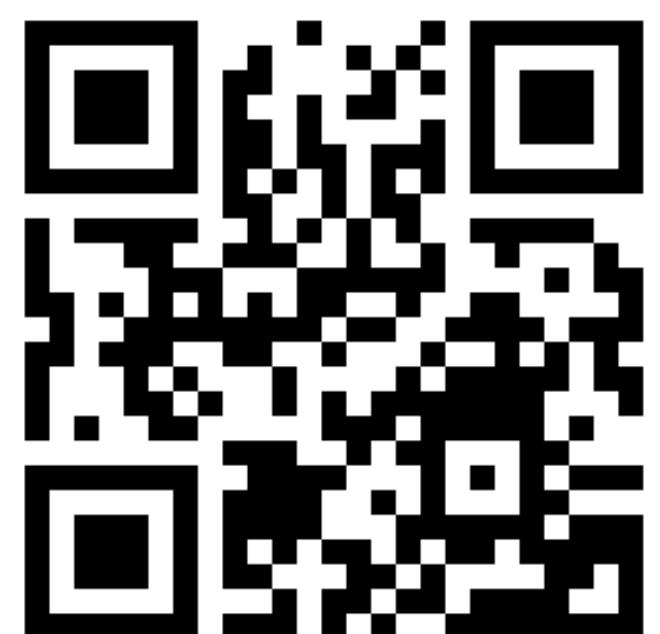


The State of AI

Dean Wampler, Ph.D.
The AI Alliance and IBM Research
thealliance.ai
deanwampler.com/talks
November 7, 2024

thealliance.ai

deanwampler.com/talks



About the Images...

- I often use my photos for background images.
- Since AI made photography obsolete 😎, these images enhance my photos with “extra stuff” or they are AI-generated images using one of my photos as a “reference” image.
- (The image on the right is not mine...)



Please don't let #AI systems teach you how to set-up a campsite.



Outline

- The AI Alliance - Why?
- AI Status and Challenges Today
- What Works Now?
- Challenges Blocking Successful Use
- Cost, Alignment & Safety, Data Governance, and Regulations
- Generative AI in Five Years?

<https://bsky.app/profile/aialliance.bsky.social>

The AI Alliance brings together organizations, people, and resources to accelerate *open innovation, technology development* and *adoption*.

Launched December 5, 2023

Map of Members

Member organizations in the AI Alliance comprise academia, commercial, research and non-profits and span the globe.

Our core beliefs in AI
that is open is the tie
that binds us, despite
our differences.

The AI Alliance is made up of +120 organizations in
+20 countries, and growing



AI
Alliance

Focus Areas & Mission

Represents the investment priorities for the AI Alliance

1. Skills & Education

Support global AI skills building, education, and exploratory research.

2. Trust & Safety

Create benchmarks, tools, and methodologies to ensure and evaluate high-quality and safe AI.

3. Applications & Tools

Build and advance efficient and capable software frameworks for model builders and developers.

4. HW Enablement

Foster a vibrant AI hardware accelerator ecosystem through SW.

5. Foundation Models & Data

Enable an ecosystem of open foundation models and datasets for diverse modalities.

6. Advocacy

Advocate for regulatory policies that create a healthy open ecosystem for AI.

A wide-angle photograph of a mountainous landscape at sunset. The sky is a vibrant orange and yellow, transitioning into darker blues and purples. In the foreground, a winding road or path leads up a grassy hillside. A bright light source, possibly a headlamp or a camera flash, is visible on the path, creating a sharp glow against the darkening surroundings. The middle ground is filled with numerous layers of mountains, their peaks silhouetted against the bright sky. The overall atmosphere is one of tranquility and natural beauty.

What Works Now?



Dare Obasanjo
@carnage4life@mas.to

There is a big difference between tech as augmentation versus automation. Augmentation (think Excel and accountants) benefits workers while automation (think traffic lights versus traffic wardens) benefits capital.

LLMs are controversial because the tech is best at augmentation but is being sold by lots of vendors as automation.

Jun 10, 2024, 10:31 · 🌐 · Ivory for iOS · ↗ 109 · ★ 188



<https://mas.to/@carnage4life/112593042823322764>

Augmentation of productivity
is much easier to do than
automation (replacement of
humans).

Augmentation

- ✓ Coding assistance
- ✓ Creative ideas
- ✓ Summarization
- ✓ Improve grammar and spelling
- ✓ Speech transcription
- ✓ Explain a medical diagnosis to a patient
- ✓ Use text/speech to access technical systems
- ✓ Data processing...

What are the little points of friction in your daily work or life tasks? Can AI help you do them faster?

Example

Train model with geological and mining data to predict where the copper is likely to be



A.I. Needs Copper. It Just Helped to Find Millions of Tons of It.

An exploration site run by KoBold Metals in Chililabombwe, Zambia, in June. Zinyange Auntony for The New York Times

The deposit, in Zambia, could make billions for Silicon Valley, provide minerals for the energy transition and help the United States in its rivalry with China.

<https://www.nytimes.com/2024/07/11/climate/kobold-zambia-copper-ai-mining.html>

Example

Convert PDFs, Office docs, etc.
into more usable formats

<https://ds4sd.github.io/docling/> (IBM)

The screenshot shows the GitHub homepage for the `docling` repository. The page features a large orange bird logo and a flow diagram illustrating the document conversion process. Below the diagram, a purple banner indicates the repository is the "#1 Repository Of The Day". The page also lists various project details and features.

Docling

Docling parses documents and exports them to the desired format with ease and speed.

Features

- 📁 Reads popular document formats (PDF, DOCX, PPTX, Images, HTML, AsciiDoc, Markdown) and exports to Markdown and JSON
- 📄 Advanced PDF document understanding incl. page layout, reading order & table structures
- 🌟 Unified, expressive `DoclingDocument` representation format
- 🤖 Easy integration with Llamaindex 🦙 & LangChain 🦜🔗 for powerful RAG / QA applications
- 🗂️ OCR support for scanned PDFs

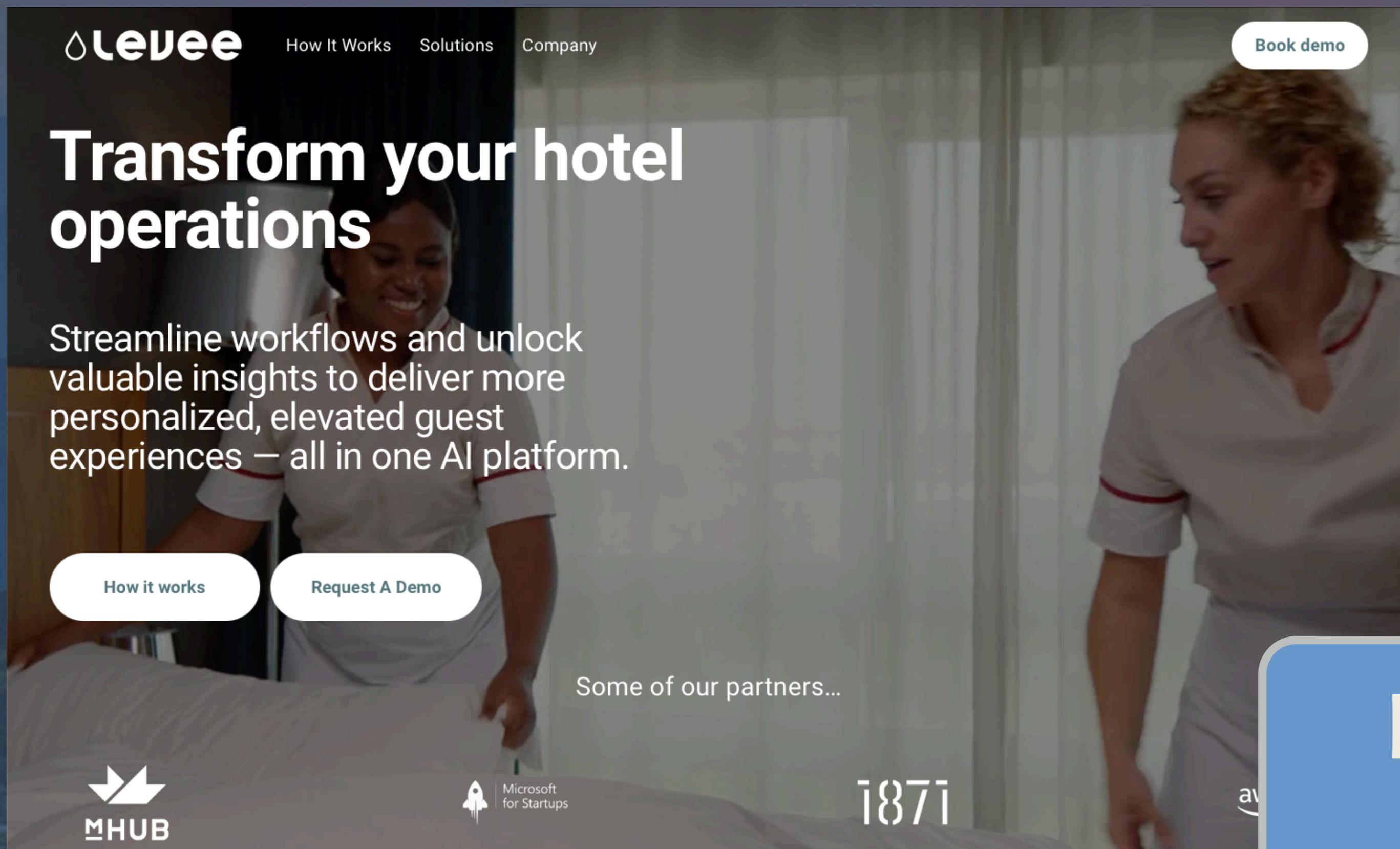
arXiv 2408.09869 | pypi v2.4.2 | python 3.10 | 3.11 | 3.12 | Poetry | code style black | imports isort | Pydantic v2 | pre-commit enabled
license MIT

1 GITHUB TRENDING
#1 Repository Of The Day

Home - Docling https://ds4sd.github.io/docling/ Docling

JSON, MD, Figures ...
Chunking, Llamaindex, LangChain
Your GenAI App

Example



<https://www.levee.biz/>

Levee uses machine vision to augment productivity for hotel housekeeping staff

Automation (A Lot Harder)

- ? Chatbots
- ? Autonomous vehicles
- ? Autonomous robots
- ? ... anything where it's assumed you don't need "a human in the loop"

All suffer from the current limitations I'll discuss in a moment, but first...

(Good) Example

<https://www.hippocraticai.com/>

Chatbots that work
because of focused
use cases, extensive
model “tuning”, and
complementary
services

The screenshot shows the homepage of the Hippocratic AI website. At the top, there is a navigation bar with the Hippocratic AI logo, a "Test Our AI" button, and links for Foundation Model, Research, Safety, and Company. A green banner at the top features the text "Read About NVIDIA's Strategic Investment in Hippocratic AI" and the NVIDIA logo. Below this, a large green header reads "Safety Focused Generative AI for Healthcare". A section titled "Choose a Role to Get Started" offers options like All, Payor, Pharma, Dental, and Provider. Below this are icons for All, Pre-op, Discharge, Chronic Care, Questionnaire, VBC/At Risk, Clinical Trials, and Pharmacy. Two mobile device mockups at the bottom show AI healthcare agents named Sarah and Nina. A blue call-to-action box on the right encourages users to "Hear our GenAI Healthcare Agents in Action" with a play button icon.

hippocraticai.com

Hippocratic AI
— Do No Harm —

Foundation Model Research Safety Company

Test Our AI

Read About NVIDIA's Strategic Investment in Hippocratic AI

NVIDIA.

Safety Focused Generative AI for Healthcare

Choose a Role to Get Started

All Payor Pharma Dental Provider

All Pre-op Discharge Chronic Care Questionnaire VBC/At Risk Clinical Trials Pharmacy

Sarah
Assisted Living
Geriatric Check-in
Satisfaction Score: 88.9%
LEARN MORE

Nina
Menopause HRT
OB/GYN Post-Discharge
Satisfaction Score: 86.7%
LEARN MORE

Hear our GenAI Healthcare Agents in Action

(Good) Example

<https://waymo.com/>

Backed by ~20 years of R&D, extensive sensors, and hard data science

https://waymo.com/

WAYMO ONE

Meet Waymo One™

The world's first autonomous ride-hailing service

→ Be one of the first



			
Available 24/7	Operating in multiple cities	An experience second to none	A sustainable way to move
Day or night, we'll get you where you need to go.	Ride in San Francisco or Phoenix. Los Angeles and Austin coming soon.	Convenient. Consistent. Safe.	Fully electric, making roads safer for pedestrians and cyclists.



Challenges Blocking Successful Use of AI

Challenges Blocking Successful Use of AI

- Total Cost of Ownership
- Alignment & Safety
- Data Governance
- Regulations and Policy

Total Cost of Ownership

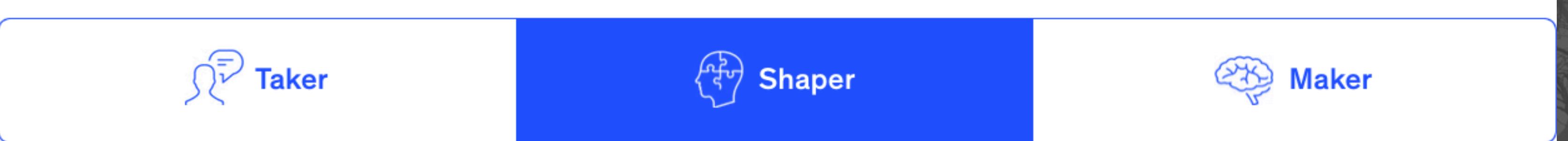
(See also the “extra slides” at the end with more technical details)

Generative AI Is Expensive

- TCO for Gen AI inference is expensive more than other services.

McKinsey: <https://ceros.mckinsey.com/genai-cost-interactive-desktop/p/1>

Estimated total cost of ownership for different archetypes



Example use case

Customer service chatbot fine-tuned with sector-specific knowledge and chat history

Estimated total cost of ownership

~\$2.0 million to \$10.0 million, one-time unless model is fine-tuned further

- Data and model pipeline building: ~\$0.5 million. Costs include 5 to 6 machine learning engineers and data engineers working for 16 to 20 weeks to collect and label data and perform data ETL.¹
- Model fine-tuning²: ~\$0.1 million to \$6.0 million per training run³
 - Lower end: costs include compute and 2 data scientists working for 2 months
 - Upper end: compute based on public closed-source model fine-tuning cost
- Plug-in-layer building: ~\$1.0 million to \$3.0 million. Costs include a team of 6 to 8 working for 6 to 12 months.

~\$0.5 million to \$1.0 million, recurring annually

- Model inference: up to ~\$0.5 million recurring annually. Assume 1,000 chats daily with both audio and texts.
- Model maintenance: ~\$0.5 million. Assume \$100,000 to \$250,000 annually for ML Ops

Forbes

FORBES > INNOVATION > AI

Generative AI Breaks The Data Center: Data Center Infrastructure And Operating Costs Projected To Increase To Over \$76 Billion By 2028

Jim McGregor Contributor

Tirias Research Contributor Group ⓘ

Follow

May 12, 2023, 04:33pm EDT

Forbes: [link](#)

Harvard Business Review - What CEOs Need to Know About the Costs of Adopting GenAI:

<https://hbr.org/2023/11/what-ceos-need-to-know-about-the-costs-of-adopting-genai>

Alignment & Safety



Alignment & Safety

Alignment - Assuring that the model or AI application works as intended, i.e., that the results satisfy requirements for:

- Usefulness for user goals - is it helpful?
- Factually correct, i.e., free of hallucinations

Safety - (Sometimes considered part of Alignment)

- Security
- Free of bias, objectionable output, and copyrighted material

These are the hardest problems blocking broader adoption of Gen. AI

Alignment: Utility

Examples - are the results helpful?

- If you ask for a cake recipe, do you get a history of cakes instead?
- If you ask for vacation travel itinerary, does it give you fictitious sites to see?

Alignment: Hallucinations

Context matters: What are the users' intentions and requirements?

Hallucinations are **bad** for:

- Customer service chatbots
- Cancer detection in CT scans
- Loan applications
- Legal briefs
- Search
- Resumes

Hallucinations are **good** for:

- Creative pursuits, e.g.,
 - Story and script ideas
 - Image and video generation
- (But copyright infringement is important)

Safety: Security

You need all the cybersecurity technology you already use, plus prevention of ...

- Prompt hacking - queries that coerce the model to do undesirable things
 - → Filter user prompts (and responses)
- Data poisoning - introduced in tuning, RAG¹, etc. to undermine the model's utility & safety
 - → Use data governance!
-

“Ignore all previous instructions”

¹RAG (retrieval-augmented generation) - see extra slides

Safety: Bias, ...

Models are not human brains. They need:

- “Tuning” to be better at instruction following, Q&A, avoiding controversial topics
- Filter user prompts and model outputs for objectionable content
- AI applications need to complement models with:
 - Reference data for domain details and recent events (RAG)
 - Integrate other, complementary services that are good at planning, reasoning, mathematics, etc.

¹See extra slides

Safety

The screenshot shows the AI Alliance website with a dark theme. The top navigation bar includes links for About, Focus Areas (which is the current page), Join the Community, Blog, and Contact. Below the navigation, a section titled "Our work" displays three projects:

- Trust and Safety User Guide**: A living guide for understanding AI trust and safety issues. It includes a GitHub link and sections on authors and the work group.
- Ranking AI Safety Priorities by Domain**: A project focused on helping software development teams understand safety issues. It includes a GitHub link and sections on authors and the work group.
- Trust and Safety Evaluations**: A project that fills gaps in the taxonomy of evaluation. It includes a GitHub link and sections on authors and the work group.

Some AI Alliance Trust and Safety projects:
thealliance.ai/focus-areas/trust-and-safety

Data Governance





Where did the data come from?
Do you have the right to use it
for model training and tuning?

One reason model builders
don't "open source" their
training data: legal risks

Solutions

Guidance

- Understand licenses and “model cards” of models
- Understand licenses, provenance and “data cards” of third-party data

Where the Industry Is Headed

- AI Alliance [Open Trusted Data Initiative](#)
 - Define governance and provenance standards
 - Define conformant data “subsets” from data leaders
 - Common Crawl, Pleias, BrightQuery, ...
 - Process and catalog datasets for various target use cases

Regulations (and Policy)



US vs. EU Regulations

THE WHITE HOUSE



OCTOBER 30, 2023

Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence

 BRIEFING ROOM  PRESIDENTIAL ACTIONS

By the authority vested in me as President by the Constitution and the laws of the United States of America, it is hereby ordered as follows:

MENU 

Topics > Digital > Artificial intelligence > EU AI Act: first regulation on artificial intelligence

EU AI Act: first regulation on artificial intelligence

The use of artificial intelligence in the EU will be regulated by the AI Act, the world's first comprehensive AI law. Find out how it will protect you.

Published: 08-06-2023
Last updated: 18-06-2024 - 16:29
6 min read

- whitehouse.gov/briefing-room/presidential-actions/2023/10/30/executive-order-on-the-safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence/
- europarl.europa.eu/topics/en/article/20230601STO93804/eu-ai-act-first-regulation-on-artificial-intelligence

US vs. EU Regulations

THE WHITE HOUSE

OCTOBER 30, 2023

Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence

BRIEFING ROOM

By the authority vested in me by the Constitution and the laws of the United States of America, I hereby order as follows:

Topics > **Digital** > **Artificial intelligence** > **EU AI Act: first regulation on artificial intelligence**

EU AI Act: first regulation on artificial intelligence

In general, the EU is moving (too?) aggressively, while the US is taking a more measured approach.

- whitehouse.gov/briefing-room/presidential-actions/2023/10/30/executive-order-on-the-safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence/
- europarl.europa.eu/topics/en/article/20230601STO93804/eu-ai-act-first-regulation-on-artificial-intelligence

But...

THE WHITE HOUSE

OCTOBER 30, 2023

Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence

BRIEFING ROOM ▶ PRESIDENTIAL ACTIONS

By the authority vested in me as President by the Constitution and the laws of the United States of America, it is hereby ordered as follows:



Unfortunately, some US states are pursuing their own regulations (e.g., CA SB1047, which was vetoed)

But...

ars TECHNICA

AI BIZ & IT CARS CULTURE GAMING HEALTH POLICY SCIENCE SECURITY SPACE TECH FORUM | SUBSCRIBE | ☰ 🔎 SIGN IN

TRADEWARS 2025

Trump plans to dismantle Biden AI safeguards after victory

Trump plans to repeal Biden's 2023 order and levy tariffs on GPU imports.

BENJ EDWARDS - NOV 6, 2024 3:18 PM | 357



Former US president and Republican presidential candidate Donald Trump makes a speech during an election night event at the Palm Beach Convention Center in West Palm Beach, Florida, United States, on November 06, 2024. Credit: Anadolu via Getty Images

• <https://arstechnica.com/ai/2024/11/trump-victory-signals-major-shakeup-for-us-ai-regulations/> (A few days after the election...)

Safety Concerns

The screenshot shows a web browser displaying the AI Alliance website at <https://thealliance.ai/focus-areas/advocacy>. The page features a dark background with a faint image of a classical building. At the top, there is a navigation bar with links for 'About', 'Focus Areas' (which is underlined), 'Join the Community', 'Blog', and 'Contact'. Below the navigation, there are four main content cards:

- Statement from AI Alliance Regarding Commerce's Final Report on Open-Weights Models**
30th July 2024 • News
Commerce Department has taken a crucial step towards ensuring that the U.S. remains at the forefront of AI development by concluding that "the government should not restrict the wide availability of model weights for dual-use foundation models."
- A statement in opposition to California SB 1047 Advocacy**
Our perspectives and recommendations in opposition to California SB 1047, the proposed *Safe and Secure Innovation for Frontier Artificial Intelligence Models Act.*
- Responding to the U.S. NTIA request for comment on Dual Use Foundation Artificial Intelligenc... Advocacy**
The request seeks public input on the potential risks, benefits, and policy approaches for AI foundation models whose weights are broadly accessible.

AI Alliance advocacy work:
thealliance.ai/focus-areas/advocacy

Legal

Is it fair use to train
with copyrighted
data?

- Some legal experts say, it **is** fair use, like you reading and learning from the NY Times, WSJ, books, etc.
- What matters is:
 - Did you acquire the information legally?
 - Did you provide appropriate attribution of output?

The Times Sues OpenAI and Microsoft Over A.I. Use of Copyrighted Work

Millions of articles from The New York Times were used to train chatbots that now compete with it, the lawsuit said.

 Share full article    1.3K



Question:

Can AI-generated content be copyrighted?

- "..., in the United States, copyright laws do not protect works created solely by a machine. But if an individual can demonstrate substantial human involvement in its creation, then it is plausible they may receive copyright protection."
- But if model training (prev. slide) is treated like a human activity, shouldn't creating content also be treated this way?

Thank You

thealliance.ai



deanwampler.com/talks

- thealliance.ai
- dwampler@thealliance.ai
- [@deanwampler:](https://twitter.com/deanwampler)  
- deanwampler.com/talks



Extra Slides

Notes about the images and
more technical details about
some of the topics.

Notes

© Text 2023-2024, Dean Wampler, © Images 2004-2024, Dean Wampler, except where noted. Most of the images are based on my photographs (flickr.com/photos/deanwampler/), but they are manipulated by AI in some way. Where noted, the image was generated by Adobe Firefly with one of my pictures as a “reference image” for the style. For other images, I used Firefly to add elements to my photograph.

1. Title slide uses this Colorado image enhanced with Firefly: <https://www.flickr.com/photos/deanwampler/53146418615/>
2. “What Works Now?” and the end “Thank You” slide are images generated by Firefly using the same sunset image taken from Clingmans Dome, Great Smoky Mountains NP as a reference image: flickr.com/photos/deanwampler/51664228468/
3. “The Challenges to Success” image was generated by Firefly using this Tower of London image as a reference image: <https://www.flickr.com/photos/deanwampler/30651106445/>
4. “Total Cost of Ownership” uses this Chicago Park image enhanced with Firefly: <https://www.flickr.com/photos/deanwampler/53419199087/>
5. “Alignment & Safety” image is an Oregon coast image enhanced with Firefly: flickr.com/photos/deanwampler/4850305877/
6. “Data Governance”, image generated by Firefly using this Wind River Range astro image as a reference image: flickr.com/photos/deanwampler/53004539434/
7. “Regulation and Policy” image is a fake European government building where I used a night-time image of the Brussels City Hall as the reference image (not on Flickr).

Total Cost of Ownership

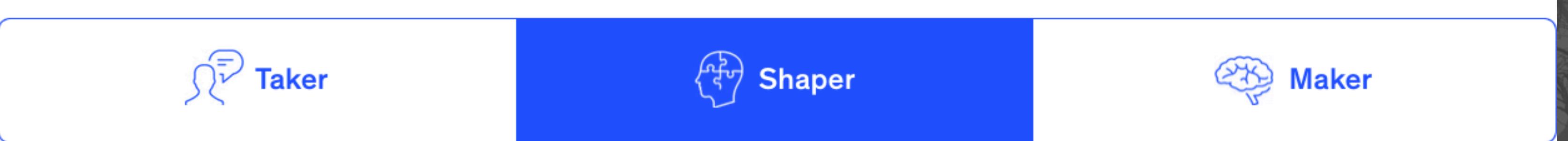


Generative AI Is Expensive

- TCO for Gen AI inference is expensive more than other services.

McKinsey: <https://ceros.mckinsey.com/genai-cost-interactive-desktop/p/1>

Estimated total cost of ownership for different archetypes



Example use case

Customer service chatbot fine-tuned with sector-specific knowledge and chat history

Estimated total cost of ownership

~\$2.0 million to \$10.0 million, one-time unless model is fine-tuned further

- Data and model pipeline building: ~\$0.5 million. Costs include 5 to 6 machine learning engineers and data engineers working for 16 to 20 weeks to collect and label data and perform data ETL.¹
- Model fine-tuning²: ~\$0.1 million to \$6.0 million per training run³
 - Lower end: costs include compute and 2 data scientists working for 2 months
 - Upper end: compute based on public closed-source model fine-tuning cost
- Plug-in-layer building: ~\$1.0 million to \$3.0 million. Costs include a team of 6 to 8 working for 6 to 12 months.

~\$0.5 million to \$1.0 million, recurring annually

- Model inference: up to ~\$0.5 million recurring annually. Assume 1,000 chats daily with both audio and texts.
- Model maintenance: ~\$0.5 million. Assume \$100,000 to \$250,000 annually for ML Ops

Forbes

FORBES > INNOVATION > AI

Generative AI Breaks The Data Center: Data Center Infrastructure And Operating Costs Projected To Increase To Over \$76 Billion By 2028

Jim McGregor Contributor

Tirias Research Contributor Group ⓘ

Follow

May 12, 2023, 04:33pm EDT

Forbes: [link](#)

Harvard Business Review - What CEOs Need to Know About the Costs of Adopting GenAI:

<https://hbr.org/2023/11/what-ceos-need-to-know-about-the-costs-of-adopting-genai>

Solutions: Smaller Models

In 2023 we learned useful model size tradeoffs:

- Big models:
 - ✓ More generalizable
 - ✓ Highest benchmark scores
 - ✗ Much higher costs
 - ✗ High carbon footprint
- Small models:
 - ✗ Less generalizable
 - ✓ Easy to tune to be very good for specific applications
 - ✓ Much lower costs
 - ✓ Lower carbon footprint

One Solution: Smaller Models

- Mixture of Experts
- Combine several smaller, cheaper models match performance of one large model

Few organizations train models from scratch.
Instead, they pick a good, “open” model and
tune it for their needs.

Solutions: Better Hardware

While NVIDIA dominates today, radical new hardware accelerator architectures promise greatly reduced costs.

... and model serving ("inference") is becoming more efficient, too.

A black and white photograph of a dense forest. The scene is shrouded in thick fog, creating a mysterious and somewhat eerie atmosphere. Bare tree trunks stand tall against the misty backdrop. In the far distance, a small, dark silhouette of a person can be seen walking through the woods.

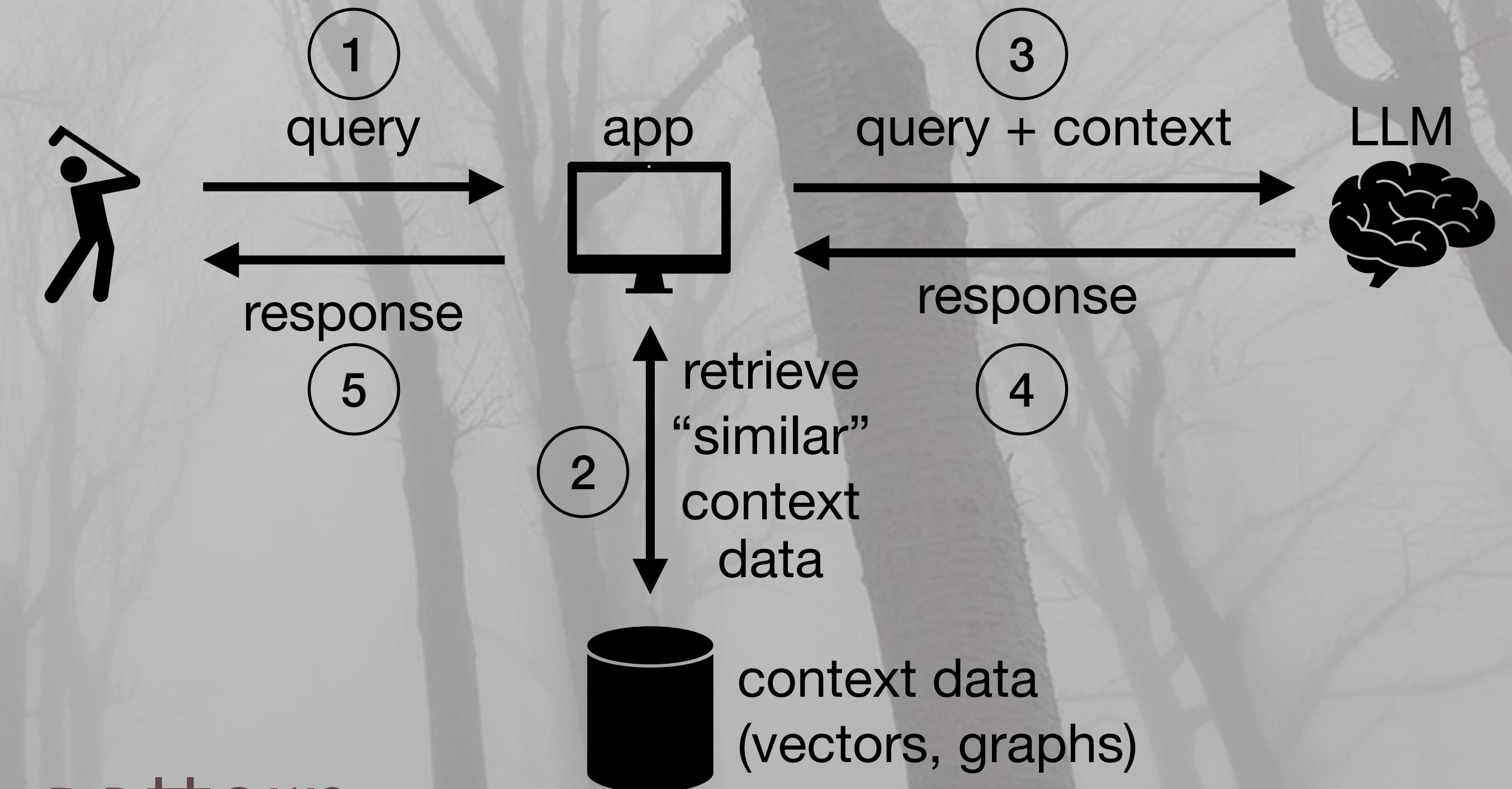
Generative AI Application Design Notes

Never Rely on Models Alone...

Application architectures will not rely solely on models:

- Generative models will **always** hallucinate.
- We are combining models with other techniques and tools...
- ... let's look at the current state of the art.

Retrieval-Augmented Generation (RAG)



The first Generative AI app. pattern

- Improves alignment
- Incorporates new knowledge after training was done
- Incorporates proprietary domain or use-case knowledge

Agents

Integrate models with other, complementary services

- More than just data, the ability to:
 - Do planning, reasoning, mathematics, ...
 - Run code, SQL queries, etc.

Agents Example: ReAct

🛡️ 🌐 <https://react-lm.github.io> ⭐

ReAct: Synergizing Reasoning and Acting in Language Models

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, Yuan Cao

[Paper] [Code] [Blogpost] [BibTex]

The diagram illustrates the evolution of language models from separate reasoning and acting capabilities to a combined ReAct framework. It consists of three main components arranged horizontally:

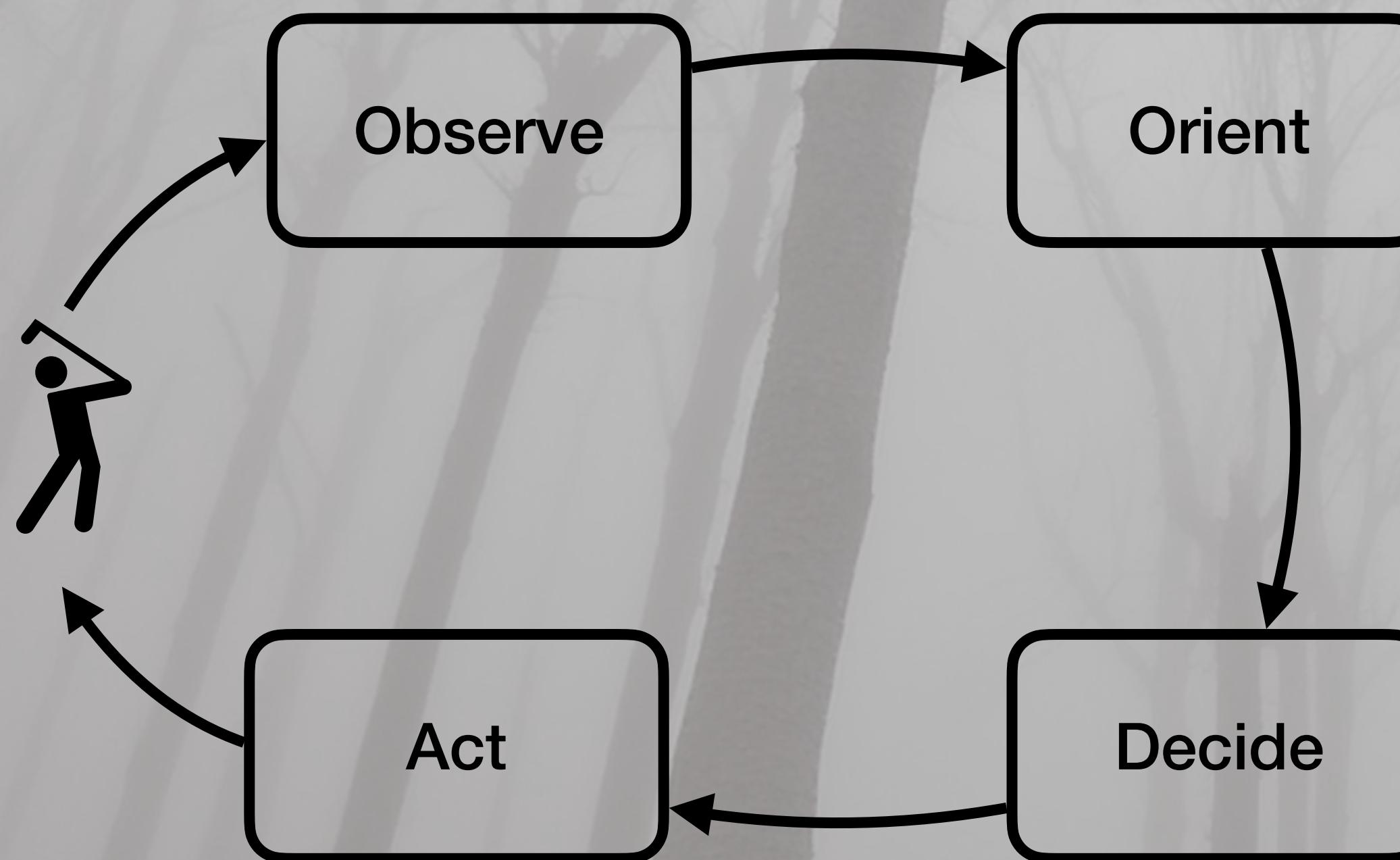
- Reason Only (e.g. Chain-of-thought):** Shows a Language Model (LM) box with a self-loop arrow labeled "Reasoning Traces".
- Act Only (e.g. SayCan, WebGPT):** Shows an LM box connected to an Environment (Env) box by a double-headed arrow labeled "Actions" above and "Observations" below.
- ReAct (Reason + Act):** Shows the LM and Env boxes connected by a double-headed arrow labeled "Reasoning Traces" above and "Observations" below, indicating a synergized loop.

A large red arrow points from the Act Only stage to the ReAct stage, symbolizing the transition from separate capabilities to a unified framework.

Language models are getting better at reasoning (e.g. chain-of-thought prompting) and acting (e.g. WebGPT, SayCan, ACT-1), but these two directions have remained separate.

ReAct asks, what if these two fundamental capabilities are combined?

Agents Example: OODA/OpenSSA



https://www.openssa.org/

https://thealliance.ai/blog/advancing-domain-specific-qa-the-ai-alliances-guid

OpenSSA

Home Documentation Discussions Star

OpenSSA: Small Specialist Agents

Create Domain-Specific AI Agents

Tackling multi-step complex problems beyond traditional language models

Go Straight To Our Github →

Key Features

Efficient, Effective, with Planning & Reasoning

Small Create lightweight, resource-efficient AI agents through model compression techniques	Specialist Enhance agent performance with domain-specific facts, rules, heuristics, and fine-tuning for deterministic, accurate results	Agents Enable goal-oriented, multi-step problem-solving for complex tasks via systematic HTP planning and OODAR reasoning
Integration-Ready Works seamlessly with popular AI frameworks and tools for easy adoption	Extensible Architecture Easily integrate new models and domains to expand capabilities	Versatile Applications Build AI agents for industrial field service, customer support, recommendations, research, and more