Deloitte.

GenAI, Agents, and Industry Applications

2025 AI Winter School

Brown Unviersity

January 16, 2025

Today's Presenters

Michael Luk
Managing Director (Partner), Deloitte Consulting, miluk@deloitte.com

CTO and co-founder, SFL Scientific

PhD, Experimental Particle Physics, Brown University

Michael leads a world-class team of data scientists, engineers, and managers, in providing clients with innovative, practical, bespoke solutions.

Alexis Johnson
Consultant -- ML Engineer
alexijohnson@deloitte.com

PhD, Math, Rice University (2019)

Alexis is a consultant at SFL Scientific, a Deloitte business, where she develops machine learning and generated AI solutions.

She is an expert in machine learning, deep learning, and GenAI with extensive experience building and integrating enterprise models at large scale and providing demonstrable ROI with these technologies.

Agenda

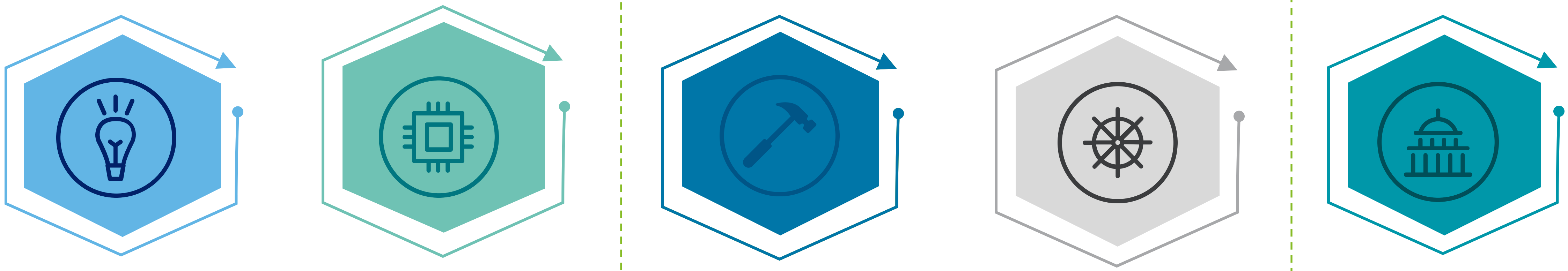
1. GenAI: Production lifecycle
2. Industry Examples
3. Large Language Models (LLMs)
4. Retrieval Augmented Generation (RAG)
5. Notebook content
6. ReAct Framework
7. Notebook content

GenAI: From Idea to Production

AI/GenAI dimensions

There are multiple considerations and dependencies when initializing AI/GenAI Opportunities

TODAY'S FOCUS



Strategy

Define the organizational GenAI vision & guiding principles in line with broader business strategy and activate the capabilities to realize this vision

Technology

Enable the technology stack with next-gen architecture design and ensure quality data is available for GenAI to work

Delivery

End-to-end GenAI solution and capability delivery in alignment with GenAI vision and business value realization

Talent, Organization, and Culture

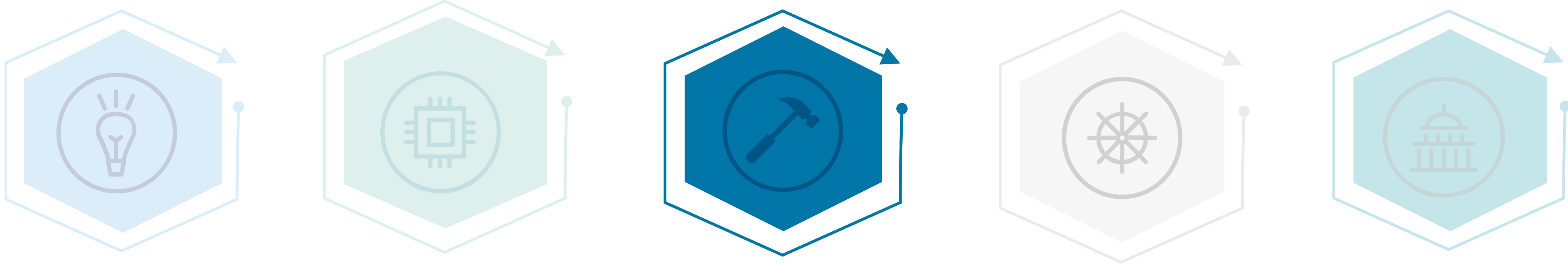
Create channels for efficient training and comms. Along with an operating model to ensure rapid propagation of GenAI across the enterprise

Governance

Define the guardrails and controls to mitigate against GenAI risks and define decision rights for the organization to enable stakeholders to deliver capabilities against defined standards

AI/GenAI dimensions

There are multiple considerations and dependencies when initializing AI/GenAI Opportunities



Strategy

Define the organizational GenAI vision & guiding principles in line with broader business strategy and activate the capabilities to realize this vision

Technology

Enable the technology stack with next-gen architecture design and ensure quality data is available for GenAI to work

Delivery

End-to-end GenAI solution and capability delivery in alignment with GenAI vision and business value realization

Talent, Organization, and Culture

Create channels for efficient training and comms. Along with an operating model to ensure rapid propagation of GenAI across the enterprise

Governance

Define the guardrails and controls to mitigate against GenAI risks and define decision rights for the organization to enable stakeholders to deliver capabilities against defined standards

Delivery: AI program lifecycle

To scale AI, organizations have started to establish a GenAI governance model that deters, detects and monitors AI specific risks as AI solutions are designed and implemented

PHASE 1:ADVISE
Ideation, Prioritization & Value Analysis

Ideation & Evaluation → Exploratory Data Analysis → AI Strategy → Solution Architecture

PHASE 2: IMPLEMENT
Development & Governance Activation

POC → Pilot → MVP

PHASE 3: OPERATE
Productionize & Scale

Full Scale Deployment & MLOps → Solution Expansion & AI Strategy

Deliverables & Outcomes

Phase 1:
- Use Case Inventory and Roadmap
- Tech Capability Assessment
- High Level Op Model & Governance Framework
- Center of Excellence (CoE) Setup Roadmap

Phase 2:
- Business Case Evaluation Framework
- Physical Architecture and Vendor Assessment
- Target GenAI Operating Model
- Communications and Change Management Plan

Phase 3:
- Pre-Defined Value Measurement Framework
- Monthly Model Performance Evaluation
- Quarterly Model Fine Tuning

Phase 1 Advise: Desirability, viability, feasibility

Use cases must be evaluated across three dimensions – desirability, viability, and feasibility – to validate the strategies and establish roadmaps

1 DESIRABILITY

What is the business value?

Is the use case aligned to enterprise strategic priorities?

2 VIABILITY & FEASIBILITY

What is the business case?

What is the return on investment?

What is the ease of implementation?

Does the talent exist to enable implementation?

Data/Tech Availability?

Does the data and technical solution exist?

VIABILITY (High / Low)

How does this use case fit into our financial goals?

BIG POTENTIAL

- *High business impact on objectives and priorities*
- *High ROI, lower time to value, and ethical implications*
- *High technology / delivery complexity and technology to implement is evolving*
- *Low business capability to support delivery and maximize value*

CLEAR WINNERS

- *High business impact on objectives and priorities*
- *High ROI, lower time to value, and low ethical concerns*
- *Low technology / delivery complexity and technology to implement is mature*
- *High business capability to support delivery and maximize value*

NOT A FOCUS NOW

- *Low business impact on objectives and priorities*
- *Low ROI, higher time to value, and ethical implications*
- *High technology / delivery complexity and technology to implement does not exist*
- *Low business capability to support delivery and maximize value*

QUICK HITS

- *Moderate business impact on objectives and priorities*
- *Moderate ROI, higher time to value, and some ethical concerns*
- *Low technology / delivery complexity and technology to implement exists*
- *High business capability to support delivery*

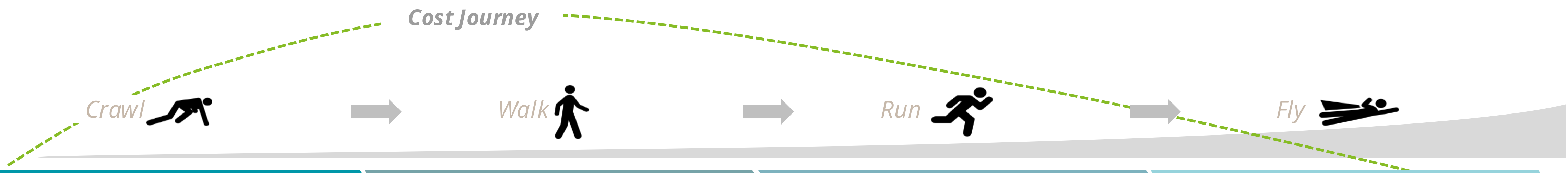
Low — High

FEASIBILITY

What is the likelihood of a successful implementation by technical complexity and functional alignment?

Phase 2 Implement: Delivering from idea to production

Exploration and application can be advanced rapidly in companies with a clear agenda and purpose



Explore	Augment	Elevate	Industrialize
3 – 6 months	*6 – 12 months*	*12 – 24 months*	*24+ months*

Explore
- Proof of concept/s
- Sandbox environments
- AI Capability development
- Risk evaluation
- Data security & controls

Augment
- Production MVPs
- Technology optimization
- Standardized deliver approach
- Governance and sustainment
- Talent Augmentation

Elevate
- Foundry model/s
- Enterprise adoption
- Functional enhancements
- Org & Talent realignment

Industrialize
- Scaled AI function
- Enterprise enablement
- Business model evaluation
- New businesses / revenue streams

Key Enablers

- *Well-defined strategy and approach w/ technology partners*
- *Technology & data maturity, availability & quality*
- *Innovation mindset, and willingness experiment and learn*
- *Talent Development (Data Science, LLMs)*

Phase 3 Operate: Application integrity and maintenance

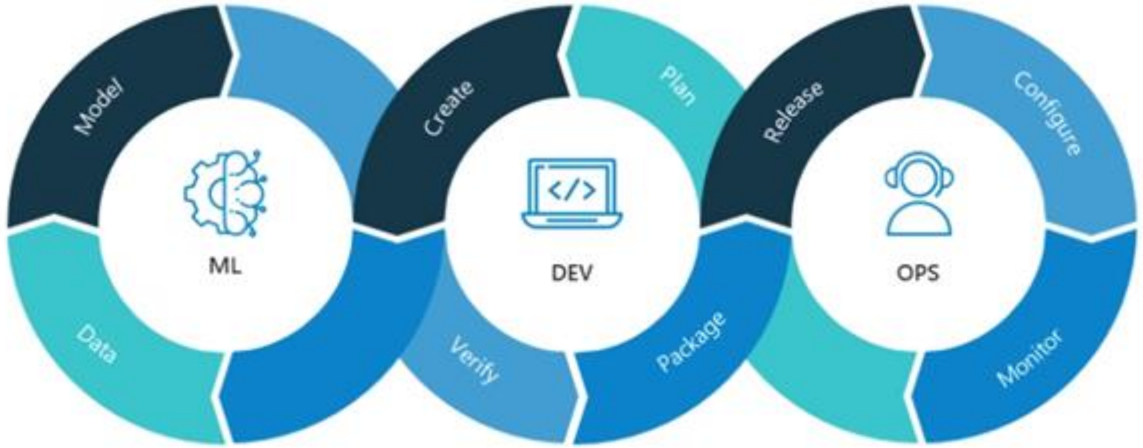
MLOps and DevOps ensures models are effectively developed, deployed, maintained, and benchmarked

What is MLOps and DevOps?

It is a combination of a robust architecture, set of tools and workflows to ensure that **models are tracked through cycles of experiments, tuning and retraining jobs in development and then maintained and benchmarked as they are rolled out to production.**

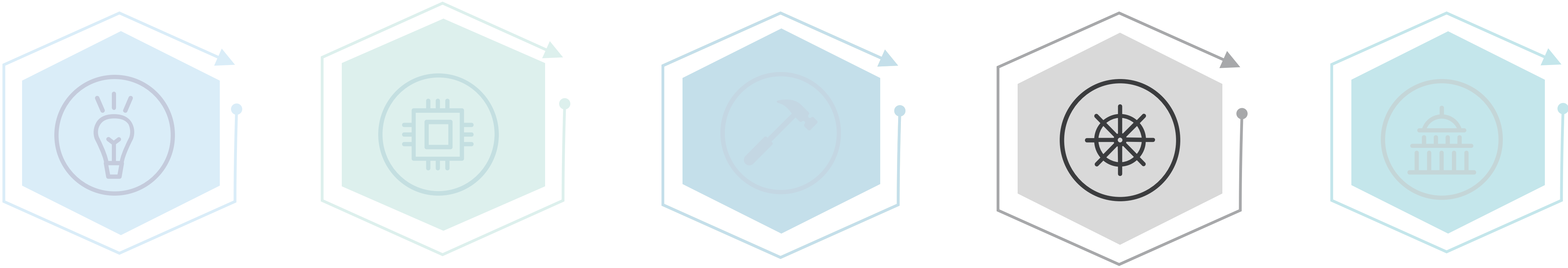
What value does it provide me?

Implementing MLOps and DevOps systems for selecting high-value data for low-resource languages, retraining models based on specific conditions, and tracking model and dataset lineage will reduce workload and maintenance requirements. These systems enhance efficiency and ensure models remain accurate and up-to-date.



AI/GenAI dimensions

There are multiple considerations and dependencies when initializing AI/GenAI Opportunities



Strategy

Define the organizational GenAI vision & guiding principles in line with broader business strategy and activate the capabilities to realize this vision

Technology

Enable the technology stack with next-gen architecture design and ensure quality data is available for GenAI to work

Delivery

End-to-end GenAI solution and capability delivery in alignment with GenAI vision and business value realization

Talent, Organization, and Culture

Create channels for efficient training and comms. Along with an operating model to ensure rapid propagation of GenAI across the enterprise

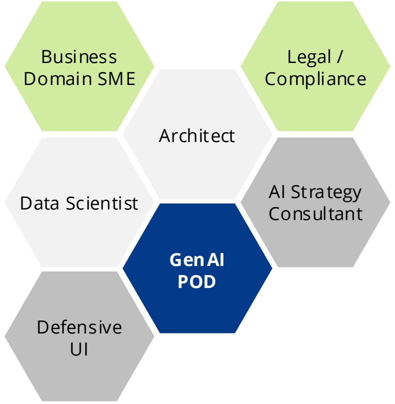
Governance

Define the guardrails and controls to mitigate against GenAI risks and define decision rights for the organization to enable stakeholders to deliver capabilities against defined standards

Talent: Different projects require different team compositions

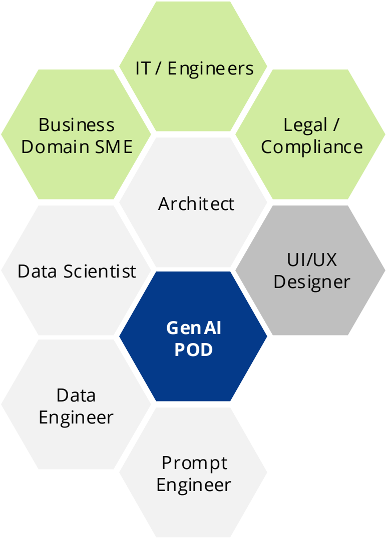
Advise | *Implement* | *Operate*

Strategy



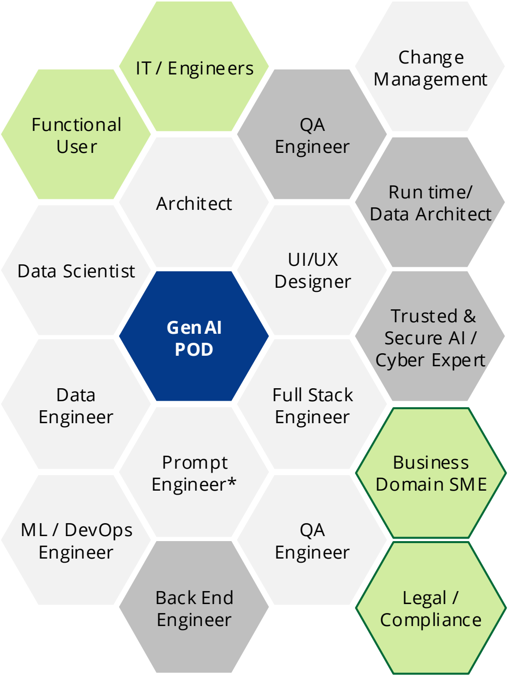
Identify use cases with focus on impact and value realization, while considering technical scalability and user adoption without legal ramifications

POC & Pilot



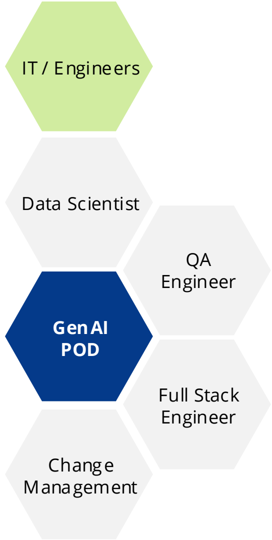
Proof of concept to understand impact, scalability consideration, and gather user feedback while ensuring approvals from risk & compliance teams

Integrated Solution



Scaled product that serves end-to-end goal. Full fledge user adoption and real value generation

Operate



Ensure smooth operation of platform and provide bug resolution support

Illustrative Roles: Client | Part Time / Flex | Core Team | Roles for MVP

Green indicates where the number of people required in the role could be higher based on the nature and complexity of the use case

Roles and Responsibilities

Exploring the critical roles for project success

Persona	Communication	Domain Knowledge	Machine Learning	Software Engineering	Compute Infrastructure
AI PM	Full time	Full time	Medium	Low	Low
DS	Low	Low	Full time	Medium	Low
DE			Low	Full time	Full time
ML Eng			Full time	Full time	Medium
ML Ops			Low	Low	Full time
SE			Low	Full time	Medium
SME	Full time	Full time	Low		

COMMITMENT PER ROLE: Low · Medium · Full time

SFL Scientific
a Deloitte business

1. AI Project Manager
Technical resource ensuring project delivery and hitting business goals.
Requirements: Business appreciation, technical communication, and deep AI understanding. Ideally, former DS/MLE background.

2. Data Scientist
Core development of models, validation, and experiments.
Requirements: Analytical background with core skills in ML. Typically MS/PhD in STEM.

3. Data Engineer
Design, architect, and create data pipelines for solutions.
Requirements: Cloud and on-prem hardware expertise.

4. ML Engineer
Train, deploy, optimize, and maintain large-scale models in production.
Requirements: Mix of SE, ML, and DE. Typically, SE and/or STEM backgrounds.

5. MLOps Engineer
Build infrastructure to make models easier to deploy, more scalable, and maintainable.
Requirements: Mix of DE/DevOps/MLE backgrounds.

6. Software Engineer
A traditional software developer that hardens products and deliverables.
Requirements: Standard software development skills. Typically, CS background.

7. SME
Business domain expert that understands the capabilities of ML broadly.
Requirements: Deep understanding of the industry, domain, problem statement, and ROI.

8. Prompt Engineer*
Engineer that injects intent into models. This may be a part-time role but will require understanding models and how they are trained.
Requirements: ML and model-specific knowledge and experience.

Industry Applications

Generative AI Projects Over The Years

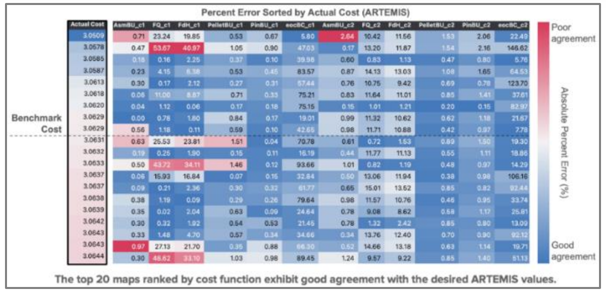
1 Pattern Configuration Generation for Nuclear Reactor

Configuration in nuclear power rods dictate the reactor efficiency and operational expenses. Building configurations that meet specifications is a costly manual process.

Generative AI techniques were utilized to refine nuclear reactor load patterns to prioritize and augment neurotronic efficiency over conventional approaches.

Broader Applications

AI-driven configuration optimization is widely applicable to many domains. Pattern generation to fit certain constraints is useful from everything from warehouse optimization to kitchen modelling.



The top 20 maps ranked by cost function exhibit good agreement with the desired ARTEMIS values.

+ Reduce human burden to manually calculate load configuration patterns and increasing workflow efficiency

+ Optimize sensitive load patterns by 1-2% in the effect of $1.3m+ annually

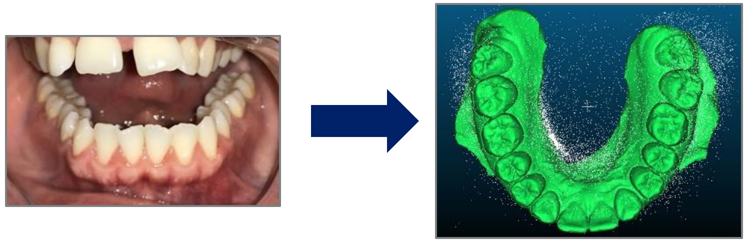
2 3D Model Generation for Tele-dentistry

Telehealth is driving the digital transformation of once-exclusive in-person services. Teledentistry, for example, often relies on physical mapping of patients' teeth, which can be inaccessible.

To overcome this, a deep learning model was developed that reconstructs 3D representations of oral cavities using stock images of upper and lower jaws.

Broader Applications

Generating 3D representations using images taken from phones and applications paves the way for intelligent design of products and spaces.



+ Decreased operational costs by eliminating the need for physical presence and associated materials, as well as related overhead expenses

+ Expanding customer reach to increase product treatment accessibility

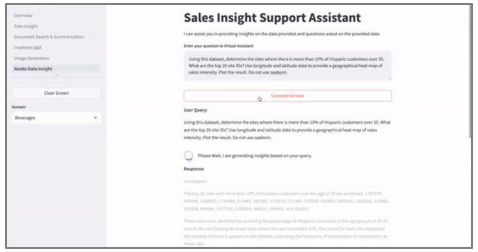
3 Sales Account Generative AI Sales Assistant

Businesses depend on comprehensive access to and understanding of their data to effectively manage diverse products and brands.

Generative AI can empower sales teams to extract insights, trends, and observations from data more efficiently by utilizing a LLM chatbot with a RAG-pattern architecture, expediting tasks such as document analysis and sales data calculation.

Broader Applications

General chat bots that can aggregate information, generate information, perform baseline analytics can be used in a multitude of different business units from costumer chatbots, support, sales assistants or internally to improve productivity and effectiveness.



+ Increased sales revenue through targeted product recommendations generated by AI-driven insights

+ Reduced document analysis allowing sales team to focus more on strategic initiatives and client relationships

Use Cases | What made these good AI use cases

SFL Scientific worked with a various clients to improve operations and deliver value using GenAI

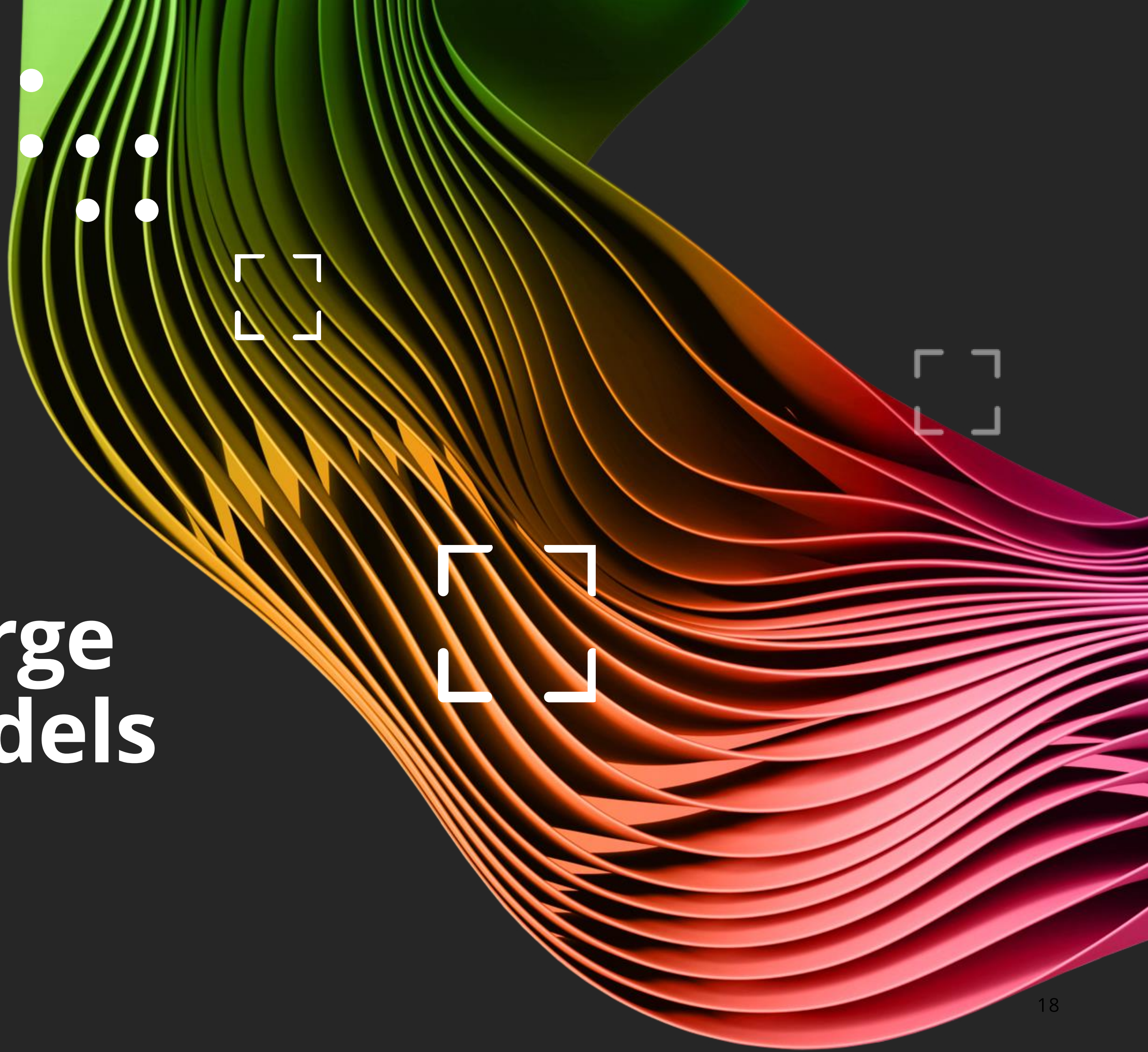
	A Well-Defined Problem Statement	**Too Complex for By-Hand Coding**	**Had High Quality, Representative Data**	**Have the best Return on Investment**
Good Use Cases	1. Optimize novel configurations to maximize yield 2. Build 3D molds from photos that fit space 3. Answer user queries through chatbots	1. Novel configurations are found by trial and error and more of an art by humans 2. 3D molds have huge variability that cannot be directly codified 3. Many possible Q&A responses	1. Historic configurations & their associated yields 2. 2D images and actual 3D molds for comparison 3. Questions and ideal answers that are hand annotated for training/validation & data for context about sales	1. Hours and days of domain SME engineers saved with higher yield 2. Impossible to perform otherwise, opening entire tele-market 3. Reduces human review and analysis by 20%
Bad Use Cases	1. Improve reactor efficiency 2. Increase 3D denture sales 3. Engage customers	1. If there was a smaller set of configurations 2. No variability in molds (e.g. gum shields) 3. Limited set of questions / responses	1. No recorded yield data 2. Only 2D images without 3D mold data 3. No access to SME for validation or no domain specific datasheets / tables	If the use case was a simpler problem traditional AI (known configuration yield estimation), programming (only 3 size molds), or human review would have been more cost effective (limited number of complex questions per day) to implement than GenAI solutions.

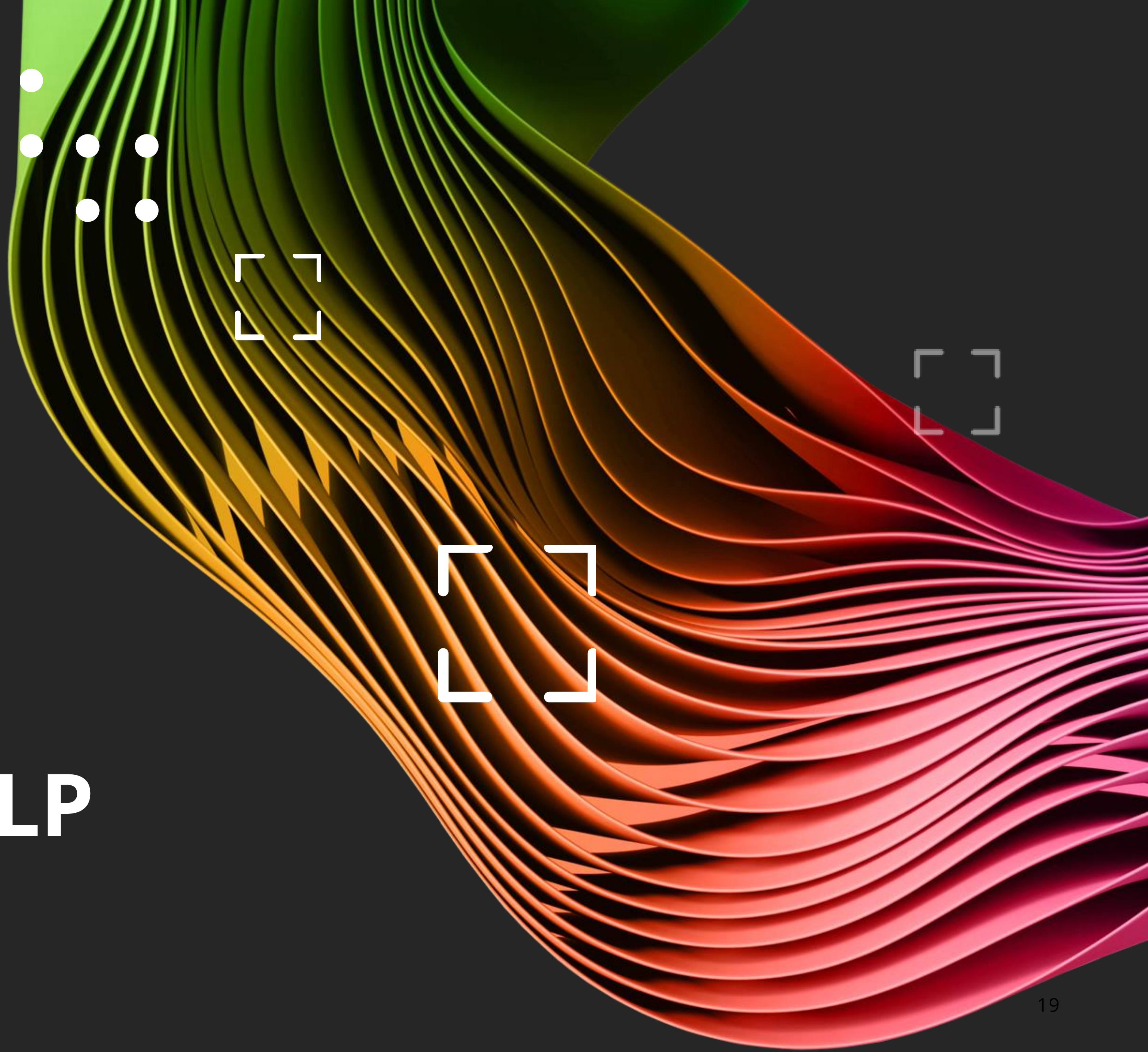
GenAI use cases across industries

Understanding the needs of the consumer will help drive use case strategy

Industry / Modalities	Energy, Resources, and Industrials	Financial Services and Insurance	Government and Public Services	Tech, Media, and Telecom	Life Sciences and Healthcare	Consumer
Audio	**Field Virtual Assistant** Enable field agents to access best practices and repair information using natural language while hands-free	**Retail Banking Transaction Support** Provide human-like support for complex retail transactions including customer applications, questions, negotiations, and more	**Intelligent Agents / Student Office Hours** Provide natural language support for government services and on-demand access to information for students	**Translations, Subtitles, and Descriptions** Translate audio into multiple languages (e.g., subtitle generation) and provide descriptions to visual media content	**Automated Follow-Ups** Ingest clinical notes to identify patients that will need follow-up and create audio messages that can be sent to schedule follow-ups and encourage healthy habits	**Conversational Retail** Provide detailed product support and guidance using human-like chatbots in retail stores focused on specific brands and/or categories
Code	**No-Code Physics-Based Environments** Allow researchers to create highly computational and accurate physics-based models of weather, fluid dynamics, and environments	**Database Search** Query massive financial transaction databases to find specific items and insights using natural language instead of database languages such as SQL	**Knowledge Management** Allow government workers to cluster, search, and filter large amounts of unstructured data from images, video, and text files through natural language	**Original Games Creation** Ideate and code novel computer and video games and accelerate the game testing process	**Clinical Trial Data Processing** Allow researchers to clean up data and generate graphs and insights for clinical trials and approvals processes using natural language	**Marketing Speed** Help marketers build websites and external collateral at the speed of natural language to go-to-market faster with new products and services
Image	**New Product Development** Create detailed schematic drawings of industrial products and parts to aid in new product development and repairs	**Fraud Detection** Generate customer signatures to enhance internal fraud models in areas such as credit card authorization, and summarize potential fraud hotspots	**Infrastructure Mapping** Enhance infrastructure mapping and planning processes by generating detailed plans and iterating using natural language	**Semiconductor Chip Design** Iterate and enhance designs based on performance parameters and reduce the development life cycle time	**Improved Medical Imaging** Generate large sets of synthetic medical images to train imaging algorithms to better identify abnormalities as well as train clinicians to better identify issues	**Product Photography and Details** Generate details and ultra-realistic photographs of new and existing products in different environments
Text	**Technical Document Summarization** Extract information from detailed documentation and synthesize field-reports in specific formats	**Customer Due Diligence Reporting** Generate reports on new customers such as KYC processes and summarize them for employees to action and make decisions for customer onboarding	**Intelligent Case Management** Parse complex government case files for actionable details which are then summarized for rapid comprehension and used to generate reports	**Cybersecurity Threat Detection** Summarize areas of high-risk, answer questions, and generate executive reports for malware, anomalies, and potential threats	**Medical History Summary** Summarize patient demographics, medical history, allergies, medications, and other relevant details from EHR clinical notes to aid hospital intake	**Personalized Supermarket** Create custom meal plans and shopping lists fine-tuned for each buyer/family specific to the store and what's available
Video (Early Stages)	**Event Identification** Absorb live video feeds of the end-to-end production chain and answer specific questions about processes and events	**Claims Footage** Review video footage of claims (e.g., car crashes) to pull out summaries and eventually generate new video of potential crash scenarios	**Citizen Support** Provide hyper-realistic, life-like personal assistants in places such as the airport, DMV, border patrol and immigration, to support citizen needs	**Virtual Anchors** Create virtual on-air anchors for high-demand events (e.g., live sports) where there are not enough people to support across languages/borders	**Digital Therapy** AR/VR content generation for assets required in digital therapy or virtual environments	**Commercial Brainstorming** Rapidly brainstorm with generated video and video storyboards for pieces such as television/online commercials
3D Models & Data	**Geological Assessments** Assess both real and synthetic data for oil exploration and the likelihood of finding resources	**Financial Model Enhancement** Generate synthetic data to improve and enhance financial models and pressure test an institution's liquidity and processes	**Disaster Recovery and Planning** Support urban planners and disaster recovery teams with synthetic data (e.g., traffic, population, 'what-if scenarios') to aid in planning and preparation	**Telecom Network Maintenance** Train digital twins on synthetic data to help identify network faults and provide remediations for on-field technicians	**New Drug Discovery** Generate the structure and function of proteins and biomolecules, accelerating the creation of new drug candidates	**Rapid Product Design / Consumer Preferences** Accelerate product prototyping lifecycle through creation of unique and high-fidelity product mock-ups, and create synthetic behavioral data of buyers

GenAI and Large Language Models (LLMs)





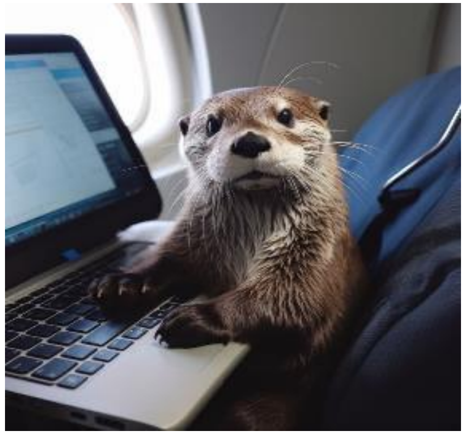
Evolution of NLP

Pace of Advancement

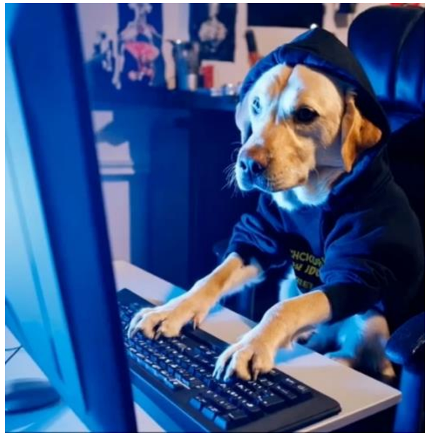
October 2022

November 2022

May 2023



Feb 2024

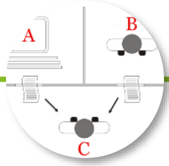


Testing the prompt: “An otter on an airplane using Wi-Fi”

Testing the prompt: “Labrador Hacker”

Stanford University has identified that AI is moving faster than Moore’s Law, doubling in power every 3 months.

Pace of Advancement



1950
Alan Turing tests for machine intelligence

1964
Chatbot ELIZA is invented

1997
AI wins against top human in chess

2011
iPhone prompts daily use of AI

2014
Alexa becomes a home-based virtual assistant

2016
Artificial creativity, AlphaGo, is introduced

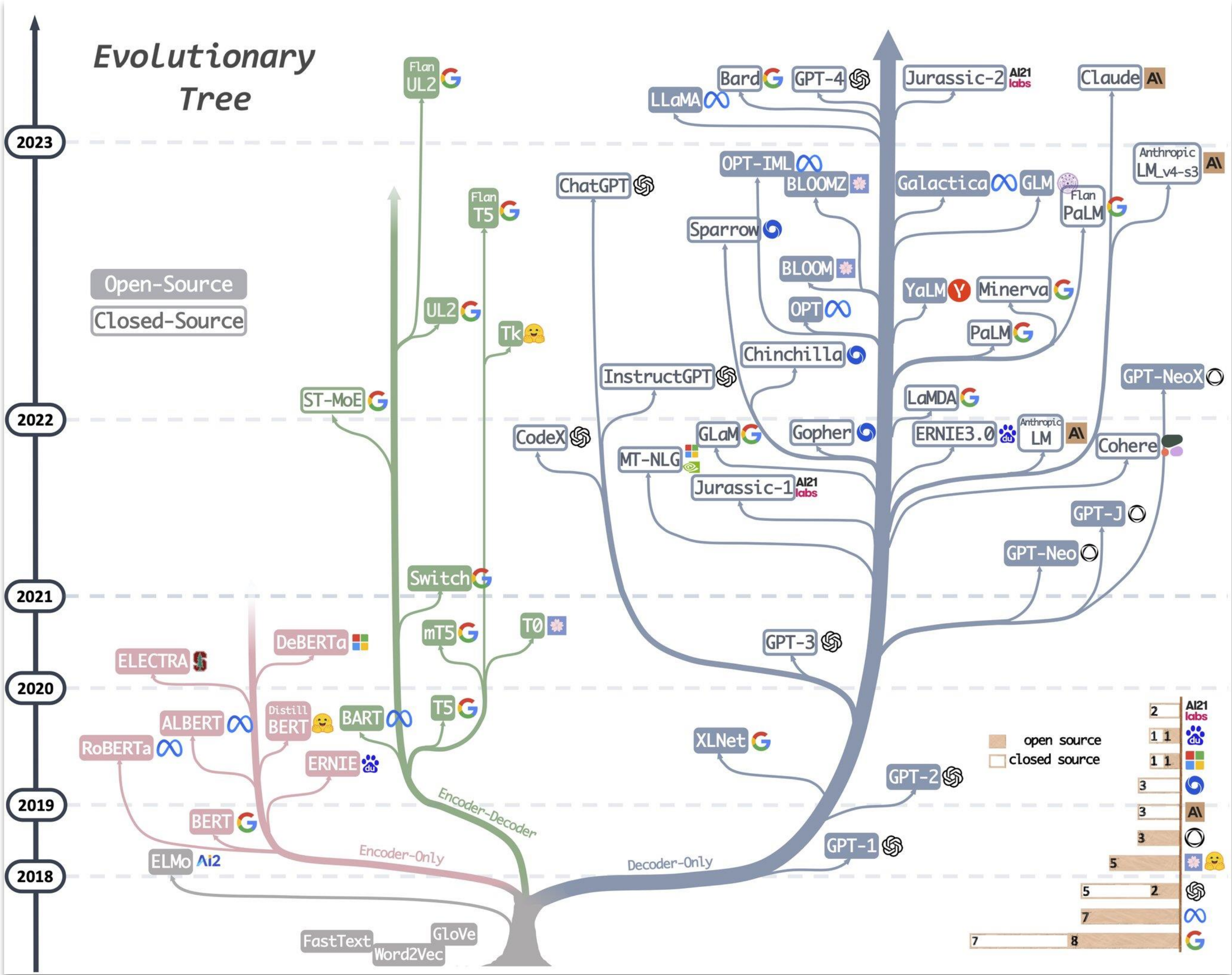
2022
Lensa creates mass social media adoption of GAI

2022
ChatGPT gains mass awareness

2023
Bing begins large-scale integration of GAI into everyday tech

AI spending will exceed **$154B in 2023** and **will double by 2026** with Microsoft, Google, and AWS leading the way

LLM Evolution Tree



Vectorization-- Bag of Words

Common Approaches:

- Count-vectorization
- TFIDF (Term Frequency-Inverse Document Frequency)

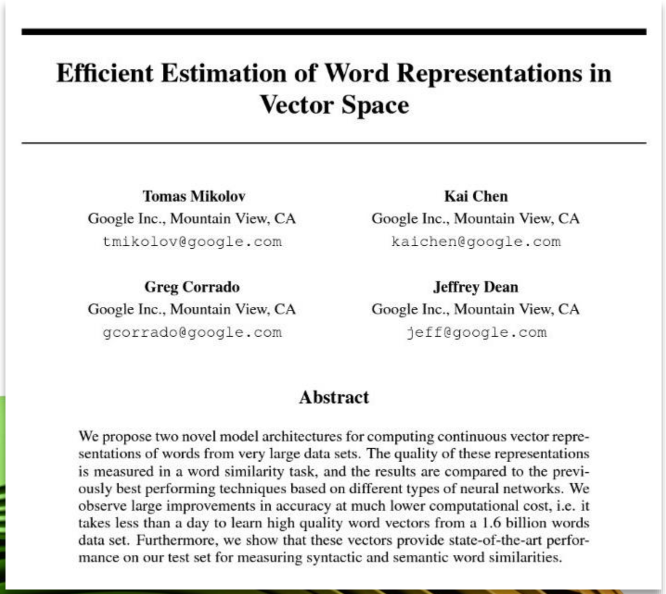
	she	loves	physics	is	coolest	a	good	person	math	are	the	best	second
She loves math, math is the coolest.	1	1	0	1	1	0	0	0	2	0	1	0	0
She loves physics, physics is the second best.	1	1	2	1	0	0	0	0	0	0	1	1	1
She is a good person.	1	0	0	1	0	1	1	1	0	0	0	0	0

Vectorization – Dense representations

State-of-the-art (self)-supervised algorithm to create dense embeddings of semantically similar words

Common Methods:

- CBOW
- Skip-gram



Efficient Estimation of Word Representations in Vector Space

Tomas Mikolov
Google Inc., Mountain View, CA
tmikolov@google.com

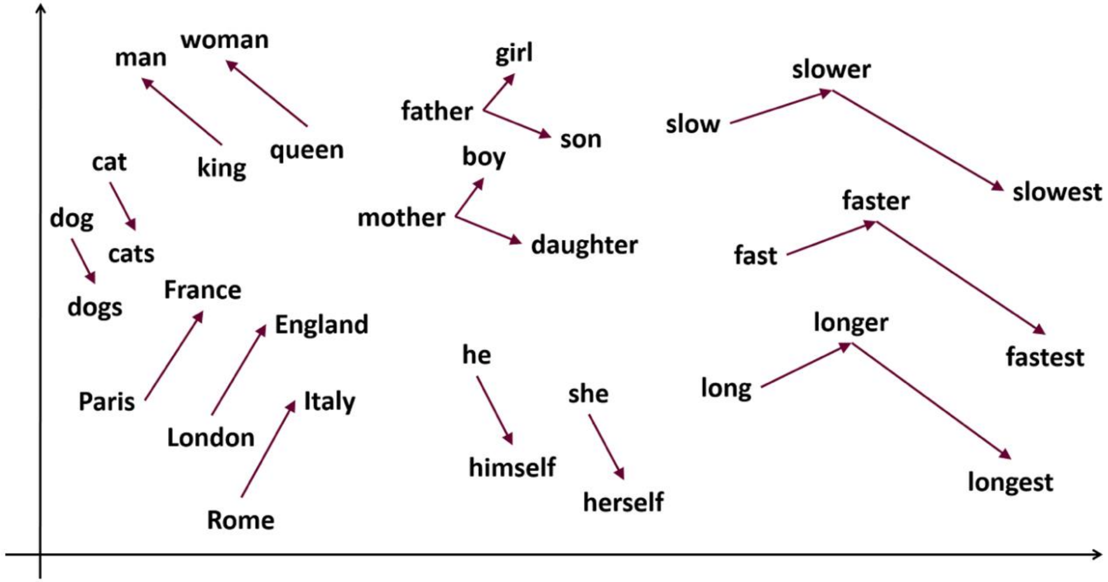
Kai Chen
Google Inc., Mountain View, CA
kaichen@google.com

Greg Corrado
Google Inc., Mountain View, CA
gcorrado@google.com

Jeffrey Dean
Google Inc., Mountain View, CA
jeff@google.com

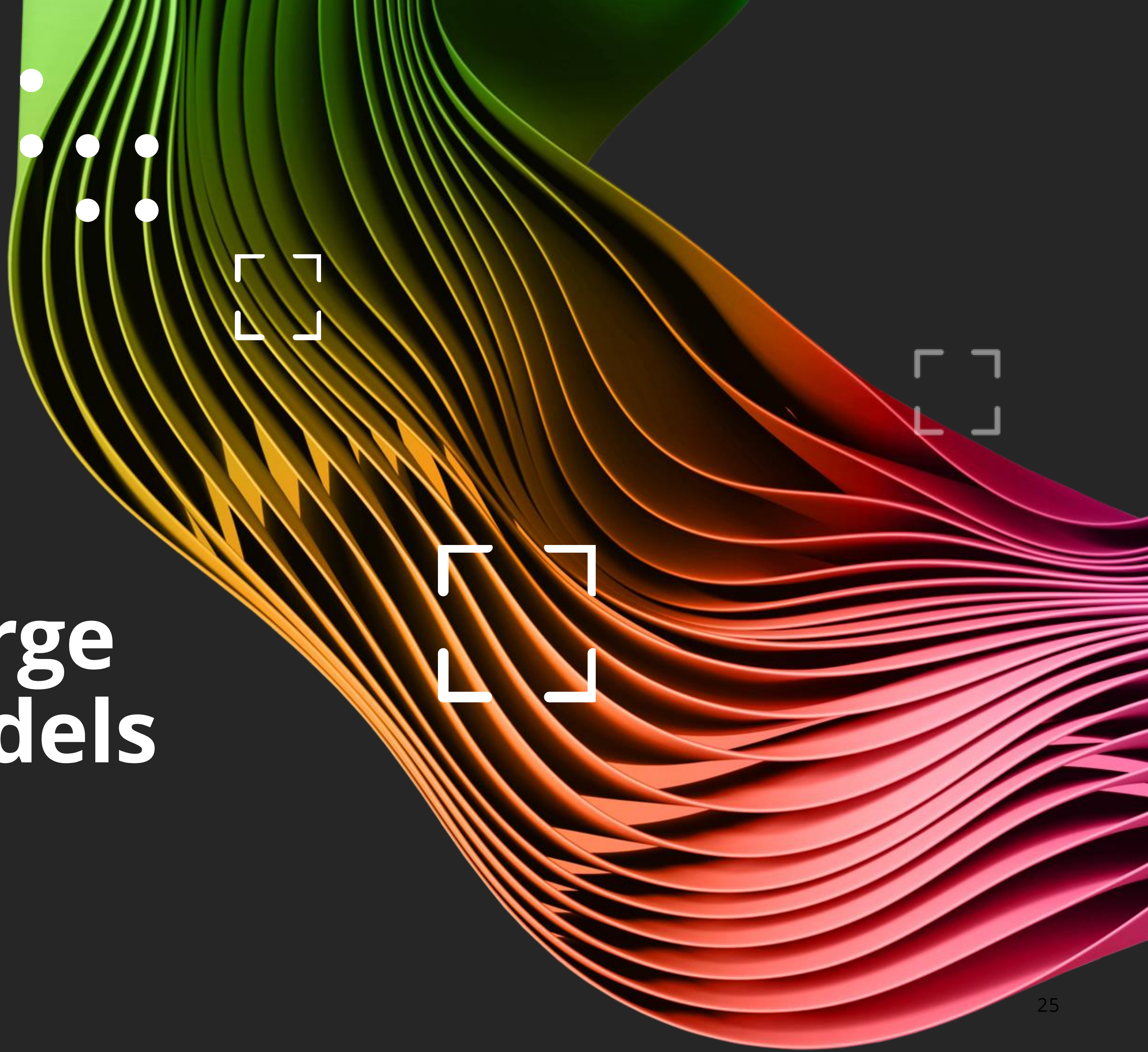
Abstract

We propose two novel model architectures for computing continuous vector representations of words from very large data sets. The quality of these representations is measured in a word similarity task, and the results are compared to the previously best performing techniques based on different types of neural networks. We observe large improvements in accuracy at much lower computational cost, i.e. it takes less than a day to learn high quality word vectors from a 1.6 billion words data set. Furthermore, we show that these vectors provide state-of-the-art performance on our test set for measuring syntactic and semantic word similarities.



Source: https://medium.com/@dube.aditya8/word2vec-skip-gram-cbow-b5e802b00390

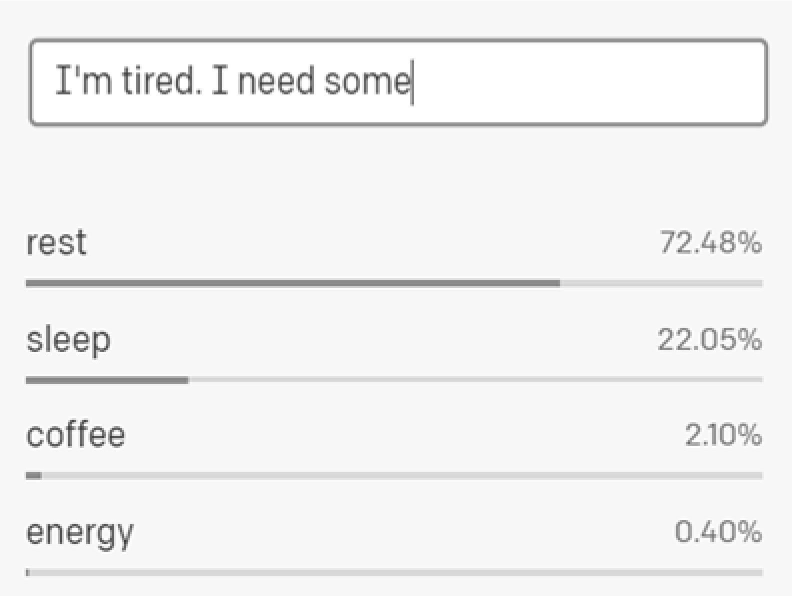
GenAI and Large Language Models (LLMs)

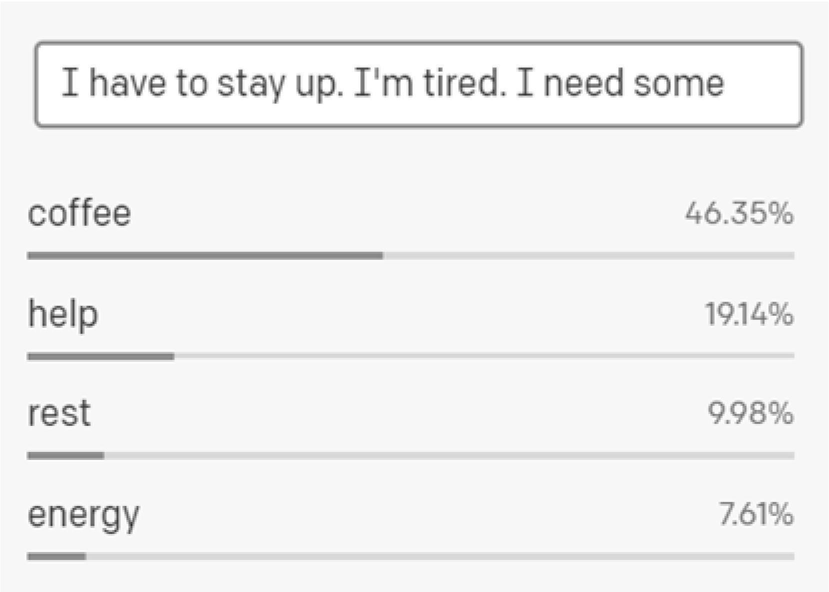


What is ChatGPT?

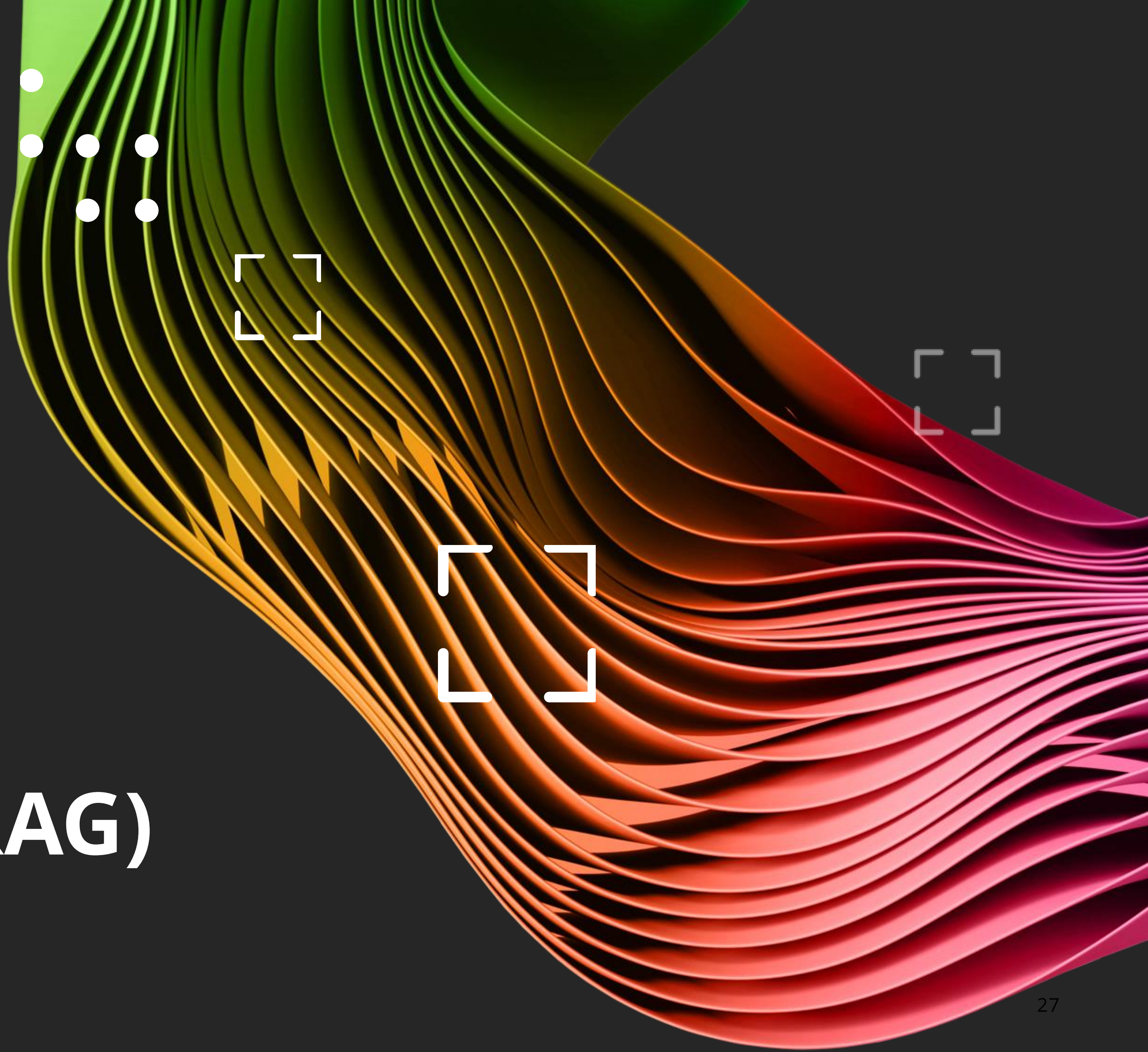
Language Models are
A probability distribution over sequences of words: $P(w_n \mid w_1, w_2, \ldots, w_{n-1})$

Given a **prompt**, a language model **predicts words** that are **likely to follow**:





Source: [Borealis AI "A High-level Overview of Large Language Models"]



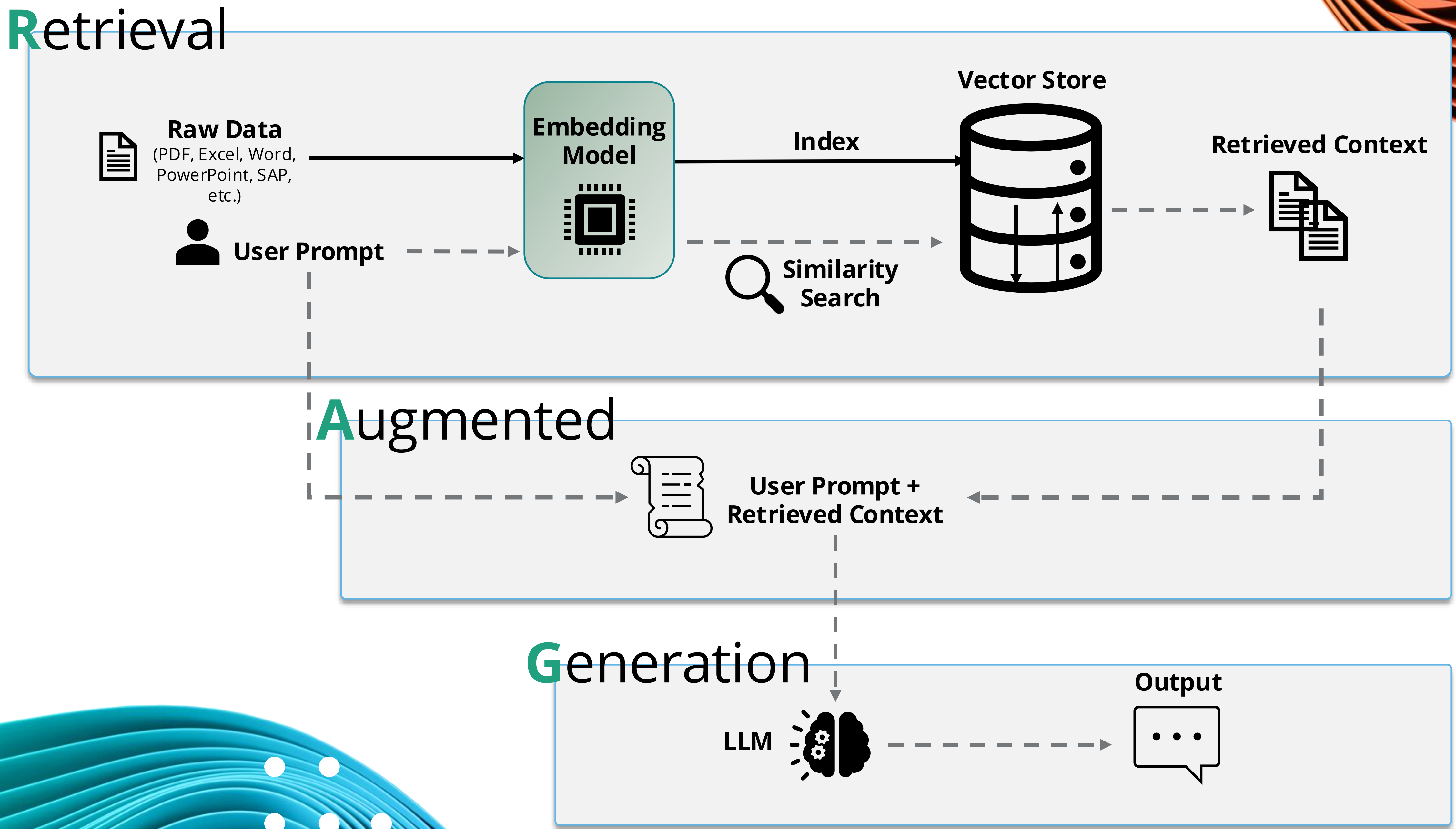
Retrieval-Augmented Generation (RAG)

What is Retrieval-Augmented Generation (RAG)?

Retrieval Augmented Generation (RAG) is a technique in the field of natural language processing (NLP) that **combines retrieval-based** and **generation-based** approaches to improve the performance of language models, particularly for tasks such as question answering, information retrieval, and text generation

Advantages of RAG

- **Improved Accuracy:** Provides more accurate and contextually relevant responses
- **Scalability:** Retrieval component allows the model to access a vast amount of information without needing to encode all knowledge within the generative model itself
- **Flexibility:** RAG models can be adapted to various tasks, including question answering, summarization, and conversational AI


Retrieval
Raw Data
(PDF, Excel, Word, PowerPoint, SAP, etc.)
User Prompt
Embedding Model
Index
Vector Store
Similarity Search
Retrieved Context
Augmented
User Prompt + Retrieved Context
Generation
LLM
Output

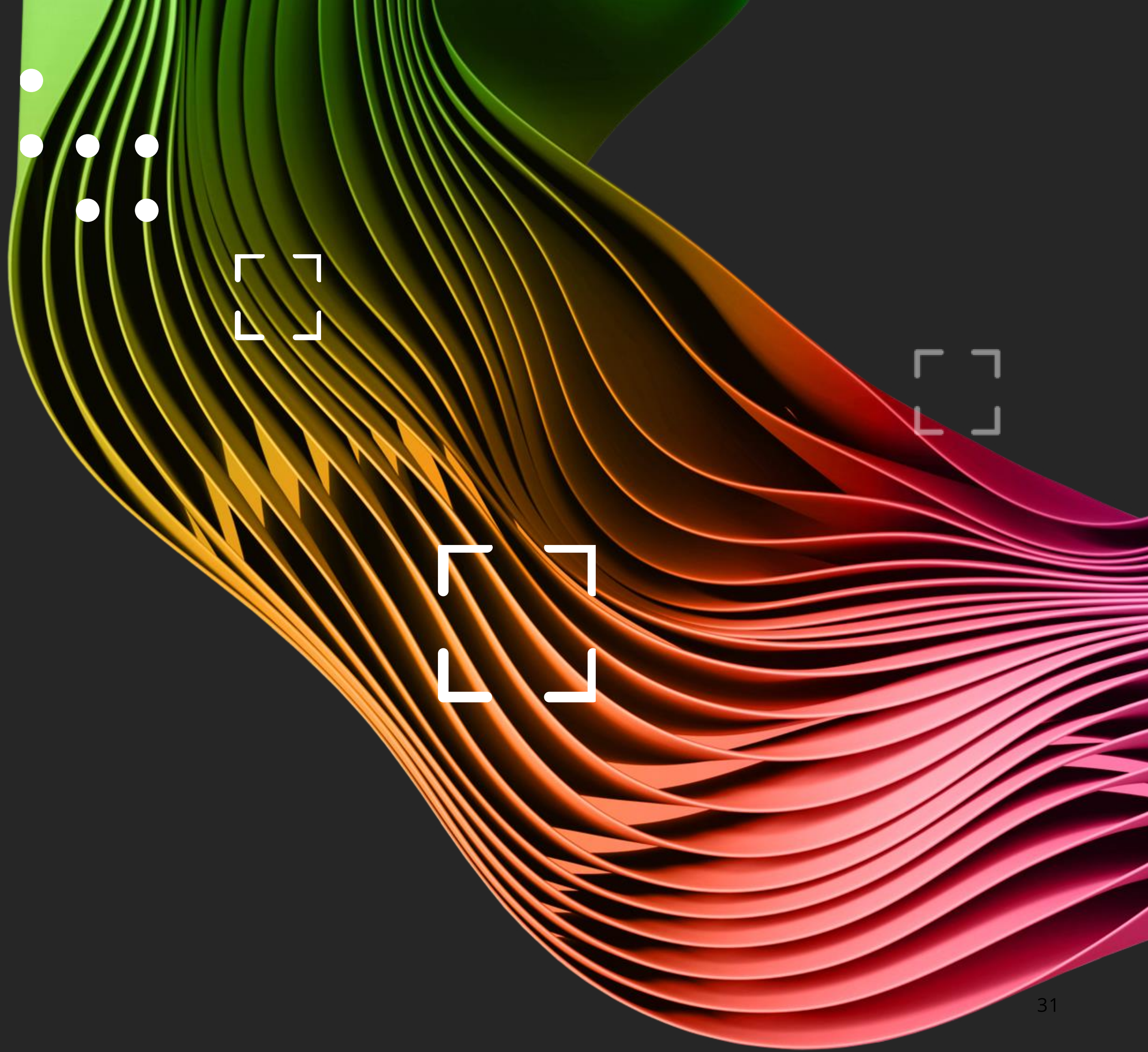
Limitations of LLM centric applications

LLMs employ neural networks with numerous layers to process extensive textual data, learning intricate patters and relationships embedded in language.

However, there are **limitations** to what LLMs are capable of

- Unable to retain information between interactions
- LLMs can't update their knowledge base in real-time
- Often hallucinate and generate responses that are nonsensical, illogical, or irrelevant to the query
- Output quality is dependent upon end user
- Not grounded in reality, no embodied experience of actions in an environment

ReAct Framework



What is the ReAct Framework?

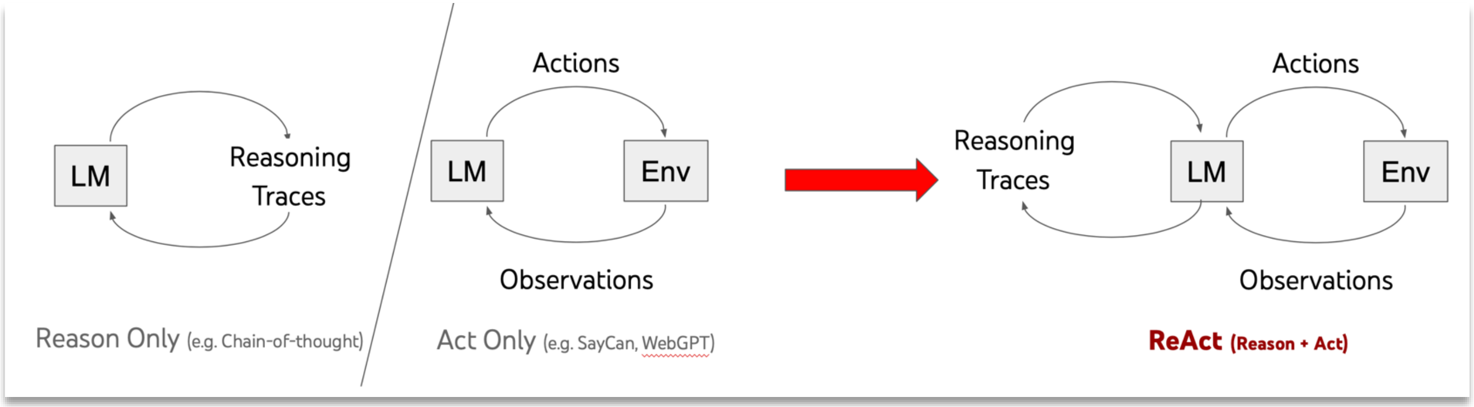
Reasoning + **Acting** = **ReAct**

ReAct Framework is a framework where LLMs are used to generate both *reasoning traces* and *task-specific actions* in an interleaved manner

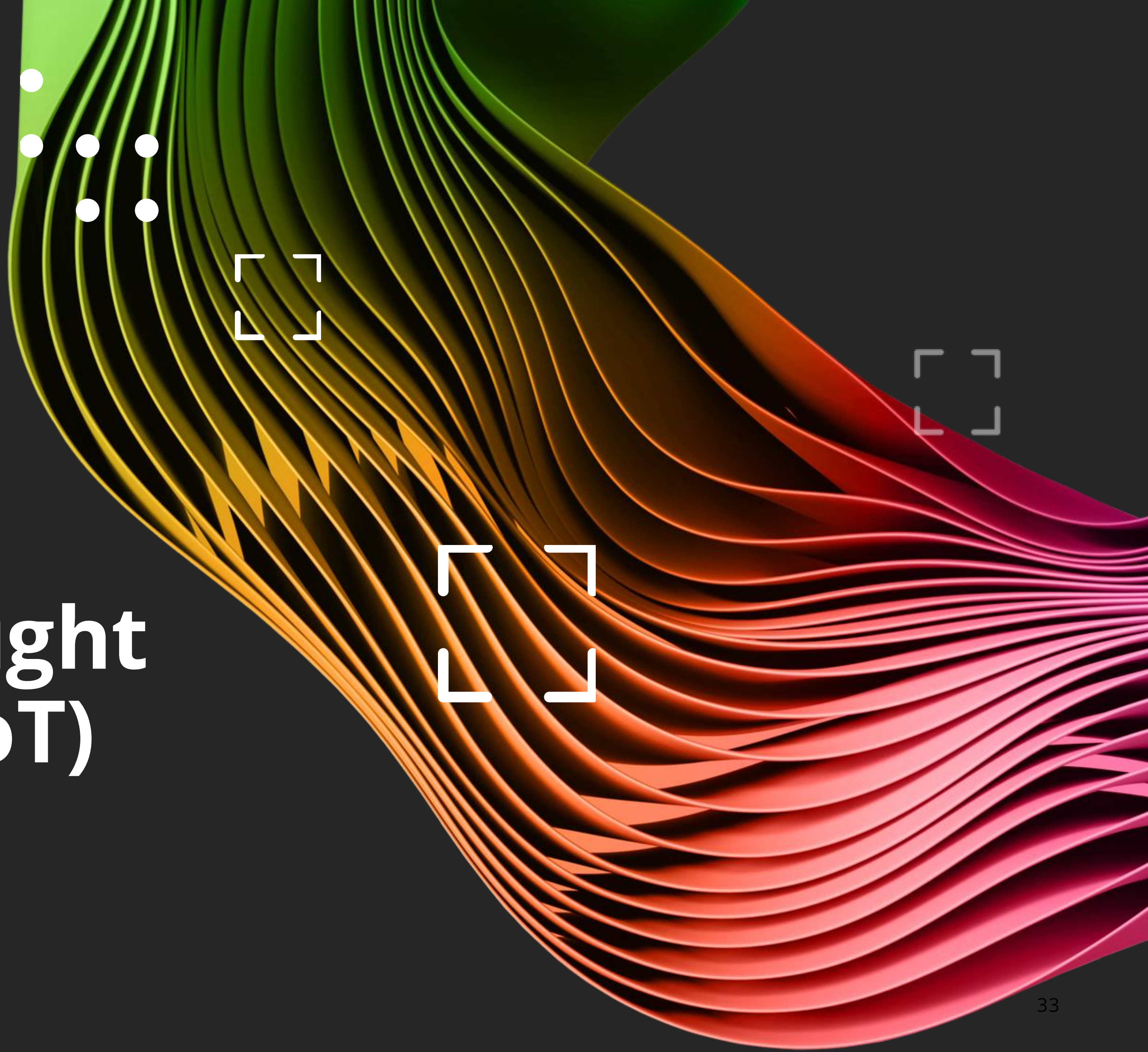
Key Components

- Induce
- Track processes
- Update plans
- Handle exceptions

- Interface with external sources (i.e. knowledge bases or environments)



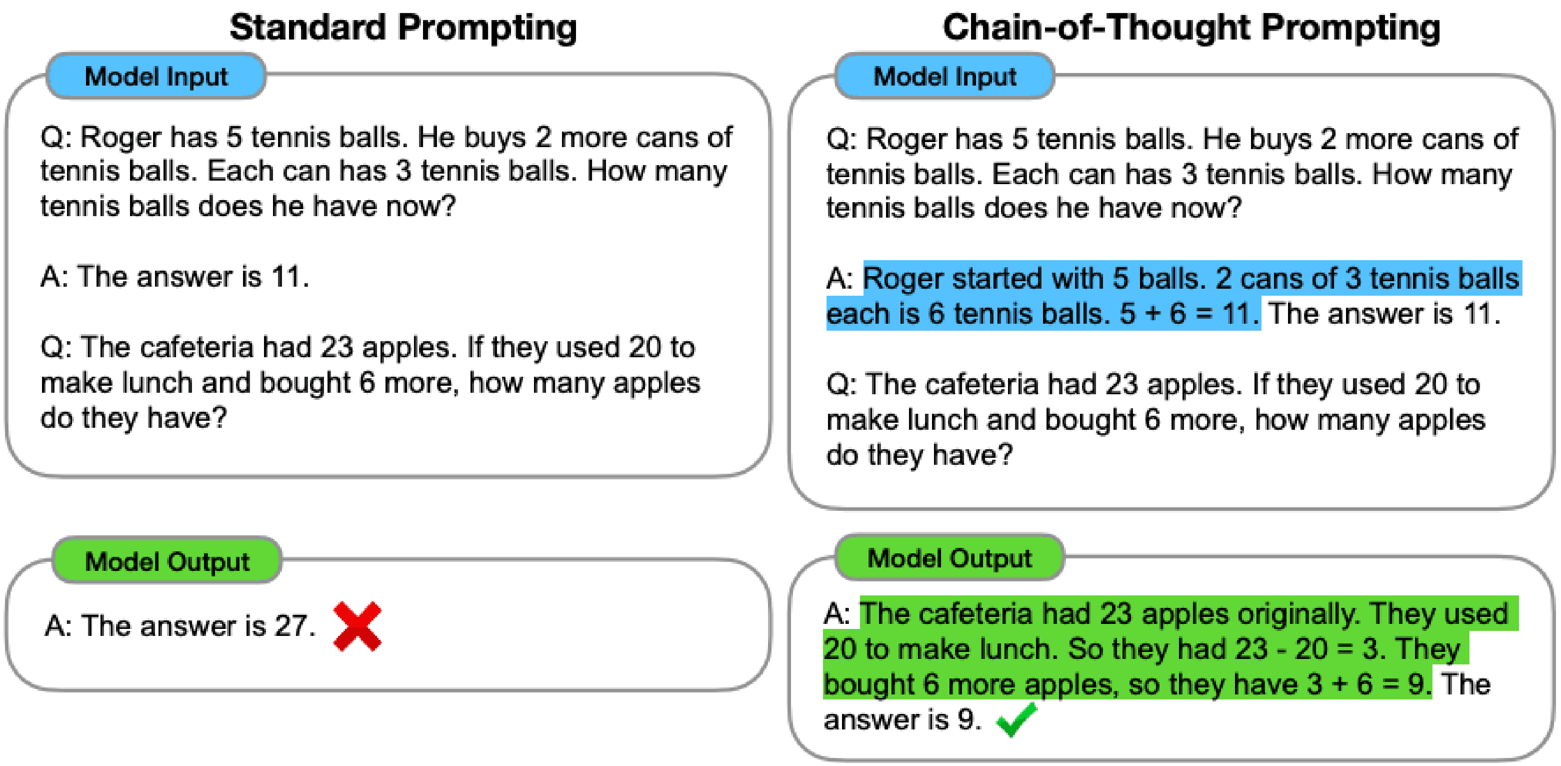
Chain of Thought Reasoning (CoT)



What is Chain of Thought (CoT)?

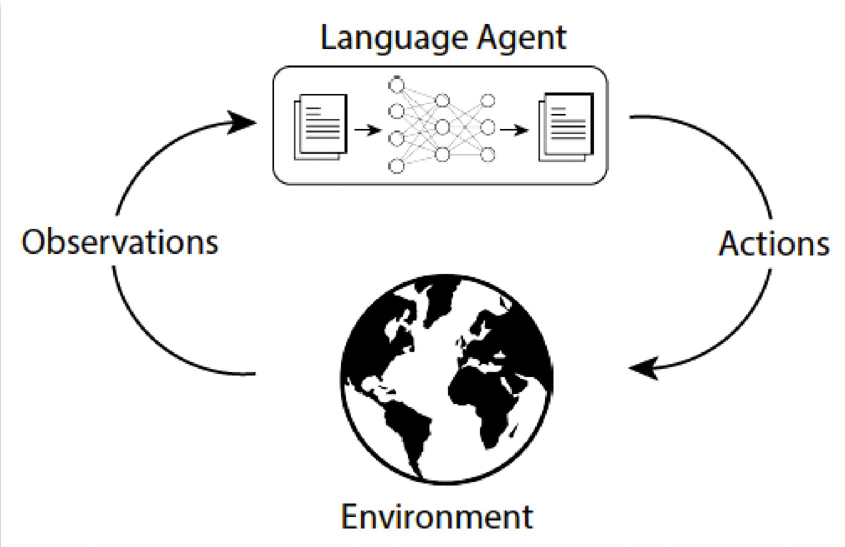
Chain of Thought (CoT)

Prompt engineering technique that aims to improve language models’ performance on tasks that require logic, calculation, and decision-making



Language Agents

What are Language Agents?



The Language Agent is in a feedback loop with its environment.

At time step t, an agent has an observation, $o_t \in \mathcal{O}$
and makes an action, $a_t \in \mathcal{A}$ based on the policy, $\pi(a_t|c_t)$
and the context, $c_t = (o_1, a_1, o_2, a_2, \ldots, o_{t-1}, a_{t-1}, o_t)$

Language Agents can store information via:

- Context window
- Implicit knowledge in the LLM weights (training date, hallucinations)
- Information retrieval from a knowledge bank (RAG)

This is a <u>reactive</u> process. The agent is always moving from action to action.

Benefits of Language Agents

Operational Efficiency

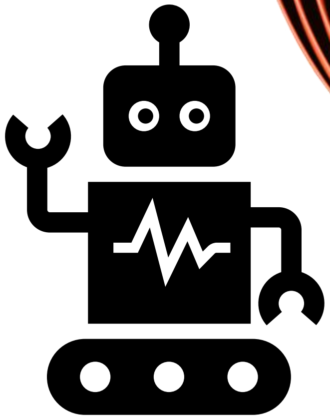
- Can operate ***independently*** and often repetitive tasks such as scheduling meetings, sending reminders, and managing emails
- Boost ***time and cost savings*** by automating routine tasks and interactions

Improved User Experience

- **Improve** user interactions by **tailoring responses** based on individual user preferences and past behavior
- With context awareness, an agent can provide more **relevant** and **timely responses**

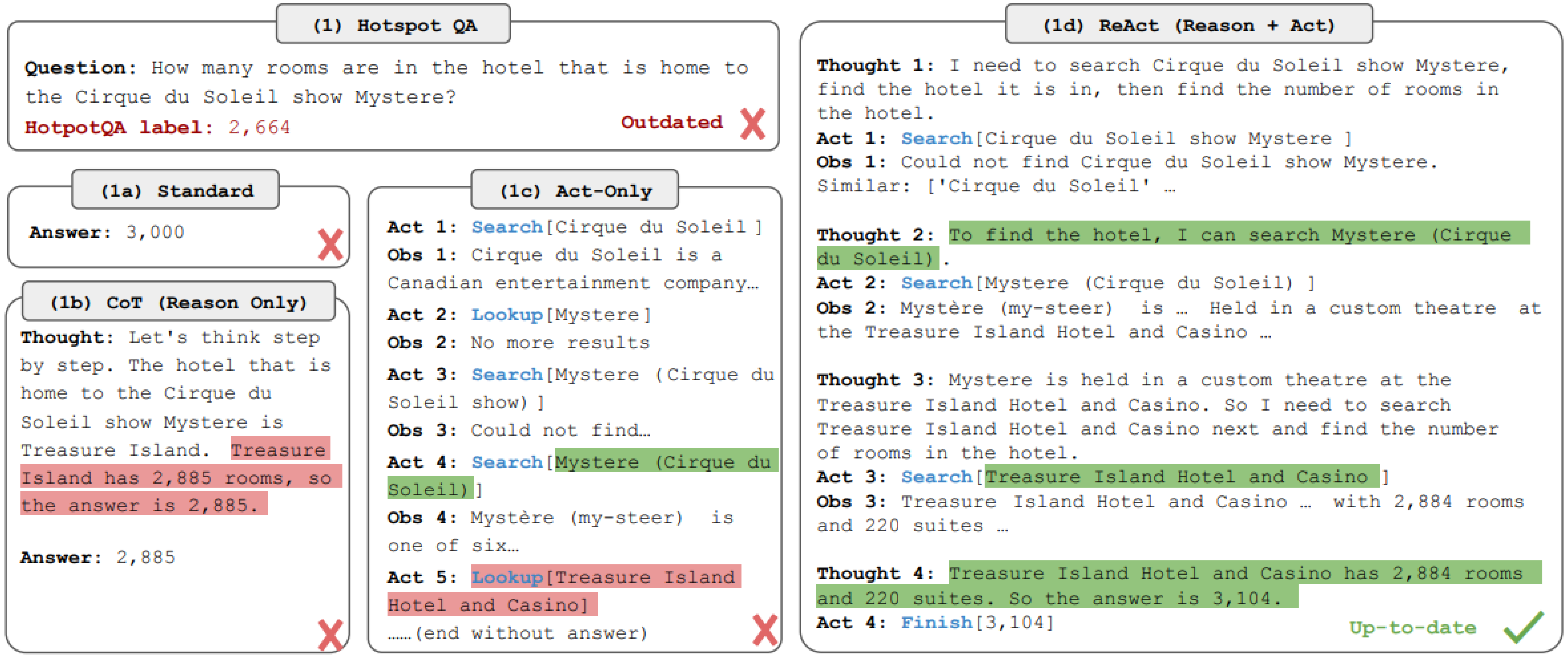
Continuous Improvement

- Receiving feedback on errors or unsuccessful interactions can help lead to more **accurate** and **reliable** responses
- **Increased insights** due to agent's ability to collect and analyze large volumes of data from user interactions



ReAct Prompting | Illustrative Example

Method that uses **few-shot learning** to teach the model how to solve problems. ReAct includes a reasoning element, but the agent is able to interact with its environment through **actions**.



*Examples from "ReAct: Synergizing Reasoning and Acting in Language Models"
Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, Yuan Cao

ReAct Agents | Illustrative Example

Typical language agent:

At time *t*, an agent has an observation, *o* and makes an action, *a* based on the policy, π and the context, *c*

$$o_t \in \mathcal{O}$$
$$a_t \in \mathcal{A}\ using\ \pi(a_t|c_t)$$
$$c_t = (o_1, a_1, \ldots, o_{t-1}, a_{t-1}, o_t)$$

ReAct agent:

The action space expanded to include internal text generation around thoughts.

$$\mathcal{A} \rightarrow \mathcal{A} \cup \mathcal{L}$$
$$c_t = (o_1, \boldsymbol{th_1}, a_1, \ldots, o_{t-1}, \boldsymbol{th_{t-1}}, a_{t-1}, o_t)$$

(2) AlfWorld

You are in the middle of a room. Looking quickly around you, you see a cabinet 6, a cabinet 1, a coffee machine 1, a countertop 3, a stove burner 1, and a toaster 1.
Your task is to: Put some pepper shaker on a drawer.

(2a) Act-Only

Act 1: Go to drawer 1
Obs 1: The drawer 1 is closed.

Act 2: Open drawer 1
Obs 2: You open the drawer 1. The drawer 1 is open. In it, you see a dishsponge 2, and a spoon 1.

Act 3: Go to sinkbasin 1
Obs 3: On the sinkbasin 1, you see a dishsponge 3, a spatula 1, and a spoon 2.

Act 4: Take peppershaker 1 from sinkbasin 1
Obs 4: Nothing happens.

Act 5: Take peppershaker 1 from sinkbasin 1
Obs 5: Nothing happens.

(2b) ReAct (Reason + Act)

Act 1: Think[First I need to find a pepper shaker...more likely to appear in cabinets (1-6), countertops (1-3), ...]
Act 2: Go to cabinet 1
Obs 2: On the cabinet 1, you see a vase 2.

(Here the agent go to cabinet 1, then cabinet 2, 3, then countertop 1 and 2)

Act 6: Go to countertop 3
Obs 6: On the countertop 3, you see a apple 1, a bread 1, ... a pepper shaker 1, and a vase 3.

Act 7: Take pepper shaker 1 from countertop 3
Obs 7: You pick up the pepper shaker 1 from the countertop 3.

Act 8: Think[Now I find a pepper shaker 1. Next, I need to put it in/on drawer 1.]
Act 9: Go to drawer 1
Obs 9: Drawer 1 is closed.

Act 10: Open drawer 1
Obs 10: You open Drawer 1 …

Act 11: Put pepper shaker 1 in/on drawer 1
Obs 11: You put pepper shaker 1 in/on the drawer 1.

*Examples from "ReAct: Synergizing Reasoning and Acting in Language Models"
Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, Yuan Cao

Strengths and Weakness of ReAct

Outperforms both CoT and Actions alone on QA test sets using exact match scoring.

Output is more interpretable as it generates a readable reasoning trace.

Prompt Method	Test Set	
	HotpotQA	Fever
Action-only	25.7	58.9
Reason-only (CoT)	29.4	56.3
Best ReAct Method	35.1	64.6

**Using PaLM-540B

ReAct does not perform better than domain-specific models trained on human annotations.

Retains many of the problems associated with LLM Agents:

- Hallucinations
- Recency bias
- Over-agreeableness
- Encoded biases in parameters
- Error propagation in reasoning
- Limited context storage

*Examples from "ReAct: Synergizing Reasoning and Acting in Language Models"
Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, Yuan Cao

Benefits of using a ReAct framework

Outperforms Imitation and Reinforcement Learning

ReAct outperforms imitation and reinforcement learning methods by a success rate of 34% and 10% respectively[1]

Overcomes Issues of Hallucination

Combines reasoning traces and task-specific actions to ensure that the AI system continuously verifies and updates its information

Dynamic Adaptation

Greater synergy between reasoning traces and task-specific actions

Interact with External Tools

ReAct Framework allows LLMs to interact with external sources (i.e. APIs) to retrieve additional information

General and Flexible

Works for diverse tasks, included but not limited to QA, fact verification, text game, and web navigation

Improved User Interaction

Ability to dynamically problem solve, which allows for more sophisticated interactions with users

Note: 1. Success rates based on two interactive decision-making benchmarks (ALFWorld and WebShop) as performed in this paper: "ReAct: Synergizing Reasoning and Acting in Language Models"

Examples of ReAct agent use cases

Patient Diagnosis and Treatment
Analyze large datasets of medical records, lab results, and imaging studies

Demand Forecasting
Predictive analytics can forecast demand for products based on historical sales data, market trends, and external factors

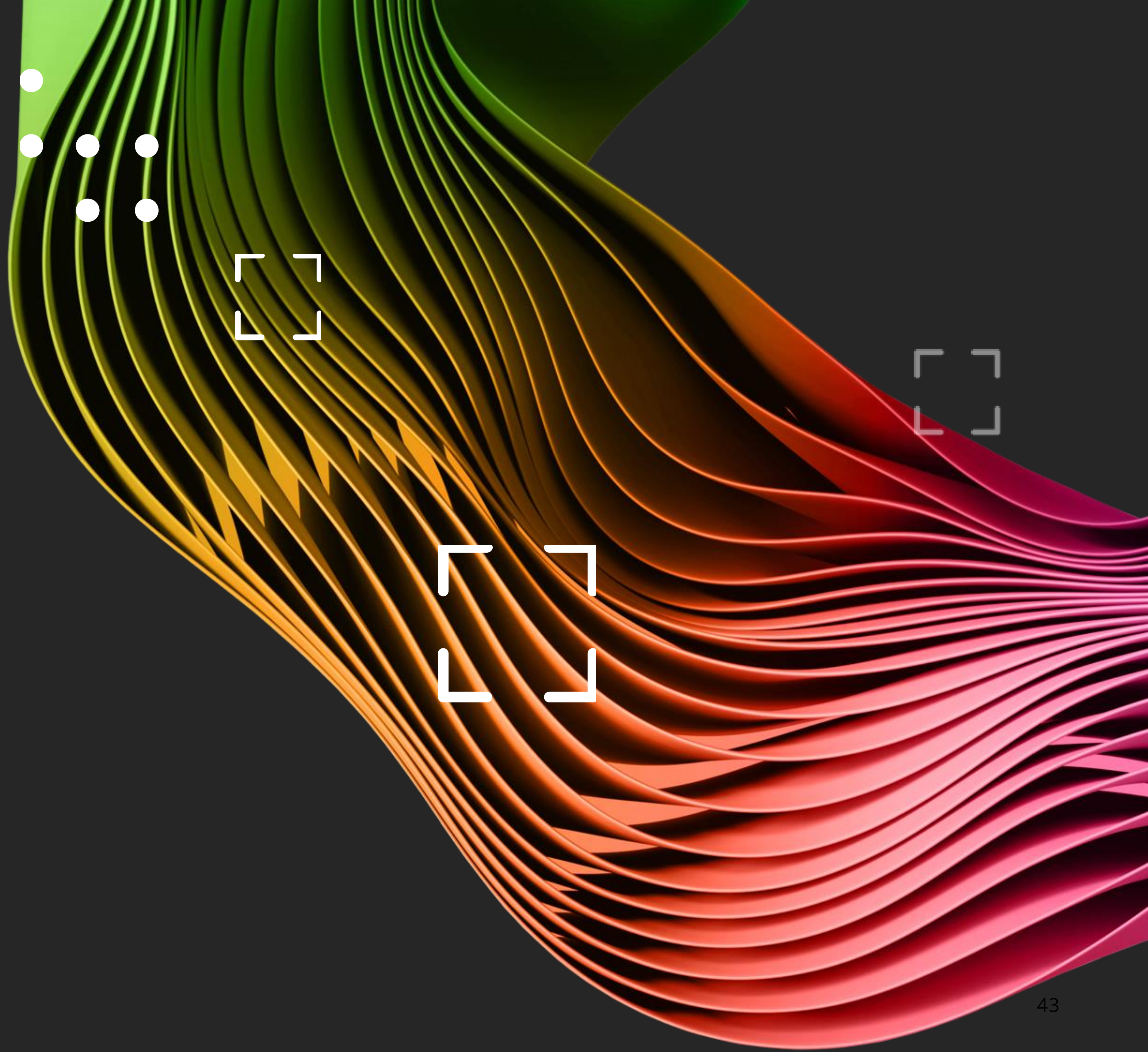
Data Collection and Analysis
Efficiently gather, cleanse, and integrate data from multiple sources, including ERP and CRM systems, social media, and market feeds

Disease Outbreak Prediction
Can be used to monitor and analyze data from various sources (i.e. social media, health records, and environmental data)

Social Media Management
Craft social media postings by retrieving and creating material based on current trends and subjects

Personalized Shopping Experience
E-commerce site can recommend products based on a customer's browsing history and past purchases

How do all of these relate?



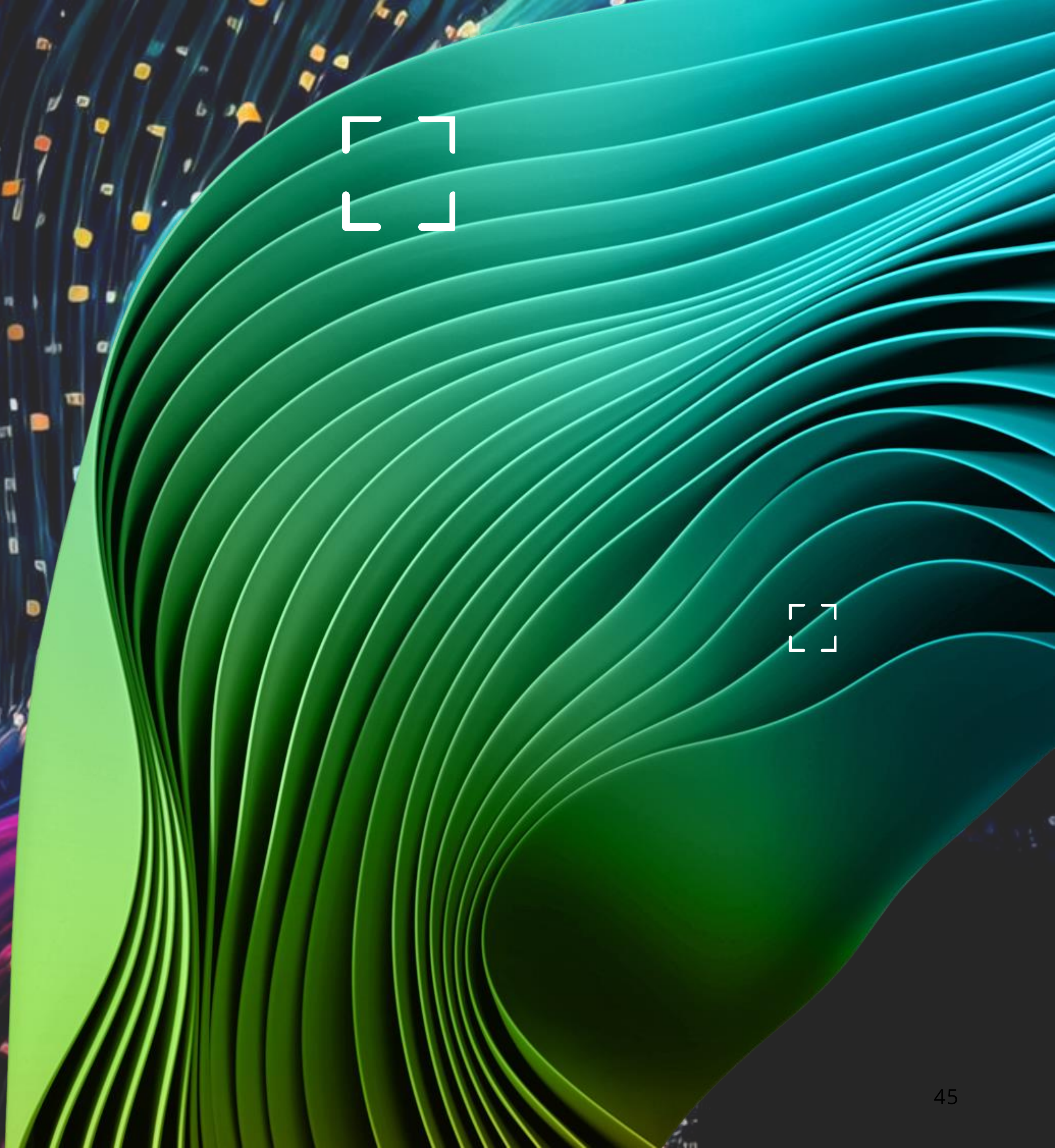
LLMs, Agents, and ReAct all work together

Due to the limitations of LLMs, we can utilize Agents to help us

By using Agents, we can perform specific tasks autonomously or with minimal human intervention

ReAct enables LLM Agents to do reasoning and take task-specific actions to provide a structured approach while developing

Questions?



THANK YOU!