AI Tools/Agents - March 2025

Goal of Document

Readers will acquire a general understanding of “where AI is today and where it might go tomorrow.”

This document provides a brief history of AI, a snapshot of the “state of AI technologies” as of late 2024, and imagines/forecasts how the technologies may evolve over time.

This document uses the term artificial intelligence or “AI” broadly.

Table of Contents

Why the Recent Excitement Over AI?

People are excited about AI today because we seem to now have the beginnings of key technologies to assemble into digital humans that complete work traditionally done by people. AI may even complete work that humans can't do ourselves. Below are some recent events that are driving the excitement.

In 2021, DALL-E, a text-to-image model, wowed the public with its ability to generate still images of text descriptions entered by a human user. Enter "Teddy bears working on new AI research underwater with 1990s technology" and it generated this. (https://en.wikipedia.org/wiki/DALL-E) People were impressed by DALL-E's ability to respond to human direction and generate images that people would have trouble creating themselves.

In 2022, ChatGPT, a Large Language Model (LLM), wowed the public with its ability to respond to text prompts submitted by human users. The prompt could be a question such as "Why is the sky blue?" or instructions such as "Tell me why the sky is blue and explain it so that a 5 year old understands." ChatGPT would attempt to answer the question and execute the instructions. ChatGPTs capabilities include summarizing text, coding, brainstorming, writing on your behalf and others. People felt that ChatGPT had human-like intelligence and felt this was a great advancement to achieving intelligence that could get work done for us.

In 2024, ChatGPT demoed version 4o to the public that made interacting with ChatGPT feel as if you were conversing with another human. Here is the video: https://www.youtube.com/watch?v=DQacCB9tDaw.

Among its new features included the ability to listen to human voice instead of relying on written input, the ability to output via a human voice, and the ability to accept visual input through a mobile phone camera. The human voice could express a variety of tones and emotions. Technologies that translate spoken language into text and vice versa have been around before ChatGPT. Combining them with ChatGPT's ability to answer questions and execute instructions opened the door for technology to interact with humans in the way we interact with each other, by speaking and listening, and have technology complete work for us. The demo had flaws, but some were still impressed.

DALL-E grew people's anticipation that AI could deliver some tools beyond image generation, but these tools were yet to be identified. ChatGPT appeared to listen to human speech and respond somewhat intelligently. It writes emails and codes for you. People are excited about AI because it seems as if new categories of work could be done by seemingly sentient computers; this would drive another wave of productivity increases.

Note About LLMs

LLMs predict the most probable next word when it generates responses to prompts. Trained on a very large set of text input, presumably factually correct input, LLMs generate sensible sounding sentences because it was trained on sensible sentences. LLMs do not actually understand their inputs or outputs.

Sometimes, the probabilities work out so that LLMs generate sentences that are actually factually incorrect. These are called “hallucinations.” The word “hallucinations” is misleading because it anthropomorphizes LLMs as technologies that have the capacity to think critically. LLMs do not think critically. They generate sensible sounding sentences by generating highly probable sequences of words. Here’s an article about Apple Intelligence hallucinations dated January 2025: https://www.macrumors.com/2025/01/06/bbc-calls-out-apple-ai-creating-fake-news-titles/ .

You’ve probably read newspaper headlines such as “A.I. Chatbots Defeated Doctors at Diagnosing Illness” and some others about advancements in AI reasoning. If these headlines are true, then how can LLMs provide only “limited insight?”

It’s a good idea to take these exciting headlines with a grain of salt.

Here’s a paper published in October 2024 in the Journal of the American Medical Association: https://jamanetwork.com/journals/jamanetworkopen/fullarticle/2825399 . This article states the way the study that resulted in chatbots defeating doctors was done was by inputting organized information into chatbots. In a real clinical environment, doctors and patients go back and forth with questions and answers in performing differential diagnoses, where illnesses are ruled out and the most likely root cause is isolated. When chatbots were asked to perform differential diagnosis, the chatbot underperformed in comparison to human doctors. In addition, LLMs frequently made incorrect treatment recommendations.

Here’s an article titled, “Are Your LLMs Capable of Stable Reasoning?”: https://arxiv.org/abs/2412.13147 . It states that there is “substantial room for improvement in LLMs’ realistic reasoning capabilities.” Part of this paper describes how to evaluate reasoning capabilities. Reason capabilities appear to be early development.

What Do AI Technologies Do?

AI technologies refer to technologies that can complete tasks that were once considered ones that only humans can do. AI technologies complete tasks by accepting input information, processing it, and then outputting information. This pattern is the same with most computing; so how are AI inputs/outputs and processing different?

Let's break down the discussion of what AI technologies do into the following components so that we gain a sense of how AI computing is different from "traditional" computing.

1. Inputs/Outputs
2. Processing
    a. Recognition
    b. Transformation with One Medium
    c. Transformation from One Medium to Others
    d. Mimicry

In this article, we'll discuss only inputs and outputs, the first bullet point above. Later articles will discuss each of the "processing" bullet points and describe some illustrative use cases. We'll learn by example.

Inputs/Outputs

Most AI accept inputs/generate outputs in the mediums of information below:

1. Visual
    a. Photos (sceneries, objects, animals, face portraits, signs, receipts, drone, etc.)
    b. Videos - selected photo frames (surveillance, movies, DIY explanations, etc.)
    c. Image Scans
        i. Documents (text, tables, charts, etc.)
        ii. MRIs, X-Rays
    d. App User Interfaces - screenshots or UI itself
    e. Satellite Images (weather, spy)
2. Aural
    a. Voice
    b. Music
3. Human Language Text (word documents, web pages, screenplays, etc.)
4. Numeric Data (CSV files, spreadsheets, parquet, etc.)
5. Programming Code

Processing

Recognition

Examples of recognition capabilities include recognizing something or someone in photos and recognizing contraband in parcels going through scanners at customs. Illustrative use cases below describe targets for recognition and how recognition can be useful.

Recognition enables computers to:

1. Transform input into human language description (overlaps with "transformation from one medium to others")
2. Find similar things
3. Find exact match
4. Find “interesting” patterns

Illustrative Use Cases

1. Photos
    a. Plant Disease - “Monitor and notify when crops have gotten disease/bug infestation so that I know to prevent its spread and treat it.”
    b. Human Face - “Verify the identity of this human so that I can let him through passport control.”
    c. Distractions - “Identify main subject and distractions in photo.”
2. Videos
    a. Human Emotion - “On video conference calls, show me who feels positive/negative about what is being discussed.”
    b. Human - “Find the suspect in 1,000,000 hours of surveillance video from different sources so that we can track him down before he gets away.”
    c. Object - “Find rug similar to the one in this TikTok video so that I can buy it.”
3. Image Scans
    a. Cancer - “Determine whether there is cancer on this x-ray so that we can rule it out.”
    b. Chart - “Determine underlying data for this chart so that we can use the data.”
    c. Aberrations - ““Highlight aberrations in sonar scans of ocean floor so that reviewers can focus attention on scans that might be plane wreckage and not have to review 100% of scans.”
4. App User Interfaces
    a. Mobile UI - “Describe this UI so that vision impaired can use it.”
5. Satellite Images
    a. Tornado - “Detect precursors of tornadoes so that people can be warned to shelter in advance.”
    b. Troop Fortifications - “Identify hidden fortifications so that we know where the enemy might be.”
6. Voice
    a. Words, Sentences - “Take dictation of this court proceedings so that we have a readable accurate history of all that was said.”
7. Music
    a. Melody - “Recognize the music that is being played now so that I can later find the song on music streaming service.”
8. Human Language Text

    a. Topic - “Identify the topic and the thesis of the article so that I can decide whether or not to read it.”
    b. Sentiment - “Identify whether an author feels positively or negatively about a subject matter.”
    c. Patient Diagnosis - “Diagnose illness so that I can get treated.”
9. Numeric Data
    a. Fraud - “Identify suspicious money transfers by finding patterns in banking history data of all account holders of bank X.”
10. Programming Code
    a. Security Holes - “Identify code that doesn’t validate data accepted from untrusted system/user so that programmer can fix security holes.”

Transformation Within One Medium

The word “transform” here does not refer to “AI transformers” that are a type of neural network architecture. “Transform” here literally means “make a change in form.”

A “medium” is a format of information such as photo, video, audio, screenshot or written human language. Transformation within one medium supports varieties of user goals. Illustrative use cases below describe user goals. The use cases typically require “recognition” capabilities to identify targets and then “transformation” capabilities to alter targets to achieve user goals.

1. Photos
    a. “Remove the distractions in my mobile phone photos so that viewers can more easily visually isolate the primary subject of the photo.”
2. Videos
    a. “Correct lighting of actors’ faces on movie shots so that viewers can see their expressions more readily.”
3. Image Scans
    a. -
4. App User Interfaces
    a. -
5. Satellite Images
    a. -
6. Voice
    a. “Isolate and amplify only the voices that the listener is likely wanting to hear.”
7. Music
    a. “Remove the main voice track in songs so that users can sing karaoke to their favorite songs.”
8. Human Language Text
    a. “Translate Korean news articles into English.”
    b. “Summarize this encyclopedia article about the Revolutionary War for a 5th grader.”
9. Numeric Data
    a. “Remove outlier data from data set.”
10. Programming Code

a. “Alter the code so that all inputs from human users are validated so that the code has fewer vulnerabilities.”

LLMs are the AI technology that makes transformation from text to another form of text possible. LLMs can generate fictional stories, essays, and poems at the request of the user. It can read programming code and generate documentation for the code.

Transformation From One Medium to Others

Today, many use cases transform mediums that are not human language into human language and vice versa.

Chaining AI technologies together to transform voice into human language text, then feeding that into LLMs, then letting LLMs respond after they query some internal/3rd party systems for information, and outputting the response as voice is a popular pattern. Companies are attempting to replace phone customer service agents with this chain of technologies and LLMs. Others are putting the voice-in-voice-out interface on top of their existing applications.

Another example of transforming one medium into another includes a technology that transforms screenshots of mobile phone UIs into human language. It takes human voice as input and attempts to execute the voiced instructions as taps on mobile phone screens. This human language is transformed into audio, a human voice. These technologies help blind users understand and interact with mobile phones.

Transforming information of paper documents/faxes, emails, and recordings of phone calls into a single human language text that LLMs can ingest has opened the door to addressing the “messy inbox problem.” Information about a patient or a client that may be spread across different mediums can be unified AI technologies and downstream workflows can be automated with Saas.

An example of technology transforming human language into video is Sora. Describe in words a video you would like to see to Sora, and Sora generates it. Sora can also take photos and video clips as input.

Further into the future, AI technologies similar to Sora may be able to take a novel or a screenplay as an input and generate a movie. Another application of video generation can be generating tutorial videos that convey ideas visually so that the student learner has an easier time learning than by learning from static illustrations.

Mimicry

Mimicry is a category of AI capabilities that span both categories of transformations. Examples of mimicry use cases include generating videos of celebrities or politicians saying and doing things that they in reality didn’t say or do. Other examples include mimicking a singer’s voice or a musical style of a rock band to generate alternative versions of existing songs.

Other use cases include generating videos mimicking human faces and speaking voices so quickly that human users can interact with mimicked entities in real time.

Some may argue that LLMs are mimicking human language.

How Are “AI” Technologies Built?

AI technologies are built differently from traditional computer programs.

Traditional Computer Programs

Traditional programs are written to run deterministic sequences of instructions by testing conditions and then branching or looping depending on the result of the test.

For example, let X be a variable that contains the value 50. The value represents the number of dollars in a bank account. When the account holder attempts to withdraw $20, a computer program checks, “Does the bank account have enough money?” If X is greater than $20, then the program allows the withdrawal. If not, then the program disallows the withdrawal. Checking the account balance is condition testing. The program can take two different branches depending on the results of the condition testing: allowing vs. disallowing.

Let’s take another example for looping. Let Y be a counter that counts from 0 to 100. The value of Y represents the distance in miles from the starting point which is at mile marker 0 to the ending point at mile marker 100. A program is printing mile markers for a highway. It starts at 0 and prints “Mile 0” onto a sign. Then it increments Y by 1 mile and checks the condition, is Y 101 yet? If it is, then it will stop. But if not, then the program loops and prints the mile marker for the current value of Y.

“AI” Computer Programs

Programming a computer to do “AI” things like recognizing a cat in a photo is different from the above approach to programming where humans write deterministic sequences of instructions. Humans write computer programs that are machine learning models and train the models feeding data such as photos with cats in it and some without, and then teaching them which photos had cats and which did not. This explanation is far from perfect but the point is that humans don’t write deterministic programs that recognize cats by first identifying two eyes, then two pointy ears, four legs, and then a furry exterior.

Humans program the different types of models that use the training data differently and learn differently. For example, different object recognition models require data to be provided in different ways. The digital data of photos can be resized, chopped up into smaller pieces, and its data altered to serve as input into the models.

Models contain a large set of numbers. “Learning” takes the input data and sets the values of these model numbers. Once learning is done, then the model can be used to calculate predictions based on inputs. In our case, the inputs are photos that may or may not have a cat in it and the model predicts whether the inputs have a cat or not. The prediction is not necessarily a black or white “yes” or “no”; it can be a probability value.

How AI Might Outperform Humans

Human Limitations

Machines/computers are not bound by human constraints.

1. Do not fatigue physically or mentally from mundane or high-stress tasks.
2. Do not need nutrition, rest, or sleep.
3. Do not grow attached to their own ideas.
4. Can be digitally copied to scale up. Scaling down is not as painful as human layoffs.

The hope is that the overall cost of building and maintaining AI technologies will cost less than humans completing the same volume of work at the same level of quality.

How AI Machines Might Learn "Better"

Some other ways AI technologies may outperform human counterparts in the future include:

1. **Learn Broadly** - Trained on a broader information base than humans can read, experience and internalize in a lifetime. This may enable AI technologies to be wiser than any single individual and make "better" decisions.
2. **Retain More Information** - Have memories that don't fade away as they do in humans. For example, it may be able to remember all evidence and witness testimonies for a murder case and find paths of investigation. Perfect memories may also enable AI to perform real time fact checks.
3. **Intake Information and Make Sense of it Faster** - Speed of information intake may be faster than top human intake speeds. This may be advantageous in real time applications such as one that helps formulate battlefield tactics where lots of information about what's going on in the air and on the ground must be integrated to make "better" decisions.
4. **Share Learnings Faster** - Learnings of one instance of an AI agent can be copied to all other instances of the AI agent. Entire populations of AI agents can get smarter faster from all their experiences. In contrast, each human individual must expend effort to identify what is worth learning and learning it.

What the above statements leave unclear is "learning." Can AI technologies learn in the same way humans can? That is, if we want AI to learn college physics, can we give it a physics textbook and lectures, and will AI transform the information into knowledge and skills? Will it be able to solve all practice problems at the end of each chapter? If yes, to what extent and how quickly? Will it get an "A" on a college physics course? Will it be able to teach a physics course to humans and customize homework so that the student masters the material?

How do we teach AI to be better at sales? Provide better psychological therapy?

How efficiently will future AI technologies be able to convert every ounce of input such as books and lectures into learnings that it can apply in real life to get work done? Is that even a relevant measure because the way AI learns is different from the way humans learn?

If there are answers to these questions today, the answers will likely change as we develop new methods that AI can learn specific skills.

What Might AI Technologies Do in the Future?

Beyond the categories of recognition, transformation, and mimicry, what other categories of capabilities will AI technologies develop? We can take a guess by looking at what faculties we humans use to complete our work. Here are some ideas.

Critical Thinking for Knowledge Work

Research for Decision Making

People often need help making decisions about actions they must take to achieve their goals. AI that can think critically can help in identifying and getting missing information.

Today, LLMs take prompts at face value and generate responses. Human users revise their prompts until they get acceptable output from LLMs.

Critical thinking capabilities will enable AI to ask questions to the human user about the human request to drive "better results."

Critical thinking enables the following:

1. Clarification of the goal of work/desired outcome and the constraints on methods of achieving the goal.
2. Identify areas of ambiguity/hidden assumptions that may have substantial impact on decision.
3. Identify methods of acquiring information to address ambiguity.

In order to think critically, AI must have foundational knowledge of the target domain. For example, primary care physicians rely on a different foundational knowledge to think critically about what to do to diagnose patient illnesses than software product managers rely on to make project prioritization decisions. Thinking critically to drive murder investigations requires foundational knowledge of investigative methods and the boundaries of law.

This means that critical thinking cannot stand alone as an AI capability but must be tied to a domain. We may have AI agents that specialize in critically thinking in narrow domains.

Argumentation

People often need help in convincing an audience to support a cause, purchase something, accept a business strategy, invest in something, or fund a venture. Argumentation requires the ability to reason and evaluate the compellingness of lines of reasoning.

Critical thinking enables the following:

1. Identification of hidden/flawed assumptions
2. Identification of methods of strengthening/weakening arguments
3. Identification of reasoning flaws
4. Constructing compelling arguments

Convincing a judge or jury of one verdict over another with a line of argument relies on different foundational knowledge from convincing managers of a business to follow a particular business strategy. Argumentation capabilities will unlikely be a general skill but will be tied to narrow domains.

Empathy for Service Work and Beyond

Empathy can help AI understand what a single person, individuals in a team or organization, and entire populations of people may be thinking and feeling. Empathy can augment AI's argumentation capabilities to convince people of something.

In a customer service/support phone setting, the service worker can complete the task that the customer requests but also relieving the customer's emotional distress by leveraging empathy.

In a therapy setting, where AI is the therapist and the human is the patient, AI must recognize the emotion expressed by the patient on his face, voice, and body movements, and also know the patient's history in order to say the right things and ask the right questions to help the patient manage his thoughts and feelings.

In a video conference setting, AI can be a sales person's tool in recognizing stakeholders' sentiment and understanding their spoken objections and concerns. Empathy enables AI to identify possible underlying concerns and motives that individuals may not be explicitly revealing intentionally or unintentionally. AI can recommend the facts to present to them and how to convince them to buy in.

In a mass media setting, AI can ingest popular social media content and popular search terms to infer people's sentiment. It can also intake results of surveys/polls and focus groups. AI can then recommend what action must be taken to convince people to take a course of action such as voting for a candidate. People are motivated not only by facts but also by emotion. Should the politician focus on financial hardship imposed on the population by inflation? What should she say to convince the population that she has their backs? AI can make recommendations.

A company called Waveforms.ai raised $40 million to make empathetic audio on December 2024. Here's a Reuter's article of the company:
https://www.reuters.com/technology/artificial-intelligence/former-openai-researcher-raises-40-million-build-more-empathetic-audio-ai-2024-12-09/.

Spatial Awareness, Predictions of Movement by Actors in Adversarial Arenas

In a battlefield of soldiers, leaders need to know where to place their weapons, soldiers and defenses. During battles, soldiers need information about enemy troop locations and movements and know how to counter them. In the heat of battle, soldiers may not have time to synthesize all information to make the best decisions to build a plan of attack or defense. Leveraging real time information across land, air, and sea into optimal battlefield tactics to strategy may be valuable.

On a soccer field, coaches need to understand their opponents' strategies and counteract them. How is the opposing team confounding our defenses? How can the soccer players improve how they work individual and as part of a team? These questions can be asked for any team sport.

Fighting wildfires in California can be more effective by understanding how the fires are spreading because of the wind. It can advise firefighters to allocate resources optimally. For example, where should planes drop water to prevent fire from spreading?

Object Recognition and Manipulation

In the context of homes, a robot that completes household chores such as laundry and dishwashing will need object recognition and know how to handle objects.

What Are AI Agents?

People who haven't followed the news of advancements in AI often encounter the label "AI Agent" on products and wonder, "What is it?"

Debate About "AI Agents"

There is no single definition of the term.

Dr. Andrew Ng, a thought leader in AI, discusses "agentic workflows" of LLM-based agents on this video: https://www.youtube.com/watch?v=sal78ACtGTc .

Here's Salesforce.com's definition: https://www.salesforce.com/agentforce/what-are-ai-agents/. Salesforce.com lists their ideas for agents here: https://www.salesforce.com/agentforce/.

Here's AWS's definition: https://aws.amazon.com/what-is/ai-agents/ .

Definitions are different depending on the source and they will undoubtedly change.

A Definition of "AI Agents"

Here's a simple definition of "AI agents."

> *AI agents complete work to achieve productive goals independently from human users who have hired the agent.*

The longer stretches of work that an AI technology can complete without human guidance, the more appropriate the label of "AI agent" is rather than the label of "AI tool." An AI technology that can answer customer inquiries over the phone takes on the label of "agent" while an AI technology that writes drafts of emails under your immediate direction is a "tool." Agents seem to fulfill our expectations that they do work that normally human individuals are expected to do.

In the near future, AI agents are suited for knowledge work rather than physical work because a lot of data exists in a digital form that AI agents can ingest. Further into the future, robots may be able to perceive its physical surroundings to the degree that humans do and be dextrous enough to handle tasks of chopping vegetables, folding laundry, or taking out the trash. Perhaps embedding knowledge work capability into these robots will make us call these robots "AI agents" too.

AI agents will complete progressively more challenging work and achieve more complex goals beyond the current question-and-answer interactions we have today. That is, the AI agents will work in ambiguous situations, identify what resources, including information and other agents, if acquired will help achieve better results, and work in collaboration with the human who hired it to clarify, redefine goals, discuss pros and cons of execution approaches.

What Not to Do With The Definition

Once, I witnessed a product management executive deliver a presentation that declared that if the company product were “bikes” then, the company would have to first understand what a “bike” is. Once the definition was fully understood, then the company could build bikes to fit the definition.

Products are never built to fit definitions. They are built to address pain points that people encounter on their way to achieving their goals.

Don’t use this article’s definition to drive the implementation of “AI Agents.” Use it to label products that help people understand what the products do.

How AI Agents Will Compete Against Each Other

AI agents will compete against each other by requiring less resources to complete the same tasks at the same level of quality.

Identifying the key contributors to the quality of completed tasks and the total costs of executing those tasks helps us better understand how agents will compete against each other. The total costs include all costs from getting training data to powering specialized hardware during training, and powering inference.

With the recent news about DeepSeek's LLM matching/beating OpenAI's LLMs across a variety of benchmarks, the world realized that useful LLMs can be trained with much less money than OpenAI was spending. There isn't just one best training method.

Model

Different models may require different types of training data, training process, and perform inference differently.

A startup called Inception claimed to have implemented a new AI model called diffusion based large language model. This model can complete tasks like LLMs but do it much faster. (Techcrunch https://techcrunch.com/2025/02/26/inception-emerges-from-stealth-with-a-new-type-of-ai-model/) Today's LLMs predict one word at a time; however, this new model generates sets of words all "at once." The training data and process for this new model is unclear. In addition, no performance benchmarks have been published.

The model itself will heavily influence the total cost of providing AI Agents to customers and the quality of the work they output.

Access to Training Data

Data is required to train agents. If some AI agents have access to training data that others don't, then it can have an advantage in more skillfully/efficiently completing tasks downstream.

Data includes online encyclopedias, data that labels cats in photos, math problems/solutions written by computer programs, or email users labeling spam email among many others.

Some ways that companies can create this advantage include:

1. Entering into exclusive licensing contracts with companies that own data so that no other company can have the data.
2. Creating its own proprietary data set by using human resources, using computer programs/AI, or collecting feedback from its users.

Computer programs and AI can generate math problems/solutions.

Training Process

Training methods, the model architecture, and method of using hardware drive training costs and the time it takes to update an AI's capabilities. AI companies can compete by using novel training methods, model architectures, and proprietary programming approaches to manage hardware resources that allow their AI agents to complete the same tasks with cheaper training.

Examples of training methods include reinforcement learning and supervised fine tuning. Question-answer training that allows LLMs to answer questions; without it, they can only predict the next word in word sequences. The invention of question-answer training opened the door to today's LLMs ability to answer your questions. The training involves showing sets of question-answer pairs that are authored by humans or some blend of programmatic automation and human labor. Reinforcement learning enables a model to attempt answering a question and get a reward proportional to the quality of the answer compared to prior answers. Deepseek uses math and coding problems that can be programmatically scored without humans.

Model architecture includes mixture-of-experts or single-expert. Deepseek uses mixture-of-experts architecture; each "expert" neural network is trained on a special domain. A task orchestrator knows which experts to ask. In contrast, single-expert architecture trains a single model for all domains. The single-expert architecture has more active parameters to calculate results than a single specialized mixture-of-experts neural network. This means that single expert architectures expend more energy to perform calculations than mixture-of-expert architectures.

Training methods for object manipulation by robots is evolving. Creating a simulation environment that substitutes for a real physical environment is a quicker way to train robots. Check out this article. https://arstechnica.com/information-technology/2024/12/new-physics-sim-trains-robots-430000-times-faster-than-reality/

Even with the same computers running with identical CPU and GPUs and ultimately resulting in the same model values, the training process can use more or less energy and time depending how hardware resources are managed programmatically.

Computing Inference

Inference refers to the model's work of calculating a value based on inputs. Some models can run on mobile phones and others need special computer hardware with high performance GPUs to run. Models that require special computer hardware will be accessible to mobile phones if the mobile phone can leverage remote computing resources. The speed of calculating inference will also make one AI agent more attractive than others.

Access to Data to Search

The information on the world wide web is accessible to most people and computers. Some information is only available behind paywalls. Some information is only available within internal IT systems for businesses.

The pages of the New York Times can be made exclusively available to AI Agents that have contracted to access the information.

Any company's internal wikis or file servers are likely proprietary as well as their employees' emails. Companies may have information stored as PDFs, emails, videos, and whatever else. If AI Agents were permitted access to all internal data sources, they must still be able to get data from those sources. How will they do it? Via an API? What will the connections between these data sources and AI Agents look like?

Access to Capabilities

AI agents may not be best-in-class for all tasks. AI agents that recognize which other agents are better at particular tasks and the ability to work with other AI agents to get work done can be valuable. The more easily an AI agent can work with other agents with better capabilities, the more likely it can complete high quality work.

Future of the AI Agent Ecosystem

Now that we have some idea of what AI agents are, we should try to point out what the environment in which AI agents will operate will look like.

One Dominant Generalist vs. Many Specialists

Some predict a future where one company, presumably the one with the most money or the one with the first mover advantage, will conquer all.

The premise behind the “most money” line of thinking is that the company with the most money will have access to the “best/most” data because some data will have a price to license from the owner of the intellectual property or some data will have to be created by people manually labeling data. The compute power necessary to train models can be expensive.

The premise behind the “first mover” taking all is that a virtuous flywheel of usage of the model will result in learnings/data that are proprietary. It will learn faster than any other company, make the best investments, and “win at the end.”

These arguments are uncompelling. Why couldn’t Microsoft win the search war against Google if more money is advantageous? OpenAI can be considered the first mover but how did Deepseek catch up?

Some predict a future where there isn’t just one clear AI technology that outperforms all others in every task. Agents will work together in teams to get work done for human users. This means that a variety of AI agents who are good at particular tasks will be asked by other AI agents to get subtasks of a larger task done. This means that agents will have to be able to discover and interact with other agents.

How will AI agents work with each other and how will they be able to interact with shopping applications and send email? Will the technical interface to other AI agents and software tools be APIs or will AI agents use existing human interfaces? Most likely, the less expensive method will win out. If AI agents are taught to use human UIs, then software tools won’t necessarily need to open APIs to serve AI agents. If AI agents’ ability to interact with human interfaces is unreliable, then opening an API will be a more reliable way. But what would be involved in making AI agents to use APIs? Will all APIs for e-commerce platforms be uniform so that AI agents simply have to learn one API or will it have to learn a new API for each e-commerce platform? As the API changes, how will AI agents be updated on the differences? The API route also have its challenges.

AI Agent Autonomy and Human in the Loop

Some predict that in the near future, AI agents will be completing purchases on behalf of human users. Some argue that they don’t trust anyone else with their credit card information so why would they trust AI agents with theirs? Very few people trust others with their money; it’s not surprising that people won’t trust AI agents anytime soon. More likely is the scenario where human users are key decision makers in AI agent workflows. AI agents will contact their human users and ask for permission whenever the next step to be taken will impact the human users’ wallet.

AI agents will most likely include the human user in key decision-making beyond just purchase decisions. In any scenario where the consequences of a decision can generate legal liability, a human must make the final decision and take responsibility to fit into our current legal system. Humans will likely be in the loop for consequential decisions.

AI agents will also have to ask their human users for permission to submit the human user's private information to get something done. For example, if a human user is shopping for car insurance, then the insurance company will need some private information to evaluate how much your premium will cost. Driver's license numbers or social security numbers won't be freely handed out by AI agents.

AI Agent Collaboration with Other Agents

In the future, AI agents will most likely collaborate with other AI agents. The manner in which AI agents will interact with other agents and software tools is still unclear. Humans will be in the loop in making key decisions and humans will be legally liable for those decisions. Furthermore, humans approval will be necessary for AI agents to share the human user's private information.