Architecting Next-Generation Generative AI/ML Models and Agentic AI Solutions*

Prof. Dr. V. David Sánchez A., Ph.D.
Brilliant Brains, Palo Alto, California
August 2025

Abstract

Agentic AI solutions are currently revolutionizing operations in broad sectors of multiple industries and business environments. As such, they are a form of, are embedded in AIOps [4]. As an indispensable foundation, the associated development of Large Language Models (LLMs) in summarized in Figure 1. In order to design real-world impactful AI-integrated solutions making use of cloud infrastructure, corresponding APIs, and enterprise platforms, solid foundations of Generative and Agentic AI are needed and are therefore presented in-depth, highlighting among others, how they differ among each other, how they are embedded in business workflows, how they are applied to realize real organizational functions, and how to solve real-world problems. In particular, gen AI agents [6], AI-powered software entities that plan and autonomously perform service and increasingly, other tasks for and/or with humans have enjoyed incredible growth. On the path to technology maturity, the prototyping, evaluation, and deployment of AI-powered solutions can be performed by using industry-leading platforms and frameworks, e.g., two of those being OpenAI's ChatGPT [7] and Ollama [8], closed-(proprietary) and open-source respectively. ChatGPT is a chatbot released in November 2022 that uses generative pre-trained transformers (GPTs) [9] to generate text, speech, and images in response to user prompts. Among multiple others, one can type or start a real-time voice conversation, search the web getting links to relevant web sources, collaborate on writing and coding, analyze data and create charts, create and talk about images, and offload complex tasks from start to finish with agents. Ollama is an open-source framework first released in September 2023 designed to facilitate the privacy-focused, customizable, and cost-efficient deployment of large language models (LLMs) [10] on local environments. It provides a straightforward API using Python and JavaScript libraries and a command-line interface for managing and interacting with models, e.g. allowing for the download, update, and removal of AI models directly on the local system eliminating the reliance on third-party services.

The transformative character of these generative and agentic AI technologies in global economic terms may not be underestimated, in the meantime as a crucial part of the World Economic Forum (WEF)'s Digital Transformation Intitiative (DTI) [11] in the amount of $100 trillion over a decade. A micro-to-macro and simulation-based approach to modeling the AI's potential impact on the global

*This abstract has been granted permission for public release. The author has among many others a 15-year tenure as founding Chief Scientist/EiC of an Elsevier Science's AI/ML journal [1] and is the youngest "Nobel"-Prize in Engineering (IEEE Fellow) in history worldwide with the following mention "for leadership in neural and parallel computation, and pioneering contributions to autonomous space robots" [2]. As a decade civil servant at the German Aerospace Center DLR, he launched with the Office of the German Federal Minister of Research and Technology the First Federal Program for AI/ML Research and Technology Development in Germany [3], which funded multiple consortia all over the country, one of them being his consortium including DLR and Siemens Corporate R&D in Munich, Germany. He has been conceiving, designing, building, and operating advanced scalable mission-critical parallel and distributed as well as secure (NSA, commercial) computer systems for federal and state governments and the commercial industry, e.g., AI pipelines deployed to and executing during NASA-ESA-DLR spaceflight missions, ML pipelines for DoD classified programs, advanced AI/ML pipelines for advanced data centers and vertical on-prem, hybrid, and in the cloud applications of the State of California.

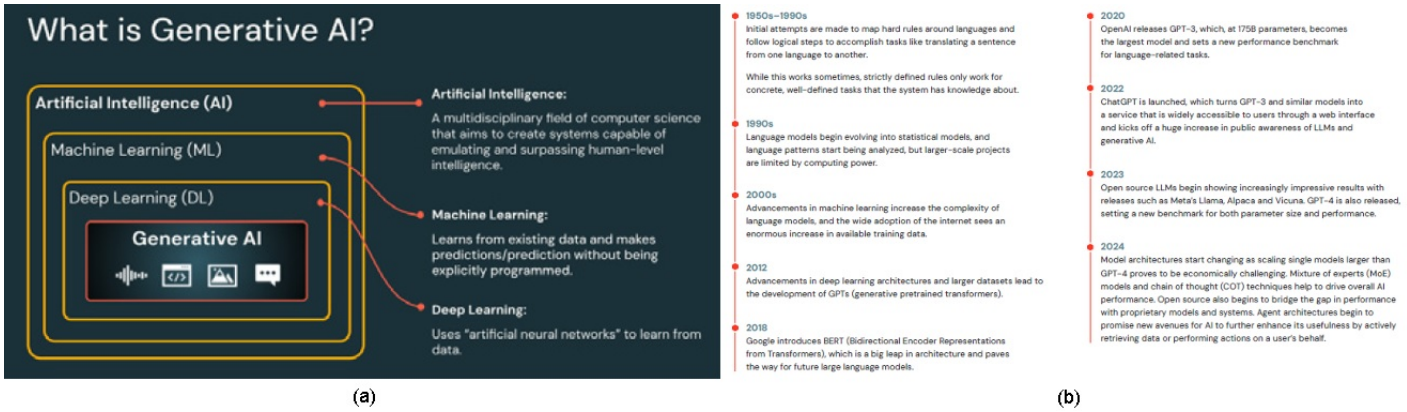


Figure 1: The evolution of Gen and Agentic AI [5] (a) Generative Artificial Intelligence (Gen AI) (b) Historical development of Large Language Models (LLMs)

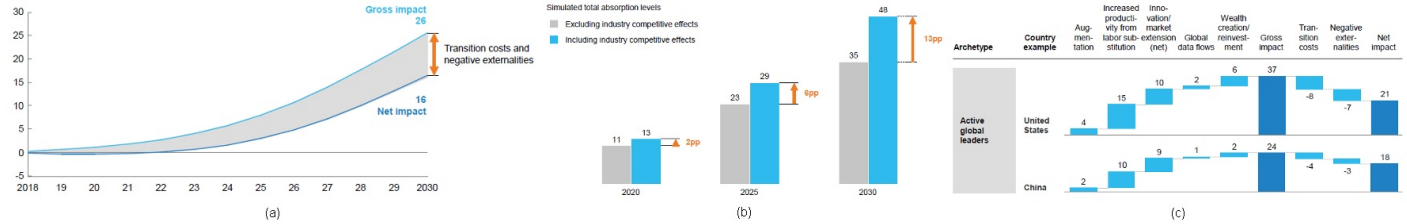


Figure 2: The impact of AI adoption and absorption [12] (a) Value-added gains of economic output: cumulative boost vs. today in % (b) Competitice pressure accelerates the pace of AI absorption (c) Country AI impact variation: economic drivers by country in % points

economy resulted in AI having a large potential to contribute to global economic activity, i.e. AI could potentially deliver additional economic output of around U.S.$13 trillion by 2030, boosting global GDP by about 1.2% a year [12]. This gradually emerging economic impact may only be visible over time, and depending on how companies and countries choose to embrace AI will likely impact corresponding outcomes and could further widen gaps between countries, companies, and workers. Figure 2(a) shows how the impact of AI, i.e., in this case, the net productivity effect, can build up at an accelerating pace, modest within the first five years to material by 2030. Rather than forecasts, the numbers shown are simulated figures to provide directional perspectives. Obviously, there are associated costs related to the implementation of AI. Companies are expected to adopt and absorb AI throughout their organizations at an accelerating pace over the years as a result of competition and improvement in complementary capabilities to use AI tools. The benefits to early adopters of these technologies increase sharply in later years at the expense of non-adopters. Figure 2(b) shows how the competitive pressure can gradually increase the absorption level by about 13% in 2030. Figure 2(c) shows the AI impact economic drivers, which favor countries most ready for associated technologies, shown here are only the global leaders, the U.S.A. and China, and with differences in the degree of impact according to the shown breakdown of economic drivers. For example, the economy of these two global leaders could gain about 10–15% of impact from labor substitution, compared with an impact of 5–10% in developing economies.

The evolution of mature AI systems and agentic AI systems has notably accelerated in recent years. The autonomy and real-time capability required in current systems was in significant part developed over decades for space missions. For example, just in recent months, I co-developed at NASA JPL in Pasadena, CA the next generation of Mars helicopters [13, 14], cf. Figure 3(a). Due to the long distance to the Red Planet, teleoperation from Earth is not an option and those flying robots need to operate 100% autonomously. For the development of those extremely advanced systems, frameworks like JPL's F Prime [15] need to support a component-driven architecture which breaks down flight software into discrete components with clearly defined interfaces facilitating rapid devel-

(a) (b) (c) (d)

Figure 3: AI solution examples (a) autonomous flying machines in space (b) data center infrastructure (c) healthcare applications (d) AI-powered material R&D.

opment, portability, reusability, and testing of such modular designs. Modeling tools like FPP [16] allow then developers to specify components and their connections, and then automatically generate the code, which saves significant development time. This type of development workflow of reusable components and systems accommodate tighter cost constraints and shorter development schedules, e.g., for commercial CubeSats, SmallSats, and instruments. Multiple AI applications for data centers, cf. Figure 3(b), exist, one example is the automated claim processing I co-developed for the State of California, which operates hundreds of large data centers via partner companies. Those solutions incorporate, e.g., Computer Vision and Artificial Intelligence, in conjunction with legacy relational database systems securely processing sensitive data of California citizens. AI applications for healthcare are actively being further developed not only as an integral part of the Project Stargate efforts [17, 18], cf. Figure 3(c). New materials could be developed faster, powered by AI, for terrestrial and space applications. For example, an AI-powered catalyst could produce oxygen using materials from Mars [19, 20], cf. Figure 3(d). The deep impact of these advanced AI technology developments as well as new government regulatory measures to protect AI innovation and provide safeguards are discussed in [21]. Success stories of the changing effects of GenAI and agentic AI in broader areas of business application have multiplied recently [23] including industries (with example companies) like manufacturing (Siemens), healthcare (Mayo Clinic), finance (JP Morgan Chase), retail (Amazon), transportation and logistics (DHL), energy (British Petroleum BP), education (Pearson), media and entertainment (Netflix), telecommunications (AT&T), government and public sector (Singapore) among others, some of them are further detailed in Figure 4 including the technology stack used, the financial impact achieved as well as additional, non-financial benefits. An annual comprehensive review of AI trends including technological advancements, performance on benchmarks, and economic impact as well as the uneven development in the responsible AI ecosystem can be consulted, e.g., in [22].

To gain an overview of current agentic AI developments and capabilities, we now focus our attention to some of the main LLM families: Anthropic's Claude, Meta's Lama, OpenAI's GPT, Google DeepMind's Gemini, and xAI's Grok as well as agentic AI tools: LangGraph, CrewAI, AutoGen, and AutoGPT, in particular in the framework of building agentic AI solutions to real-world problems. An example of how to use LLMs to build AI agentic solutions follows. Anthropic's Claude 3 large language models (LLMs) family: Opus, the most powerful, Haiku for speed and responsiveness, and Sonnet, the balanced model, cf. Figure 5(a), and Claude 4 successors are well-suited for that purpose, cf. Figure 5(b). Compared to previous models, their reduced hallucination and increased accuracy minimizes the risk of built agents making decisions based on incorrect information. For enhanced reliability and adaptability, Claude agents can be built using modular code, prompt engineering, and structured APIs. Custom tools can be defined and integrated into the agent's conversational logic. Multiple Claude agents can be run in parallel, this ability is utilized to build multi-agent systems by synthesizing the outputs of the individual agents to tackle larger and more complex projects. Claude models enjoy a simplified integration into existing workflows and applications by being accessed through the Anthropic API, AWS Bedrock, and GCP Vertex AI, simplifying integration into existing workflows and applications. The next generation of models continue to be developed at a fast pace as shown in Figure 5, more recently with Claude Opus 4.1 and Sonnet 4 [30]. The flagship

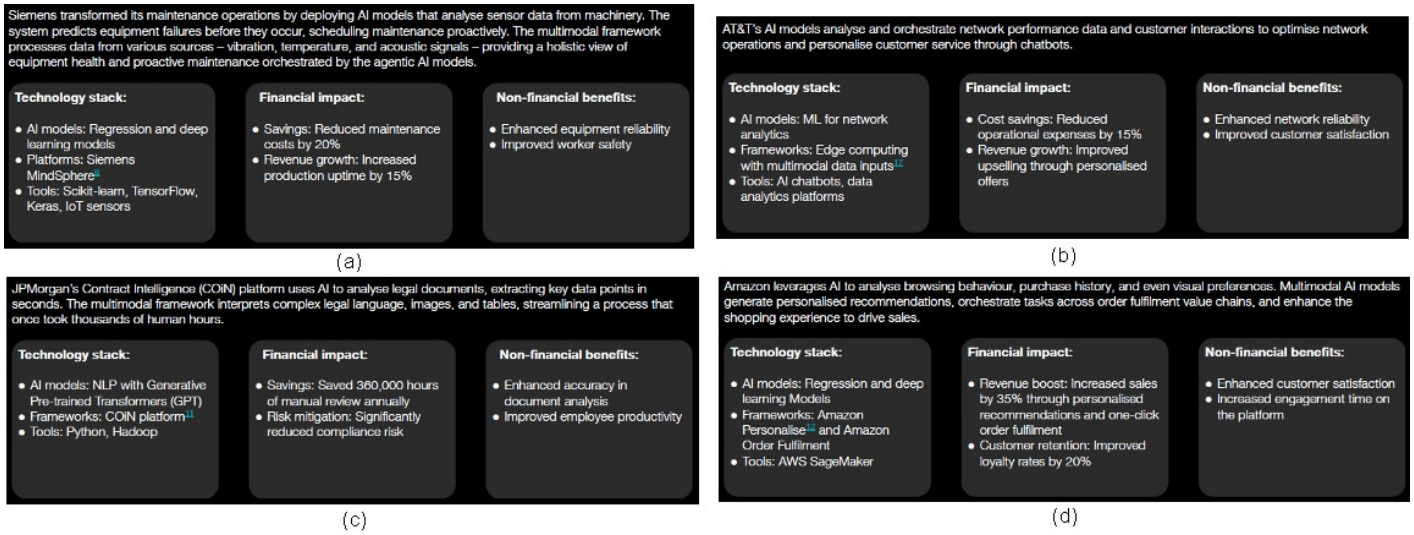
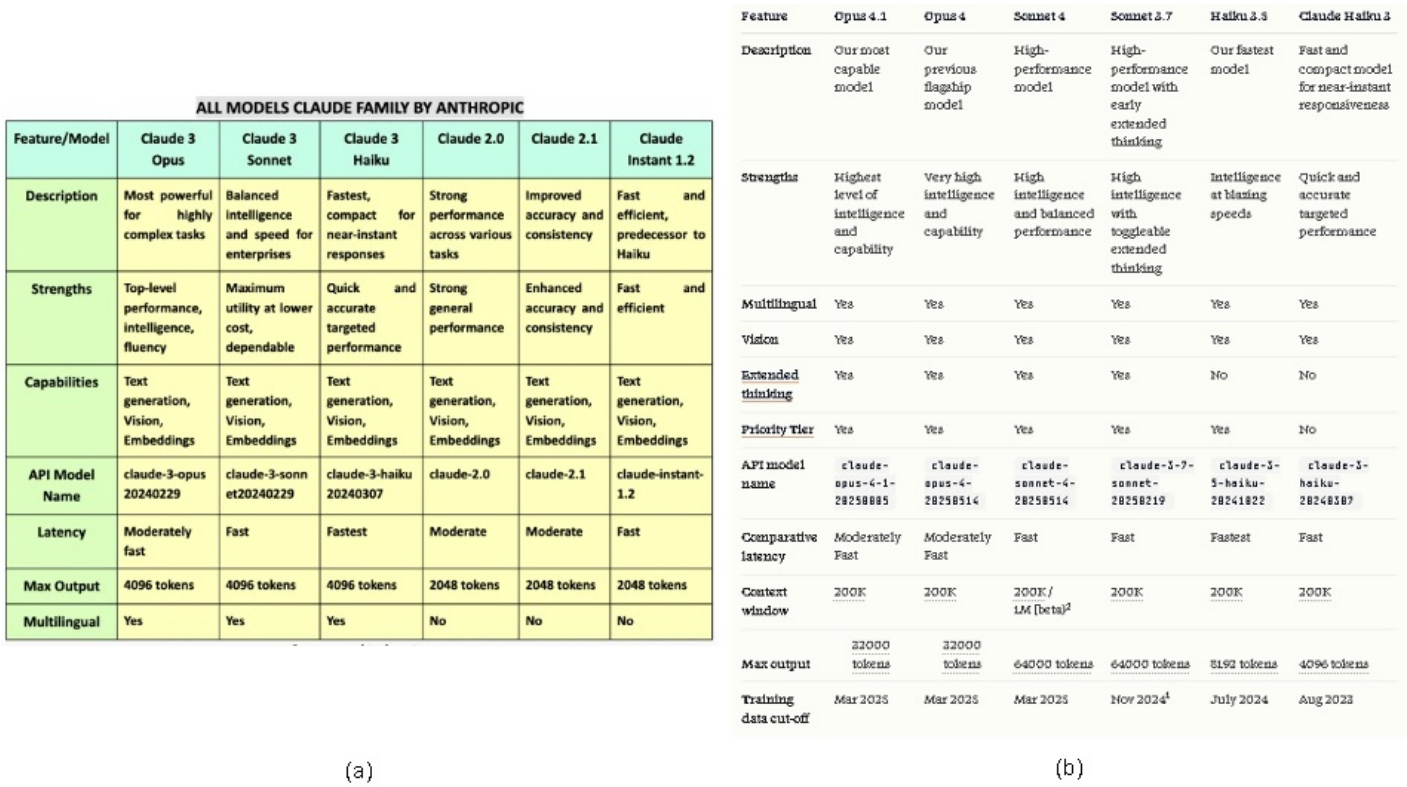


Figure 4: GenAI and agentic AI changing industries [23], a few examples: (a) Manufacturing: Siemens [24] (b) Telecommunications: AT&T [25] (c) Finance: JPMorgan Chase [26] (d) Retail: Amazon [27].

model Opus 4.1 is used for high complex tasks in coding, advanced reasoning, and orchestrating sophisticated, long-running agentic workflows. Designed for scale, Sonnet 4 balances intelligence with speed and efficiency, powers enterprise applications, and acts as a capable sub-agent within larger systems. Input and ouput Claude 4 model tokens per million are priced as follows: Opus ($15 and $75 respectively) and Sonnet ($3 – $6 for prompts $>$ 200K tokens – and $15, respectively).

Llama (Large Language Model Meta AI), a family of LLMs, was initially released by Meta as a foundation model (FM) and later, starting with Llama 2, also as instruction fine-tuned versions alongside FMs. Figure 6(a) shows a comparison of Llama, LLama2, and Llama3. The latest generation Llama 4 includes the Behemoth, Maverick, and Scout models with the following amounts of active parameters, experts, and total parameters: {288B, 16, 2T}, {17B, 128, 400B}, and {17B, 16, 109B}, respectively. Figure 6(b) to (d) shows a comparison between Llama 3.3 70 B and Llama 4 Scout. Llama 3.3 70 B is a model that focuses on efficiency, coding performance, and lower hardware requirements, suitable for general-purpose applications and local deployment. Llama 4 Scout was designed for advanced reasoning and extended memory, based on a mixture-of-experts (MoE) architecture with multimodal capabilities: capable of understanding text and images, and long context: 10 million token context window.

Best practices for LLMs in production include of course evaluating LLMs [33], industry benchmarks for model evaluation [34], and doing it not only by evaluating the basic capability of language modeling, but performing more specialized activities like language translation, text summarization, programming, question answering based on pre-training or on evidence, commonsense reasoning, math, or after applying fine tuning, Retrieval Augmented Generation (RAG) or human alignment [35] the latter three of which are all techniques for LLM improvement. Fine-tuning adapts a pre-trained LLM to specific tasks or domains. RAG enhances the ability of LLMs to access and utilize external information at runtime. Human alignment ensures that the LLM outputs align with human values, promoting safety and ethical behavior. Figure 6(d) shows the evaluation of Llama 3.3 70 B and Llama 4 Scout using three industry benchmarks apart from the price comparison. LiveCodeBench [36] is a benchmark for evaluating LLMs for code that mitigates contamination issues by introducing live evaluations and emphasizing scenarios beyond code generation. MMLU-Pro [37] is a benchmark for evaluating the multi-task language understanding and reasoning capabilities of LLMs. It is more difficult and robust than the original Massive Multitask Language Understanding (MMLU) benchmark [38] after incorporating more challenging, reasoning-intensive questions and increasing the number of answer choices. The Graduate-Level Google-Proof Q&A (GPQA) benchmark [39] is a dataset of challenging 448 multiple-choice questions designed to evaluate the reasoning and



ALL MODELS CLAUDE FAMILY BY ANTHROPIC

Feature/Model	Claude 3 Opus	Claude 3 Sonnet	Claude 3 Haiku	Claude 2.0	Claude 2.1	Claude Instant 1.2
Description	Most powerful for highly complex tasks	Balanced intelligence and speed for enterprises	Fastest, compact for near-instant responses	Strong performance across various tasks	Improved accuracy and consistency	Fast and efficient, predecessor to Haiku
Strengths	Top-level performance, intelligence, fluency	Maximum utility at lower cost, dependable	Quick and accurate targeted performance	Strong general performance	Enhanced accuracy and consistency	Fast and efficient
Capabilities	Text generation, Vision, Embeddings	Text generation, Vision, Embeddings	Text generation, Vision, Embeddings	Text generation, Vision, Embeddings	Text generation, Vision, Embeddings	Text generation, Vision, Embeddings
API Model Name	claude-3-opus 20240229	claude-3-sonn et20240229	claude-3-haiku 20240307	claude-2.0	claude-2.1	claude-instant-1.2
Latency	Moderately fast	Fast	Fastest	Moderate	Moderate	Fast
Max Output	4096 tokens	4096 tokens	4096 tokens	2048 tokens	2048 tokens	2048 tokens
Multilingual	Yes	Yes	Yes	No	No	No

(a)

Feature	Opus 4.1	Opus 4	Sonnet 4	Sonnet 3.7	Haiku 3.5	Claude Haiku 3
Description	Our most capable model	Our previous flagship model	High-performance model	High-performance model with early extended thinking	Our fastest model	Fast and compact model for near-instant responsiveness
Strengths	Highest level of intelligence and capability	Very high intelligence and capability	High intelligence and balanced performance	High intelligence with toggleable extended thinking	Intelligence at blazing speeds	Quick and accurate targeted performance
Multilingual	Yes	Yes	Yes	Yes	Yes	Yes
Vision	Yes	Yes	Yes	Yes	Yes	Yes
Extended thinking	Yes	Yes	Yes	Yes	No	No
Priority Tier	Yes	Yes	Yes	Yes	Yes	No
API model name	claude-opus-4-1-20250805	claude-opus-4-20250514	claude-sonnet-4-20250514	claude-3-7-sonnet-20250219	claude-3-5-haiku-20241022	claude-3-haiku-20240307
Comparative latency	Moderately Fast	Moderately Fast	Fast	Fast	Fastest	Fast
Context window	200K	200K	200K / 1M (beta)[2]	200K	200K	200K
Max output	32000 tokens	32000 tokens	64000 tokens	64000 tokens	8192 tokens	4096 tokens
Training data cut-off	Mar 2025	Mar 2025	Mar 2025	Nov 2024[1]	July 2024	Aug 2023

(b)

Figure 5: Anthropic Claude models (a) Comparative Overview of Claude 3 Models and before [28] (b) Current Claude family models [29]

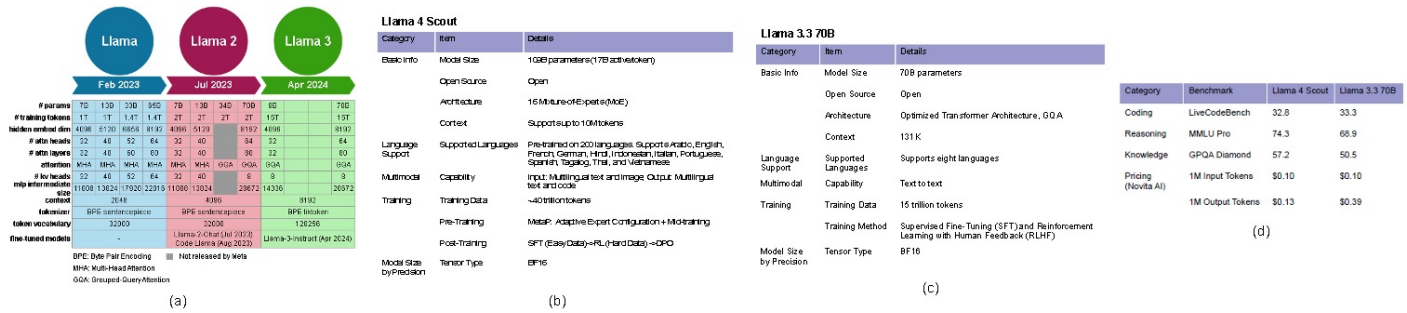


Figure 6: Meta Llama models (a) First three Llama generations [31] (b) Llama 4 Scout (c) Llama 3.3 70B (d) Benchmark comparison [32].

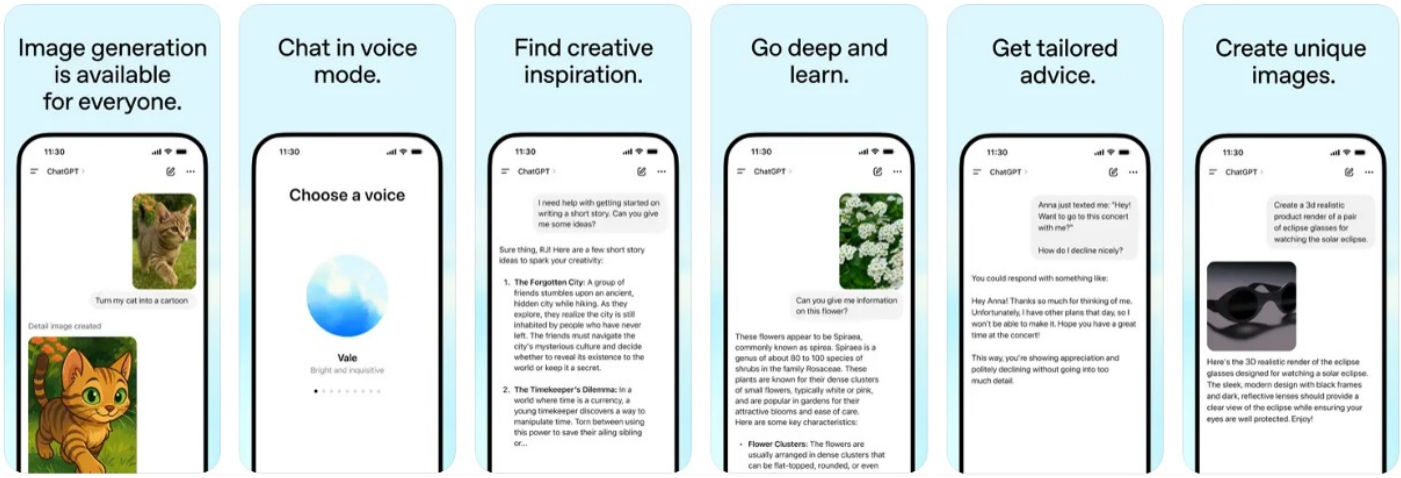


Figure 7: ChatGPT for the iPhone on the App Store [43]

GPT Model Comparison Overview

Model	Release Year	Parameters	Context Window	Architecture	Key Capabilities
GPT-3	2020	175B	2,048 tokens	Dense transformer	Basic language modeling, no alignment tuning. Struggled with coherence in long texts.
GPT-3.5	2022	~175B (likely)	~4,096 tokens (est.)	Dense + RLHF	RLHF + instruction tuning. Dramatically better coherence, task-specific reliability.
GPT-4	2023	~1T (est.)	8K–32K tokens	Mixture of Experts (MoE)	Strong reasoning, few-shot performance, multimodal via bolt-ons.
GPT-4 Turbo	Late 2023	~1T (est.)	128K tokens	Optimized MoE	Faster/cheaper inference. Efficient with long documents, stable latency.
GPT-4o	2024	~1T (est.)	128K tokens	Native multimodal	Text, image, and audio understanding. Real-time interaction and low latency.
GPT-5 *(Expected)*	2025 (est.)	1.5–3T (expected)	**1M+ tokens (expected)**	Sparse MoE (10–30% active)	Long-term memory, personalization, multi-agent planning, and deeper general reasoning.

(a)

Previous model	GPT-5 model
GPT-4o	gpt-5-main
GPT-4o-mini	gpt-5-main-mini
OpenAI o3	gpt-5-thinking
OpenAI o4-mini	gpt-5-thinking-mini
GPT-4.1-nano	gpt-5-thinking-nano
OpenAI o3 Pro	gpt-5-thinking-pro

(b)

Figure 8: GPT Model Comparison (a) Overview [44] (b) GPT 5 models from system card [45] .

problem-solving abilities of both humans and AI systems in complex scientific domains. The questions are not easily solved by searching the web. Thus, they are Google-proof. The GPQA Diamond benchmark is a subset of the GPQA benchmark containing the most challenging questions: 198 multiple-choice questions in biology, physics, and chemistry.

Open AI's current flagship model, GPT-5 [40] is a fast, high-throughput, multimodal (text, images, code, voice, short video clips) LLM, a deeper reasoning model that uses a real-time router to decide which model, current or previous, to use, based on conversation type, complexity, tool needs, and explicit intent. GPT-5 can be accessed by users through ChatGPT [41], cf. Figure 7, and by developers through the OpenAI API [42]. A GPT model comparison is shown in Figure 8(a). New improvements include faster response time, better coding and writing skills, and lower levels of hallucination. ChatGPT Voice replaces ChatGPT's Advanced Voice Mode to enable more natural-sounding conversations. The two fast, high-throughput models are gpt-5-main and gpt-5-main-mini. The two thinking models are gpt-5-thinking and gpt-5-thinking-mini, cf. Figure 8(b). gpt-5-thinking-nano is an even smaller and faster nano version of the thinking model in the API. gpt-5-thinking-pro allows the access to gpt-5-thinking in ChatGPT, used as an expert assistant for tasks demanding the absolute highest level of intelligence and accuracy. Some initial benchmarking of GPT 5 is shown in Figure 9. Software engineering and multi-language code editing tasks are evaluated using the SWE-bench Verified and the Aider Polyglot benchmarks scoring {74.9%, 52.8%} and {88.0%, 26.7%}

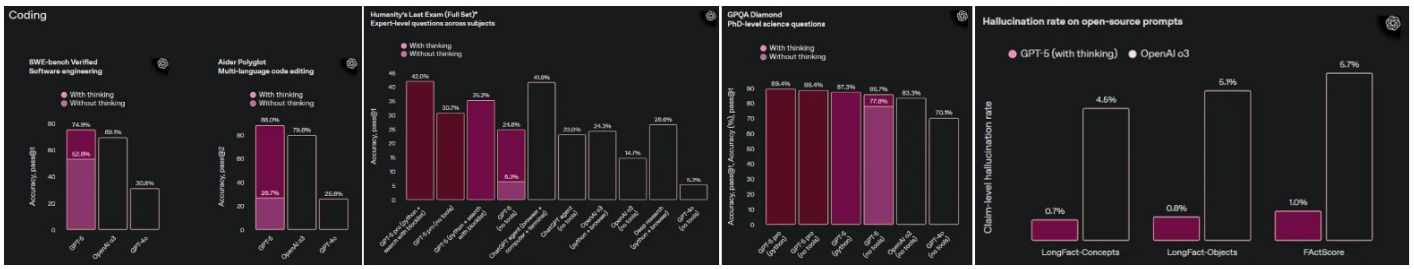
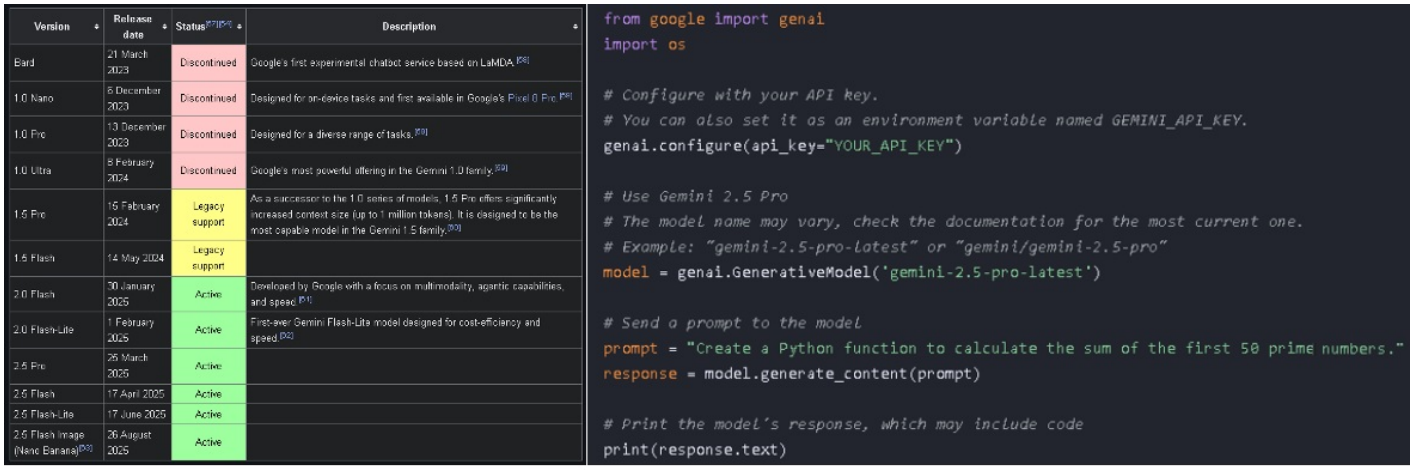


Figure 9: Benchmarking GPT 5 [46] (a) Coding (b) Humanity's Last Exam (c) GPQA Diamond (d) Hallucination rate on open-source prompts.

with and without thinking, respectively. In the Humanity's Last Exam test and the GPQA Diamond (PhD-level science questions) benchmarks, GPT 5 Pro with tools access scored 42.0% and 89.4%, respectively. GPT 5 also hallucinates much less than previous OpenAI models. For example, in LongFace-Concepts, its hallucination rate is 0.7% compared to 4.5% with OpenAI o3 .

Google DeepMind's Gemini is a family of multimodal LLMs and an equally named chatbot, successor of the former LaMDA, PaLM 2, and Bard. It was designed to be capable of processing multiple types of data simultaneously: text, images, audio, video, and computer code. Gemini 1.0 was launched on December 6, 2023 comprising 3 models: Gemini Ultra for highly complex tasks Gemini Pro for a wide range of tasks, and Gemini Nano for on-device tasks. Gemini 1.5 was launched in February, 2024 as a more powerful and capable model than Gemini 1.0 Ultra. Gemini 1.5 used a mixture-of-experts (MoE) approach and offered a larger one-million-token context window, which allowed for roughly an hour of silent video, 11 hours of audio, 30,000 lines of code, or 700,000 words. The same month, a family of free and open-source LLMs: Google Gemma debuted as a lightweight version of Gemini. On January 30, 2025 and on February 5, 2025, Gemini 2.0 Flash and Gemini 2.0 Pro were released, respectively. The general availability of Gemini 2.5 Pro and Gemini 2.5 Flash was announced on June 17, 2025. A model optimized for speed and cost-efficiency, Gemini 2.5 Flash-Lite, was introduced that same day. The Gemini 2.5 Pro [47] model is best for coding and highly complex tasks. Figure 10(a) shows a table listing the main Gemini model versions and describing the significant changes included with each version. After obtaining an API key from Google AI Studio, developers can use the Gemini API client libraries for Python or JavaScript, or make direct REST calls, to send prompts and receive responses that can include text, executable code, and code execution results. Figure 10(b) shows how to configure the environment and use the gemini-2.5-pro model in Python as an example of that.

The evolution timeline of xAI Grok, a family of large language models (LLMs) and the AI-powered chatbot has been as follows: the 33B parameter dense transformer architecture Grok-0 was completed in late 2023. followed by Grok-1, which was launched in November 2023. Grok 1.5 was released on May 15, 2024 with improved reasoning capabilities and a context length of 128k tokens. Grok-2 and Grok-2 mini were released on August 14, 2024 with upgraded performance and reasoning as well as image generation capability. Grok 3 was released on February 17, 2025 after being trained with 10x more computing power than its Grok-2 predecessor at xAI's Colossus data center. The new flagship multimodal models Grok 4 [48] and Grok 4 Heavy were released on July 9, 2025, claiming being able to achieve PhD-level academic performance. To build own applications, the Grok 4 API provides developers programmatic access to the model. A fast reasoning model variant built for agentic coding is grok-code-fast-1. Its most powerful version is Grok 4 Heavy, which is based on multi-agent collaboration, cf. Figure 11(a), and requires significant computational resources. It uses several independent AI agents that tackle different parts of a prompt simultaneously. These AI agents conference together to compare their findings and synthesize the optimal response. The Heavy variant's GPU consumption is double that of the standard Grok-4. One of the latest LLM benchmarks introduced is the Humanity's Last Exam (HLE) [49], an LLM benchmark created jointly by the Center for AI Safety and Scale AI consisting of 2,500 challenging questions across over a hun-

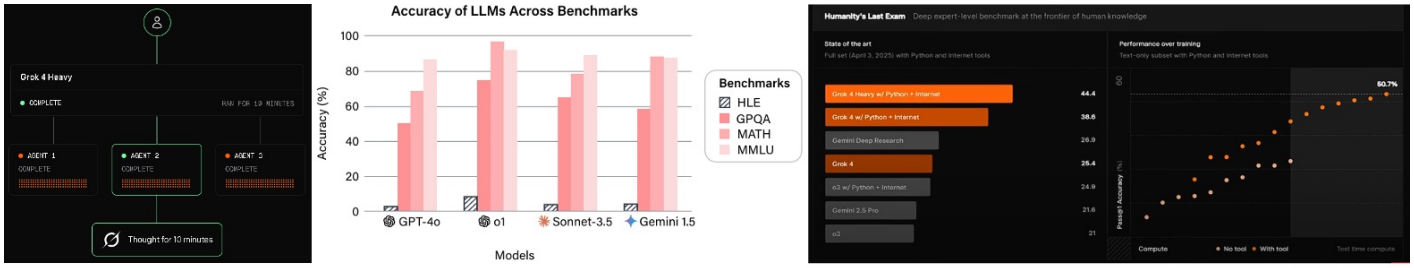


Version	Release date	Status[97][98]	Description
Bard	21 March 2023	Discontinued	Google's first experimental chatbot service based on LaMDA.[95]
1.0 Nano	6 December 2023	Discontinued	Designed for on-device tasks and first available in Google's Pixel 8 Pro.[96]
1.0 Pro	13 December 2023	Discontinued	Designed for a diverse range of tasks.[97]
1.0 Ultra	8 February 2024	Discontinued	Google's most powerful offering in the Gemini 1.0 family.[98]
1.5 Pro	15 February 2024	Legacy support	As a successor to the 1.0 series of models, 1.5 Pro offers significantly increased context size (up to 1 million tokens). It is designed to be the most capable model in the Gemini 1.5 family.[99]
1.5 Flash	14 May 2024	Legacy support	
2.0 Flash	30 January 2025	Active	Developed by Google with a focus on multimodality, agentic capabilities, and speed.[91]
2.0 Flash-Lite	1 February 2025	Active	First-ever Gemini Flash-Lite model designed for cost-efficiency and speed.[92]
2.5 Pro	26 March 2025	Active	
2.5 Flash	17 April 2025	Active	
2.5 Flash-Lite	17 June 2025	Active	
2.5 Flash Image (Nano Banana)[93]	26 August 2025	Active	

```python
from google import genai
import os

# Configure with your API key.
# You can also set it as an environment variable named GEMINI_API_KEY.
genai.configure(api_key="YOUR_API_KEY")

# Use Gemini 2.5 Pro
# The model name may vary, check the documentation for the most current one.
# Example: "gemini-2.5-pro-latest" or "gemini/gemini-2.5-pro"
model = genai.GenerativeModel('gemini-2.5-pro-latest')

# Send a prompt to the model
prompt = "Create a Python function to calculate the sum of the first 50 prime numbers."
response = model.generate_content(prompt)

# Print the model's response, which may include code
print(response.text)
```

(a) (b)

Figure 10: Google DeepMind's Gemini models [GOO] (a) List of model versions and main changes introduced (b) Use of the Gemini API with the gemini-2.5-pro model in Python.

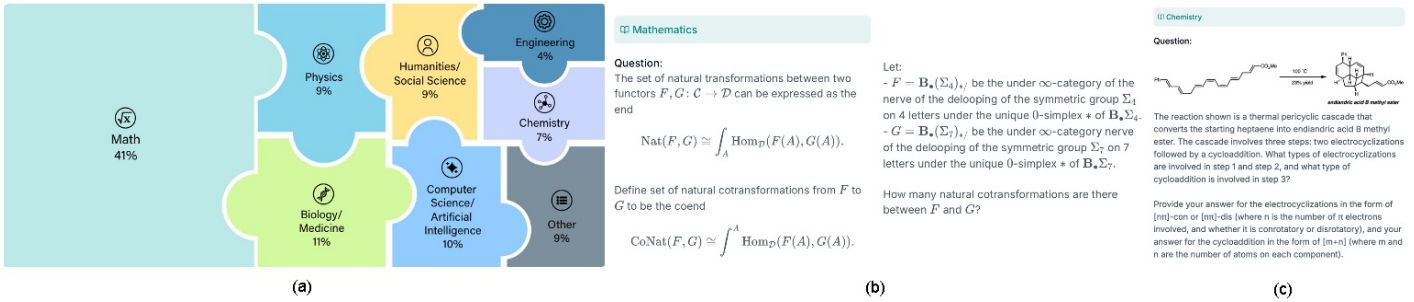


(a) (b) (c)

Figure 11: xAI Grok models [XAI] (a) Multi-agent models (b) LLM benchmark performance by some frontier models [49] (c) Grok 4 models versus Gemini, o3 on HLE benchmark.

dred subjects, including mathematics, humanities, and the natural sciences. Figure 12(a), (b), and (c) show the percentage composition of the high level categories of the subjects covered, a sample Mathematics question, and a sample Chemistry question, respectively. The HLE benchmark was developed in response to the fact that popular AI benchmarks having reached saturation, i.e., the leading LLMs were performing with a high accuracy on those benchmarks.

Figure 11(b) shows the saturation of some popular LLM benchmarks {Graduate-Level Google-Proof Q&A (GPQA) [39], the MATH dataset benchmark [50], Massive Multitask Language Understanding (MMLU) [38]} achieved by some frontier models {OpenAI GPT-4o, OpenAI o1, Anthropic Claude Sonnet-3.5, Google Gemini 1.4} compared with the newly introduced Humanity's Last Exam (HLE) [49]. As an example, OpenAI o1 was introduced in September 2024, designed for complex

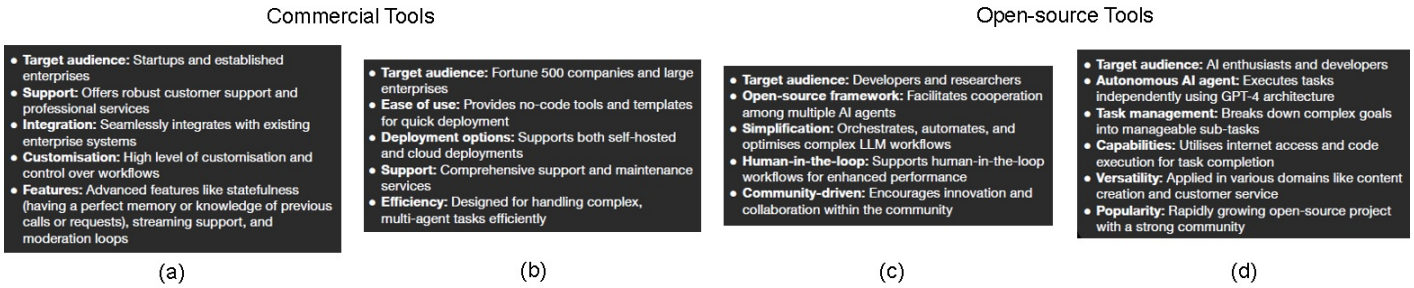


(a) (b) (c)

Figure 12: Humanity's Last Exam (HLE) [49] (a) High level categories of subjects covered (b) sample Mathematics question (c) sample Chemistry question.

(a) (b) (c)

Figure 13: xAI Collossus AI Supercomputer [XAI] (a) AI training data center facility in Memphis, TN (b) Server H100/H200 GPU cluster infrastructure (c) World's most powerful computer clusters in June 2025.

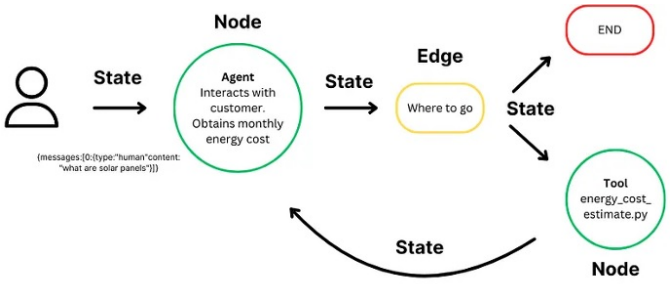


(a) (b) (c) (d)

Figure 14: Agentic AI Tools [23]: (a) LangGraph [52] (b) CrewAI [53] (c) AutoGen [54] (d) AutoGPT [55].

reasoning in science, math, and coding, aimed to solve harder problems than previous models by spending more time deep, step-by-step thinking and generating long Chain-of-Thought (CoT) output. On the other hand, two MATH problems and solutions are shown to illustrate the complexity the LLMs are capable of handling with that benchmark. Problem 1: Tom has a red marble, a green marble, a blue marble, and three identical yellow marbles. How many different groups of two marbles can Tom choose?. Solution of Problem 1: There are two cases here: either Tom chooses two yellow marbles (1 result), or he chooses two marbles of different colors ($\binom{4}{2} = 6$ results). The total number of distinct pairs of marbles Tom can choose is 1+6 = $\boxed{7}$. Problem 2: The equation $x^2 + 2x = i$ has two complex solutions. Determine the product of their real parts. Solution of Problem 2: Complete the square by adding 1 to each side. Then $(x+1)^2 = 1+i = \sqrt{2}\cdot e^{\frac{i\cdot\pi}{4}}$, so $x+1 = \pm\sqrt[4]{2}\cdot e^{\frac{i\cdot\pi}{8}}$. The desired product is then $(-1+\sqrt[4]{2}\cdot\cos(\frac{\pi}{8}))(-1-\sqrt[4]{2}\cdot\cos(\frac{\pi}{8})) = 1-\sqrt{2}\cdot\cos^2(\frac{\pi}{8}) = 1-\sqrt{2}\cdot\frac{1+\cos(\frac{\pi}{4})}{2} = \boxed{\frac{1-\sqrt{2}}{2}}$. Figure 11(c) shows Grok 4 Heavy w/ Python + Internet outperforming (April 3, 2025) Grok 4 w/ Python + Internet, Gemini Deep Research, Grok 4, o3 w/ Python + Internet, Gemini 2.5 Pro, and o3 on the HLE benchmark. Some lack of maturity of the emerging benchmarks appears to be given. For example, it has been reported that about 30% of Humanity's Last Exam (HLE) text-only chemistry/biology answers are likely to be wrong [51], i.e., with directly conflicting evidence in peer reviewed literature. Figure 13 shows the servers, NVIDIA H100/H200 GPU clusters hosted inside the Collossus AI supercomputer at xAI's training data center facility in Memphis, TN.

Four agentic AI tools, two commercial and two open-source, are shown in Figure 14. For example, LangGraph is an open-source platform for deploying AI agents that can scale with production volume. It provides easy-to-use APIs for managing agent state, memory, and user interactions. When ready to ship to production, it can gracefully handle large workloads, with features like retries and cost-efficient execution for reliable performance. A small example illustrates some of the main steps in LLM application development using LangGraph, cf. Figure 15 and Figure 16. The goal is to generate an AI agent assistant for customer support hat can calculate solar panel energy savings based



(a)

```python
from langchain_core.tools import tool
from langchain_community.tools.tavily_search import TavilySearchResults
from langchain_core.prompts import ChatPromptTemplate
from langchain_core.runnables import Runnable
from langchain_aws import ChatBedrock
import boto3
from typing import Annotated
from typing_extensions import TypedDict
from langgraph.graph.message import AnyMessage, add_messages
from langchain_core.messages import ToolMessage
from langchain_core.runnables import RunnableLambda
from langgraph.prebuilt import ToolNode
from langgraph.prebuilt import tools_condition
```

(b)

Figure 15: Building AI Agents with LangGraph [56]: (a) Nodes, states, and edges of a stateful graph. (b) Importing Python libraries and modules.

```python
def get_bedrock_client(region):
    return boto3.client("bedrock-runtime", region_name=region)

def create_bedrock_llm(client):
    return ChatBedrock(model_id='anthropic.claude-3-sonnet-20240229-v1:0',
    client=client, model_kwargs={'temperature': 0}, region_name='us-east-1')
llm = create_bedrock_llm(get_bedrock_client(region='us-east-1'))
```

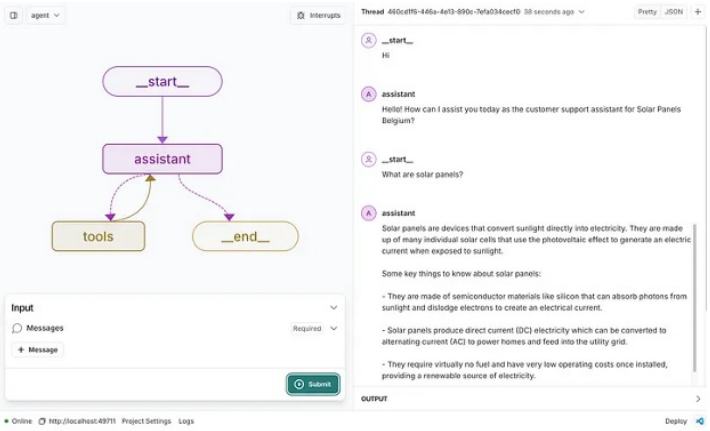
(a)

```python
# Define the tools the assistant will use
part_1_tools = [
    compute_savings
]
# Bind the tools to the assistant's workflow
part_1_assistant_runnable = primary_assistant_prompt | llm.bind_tools(part_1_tools)
                                                                    tools)
```

(b)

```python
builder = StateGraph(State)
builder.add_node("assistant", Assistant(part_1_assistant_runnable))
builder.add_node("tools", create_tool_node_with_fallback(part_1_tools)) ))
builder.add_edge(START, "assistant")  # Start with the assistant
builder.add_conditional_edges("assistant", tools_condition)  # Move to tools after input
builder.add_edge("tools", "assistant")  # Return to assistant after tool execution
memory = MemorySaver()
graph = builder.compile(checkpointer=memory)
```

(c)

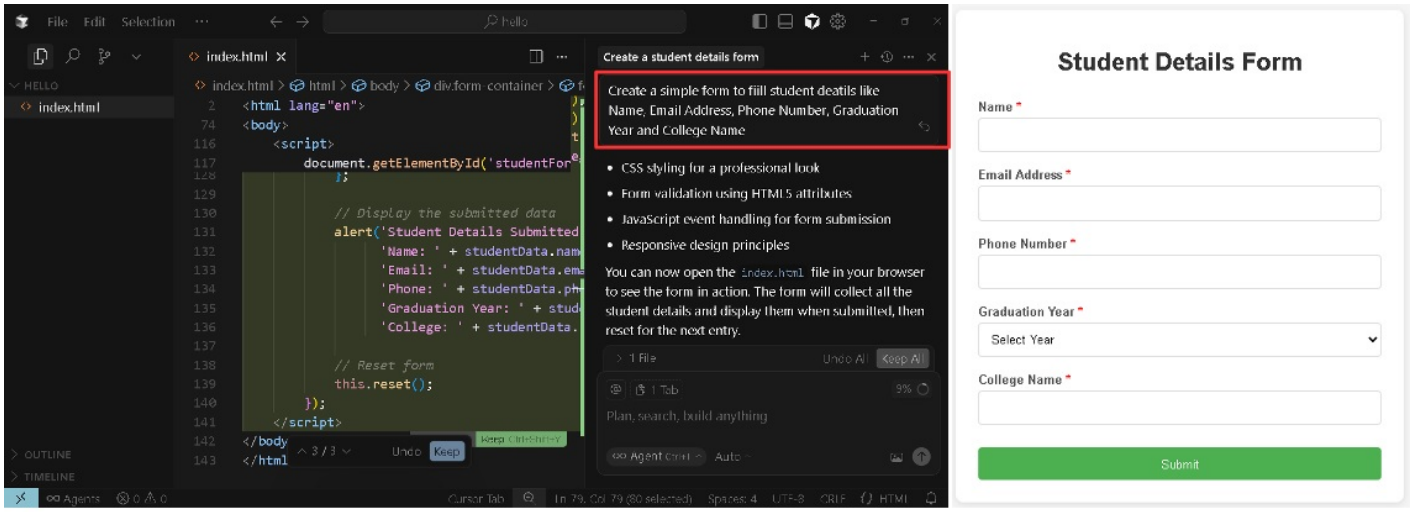


(d)

Figure 16: Building AI Agents with LangGraph [56]: (a) Using AWS Bedrock to configure LLM: Anthropic Claude3 Sonnet (b) Binding the tools to the assistant (c) Defining nodes (assistant, tools), edges, and using MemorySaver to retain the conversation state (d) AI agent graph structure, running and testing it with LangGraph Studio

on user inputs [56]. LangChain is an open-source software framework for LLM application development that facilitates the integration of LLMs. It allows the creation of Directed Acyclic Graphs (DAGs) for linear workflows. LangGraph, an advanced library built on top of LangChain, enhances that capability by enabling the addition of cycles, essential for developing complex, agent-like behaviors which allow LLMs to continuously loop through a process, dynamically deciding what action to take next based on evolving conditions. Thus, it helps to build advanced AI agents by enabling stateful, multi-actor applications with cyclic computation capabilities. LangGraph Studio is an Integrated Development Environment (IDE) built specifically for visualizing, testing, and debugging LangChain-based AI agents. It offers a mix of visual tools and real-time code editing to easily manage graph execution, make changes to the agent's logic, and see the results instantly as opposed to doing all this only with code.

To gain further insight into recent developments, capabilities of additional agentic AI tools and frameworks are showcased. AI coding tools like Cursor [57], Windsurf [58], and Copilot [59] provide an IDE to turn LLMs into coding assistants to boost the productivity of human software developers. A preliminary example shown in Figure 17 conveys some basic ideas of using the Anysphere Cursor code editor. To provide some initial context a minimal project (folder) has been created with the html code for hello world. The requirements are then provided for a coding task: a website form is to be developed to fill student details. In this basic example, a "New Chat" is entered with the prompt "Create a simple form to fill student details like Name, Email Address, Phone Number, Graduation Year and College Name". Figure 17(a) shows the three panels of Cursor: on the left, the File Explorer showing the only one file in this project: index.html. In the middle, part of the autonomously



(a) (b)

Figure 17: Anysphere Cursor Code Editor based on Microsoft Visual Studio Code (a) index.html file autonomously created on top of the original hello world code, prompt inside the red box (New Chat) (b) Requested form to fill student details in the browser.

generated html code is shown. On the right, the Chat prompt and part of the results' explanations are shown. Figure 17(b) shows the form displayed in the browser when running the code. Some of the Cursor Settings used from a large number include: Default Mode (Mode for new chats) [Agent, the other option is Ask], Web Search Tool (Allow Agent to search the web for relevant information) [Yes], Auto-Fix Lints (Automatically fix lint errors in the chat) [Yes], Auto-Accept on Commit (Automatically accept all changes when files are committed and no longer in the worktree) [Yes]. At the bottom of the chat panel, next to "Agent" (agent mode), "Auto" (AI model) is the default selection, which indicates "Balanced quality and speed, recommended for most tasks". Enabling Auto allows Cursor to select the premium model best fit for the immediate task and with the highest reliability based on current demand. This feature can detect degraded output performance and automatically switch models to resolve it. Cursor supports all frontier coding models from all major model providers {OpenAI, Anthropic, Google, xAI} including Claude 4 Sonnet, Claude 4.1 Opus, Gemini 2.5 Flash and Pro, GPT 4.1 and GPT-5, Grok, and variants. For more complex software development tasks, asking for a plan before action including waiting for approval before coding is often the way to go. Additional measures include the use of Product Requirements Documents (PRDs) and controlling the tech stack with such details to be incorporated in a document that becomes a part of the prompt context. To avoid misleading guessing and keep the agent focused, the agent needs to be provided contextual awareness of any domain-specific terms that are not obvious from the codebase alone by including associated details in README.md, JSON, and other files which the agent can read.

Moving from single-use LLMs to networks of intelligent agents, those agents need to communicate with one another using standardized protocols like Google's Agent2Agent (A2A) [60] and Anthropic Model Context Protocol (MCP) [61], which often serves as a bridge to traditional APIs defined by another open standard: OpenAPI [62]. See Figure 18(a) for an example of how agents cooperate by building networks and using those standardized protocols in agentic AI systems. An event-driven architecture (EDA) provides a natural foundation for agentic AI systems that thrive in environments that are dynamic, distributed, and data-rich. Combining event-driven principles with the aforementioned collaboration between AI agents leads to a reference architecture for event-driven agentic AI [63], cf. Figure 18(b). W.r.t. to the further development of foundation models (FMs) and Large Language Models (LLMs) in general, the tradeoff of training models with generalist versus specialist domain-specific data or expanding on this trend of projection, the associated question of how to improve the trustworthiness of the model's responses continues to be addressed, i.e., the chatbot's and assistant's training data annotation strategies. Another hot topic for expansion is multimodal-

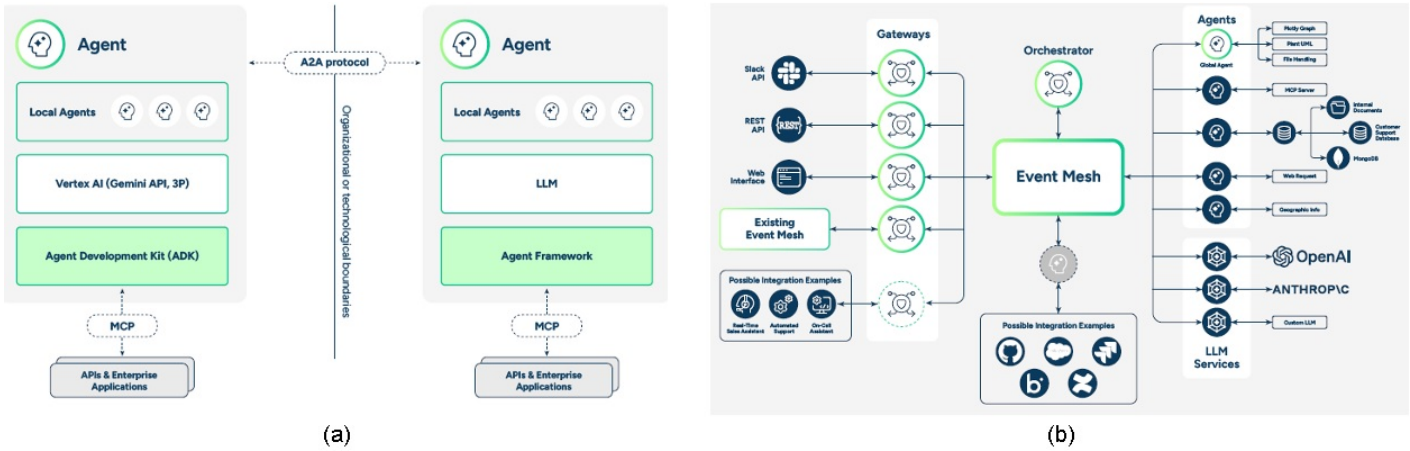


Figure 18: Event-driven Agentic AI systems [63] [SOL] (a) Emergent agent protocols use agent collaboration (b) Reference architecture for event-driven agentic AI.

ity, e.g., models need to perform better on images and video. More specific issues to be addressed involve transparency and hallucination mitigation, for example. At the cutting edge of innovative agentic AI systems that can assist in solving even more complex problems, reserved so far for humans, are those emerging to attack unsolved problems in Mathematics, as described in [64]. In this report, the efficient automation and streamlining of business workflows is handled in detail. The adequate use of AI-powered insights can substantially enhance real-time decision-making in organizations and the adequate use of AI agents can vehemently position organizations with clear competitive advantages. These facts are meticulously elaborated with practical and diverse business use cases. Multi-agent AI systems can provide responsible scalable flexibility to business, industrial workflow solutions when advanced security measures to protect data are incorporated as an integral part of those agentic AI solutions.

References

[1] V.D. Sánchez. *Neurocomputing 50th volume anniversary*. Neurocomputing, 50, ix, 2003. https://profdrvdsaphd.lima-city.de/documents/Neurocomputing50thAnniversary.pdf

[2] V.D. Sánchez. *IEEE Fellow Award – "for leadership in neural and parallel computation, and pioneering contributions to autonomous space robots"*. 1995. https://profdrvdsaphd.lima-city.de/documents/IEEEFellow.pdf

[3] V.D. Sánchez. *Personal written communication with the Office of the German Federal Minister of Research and Technology BMFT – Start of the First German National AI/ML Research and Technology Development Program*. 1988.

[4] V.D. Sánchez. *Advanced Automation for mission-critical Information Technology past AIOps*. November 2024. https://profdrvdsaphd.lima-city.de/documents/AdvancedAutomationMisCriITPastAIOps.pdf

[5] Databricks. *A Compact Guide to Large Language Models*, 2025.

[6] McKinsey & Company. *The promise and the reality of gen AI agents in the enterprise*, May 2024.

[7] OpenAI. *Chat GPT – Overview*. May 2025. https://openai.com/chatgpt/overview/

[8] Ollama. *Ollama – Get up and running with Llama 3.3, DeepSeek-R1, Phi-4, Gemma 3, Mistral Small 3.1 and other large language models.*. May 2025. https://github.com/ollama/ollama

[9] A. Radford et al. *Improving Language Understanding by Generative Pre-Training*. OpenAI, June11, 2018.

[10] V.D. Sánchez. *Advanced Gen AI & LLM Foundations and Applications – Paving the way to a more powerful and diverse ML –*. December 2023. https://profdrvdsaphd.lima-city.de/documents/AdvancedGenAILLMs.pdf.

[11] World Economic Forum. *Digital Transformation Initiative, in collaboration with Accenture: Unlocking $100 Trillion for Business and Society from Digital Transformation*, Executive Summary, January 2017.

[12] McKinsey & Company. *McKinsey Global Institute: Notes from the AI frontier – Modeling the impact of AI on the world economy*, Discussion paper, September 2018.

[13] V.D. Sánchez. *Colonizing the Red Planet*. November 2022. https://profdrvdsaphd.lima-city.de/documents/MarsColonization.pdf

[14] V.D. Sánchez. *Enabling Robots to become more human and obtain superhuman capabilities for terrestrial and space applications using advanced AI/ML*. November 2024. https://profdrvdsaphd.lima-city.de/documents/EnablingSuperhumanRobotsUsingAdvancedAIML.pdf

[15] R.L. Bocchino Jr. et al. *F Prime: An Open-Source Framework for Small-Scale Flight Software Systems*. Small Satellites Conference (SmallSat 2018), Logan, Utah, August 4–9, 2018.

[16] R.L. Bocchino Jr et al. *FPP: A Modeling Language for F Prime*. 2022 IEEE Aerospace Conference, Big Sky, Montana, March 5-12, 2022.

[17] OpenAI. *Announcing The Stargate Project*. January 21, 2025. https://openai.com/index/announcing-the-stargate-project/

[18] V.D. Sánchez. *On Artificial General Intelligence (AGI) and Artificial Super Intelligence (ASI)*. January 2025. https://profdrvdsaphd.lima-city.de/documents/OnArtificialGeneralIntelligenceAndArtificialSuperIntelligence.pdf

[19] Q. Zhu et al . *Automated synthesis of oxygen-producing catalysts from Martian meteorites by a robotic AI chemist*. Nature Synthesis 3 (2024), 319–328.

[20] Statworx. *AI Trends Report 2025*. 2025. https://www.statworx.com/en/content-hub/whitepaper/ai-trends-report-2025

[21] V.D. Sánchez. *Deep Impact of Advanced AI Technology Developments – Government Regulatory Measures –*. September 2023. https://profdrvdsaphd.lima-city.de/documents/ArtificialIntelligenceRegulation.pdf

[22] Stanford University, Institute for Human-Centered Artificial Intelligence. *The AI Index 2025 Annual Report*. 8th Edition, 2025.

[23] PwC. *Agentic AI – the new frontier in GenAI, An executive playbook*. 2024.

[24] Siemens. *Insights Hub – MindSphere has evolved into Insights Hub*. 2025. https://plm.sw.siemens.com/en-US/insights-hub/

[25] AT&T. *Analytics and AI-based automation*. 2025. https://about.att.com/sites/labs/our-work/analytics-ai-automation

[26] Medium. *How JPMorgan Chase's COIN is Revolutionizing Financial Operations with AI*. June 26, 2024. https://medium.com/@ishan_dhodu/how-jpmorgan-chases-coin-is-revolutionizing-financial-operations-with-ai-120a2938dab7

[27] Amazon. *Amazon Personalize*. 2025. https://aws.amazon.com/personalize/

[28] Marktechpost. *Developments in Family of Claude Models by Anthropic AI: A Comprehensive Review*. May 26, 2024. https://www.marktechpost.com/2024/05/26/developments-in-family-of-claude-models-by-anthropic-ai-a-comprehensive-review/

[29] Anthropic. *Models & pricing – Model comparison table*. May 2025. https://docs.anthropic.com/en/docs/about-claude/models/overview

[30] Anthropic. *AI research and products that put safety at the frontier*. May 2025. https://www.anthropic.com/

[31] Devopedia . *Llama (LLM)*. 2024. https://devopedia.org/llama-llm

[32] Novita AI. *Llama 4 Scout vs. Llama 3.3 70B: Multimodal Excellence or Coding Efficiency?*. April 23, 2025. https://blogs.novita.ai/llama-4-scout-vs-llama-3-3-70b/

[33] P. Iusztin and M. Labonne. *LLM Engineer's Handbook: Master the art of engineering large language models from concept to production*. Packt Publishing, 2024.

[34] C. Brousseau and M. Sharp. *LLMs in Production: From language models to successful products*. Manning, 2025.

[35] B. Singh. *Building Applications with Large Language Models: Techniques, Implementation, and Applications*. APress, 2024.

[36] N. Jain et al. *LiveCodeBench: Holistic and Contamination Free Evaluation of Large Language Models for Code*. June 6, 2024. https://arxiv.org/abs/2403.07974v2

[37] Y. Wang et al. *MMLU-Pro: A MoreRobust and Challenging Multi-Task Language Understanding Benchmark*. November 6, 2024. https://arxiv.org/abs/2406.01574

[38] D. Hendrycks et al. *Measuring Massive Multitask Language Understanding*. January 12, 2021. https://arxiv.org/abs/2009.03300v3

[39] D. Rein et al. *GPQA: A Graduate-Level Google-Proof Q&A Benchmark*. November 20, 2023. https://arxiv.org/abs/2311.12022

[40] OpenAI. *GPT-5 is here*. 2025. https://openai.com/gpt-5/

[41] ChatGPT. *ChatGPT – Ask anything*. 2025. https://chatgpt.com/

[42] OpenAI. *OpenAI developer platform*. 2025. https://platform.openai.com/docs/overview

[43] Apple. *ChatGPT – The official app by OpenAI*. https://apps.apple.com/us/app/chatgpt/id6448311069?platform=iphone

[44] Felloai. *GPT-5 Release Date Confirmed: Here Is All We Know And What to Expect!*. https://felloai.com/2025/06/gpt-5-release-date-confirmed-here-is-all-we-know-and-what-to-expect/

[45] OpenAI. *GPT-5 System Card*. August 7, 2025. https://openai.com/index/gpt-5-system-card/

[46] Beebom. *OpenAI Launches GPT-5, Makes It Free for All ChatGPT Users*. August 7, 2025. https://beebom.com/openai-launches-gpt-5-makes-it-free-for-all-chatgpt-users/

[47] Google DeepMind. *Gemini 2.5 Pro*. 2025. https://deepmind.google/models/gemini/pro/

[48] xAI. *Grok – Models and Pricing*. 2025. https://docs.x.ai/docs/models

[49] Center for AI Safety and Scale AI. *Humanity's Last Exam*. April, 2025. https://agi.safe.ai/

[50] D. Hendrycks et al. *Measuring Mathematical Problem Solving with the MATH Dataset*. November 9, 2021. https://arxiv.org/pdf/2103.03874

[51] FutureHouse. *About 30% of Humanity's Last Exam chemistry/biology answers are likely wrong*. July 23, 2025. https://www.futurehouse.org/research-announcements/hle-exam

[52] LangChain Academy. *Foundation: Introduction to LangGraph*. 2025. https://academy.langchain.com/courses/intro-to-langgraph

[53] CrewAI. *The Leading Multi-Agent Platform*. 2025. https://www.crewai.com/

[54] Microsoft. *AutoGen – A framework for building AI agents and applications*. 2025. https://microsoft.github.io/autogen/stable/

[55] AutoGPT. *Empower your digital tasks with AutoGPT*. 2025. https://agpt.co/

[56] Medium. *How to Build AI Agents with LangGraph: A Step-by-Step Guide*. September 6, 2024. https://medium.com/@lorevanoudenhove/how-to-build-ai-agents-with-langgraph-a-step-by-step-guide-5d84d9c7e832

[57] Anysphere. *Cursor: The AI Code Editor*. 2025. https://cursor.com

[58] Windsurf. *The Windsurf Editor*. 2025. https://windsurf.com/editor/

[59] Microsoft. *Microsoft Copilot: Your AI companion*. 2025. https://copilot.microsoft.com/

[60] A2AProject. *Agent2Agent (A2A) Protocol Specification – Version: v0.2.5*. June 2025. https://a2a-protocol.org/v0.2.5/specification/

[61] Anthropic. *Model Context Protocol*. June 2025. https://modelcontextprotocol.io/specification/2025-06-18

[62] OpenAPI. *OpenAPI Specification v3.1.1*. October 2024. https://spec.openapis.org/oas/latest.html

[63] Solace. *The Architect's Guide to Event-Driven Agentic AI*. 2025.

[64] V.D. Sánchez. *Künstliche Intelligenz und Maschinelles Lernen zur Lösung ungelöster Probleme in der Mathematik (in German)*. March 2025. https://profdrvdsaphd.lima-city.de/documents/ArtificialGeneralIntelligenceMachineLearningOnMathematics.pdf