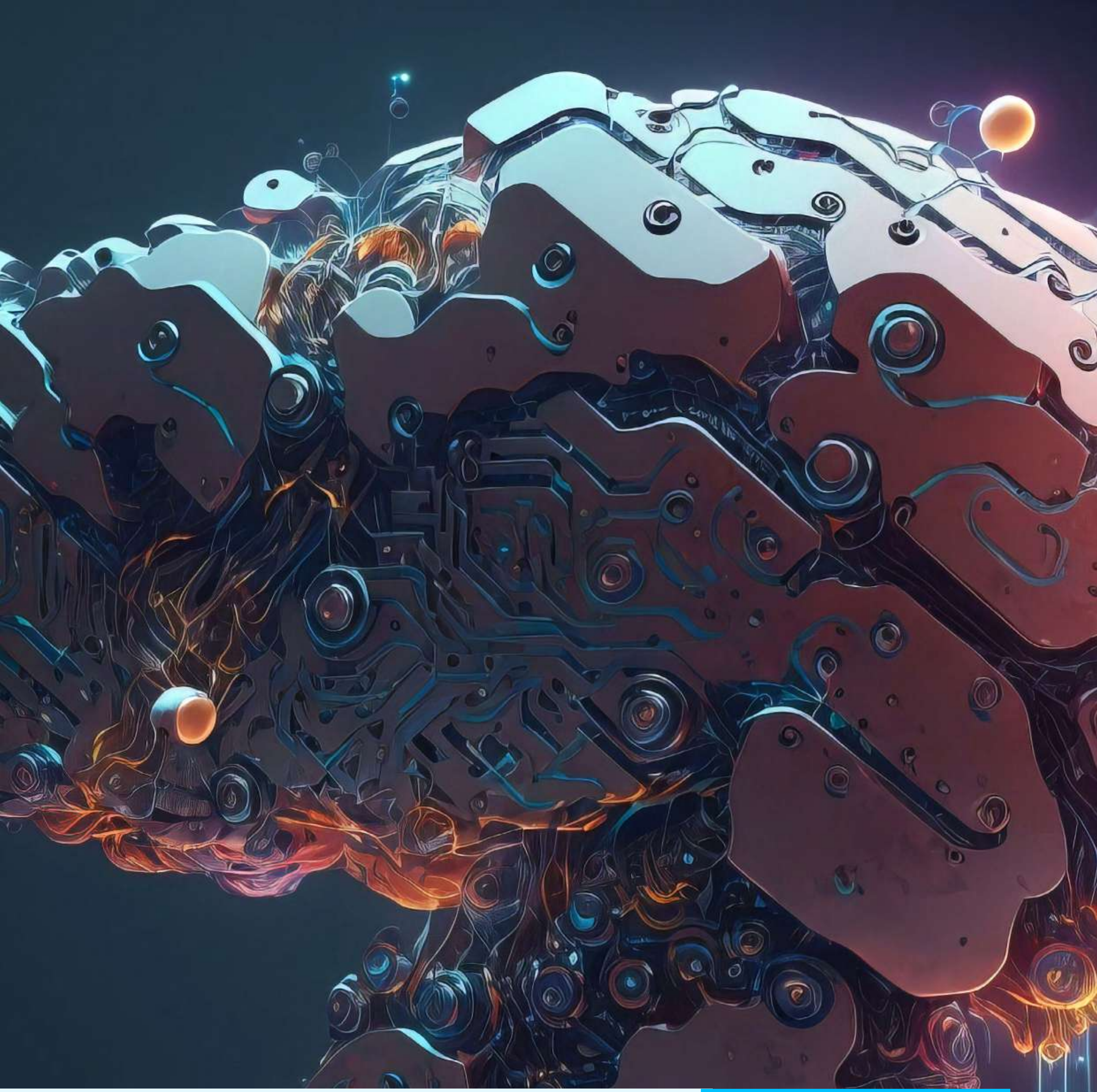


avenga

Generative AI in life sciences: staging a renaissance of biomedical discovery

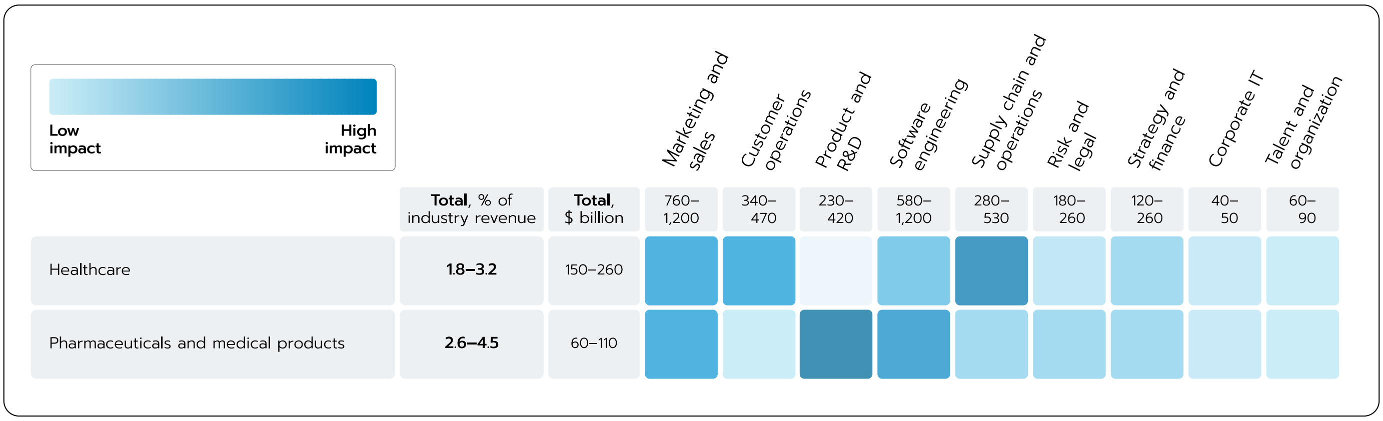
Content

Before we begin

Generative AI is riding a crest of popularity in all major domains. A McKinsey study highlights that the technology offers a value of **$2.6** trillion to **$4.4** trillion of potential improvements across **63** use cases. And today, life sciences is one of the major areas where essential progress can take place. Long burdened with the demand for accelerated research and development, the industry finds itself at the ideal juncture for an impactful change. In the pharma and medical products industries alone, the potential impact of generative AI implementation reaches **$60-110 billion**, which accounts for **3-5%** of their current revenue (see Fig. 1).

As generative AI promises to reshape industry approaches, it gives rise to a variety of new possibilities. From accelerated drug development to novel approaches in disease modeling, it has the potential to expedite the delivery of life-saving medications to patients and transform the way scientists explore life. While this technology brings challenges that require ethical foresight, its benefits feel undeniably far-reaching. That is why generative AI warrants substantial investment and research by leaders seeking to future-proof their organizations. Its moment in life sciences has arrived.

Generative AI Productivity impact by business functions



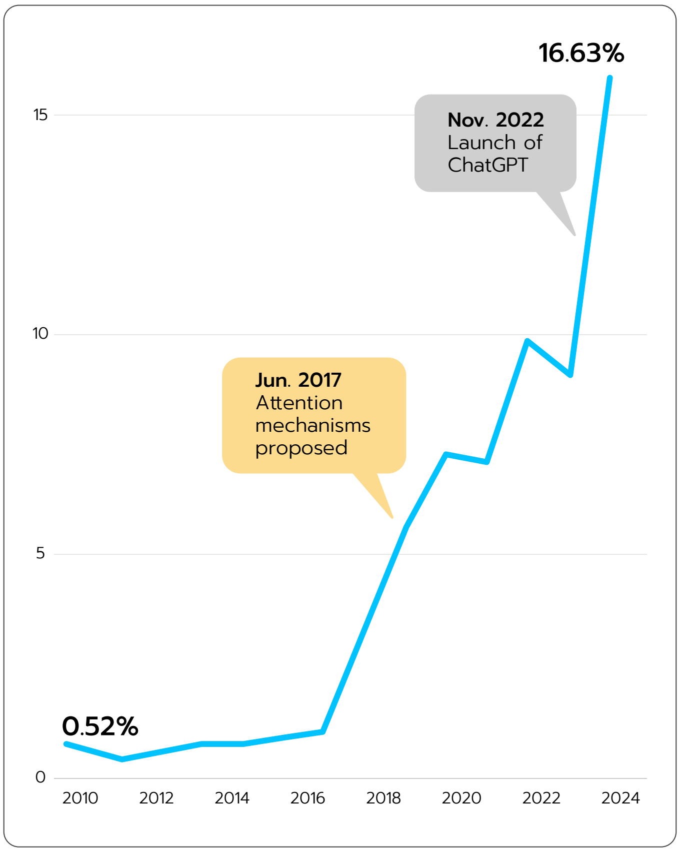
			Marketing and sales	Customer operations	Product and R&D	Software engineering	Supply chain and operations	Risk and legal	Strategy and finance	Corporate IT	Talent and organization
	Total, % of industry revenue	**Total**, $ billion	760–1,200	340–470	230–420	580–1,200	280–530	180–260	120–260	40–50	60–90
Healthcare	1.8–3.2	150–260									
Pharmaceuticals and medical products	2.6–4.5	60–110									

Figure 1. Generative AI use cases across different industries and operations.

What is generative AI and what it’s not

To put it simply, generative AI is a subfield of AI that can create new content. It is based on deep learning models that are pre-trained on large volumes of data in order to come up with new content, be it text, sound, video, or computer code. One of the earliest examples of generative AI can be traced back to the **1960s** when computer scientist and MIT professor Joseph Weizenbaum built the Eliza chatbot. Although basic by today’s standards, Eliza marked the beginning of AI systems that could generate human-like responses.

AI has been considered one of the most popular fields in computer science for decades, a rock star that was previously in the spotlight through computer vision advancements. A new wave of popularity hit when OpenAI, an AI research and deployment company, released its viral chatbot **ChatGPT** in November 2022. Although OpenAI previously offered access to its Large Language Models (LLMs) models like GPT-3, it was ChatGPT that spurred an insatiable curiosity towards the potential of generative AI among the general public (see Fig. 2).



The progress in generative AI had been a mix of slow evolution and sudden leaps, all the way until that specific moment in 2022. Throughout the 20th century, early pioneers like Joseph Weizenbaum, with his Eliza, laid the foundation for the concept of machines generating human-like responses. However, it wasn’t until the advent of deep learning techniques that generative AI truly began to flourish. Deep learning algorithms, especially generative **Adversarial Networks (GANs)** introduced by Ian Goodfellow and his colleagues in 2014, played a pivotal role. GANs consist of two neural networks, a generator and a discriminator, which work together in a competitive process to produce increasingly convincing synthetic data.

Figure 2. The rise of generative AI’s popularity can be partly attributed to the launch of ChatGPT, as a graph by MIT Technology Review shows.

In parallel, the development of foundation models, particularly LLMs, marked a significant milestone. These models, like OpenAI’s **Generative Pre-trained Transformer (GPT) series**, were pre-trained on massive datasets in an unsupervised manner. This pre-training enabled them to learn intricate patterns of language and context. Another aspect of this progress came with tuning and prompt engineering, which make it possible for foundation models to be adept at generating coherent and contextually relevant content.

This story has another crucial aspect: **discriminative AI models**. Knowing the difference between discriminative and generative AI is essential to understanding generative AI's evolution. Discriminative models, previously prevalent in AI, were primarily used for classification tasks. They learned how to discriminate between different classes of data and were proficient in tasks like image recognition or spam email detection. Generative AI diverges from discriminative AI in its ability to create new data points instead of simply recognizing patterns in existing data. While discriminative models are limited to categorizing given data, generative models can synthesize completely novel samples similar to the data they are trained on.

Here is a closer look at various functions generative and discriminative models bring to the table:

Battle of AIs!

This is what we have been doing

Discriminative AI

AI models which discriminate data points to:

- Classify emails as spam
- Predict customer churn
- Segment customers
- Calculate a propensity of a risk

This is where the shift is happening

Generative AI

AI models which generate data points to:

- Generate image from text
- Generate sound from text
- Generate text from sound
- Generate summary of a meeting

The coexistence of generative and discriminative models reflects the intricate nature of AI applications. And, although generative AI has seized the popular imagination, discriminative models remain relevant to this day. The strengths of each approach are complementary. Although generative AI expands possibilities through its ability to create, discriminative models analyze what is already there with nuance and rigor. Together, they form a full toolbox that AI practitioners can draw upon based on the demands of the problem at hand. Both will continue playing crucial and often complementary roles across industries to extract maximal utility from data.

Domains in generative AI

The introduction of multi-modal foundation models marked a paradigm shift in the use of AI. It opened up opportunities for the generation of different types of content, multiplying the applications of generative AI. Below you will find an overview of various domains in generative AI.

Domains in generative AI

The advent of generative AI enables a whole set of capabilities towards the generation of content, images, speech, and videos.

Content generation

Generate content through text and voice inputs; ChatGPT, BLOOM, etc.

Image generation

Generate images through text, images, and/or audio; Stable Diffusion, Dalle-e, etc.

Audio/video generation

Generate audio and/or video through text, images, and/or voice; WhisperAI, etc.

Code generation

AI-assisted programming partners to develop code faster; GitHub Copilot, etc.

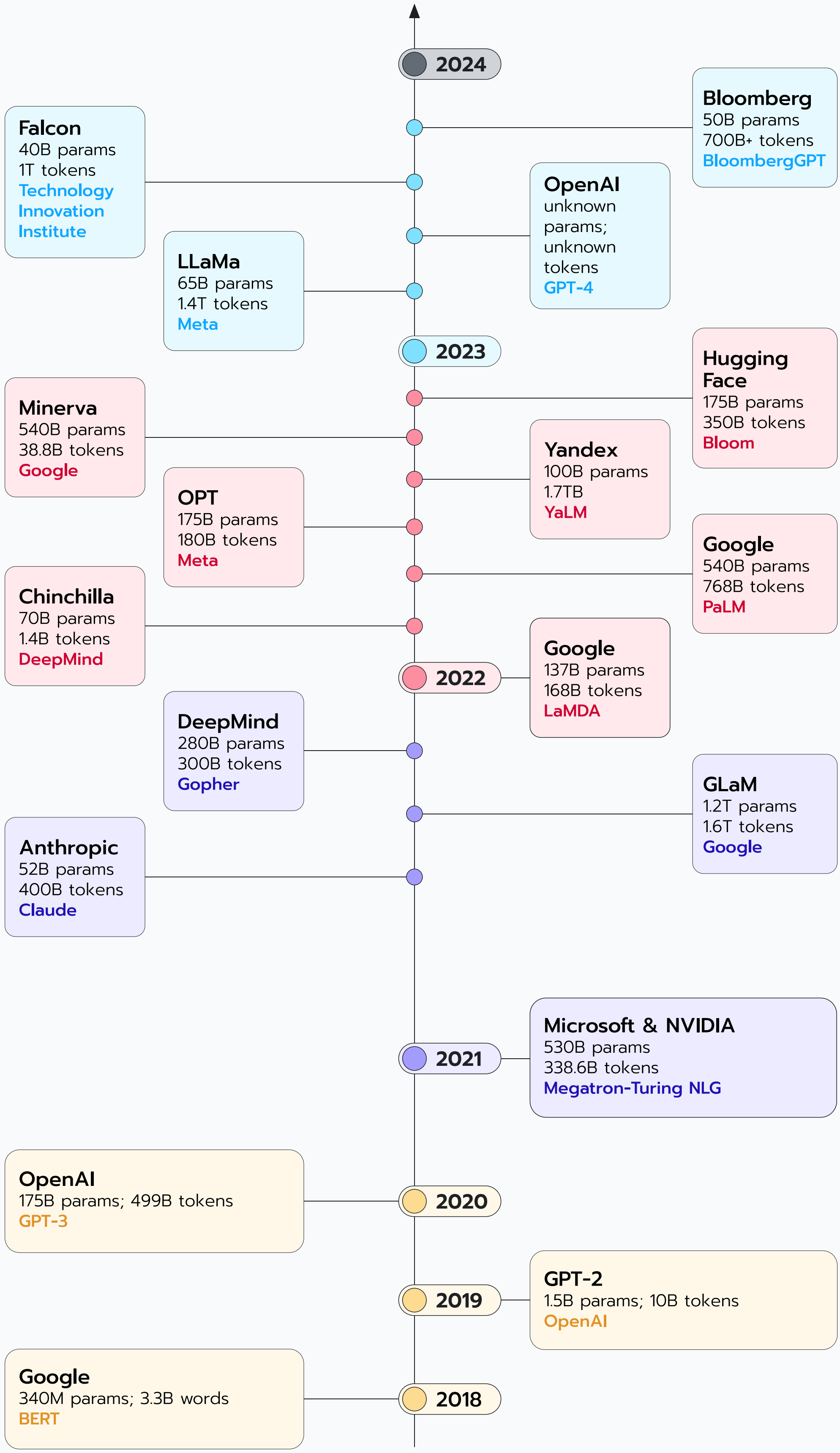
Synthetic data

Synthetic clone for tabular datasets; i.e., Hazy.

Starting from 2018, there was a noticeable increase in the number of released LLMs and large multimodal models worldwide. This proliferation of LLMs can be attributed to several key factors. First, the availability of massive datasets and increased computational power enabled the training of complex neural network models with billions of parameters. And, tech companies like Google, Microsoft, or Meta invested significant resources into developing LLMs. Second, the innovative transformer architecture allowed models to learn contextual representations of language. It turned out to be dramatically better at language tasks compared to previous statistical NLP techniques. Finally, LLMs have proved incredibly versatile, for example, they are useful for summarization, translation, question answering, and creative applications like story generation.

Maestros of Linguistic Symphony

Selected LLMs



2024

Bloomberg
50B params
700B+ tokens
BloombergGPT

Falcon
40B params
1T tokens
Technology Innovation Institute

OpenAI
unknown params; unknown tokens
GPT-4

LLaMa
65B params
1.4T tokens
Meta

2023

Hugging Face
175B params
350B tokens
Bloom

Minerva
540B params
38.8B tokens
Google

Yandex
100B params
1.7TB
YaLM

OPT
175B params
180B tokens
Meta

Google
540B params
768B tokens
PaLM

Chinchilla
70B params
1.4B tokens
DeepMind

Google
137B params
168B tokens
LaMDA

2022

DeepMind
280B params
300B tokens
Gopher

GLaM
1.2T params
1.6T tokens
Google

Anthropic
52B params
400B tokens
Claude

Microsoft & NVIDIA
530B params
338.6B tokens
Megatron-Turing NLG

2021

OpenAI
175B params; 499B tokens
GPT-3

2020

GPT-2
1.5B params; 10B tokens
OpenAI

2019

Google
340M params; 3.3B words
BERT

2018

Figure 3. Releases of LLMs from 2018 from 2018 to 2023.

Here is a detailed snapshot of what's possible today with LLMs and which essential generative models are contributing to the landscape:

Content type	Mechanism	Notable models
Text generation	Generates human-like text based on input prompts	GPT-3.5 and GPT-4 by OpenAI PaLM 2 by Google Claude by Anthropic Inflection-1 by Inflection Cohere Command by Cohere LLaMA 2 by Meta Amazon Titan foundation model family by Amazon Jurassic by AI21 Labs Luminous by Aleph Alpha
Image generation	Creates images from textual or visual input	DALL·E 3 by OpenAI Midjourney by Midjourney Stable Diffusion by Stability AI Firefly by Adobe Magic Design by Canva Jasper Art by Jasper AI
Audio generation	Generates synthetic sound, such as speech or music, through text-to-audio and audio-to-text model	AudioCraft by Meta MusicLM and AudioPaLM by Google Whisper by OpenAI VALL-E by Microsoft
Video generation	Makes videos from textual input	Imagen Video and Phenaki by Google Make-A-Video by Meta
Code generation	Writes code snippets or entire programs	Codex by OpenAI Copilot by Microsoft and OpenAI
Synthetic data generation	Constructs artificial data	DataSynthesizer

The foundation models' ability to work with more modalities of data is opening up new vistas of innovation for life sciences. This includes everything from accelerated drug discovery to personalized medicine, automated lab workflows, and advancements in our comprehension of life itself. Across biotechnology, pharmaceuticals, and healthcare, this new wave of AI adoption promises to elevate the work of protecting human life.

One of the most striking aspects of generative AI's deployment lies in its democratization. Traditionally confined to research labs and tech giants, the technology distanced itself from exclusivity, which made advanced AI tools accessible to a broader audience. And, open-source initiatives and community-driven platforms have paved the way for aspiring developers and entrepreneurs to experiment with novel capabilities of this technology.

However, this democratization comes with a caveat. Foundation models rely upon extensive computing costs and require massive datasets to learn from. This is one of the reasons why there are so many technology-heavy lifters in the table above. Small enterprises can rarely find sufficient funding to train their foundation models, in spite of the fact that in the first half of 2023 alone, more than $40 billion in venture capital was directed towards AI firms. This means that we are likely to witness an increased generative AI arms race across major technology companies in the future.

Generative AI in life sciences

A widespread transformation is underway in the field of life sciences. Below, we've highlight just a fraction of the diverse areas within this field where generative AI can bring notable impact:

1 Drug development and discovery

- **Compound generation**. Through the design and discovery of novel chemical compounds with specific properties, generative models can beef up drug discovery. For instance, InSilico Medicine, a biotechnology company based in Hong Kong, utilizes generative AI techniques to expedite drug discovery processes. The company's potential drug candidate for idiopathic pulmonary fibrosis has already entered phase 2 in its clinical trials. It took less than 18 months for the company to achieve this milestone with its Chemistry42 engine.

- **Virtual screening**. AI-driven virtual screening methods can predict how potential drug candidates will interact with biological targets, thereby reducing the number of physical tests required.

- **Drug repurposing**. Generative algorithms can identify existing drugs that are suitable for new therapeutic purposes and save resources for more efficient drug development.

- **Patient recruitment for clinical trial optimization**. Generative AI models can analyze diverse datasets so as to identify new candidates for clinical trials or diversify trial populations.

2 Biomolecular engineering

- **Protein design**. With generative AI, researchers can optimize protein structures for specific functions, facilitating enzyme design, vaccine development, and protein therapeutics.
- **Peptide design**. Generative models can design peptides for targeted drug delivery and cancer therapy.
- **Antibody design**. AI algorithms can predict potential antibody candidates with high specificity and efficacy. Below you will find an example of a de novo antibody design.

Zero-shot generative AI for de novo antibody design

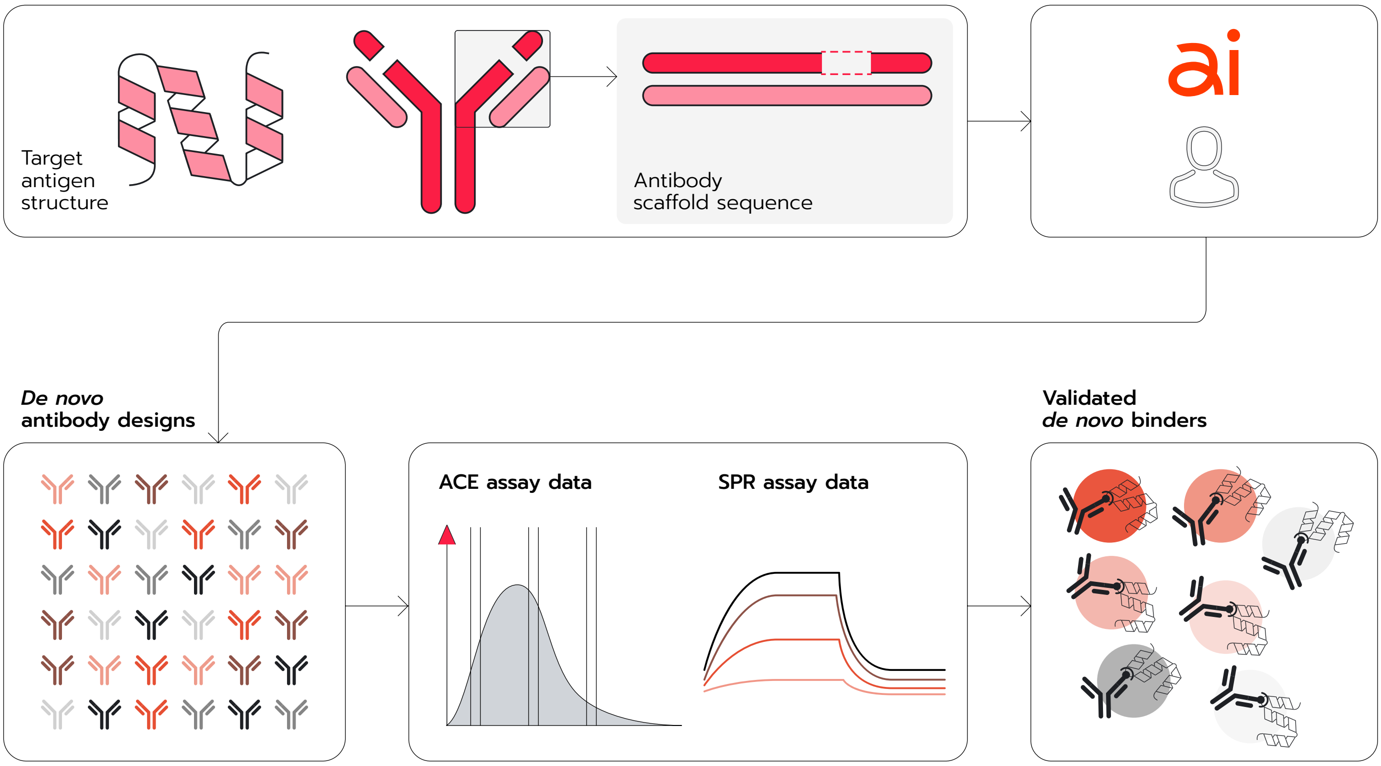

Target antigen structure
Antibody scaffold sequence
ai
De novo antibody designs
ACE assay data
SPR assay data
Validated *de novo* binders

Figure 4. Zero-shot generative AI for de novo antibody design.

3 Disease modeling and epidemiology

- **Disease simulation**. It is possible to use AI-driven models to simulate the progression of diseases and empower researchers to study disease mechanisms and test potential treatments.

- **Disease spread prediction**. Predicting disease outbreaks and planning public health interventions always present a great challenge for scientists. With generative AI, scientists can monitor vast datasets in real-time, and try to strategize and implement timely interventions.

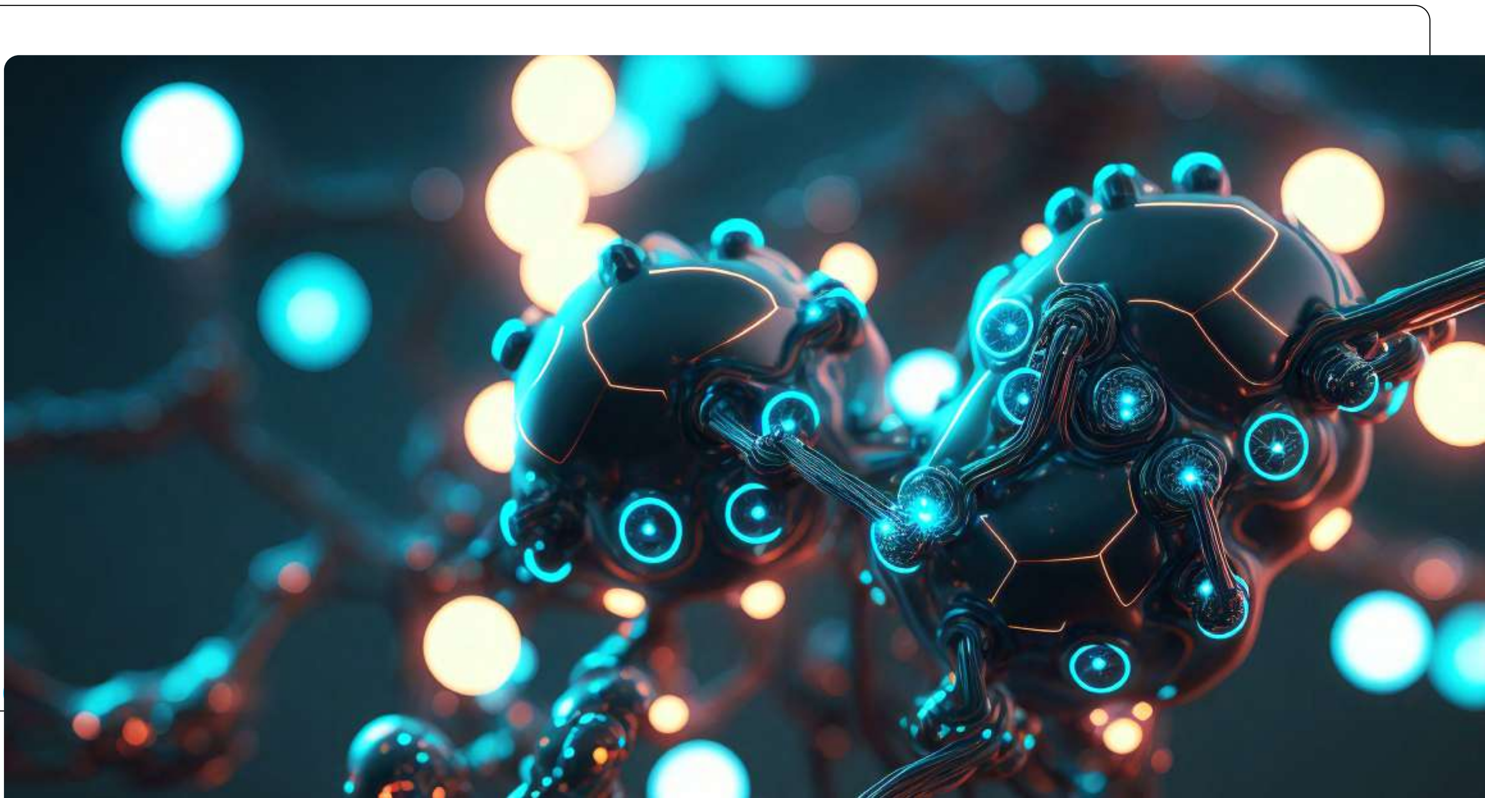
4 Biomedical imaging

- **Image synthesis**. AI-driven models can generate synthetic medical images for training Machine Learning (ML) algorithms and serve as valuable tools for tasks like image segmentation and disease detection.

- **Image enhancement**. With AI, image processing can enhance the resolution and quality of biomedical images which, in turn, improves the decision-making of healthcare professionals.

5 Laboratory automation

- **Experiment design**. Generative AI models are adept at suggesting optimal experimental designs, which can be leveraged for the efficiency of laboratory research.



The field of omics represents yet another pivotal frontier in scientific progress. Below, we explore how the infusion of generative AI can shape the landscape of major omics sub-branches.

1 Genomics

- **Variant analysis**. Generative models can predict the impact of genetic variants on protein structures and functions.
- **Genomic sequence generation**. AI can generate synthetic genomic sequences for research and create diverse datasets without compromising privacy.
- **Personalized genomics**. Generative algorithms can analyze individual genomic data to pinpoint personalized drug targets and disease risk factors.

2 Proteomics

- **Protein structure prediction**. Generative AI enables a more accurate prediction of 3D protein folding and tertiary structures from amino acid sequences.
- **Post-translational modification (PTM) prediction**. AI models can foresee PTM sites on proteins and elucidate their functional roles and regulatory mechanisms.
- **Protein-protein interaction networks**. AI-supported systems can predict protein interactions and contribute to the understanding of complex biological processes and disease pathways.

3 Transcriptomics

- **Gene expression prediction**. Through generative AI, researchers can predict gene expression levels under different conditions allowing them to better understand cellular responses and disease pathways.
- **Single-cell RNA sequencing (scRNA-Seq)**. AI-driven models can analyze and cluster single-cell data, and reveal cell types or transitions in complex tissues.
- **Alternative splicing prediction**. Generative AI models can forecast alternative splicing events, and offer insights into protein diversity and disease-associated isoforms.

4 Epigenomics

- **DNA methylation analysis**. AI algorithms can predict DNA methylation patterns and provide new knowledge of gene regulation and disease epigenetics.

- **Chromatin state prediction**. Generative AI can predict chromatin states, and contribute to the understanding of gene expression regulation and cell differentiation.

- **Histone modification prediction**. AI techniques can predict histone modification patterns and increase awareness of the epigenetic regulation in various biological processes.

5 Metabolomics

- **Metabolite identification**. With AI, scientists can identify unknown metabolites from mass spectrometry data, which is essential for the understanding of metabolic pathways and disease biomarkers.

- **Metabolic flux analysis**. Generative AI models can model and optimize metabolic fluxes, and help with metabolic engineering and bioprocess optimization.

6 Microbiomics

- **Microbial community profiling**. There is a strong need for better analysis of microbiomes. Generative AI can analyze microbiome data and identify microbial species and their functional potential, which is crucial for tracking host-microbiome interactions.

- **Predictive metagenomics**. Through generative AI adoption, researchers can predict the metabolic capabilities of microbial communities and propel biogeochemical cycle studies.

Last but not least, there is another layer of crucial developments in the field of multi-omics. For instance, researchers are investigating the potential of generative AI techniques to create scGPT (single-cell third-generation), a foundation model that could help grasp the complexities of individual cells at the molecular level. Here is schematic model of the scGPT workflow:

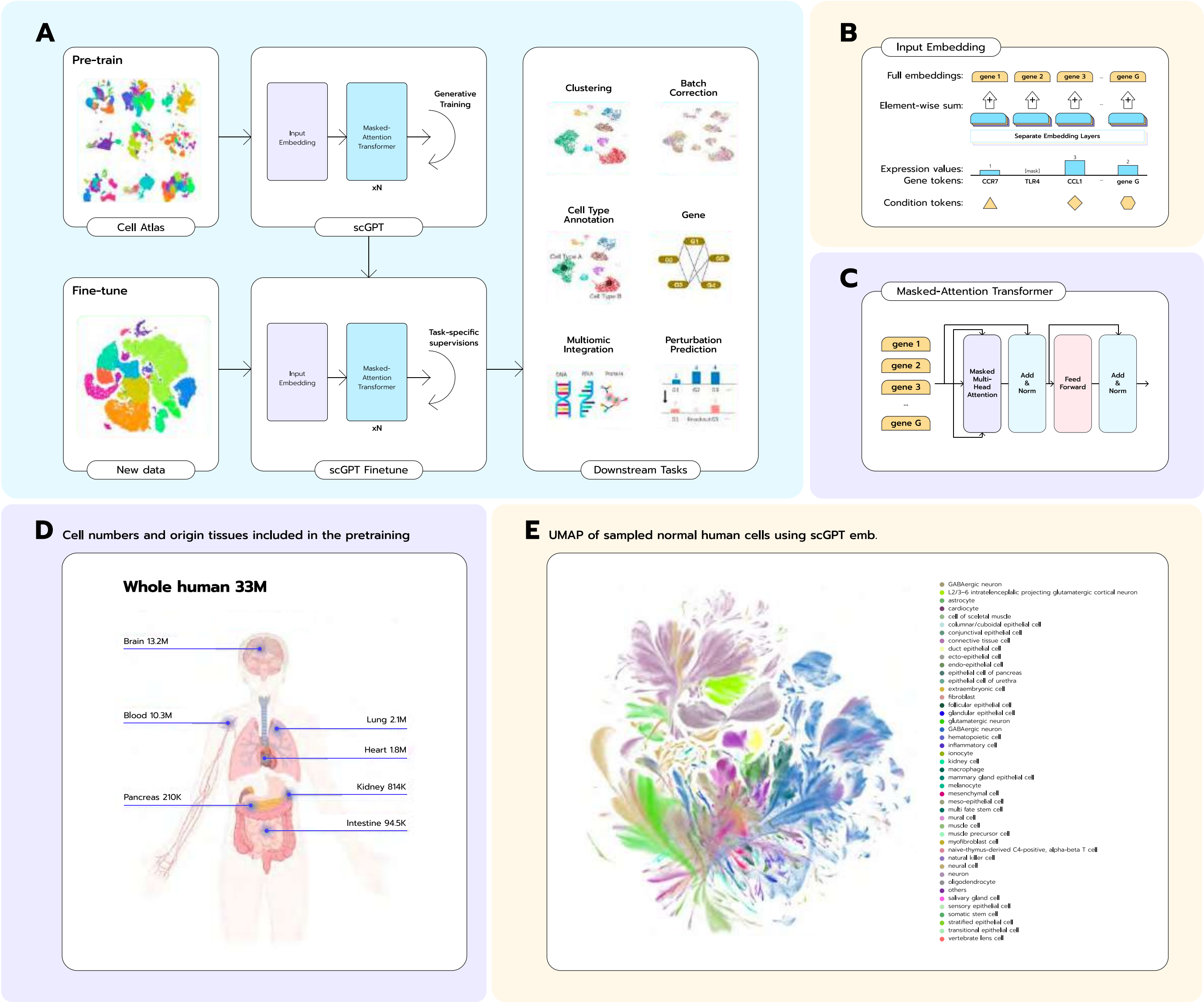
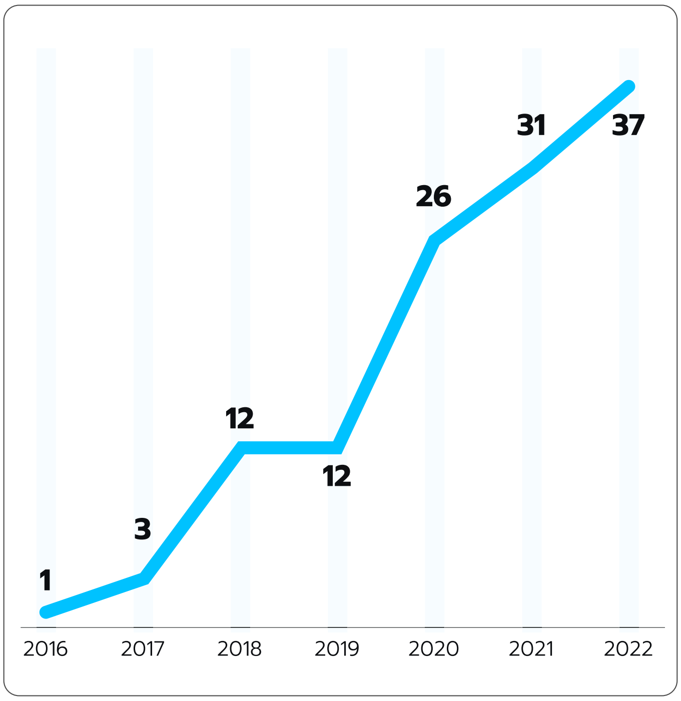


Figure 5. What it looks like to build a foundation model for single-cell multi-omics, as demonstrated in a 2023 research project.

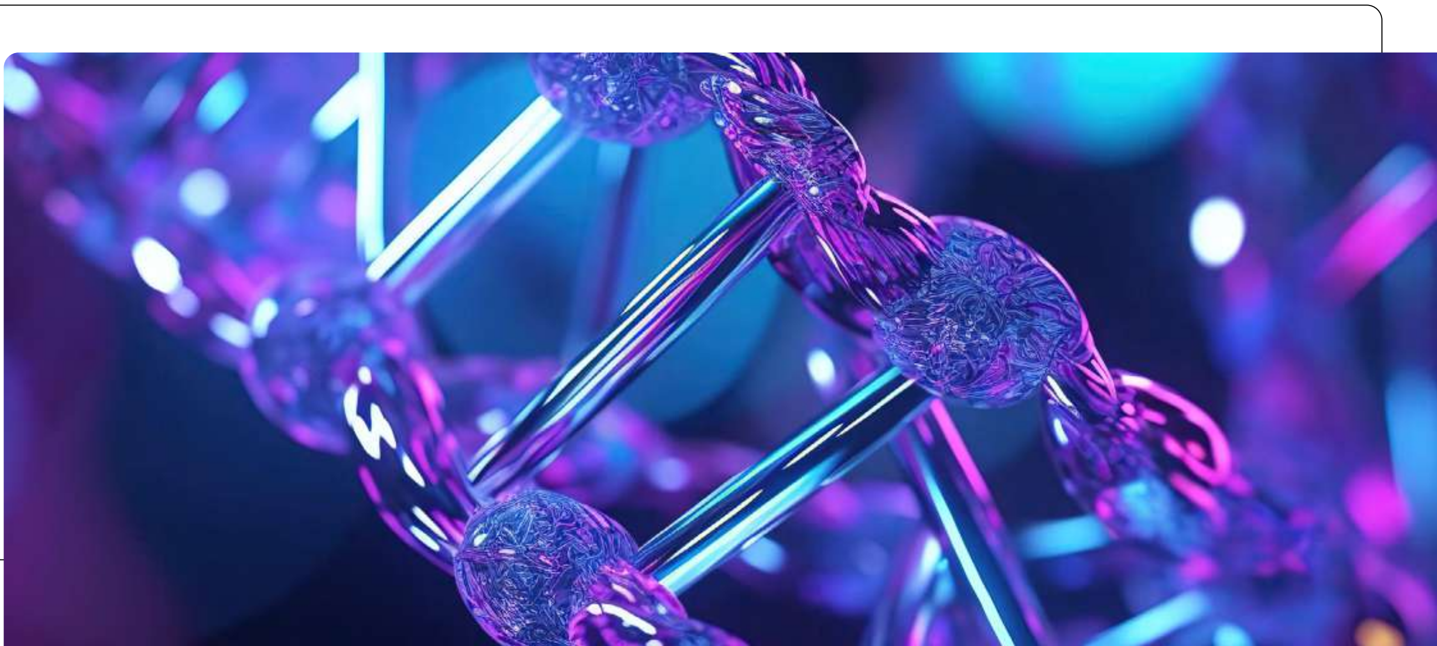
Although the stakes are high, the rapid advancements in generative AI within the life sciences have outpaced the development of comprehensive policies and regulations. And, governments worldwide are moving to implement policies and regulations to both support innovation and mitigate risks (see Fig. 6). In the European Union (EU), lawmakers have proposed the AI Act, which would impose requirements on high-risk AI systems related to data quality, documentation, transparency, human oversight, and accuracy. The EU also formed the High-Level Expert Group on AI to develop ethical guidelines for AI development and use.



The US released its own guidance on regulation and oversight of AI that focused on principles such as public trust, regular evaluation, and the protection of rights and safety. The country has generally taken a light-touch sector-specific approach though policy discussions continue. In July 2023, the administration of US President Joe Biden worked on voluntary commitments with seven leading tech companies that were supposed to balance the development of generative AI. The country also seeks to prioritize AI research funding and export controls.

Figure 6. The number of laws regarding AI, which have been passed globally, has risen greatly from 2016, according to a reportby MIT Technology Review.

Ongoing debates persist around striking the right balance between supporting innovation and managing risks across different political systems and cultural values. International alignment and cooperation on AI ethics and governance will become increasingly important as these exponential technologies grow more powerful and widespread. The opportunities feel boundless, but responsible development and deployment are undeniably vital.



Avenga generative AI use cases

1

AI and robotics for IVF challenges

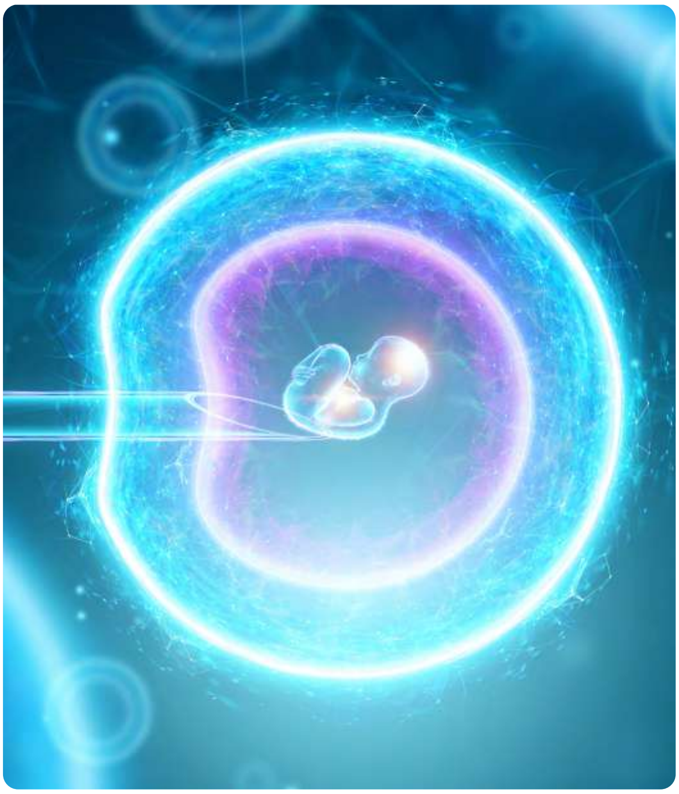
Making the impossible possible

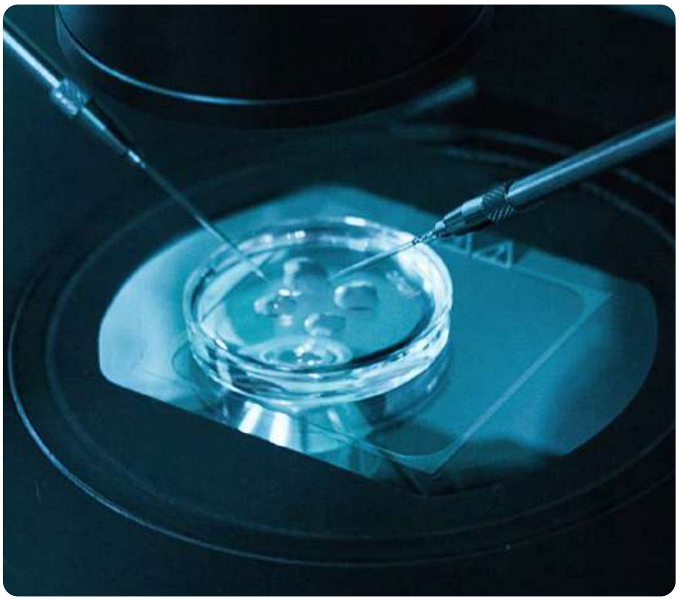
In vitro fertilization (IVF) enables pregnancies once thought impossible, but IVF companies should offer excellence and state-of-the-art technologies for patients' guidance. Efficiency at the highest level is required at all times. Our partner approached Avenga with an idea to create a full-fledged solution that would streamline multiple processes related to assisted reproduction, and through this, beef up tedious procedures and cater a more supportive environment.

Transforming IVF through integrated AI and robotics

Our collaboration involved numerous complex undertakings. Initially, we designed a robust framework for the future cell segmentation application. The team thoroughly researched optimal architectures, loss functions, data augmentation strategies, and hyperparameter tuning methods. This critical planning phase set the stage for efficient development down the road.

Then, our team integrated ML algorithms that covered, including but not limited to, object detection, tracking, and classification. Avenga's team also combined AI and robotics technologies to streamline our partner's workflows and optimize a range of IVF procedures. For these tasks, we leveraged expertise in Qt, Python, OpenCV, TensorRT, PyTorch, YOLO, and UNet.





Ingenuity meets technology

At the end of the collaboration, specialists at the company could take advantage of an advanced IVF microscope that enabled embryologists to analyze 20 times more cells in comparison to its traditional version. Through the combination of AI and robotics, our experts could automate sperm selection and extraction, as well as intracytoplasmic sperm injection. These advancements were supported by a robust infrastructure that can be adapted for various purposes, including fertilization, incubation, or biopsy.

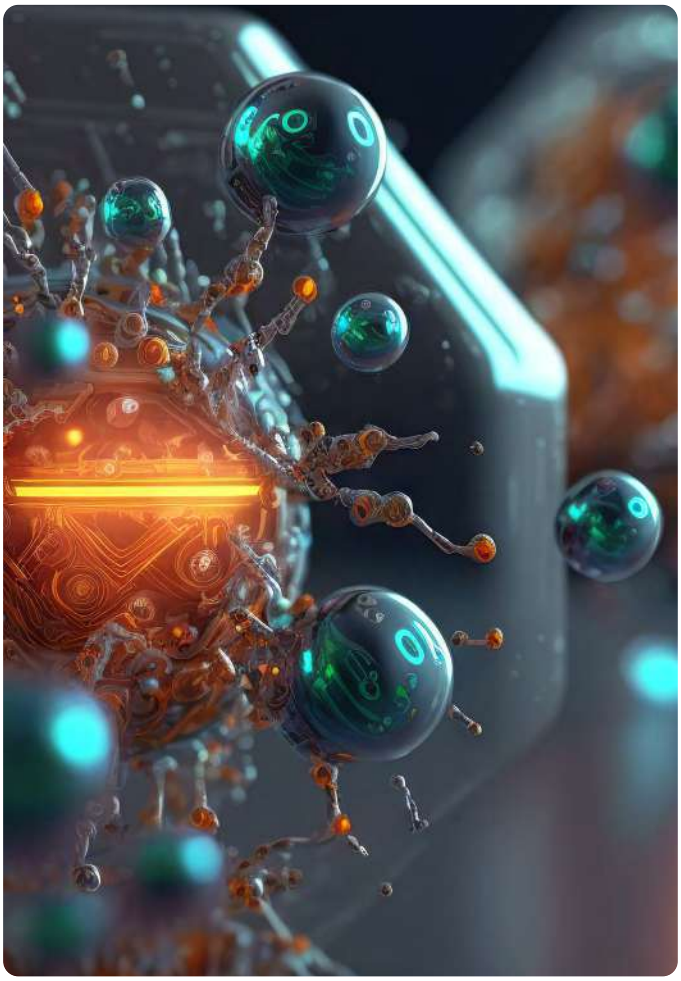
Cell segmentation for clinical trials

2

Addressing the complexities of accurate cell segmentation

Cell segmentation uses image processing techniques to automatically delineate and measure individual cells in microscope images of cell populations. This allows for studying morphology changes, cell counting, or biomarker co-localization.

Proper segmentation is crucial for accurate downstream quantitative biometrics. But, the process is inherently complex. The variability in cell morphology, fuzzy boundaries, low signal-to-noise ratios, or a large set of interacting parameters make reliable segmentation challenging. The complexities of cell microscopy images require nuanced combinations of multiple approaches.



Precision bioimaging analysis via tailored Computer Vision (CV) techniques

In this R&D scientific consulting project, our team worked with microscope cell population images as input. Specifically, cell segmentation served as a basic tool to measure damaged DNA using a technique that generates a 'comet assay.' In this case, cells are lysed to release their DNA, and electrophoresis is used to pull broken DNA strands out of the nucleus. The intact nucleus then looks like the 'head' of a comet, while the broken pieces of DNA stream outward, forming the 'tail.'

Our experts classified each pixel within the neural network architecture into three different classes: background, head, and tail. Precisely delineating the nuclear and body regions was critical for the analysis of biological features and DNA damage markers. Our team used Deep Learning techniques, computer vision (CV), segmentation and image recognition, TensorFlow, OpenCV, U-Net, and ResNet.

Enhancing clinical trials with precise segmentation

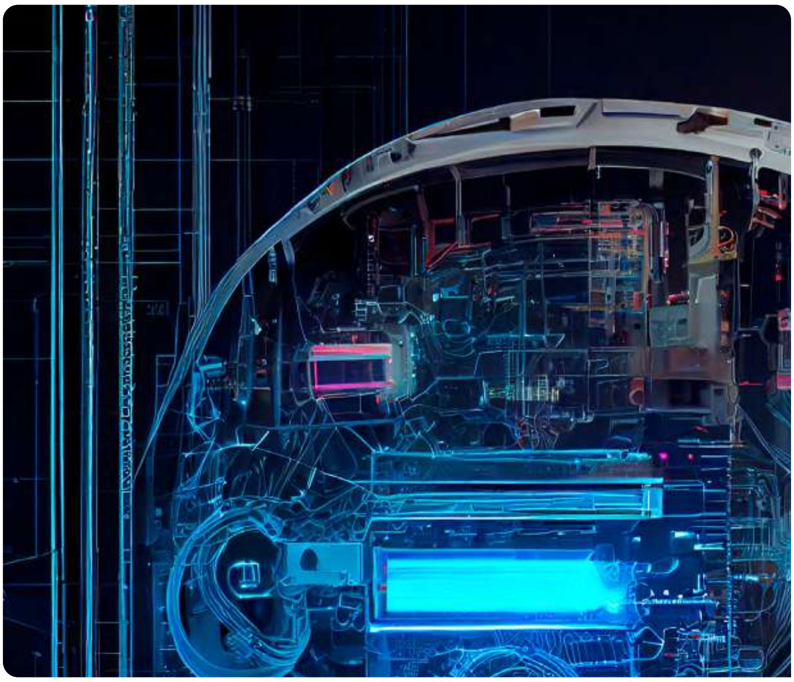
The cell segmentation solution that accurately delineated the head and tail of comets in these assay images provided an essential tool for high-throughput and automatic measurement of DNA damage. This kind of bioinformatics analysis had essential applications in clinical trials.

Real-time claim resolution with Machine Learning

The challenge of automating insurance claims resolution

Our partner was facing a common issue in the insurance industry: the time-consuming and manual process of resolving claims. With a growing number of claims each year, the legacy systems and manual workflows were not scalable. The company needed a way to increase efficiency and automation in claims resolution.

Real-time claims resolution, transfused with automation and ML, brings multiple advantages previously unavailable to insurers and claimants alike. Specifically, it enables insurers to prevent and detect fraud in real time. For claimants, the seamless customer experience increases trust towards the insurance provider.

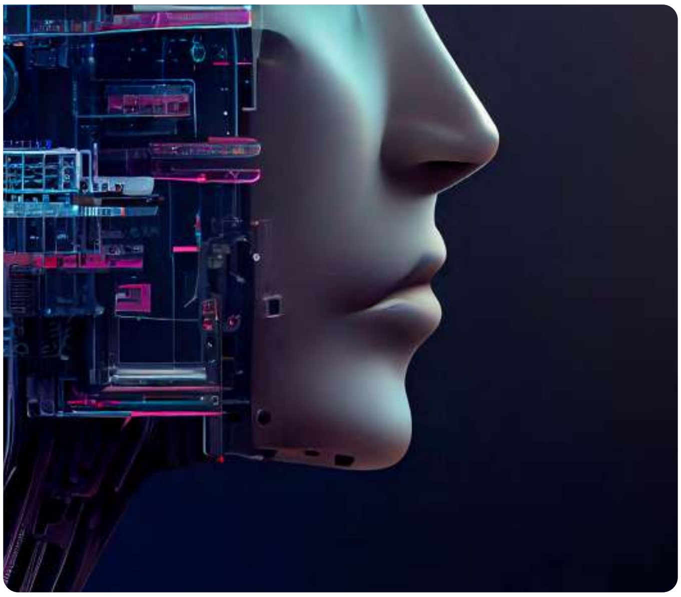


Transfusing insurance claims resolution with speed and velocity

Avenga's team used Python for core development, Azure Blob Storage for scalable cloud data storage, Azure ML Studio for rapid prototyping, robust feature engineering to extract key predictive insights from the data, and XGBoost for its state-of-the-art classification algorithms. The technical architecture we chose allowed rapid delivery of a performant and an adaptable solution tailored to the client's needs.

An intelligent platform driving the next-generation claims experience

Our solution was a fully automated end-to-end ML platform that was custom-built to improve the client's claims resolution process. This end-to-end system automates the entire process: it queries relevant data, processes the data for model input, generates predictions, and delivers results in real time. We used XGBoost's powerful classification capabilities to build a predictive model using historical claims data. One of the notable aspects of this solution is its adaptability and responsiveness to evolving data patterns. This adaptive nature enriches the solution's utility.



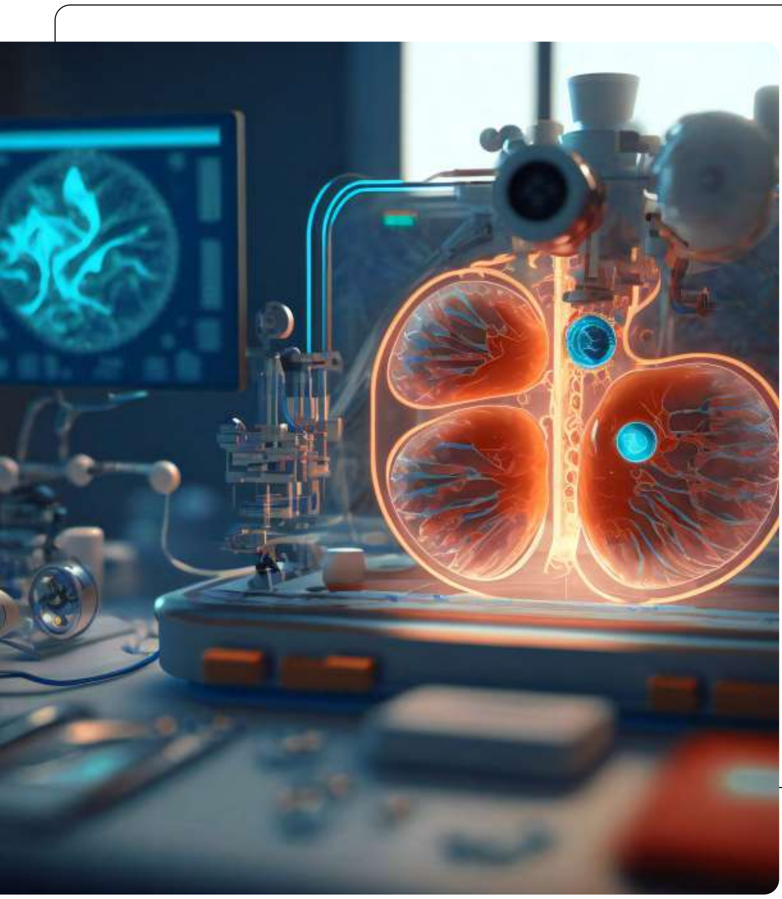
Lung disease prediction

Mapping out a way to make life-saving predictions

Numerous perilous lung diseases have a dramatically improved prognosis when identified at the initial treatable stages before rampant progression. Predictive models that analyze risk factors and slight symptom changes can pinpoint patients in need of preemptive screening and testing. This facilitates diagnosis when intervention can still prevent irreparable damage and preserve lung function. Enabling this kind of early detection through predictive analytics can thus be life-saving, sparing patients from invasive treatments later on.

Building a multi-step system for robust lung assessments

Our team leveraged weakly supervised semantic segmentation to build an AI system for early lung disease screening. Having only image-level labels at hand during the model training, Avenga's specialists segmented the various classes of objects. The proposed approach consisted of three consecutive steps: classification followed by CAM generation, IR-Net for activation map improvement, and segmentation. Each of these steps was followed by post-processing so as to improve the result of the previous one.



Crafting a solution that intercepts progressive lung conditions

We trained a model that could automatically predict the presence of a lung disease. It learned on 12,000 X-ray images. Our experts worked with an imbalanced dataset, with 70% of the dataset representing disease-free cases. In addition, data was distributed across devices by a random stratified split. We leveraged a range of technical tools, including Python, PyTorch, Class Activation Maps (CAM) generation using VGG16 architecture, map enhancement with Inter-pixel Relation Network (IRNet), and segmentation with U-Net. As a result, we gained a cross-silo, horizontal, model-centric, and centralized federated learning set up.

Key opinion leaders (KOLs) identification for optimized pharma research

Selecting KOLs through data-driven approaches

KOLs serve as trusted advisors within the medical field. Their expertise provides guidance to pharmaceutical companies during drug development, clinical trials, or marketing campaigns, while their reputation is critical for the drug's promotion and uptake after regulatory approval. Identifying and engaging the right KOLs is crucial for pharma success. However, manual identification based on credentials and publications is inefficient.

A portion of technicalities

Avenga developed a system built on networks that were based on target locations, patients, and referrals. We also considered a person's influence, leaning on the number of patients and referrals they had associated with them. Then, through clustering algorithms, KOLs were determined based on centrality and a dominant Rx (prescription) in the cluster. Our technology stack encompassed Python, AWS, SQL, NetworkX, PCA, t-SHE, Graphs, and Clustering ML.

How technology serves strategic KOL relationship building

Our team developed and implemented a data science approach in order to explore the diversity of KOLs among healthcare providers. We used a whole set of different data, including prescriptions, claims, and referral data, as well as data enrichment through public sources. This had a positive effect on the quality of partnerships for our client operating in the pharmaceutical industry.

6

Text summarization and entity extraction for better text analysis

Accelerating pharma R&D with text mining technologies

Text summarization and entity extraction offer informative insights for pharmaceutical and life sciences research. Summaries allow researchers to rapidly distill essential information from massive volumes of literature. At the same time, entity extraction can pinpoint relevant biomedical concepts like drug names, protein interactions, and disease terms within unstructured text. Combining these two techniques enables intelligent analysis of publications to highlight connections between extracted entities. For example, automated tools can scan recent papers to assemble summary reports associating discovered genes and pathways with potential therapeutic targets or disease associations.

Reviewing literature with intelligent text analytics

In this project, Avenga's experts worked on two primary tasks: named entity recognition (NER) and literature clustering. NER implied information extraction is used to locate and classify named entity mentions in unstructured text into predefined categories.

Literature clustering relies upon the process of grouping similar textual documents or pieces of literature based on their content. This procedure helps organize vast amounts of textual data into coherent and meaningful clusters, making it easier to analyze and understand the underlying patterns within the text.

From unstructured text to actionable decisions

Avenga's solution effectively processed massive volumes of pharmaceutical literature so as to accelerate drug discovery. The multi-step workflow enabled efficient preprocessing and vectorization of documents, and dimensionality reduction, followed by intelligent clustering and granular topic modeling. As a result, researchers could now use new tools to rapidly pinpoint connections in literature, like associations between genes, protein pathways, therapeutic targets, and disease mechanisms. Automated analysis of these connections facilitated hypothesis generation and informed downstream experiments. The text mining capabilities unlocked hidden insights from publications that transformed manual research workflows.

What Avenga recommends

What Avenga recommends

We are in the early days of generative AI adoption. While the technology is making substantial strides within the life sciences domain, a multitude of both opportunities and challenges have become increasingly evident. So, what are the most promising use cases of generative AI in life sciences today? Let's explore the current landscape in a snapshot.

Generative AI for the pharmaceutical industry

The pharmaceutical industry has historically used AI to speed up a number of processes, but the rise of generative AI caused a renewed curiosity about what's possible in the productivity domain. Although significant attention is allocated to drug discovery and development, which have been time- and cost-consuming processes, other areas can also benefit from generative AI implementation. Perennial challenges in drug manufacturing and supply chain management, as well as commercial operations, could be improved with new AI models.

More efficient drug discovery looming on the horizon

Drug discovery is often like searching for a needle in a haystack. The traditional approach to drug discovery relies heavily upon chance discoveries and a limited understanding of complex biological systems. It can take years for a candidate drug to end up in a healthcare facility, with rounds of clinical trials and extensive trial-and-error along the way. Generative AI, however, offers the potential to systematically explore vast chemical spaces, predict molecular interactions, and rapidly identify promising drug candidates.

A recent breakthrough centers around AlphaFold, a deep learning system that predicts protein structures. Widely adopted in life sciences, it has been utilized by over 1.2 million researchers. Recently, AlphaFold helped researchers identify a critical protein in mosquitoes linked to malaria parasites and revealed the structure of a key protein associated with liver cancer proliferation. These advancements hint at the possibility of accelerated drug discovery in the foreseeable future.

In this situation, generative AI is a transformative force that not only changes the research methodology, but also reshapes the very questions asked in the field. It shifts the narrative from “What can we find?” to “What can we create?” This shift empowers researchers to design molecules with desired properties, accelerates the drug development pipeline, and offers more precise and targeted treatments for various diseases. It’s not surprising that biotech companies are likely to contribute $50 billion a year to AI-based drug development initiatives in the next decade, Morgan Stanley underscores.

Challenges to generative AI adoption

Despite progress, there is still room for generative AI to strengthen its presence in science. For example, chemistry is one of the fields where further advancement is still being awaited. As a recent **Nature** article highlights, thousands and even millions of data points are necessary to galvanize action, but often researchers can only obtain a few thousand data points. The interpretation of structures and properties of molecules heavily relies upon reliable and easily accessible training data, which is scarce.

The issue of data quality and integration implies the complexity of biological systems and the heterogeneity of available data types. Biological systems are intricate and it is difficult to model their behavior accurately. AI models must grapple with this complexity in order to provide meaningful insights. What's more, data in life sciences often originates from diverse sources, each with unique formats and scales. Integrating these disparate datasets cohesively is a substantial challenge.

To make things even more complicated, the availability of high-quality well-annotated data is vital for training robust AI models. In many cases, acquiring such data is a significant obstacle due to cost constraints and limited resources. Addressing these challenges involves developing innovative techniques for efficient data annotation. Successful solutions in these areas are pivotal for harnessing the full potential of generative AI in advancing our understanding of life sciences.

Another major challenge lies in bridging the gap between the expertise of data scientists developing generative AI models and the domain-specific knowledge of life scientists and clinicians. Effective implementation of generative AI in life sciences requires interdisciplinary collaboration. Researchers often have complex and nuanced questions that may not be fully understood or translated into ML tasks. Integrating domain-specific knowledge into AI models, interpreting AI-generated results in biological contexts, and effectively communicating findings between different disciplines are new challenges. Collaborative efforts between data scientists, biologists, clinicians, and software developers are essential for handling these complications.

For these reasons, implementing generative AI in life sciences raises trust and ethical challenges. The generated data and conclusions must be trustworthy and reliable for scientific and medical decision-making. That's why it is crucial to guarantee that the AI-generated results are accurate and unbiased. Transparent methodologies, rigorous validation, and continuous monitoring are vital for trust-building in generative AI applications. Furthermore, addressing concerns related to data privacy, consent, and the responsible use of AI-generated data is paramount to maintaining ethical standards in life sciences' research and applications.

Generative AI for the better understanding of proteins

If, in some cases, generative AI 'hallucinations,' situations where AI systems produce content that is fabricated and does not reflect factual reality, represent a challenge, in others, they embody a creative opportunity to test new hypotheses. As we mentioned before, deep learning systems are adept at predicting protein structures from their sequences. In one of the recent studies, researchers explored the capabilities of deep neural networks designed for protein structure prediction. They questioned whether these networks could generate entirely new folded proteins unrelated to any naturally occurring ones used for training. This groundbreaking work suggests that deep learning networks originally trained to predict native protein structures can be reversed to design entirely new proteins. These methods, alongside traditional physics-based models, hold great promise for the de novo design of proteins.

Generative AI for personalized healthcare

More personalized approaches to healthcare can be supported by the implementation of generative AI. One such area is the collection of additional information from patients and wearable devices. Generative AI algorithms can sift through vast datasets, including real-time health metrics from wearables, patient medical histories, and lifestyle data, so as to create detailed and nuanced profiles of individuals' health. Then, based on the patient's input, they can analyze measurements taken by wearable technology, which can boost the quality of personalized care.

In addition, generative AI can contribute to remote patient monitoring. Patients with chronic conditions or those recovering from surgeries can be continuously monitored through wearables. In their role as data interpreters, generative algorithms analyze this information instantaneously, meticulously comparing it against established baseline patterns. Any anomalies or deviations from the norm are promptly flagged, and special signals trigger timely alerts to inform both patients and healthcare providers.

GitHub Copilot for AI-powered programming in life sciences

In life sciences, where computational methods are paramount to data analysis, simulations, and experimental design, GitHub Copilot embodies a powerful asset. This AI-driven code assistant, a collaborative effort between GitHub and OpenAI, can change the way scientists and researchers in the domain approach coding processes and software development in general. According to a study by GitHub Copilot, 88% out of more than 2000 respondents agree that using Copilot increased their perceived productivity (see Fig. 7).

Copilot has found numerous applications in life sciences. When it comes to bioinformatics, researchers can now rapidly generate complex algorithms for DNA sequence analysis, protein structure prediction, and genomic data processing. Copilot's contextual understanding and ability to discern the code's intent enables it to offer accurate and highly efficient solutions, significantly expediting bioinformatics workflows. Moreover, in computational biology, where sophisticated modeling and simulation are essential, Copilot aids researchers in constructing intricate simulation frameworks and scripting for statistical analysis.

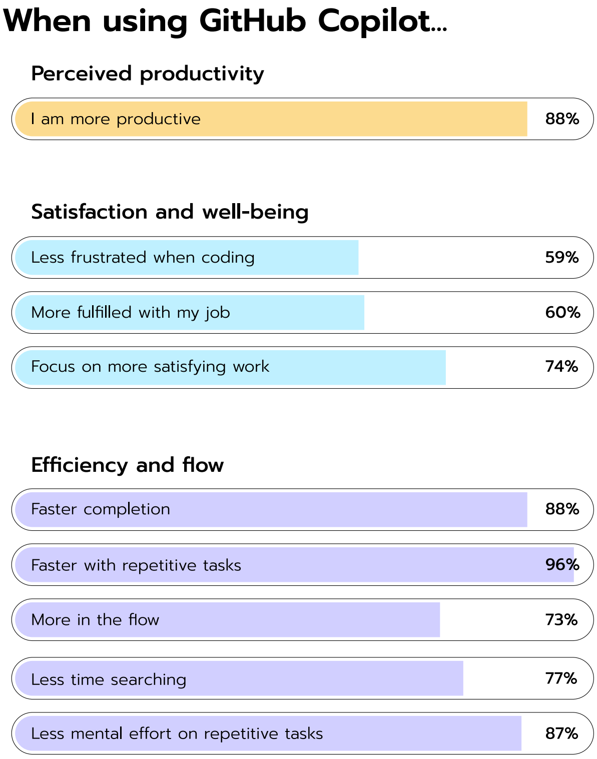


Figure 7. Data shows that GitHub Copilot has an overall positive impact on developers' productivity, well-being, and efficiency.

The impact of GitHub Copilot extends to pharmaceutical research and drug discovery, where computational methods are fundamental. In an industry heavily reliant on molecular modeling, virtual screening, and chemical informatics, it provides a beneficial advantage. It facilitates the generation of code for different purposes, including molecular dynamics simulations or large-scale data mining. Copilot also helps establish interdisciplinary collaboration within life science projects. Biologists, data scientists, and software developers can collaborate seamlessly, leveraging Copilot's contextual intelligence to bridge the gap between domain-specific knowledge and technical implementation.

Final words

What began decades ago with pioneering work on expert systems and bioinformatics has led to versatile generative models like AlphaFold. Today, generative AI appears to have the ability to overcome long-standing challenges from the lab bench to the bedside. Its future applications feel boundless – from illuminating the intricacies of life to democratizing access to personalized therapeutics globally. Yet, realizing this potential requires negotiating complex technical and ethical terrain. If handled properly, generative AI could catalyze a new stage of discovery that explores life's mysteries and expands the frontiers of human knowledge.

avenga