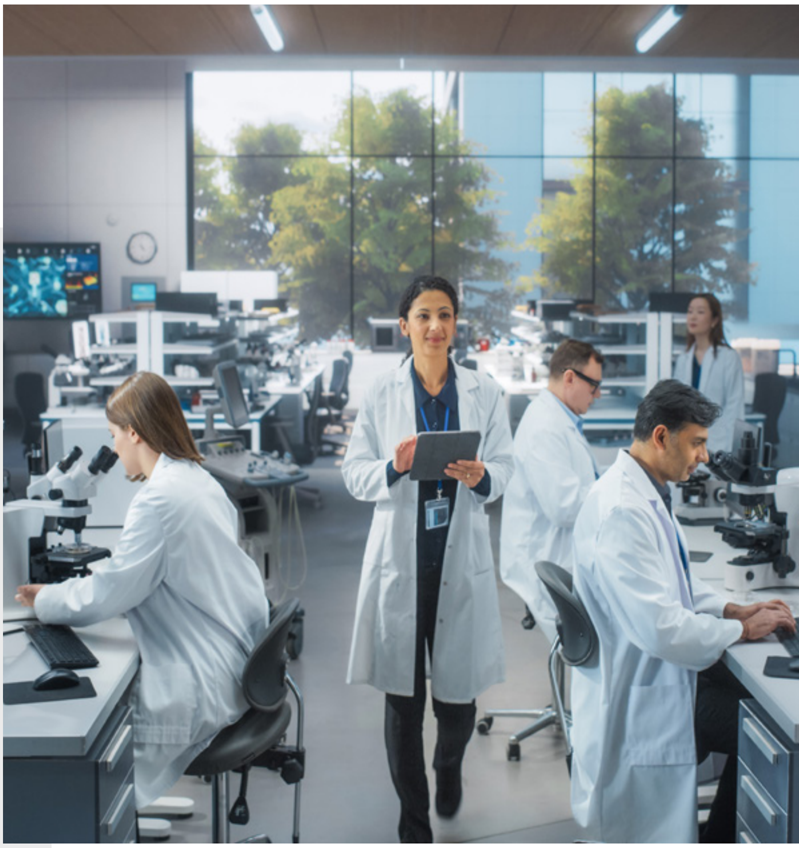
Generative AI for R&D discovery

transformation of the research journey

Capgemini

Generative AI for R&D discovery - *transformation of the research journey*



In 1902, four years after discovering radium, Marie Skłodowska-Curie and her husband, Pierre, purified several tons of pitchblende ore with 400 tons of washing water to produce a mere 0.1 grams of the element. In the process of their ground-breaking research into radiation, they damaged their health and risked their lives.[1] The modern research process is light years from the physical experience that the Curies endured – it is now in its latest transformative phase.

Formulation-based industries combine materials to create complex products from molecules with unique characteristics. Life sciences, consumer products, transportation, and synthetic chemicals are all industries that use these processes, which are applied in typical research workflows. Their workflow structures are broadly similar in principle but are executed with specific adaptations to each industry's technical domain.

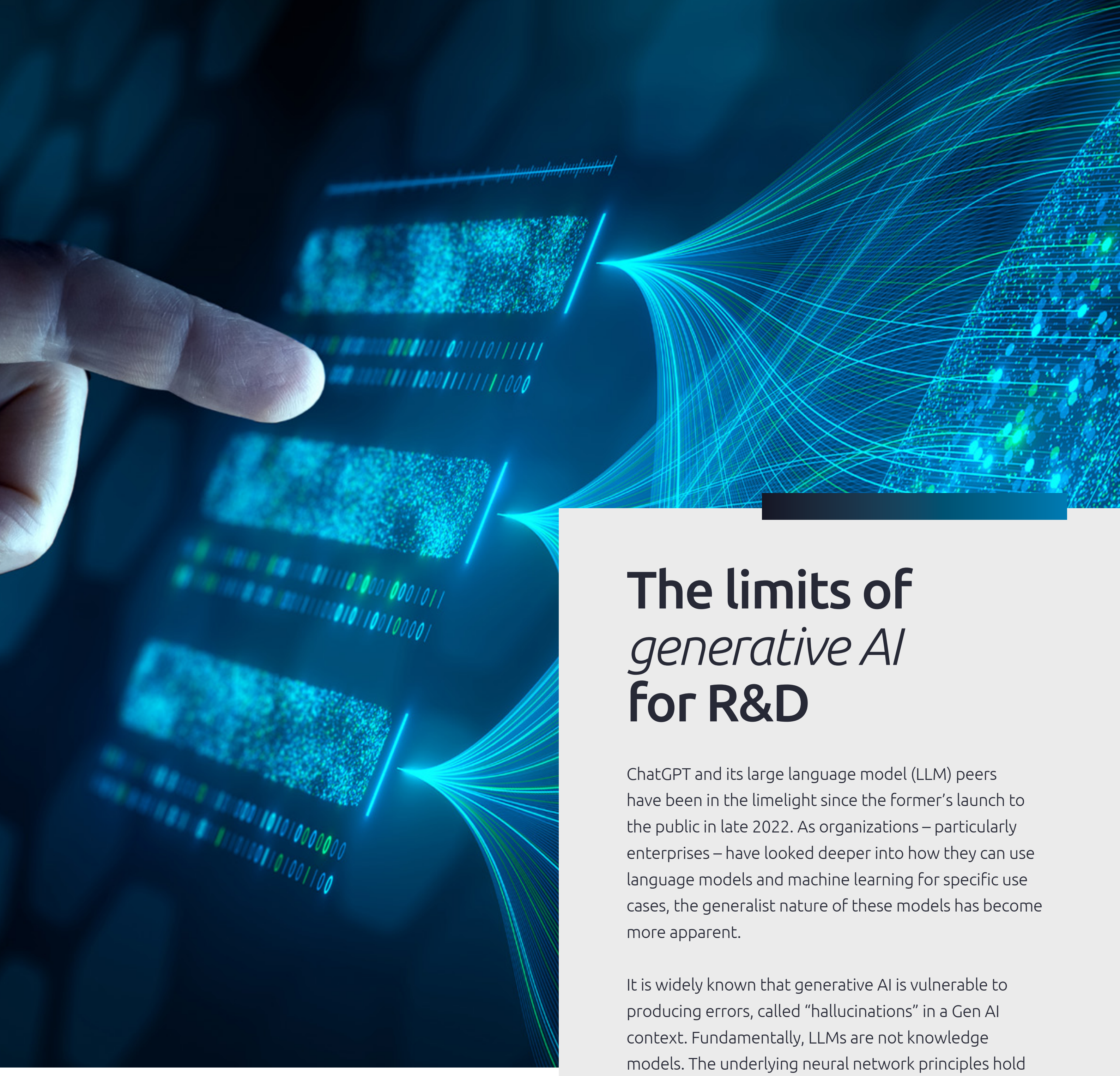
Augmented scientific assistance (ASA) using generative AI in these industries' research and development has already increased efficiency, product usability, and cost effectiveness. It has also opened new possibilities for innovation.

In the pharmaceutical industry, Insilico Medicine used AI in its search for a treatment for idiopathic pulmonary fibrosis, a lung disease. Their scientists used ASA to generate an initial group of 30,000 novel small molecules, which was refined to the six most promising candidates until singling out the strongest for clinical trials. Choosing candidates with strong supporting data to maximize success rates is critical when 90 percent of drugs that go into clinical trials on volunteers fail.[2]

Other industries where Gen AI has already proven its value are:

- **Consumer products** – component substitution in food, beverages, and cosmetics
- **Aerospace and defense** – aircraft fuel composition
- **Transportation** – new tire/rubber properties discovery
- **Energy** – battery design
- **Chemicals** – new fertilizers in agrochemicals.

[1] Mukherjee, S., The Emperor of all Maladies, p. 74
[2] Guardian, *Speedier drug trials and better films: how AI is transforming businesses,* January 17, 2025



The limits of *generative AI* for R&D

ChatGPT and its large language model (LLM) peers have been in the limelight since the former's launch to the public in late 2022. As organizations – particularly enterprises – have looked deeper into how they can use language models and machine learning for specific use cases, the generalist nature of these models has become more apparent.

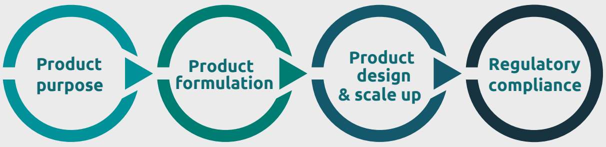
It is widely known that generative AI is vulnerable to producing errors, called "hallucinations" in a Gen AI context. Fundamentally, LLMs are not knowledge models. The underlying neural network principles hold up, but a process of honing, what is known as "fine-tuning" in language model training, is necessary to significantly reduce the error rate.

LLMs work without a structured scientific methodology, i.e., they may not carry out structured scientific tasks or act according to the specific scientific context of an industry's research rules. To address these vulnerabilities, a hybrid form of AI combines an LLM with other AI models and related scientific tools to produce an optimized tool.

The conventional *research process*

Formulation-based industries rely on R&D teams following specific scientific workflows to generate innovation. The typical research flow involves exploring possibilities, making decisions, and coming to conclusions across various scientific disciplines: biology, chemistry, and materials science.

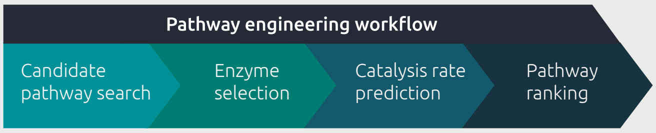
A typical process in synthetic biology research, for example, follows these steps:[3]



1. **Establishing product purpose**
   a. Economic assessment of bio-based material or ingredient
   b. Technical feasibility assessment, with host strain selection
2. **Deciding on product formulation**
   a. Pathway engineering
   b. Iterative optimization through design-build-learn cycles
3. **Setting product design and scaling process**
   a. Fermentation[4] and scaling up
   b. Recovery and purification
   c. Analytical testing
4. **Achieving regulatory compliance**
   a. Documentation and submission for approval.

Pathway **potential**

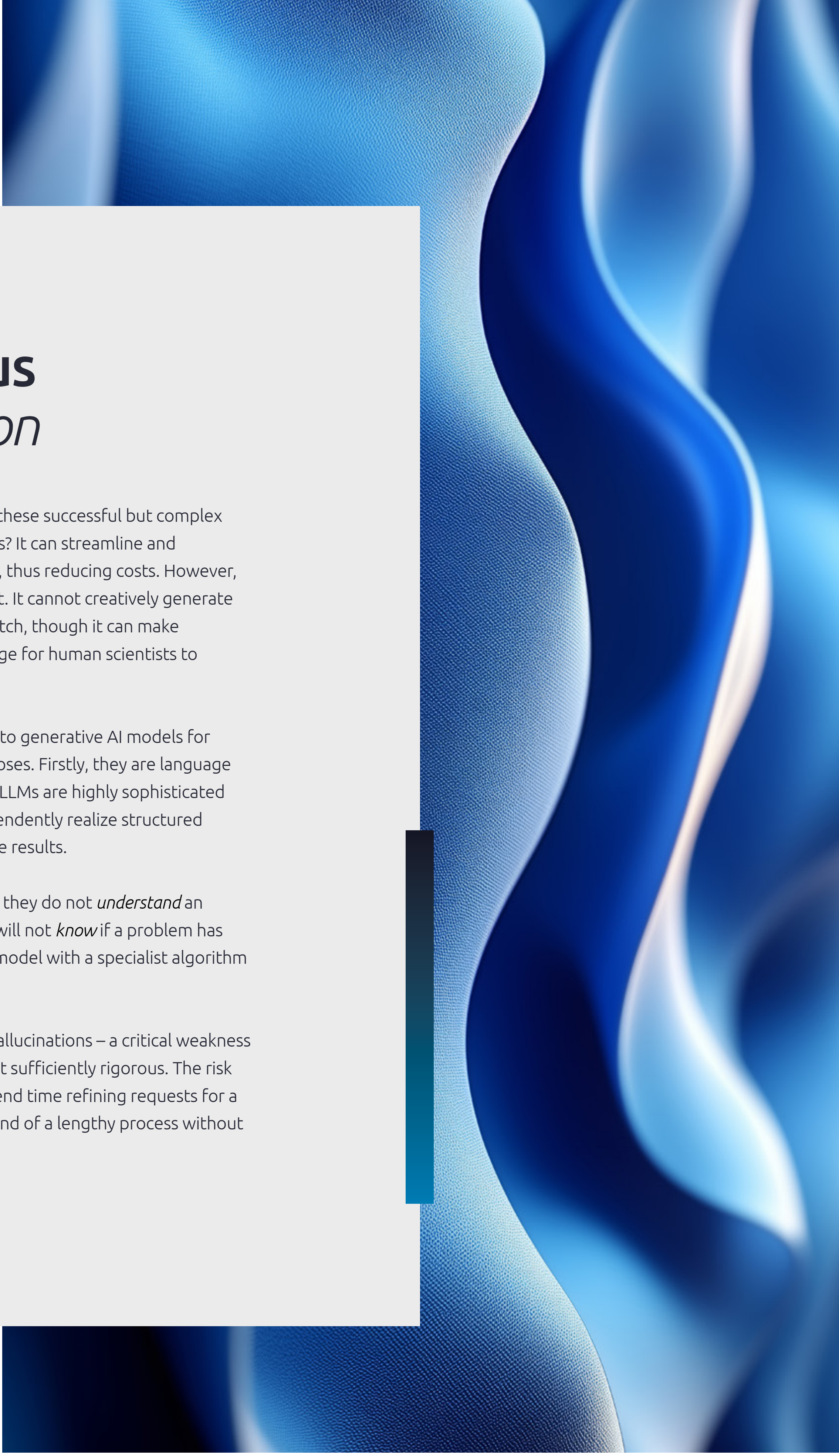
During formulation, the scientific team applies its knowledge and experience to refine existing formulations or discover new ones. The search for the most promising pathway involves enzyme selection, catalysis rate prediction, and a ranking of pathways according to their historical success rates. The process involves using a host organism to test its reaction to the new molecule, potentially generating dozens or even hundreds of candidate strains. Each of these must be evaluated for their potential growth and product profiles.



The research team follows a workflow that involves literature review, component selection, experimentation, and data analysis. This process has built a successful track record over decades. However, although highly sophisticated in the knowledge, experience, and technology required, the workflow is slow and costly because of the high degree of input needed from specialist research teams. It relies on iterative experimentation (meaning repeated) to eventually achieve a desired outcome. In quantifiable terms, that can mean a research project running for a decade and costing over $1 billion.

In a consumer product industry such as food production, a company may decide to re-formulate a product after conducting market research, consumer analysis, product history, and sales analysis, for marketing purposes (ingredients with more appeal), or economics (cheaper alternatives). The workflow is different to that in synthetic biology, but it is also heavily reliant on intensive human input.

[3] Simplified for illustration.
[4] Modern fermentation practices are used to produce "industrially, medically,and nutritionally important compounds." *Antheia, Demystifying Synthetic Biology IV: Fermentation Bioprocess Development,* 2022



Gen AI for harmonious *collaboration*

What can generative AI bring to these successful but complex and laborious scientific processes? It can streamline and accelerate production processes, thus reducing costs. However, it is not a scientist in its own right. It cannot creatively generate new product formulae from scratch, though it can make connections to existing knowledge for human scientists to assess.

There are three main limitations to generative AI models for research and development purposes. Firstly, they are language models, not knowledge models. LLMs are highly sophisticated text networks and cannot independently realize structured scientific tasks and synthesize the results.

Secondly, they are not scientists: they do not *understand* an existing scientific context. They will not *know* if a problem has already been solved using an AI model with a specialist algorithm or a specific routine.

Finally, since they are prone to hallucinations – a critical weakness and risk in this field – they are not sufficiently rigorous. The risk is that a scientific team could spend time refining requests for a Gen AI model only to reach the end of a lengthy process without a productive outcome.

Using generative AI to *the best of* its R&D abilities

Once its limitations are understood, a generative AI toolkit can offer definitive enhancements to R&D processes. The scientific team begins integrating a suitable model by designing it for operation in a specific research context. At the top of the multi-agent system pyramid is a reasoning agent, which can execute a workflow or create a new, optimized version.

In executing the process, an orchestrating agent analyzes the problem, divides it into subtasks, establishes the order of execution, then executes it under human supervision, or autonomously, subject to human oversight. Specialist AI agents run using the output of each other's subtasks. For example, the orchestrator agent can collate and re-combine sub-task results to generate re-formulations. The results are verifiable against norms, standards, and state-of-the-art science, with transparent information provenance for methodologies.

Guiding principles for augmented scientific assistant architecture:

1. Apply knowledge specific to an organization, including its scientific models, AI models, and curated datasets,
2. Customize with specific scientific contextual reasoning,
3. Make results verifiable.

Agents are divided between three pillars:

Gen AI expert research assistant

This is the LLM component of the toolkit that can collaborate with scientists through natural language, while maintaining trust controls over its output.

It integrates existing tools and databases using specific agents to generate data in the form of responses easily understood by its human collaborators. There is no sophisticated reasoning engine within this pillar.

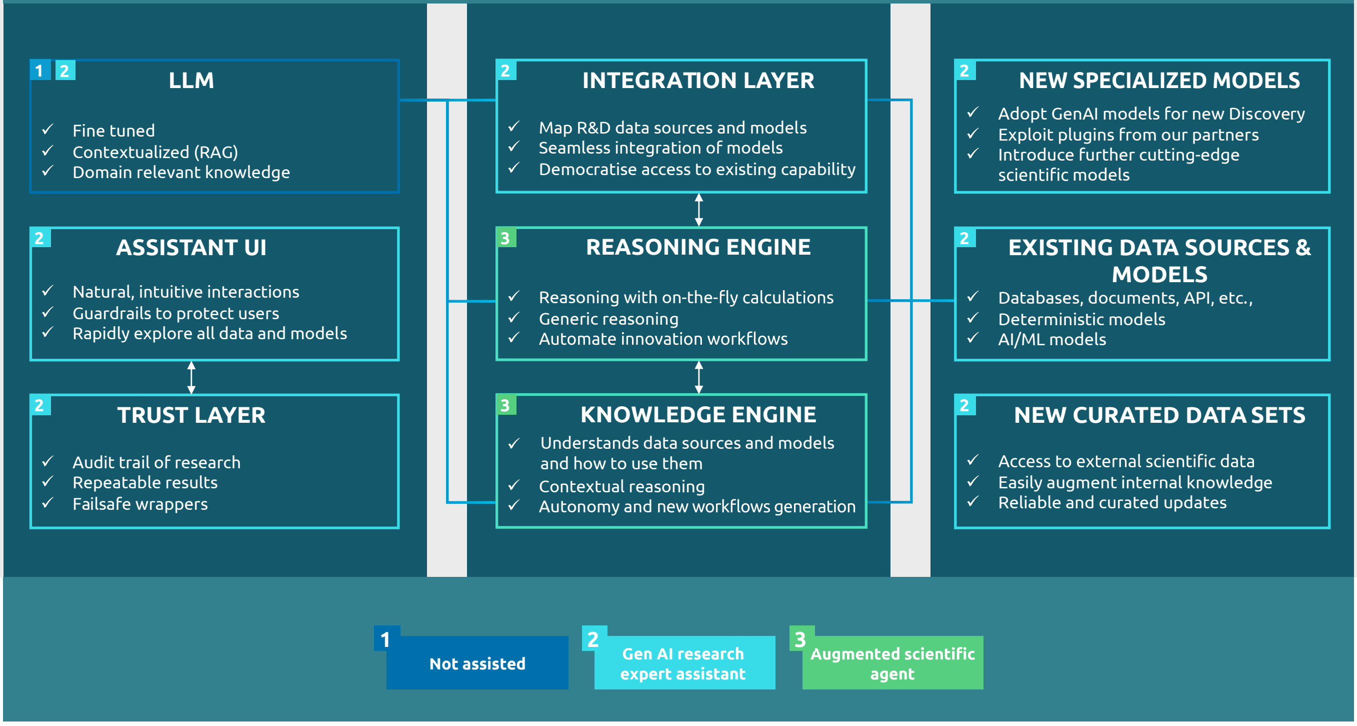
Scientific AI multi-agent system

The second pillar of the toolkit introduces reasoning, orchestration, and a knowledge engine. This pillar can understand complex questions, reason according to a workflow, and employ scientific tools and specialized data as sources for its answers. The agent stores and builds on its interaction with scientists, suggesting new solutions to the problems that they offer with dynamic modifications, or original proposals. In this way, solutions are flexible and adaptive.

Scientific backbone

This pillar integrates existing scientific data sources, curated data sets, and published scientific papers with Gen AI scientific tools and deterministic (i.e., specific purpose) algorithms. All of these sources, including autonomous independent AI agents, can be incorporated in a foundational language model that can answer complex questions. The algorithms are activated in response to queries (or prompts) submitted to the LLM.

Augmented scientific assistance (ASA) architecture



1 2 **LLM**
- ✓ Fine tuned
- ✓ Contextualized (RAG)
- ✓ Domain relevant knowledge

2 **ASSISTANT UI**
- ✓ Natural, intuitive interactions
- ✓ Guardrails to protect users
- ✓ Rapidly explore all data and models

2 **TRUST LAYER**
- ✓ Audit trail of research
- ✓ Repeatable results
- ✓ Failsafe wrappers

2 **INTEGRATION LAYER**
- ✓ Map R&D data sources and models
- ✓ Seamless integration of models
- ✓ Democratise access to existing capability

3 **REASONING ENGINE**
- ✓ Reasoning with on-the-fly calculations
- ✓ Generic reasoning
- ✓ Automate innovation workflows

3 **KNOWLEDGE ENGINE**
- ✓ Understands data sources and models and how to use them
- ✓ Contextual reasoning
- ✓ Autonomy and new workflows generation

2 **NEW SPECIALIZED MODELS**
- ✓ Adopt GenAI models for new Discovery
- ✓ Exploit plugins from our partners
- ✓ Introduce further cutting-edge scientific models

2 **EXISTING DATA SOURCES & MODELS**
- ✓ Databases, documents, API, etc.,
- ✓ Deterministic models
- ✓ AI/ML models

2 **NEW CURATED DATA SETS**
- ✓ Access to external scientific data
- ✓ Easily augment internal knowledge
- ✓ Reliable and curated updates

1 Not assisted
2 Gen AI research expert assistant
3 Augmented scientific agent

Levels of *Gen AI* research

While formulation industries share research practices, each organization has its own requirements for specialization that will determine the level of Gen AI-augmented scientific assistance they can apply to improve existing processes. Generally, these levels can be divided into three categories:

3 **Augmented scientific agent:** This is the most advanced level of Gen AI for research and development. It can automate the execution of complex workflows and recommend relevant research content and solutions. The LLM is combined with explicit models of knowledge to answer complex queries, meaning concepts and categories of data are intentionally connected to improve the agent's reasoning. It also learns from researchers' questions so that the knowledge between humans and AI assistants evolves as with long-term memory.

2 **Gen AI research expert assistant:** The next level of complexity and capability, also known as the "librarian stage". The assistant integrates existing data sources (databases, web services, specific calculation functions, or AI models) or other generative models (specialized for particular tasks), which means it can be used as a reliable tool, developing in experience and sophistication, with minimal risk of hallucinations. It can automate simple and static workflows.

1 **Unassisted:** Adopting an LLM with RAG (retrieval-augmented generation) that has been fine-tuned (trained) on a specific domain and dataset. This can answer, in natural language, relatively simple questions using scientific knowledge.

How can an *augmented scientific assistant* transform research?

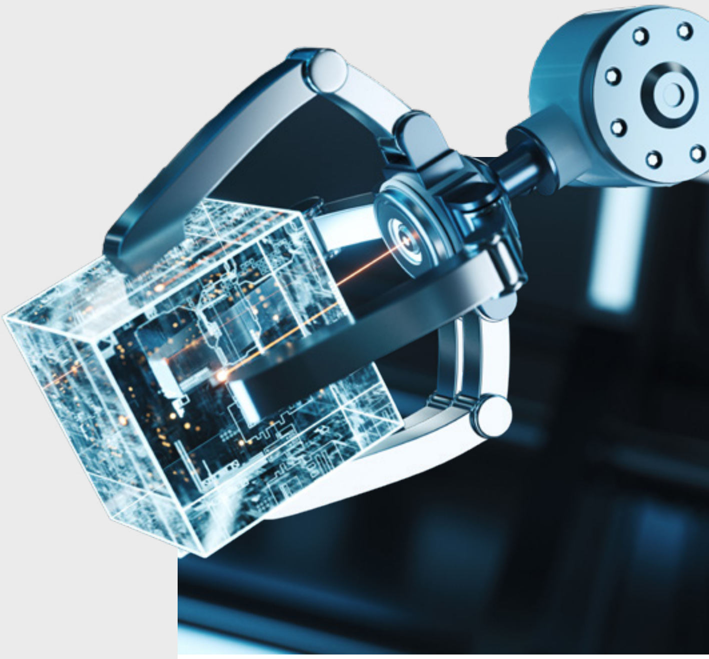
Synthetic biology

At the execution level, a synthetic biology scientist could research genes to produce specific properties in a target molecule. This involves using computational algorithms to design novel metabolic pathways with enzyme combinations to create synthetic products.[5] Pathway construction may involve integrating structural and regulatory genes into the host microorganism.

A generalist LLM will not be up to this level of sophisticated query. With an augmented scientific assistant (ASA), the scientific backbone contributes knowledge from its catalyzed reactions data, generative candidate pathway model, enzyme application programming interface, and populated pathways[6] ranked by a generative model. "Behind the scenes", an orchestrator agent coordinates and automates the entire process by deploying independent agents that dynamically collect relevant data and collaborate to discover the desired product

Food production

In the equivalent process in consumer products, for example, food production, a research team may seek to reformulate a product. This could involve substituting a petroleum-based ingredient with a more sustainable, non-fossil fuel-based alternative to deliver the same product performance and properties with similar shelf-life stability.



[5] Dawn T. Eriksen, Sijin Li and Huimin Zhao, *Pathway Engineering as an Enabling Synthetic Biology Tool*, p.43

[6] Pathways that have been optimized for efficiency and reliability.

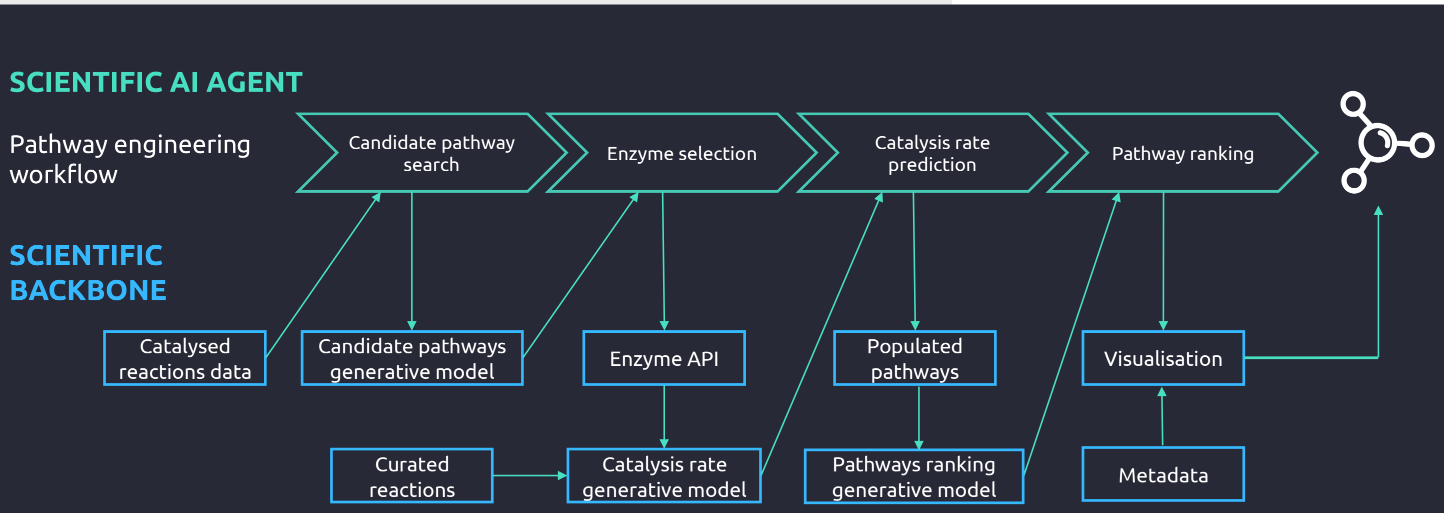


Figure 1 Illustration of an ASA in synthetic biology research

The scientific *payoff*

For the previously mentioned R&D scenarios, using an AI-augmented scientific assistant can significantly improve cost and duration.

For a synthetic biology R&D workflow, incorporating an AI assistant can reduce R&D time:

- 35% reduction in production cost and duration
- Establishing product purpose – from three to one and a half months
- Product formulation – from 12 to 9 months
- Product design and scaling – from 18 to 12 months
- Regulatory compliance – from six to three months.

In a consumer product scenario, significant improvements can be made in:

- Accelerated identification of alternative, sustainable ingredients
- Testing, with fewer physical product iterations and costly experiments
- Waste reduction, by reaching scaled-up production first time
- Automated documentation and regulatory compliance checks.

Augmented scientific research is already making an impact

In scientific endeavors such as cancer research, Gen AI assistance should significantly accelerate drug discovery. Taking a historical example from medical science, in the early stages of chemotherapy research in the 1970s, following the US National Cancer Act (1971), the American National Cancer Institute (NCI) tested hundreds of thousands of chemicals each year in its quest to discover new cytotoxic drugs for use in chemotherapy. It was "trial and error on a giant human scale, with the emphasis... distinctly on error."[7]

The demonstrable gains in speed that generative AI brings to research processes are likely to lead to life-saving medical compounds being discovered faster.

The following are examples of how Gen AI has been incorporated to some degree to improve workflows and represent only the beginning of what could be fully achieved as the technology matures.

- In pharma and biotech, Gen AI R&D assistants (or ASAs) have been key to predicting adverse drug reactions prior to market release with 75% accuracy.
- In the food and beverage industry, Gen AI assistants have cut product launch times by 50% by using better technical, consumer, and market analysis.
- For luxury goods, a perfume producer has used its Gen AI R&D assistant to generate 500,000 fragrance formulae and streamline product development.[8]

[7] Mukherjee, S., *The Emperor of all Maladies,* p. 207
[8] Capgemini client engagements

Capgemini's *vision for Gen AI* in R&D discovery

Generative AI for R&D discovery is a solution to reduce research effort and cost and optimize resource allocation.

Augmented scientific research can inspire entirely new discoveries and ways of working by unlocking in-house experience and cutting-edge academic research. By introducing AI to research processes, scientists can focus their expertise where it is most needed in their search for innovation.

Expert to contact

Fabio Fusco
Head of Hybrid Intelligence Center of Excellence
Capgemini Engineering
fabio.fusco@capgemini.com

About Capgemini

Capgemini is a global business and technology transformation partner, helping organizations to accelerate their dual transition to a digital and sustainable world, while creating tangible impact for enterprises and society. It is a responsible and diverse group of 340,000 team members in more than 50 countries. With its strong over 55-year heritage, Capgemini is trusted by its clients to unlock the value of technology to address the entire breadth of their business needs. It delivers end-to-end services and solutions leveraging strengths from strategy and design to engineering, all fueled by its market leading capabilities in AI, cloud and data, combined with its deep industry expertise and partner ecosystem. The Group reported 2024 global revenues of €22,096 million.

Get the future you want | www.capgemini.com