Can language model agents automate biomedical research?

Topic for: Masterarbeit, Projektstudium, KfK-Praktikum, KfK-Seminar
Contact: Assoc.Prof. Matthias Samwald, matthias.samwald@meduniwien.ac.at, https://samwald.info/

Overview

The rapid evolution of artificial intelligence (AI) systems such as large language model (LLM) agents and future superintelligent systems, holds the potential to revolutionize biomedical research. These advancements promise breakthroughs, such as combating major diseases, enhancing crop resilience, and improving pandemic preparedness. However, they also pose substantial risks, including the creation of novel pathogens and the proliferation of biological weapons. Although there is consensus on the transformative potential of advanced AI systems in biomedical research, the specific mechanisms, opportunities, challenges, and risks remain largely unexplored.

We aim to:

- Develop a **novel conceptual model** of human intelligence processes in biomedical research that are amenable to augmentation or replacement by highly capable, agentive AI systems.
- Operationalize and **measure or estimate relevant AI capabilities** through critical analysis of AI benchmarks, expert elicitation, and analysis of real-world adoption.
- **Identify high-impact areas** at the intersection of biomedical research and highly capable, agentive AI systems that might result in discontinuous progress.
- Identify risks and formulate actionable recommendations to **inform practice and policy making**.

Process BIO+AI	Capability	Uncertainty ?	Velocity	Impact
Hypothesis generation	■■■■	■■	■	■■■
Experiment design	■■■	■■■	■■■	■■■
Experiment execution	■■	■■■	■■	■■■■
└ Fully autonomous robotics	■■	■	■■■	■■■■■
└ Cloud lab	■■	■■	■■■	■■■
Data analysis	■■■	■■	■■■■	■■■
…	…	…	…	…

Fig. 1: The BIO+AI model tracks how LLM agents and future superintelligent systems may augment and replace human intelligence processes and lead to discontinuous acceleration of biomedical research progress.

Methodology

Developing the BIO+AI model

To radically accelerate progress, future AI systems will need to address a wide variety of processes that constitute biomedical research and development. We will build a model of these key processes and annotate them along four dimensions (Fig. 1):

1. **Capability:** The capability level of AI systems for the selected process, operationalized as performance on benchmarks and representative real-world tasks. We will base this measure on the AGI levels recently proposed by (Morris et al. 2023).
2. **Uncertainty:** The uncertainty surrounding current AI capabilities, operationalized as the lack of appropriate benchmarks, insufficient real-world evaluation, or limited access and transparency of frontier models.
3. **Velocity:** The current rate of growth in relevant AI capabilities.
4. **Potential impact:** The potential impact of AI reaching human-level capability for a specific process on the overall speed of biomedical research. We operationalize this by estimating the effect of AI reaching a capability level at the 99th percentile of skilled adults ('capability level 4' in the Morris *et al.* terminology) on time required, compared to the current time required by human researchers.

Four **data sources** form the basis of the BIO+AI model, providing a mix of qualitative and quantitative indicators:

1. An initial comprehensive review of relevant **scientific literature**, followed by continuous literature monitoring. This review will shape the initial list of key processes, a rough initial positioning of each process along the four dimensions, and a deeper insight into opportunities and risks.
2. Analysis of relevant **benchmarks**, their coverage of relevant AI capabilities, their ecological validity and quality, benchmark results achieved by current AI systems, and the ongoing evolution of the benchmarking landscape. Where dedicated biomedical benchmarks are not yet available, we will extrapolate AI performance on relevant proxy tasks outside the biomedical domain (e.g., code generation, general-domain agent benchmarks).
3. **Expert elicitation** and feedback through structured interviews with a diverse group of experts working in AI-automated science and adjacent fields, including leading academic research groups, pharmaceutical R&D units, and novel institutions focused on AI-driven discovery such as Future House. We will ensure the participation of highly qualified experts by offering suitable honoraria. While we will strive to find consensus views where feasible, we expect to also find and document significant disagreements.
4. Data on the **real-world adoption** of general AI systems in biomedical research. This will provide a critical complement to the analysis of benchmark results, which can sometimes give distorted impressions of performance and significance in realistic practical settings.

> Importantly, we will focus on general AI systems that are able to replace human intelligence processes. Narrow AI systems — such as biological design tools or deep-learning based algorithms for structure prediction — are *not* within the scope of this project, except for analyzing the ability of general AI systems to utilize and orchestrate such tools.

The resultant BIO+AI model will help address crucial questions, such as:

- How close are we to achieving truly **general AI** systems capable of tackling all relevant research processes?
- What is the **current uncertainty about AI capabilities**, and how can we improve instruments like benchmarks to better estimate these capabilities?
- Which biomedical processes have the **greatest impact** on overall research progress?
- What are the current **blockers**, and are blockers associated with particular classes of capabilities, e.g., would a burst in autonomous robotics capabilities lead to a significant unblocking across the entire model?
- Which high-impact processes should receive **special scrutiny**, due to high current AI capability, high velocity of capability development, or high uncertainty?