Evaluating scientific papers with AI

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

Transformative AI systems are rapidly approaching near-human performance in several cognitive tasks, with profound implications for scientific research and knowledge creation. Recent studies show a **rapid increase in AI involvement in scientific paper creation and growing interest in leveraging AI to support the evaluation of research,** especially in the paper review process.

A critical challenge emerges: balancing AI systems that generate research with those that evaluate it. This "generate-evaluate dynamic" presents both opportunities and risks. While AI has the potential to accelerate scientific discovery and enhance the efficiency of paper review, it also introduces the risk of an influx of flawed research, potentially undermining the scientific system.

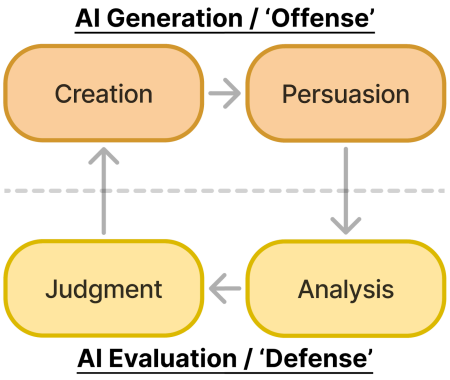

Figure 1: Our basic conceptualization of the "generate-evaluate dynamic" and its components

A recent paper introduced the "AI Scientist", an LLM-based framework for fully automated scientific discovery. While demonstrating the potential of AI in research, it also raised concerns

about potentially credible-looking papers with hidden flaws, as some AI-generated papers contained severe methodological errors that went undetected by the study's automated reviewer.

Our project aims to investigate this dynamic and explore the potential and limitations of AI-driven evaluation in scientific research. We have two primary objectives:

1. Develop a unified framework for understanding AI-driven research evaluation
2. Collect empirical data on the performance of current AI models in generate-evaluate scenarios

We will design and conduct experiments using existing paper review datasets. We will assess the robustness of AI-driven evaluation across different scenarios and propose potential improvements.

Key research questions our project addresses include:

- What are the capabilities and limitations of AI in enhancing or potentially undermining paper review?
- How can we develop effective oversight mechanisms for increasingly complex AI systems in scientific contexts?
- What ethical considerations arise from the use of AI in research creation and evaluation?

Examples of current project ideas:

- Design and implement AI models capable of reviewing research-like content and evaluate their performance against human-written reviewers.
- Develop metrics and benchmarks for assessing the quality and reliability of AI-driven evaluation.
- Investigate the generate-evaluate dynamic in different scientific domains, exploring how it mirrors concepts like the offense-defense balance in cybersecurity.