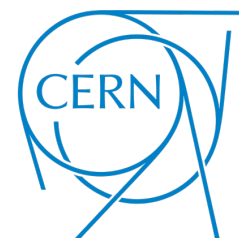# ADVANCED TEXT ANALYSIS FOR KNOWLEDGE DISCOVERY

45TH CERN SCHOOL OF COMPUTING, STUDENT LIGHTNING TALK
17TH SEPTEMBER 2024, HAMBURG

ALEKSANDRA KOWALCZUK

INTERDISCIPLINARY CENTRE FOR MATHEMATICAL AND COMPUTATIONAL MODELLING, UNIVERSITY OF WARSAW

INTRODUCTION

- Master's Thesis

- Core Partnership

- Keywords:

  Novel algorithms, AI applications powered by HPC Technologies, Large Language Models and Generative AI in HPC, ML Systems and Tools, Bioinformatics and Life Sciences, Science of Science

MOTIVATION

- Challenge:

  Because of the constantly growing amount of scientific literature, there is a **real need to make it easier for experts like medical doctors to access resources published worldwide** and quickly draw conclusions from them.
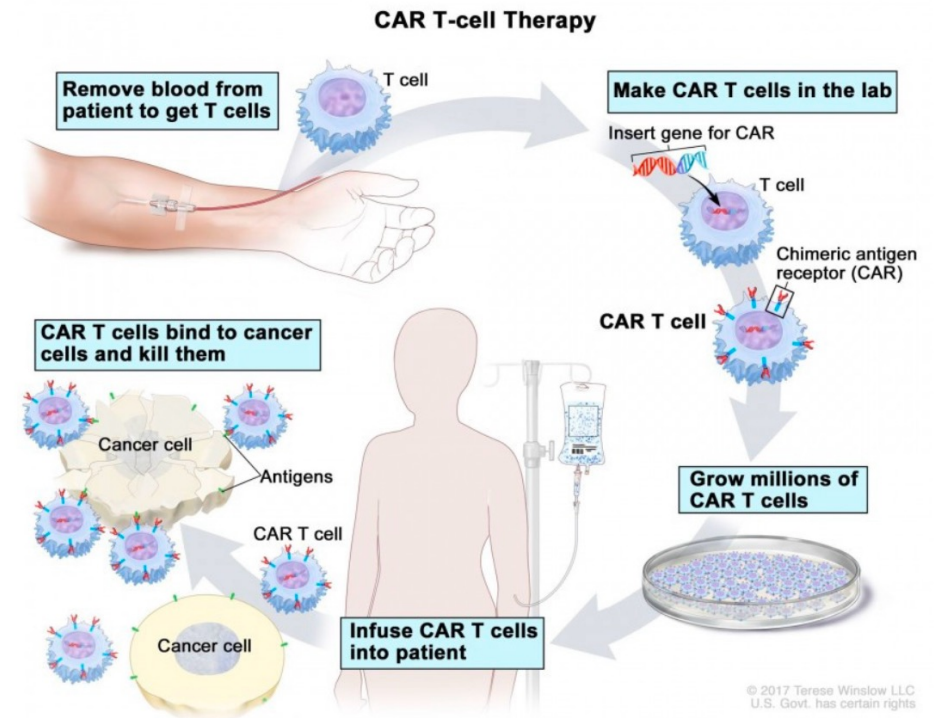
- Where are we now:

  We can observe a dynamic development of large language models and ready-to-use tools dedicated to text analysis, processing, and generation.
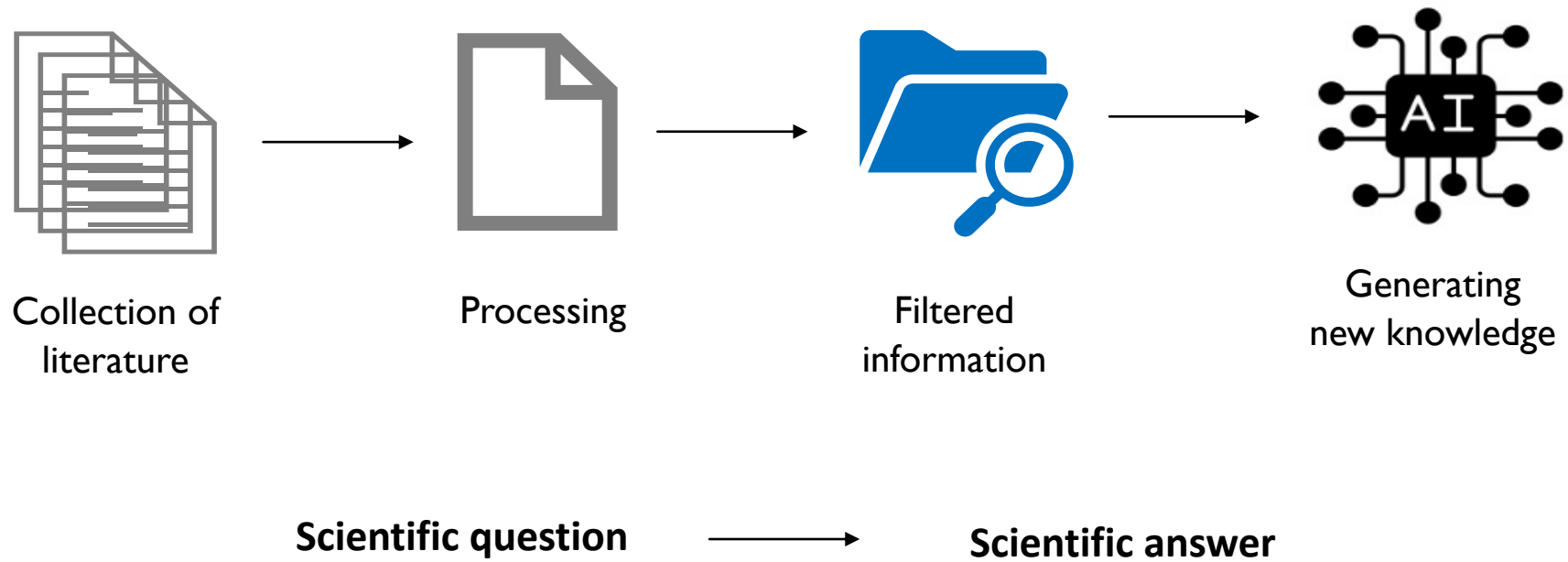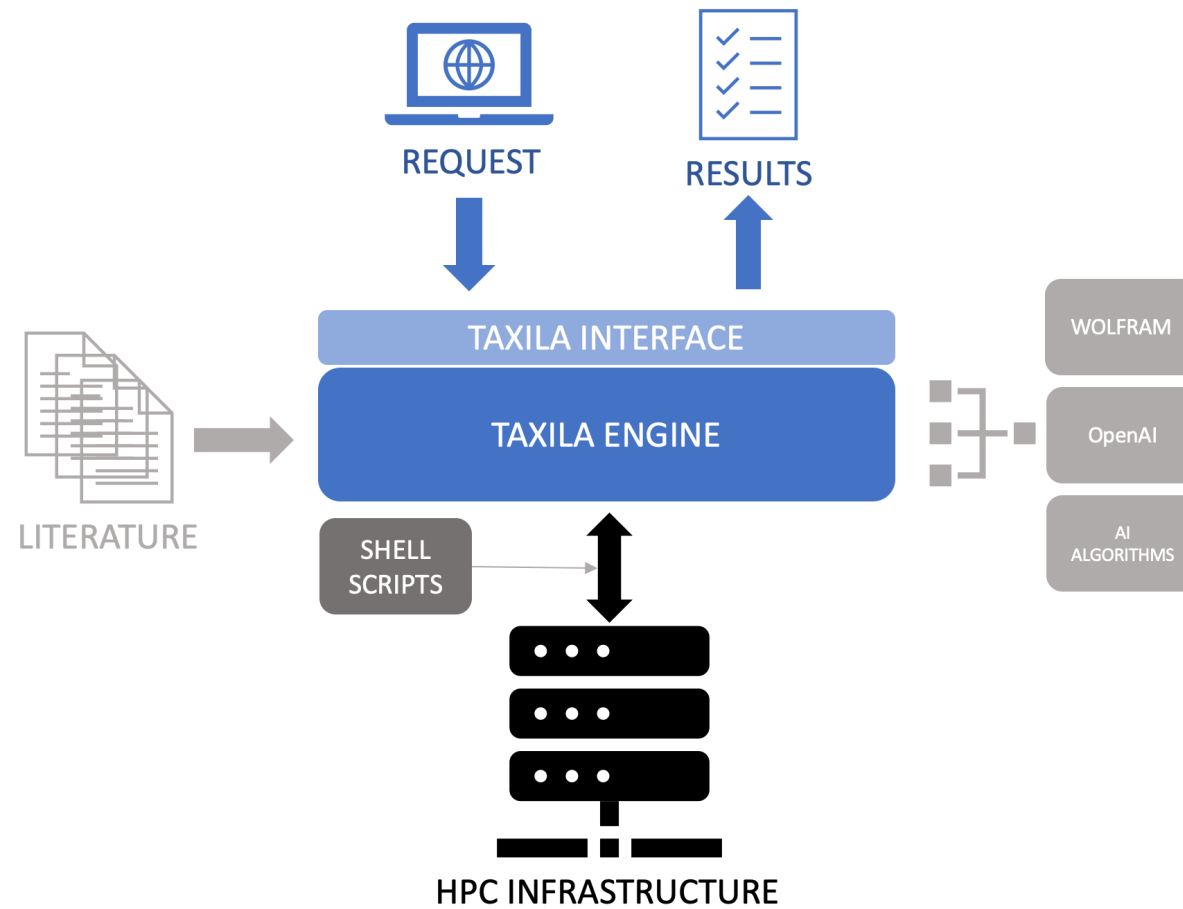
KNOWLEDGE EXTRACTION
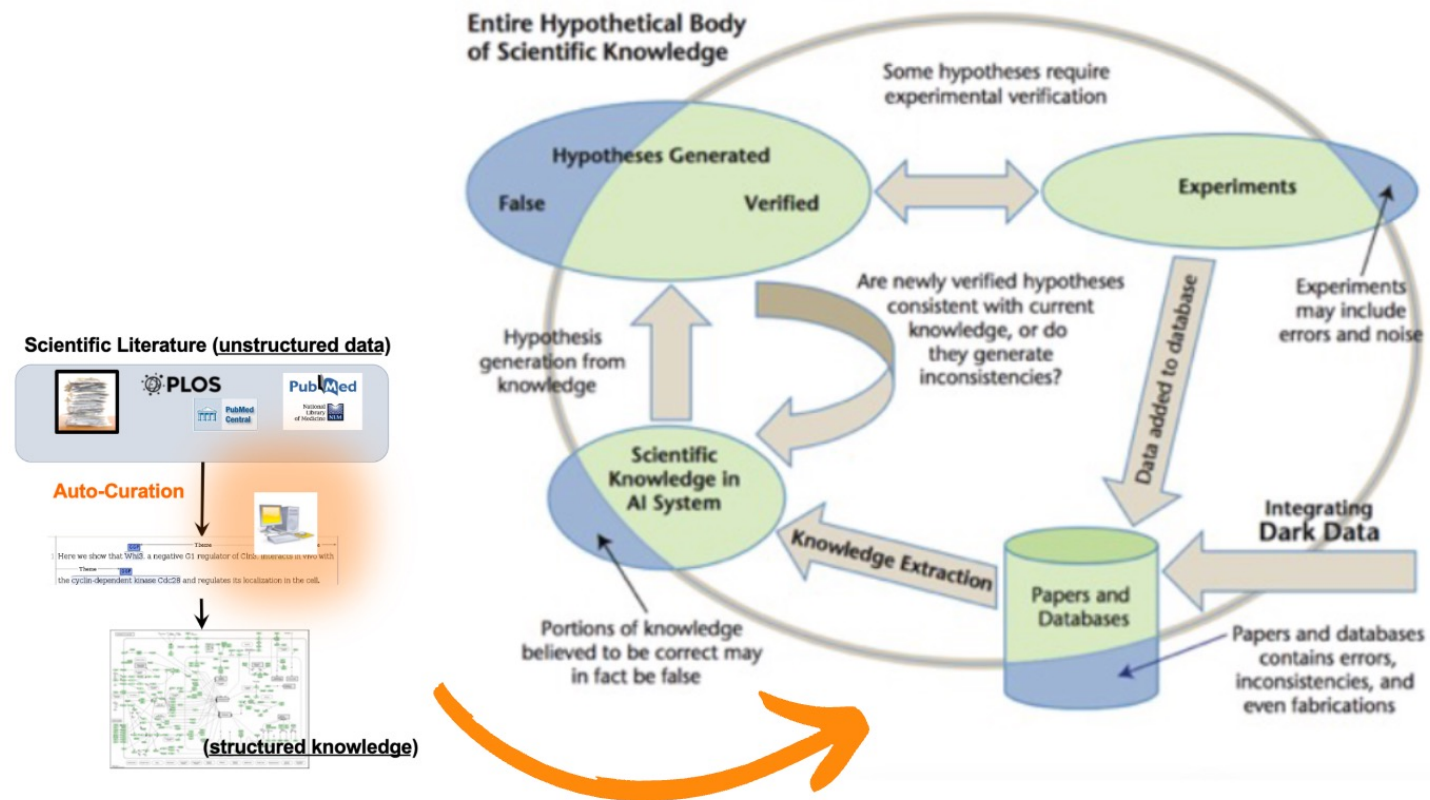
MATERIALS

- Materials:
  - Collection of scientific papers within the biomedical context
  - Focus on novel **CAR-T therapy**
  - Abstracts in PDF format
  - Collected from relevant sources
  - Updated daily
- CAR-T cancer therapy's mechanism is still **not fully understood** and encounters various obstacles.

METHODS

Collection of literature → Processing → Filtered information → Generating new knowledge

Scientific question → **Scientific answer**
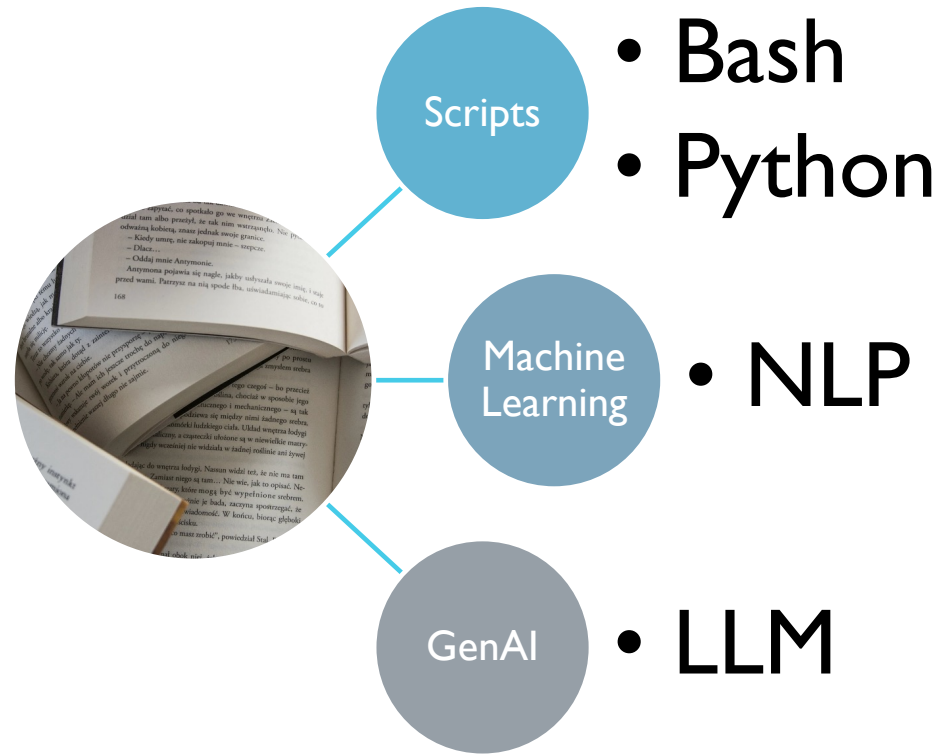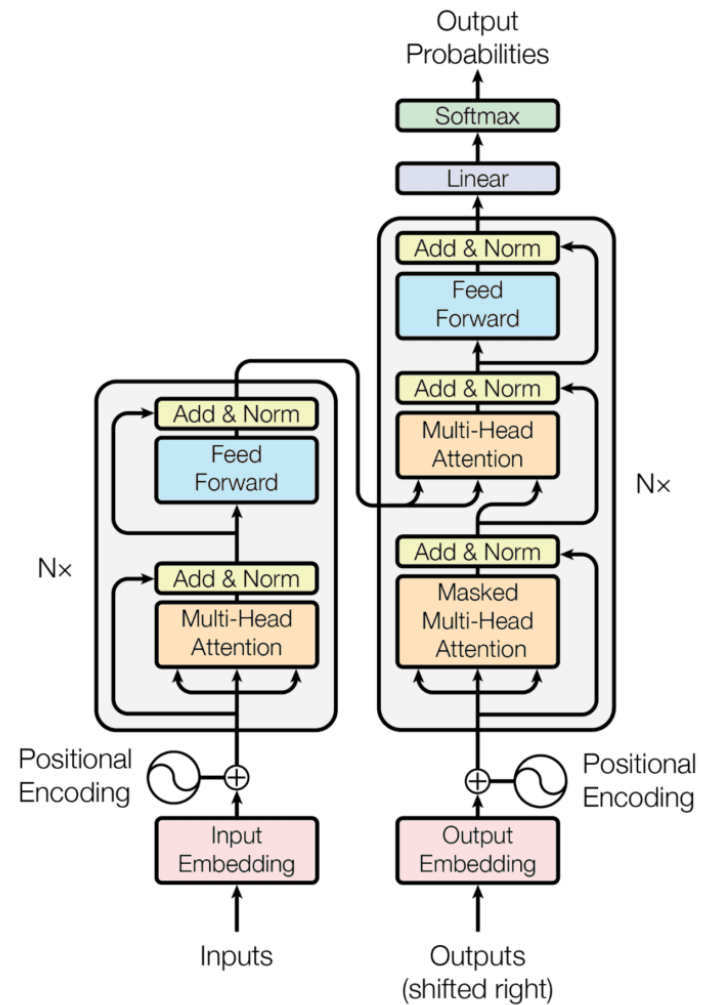
ARCHITECTURE

COMPUTING MACHINES

- HPC (High Performance Computing) environment:

  Okeanos system (Cray XC40) with 1084 specific compute, each node with 24 Intel Xeon E5-2690v3 cores, 128 GB RAM.

- Cloud

Source: icm.edu.pl

ALGORITHMICS BEHIND THE SCENE

Scripts
- Bash
- Python

Machine Learning
- NLP

GenAI
- LLM

TRANSFORMER MODELS

INSIGHTS

INSIGHTS

INSIGHTS

CONCLUSION AND FUTURE WORK

- Core questions:

  - Can a machine write a good scientific paper?

  - Can a machine win the Nobel Prize?

- What Next?

THANK YOU!

CONTACT:

ALEXANDRA.KOWALCZUK@GMAIL.COM