

Large Language Models as infrastructure for Open Science

Open Science Fair 2023
Sept 25-27, Madrid, Spain

@athenaRICinfo



Haris Papageorgiou

Athena Research & Innovation Center

OPIX



Outline

- State-of-Play in AI and Large Language Models
- LLMs as agents
- Case study – SciNoBo for Research Analysis
- Case study - Policy Intelligence
- LLMs as an infrastructure (Societal Impact)



1/ State of Play



What happened over the past ten years in AI research?

“The
Industrialization of
AI”

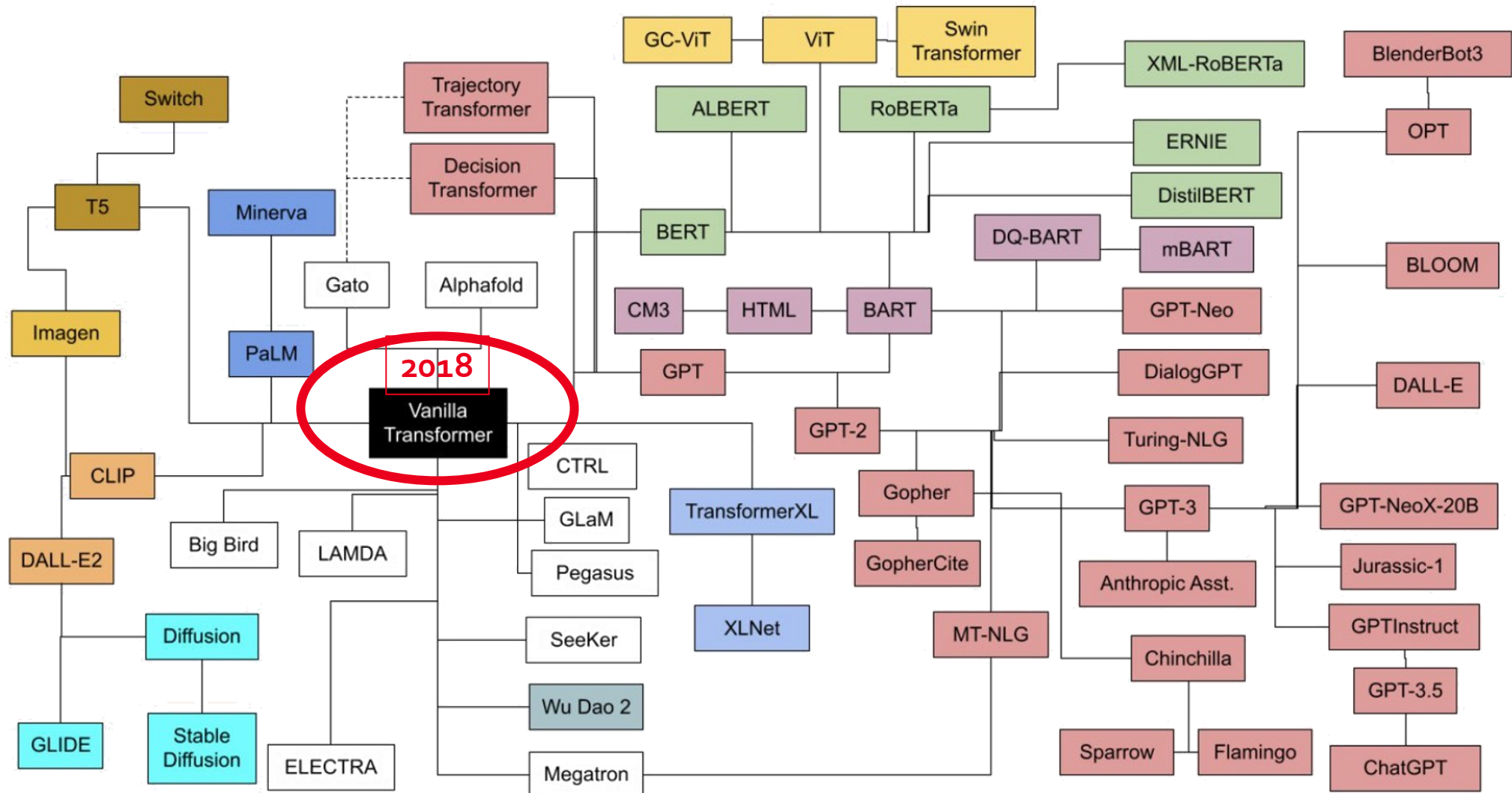
Scaling up of AI
systems (compute,
data)

Shift from academia
and government to
industry in terms of
research.

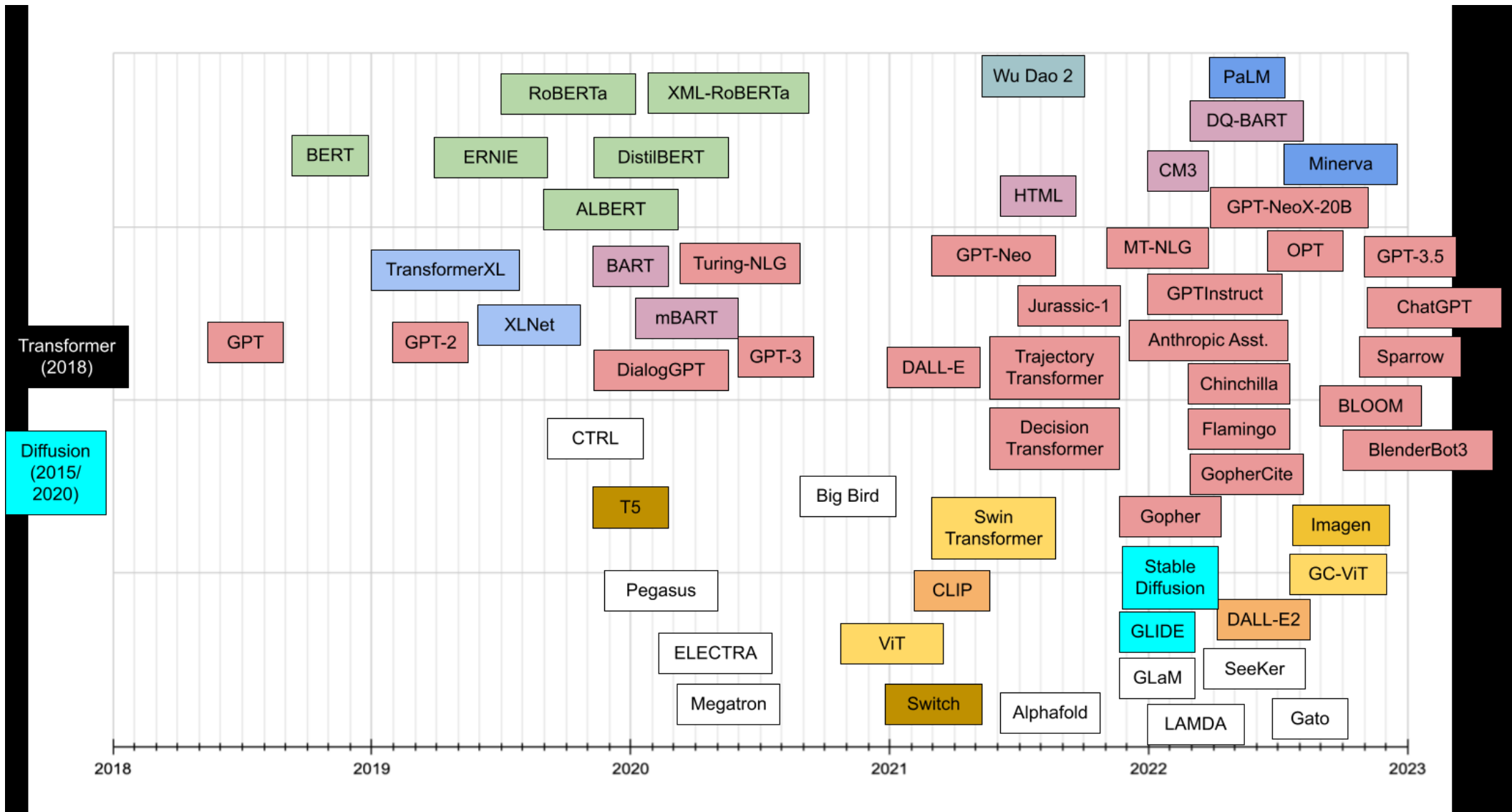
Homogenization
effect by fine-tuning
a few proprietary
FMs.

Large-scale
deployment of AI
systems by
companies.

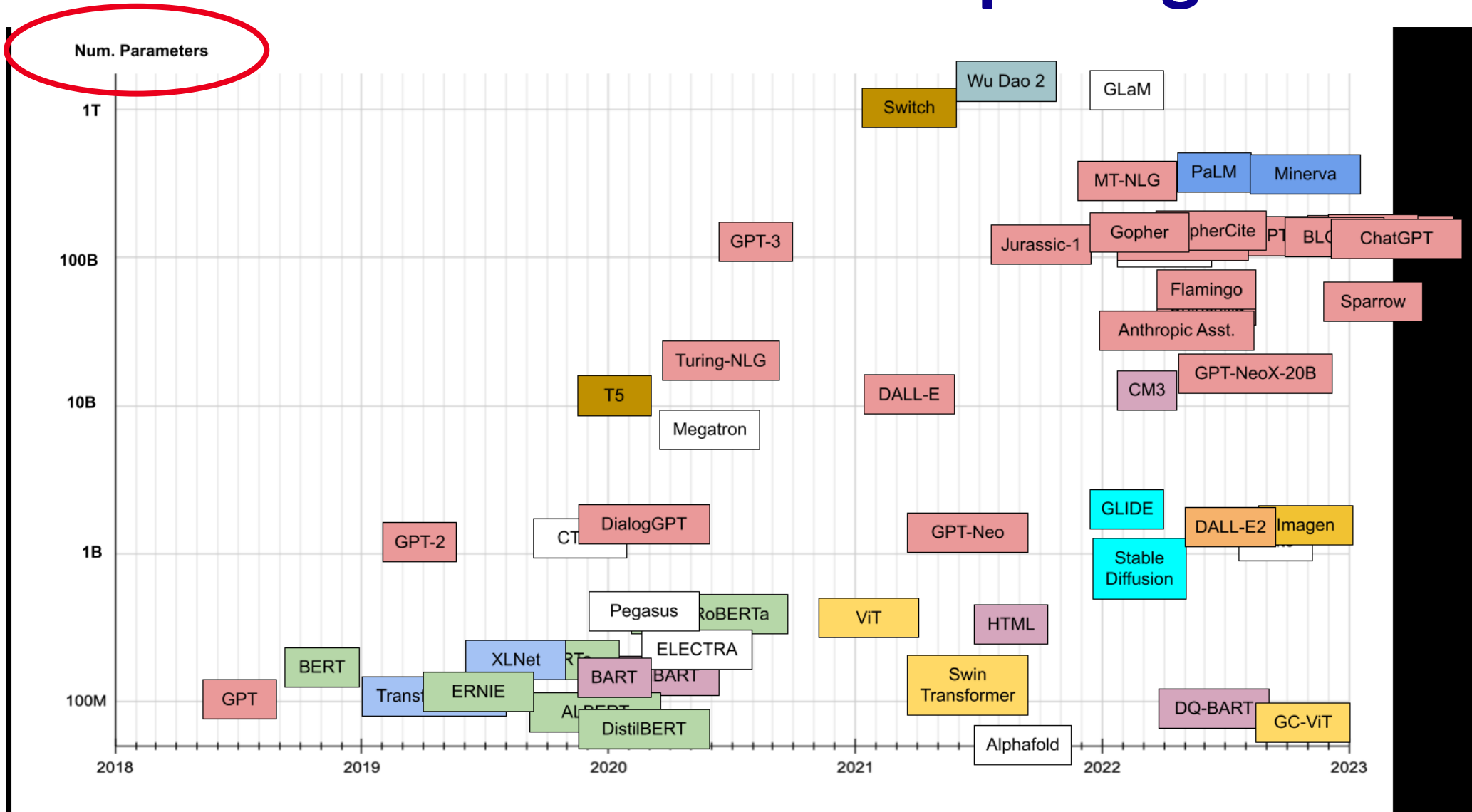
Rapid developments



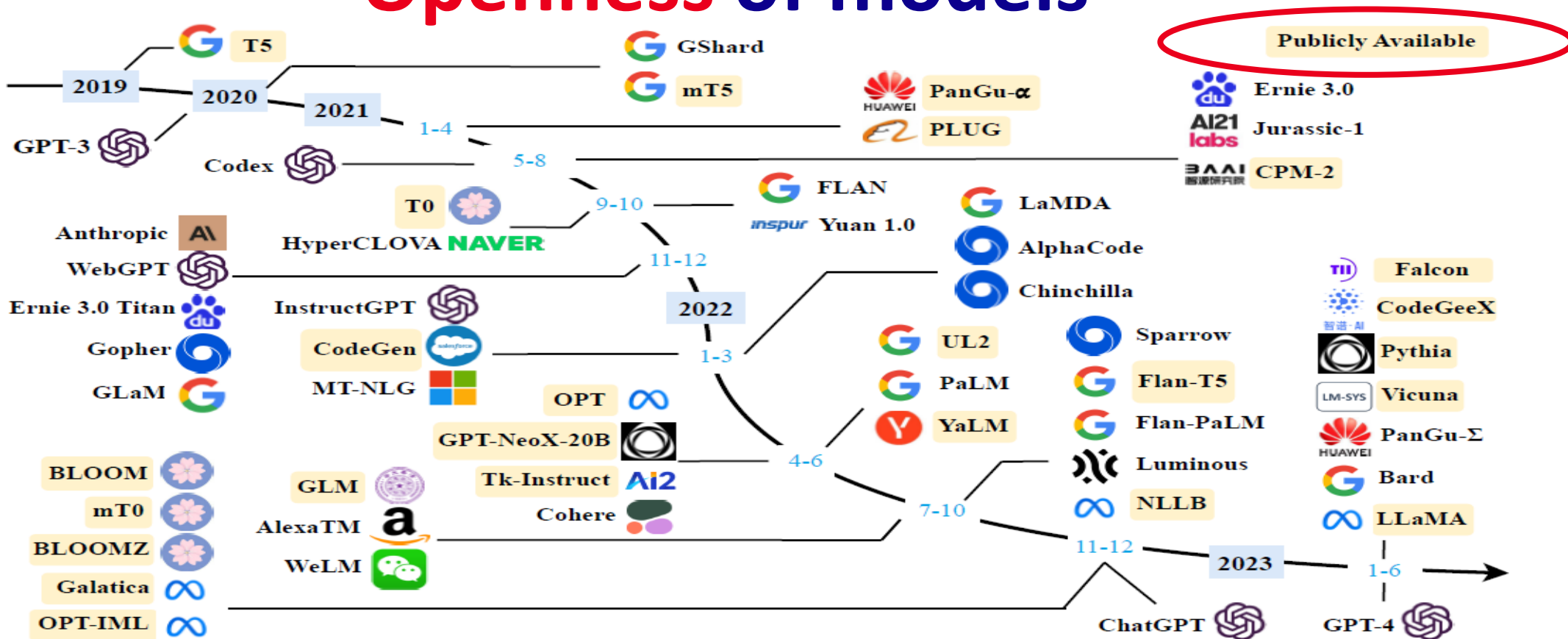
Community convergence



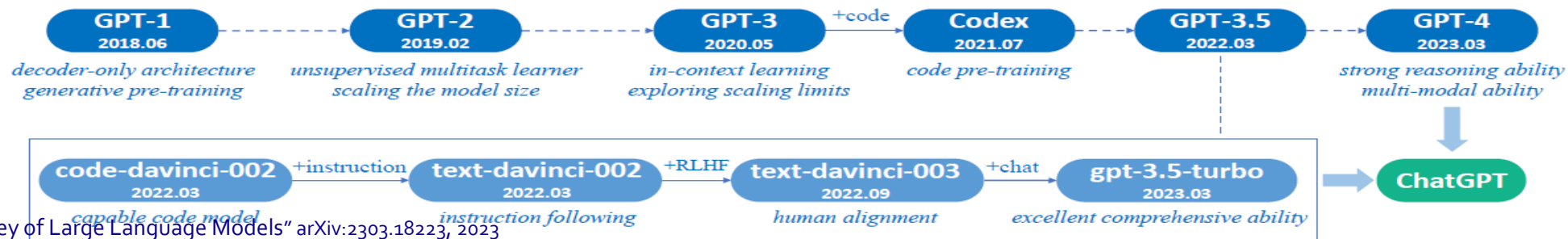
Increase in data & computing



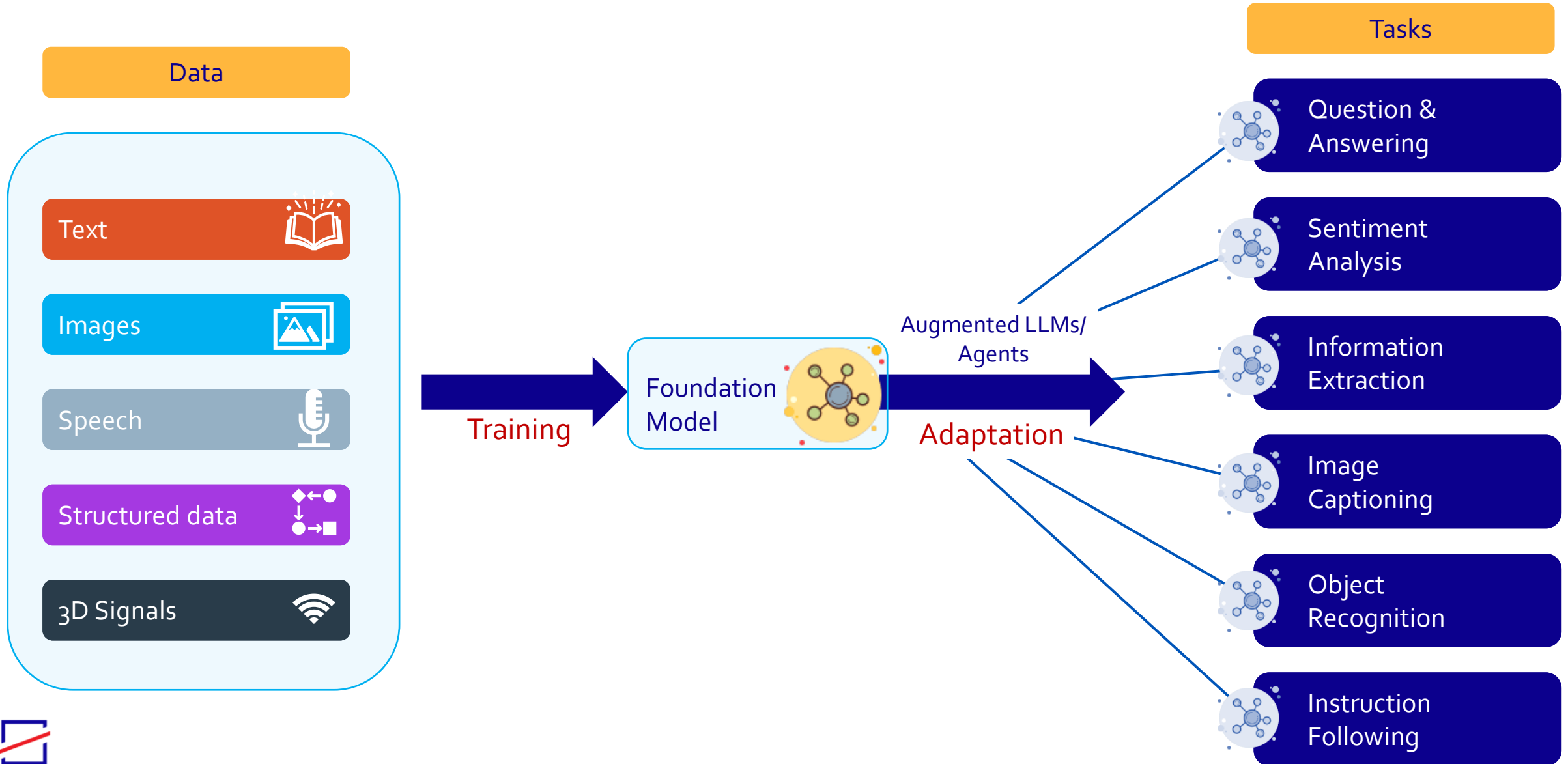
Openness of models



OpenAI



Foundation LLMs



Data

Text



Images



Speech



Structured data



3D Signals



Training

Foundation Model



Augmented LLMs/
Agents

Adaptation

Tasks

Question & Answering

Sentiment Analysis

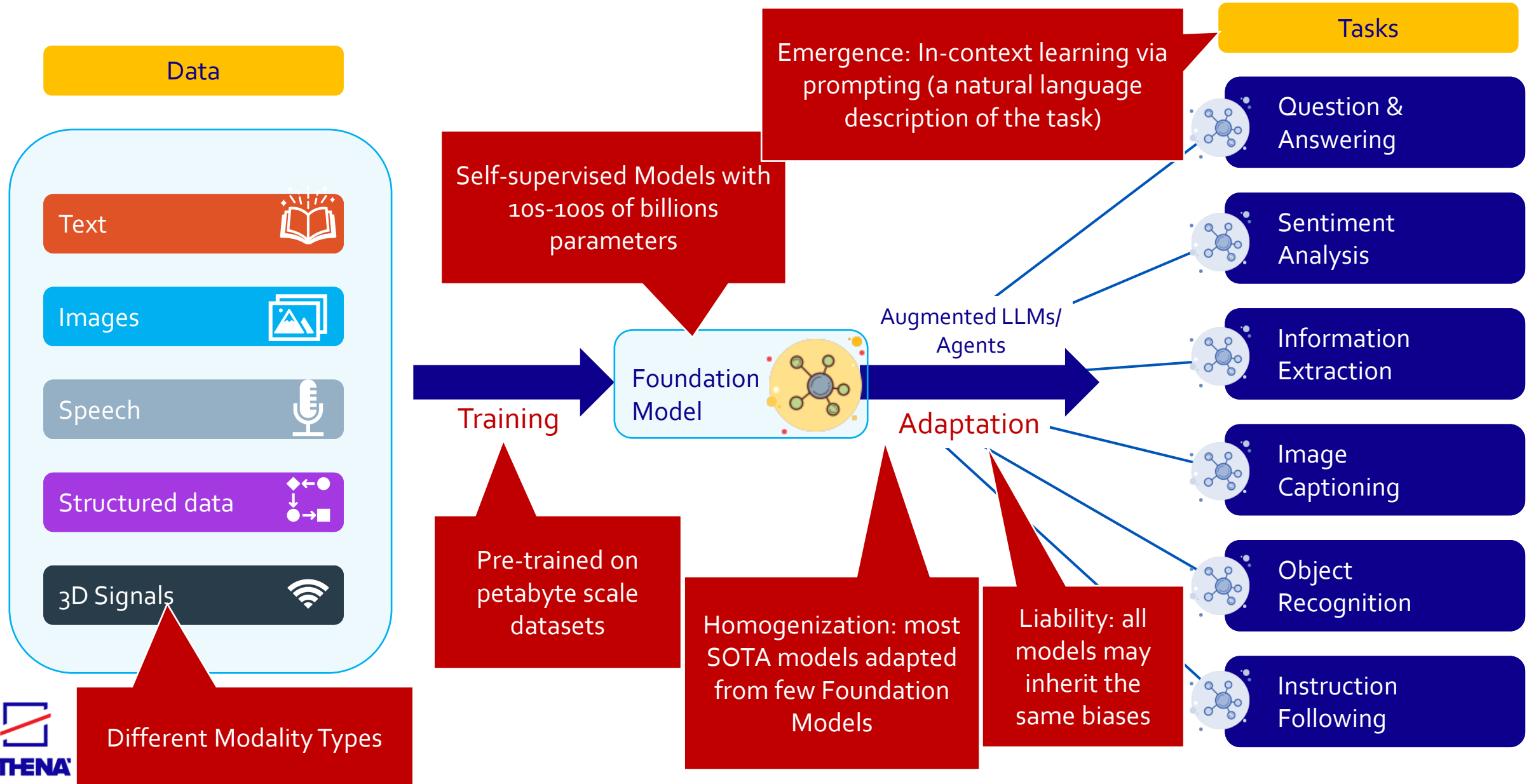
Information Extraction

Image Captioning

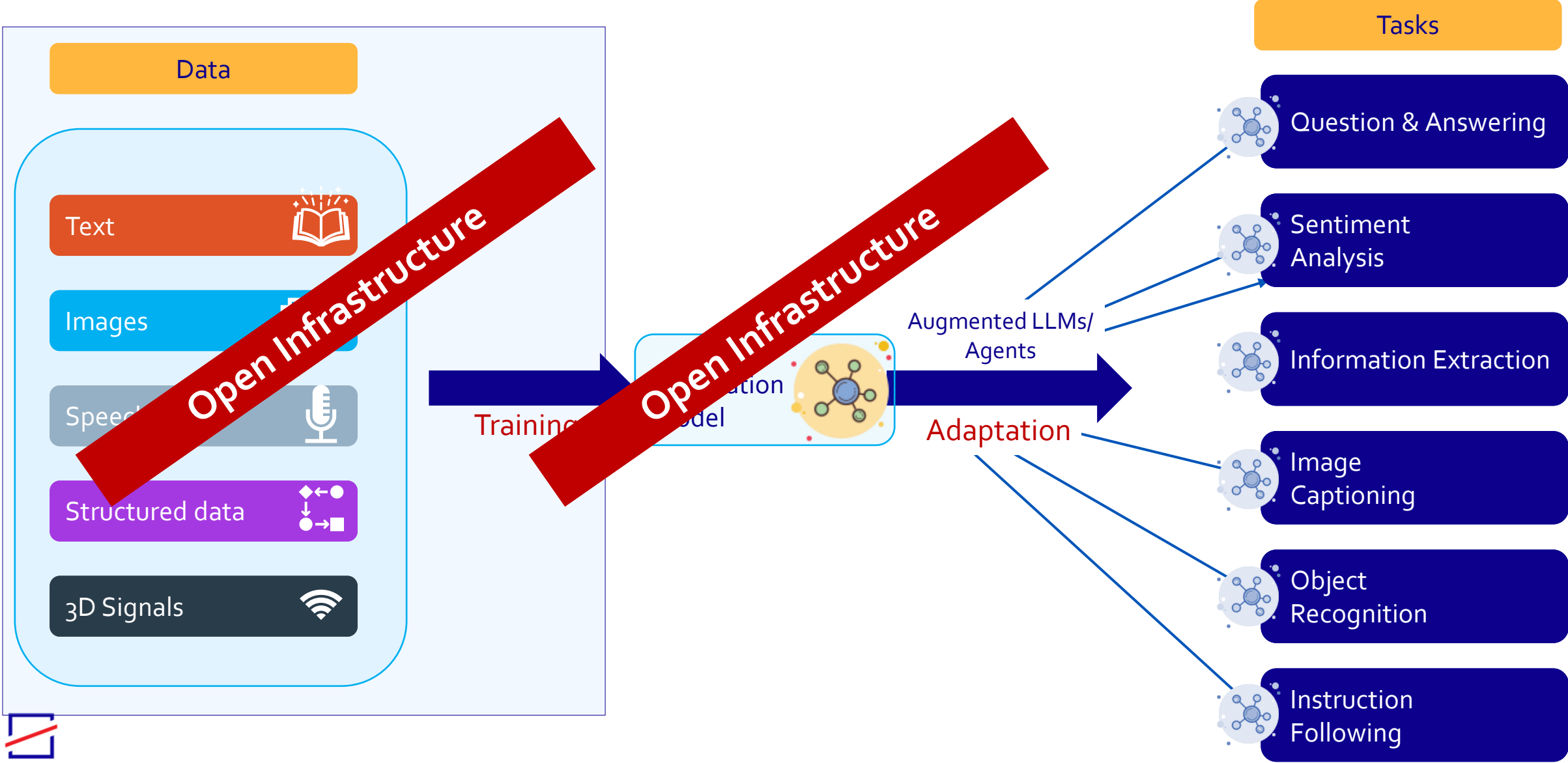
Object Recognition

Instruction Following

Foundation LLMs



The new Infrastructure

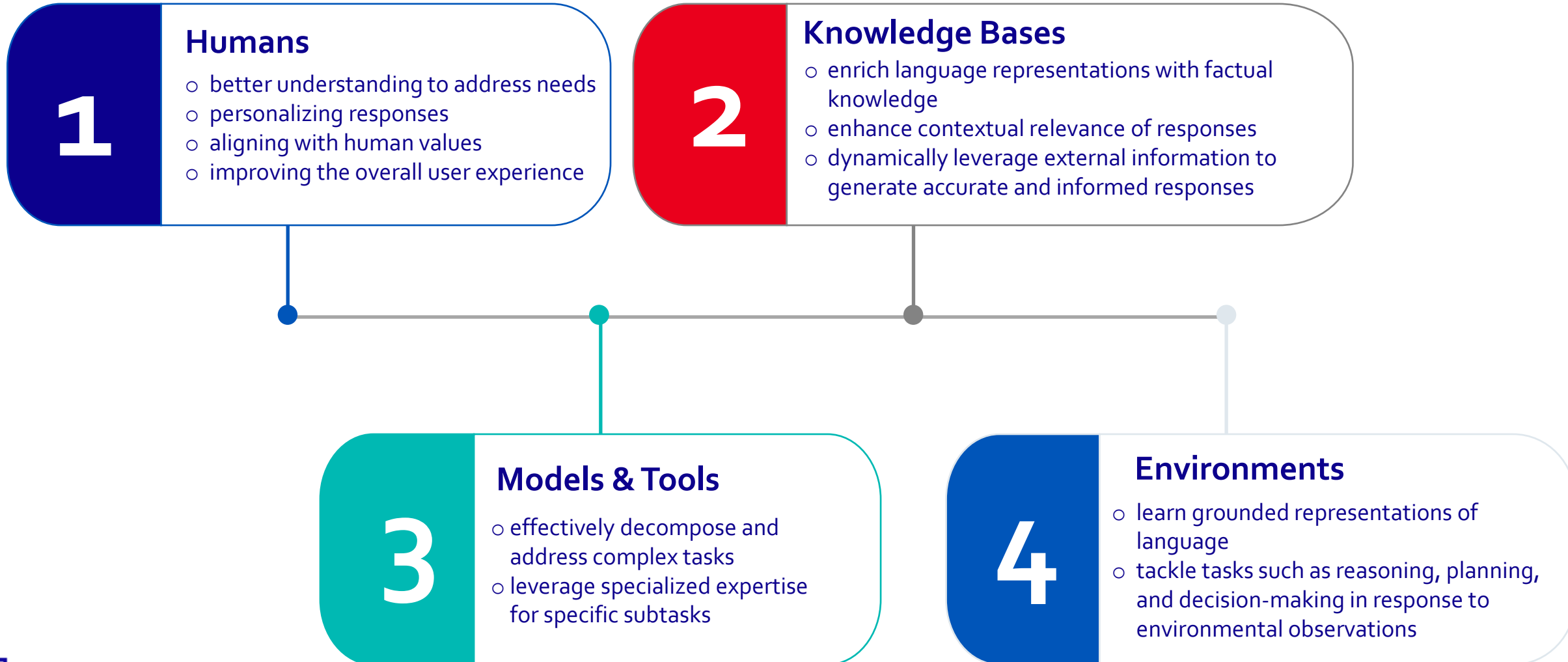




2/ LLMs as Agents



LLMs as **Agents** interacting with





Case study 1

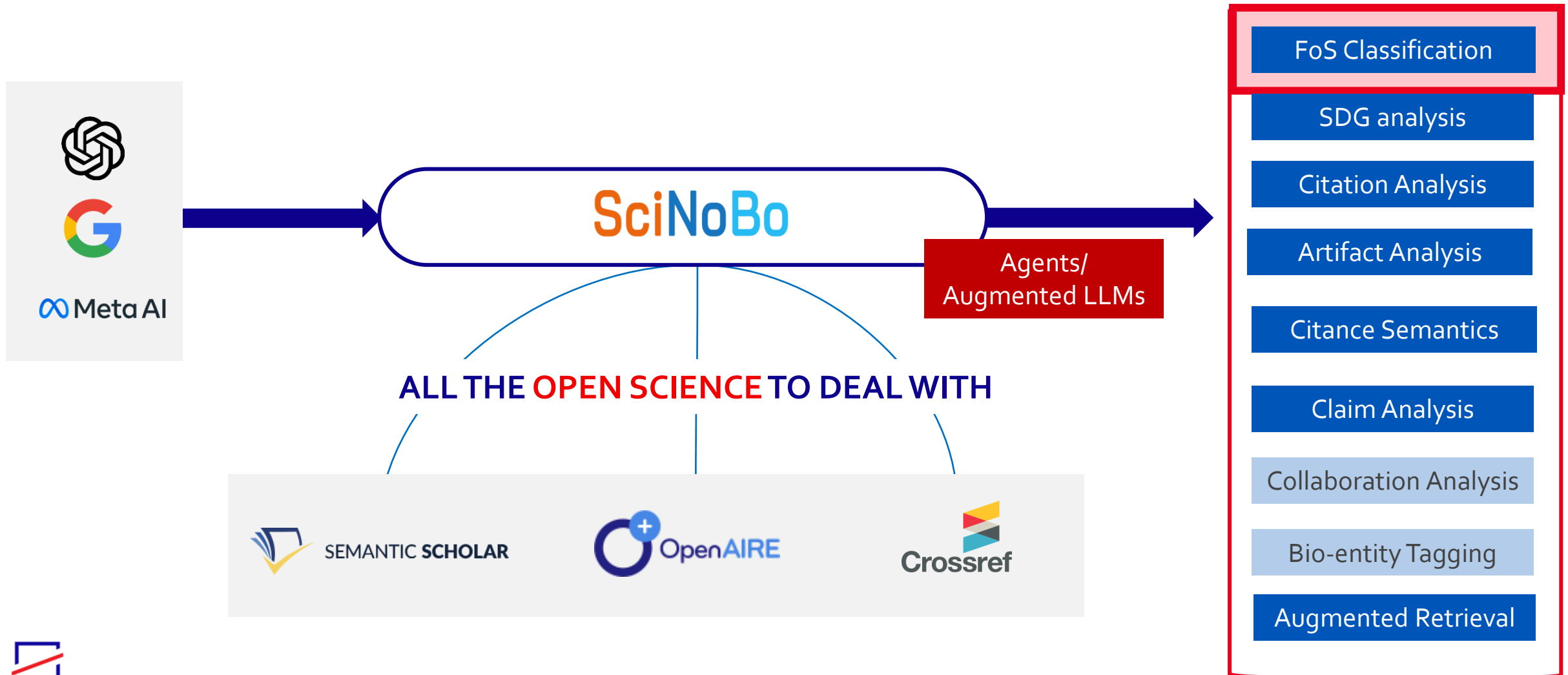
Augmented LLMs for Research Analysis

Science No Borders

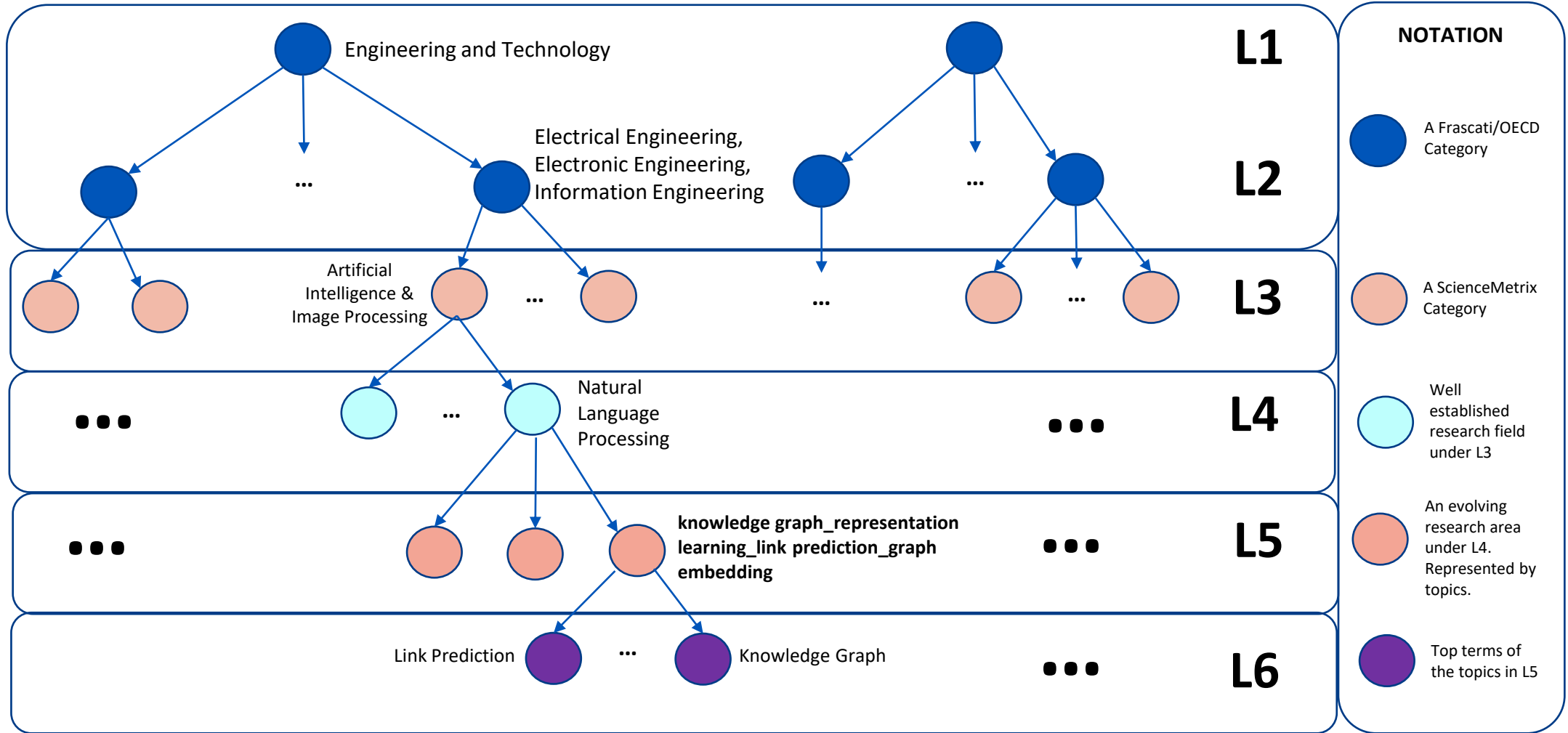
SciNoBo



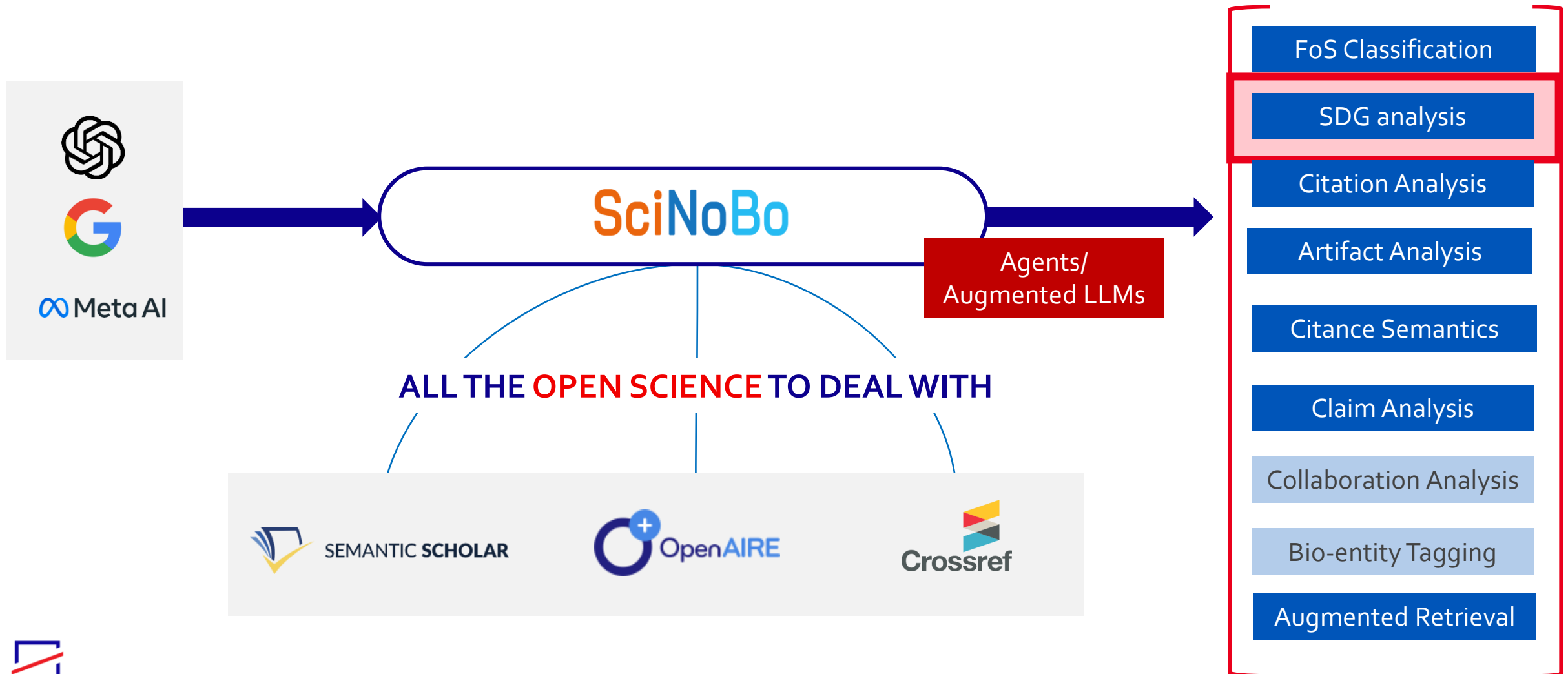
Enrich language representations with **factual knowledge**



LLM Agent: Field of Science Classification



Enrich language representations with **factual knowledge**



SDG Classification on Publications

Gender Differences in Job Search: Trading off Commute against Wage

Publication » Article • 01 Jan 2019 • Italy • Elsevier BV • SSRN Electronic Journal (eissn: 1556-5068, Copyright policy) • EC | ESEARCH

Authors: Thomas Le Barbanchon; Roland Rathelot; Alexandra Roulet;

DOI: 10.2139/ssrn.3467750, 10.1093/qje/qjaa033

Summary Subjects Metrics

Abstract
ABSTRACT We relate gender differences in willingness to commute to the gender wage gap. Using French administrative data on job search criteria, we first document that unemployed women have a lower reservation wage and a shorter maximum acceptable commute than their male counterparts. We identify indifference curves between wage and commute using the joint distributions of reservation job attributes and accepted job bundles. Indifference curves are steeper for women, who value commute around 20% more than men. Controlling in particular for the previous job, newly hired women are paid after unemployment 4% less per hour and have a 12% shorter commute than men. Through the lens of a job search model where commuting matters, we estimate that gender differences in commute valuation can account for a 0.5 log point hourly wage deficit for women, that is, 14% of the residualized gender wage gap. Finally, we use job application data to test the robustness of our results and to show that female workers do not receive less demand from far-away employers, confirming that most of the gender gap in commute is supply-side driven.

Country
Italy



Beta
SDGs Suggest
5. Gender equality
8. Economic growth

Beta
Fields of Science (4) View all & suggest >
social sciences
economics and business

Funded by
EC | ESEARCH

SDG Classification on HE Projects

HORIZON
EUROPE

Emission control system for wastewater treatment

Fact Sheet

Objective

Wastewater treatment plants are major contributors of greenhouse gasses emission and responsible for 1-2% of emissions worldwide. VARIO "emission control" is the first system that enables online monitoring of greenhouse gases (nitrous oxide and methane) and chemical pollutants in wastewater treatments plants. The VARIO solution measures relevant substances automatically and adjusts the wastewater plants aeration system via AI models. The technology combines sensorics (mass spectrometry) and data driven models (AI), for adjusting wastewater treatment process parameters. It does this in multiple aeration basins with one analyzer via its multiplexing capabilities to save costs for operators. By using VARIO the benefits for wastewater treatment plants are: 20% costs saving for energy and chemical consumption, as well as a tremendous reduction of emissions by up to 50%. For society, these savings are equivalent to millions of vehicles off the street or a 25% reduction in airline travel.

Fields of science

[engineering_and technology](#) > [environmental engineering](#) > [water treatment processes](#) > [wastewater treatment processes](#)

[natural sciences](#) > [chemical sciences](#) > [organic chemistry](#) > [aliphatic compounds](#)

[natural sciences](#) > [chemical sciences](#) > [analytical chemistry](#) > [mass spectrometry](#)

Project Information

VARIO

Grant agreement ID: 190116867

DOI

[10.3030/190116867](https://doi.org/10.3030/190116867)

Start date

1 April 2023

End date

31 March 2025

SDGs

6. Clean water
13. Climate action

Total cost


€ 2 463 750,00

EU contribution

€ 1 724 625,00

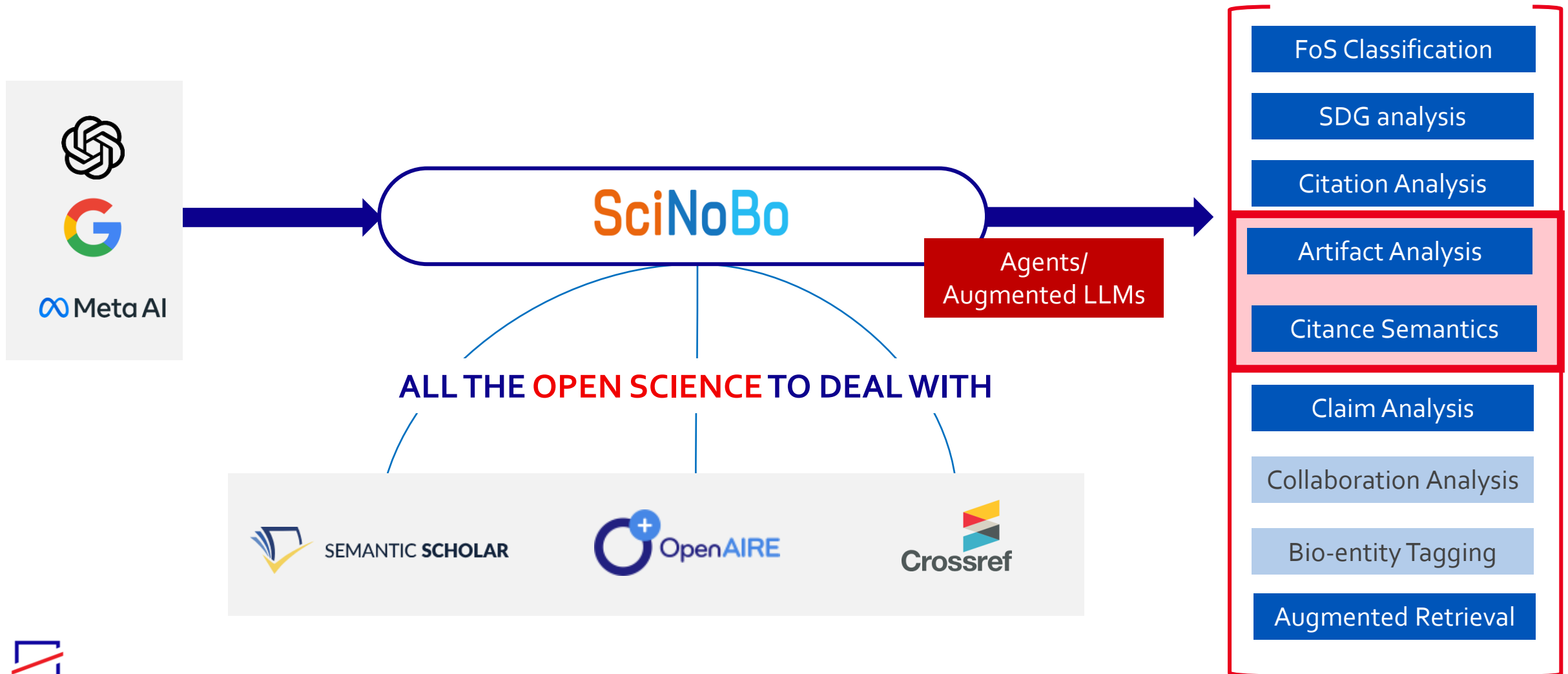
Coordinated by

VARIOLYTICS GMBH

 Germany



Enrich language representations with **factual knowledge**



Reproducibility - Artefact Detection

1. Abstract

As DNA sequencing and synthesis become cheaper and more easily accessible, the scale and complexity of biological engineering projects is set to grow. Yet, although there is an accelerating convergence between biotechnology and computing science, a deficit in software and laboratory techniques diminishes the ability to make biotechnology more agile, reproducible and transparent while, at the same time, limiting the security and safety of synthetic biology constructs. To partially address some of these problems, this paper presents an **approach** for physically linking engineered cells to their digital footprint - we called it **digital twinning**. This enables the tracking of the entire engineering history of a cell line in a specialised version control system for collaborative strain engineering via simple barcoding protocols.

In this paper we present a biotechnology specific version control **system**, **CellRepo**, that provides both the genetic toolkits and cloud-based software to physically link living samples to their digital footprint history. CellRepo is based on small, unique and bio-orthogonal DNA sequences inserted in specific genomic locations of a strain. By a single sequencing reaction, the DNA sequence can be retrieved, and hence the strain user can track down the entire digital footprint history of a strain via our web server: strain creators, parental and derivative strains, strain design documentation, related papers, experimental protocols, computer models, etc can all be retrieved via the cloud computing component of our system (Fig. 1). Put together, the biotechnology kits and the software repository move the digitalization of biotechnology a step closer making it more collaborative, scalable, transparent, trackable and reproducible.

Research Artifact Type	Method
Trigger	Approach
Name	Digital twinning
Class	Owned

Research Artifact Type	Software
Trigger	System
Name	CellRepo
Class	Owned

Reproducibility - Citance Analysis

DOI: 10.1101/786111

3.5.1. Toxin/Antitoxin

Using SOE-PCR a cassette containing both homologous arms, a mazF-ZeoR cassette and the barcode sequence was created and amplified following an adaptation of the protocol described in (Lin et al., 2013). After transformation colonies were restreaked on LB/Zeocin (20 µg/mL) plates and tested for the integration of the recombinant DNA by PCR. A positive clone was grown with xylose (1%) and the toxin gene induced. Cells were plated in xylose supplemented media. Individual colonies were restreaked on LB and LB/zeocin plates and colonies were tested positive by PCR and sequencing.

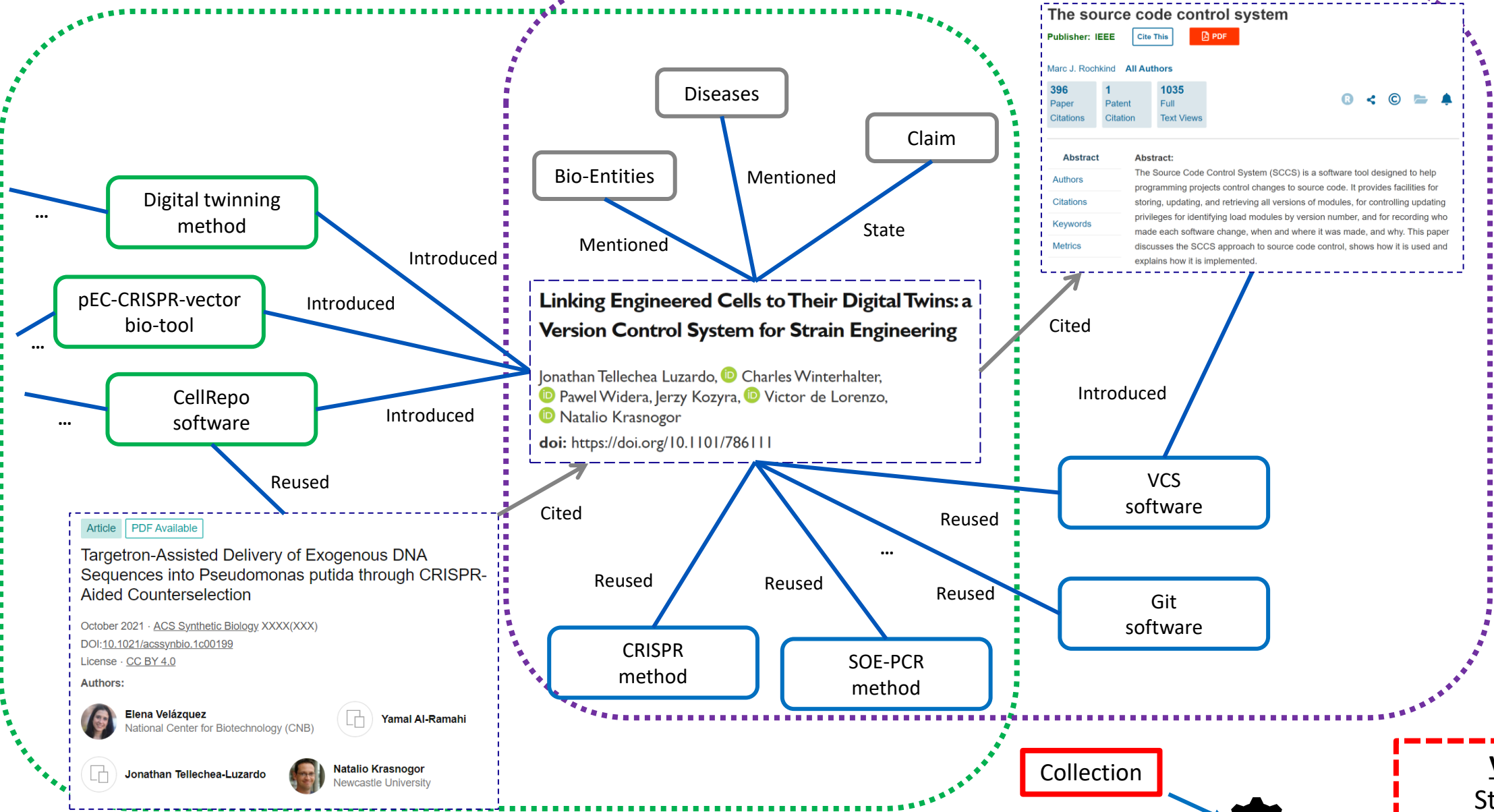
Intent	Reuse
Polarity	Supporting
Semantics	Methodology

Introduced by

Lin et al, 2013
10.1016/j.mimet.2013.07.020

Medical and Health Sciences

Engineering and Technology



Linking Engineered Cells to Their Digital Twins: a Version Control System for Strain Engineering
 Jonathan Tellechea Luzardo, Charles Winterhalter, Pawel Widera, Jerzy Kozyra, Victor de Lorenzo, Natalio Krasnogor
 doi: <https://doi.org/10.1101/786111>

The source code control system
 Publisher: IEEE [Cite This] [PDF]
 Marc J. Rochkind All Authors
 396 Paper Citations | 1 Patent Citation | 1035 Full Text Views
 Abstract: The Source Code Control System (SCCS) is a software tool designed to help programming projects control changes to source code. It provides facilities for storing, updating, and retrieving all versions of modules, for controlling updating privileges for identifying load modules by version number, and for recording who made each software change, when and where it was made, and why. This paper discusses the SCCS approach to source code control, shows how it is used and explains how it is implemented.

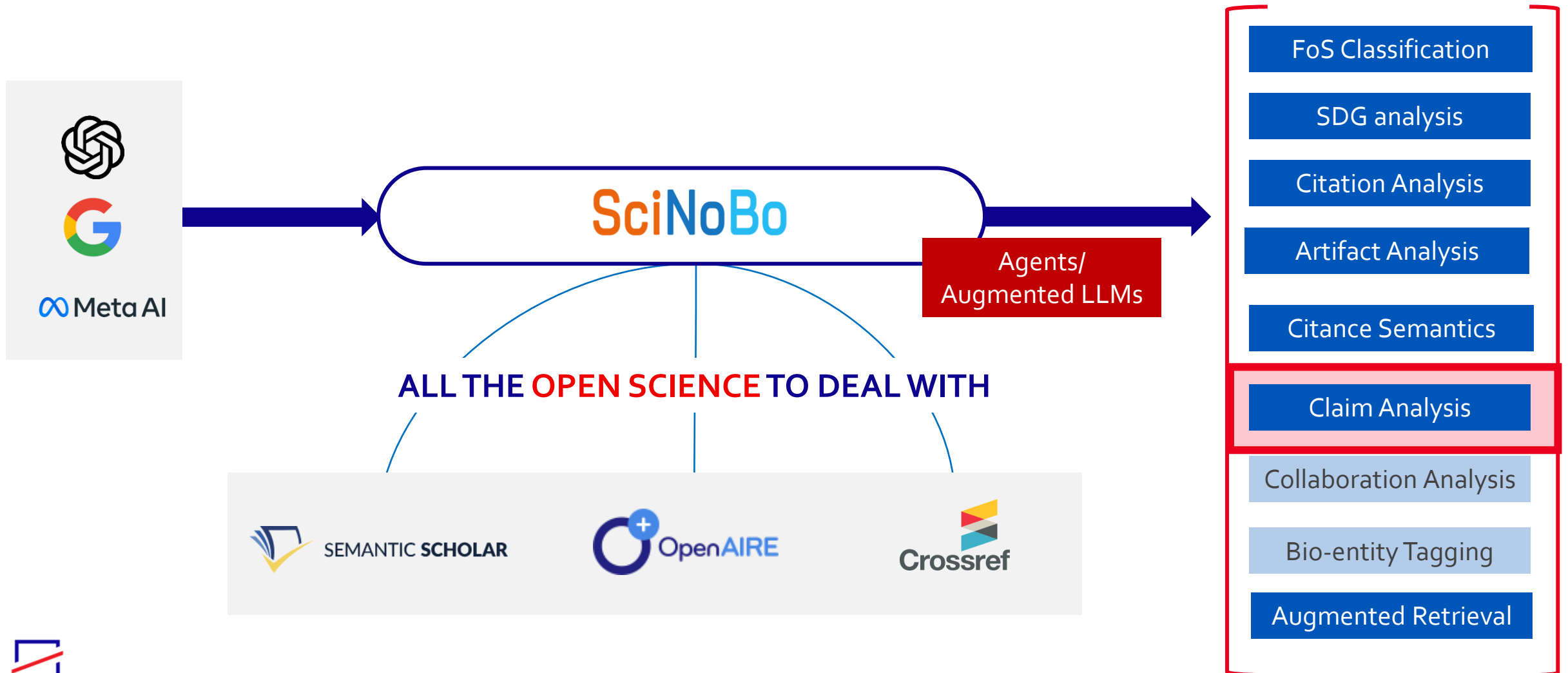
Article [PDF Available]
Targetron-Assisted Delivery of Exogenous DNA Sequences into Pseudomonas putida through CRISPR-Aided Counterselection
 October 2021 · ACS Synthetic Biology XXXX(XXX)
 DOI:10.1021/acssynbio.1c00199
 License · CC BY 4.0
 Authors:
 Elena Velázquez National Center for Biotechnology (CNB) Yamal Al-Ramahi
 Jonathan Tellechea-Luzardo Natalio Krasnogor Newcastle University

Collection



Vision
 Statistics
 Analytics
 Summary

Enrich language representations with **factual knowledge**



News Claim Analysis & Scientific Claim Verification

News Article

Health effects of aspartame draw new scrutiny from WHO experts



Claim Analysis

Claim: Decades after aspartame was approved for use in the United States, the sweetener's safety is getting another look by **global health bodies** assessing its **potential links to cancer**.

Claim Object: Aspartame

Claimer: The World Health Organization's International Agency for Research on Cancer

(CNN) — Decades after aspartame was approved for use in the United States, the sweetener's safety is getting another look by global health bodies assessing its potential links to cancer.

The World Health Organization's International Agency for Research on Cancer analyzed the potential carcinogenic effects of the sweetener this month. A separate WHO and United Nations committee, the Joint Expert Committee on Food Additives, is now updating its risk assessment, including what it considers to be an acceptable daily intake. Their findings have not been made public; they will be released together July 14.

Scientific Claim Verification

Results

Info

"doi": "[10.1007/s10616-013-9681-0](https://doi.org/10.1007/s10616-013-9681-0)",

"sentence": "So, consumers should be aware of the potential side effects of aspartame before they consume it."

"doi": "[10.1039/c5tx00269a](https://doi.org/10.1039/c5tx00269a)",

"sentence": "Cell viability was significantly altered following a higher concentration of aspartame exposure."

"doi": "[10.3390/nu13061957](https://doi.org/10.3390/nu13061957)",

"sentence": "Further research should be conducted to ensure clear information about the impact of aspartame on health."

No Support or Refute Claims/Evidence



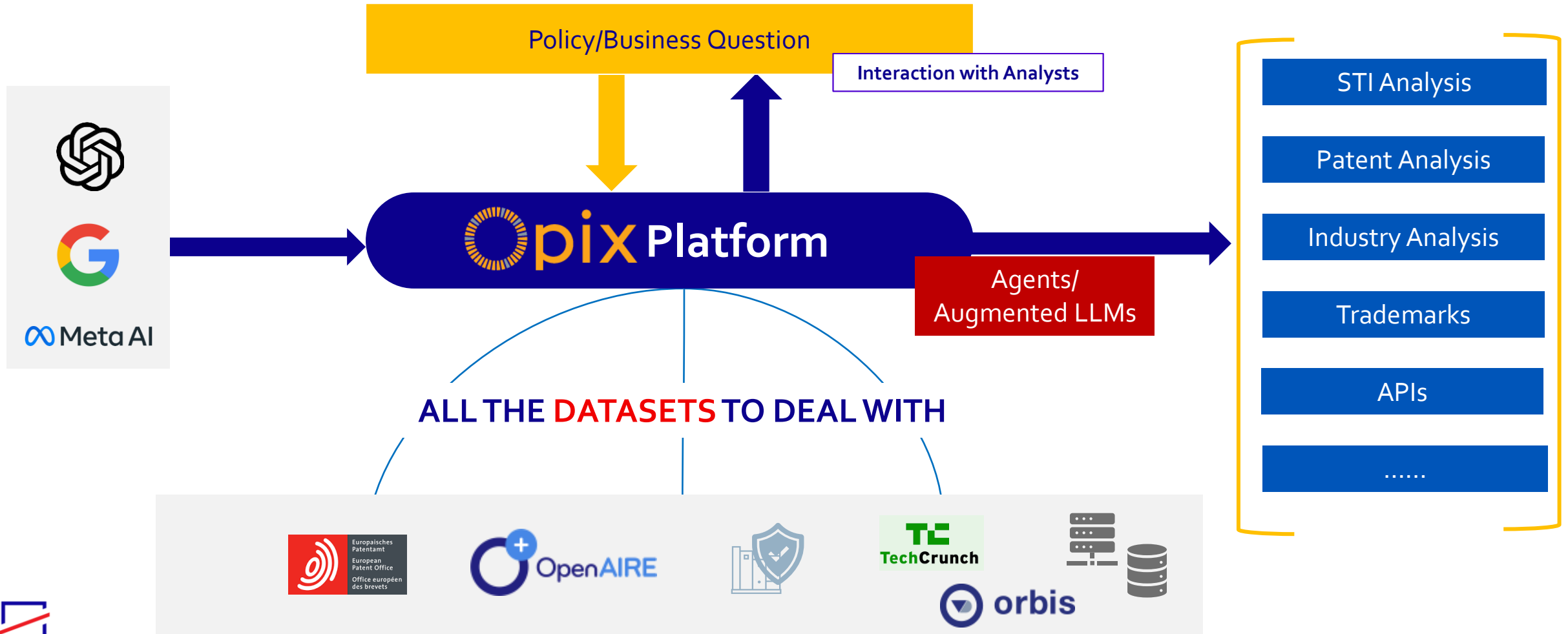
Case Study 2

LLMs – Policy Intelligence

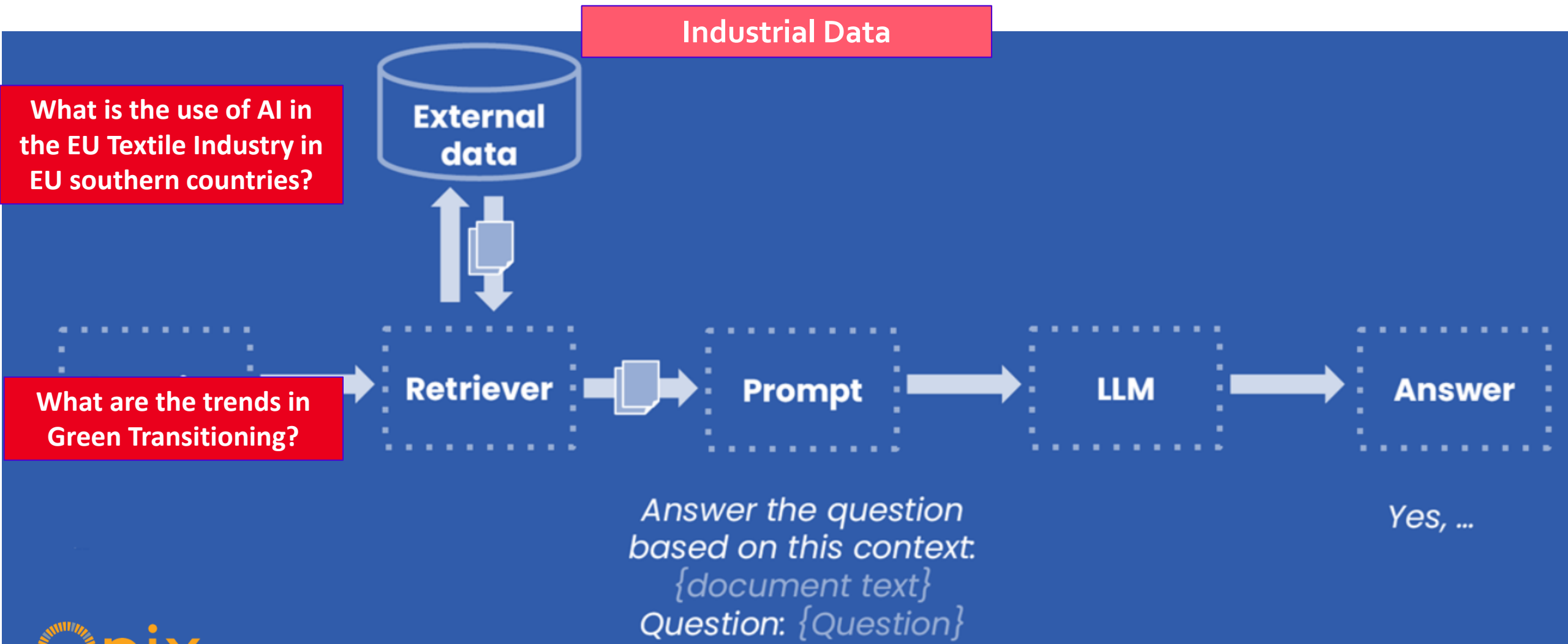
The OPIX Platform



Combining data from **different sources**







Data Quality - Retrieval Augmentation



Tech Innovation – (a) on scientific data

Articles

Yarn evenness parameters optimization in jetring spinning process

Ekrem Gulsevincler  , Mustafa Resit Usal  & Demet Yilmaz 

Abstract

In this study, the effects of air nozzles called jetring or nozzlering, on yarn quality which is used as an additional process in conventional ring spinning machines have been investigated. The response surface methodology (RSM) was used to investigate the yarn properties value. In the experiment, Ne30/1 100% cotton fibers were used. In all jetring yarn productions, the air pressure was kept at 125 kPa (gauge). In all samples, the nozzle length was kept 27 mm and twisting chamber diameter was kept Ø2mm. The number of injectors has been kept constant as 4 pieces. Giving the best yarn quality as a result of RSM, injector diameter Ø0.5 mm and injector angle 35° as determined. With this nozzle structural configuration, yarn hairiness values were reduced by 9.2% but yarn irregularity values were increased by 0.7%, yarn elongation values were decreased by 6.2% and yarn tenacity (cN/tex) values decreased to 5.6.

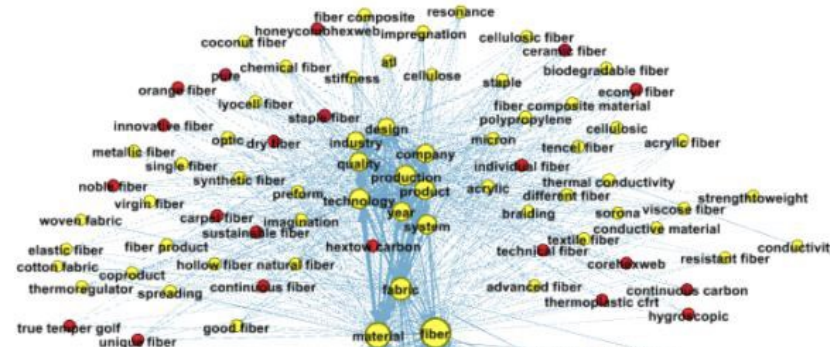
Tech Innovation – (b) on industrial data

WATER SAVINGS THROUGH SYMPATEX LAMINATES

At the beginning of 2017, Sympatex has committed itself to using more and more recycled polyester for its uppers and lining materials as part of its "Agenda 2020". The advantage: if you laminate the Sympatex membrane made of pure polyether/ester with recycled polyester uppers and liners, you will get pure laminates that can easily be added to the closed textile loop at the end of their product life cycle. Furthermore, recycled polyester consumes much more water compared to virgin polyester. In fact: If you compare the production of 1 kg recycled "New Life" fibres with 1 kg oil-based polyester fibres, then the 90 percent water savings are excellent or in other words: you only need around 3 instead of 60 litres water for 1 kg fibres.

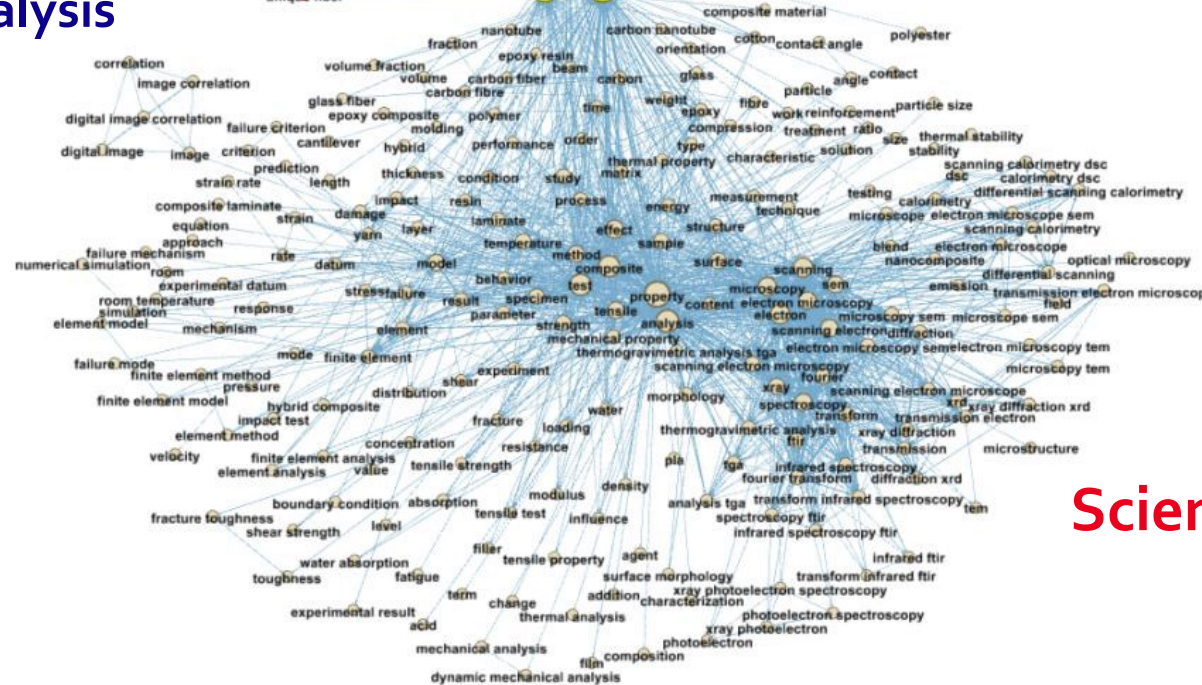
Technology Uptake Indicator

Industrial Graph



Matching terms via network analysis

- Node coupling strength
- Edge coupling strength
- Technological Uptake



Science Graph

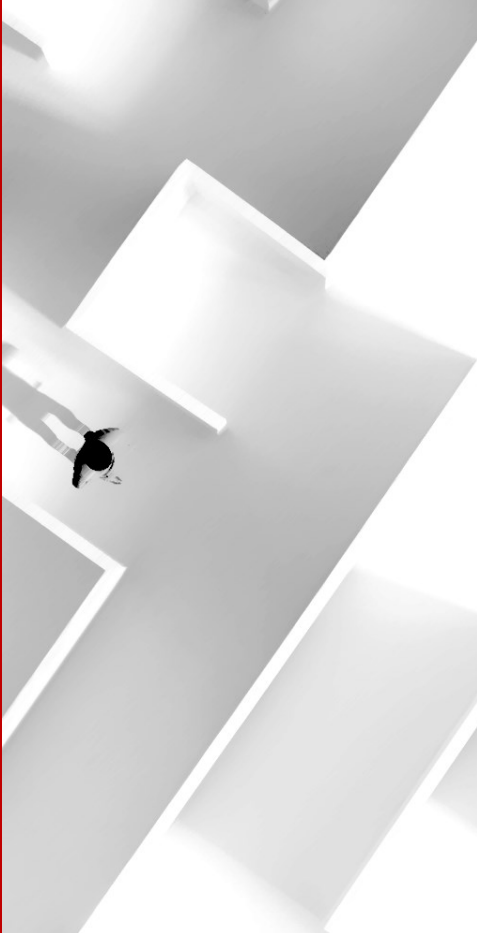


Open Science at rescue
Legality, Ethics



Limitations

- **Misalignment with human needs**
- **Lack of interpretability**
- Provide nonfactual but seemingly plausible predictions (**hallucinations**)
- **Recency – keeping LLMs up-to-date**
- **Attribution – providing citations**
- **Limited ability for complex reasoning**



LLMs & Societal Impact: Legal & Ethics

- **Data collection:** web scraping/crawling, copyrighted & IPR protected material, privacy law (GDPR)
- **Output Liability: Deployment** in sensitive domains or governmental entities – decisions rooted in the same FMs weaknesses and biases
- **Concentration effect:** increased inequality due to potential centralization
- **Ethics:** amplification of weaknesses and biases of the FMs (same FM model used in a variety of domains)

Rishi Bommasani et al; "On the opportunities and Risks of Foundation Models", ArXiv, abs/2108.07258, 2022

What do we need **to do**

- **Source tracing for attributing responsibility**
 - Training data, adaptation data, test-time user data/interaction
- **(Good) practices in data/model management**
 - Data curation, data selection, data weighting, documentation,
 - affecting modeling decisions (training objective, model architecture, adaptation method), transparency
- **Community values and norms**
 - User studies with diverse user groups
 - User feedback to model design,
 - Pro-active vs re-active interventions,
 - Other interventions addressing bias mitigation, auditing

Rishi Bommasani et al; "On the opportunities and Risks of Foundation Models", ArXiv, abs/2108.07258, 2022

A risk-based approach under which different obligations apply according to the risk posed by the AI system

What do we need – EU AI Act

- **Unacceptable risk AI systems** are prohibited (Title II), for example, because they are contrary to Union values.
- **High-risk AI systems** are permitted, albeit subject to obligations (Title III). High risk refers to AI systems creating a high risk to the health and safety or fundamental rights of natural persons. They are subject to certain mandatory requirements and an ex-ante conformity assessment.
- **Limited risk AI systems** are allowed if they meet minimal transparency requirements, such as making users aware that they are interacting with AI (Art. 52).
- **Low risk AI systems** can be developed, produced and used freely. However, providers can choose to voluntarily conform to the AI Act.



LLMs as Infra supported by **Open Science community**

Open Science accelerates AI Diffusion in almost all sectors

The ability of a country or region to capitalize on scientific and technological progress is diffusion efficiency, i.e., the speed at which the technology diffuses in the economy.

Jeff Ding, Assistant Professor of Political Science
George Washington University, testimony on US Senate Committee on "Intelligence", 19 Sept. 2023

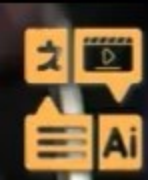


UNITED STATES SENATE
COMMITTEE HEARING CHANNELS

Yann LeCun
Meta
Turing Award



ATHENA



AI assisted translation
<https://targum.video>



EΥΧΑΡΙΣΤΩ/**GRÀCIES**/HVALA/DĚKUJI/TAK/**DANK**
JEWEL/AITÄH/KIITOS/MERCI/**DANKE**/KÖSZNÖNÖ/
GRAZIE/PALDIES/**AČIŮ**/GRAZZI/TAKK/DZIĘKUJĘ
OBRIGADO/MULTUMESC/**СПАСИБО**/ХВАЛА
HVALA/**БЛАГОДАРЯ**/THANKYOU/TAK/GRACIAS
/KIITOS/TACK/TEŞEKKÜREDERİM/**СПАСИБИ**/
JUFALEMINDERIT/**ΕΥΧΑΡΙΣΤΩ**/DANKJEWEL/TAK
TACK/**GRAZZI**/DANKJEWEL/MULTUMESC/AITÄH
KÖSZNÖNÖ/СПАСИБО/**ХВАЛА**/AČIŮ/**THANKYOU**

