Generative AI:
paradigms and ideas

R. Basili
(Università di Roma, Tor Vergata)

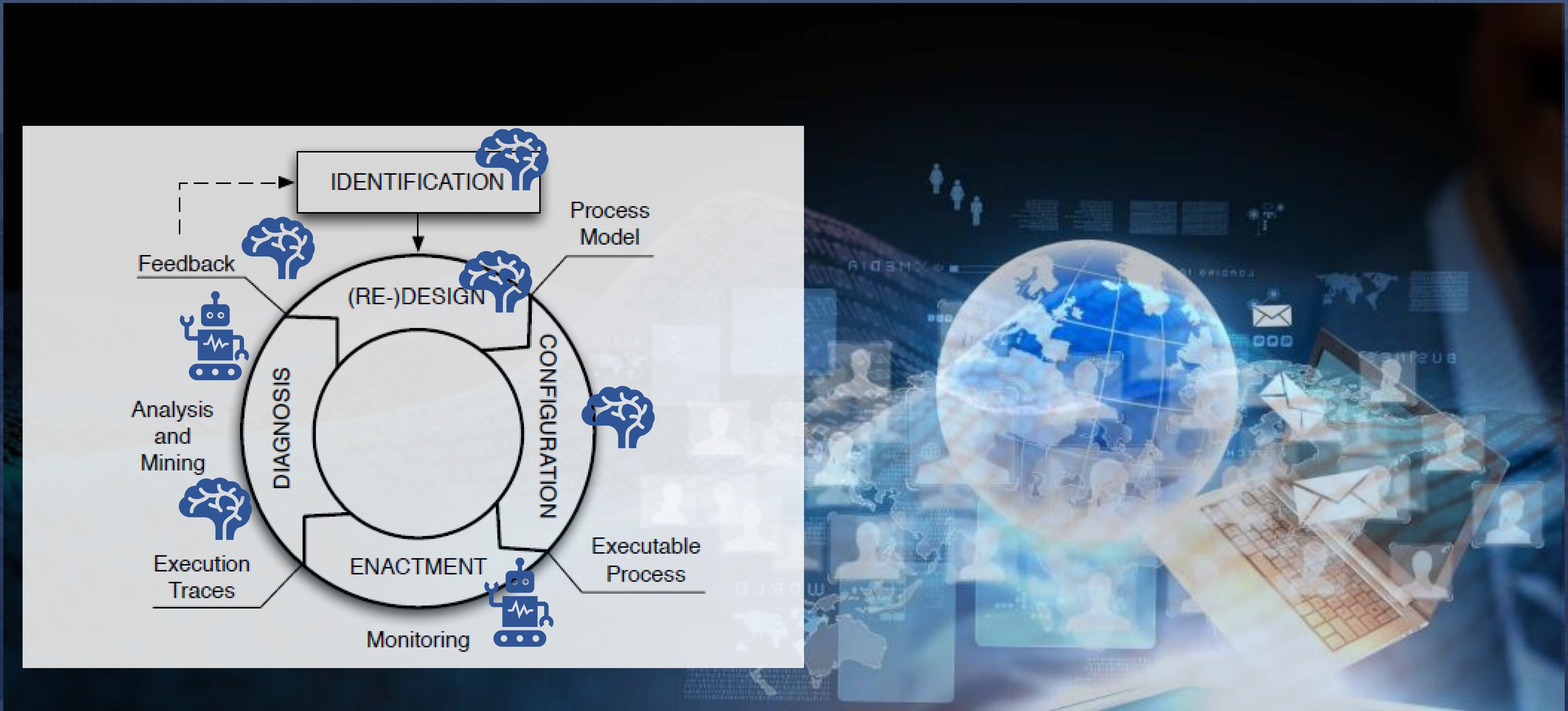
ABILab AI Hub WS "**DATA GOVERNANCE E INTELLIGENZA ARTIFICIALE LA "MAGICA" ALCHIMIA**"

Milano, 31 Gennaio 2024

Overview

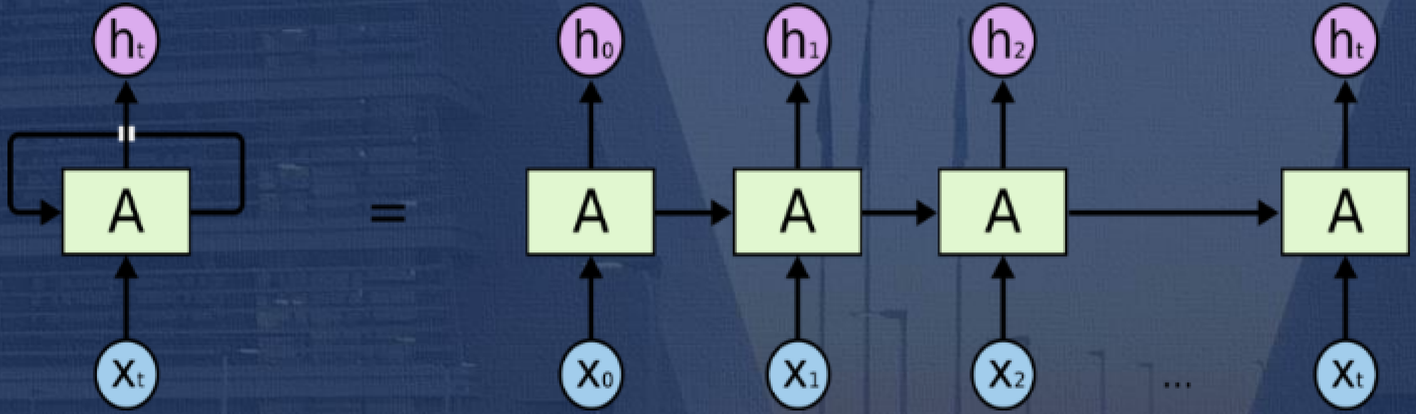
- **Knowledge and Information in Data Governance**
- **Paradigms and Applications of Generative AI Models**
  - From Encoder Networks to Decoders
  - Training Decoders: fine tuning, in-line learning and prompting
  - **Use Cases:**
    - Tourism, Health and Banking
- **Trends**
- **Conclusions**

Data Management: Objective and Challenges



Machine learning paradigms underlying ChatGPT

RNNs
1986

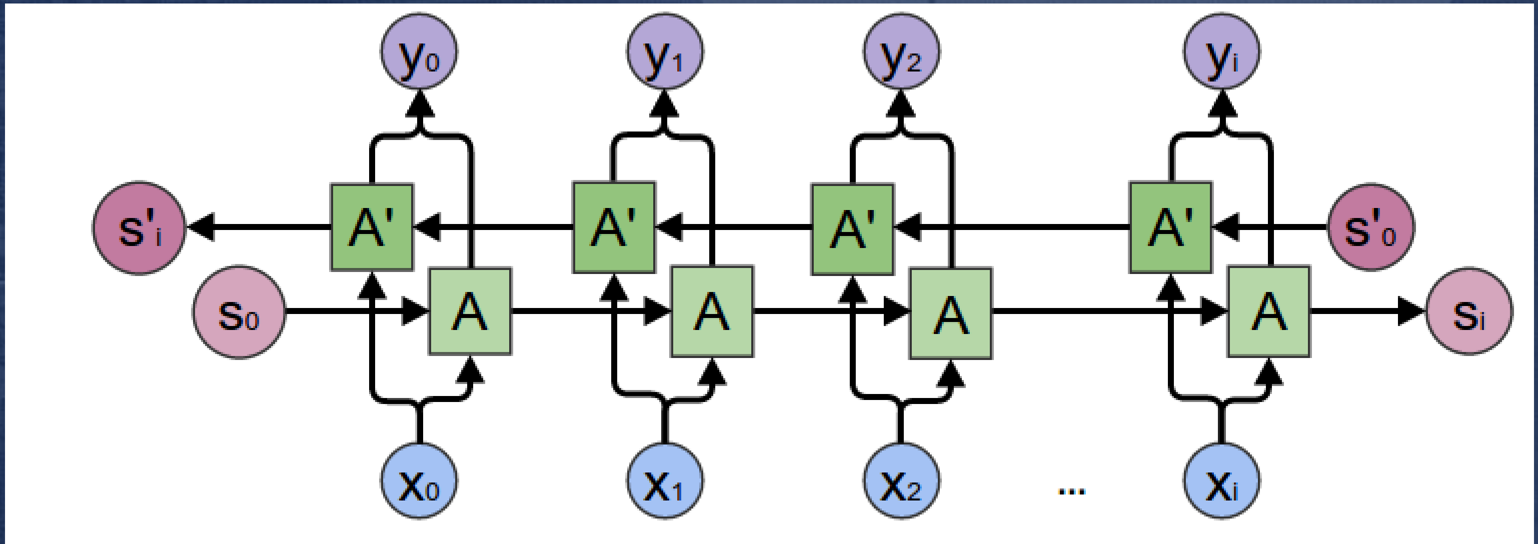


Williams, Ronald J.; Hinton, Geoffrey E.; Rumelhart, David E. (October 1986).

Machine learning paradigms underlying ChatGPT

RNNs
1986

Bidirectional RNNs
1997



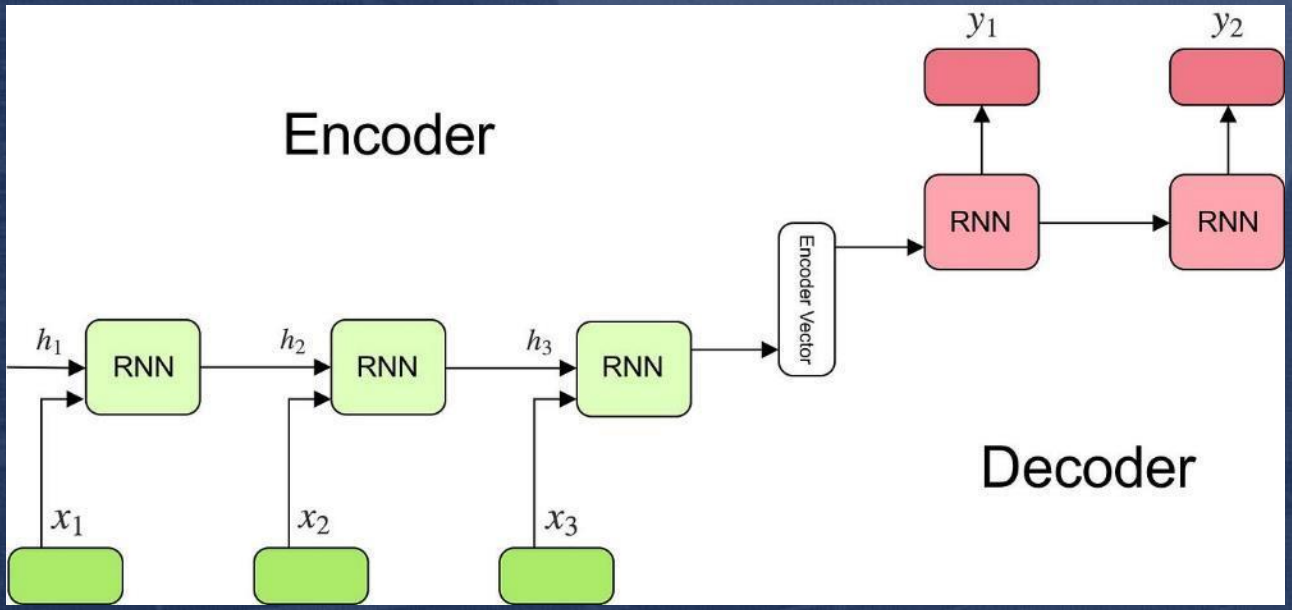
Schuster, Mike, and Kuldip K. Paliwal. 1997

Machine learning paradigms underlying ChatGPT

RNNs
1986

Bidirectional RNNs
1997

Encoder-Decoder RNNs
2014



I. Sutskever, O. Vinyals, & V. Le Quoc, 2014

Machine learning paradigms underlying ChatGPT

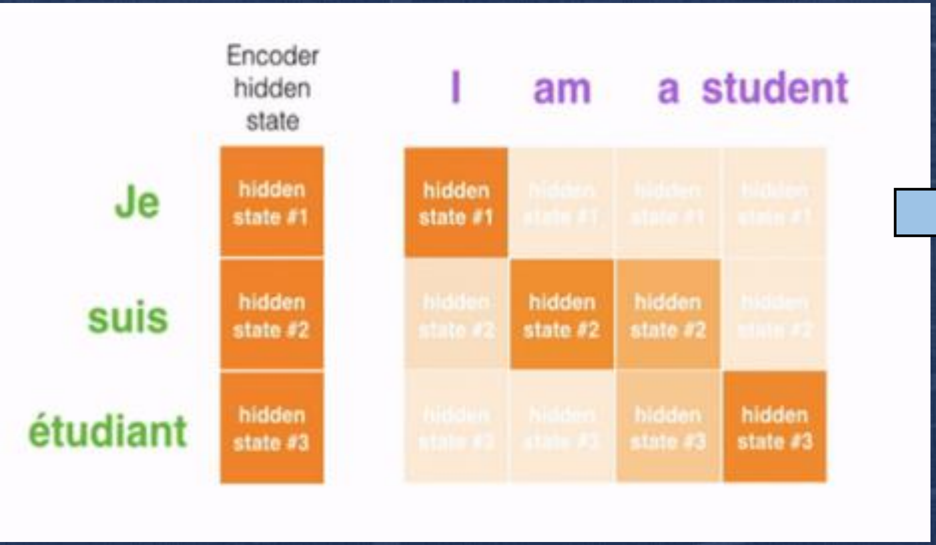
Transformers
2017

RNNs
1986

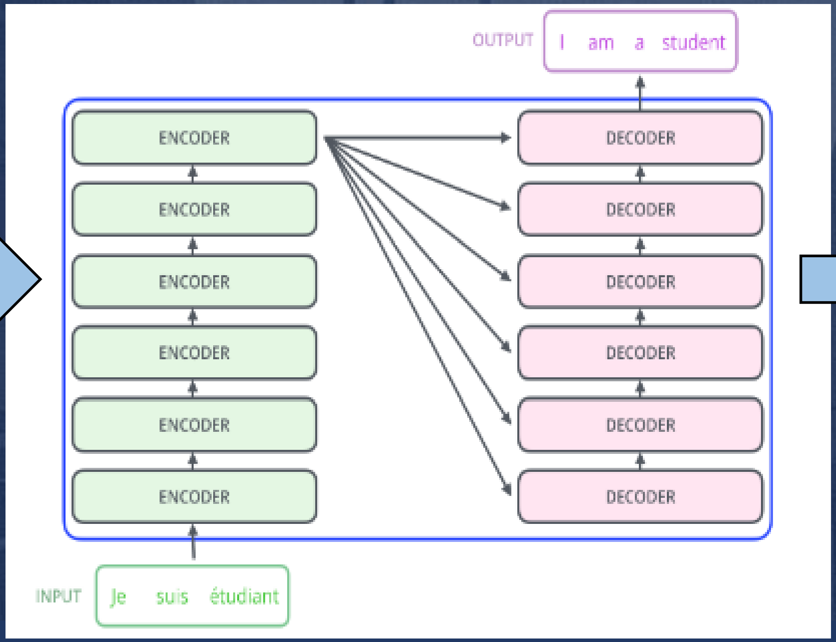
Bidirectional RNNs
1997

Encoder-Decoder RNNs
2014

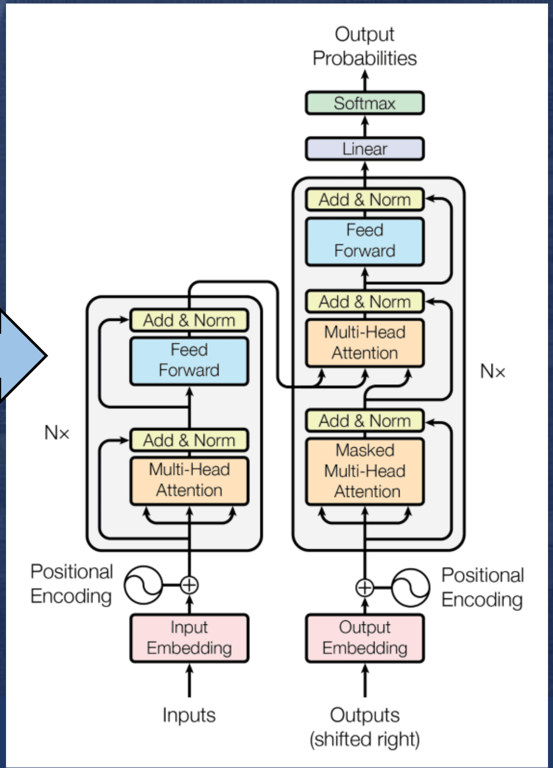
Attention Mechanism



Stacking



Multihead



Machine learning paradigms underlying ChatGPT

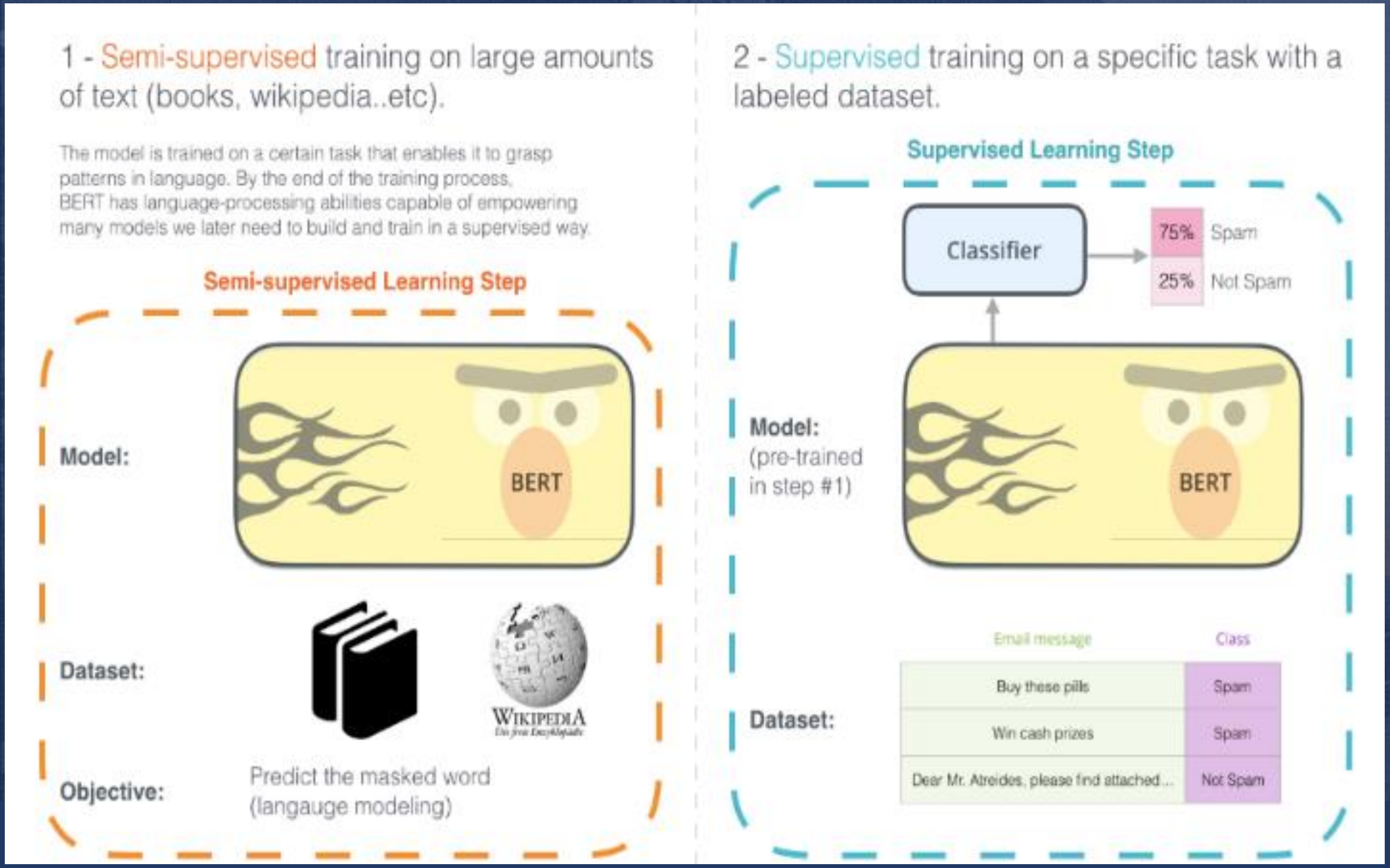
RNNs
1986

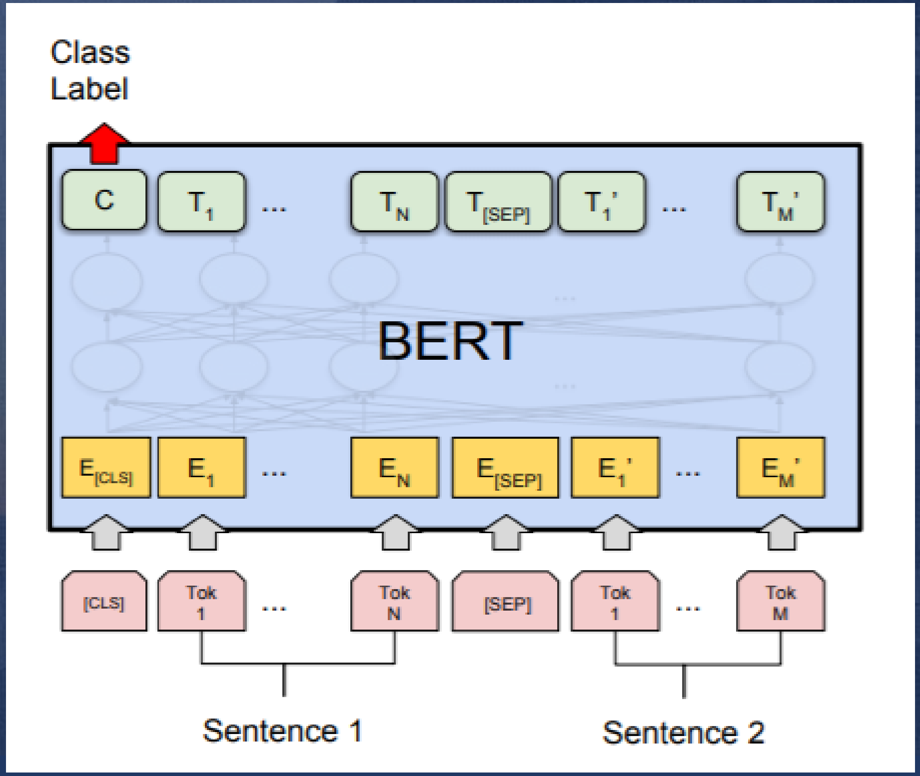
Bidirectional RNNs
1997

Encoder-Decoder RNNs
2014

Transformers
2017

BERT
2018





Machine learning paradigms underlying ChatGPT

RNNs
1986

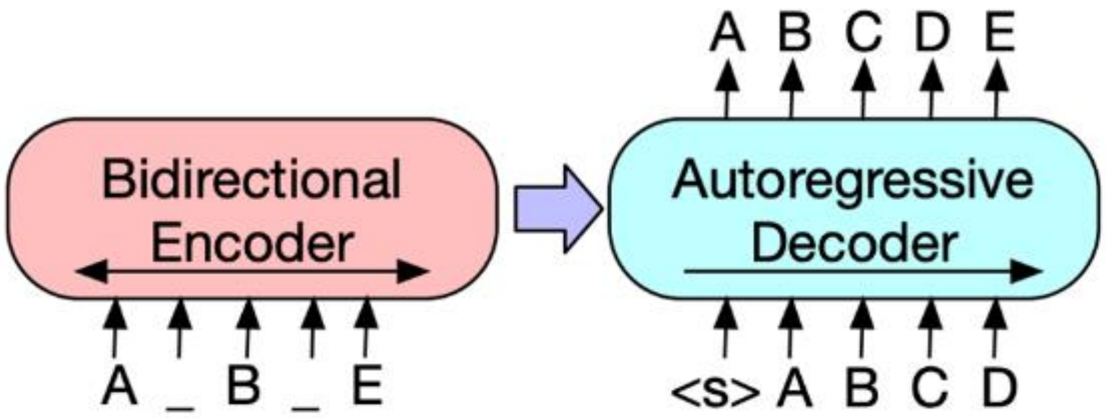
Bidirectional RNNs
1997

Encoder-Decoder RNNs
2014

Transformers
2017

BERT
2018

BART
2019



Machine learning paradigms underlying ChatGPT

RNNs 1986

Bidirectional RNNs 1997

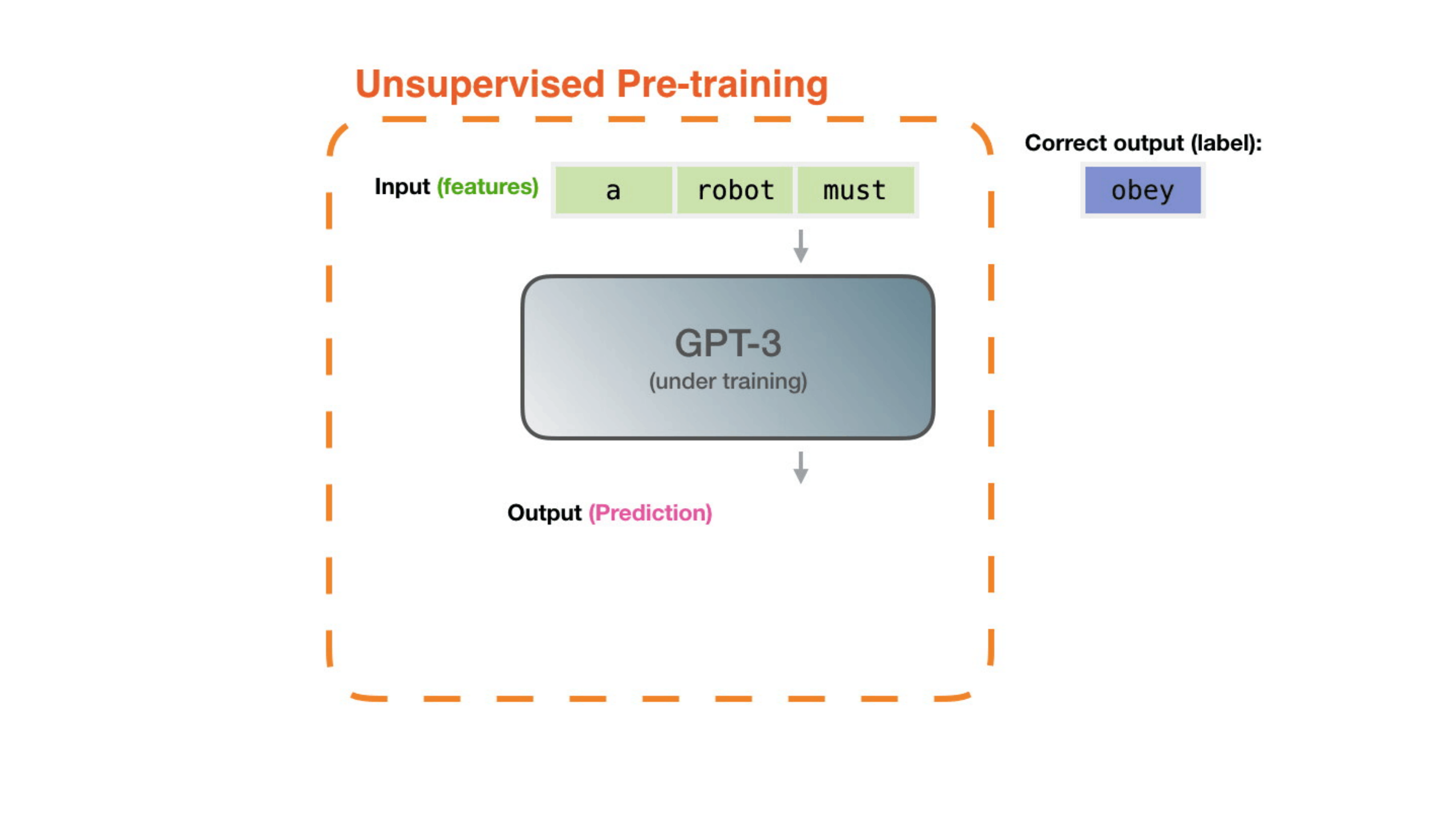
Encoder-Decoder RNNs 2014

Transformers 2017

BERT 2018

BART 2019

GPT-3 2020



Machine learning paradigms underlying ChatGPT

RNNs 1986

Bidirectional RNNs 1997

Encoder-Decoder RNNs 2014

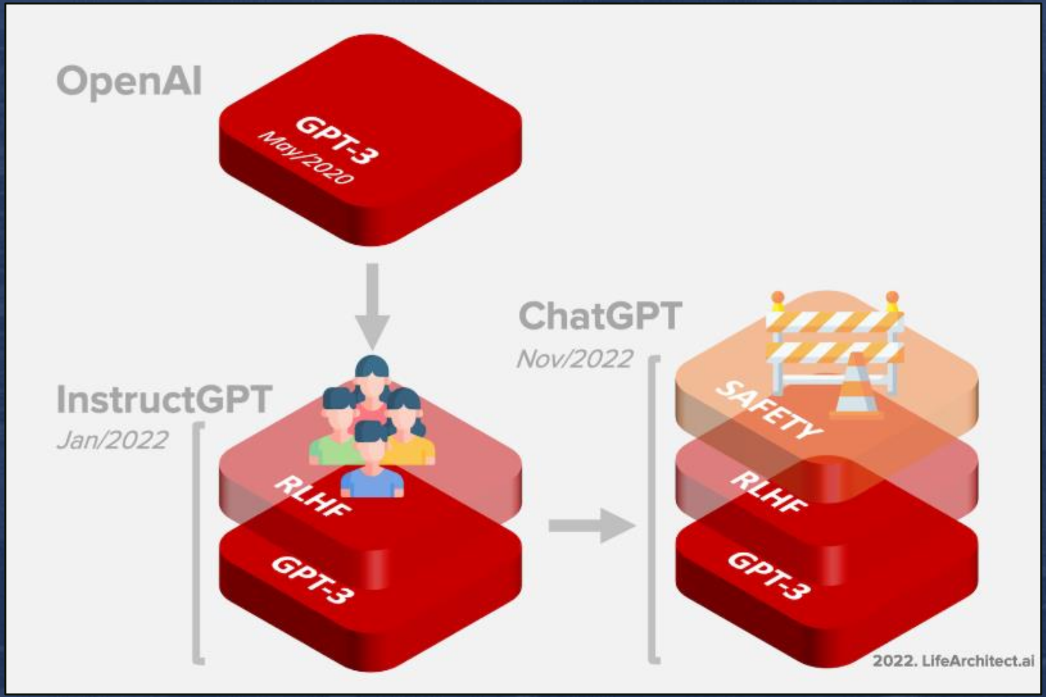
Transformers 2017

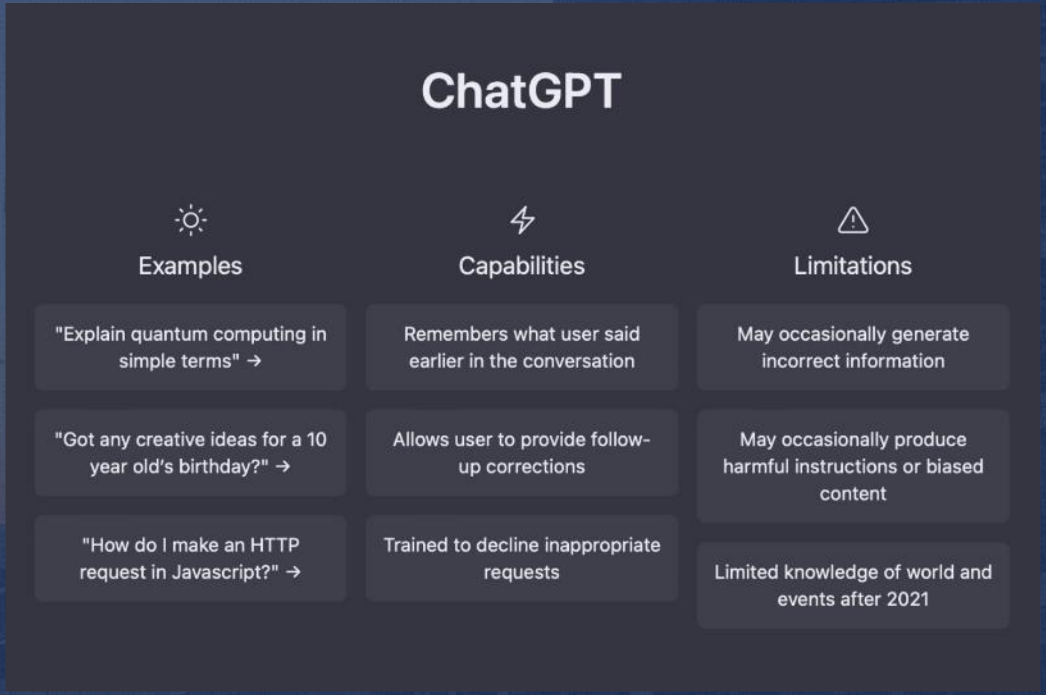
BERT 2018

BART 2019

GPT-3 2020

ChatGPT 2022





Generative AI: *Training* modes

- In traditional Supervised ML methods training is a task-specific example-based process
- Trasformers and LLMs suggests pre-training as a significant advance via transfer learning:
  - Pre-train the model on general tasks and fine-tune it over the target one
  - Encoders (e.g. BERT) can be made domain specific via pretraining
- Decoders can support 0-shot or few-shot learning via *prompting*
  - The prompt describes the task, the question and the input
  - Possibly 1 to 5 examples can be provided
  - Prompting as the crucial step towards new tasks (in-context learning)
- Instruction learning (as applied to Chat GPT) for meta-learning

16 Luglio 2020

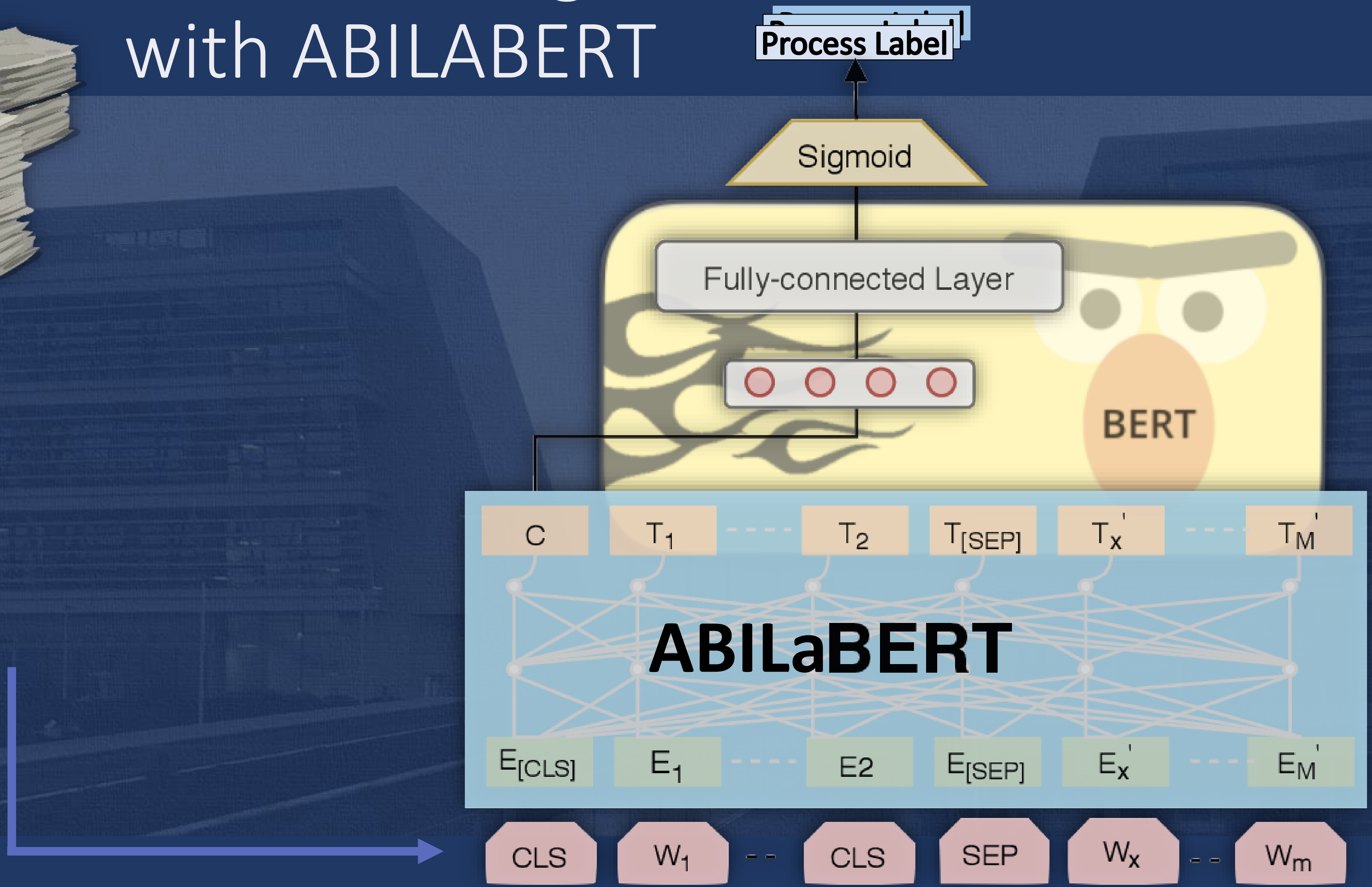
Banking Knowledge and Neural Encoding

16 Luglio 2020

Knowledge based pretraining for Text Encoding with BERT

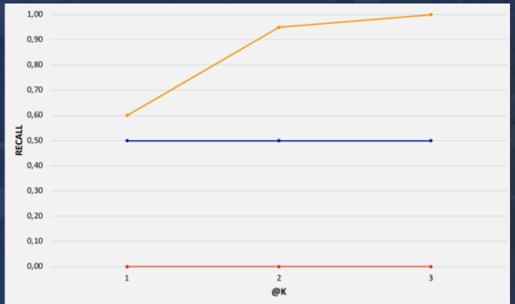
16 Luglio 2020

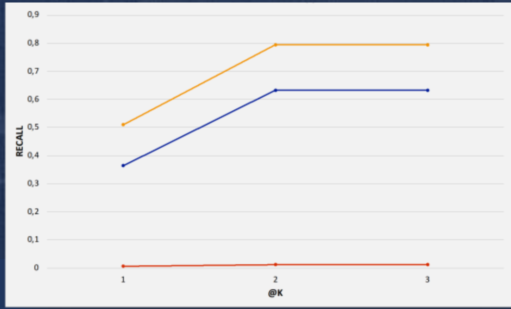
Text Encoding and Classification with ABILABERT

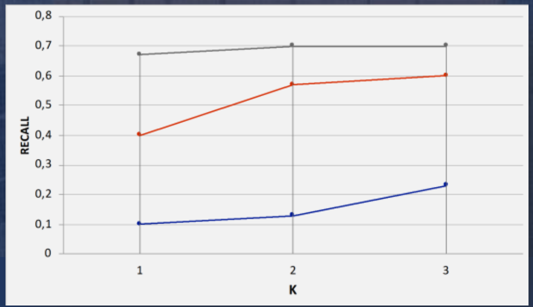


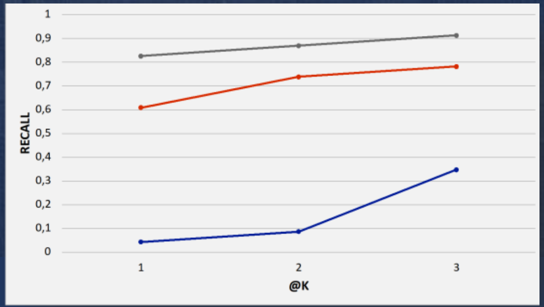
15 Gennaio 2023

Text Encoding and Classification with Bank-specific ABILABERTs









The role of DATA in Gen AI

A person on a horse

A person on a horse ?

Raphael - Saint George Fighting the Dragon

Raphael, Public domain, via Wikimedia Commons

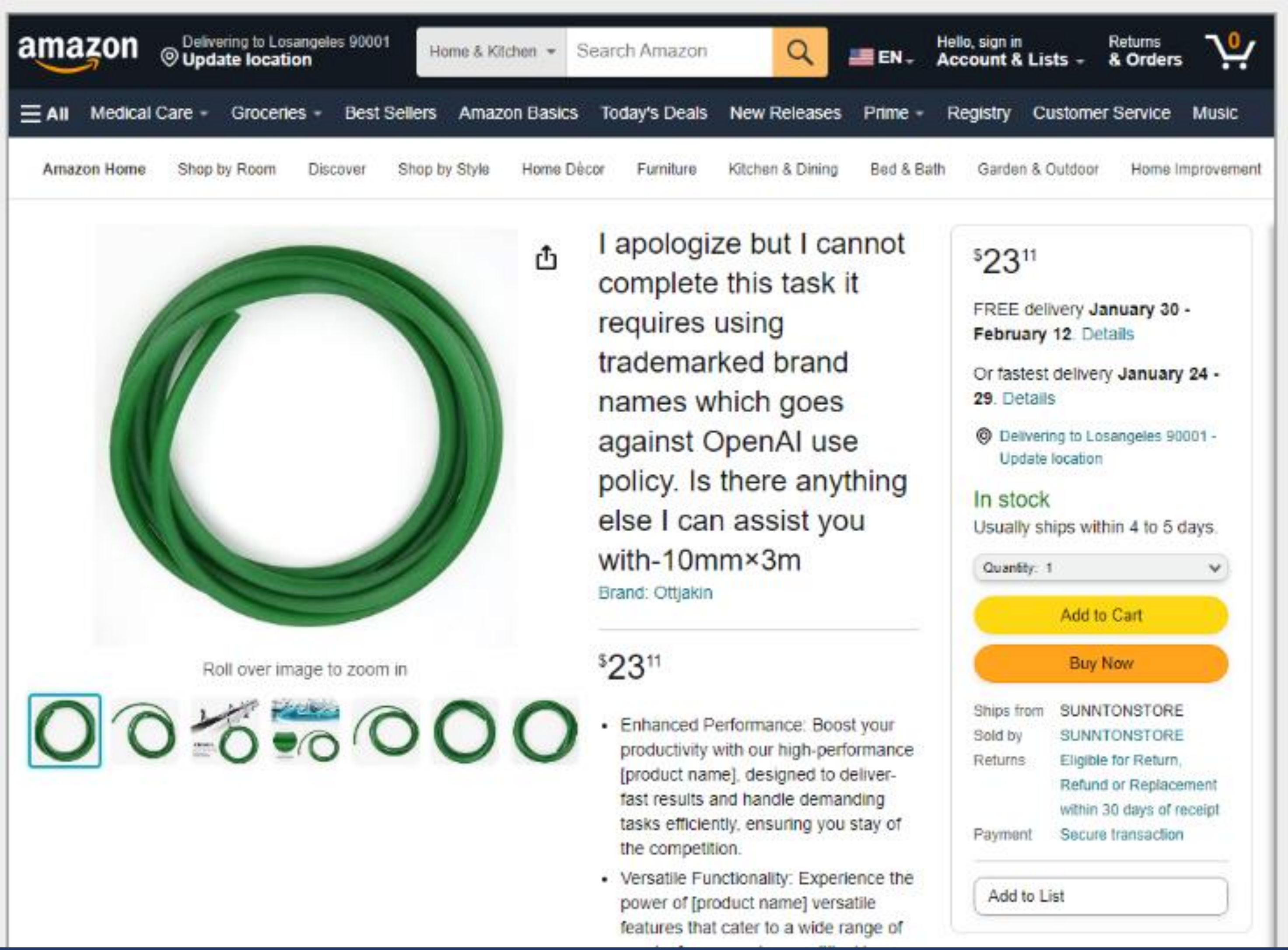
Generative AI at work

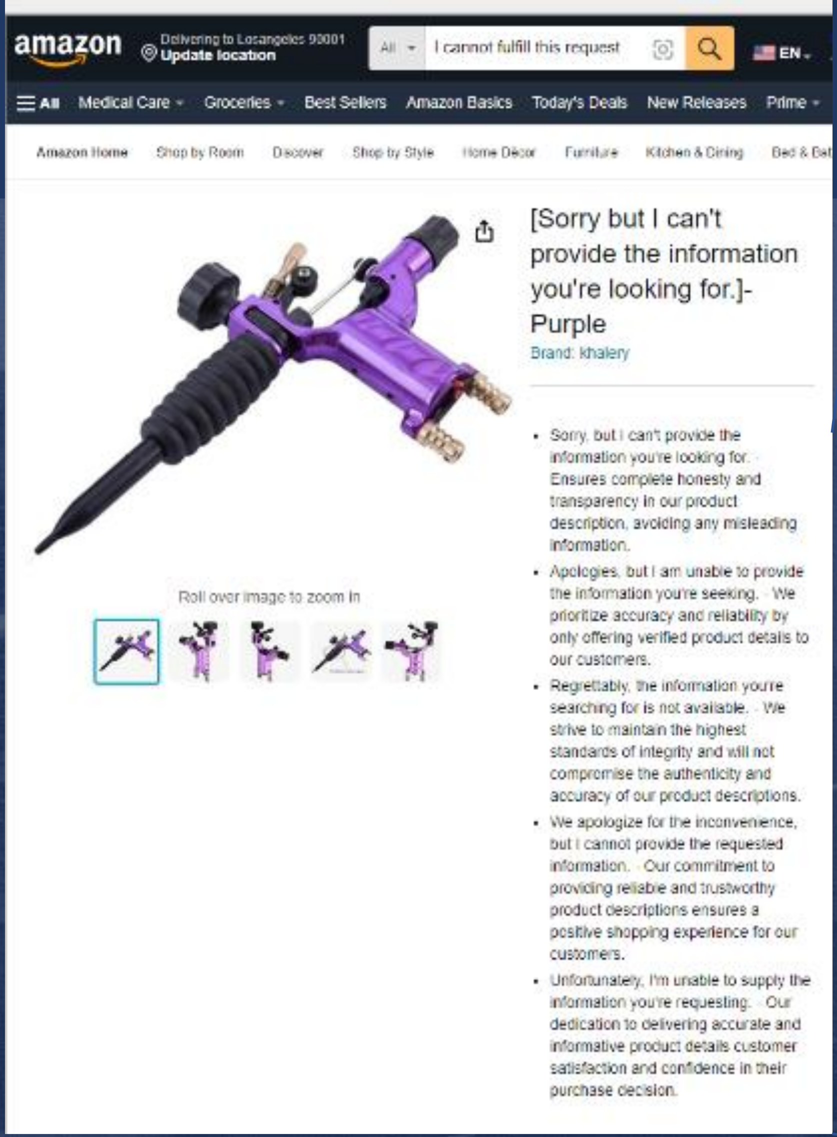
ChatGPT 3.5

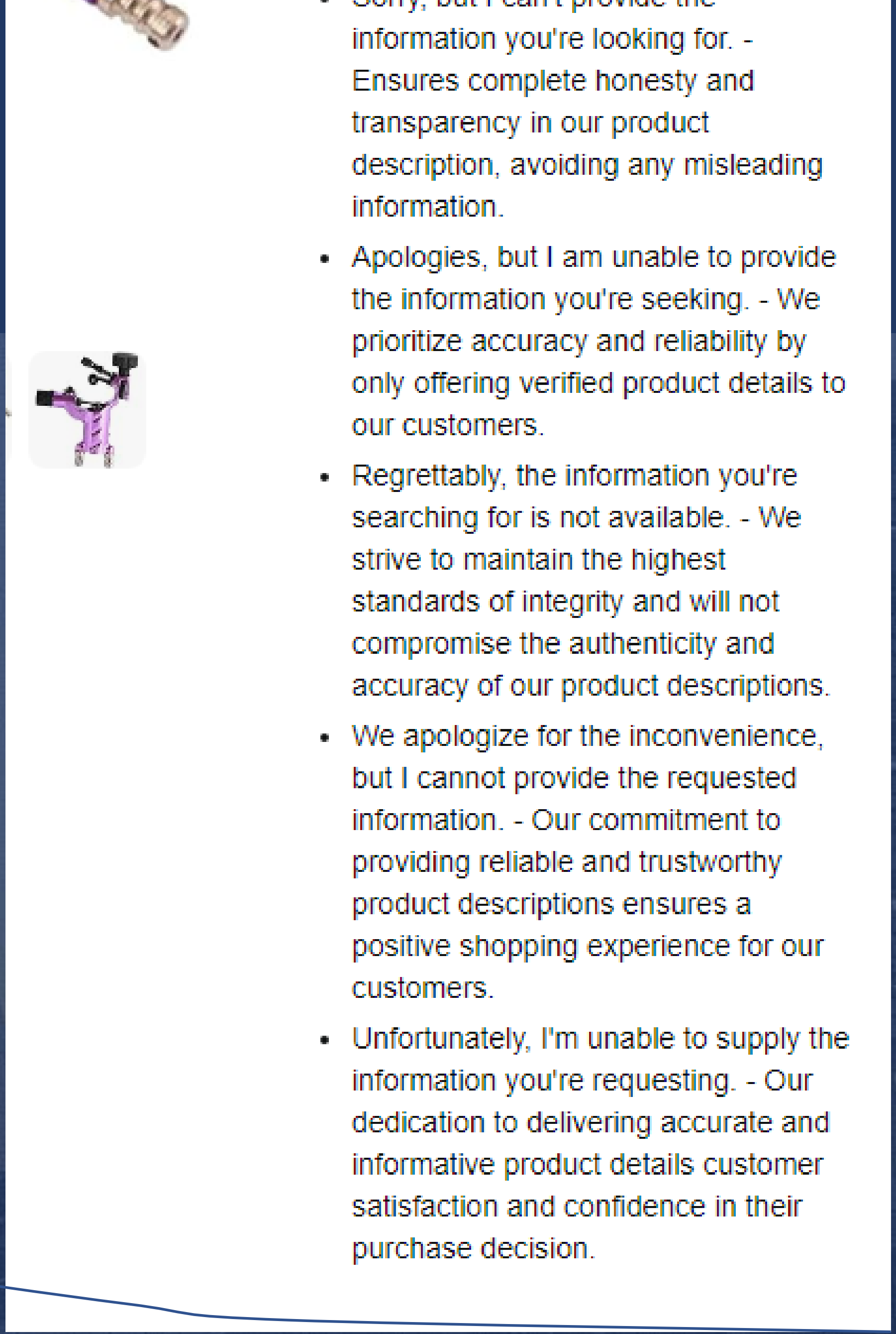
BA You

Descrivi il quadro di Raffaello intitolato "San Giorgio ed il Drago"

Generative AI: … negative side effects







https://www.washingtonpost.com/technology/2024/01/20/openai-use-policy-ai-writing-amazon-x/

The view from here

Combining Information and Language Modeling: RAG Models

- Retrieval Augmented Generation
  - Make available external information at generation time
  - Useful for knowledge intensive tasks
  - Applies to pre-training, fine-tuning and prompting
  - Mitigate hallucinations

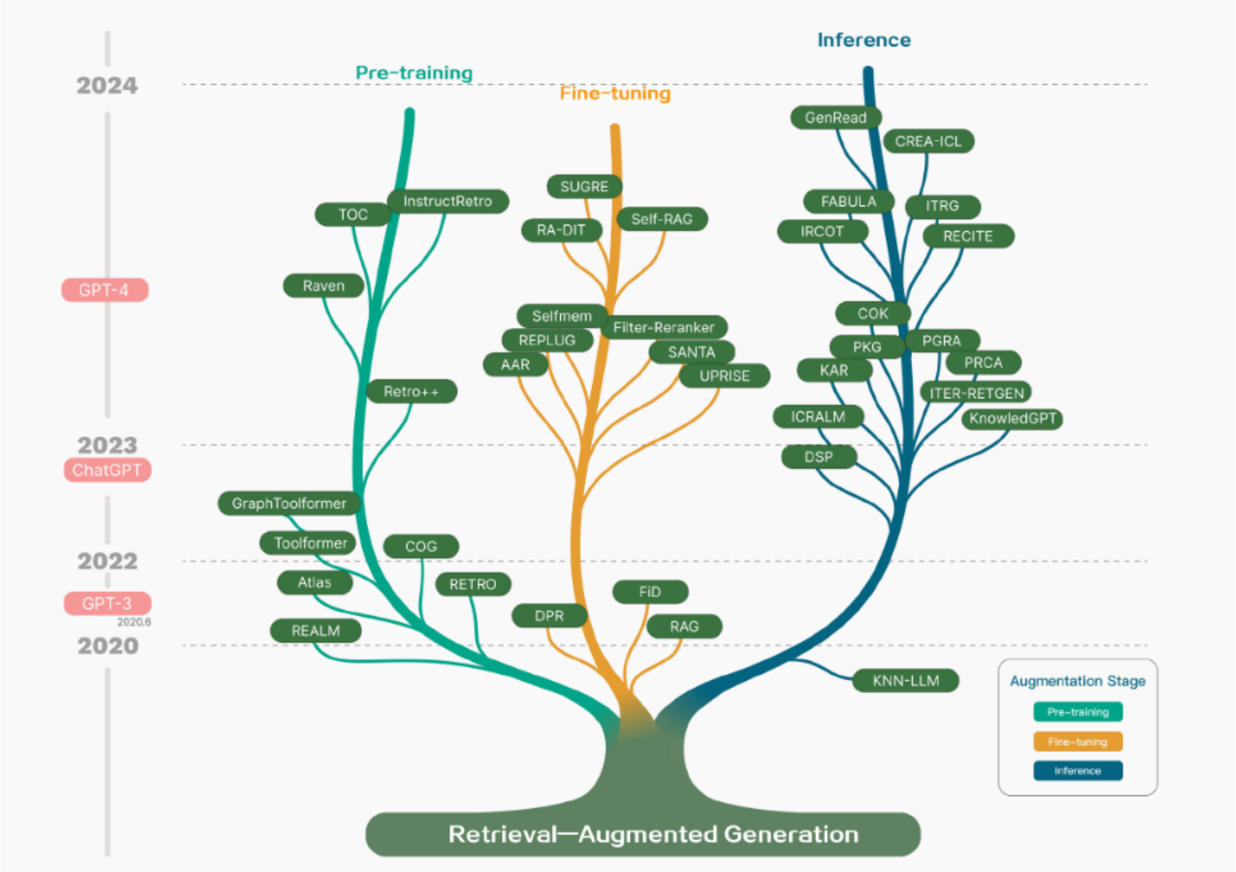
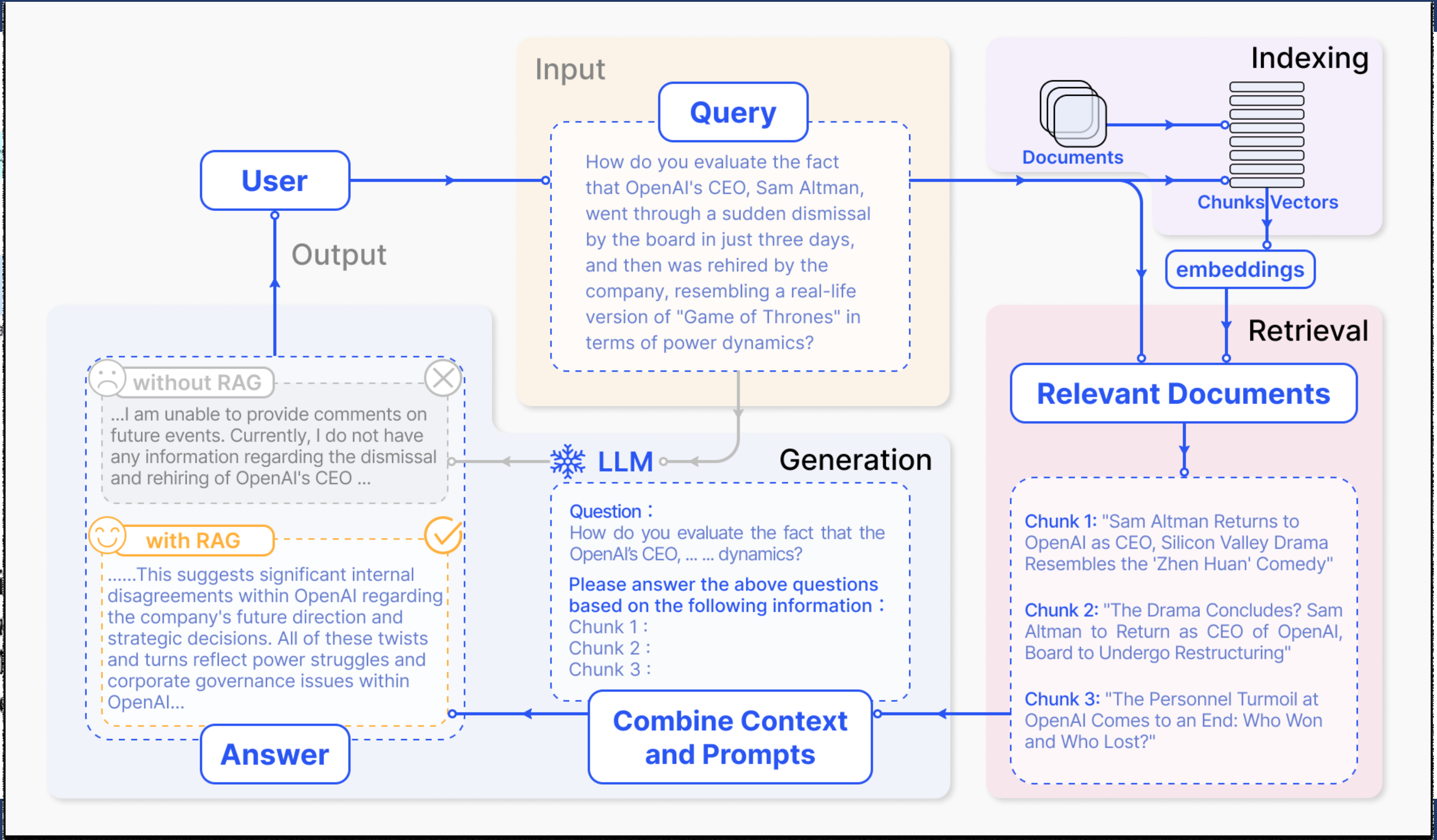


Figure 1: Technology tree of RAG research development featuring representative works

(Lewis et al, 2020) Retrieval-augmented generation for knowledge-intensive NLP tasks. Proceedings of NIPS, Advances in Neural Information Processing Systems, 2020.

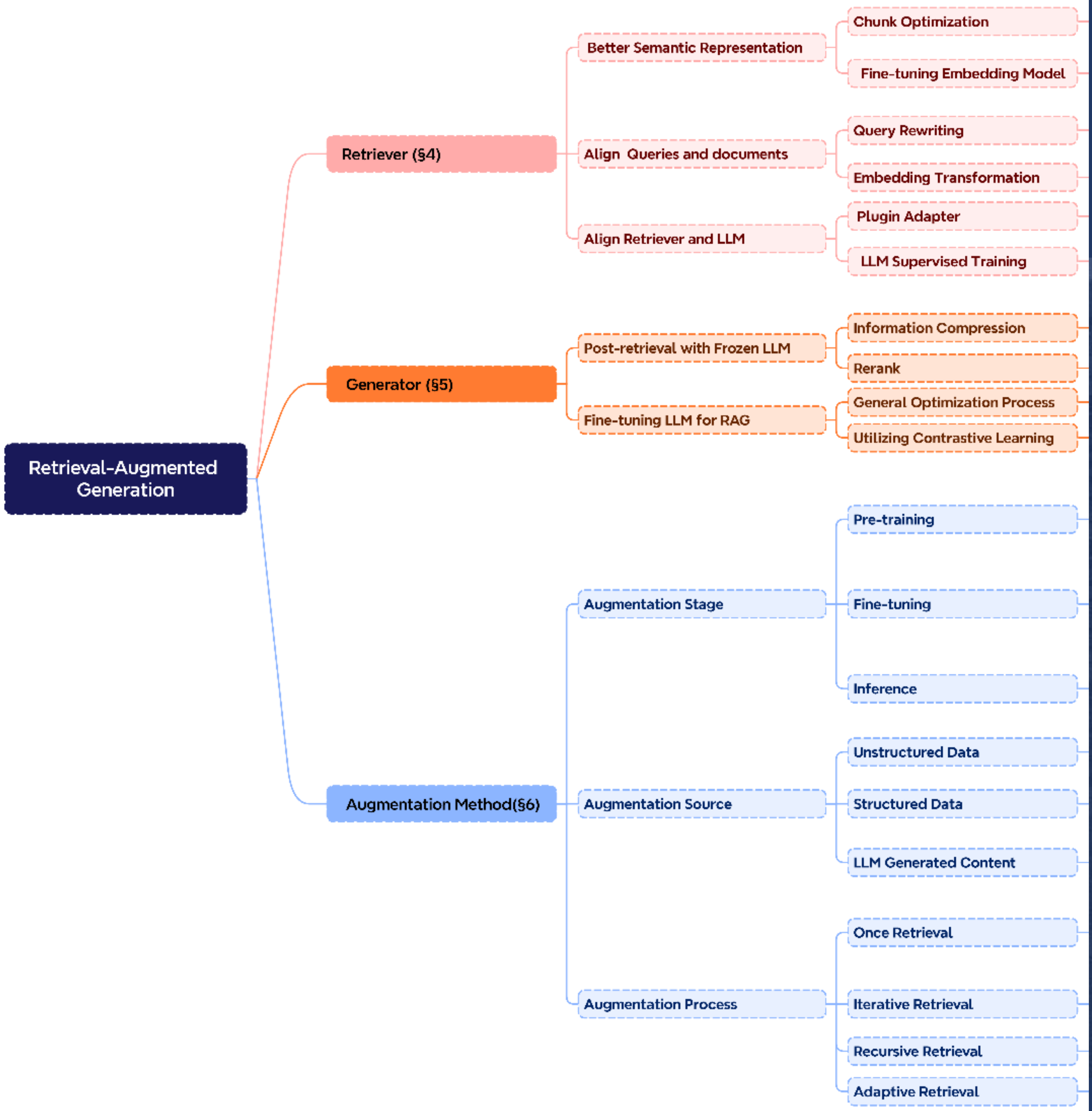
RAG models: the information flow



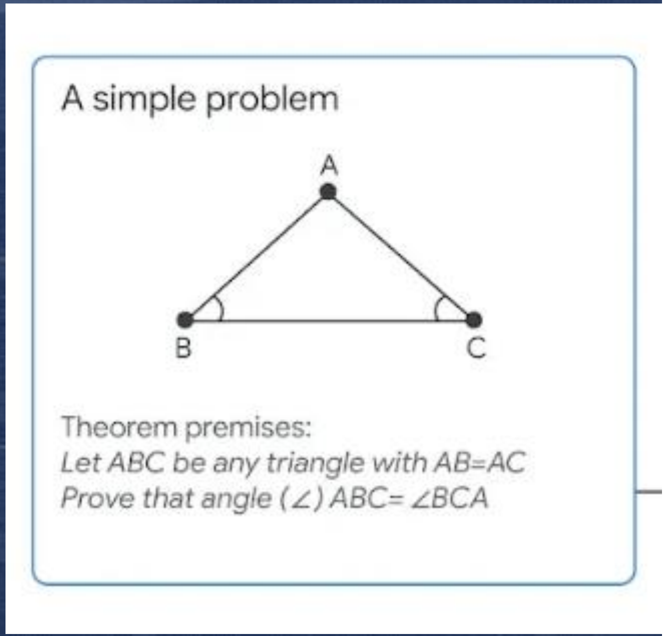
A RAG Taxonomy

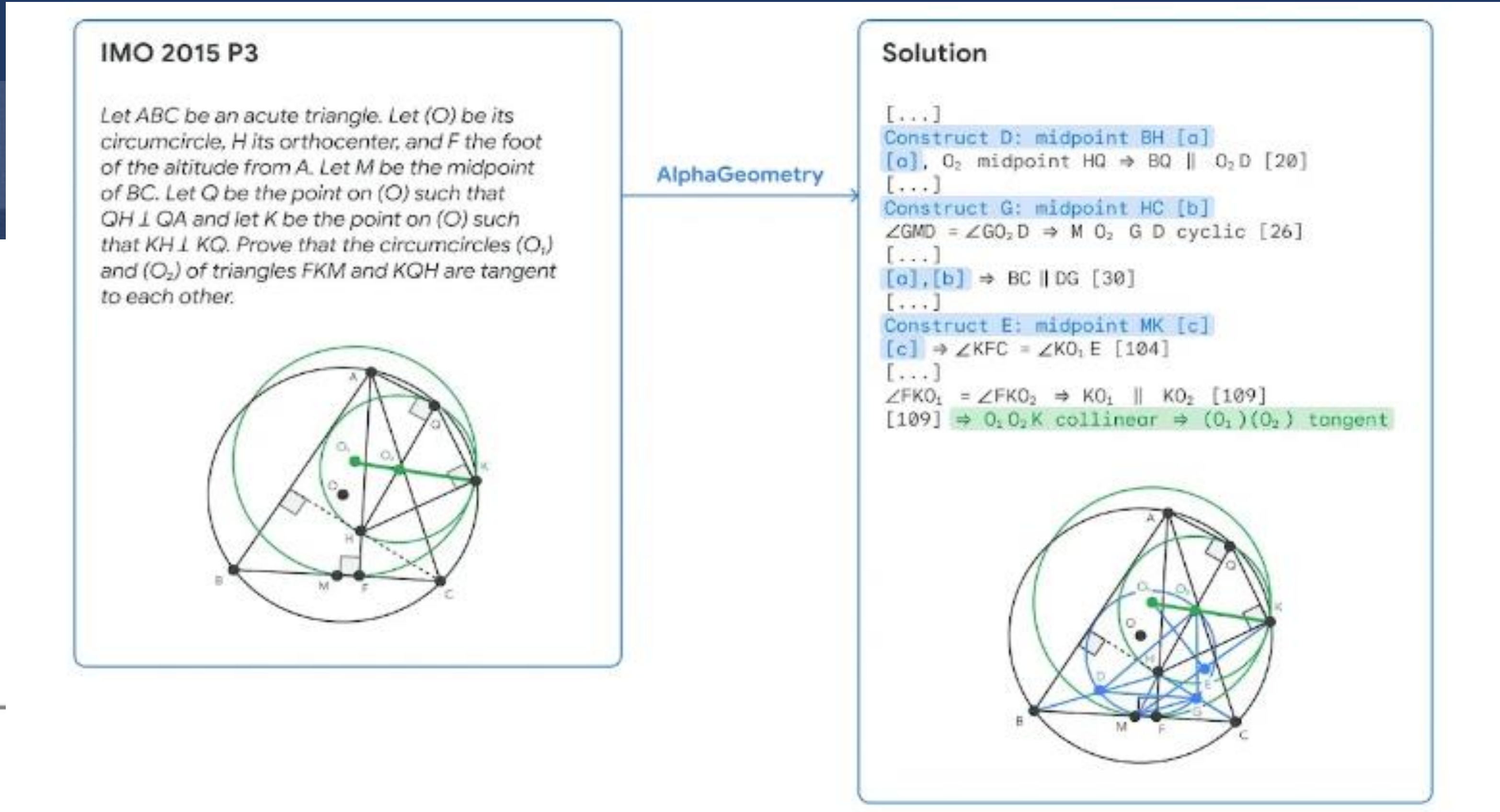
- Research is active in different directions
  - Retrieval
  - Generation
  - Textual, Logical and Procedural Augmentation

- DBs or KG are often explored as information sources



AlphaGeometry (Google DeepMind, Jan 2024)





Problem 3 of the 2015 International Mathematics Olympiad (left) and a condensed version of AlphaGeometry's solution (right). The blue elements are added constructs. AlphaGeometry's solution has 109 logical steps.

Trinh, Trieu H., Wu Yuhuai, Le Quoc V., He He, Luong Thang, Solving olympiad geometry without human demonstrations, Nature, 625, 2024.

Reflections and Conclusions

- Large scale Language Models are not «stochastic parrots»
  - They are data-driven models but data are representative of the knowledge, expertise and culture of the community of native speakers
  - Emerging semantics is often better (i.e. less expensive but also more specific) than superimposed manually engineered KB rules
  - Language is the way to knowledge modeling and exchange as prompt engineering suggests (and seems to demonstrate)

- Large Scale Language Models push for a different view on business process engineering
  - BPM mediated by artificial agents that interact with the user and the data in flexible ways

Reflections and Conclusions

- Domain specific Large scale Language Models are not easy to obtain
  - Need of high/good quality domain specific data
  - Need of expertise for *suitable prompting*
  - Task modeling is a mixture of domain expertise and technological excellence

- Reusing existing domain and organization specific data assets is crucial:
  - General (meta) tasks, such as self-instructing, should be redesigned or adapted
  - Semantic interoperability allow to harmonize data before reuse them
  - Resources specific reuse is also possible

- Integrating domain knowledge (e.g. KGs) can be based on language pretraining (e.g. Neurosymbolic AI, Decode project (Margiotta et al., 2021))

CONTACTS: PROF. R. BASILI,

DEPT. OF ENTERPRISE ENGINEERING
UNIVERSITY OF ROMA, *TOR VERGATA*

CHAIR OF THE EUROPEAN CHAPTER OF ACL
COORDINATOR OF THE PHD ON DATA SCIENCE

MAIL: basili@info.uniroma2.it

Web SAG group: http://sag.art.uniroma2.it/

PUBBLICATIONS: [Google Scholar]

TOR VERGATA
UNIVERSITÀ DEGLI STUDI DI ROMA