Phil_AI Lab

PHILOSOPHY OF LANGUAGE
LINGUISTICS SEMIOTICS
ARTIFICIAL INTELLIGENCE

La Babele dei Golem

Conversazioni su linguaggi umani e IA

Mercoledì 13 novembre h 15.30
«Sapienza» Università di Roma
Dipartimento di Filosofia, Villa Mirafiori, Aula seminari, I piano

DANILO CROCE

Tenure Track Researcher
Department of Enterprise Engineering
«Tor Vergata» University of Rome

Hallucinations in LLMs

This talk provides an exploration of the nature, causes, and mitigation of hallucinations in state-of-the-art language models. Starting with an introduction to the foundations of LLMs, it emphasizes the transformative role of technologies like transformers and pre-training in language understanding and generation.

The main focus is on understanding the "hallucination" phenomenon, where models produce factually incorrect or contextually inappropriate outputs. We classify and dissect different types of hallucinations, exploring both factual inaccuracies and faithfulness issues in responses. We then address the causes of these hallucinations, spanning data limitations, training misalignments, and inference challenges.

The talk also summarizes current strategies for detecting hallucinations, alongside methods for mitigating them through improved data handling or refined training techniques.

Finally, we discuss open challenges and questions surrounding the reliability, knowledge boundaries, and ethical responsibilities of deploying LLMs.

This presentation aims to foster an understanding of LLM limitations and engage participants in deeper reflection on the implications of these powerful AI systems in shaping human knowledge and ethics.

Introducono e moderano

LAB SIL

FILOMENA DIODATO
filomena.diodato@uniroma1.it

FRANCO CUTUGNO
cutugno@unina.it

Urban/Eco

Phil_AI Lab

Comitato scientifico: Federico Albano Leoni, Marina De Palo, Franco Cutugno, Filomena Diodato
Comitato organizzativo: LabSil Sapienza: Flaminia Carocci, Claudia Cicerchia, Sara Dellino, Edoardo Moré, Federica Ruggiero, Nicola Sighinolfi, Francesco Verde – Urban/ECO: Loredana Schettino, Norman Vitale

zoom https://uniroma1.zoom.us/j/88081003648?pwd=fK3obDsDhM2w4iuEGnmBXUITRZ7wLg.1

Sullo sfondo Éidola#40 gentilmente concessa da Damiano Errico