ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Distributed Computing

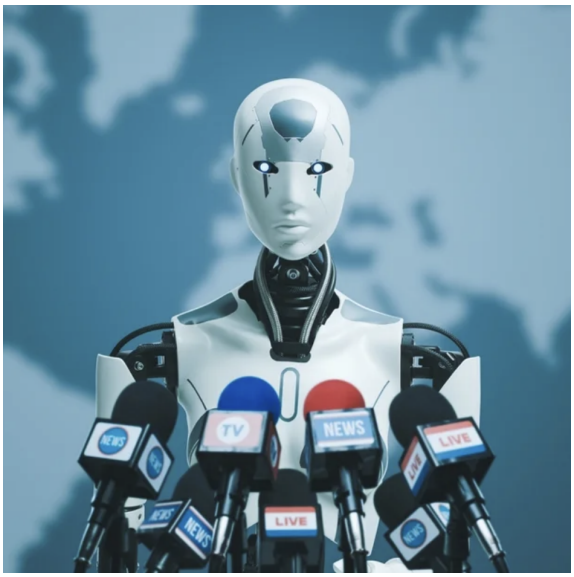
Prof. R. Wattenhofer

Evaluating and Controlling the Political Bias of LLMs

Large Language Models (LLMs) are rapidly integrating into our daily lives. While they are generally safe for routine tasks such as trip planning, coding, and general knowledge seeking, their use in critical applications like social decision-making processes (e.g., drafting fair policies) has been limited due to a lack of transparency regarding their true objectives. In this project, you will explore one of the many open problems in this vast area. Some possible research questions include:

- Do LLMs exhibit a "political opinion" (i.e., consistent bias)? How can it be measured?
- Is it possible to control this political opinion?
- Can LLMs be fine-tuned to accurately reflect the political opinion of a given individual?
- Any other question you might propose...

Depending on your interests, you will develop empirical tools and/or a theoretical framework to evaluate the political bias of existing (and potentially future) LLMs. You are encouraged to bring your own ideas to address this problem. We will hold weekly meetings to monitor the progress of the project, culminating in a presentation of your findings to the group.



Requirements Strong motivation, ability to work independently, and an interest in conducting innovative theoretical and/or empirical research. A solid background in mathematics and artificial intelligence (ideally in NLP). Good programming skills, in particular with python. An interest in politics is a significant plus.

Interested? Please get in touch with us for more details!

Contact

- Frédéric Berdoz: fberdoz@ethz.ch, ETZ G60.1