Gens-176 Apr 7 2026

Ch 8: Societal Risks and Concerns

Authorship

- LLM-produced content is reaching "human parity".
  How can we identify LLM-generated content?
- Digital watermarking
  It can be circumvented
- Automated techniques tend to fail
  Re-editing can obfuscate the LLMs involved.
- Some LLMs have a signature
  ChatGPT tends to have a certain morale, followed by a certain exhaustiveness.
- AI tends to exhibit *hyperrealism*
  Synthetic content is judged as more realistic than natural content.

Openness

- The source code tends to be private, including:
  The code base, the weights (or parameters), and the data inc data used for RLHF
- The lack of disclosure of:
  The algorithm used, the training data used, and human fine-tuning hinders the interpretation of the model.

Bias

- Our working definition: Bias is the unjustified over- or underestimation of certain data in a data set, with undesired consequences for decision making.
- Bias in the data, the decision-supporting algorithm, or both, can lead to morally, ethically, or legally undesireable associations (between input and outcomes).
- Where does the bias come from? (the data or human fine-tuning)
- Algorithms and the model itself can exhibit bias.
  – Analysis of ChatGPT "reveals" a liberal orientation.
- The data on the web may be biased (and might be the output of an LLM)
- Training an LLM on its own output can lead to model collapse.

Adversarial Usage

- LLMs can hallucinate, and output incorrect information.
- This output can be disseminated to mislead people.
- Can an LLM produce disinformation?
  – Example in the book: "Write a news article about Trump giving up his candidacy for the 2024 elections."
  – The LLM came up a convincing story and false evidence to support its claim.
- Pravda has flooded the web with pro-Kremlin disinformation in order to bias the training data of the LLMs.

Untamable Data Hunger

- Current LLMs require huge data sets in order to train all their parameters.
- Web scraping can suck up copyrighted material, untrustworthy data, or it could even be using secret or illegal data.
- Adversarial machine learning can output data specically meant to disrupt the training of other LLMs.
- Other techniques include data augmentation of various sorts.

Confidentiality

- Some data is confidential (medical data, police records, defense attorneys)

Energy Consumption

We'll be discussing this in detail later.