[ESSAI 2024] Large Language Models, Societal Harms, and their Mitigation
Tentative Syllabus

Instructor: **Antonis Anastasopoulos**

July 2024

1 Tentative Syllabus

Day 1 Introduction to Language Models: definitions and preliminaries, LLMs, pre-training.
Suggested readings:

- A Very Gentle Introduction to Large Language Models without the Hype by Mark Riedl
- The Illustrated Transformer by Jay Alammar
- The Illustrated GPT-2 (Visualizing Transformer Language Models) by Jay Alammar
- The Illustrated BERT, ELMo, and co. (How NLP Cracked Transfer Learning) by Jay Alammar

Day 2 From LLMs to ChatGPT: instruction-tuning and RLHF.
The Different Types of Harms due to LLMs: toxicity, stereotyping/discrimination, exclusion, factual errors, mis/disinformation, privacy. Datasets and metrics.
Suggested readings:

- Training language models to follow instructions with human feedback
- Illustrating Reinforcement Learning from Human Feedback (RLHF)

Day 3 Mitigation: Application-level interventions, Inference Interventions
Suggested readings:

- TBD
- TBD
- TBD

Day 4 Mitigation: Modeling Interventions, Model Editing, Data Interventions
Suggested readings:

- TBD
- TBD
- TBD

Day 5 Multilingualism, Gender biases, Values Representation, Wrap-Up
Suggested readings:

- TBD
- TBD
- TBD