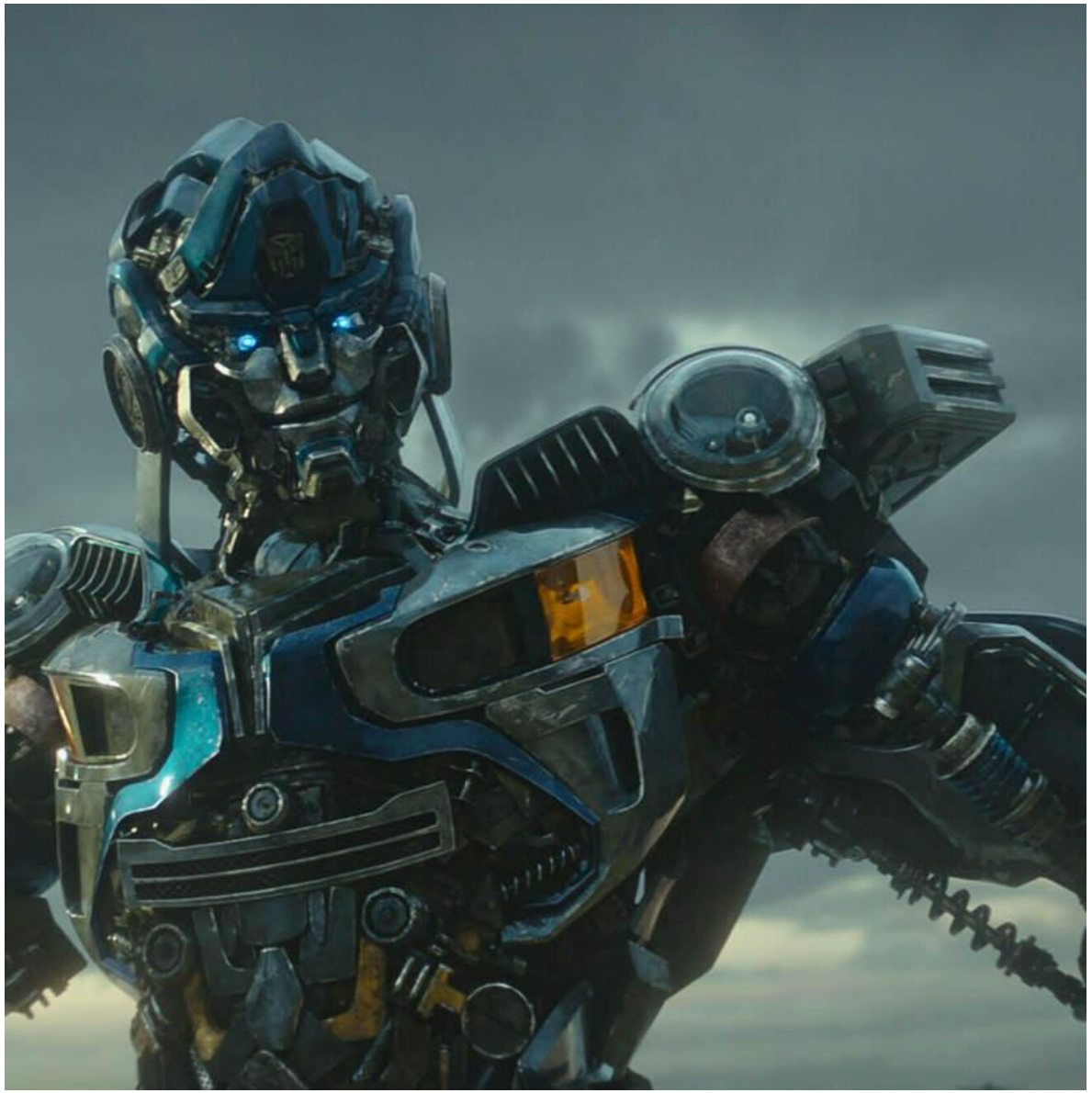
transformers and beyond

GESIS Fall Seminar 2023

“From Embeddings to Transformers: Advanced Text Analysis in Python”

[day 5, GPT & ethics]



paper:[2203.05794] BERTopic: Neural topic modeling with a class-based TF-IDF procedure (arxiv.org)

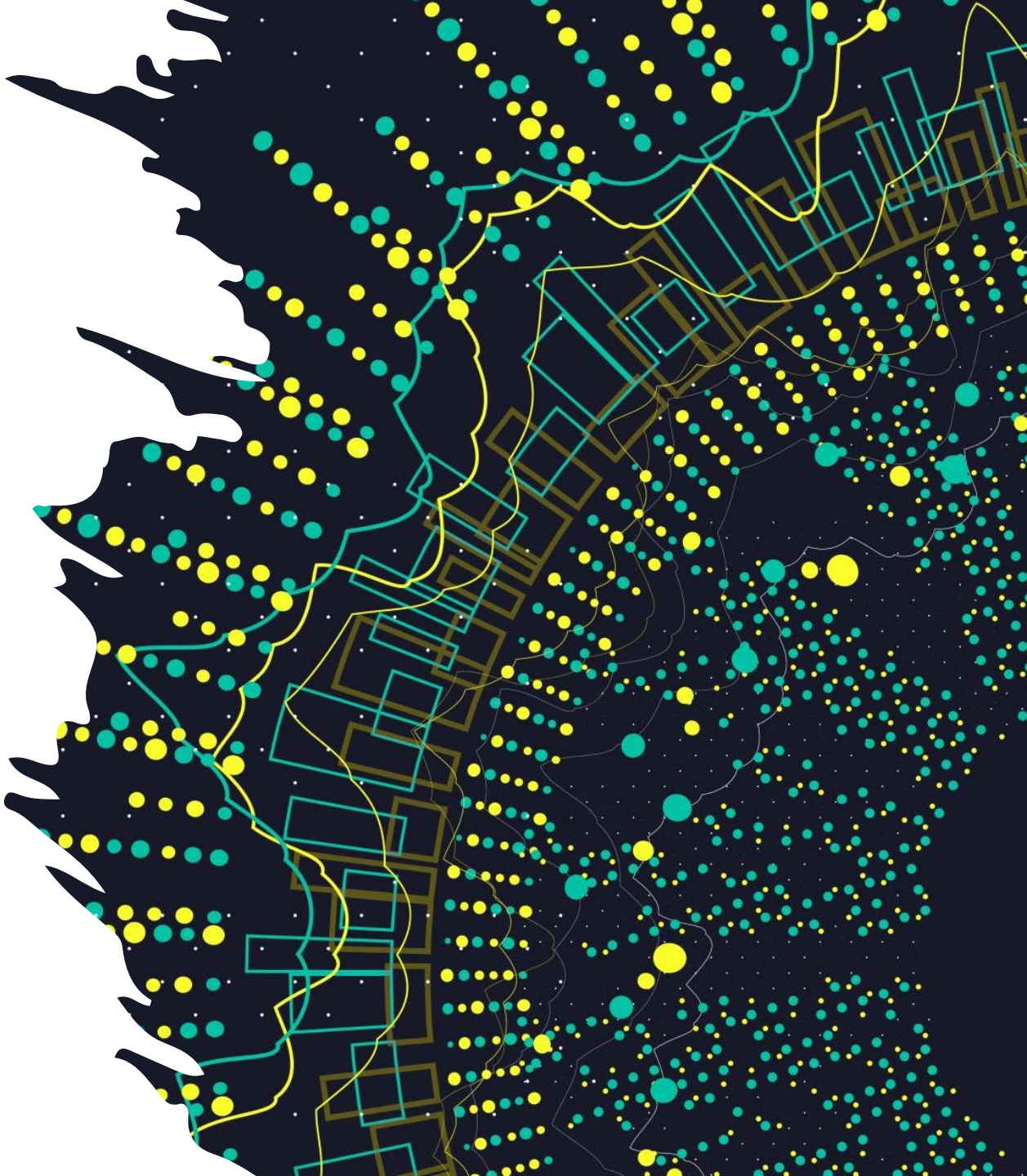
BERTopic github:MaartenGr/BERTopic: Leveraging BERT and c-TF-IDF to create easily interpretable topics. (github.com)

more variants: Quick Start - BERTopic (maartengr.github.io)

notebook: BERTopic.ipynb - Colaboratory (google.com)

comparison BERTopic to LDA: Comparison of LDA vs BERTopic (hashnode.dev)

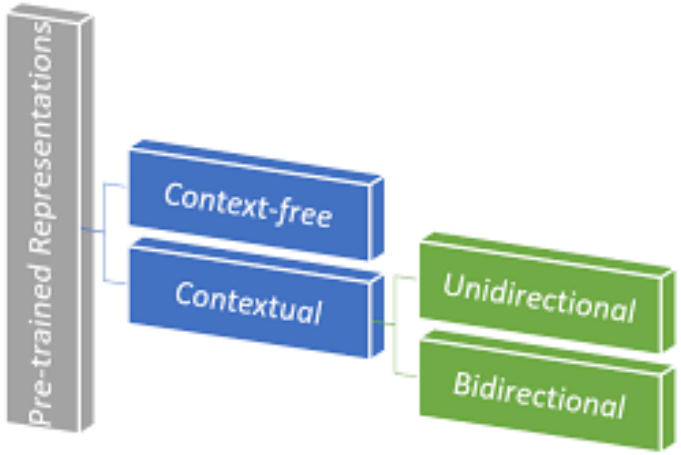
BERTopic



GPT

Generative Pre-Trained Transformer

bidirectional vs unidirectional



GPT = unidirectional:

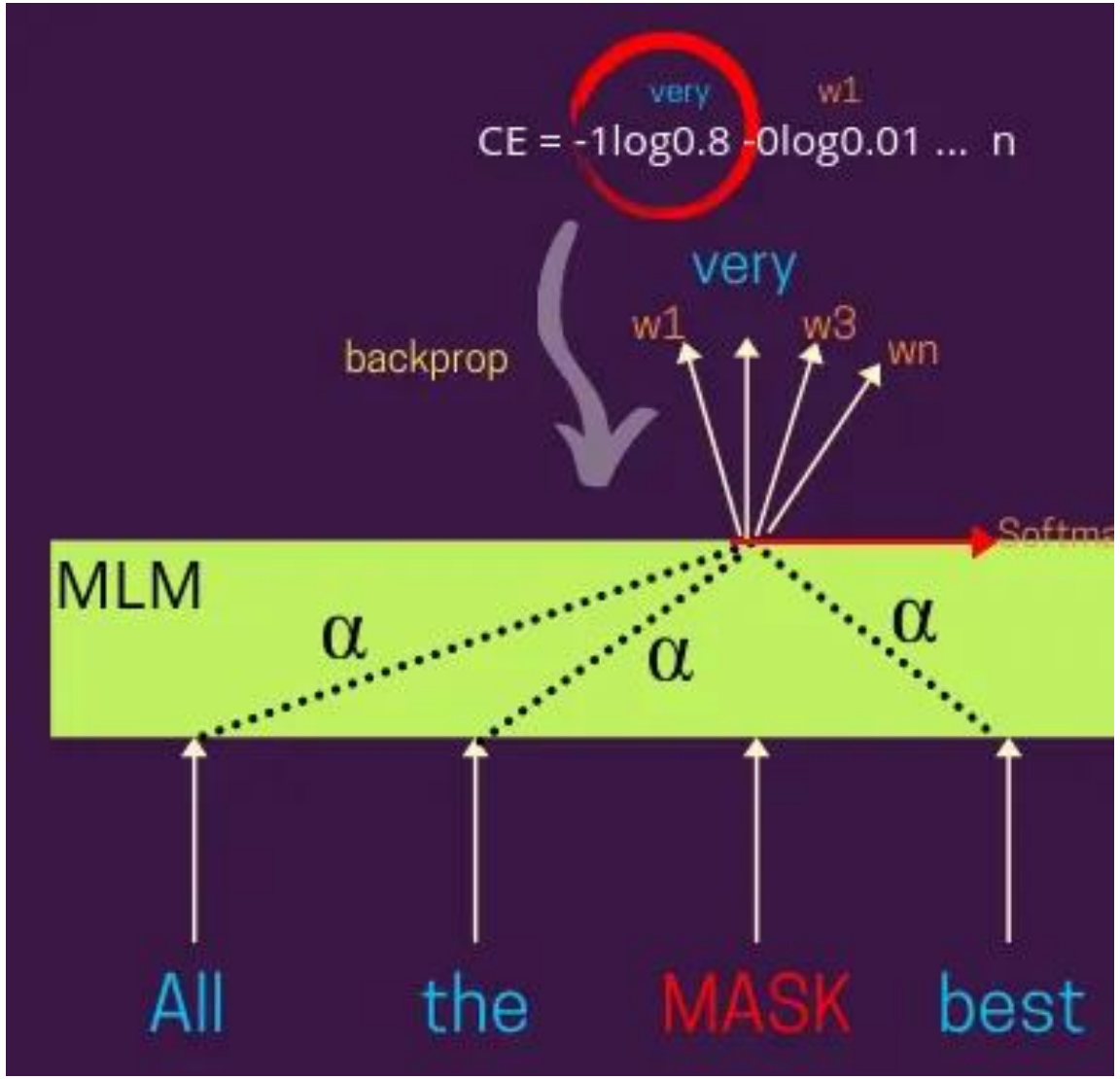
autoregressive language modeling

BERT = bidirectional:

masked language modeling

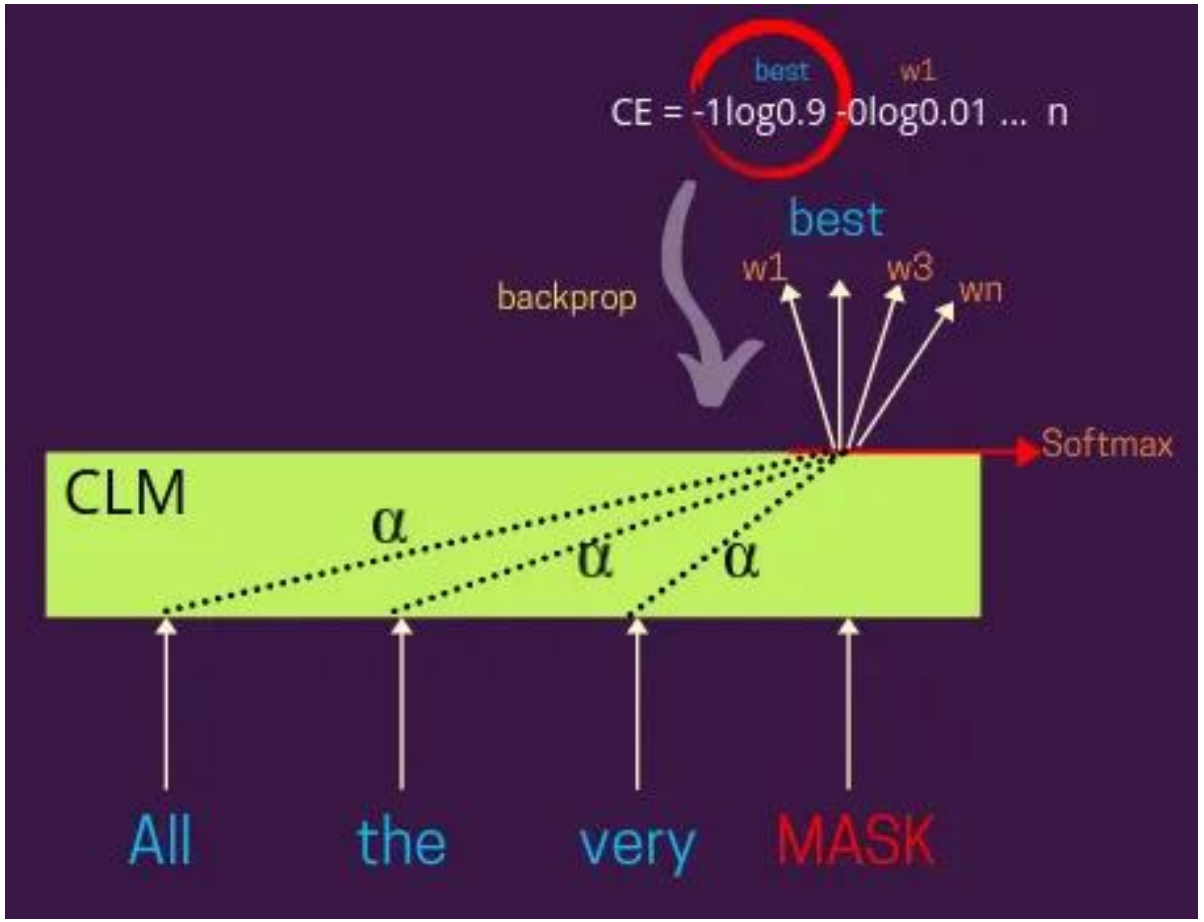
BERT

- [Understanding Masked Language Models (MLM) and Causal Language Models (CLM) in NLP | by Prakhar Mishra | Towards Data Science]



GPT

- [Understanding Masked Language Models (MLM) and Causal Language Models (CLM) in NLP | by Prakhar Mishra | Towards Data Science]



What is GPT good for?

- Text generation
- Text completion
- Translation
- Summarization
- Question answering

- Sentiment analysis
- Classification tasks

ChatGPT or API

Fine-tuning? Well, no: Prompting!

(the future is interactive)

Zero-Shot Learning:

Zero-shot learning refers to a scenario in which a model performs a task without any specific training data or examples for that task. Instead, it relies on a textual description or context provided to understand and perform the task. Zero-shot learning demonstrates the model's ability to generalize from textual descriptions.

Example: Asking a model to translate a sentence from English to French without prior translation training data.

Fine-tuning? Well, no: Prompting!

(the future is interactive)

Few-Shot Learning:

Few-shot learning involves training a model with a very small amount of task-specific data, often just a few examples or examples for a few classes. The model learns to perform the task by leveraging this limited training data.

Example: Training a sentiment analysis model with only a few positive and negative reviews for each sentiment class.

Fine-tuning? Well, no: Prompting!

(the future is interactive)

One-Shot Learning:

One-shot learning is an extreme form of few-shot learning where the model is trained with just one example per class. It aims to recognize or perform tasks with minimal training data.

Example: Training a facial recognition model to recognize a person's face with only one image of that person.

playground or chat interface

Playground - OpenAI API

ChatGPT (openai.com)

so what are good prompts?

…and if it doesn’t do what I want?

ChatGPT outperforms crowd workers for text-annotation tasks (Gilardi et al. 2023)

ChatGPT outperforms crowd workers for text-annotation tasks | PNAS

Many NLP applications require manual text annotations for a variety of tasks, notably to train classifiers or evaluate the performance of unsupervised models. Depending on the size and degree of complexity, the tasks may be conducted by crowd workers on platforms such as MTurk as well as trained annotators, such as research assistants. Using four samples of tweets and news articles (n = 6,183), we show that ChatGPT outperforms crowd workers for several annotation tasks, including relevance, stance, topics, and frame detection. Across the four datasets, the zero-shot accuracy of ChatGPT exceeds that of crowd workers by about 25 percentage points on average, while ChatGPT's intercoder agreement exceeds that of both crowd workers and trained annotators for all tasks. Moreover, the per-annotation cost of ChatGPT is less than $0.003—about thirty times cheaper than MTurk. These results demonstrate the potential of large language models to drastically increase the efficiency of text classification.

Automated Interviewer or Augmented Survey? Collecting Social Data with Large Language Models (Cuevas et al. 2023, submitted to CHI)

[2309.10187] Automated Interviewer or Augmented Survey? Collecting Social Data with Large Language Models (arxiv.org)

Qualitative methods like interviews produce richer data in comparison with quantitative surveys, but are difficult to scale. Switching from web-based questionnaires to interactive chatbots offers a compromise, improving user engagement and response quality. Uptake remains limited, however, because of differences in users' expectations versus the capabilities of natural language processing methods. In this study, we evaluate the potential of large language models (LLMs) to support an information elicitation chatbot that narrows this "gulf of expectations" (Luger & Sellen 2016). We conduct a user study in which participants (N = 399) were randomly assigned to interact with a rule-based chatbot versus one of two LLM-augmented chatbots. We observe limited evidence of differences in user engagement or response richness between conditions. However, the addition of LLM-based dynamic probing skills produces significant improvements in both quantitative and qualitative measures of user experience, consistent with a narrowing of the expectations gulf.

resources

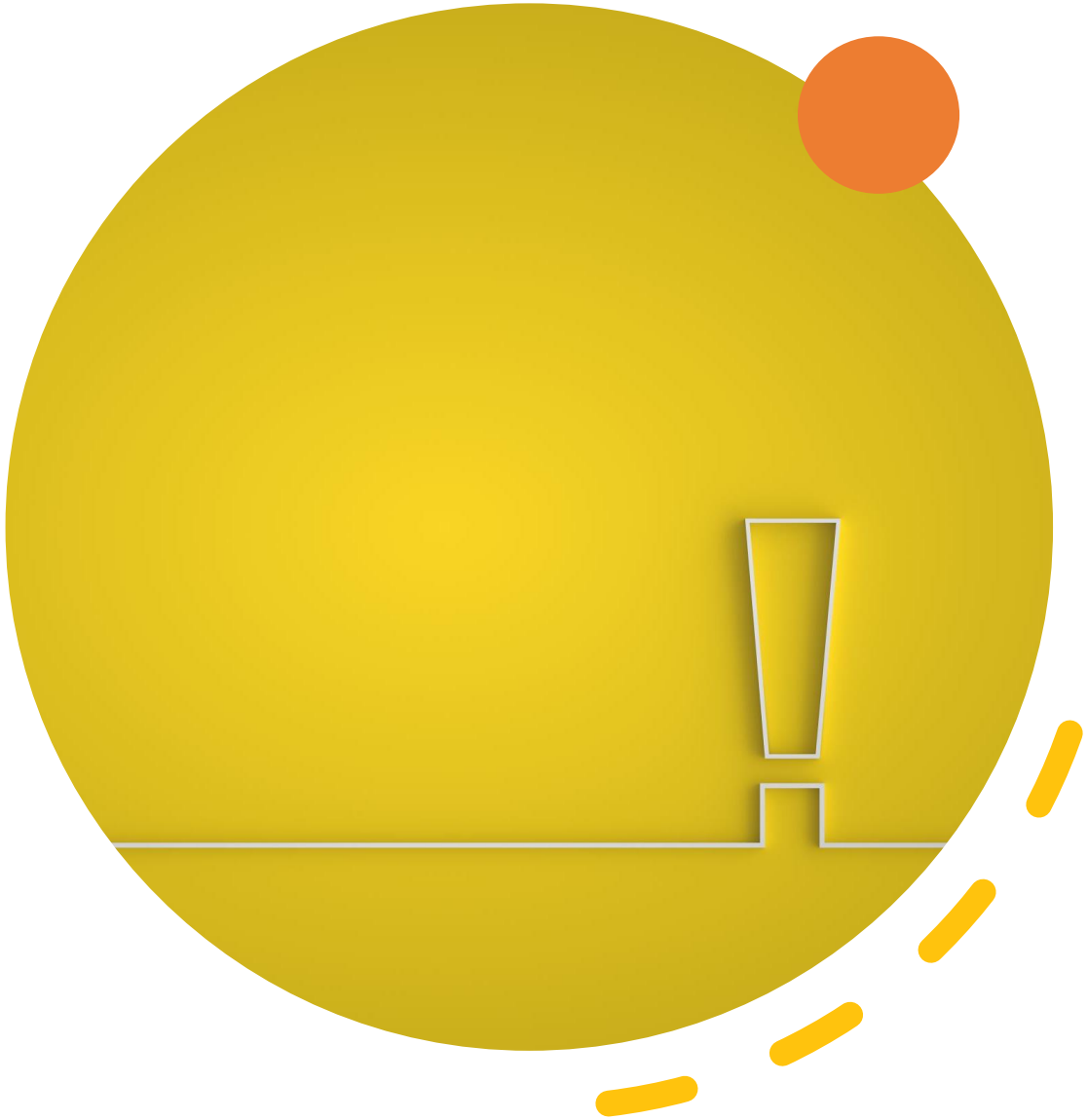
Short Courses | Learn Generative AI from DeepLearning.AI (Andrew Ng)

Prompt Engineering Guide | Prompt Engineering Guide (promptingguide.ai)

f/awesome-chatgpt-prompts: This repo includes ChatGPT prompt curation to use ChatGPT better. (github.com)

Introduction - OpenAI API

Quickstart tutorial - OpenAI API



Ethics in AI

ethics in AI – common concepts

responsible AI

AI alignment

AI safety

Responsible AI: Responsible AI is a broader concept that encompasses the ethical, legal, and moral considerations associated with the development, deployment, and use of artificial intelligence technologies. It involves designing AI systems that are accountable, transparent, fair, and respectful of human rights and values. Responsible AI also entails addressing issues related to bias, privacy, safety, and the societal impact of AI technologies. It aims to ensure that AI is developed and used in ways that benefit society as a whole while minimizing risks and harms.

AI Alignment: AI alignment refers to the process of ensuring that artificial intelligence systems, particularly advanced and autonomous AI, act in ways that are beneficial and aligned with human values and goals. It addresses the challenge of making AI systems understand and pursue the intended objectives while avoiding undesirable or harmful outcomes. The goal of AI alignment is to narrow the gap between the objectives of AI systems and the values and intentions of their human creators.

AI Safety: AI safety is a related concept that focuses on developing AI systems that are safe to use and operate, with an emphasis on avoiding catastrophic or unintended consequences. It includes techniques and research aimed at preventing AI systems from causing harm to humans, whether intentionally or accidentally. AI safety encompasses various aspects, such as robustness, security, and ensuring that AI systems do not exhibit harmful behavior, even in unforeseen situations.

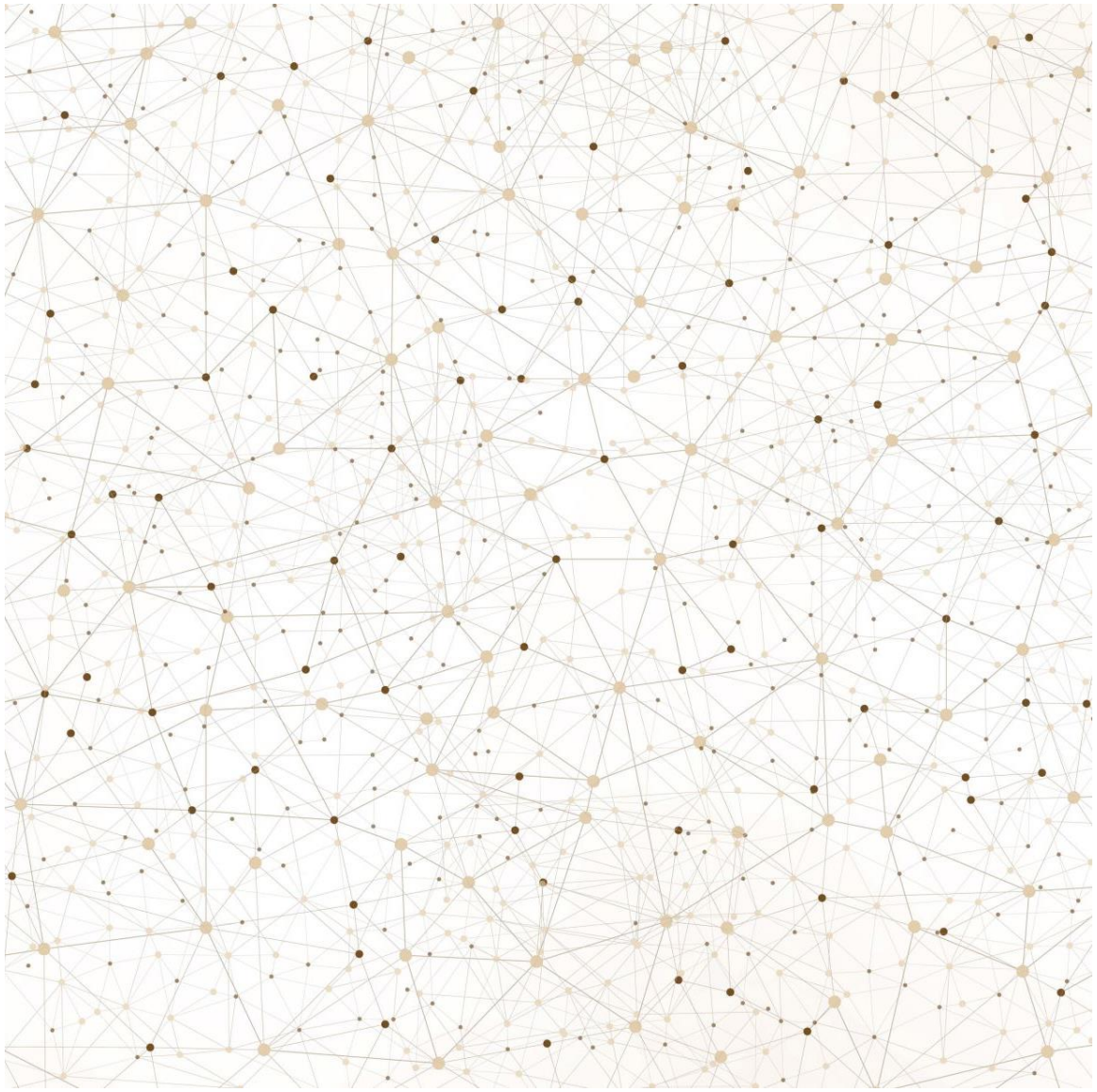
Exercise

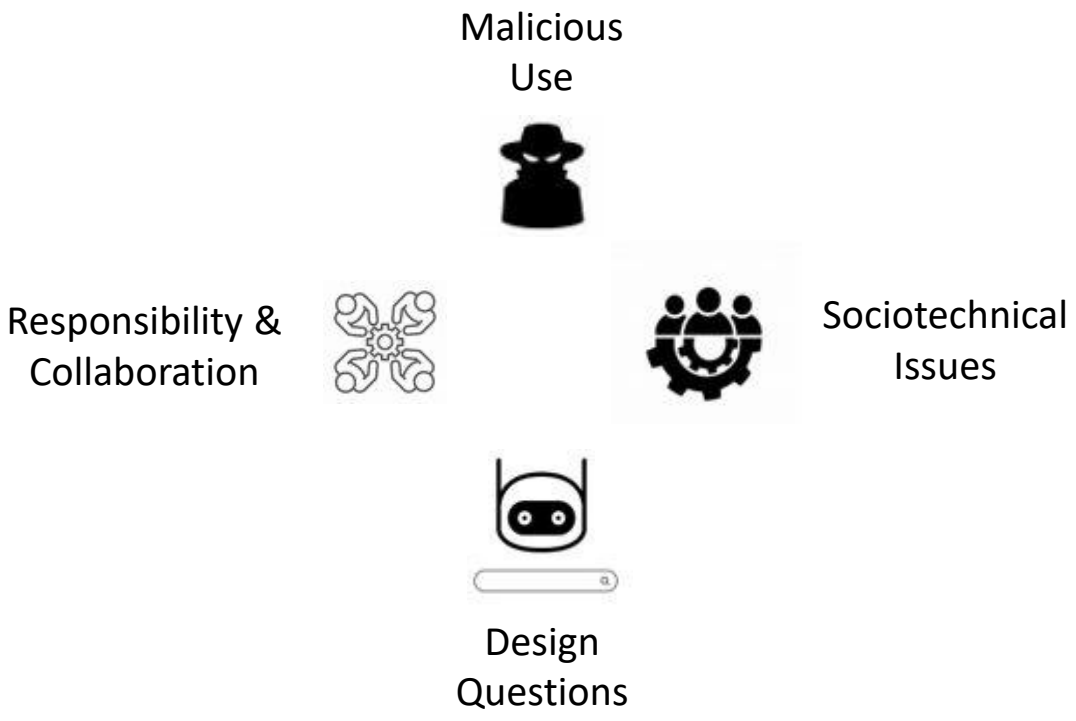
1. Have a look at the responsible AI principles of

Microsoft,
OpenAI,
Meta,
Google,
Anthropic,
Inflection.

Please take notes on which principles are mentioned. Any surprises? Anything missing?

2. Please discuss the implications of your findings for your research and research field.





4 Sets of Concerns

[Scurrell & Daepp (forthcoming)]

Ethics Explainers - The Ethics Centre

The-Ethics-Centre_PRINCIPLES-FOR-GOOD-TECHNOLOGY-29JAN.pdf

TEC_PRINCIPLES-FOR-DATA-DEVELOPMENT_UPDATE_SINGLES.pdf (ethics.org.au)

DAIR.AI

reading & source recommendations

poll before

When you hear "ethics in AI", what comes into your mind?

poll after

If you think about "AI ethics", what comes into your mind now? :)

Big companies fooling around
Now I am getting much more pessimistic after the lecture
Geographical differences
Really difficult to judge what is really done
Regulation
bias in engeneering
Marketing
No human in AI
Diversity issues
Principles
reliability
Complex
cheap talk
Important
Input bias
openness
Aiwashing
Pandoras box is open
US hand's off approach
Legal protection
trustworthiness
Alignment
complexity
EU AI act
computers taking over and causing disasters
Hope
Accountability
Same problems, but even less optimism
Constitutional AI
Personlization is a problem
China's AI governance framework