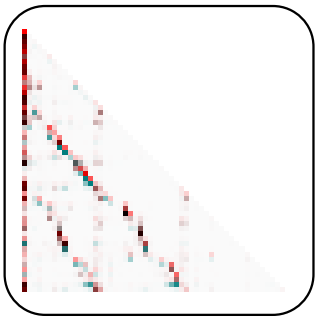
# Decoding GPT:

## Mathematical foundations of Interpretability and Alignment for Large Language Models

A course offering an in-depth exploration of the mathematical foundations, current challenges, and advanced topics in AI Alignment and Interpretability. Understand the risks and learn how to steer the future of AI towards safe and beneficial outcomes. Scan the QR code below to visit the course website:

mivanit.github.io/decoding-gpt

DeepDream visualization of layer `fc7` of VGG-19 vision model. From OpenAI microscope.



Attention pattern of a simplified model. From Transformer Circuits – Induction Heads.

---

**Course Information:**

**Term:** Spring 2023 Semester
**Time:** 12:30pm-1:45pm, Tuesdays/Thursdays
**Course:** MATH 598B, CRN 12406
**Location:** 141 Alderson Hall
**Instructors:** Samy Wu Fung, Daniel McKenzie, Michael Ivanitskiy
**Contact:** mivanits@mines.edu

Please see the course website for the syllabus, updated information, course materials, and an office hours poll.

---

APPLIED MATHEMATICS & STATISTICS
COLORADO SCHOOL OF MINES