# Decoding GPT: Mathematical foundations of Interpretability and Alignment for Large Language Models

Colorado School of Mines, Department of Applied Mathematics and Statistics

Spring 2024

| | |
|---|---|
| Course Code | MATH 598B |
| Credit Hours | 3 Credit Hours |
| Meeting Times | 12:30pm-1:45pm, Tuesdays/Thursdays |
| Location | 141 Alderson Hall |
| Instructors | Samy Wu Fung, Daniel McKenzie, Michael Ivanitskiy |
| Contact | mivanits@mines.edu |
| Office Hours | 269 Chauvenet Hall or Zoom by request. Time TBD, poll here |
| Course materials | github.com/mines-opt-ml/decoding-gpt |
| Course website | miv.name/decoding-gpt |

## Course Description

Since the public release of GPT-3 in 2020, Large Language Models have made drastic progress across a wide variety of tasks thought to be exclusively in the domain of human reasoning. However, the internal mechanisms by which these models are capable of performing such tasks is not understood. A large fraction of machine learning researchers believe that there are significant risks from training and deploying such models, ranging from mass unemployment and societal harms due to misinformation, to existential risks due to misaligned AI systems. This course will explore the mathematical foundations of Transformer networks, the issues that come with trying to impart human values onto such systems, and the current state of the art in interpretability and alignment research.

## Learning outcomes

Over the duration of the course, students will gain:

1. A solid theoretical understanding of the mechanics of a transformer networks and attention heads
2. Practical experience with implementing, training, and deploying GPTs for simple tasks
3. Understanding of the fundamentals of the AI alignment problem, present and future risks and harms, and a broad overview of the current state of the field
4. Familiarity with current results and techniques in interpretability research for GPT systems

## Prerequisites

- **Linear Algebra:** Students should have a strong grasp of linear algebra, including matrix multiplication, vector spaces, matrix decompositions, and eigenvalues/eigenvectors. MATH 500 recommended.
- **Machine Learning:** Students should be familiar with basic Deep Neural Networks and stochastic gradient descent via backpropagation. CSCI 470 or above recommended.

- **Software:** Students should be very comfortable writing software in python. Familiarity with setting up virtual environments, dependency management, and version control via git is recommended. *Experience with PyTorch or another deep learning framework is highly recommended.*
- **Research Skills:** Students should be comfortable finding and reading relevant papers in depth. How you read papers, whether you take notes, etc. is up to you, but you should be able to understand novel material from a paper in depth and be able to explain it to others.

## Course Materials

This field moves too quickly for there to currently be an up-to-date textbook on interpretability and alignment for transformers. Below are provided some useful introductory materials which we will be going over in part. Reading or at least skimming some these before the start of the course is recommended – they are listed in a rough order of priority, but feel free to skip around. We will also be reading a wide variety of papers throughout the course, and you will be expected to find interesting and useful ones.

- Stanford's CS324 course on Large Language Models
- The Illustrated GPT-2 by Jay Alammar, along with other blog posts in the series. Introduces the basics of attention heads, transformers, and autoregressive GPT-style models.
- Andrej Karpathy's `nanogpt` implementation. A simple and minimal implementation of a GPT architecture.
- Mechanistic Interpretability Quickstart Guide by Neel Nanda. A compressed guide on how to get started doing interpretability research, with many useful links (follow them!). The prereqs are also useful.
- The Transformer Circuits Thread by Nelson Elhage, Neel Nanda, Catherine Olsson, Chris Olah, and many others. Series of posts on some of the top results in transformer interpretability research.
- AGI safety from first principles. Standard introduction for why AI systems might pose catastrophic risks, and overview of the kinds of work that need to be done to mitigate them. Of particular importance is post 4 (Alignment).
- Risks from Learned Optimization in Advanced Machine Learning Systems by Evan Hubinger et al. A paper on the risks of training AI systems to optimize for a particular objective, and the potential for such systems to become misaligned – defines the inner alignment problem more formally.
- Maxime Labonne's LLM course. A hands-on set of notebooks and resources on the more practical aspects of implementing, training, and using transformers.

## Evaluation

(Subject to change)

- **Paper presentations: (30%)** Students will be expected to select and present elevant papers throughout the course of the semester. Presentations should be ~30 minutes long. A further 15 minutes will be alloted for questions, and students should participate in paper discussions. These papers should be selected with the aim of giving background for the final projects.
- **Mini Project (10%)** Working on their own, students will be expected to work through a tutorial on how to use transformers, and write a short report on their findings. Further details will be provided.
- **Final Project: (50%)** Students working in groups should select a topic related to the course material, and write a 10-15 page report on their findings. Example topics will be provided, but topic selection is flexible, as long as it relates to alignment or interpretability for ML systems.
- **Class participation (10%):** Students will be expected to attend course lectures, participate in discussions, and ask questions. Allowances for absences will be made.

## Tenative Course Outline

- Background
  - Neural Networks
  - Optimization theory
  - Architectures
  - Language Modeling
- Attention Heads & the Transformer
  - attention heads
  - positional encodings, causal attention
  - Transformers
  - Lab: using transformers
- Interpretability & Alignment
  - the AI Alignment problem
  - AI safety, ethics, and policy
  - Intro to interpretability
  - Interpretability papers
- Student Presentations
  - paper presentations
  - final project presentations

# Policies and Campus Resources

## Absences:

Mines students are expected to fulfill their academic requirements through attendance and/or participation. Class attendance is required of all students unless the student has an excused absence granted by the school or the student's professor. An excused absence awarded by the school or professor comes after a student's request or initiative. To review the Excused Absence Policy and/or to request an excused absence, please visit https://www.mines.edu/student-life/student-absences/.

## Sexual Misconduct, Discrimination, and Retaliation:

Discrimination, Harassment, and Sexual Misconduct of any type, including sexual harassment, sexual assault, dating violence, domestic violence, and stalking, are prohibited under the Policy Prohibiting Sexual Misconduct, Discrimination, and Retaliation. Please see the Office for Institutional Equity website for information on Sexual Misconduct and Discrimination.

## Preferred First Name Project:

Please feel free to let me know either in class, office hours, or via email what your preferred name and pronouns are.

Mines recognizes members of the campus community may prefer to use a first name other than their legal name to identify themselves. Many services on campus, like Canvas, utilize and display preferred first names. Additional information on preferred name, including how to update your preferred name, is available at the Office For Institutional Equity website.

## Academic Integrity:

Colorado School of Mines affirms the principle that all individuals associated with the Mines academic community have a responsibility for establishing, maintaining, and fostering an understanding and appreciation for academic integrity. In broad terms, this implies protecting the environment of mutual trust within which scholarly exchange occurs, supporting the ability of the faculty to fairly and effectively evaluate every student's academic achievements, and giving credence to the university's educational mission, its scholarly objectives, and the substance of the degrees it awards. We desire an environment free of any and all forms of academic misconduct and expects students to act with integrity at all times. Please read the full academic misconduct/integrity policy for full definitions of academic misconduct. Additionally, please use resources provided by the Office of Community Standards for guidance should you need to know more about the procedures of the policy for academic misconduct/integrity.

## Generative Artificial Intelligence:

Generative Artificial Intelligence (genAI) tools such as ChatGPT are important resources in many fields and industries. This course, in particular, is *about generative AI*. As such, students are encouraged to use generative AI tools (particularly frontier language models) for better understanding course material and completing course projects, except where explicitly prohibited by the instructor. In particular, students are required to thoroughly document their use of generative AI tools in their course projects, since the skill of using genAI tools is a learning outcome of this course.

The Office of the Provost encourages the entire University community to explore the uses and impacts of GenAI technologies, whether through critical discussions or creative applications. See the most recent Guidelines for Using Generative Artificial Intelligence at Colorado School of Mines.

## Course Issues and Concerns:

As part of good professional practice, students are encouraged to speak with the faculty directly to raise issues and concerns with regards to the course professionally in compliance with the student code of conduct. Students can also reach out to the Department Head or college dean.

## Disability Support Services:

Disability Support Services (DSS) works collaboratively with students, faculty, and staff to minimize barriers and support an accessible campus community. When barriers to access occur, Disability Support Services works one-on-one with students to determine accommodations and facilitate access to programs and services. If you've been approved for accommodations through Disability Support Services, please contact your professor to confirm receipt of your accommodation letter and to discuss the implementation of accommodations in this course. Please visit mines.edu/disability-support-services for more information or to request accommodations.

## Digital Accessibility:

The Colorado School of Mines is committed to supporting an accessible digital environment for all members of our community, including students with disabilities. If you have an accessibility concern with Canvas or any digital materials or software used in this course, please contact your professor or request support from Information & Technology Solutions. Please visit https://www.mines.edu/accessibility/ for more information.

## Student Outreach & Support (SOS) Resources:

If you feel overwhelmed, anxious, depressed, distressed, mentally or physically unhealthy, or concerned about your wellbeing overall, there are resources both on- and off-campus available to you. If you need assistance, please ask for help from a trusted faculty or staff member, fellow student, or submit a referral for yours. As a community of care, we can help one another get through difficult times. If you are concerned for another student, offer assistance and/or ask for help on their behalf. Students seeking resources for themselves or others should visit mines.edu/sos.

Student Outreach and Support can help connect you with a variety of resources; some of those might include:

- Counseling Center – https://www.mines.edu/counseling-center/ or students may call to make an appointment. There are also online resources for students on the website. Located in the Wellness Center 2nd floor. Located at 1770 Elm St.
- Health Center - https://www.mines.edu/student-health/ or students may call to make an appointment. Located in Wellness Center 1st floor.
- Colorado Crisis Services - For crisis support 24 hrs/7 days, either by phone, text, or in person, Colorado Crisis Services is a great confidential resource, available to anyone. http://coloradocrisisservices.org, 1-844-493-8255, or text "TALK" to 38255. Walk-in location addresses are posted on the website.

In an emergency, you should call 911, and they will dispatch a Mines or Golden PD officer to assist.

## Diversity and Inclusion:

At Colorado School of Mines, we understand that a diverse and inclusive learning environment inspires creativity and innovation, which are essential to the engineering process. We also know that in order to address current and emerging national and global challenges, it is important to learn with and from people who have different backgrounds, thoughts, and experiences.

Our students represent every state in the nation and more than 90 countries around the world, and we continue to make progress in the areas of diversity and inclusion by providing Diversity and Inclusion programs and services to support these efforts.