INFO 7375
Prompt Engineering for Generative AI

Welcome to the world of Prompt Engineering and Fine-Tuning for Generative AI with Large Language Models (LLMs), where we delve deep into the art and science of crafting prompts that drive LLMs to create captivating and context-aware content. In this comprehensive course, you'll not only master the essential techniques for effective prompt engineering but also gain expertise in the fine-tuning and configuration of LLMs. This dual skill set will empower you to harness the full potential of AI-driven creativity and problem-solving across a wide range of domains.

Course Highlights

Prompt Engineering Mastery: Learn the principles of creating prompts that elicit desired responses from LLMs, whether you're generating text, code, or creative content.

Fine-Tuning Expertise: Explore the intricate process of fine-tuning LLMs, optimizing them for specific tasks, domains, and applications.

Real-World Applications: Apply your skills to real-world scenarios, from content creation and decision support to interactive media and beyond.

Ethical Considerations: Discuss the ethical implications of AI-generated content and responsible AI usage in media production.

Hands-On Experience: Engage in practical exercises, assignments, and projects to reinforce your learning and gain practical experience in prompt engineering and LLM fine-tuning.

By the end of this course, you will not only be proficient in the art of prompt engineering for generative AI but also equipped with the skills to configure and fine-tune LLMs, enabling you to unleash the power of AI-driven creativity and problem-solving across diverse domains. Join us on this transformative journey into the realm of AI-driven creativity and problem-solving.

Learning Objectives

Module 1: Introduction to LLMs and Prompting

- Unveiling Large Language Models (LLMs): Their capabilities, use cases, and historical context.

- Understanding randomness in LLM output and setting the stage for effective prompt engineering.
- Creating Your First Prompts: A hands-on initiation into the world of AI-powered content generation.

Module 2: The Art of Prompt Engineering

- Deciphering the Essence of a Prompt: What is a prompt, and how can it be tailored to your needs?
- Exploring Prompt Patterns: Unraveling the Persona Pattern, Question Refinement Pattern, Cognitive Verifier Pattern, Audience Persona Pattern, and more.
- Applying Prompt Patterns: Crafting prompts for various scenarios, including Few-shot Examples, Chain of Thought Prompting, and Game Play Patterns.

Module 3: Advanced Integration Techniques for LLMs

- This module delves into sophisticated methods for augmenting LLMs with vector databases and LangChain. It covers the essentials of vector databases, including their significance in handling high-dimensional data and the process of embedding textual data for enhanced semantic search capabilities. Additionally, the module explores the creation of semantic search applications, focusing on best practices for indexing and querying.
- In the segment on LangChain, learners will explore its architecture and utility in building advanced LLM applications, alongside practical setup guidance and development of applications that integrate LLM capabilities with external data sources and APIs. The module also addresses advanced techniques and best practices for LangChain use, culminating in case studies that highlight real-world implementations and solutions to encountered challenges.

Module 4: Fine-Tuning and Configuring LLMs

- Pre-training Large Language Models: Unpacking the computational challenges, scaling laws, and domain-specific training.
- Instruction Fine-Tuning: Mastering single and multi-task instruction fine-tuning, scaling instruct models, and evaluating model performance.
- Reinforcement Learning and LLM-Powered Applications: Aligning models with human values, obtaining feedback, and optimizing for deployment.

Module 5: Beyond the Basics

- Interacting with External Applications: Integrating LLMs into real-world scenarios and applications.
- Program-Aided Language Models (PAL): Enhancing reasoning and action with LLMs.
- Model Application Architectures: Exploring advanced architectures for deploying LLMs in practical projects.

Weekly Schedule

Week 1: Introduction to Large Language Models

- Overview of Large Language Models (LLMs)
- Randomness in LLM Outputs
- Crafting Your First Prompts
    - Understanding Prompts
    - Introduction to Prompt Patterns
        - The Persona Pattern
        - Reading and Formatting Prompt Patterns

Week 2: Advanced Prompt Engineering

- Prompts as Tools for Repeated Use
- Advanced Prompt Patterns:
    - Root Prompts
    - Question Refinement
    - Cognitive Verifier
    - Audience Persona
    - Flipped Interaction
- Writing Effective Few-Shot Examples

Week 3: Advanced Prompt Techniques Continued

- Expanding Prompt Strategies:
    - Chain of Thought Prompting
    - ReAct Prompting
    - Using LLMs for Peer Grading
- Combining Prompt Patterns:
    - Game Play
    - Template Creation
    - Meta Language Creation
    - Recipe and Alternative Approaches
    - Input Solicitation
    - Outline Expansion
    - Menu Actions
    - Fact Check Lists
    - Tail Generation
    - Semantic Filtering

Week 4: Understanding Large Language Models

- Generative AI and LLMs: Foundations and Use Cases
- Before Transformers: Evolution of Text Generation
- Deep Dive into Transformer Architecture

- Generating Text with Transformers
- Prompt Engineering and Its Importance
- Lifecycle of a Generative AI Project

Weeks 5 & 6: Integrating Vector Databases with LLMs

- Introduction to Vector Databases
- Embedding Textual Data for Vector Databases
- Building Semantic Search Applications
- Enhancing LLM Responses with Vector Database Queries

Weeks 7 & 8: Leveraging LangChain for Advanced LLM Applications

- Getting to Know LangChain
- Setting Up and Configuring LangChain
- Developing LangChain Applications
- Advanced Techniques and Best Practices in LangChain Use
- Case Studies on LangChain Implementation

Weeks 9 & 10: Fine-Tuning and Configuring LLMs

- Pre-training LLMs: Challenges and Scaling Laws
- Instruction Fine-Tuning: Single and Multi-task Approaches
- Reinforcement Learning in LLM-Powered Applications
- Techniques for Parameter-Efficient Fine-Tuning (PEFT)

Weeks 11 & 12: Reinforcement Learning and LLM Applications

- Reinforcement Learning and Its Application in LLMs
- Aligning LLMs with Human Values
- Detailed Look at RLHF: Feedback, Reward Models, Fine-tuning
- Understanding Policy Optimization and Reward Hacking

Weeks 13 & 14: Deployment and Advanced Topics

- Optimizing Models for Deployment
- Utilizing LLMs in Real-World Applications
- Integrating LLMs with External Applications
- Advanced Deployment Strategies: PAL, ReAct, and LLM Architectures

Course Materials

Textbook

Title: "Prompt Engineering for Generative AI" by Nik Bear Brown (Free Online)

Publisher: Abecedarian, LLC
Publication Date: January 2023

Course GitHub

The course GitHub (for all lectures, assignments and projects):

https://github.com/nikbearbrown

nikbearbrown YouTube channel

Over the course of the semester, I'll be making and putting additional data science and machine learning related video's on my YouTube channel.

https://www.youtube.com/@BearBrownCo

The purpose of these videos is to put additional advanced content as well as supplemental content to provide additional coverage of the material in the course. Suggestions for topics for additional videos are always welcome.

Teaching assistants

The Teaching assistants are:

TBA

Programming questions should first go to the TA's. If they can't answer them then the TA's will forward the questions to the Professor.

Learning Assessment

Achievement of learning outcomes will be assessed and graded through:

- Quizzes
- Completion of assignments
- Completion of term projects

Reaching out for help

A student can always reach out for help to the Professor, Nik Bear Brown ni.brown@neu.edu. In an online course, it's important that a student reaches out early should he/she run into any issues.

Grading Policies

A point system is used for grading. Every assignment, project, and exam you are expected to complete is assigned a point value, ranging from 1 to 1000 points.

- Late submissions incur a 10% deduction per day, rounded up.
- Exams cannot be made up unless prior arrangements have been made.

Ai Based Grading Approach

Due to the widespread use of Generative AI, grading is adjusted as follows:

- Students scoring 80% or higher (absolute scale) will be graded relatively based on class performance:
    - Top 25% (rounded to the nearest integer): A
    - Next 25%: A-
    - Next 25%: B+
    - Final 25%: B

- Students scoring below 80% (absolute scale) will be graded based on the traditional absolute grading scale below:

Score	Grade
78–79	C+
73–77	C
70–72	C-
60–69	D
Below 60	F

- Important: Students below 80% cannot earn a grade higher than B-, even if the relative curve would otherwise place them higher.

Notes

- The instructor reserves the right to make minor adjustments for fairness.

- Students are encouraged to complete all work independently and ethically, especially in light of Generative AI use.

Canvas

You will submit your assignments via Canvas. Click the title of assignment (Canvas -> assignment -> <Title of Assignment>), to go to the submission page. You will know your score on an assignment, project or test via Canvas. Canvas only represents only the raw scores. Not normalized or curved grades. A jupyter notebook file ALONG with either a .DOC or .PDF rendering of that jupyter notebook file must be submitted with each assignment.

Your name MUST be part of your submission, for example Sanchez_Rick_Assignment_1.zip

Multiple files must be zipped. No .RAR, .bz, .7z or other extensions.

Assignment file names MUST start with students last name then first name OR the groups name and include the class number and assignment number.

Assignment MUST estimate the percentage of code written by the student and that which came from external sources.

Assignment MUST specify a license at the bottom of each notebook turned in.

All code must adhere to a style guide and state which guide was used.

Due dates

Assignments are due by 11:59pm on the due date marked on the schedule. It is your responsibility to determine whether or not it is worth spending the extra time on an assignment vs. turning in incomplete work for partial credit without penalty. Any exceptions to this policy (e.g. long-term illness or family emergencies) must be approved by the professor.

Assignments will receive NO CREDIT if submitted after the solutions are posted. Any extensions MUST be granted via e-mail and with a specific new due date.

Due dates for assignments at midnight on the due date of the assignment.

Five percent (i.e. 5%) is deducted for each day an assignment is late. Solutions will be posted the following Monday. Assignments will receive NO CREDIT if submitted after the solutions are posted. Any extensions MUST be granted via e-mail and with a specific new due date.

Participation Policy

Participation in discussions is an important aspect on the class. It is important that both students and instructional staff help foster an environment in which students feel safe asking questions, posing their opinions, and sharing their work for critique. If at any time you feel this environment is being threatened—by other students, the TA, or the professor—speak up and make your concerns heard. If you feel uncomfortable broaching this topic with the professor, you should feel free to voice your concerns to the Dean's office.

Collaboration Policies

Students are strongly encouraged to collaborate through discussing strategies for completing assignments, talking about the readings before class, and studying for the exams. However, all work that you turn in to me with your name on it must be in your own words or coded in your own style. Directly copied code or text from any other source MUST be cited. In any case, you must write up your solutions, in your own words. Furthermore, if you did collaborate on any problem, you must clearly list all of the collaborators in your submission. Handing in the same work for more than one course without explicit permission is forbidden.

Feel free to discuss general strategies, but any written work or code should be your own, in your own words/style. If you have collaborated on ideas leading up to the final solution, give each other credit on what you turn in, clearly labeling who contributed what ideas. Individuals should be able to explain the function of every aspect of group-produced work. Not understanding what plagiarism is does not constitute an excuse for committing it. You should familiarize yourself with the University's policies on academic dishonesty at the beginning of the semester. If you have any doubts whatsoever about whether you are breaking the rules – ask!

Any submitted work violating the collaboration policies WILL BE GIVEN A ZERO even if "by mistake." Multiple mistakes *will be sent to OSCCR for disciplinary review.*

To reiterate: **plagiarism and cheating are strictly forbidden. No excuses, no exceptions.** *All incidents of plagiarism and cheating will be sent to OSCCR for disciplinary review.*

Assignment Late Policy

Assignments are due by 11:59pm on the due date marked on the schedule. It is your responsibility to determine whether or not it is worth spending the extra time on an assignment vs. turning in incomplete work for partial credit without penalty. Any exceptions to this policy (e.g. long-term illness or family emergencies) must be approved by the professor.

Assignments will receive NO CREDIT if submitted after the solutions are posted. Any extensions MUST be granted via e-mail and with a specific new due date.

Student Resources

Special Accommodations/ADA: In accordance with the Americans with Disabilities Act (ADA 1990), Northeastern University seeks to provide equal access to its programs, services, and activities. If you will need accommodations in this class, please contact the Disability Resource Center (www.northeastern.edu/drc/) *as soon as possible* to make appropriate arrangements, and please provide the course instructors with any necessary documentation. The University requires that you provide documentation of your disabilities to the DRC so that they may identify what accommodations are required, and arrange with the instructor to provide those on your behalf, as needed.

Academic Integrity: All students must adhere to the university's Academic Integrity Policy, which can be found on the website of the Office of Student Conduct and Conflict Resolution (OSCCR), at http://www.northeastern.edu/osccr/academicintegrity/index.html. Please be particularly aware of the policy regarding plagiarism. As you probably know, plagiarism involves *representing anyone else's words or ideas as your own*. It doesn't matter where you got these ideas—from a book, on the web, from a fellow-student, from your mother. It doesn't matter whether you quote the source directly or paraphrase it; if you are not the originator of the words or ideas, *you must state clearly and specifically where they came from*. Please consult an instructor if you have any confusion or concerns when preparing any of the assignments so that together. You can also consult the guide "Avoiding Plagiarism" on the NU Library Website at http://www.lib.neu.edu/online_research/help/avoiding_plagiarism/. If an academic integrity concern arises, one of the instructors will speak with you about it; if the discussion does not resolve the concern, we will refer the matter to OSCCR.

Writing Center: The Northeastern University Writing Center, housed in the Department of English within the College of Social Sciences and Humanities, is open to any member of the Northeastern community and exists to help any level writer, from any academic discipline, become a better writer. You can book face-to-face, online, or same day appointments in two locations: 412 Holmes Hall and 136 Snell Library (behind Argo Tea). For more information or to book an appointment, please visit http://www.northeastern.edu/writingcenter/.