CS795/895 Special Topics: Large Language Model Architectures and Applications
Spring 2026

Instructor

Dr. Mahmoud Nazzal, Assistant Professor, CS Department
Office: Room 3210, ENGR & COMP SCI BLDG, Norfolk, VA 23529, USA
Email: mahmoud [dot] nazzal [at] odu [dot] edu
Personal site: fs.wp.odu.edu/mnazzal
Research profile: mahmoudkanazzal.github.io
Phone:
Class times: 3 credit hours. W: 07:00 pm-10:40 pm
Office Hours: W: 05:00 pm-06:00 pm and/or by appointment
Class location:

Teaching Assistant (TA)

TA Role
The TA assists with grading

Prerequisites
Linear algebra, basic probability theory, basic data structures and algorithms, computer programming (Python), basic familiarity with machine learning and deep learning concepts

Textbook and References

- Primary Course Material: **Lecture slides and instructor-prepared notes**.
- Textbook (assumed):
  Afshine Amidi and Shervine Amidi, Super Study Guide: Transformers & Large Language Models, 2024. ISBN: 979-8836693312

Optional References (for depth and additional reading)

- ***Lewis Tunstall, Leandro von Werra, and Thomas Wolf****, Natural Language Processing with Transformers, O'Reilly Media, 2022.*
  ISBN: ***9781098103248***
- **Sebastian Raschka**, *Build a Large Language Model (From Scratch)*, Manning Publications, 2024.
  ISBN: **9781633437166**

Brief Course Description
This course provides a principled, systems-level introduction to large language models, tracing their evolution from general learning machines to full-fledged computational systems whose behavior emerges from architectural, training, and inference choices. The course is organized into four blocks that progressively address what transformer-based language models are, how they are trained and adapted at scale, why their behavior is constrained and sometimes unreliable, and how they function as components within larger agentic systems. Students begin by revisiting foundational learning concepts and examining the architectural shift from sequential modeling to attention, developing a deep understanding of transformer structure, tokenization, context, and decoding behavior. The course then

explores pretraining objectives, fine-tuning strategies, scaling limits, and system-level constraints that shape real-world deployment. Building on this foundation, students analyze grounding, alignment, and reasoning, with emphasis on hallucination, retrieval-augmented generation, and test-time computation. The final block focuses on agentic language-model systems, evaluation methodologies, and fundamental limitations, equipping students with a rigorous conceptual framework for understanding, analyzing, and designing reliable large language model–based systems.

Upon successful completion of this course, learners will be able to:

1. Explain the learning principles and architectural motivations behind transformer-based language models.
2. Describe Transformer components and information flow, including tokenization, embeddings, self-attention, and encoder–decoder variants.
3. Analyze how context length, tokenization, and decoding strategies shape model behavior, variability, and hallucination.
4. Explain pretraining, fine-tuning, and parameter-efficient adaptation methods, along with their scaling trade-offs.
5. Evaluate system constraints affecting LLM deployment, including compute, memory, latency, and context limits.
6. Identify causes of grounding and hallucination failures and assess retrieval-augmented generation designs.
7. Explain alignment and preference-optimization methods and their trade-offs.
8. Compare reasoning approaches and test-time scaling strategies such as chain-of-thought and self-consistency.
9. Analyze agentic LLM systems, including tool use, orchestration, and cascading failure modes.
10. Critically assess evaluation methods for LLMs and LLM-based systems, including robustness and cost-aware metrics.
11. Articulate key limitations, safety challenges, and open research problems in large language models.

Grading Policy

Item	Grade Allocated
Quizes	10
Final Exam	30
Homework Assignments	25
Term Project*	35

*The project grade entails the grade of paper and proposal presentation, project presentation, and project report.

Late Submission Policy

Homework assignments submitted after the deadline will incur a 10% grade deduction per late day. Submissions more than 3 days late will receive a grade of zero, unless prior arrangements have been approved by the instructor.

Weekly Topic Organization

Block I (Weeks 1–3): Foundations of Transformer-Based Language Modeling
Focus: ***What is the machine, and why did this architecture prevail?***

Week	Date (Wed)	Topics Covered (Complete)
Week 1	Jan 21, 2026	**From Learning Models to Attention:** Basic revision of ML, DL, and NLP concepts (high level). Learning models: inputs, parameters, outputs. Training at a conceptual level: loss, optimization, generalization. Representations and embeddings. Sequence data and autoregressive generation. Likelihood versus truth in generative models. Architectural pivot to attention: limitations of naïve sequence modeling at scale; conceptual limitations of recurrence for long-range dependencies; attention as a mechanism for content-based information routing; self-attention and its role in parallelism and scalability.
Week 2	Jan 28, 2026	**Transformer Architecture:** Tokens and embeddings in Transformer models. Self-attention blocks and information flow. Feed-forward layers, residual connections, and normalization. Encoder, decoder, and encoder–decoder architectures. Autoregressive versus masked language modeling.
Week 3	Feb 4, 2026	**Tokens, Context, and Inference Behavior:** Tokenization strategies and their implications. Context windows and information retention. Autoregressive decoding strategies: greedy decoding; beam search; top-k and top-p sampling; temperature control. Effects of decoding choices on hallucination, verbosity, refusal behavior, and reasoning variability.

Block II (Weeks 4–7): Training, Adaptation, and Scaling Constraints
Focus: ***How large language models are built, adapted, and constrained in practice.***

Week	Date (Wed)	Topics Covered / Activities (Complete)
Week 4	Feb 11, 2026	**Pretraining Objectives, Data, and Scale:** Next-token prediction as the core pretraining objective. Structural consequences of likelihood-based training. Dataset composition, filtering, and contamination. Conceptual understanding of scaling behavior. Mixture-of-Experts as a scaling strategy.
Week 5	Feb 18, 2026	**Base Paper & Project Proposal Presentations:** Student teams present an approved base paper and their project proposal, including motivation, related work positioning, and planned methodology.
Week 6	Feb 25, 2026	**Fine-Tuning and Parameter-Efficient Adaptation:** Supervised fine-tuning. Parameter-efficient fine-tuning methods: Low-Rank Adaptation (LoRA); adapters; prefix tuning. Trade-offs involving capacity, stability, and forgetting. Conditions under which fine-tuning improves or degrades performance.
Week 7	Mar 4, 2026	**Efficiency, Context Limits, and System Constraints:** Context length limits and attention cost. Memory and compute trade-offs. Quantization and numerical precision. Latency, cost, and deployment constraints. Limitations of naïve scaling assumptions.

Block III (Weeks 8–10): Grounding, Alignment, and Reasoning
Focus: ***Why models fail and how their behavior is constrained.***

Week	Date (Wed)	Topics Covered (Complete)
Week 8	Mar 11, 2026	**Grounding and Retrieval-Augmented Generation:** Structural causes of hallucination. Retrieval-augmented generation architectures. Retrieval errors and brittleness. Failure-inducing interactions in retrieval-augmented systems.
—	Mar 18, 2026	**NO CLASS** – Spring Holiday
Week 9	Mar 25, 2026	**Alignment and Preference Optimization:** Motivation for alignment in deployed language models. Human preference modeling. Reinforcement Learning from Human Feedback (conceptual overview). Direct Preference Optimization. Alignment trade-offs and unintended behaviors.
Week 10	Apr 1, 2026	**Reasoning and Test-Time Scaling:** Definitions and interpretations of reasoning in language models. Chain-of-thought prompting and reasoning regimes. Self-consistency and verification-based approaches. Test-time computation versus training-time computation.

Block IV (Weeks 11–14): Agentic Systems, Evaluation, and Limitations
Focus: ***When language models become systems and where they break down.***

Week	Date (Wed)	Topics Covered / Activities (Complete)
Week 11	Apr 8, 2026	**Agentic Language Models and Tool-Using Systems:** Tool use and function calling. Plan–act–observe (ReAct-style) interaction loops. State, memory, and orchestration. Failure modes in agentic systems, including cascading errors.
Week 12	Apr 15, 2026	**Evaluation of Language Models and Language-Model-Based Systems:** Task-based evaluation methodologies (e.g., pass@k, constraint satisfaction). Model-based evaluation using language models as judges. Robustness under distribution shift. Cost-aware evaluation: quality, latency, and computational expense.
Week 13	Apr 22, 2026	**Limitations, Safety Perspectives, and Open Problems:** Hallucination mitigation strategies and their limitations. System brittleness and over-trust. Calibration gaps and uncertainty. Open research challenges and unresolved questions in large language models.
Week 14	Apr 29, 2026	**Final Project Presentations:** Student teams present final project outcomes, evaluation results, limitations, and lessons learned.

Final Exam (Official Exam Week)

Period	Dates	Activity
Final Exam Week	May 6–13, 2026	**Final Comprehensive Exam** (scheduled by the Registrar; not held during a regular Wednesday session).

Student Mental Health and Wellbeing
ODU is committed to supporting your mental health. The Office of Counseling Services offers free, confidential support including virtual and in-person counseling, group sessions, and crisis services. Same-day or next-day appointments can be scheduled at odu.edu/counseling services. In case of a mental health emergency—such as thoughts of self-harm, harming others, or recent assault—call 911, or contact a crisis counsellor 24/7 at 757-683-4401 (press Option 2). You may also call the Suicide & Crisis Lifeline at 988 or text "HOME" to 741741.

Academic Integrity
Old Dominion University expects honesty in all academic work. In this course, your work and conduct must comply with the **Code of Student Conduct** (odu.edu/oscai). Academic dishonesty includes:

- **Cheating:** using unauthorized assistance, materials, study aids, or information.
- **Plagiarism:** using others' language or ideas without proper acknowledgment.
- **Fabrication:** inventing, altering, or falsifying data, citations, or information.
- **Facilitation:** helping another student commit a violation or failing to report suspected violations.

Classroom disruptions that undermine instruction are also prohibited. **Suspected violations will be reported to the Office of Student Conduct & Academic Integrity and may result in sanctions up to and including expulsion.**

Honor Code
Students must follow the ODU Honor Code for all assignments and exams. Collaboration on ideas is encouraged, but **all submitted work (code, text, analysis) must be your own**. Limited use of tools (e.g., ChatGPT) for brainstorming is acceptable; **do not submit AI-generated text/code as your own**. Suspected violations will be handled under University policy.

Drop Policy
As per University guidelines. See the University Calendar for drop dates.

Accessibility & Accommodations
ODU provides reasonable accommodations under the ADA. If you need accommodations, **obtain an accommodation letter from the Office of Educational Accessibility (OEA)** and share it with me early so we can implement the approved arrangements. If you anticipate barriers but don't yet have a letter, **contact OEA** to discuss eligibility.

- **OEA:** 1021 Student Success Center, (757) 683-4655, odu.edu/educationalaccessibility
- Accommodations begin **once** I receive your official OEA letter.

Attendance Policy
Students are expected to attend classes regularly.