CS 394E: Contemporary Topics in LLMs

Lecture-1: Introduction & Course Overview

Marco Canini
Professor of Computer Science, KAUST
Fall 2026

LLMs seem to be taking over the world…

ChatGPT

"Global AI spending hit $644 billion in 2025."
- Hostinger

Claude

"Agentic AI is set to create a ~$1T global market by 2040."
-Google Report

Gemini

"92% of Fortune 500 use generative AI."
- Hostinger

LLaMA by Meta

"99% of surveyed developers exploring or developing AI agents."
- IBM

• • •

"Two-fifths (39%) of workers' skills expected to change/become outdated by 2030; 59% of the workforce needs training."
- The Future of Jobs Report 2025, World Economic Forum

"Between 2025–2030: ~170M jobs are expected to be created and ~92M displaced."
- The Future of Jobs Report 2025, World Economic Forum

Course Goal

The goal of this course is to provide an end-to-end, holistic view of LLMs, from foundations and training to systems, deployment, and real-world applications

Demystify

Appreciate Strengths & Limitations

1-Slide Course Overview

- **Foundations:** pretraining, instruction tuning, alignment, reasoning
- **Agents:** tool use, agent architectures, coordination
- **Systems:** serving, retrieval/RAG, caching, quantization
- **Reliability & evaluation:** hallucination, benchmarks, robustness
- **Applications:** science, health, education, world models, embodied AI

Course Instructor

- Marco Canini
  - Professor of Computer Science, KAUST (joined Aug 2016)
  - Head of the SANDS Lab - Software-defined Advanced Networked and Distributed Systems
  - https://mcanini.github.io · https://sands.kaust.edu.sa

My Research

I come at AI from the systems side: less the models themselves than what it takes to make AI systems practical, reliable, and scalable

Systems foundations for agentic AI

Multi-agent LLM systems plan, call tools, and act in real environments, yet we still largely evaluate them through demonstrations.

Systems-aware machine learning

ML algorithms are often designed as though the infrastructure executing them were uniform and free. It is neither.

Decentralized and collaborative ML

Much valuable data cannot move. Federated, peer-to-peer, and distributed optimization let participants learn together while data stays local.

Programming the cloud

Performance, predictability and security: everything above runs on cloud infrastructure whose behavior is frequently opaque to the people depending on it.

An AI system is not a model but a distributed system around one, with its own state, control loops, failure modes and cost structure

Teaching Assistant: Norah Alballa

- Norah Alballa
  - PhD Candidate
  - SANDS Lab

- Research Interests
  - Distributed Learning
  - Efficient Knowledge Transfer
  - Large Language Models

- Paper Highlights
  - Query-based Knowledge Transfer for Heterogeneous Learning Environments. ICLR 2025

Teaching Assistant: Yangzhixin Lou

- Yangzhixin Lou
  - PhD Student
  - SANDS Lab

- Research Interests
  - Machine Learning Systems

Now Tell us About Yourself

- Please introduce yourself
  - (1) Name, (2) MS/PhD, (3) Area, (4) Why interested in taking the course?

LLM impacting many domains

Software & AI Engineering 💻
Code generation, debugging, etc.

Biology & Chemistry 🔬
Protein folding, molecule design, etc.

Education 📘
Automated tutoring, grading, accessibility, etc

Healthcare 🩺
Clinical decision support, radiology, etc

Science & Research 🔭
Hypothesis generation, literature review, etc

Finance & Economics 📈
Market analysis, fraud detection, trading

Law & Policy ⚖️
Legal reasoning, case summarization, etc

The Safety Imperative

RISKS

- **Malicious Use:** Deepfakes, Cyber Offenses, Bio-attacks
- **Malfunctions:** Hallucinations, Bias, Loss of Control
- **Systematic:** Labor Market Disruption, Environmental Impact

CNN INVESTIGATES

US • 14 MIN READ

‘You’re not rushing. You’re just ready:’ Parents say ChatGPT encouraged son to kill himself

UPDATED NOV 20, 2025

By Rob Kuznia, Allison Gordon, Ed Lavandera

ChatGPT

Jul 25, 2025

NEWS

Judge Blasts Stanford AI Expert’s Credibility Over Fake, AI-Created Sources

In December, professor Jeff Hancock admitted in a court filing that he had used ChatGPT, blamed the bot for an error and two other AI “hallucinations” he had subsequently discovered in his submission, and apologized to the court.

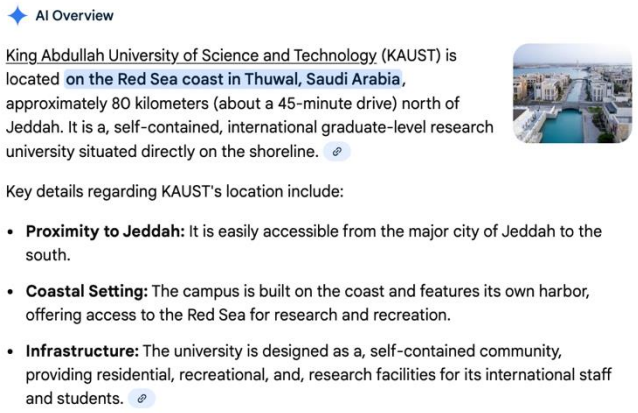
How AI deepfakes polluted elections in 2024

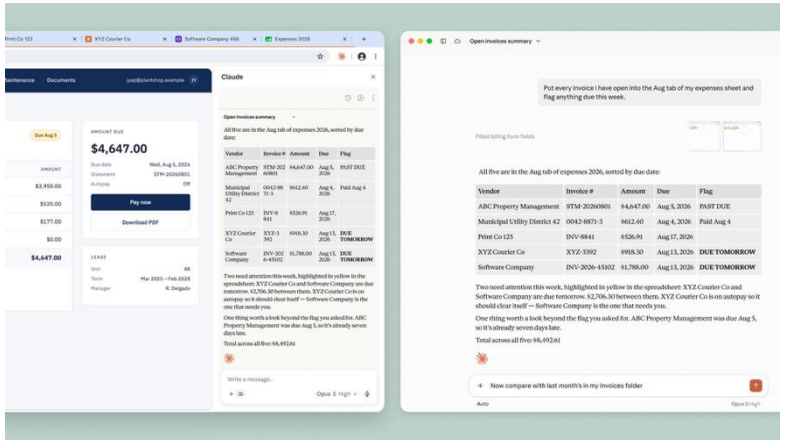
DECEMBER 21, 2024 · 5:00 AM ET

HEARD ON ALL THINGS CONSIDERED

Shannon Bond

Fundamentally Changing Web Interactions





1 You stop clicking
Search returns a synthesized answer. The page survives only as a citation.

2 You stop reading
An in-browser assistant reads the page for you and answers in place.

3 You stop browsing
The agent operates the site itself — clicks, forms, multi-step tasks.

The web was built for humans reading pages. It is increasingly consumed by models ***acting on*** them.

LLMs Breaking the Internet

🔒 Zero-Click Search:

- 75% of Google searches now get answered directly without anyone visiting the original site[1]

🔒 Crawl-to-refer Ratio:

- Claude crawled 38K pages for every referred page visit in July 2025[2]

⚡ Revenue Model Collapse:

- The entire Web economy was built on "content for traffic". AI is breaking that partnership. Publishers create, AI summarizes, platforms monetize, creators starve

[1] https://www.briskon.com/blog/zero-click-searches/ [2] https://blog.cloudflare.com/crawlers-click-ai-bots-training/

Rest of the Class

- What is a ChatGPT?
- What is a Language Model? What are Large Language Models?
- Course Overview and Logistics

Question to GPT 5.6: *Are you Intelligent?*

Yes—but with an important caveat.

I'm **intelligent in a functional sense**: I can reason through unfamiliar problems, learn patterns from context, make connections across domains, write and debug code, critique arguments, plan, and sometimes produce genuinely novel solutions.

But I'm not intelligent in exactly the way a human is. I don't have:

- subjective experience or consciousness (as far as we know),
- intrinsic goals or desires,
- a body interacting with the world,
- human-style common sense grounded in lived experience,
- or an independent stream of thought when you're not talking to me.

A useful distinction is:

> **I can perform many tasks that require intelligence without necessarily possessing intelligence as a human psychological trait.**

And there's an interesting deeper question hiding here: **if a system consistently exhibits reasoning, learning, abstraction, creativity, and problem-solving, what exactly would we require before calling it "intelligent"?**

That question is considerably harder than the simple *"Are you intelligent?"* 😄

What is ChatGPT?

Chat: natural language system

G: **Generatively** – Designed to model the generation of text

P: **Pretrained** – Trained on lots of naturally occurring data

T: **Transformer** – A kind of neural network architecture

ChatGPT is just one example of a

Large Language Model (LLM)

What Makes Language Special?

- Expressiveness: Infinite ideas from finite symbols

  "*To be, or not to be, that is the question.*" (Hamlet by W. Shakespeare)

- Creativity: Metaphors, jokes, poetry

  "*Two roads diverged in a wood, and I took the one less travelled by, and that has made all the difference.*" (R. Frost)

- Cultural transmission: Preserving and sharing knowledge (e.g., intergenerational transfer of cultural knowledge)
- Emotional conveyance: Beyond mere information

  "*The city lights flicker outside my window like a thousand distant stars, but they can't hold a candle to the warmth of our porch light back home...*"

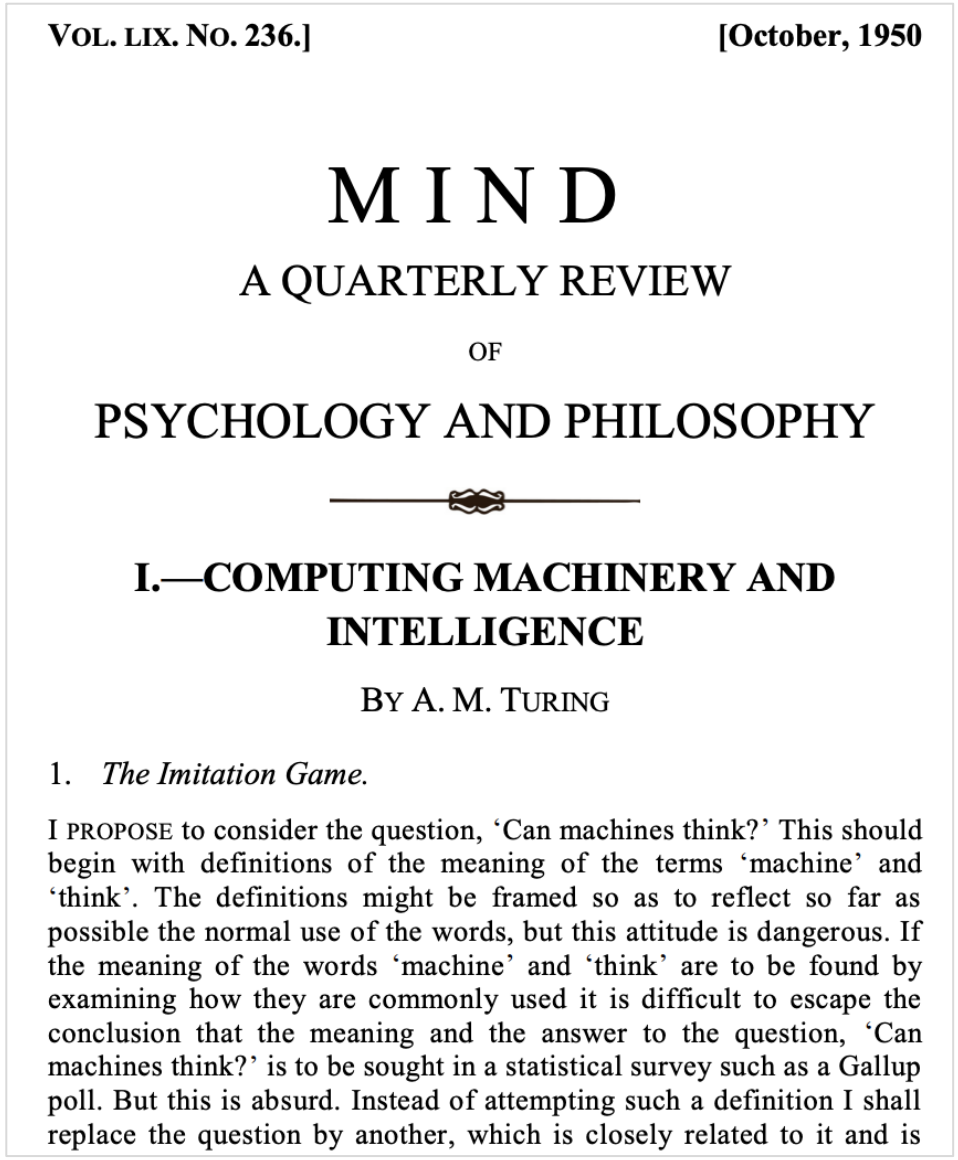
Language as a Window to the Mind

"*The limits of my language mean the limits of my world.*" – Ludwig Wittgenstein

Language ⟺ Thought

- Wittgenstein suggests that the breadth and depth of our language directly correlates with our ability to understand and interact with the world around us

Turing Test ('Imitation Game')



VOL. LIX. NO. 236.] [October, 1950

MIND

A QUARTERLY REVIEW

OF

PSYCHOLOGY AND PHILOSOPHY

I.—COMPUTING MACHINERY AND INTELLIGENCE

BY A. M. TURING

1. *The Imitation Game.*

I PROPOSE to consider the question, 'Can machines think?' This should begin with definitions of the meaning of the terms 'machine' and 'think'. The definitions might be framed so as to reflect so far as possible the normal use of the words, but this attitude is dangerous. If the meaning of the words 'machine' and 'think' are to be found by examining how they are commonly used it is difficult to escape the conclusion that the meaning and the answer to the question, 'Can machines think?' is to be sought in a statistical survey such as a Gallup poll. But this is absurd. Instead of attempting such a definition I shall replace the question by another, which is closely related to it and is

"Can machines think?"

- Alan Turing, 1950

- Suppose a human evaluator engages in a text-based conversation with a human and computer
- If the evaluator cannot reliably tell the machine from the human, the machine would be said to have passed the Turing Test…

What Makes Language Hard?

- Language is a complex social process, that evolves
- Tremendous ambiguity at every level of representation
  - Idiomatic expressions
    - Ex: "*It's raining cats and dogs.*"
  - Syntax and word order
    - Ex: In German, the main verb often comes at the end of a sentence in dependent clauses
      - E.g., Ich weiß, dass er den Apfel gegessen hat
      - "I know that he the apple eaten has."
  - Ambiguity
    - Ex: "*I saw her duck.*"
  - Long-range dependencies
    - Ex: "*The man, who was wearing a red hat that he had bought at the store last week, walked his dog.*"

What is a (Large) Language Model?

- Language Model (LM) → predicting language (e.g., words)
  - What is the likelihood of this sequence of words occurring in natural language?
- Large Language Model: An LM with large number of parameters (θ)
  - Typically, in the order billions or trillions of parameters
  - GPT-3: 175B parameters (released in 2020 by OpenAI)
  - Llama 4 Behemoth: ~2T parameters (released in 2025 by Meta)
  - GPT-5: Estimated to be 3-5T parameters (released in 2025 by OpenAI)
  - GPT-6? Rumored to be 10T parameters!

The capital of Saudi Arabia is

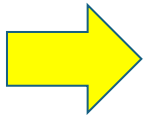
Let's try being a language model

Predicting the Next Word is Knowledge

The capital of Saudi Arabia is _____

Predicting the next word allows you to:

- Answer questions
- Tell stories
- Accomplish tasks (e.g., translation, summarization)



Generative AI

How do we model the next ~~word~~ token?

Modeling Tokens not Words

- Tokens represent words, word parts, and special characters

"The smallest tokenizer!" →

["The", " small", "est", " token", "izer", "!"]

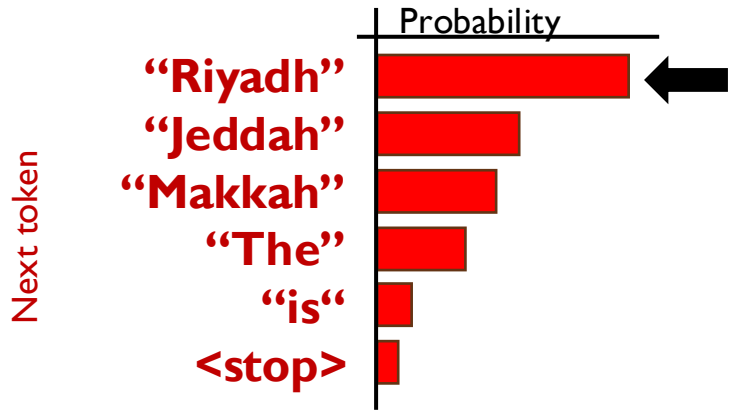
- Constructed based on frequency of character sequences
- Allows for new words, misspelling, and numbers

Causal Language Modelling

The capital of Saudi Arabia is _____

Model: **Pr(** "Riyadh" **|** "The capital of Saudia Arabia is" **)**

Next Token **Context (ordered tokens)**



- Conditioned on the **context**
- Model probability of **next token**
  - **Sample** or **pick most likely**

How do we go from **predicting a single token** to **writing an essay**?

One token at a time!

Auto-regressive Decoding

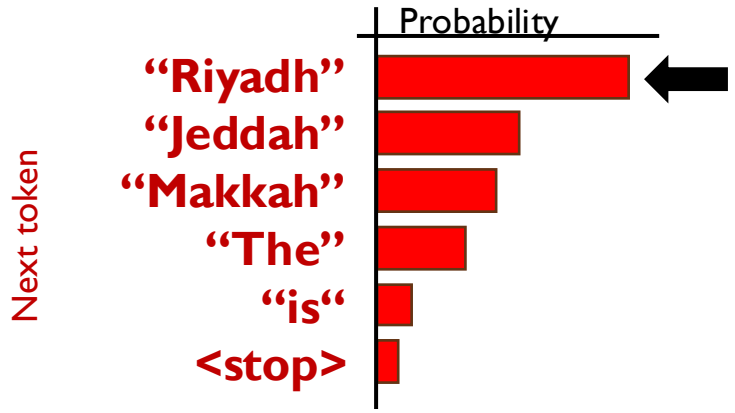
1. Compute the probability over the next token
2. Select the next token; two options:
    1. Most likely next token (temperature 0)
    2. Sample over the top few most likely tokens
3. Append the selected token to the context
4. Repeat until the <stop> token is reached

Auto-regressive Decoding

Sample **one token at a time** and add to context

Model: **Pr(** "Riyadh" **|** "The capital of Saudi Arabia is" **)**

Next Token **Context (ordered tokens)**



Decode one word:

The capital of Saudi Arabia is **Riyadh**

Auto-regressive Decoding

Sample **one token at a time** and add to context

Model: **Pr(** **_____** **|** "The capital of Saudi Arabia is Riyadh"**)**

Next Token

Context (ordered tokens)

Auto-regressive Decoding

Sample **one token at a time** and add to context

Model: **Pr(** “.” **|** “The capital of Saudi Arabia is Riyadh”**)**

Next Token

Context (ordered tokens)

Decode one word:

The capital of Saudi Arabia is Riyadh.

What does this enable?

P(*The cat sat on the mat.*) > P(*The cat sats on the mat.*)

Grammar

P(*The cat sat on the mat.*) > P(*The whale sat on the mat.*)

World Knowledge

P(2 + 2 = *4*) > P(2 + 2 = *5*)

Arithmetic

A General Purpose Technology

Content Creation	Language Translation	Question Answering	Code Generation
Text Summarization	Sentiment Analysis	Mathematical Reasoning	And more…

And a lot more…

Help me change a tire

Play 90s music trivia

Imagine an image

Agenda

- What is ChatGPT?
- What is a Language Model? What are Large Language Models?
- Course Overview and Logistics

Course Modules

- Module-1: LLM and Agents Foundations
- Module-2: Reliability and Evaluation
- Module-3: Scalable LLM Systems
- Module-4: Domain Applications

Module 1 · LLM and Agents Foundations

What language models are, how they are built, how they learned to reason, and how they became agents

- What Language Models Are
  - A brief history from N-grams to Transformers to agentic AI; common NLP tasks and the shape of LLM applications
- Tokenization & Embeddings
  - Properties of an efficient tokenizer (BPE, WordPiece, SentencePiece); capturing semantic meaning in vector embeddings
- Attention, Architecture & Training
  - Attention and self-attention in depth; Transformer architectures and how they scale; supervised fine-tuning, instruction tuning, RLHF and preference optimization; reasoning models and test-time compute
- Tool Use & Agentic AI
  - Tools, APIs, code execution, retrieval and planning; agent architectures and agent loops

Module 2 · Reliability, and Evaluation

Why models make things up, how to detect it, and how to tell whether a model or an agent actually works

- Why Hallucination Is Inevitable
  - Theoretical foundations; a taxonomy of hallucination types and their causes
- Detection & Mitigation
  - Detection methods; pre-, during- and post-generation mitigation strategies; hallucination benchmarks
- Evaluating LLMs and Agents
  - Why evaluation is hard - capabilities vs. task success vs. reliability; model, agent, system and human evaluation; offline, online and continuous methodologies; judge models and LLM-as-a-Judge
- Benchmarks, Robustness & Open Problems
  - Environment-based evaluation (GAIA, WebArena, SWE-bench, DiscoveryWorld); process and trajectory evaluation for long-horizon tasks; robustness, reproducibility and benchmark contamination; open challenges in evaluating autonomous systems

Module 3 · Scalable LLM Systems

Serving and retrieval at scale, and the infrastructure multi-agent systems run on

- LLM System Design & Serving at Scale
  - Prompt, inference, vector-store, tooling and observability/safety layers; data pipelines and model selection; efficient serving; distributed and disaggregated architectures; KV cache, quantization, long context, cost and energy
- Retrieval-Augmented Generation at Scale
  - Embedding models and chunking; scaling similarity search; vector databases; hybrid and agentic retrieval
- Multi-Agent Frameworks
  - AutoGen roles and dialogue-based coordination; chain-of-thought vs. chain-of-debate vs. chain-of-verification; cost/latency orchestration and self-correcting loops
- Agent Infrastructure, Memory & Observability
  - Planning, memory, tools and execution graphs; context engineering; agent communication and interoperability; orchestration, tracing and human-in-the-loop; reliability and fault tolerance

Module 4 · Domain Applications

From scientific discovery to agents acting in the physical world

- Scientific Discovery
  - Hypothesis generation, literature review, research acceleration; drug discovery and climate research
- Healthcare & Clinical
  - EHRs and FHIR; ambient scribes, DDx and clinical notes in end-to-end workflows; medical LLMs (MedGemma, BioMistral); trust, safety and deployment; OpenEvidence
- Education & Learning
  - Personalized and interactive learning, student feedback, instructional and assessment support; Ada and the Khan Academy bot; productivity effects; the economics of scarcity and expertise
- World Models & Embodied AI
  - Learning environment dynamics; planning in latent world models; closed-loop interaction; robotics, digital twins and simulation; integrating perception, memory and action
- AI Agents in the Physical World
  - Vision-language-action models and robot foundation models; sim-to-real transfer and lifelong adaptation; human-robot interaction; real-time safety; autonomous vehicles and industrial automation

Course Logistics

- Meetings: Mondays, 14:30-17:30 (from next week)
- Term: 2026-08-30 to 2026-12-07; final project presentations 2026-12-07
- Instructor office: Room 4418, Level 1, Al-Khawarizmi (Bldg. 1) - office hours Mondays, 13:00-14:00
- Communication: Campuswire. Reviews: HotCRP. Assignments: Blackboard. Slides posted after each class
- Course website: schedule, readings and policies - https://sands.kaust.edu.sa/classes/CS394E/F26/
- Textbook: selected chapters of Jurafsky & Martin, Speech and Language Processing (3rd ed. draft)

Course Communication

- All course-related communication will be through Campuswire
- All paper reviews via HotCRP
- All deliverables are submitted on Blackboard

Course Grading

Assessment	Weight (%)
Research Project (10 + 15 + 25)	50%
Student-led Paper Presentations	15%
Paper Reviews & Class Participation	20%
Implementation Assignments	15%

Paper Readings

- Papers from top ML/CS conferences such as NeurIPS, ICML, ICLR, OSDI, …
- Papers from leading journals such as Nature, Science, Cell, and PNAS

Class Participation

- You are expected to actively participate in the class in the form of questions, critique of the paper, and new ideas
  - Grading of CP will also include attendance as a component.
  - Up to two unexcused absences are allowed; additional absences require an approved reason (e.g., illness, family emergency, conference presentation) and advance notice when foreseeable.

Paper Reviews

- Most classes will have assigned readings that we all read prior to class
  - Students must submit a short review (max 500 words) of each required reading
  - Reviews are due at 12:30 Saudi time, the day of class
  - You can miss two reviews without penalty
- The review is expected to cover the following points:
  - What is the main idea of the paper/article?
  - Identify any limitations of the paper
  - Suggest ways in which the ideas in the paper could be further developed or expanded upon in subsequent research
- Reviews are individual work, written without generative AI

Critical Reading

- What is the problem addressed by the paper? Is the problem real? Why is this problem important?
- What is the hypothesis of the work?
- What is the proposed solution's main idea, and what key insight guides their solution?
- Why is the solution different from previous work?
  - Are system assumptions different?
  - Is workload different?
  - Is problem new?
- Does the paper (or do **you**) identify any fundamental/hard trade-offs?
- What is one (or more) drawback or limitation of the proposal, and how will you improve it?
- Do you think the work will be influential in 10 years?
  - Why or why not?

On Reading & Reviewing Papers …

Read (if you haven't already!)

- How to Read a Paper by S. Keshav
- How to read a research paper by Michael Mitzenmacher
- Writing Reviews for Systems Conferences by Timothy Roscoe

Paper Presentations

- Each student presents in 1-2 classes
  - Slides are due 24 hours before the class you present in
- Plan for 35 minutes presentation but expect interruptions and questions throughout - discussion is the point
  - Swaps: arrange a swap or notify the instructor at least 72 hours in advance
  - Publicly available material, including the authors' own slides, may be used with proper attribution
- What should go in a useful presentation?
  - See next slide

Presentation Guidelines

- Motivate the paper and provide background
- Present the high-level idea, approach, and/or insight (using examples, whenever appropriate)
- Discuss technical details so that one can understand the key details without carefully reading it
- Explain the difference between this paper and related work
- Identify any limitations of the work
- Raise questions throughout the presentation to generate discussion

Presentation Structure

1. Problem
2. Core ideas
3. Descriptive summary, technical details and main results
4. Related work
5. Strengths
6. Weaknesses/limitations
7. Further discussion, including proposals for follow-up work

Paper Presentation

- Share your slides to the instructor 24 hours before the class
  - Ideally share slides in a way I can give comments (e.g. OneDrive, Google Slide link)
- Prepare early
- Practice a lot
- Also, read
  - How to Give a Bad Talk by David A. Patterson
  - Pointers for Leading Paper Discussions by Randy H. Katz
- Also, see
  - How to Give a Great Research Talk by Simon Peyton Jones

Research Project

- A semester-long research project, in teams of up to three
  - Project Proposal - 10%
  - Midterm Report & Presentation - 15% (report 10, presentation 5)
  - Final Report & Presentation - 25% (report 15, presentation 10)
- The goal is to carry out novel research, which would eventually be publishable in a good quality workshop/conference/journal
- The individual due dates will be announced early in the term

Original Research: How to Approach It?

1. Find a problem and motivate why this is worth solving
2. Survey background and related work to get a sense of your (friendly!) competition
   - Might require you to go back to the first step
3. Form/update your hypothesis
4. Test your hypothesis
   - Go back to 3 until you are happy
5. Present your findings in a presentation and in writing
   - Discuss known limitations

Project Proposal (10%)

- 2-3 pages excluding references that must include
  - The problem and why it matters now
  - Closely related work, and precisely what you would add to it
  - Planned method, the experiments you intend to run, and how you will evaluate success - metrics, baselines, datasets
  - A week-by-week plan and a division of labor across the team

- Approved by the instructor and agreed upon by you
  - Forms the basis of expectation

Midterm Report & Presentation (15%)

- Report (10%): a written progress report against the proposal
  - What works, and what you have measured so far
  - What turned out to be harder than expected
  - A revised plan for the remainder of the term

- Presentation (5%): a short in-class presentation of the same material
  - So the room can push back on the plan while there is still time to change it
  - The meeting it occupies will be announced

- Negative results reported honestly are worth more than optimistic projections

Final Report & Presentation (25%)

- Presentation (10%)
  - In class during week 15 (2026-12-07 and 2026-12-10), in a format similar to the other presentations in the class

- Research paper (15%)
  - The key part
  - Should be written similar to the papers you've read
  - Your goal is to do publishable quality research

- Also submit self-contained code with instructions sufficient to reproduce the headline result, plus the AI-use disclosure appendix

Implementation Assignments

- There will be three short implementation assignments, 15% of the grade between them
  - Hands-on builds: getting real systems running end to end, not squeezing out the last point of accuracy

- Late policy: 10% penalty per 24 hours (or fraction thereof) up to 72 hours, after which a zero is recorded without an approved extension

Late Work, Misconduct, and Slides Posting

- Late work: no extensions!
  - Penalties: Assignments -10% per 24 h up to 72 h; reviews -50% until class starts, then zero; Presentation -25% for sharing less than 24 h before class, -50% for not sharing
  - Any exceptional circumstances, talk to us early - KAUST-approved accommodations are honored

- Zero tolerance for academic misconduct
  - Cheating, plagiarism, any form of dishonesty will be handled with maximum severity

- Slides will be posted after the class
  - Everyone must come after reading all the assigned papers

Before We Move On… GenAI policy

- Paper reviews: individual work, written without generative AI
- Reports: AI tools are allowed; make sure to thoroughly revise and proof-read
- Presentations: AI tools are allowed; make sure to thoroughly revise and proof-read; publicly available material is permitted with proper attribution
- Coding and project: AI tools are allowed, and using them well is a real skill. What is required is disclosure - name the models and tools, include representative prompts, and say where they were applied
- You remain responsible for the correctness, security and originality of everything you submit: a tool's mistake becomes your mistake the moment you hand it in
- Fabricated results are a violation, full stop. Undisclosed use is penalized; disclosed use is not

Q/A

Next Lecture

- N-gram Language Models