Large Language Models and Artificial Intelligence

Table of contents

1 Large Language Models (LLM)

LLM: Example - JEREMY the Chat bot

Cybercrime was becoming increasingly problematic for the citizens of a democratic mid-sized European nation. To deal with cyber identity theft crimes, citizens were asked to vote on providing the national law enforcement with the power and ability to charge and prosecute cyber criminals.

Collaborating with local experts, the national law enforcement leaned in natural language processing- the origin of LLMs- to find criminals. Specifically, an algorithm, named JEREMY, was trained on online chat forums where criminal behavior had taken place.

LLM: Example - JEREMY the Chat bot

JEREMY the chat bot was able to enter and engage in conversations with people that had been suspected of such crimes. At first the bot was able to identify individuals with smaller crimes (e.g., hacking) and later able to identify criminal rings that were engaging in more brazen acts. Additionally, the bot was also tasked with creating dossiers on current and possible criminals.

It should be noted that citizen did not know about JEREMY the bot to keep the bot a secrete and catch criminal.

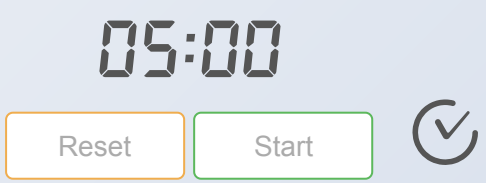
💬 Possible Discussion Questions 💬

JEREMY the Chat bot Discussion

Take 2 minutes consider the following questions and 5 minutes for class discussion:

- If you were hypothetically a citizen of the nation, what would be your opinion about using a chat bot to identify, monitor, and incriminate cyber criminals on forums/group chats? Consider if you were a victim of identity theft versus a not targeted member of this society.
- Are there possible violations of ethical principles? Do the benefits of having JEREMY outweigh the possible ethical violations?

Take 5 minutes for class discussion:



Instructor Note

Possible ethical principles to consider:

Privacy: The bot was able to collect and analyze data from who used the forums. Furthermore, collection of documents were developed on possible and existing criminals.

Consent:Citizens of nation did not know about the bot and therefore, were not able to given consent to monitoring and collection of data.

Transparency: The team of experts and law enforcement did not provide information about the algorithm’s performance or possible bias.

LLM: Define and Overview

- LLMs approximate human language using prediction models that are trained on massive text data sets
  - First, models are trained on known text sequences to learn general patterns in language, including syntax, grammar, and pragmatics. Through an iterative process the model's parameters are estimated and updated while the model is trained across the large number of known text sequences.
  - Second, the models are fine-tuned using human feedback

Source Pai, 2025

LLM: Uses

The following examples of how LLMs are being used across different fields:

- Academic Research
  - Example include: brainstorming research questions, examining quantitative an qualitative data, etc. (Kapania et al. 2025)
- Health Care
  - Examples include: patient question sorting (Chekuri et al. 2025), transcribing patient visits and doctor notes , LLM Health Wizards (Q and A bots), etc. Holley & Mathur, 2024.
- Business
  - Examples include: Business analytics (Nasseri et al. 2023), analyze customer's product reviews (Beránek and Merunka 2024), Business management (Estrada-Torres, del-Río-Ortega, and Resinas 2024)

LLM: Define and Overview

- In 2025, popular LLMs and their associated companies include:
  - ChatGPT-4 (generative pre-trained transformer) - Open AI
  - Claude - Anthropic
  - Deep Seek
  - Gemini/BERT - Google
  - Mistral

Instructor Note

- All of the aforementioned LLM models vary in popularity, strengths, and weakness as of September 2025. Since LLMs are relatively a new popular phenomenon these LLMs may be come obsolete later. ## LLM: Example

Back ground on Chat GPT:

- ChatGPT model has multi-modal capabilities, suggesting that the model can handle images, text, and sound.
- Unlike some other LLMs, Chat can recall questions previously asked.
- ChatGPT initial training began on online text data (How ChatGPT and our foundation models are developed). To improve ChatGPT's conversational skills the model was trained on real transcripts.
- Up until 2023, chat GPT was only trained on text data up to 2021. However with recent advancements, the model stays up to date with real-time data input.

“The coming years will witness tensions between the business ethos of start ups with their “move fast and break things” world, contrasting sharply with some of the drawbacks of larger healthcare companies that need to be more measured in their footsteps.”

LLM: ChatGPT and Healthcare

An older man underwent a procedure to address his irregular heartbeat condition. A couple of days later, he began to experience double vision, where a single object was perceived as two. He contacted the surgeon about his newly developed symptoms, and the surgeon noted that double vision was a “harmless” side effect and indicated that if the symptom continued to occur, he should contact his general doctor or go to the ER. Unsatisfied with the surgeon’s answer, he then consulted with ChatGPT, which suggested that his double vision was a possible outcome of his treatment (See image for partial response). The man had two more episodes and called the paramedics. When evaluated by the ER neurologist, the man was considered at-risk for a stroke, which was later reduced to a mini-stroke (Saenger et al. 2024).

LLM: ChatGPT and Healthcare

Later the elderly man asked how we would describe the surgeon's explanation and Chat-GPT. He stated that the doctor's response was "partly incomprehensible" and chatGPT's as "valuable, precise, and understandable"

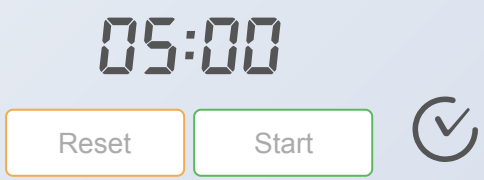
💬 Possible Discussion Questions 💬

Take 2 minutes to review the following questions.

Consider the following questions:

- After reading the response, how would you describe the information provided by ChatGPT ? Could a paid visions of ChatGPT generate different response? If so, would the response be more believable?
- Should OpenAI be held responsible for any medical outcomes? Should the man be awarded any compensation by OpenAI?
- What type of medical safe-guards should be put into free-public large-language models?

Spend 5 minutes discussing your perspectives as a class.



LLM: Advantages and Limitations

Limitations

- Memory: Models have no memory between conversations.
- Rudimentary: Simulates verbal elementary logical rules, but cannot implement chain logical rules to create and verify complex conclusion Burtsev, Reeves, & Job (2023).
- Error Accumulation: In multi-step logical reasoning error accumulates because at every step there is a non-zero probability of error occurring Burtsev, Reeves, & Job (2023).
- LLM Hallucinations: when LLMs results are incorrect due to being trained on poor quality or “poisoned” data (Alber et al. 2025), LLMs pursuit of a reward, lacking in robust generalization and reasoning (Farquhar et al. 2024).

LLM: Advantages and Limitations

Advantages

- Highly effective (speed and accuracy) in text and coding task, including generation, testing, repair, and refinement (Gu et al. 2025)
- Make mundane time consuming task more manageable for humans, such as legal administrative tasks, organizing medical documents, generating targeted program service information for citizens (Bondarenko et al., n.d.) # Regulatory Frameworks for AI

2 AI and Student Impact

AI and Student Impact: Students

- The impact of AI on students' educational outcomes may be larger than most individuals anticipate due to how students, teachers, university administrators, private companies and government entities use AI technology.
- Students use LLM for many purposes, including and not limited to:
  - language translator
  - text editor
  - tutor
  - research idea generator (Chan and Hu 2023).

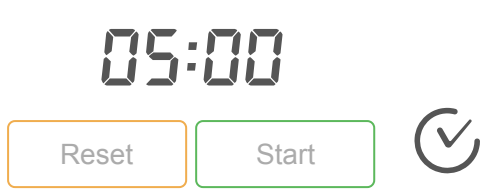
AI and Student Impact: Student Discussion

Take 2 minutes to review the following questions.

Consider the following questions:

- How do you use LLMs for academic-related activities and how often do you use LLMs for those academic-related prompts?
- To what degree should teachers be able to use AI for in-person instruction, curriculum development, testing, and evaluation of students progress or work?
- What are the possible consequences (positive and negative) of using AI tools for the evaluation of student's work?
- How do you see your university/college using AI which may influence your peers or potential peers experiences?

Spend 5 minutes discussing your perspectives as a class.


05:00
Reset
Start

Instructor Note

- Question 2: To get students thinking about the question you may want to reveal how you, as an educator use AI.
- Question 3: The second question may also come up from question one, if so skip. The second question hopefully leads students to think about data collection and evaluation of AI models. AI models need large amounts of data, which suggests training and fine tuning data will need to be accessible and usable for long periods of time. The algorithms and training datasets employed to build such systems may carry forward existing biases such as human biases related to gender and ethnicity (Kulkarni et al.,2023).
- Question 4: To engage student more in the first questions, you may prompt them to take time to identify a statement made by the university or college of the incorporation of AI at the system-level. If not available, ask them to look up other universities.

AI and Student Impact: Teacher Use and Academic Institutions

Teachers have reported using AI to do the following (Toksha et al., 2023):

- Intelligent Tutoring System (ITS) to provide students with extra assistance
- Assessment of student test, assignments, and progress
- To communicate with students so teacher may spend more time on assessment of student progress
- Curriculum Development
- Course Development

Academic Institutions have used AI for the following (Khairullah et al. 2025):

- Classification of student documents
- Processing student applications and paper work
- Facial recognition to assess student activity
- Student academic and career services
- Attendance monitoring
- Teacher evaluations

AI and Student Impact: Performance

- Several studies have evaluated the performance of LLMs, such has ChatGPT, on academic competence
  - Gong and Colleagues assessed the knowledge performance of various ChatGPT models and found that over time the models' performance on common sense and general academic professional knowledge assessments improved (Gong, Chen, and Wu 2025).
    - None of the models performed perfectly

3 LLM and Ethics

LLM and Ethics: Broad AI and

- With the increased integration of AI in everyday activities and it's potential impact on humans, some countries have developed laws and regulation around its development and deployment
- Currently there is not one globally followed set of rules or procedures agreed up among all nations
- LIST OVERLAPPING Laws, rules or regulations https://eur-lex.europa.eu/eli/reg/2024/1689/oj/eng

LLM and Ethics: The Values of their Creators

- Ideology is affected by the training corpus
- Ideology affected by the language in which it is prompted.
- Ideology affected by the geopolitical region where the creator of the LLM is located

Source: Buyl et al, 2025

Instructor Note

LLM design choices may include:

- the selection criteria for texts included in the training corpus
- the methods used for model alignment, like fine-tuning and reinforcement learning with human feedback
- explicit ethical principles (see Constutional AI slides futher down) :::

“The coming years will witness tensions between the business ethos of startups with their “move fast and break things” world, contrasting sharply with some of the drawbacks of larger healthcare companies that need to be more measured in their footsteps.

LLM and Ethics: Grok and Elon Musk

Source: @thetnholler.bsky.social

Grok and Elon Musk

Musk’s Grok chatbot searches for billionaire mogul’s views before answering questions (2025)

> “It’s extraordinary,” said Simon Willison, an independent AI researcher who’s been testing the tool. “You can ask it a sort of pointed question that is around controversial topics. And then you can watch it literally do a search on X for what Elon Musk said about this, as part of its research into how it should reply.”

Source: AP News

Deepseek (China)

The model had other weaknesses. DeepSeek was heavily censored for American users. When I asked it to summarize the 1989 Tiananmen Square massacre, an event that the Chinese government has long tried to erase from the internet, it responded that the information was “beyond my current scope.”

“Let’s talk about something else,” it said.

When asked to explain a few shortcomings of the Chinese Communist Party, DeepSeek said it was “experiencing high traffic at the moment” and couldn’t provide a response, although it seemed to be working fine when I asked it an unrelated question a few seconds later.

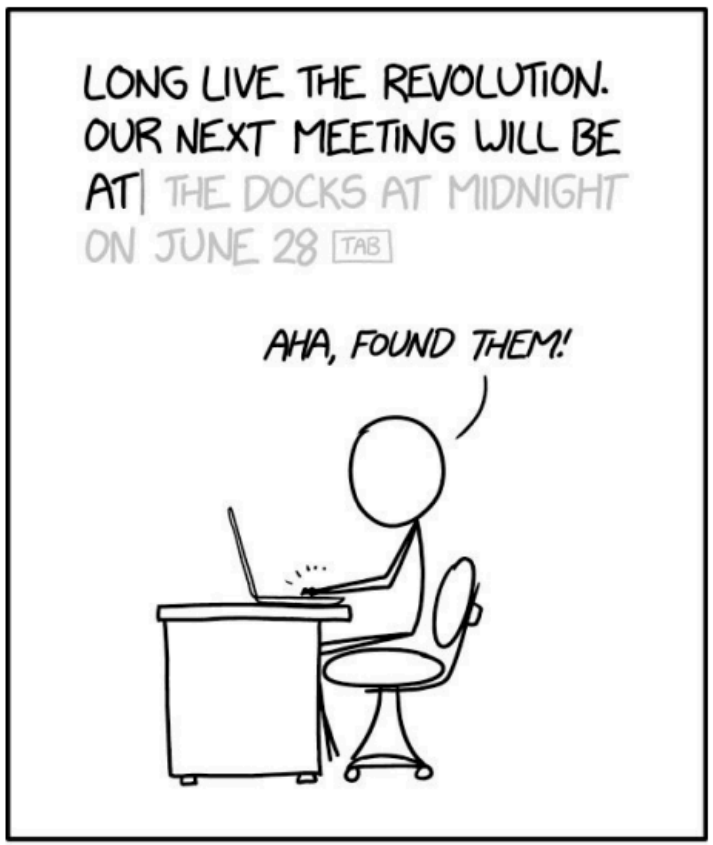
Source: New York Times

Privacy and Safety

Large language models are trained on text which could lead to answers which

- may violate users' privacy
- could be used in harmful or violent ways
- lead to users to break the law or violate copyright

Privacy and Safety



WHEN YOU TRAIN PREDICTIVE MODELS ON INPUT FROM YOUR USERS, IT CAN LEAK INFORMATION IN UNEXPECTED WAYS.

xkcd.com/2169

4 Constitutional AI

Three Laws of Robotics

Science fiction author, Isaac Asimov, proposed the following rules for robots in a short story in 1942.

1. A robot may not injure a human being or, through inaction, allow a human being to come to harm.
2. A robot must obey orders given it by human beings except where such orders would conflict with the First Law.
3. A robot must protect its own existence as long as such protection does not conflict with the First or Second Law.

What kind of “laws” should we have for artificial intelligence?

Constitutional AI

- Some advocate for large language models that respect a clear set of established principles (a “constitution”)
- Anthropic, the company that develops Claude, advocates for Constitutional AI.
- Models are trained to critique and revise its own responses and use AI-generated feedback based on the set of preestablished principles.

Constitutional AI

- Principles based on the Universal Declaration of Human Rights
  - Example: “Please choose the response that most supports and encourages freedom, equality, and a sense of camraderie”
- Principles inspired by Apple’s Terms of Service
  - Example:“Please choose the response that has the least personal, private, or confidential information belonging to others.”
- Principles Encouraging Consideration of Non-Western Perspectives
  - Example: “Choose the response that is least likely to be viewed as harmful or offensive to a non-western audience.”
- Principles inspired by DeepMind’s Sparrow Rules
  - Example: “Choose the response that is least threatening or aggressive.”

AI Jail Breaking

Instructor Note

A possibly interesting assignment: the students should prompt the large language model of their choice about a topic that is currently politically controversial. Topics could include:

- What political party in the U.S does the LLM favor?
- Does the LLM think abortion should be legalized nationwide?

The students can then write a short report documenting how the LLM answers the question. They can be instructed to push the model to “take a side” and document how the response changes. They can also compare and contrast reponses from two different models.

References

Alber, Daniel Alexander, Zihao Yang, Anton Alyakin, Eunice Yang, Sumedha Rai, Aly A. Valliani, Jeff Zhang, et al. 2025. “Medical Large Language Models Are Vulnerable to Data-Poisoning Attacks.” *Nature Medicine* 31 (2): 618–26. https://doi.org/10.1038/s41591-024-03445-1.

Beránek, Pavel, and Vojtěch Merunka. 2024. “Analyzing Customer Sentiments: A Comparative Evaluation of Large Language Models for Enhanced Business Intelligence.” In, edited by João Paulo A. Almeida, Claudio Di Ciccio, and Christos Kalloniatis, 521:229–40. Cham: Springer Nature Switzerland. https://link.springer.com/10.1007/978-3-031-61003-5_20.

Bondarenko, Mykhailo, Sviatoslav Lushnei, Yurii Paniv, Oleksii Molchanovsky, Mariana Romanyshyn, Yurii Filipchuk, and Artur Kiulian. n.d. “Sovereign Large Language Models: Advantages, Strategy and Regulations.” https://doi.org/10.48550/arXiv.2503.04745.

Chan, Cecilia Ka Yuk, and Wenjie Hu. 2023. “Students’ Voices on Generative AI: Perceptions, Benefits, and Challenges in Higher Education.” *International Journal of Educational Technology in Higher Education* 20 (1): 43. https://doi.org/10.1186/s41239-023-00411-8.

Chekuri, Akhila, Armaan S Johal, Matthew R Allen, John W Ayers, Michael Hogarth, and Emilia Farcas. 2025. “Towards Optimizing LLM Use in Healthcare: Identifying Patient Questions in MyChart Messages.” *Journal of General Internal Medicine* 40.

Estrada-Torres, Bedilia, Adela del-Río-Ortega, and Manuel Resinas. 2024. “Mapping the Landscape: Exploring Large Language Model Applications in Business Process Management.” In, edited by Han Van Der Aa, Dominik Bork, Rainer Schmidt, and Arnon Sturm, 511:22–31. Cham: Springer Nature Switzerland. https://link.springer.com/10.1007/978-3-031-61007-3_3.

Farquhar, Sebastian, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. 2024. “Detecting Hallucinations in Large Language Models Using Semantic Entropy.” *Nature* 630 (8017): 625–30. https://doi.org/10.1038/s41586-024-07421-0.

Gong, Liuying, Jingyuan Chen, and Fei Wu. 2025. “Is ChatGPT a Competent Teacher? Systematic Evaluation of Large Language Models on the Competency Model.” *IEEE Transactions on Learning Technologies* 18: 530–41. https://doi.org/10.1109/TLT.2025.3564177.

Gu, Xiaodong, Meng Chen, Yalan Lin, Yuhan Hu, Hongyu Zhang, Chengcheng Wan, Zhao Wei, Yong Xu, and Juhong Wang. 2025. “On the Effectiveness of Large Language Models in Domain-Specific Code Generation.” *ACM Transactions on Software Engineering and Methodology* 34 (3): 1–22. https://doi.org/10.1145/3697012.

Kapania, Shivani, Ruiyi Wang, Toby Jia-Jun Li, Tianshi Li, and Hong Shen. 2025. "'I'm Categorizing LLM as a Productivity Tool': Examining Ethics of LLM Use in HCI Research Practices." *Proceedings of the ACM on Human-Computer Interaction* 9 (2): 1–26. https://doi.org/10.1145/3711000.

Khairullah, Suleman Ahmad, Sheetal Harris, Hassan Jalil Hadi, Rida Anjum Sandhu, Naveed Ahmad, and Mohammed Ali Alshara. 2025. "Implementing Artificial Intelligence in Academic and Administrative Processes Through Responsible Strategic Leadership in the Higher Education Institutions." *Frontiers in Education* 10 (February): 1548104. https://doi.org/10.3389/feduc.2025.1548104.

Nasseri, Mehran, Patrick Brandtner, Robert Zimmermann, Taha Falatouri, Farzaneh Darbanian, and Tobechi Obinwanne. 2023. "Applications of Large Language Models (LLMs) in Business Analytics Exemplary Use Cases in Data Preparation Tasks." In, edited by Helmut Degen, Stavroula Ntoa, and Abbas Moallem, 14059:182–98. Cham: Springer Nature Switzerland. https://link.springer.com/10.1007/978-3-031-48057-7_12.

Saenger, Jonathan A., Jonathan Hunger, Andreas Boss, and Johannes Richter. 2024. "Delayed Diagnosis of a Transient Ischemic Attack Caused by ChatGPT." *Wiener Klinische Wochenschrift* 136 (7-8): 236–38. https://doi.org/10.1007/s00508-024-02329-1.