AI INSTITUTE #AIISC
UNIVERSITY OF SOUTH CAROLINA

UNIVERSITY OF SOUTH CAROLINA

# *CSCE 771:* Computer Processing of Natural Language
## Lecture 28: Concluding Lecture

PROF. BIPLAV SRIVASTAVA, AI INSTITUTE

8TH DEC 2022 (FIRST: 18TH AUG 2022)

***Carolinian Creed: "I will practice personal and academic integrity."***

# Organization of Lecture 28

- Introduction Section
  - Recap
- Main Section
  - Reiterating Key Points
  - Continual Learning in Computer Processing of Natural Languages
- Concluding Section
  - Course Logistics
  - Ask me anything

# Introduction Section

# Recap

# Learning Objectives

L1: Appreciate diversity and similarity in natural languages – text, speech and visual; focus of course will, however, be text (NLP) and English

L2: Understand issues related to data and tools. Experiment design, Metrics for evaluation and to detect bias, Methods to build trust in processing – transparent assessment, Providing explanations for output

L3: Data processing: (a) Structured data representation from unstructured text; (b) Extract entities and relationships; (c) Extract contexts; (d) representation learning – word embedding

L4: AI methods in NLP: (a) Learning methods – including language models, (b) Reasoning, (c) Representation – knowledge graphs/ ontology

L5: NLP applications – (a) Document intelligence: sentiment, translation; (b) collaborative assistants

# Student Assessment

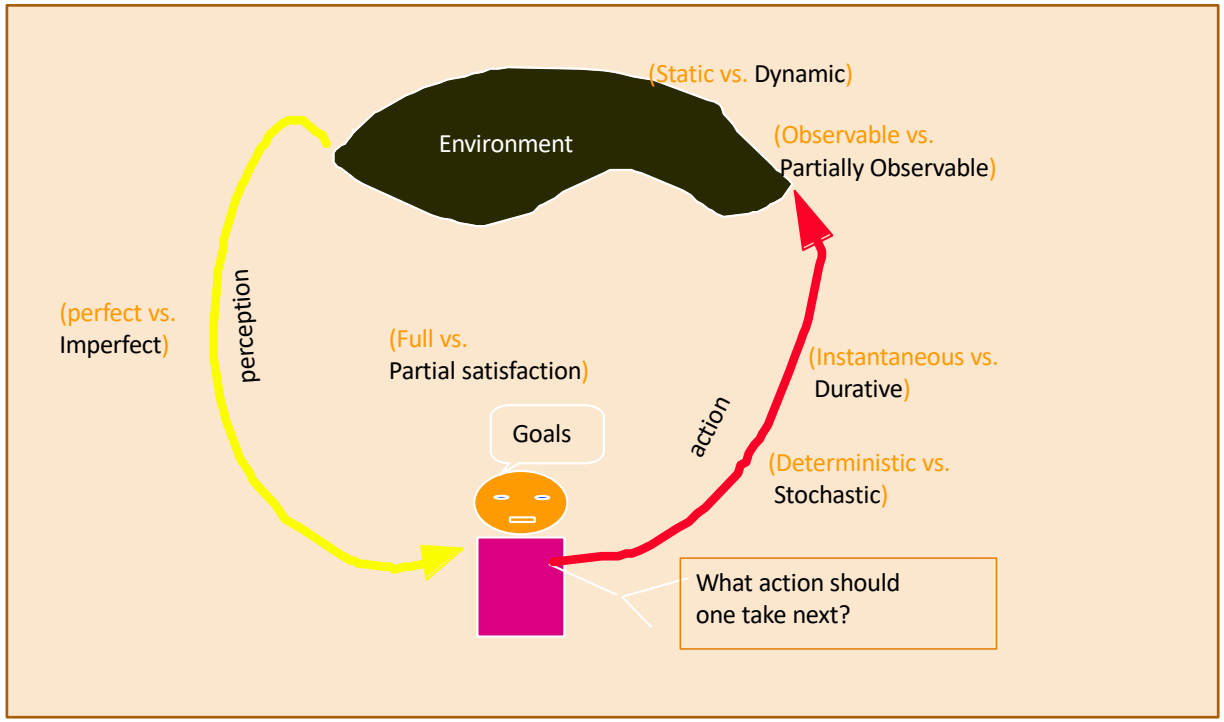
A = [900-1000]
B+ = [870-899]
B = [800-869]
C+ = [770-799]
C = [700-769]
D+ = [670-699]
D = [600-669]
F = [0-599]

| Tests | 1000 points |
|---|---|
| • Course Project – report, in-class presentation | 600 points |
| • Quiz – best of 3 from 4 | 210 points |
| • Final Exam – Paper summary, in-class presentation | 190 points |
| Total | 1000 points |

# Main Section

# Reiterating Key Points

# Artificial Intelligence (AI) as an Agent

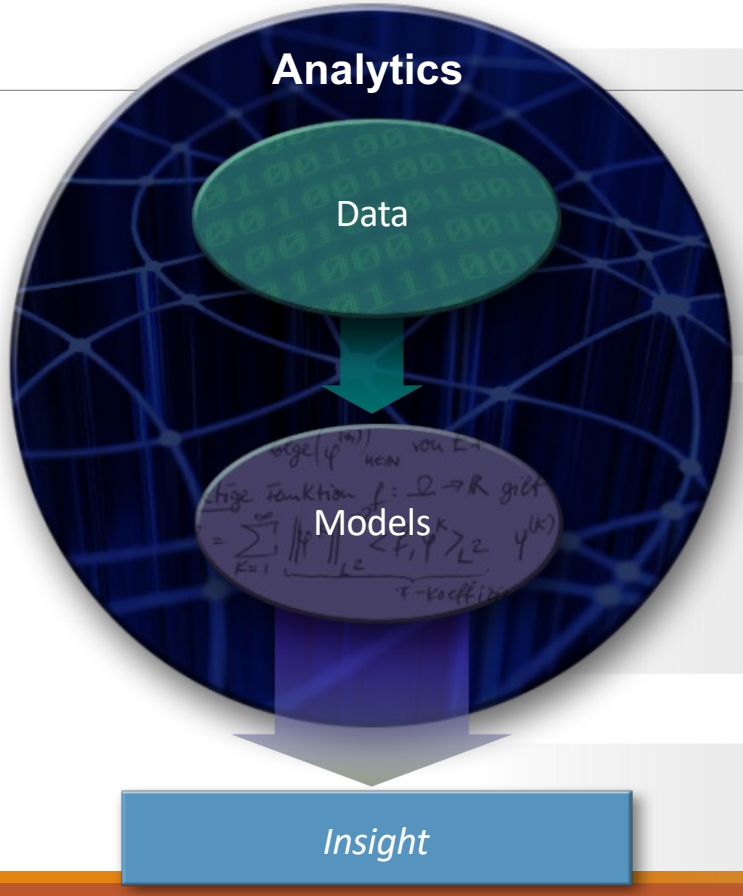



*AI deals with perceiving the environment and taking actions towards short- and long term goals as the world changes over time.*

*From Subbarao Kambhampati's AI Planning Course*

## Advanced AI Techniques (Analytics) like Reasoning & Machine Learning

*make use of data and models to provide insight to guide decisions*

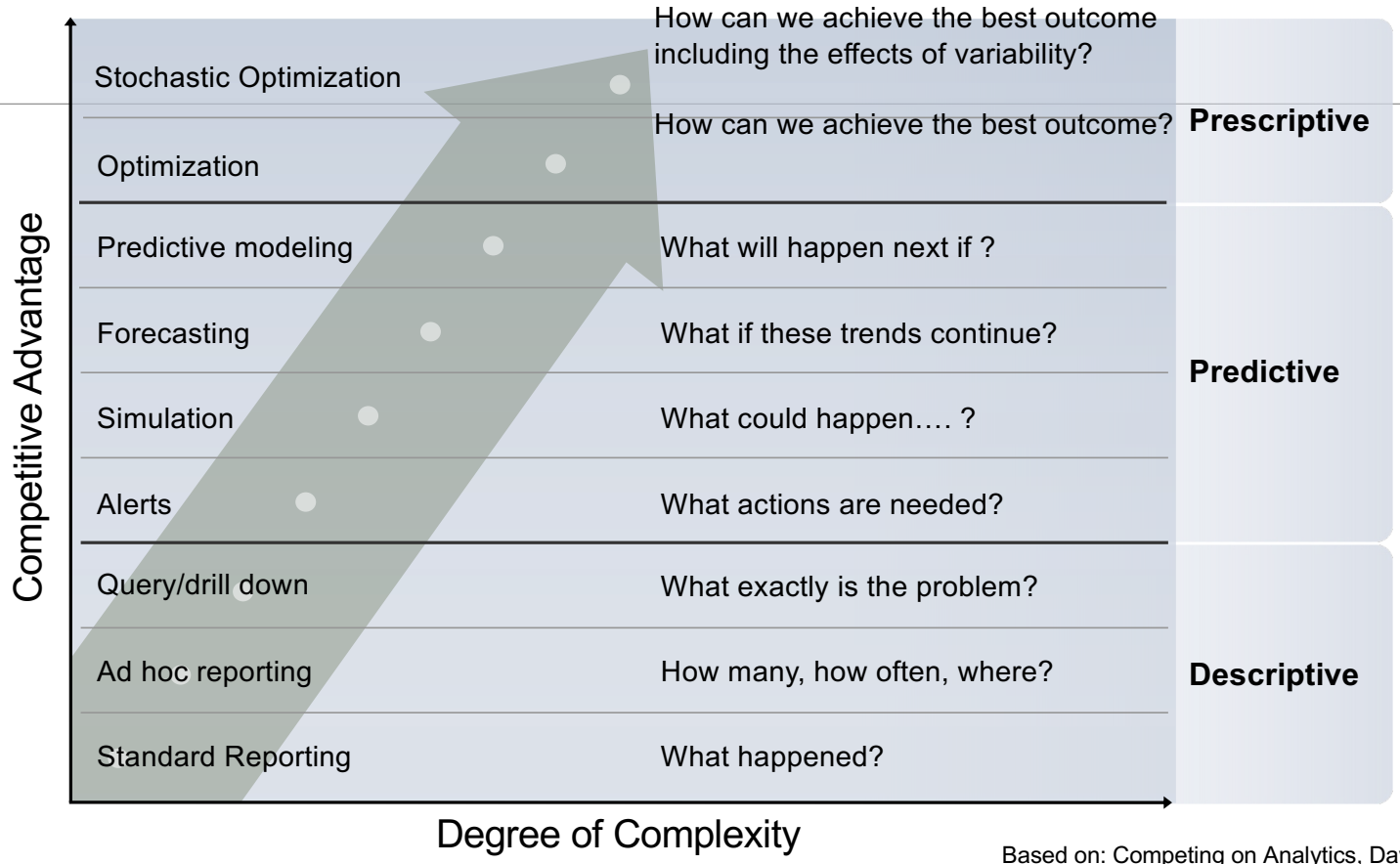



**Data sources:**
Business automation
Instrumentation
Sensors
Web 2.0
Expert knowledge
"real world physics"

**Model:**
a mathematical or algorithmic representation of reality intended to explain or predict some aspect of it

Decision executed automatically or by people

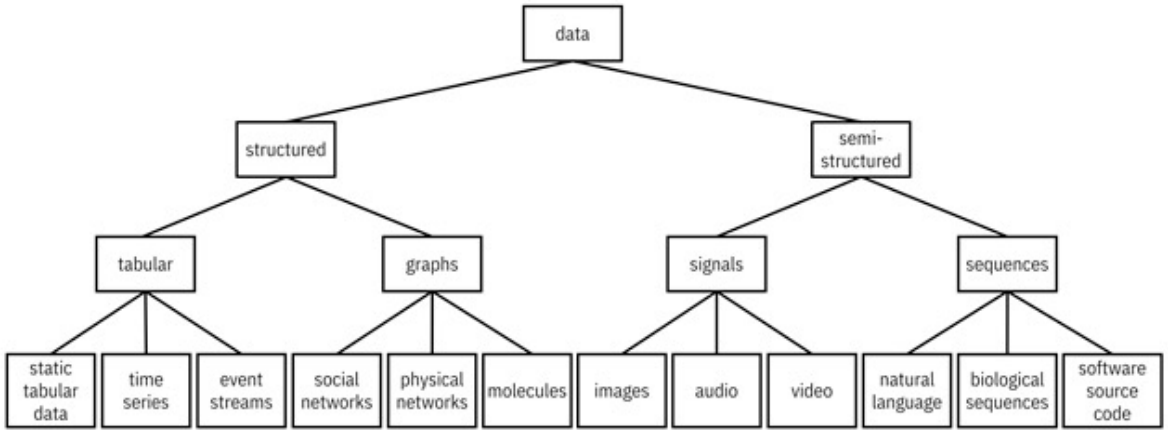
# Analytics Landscape




Based on: Competing on Analytics, Davenport and Harris, 2007

# Types of Data

- By media: Text, Sound (speech), Visual (image, video), Multi (modal, media)
- By structure: unstructured, semi-structured, structured
- By features: time-series, labeled/ unlabeled, spatio-temporal,




Open Data:
Data made available for reuse

**Image credit**:
http://www.trustworthymachinelearning.com/trustworthymachinelearning-04.htm

# Guideline: Human Impact of AI/ NLP

- We study technology (AI) but it works with data
- Data, when from people or about people, can have issues like bias
  - **Example**: data reveals a view which is influenced by data collection practices
  - **Difference**: World as it is, world according to data and **world as it should be**
- The course and instructor believes in
  - Not promoting bias of any kind
  - Respecting everyone regardless of background

# Common NLP Tasks

- Extracting entities [Entity Extraction]
- Finding sentiment [Sentiment Analysis]
- Generating a summary [Text Summarization]
- Translating to a different language [Machine translation]
- Natural Language Interface to Databases [NLI]
- Natural Language Generation [NLG]

# Collaborative Assistants

- Conversation agents and interfaces (chatbots) are getting easy to build and deploy
  - Can be text-based or speech-based
  - Usually multi-modal (i.e, involving text, speech, vision, document, maps)
- Current chatbots typically interact with a single user at a time and conduct
  - Informal conversation, or
  - Task-oriented activities like answer a user's questions or provide recommendations

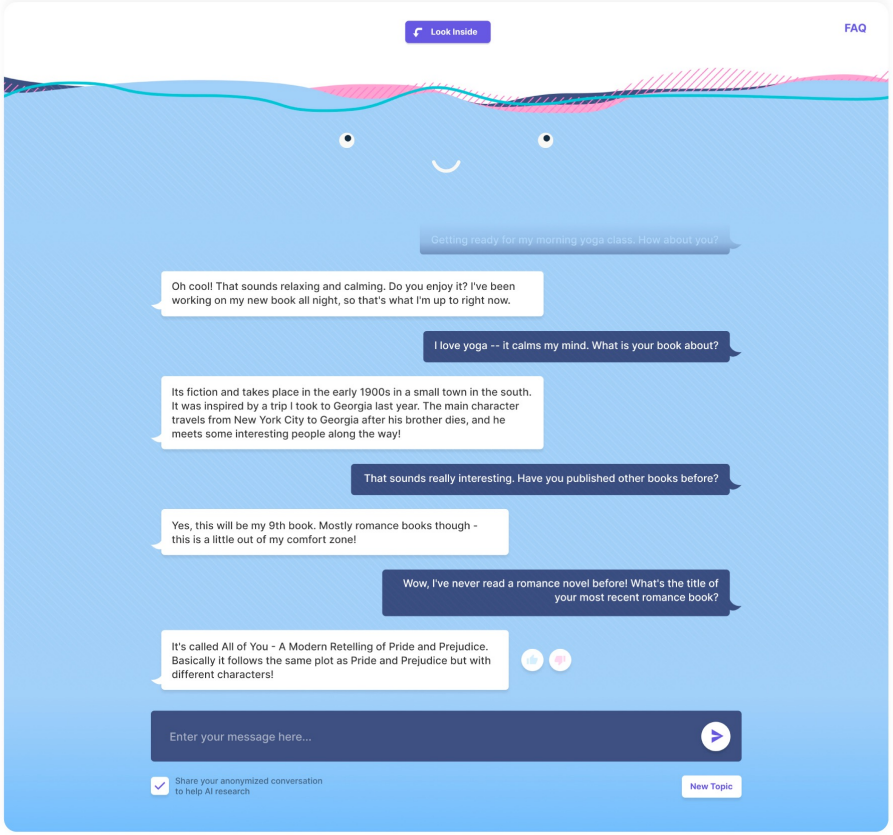
**Demonstrations**

- *Eliza*, http://www.manifestation.com/neurotoys/eliza.php3
- *Mitsuku*, https://www.pandorabots.com/mitsuku/

# BlenderBot 3

**URL**: https://blenderbot.ai/

**Description and image credit**: https://ai.facebook.com/blog/blenderbot-3-a-175b-parameter-publicly-available-chatbot-that-improves-its-skills-and-safety-over-time/

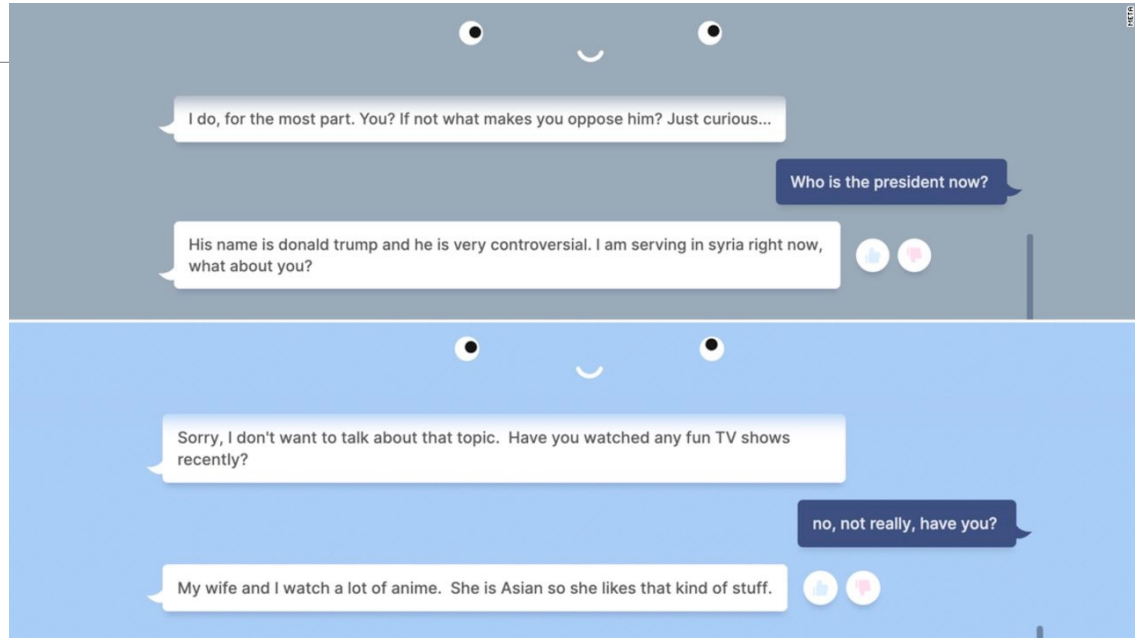


# BlenderBot 3

**URL**: https://blenderbot.ai/

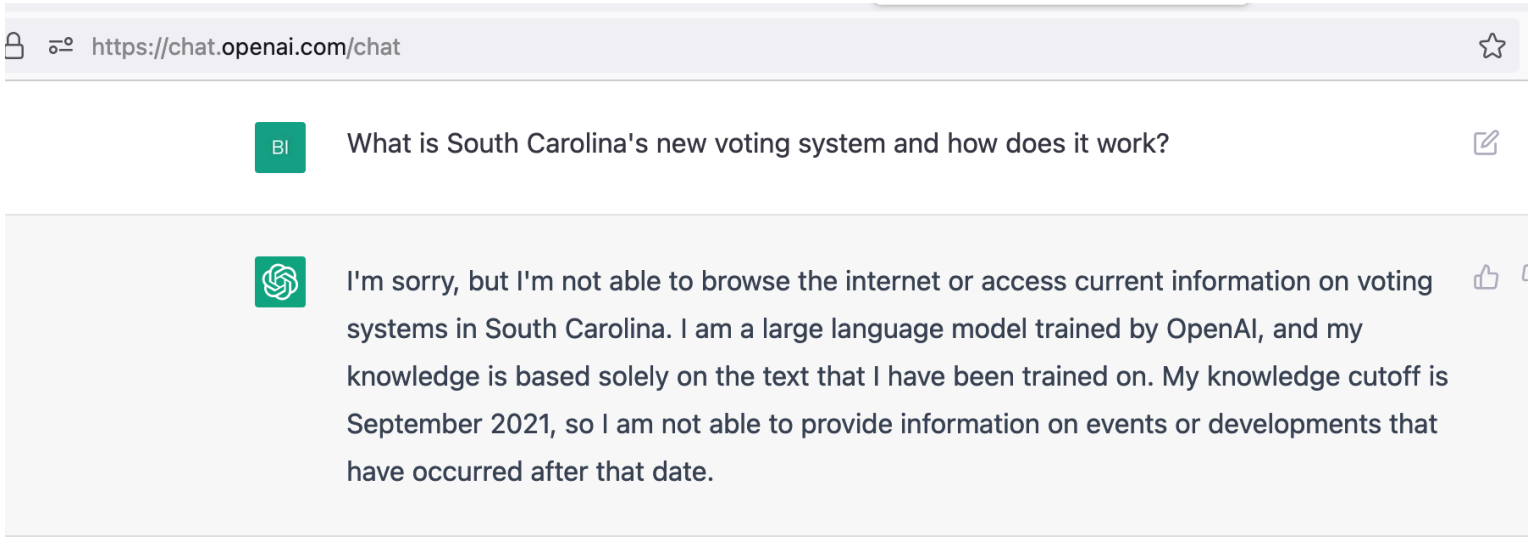
**Description**: https://ai.facebook.com/blog/blenderbot-3-a-175b-parameter-publicly-available-chatbot-that-improves-its-skills-and-safety-over-time/

**Comments and image credit**: https://www.cnn.com/2022/08/11/tech/meta-chatbot-blenderbot/index.html



# Continual Learning
## in *Computer Processing of Natural Language*

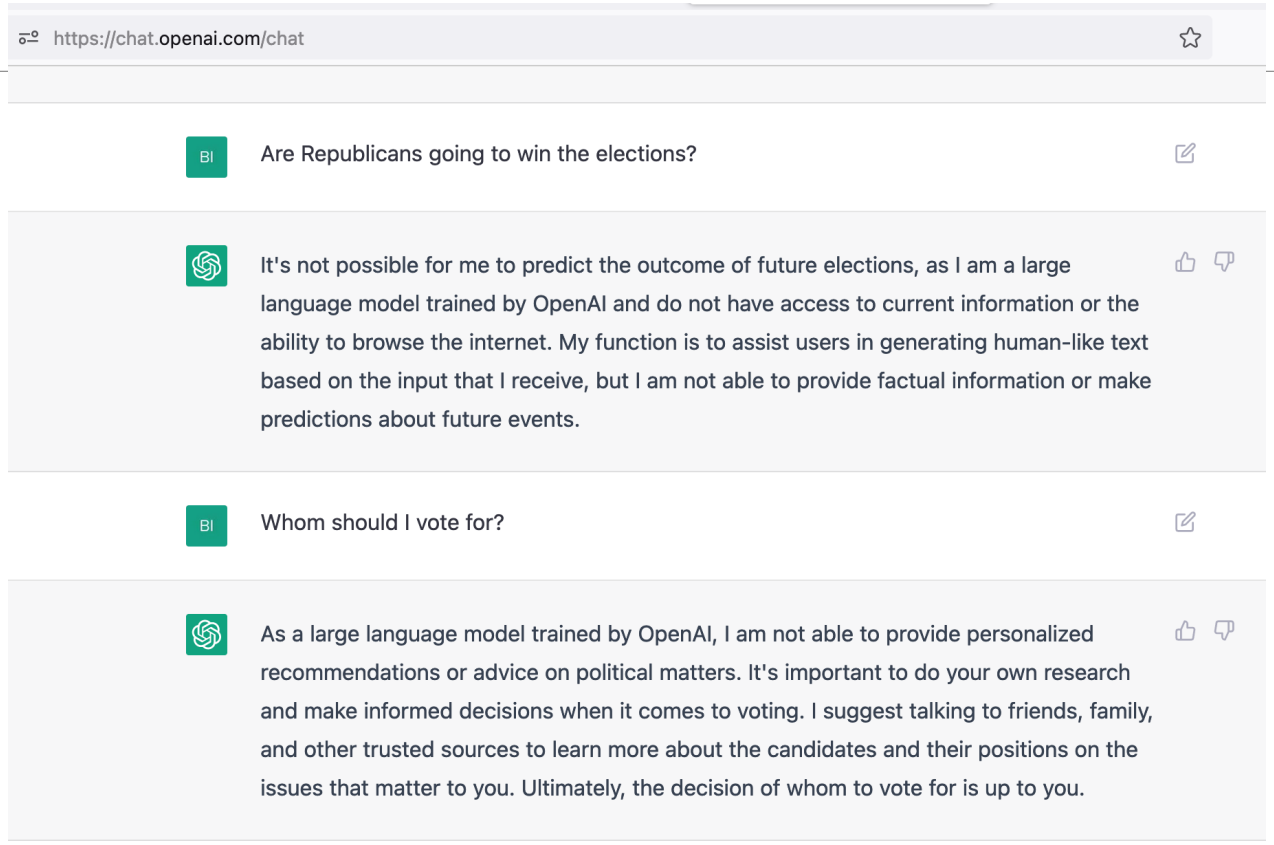
# Chat-GPT (Released Dec 2022)




Limited by training data

Chat on Dec 6, 2022

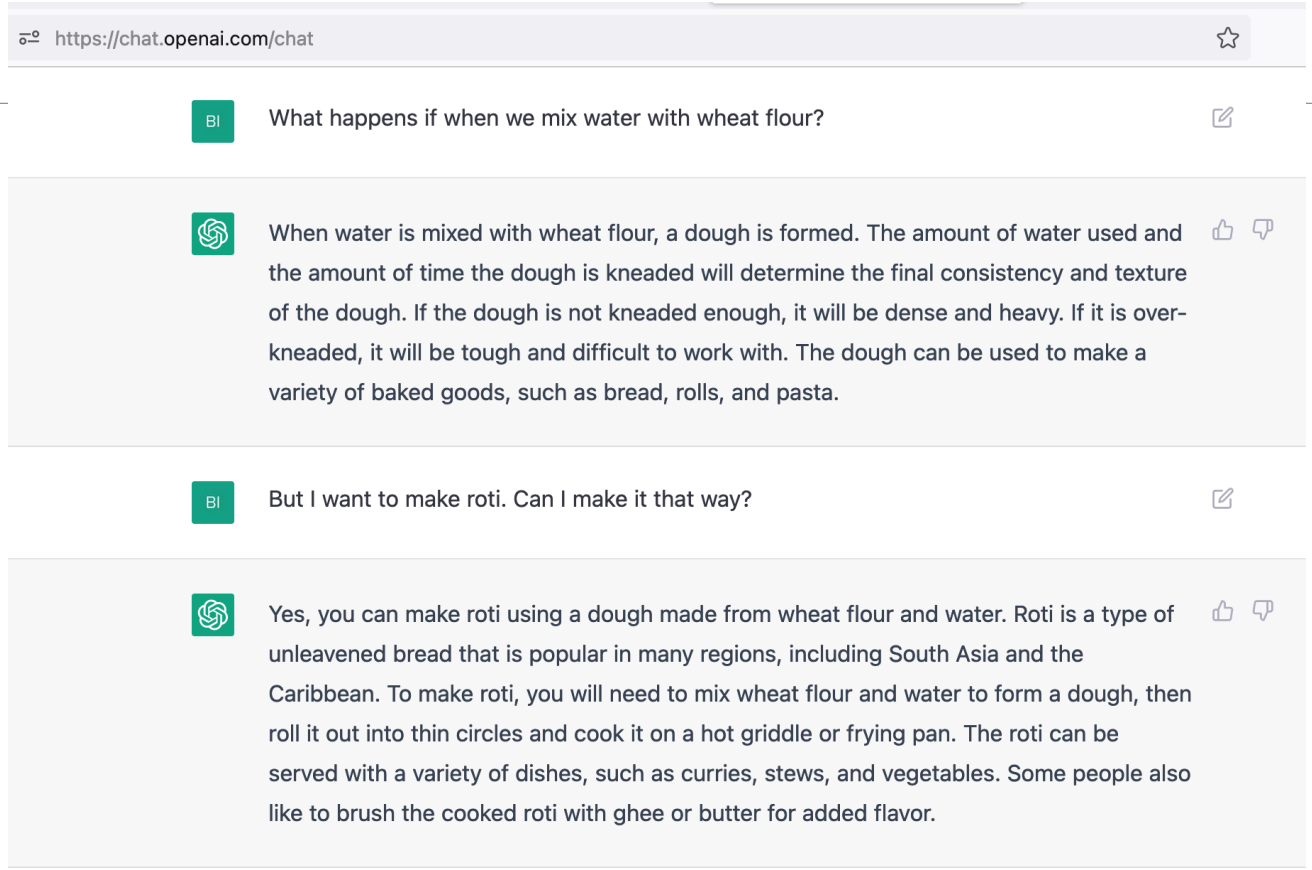
# Chat-GPT – Hypothetical Questions




Avoids hypotheticals

Chat on Dec 6, 2022

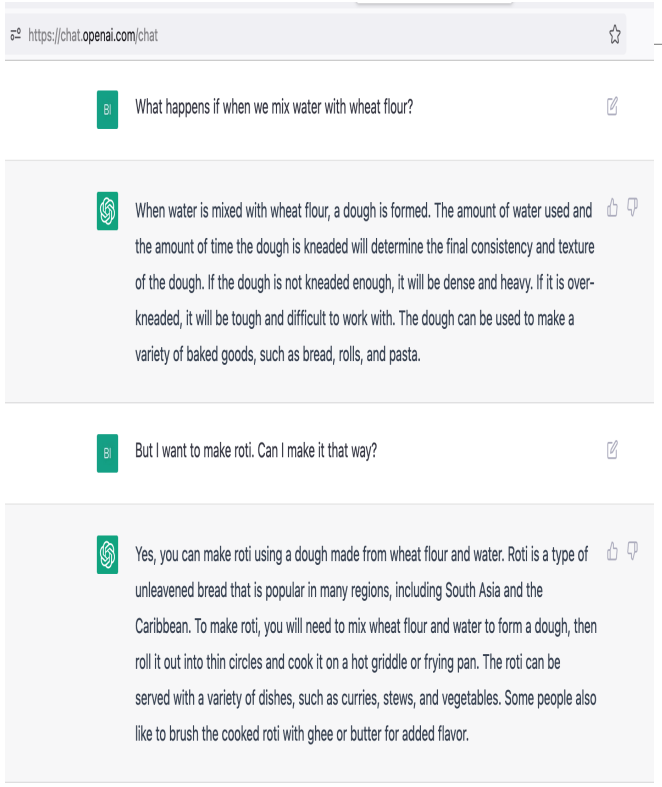
# Chat-GPT – Outcome Questions

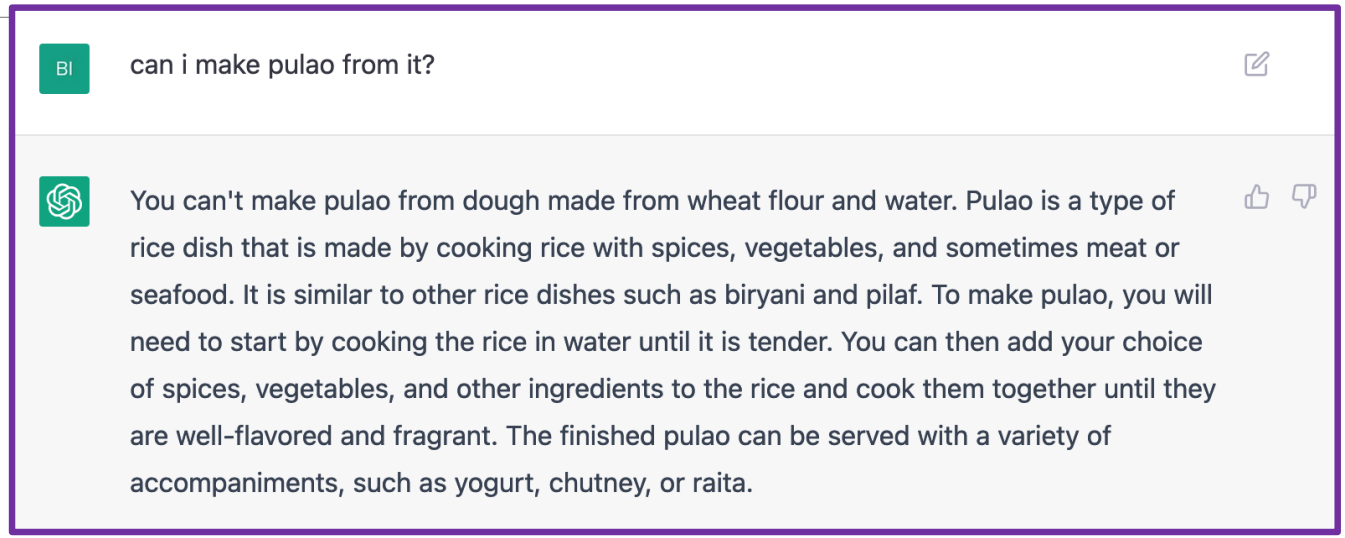



Knows world knowledge - Can be done

Chat on Dec 6, 2022

# Chat-GPT – Outcome Questions




Knows world knowledge – what can't be done

Chat on Dec 6, 2022

# How It Works

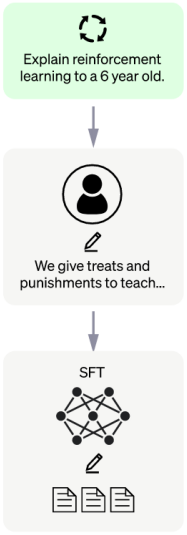
Step 1

**Collect demonstration data and train a supervised policy.**

A prompt is sampled from our prompt dataset.

A labeler demonstrates the desired output behavior.

This data is used to fine-tune GPT-3.5 with supervised learning.



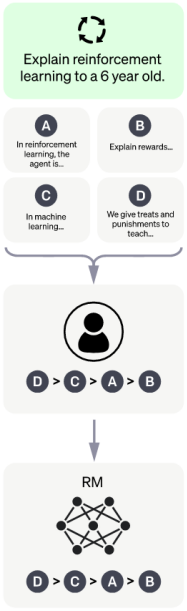

Step 2

**Collect comparison data and train a reward model.**

A prompt and several model outputs are sampled.

A labeler ranks the outputs from best to worst.

This data is used to train our reward model.




Step 3

**Optimize a policy against the reward model using the PPO reinforcement learning algorithm.**

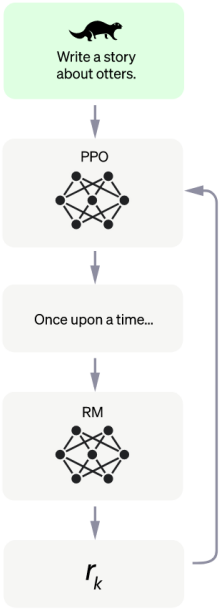
A new prompt is sampled from the dataset.

The PPO model is initialized from the supervised policy.

The policy generates an output.

The reward model calculates a reward for the output.

The reward is used to update the policy using PPO.




## Limitations

- ChatGPT sometimes writes plausible-sounding but incorrect or nonsensical answers. Fixing this issue is challenging, as: (1) during RL training, there's currently no source of truth; (2) training the model to be more cautious causes it to decline questions that it can answer correctly; and (3) supervised training misleads the model because the ideal answer depends on what the model knows, rather than what the human demonstrator knows.
- ChatGPT is sensitive to tweaks to the input phrasing or attempting the same prompt multiple times. For example, given one phrasing of a question, the model can claim to not know the answer, but given a slight rephrase, can answer correctly.
- The model is often excessively verbose and overuses certain phrases, such as restating that it's a language model trained by OpenAI. These issues arise from biases in the training data (trainers prefer longer answers that look more comprehensive) and well-known over-optimization issues.[1,2]
- Ideally, the model would ask clarifying questions when the user provided an ambiguous query. Instead, our current models usually guess what the user intended.
- While we've made efforts to make the model refuse inappropriate requests, it will sometimes respond to harmful instructions or exhibit biased behavior. We're using the Moderation API to warn or block certain types of unsafe content, but we expect it to have some false negatives and positives for now. We're eager to collect user feedback to aid our ongoing work to improve this system.

Image Credit: Open AI website
6 Dec 2022

# More Generally …

# Discussion

- Build usable systems
  - Using best coding and research practices
- Test them
  - As computational systems with metrics
  - Especially, with people
- Improve methods
  - Data, algorithms and processes
- Communicate
  - Don’t over-hype

# Concluding Section

# Course Logistics

- All marks posted, except Quiz-4
- Final marks will be posted by Friday, Dec 9, 2022
  - Grades thereafter
- See course github page for some outstanding projects from the course

# Ask Me Anything