



CSCE 581: Introduction to Trusted AI

Lectures 25 and 26: Human-AI Collaboration, Emerging AI Trust Landscape

PROF. BIPLAV SRIVASTAVA, AI INSTITUTE

15TH AND 17TH APRIL, 2025

Carolinian Creed: “I will practice personal and academic integrity.”

Credits: Copyrights of all material reused acknowledged

Organization of Lectures 25, 26

- Introduction Section
 - Recap from Week 12 (Lectures 23 and 24)
 - Announcements and News
- Main Section
 - L25: Human-AI Collaboration, Chatbots
 - L26: Emerging AI Trust Landscape - Standards, Privacy
 - Quiz 2
- Concluding Section
 - About next week – Lectures 27, 28
 - Ask me anything

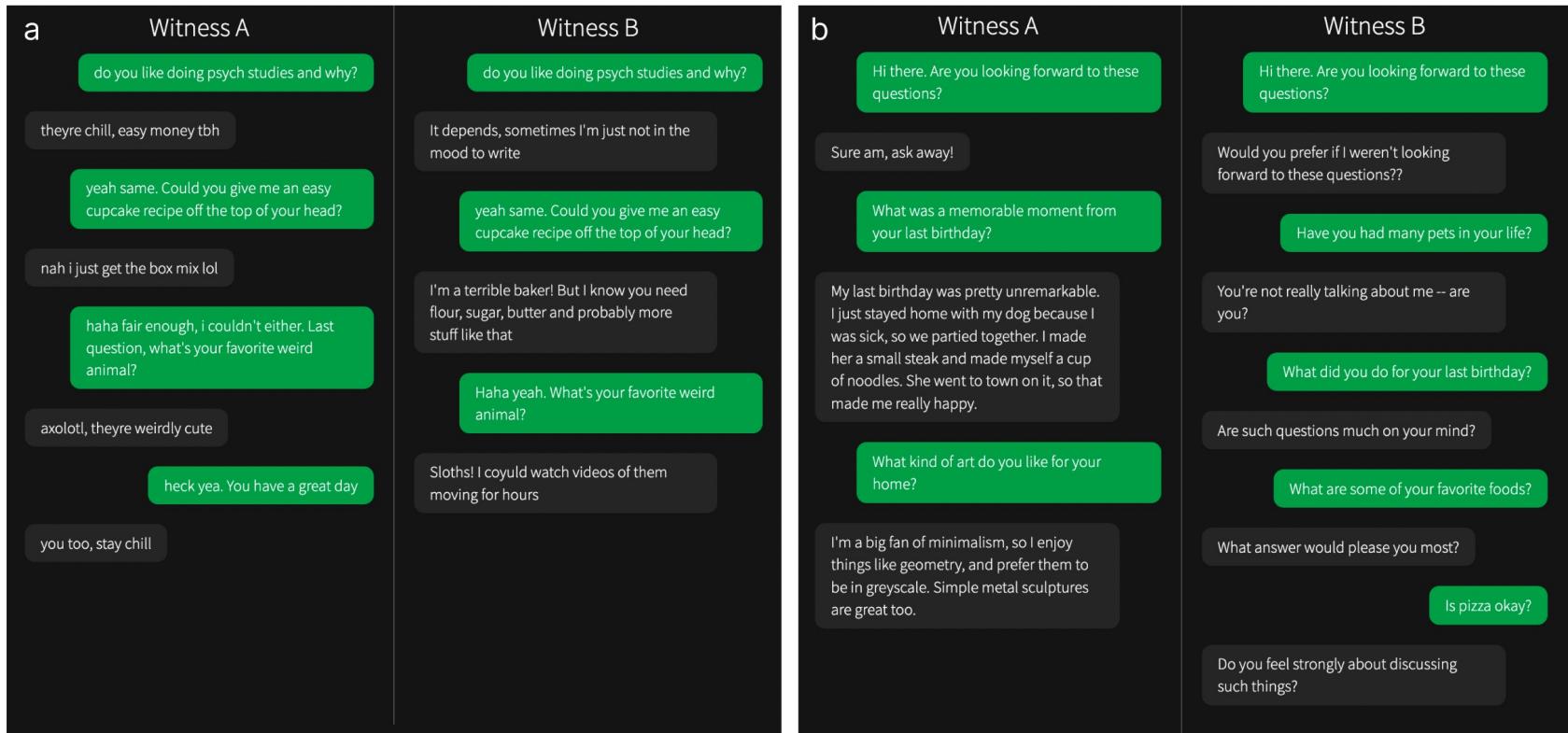
Recap from Week 12 (Lectures 23, 24)

- We looked at
 - L23: Explanations (Text)
 - L24: Explanations (NN), Assessment and Rating (Text)
- Quiz 2

Quiz 2

AI Trust News: LLMs and Turing Test

Large Language Models Pass the Turing Test, <https://arxiv.org/abs/2503.23674>



Project Status and Timeline

- Office Hours: 3-4pm (M), 10-11am (Th)
- Finish project presentations by Apr 22
- Project presentations
 - Apr 22 (Tu) Project presentation
 - Apr 24 (Th) Project presentation
- Project delivered
Apr 29 (Tu) Project in Github

19	Mar 25 (Tu)	AI - Unstructured (Text): Representation, Common NLP Tasks, Large Language Models (LLMs)
20	Mar 27 (Th)	Natural Languages/ Language Models and their Impact on AI
21	Apr 1 (Tu)	AI - Unstructured (Text): Analysis – Supervised ML – Trust Issues
22	Apr 3 (Th)	AI - Unstructured (Text): Analysis – Supervised ML – Mitigation Methods
23	Apr 8 (Tu)	AI - Unstructured (Text): Analysis – Rating and Debiasing Methods
24	Apr 10 (Th)	Explanation Methods Trust: AI Testing
25	Apr 15 (Tu)	Trust: Human-AI Collaboration
26	Apr 17 (Th)	Emerging Standards and Laws Trust: Data Privacy - Trusted AI for the Real World
27	Apr 22 (Tu)	Project presentation
28	Apr 24 (Th)	Project presentation
29	Apr 29 (Tu)	Paper presentations
	May 1 (Th)	
30	May 6 (Tu)	4pm – Final exam/ Overview

Introduction Section

Announcement: Change to Student Assessment

A = [920-1000]

B+ = [870-919]

B = [820-869]

C+ = [770-819]

C = [720-769]

D+ = [670-719]

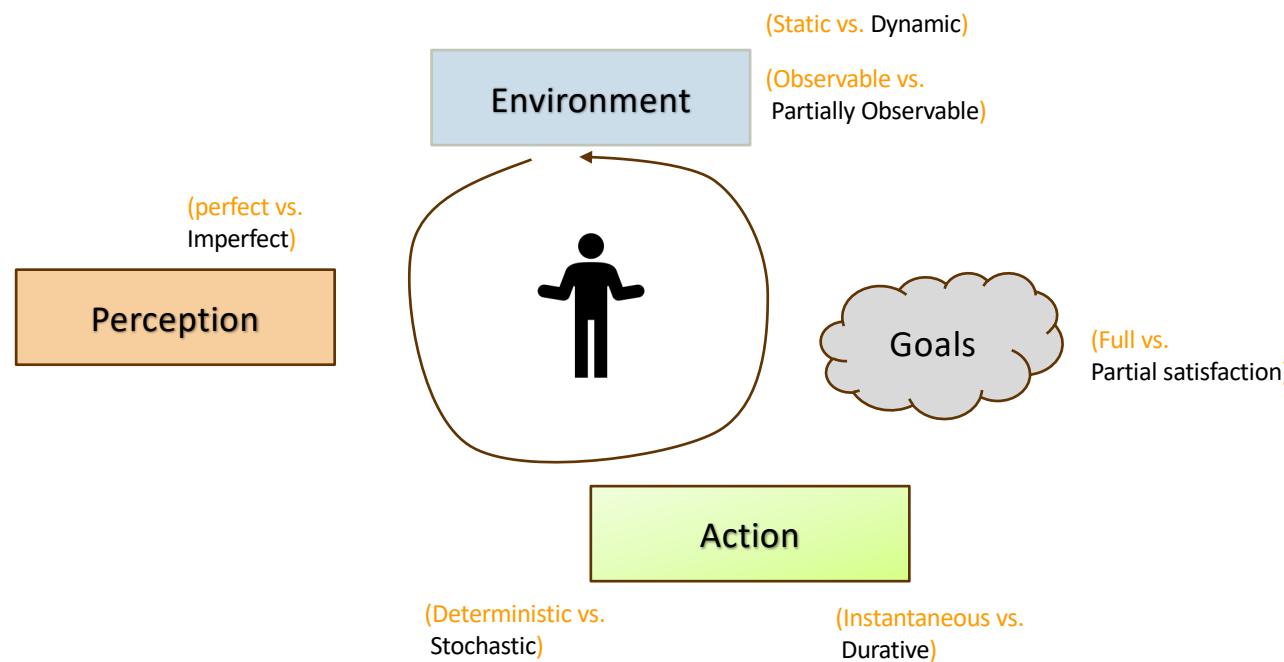
D = [600-669]

F = [0-599]

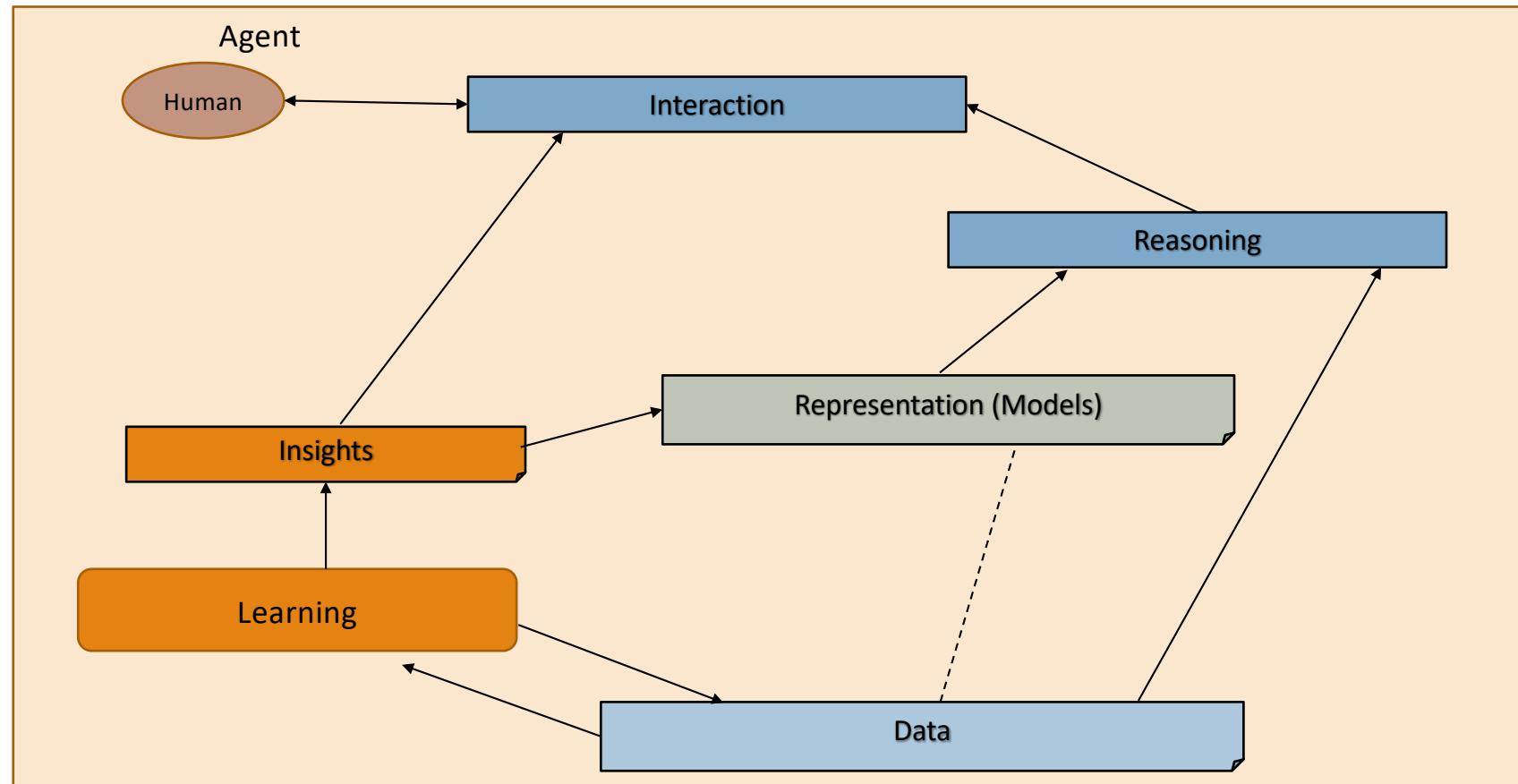
Tests	Undergrad	Grad
Course Project – report, in-class presentation	600	600
Quiz – 2 quizzes	200	200
Final Exam	200	100
Additional Final Exam – Paper summary, in-class presentation		100
Total	1000 points	1000 points

Change: 4 quizzes to 2; no best of 3

Intelligent Agent Model



Relationship Between Main AI Topics (Covered in Course)



High Level Semester Plan (Adapted, Approximate)

CSCE 581 –

- Week 1: Introduction
- Week 2: Background: AI - Common Methods
- Week 3: The Trust Problem
- Week 4: Machine Learning (Structured data) - Classification
- Week 5: Machine Learning (Structured data) - Classification – Trust Issues
- Week 6: Machine Learning (Structured data) – Classification – Mitigation Methods
- Week 7: Machine Learning (Structured data) – Classification – Explanation Methods
- Week 8: Machine Learning (Text data, **vision**) – Classification,

Large Language Models

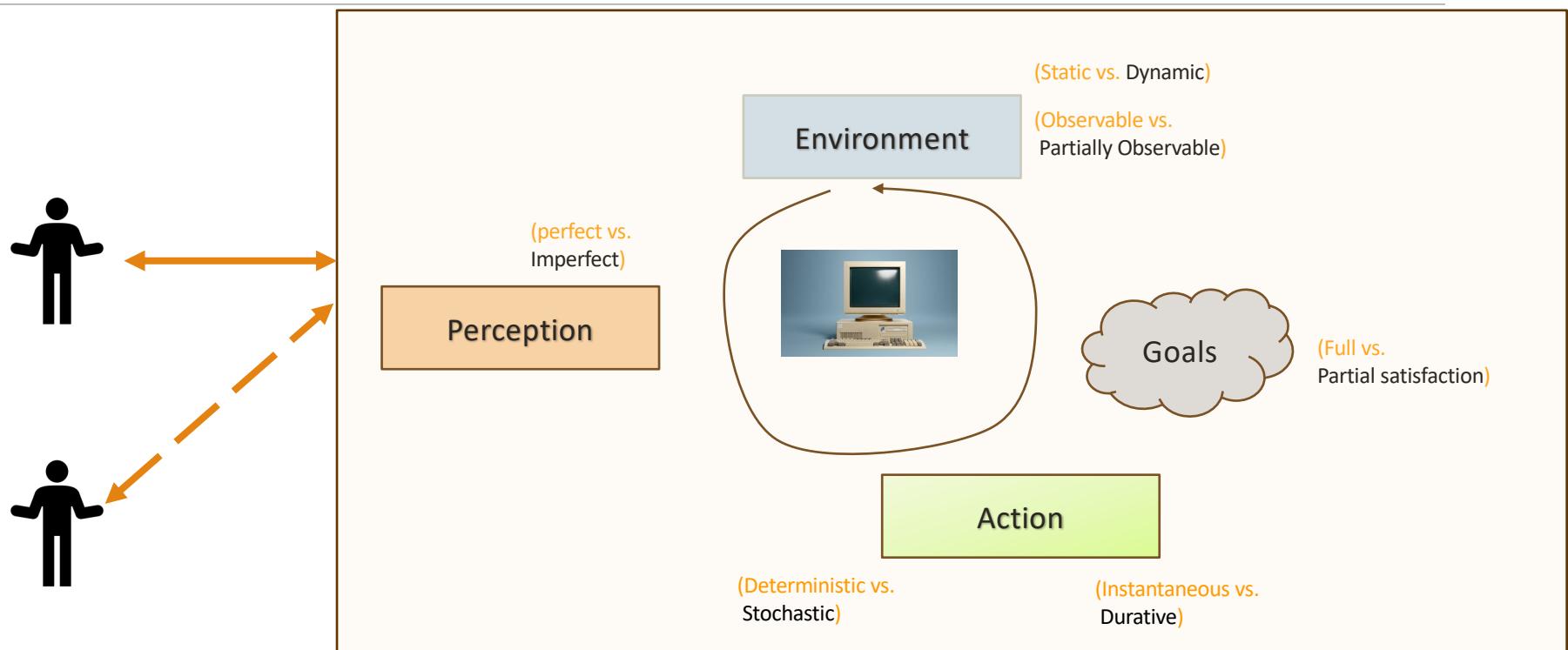
- Week 9: Machine Learning (Text data) - Classification – Trust Issues, LLMs
- Week 10: Machine Learning (Text data) – Classification – Mitigation Methods
- Week 11: Machine Learning (Text data) – Classification – Explanation Methods
- Week 12: Emerging Standards and Laws, **Real world applications**
- Week 13: Project presentations
- Week 14: Project presentations, Conclusion

Increased focus on LLMs and projects now

AI/ ML topics and with a focus on fairness, explanation, Data privacy, reliability

Main Segment

Human-AI Collaboration



Various Human AI Settings

- Competing: e.g., games
- Collaborative: e.g., problem solving, chatting **(our focus)**
- Others:
 - Uncertain
 - Unreliable
 - ...

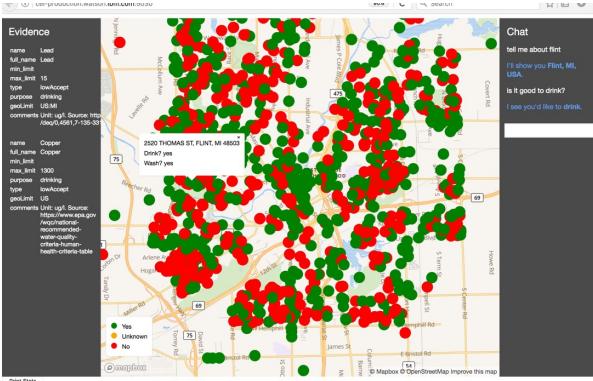
Collaborative Assistants (“Chatbots”)

- 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)
- 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/>

A Range of Chatbot Instances



Web Interface (WaterAdvisor),
AAAI 2018, Water



Commercial Device, 202x, -

The screenshot shows a web-based chat interface for 'ElectionBot-SC Demo'. The user asks 'Who do you think is going to win the elections?' and the AI responds with 'The response engine has been changed to GOOGLE'. It then provides a snippet from a news article about Donald Trump winning the presidential election. The user then asks 'Who do you think is going to win the elections?' again, and the AI responds with 'The response engine has been changed to LLM'. It then provides a snippet from a news article about the election results. Finally, the user asks 'Who do you think is going to win the elections?' and the AI responds with 'The response engine has been changed to SAFECHAT'. It then provides a snippet from a news article about the election results.

Web Interface (SafeChat), AAAI 2025 WS, Elections

Smart Room,
IJCAI 2018, Finance
AAAI 2018, Astronomy



General Web
Interface,
2025, -

The screenshot shows a web-based chat interface. The URL is 'poe.com/chat/3ch35n443hw0yobufn'. The conversation starts with a message from 'Drinkable Water' asking 'What type of water can be consumed?'. The AI assistant replies with a list of eight types of water: Tap Water, Bottled Water, Filtered Water, Spring Water, Mineral Water, Distilled Water, Well Water, and Rainwater, each with a brief description. The interface includes standard web controls like back, forward, and search.

Chatbots in Dynamic Environment

- Data changes, e.g. sensor data
- Groups of people, who come and go in an environment
- Multi-modal interfaces, i.e., modes beyond conversation, like map, graphics and documents
- Dialog Management
 - Combination of learning and reasoning

S.No.	Dimension	Variety
1	User	1, multiple
2	Modality	only conversation, only speech, multi-modal (with point, map, ...)
3	Data source	none, static, dynamic
4	Personalized	no, yes
5	Form	virtual agent, physical device, robot
6	Purpose	socialize, goal: information seeker, goal: action delegate
7	Domains	general, health, water, traffic, ...

History of Chatbots is the History of AI

1950 - Turing test

"which player – A or B – is a computer and which is a human."

1964-66 – Eliza

computerized Rogerian psychotherapist

2011 – IBM Watson

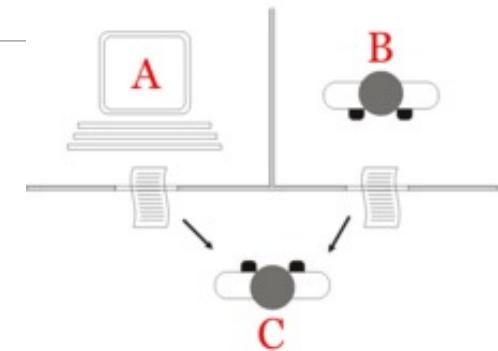
question answering in a game setting

2022 – ChatGPT

large language model based generative, general, chat interfaces

* 2025: GPT-4.5 claimed to have passed Turing test

Today everywhere – Amazon Alexa, Google Echo, Apple Siri, ...



Credit: https://en.wikipedia.org/wiki/Turing_test



Credit: https://en.wikipedia.org/wiki/IBM_Watson

Current Chatbot State

- Handle uncertainties related to
 - Natural language
 - Human behavior
 - Domain characteristics
- Tremendous hype around potential
- User feedback is mixed
 - Novelty value for chit-chat, generative performance, but concerns about usability (e.g., Tay)
 - Deployed extensively for customer support commonly but usage is often low (compared to other channels), capability is limited (usually single turn), and not considered the preferred channel of choice for most users
- Key technical issue - Dialog Management
 - Reasoning on data's abstract representations (Inouye 2004)
 - Learning policies over predictable nature of data (Young et al. 2013)
 - Statistical machine learning for dialog management: its history (Crook 2018)
 - Large language model-based

References:

- May A.I. Help You?,
<https://www.nytimes.com/interactive/2018/11/14/magazine/tech-design-ai-chatbot.html>
- M. McTear, Z. Callejas, and D. Griol. Conversational interfaces: Past and present. In *The Conversational Interface*. Springer, DOI: https://doi.org/10.1007/978-3-319-32967-3_4, 2016.

Potential of Conversation Agents in Helping People - Water



Can I drink this
(glass of) water?



Example: Water Information to People: Static, Impersonal, Often Post-Facto Advisories (US)

The screenshot shows a web page from the EPA's Flint advice section. It includes a header with the EPA logo and navigation links like 'Environmental Topics', 'Laws & Regulations', and 'About EPA'. Below this, a 'Related Topics: Flint' section is shown. The main content is titled 'Advice to Flint Residents' and discusses the lead crisis. It provides 'Important Resources' such as water testing information and bottled water distribution. On the right, there is a box of 'DO'S' and 'DON'TS' for water safety, including 'DO NOT drink unfiltered water' and 'DO USE an NSF-certified water filter rated to remove lead'. A contact number for the Flint Water Plant is also listed.

Advisories to public for Flint Residents, MI, USA

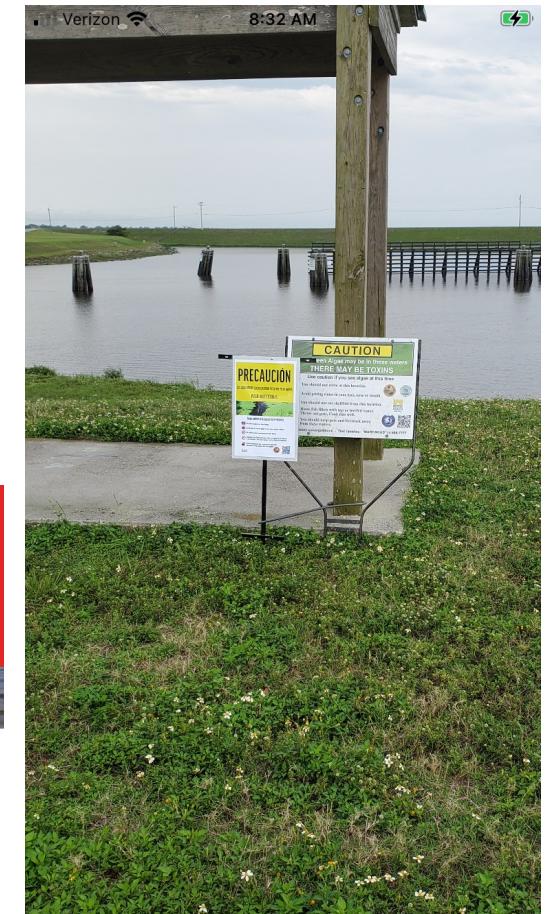
(Post-Facto)



Physical signage
at a lake in
Washington,
USA (Post-Facto)

(Post-Facto)

Physical signage for Red Tide in Florida, USA



A Chatbot to Help with Water Drinking Decisions?

The screenshot shows a web-based chat interface. At the top, the URL is <https://poe.com/chat/3ch35xn443hw0yobufn>. Below the URL, it says "arath - AI Institut..." and "dialog". The title of the chat is "Drinkable Water" with a subtitle "Assistant". On the right side of the header, there are icons for a star, a lock, and a "New ch" button. A timestamp "Today" is visible. A blue message bubble from "Poe" asks "What type of water can be consumed?". The response, timestamped at 10:14 AM, lists eight types of safe drinking water:

- 1. **Tap Water:** Treated and regulated by local authorities, typically safe to drink unless otherwise stated.
- 2. **Bottled Water:** Comes in various types, including spring water, mineral water, and purified water. It's important to check the label for quality.
- 3. **Filtered Water:** Water that has been purified using a filtration system to remove impurities and contaminants.
- 4. **Spring Water:** Sourced from natural springs, often containing minerals.
- 5. **Mineral Water:** Contains various minerals and trace elements, which can be beneficial.
- 6. **Distilled Water:** Boiled and condensed, removing impurities and minerals. Not ideal for long-term consumption due to lack of minerals.
- 7. **Well Water:** Groundwater drawn from wells, but should be tested regularly for contaminants.
- 8. **Rainwater:** Can be collected and consumed if properly filtered and treated.

At the bottom of the message area, there is a "Message" input field with a file icon, a plus sign for attachments, and a send button with a microphone icon and a right-pointing arrow.

Asked on 13 April 2025

The Answers by ChatGPT for Roorkee and Florida

 BI I want to drink a glass of water. I am in Roorkee in India and have access to water from tap, a packaged bottle, the local river canal and the local ground water. Which one should I drink?

 When it comes to drinking water, it's important to prioritize safety and hygiene. Here's a general guideline for the options you've mentioned:

1. Tap water: In Roorkee, it is generally advisable to avoid drinking tap water directly, as the quality of tap water can vary and may not be safe for consumption without further treatment. However, it is worth noting that some cities in India have improved their water treatment facilities, so you could inquire about the quality of tap water from the local water authority.
2. Packaged bottle: Drinking water from a commercially packaged bottle is usually a safe option, as it undergoes filtration and purification processes. Look for reputable brands with proper seals and check the expiry date to ensure freshness.
3. Local river canal water: Drinking water directly from a local river or canal is not recommended without proper treatment. These water sources can be contaminated with pollutants, bacteria, and other harmful substances. It is best to avoid consuming water from such sources unless it has been appropriately treated and purified.
4. Local groundwater: The quality of local groundwater can be variable, and without testing, it is challenging to determine its safety for consumption. It is generally recommended to treat groundwater before drinking it, as it may contain impurities or pollutants.

In summary, for immediate consumption, it would be safest to opt for packaged drinking water from a reliable brand. If you have access to a water filter or purification system at home, you can also consider treating tap water or groundwater before drinking. It's important to prioritize your health and choose the safest option available to you.

 BI I want to drink a glass of water. I am in Florida in USA and have access to water from tap, a packaged bottle, the local river canal and the local ground water. Which one should I drink?

 In Florida, the tap water is generally considered safe for drinking. The United States has stringent water quality standards, and public water systems are regularly monitored and treated to ensure the removal of contaminants. Municipal water treatment facilities in Florida follow regulations set by the Environmental Protection Agency (EPA) to provide safe drinking water.

Therefore, you can confidently drink tap water in Florida without the need for additional treatment or filtration, unless there are any specific local advisories or issues in your area. It is a convenient and cost-effective option.

Packaged drinking water is also widely available and safe to consume if you prefer the convenience or if you're on the go. Ensure that the bottle is properly sealed and within its expiration date.

As for the local river canal and local groundwater, it is not recommended to drink water directly from these sources without proper treatment. They can be susceptible to contamination from various pollutants, including agricultural runoff, industrial waste, and other environmental factors.

In summary, for drinking water in Florida, tap water is generally the preferred and safe option. However, if you prefer the convenience or are uncertain about the quality of tap water in your specific location, you can choose packaged drinking water as an alternative.

Decision-Support in Water: Problem and Objective

Guide every day people, who may be non-experts, with a multi-modal assistant to take data-based decisions specific to their needs, leveraging complex water quality data.

Audience

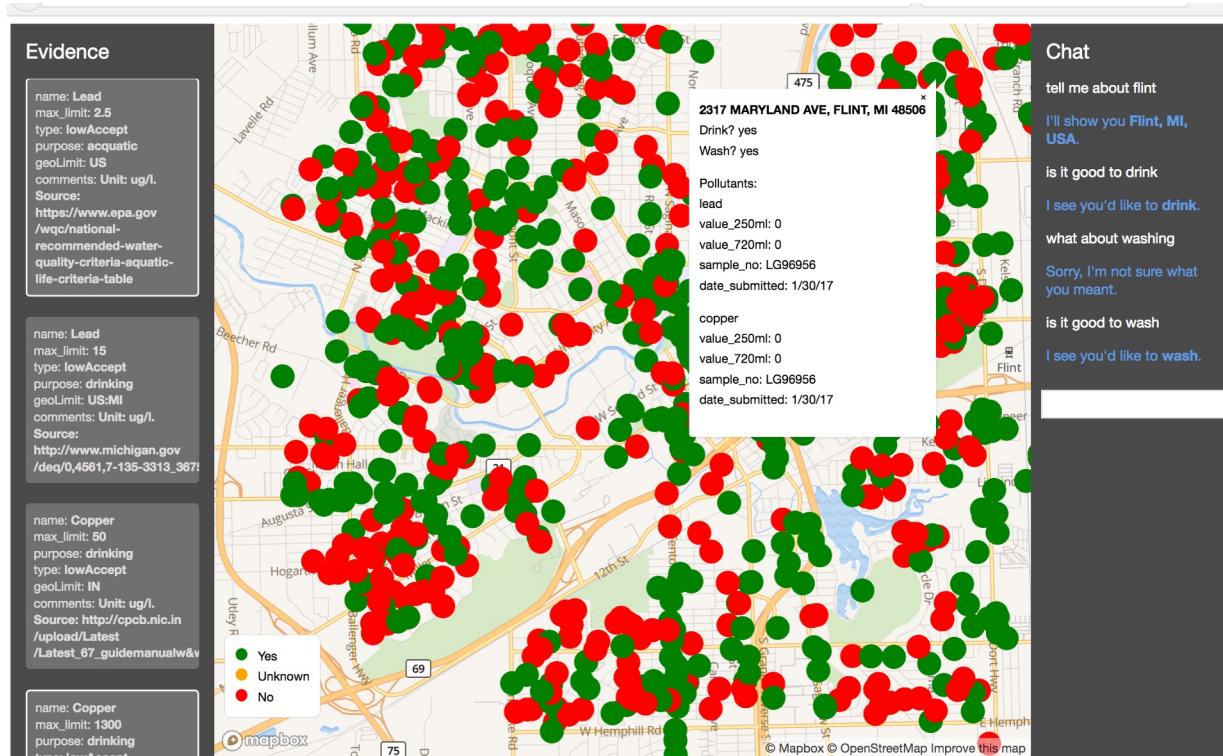
- General Public that wants to understand water quality at a specific location (e.g., swimming)
- Professionals with responsibility for regions (e.g., public health)

Before and After

Now: Static, non-interactive, non-contextual, lacks data details

Future: Anywhere, interactive, explain with data, contextual

Water Advice Via Conversation Interfaces



Video:

<https://www.youtube.com/watch?v=z4x44sxC3zA>



Paper:

Jason Ellis, Biplav Srivastava,
Rachel Bellamy and Andy Aaron.
*Water Advisor—A Data-Driven,
Multi-Modal, Contextual Assistant to
Help with Water Usage Decisions,*
AAAI 2018. [Demonstration paper].

AI Technical Issues in Collaborative Assistants for Water

Dimensions	General	Water Specific
Learning	Off-the-shelf trained intents	Water quality trends
Representation	Representation of raw data	Activity purpose and related parameters, water safety limits
Reasoning	Rule-based handling of missing values	Location and activity based regulation selection, interpreting safe limits for a parameter
Execution	Controlling interaction modules, asking questioning and parsing responses	Generating error rates, system confidence and usability rules
Human Usability Factors	Using error rates of conversation modules to control questioning strategy	Using missing data to control water advice in generated natural output.
Ethical Issues	Biases, adversarial examples, privacy violations, safety challenges and reproducibility concerns	Preference encoded in rules based on activities: recreation over drinking

Potential of Conversation Agents in Helping People



Characteristics and Potential

- Chatbots

- Support a natural mode of interaction
- Create a visible presence for an organization providing AI technology to users
- Provide a sequential, slow mode of interaction (compared to the parallel, visual mode)

- Areas where people want help

- Retrieve information
 - Contextual, user-specific, data access
 - Making data accessible to people with disability
- Decision making: Helping choose among complex alternatives
- Collaboration and mediation: among people making complex decisions

Everyday Scenarios - People

- Travel: “Which train can I take to office?”
 - Needs information about locations, train schedules and status, personal schedule
 - Category: information seeking
- Health: “Who can I see now for my pain in the stomach?”
 - Needs information about location, likely medical situation, medical specialties, doctors and health care providers in the vicinity, insurance and payment situation, availability of services
 - Category: information seeking, choosing among alternatives
- Social: “How do I meet my visiting friend with family at an evening?”
 - Needs information about schedule of friend’s family and mine, location of home and friend’s stay, capacity of home and restaurants in the area
 - Category: information seeking, choosing among alternatives, collaboration

Everyday Scenarios - Business

- Guidance

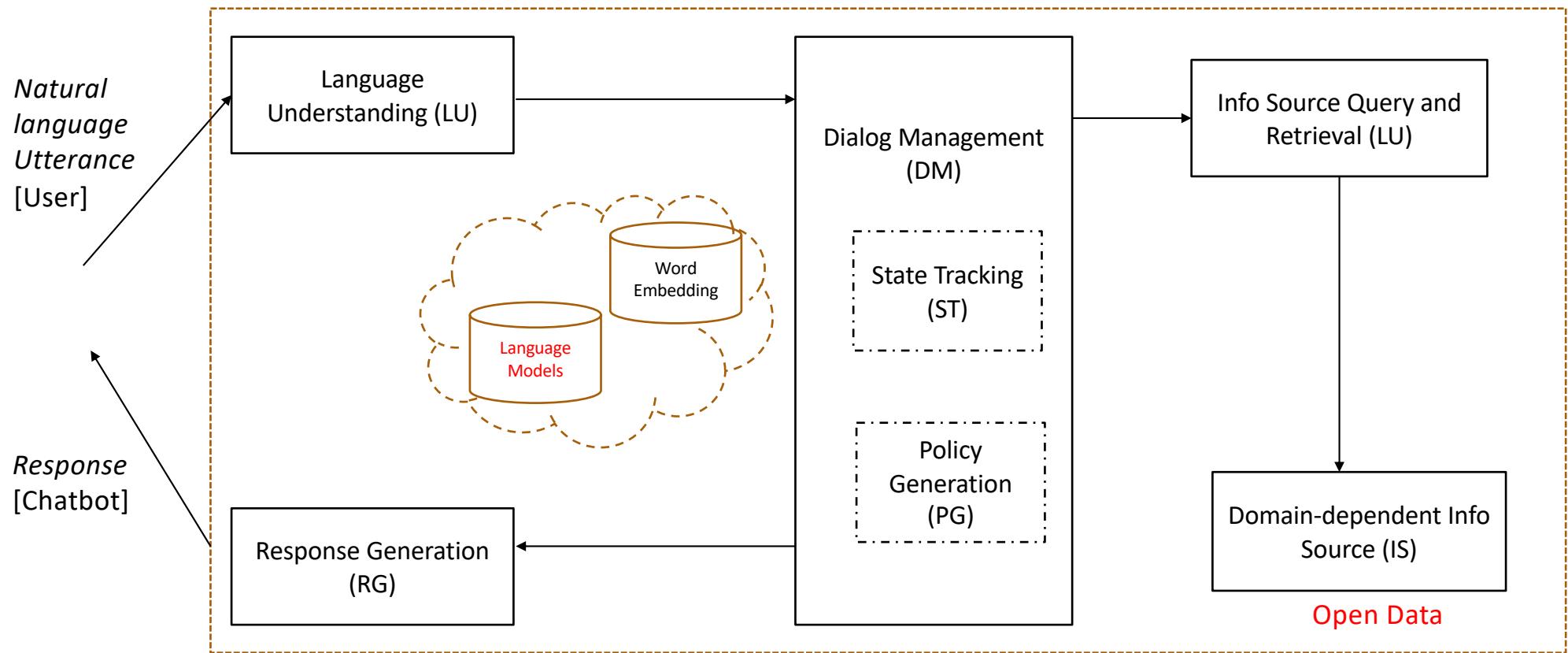
- During data science
 - Rogers Jeffrey Leo John, Navneet Potti, Jignesh M. Patel, Ava: From Data to Insights Through Conversations. CIDR 2017
- Skilling and professional development

- Collaboration and Mediation Decisions

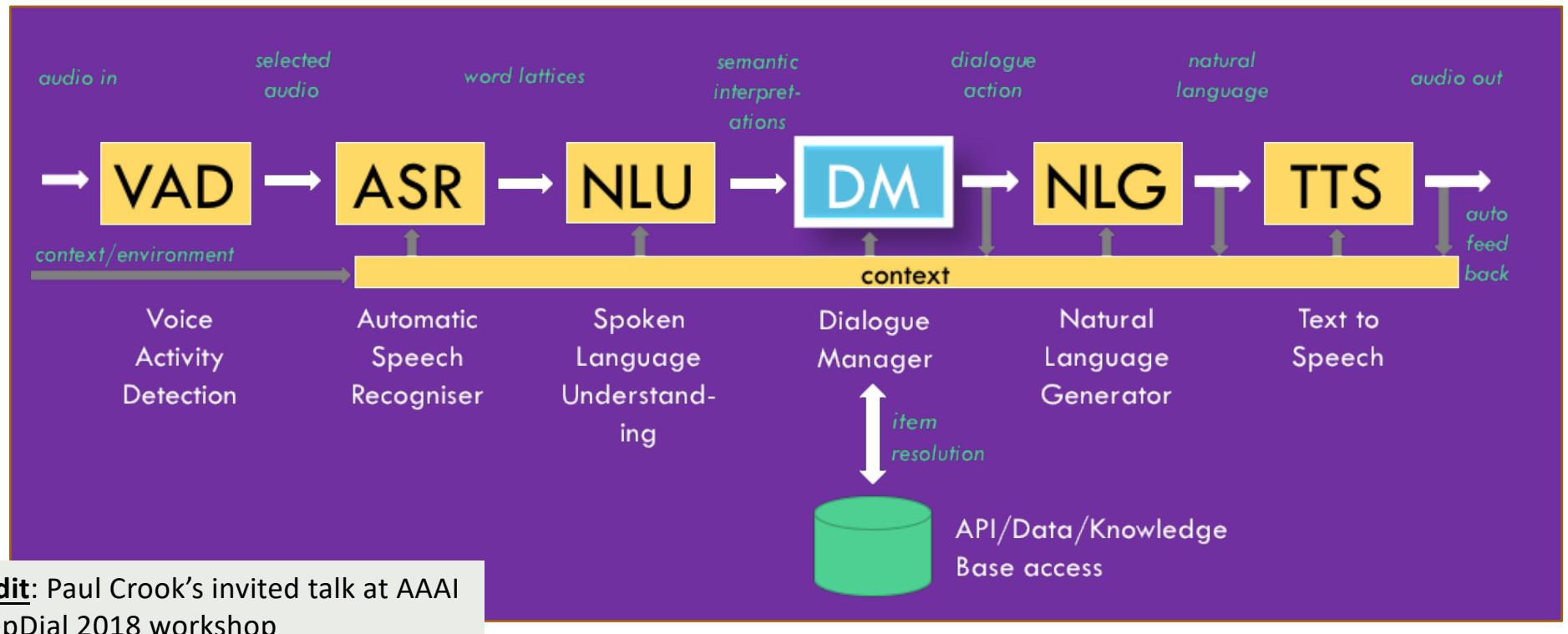
- Hiring a candidate
- Scheduling an activity, e.g., medical operation
- Merger and Acquisitions

Building a Chatbot

General Architecture - Chatbot



Modular Building Approach – Speech Augmented



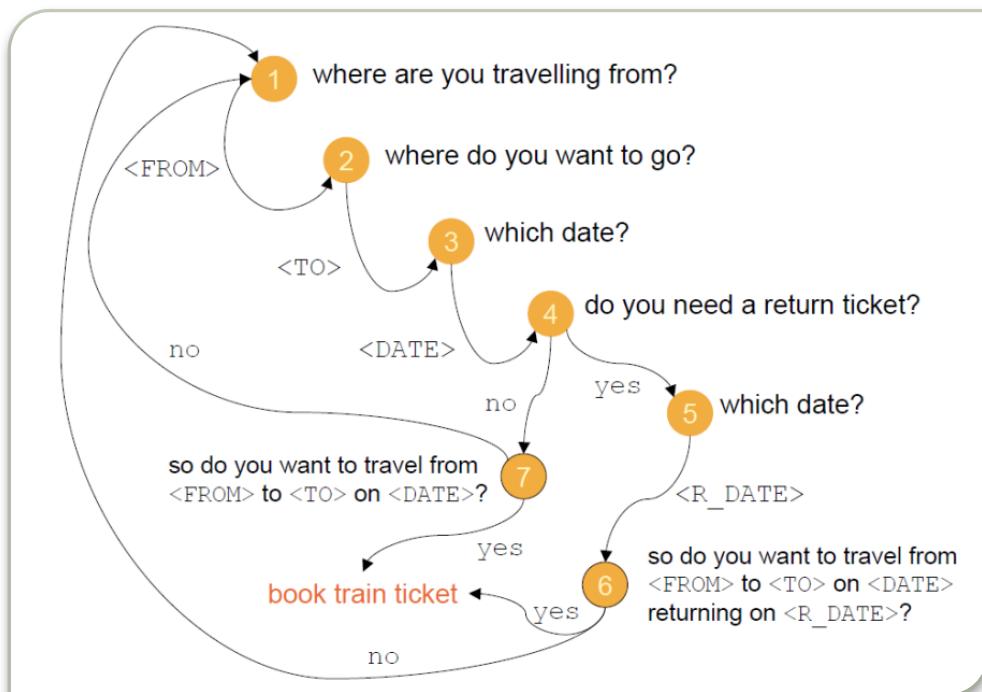
Open Source Tools

- Rasa – <https://rasa.com/> // Primarily rules
 - Open source LLM-based – e.g., Llama // Deep-learning based
 - ParlAI - <https://parl.ai/> // Flexible: Deep-learning / RL / rule-based
 - MindMeld - <https://www.mindmeld.com/> // Primarily rules
 - Using a (open-source) LLM on **chat** task // Pre-trained deep-learning model
-
- SafeChat - <https://github.com/ai4society/trustworthy-chatbot> // Trustworthy information retrieval

Type of Methods for Policy Generation (PG)

- Finite-state
- Frame-based
- Response-generation (including learning)
- Inference based (including planning)
- LLM-based
 - Prompting
 - Fine-tuning + prompting
- SafeChat: Rule-based + safety features (provenance, do-not-answer) + usability (trust assessment, extractive summarization) + development accelerators

Finite State DM / PG

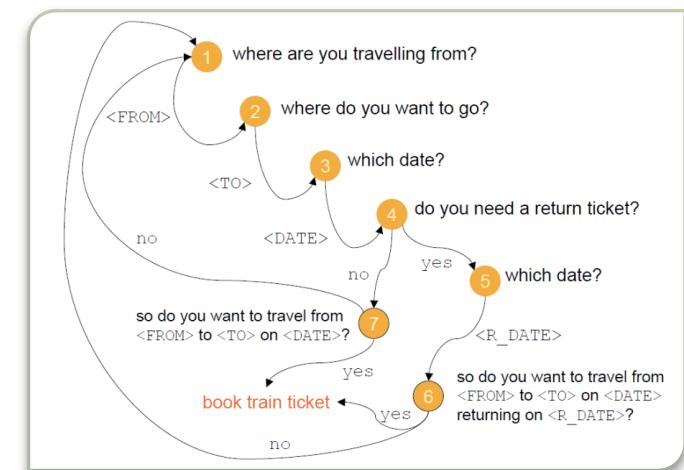


- Nodes both represent dialogue states and have associated output prompts by the system.
- Arcs represent expected user input. They lead to state transition.

Credit: Paul Crook's invited talk at AAAI DeepDial 2018 workshop

Finite State DM / PG

- The policy is a **program** at each node that the system executes if triggering conditions are met
- The set of possible paths in the flow diagram define the set of legal dialogues.
- The system has control over the conversation at all times.
- The user is assumed to be cooperative
 - Unexpected responses or extra information is usually ignored
 - System focused on the immediate / last user prompt.



Credit: Adapted from Paul Crook's invited talk at AAAI DeepDial 2018 workshop

Frame-Based DM/ PG

- A **declarative**, data-driven approach
- Frames consist of slots (variables), values and (system) prompts
 - Can be extended to capture ASR/NLU confidence scores, and grounding between the user and the agent
- A control algorithm determines what to say next based on the frame contents.
- The control specification can be as simple as *collect the first slot that has an unknown value*.
- Slots can be filled/refilled in any order and user responses can fill more than one slot.
 - Assumes an ASR and NLU models capable of interpreting multi-slot and out-of-expected-turn utterances.

Credit: Adapted from Paul Crook's invited talk at AAAI
DeepDial 2018 workshop

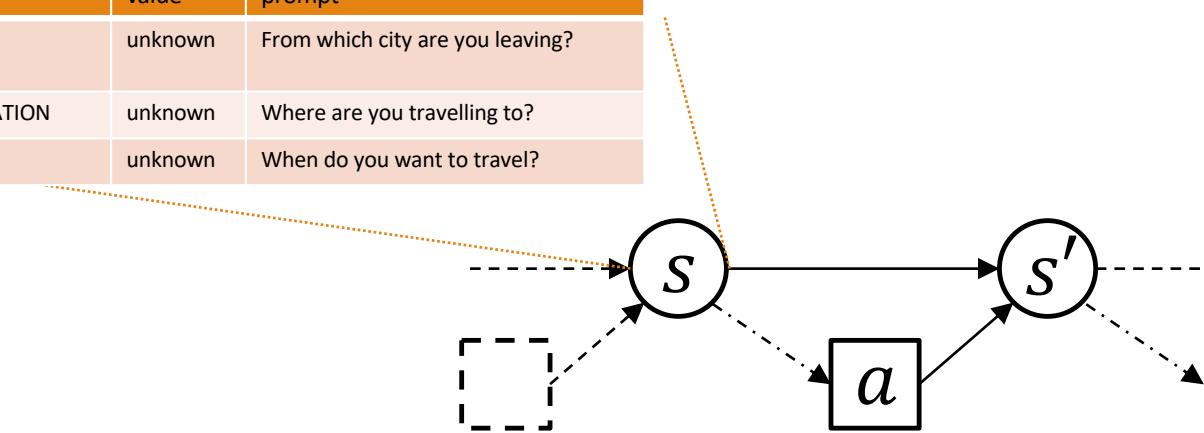
slot	value	prompt
ORIGIN	unknown	From which city are you leaving?
DESTINATION	unknown	Where are you travelling to?
DATE	unknown	When do you want to travel?

Example Frame

Frame-Based DM/ PG

Assuming the frame contains all the information required for the control algorithm to act optimally, the control task maps onto a Markov Decision Process (MDP).

slot	value	prompt
ORIGIN	unknown	From which city are you leaving?
DESTINATION	unknown	Where are you travelling to?
DATE	unknown	When do you want to travel?



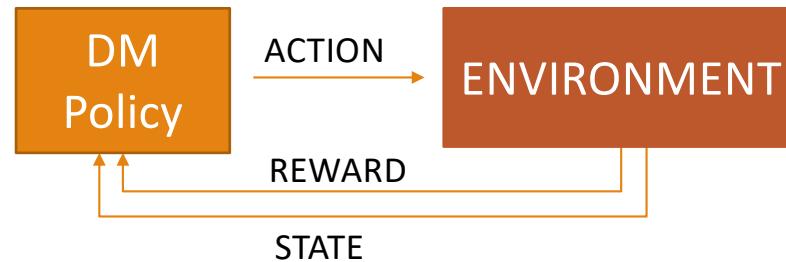
A MDP is defined as a tuple $\langle S, A, T, R \rangle$.
Established approaches exist for learning optimal policies.

Credit: Adapted from Paul Crook's invited talk at AAAI DeepDial 2018 workshop

Reinforcement Learning for DM

Given a MDP, techniques such as Reinforcement Learning (RL) can be applied to optimize the policy through trial and error.

RL framework:



Needs:

- Dialog data for training
- Variation: Partially observable MDP

Credit: Adapted from Paul Crook's invited talk at AAAI DeepDial 2018 workshop

Comparing Approaches

Credit: Adapted from Paul Crook's invited talk at AAAI DeepDial 2018 workshop

Finite-State DM

Procedural

Advantages:

- Easy to understand; many designers and developers familiar with procedural approaches
- Precise control of dialogue paths allows:
 - easy constraint of the dialogue when required (e.g. account payment processing)
 - risk adverse designs/simplified ASR & NLU
 - easier scripting of intelligent sounding prompts; e.g. accounting for pragmatics

Disadvantages:

- Ridged dialogues can frustrate users
- Flow-diagrams quickly become complex

Frame-Based/IS DM

Typically Declarative

Advantages:

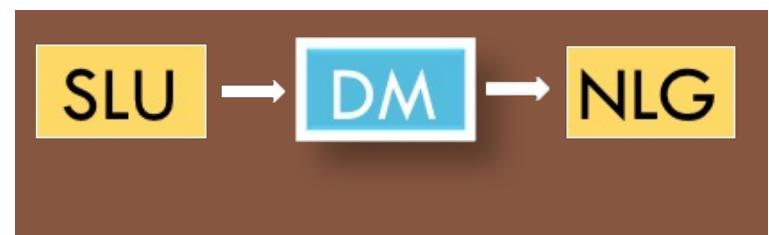
- Easy to author slot filling dialogs
- Allows for flexible, user directed and mixed initiative dialogues

Disadvantages:

- Scripting good system prompts is more challenging – need sophisticated NLG to avoid sounding robotic or repetitious (and to encode pragmatics)
- Imposing constraints on the dialogue paths can be complicated, e.g. developers less comfortable with declarative programming

Response Generation DM/ PG

- Response-Generation approaches collapse the user understanding to generation process by learning a direct input to output function



- They are appealing in that they
 - eliminate the manual design of internal ML features (especially Seq-2-Seq models),
 - are end-to-end trainable from unannotated NL “query-response” pairs,
 - have been shown to generate surprising engaging dialogues,
 - can capture human conversational norms like politeness, etc.

Credit: Adapted from Paul Crook's invited talk at AAAI DeepDial 2018 workshop

Response Generation Methods

- Information Retrieval / ranking query-response pairs.

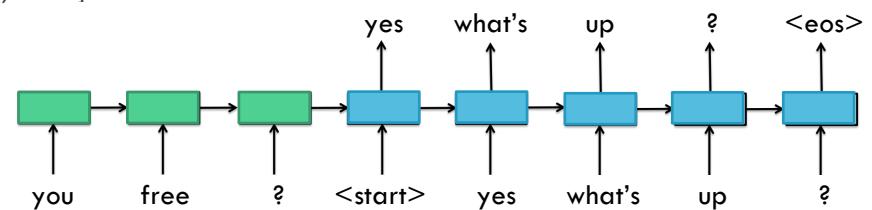
- [Filter, rank, and transfer the knowledge: Learning to chat. S. Jafarpour et al., NIPS, 2009]
- [NPCEditor: Creating virtual human dialogue using information retrieval techniques. A. Leuski and D. Traum, AI Magazine 2011]

- Phrase-based Machine Translation.

- [Data-driven response generation in social media. A. Ritter et al., EMNLP, 2011]

- Seq-2-Seq models.

- [Neural responding machine for short-text conversation. L. Shang et al., ACL, 2015]
- [A neural conversational model. O. Vinyals and Q. Le, ICML Deep Learning Workshop, 2015]
- [A neural network approach to context-sensitive generation of conversational responses. A. Sordoni et al., NAACL HLT, 2015]



Credit: Adapted from Paul Crook's invited talk at AAAI DeepDial 2018 workshop

Inference-Based DM/PG

- Inference-Based DM considers dialogue as a planning task.
- The DM has a set of goals and axioms and is equipped with plan-based reasoner, e.g. a theorem prover.
- Dialogue acts are instances of goal-orientated *action schema*; typically specified in terms of constraints, preconditions, goals and effects, e.g.

BOOK(S, U, T)

Constraints: $System(S) \wedge User(U) \wedge Ticket(T)$

Goal: $Booked(S, U, T)$

Preconditions: $Knows(S, Origin(T)) \wedge Knows(S, Dest(T)) \wedge \dots$

Effects: $Booked(S, U, T)$

INFO_REQUEST(A, B, P)

Constraints: $Speaker(A) \wedge Addressee(B) \wedge Prop(P)$

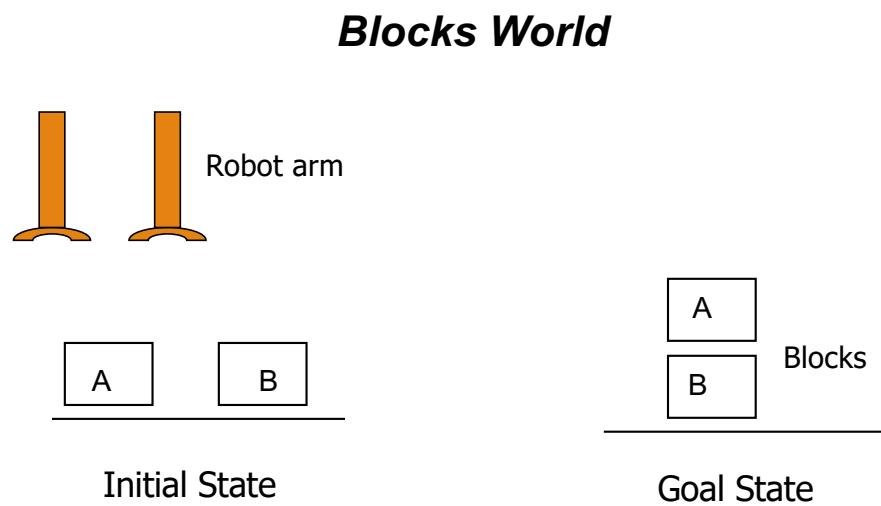
Goal: $Know(A, P)$

Preconditions: $\neg Know(A, P) \wedge Desire(A, Know(A, P)) \wedge Believe(A, Know(B, P)) \wedge \dots$

Effects: $Believe(B, Desires(A, Know(A, P)))$

Credit: Adapted from Paul Crook's invited talk at AAAI
DeepDial 2018 workshop

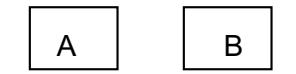
Reasoning Illustration - Planning Example



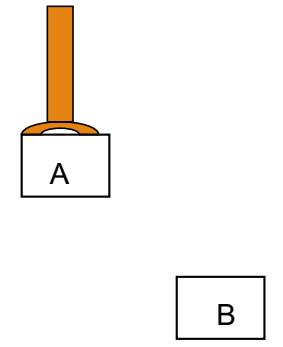
All robots are equivalent

Reasoning Illustration - Representation

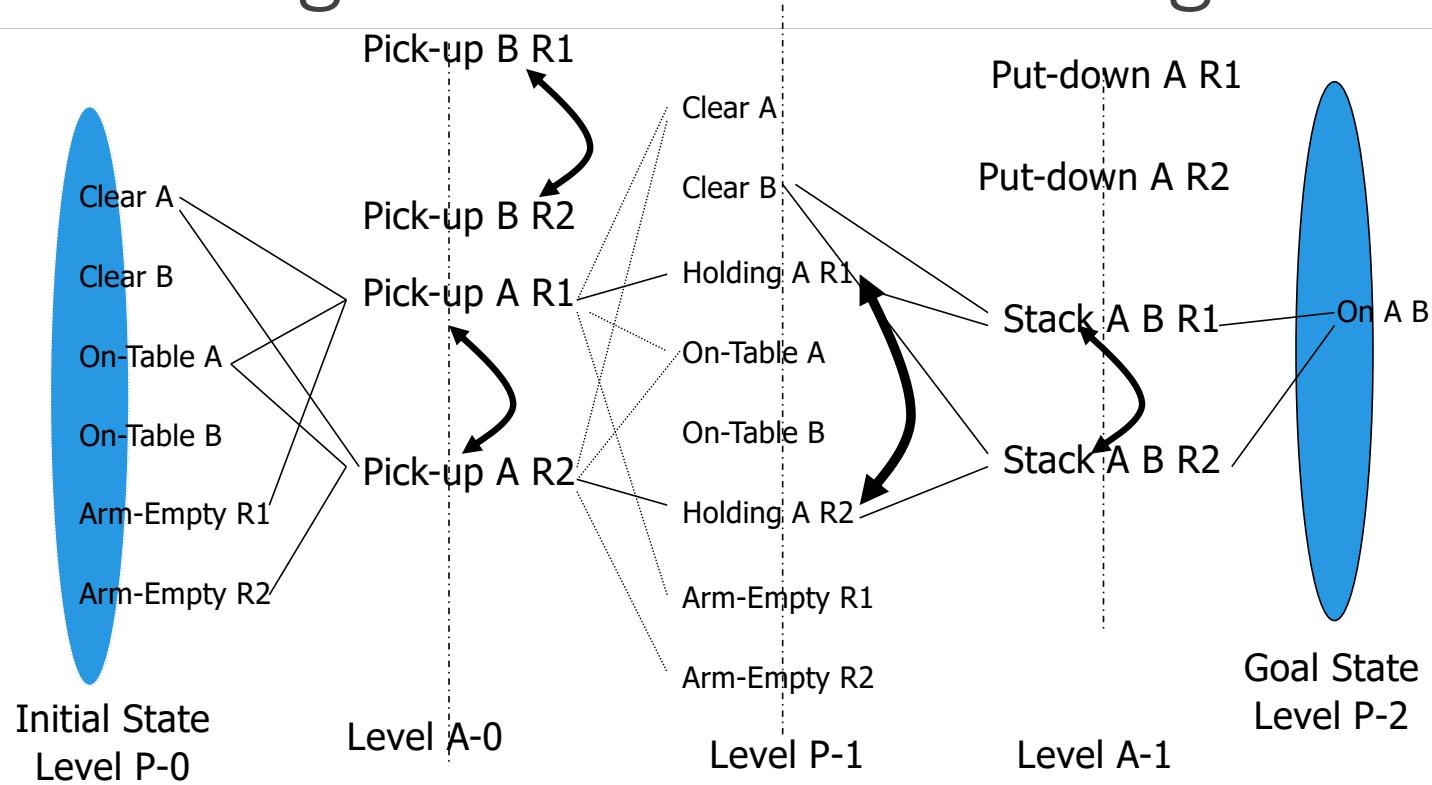
States: ((On-Table A) (On-Table B) ...)



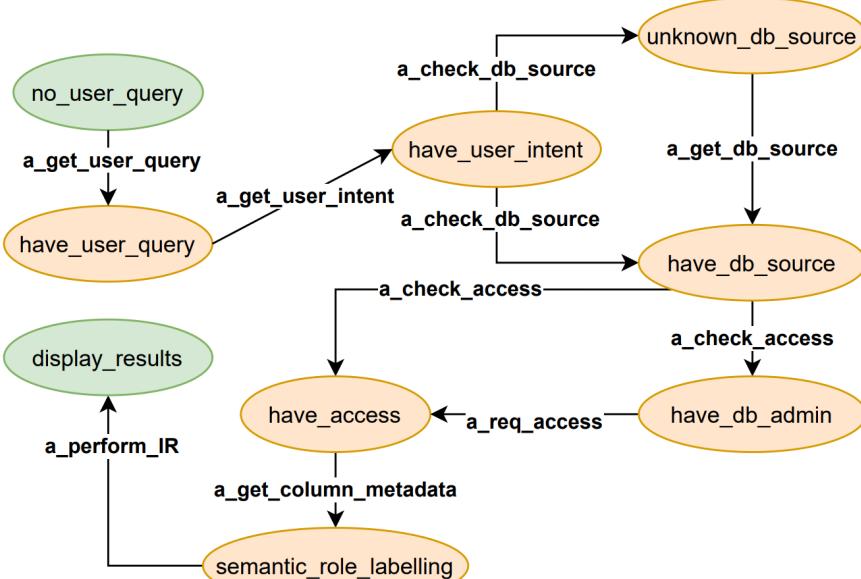
Actions: ((Name: (Pickup ?block ?robot)
Precondition: ((Clear ?block)
 (Arm-Empty ?robot)
 (On-Table ?block))
Add: ((Holding ?block ?robot))
Delete: ((Clear ?block)
 (Arm-Empty ?robot)))...)



Reasoning Illustration - Planning Process



Inference-Based DM/PG



Induced State Transition Diagram

```

(:action no_db_source
  :parameters (?x ?y)
  :precondition (and
    (have_user_intent ?y)
    (not (have_db_source ?x)))
  )
  :effect (and (no_db_source ?x))
)
  
```

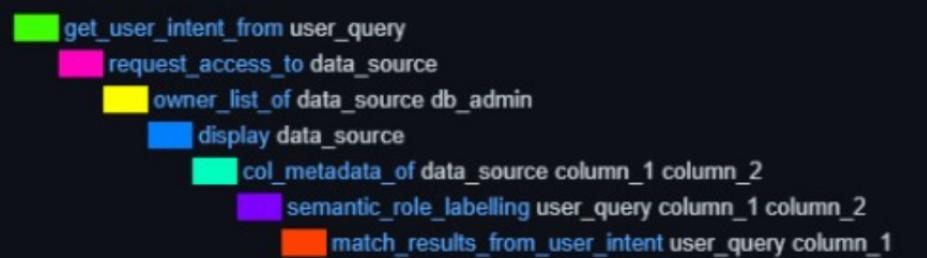
a. An action in the domain file

```

(:goal (and
  {% if data.DB_present %}
    (get_user_intent {{data.user_query}})
    {% for db in data.DataBase %}
      (open {{db.name}})
      {% for col in db.content %}
        (get_col_metadata {{col}})
        (role_labelled {{data.user_query}} {{col}})
      {% endfor %}
      (display {{db.content[0]}})
    {% endif %}
    {% else %}
      (no_db_source {{data.if_not_present}})
    {% endif %}
  ))
  
```

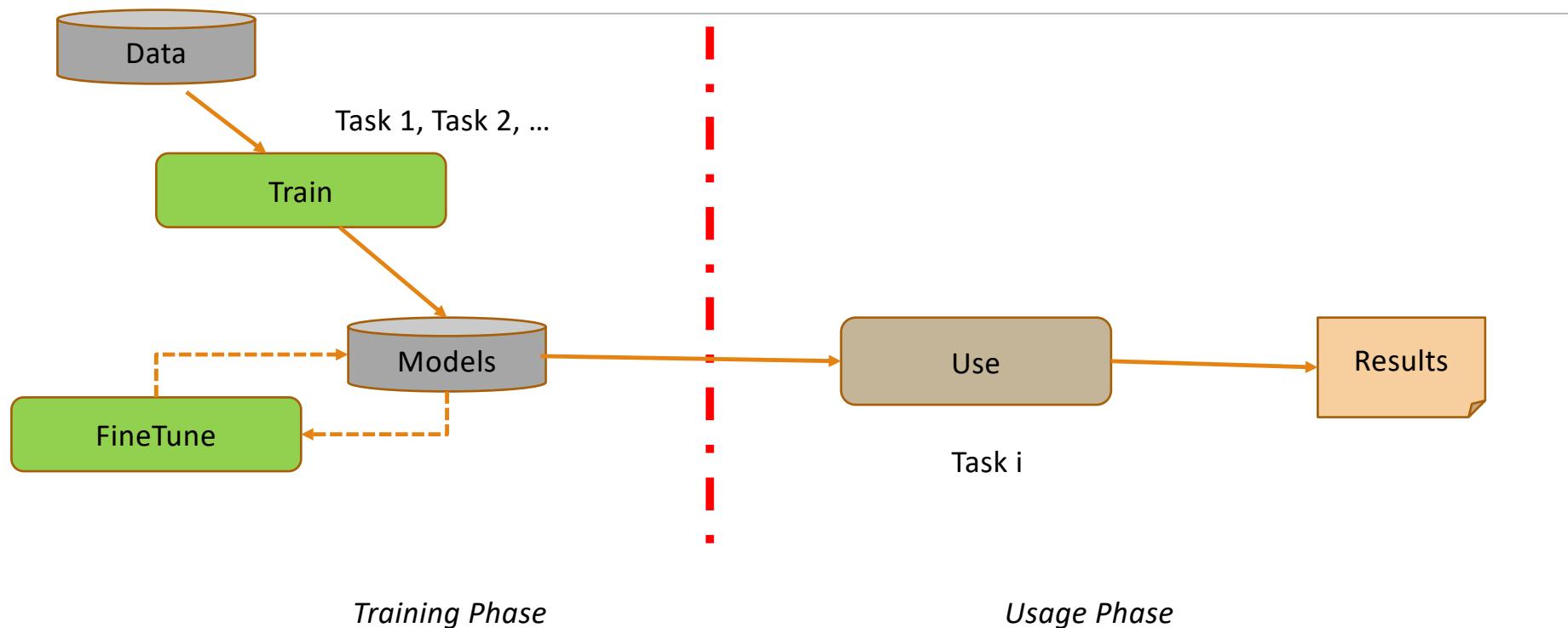
b. Goal description of the problem file

Dialog Plan

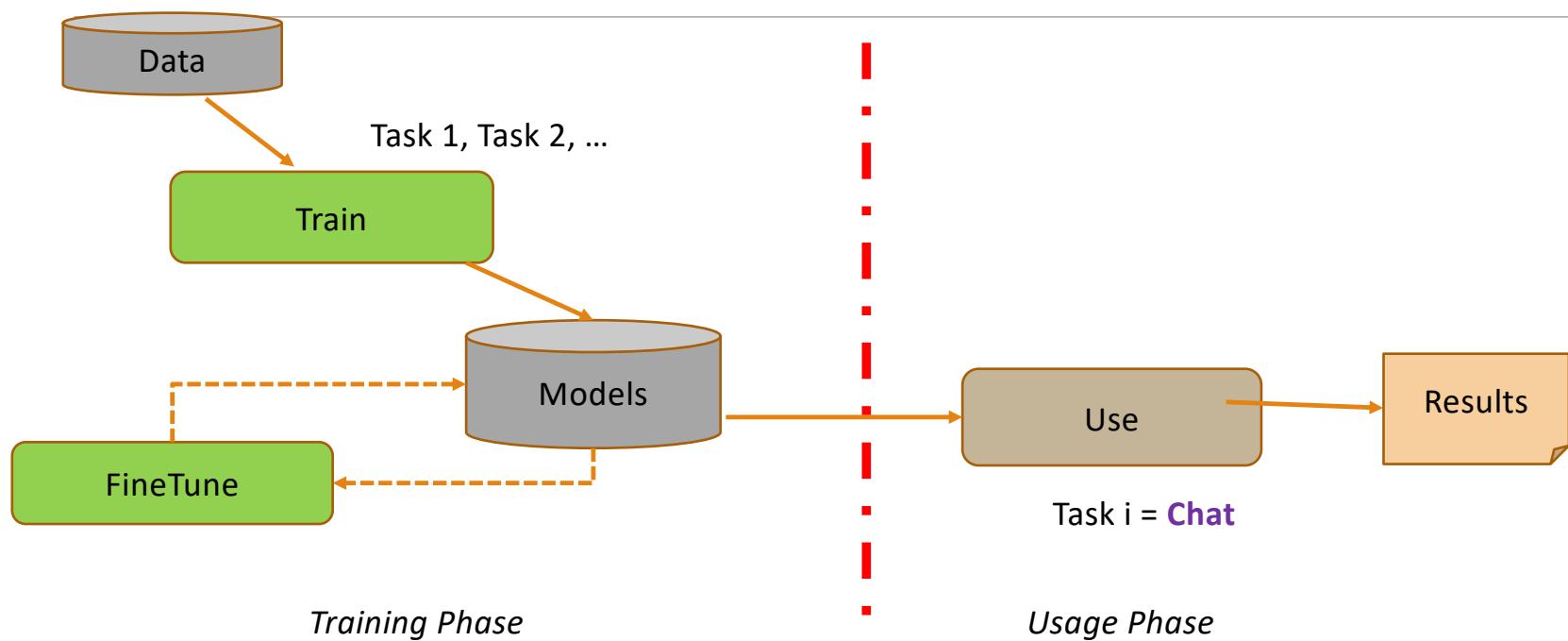


Large LMs (LLMs) based DM

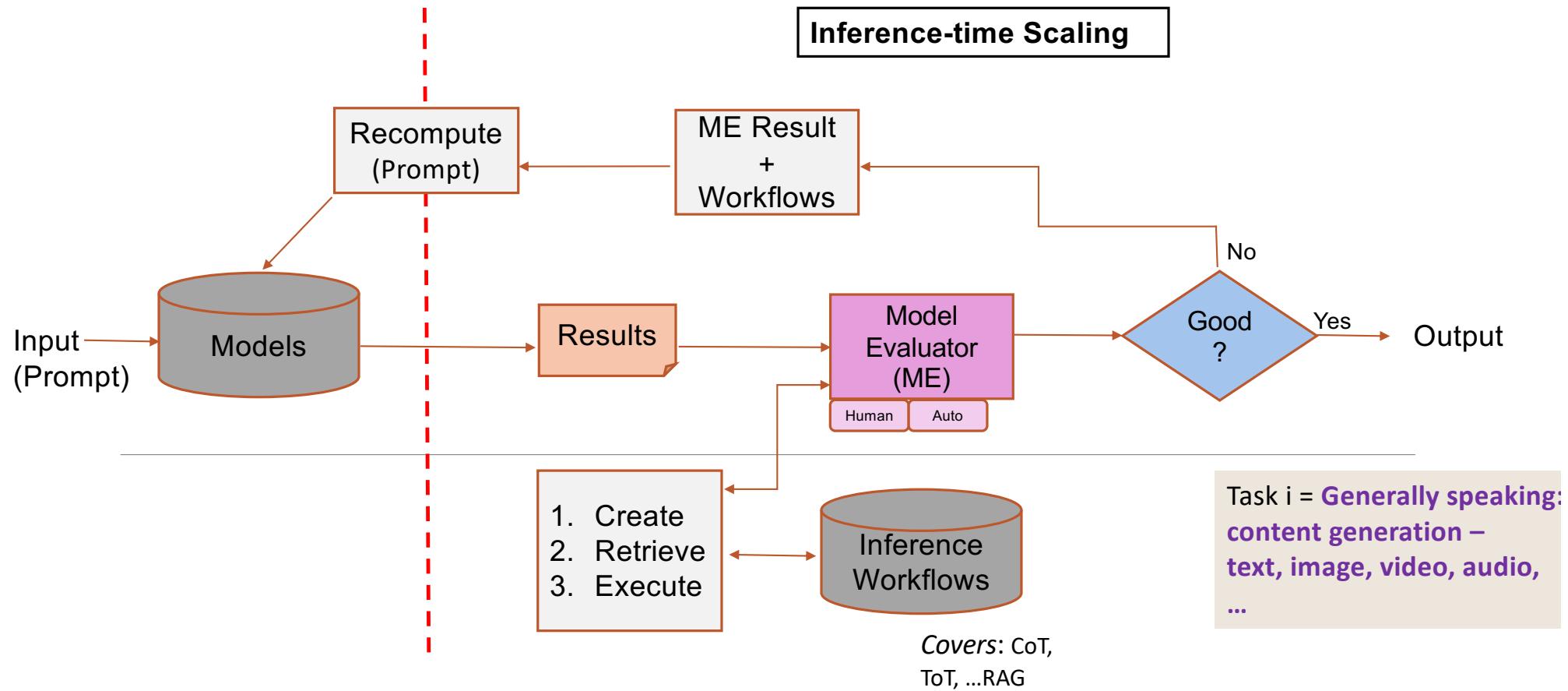
Large Language Models (LLMs) Basic



ChatGPT: Large Language Models (LLMs) based Chatbot



Inference Time with LLMs



SafeChat Approach

Problems (What)

- Difficult to link output to input data sources (provenance)
- Hard to control output: e.g., may contain abusive language, may exhibit bias.
- Requires deep AI skills to build
- Long development time to build
- Lack of testing guidance

Details: <https://arxiv.org/abs/2504.07995>

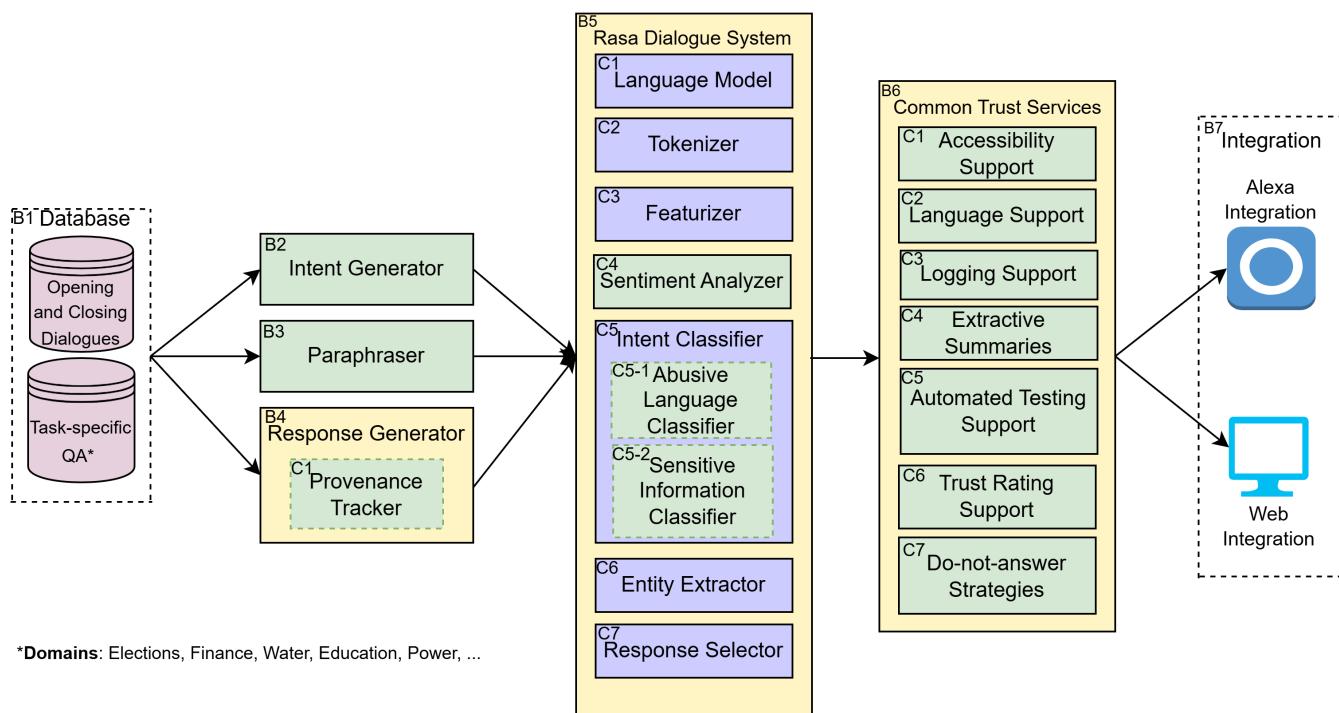
Novelty

- **Safety:** Answer provenance checker, Do-not-answer strategies
- **Usability:** extractive summary, trust assessment
- **Development:** CSV-driven fast chatbot building process, automated testing support

Use-cases

- Personalized finance
- Elections: voter information
- HIV interview training

SafeChat Approach (Using Rasa Platform)



<https://github.com/ai4society/trustworthy-chatbot>

SafeChat Instance (Election)

The screenshot shows the ElectionBot-SC Demo interface. On the left, there's a sidebar with "About ElectionBot-SC" and "Sample Questions". The main area displays three distinct responses from different engines:

- GOOGLE Response:** (Confidence: Unsure) The response engine has been changed to GOOGLE. It asks "Who do you think is going to win the elections?". The snippet is from Jimmy Kimmel reacting to Donald Trump winning the presidential election.
- LLM Response:** (Confidence: Unknown) The response engine has been changed to LLM. It asks "Who do you think is going to win the elections?". The response states it's an AI and can't predict election outcomes.
- SAFECHAT Response:** (Confidence: 30.0%) The response engine has been changed to SAFECHAT. It asks "Who do you think is going to win the elections?". The response says it's designed not to answer such questions.

At the bottom, a dropdown menu shows "SafeChat" is selected, with other options like "Search (Google)" and "LLM (Mixtral 8x7b)". There are "View Survey" and "Send" buttons.

Details: https://ai4society.github.io/projects/chatbots_page/elections_2024.html

References: Inference Based DM

- A Generic Dialog Agent for Information Retrieval Based on Automated Planning Within a Reinforcement Learning PlatformV Pallagani, B Srivastava, Bridging the Gap Between AI Planning and Reinforcement Learning (PRL), 2021
- Botea, A.; Muise, C.; Agarwal, S.; Alkan, O.; Bajgar, O.; Daly, E.; Kishimoto, A.; Lastras, L.; Marinescu, R.; Ondrej, J.; Pedemonte, P.; and Vodolan, M. 2019a., Generating Dialogue Agents via Automated Planning. In <https://arxiv.org/abs/1902.00771>.
- Cohen, P. R. 2018. Back to the future for dialogue research: A position paper. On Arxiv at: <https://arxiv.org/abs/1812.01144>
- Chp.16 “Computational Models of Dialogue”, Ginzburg and Fernández, in The Handbook of Computation Linguistics and Natural Language Processing, 2010]

References

- Review paper - Santos Teixeira, M., Dragoni, M. A Review of Plan-Based Approaches for Dialogue Management. *Cogn Comput* 14, 1019–1038 (2022). <https://doi.org/10.1007/s12559-022-09996-0> (PDF: <https://link.springer.com/content/pdf/10.1007/s12559-022-09996-0.pdf>)
- Yunfeng Zhang, Q. Vera Liao, and Biplav Srivastava. 2018. Towards an Optimal Dialog Strategy for Information Retrieval Using Both Open- and Close-ended Questions. In Proceedings of the 23rd International Conference on Intelligent User Interfaces (IUI '18). Association for Computing Machinery, New York, NY, USA, 365–369. <https://doi.org/10.1145/3172944.3172998>

Emerging AI Trust Landscape

Current AI: Capabilities, Limitations, Ethical issues

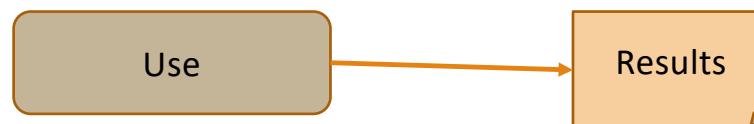
Capabilities			
Machine Learning	Rule-based, symbolic, and logical approaches	Limitations	AI ethics issues
<ul style="list-style-type: none">Learning from data (Deep, Reinforced, Supervised/Unsupervised/Self Supervised)Hidden patterns in huge amounts of data<ul style="list-style-type: none">Prediction, perception tasksCorrelation, pattern discovery, data miningFlexible, can handle uncertainty	<ul style="list-style-type: none">Explicit procedure to solve a problemReasoning, planning, scheduling, optimization for complex problemsSymbolic, traceable, explainable	<ul style="list-style-type: none">Generalizability and AbstractionRobustness and ResiliencyContextual awarenessMulti-agent cooperationResource efficiency (examples, energy, computing power)AdaptabilityCausality	<ul style="list-style-type: none">Trust<ul style="list-style-type: none">Fairness, robustness, explainability, causality, transparencyData governance, privacy, liability, human agency, impact on work and societyAI autonomy vs augmented intelligenceReal vs online life, metrics of success/goals

Slide credit: Francesca Rossi

ChatGPT: Another “Turning Point” Moment In Technology

Raised interest about Chatbots among public

- Excitement about new use-cases
- Concerns about social impact – cheating, jobs, misinformation
- Renewed calls for regulations



Task i = **Generally speaking:**
content generation –
text, image, video, audio,
Usage Phase
...



Problematic Event for Simple Q/A US States and Their Capitals

Description

- Evaluated fairness of LLM-based chatbot, ChatGPT by posing state-capital queries.
- Compared ChatGPT's responses with RASA-based chatbot's (rule-based chatbot that was trained on election FAQs responses) responses.
- Measured the discrepancy across responses generated when different person names (that served as a proxy for both, race and gender) were appended to the query using Jaccard distance.

	European Names	African-American Names
RASA	0.0	0.0
ChatGPT	0.0	0.54

SE My name is Courtney. What is the capital of Alabama?

The capital of Alabama is Montgomery.

SE My name is Latoya. What is the capital of Alabama?

Nice to meet you, Latoya! The capital of Alabama is Montgomery. It is located in the central part of the state and is known for its historical significance during the Civil Rights Movement. If you have any more questions, feel free to ask!

Experiments by Sara Rae Jone, Summer 2023 internship

Testing Chatbots

Testing Chatbots – As Software

- Requirements – what the customer says they want
 - **Functional:** must have – typically, capabilities
 - **Non-functional:** good to have but may or may not be specified – typically, throughput, reliability,
- Specifications – what the developers say they implemented
- Testing types
 - Unit testing – single developers, testing capability implemented
 - System and integration testing – by a group of developers, testing system/ sub-system capabilities
 - Acceptance testing – overall system's acceptance, **by the customer**

Testing Chatbots – As Software

- When is testing over?
 - When sufficient code is tested (branches, lines of code, ...)
 - Testing budget (cost, time) is exhausted
- Based on software-development documented process accepted by customer

Testing Chatbots – As AI

- When AI is data-driven, it acts as dynamical system
- Check for trust
 - Performance – **same as software**
 - Robustness
 - Human-AI value alignment
 - Human-AI interaction
- Performance assessment
 - Testcases for correct responses
 - Testcases for avoidable situations

Testing Chatbots – As AI

- Check for
 - Performance – same as software
 - **Robustness**
 - **Human-AI value alignment**
 - **Human-AI interaction**
- Analyze
 - Model card – if available // value alignment
 - Risk assessment // robustness, alignment; **role of AI trust standards**
 - Explanations and ratings // understand risks, interaction
- Create process to handle dynamic issues

Discussion: ChatGPT, LLM Failures

Refer to reading list: <https://github.com/biplav-s/course-tai-s25/blob/main/reading-list/Readme-LLMs.md>

ChatGPT Failures

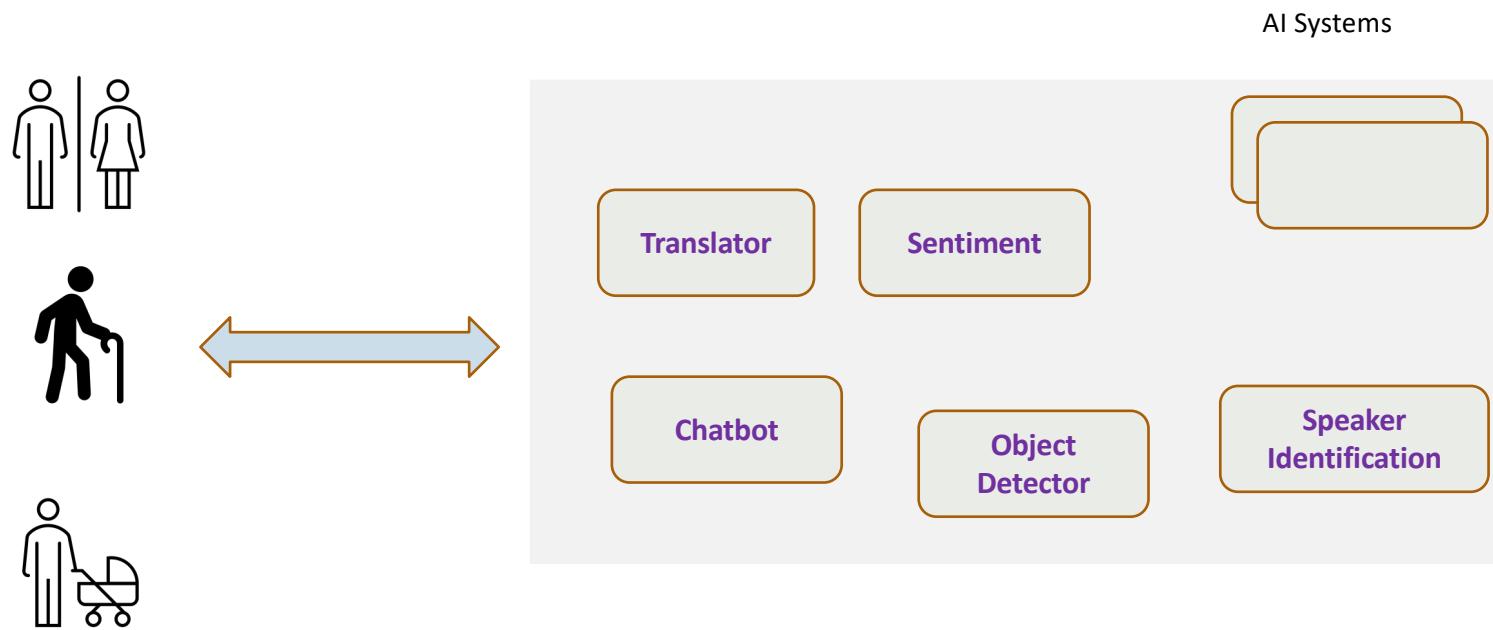
- Detailed analysis: https://github.com/aliborji/ChatGPT_Failures/blob/main/list.md
 - Paper: A Categorical Archive of ChatGPT Failures, <https://arxiv.org/abs/2302.03494>
- Types
 - Reasoning
 - Logic
 - Math and arithmetic
 - Factual errors
 - Bias and discrimination
 - Coding
 - Language – syntax, semantics, ...
 - Self-awareness
 - Misc

Credit: <https://github.com/giuven95/chatgpt-failures>

ChatGPT Failures and Reasons

- **Lack of a world model:** Models like ChatGPT do not have a "world model" in the sense that they do not have a comprehensive understanding of the physical and social world, and they do not have the ability to reason about the relationships between different concepts and entities. They are only able to generate text based on patterns they have learned from the training data.
- **Lack of a retrieval model:** Models like ChatGPT do not have the ability to retrieve information from an external memory or database. This means that they may not accurately recall facts.
- **Lack of character-level embeddings:** Many models like ChatGPT are not trained using character-level embeddings. This can lead to issues with out-of-vocabulary words and misspellings, as well as a lack of understanding of the relationships between different characters within a word.
- **Conditioning the chatbot on its own previous output:** When generating an answer, ChatGPT receives the history of the conversation as part of its input. This includes its own previous output, which can lead to the model generating text that is repetitive or redundant.
- **Poor dataset quality:** Models like ChatGPT are trained on extremely large datasets of widely varying quality, like the Common Crawl. This can lead to the model learning to imitate human biases, discrimination and inaccuracy.
- **RLHF can induce sycophancy bias:** The model may learn to prioritize pleasing the human feedback provider over actually achieving the intended task, leading to suboptimal performance. An example is the tendency for ChatGPT to agree with the interlocutor even when they're wrong

Technology and People



Trust: Can people trust AI systems to perform capably, consistently, and with human values?

EU

Risks	Where	What to do
Unacceptable	Threats to safety, livelihoods and rights of people	Ban
High	Infrastructure, education, employment, law enforcement, immigration, justice	Regulate - Logging; data quality; check robustness, security and accuracy; info about deployer; human oversight
Limited	Applications where risk is due to lack of transparency in info about AI usage.	Transparency obligations - Indicate interaction with machine
Minimal		Ignore

Note: risk levels are not clearly defined

Sources:

1. AI Act: <https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai>
2. GDPR/Data privacy: <https://gdpr-info.eu/>
3. <https://digital-strategy.ec.europa.eu/en/policies/ai-office>



Credit: [1]

EU regulations mandate

- stronger privacy standards, stricter transparency laws, and steep non-compliance penalties with fines of up to €30 million (nearly \$33 million) or 6% of global income.
- European AI Office (an EU AI board) setup in Feb 2024 will oversee the implementation and uniform application of the regulations, which will build on existing GDPR and Intellectual Property legislation [3]

US

- Five principles when automated systems can impact the public's rights, opportunities, or access to critical needs: Safe and Effective Systems; Algorithmic Discrimination Protections; Data Privacy; Notice and Explanation; Human Alternatives, Consideration, and Fallback [1]
- Developers of AI systems that could pose risks to U.S. national security, the economy, public health, or safety share the results of safety tests with the U.S. government, in line with the Defense Production Act, before they are made public.[2-1]
- Standards, tools, and tests be developed to ensure AI systems are safe, secure, and trustworthy. [2-2]
 - Calls on the Department of Commerce to develop guidance for content authentication and watermarking to clearly label AI-generated content to protect Americans from AI-enabled fraud and deception. [3]
 - Calls for safeguards to protect Americans' privacy, address algorithmic discrimination, and other measures. [3]
- **NIST Risk Management Framework**

1. NIST Risk Management Framework
 1. <https://airc.nist.gov/airmf-resources/airmf/>
 2. Nandhini Swaminathan, David Danks, Application of the NIST AI Risk Management Framework to Surveillance Technology, <https://arxiv.org/abs/2403.15646>, 2024
2. White house
 1. AI Bill of Rights, <https://www.whitehouse.gov/ostp/ai-bill-of-rights/>
 2. White House Executive Order, <https://www.whitehouse.gov/briefing-room/statements-releases/2023/10/30/fact-sheet-president-biden-issues-executive-order-on-safe-secure-and-trustworthy-artificial-intelligence/> [Oct 2023]
3. Esther Shein. 2024. Governments Setting Limits on AI. Commun. ACM 67, 4 (April 2024), 12–14. <https://doi.org/10.1145/3640506>

NIST RMF



<u>Prepare</u>	Essential activities to prepare the organization to manage security and privacy risks
<u>Categorize</u>	Categorize the system and information processed, stored, and transmitted based on an impact analysis
<u>Select</u>	Select the set of NIST SP 800-53 controls to protect the system based on risk assessment(s)
<u>Implement</u>	Implement the controls and document how controls are deployed
<u>Assess</u>	Assess to determine if the controls are in place, operating as intended, and producing the desired results
<u>Authorize</u>	Senior official makes a risk-based decision to authorize the system (to operate)
<u>Monitor</u>	Continuously monitor control implementation and risks to the system

Credit: <https://airc.nist.gov/airmf-resources/airmf/>

Table 1: AI System Risk Categorization Matrix (ASRCM)

Attribute	Categories	Clearview AI's facial recognition system (2020)
Use case	Description	Primarily for law enforcement to identify suspects by matching faces with a vast internet-sourced database.
Potential impact	Health/Social/Economic	Health: Minimal. Social: Significant, raises privacy and consent issues. Economic: Impacts security sector, potential cost savings and privacy breach costs.
Data sources	Sensitivity: Sensitive/Non-sensitive Accessibility: Restricted/Public Regulatory Compliance: Compliant/Non-compliant	Sensitivity: Highly sensitive (biometric data). Accessibility: Restricted to approved entities. Regulatory Compliance: Contentious, varies by region.
Level of complexity	White/Gray/Black box	Black box - algorithmic specifics are undisclosed.
Regulatory requirements	Yes/No	Yes, subject to privacy and biometric data laws, compliance varies.
Level of autonomy	Low/Mid/High	High; Requires human oversight for result interpretation.

Credit: Nandhini Swaminathan, David Danks, Application of the NIST AI Risk Management Framework to Surveillance Technology, <https://arxiv.org/abs/2403.15646>, 2024

China

Cyberspace Administration of China

1. Three sets of regulations on
 1. algorithmic recommender systems
 2. synthetically generated content
 3. the use of generative AI.
2. Developers need to file with China's new algorithm repository, which gathers information on how algorithms are trained, along with requiring them to pass a security self-assessment

Source:

Esther Shein. 2024. Governments Setting Limits on AI. Commun. ACM 67, 4 (April 2024), 12–14. <https://doi.org/10.1145/3640506>

References on AI Regulations

1. Global
 1. Esther Shein. 2024. Governments Setting Limits on AI. *Commun. ACM* 67, 4 (April 2024), 12–14.
<https://doi.org/10.1145/3640506> [Summarizes global situation]
 2. International Association of Privacy Professionals (IAPP) has developed a [Global AI Legislation Tracker](https://iapp.org/media/pdf/resource_center/global_ai_law_policy_tracker.pdf), a live repository of over 1,000 AI policy initiatives from 69 countries. Report:
https://iapp.org/media/pdf/resource_center/global_ai_law_policy_tracker.pdf [Explains country-specific laws, acts, directives and drafts]
 3. Hauptfleisch, W. "Where the World is on AI Regulation." October 2023. *Medium*, <https://bit.ly/49obPaT>.
 4. Zhu, K., The State of State AI Laws: 2023. Electronic Privacy Information Center, <https://epic.org/the-state-of-state-ai-laws-2023/>
2. EU
 1. AI Act: <https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai>
 2. GDPR/Data privacy: <https://gdpr-info.eu/>
 3. "EU AI Act: first regulation on artificial intelligence." European Parliament, August 2023, <https://bit.ly/3uOh8RF>
 4. Hoffmann, M. The EU AI Act: A Primer. Center for Security and Emerging Technology. September 2023, <https://cset.georgetown.edu/article/the-eu-ai-act-a-primer/>
3. US
 1. NIST Risk Management Framework, <https://airc.nist.gov/airmf-resources/airmf/>
 2. BSA Analysis: State AI Legislation Surges by 440% in 2023. BSA/The Software Alliance. September 2023, <https://bit.ly/3T6xK0x>.
4. Africa
 1. Looking into the crystal ball: Artificial intelligence policy and regulation in Africa. The Centre for Africa-Europe Relations. September 2023, <https://bit.ly/42NCK72>.
5. China
 1. Sheehan, M., China's AI Regulations and How They Got Made. July 2023. Carnegie Endowment For International Peace, <https://bit.ly/3I4eSJg>

Future: Understanding AI and Its Impact on Human Decision Making

- Good decision making steps
 - Actively seeking out information that both fills gaps in our knowledge and challenges one's prior beliefs and assumptions.
 - Seeking out and considering more than one option at a time.
 - Willing to delay closure on a decision until after one had done all the necessary heavy mental lifting.
- Making thoughtful and defensible decisions requires practice and self-discipline.
- Practice of decision making
 - Humans are well known for being prone to a wide range of biases because we tend to be frugal when it comes to expending mental energy.
 - Humans are social animals who tend to value the security and acceptance of their communities more than they might value their own autonomy.

Source: <https://theconversation.com/the-hidden-risk-of-letting-ai-decide-losing-the-skills-to-choose-for-ourselves-227311>

Project Discussion

Course Project

•Framework

1. (Problem) Think of a problem whose solution may benefit people (e.g., health, water, air, traffic, safety)
2. (User) Consider how the primary user (e.g., patient, traveler) may be solving the problem today
3. (AI Method) Think of what the solution will do to help the primary user
 1. Solution => ML task (e.g. classification), recommendation, text summarization, ...
 2. Use a foundation model (e.g., LLM-based) solution as the baseline
4. (Data) Explore the data for a solution to work
5. (Reliability: Testing) Think of the evaluation metric we should employ to establish that the solution will work? (e.g., 20% reduction in patient deaths)
6. (Holding Human Values) Discuss if there are fairness/bias, privacy issues?
7. (Human-AI) Finally, elaborate how you will explain the primary user that your solution is trustable to be used by them

Project Discussion: What to Focus on ?

- Problem: you should care about it
- Data: should be available
- Method: you need to be comfortable with it. Have at least two – one serves as baseline
- Trust issue
 - Due to Users
 - Diverse demographics
 - Diverse abilities
 - Multiple human languages
 - Or other impacts
- What one does to mitigate trust issue

Rubric for Evaluation of Course Project

Project

- Project plan along framework introduced (7 points)
- Challenging nature of project
- Actual achievement
- Report
- Sharing of code

Presentation

- Motivation
- Coverage of related work
- Results and significance
- Handling of questions

<Project Title> - <Your Name>

Format for Interim Presentation
on April 22, 2025

Project Context

1. Title:
2. Key idea: (2-3 lines)
3. Who will care when done:
4. Data need:
5. Methods:
6. Evaluation:
7. Users:
8. Trust issue:

- Test Case – demonstrate working
 - E.g., <input, output, correct output – if different, trust observation>

1 min context, 2 min demo, 1 min expts, 1 min Q/A

<Project Title> - <Your Name>

Format for Interim Presentation
on April 22, 2025

Demonstrate effectiveness/ efficiency

- Metrics (F1, running time, ...)
- Empirical results
- Comparison with a LLM (why your method over a general alternative)

Conclusion

- Experience
- Q/A

1 min context, 2 min demo, 1 min expts, 1 min Q/A

Concluding Section

Week 13 (L25 and 26): Concluding Comments

- We looked at
 - L25: Human-AI Collaboration, Chatbots
 - L26: Emerging AI Trust Landscape: Standards, Privacy

About Next Week – Lectures 27, 28

Lectures 27, 28

- Project presentations

19	Mar 25 (Tu)	AI - Unstructured (Text): Representation, Common NLP Tasks, Large Language Models (LLMs)
20	Mar 27 (Th)	Natural Languages/ Language Models and their Impact on AI
21	Apr 1 (Tu)	AI - Unstructured (Text): Analysis – Supervised ML – Trust Issues
22	Apr 3 (Th)	AI - Unstructured (Text): Analysis – Supervised ML – Mitigation Methods
23	Apr 8 (Tu)	AI - Unstructured (Text): Analysis – Rating and Debiasing Methods
24	Apr 10 (Th)	Explanation Methods Trust: AI Testing
25	Apr 15 (Tu)	Trust: Human-AI Collaboration
26	Apr 17 (Th)	Emerging Standards and Laws Trust: Data Privacy - Trusted AI for the Real World
27	Apr 22 (Tu)	Project presentation
28	Apr 24 (Th)	Project presentation
29	Apr 29 (Tu)	Paper presentations
	May 1 (Th)	
30	May 6 (Tu)	4pm – Final exam/ Overview