CS224v

Conversational Virtual Assistants with Deep Learning

Lecture 1: Introduction

Monica Lam

FOCUS OF THE COURSE

FOUNDATION TO MAKE LLMS USEFUL AS A TRUSTWORTHY GENERAL VIRTUAL ASSISTANT

90+% ACCURACY

APPLICABLE TO ALL DOMAINS (FOR NON-AI EXPERTS)

A Project/Research Course

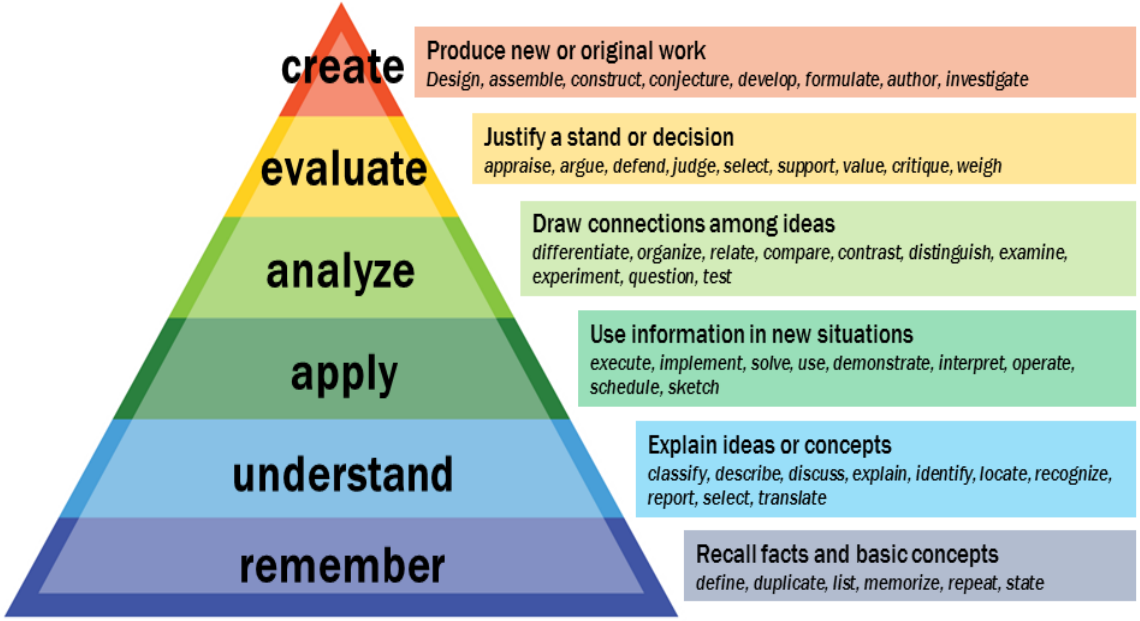
Actively Create and Advance New Technologies
Adopted by Real Users

Select Class Projects Continue as Research (2026)

Course Publications: 2023(2), 2024(5), 2025(6)

Secret to Success: **Working Software**
Supports Increasingly Complex Projects

Goal: an AI-Based Research Assistant



Produce new or original work
Design, assemble, construct, conjecture, develop, formulate, author, investigate

Justify a stand or decision
appraise, argue, defend, judge, select, support, value, critique, weigh

Draw connections among ideas
differentiate, organize, relate, compare, contrast, distinguish, examine, experiment, question, test

Use information in new situations
execute, implement, solve, use, demonstrate, interpret, operate, schedule, sketch

Explain ideas or concepts
classify, describe, discuss, explain, identify, locate, recognize, report, select, translate

Recall facts and basic concepts
define, duplicate, list, memorize, repeat, state

Vanderbilt University Center for Teaching

Highest Level of Cognition
Subsumes all other Levels

Bloom's Taxonomy
Education Objectives in Cognition Domain

WHY AN AI-BASED RESEARCH ASSISTANT?

TO ADVANCE KNOWLEDGE DISCOVERY

FASTER

BETTER RESULTS

SCALABILITY

BROADER (BREAKING DOWN DOMAIN & LANGUAGE BARRIERS)

RESEARCH ACCESSIBLE TO MORE PEOPLE

Knowledge As We Know It

BBC 2025

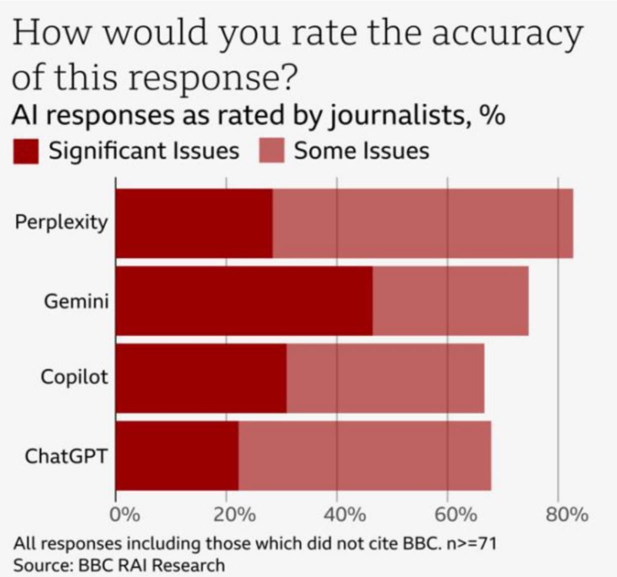
Perplexity, Gemini, Copilot, ChatGPT Still Hallucinate With RAG

(Retrieval Augmented Generation)

51% of AI answers to news problems have significant issues.

19% of AI answers which cited BBC content introduced factual errors.

13% of the quotes sourced from BBC articles were either altered or didn’t actually exist in that article.



How would you rate the accuracy of this response?

AI responses as rated by journalists, %

All responses including those which did not cite BBC. n>=71

Source: BBC RAI Research

Representation of BBC News content in AI Assistants. 2025.

The Problem
The Enterprise AI Paradox

Enterprises are rushing to deploy AI,
but off-the-shelf LLMs are a ticking time bomb.

Legal and Financial Liability

Air Canada forced to honor fictitious refund policy

Reputational Damage

Lawyer sanctioned over ChatGPT-fabricated legal precedents

Regulatory Risk

NYC's AI chatbot advised businesses to break laws

The result: A crisis of trust that blocks AI adoption for high-value use cases.

THE 70% ACCURACY PROBLEM

IT'S DECEIVINGLY EASY!

IMPOSSIBLE TO GET TO NEAR 100%
WITH NEXT-WORD PREDICTION LLMS!

NEED OF HUMAN FILTERING → NOT SCALABLE!

FUNDAMENTAL PROBLEMS FOR HARDER TASKS

LLMS TRAINED ON NEXT-WORD PREDICTION:

NOT **PRECISE** ENOUGH FOR RESEARCH

CANNOT PERFORM **COMPLEX** TASKS RELIABLY

LIMITED **CONTEXT LENGTH**

Approach

Augment LLMs with **Computational Thinking**

Algorithms

Data Representation (Pattern Recognition)

Abstraction (generalization)

Decomposition

Universal to All Disciplines

COMPUTATIONAL THINKING

BASIC CONCEPTS: ALAN PERLIS, DON KNUTH (1950S)

COINED TERM: SEYMOUR PAPERT (1980)

EDUCATION OBJECTIVE: JEANNETTE WING, NSF (2006)

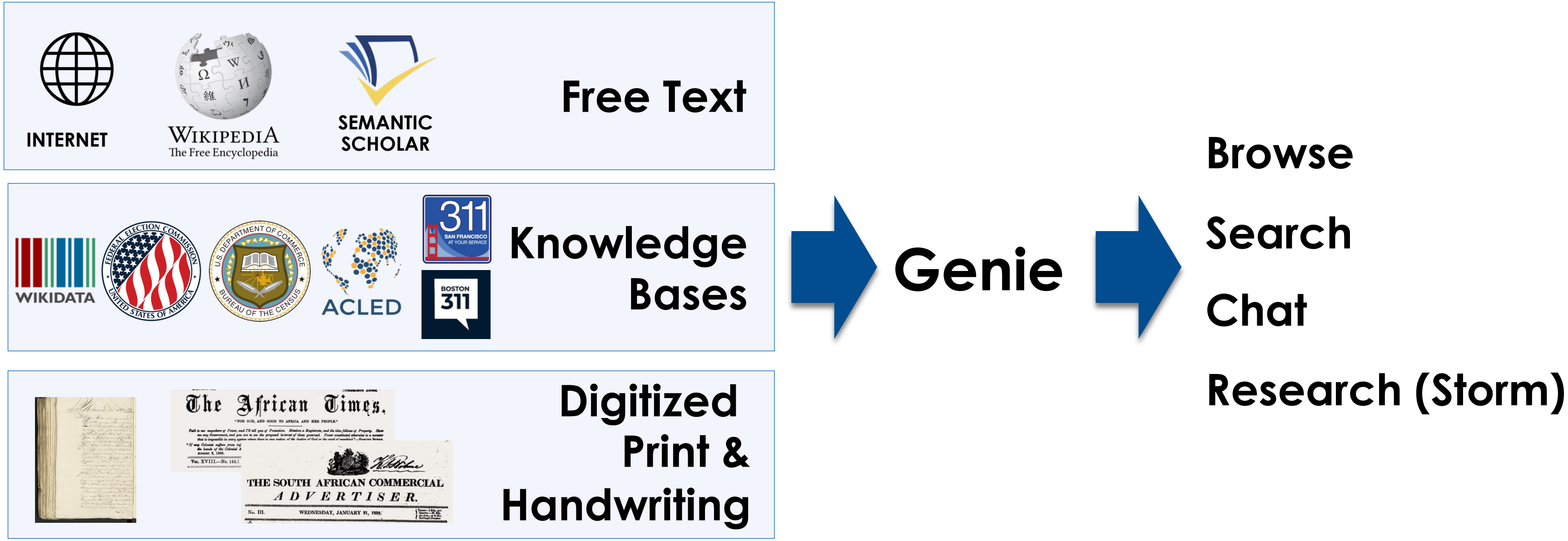
Lecture Outline

1. Course Material Overview
   - Stage 1 (2022 -- 2025): Computational thinking → general research assistant
   - Stage 2 (2025 --): Computational thinking → scientific research assistant

2. Course overview
   - Project choice
     - Build on technology for a general assistant
     - Research the bleeding edge of a technical research assistant

Just a high-level picture! Do not worry if you can't follow all the ideas
(Details in the lectures)

Genie: A Prototype of a Public AI Assistant to World Wide Knowledge (WWK)


INTERNET
WIKIPEDIA
The Free Encyclopedia
SEMANTIC SCHOLAR
WIKIDATA
FEDERAL ELECTION COMMISSION · UNITED STATES OF AMERICA
U.S. DEPARTMENT OF COMMERCE · BUREAU OF THE CENSUS
ACLED
311 SAN FRANCISCO AT YOUR SERVICE
BOSTON 311
The African Times.
THE SOUTH AFRICAN COMMERCIAL ADVERTISER.

Free Text

Knowledge Bases

Digitized Print & Handwriting

Genie

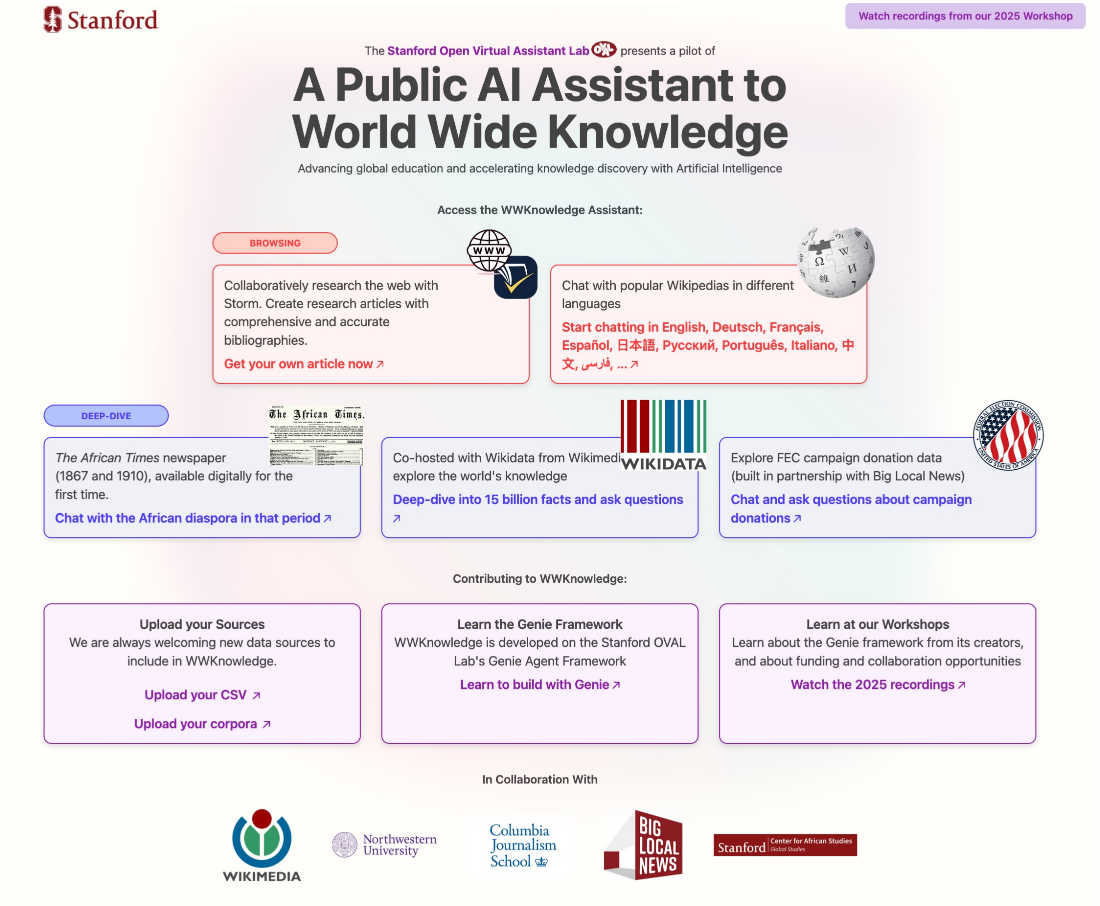
Browse

Search

Chat

Research (Storm)

Live Demos: wwknowledge.org



Stanford

Watch recordings from our 2025 Workshop

The **Stanford Open Virtual Assistant Lab** presents a pilot of

A Public AI Assistant to World Wide Knowledge

Advancing global education and accelerating knowledge discovery with Artificial Intelligence

Access the WWKnowledge Assistant:

BROWSING

Collaboratively research the web with Storm. Create research articles with comprehensive and accurate bibliographies.

Get your own article now ↗

Chat with popular Wikipedias in different languages

Start chatting in English, Deutsch, Français, Español, 日本語, Русский, Português, Italiano, 中文, فارسی, ... ↗

DEEP-DIVE

The African Times newspaper (1867 and 1910), available digitally for the first time.

Chat with the African diaspora in that period ↗

Co-hosted with Wikidata from Wikimedia, explore the world's knowledge

Deep-dive into 15 billion facts and ask questions ↗

Explore FEC campaign donation data (built in partnership with Big Local News)

Chat and ask questions about campaign donations ↗

Contributing to WWKnowledge:

Upload your Sources

We are always welcoming new data sources to include in WWKnowledge.

Upload your CSV ↗

Upload your corpora ↗

Learn the Genie Framework

WWKnowledge is developed on the Stanford OVAL Lab's Genie Agent Framework

Learn to build with Genie ↗

Learn at our Workshops

Learn about the Genie framework from its creators, and about funding and collaboration opportunities

Watch the 2025 recordings ↗

In Collaboration With

WIKIMEDIA

Northwestern University

Columbia Journalism School

BIG LOCAL NEWS

Stanford | Center for African Studies

COMPUTATIONAL THINKING APPROACH

FORMALIZE HUMAN COGNITIVE PROCESS

PRODUCE STEP-BY-STEP INSTRUCTIONS

1. READING DOCUMENTS

RAG-BASED LLMS HALLUCINATE

ESPECIALLY WHEN RETRIEVED INFORMATION DOES NOT ANSWER THE USERS' QUESTION

Genie: Automating Question Answering

Ansatze (an educated guess)

3. Ask LLM
4. Dissect into claims
5. Fact-check each claim
    - Search Wikipedia with claim
    - Filter incorrect claim

Retrieve

1. Search Wikipedia with query
2. Filter irrelevant info

6. Draft
7. Refine

Use multiple, easy LLM steps to consult external data

WikiChat

- Accuracy: 97% (English)
- Best Research Award of the Year, Wikimedia Foundation
- Now in 25 languages: Breaking the language barrier!

Arabic
Chinese
Czech
Danish
Dutch
English
Finnish
French
German
Hebrew
Indonesian
Italian
Japanese
Korean
Norwegian
Persian
Polish
Portuguese
Romanian
Russian
Serbian
Spanish
Swedish
Turkish
Ukrainian

WikiChat: Stopping the Hallucination of Large Language Model Chatbots by Few-Shot Grounding on Wikipedia Sina J. Semnani, Violet Z. Yao*, Heidi C. Zhang*, Monica S. Lam
In EMNLP 2023, Singapore, December 6-10, 2023.

The problem is Harder than Expected

3 students, 4 months for Development

Evaluation Was Also Hard

The Devil is In the Details

Solid Building Blocks are Critical

Applicable to General Cognitive Processes

2. READING DATA

Example:

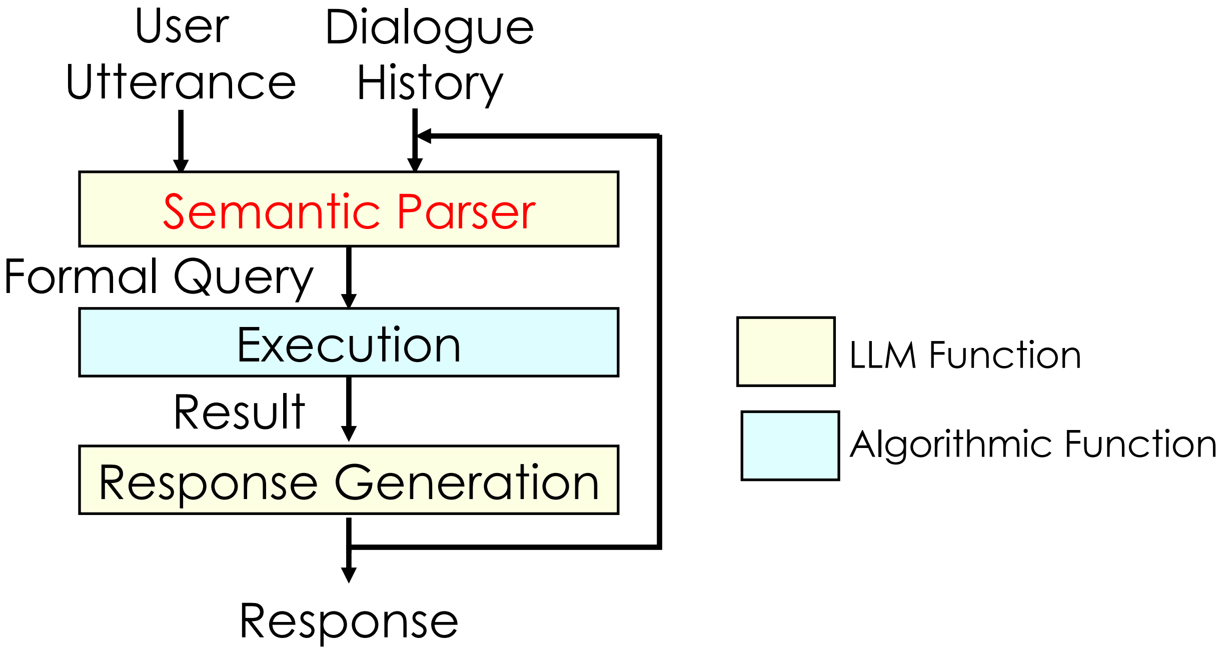
WIKIDATA

World’s Largest Live Knowledge Graph

- 15B facts, 100M entities, 10K properties, 25K contributors
- Wikidata could only be accessed by **SPARQL query language**

Key Concept:

A Semantic Parser Translates NL to Formal Semantics



User Utterance

Dialogue History

Semantic Parser

Formal Query

Execution

Result

Response Generation

Response

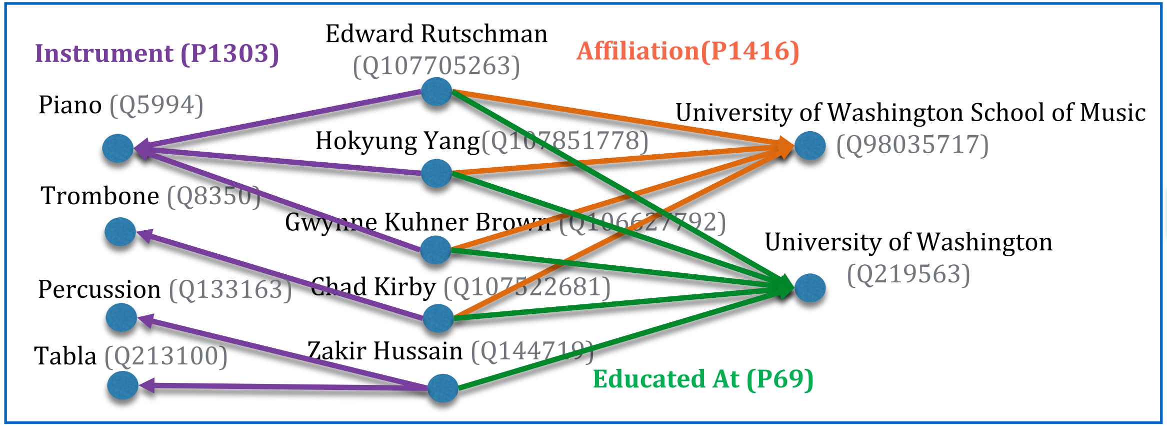
🟨 LLM Function

🟦 Algorithmic Function

What are the musical instruments played by people
who are affiliated with the U. of Washington School of Music
and have been educated at the U. of Washington, and how many people play each instrument?

Genie

```
SELECT ?instrument ?instrumentLabel (COUNT(?student) AS ?count) WHERE {
  ?student wdt:P1303 ?instrument;
    wdt:P1416 wd:Q98035717;
    wdt:P69 wd:Q219563.
  SERVICE wikibase:label { bd:serviceParam wikibase:language "en". }
}
GROUP BY ?instrument ?instrumentLabel
```



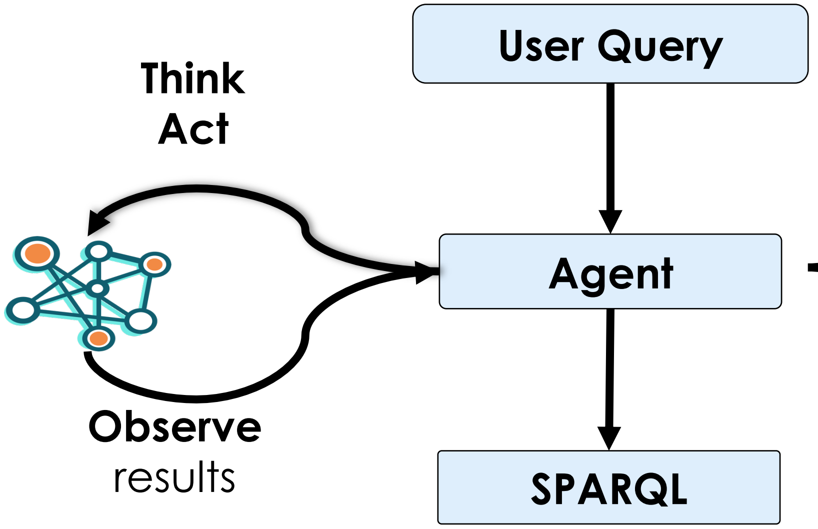
A Small Subset of Wikidata

```
| instrument    | instrumentLabel  | count |
|---------------|------------------|-------|
| Q5994         | piano            | 99    |
| Q1467960      | mbira            | 2     |
| Q8350         | trombone         | 11    |
| Q8338         | trumpet          | 8     |
| Q17172850     | voice            | 32    |
| ...           | ...              | ...   |
| Q302497       | mandolin         | 1     |
| Q187851       | recorder         | 1     |
| Q185041       | cor anglais      | 1     |
| Q83509        | piccolo          | 1     |
```

Result

Break Down the Cognitive Process

Go acquire knowledge to complete a task!



Think
Act

Observe
results

User Query

Agent

SPARQL

```
Possible Actions
search_wikidata(string)
   returns QIDs and PIDs
get_wikidata_entry(QID)
   returns Wikidata entity page
execute_sparql(SPARQL)
   returns SPARQL execution results
get_property_examples(PID)
   example of how properties are used
stop()
```

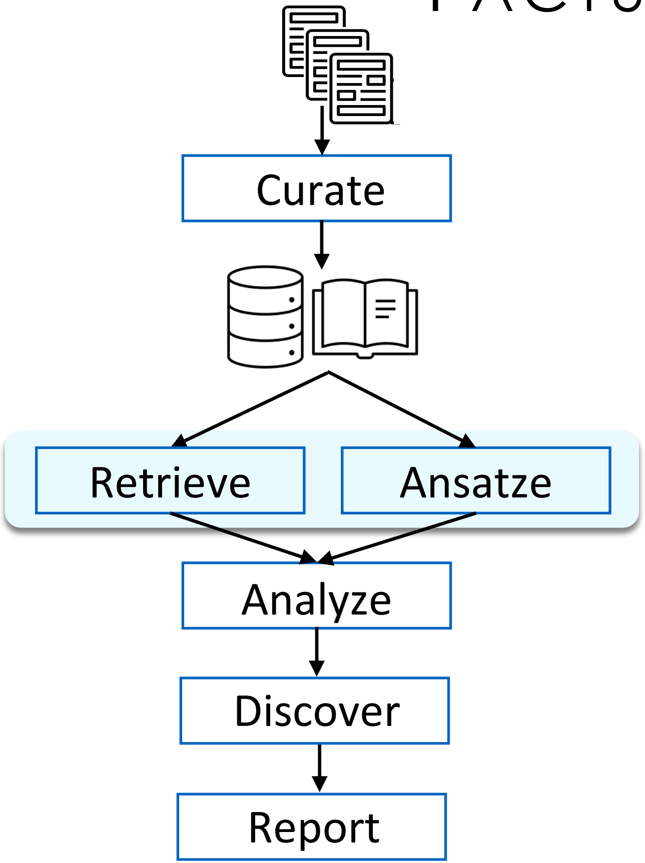
Deployed at Wikidata Query Forum

https://www.wikidata.org/wiki/Wikidata:SPARQL_query_service/Wikidata_Query_Help

- Help users with their complex queries
  - 1700 conversations
  - In 198 samples, success rate: 78%
  - Interpretability:
    - Genie shows the query (in English) and its answer

SPINACH: SPARQL-Based Information Navigation for Challenging Real-World Questions
Shicheng Liu*, Sina J. Semnani*, Harold Triedman, Jialiang Xu, Isaac Dan Zhao, Monica S. Lam
In Findings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP) , Miami, Florida, November 12 –16, 2024.

FACTUAL INFO RETRIEVAL



Curate

Retrieve

Ansatze

Analyze

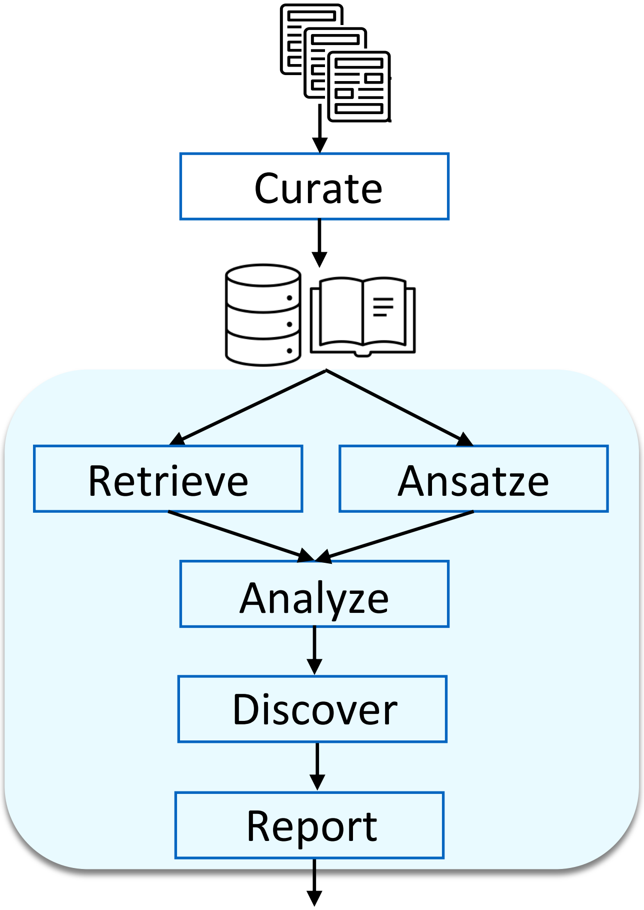
Discover

Report

Research Pipeline

Input: Text, Databases, Knowledge Bases (Wikipedia, Wikidata, FEC …)

2023/12 EMNLP	WikiChat: **Control hallucination** by grounding in **Wikipedia** - **Best chat: achieving 97% factuality** - Now grounded in Wikipedia in 25 languages - **Best Research of the Year Award by Wikimedia Foundation, 2024**
2024/ 6 NAACL	1st assistant that supports **full hybrid queries on free-text + databases** - Translate NL into our **novel SUQL query language** with optimizing compiler - **Used by investigative journalists (FEC campaign donation), 311**
2024/12 EMNLP	Best assistant for **complex queries** on **Wikidata** - Agentic approach that learns about the schema on its own - **Deployed on Wikidata**
2025/ 7 ACL	Genie Worksheets: **Knowledge-Intensive Task-Oriented Agents** - Accuracy: 80% vs 0-10% with GPT4 with function calling - **Validation with industry in progress**



Curate

Retrieve

Ansatze

Analyze

Discover

Report

DISCOVERY OF GENERAL KNOWLEDGE

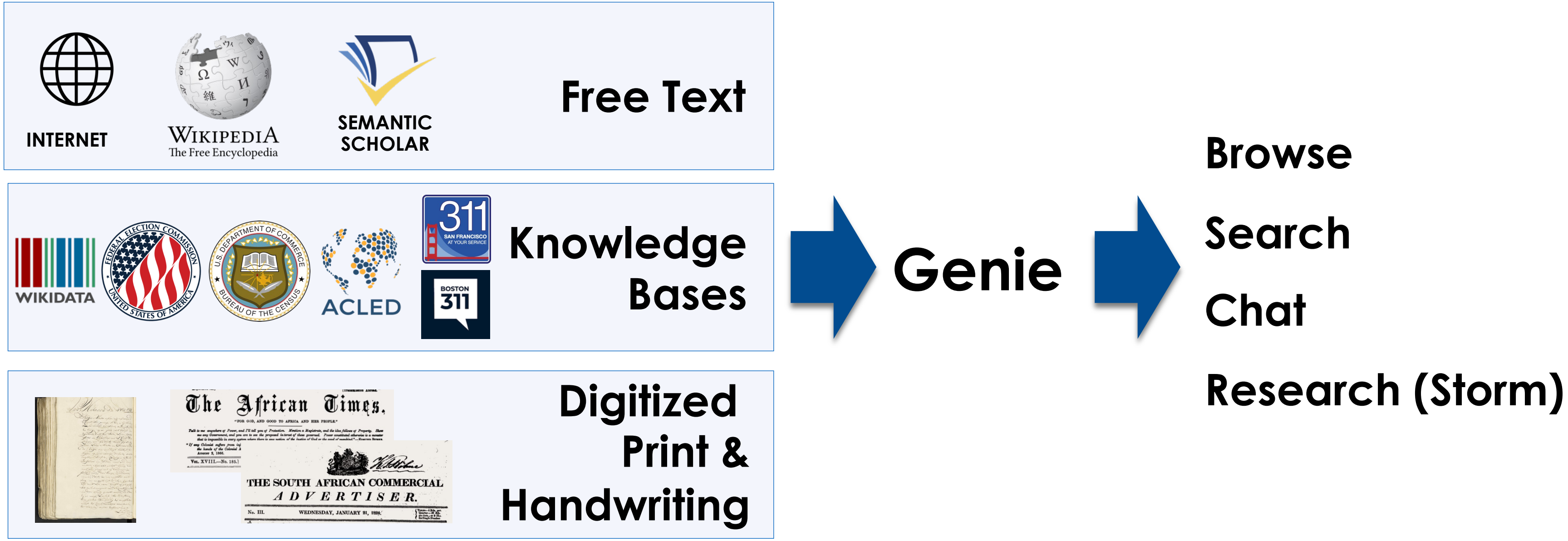
Input: Text, Databases, Knowledge Bases (Wikipedia, Wikidata, FEC …)

2024/ 6 NAACL	STORM: 1st **deep research** assistant by searching the **Internet** - Comprehensive Wikipedia-like articles with citations to full references - 800K organic users, 1.4M articles written - **Used by history students** to discover African history newspapers (1800s) - **Inspired OpenAI Deep Research, Google Gemini Deep Research, Databricks Genie Deep Research Model**
2024/12 EMNLP	1st assistant that lets users **interactively direct the deep research** - Uses a novel round-table of experts to explore unknown unknowns

New: DATASTORM discovers insights by combining Storm + DataTalk

GENIE: A PROTOTYPE OF A PUBLIC AI ASSISTANT TO WORLD WIDE KNOWLEDGE (WWK)

Demos: wwknowledge.org



Free Text

Knowledge Bases

Digitized Print & Handwriting

Genie

Browse

Search

Chat

Research (Storm)

NLP Research Can Create Useful Tools

Usage Also Helps Generate Research Ideas

Lecture Outline

1. Course Material Overview

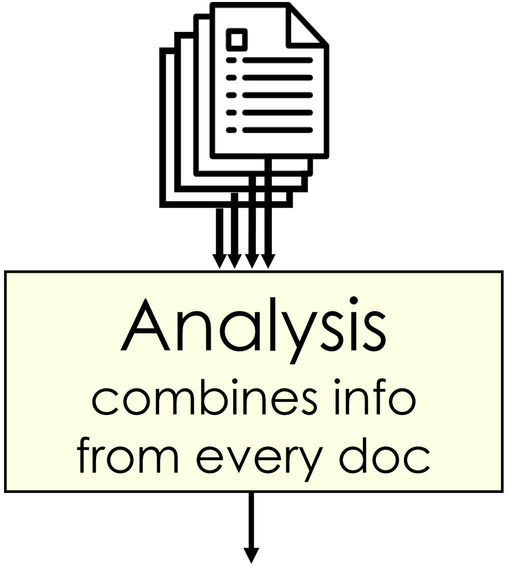
   - Stage 1 (2022 -- 2025): Computational thinking → general research assistant
   - Stage 2 (2025 --   ): Computational thinking → scientific research assistant

2. Course overview
   - Project choice
     - Build on technology for a general assistant
     - Research the bleeding edge of a technical research assistant

SCIENTIFIC RESEARCH
REQUIRES CHANGING THE FUNDAMENTALS OF NLP
WITH COMPUTATION TRAINING (CT)

1. INFORMATION RETRIEVAL (EMBEDDING SIMILARITY)
   → FORMAL SEMANTICS
2. REASONING: CHAIN OF THOUGHT
   → COMPUTATIONAL THINKING (CT)
3. TRAINING: UNSUPERVISED NEXT-WORD PREDICTION TRAINING
   FINE-TUNED WITH CHAIN OF THOUGHT
   → FINE-TUNED WITH CT TRAINING

Knowledge in Large Sets of Long Documents



Analysis

combines info from every doc

Decision, Stats, Trend, …

Resumes Paper Reviews Proposals Financial Reports	Pubmed Articles Human Genome Drug/Protein/Disease Event News	News Social Media Interviews Medical Rec.	Requirements Regulations Clinical Trials Invoice Match
Ranking (Top n) of ALL documents	Accumulation	Stats/trends	Satisfiability
Finalists	Knowledge Bases	Discover Knowledge	Compliance to Constraints

An Example: Clinical Trials

- Motivation
  - Patients cannot navigate clinical trial websites
  - Clinical trials fail for a lack of finding patients
- Challenges: matching millions of trials with millions of patients
  - Informal Retrieval (IR)
    - Similarity → false positives; false negatives
  - Matching
    - LLM “reasoning” → inaccurate

Missing one appropriate trial is fatal!

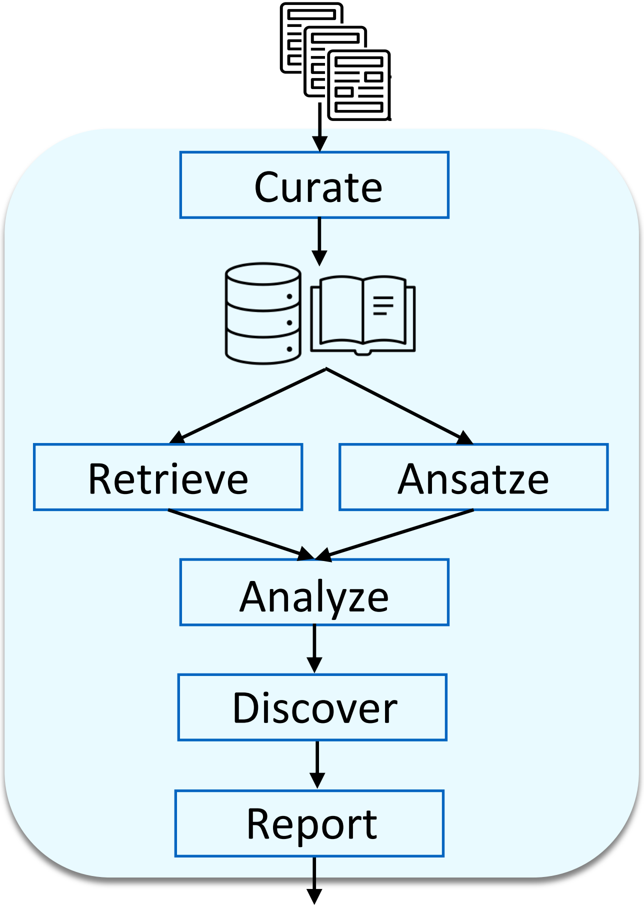
Approach for Effective Clinical Trial Matching

1. Representation
    - Translate free-text trials/patient records to formal semantics
        - SMT (Satisfiability Modulo Theories)
        - UMLS (Universal Medical Language System) with 3M terms
2. Retrieval algorithm
    - Store trials in DBs with columns representing simpler predicates
    - Retrieve with DB queries
3. Match algorithm: SMT theorem prover

Curating Knowledge

Semantic Parsing from Free-Text to:

- **Databases**:
  for analyzing large & long document sets
  - Sets of papers, articles, reports, transcripts
- **Knowledge graphs**: a semantic web
  - NSF Proto-OKN project
- **SMT: constraint satisfaction**
  - Degree satisfaction, regulation compliance, invoice matching



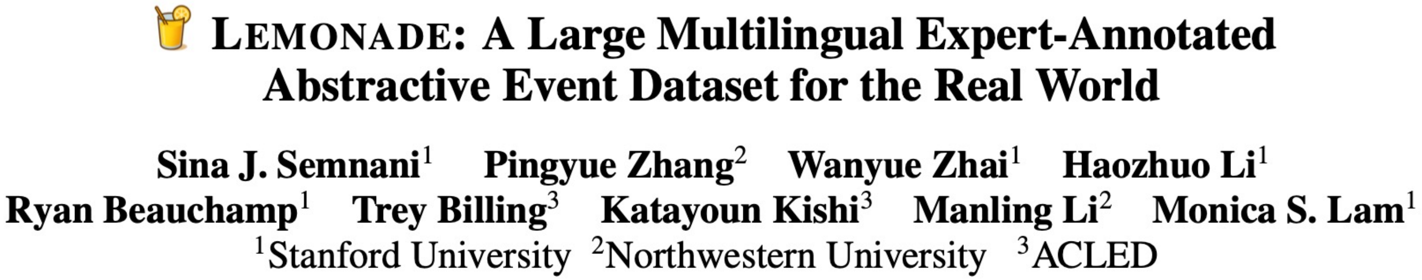
Full Research Pipeline

Knowledge Curation—Important but Hard Example:

ACLED is an independent, impartial conflict monitor providing real-time data and analysis on violent conflict and protest in all countries and territories across the world.

ACLED: Armed Conflict Location & Event Data

- Best dataset of qualitative coding (Free-text → DB)
- Human annotators, with rounds of reviews
- Over 200 territories, 80 different languages



Lemonade: A Large Multilingual Expert-Annotated Abstractive Event Dataset for the Real World

Sina J. Semnani[1] **Pingyue Zhang**[2] **Wanyue Zhai**[1] **Haozhuo Li**[1]
Ryan Beauchamp[1] **Trey Billing**[3] **Katayoun Kishi**[3] **Manling Li**[2] **Monica S. Lam**[1]
[1]Stanford University [2]Northwestern University [3]ACLED

In Findings of ACL, 2025

Academic Formulation Falls Short

Extractive → Abstractive

Canonicalization of Entities

Needed: Research on Entities in External Literature

SOTA, but Not Accurate Enough

Goal: Create A Universal Semantic Parser

Free Text, Sample Annotations → Formal Representation

1. Ontology: Automatic standardization of terminology
2. Translation to formal semantics. Challenges:
   - Classification among many entities
   - Acquisition of missing knowledge
   - Long documents exceeding LLM contexts
   - Consistency checking
   - Refinement with annotated samples
3. Information retrieval optimizations

3 Case Studies of Scientific Assistants

All the technology developed is generally applicable to any domain

You can join these new projects from the ground up.

1. **Clinical Trial Matching** (Mayo Clinic, Stanford Medicine) using **SMT**
    - Off-line matching & a conversational agent for patients looking for trials
2. **Cancer drug resistance**: Data Synthesis with literature in RNA-Seq Analysis (Stanford Oncology)
    - Analysis of experimental results
    - Extraction of **knowledge graph** of pubmed articles
    - Synthesizing data with literature

    Similar use case: **Repurposing drugs for rare diseases** (Everycure, Rare Genomics Inst.)
3. **Sustainability Modeling** (Stanford Cornerstone Initiative)
    - Synthesizing model from many different raw **databases** with judgement

SCIENTIFIC RESEARCH
REQUIRES CHANGING THE FUNDAMENTALS OF NLP
WITH COMPUTATION TRAINING (CT)

1. INFORMATION RETRIEVAL (EMBEDDING SIMILARITY)
   → FORMAL SEMANTICS
2. **REASONING: CHAIN OF THOUGHT**
   → COMPUTATIONAL THINKING (CT)
3. TRAINING: UNSUPERVISED NEXT-WORD PREDICTION TRAINING
   FINE-TUNED WITH CHAIN OF THOUGHT
   → FINE-TUNED WITH CT TRAINING

LLMs Lack Computational Thinking (CT)

Example: Composition

Question	OpenAI-03 Model
Who is the wife of Benjamin Harrison?	Caroline Harrison
Who is the grandfather of Caroline Harrison?	George Scott
Who's the grandfather of the wife of Benjamin Harrison?	Dr. John Witherspoon

Idea

Do not use LLMs to answer complex questions directly

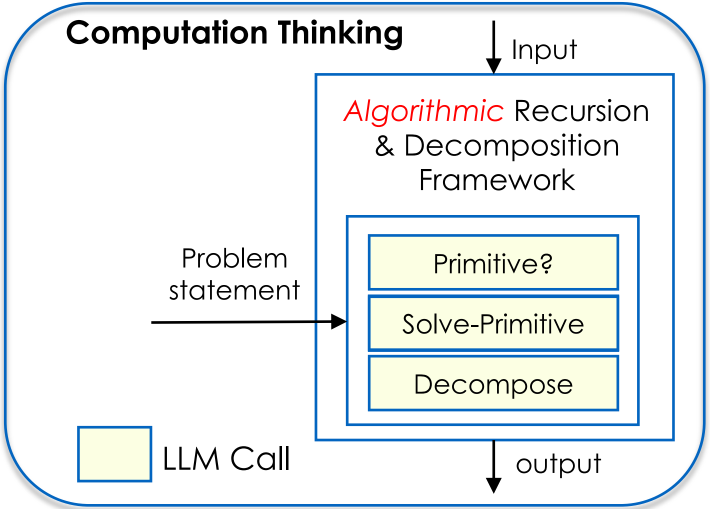
We propose:

1. Use LLM to perform simple functions
2. Use an Algorithmic Engine to invoke LLM functions and compose them

LLMs Lack Computational Thinking (CT)

Example: Composition

Question	OpenAI-03 Model
Who is the wife of Benjamin Harrison?	Caroline Harrison
Who is the grandfather of Caroline Harrison?	George Scott
Who's the grandfather of the wife of Benjamin Harrison?	



Computation Thinking

Input

Algorithmic Recursion & Decomposition Framework

Problem statement

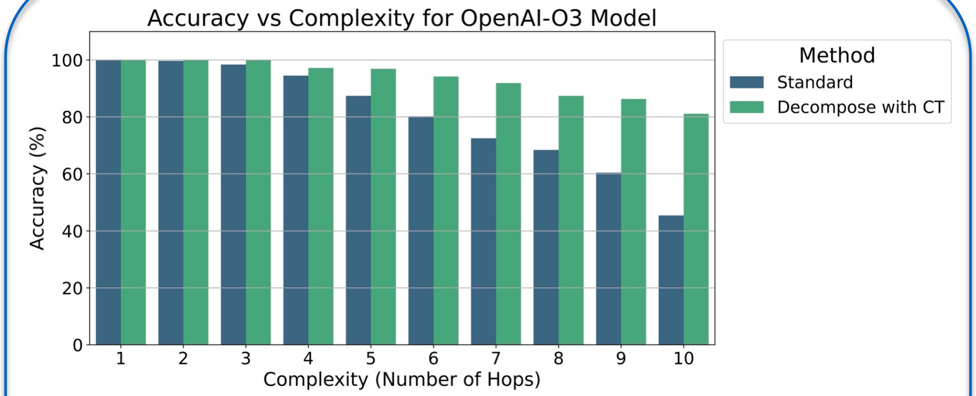
Primitive?

Solve-Primitive

Decompose

LLM Call

output



Dataset: LLM knows every hop of the question

Potential Applications

- Semantic Parser

This just in: Beats SOTA on Spider-2 Text-to-SQL

- Data Synthesis
- Coding
- Scientific and Engineering Workflows

Scientific Research
Requires Changing the Fundamentals of NLP
With Computation Training (CT)

1. Information Retrieval (Embedding Similarity)
   → Formal Semantics
2. Reasoning: Chain of Thought
   → Computational Thinking (CT)
3. **Training: Unsupervised Next-word Prediction Training**
   Fine-tuned with Chain of Thought
   → Fine-tuned with CT Training

Training CT

- Why?
  - To improve ansatze (educated guess) to aid knowledge discovery
    - Example: Ask LLM "What drug may cure this rare disease?" Literature search on the answer.
- How?
  - Fine-tune with training samples: (Natural language question → Formal Query; Answer)
  - Teach it formal semantics representation
  - Improve its latent representation of knowledge in LLMs
- Expected result: improved LLMs recall and ansatzes

THE CS224V COURSE

COURSE OBJECTIVES

FOUNDATION TO MAKE LLMS USEFUL AS A TRUSTWORTHY GENERAL VIRTUAL ASSISTANT

HANDS-ON PROJECT/RESEARCH EXPERIENCE
BUILDING ON THE STATE-OF-THE-ART TOOLS
OR
ADVANCING THE STATE-OF-THE-ART

Course Design

1. 2 homeworks to bring everybody onboard with SOTA tools
2. Lectures on techniques of LLM-based conversational agents
3. Supervised quarter-long project
   - Create a new application on existing tools
   - Enhance existing tools (with an application)
   - Develop new tools (with an application)

Purpose of the 2 Assignments

Prepare you for your project proposal

1. How to create an autonomous LLM-based research agent?
2. How to create a conversational agent using Genie Worksheet
   - Performs tasks; Answers question
   - Does not hallucinate

Project Apprenticeship

- Assistance with project selection: Hardest part in research!
  - We suggest over 20 research-oriented projects on the website
  - Student-initiated projects are hugely welcome

- Weekly group mentorship meeting
  - We want to make you succeed!

Project Mentorship

All homeworks and projects are to be done in pairs

- Week 4: Project proposal, with a weekly plan
- Weeks 5-10 (excluding Thanksgiving break):
    - Submit a written weekend update (every Monday)
    - Group meeting with mentors during the week
- Week 11: Poster presentation (Dec 4)
- Final project report due Dec 9, 2025.

Course Schedule at a Glance

Dates	Lectures / Homeworks	Projects
9/22 - 10/ 6	Introduction; Autonomous Research Agents (HW1) Task-oriented agents (HW2) Grounding on free text	Research Project Ideas
10/ 8 - 10/20		Student-initiated ideas Project discussions Project proposals (2)
10/22 - 11/3	Knowledge retrieval: databases, knowledge graphs, hybrid data, long documents	Weekly meetings with mentor
11/ 5 - 11/12	Reasoning with computational thinking; Formal semantics; NLP building blocks	Weekly meetings with mentor
11/17 – 11/19	Misc: Multimodal apps; Training LLMs	Weekly meetings with mentor
11/25 - 11/27	*Thanksgiving*	
12/3		Final project posters (3:00-5:40)

This Course

	Grade
Participation	15%
Assignment	25%
Final Project	60%

Participation includes

- Class attendance and participation
- Ed discussion
- Meetings with project mentors