

# Triton AI Platform & Triton GPT

Strategy, use cases, technical architecture, collaborations

# Contents

- 1 Strategy
- 2 Use cases
- 3 Architecture
- 4 Collaborations

# UC San Diego AI Strategy

1. Leverage SDSC AI infrastructure and capabilities
  - Hardware platform, operating knowledge, design patterns
2. Focus on 'Vertical AI' instead of 'Horizontal AI'
  - Point GenAI at institutional content, public and proprietary
3. Implement more 'Small AI' and less 'Big AI'
  - Leverage open-source models and software
4. Democratize AI
  - Make it affordable, assist novices

Can we all take an AI breath now? <https://www.linkedin.com/pulse/can-we-all-take-ai-breath-now-vince-kellen-ph-d--chedc/>

Horizontal and Vertical AI: <https://www.linkedin.com/pulse/future-education-horizontal-vertical-ai-knowledge-vince-kellen-ph-d-/>

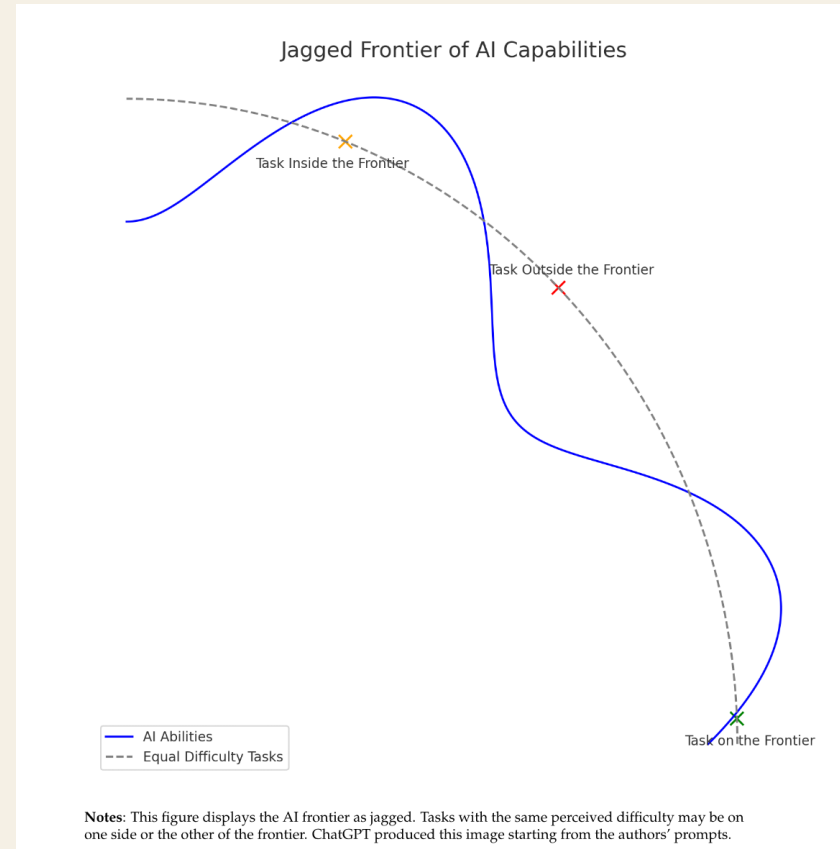
# AI and the ‘Jagged Frontier’

Tasks “outside the frontier” are tasks the researchers designed that would create wrong GPT responses. These are more complex, interpretive tasks. **The consultants who blindly adopted the GenAI output did worse. Those who used critical thinking skills do better.**

This means two things:

1. Selecting tasks that we think fit inside the capabilities of the AI
2. Training of staff users that slavishly copying from the AI especially on complex tasks without spending mental effort to think through it could result in worse results

Navigating the Jagged Technological Frontier: Field Experimental Evidence of the Effects of AI on Knowledge Worker Productivity and Quality. [Dell’Acqua, et al 2023](#)



# AI and Business Strategy

## Horizontal AI

Can be purchased and configured to be applicable to many organizations in one or more industries. Examples include ChatGPT, GitHub CoPilot, KhanAmigo Tutor

This AI makes available knowledge to many organizations in a replicable end-to-end solution

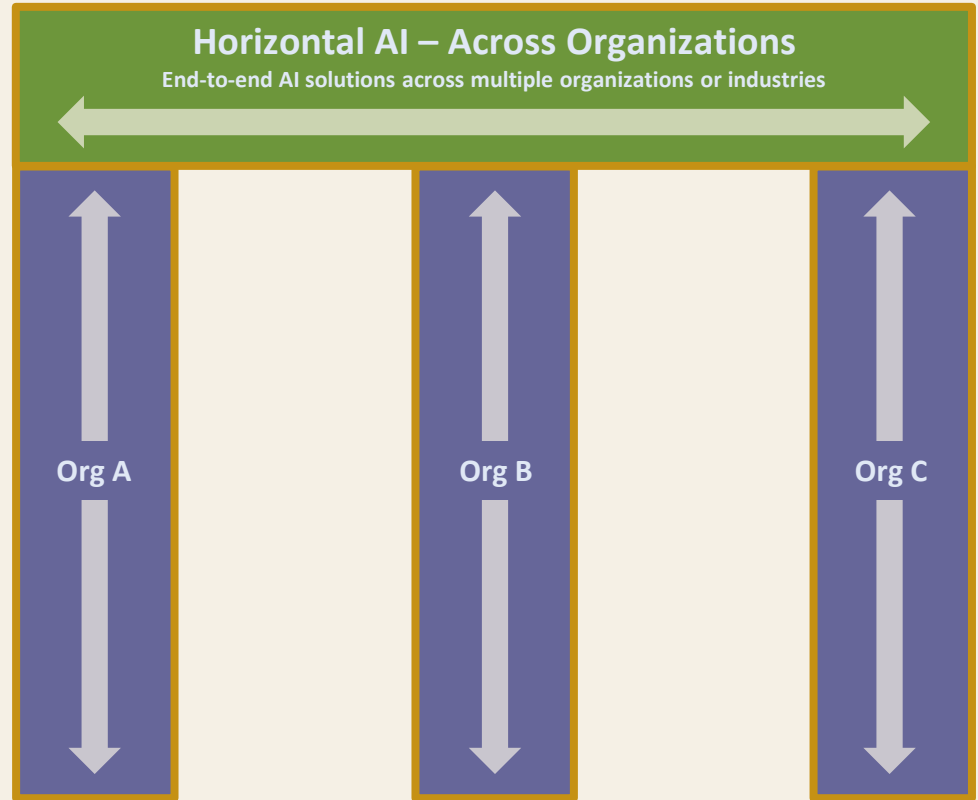
Adopting horizontal end-to-end AI is like adopting Microsoft Office. All organizations benefit similarly with little strategic advantage

## Vertical AI

Must be configured specific to an organization to be to be useful. Examples: ChatGPT that uses proprietary data unique to the organization for embedding or fine tuning in a single solution

This AI makes available knowledge to employees inside an organization and is not replicable to other organizations

Vertical end-to-end AI solutions use knowledge within an organization. Each organization benefits from its own unique combinations of information, systems, processes, people, culture, enhancing organizational differences, enhancing strategic advantage



# Where and AI/Gen AI help?

1. Generation/Concision/Classification (recruit words to words, images to/from words)
2. Fact retrieval
3. Reasoning

- **Summarization.** Generate a concise and easy-to-read summary of text content. E.g., Summarize a book or a transcribed lecture
- **Q&A.** Use natural language to ask and receive answers about general knowledge or institution-specific information. E.g., Answer IT, HR, finance or other policy questions via a chatbot
- **Content transformation.** Create new content and transform existing content into new formats, styles, perform language translations. E.g., Write programming code, put old content into a new format or structure
- **Classification.** Assign one or more labels to any input documents, images or text. E.g., Assign contract submissions a category of high, medium or low risk.
- **Text generation.** Create new content based on input examples or direction. E.g., Create a new web page, proposal, job description, etc.
- **Image generation.** Create static images or videos from textual prompts and direction. E.g., Create a website or presentation image
- **Reasoning.** Create logical analysis, explanations of relations, inference, in tandem with a knowledge base and prompting methods. E.g., Impact of policy on organization operations, tutoring students on math reasoning
- **Extraction.** Pull out of or highlight specific concepts, information, terms from input documents, text, images. E.g., find customer or other numbers from a scanned purchase order or invoice, highlight areas in a contract that relate to leasing terms

[Deloitte higher education webinar, October 18, 2023](#)

# How did we get here?

- Over one year ago, ITS started doing skunkworks POCs involving dozens of ITS staff; CIO held “Road Ahead” small-group sessions with all ITS staff, sharing the direction change
- ITS cultivated an internal community of practice and expanded it to include members of the community
- Initial hardware orders were placed in summer 2023. Governance and incentive fund established Q4 2023
- Onboarded ~400 staff in Q4 2023
- Expanding to all staff and faculty (~20,000) this spring

# Incentives, initial ideas

- Established a \$2 million one-time fund to support academic and administrative AI initiatives
- Implemented on on-premise, open-source, free to departments, centrally funded GenAI platform partnering with our supercomputer center, SDSC
- Various academic departments piloted several teaching and learning GenAI tools (e.g. OpenAI, etc.), but coordinating with the larger AI initiative

<b>Administration ideas</b>	<b>Status</b>
TritonGPT	V1
JDHelper	V1
Fund Manager Coach	Testing
International Students Work Study	Waiting
Categorize GASB 96 contracts	POC
Risk-Based Proposal Review	POC
SQL access to Activity Hubs, Fund Mgr V2	Pilot
Contract doc to lease agreements in Tririga	
Review Lease Contracts	
Staff Equity review	
Staff/Faculty onboarding	
<b>Teaching and learning ideas</b>	<b>Status</b>
Assessment, student support pre-calc classes	Pilot
TA support for Stats, Math, Programming classes	Pilot
Nano-engineering textbook augmentation	Pilot
Lecture video summary and notes	POC



# AI Training & Support offered by OSI

- ❖ Administrative transformation using AI tools
- ❖ Scope includes change management, process standardization

- ❖ AI + Continuous Improvement/LSS
- ❖ In-person/live training 4hr in 1 day

- ❖ Provides overview of AI, AI tools, and how to use AI safely

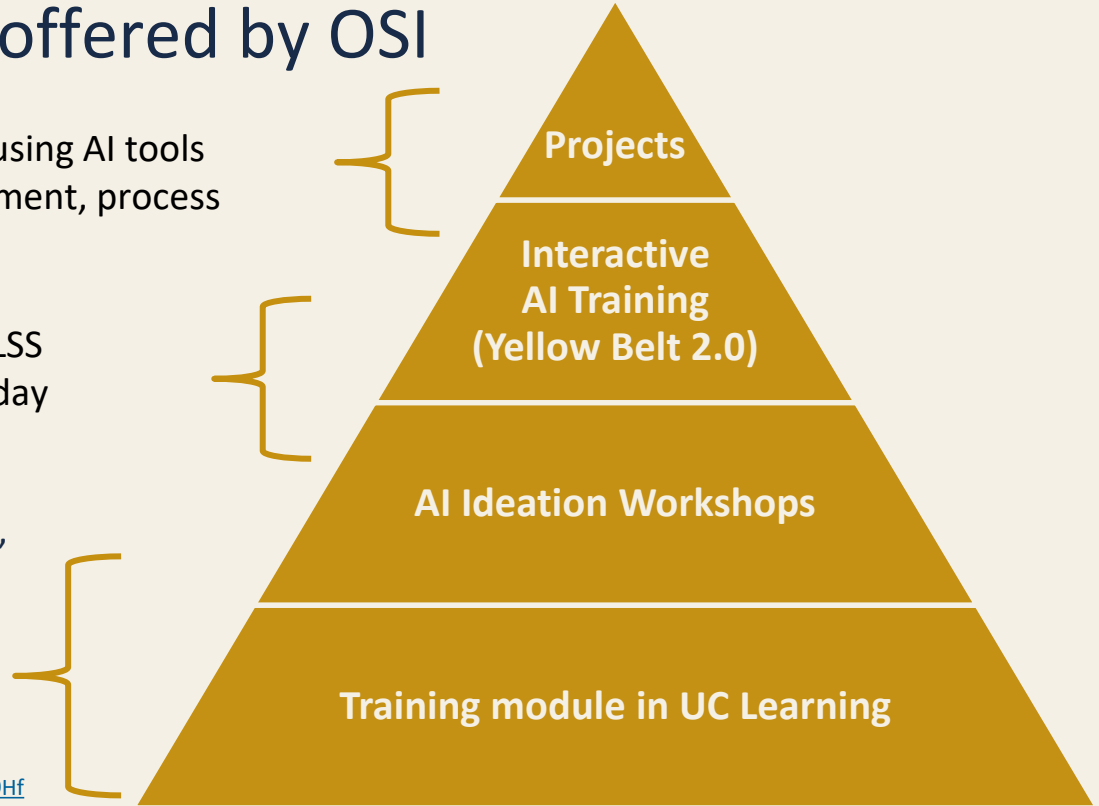
- ❖ Free for UCSD ~30mins

**Course Title:** *AI Essentials at UCSD*

**Duration:** Approximately 20-40 minutes

**Access:** UCSD Staff

**Get Started within UCLC:** <https://go.ucsd.edu/3FvH9Hf>



OSI offers AI training/in-depth process review services to fully leverage the power of AI tools

# First two administrative assistants in production

AI Assistant	
Overview	Expands beyond UC San Diego's scope by accommodating larger information exchanges. AI Assistant interacts with Large Language Models for tasks like document summarization, idea generation, and creating various content such as emails and reports.
Trained on	Uses Llama2 model pretrained on a massive corpus of datasets worldwide.
Benefits	Reduces staff time and improves communication workflow. Provides a secure on-premise alternative to public LLMs.

AI Assistant gives users access to 'vanilla' Llama2. No UC San Diego content is included. This is desirable for some use cases

Job Description Helper	
Overview	Engineered to streamline the job description creation process for hiring managers, JD Helper uses a pre-defined flow that engages hiring managers in a dialogue, capturing the specific requirements of the job.
LLM and sources	Uses Llama2 to process over 1,300 career tracks job standard templates and supporting HR documentation.
Benefits	Reduces staff time and administrative overhead. Accelerates time-to-hire.

JD Helper prioritizes curated UC San Diego content

# Soon to be released assistant – grant accounting V1

Fund Manager Coach	
Overview	Recognizing the crucial role of Fund Managers in overseeing grants and managing departmental finances, this focused AI coach enhances their understanding of UC San Diego's financial policies and procedures.
LLM and sources	Uses Llama2 to process documentation on how to develop budgets for research proposals, federal grant guidelines, reviewing and approving financial transactions, and managing payroll and expenses.
Benefits	Reduces staff time and administrative overhead. Accelerates employee onboarding.

V1 prioritizes curated UC San Diego content specific to fund management

V2 features include Finance Activity Hub data retrieval, including visualization via natural language with Wolfram Alpha. V3 features include advanced analysis of data via natural language with Wolfram Alpha, Mathematica

AI Study Aide	
Overview	Leverages UC San Diego's lecture capture system to create class summaries, study guides, and new ways to interact with course content.
LLM and sources	Uses Llama2 to process recorded course lecture capture audio files (with faculty permission).
Benefits	Saves instructor and student time. Focuses on course learning objectives.

Accessibility Enhancer	
Overview	Recognizing that auto-captioning from enterprise video platforms is often subpar, Accessibility Enhancer improves lecture capture captioning and transcription accuracy for all students.
LLM and sources	Uses WhisperAI to process recorded course lecture capture audio files for transcription and captioning.
Benefits	Greatly improves word accuracy over traditional auto-captioning. Benefits all students. Reduces staff costs for campus accessibility centers.

Assessment Builder	
Overview	Allows instructors to create exam and quiz assessments tailored to their course content and materials. Numerous assessment types are supported, including knowledge checks, multiple choice tests, short answer exams, and written essay prompts.
LLM and sources	Uses Llama2 to process instructors' own course materials (lectures, syllabi, readings, etc.).
Benefits	Saves instructor time. Creates assessments that directly measure comprehension of course content. Helps create learning objectives for new courses.

# Other assistants

POCs have determined feasibility.  
Timing in production TBD.

# Larger list of popular questions

Curated from current TritonGPT users and workshops

## Campus Services and Amenities

- What are some places to eat on campus?
- Which restaurants serve breakfast?
- What are the hours for <restaurant name>?
- What are the hours for Geisel Library?
- How do I book a study space?
- How do I navigate the campus and find information about parking, transportation, and dining options?
- Can I get a daily parking permit?
- What are some upcoming events?
- What are some events happening on campus this week for staff?
- When will UC San Diego be closed for the holidays?
- What fitness facilities are available on campus, and how do I access them?
- Are there child care services available for students and staff, and how do I apply?
- How do I rent equipment or reserve spaces for events on campus?
- What transportation services are available for getting around campus and the local area?
- Where can I find a campus map with accessible routes and facilities?
- How can I get involved in campus recreational sports or join a fitness class?

## HR and Professional Development

- What are some ways I can get a raise?
- I am a full-time career staff member working in IT Services. What resources are available to support my professional development and career growth at UC San Diego?
- What type of discounts are available for staff?
- What are the policies and procedures for requesting time off, such as vacation, sick leave, or family medical leave?
- What are the available resources for addressing work-related concerns or personal issues that may affect my job performance?
- How do I access employee wellness programs and mental health resources to support my overall well-being?
- What are the benefits offered to UC San Diego staff members, such as health insurance, retirement plans, and paid time off?

# Larger list of popular questions

Curated from current TritonGPT users and workshops

## **Policies and Procedures**

- Explain our travel policy in plain language.
- Can I expense alcohol per the UCSD Travel policy?
- What are the UCSD Holidays?
- What are the guidelines for remote work or telecommuting?
- What is the policy on conflict of interest for university employees?
- How do I file a grievance if I feel I've been treated unfairly at work?
- What are the procedures for taking a leave of absence?
- How does the university handle complaints of discrimination or harassment?
- What is the policy regarding intellectual property and research conducted at the university?
- How do I report a violation of the university's code of conduct?
- What are the guidelines for ethical conduct in research?
- What is the university's policy on data protection and privacy?
- How do I request access to university administrative records?
- What are the university's travel reimbursement procedures?
- How are conflicts resolved within academic departments?
- What are the requirements for mandatory training for university employees?

## **IT/Finance/Contracts**

- What is the process for requesting access to financial systems, and who should I contact for assistance?
- How do I determine the Oracle and Concur standard roles that departmental preparers need?
- What team supports Microsoft 365?
- How can I purchase a Tableau license?
- My computer is broken. Where can I take it to get fixed?
- How do I order a new computer?
- How do I log into Concur?
- What is the difference between AD and Business Systems?
- How can I update payroll distributions to reflect actual effort expended on a sponsored project?
- What are the consequences if I do not certify my effort report on time?
- How can I ensure that my effort report is certified within the 120-day deadline?
- How can I access the ECERT system to manage certifications and effort reports?

# TritonGPT – UC San Diego's AI Assistant

**TritonGPT**  
Powered by Danswer

+ New Chat

List Helpers

- Requirements Tracea...
- Chat Session Name: "I..."
- Director-Level Position
- TritonGPT Explained
- Purchasing Card Rene...
- Business Systems An...
- Sure, here's a short a...

maamini@ucsd.edu

UC San Diego

TritonGPT

Show me what you can do  
Get an idea of what TritonGPT can do

Explore Topics  
Get answers to questions

Job Description Helper  
Get help writing job descriptions

Fund Manager Coach  
(coming soon)

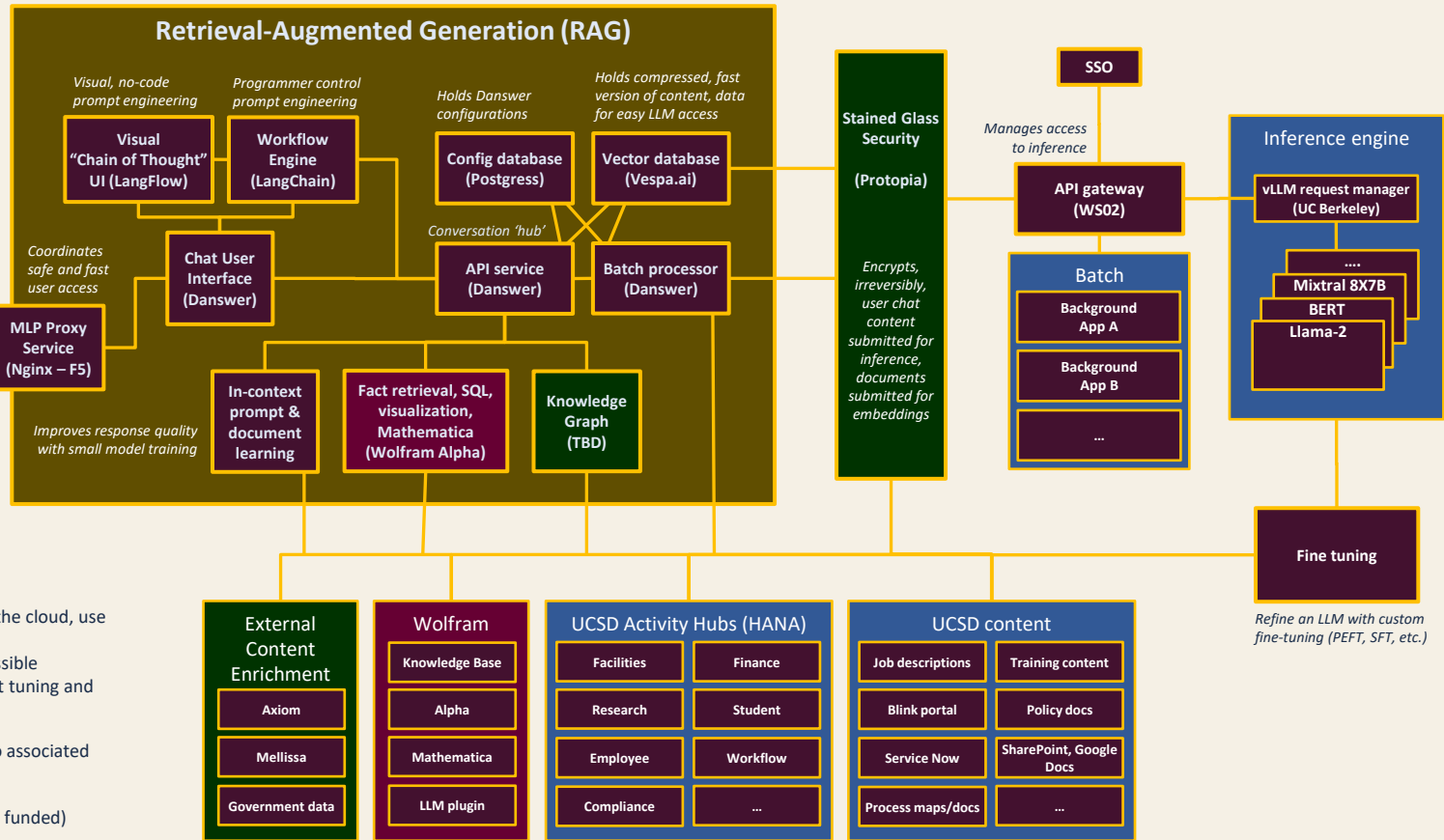
Ask me anything...

Find helpful hints on using TritonGPT as well as the [Terms of Use](#)

[TritonGPT](#) is an artificial intelligence tool designed to assist UC San Diego staff and students by providing information and generating human-like text. It can be used to answer questions, summarize documents, and even help with tasks such as writing emails or creating content. It knows about our university services, policies, procedures, and initiatives. It serves as a powerful foundation upon which task-specific assistants can be built, such as the Job Description Helper and Fund Manager Coach.

- [TritonGPT](#) available on a limited pilot release
- The "AI Essentials at UCSD" Training <https://go.ucsd.edu/3FvH9Hf> is a precursor to getting access.
- Workgroup members will be given priority.
- Initial users of TritonGPT will be expected to provide feedback to continually refine its accuracy and effectiveness.

# Triton AI Platform Architecture



## Key features:

- On-premise (inference can burst to the cloud, use cloud AI services if needed)
- Open-source software wherever possible
- Prompt-tuning, in-context document tuning and fine-tuning capabilities

Most components involve UC San Diego associated start-ups and UC Berkeley software:

- Danswer (UCSD student start-up, VC funded)
- Protopia (UCSD faculty startup)
- vLLM (UCB open-source project)
- Hardware is based on SDSC designs using NVIDIA H100s

  Spring 2024

  Summer 2024 – Winter 2025



# Danswer: Multi-step retrieval with semantic prompt expansion and reference management

- Contextual prompt expansion based on retrieved documents
- Reference retrieval for quality and relevance verification
- UI for directed "conversations" with specific context document collections

The screenshot displays the Danswer user interface. On the left, a chat window shows a user asking "What is Danswer?". The system response, labeled "Danswer", indicates it is searching for "Danswer definition". The answer text describes Danswer as an Open Source Enterprise Question Answering Tool, detailing its capabilities and licensing. Below the text, a "Sources:" section lists a reference: "[1] Introduction - Danswer Documentation". On the right side of the interface, a "Retrieved Documents" panel lists three relevant documents: "Introduction - Danswer Docume...", "Enterprise Setup - Danswer Doc...", and "System Overview - Danswer Doc...". Each document entry includes a relevance score (e.g., 0.88, 0.82, 0.81) and a "Select" button.

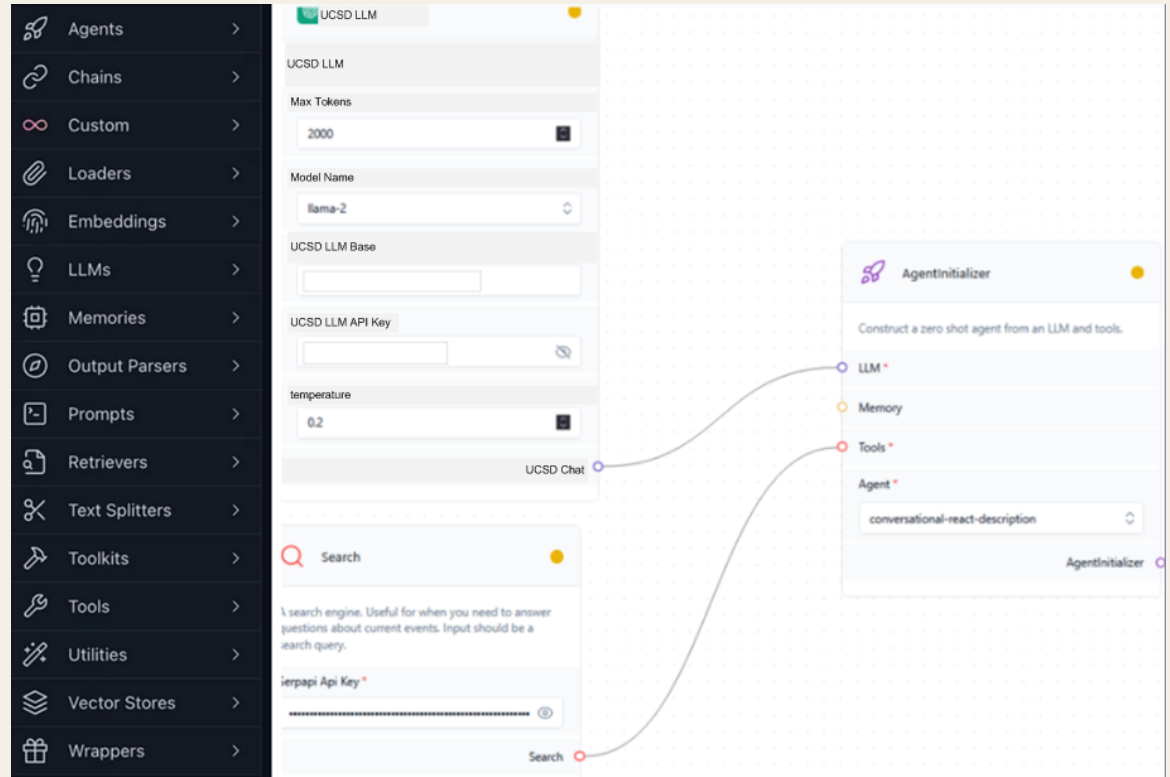
# LangChain

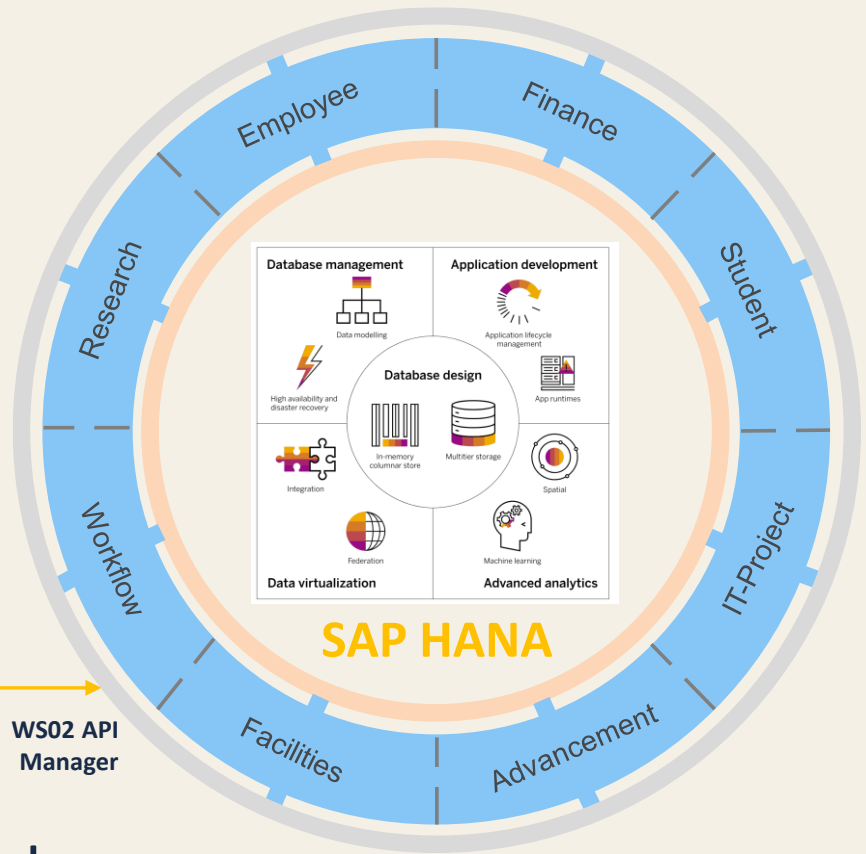
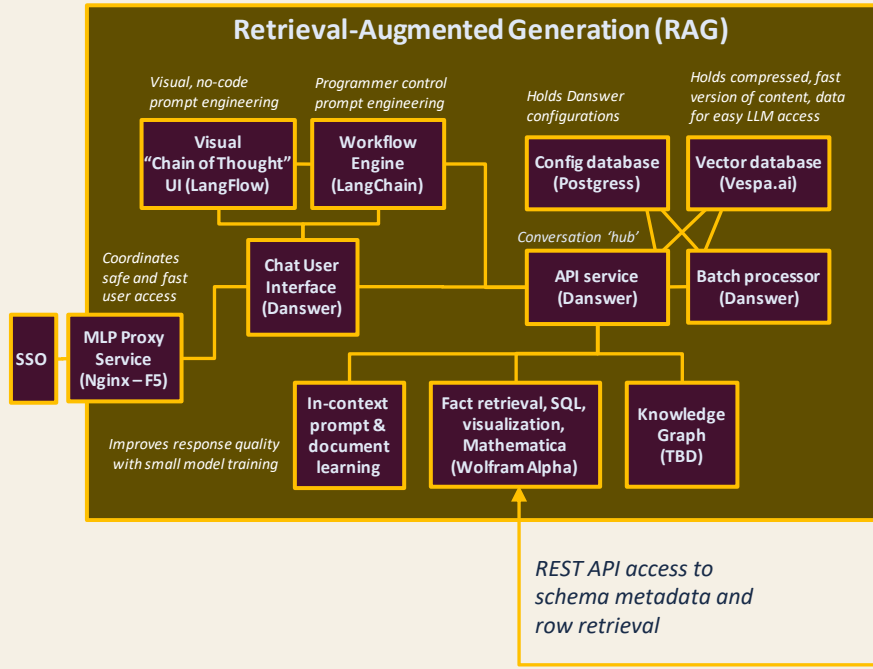
Gen AI orchestration, libraries and templates for various tasks, creating APIs,

Tools	Toolkits
<p>Resources are utilized by the agent to accomplish tasks eg. LLMs or APIS</p> <pre>Tool(   name="Wikipedia",   func=wiki.run,   description="useful when you need an answer about encyclopedic general knowledge" )</pre>	<p>Groups of tools assembled for specific functions e.g. toolkits for question answering, text generation, and natural language processing</p> <pre>const db = await SQLiteDatabase.fromDataSourceParams({   appDataSource: datasource, }); const model = new OpenAI({ temperature: 0 }); const toolkit = new SqlToolkit(db, model); const executor = createSqlAgent(model, toolkit);  const input = `How many students enrolled in Spring semester courses`;  console.log(`Executing with input "\${input}"...`);</pre>

# LangFlow:

A GUI tool for developing agents, vector stores and API tools

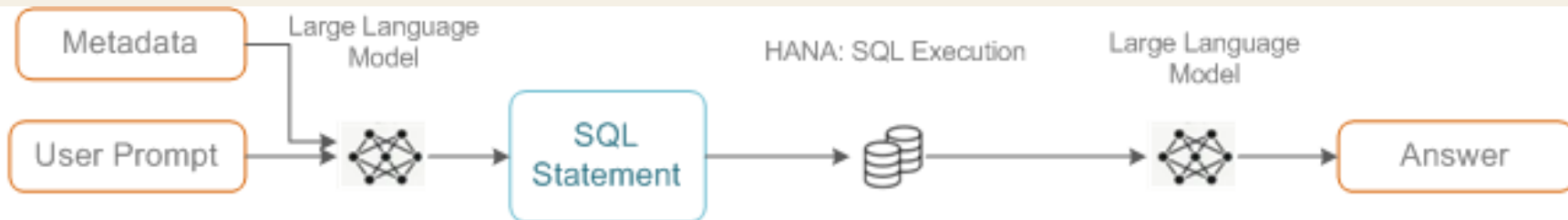




# Wolfram, TritonGPT & Activity Hubs

# Reasoning with SQL and structured data

Wolfram Alpha + SAP HANA Activity Hubs with 8 billion rows of data

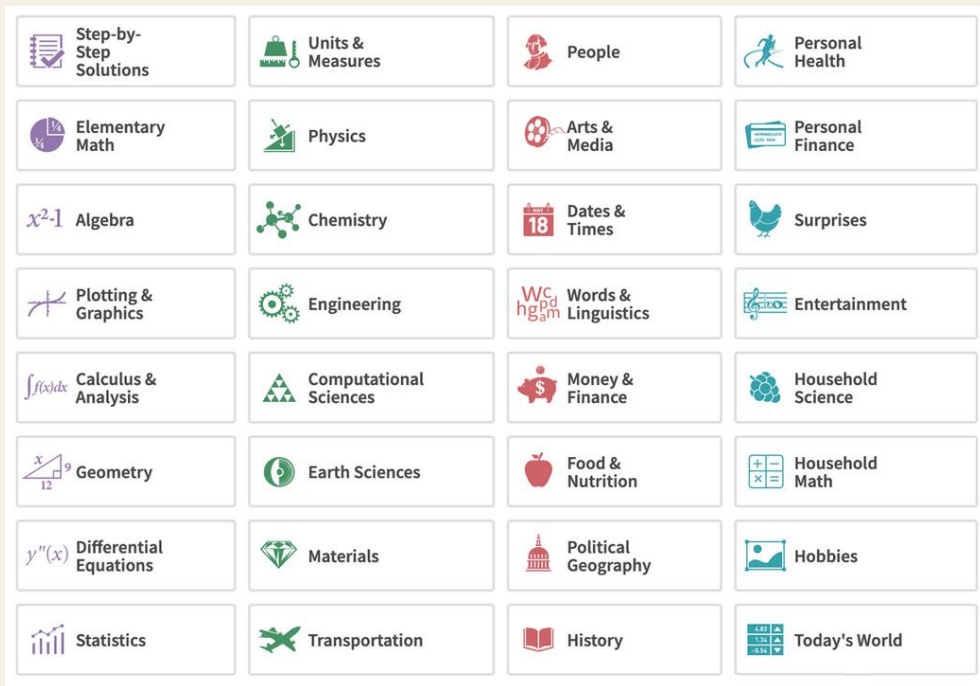


Write an SQL statement that answers the following question:  
Look into: UCSD\_BALANCES\_Sample  
What is the total debit balance?"

```
SELECT SUM(CAST('Fiscal Period Net Debit Balance' AS INT))  
AS TotalDebitBalance  
FROM UCSD_BALANCES_Sample;
```

# Wolfram Alpha, Mathematica

Tool chains for reasoning with curated knowledge bases



S Find the alkali metal with the highest resistivity

Used Wolfram

Among the alkali metals, cesium (Cs) has the highest resistivity. Here is the Wolfram Language code I used to find this result:

```
wolfram Copy code  
alkaliMetals = {Entity["Element", "Lithium"], Entity["Element", "Sodium"], Entity["Element", "Potassium"], Entity["Element", "Rubidium"], Entity["Element", "Francium"]};  
resistivities = EntityValue[alkaliMetals, "Resistivity"];  
maxResistivity = Max[resistivities];  
position = Position[resistivities, maxResistivity];  
alkaliMetals[[position[[1, 1]]]]
```

This code defines a list of alkali metal entities (excluding francium due to its instability and lack of data), retrieves their resistivity values, finds the maximum resistivity, and then identifies the alkali metal with the highest resistivity.

# Additional platform development

- From RAG to RAGE - RAG with Execution: workflow triggering (Kuali Build), enterprise application process and data integration, batch processing scheduling and execution
- Fully automated, customizable and continuous quality assurance testing
- Ability to leverage prompt engineering across multiple LLMs
- Supporting instructional faculty use cases
- Adding cloud bursting and/or cloud high availability
- Experimentation with different chip designs, full-stack LLMs

# Costs and collaboration

- Our engineering target is 20%-35% of commercial offerings. This ongoing cost includes hardware, power, facility rental cost and ongoing support costs. This does not include one-time project and start-up costs
- We have received interest from other educational institutions (higher education and K12) in leasing the platform and we are in early discussions
- We are actively looking for collaborators who want to help in using and developing lower-cost, strategic, generative AI platforms that can be afforded by everyone

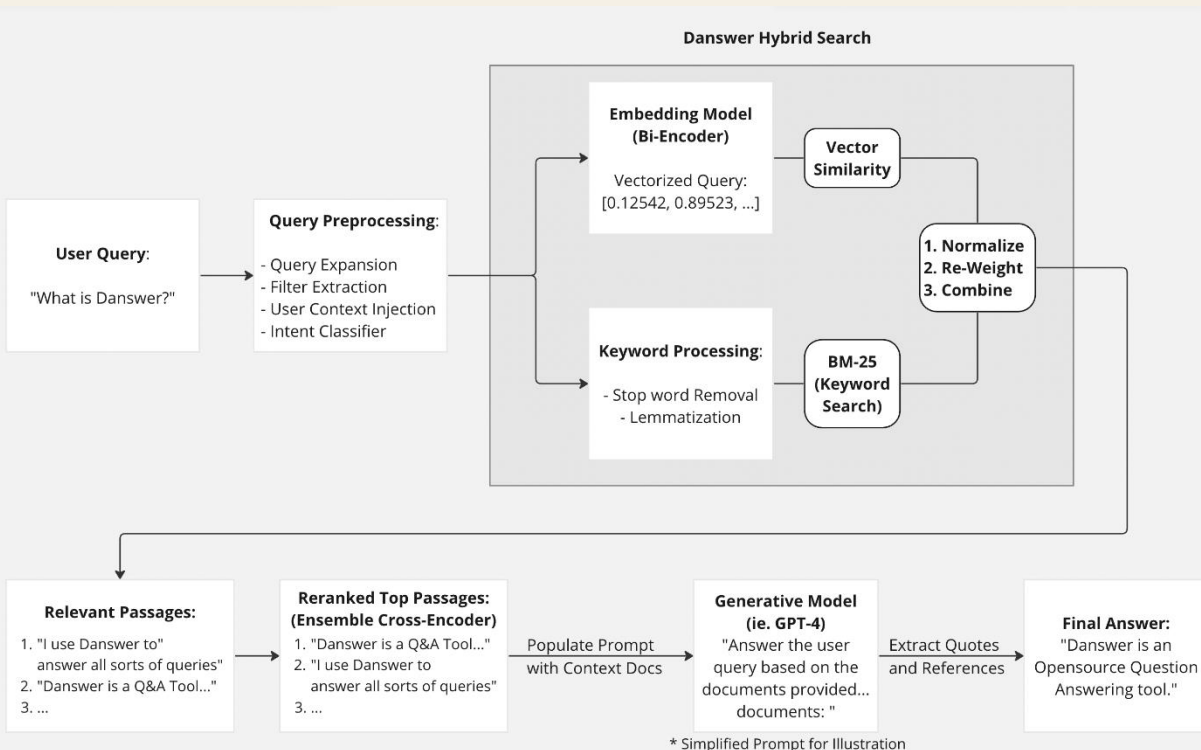


# Whew!

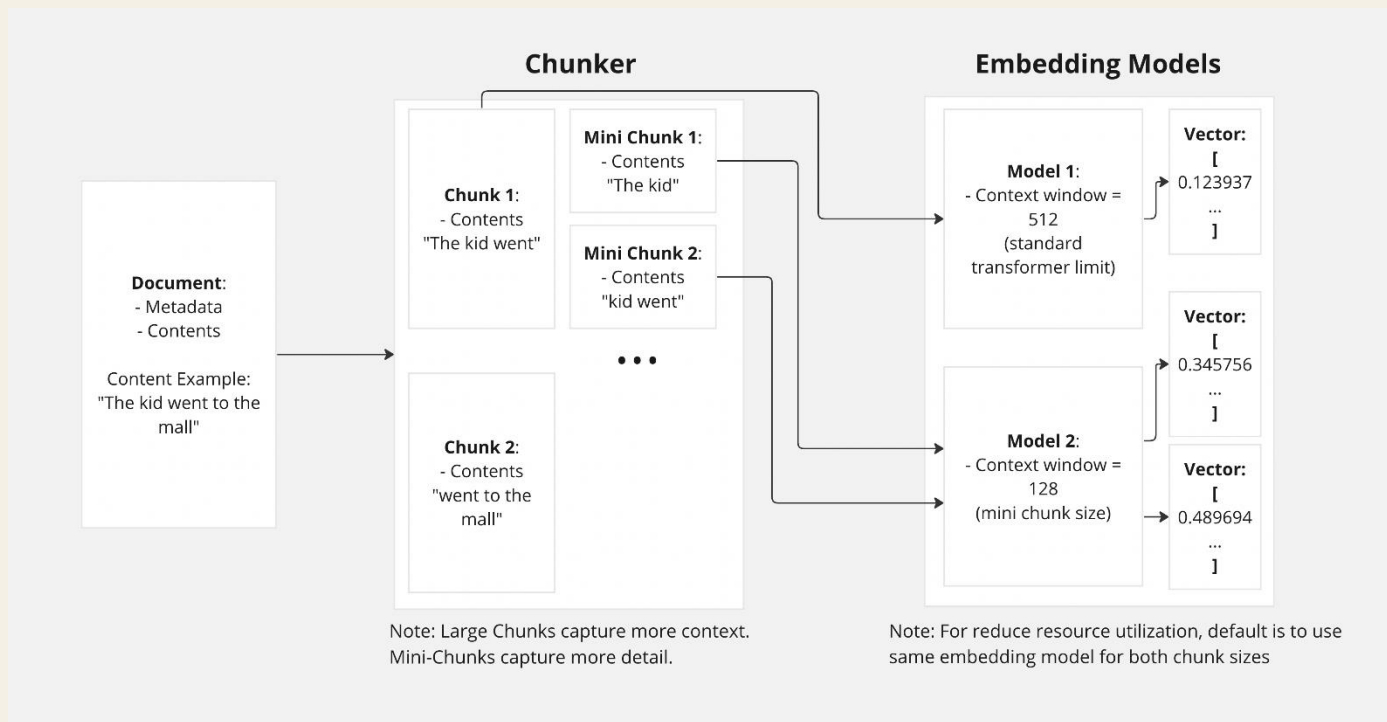
If you want to learn more, we are happy to share what we have learned. We have Zoom webinars ongoing and can invite you to them!

Just email [bpollak@ucsd.edu](mailto:bpollak@ucsd.edu) or [vkellen@ucsd.edu](mailto:vkellen@ucsd.edu)!

# Danswer query flow



# Danswer embedding flow



# Tasks retrieving curated knowledge using LLMs

## Wolfram Alpha

Altair (star) | Betelgeuse (star)

Properties More Show history

	Altair	Betelgeuse
distance from Earth	16.77 ly	548 ly
apparent magnitude	+ 0.76 <small>(easily visible to the naked eye, variable)</small>	+ 0.45 <small>(easily visible to the naked eye, variable)</small>
absolute magnitude	+ 2.2 <small>(visual)</small>	- 6.02 <small>(visual)</small>
spectral class	A7V <span style="color: blue;">■</span> <small>(main sequence)</small>	M1.5Iab <span style="color: orange;">■</span> <small>(supergiant)</small>
effective temperature	7800 K	3600 K
mass	$3.6 \times 10^{30}$ kg 1.8 $M_{\odot}$	$3.6 \times 10^{31}$ kg 18 $M_{\odot}$
main sequence lifetime	2.2 billion yr	7.3 million yr
end state	carbon-oxygen white dwarf	neutron star or black hole
Bayer name	$\alpha$ Aquilae (Alpha Aquilae)	$\alpha$ Orionis (Alpha Orionis)

Write an essay comparing altair and betelgeuse using data from wolfram

Used **Wolfram** ▼

Title: A Comparison of Altair and Betelgeuse

### Introduction

Altair and Betelgeuse are two well-known stars in the night sky, each with distinct characteristics. Altair is the brightest star in the constellation Aquila, while Betelgeuse is a prominent star in the constellation Orion. This essay will compare the properties of these two stars, including their distance from Earth, apparent and absolute magnitudes, spectral classes, effective temperatures, masses, main sequence lifetimes, end states, and variable star properties.

### Distance from Earth

Altair is much closer to Earth than Betelgeuse, with a distance of 16.77 light-years (ly) compared to Betelgeuse's 548 ly.

# Collaboration Approach

Service to the community. Help democratize access to information in education through Large Language Models

Be frugal and affordable. Commercial Generative AI products are expensive. Leverage our infrastructure investment at San Diego Computer Center

Use leading edge open-source technology or low-cost solutions wherever possible

Use partners! We know we can't do it alone, so we are partner-friendly

Use a very rigorous, disciplined, and clever engineering approach

Be transparent with everything, including pricing