

TritonGPT

Strategy, Architecture, Use cases, Roadmap, Collaborations

Discussion outline

- 1 Strategy
- 2 Architecture
- 3 Assistants, use cases
- 4 Roadmap
- 5 Collaborations

UC San Diego AI Strategy

- Focus on 'Vertical AI' instead of 'Horizontal AI'
- Point GenAI at UC San Diego content, public and proprietary
- Implement more 'Small AI' and less 'Big AI'
- Leverage open-source models and software
- Given market volatility, invest in options

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 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

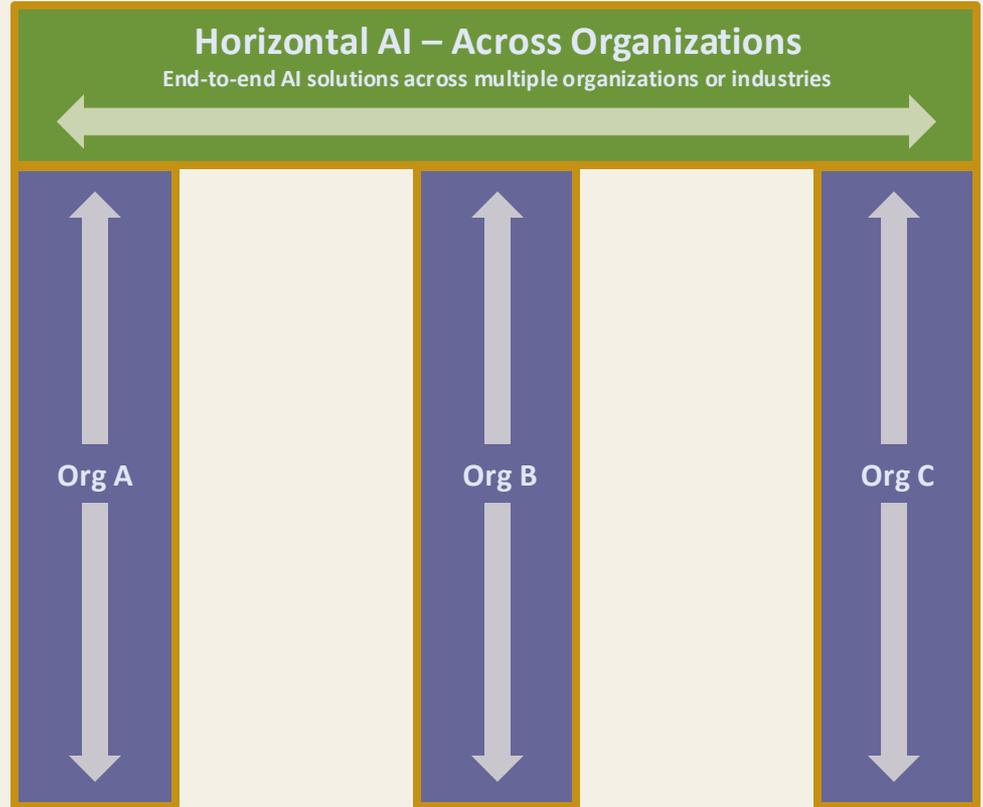
Adopting horizontal end-to-end AI is like adopting Microsoft Office. All organizations benefit similarly with little strategic advantage

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

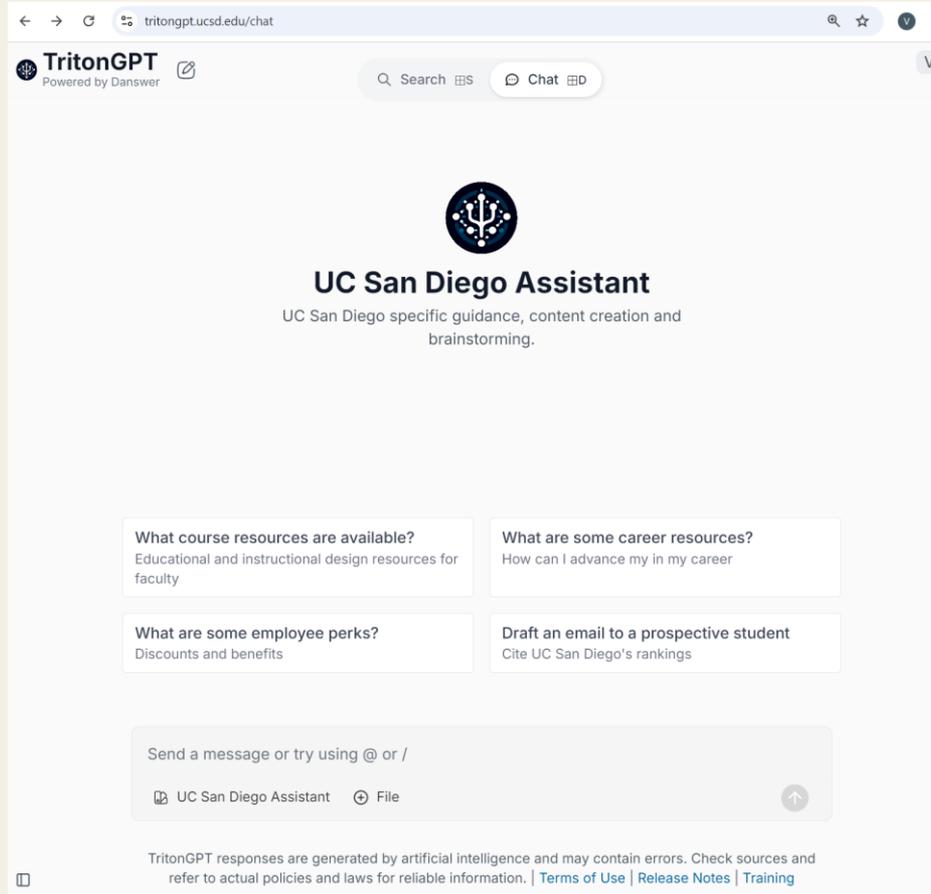
This AI makes available knowledge to employees inside an organization and is not replicable to other organizations

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



Discussion outline

- 1 Strategy
- 2 Architecture
- 3 Assistants, use cases
- 4 Roadmap
- 5 Collaborations

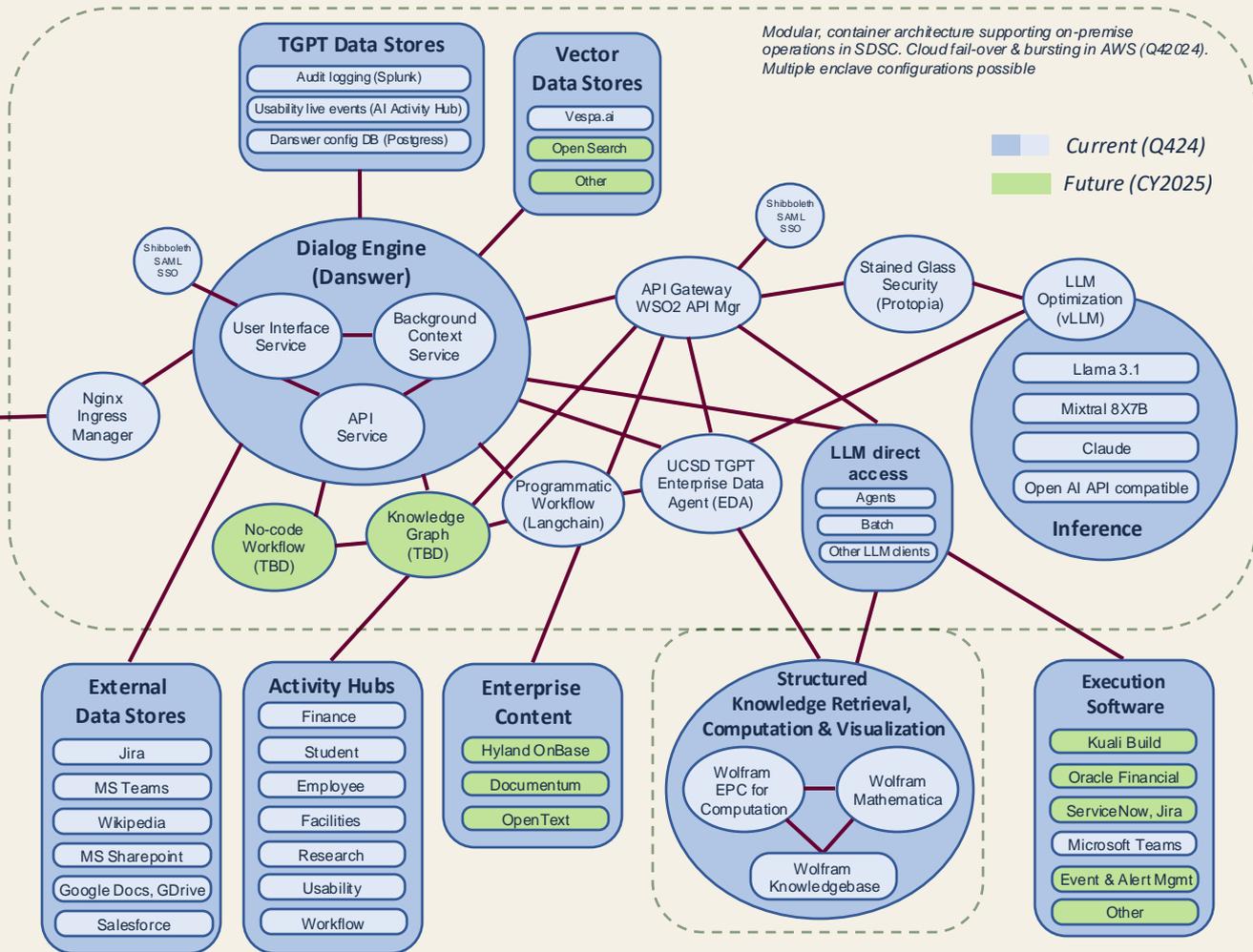
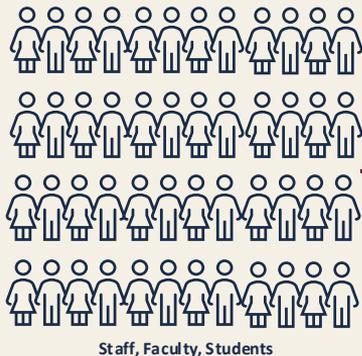


The screenshot shows a web browser window with the URL `tritongpt.ucsd.edu/chat`. The page header includes the TritonGPT logo (Powered by Danswer) and navigation buttons for Search and Chat. The main content area features the UC San Diego Assistant logo and the text: "UC San Diego Assistant" and "UC San Diego specific guidance, content creation and brainstorming." Below this, there are four suggested prompts in a grid:

- What course resources are available?**
Educational and instructional design resources for faculty
- What are some career resources?**
How can I advance my in my career
- What are some employee perks?**
Discounts and benefits
- Draft an email to a prospective student**
Cite UC San Diego's rankings

At the bottom, there is a text input field with the placeholder "Send a message or try using @ or /". Below the input field, there is a dropdown menu showing "UC San Diego Assistant" and "File". A small icon of a person is visible in the bottom left corner of the chat area. At the very bottom, a disclaimer reads: "TritonGPT responses are generated by artificial intelligence and may contain errors. Check sources and refer to actual policies and laws for reliable information. | [Terms of Use](#) | [Release Notes](#) | [Training](#)"

TritonGPT Architecture



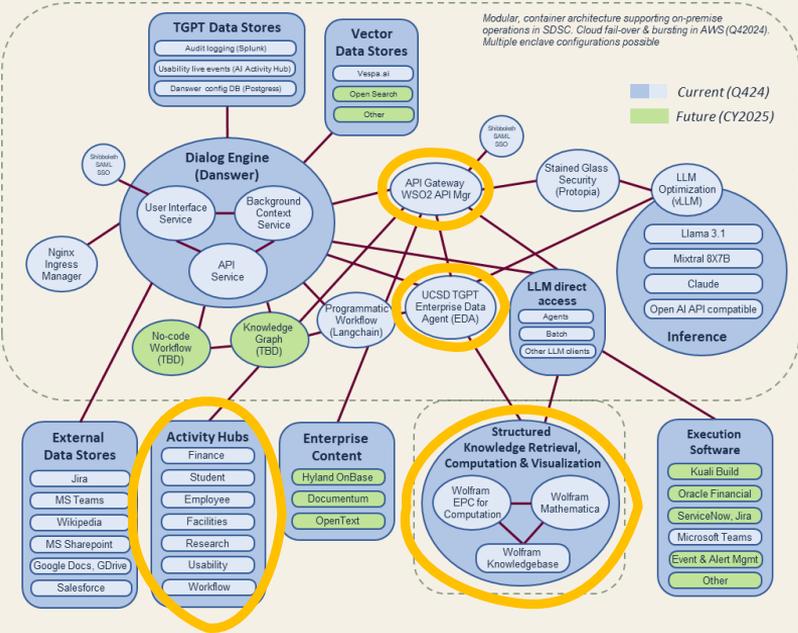
Key features

- Open-source software wherever possible
- Agentic approach
- Enhanced multi-level privacy, security, and IP protection
- Secure enterprise data access, SQL code generation
- Custom private cloud/network configurations supported

Most components involve University of California startups

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

Accessing relational databases (Activity Hubs)

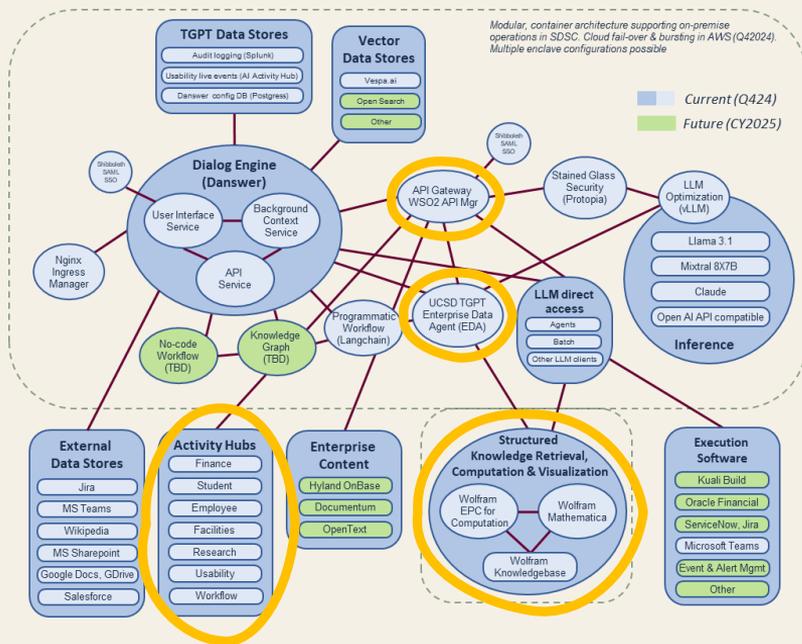


Activity hubs have 15,000+ columns surfaced in 550 views, with 14 billion rows of data

- TritonGPT SQL generation
 - Wolfram EPC product is used for access to the Wolfram visualization, analysis tools
 - No ODBC/JDBC access. Access is controlled through levels: WSO2 API manager and a single Activity Hub access point
- Enhanced metadata techniques
 - View and column descriptions elicited from staff experts, enriched with LLMs
 - Metadata describes column groupings
 - Column sampling adds to the metadata data concrete examples of the data
 - Sampling methods: unique values, random values, selected values
 - Metadata curation automation
 - User searching and replaying previously executed queries, analytic workflows
- Row and column context filtering
 - TGPT only sees rows and columns that the user can see normally
 - All columns have valid 'passport' and 'anti-passport' lists
 - Via AD groups, all TGPT users are assigned column passports
- Enhanced security, privacy protection methods in Activity Hubs
 - k-anonymity, l-diversity, and differential privacy
 - Privacy protection configurations can be added to any Activity Hub view
 - Multi-level row limits (TGPT, metadata and AH access point levels)
 - Narrow, not broad, access to activity hubs, views, and columns within views
 - SQL injection prevention, significant limits on SQL language use (DDL, update, insert, delete operations, multiple statements, referencing out-of-scope view/tables/columns, prohibition on joins, etc.)
 - Extensive logging and monitoring capabilities

Fact retrieval and Wolfram EPC

- TritonGPT generates the SQL query leveraging rich Activity Hub metadata, retrieves rows
- UC San Diego EDA can choose to use Wolfram EPC to further analyze and visualize data



Calculate the remaining funds on my project and project how long they will last based on my current spend rate. What are the projected financial needs for the next quarter based on current trends?

To calculate the remaining funds on your project and project how long they will last based on your current spend rate, we can use the following steps:

1. Calculate the remaining funds:

Total Amount on the Grant: \$1,200,000 Total Spent so far: \$645,235 Remaining Funds: \$1,200,000 - \$645,235 = \$554,765

2. Calculate the number of months the remaining funds will last:

Current Spend Rate: \$13,654 per month Remaining Funds: \$554,765 Number of months remaining funds will last: $\$554,765 / \$13,654 = 40.6$ months

Therefore, based on your current spend rate, the remaining funds on your project will last for approximately 40.6 months.

To project the financial needs for the next quarter based on current trends, we can use the following steps:

1. Calculate the total spend for the next quarter:

Current Spend Rate: \$13,654 per month Number of months in a quarter: 3 Total spend for the next quarter: $\$13,654 \times 3 = \$40,962$

2. Calculate the remaining funds after the next quarter:

Remaining Funds: \$554,765 Total spend for the next quarter: \$40,962 Remaining Funds after the next quarter: $\$554,765 - \$40,962 = \$513,803$

Therefore, based on current trends, the projected financial needs for the next quarter are \$40,962, and the remaining funds after the next quarter will be \$513,803.

Send a message or try using @ or /

🔗 Tool Calling Test 🗨️ File

remaining funds, \$40,962. The remaining funds after the next quarter: \$554,765 - \$40,962 = \$513,803

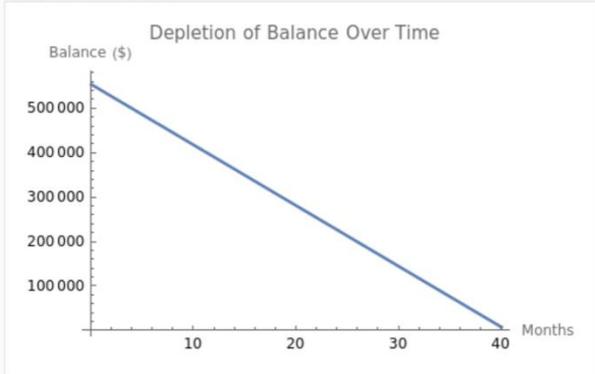
Therefore, based on current trends, the projected financial needs for the next quarter are \$40,962, and the remaining funds after the next quarter will be \$513,803.

It's important to note that this is just a projection based on current trends, and actual spend may vary. It's always a good idea to regularly review and update your budget to ensure that you are on track to meet your financial goals.

Develop a visual that shows the depletion of the balance over time.

< 3 / 3 >

Sure! Here's a simple visual representation of the depletion of the balance over time, based on the data provided:



Send a message or try using @ or /

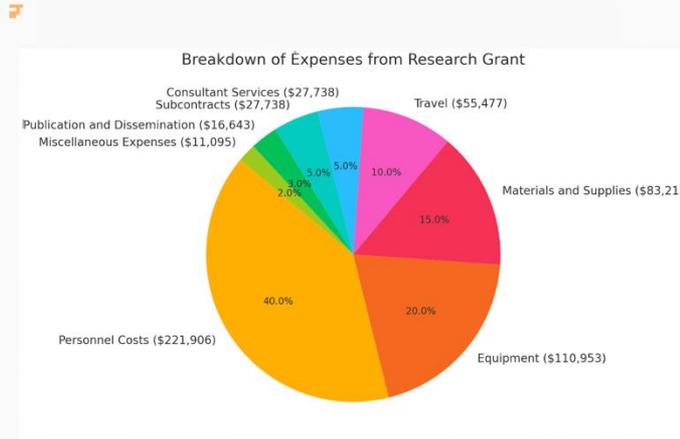
Tool Calling Test File

TritonGPT responses are generated by artificial intelligence and may contain errors. Check sources and refer to actual policies and laws for reliable information. | [Terms of Use](#) | [Release Notes](#) | [Training](#)

TritonGPT and Wolfram EPC

Q4 2024

create a visual based on the expenditure categories.

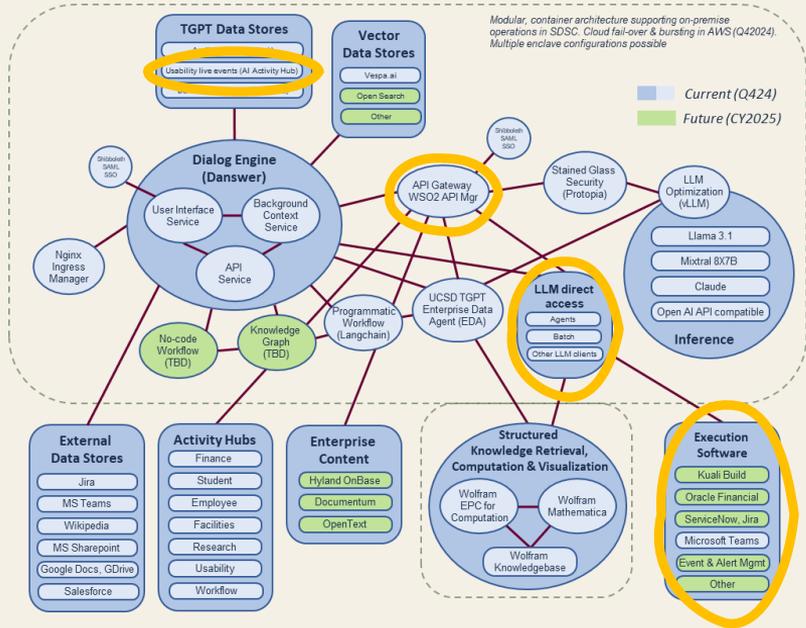


Send a message or try using @ or /

Tool Calling Test File

TritonGPT responses are generated by artificial intelligence and may contain errors. Check sources and refer to actual policies and laws for reliable information. | [Terms of Use](#) | [Release Notes](#) | [Training](#)

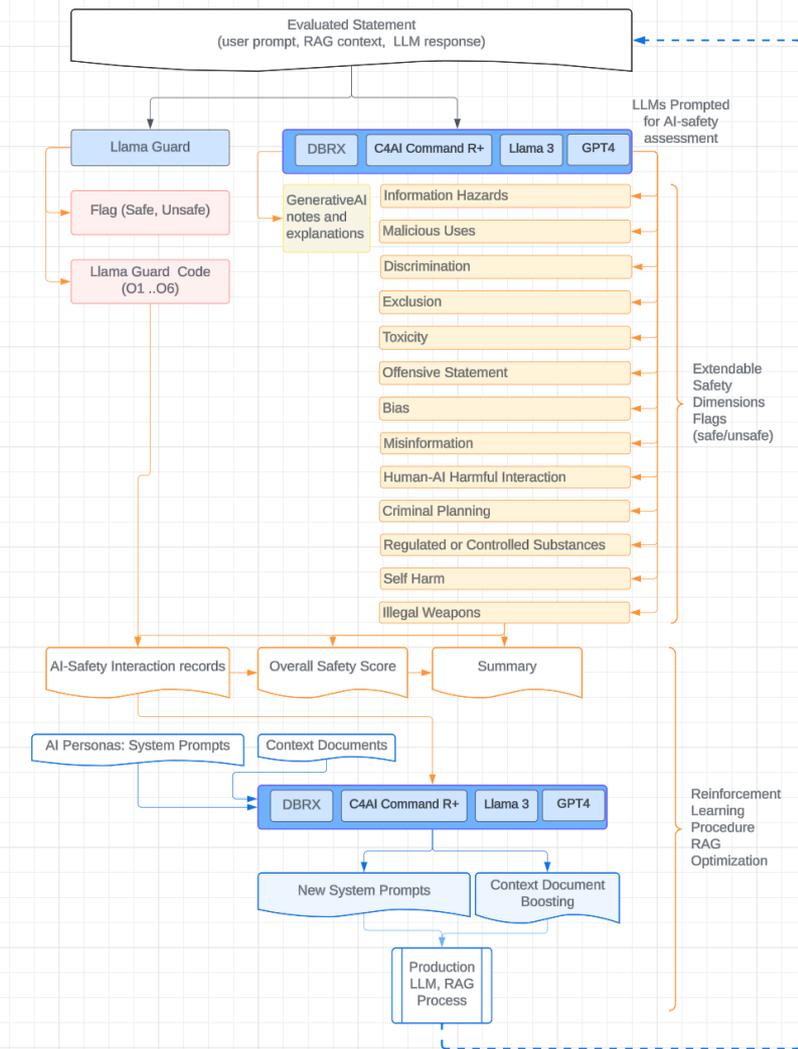
TritonGPT Execution Framework



- Legal document annotation process
 - Legal staff initiates a document uploading via Teams
 - Using Kafka Airflow to integrate Teams “dialog” with TGPT LLMs
 - Output includes annotations, templated language inclusion
- Roadmap execution features
 - Bi-directional Kuali Build integration
 - Start a workflow in Kuali Build, reach out to TGPT for augmentation
 - Start a dialog in TGPT, reach out to Kuali Build to trigger a workflow
 - Similar design patterns for Jira, ServiceNow, etc.
 - Automation of metadata curation for Enterprise Data Agent
 - Elicit knowledge from experts via Zoom session
 - Trigger workflow to automatically enrich Zoom content with LLM
 - Trigger workflow to elicit peer-review and comments
 - Final review by the BI and AI team
 - A form of reinforcement learning from human feedback
- Usability Live Event logging (real-time)
 - Captures usage patterns with the TGPT user interface in the Usability Activity Hub
 - Helpful for analyzing and guiding user interface improvements, adoption planning
 - Can be used for identification of adverse usage (e.g., cybersecurity risk) and general alerting

TritonGPT safety

- We use an ensemble of experts approach to evaluate LLM safety
 - DBRX
 - C4AI Command R+
 - Llama 3
 - GPT4
 - Llama Guard, Llama Guard Code
- Each model votes in each safety dimension based on how it evaluates the input document and the safety area description
- An overall safety score is collected
- Automated daily execution of this safety evaluation is controlled via Kafka AirFlow
- Several safety actions are possible
 - Identify users performing unsafe prompting
 - Identify unsafe context documents and uploads
 - Identify unsafe user feedback, etc.
 - Guide data ingestion, RAG development
 - Guide overall TGPT design



Discussion outline

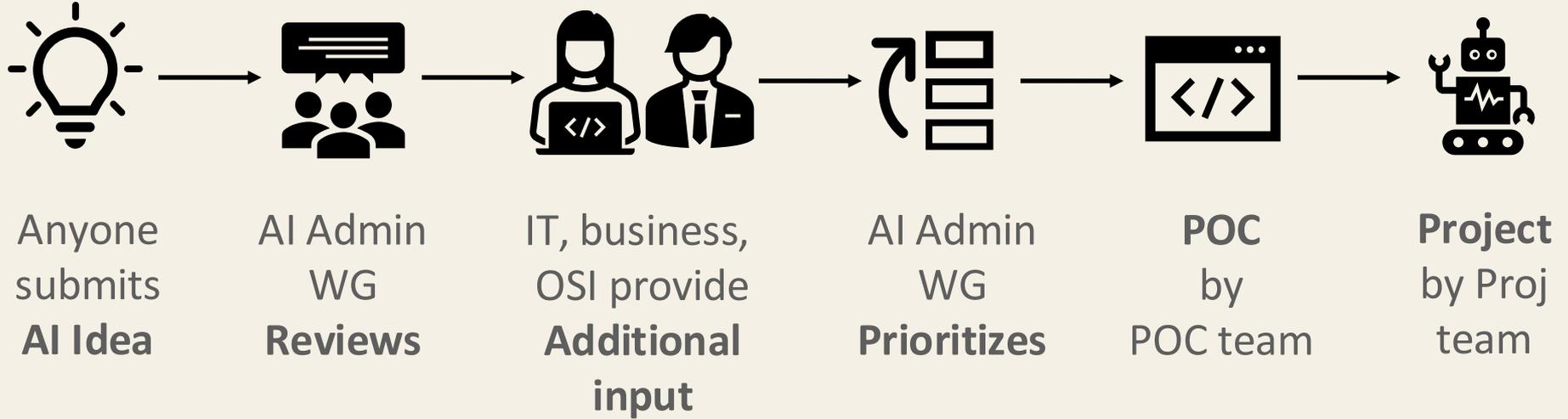
- 1 Strategy
- 2 Architecture
- 3 Assistants, use cases
- 4 Roadmap
- 5 Collaborations

TritonGPT Task-Specific AI Assistants

Other assistants in development or testing: Policy Assistant, Data Dictionary Assistant, Suspicious Email Reviewer, and Coding Assistant

 UC San Diego Assistant 	Live. Centralizes UC San Diego policies, processes, and help documentation, providing direct answers to your questions. Also ideal for generating new content and brainstorming ideas with UC San Diego context.
 Job Description Helper 	Live. Streamlines job description creation for hiring managers using 1,300+ career track templates. Engages managers in a dialogue to capture job requirements, then crafts compliant and tailored job descriptions, reducing drafting time and effort.
 General AI Assistant 	Live. Expands beyond UC San Diego, facilitating larger information exchanges. Interacts with a Large Language Model for tasks like document summarization, idea generation, and content creation (e.g., emails, reports).
 Fund Manager Coach 	Live. Enhances understanding of UC San Diego's financial policies and procedures for Fund Managers. Trained on documentation for research proposal budgets, contract and grant guidelines, financial transactions, payroll management, and compliance.
 Contract Reviewer	Proof of Concept. Review, highlight, summarize and accelerate contract and legal document reviews based on playbook of rules.
 Fact Retrieval	Proof of Concept. An internally developed Enterprise Data Agent integrates Danswer, our data warehouse, and Wolfram's engine to streamline SQL generation, data retrieval, and analysis.

Idea Intake and Review Process



Idea Intake and Review Process

AI Idea Form		Intake	
<p>Enter your idea for an AI-powered business process improvement. Submitted ideas will be placed into a holding tank for review by the Gen AI workgroup or a relevant body. Not all ideas will be considered for implementation, however they may lead to identifying patterns or use cases that will help solve multiple problem areas.</p>			
Submitted By		Submission Date	
Amini, Mojgan		December 4, 2023	
Requestor (if different than Submitter)	VC Area *	Department	
<input type="text"/>	<input type="text"/>	<input type="text"/>	
<p>Idea Title *</p> <p>Short name of idea</p>			
<p>Idea Description *</p> <p>Description of the AI idea including problem being addressed, stated clearly and with enough contextual detail and data to establish why it is important, the specific problem it would solve, when and how this problem occurs, how much impact the problem has, and the requirements that are failing to be met.</p> <p>Start typing</p>			
<p>Attachments (optional) ?</p> <p>Select a File</p>			



Workgroup Assessment			Review
<p>Status * ?</p> <p>New</p>			
AI Category ?		Business Domain	
Tutor		IT	
Benefiting Audience			
Staff			
Impact/Benefit (1=low; 3=high) ?	Tech Alignment (1=low; 3=high) ?	Non-Tech Alignment (1=low; 3=high) ?	
2	3	3	
<p>1 – Low – Limited savings and value to university. Impacts/benefits few individuals in one business unit</p> <p>2 – Med – Moderate impact and value/savings. Impacts/benefits many individuals or multiple business units</p> <p>3 – High – Furthers the mission, large campus impact, significant value/savings. Impacts/benefits many individuals and multiple business units, university wide and/or system wide.</p>	<p>1 – Low. Not aligned with capabilities of existing architecture/tools. Will require research and new design solutions. Unknown resourcing needs.</p> <p>2 – Med. Existing architecture can be extended to add needed data/logic. Have done parts of this type of project before but not the entire scope.</p> <p>3 – High. Existing architecture can inherently support needed data/logic. Can leverage previous designs and processes.</p>	<p>1 – Low. Less than 50% of business processes are known. New processes to be defined. Unknown resourcing needs.</p> <p>2 – Med. 50%+ of the business processes are known. Moderate understanding of effort, moderate cost.</p> <p>3 – High. 80%+ of the business processes are known. Low investment cost, can leverage existing designs/tools/processes.</p>	
Decision ?		Decision Notes (viewable by submitter)	
Internal Notes (viewable by workgroup only)			
Score ?			

Sample AI Project Evaluation Charter

Process Overview

Purpose	Enhance the quality and accessibility of lecture recordings for all students.
Description	Create an Accessibility Enhancer tool that improves lecture captioning and transcription accuracy, with a focus on precise representation of academic terminology, to support and empower diverse student populations.
Key Points	<ul style="list-style-type: none"> Existing tools, such as Zoom's automatic captioning frequently struggle with accurately transcribing academic terminology, resulting in subpar performance for educational content. Utilizing advanced open-source AI tools like WhisperAI, we can efficiently transcribe the recordings from individual class sessions and entire audio/video series of past courses, significantly enhancing the accuracy and comprehensiveness of the captions. All students, especially those requiring accessibility accommodations (5% of the population may benefit from an accommodation but do not voice their needs) could benefit from this tool. In a pilot study conducted by ETS, the WhisperAI accessibility tool demonstrated a 30-40% improvement in performance compared to Zoom's automatic captioning when used with one professor's course material.

AI Solution Considerations

Potential Impact	<ul style="list-style-type: none"> Improved learning outcomes for students who use transcripts to study or review lectures. Increased compliance with Title II regulations. Cost savings when compared to external/vended solutions.
Questions & Considerations	<ul style="list-style-type: none"> Academic ideas should be aligned with Academic Senate and AI Admin Workgroup priorities Academic ideas should be aligned with the Academic Senate report on <i>AI in Education</i> (expected completion on/aaround Aug 31, 2024) Office of Students with Disabilities (OSD) should be consulted throughout the development process How would students be able to access the AI-generated captions? 60% of classes are either recorded or made available as podcasts, the remaining 40% lack any form of transcription, highlighting a significant accessibility gap. Could this solution help address this gap?

Desirability

Customer Size	One department multiple departments multiple VC areas campuswide
Customer Need	Low Medium High

Viability

Process Volume	<100 per month 100-1000 per month >1000 per month
Potential Savings per Process Cycle	<1 hour 1-4 hours >4 hours
Potential Savings per Month	<2,000 hours 2,000-20,000 hours >20,000 hours

Feasibility

Alignment with Existing Solutions	Addition to existing TGPT tool new TGPT tool new tool (non-TGPT)
Data Availability	Readily available Requires new integration Requires new data source
Complexity	Low Medium High

VC Area: Academic Affairs
Submitted by: Dan Suchy
On behalf of: Dan Suchy

AI in Administration Workgroup Charge

Purpose:

Artificial Intelligence and other advanced technologies are fast-becoming part of our everyday work processes. As with any new technology, there are risks, benefits, and challenges to consider. This workgroup will **build awareness** of UC San Diego's **AI**-related strategies, **understand** and **promote** published **guidelines**, and **influence** the **safe application of AI** across the university.

- **Co-chairs:** Allorah Pradenas, Mojgan Amini
- **Sponsors:** CFO, CIO Vince Kellen
- **Participants:** ~20 representatives from university units, sanctioned by Enterprise Information Services Committee
- **Cadence:** Bi-monthly meetings and ongoing asynchronous communication
- **Key Artifacts:** List of AI initiatives & AI ideas, approach for showcasing AI outcomes & successes

Discussion outline

- 1 Strategy
- 2 Architecture
- 3 Assistants, use cases
- 4 Roadmap
- 5 Collaborations

April - June

- **April 15:** Implemented Single Sign
- **April 22:** Installed additional H100 in the environment at SDSC
- **April 29:** Upgrade from Llama 2 to Llama 3 LLM
- **May 13:** Launched Fund Manager Coach to beta testers
- **May 15:** Deployed TritonGPT to all 37,000 Campus and Health Sciences employees
- **June 14:** Upgraded to major new release of the Danswer framework

July - September

- **July 8:** Kicked off engagement with Wolfram for fact retrieval
- **July 15:** Developed pilot referred from AI Governance Work Group – Document classification
- **July 29:** Upgrade to Llama 3.1 LLM
- **August 13:** Upgraded to major new release of the Danswer Framework. Fund Manger Coach broadly available
- **August 26-28:** Signed UC Berkeley & San Diego State as TritonGPT customers
- **September 12:** Getting the most out of TritonGPT Webinar
- **September 23:** Pilot fact retrieval using Enterprise Data Agent and Wolfram

October - December

- **October 1-4:** UC Berkeley & San Diego State deployments
- **October 21:** End user pilot for Document Classification
- **November 1:** Begin Protopia implementation
- **November/December:** Implement additional projects from AI Work Group
- **November/December:** Support instructional use cases
- **November/December:** Enterprise content pilot

Collaboration values and approach

Service to the community. Help democratize access to information and knowledge using GenAI. Accelerate community adoption. Share practical knowledge

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

Current and potential collaborations

- Institutional collaborators
 - San Diego Supercomputer Center (SDSC)
 - San Diego State University, UC Berkeley
 - In discussions with 5 other universities, state government
- Software provider collaborators
 - Danswer
 - Wolfram
 - Protopia
- Implementation collaborators
 - Slower, Inc.
 - Moran Technology Consulting
- Extramural funding
 - Multi-institutional grant opportunities
- Cloud and implementation partners
 - AWS – TGPT cloud fail-over and cloud bursting
 - Slower Inc. – Kafka Airflow and execution pilots
 - DataBank – Enterprise content system integrations
- Topics for additional collaboration
 - Knowledge elicitation from experts automation
 - Structured data metadata curation automation
 - Improving dialog quality with live-event usability analysis
 - TritonGPT and Wolfram Mathematica for students
 - Local ensemble approach for LLM safety
 - Open Educational Resource content generation

Questions?

Feel free to contact the team here!

Brett Pollak (bpollak@ucsd.edu)

Vince Kellen (vkellen@ucsd.edu)

Kevin Chou (kcchou@ucsd.edu)

Mojgan Amini (maamini@ucsd.edu)

Dan Suchy (dsuchy@ucsd.edu)

Extra slides

Where and AI/Gen AI help?

- **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](#)

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?