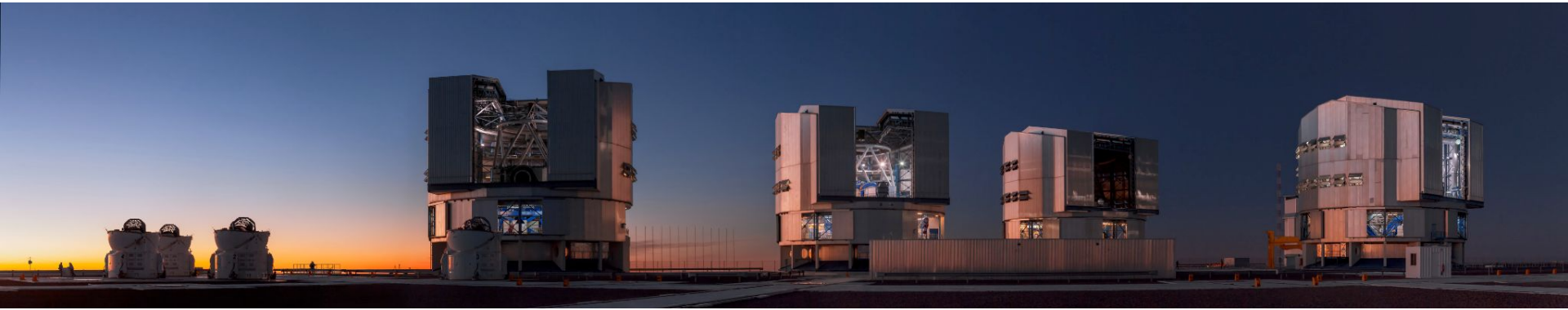
Building LLM-based Systems & Products at Paranal



Engineering with prompts, rather than prompt engineering

Cristóbal Alcázar
MSc in Finance & (not yet) MSc in Data Science

Cristóbal Alcázar

alcazar.cristobal90@gmail.com - alkzar.cl

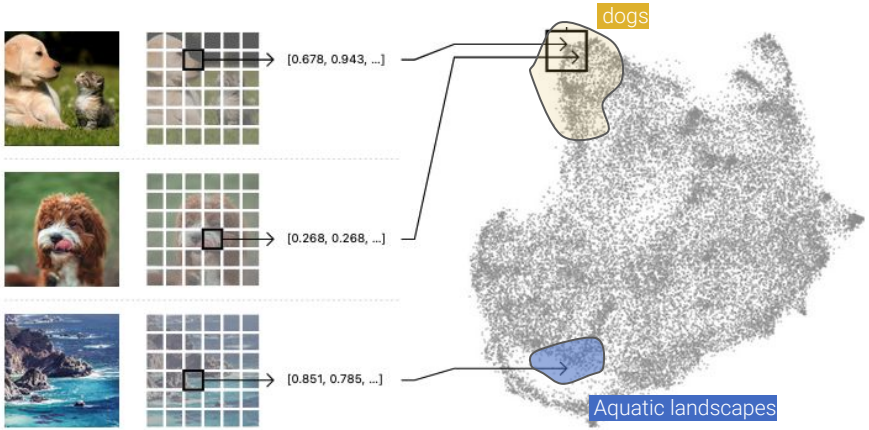
- MSc (not yet) in Data Science @U. de Chile
- Current research: diffusion generative models, RLHF. Supervised by Felipe Tobar.
- MDS7203 Deep Generative Models: Teaching assistant.
- Hugging Face student ambassador program.
- Finance and Economic background.
- Work Experience: 3 years in Macroeconomic Statistics area at the Central Bank of Chile. 2 years at a Fintech joint venture on payments (currently the payment subsidiary of BancoEstado).
- Currently at @ ESO Paranal Software group (winter-internship 2023).

Context

LLM, embeddings, generative models, context window

Neural Networks and Representation Learning

Learning a simplified numerical representation of complex data



- **Embeddings:** Vectors representing complex data in a numerical and dense format.
- **Vector space representation:** Imagine a space where data resides. Within this space, dogs cluster together, forming a distinct neighborhood than cats, but both classes lives in a different district than humans.
- **Embeddings as a features:** These vectors serve as a foundation for constructing various systems (object recognition, scene detection).

Image Source: Exploring Neural Network with Activation Atlasses (Carter et al. 2019)

Neural Networks and Representation Learning

End-to-end system, from raw data to task mapping layers.

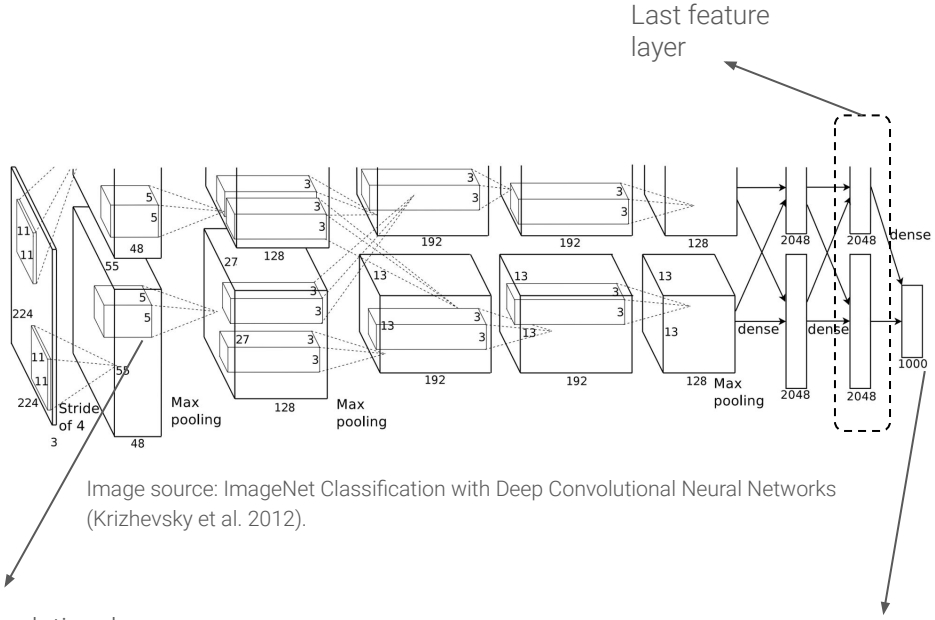
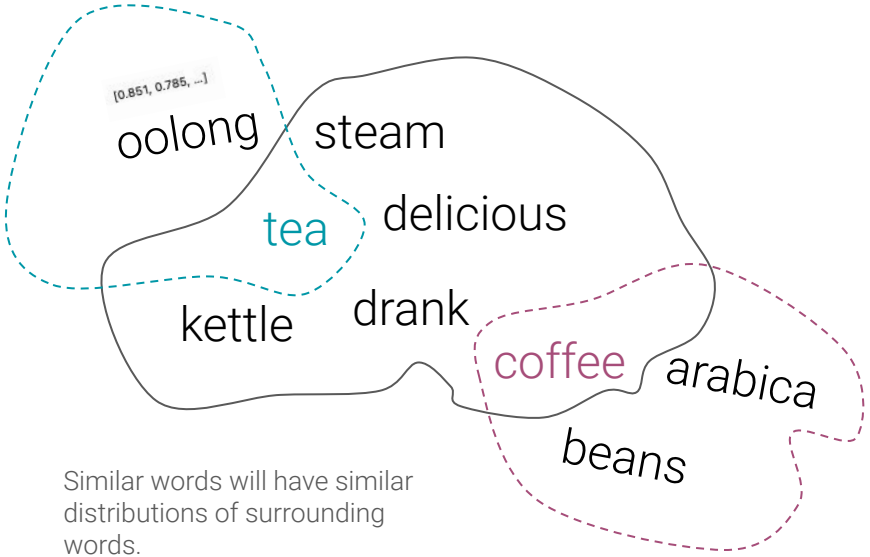


Image source: ImageNet Classification with Deep Convolutional Neural Networks (Krizhevsky et al. 2012).

Image source: t-SNE visualisation of CNN Codes (Karpathy)

Vector Semantics and Embeddings

"You shall know a word by the company it keeps" .- Firth, 1957

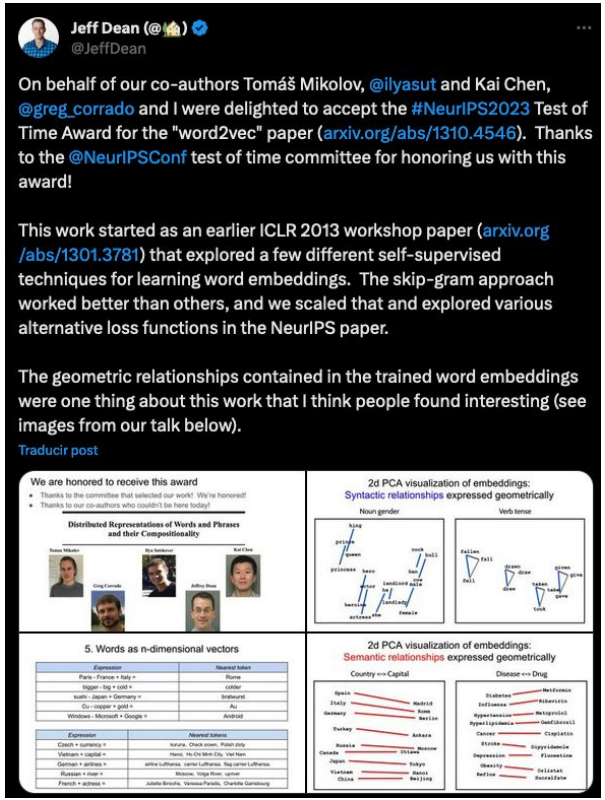


Similar words will have similar distributions of surrounding words.

- **Distributional hypothesis:** the meaning of a word can be derived from the distribution of contexts in which it appears.
- **Contextual embeddings:** to capture the semantic of a word or token, context is critical. It's necessary a way to resignify a fix embedding given their context.
- **More about sequence models:** *Sequence Models with Transformers. A Journey Through Language Models, Current Insights, and Practical Applications at Paranal Observatory (Carvajal, Alcázar 2023).*

2023 NeurIPS Test of Time Award

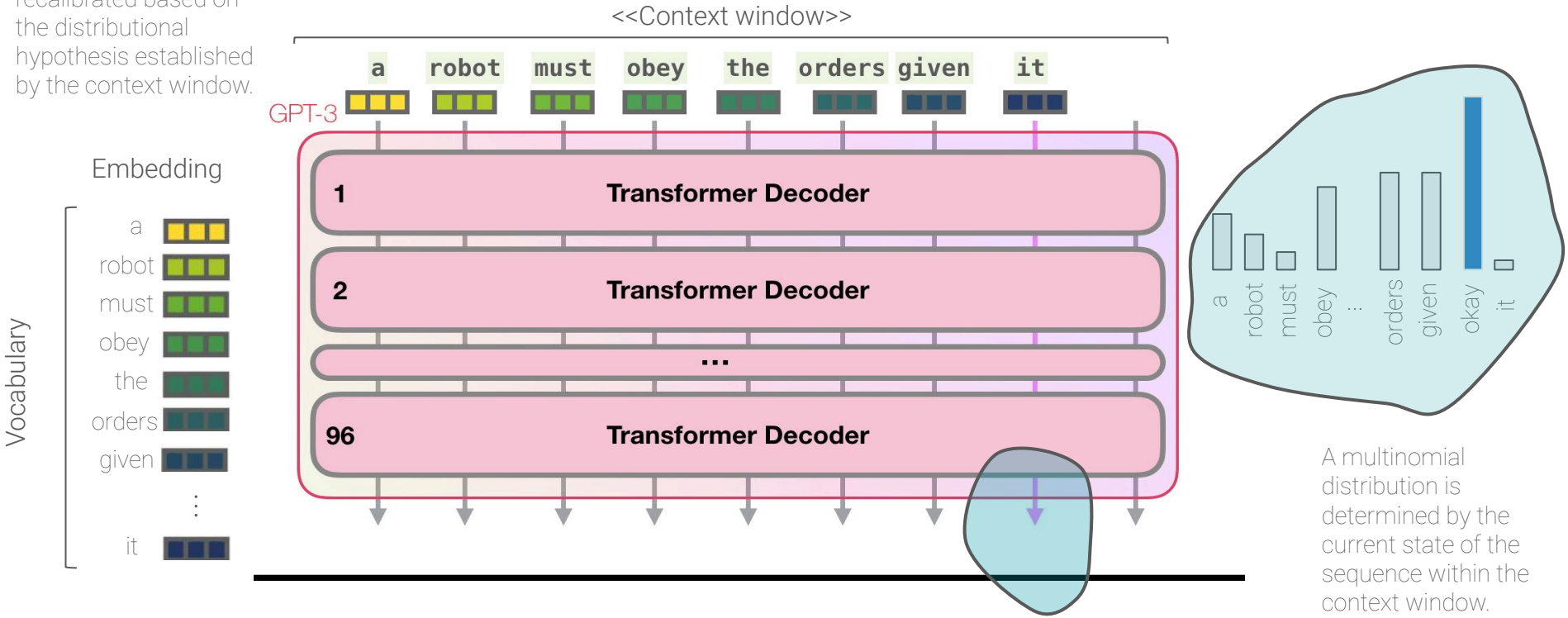
Distributed Representations of Words and Phrases and their Compositionality (word2vec)



A Neural Network to predict the next token in the sequence

Generative Pre-trained Transformer (GPT)

Embeddings are recalibrated based on the distributional hypothesis established by the context window.



A multinomial distribution is determined by the current state of the sequence within the context window.

Image Source: [How GPT3 Works - Visualizations and Animations (Alammar 2020)]

How to Achieve a ChatGPTish Model?

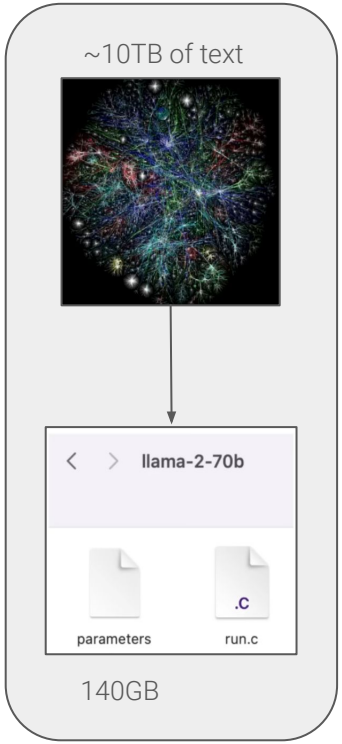
Involved at least 2 training stages and result in a lossy compression about 5 order of magnitude

every ~year

Stage 1: Pretraining
1. Download ~10TB of text.
2. Get a cluster of ~6,000 GPUs.
3. Compress the text into a neural network, pay ~$2M, wait ~12 days.
4. Obtain **base model**.

every ~week

Stage 2: Finetuning
1. Write labeling instructions
2. Hire people (or use scale.ai!), collect 100K high quality ideal Q&A responses, and/or comparisons.
3. Finetune base model on this data, wait ~1 day.
4. Obtain **assistant model.**
5. Run a lot of evaluations.
6. Deploy.
7. Monitor, collect misbehaviors, go to step 1.



Source: Intro to Large Language Models (Karpathy 2023)

Resource - Llama 2: Open Foundation and Fine-Tuned Chat Models (paper)

Zero-shot prompt

Just feed the task text to the model and ask for results.

Text: i'll bet the video game is a lot more fun than the film.
Sentiment: [MODEL COMPLETION]

Few-shot prompt

High-quality task demonstrations, showcasing input-output pairs. **Pros:** enhance comprehension of human intent and expected answer criteria. **Cons:** increase token consumption and context length limitations.

Text: (lawrence bounces) all over the stage, dancing, running, sweating, mopping his face and generally displaying the wacky talent that brought him fame in the first place.
Sentiment: positive

Text: despite all evidence to the contrary, this clunker has somehow managed to pose as an actual feature movie, the kind that charges full admission and gets hyped on tv and purports to amuse small children and ostensible adults.
Sentiment: negative

Text: i'll bet the video game is a lot more fun than the film.
Sentiment: [MODEL COMPLETION]

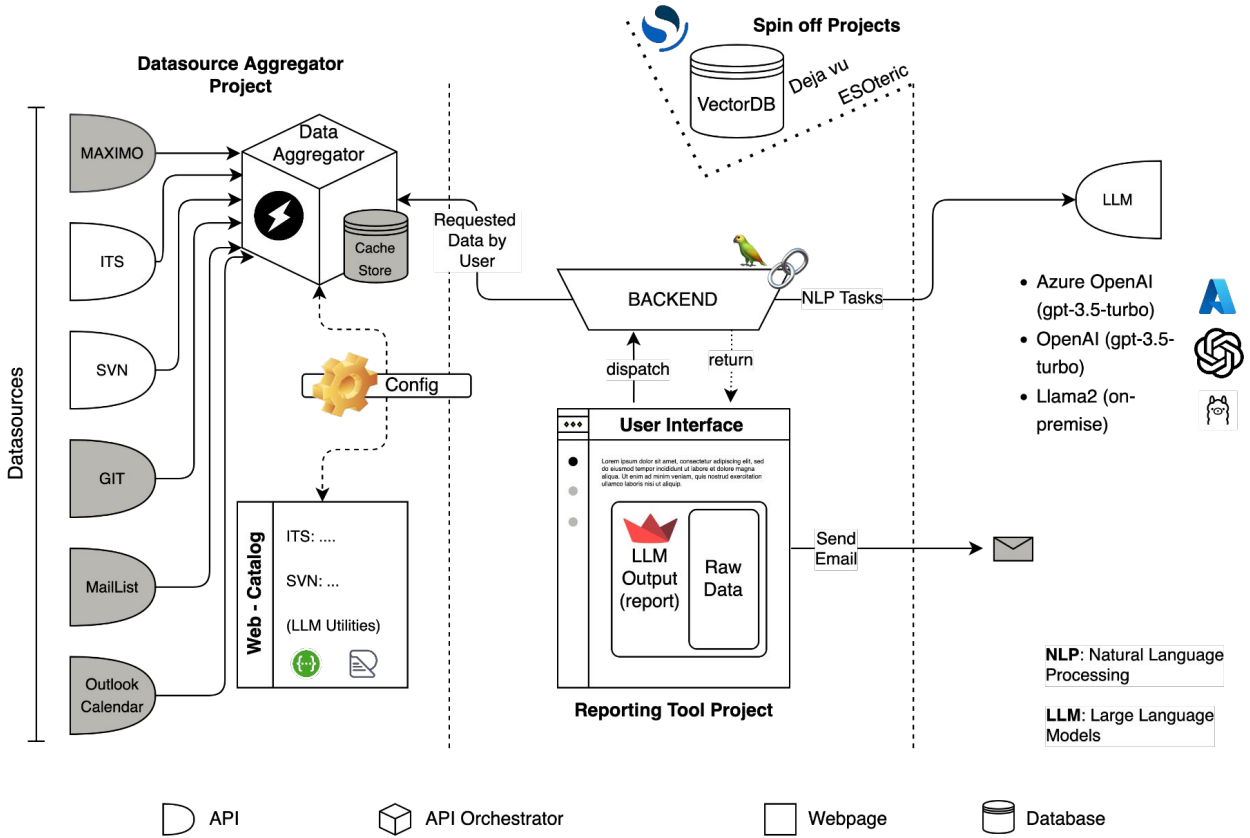
Source: Prompt Engineering (Weng 2023)

The Internship

In a nutshell

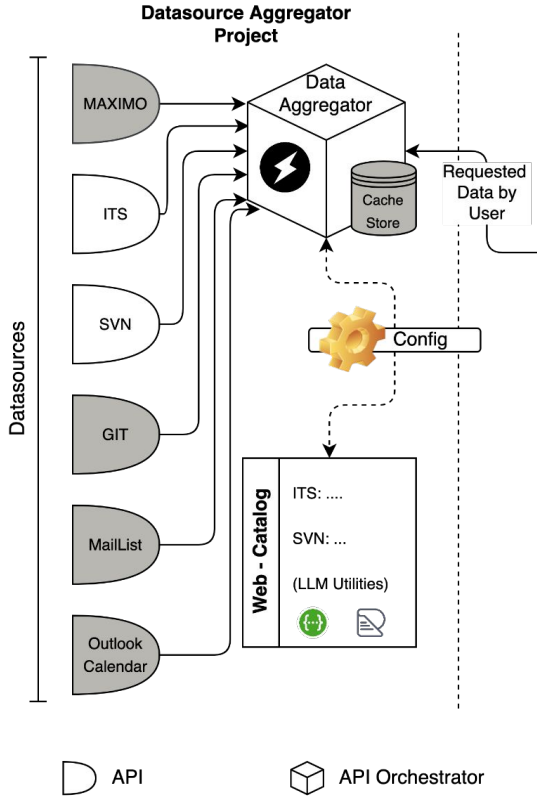
Internship Overview

Datasource Aggregator - Reporting Tool - Spin off Projects



Datasource Aggregator

A comprehensive API featuring various endpoints for retrieving specific data from diverse sources



- **Aggregator:** connecting different data sources.
- Specific interaction with **data sources** (e.g., obtaining a ticket for a user).
- Independent project; Reporting Tool is the first service...

Get Issues By User Between

Returns issues related to a specific ITS (Jira) user within a time range between from_date and to_date.

Parameters:

- `jira_user` (str): The ITS (Jira) user to retrieve issues for.
- `from_date` (str): The date to retrieve issues from. Format: YYYY-MM-DD.
- `to_date` (str): The date to retrieve issues to. Format: YYYY-MM-DD or 'HEAD' for the current date.

Returns:

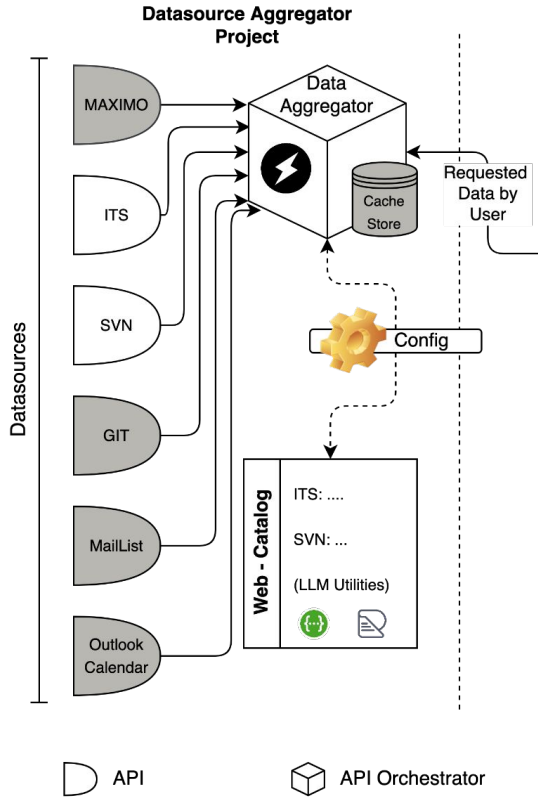
- `List[JiraIssue]`: A list of parsed ITS (Jira) issues.

QUERY PARAMETERS

jira_user	string (Jira User) Default: `"jaraneda"` The ITS (Jira) user to retrieve issues for.
from_date	string (From Date) Default: `"2023-11-25"` The date to retrieve issues from.
to_date	string (To Date) Default: `"2023-12-25"` The date to retrieve issues to. You can use 'HEAD' to retrieve issues up to the current date.

Responses

> **200** Successful Response

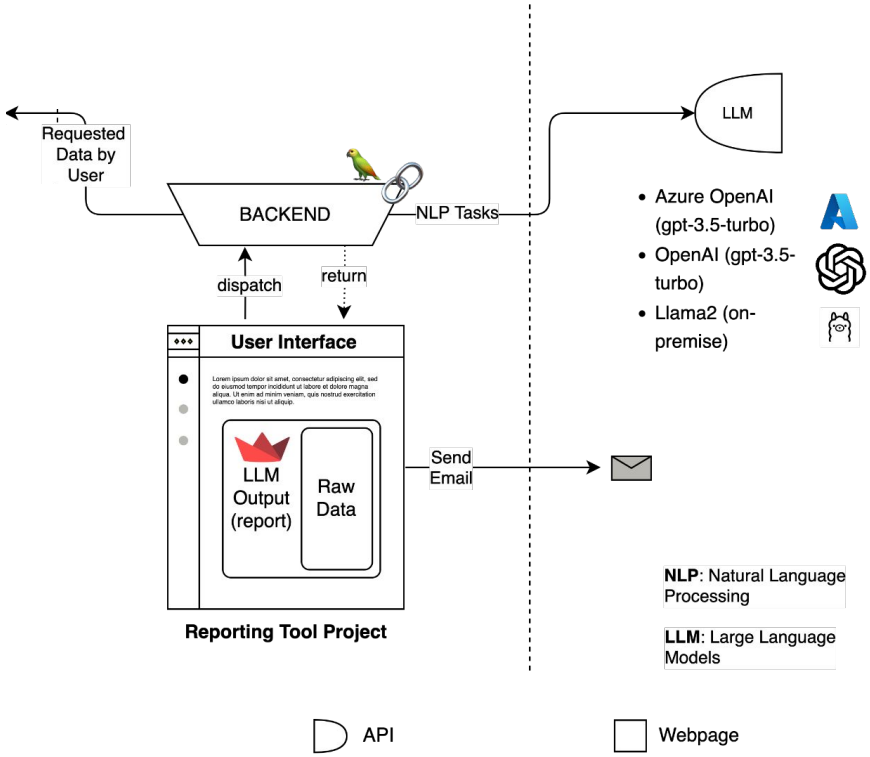


Source: https://scgitlab.sc.eso.org/psw/projects/datasource-aggregator

Reporting Tool

Application to Generate Reports Automatically

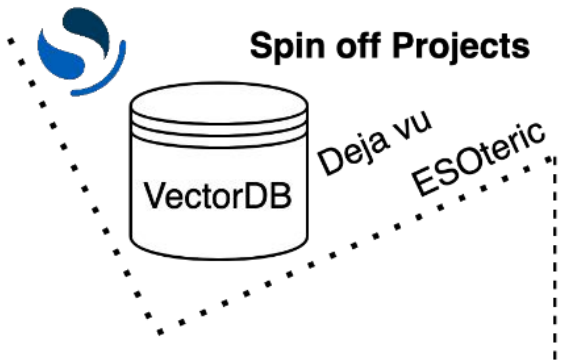
- Automatic generation of reports through summarization using LLM.
- Builds the domain of information requested by the user.
- User serves as an editor with the option to edit and remove the information to be used by the LLM (augmented tool).



Spin off Projects

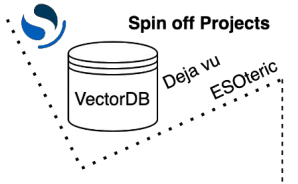
Vector Similarity Search & Retrieval Augmented Generation

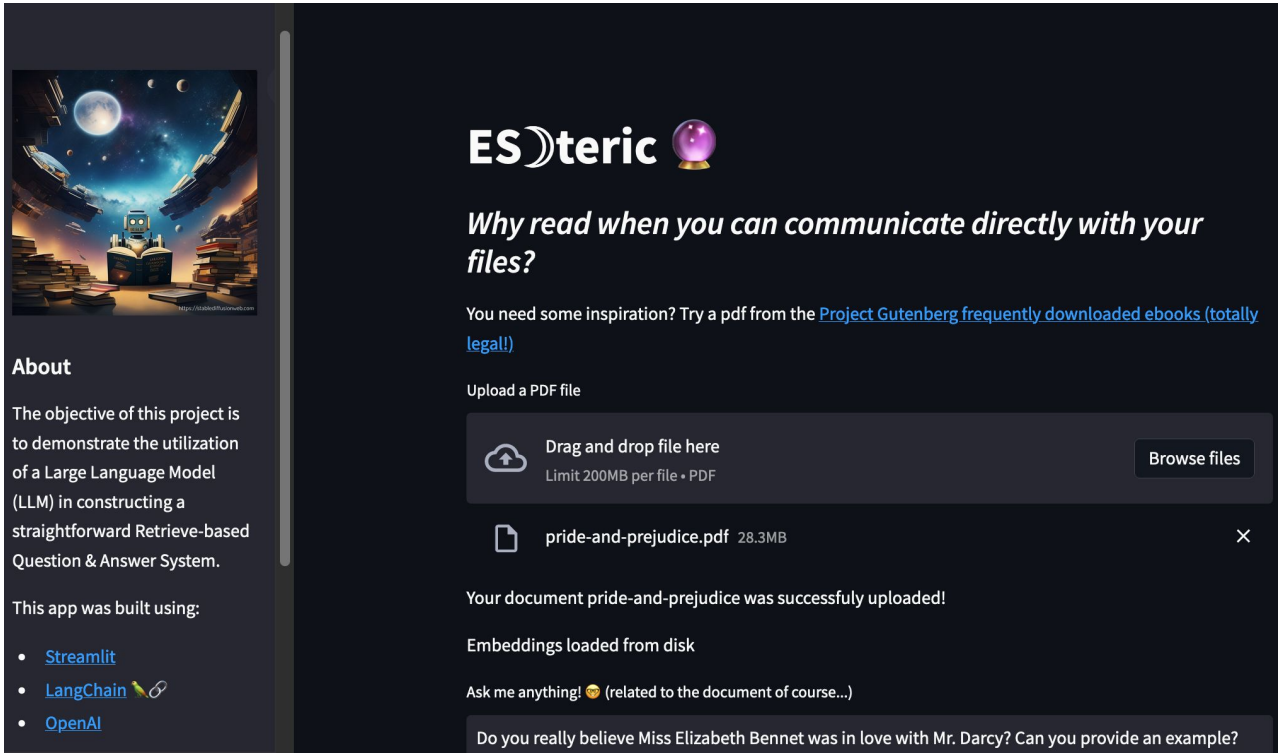
- **ESOteric**: a “talk with your files” chatbot.
- **DejaVu** (by NMI).
- Two core concepts behind these spin off projects.
  - Vector Similarity Search
  - Retrieval Augmented Generation (RAG)



ESOteric: a “Talk with your Files” Q&A

Retrieval Augmented Generation (RAG) Pattern





https://gitlab.eso.org/calcazar/esoteric-qa

ESOteric: a “Talk with your Files” Q&A

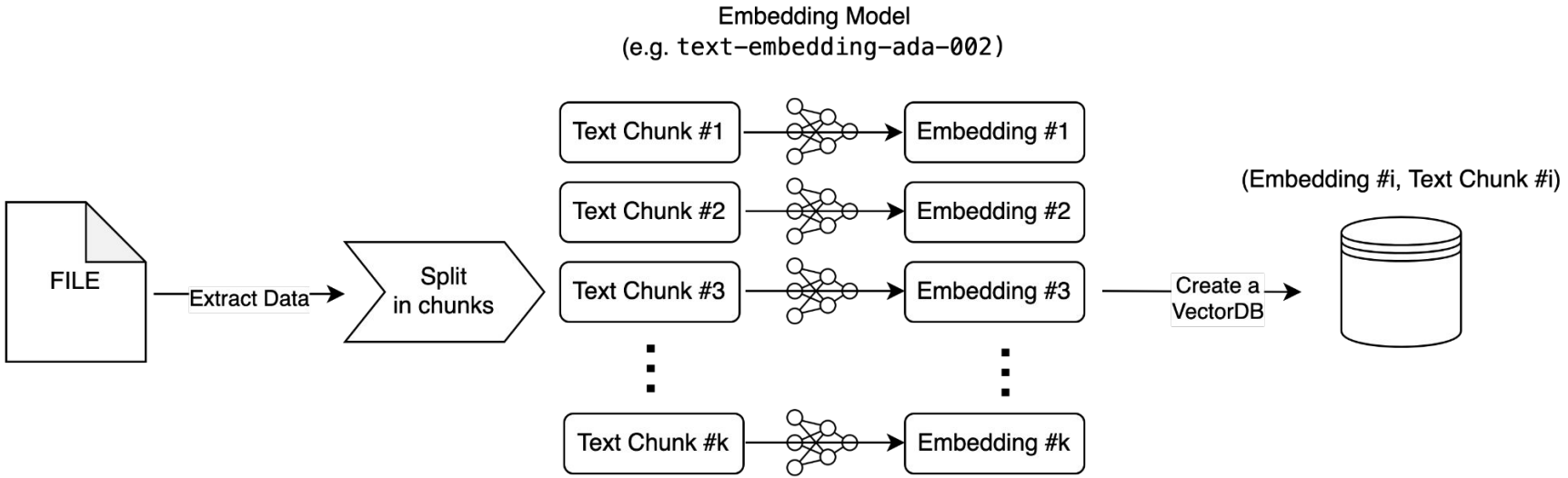
Retrieval Augmented Generation (RAG) Pattern

https://gitlab.eso.org/calcazar/esoteric-qa

- A reader assistant agent via Question & Answering (QA) to interact with pdf documents such as Pride and Prejudice.
- A retrieve-based QA System:
  1. A large document, or a collection of documents, is transform into a knowledge base using embeddings.
  2. User ask questions that retrieve information from the knowledge data using a large language model.

How to Create a Vector Database?

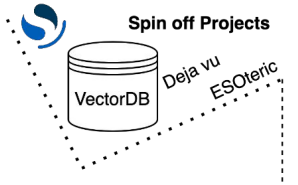
A simple split-apply pattern to create a vector database



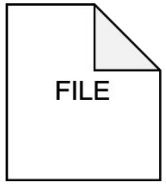
Create your own vector database with OpenSearch
https://gitlab.eso.org/calcazar/langchain-series/-/tree/main/examples/embedder

How to Create a Vector Database?

A simple split-apply pattern to create a key, value database



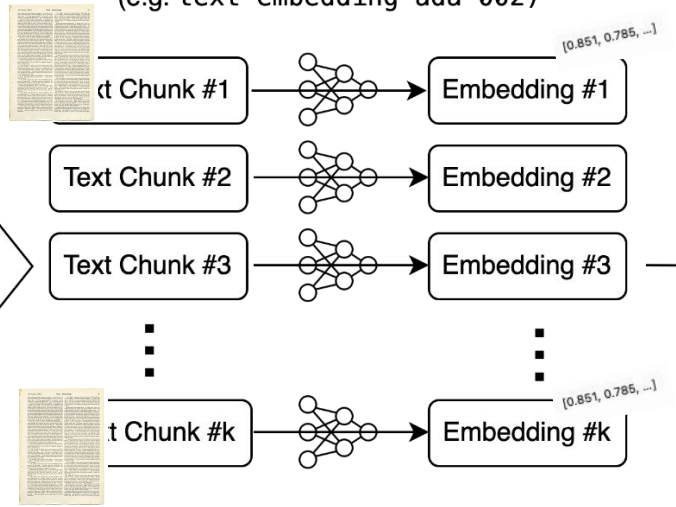
Get a book in a readable format



Extract Data

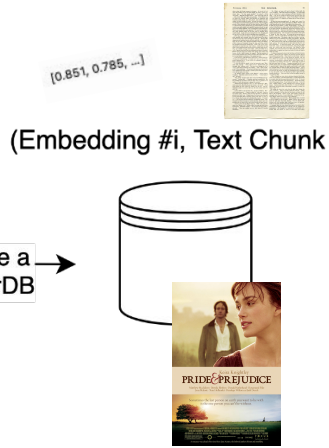
Split in pages

Embedding Model
(e.g. `text-embedding-ada-002`)



Get embeddings

Create a VectorDB

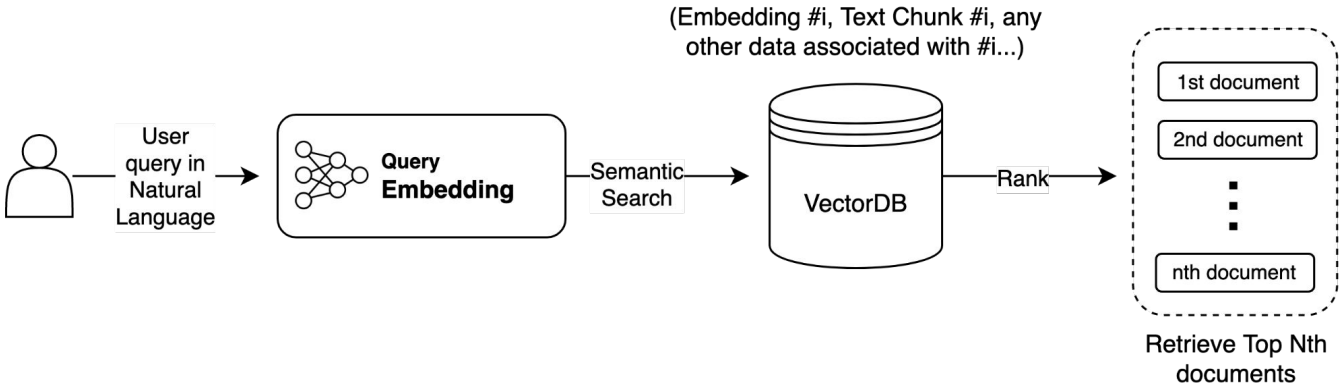


A Pride & Prejudice "knowledge database"

Create your own vector database with OpenSearch
https://gitlab.eso.org/calcazar/langchain-series/-/tree/main/examples/embedder

Retrieve Documents using Natural Language

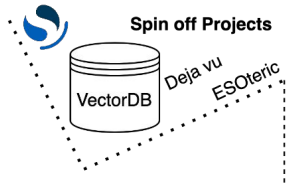
Embeddings deliver similarity metrics, while the remainder involves ranking and retrieval



Create your own vector database with OpenSearch
https://gitlab.eso.org/calcazar/langchain-series/-/tree/main/examples/embedder

Retrieval Augmented Generation

Improve LLM input by retrieving and incorporating pertinent data from a vectorDB



User: Do you really believe Miss Elizabeth Bennet was in love with Mr. Darcy? Can you provide and example?

User query in Natural Language

Query Embedding

Semantic Search

A Pride & Prejudice "knowledge database"

(Embedding #i, Text Chunk #i, any other data associated with #i...)

VectorDB

Rank

1st document

2nd document

nth document

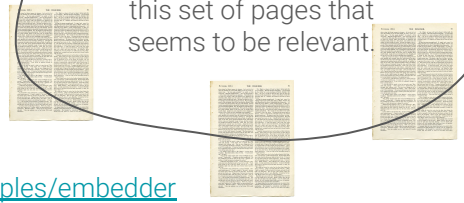
Retrieve Top Nth documents

LLM

Input - Prompt

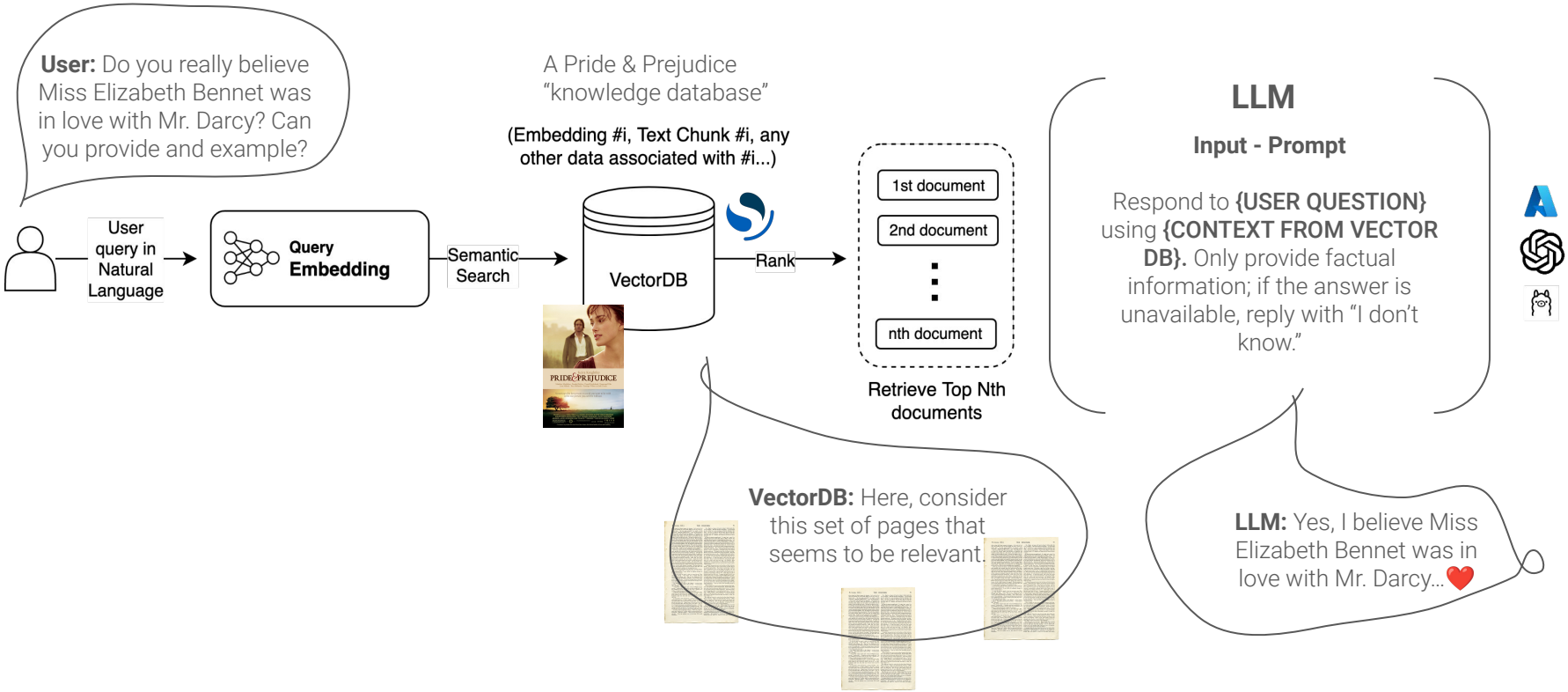
Respond to **{USER QUESTION}** using **{CONTEXT FROM VECTOR DB}.** Only provide factual information; if the answer is unavailable, reply with "I don't know."

VectorDB: Here, consider this set of pages that seems to be relevant.

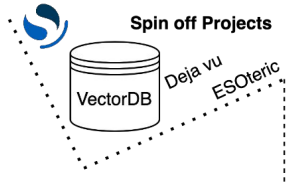


LLM: Yes, I believe Miss Elizabeth Bennet was in love with Mr. Darcy... ❤️

Create your own vector database with OpenSearch
https://gitlab.eso.org/calcazar/langchain-series/-/tree/main/examples/embedder


User: Do you really believe Miss Elizabeth Bennet was in love with Mr. Darcy? Can you provide and example?
User query in Natural Language
Query Embedding
Semantic Search
A Pride & Prejudice "knowledge database"
(Embedding #i, Text Chunk #i, any other data associated with #i...)
VectorDB
Rank
1st document
2nd document
nth document
Retrieve Top Nth documents
LLM
Input - Prompt
Respond to {USER QUESTION} using {CONTEXT FROM VECTOR DB}. Only provide factual information; if the answer is unavailable, reply with "I don't know."
VectorDB: Here, consider this set of pages that seems to be relevant.
LLM: Yes, I believe Miss Elizabeth Bennet was in love with Mr. Darcy...❤️

ESOteric sample responses



Ask me anything! 🤓 (related to the document of course...)

Do you really believe Miss Elizabeth Bennet was in love with Mr. Darcy? Can you provide an example?

Do you really believe Miss Elizabeth Bennet was in love with Mr. Darcy? Can you provide an example?

Yes, I believe Miss Elizabeth Bennet was in love with Mr. Darcy. An example of this is when Mr. Darcy tells Miss Bingley that Miss Elizabeth Bennet is his favorite lady, and Miss Bingley is surprised by this revelation.

Ask me anything! 🤓 (related to the document of course...)

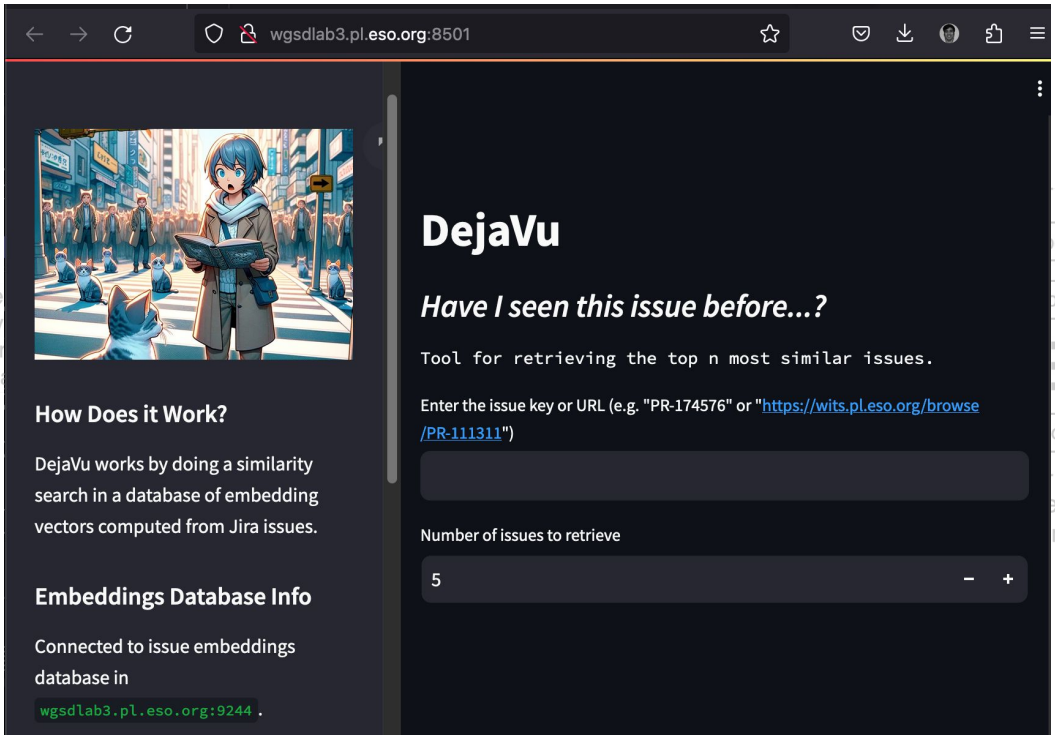
Why is Lizzy Mr. Bennet's favorite daughter?

Why is Lizzy Mr. Bennet's favorite daughter?

Based on the given context, it is not explicitly stated why Lizzy is Mr. Bennet's favorite daughter.

DejaVu: A Vector Similarity Search for ITS Issues

A practical and clean application of vectorDB, by NMI



Source: https://scgitlab.sc.eso.org/nmiranda/dejavu

Automate Report Generation

How we build a system that take relevant information, categorize, and summarize from the users' perspective?

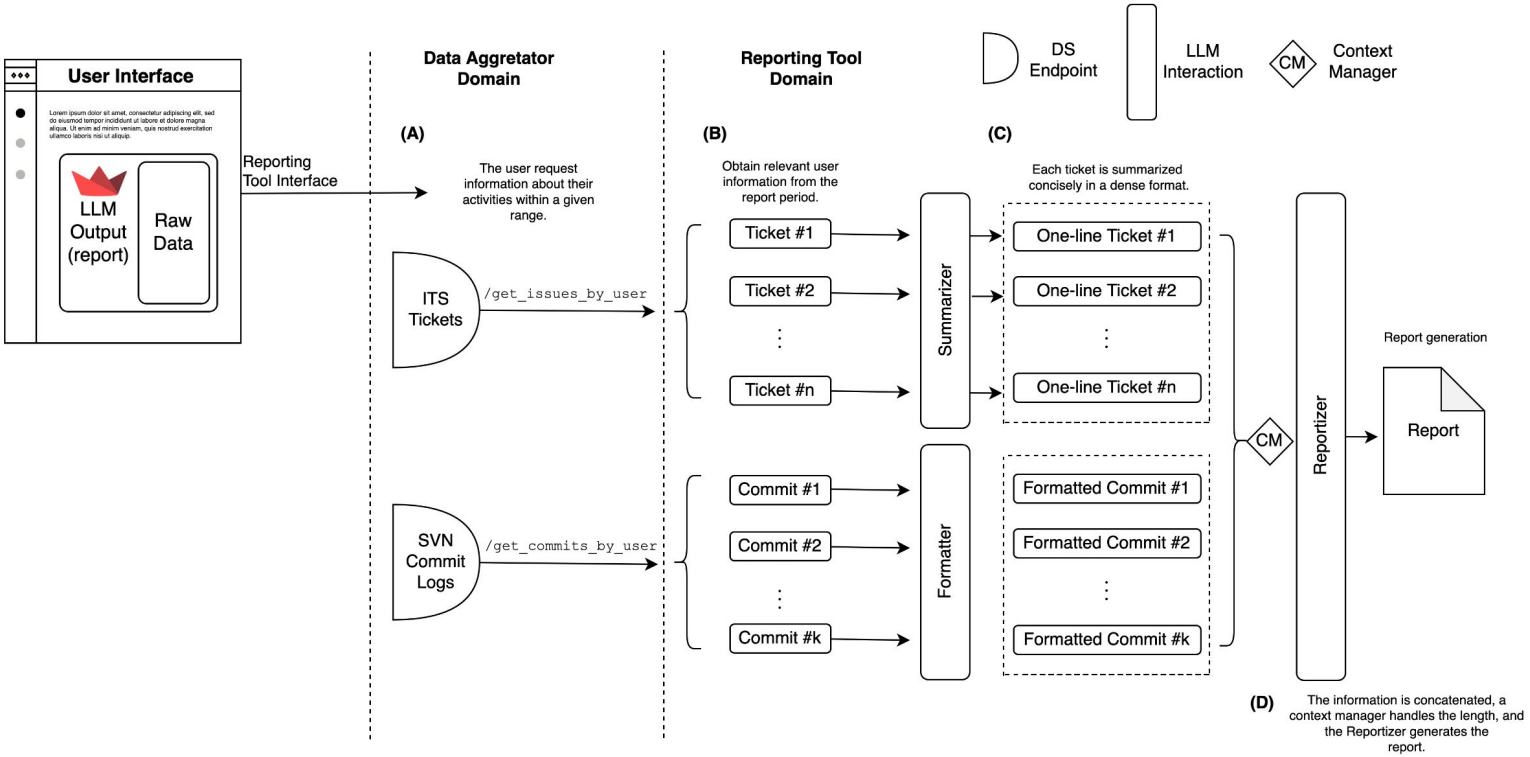
Reporting Tool

Application to Generate Reports Automatically

Source: https://scgitlab.sc.eso.org/psw/projects/reporting-tool

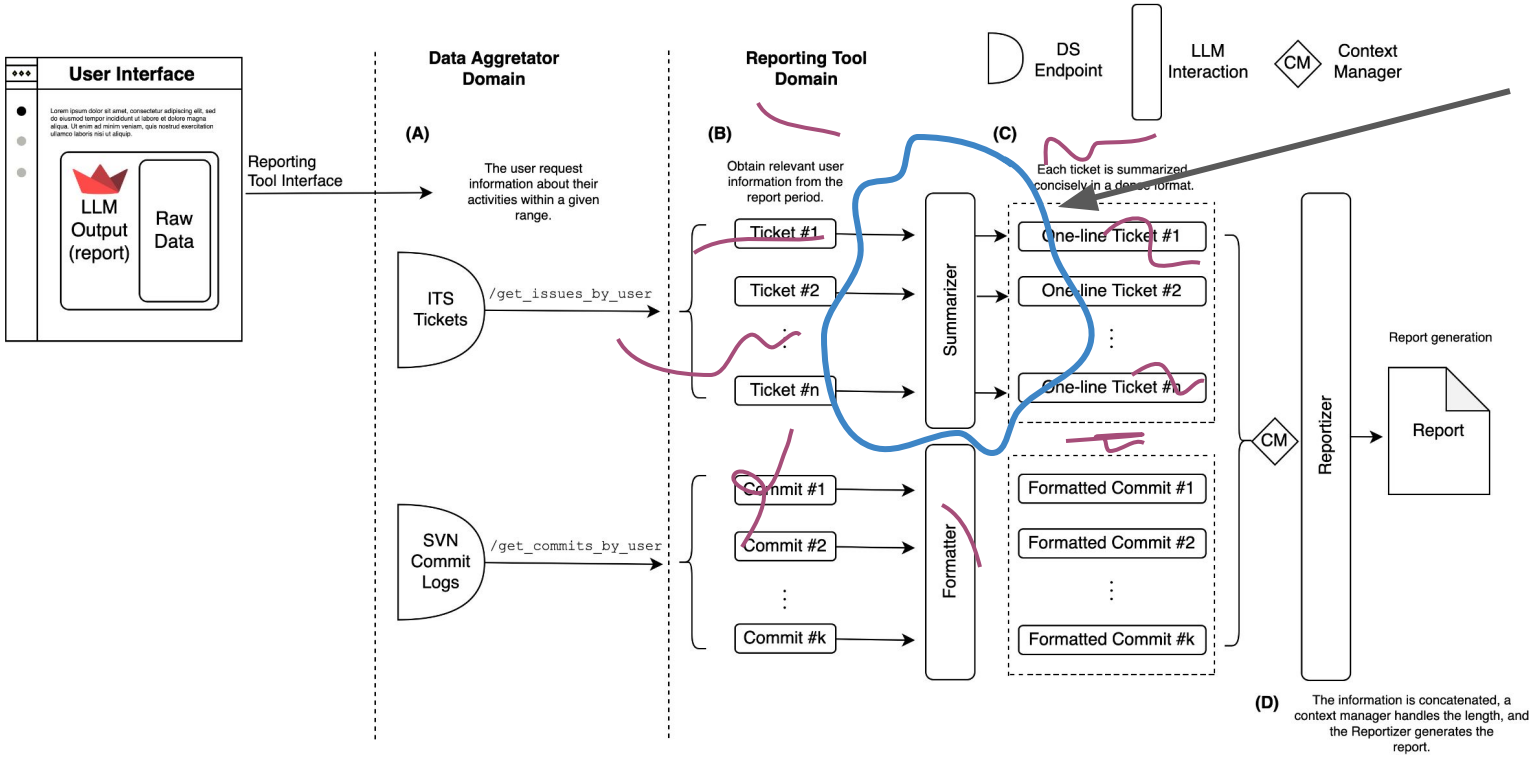
Report Generation

User Domain Information & Impersonation



How does the LLM Interaction Work?

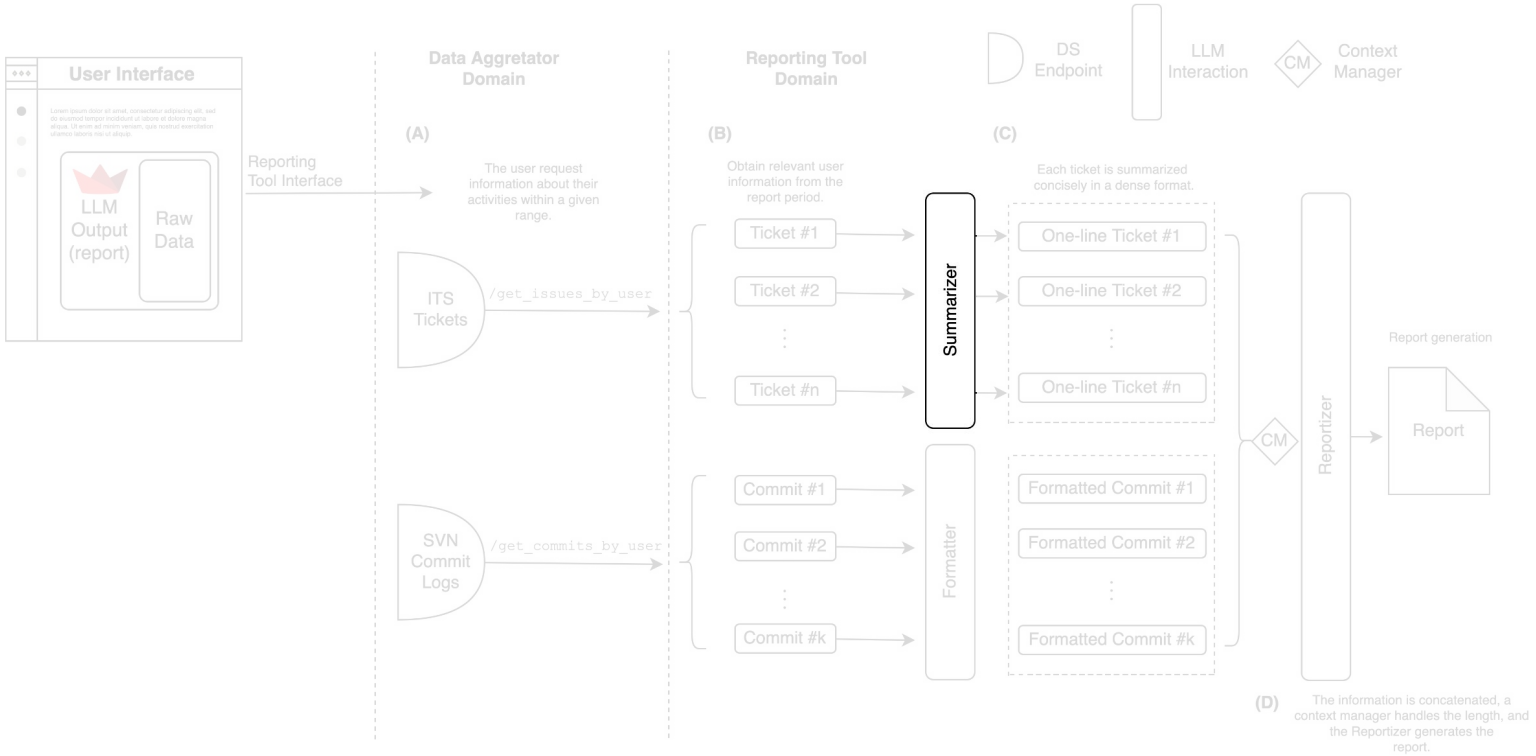
Let's Break Down the Summarizer…



Here comes the "AI", or rather, the **stochastic parrot** 🦜 !

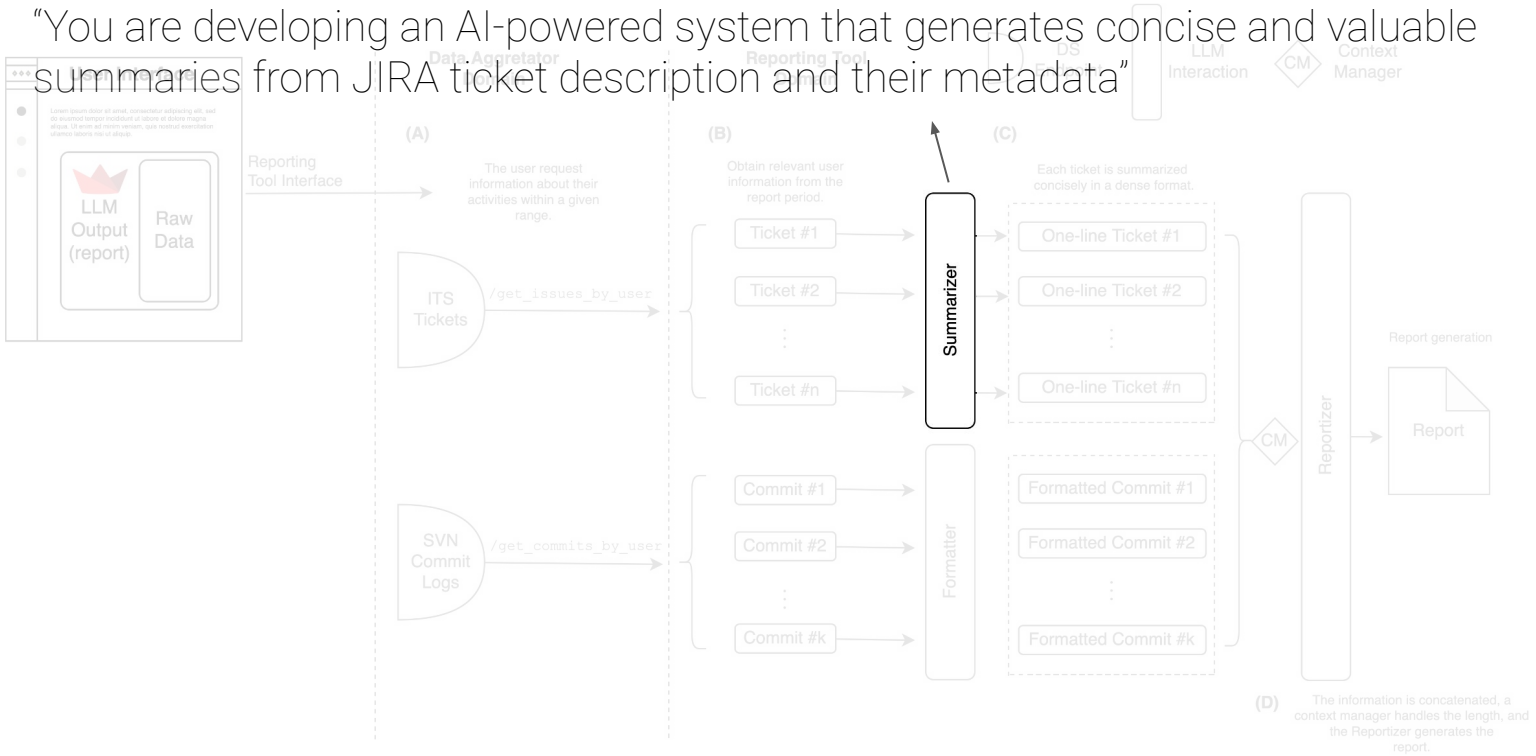
How does the LLM Interaction Work?

Let’s Break Down the Summarizer…



The Summarizer Goal is Conditioning via the System Prompt

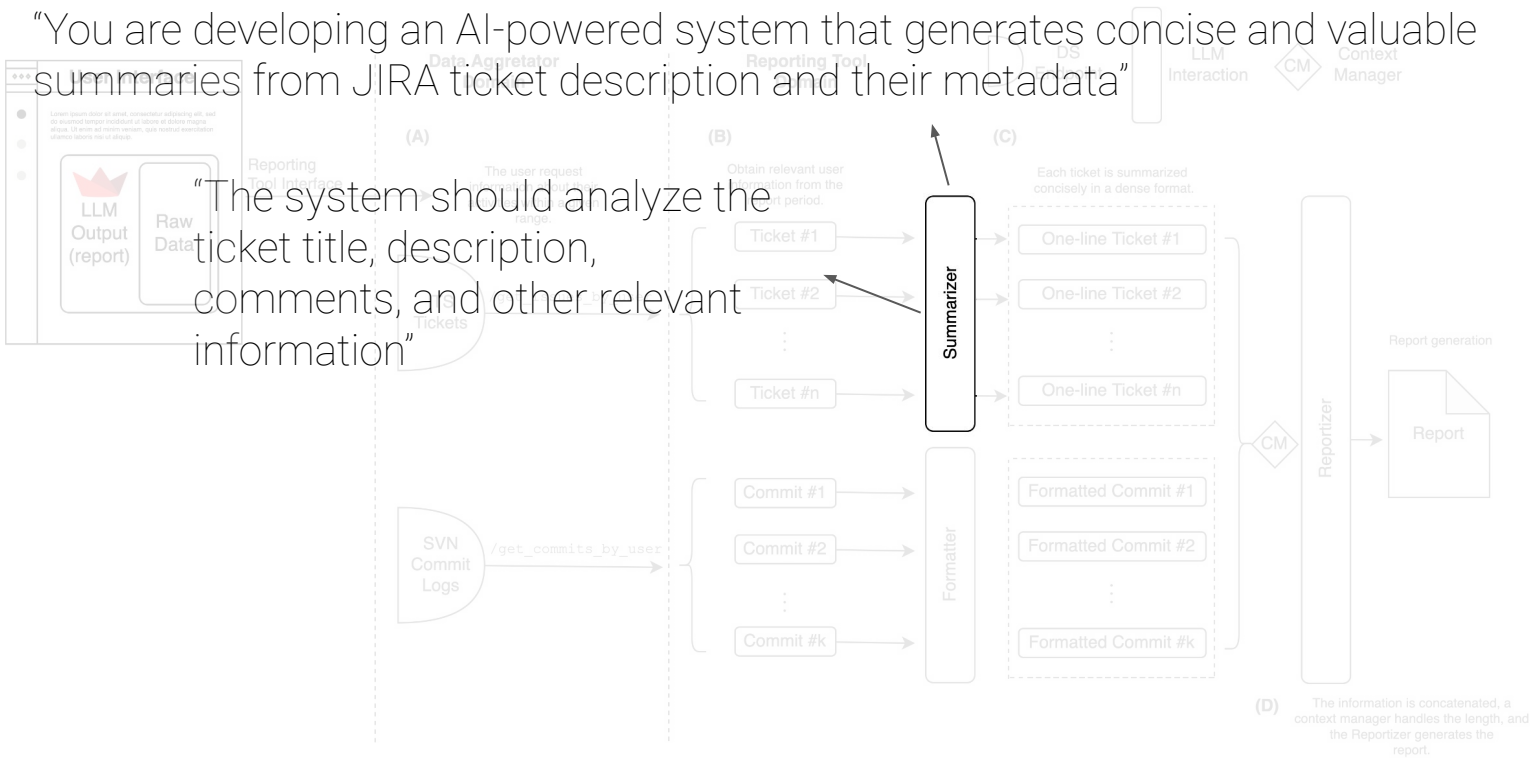
"You are developing an AI-powered system that generates concise and valuable summaries from JIRA ticket description and their metadata"



The Summarizer Goal is Conditioning via the System Prompt

"You are developing an AI-powered system that generates concise and valuable summaries from JIRA ticket description and their metadata"

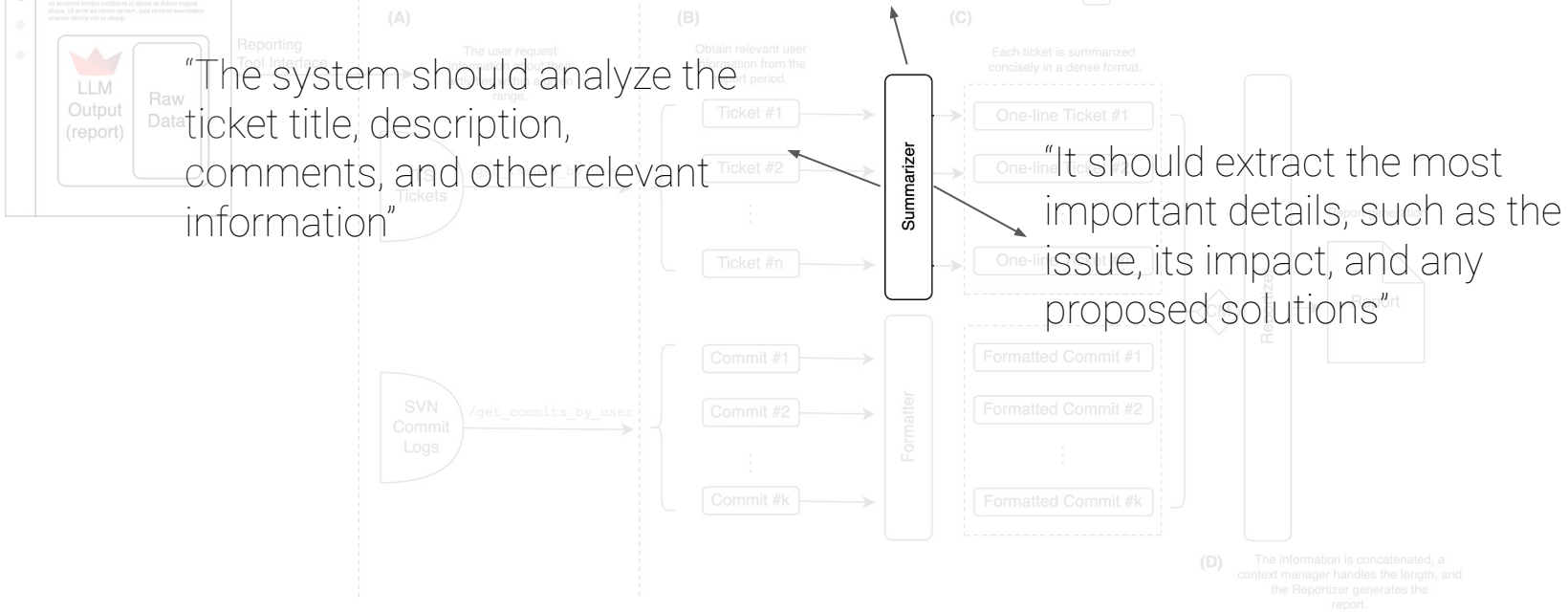
"The system should analyze the ticket title, description, comments, and other relevant information"



The Summarizer Goal is Conditioning via the System Prompt

"You are developing an AI-powered system that generates concise and valuable summaries from JIRA ticket description and their metadata"

"The system should analyze the ticket title, description, comments, and other relevant information"



"It should extract the most important details, such as the issue, its impact, and any proposed solutions"

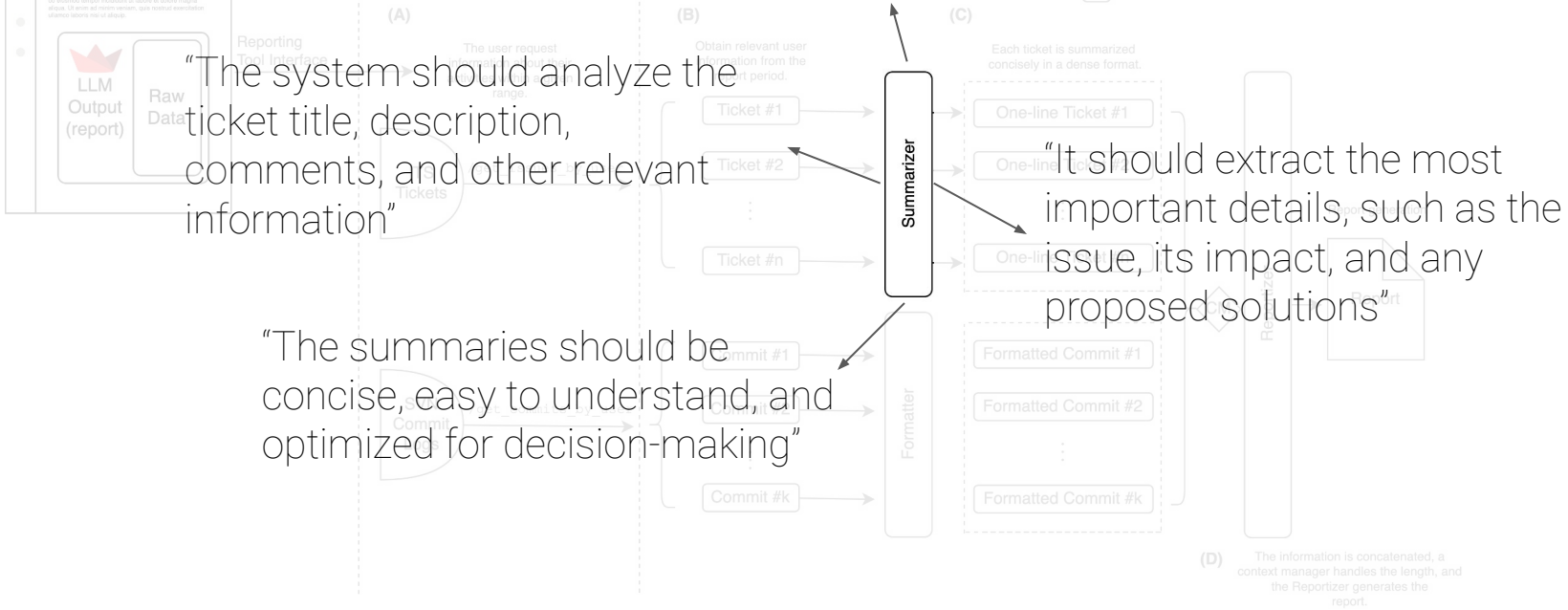
The Summarizer Goal is Conditioning via the System Prompt

"You are developing an AI-powered system that generates concise and valuable summaries from JIRA ticket description and their metadata"

"The system should analyze the ticket title, description, comments, and other relevant information"

"It should extract the most important details, such as the issue, its impact, and any proposed solutions"

"The summaries should be concise, easy to understand, and optimized for decision-making"



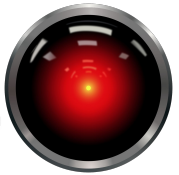
The Summarizer Goal is Conditioning via the System Prompt

"You are developing an AI-powered system that generates concise and valuable summaries from JIRA ticket description and their metadata"

"The system should analyze the ticket title, description, comments, and other relevant information"

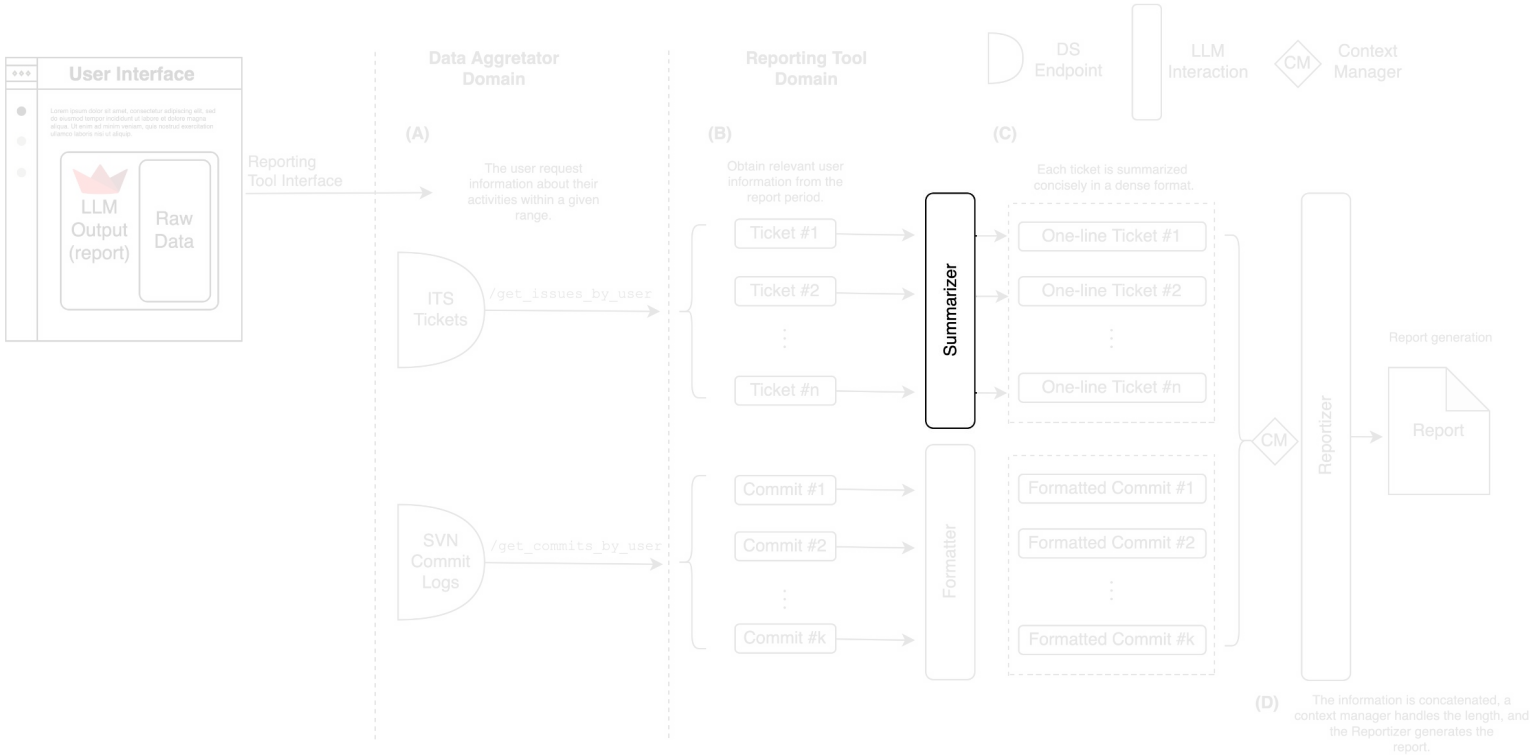
"It should extract the most important details, such as the issue, its impact, and any proposed solutions"

"The summaries should be concise, easy to understand, and optimized for decision-making"



"Your goal is to create a system that helps users quickly grasp the key information from JIRA tickets, enabling efficient collaboration and problem-solving."

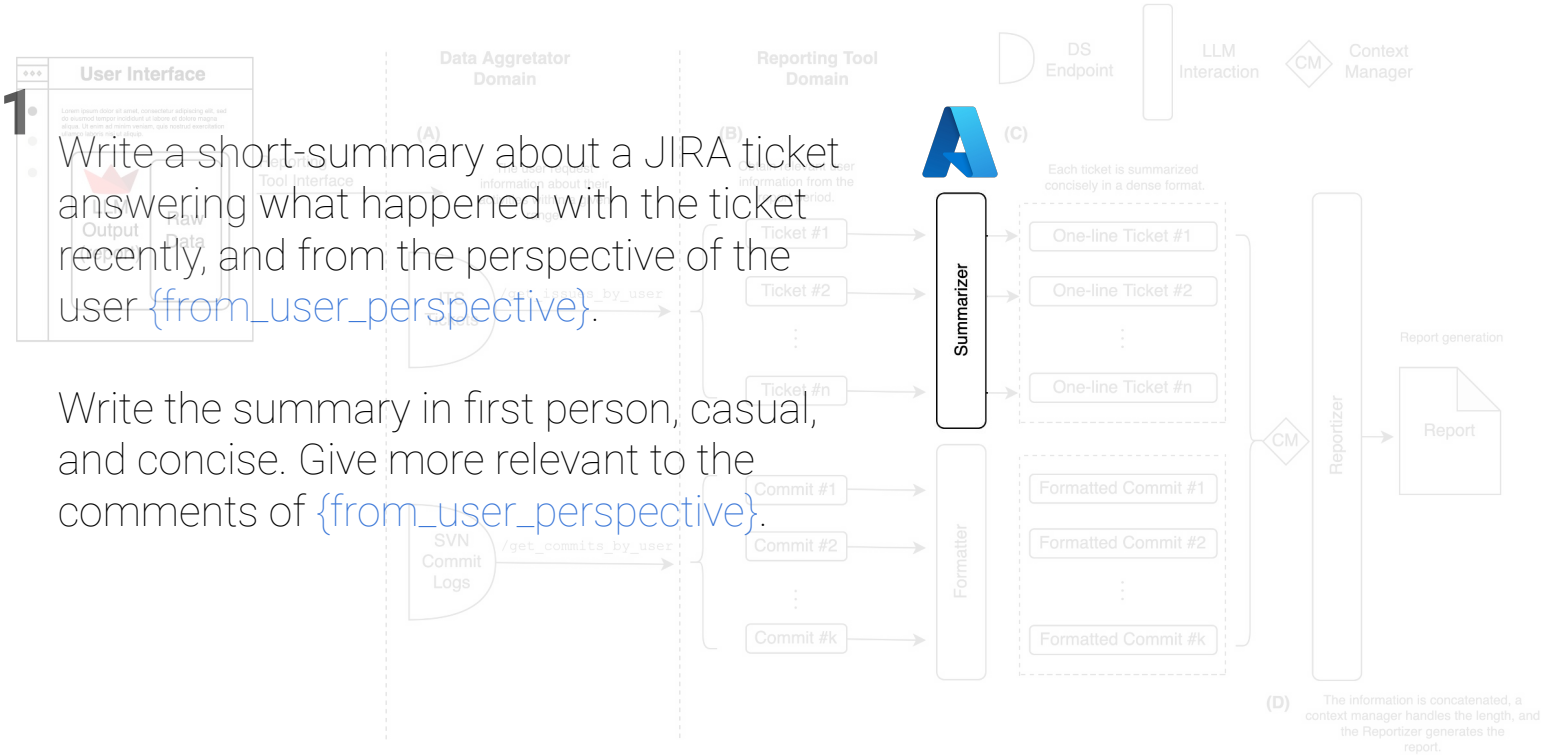
Now how the Summarize perform their task…summarize!



Now how the Summarize perform their task…summarize!

1. Write a short-summary about a JIRA ticket answering what happened with the ticket recently, and from the perspective of the user {from_user_perspective}.

   Write the summary in first person, casual, and concise. Give more relevant to the comments of {from_user_perspective}.

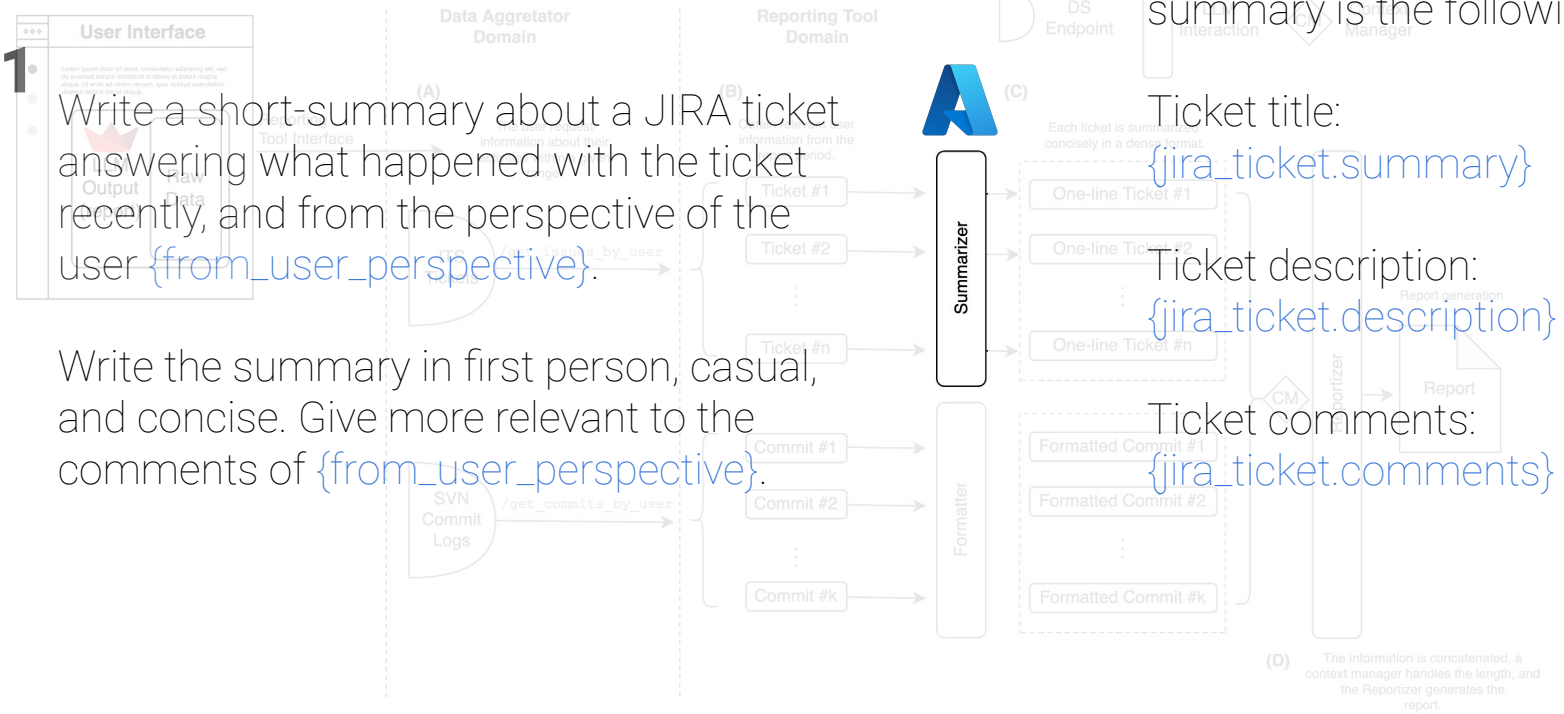


Now how the Summarize perform their task…summarize!

1

Write a short-summary about a JIRA ticket answering what happened with the ticket recently, and from the perspective of the user {from_user_perspective}.

Write the summary in first person, casual, and concise. Give more relevant to the comments of {from_user_perspective}.



2

The information used by the summary is the following:

Ticket title:
{jira_ticket.summary}

Ticket description:
{jira_ticket.description}

Ticket comments:
{jira_ticket.comments}

Now how the Summarize perform their task…summarize!

1

Write a short-summary about a JIRA ticket answering what happened with the ticket recently, and from the perspective of the user {from_user_perspective}.

Write the summary in first person, casual, and concise. Give more relevant to the comments of {from_user_perspective}.

Summarizer

2

The information used by the summary is the following:

Ticket title:
{jira_ticket.summary}

Ticket description:
{jira_ticket.description}

Ticket comments:
{jira_ticket.comments}

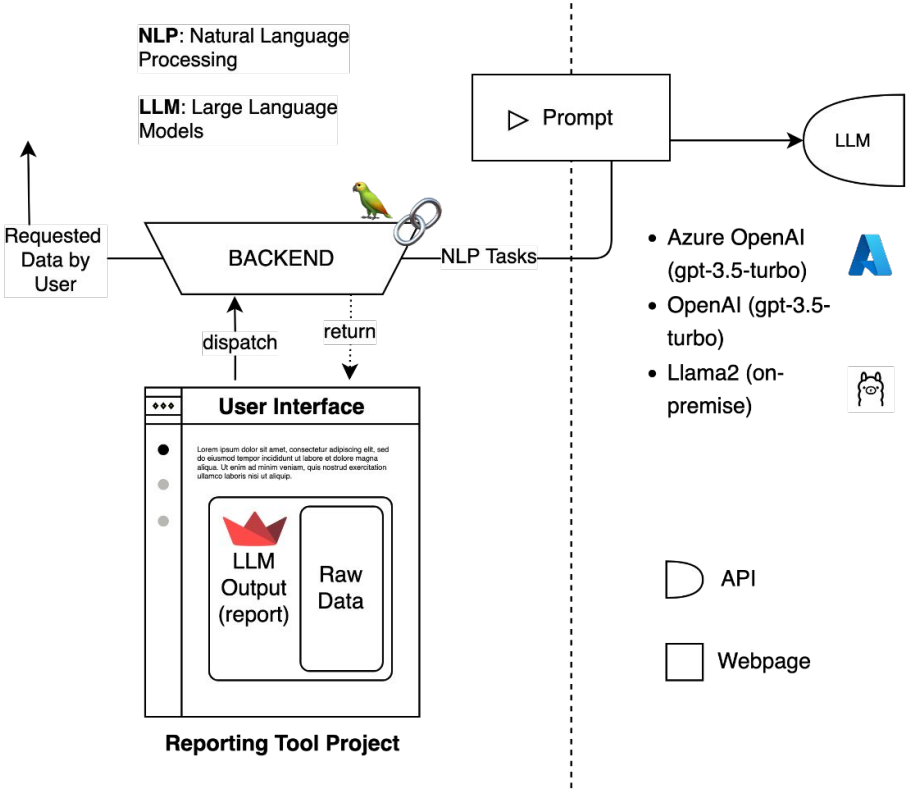
3

Remember to follow these rules:

- The summary must be concise.
- Avoid mentioning the {from_user_perspective} name in the summary (e.g use “I” instead of “I, jgil,”)

Reporting Tool Challenges

Measure and Evaluate Performance



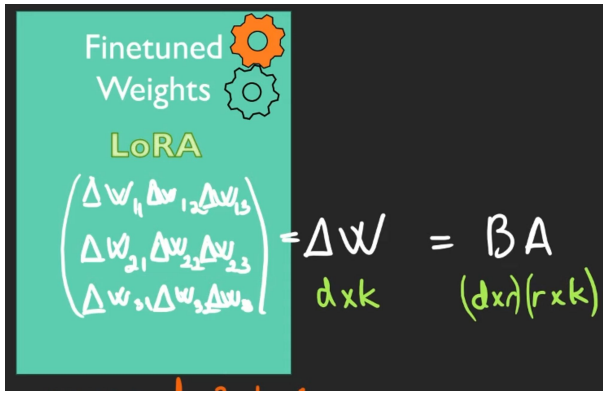
Measure Prompt - LLM Performance			
Prompt #1	0.71	0.82	0.52
Prompt #2	0.51	0.89	0.44
Prompt #3	0.84	0.82	0.71

DESIRABLE

Reporting Tool Challenges

Proprietary versus In-house models

- **Fine-tuning:** Adjusting the models parameters with their own data.
  a. Pros: privacy and more control about the model.
  b. Cons: requires infrastructure and demand dedicated expertise.
- **LoRA:** The idea is to represent the finetuned weight as a multiplication of two matrices that use a low rank decomposition. Only finetune a fraction of the model.



LoRA: Low-Rank Adaptation of Large Language Models (Hu et al. 2021)

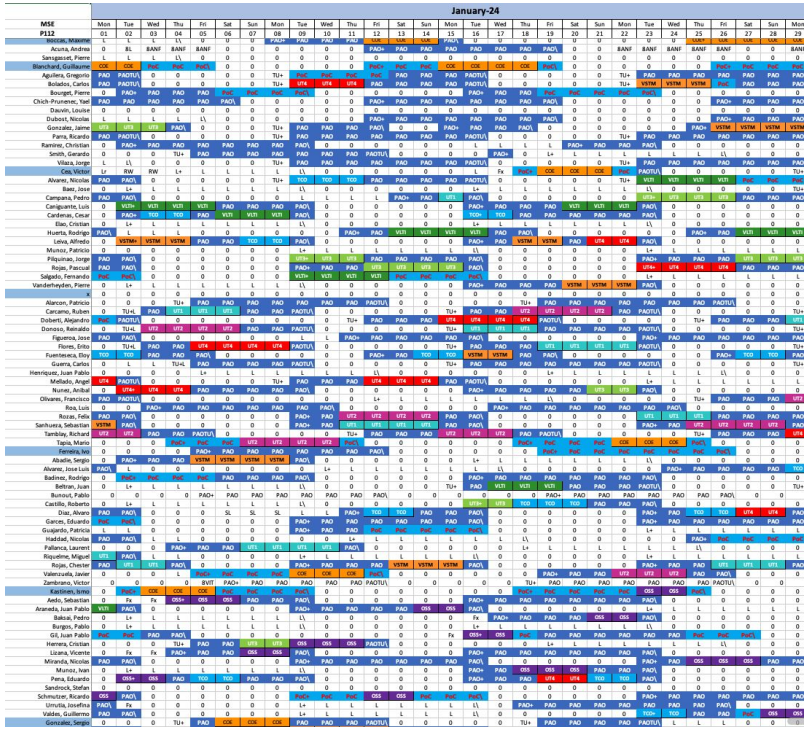
Other Projects

Non-LLM

MSE Ops Role Assignation

(MORA)

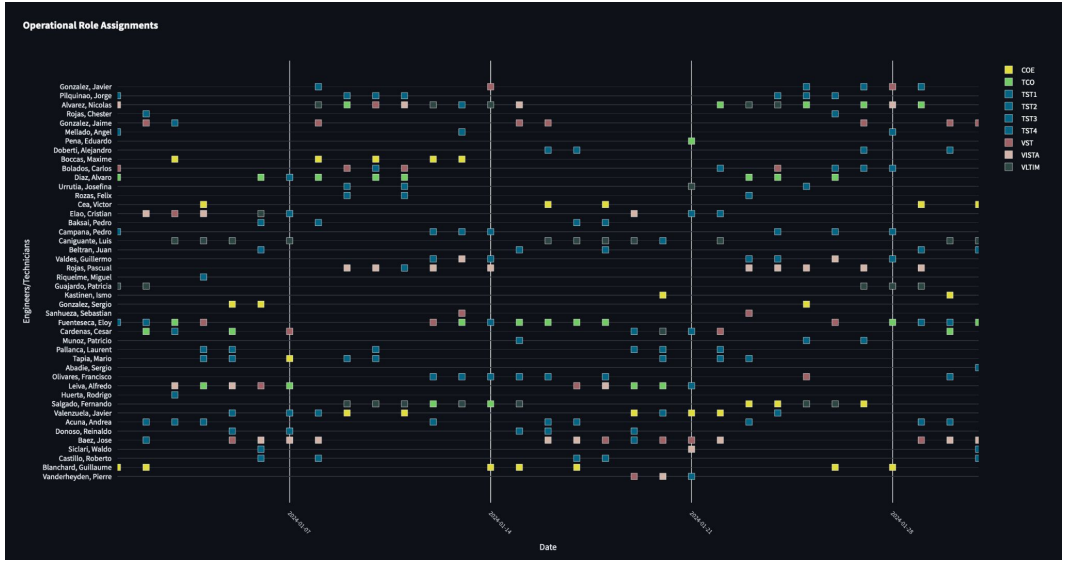
- **Rostering Problem:** Use a constraint programming solver (by ORTools) to solve the operational role assignment (i.e. TCO, VLTI).
  a. _set_one_engineer_per_role_assigned ✅
  b. _set_at_most_one_assignation_per_day ✅
  c. _set_availability_and_certification_constraints ✅
  d. _set_consecutive_days_constraint ⚠️
  e. _set_no_last_day_assignment_constraint ⚠️
  f. _set_mandatory_role_constraint ⚠️
  g. _set_balance_objective ⚠️
- **Format & Report results:** automate the report with the assignment results.

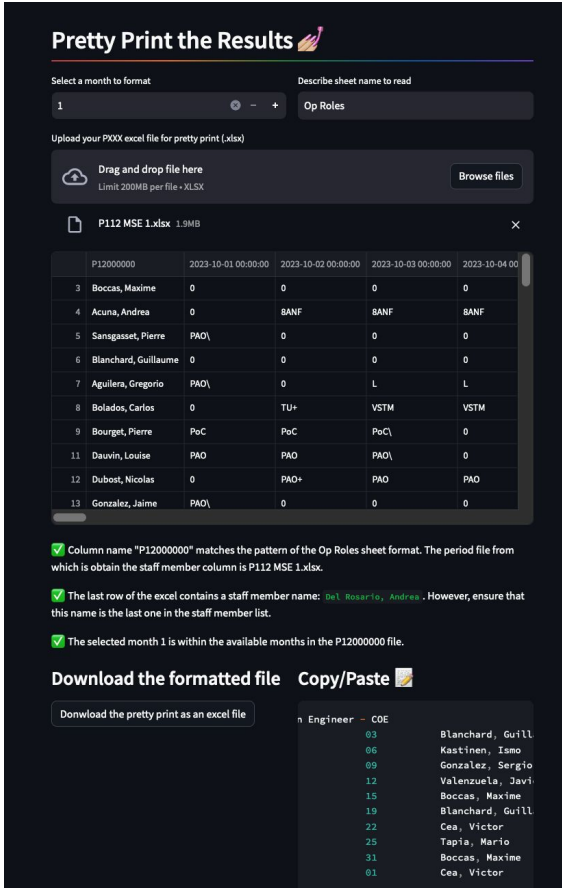


Source: 'Op Roles' sheet, from P112.xlsx

MSE Ops Role Assignation

User interface with Streamlit

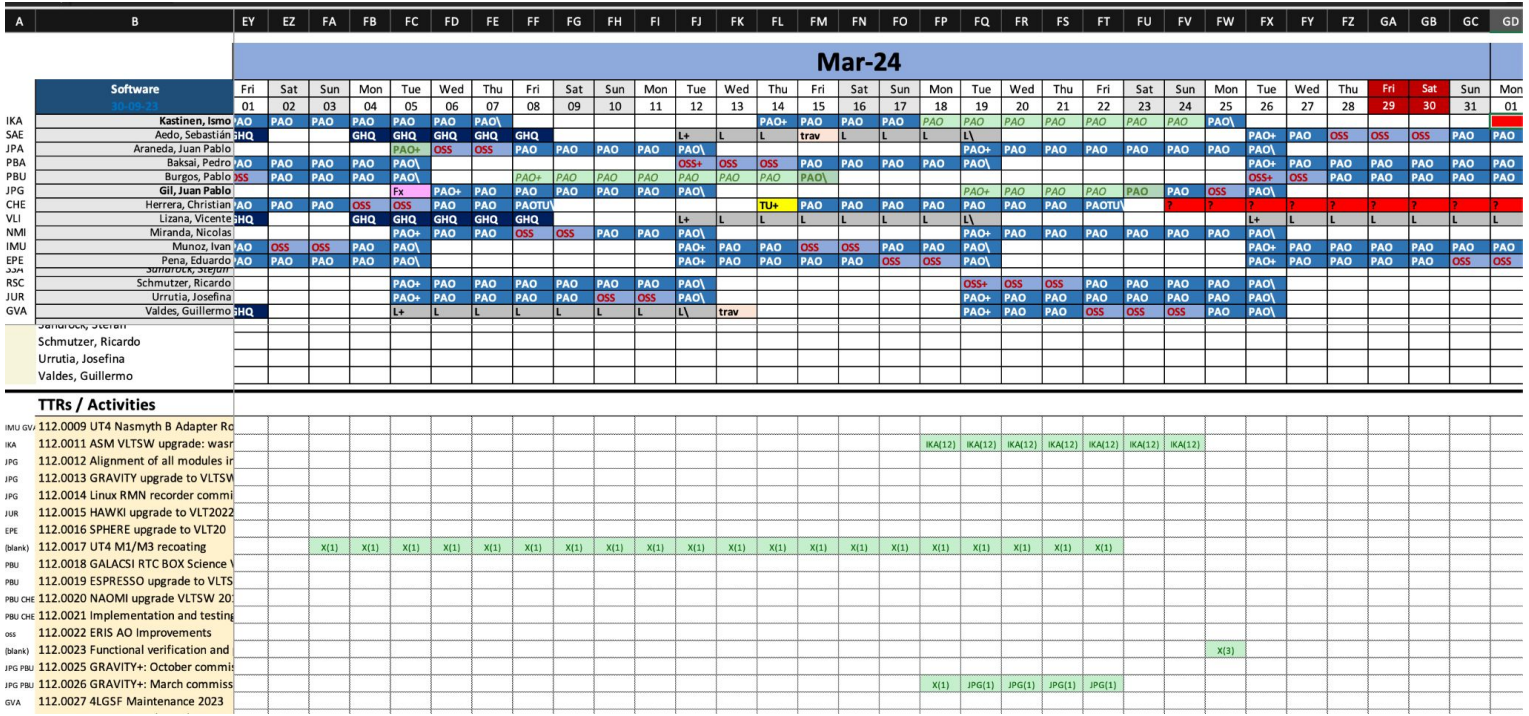




Source: https://gitlab.eso.org/calcazar/mse-planning

PSW Scheduling Tool

Collaborate and Test



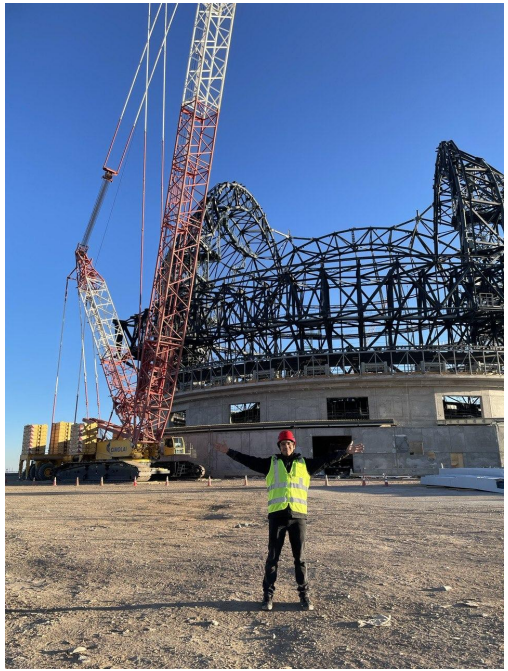
	Software	Fri	Sat	Sun	Mon	Tue	Wed	Thu	Fri	Sat	Sun	Mon	Tue	Wed	Thu	Fri	Sat	Sun	Mon	Tue	Wed	Thu	Fri	Sat	Sun	Mon	Tue	Wed	Thu	Fri	Sat	Sun	Mon
	30-09-23	01	02	03	04	05	06	07	08	09	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	01
		Mar-24																															
IKA	Kastinen, Ismo	AO	PAO	PAO	PAO	PAO	PAO	PAO\							PAO+	PAO	PAO	PAO	PAO	PAO	PAO	PAO	PAO	PAO	PAO	PAO\							
SAE	Aedo, Sebastián	HQ			GHQ	GHQ	GHQ	GHQ	GHQ				L+	L	L	trav	L	L	L	L\							PAO+	PAO	OSS	OSS	OSS	PAO	PAO
JPA	Araneda, Juan Pablo					PAO+	OSS	OSS	PAO	PAO	PAO	PAO	PAO\							PAO+	PAO	PAO	PAO	PAO	PAO	PAO	PAO\						
PBA	Baksai, Pedro	AO	PAO	PAO	PAO	PAO\							OSS+	OSS	OSS	PAO	PAO	PAO	PAO	PAO\							PAO+	PAO	PAO	PAO	PAO	PAO	PAO
PBU	Burgos, Pablo	SS	PAO	PAO	PAO	PAO\			PAO+	PAO	PAO	PAO	PAO	PAO	PAO	PAO\											OSS+	OSS	PAO	PAO	PAO	PAO	PAO
JPG	Gil, Juan Pablo					Fx	PAO+	PAO	PAO	PAO	PAO	PAO	PAO\							PAO+	PAO	PAO	PAO	PAO	PAO	OSS	PAO\						
CHE	Herrera, Christian	AO	PAO	PAO	OSS	OSS	PAO	PAO	PAOTU\						TU+	PAO	PAO	PAO	PAO	PAO	PAO	PAO	PAOTU\		?	?	?	?	?	?	?	?	?
VLI	Lizana, Vicente	HQ			GHQ	GHQ	GHQ	GHQ	GHQ				L+	L	L	L	L	L	L	L\							L+	L	L	L	L	L	L
NMI	Miranda, Nicolas					PAO+	PAO	PAO	OSS	OSS	PAO	PAO	PAO\							PAO+	PAO	PAO	PAO	PAO	PAO	PAO	PAO\						
IMU	Munoz, Ivan	AO	OSS	OSS	PAO	PAO\							PAO+	PAO	PAO	OSS	OSS	PAO	PAO	PAO\							PAO+	PAO	PAO	PAO	PAO	PAO	PAO
EPE	Pena, Eduardo	AO	PAO	PAO	PAO	PAO\							PAO+	PAO	PAO	PAO	PAO	OSS	OSS	PAO\							PAO+	PAO	PAO	PAO	PAO	OSS	OSS
SJA	~~Sandrock, Stefan~~																																
RSC	Schmutzer, Ricardo					PAO+	PAO	PAO	PAO	PAO	PAO	PAO	PAO\							OSS+	OSS	OSS	PAO	PAO	PAO	PAO	PAO\						
JUR	Urrutia, Josefina					PAO+	PAO	PAO	PAO	PAO	OSS	OSS	PAO\							PAO+	PAO	PAO	PAO	PAO	PAO	PAO	PAO\						
GVA	Valdes, Guillermo	HQ				L+	L	L	L	L	L	L	L\	trav						L+	PAO	PAO	OSS	OSS	OSS	PAO	PAO\						
	~~Sandrock, Stefan~~																																
	Schmutzer, Ricardo																																
	Urrutia, Josefina																																
	Valdes, Guillermo																																
	TTRs / Activities																																
IMU GVA	112.0009 UT4 Nasmyth B Adapter Ro																																
IKA	112.0011 ASM VLTSW upgrade: wasr																		IKA(12)	IKA(12)	IKA(12)	IKA(12)	IKA(12)	IKA(12)	IKA(12)								
JPG	112.0012 Alignment of all modules ir																																
JPG	112.0013 GRAVITY upgrade to VLTSW																																
JPG	112.0014 Linux RMN recorder commi																																
JUR	112.0015 HAWKI upgrade to VLT2022																																
EPE	112.0016 SPHERE upgrade to VLT20																																
(blank)	112.0017 UT4 M1/M3 recoating			X(1)	X(1)	X(1)	X(1)	X(1)	X(1)	X(1)	X(1)	X(1)	X(1)	X(1)	X(1)	X(1)	X(1)	X(1)	X(1)	X(1)	X(1)	X(1)	X(1)										
PBU	112.0018 GALACSI RTC BOX Science V																																
PBU	112.0019 ESPRESSO upgrade to VLTS																																
PBU CHE	112.0020 NAOMI upgrade VLTSW 20																																
PBU CHE	112.0021 Implementation and testing																																
oss	112.0022 ERIS AO Improvements																																
(blank)	112.0023 Functional verification and																									X(3)							
JPG PBU	112.0025 GRAVITY+: October commis																																
JPG PBU	112.0026 GRAVITY+: March commiss																		X(1)	JPG(1)	JPG(1)	JPG(1)	JPG(1)										
GVA	112.0027 4LGSF Maintenance 2023																																

Internship Memoirs

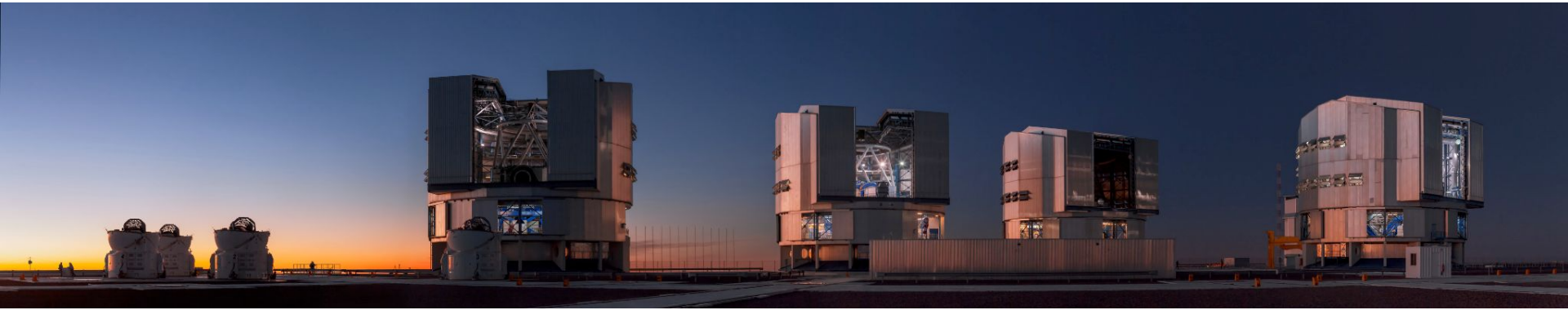
- Friends and camaraderie! 💚
- The sunset at the platform 🌅!
- Volleyball 🏐!
- Mind 🧘 & body 🏋️
- The desert 🌵 and dessert 🍨!
- The nostalgic short dose when you arrive at the airport 🛬
- Thanks to PSW! and JPG for being excellent advisors!
- And everyone I know and shared moments with, there was a pleasure!

An excellent and lovely group of people

Practicing jump serve with Nico!


Me visiting the ELT (thanks Matteo!)

Thank you!



Engineering with prompts, rather than prompt engineering

Cristóbal Alcázar
MSc in Finance & (not yet) MSc in Data Science